Argus: Multi-View Egocentric Human Mesh Reconstruction Based on Stripped-Down Wearable mmWave Add-on

Di Duan[1*], Shengzhe Lyu[2*], Mu Yuan[1], Hongfei Xue[3], Tianxing Li[4], Weitao Xu[2], Kaishun Wu[5], Guoliang Xing[1†]

[1]The Chinese University of Hong Kong, [2]City University of Hong Kong, [3]University of North Carolina at Charlotte, [4]Michigan State University, [5]Hong Kong University of Science and Technology (GZ)

ABSTRACT

In this paper, we propose *Argus*, a wearable add-on system based on stripped-down (*i.e.*, compact, lightweight, low-power, limited-capability) mmWave radars. It is the first to achieve egocentric human mesh reconstruction in a multi-view manner. Compared with conventional frontal-view mmWave sensing solutions, it addresses several pain points, such as restricted sensing range, occlusion, and the multipath effect caused by surroundings. To overcome the limited capabilities of the stripped-down mmWave radars (with only one transmit antenna and three receive antennas), we tackle three main challenges and propose a holistic solution, including tailored hardware design, sophisticated signal processing, and a deep neural network optimized for high-dimensional complex point clouds. Extensive evaluation shows that *Argus* achieves performance comparable to traditional solutions based on high-capability mmWave radars, with an average vertex error of 6.5 cm, solely using stripped-down radars deployed in a multi-view configuration. It presents robustness and practicality across conditions, such as with unseen users and different host devices.

CCS CONCEPTS

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

KEYWORDS

Wireless Sensing, Wearable Sensing, Millimeter Wave, Human Mesh Reconstruction, Virtual Reality

1 INTRODUCTION

Human pose estimation (HPE), including skeleton tracking and 3D human mesh reconstruction (HMR), remains a perennial research topic due to its broad application value in tasks such as fitness coaching [60], health monitoring [35], and virtual reality [78]. Consequently, it has received significant attention from researchers [55]. Existing solutions attempted to solve this problem use various modalities, such as visual modalities [6, 21, 28–30] (*e.g.*, RGB, depth), wearable modalities [16, 42, 72] (*e.g.*, IMU, electromyography), and wireless modalities [27, 32, 40, 49, 51, 68, 69, 76] (*e.g.*, ultrasound, Wi-Fi, mmWave). However, vision-based solutions are highly dependent on light conditions and struggle to darkness or smoke; while wearable-based solutions suffer from cumbersomeness and lack of user-friendliness. Based on this, many wireless solutions for HMR focus on using wireless signals. Among them, human sensing and reconstruction based on mmWave [74] is a representative research direction because it offers high precision, better penetration compared to Wi-Fi, and better interference resistance compared to low-frequency ultrasound. Based on this fact, it has garnered significant attention and led to many representative studies [11, 32, 66–68, 70, 73, 75] in the field.

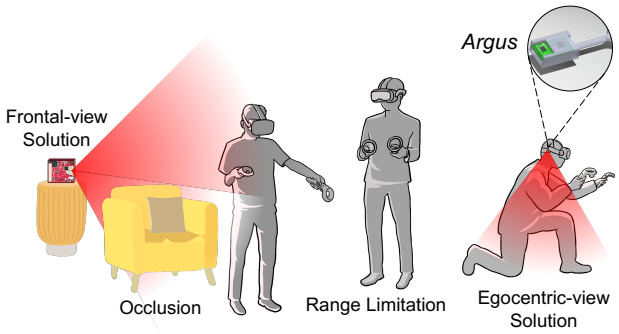


Fig. 1: *Argus* **is the first multi-view, egocentric mmWave sensing system enabling continuous HMR, breaking through the limitations of frontal-view solutions.**

Radio frequency (RF) signals are renowned for their non-contact, imperceptible, and user-friendly characteristics, making RF-based human sensing a subject of significant interest and leading to numerous practical applications [2, 15, 22, 48, 52, 64, 65]. Previously, researchers successfully used Wi-Fi for human skeleton tracking [77] and mesh reconstruction [76]. However, because of the ubiquity of Wi-Fi signals, they are prone to interference, and the sensing granularity is coarse. As a result, researchers have recently shifted towards using mmWave for human pose estimation and reconstruction, leading to the development of a series of studies. Xue *et al.* proposed mmMesh [68], which is the first human mesh reconstruction solution based on a commodity mmWave radar. Later, several follow-up studies have been proposed to address remaining challenges or to improve the solution from different perspectives. For example, SynMotion [75] and mmGPE [67] were later proposed to improve generalization by synthesizing mmWave signals; M^{4}esh [66] and m^{3}Track [31] were proposed for multi-target tracking and reconstruction; several mmWave-native studies try to improve HPE performance by introducing an additional mmWave radar [32, 73] or employing advanced deep learning methods [70]. However, all of the above works focus solely on mmWave-based

* Co-primary authors.
† Guoliang Xing is the corresponding author.

SenSys '25, May 6–9, 2025, Irvine, CA, USA

ACM ISBN 979-8-4007-1479-5/25/05
https://doi.org/10.1145/3715014.3722045

Table 1: Comparison with other solutions (❍–Not Available, ●–Available; Cons.: Consumption, TX #: Number of Transmitters, RX #: Number of Receivers, Weight: Radar Weight).

Solutions	Radar	TX #	RX #	Board Size	Weight	Power Cons.	Multi-View	Sensing View	Body Part
mmMesh [68]	AWR1843	3	4	8.3 cm × 6.4 cm	~245 g	~2.08 W	❍	Frontal	Full
SynMotion [75]	IWR1443	3	4	8.3 cm × 6.4 cm	~245 g	~2.1 W	❍	Frontal	Full
mmGPE [67]	AWR1843	3	4	8.3 cm × 6.4 cm	~245 g	~2.08 W	❍	Frontal	Full
M^{4}esh [66]	AWR1843	3	4	8.3 cm × 6.4 cm	~245 g	~2.08 W	❍	Frontal	Full
m^{3}Track [31]	AWR1443	3	4	7.8 cm × 6.4 cm	~245 g	~2.1 W	❍	Frontal	Full
HUPR [32]	IWR1843	3	4	7.8 cm × 6.4 cm	~245 g	~2.08 W	❍	Frontal	Full
SUPER [73]	IWR6843	3	4	6.8 cm × 5.5 cm	~137 g	~1.75 W	❍	Frontal	Upper
mmEgo [33]	IWR6843	3	4	6.8 cm × 5.5 cm	~137 g	~1.75 W	❍	Egocentric	Full
Ours	BGT60TR13C	1	3	3.9 cm × 2.4 cm	~7.5 g × 2	~0.35 W × 2	●	Egocentric	Full

HPE/HMR from a frontal view perspective, neglecting an intriguing perspective—*mmWave-based HMR from an egocentric view.*

This interesting sensing view has several advantages over the frontal-view solutions. First, frontal-view solutions are subject to various limitations, such as sensing range and susceptibility to interference from others' movements (*i.e.*, multipath effect) or occlusion. In contrast, using mmWave signals to sense a user from an egocentric perspective is highly promising to circumvent these challenges. Due to its on-body setup, the mmWave sensing field moves with the target, and the controllable sensing range can effectively avoid interference from others. Recently, the discovery of this ingenious research perspective brought about the first mmWave-based HMR solution from an egocentric view called mmEgo [33]. However, using such an egocentric view and mmWave signals to sense the user's human body typically implies a solution lies at the intersection of wearable and wireless sensing. Therefore, *how to elegantly combine the characteristics of both remains a significant challenge.*

Although the mmWave radar (*i.e.*, IWR6843ISK-ODS [25]) used in mmEgo offers excellent sensing performance due to its multiple transmit and receive antennas, its size, weight, and high power consumption make it unsuitable for wearable solutions, rendering mmEgo far from practical. Motivated by this, we propose our solution—a pair of compact, lightweight, low-power mmWave-based add-ons named *Argus* that can be magnetically attached to various common host devices, such as VR headsets and headphones.

However, realizing this idea presents numerous unique challenges: **(1) How can multiple factors be fully considered in hardware design?** All previous studies are based on well-developed Texas Instruments (TI) mmWave radars, such as the IWR6843, IWR1843, IWR1443 series. These radars come with comprehensive development kits, including the mmWave SDK and mmWave Studio, which provide high-precision point cloud data, essential for advanced applications. However, these radars are relatively large in size and energy consumption, making them unsuitable for wearable devices that offer mmWave sensing in egocentric view. The development of *Argus* adopts compact sensors that help reduce the size of the hardware by significantly compromising the radar's capability, such as having fewer transmit and receive antennas. In addition, these compact radars often lack the sophisticated development kits needed to achieve high-precision point clouds, posing a major challenge of this paper. **(2) How to obtain the ground truth label in a more practical and user-friendly manner?** Previous solutions all rely on expensive and high-precision Motion Capture (MoCap) systems (*e.g.*, VICON [59], OptiTrack [44], Azure Kinect [41]) to provide high-quality labels for model training. However, such expensive MoCap systems are not commonly found in personal applications, and their deployment is often constrained by the available space, making mmWave-based HMR applications less widespread and challenging to implement practically. Therefore, acquiring high-quality labels for training using cost-effective commodity devices is a significant challenge in enhancing the practicality of the system. **(3) How does *Argus* overcome the dual challenges of self-occlusion and specular reflection?** Li *et al.* has already mentioned the issue of self-occlusion of the lower body by the upper body in egocentric views [33]. However, they used a tricky approach by deploying the mmWave radar extended far in front of the user's head. Although this approach alleviates the self-occlusion problem and avoids specular reflections from the shoulders, it makes the system cumbersome and unsuitable for common small devices such as headphones. Therefore, effective integration of form factor design, signal processing pipeline, and deep neural network is essential to effectively address this intractable challenge and provide a better user experience.

To address these challenges, we proposed a series of solutions: **(1) To holistically consider the multiple factors inherent in hardware design**, we propose an innovative solution (Fig. 1) that considers multiple factors by employing a pair of compact mmWave radars (*i.e.*, BGT60TR13C [56]) as an add-on, magnetically attached to a host device. *Argus* is the first portable mmWave sensing system for egocentric-view HMR that analyzes multi-view mmWave data (*i.e.*, left and right), featuring a small size, lightweight, and low power consumption (details shown in Table 1). Furthermore, we have overcome the limitations imposed by the stripped-down mmWave radar through advanced signal processing techniques and deep learning. **(2) To facilitate the widespread application of *Argus***, we employ monocular-based human mesh estimation using a single RGB-only camera (*e.g.*, web camera, front camera of a smartphone) to acquire training labels, instead of relying on cumbersome and expensive MoCap systems. **(3) To overcome the challenges of self-occlusion and specular reflection**, we leverage multi-view sensing, namely, using dual egocentric-view mmWave sensing fields from left ear and right ear to collaboratively construct body meshes for the first time. Furthermore, we propose a tailored range-gating and energy-compensation approach for *Argus*.

Moreover, Kolmogorov–Arnold Networks (KAN) [36] is introduced to improve learning efficiency due to its superior capability in handling non-linearities, which are essential for modeling complex high-dimensional relationships (*e.g.*, multi-view egocentric HMR). The contributions of this paper can be summarized as follows:

- To the best of our knowledge, *Argus* is the first system realize the multi-view egocentric HMR by proposing a holistic solution, including prototype design, FMCW signal design and processing, and deep learning, etc.
- We first achieve multi-view mmWave sensing based on our wearable prototype. *Argus* is based on the joint analysis of mmWave fields with complementary fields of view, effectively addresses the self-occlusion and shoulder specular reflection issues.
- We propose and adopt a series of tailored techniques, such as clutter removal, range gating, energy compensation, and the introduction of KAN to enhance the system's learning ability for high-dimensional non-linear representations. These techniques make it possible to reconstruct human meshes using compact and limited-capability radars.
- We perform a comprehensive evaluation of *Argus*, including its performance on unseen users and several micro-benchmarks. The evaluation results show that *Argus* outperforms two state-of-the-art (SOTA) baselines, demonstrating both robustness and practicality.

2 RELATED WORK

mmWave-Based HPE and HMR. Driven by the contactless nature and high precision of mmWave sensing, HPE and HMR based on mmWave have received significant attention in recent years. Xue *et al.* proposed mmMesh [68], using a commodity mmWave radar (*i.e.*, AWR1843) for frontal-view HMR. However, there are still serveral challenges in this research area, such as the generalization to unseen activities, multi-target effect. Motivated by this, Xue *et al.* proposed their upgraded solutions M^{4}esh [66] and mmGPE [67], which solve the problem of reconstructing multiple human meshes simultaneously and the generalization problem for unseen activities, respectively. To address the same problems in a different way, Zhang *et al.* proposed SynMotion [75] which synthesizes mmWave sensing signals to construct a mmWave dataset for generalization; m^{3}Track [31] is a contemporaneous work with M^{4}esh, it transforms the multi-target tracking task into a single-target HPE task. Furthermore, numerous studies have been proposed to improve performance by introducing multi-modality [11], extra mmWave radar from another sensing view (*i.e.*, vertical and horizontal) [32, 73], or more advanced deep learning techniques [61, 70].

However, it seems that all the aforementioned works fall into a cognitive bias: *assuming that the mmWave sensing for HPE or HMR must be performed from a frontal view.* In fact, not all studies have ignored egocentric-view sensing; mmEgo [33] was a pioneer in this idea. It prototyped a conventional mmWave radar (*i.e.*, IWR6843) into a wearable device and made initial explorations of egocentric mmWave sensing. However, because of its size, weight, and power consumption, the design of mmEgo as a portable device was not usable, and the single-view sensing solution proved inadequate in addressing the challenges of self-occlusion and specular reflection.

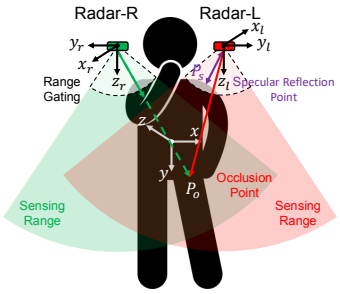


Fig. 2: Illustration of different coordinate systems and the theoretical model of *Argus*.

To fill this research gap, we propose *Argus*, the first multi-view sensing solution based on stripped-down mmWave radars. By introducing several signal processing and deep learning techniques, it achieves acceptable performance despite significant hardware limitations, thereby improving the usability of egocentric HMR.

Earable Sensing. Recently, mobile sensing based on earable devices has garnered widespread attention from researchers due to portability and informative sensing positions, which serve as a platform for multiple modalities, such as speech, ultrasound, electromyography (EMG), and photoplethysmogram (PPG). Numerous applications based on earable devices have been explored, such as speech enhancement [10, 17, 58], behavior recognition [26, 38, 39, 46], health monitoring [4, 5, 8, 9, 12, 24], and user authentication [19, 62, 63]. In addition, there have been security studies based on earable devices, covering both attack [34] and defense [18]. A recent study, TinyssimoRadar [50], successfully integrated mmWave modality with earable devices, achieving impressive performance in a simple gesture recognition task. However, the sensing capabilities of mmWave extend far beyond this. To the best of our knowledge, there has been no exploration of full-body pose estimation or mesh reconstruction using earable devices equipped with stripped-down mmWave radars. *Argus* is the first to tackle this challenging task under hardware limitations, filling this research gap.

3 MODELING OF MULTI-VIEW SENSING

Before we delve into the detailed system design, we first model the complementary multi-view sensing of the proposed system to better elucidate our core idea and the rationale behind the hardware configuration.

As Fig. 2 shows, we consider two representative sensing positions, P_o (occlusion) and P_s (specular reflection), in the pose estimation task when using two downward-directed mmWave radars with different fields of view (FOVs). We define the set of features extracted from the two radars without occlusion at time t as $F_{R_r}(t)$ and $F_{R_l}(t)$, respectively. When line-of-sight occlusion occurs, such as the raised right arm, it will affect the FOV of the right-side radar, and further result in the loss of informative features (F_o) and inferior estimation at P_o, which can be formulated as follows:

$$\theta_s(t) = \mathcal{M}(F'_{R_r}(t)) = \mathcal{M}(F_{R_r}(t) \setminus F_o),$$

where $\mathcal{M}$ represents the mapping function from extracted mmWave features to joint rotations θ_s (single-view). To address the issue of information loss, the features extracted from the other mmWave radar

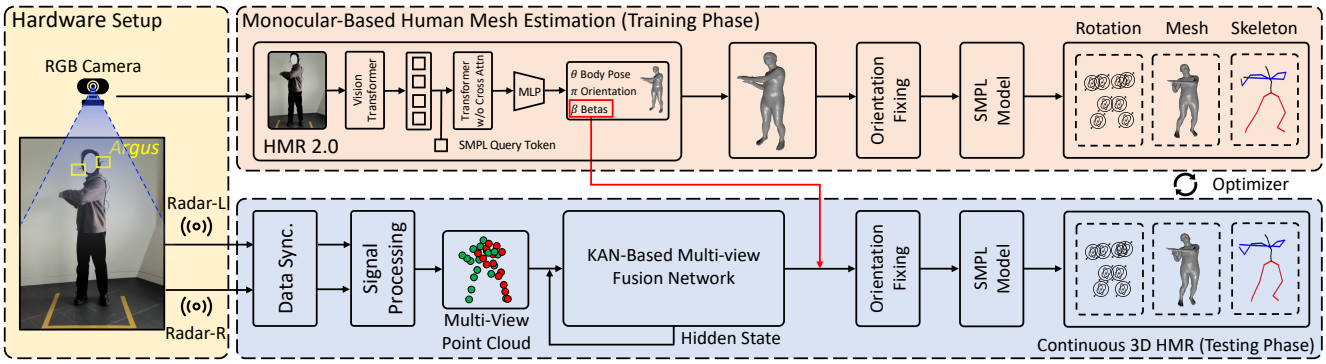


Fig. 3: System overview. By transferring the knowledge from RGB images as pseudo-labels, a well-trained model will be deployed in the testing phase to perform continuous 3D human mesh reconstruction.

(*i.e.*, $F_{R_l}(t)$) can compensate for missing F_o, thus improving the robustness and performance of the estimation. Furthermore, to align the spatial semantic information from radars in different coordinate systems (illustrated in Fig. 2), we perform coordinate transformation on the positional features extracted from both mmWave radars; the converted features are aligned to the wearer's body coordinate system before combination. The enhanced multi-view estimation can be be expressed as:

$$\begin{aligned}\theta_m(t) &= \mathcal{M}(F_{combined}(t)) \\ &= \mathcal{M}(\mathcal{T}_r(F_{R_r}(t) \setminus F_o) \cup \mathcal{T}_l(F_{R_l}(t))),\end{aligned}$$

where $F_{combined}$ denotes combined features, $\mathcal{T}_r$ and $\mathcal{T}_l$ represent the rotation operations for coordinate transformation.

Given that mmWave sensing usually uses selected point clouds to reduce computational overhead and delay, the accumulation of high-energy points caused by specular reflection (*e.g.*, P_s on a shoulder) will eclipse the informative features caused by the torso. Hence, to force the extracted features to focus on the wearer's torso, we need to specially design the FOV of the radars and propose an ad hoc range-gating approach. Theoretically, it can be formulated as follows:

$$F_{rg,R_i}(t) = \text{Sel}(\mathcal{T}_i(F_{R_i}(t) \setminus F_{s_i})), \quad i \in \{r, l\},$$

where Sel denotes the operation of selecting informative features, F_{s_i} indicates the specular reflection features, $F_{rg,R_i}(t)$ represents the refined features by using range gating; the overall theoretical model of *Argus* can be formulated as:

$$\theta_m(t) = \mathcal{M}(\mathcal{T}_r(F_{rg,R_r}(t) \setminus F_o) \cup \mathcal{T}_l(F_{rg,R_l}(t))).$$

Overall, we have conducted a clear theoretical modeling of the complementary multi-view sensing rationale of *Argus* in representative cases (*i.e.*, occlusion and specular reflection).

4 SYSTEM OVERVIEW

The system overview of *Argus* is shown in Fig. 3, which contains two phases: a training phase in which the effective pseudo-labels obtained from RGB frames serve as the supervision to develop a KAN-based multi-view fusion network, and a testing phase where the well-trained deep neural network can continuously reconstruct multi-view mmWave frames into SMPL [45] parameters.

Specifically, in the training phase, we collect visual stream data (RGB images) and mmWave stream data (data frames) using a commodity RGB camera and self-designed wearable add-ons (details in Sec. 5.1), respectively. Then, we perform cross-modality data alignment using a two-tier method, which is elaborated in Sec. 5.3. Subsequently, the RGB frames are fed into the HMR 2.0 [21] neural network to estimate the SMPL parameters of these frames. Meanwhile, the corresponding mmWave data frames are processed by the proposed signal processing module (Sec. 5.4) and further translated into the predicted SMPL parameters by the deep neural network (Sec. 5.5). Since *Argus* is designed for on-body usage, an orientation-fixing module sets the matrix that represents the global orientation to the 3×3 identity matrix. Finally, the monocular-estimated and mmWave-predicted SMPL parameters are rendered into outputs (*i.e.*, joint rotations, body meshes, skeletons). During training, the losses between these outputs are optimized together.

In the testing phase, the well-trained deep neural network will translate the processed multi-view mmWave features into SMPL parameters. The parameters, arranged as a time series, will be rendered into continuous body meshes or skeletons to support a corpus of applications such as e-fitness yoga instructor, avatar in the meta-universe, and virtual videoconferencing.

5 SYSTEM DESIGN

5.1 mmWave Stream & Hardware Design

As mentioned in Sec. 1, the hardware design of *Argus* should consider multiple factors in a comprehensive way, which causes specific difficulties that can be abstracted as follows:

- The hardware prototype needs to support stable sensor configuration and data acquisition while maintaining a lightweight form factor as an add-on for common host devices, such as VR headsets and headphones.
- To facilitate effective sensor synchronization and efficient data transmission, a dedicated intermediary should be introduced between a mobile device and the two radars.
- It should be taken into account that user-friendliness and orientation-invariance of the device play a vital role in the performance and usability of the system.

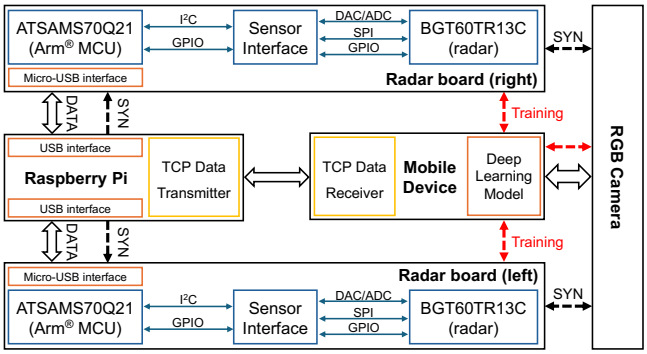


Fig. 4: Block diagram of *Argus* hardware.

Taking into account the factors mentioned above, we designed the hardware shown in Fig. 4 to implement our prototype. It consists of four main components: two radar boards for mmWave data acquisition, a Raspberry Pi for sensor synchronization and data transmission, and a mobile device for data receiving, processing, and rendering output results. Additionally, an RGB camera is utilized only during the training phase to obtain pseudo-labels for training. We now break down each of the components in detail.

Radar boards. The radar board is designed to configure and transfer data from the mmWave sensor. It includes a baseboard with a Microchip ATSAMS70Q21 32-bit Arm Cortex-M7 MCU, and a daughter board hosting a BGT60TR13C [56] 60 GHz mmWave radar sensor. The baseboard features two interfaces: a high-speed USB 2.0 connection for Raspberry Pi communication and a Serial Peripheral Interface (SPI) for mmWave sensor data transmission. It also integrates circuits for power management and debugging and the daughter board supports the radar chipset. The printed circuit board (PCB) measures $17 \times 12.7\,\text{mm}^2$, with the mmWave sensor's Antenna-in-Package (AIP) measuring $6.5 \times 5.0 \times 0.85\,\text{mm}^3$. The setup of the radar chip consists of one transmit antenna and three receive antennas arranged in an L-shape. The receiving antennas are grouped into two pairs, while the antennas are separated by half the wavelength (*i.e.*, $\lambda/2$) in each pair. Low-pass filters are employed to minimize the impact of noise and crosstalk on the supply domains. Furthermore, an EEPROM connected via an I2C interface stores data such as the board identifier, while an 80 MHz quartz oscillator ensures accurate timing for operations.

Raspberry Pi. The synchronization and transmission center, built on a Raspberry Pi 4 with a Linux OS running specified radar Software Development Kit (SDK), acts as an intermediary between a mobile device and two radar boards. Powered by a power bank, the Raspberry Pi manages the synchronization, maintaining a time offset between the same frame from the left and right radars below 10 ms. It connects to the radar boards via USB interfaces to transmit configuration parameters and receive data. Wireless connectivity is facilitated by the built-in Wi-Fi module of the Raspberry Pi, allowing real-time communication with the mobile device. The multi-view mmWave data is transferred to the mobile device for computing in real time via a TCP protocol, ensuring the reliability and integrity of data transfer.

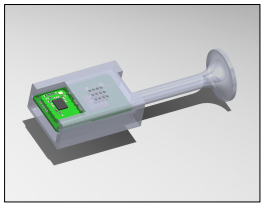


Fig. 5: *Argus* 3D.

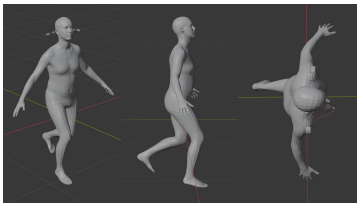


Fig. 6: 3D sensing model.

Mechanical Design and System Integration. The two radar boards are housed within 3D-printed enclosures (Fig. 5), which are part of the integral *Argus* device designed to function as an add-on for general host devices (*e.g.*, headphones). These transparent enclosures are custom-made for the radars of *Argus*, with one side featuring an open window for the emission of mmWave signals. Furthermore, the enclosures are designed to magnetically attach to host devices. As Fig. 15(d) shows, by carefully designing the pole arrangement of the four pairs of magnets on each side, it ensures that *Argus* add-ons are attached precisely to the same position each time and that left and right can be strictly distinguished. When the radar is placed on the wrong side, the pole arrangement creates repulsion rather than attraction. We modeled the sensing scenario in Blender [13], as shown in Fig. 6, where the add-ons are fixed to both sides of the user's head, with the mmWave sensors positioned approximately 8 cm horizontally from the user's ears. Each mmWave sensor has a maximum FOV of 90° and a maximum sensing distance of 3.2 m. From the multiple perspectives of the simulated scenario shown in Fig. 6, it is easy to see that the sensing range of *Argus* can effectively cover and detect the movements of the user's limbs.

In this section, we propose a holistic hardware solution that considers multiple factors such as stability, synchronization, and usability, addressing **the first key challenge** mentioned in Sec. 1.

5.2 Visual Stream & Pseudo-Label

Motivated by the high deployment overhead and the space limitation of cumbersome MoCap systems used in existing solutions [66, 68, 75], we pioneeringly propose a more practical approach to obtain pseudo-labels for training—*estimating pose parameters from monocular images.*

In practice, we implement HMR 2.0, a cutting-edge human mesh estimation network based on monocular images, which includes a Vision Transformer (ViT) and a transformer decoder, on a commodity RGB camera. The ViT dissects the image into patches and processes these through self-attention mechanisms, allowing the model to capture global dependencies and intricate details across the entire image; while the cross-attention-based transformer decoder further refines the process by selectively focusing on relevant features extracted by the ViT. It dynamically adjusts its attention to specific image regions that are more informative in predicting human meshes, effectively dealing with occlusions and complex poses. By incorporating the ViT and transformer decoder, the network is designed to efficiently parse and understand the complexities of human poses and shapes from a monocular RGB image. The implementation of HMR 2.0 follows the open-source repository [20]. However, to effectively use the pseudo-labels, we need to address

Algorithm 1 Cross-Modality Data Alignment

Require: N: Number of frames, τ: Threshold
Ensure: Data alignment within τ ms
$t_{I_m}, t_{L_m} \leftarrow$ Query NTP, Record local time
$O_m \leftarrow t_{I_m} - t_{L_m}$ ▷ mmWave offset
$t_{I_i}, t_{L_i} \leftarrow$ Query NTP, Record local time
$O_i \leftarrow t_{I_i} - t_{L_i}$ ▷ RGB offset
for $f = 1$ to N **do**
  $t'_{I_m f} \leftarrow t_{L_m f} + O_m$ ▷ Calibrate mmWave modality to NTP
  $t'_{I_i f} \leftarrow t_{L_i f} + O_i$ ▷ Calibrate visual modality to NTP
end for
for each continuous sequence **do** ▷ Across N frames
  if Avg $|t'_{I_m} - t'_{I_i}| > \tau$ **then**
    Discard all N frames
  else
    Use data for training
  end if
end for

another challenge—*the accurate alignment between mmWave frames and RGB frames with different sampling rates.*

5.3 Cross-Modality Data Alignment

We propose two countermeasures to achieve the goal of data alignment, as shown in Algorithm 1. First, before capturing the mmWave data frames from both sides, the Raspberry Pi queries Network Time Protocol (NTP) servers using the ntplib library [7] to obtain the NTP timestamp t_{I_m} and records the local timestamp t_{L_m} at the same time. By doing so, we can obtain the offset O_m between the Raspberry Pi's local timestamps and the NTP timestamps, and further convert the local timestamps of the recorded mmWave frames into NTP timestamps. Similarly, we also calculate the time offset of the image modality that is represented as O_i. By converting the local timestamps of both modalities into NTP timestamps, the data from the two modalities can be aligned using the NTP timestamps. Second, during the model development phase, if the difference between the NTP timestamps of the mmWave data and the image data for a certain frame exceeds a threshold τ, the data from that frame will be discarded and will not participate in the deep learning model training.

In scenarios involving time-series input, such as using a continuous sequence of N frames as a sample, it is necessary to evaluate whether the average timestamp difference across the N frames exceeds the threshold τ. By implementing these two countermeasures together, the two modalities can be precisely aligned, and the synchronization error will be controlled within τ. In this paper, we set τ to 20 ms. At this stage, by integrating the approaches from Sec. 5.2 and Sec. 5.3, we address **the second key challenge** in Sec. 1.

5.4 Signal Processing

To extract informative features for HMR from Frequency Modulated Continuous Wave (FMCW), we tailored a signal processing pipeline as illustrated in Fig. 7. Specifically, the pipeline can be divided into three main components: **(1)** Moving Target Indicator (MTI), Clutter Removal, and Range-Doppler Processing for dynamic object attention; **(2)** Energy Compensation, Digital Beamforming (DBF), and Range-Gating for spatial attention; **(3)** Energy-Based

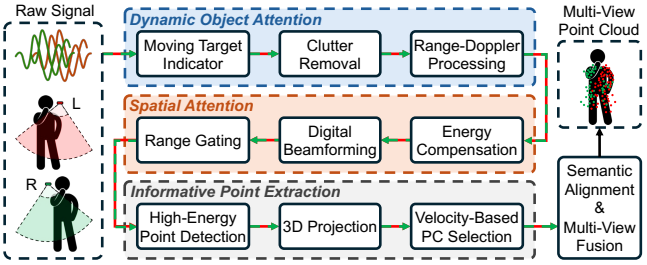


Fig. 7: Signal processing pipeline.

and Velocity-Based Selection for informative point extraction. Next, we elaborate on these components, starting with the dynamic object attention, and present an example of the point cloud estimated by the proposed pipeline in Fig. 8, Fig. 9, and Fig. 10.

MTI, Clutter Removal, and Range-Doppler Processing. The radar configuration of *Argus* includes a frame period T_f of 100 ms, containing N_c of 128 chirps per frame. Each chirp is a linearly modulated continuous wave with a bandwidth B of 3 GHz (60–63 GHz), consisting of N_s samples and lasting T_c time. In this paper, N_s and T_c are set to 128 and 700 µs, respectively, to achieve 5 cm fine-grained spatial resolution for sensing. Before processing the raw FMCW signal into range-Doppler maps, the impact of static objects (*e.g.*, walls within the radar FOV) should be carefully considered, as they can interfere with the detection of moving targets and reduce the effectiveness of the radar.

To mitigate interference from static objects, the MTI [53] method is employed on raw data before Range FFT to suppress stationary clutter signals. Fig. 13 shows the results of using MTI with the α_{MTI} set to 0.3 and the previous five mmWave frames for static object removal, before using MTI, the resulting point cloud is densely located in regions with smaller Z values (≈ -0.3) due to specular reflections from the shoulders and acromia. After using MTI, these less active regions are partially removed by considering them as static components. Since static components are effectively filtered out, employing a velocity-based selection method allows the selected fixed-count points to more accurately reflect the motion state of the target. Furthermore, *clutter removal* is a critical step following the range Fast Fourier Transform (FFT) on the Intermediate Frequency (IF) signal. By capturing the average signal level across all chirps at each range bin and subtracting this mean from the original input data, the impact of static components can be effectively removed, thereby enhancing the perception of human motions.

After applying MTI and clutter removal techniques, we have removed most of the interference from static objects and clutter. Next, we perform a Doppler FFT on the result of the range FFT (for velocity information) to complete the range-Doppler processing. Specifically, a *Range FFT* along the sample axis of each radar frame results in 2D frames where the magnitudes signify the reflected energy of targets at different distances; a *Doppler FFT* applied along the chirp axis yields range-Doppler maps, where the magnitude of each cell indicates the reflective energy of targets at specific ranges and velocities.

Energy Compensation, DBF, and Range Gating. As discussed in Sec. 1, a significant challenge in egocentric mmWave sensing is

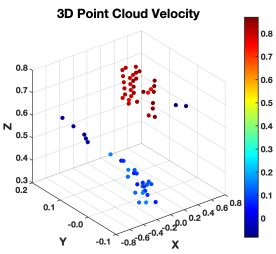


Fig. 8: PC (Velocity).

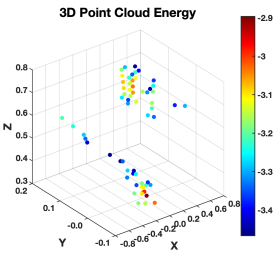


Fig. 9: PC (Energy).

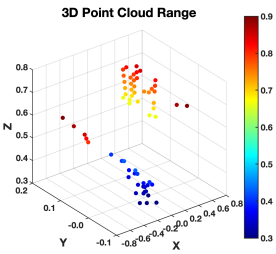


Fig. 10: PC (Range).

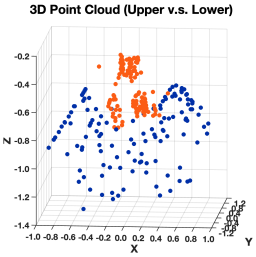


Fig. 11: Range-gating.

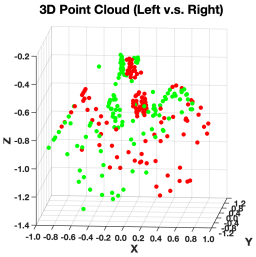
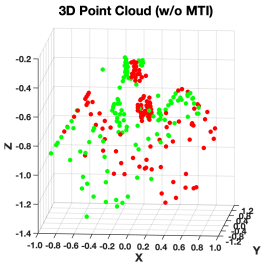
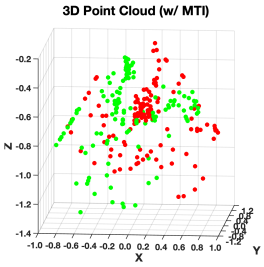


Fig. 12: Multi-view.



(a) Point cloud w/o MTI



(b) Point cloud w/ MTI

Fig. 13: Point clouds generated from the same mmWave frame with and without MTI.

Algorithm 2 Energy Compensation for Range-Doppler Map

Require: rd_s: Range-Doppler spectrum
Ensure: Compensated range-Doppler spectrum

$R, C \leftarrow \text{shape}(rd_s)$ ▷ The number of range bins and channels
for $c = 1$ to C **do**
  $M_r \leftarrow \text{mean}(\text{abs}(rd_s[:, :, c], \text{axis=1}))$ ▷ Range bins energy
  $M \leftarrow \text{mean}(M_r)$ ▷ Mean energy across bins
  for $r = 1$ to R **do**
    $m \leftarrow \text{mean}(\text{abs}(rd_s[r, :, c]))$
    $compensation \leftarrow M/m$ ▷ Scaling factor
    $\tilde{rd_s}[r, :, c] \leftarrow rd_s[r, :, c] \times compensation$
  end for
end for
return $\tilde{rd_s}$

the dual obstacle posed by self-occlusion and specular reflection. To address this challenge, it is essential to enhance the pipeline's spatial attention capability through signal processing techniques. To this end, we apply DBF to align spatial features within defined spatial grids, reducing randomness and dispersion while balancing computational overhead and spatial resolution. Furthermore, we propose two ad hoc methods, Energy Compensation and Range Gating, which take effect before and after DBF, respectively.

Due to the energy attenuation of mmWave signals during transmission, the magnitudes representing the energy level in range-Doppler maps for the lower body are significantly weaker than those for the upper body. This further leads to the point cloud, estimated by detecting high-energy points from range-Doppler maps, being concentrated in the upper body. To mitigate the impact of energy attenuation, we designed an energy compensation method for range-Doppler maps, as shown in Algorithm 2. We apply different scaling factors to range bins to make sure that the energy levels of them are the same after compensation. The proposed energy compensation increases the perception of the lower body range by improving spatial attention. Subsequently, the DBF technique [54] is applied to the compensated results as an alternative to the conventional third FFT (*i.e.*, angle FFT), reducing the ambiguity of the projected points. Finally, we empirically gate range-Doppler maps into two ranges (*i.e.*, 0.3 m–0.9 m and 0.9 m–1.5 m) and estimate point clouds from each range, respectively. By doing so, it can eliminate the specular reflection from the shoulders and acromia (usually within 0.3 m), and focus the lower body separately (Fig. 11).

Energy-Based and Velocity-Based Selection. With the support of the aforementioned components, the radar's sensing ability for dynamic targets and spatial information has been enhanced; however, refining the detected features to remove redundant information is essential. Specifically, we first detect high-energy points over a threshold of −3.5 dB in the aggregated range-Doppler map and project these points into a 3D space. Then we filter the projected point cloud based on velocity, retaining only the top N_{pn} points with the highest and lowest velocities; here, we set N_{pn} to 32 based on experimental trials. Therefore, we can obtain 64 points from the upper/lower body in one view. Finally, we can obtain 64×2 body parts $\times 2$ views $= 256$ points from each mmWave frame. Using the energy-based and velocity-based selections above, the estimated point clouds provide valuable information related to motion. Fig. 12 provides an example of the estimated point clouds from different views that complement each other, alleviating the problem of self-occlusion and enhancing the system's sensing capability.

5.5 Deep Neural Network Design

After we obtain multi-view point clouds from the upper and lower body separately, we elaborate on the design of the deep neural network used to translate multi-view point clouds into joint rotation matrices in this section. As Fig. 14 shows, the neural network includes three main components: (1) PointNet++; (2) Long Short-Term Memory (LSTM) and Kolmogorov-Arnold Networks (KAN); (3) Multi-Head Attention. These components are designed to cooperate seamlessly to capture and process the complex spatial and temporal relationships present in the point clouds.

Multi-Scale Point Cloud Feature Extractor. For the specific task of HMR, it is crucial to extract point cloud features across multiple scales. The features extracted at different scales correspond to different levels of physical significance. As the scale radius increases

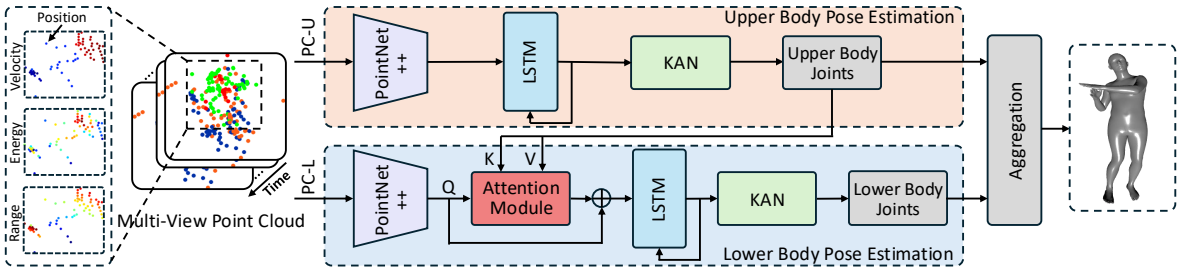


Fig. 14: KAN-based multi-view fusion network.

from small to large, the extracted features transition from local to global, enabling more precise HMR. To achieve this goal, we use PointNet++ [47] as the point cloud feature extractor to capture both local and global geometric features. It employs a series of Set Abstraction layers that progressively group points based on their spatial proximity and extract features from each group. This approach allows the model to learn multi-scale features effectively, making it well-suited for tasks that require detailed and hierarchical understanding of point cloud, such as HMR.

Specifically, we use three SA layers to progressively downsample the point cloud and extract multi-scale features. The first SA layer samples 128 points using radii of 0.1 m, 0.2 m, and 0.4 m, with each radius having 16, 32, and 64 neighbors, respectively. The second SA layer further downsamples to 64 points with radii of 0.2 m, 0.4 m, and 0.8 m, and the final SA layer aggregates global features without downsampling. Feature Propagation layers are then used to propagate the features back to the original point resolution, enabling the network to learn both local and global features effectively.

LSTM and KAN. Besides a powerful point cloud feature extractor, LSTM modules and KAN modules are also included in our deep neural network to serve as important roles.

Since *Argus* is designed to accept a series of sequential mmWave frames as inputs, we use the LSTM module to learn the relationships between adjacent frames to smooth the output. The hidden state from the previous time step (*i.e.*, frame) in an LSTM is passed to the next time step and used, along with the current input, to compute the next hidden state. This allows the LSTM to capture and maintain temporal dependencies across the sequence, enabling the modeling of long-term dependencies. After obtaining the deep features of the current frame, we use KAN [36] to transform these deep features into the dimensions required by the SMPL model. The core principle of KAN can be represented as follows:

$$f(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{2n+1} \Phi_j \left(\sum_{i=1}^{n} \phi_{ij}(x_i) \right),$$

where any multivariate continuous function $f(x_1, x_2, \ldots, x_n)$ can be decomposed into a sum of compositions of univariate continuous functions $\phi_{ij}(x_i)$ and Φ_j. It offers two significant advantages: (1) KAN, based on the Kolmogorov-Arnold representation theorem [1, 43], can theoretically approximate any complex multivariate function more effectively. For tasks like HMR, which require high-precision multivariate function approximation, KAN provides a more refined function mapping capability with fewer parameters than fully connected layers, thereby improving reconstruction performance; (2) The structure of KAN allows for better decoupling of different input dimensions, which is particularly advantageous when handling high-dimensional point clouds (*e.g.*, view, body part). This reduces model complexity, mitigates the risk of overfitting, and benefits HMR, where joint rotations involve highly non-linear and complex dependencies.

Multi-Head Attention. As mentioned in Sec. 1, a main challenge in achieving egocentric mmWave sensing is the problem of self-occlusion. To address this challenge, *Argus* features a hardware design that enables multi-view sensing, and we further fuse the information from the upper body to promote a more precise construction of the user's lower body. Specifically, we use the prediction of the upper body pose P_u as the K and V vectors in the attention module and feed the features of the lower body F_l as the Q vector. This process can be formulated as:

$$\mathrm{MHA}(Q_l, K_u, V_u) = \mathrm{softmax}\left(\frac{F_l W_{Q_l} P_u W_{K_u}^{\top}}{\sqrt{d_{kl}}} \right) P_u W_{V_u},$$

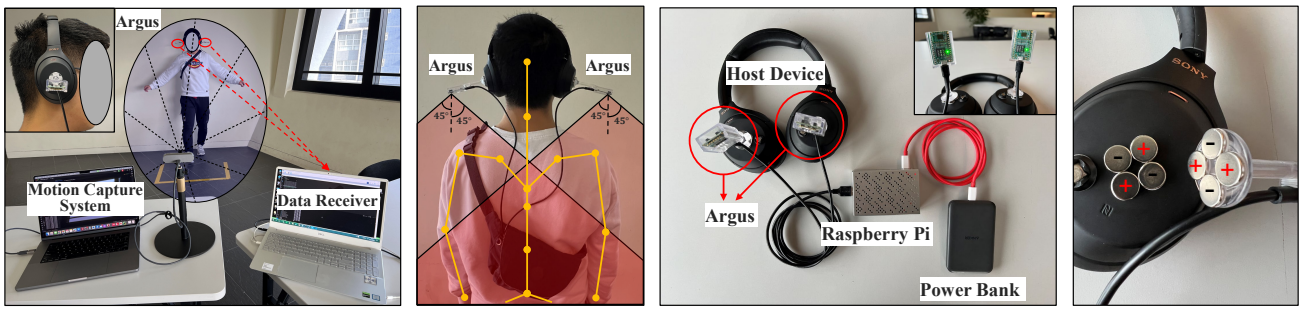
where W_{Q_l}, W_{K_u}, and W_{V_u} are the projection matrices for the Q, K, and V vectors, and d_{kl} is the scaling factor. The reason behind this design is to leverage the global information encoded in the upper body prediction to guide the lower body pose estimation, ensuring consistency and coherence between the two. The attention mechanism can effectively query the relevant global features from the upper body, helping to refine and adjust the lower body prediction. This method preserves the hierarchy of information, where the prediction of the lower body is guided by the upper body, leading to more reasonable and precise pose estimation. Our deep neural network, with 42.4K parameters, is well-designed to run smoothly on mobile devices.

With this combination of the signal processing pipeline (Sec. 5.4) and the multi-view fusion network (Sec. 5.5), **the third key challenge** in Sec. 1 is well addressed.

6 EVALUATION

6.1 Testbed and Experimental Settings

Testbed. We use a monocular RGB camera (Logitech BRIO Ultra HD Pro [37]) combined with the advanced image-to-mesh approach (HMR 2.0 [21]) to extract pose parameters for the SMPL model from monocular images. Specifically, as Fig. 15 shows, we place the camera in front of the participant and delineate a 1 m × 1 m area. Note that, participants' activities can go beyond the area. The purpose



(a) Experimental scenario. (b) Wearing diagram. (c) Details of *Argus*. (d) Detachable design.

Fig. 15: The testbed of *Argus*. The camera is used only to obtain pseudo-labels for training; once the model is trained and deployed on mobile devices, user movements are no longer restricted by it.

of setting this boundary is to assist participants in self-correction when they are in motion, preventing them from straying significantly from the camera's FOV. The camera can capture images from the real world up to 4K resolution. However, considering the balance between image quality and storage space, we use 640×480 as our implementary resolution, which means that low-cost RGB cameras with lower resolution are also applicable. For a better comparison, we follow previous studies [68] and extract the estimated rotation matrices of the unlocked joints (*i.e.*, except the wrist, hand, ankle and foot joints) as the target of training while setting the rotation matrices of the rest joints as identity matrix.

Experimental settings. We comprehensively evaluate the performance of *Argus* in various situations and compare it with open-source SOTA baselines. Unless otherwise specified, all experiments were carried out with a time-ordered split of 70%, 20%, and 10% for the training, validation, and testing datasets, respectively. Note that although the training, validation, and testing data come from the same collection session (per participant per activity), the data splitting follows a causal structure (without overlapping frames). This ensures that the prediction of a sample (N consecutive frames) does not depend on any future data. The Adam optimizer, with a default learning rate of 0.0003 and a batch size of 16, was adopted. The learning rate was decayed by a factor of 0.9 after each epoch, with a maximum of 50 training epochs and a patience of 5 epochs. Only the training dataset was segmented with a 1-frame overlap in temporal order, while the validation and test datasets were segmented without overlap.

6.2 Data Collection

In this paper, we invited 16 participants (9 males and 7 females, aged 21 to 32, with heights ranging from 1.58 to 1.90 m and weights from 49 to 83 kg) to perform 10 daily activities[1]. To comprehensively assess the perception capabilities of *Argus* across different body regions (*i.e.*, upper limb, lower limb, whole body), as Fig. 16 shows, the 10 activities include: (a) torso rotation; (b) arm swings; (c) bicep curl; (d) deep squat; (e) lunges; (f) leg swings; (g) marching in place; (h) walking back and forth; (i) side balance reach; (j) at ease. For each activity, each participant continues to perform it for 2 min. The sampling rates of mmWave radars and the RGB camera are 10 Hz and 30 Hz, respectively. We collected data over a long timespan (half a month) across three different locations to introduce environmental diversity. However, to maintain consistency in the main dataset, we ensured open surroundings free of clutter during data collection. As a result, the dataset contains 12,000 frames per participant, and the entire dataset contains more than **200,000** image-mmWave frames, including the data for the evaluation of micro-benchmarks.

6.3 Evaluation Metrics

We use the following metrics to evaluate the meshes reconstructed by our system for the activities mentioned above:

- **Average Vertex Error (V)** [3, 77]. The average vertex error by averaging the Euclidean distance between the vertices of reconstructed meshes and that of the corresponding GT[2].
- **Average Joint Localization Error (S)** [27, 77]. The average skeleton error by averaging the Euclidean distance between the joint locations of reconstructed skeletons and the corresponding GT.
- **Average Joint Rotation Error (Q)** [68]. The average joint rotation error between the predicted joint rotations and the corresponding GT.

Note that we did not include some metrics (*e.g.*, mesh location error, gender prediction accuracy) used in previous studies [66–68]. The reason is that *Argus* is a wearable system designed for self-sensing rather than sensing others. In such scenarios, users can provide accurate gender information and can use the body shape parameters (*i.e.*, betas) estimated by the monocular camera directly. Furthermore, given that *Argus* is attached to a head-mounted host device, it shares the same coordinate system with the user. Therefore, these metrics (*e.g.*, global location) are not included.

6.4 Overall Performance

We first exclude the data of four randomly selected users to serve as unseen users for further evaluation; their data are not included in any training or testing process except in Sec. 6.6. Subsequently, we train a user-specific model for each user and use the set of these

[1]This study has received the ethical approval from the authors' institution.

[2]GT means the labels generated from the ground-truth RGB images.

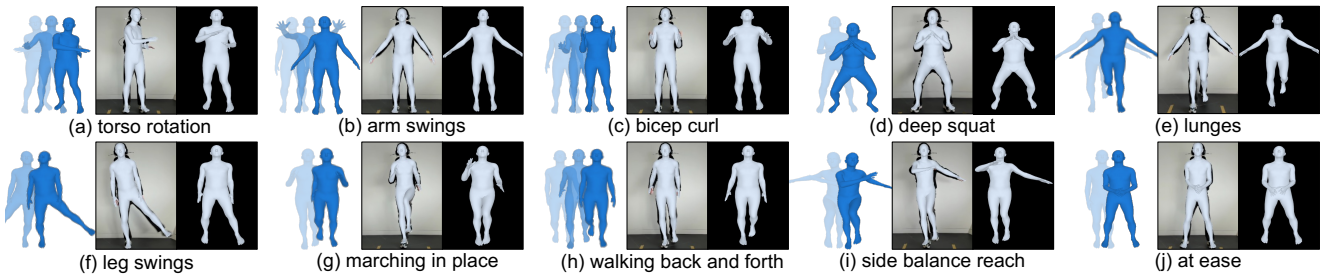


Fig. 16: Illustration of activities (left), ground truth (middle), and prediction (right) by *Argus*. (a)–(c) are upper-limb activities; (d)–(f) are lower-limb activities; (g)–(j) are whole-body activities.

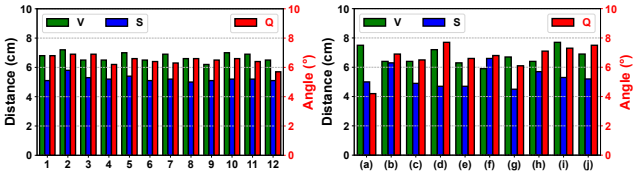


Fig. 17: Overall performance. **Fig. 18: Different activities.**

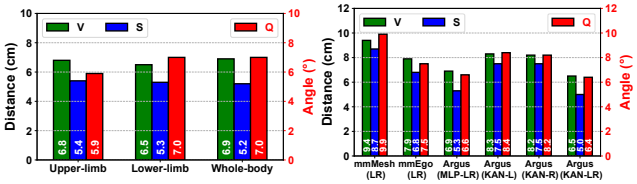


Fig. 19: Different regions. **Fig. 20: Baseline comparisons.**

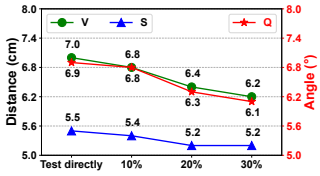


Fig. 21: Unseen users.

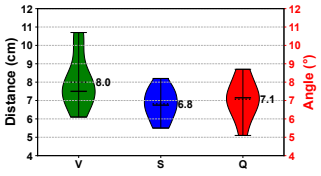


Fig. 22: Unseen poses.

training data to train a basic model to evaluate performance on unseen users. As Fig. 17 shows, the performance of *Argus* across all participants is stable, with average V, S, and Q errors of 6.8 cm, 5.3 cm, and 6.6°, respectively. However, the errors for Participant 2 are the highest, which can be attributed to the participant's height of 1.90 m. Following the signal processing described in Sec. 5.4, the mmWave features over 1.5 m will be eliminated by the Range-Gating; it does not adequately cover the lower body range. Furthermore, the absolute errors of the reconstructed meshes are positively correlated with the user's height.

We also conduct an in-depth analysis to investigate the performance of *Argus* in different activities and regions. The average errors for each activity are shown in Fig. 18; we find that the performance varies significantly. In general, the errors are higher for actions with greater complexity. For example, the side balance reach activity presents the highest V error, averaging 7.7 cm. The performance of *Argus* on the three different evaluation metrics does not show a significant correlation among activities. Although V error is as large as 7.5 cm during torso rotation activity, Q error is only 4.2°. For different regions (Fig. 19), the V and S errors are relatively stable, while the Q error for lower-limb and whole-body activities is 1.1° higher than that for upper-limb activities.

6.5 Baselines

We compare *Argus* with two egocentric SOTA baselines [33, 68] **declared in a recent study** [33]; the implementation of these baselines adopts the open-source code [23, 71]. Furthermore, to assess the effectiveness of KAN and the gain brought by the multi-view configuration, we introduce three ablation baselines. Specifically, we replace all KAN modules with MLP modules (of the same dimensions) to create a baseline named MLP-LR. Additionally, to validate the benefit of the multi-view configuration for HMR under self-occlusion, we have added two baselines (KAN-L and KAN-R). In these baselines, we only use the mmWave data from one-side radar and discard the input from the opposite side to mimic a single-view solution. For fair comparison, we use data from the 12 participants in Sec. 6.4 to train models with three different neural networks, and test them on the testing data.

Fig. 20 shows that *Argus* outperforms the two SOTA baselines (*i.e.*, mmMesh and mmEgo), achieving a 17.7–30.9% reduction in V error and a 14.7–35.4% reduction in Q error. The improvement can be attributed to the collaboration between PointNet++ and KAN, which are optimized for data involving hierarchical features. Furthermore, the multi-view configuration achieves an average gain of 21.2% and 22.9% in V and Q, respectively, over the single-view baselines. Although the performance does not reach the levels achieved by a previous study [33], it is important to note that the previous study was based on a high-capability, high-power, and high-cost radar, which features an equivalent 4×4 antenna array. In contrast, *Argus* utilizes stripped-down radars that feature an equivalent 2×2 antenna array. We overcome hardware limitations and, for the first time, achieve comparable performance with a system based on high-capability radars.

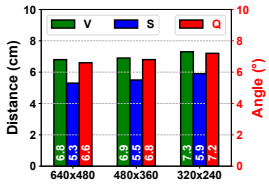


Fig. 23: Resolutions.

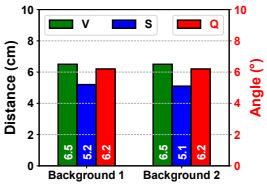


Fig. 24: Backgrounds.

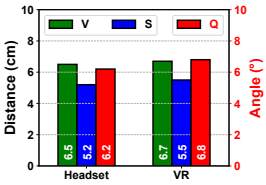


Fig. 25: Host devices.

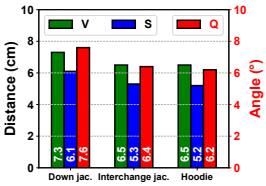


Fig. 26: Clothes.

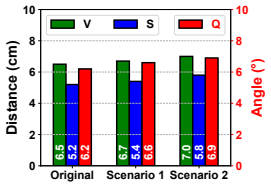


Fig. 27: Multipath.

6.6 Unseen Performance

Considering real-world application scenarios, we evaluate the performance of *Argus* on unseen users (2 males and 2 females) and unseen poses using leave-one-out method. Specifically, we first evaluate the basic model's performance on them by testing directly (without fine-tuning). Then, we fine-tune the basic model using a sub-dataset that accounts for x% of all training data for each participant. By increasing the value of x, our goal is to explore how much data is required to fine-tune the basic model to achieve satisfactory results for an unseen user. As Fig. 21 shows, when the training data used for fine-tuning reaches 30% of the total, the model achieves remarkable performance. Furthermore, owing to the generalization obtained from training data across multiple users, the performance of the model fine-tuned with 30% of the new user data surpasses the average performance of the user-specific models. Therefore, when deploying *Argus* for a new user, theoretically only $14 \times 0.3 = 4.2$ min of the new data collection is needed to deploy. For unseen poses, the Violin plot (Fig. 22) shows unstable performance, with average V and Q errors of 8.0 cm and 7.1°, respectively. Compared to vertices estimation, the results also reveal greater stability in rotation estimation, which may be attributed to the kinematic representations embedded in the trained model.

6.7 Other Micro-benchmarks

We also evaluate several factors as micro-benchmarks.

Image resolution. Most current mobile devices can capture RGB images with resolutions higher than 640×480. To simulate budget devices with lower-quality cameras, we downsample the RGB images to 480×360 and 320×240, evaluating the performance under lower-quality labels.

Image background. To evaluate the impact of the experimental background on the effectiveness of labels extracted by the proposed MoCap system, we invite two participants to record a new set of data at different experimental sites.

Host device. We select two participants and replace the host device with a VR (*i.e.*, HTC VIVE Pro Eye [14]), using the same magnetic attachment method as shown in Fig. 15(d). A new set of data is collected for each participant to evaluate the performance of *Argus* on different host devices.

Type of clothing. To evaluate the robustness of *Argus*, we select two participants and collect data while they wear different clothing. Specifically, for each participant, in addition to the data (wearing a hoodie) collected in Sec. 6.4, we also collect data while they wear an interchangeable jacket and a down jacket, with all other experimental settings the same.

Multipath effect. Since mmWave sensing suffers from the multipath effect, we evaluated the robustness of *Argus* to scatters and reflectors, such as clothes and furniture. Two participants collected additional testing data with sundries placed close by without hindering activities. In Scenario 1, a chair with a backpack was placed next to the participant. In Scenario 2, an additional chair with a down jacket was added. Testing the model trained in a clutter-free scenario allowed us to assess the robustness to multipath effects.

The results are listed in Fig. 23–27, *Argus* is robust to image resolution and background. For different devices, the variation in geometry directly affects the deployment angle and spatial position of *Argus*, and the results show that deploying *Argus* on a VR system leads to V and S errors that are approximately 3% and 5.7% higher, respectively, compared to deploying it on a headset. The user's clothing significantly affects system performance. Clothing with large reflective surfaces, such as down jackets, exacerbates both self-occlusion and specular reflection issues. As a result, the inability to precisely estimate joint rotations and positions leads to substantial errors in the results.

6.8 Computational Delay

To verify the practicality of *Argus*, we evaluate its computational delay. Unlike previous studies [33, 68], which used TI xWRxx43 radars with well-supported C-based signal processing to obtain point clouds, our work involved engineering efforts to process raw signals from BGT60TR13C radars, which lack such support. We developed a Python-based signal processing pipeline, allowing for greater customization of processing details. Despite the current signal processing time being 0.880 s on an Intel Core i7 CPU, the model inference and SMPL rendering times are 11.6 ms and 1.5 ms, respectively, leading to an overall algorithmic delay of 0.89 s.

Based on a previous study [68] using a AWR1843BOOST radar, which reported a 28 ms processing time for converting raw signals to point clouds, we reasonably estimate that transitioning our signal processing to the C programming language and leveraging hardware acceleration could reduce the delay to within 80 ms (approximately a 10x speedup from Python). *Argus* demonstrates the feasibility of egocentric HMR using stripped-down radars. If well-supported C-based signal processing and hardware acceleration are supported for the radar, real-time HMR with an algorithmic delay of less than 100 ms per frame is achievable.

7 LIMITATIONS & FUTURE WORK

While our study shows the effectiveness of *Argus* for egocentric HMR using stripped-down radars in a multi-view manner, it is subject to the assumption that the user's poses will not occlude the

radars at close range, and there are limitations that require further investigation.

Restricted activities. The head-mounted form factor of *Argus* and significant head movements limit the detection of activities above the ear level, such as raising arms overhead, resulting in restricted activities in HMR. However, the supported sensing range covers most daily activities. To overcome this limitation, adding additional sensors, oriented differently, could be a promising solution to cover the sensing blind spots. Furthermore, existing solutions, including *Argus*, are mainly focused on predefined actions or their combinations [33, 67, 68], where achieving good performance on arbitrary actions remains a challenge in mmWave-based HMR. The weak generalization ability is mostly due to insufficient data volume and dataset diversity, which could be improved through few-shot learning or meta-learning techniques.

More sophisticated hardware. *Argus* is a prototype featuring a Raspberry Pi as an intermediary between mobile devices and radars. Future work could involve designing a dedicated, more compact, and faster PCB as the intermediary. Moreover, a more powerful radar model has been released recently (*i.e.*, BGT60ATR24C [57]) with 2 transmit antennas and 4 receive antennas are available, which could further enhance the system's performance in HMR task.

8 CONCLUSION

This paper introduces a novel mmWave-based add-on system named *Argus*, which is the first wearable add-on based on stripped-down mmWave radars deployed in a multi-view configuration for egocentric HMR. The key idea of *Argus* is that the compact, limited-capability mmWave radars on both the left and right sides can form a multi-view sensing system that mitigates self-occlusion and specular reflection issues in such an egocentric view through complementary viewpoints. By addressing three unique challenges, the limitations caused by the stripped-down radar are mitigated, achieving performance comparable with solutions based on high-capability radars. Extensive evaluation demonstrates the robustness and practicality of *Argus*, making it a promising alternative to existing HMR solutions.

ACKNOWLEDGMENTS

We thank the shepherd and the anonymous reviewers for their insightful and valuable comments. The work was partially supported by the Research Grants Council (RGC) of Hong Kong under GRF 11201422, CRF C4072-21G, STG 1/E-403/24-N.

REFERENCES

[1] Vladimir Igorevich Arnol'd. 1957. On functions of three variables. In *Doklady Akademii Nauk*.

[2] Kang Min Bae, Hankyeol Moon, and Song Min Kim. 2024. SuperSight: Sub-cm NLOS Localization for mmWave Backscatter. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*.

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*.

[4] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *Proceedings of the 25th annual international conference on mobile computing and networking (MobiCom)*.

[5] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2023. hEARt: Motion-resilient Heart Rate Monitoring with In-ear Microphones. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom)*.

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

[7] cf-natali et al. 2023. ntplib - Python NTP library. https://pypi.org/project/ntplib Accessed: 2024-02-21.

[8] Justin Chan, Nada Ali, Ali Najafi, Anna Meehan, Lisa R Mancl, Emily Gallagher, Randall Bly, and Shyamnath Gollakota. 2022. An off-the-shelf otoacoustic-emission probe for hearing screening via a smartphone. *Nature biomedical engineering (NAT BIOMED ENG)* (2022).

[9] Justin Chan, Antonio Glenn, Malek Itani, Lisa R Mancl, Emily Gallagher, Randall Bly, Shwetak Patel, and Shyamnath Gollakota. 2023. Wireless earbuds for low-cost hearing screening. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*.

[10] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*.

[11] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. 2022. mmBody benchmark: 3D body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia (MM)*.

[12] Tao Chen, Yongjie Yang, Xiaoran Fan, Xiuzhen Guo, Jie Xiong, and Longfei Shangguan. 2024. Exploring the Feasibility of Remote Cardiac Auscultation Using Earphones. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom)*.

[13] Blender Online Community. 2024. Blender - a 3D modelling and rendering package. http://www.blender.org Version 4.1.1, Accessed: 2024-08-05.

[14] HTC Corporation. 2024. HTC VIVE Pro Eye. https://www.vive.com/hk/product/vive-pro-eye/overview Accessed: 2024-08-07.

[15] Yimin Dai, Xian Shuai, Rui Tan, and Guoliang Xing. 2023. Interpersonal distance tracking with mmWave radar and IMUs. In *Proceedings of the International Conference on Information Processing in Sensor Networks (IPSN)*.

[16] Nathan DeVrio, Vimal Mollyn, and Chris Harrison. 2023. SmartPoser: Arm Pose Estimation with a Smartphone and Smartwatch Using UWB and IMU Data. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*.

[17] Di Duan, Yongliang Chen, Weitao Xu, and Tianxing Li. 2024. EarSE: Bringing Robust Speech Enhancement to COTS Headphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2024).

[18] Di Duan, Zehua Sun, Tao Ni, Shuaicheng Li, Xiaohua Jia, Weitao Xu, and Tianxing Li. 2024. F2Key: Dynamically Converting Your Face into a Private Key Based on COTS Headphones for Reliable Voice Interaction. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*.

[19] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard Howard. 2021. HeadFi: bringing intelligence to all headphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*.

[20] Shubham Goel. 2023. 4DHumans: Reconstructing and Tracking Humans with Transformers. https://github.com/shubham-goel/4D-Humans

[21] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[22] Mingda Han, Huanqi Yang, Tao Ni, Di Duan, Mengzhe Ruan, Yongliang Chen, Jia Zhang, and Weitao Xu. 2024. mmSign: mmWave-based few-shot online handwritten signature verification. *ACM Transactions on Sensor Networks (TOSN)* (2024).

[23] HavocFiXer. 2024. mmMesh. https://github.com/HavocFiXer/mmMesh Accessed: 2024-07-10.

[24] Changshuo Hu, Thivya Kandappu, Yang Liu, Cecilia Mascolo, and Dong Ma. 2024. BreathPro: Monitoring Breathing Mode during Running with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2024).

[25] Texas Instruments. 2024. IWR6843ISK-ODS. https://www.ti.com.cn/tool/cn/IWR6843ISK-ODS Accessed: 2024-07-13.

[26] Nan Jiang, Terence Sim, and Jun Han. 2022. EarWalk: towards walking posture identification using earables. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications (HotMobile)*.

[27] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using WiFi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom)*.

[28] Taeho Kang, Kyungjin Lee, Jinrui Zhang, and Youngki Lee. 2023. Ego3dpose: Capturing 3d cues from binocular egocentric views. In *Proceedings of the SIGGRAPH Asia 2023 Conference Papers (SIGGRAPH Asia)*.

[29] Taeho Kang and Youngki Lee. 2024. Attention-propagation network for egocentric heatmap to 3d pose lifting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

[31] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. m3track: mmwave-based multi-user 3d posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*.

[32] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. 2023. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[33] Wenwei Li, Ruofeng Liu, Shuai Wang, Dongjiang Cao, and Wenchao Jiang. 2023. Egocentric Human Pose Estimation using Head-mounted mmWave Radar. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems (SenSys)*.

[34] Qianru Liao, Yongzhi Huang, Yandao Huang, Yuheng Zhong, Huitong Jin, and Kaishun Wu. 2022. MagEar: eavesdropping via audio recovery using magnetic side channel. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*.

[35] Shuangjun Liu, Xiaofei Huang, Nihang Fu, Cheng Li, Zhongnan Su, and Sarah Ostadabbas. 2022. Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2022).

[36] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. 2024. KAN: Kolmogorov-Arnold Networks. *arXiv preprint arXiv:2404.19756 (arXiv)* (2024).

[37] Logitech. 2023. BRIO Ultra HD Pro Business Webcam. https://www.logitech.com/en-ca/products/webcams/brio-4k-hdr-webcam.960-001105.html Accessed: 2024-01-25.

[38] Shengzhe Lyu, Yongliang Chen, Di Duan, Renqi Jia, and Weitao Xu. 2024. EarDA: Towards Accurate and Data-Efficient Earable Activity Sensing. *arXiv preprint arXiv:2406.16943 (arXiv)* (2024).

[39] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. 2021. OESense: employing occlusion effect for in-ear human sensing. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*.

[40] 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 (IMWUT)* (2023).

[41] Microsoft. 2024. Azure Kinect DK. https://www.microsoft.com/en-us/p/azure-kinect-dk/8pp5vxmd9nhq Accessed: 2024-07-14.

[42] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI)*.

[43] Kolmogorov, Andreĭ Nikolaevich. 1963. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Translations American Mathematical Society* (1963).

[44] OptiTrack. 2024. OptiTrack. https://optitrack.com Accessed: 2024-07-14.

[45] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*.

[46] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom)*.

[47] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems (NeurIPS)* (2017).

[48] Kun Qian, Zhaoyuan He, and Xinyu Zhang. 2020. 3D point cloud generation with millimeter-wave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2020).

[49] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2021. Winect: 3d human pose tracking for free-form activity using commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2021).

[50] Andrea Ronco, Philipp Schilk, and Michele Magno. 2024. TinyssimoRadar: In-Ear Hand Gesture Recognition with Ultra-Low Power mmWave Radars. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation (IoTDI)*.

[51] Yuto Shibata, Yutaka Kawashima, Mariko Isogawa, Go Irie, Akisato Kimura, and Yoshimitsu Aoki. 2023. Listening human behavior: 3d human pose estimation with acoustic signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[52] Xian Shuai, Yulin Shen, Yi Tang, Shuyao Shi, Luping Ji, and Guoliang Xing. 2021. millieye: A lightweight mmwave radar and camera fusion system for robust object detection. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation (IoTDI)*.

[53] Merrill Ivan Skolnik et al. 1980. *Introduction to radar systems*.

[54] Warren L Stutzman and Gary A Thiele. 2012. *Antenna theory and design*.

[55] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2022. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)* (2022).

[56] Infineon Technologies. 2023. BGT60TR13C: 60GHz Radar Sensor. https://www.infineon.com/cms/en/product/sensor/radar-sensors/radar-sensors-for-iot/60ghz-radar/bgt60tr13c/ Accessed: 2024-07-16.

[57] Infineon Technologies. 2024. BGT60ATR24C: 60 GHz Radar Sensor. https://www.infineon.com/cms/en/product/sensor/radar-sensors/radar-sensors-for-automotive/60ghz-radar/bgt60atr24c Accessed: 2024-08-15.

[58] Bandhav Veluri, Malek Itani, Justin Chan, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Semantic hearing: Programming acoustic scenes with binaural hearables. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*.

[59] Vicon. 2023. Vicon. https://www.vicon.com Accessed: 2024-07-14.

[60] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. 2019. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *Proceedings of the 27th ACM international conference on multimedia (MM)*.

[61] Shuai Wang, Dongjiang Cao, Ruofeng Liu, Wenchao Jiang, Tianshun Yao, and Chris Xiaoxuan Lu. 2023. Human Parsing with Joint Learning for Dynamic mmWave Radar Point Cloud. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2023).

[62] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Toothsonic: Earable authentication via acoustic toothprint. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2022).

[63] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (2021).

[64] Yucheng Xie, Ruizhe Jiang, Xiaonan Guo, Yan Wang, Jerry Cheng, and Yingying Chen. 2022. mmFit: Low-effort personalized fitness monitoring using millimeter wave. In *Proceedings of the International Conference on Computer Communications and Networks (ICCCN)*.

[65] Xiangyu Xu, Jiadi Yu, Chengguang Ma, Yanzhi Ren, Hongbo Liu, Yanmin Zhu, Yi-Chao Chen, and Feilong Tang. 2022. mmECG: Monitoring human cardiac cycle in driving environments leveraging millimeter wave. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*.

[66] Hongfei Xue, Qiming Cao, Yan Ju, Haochen Hu, Haoyu Wang, Aidong Zhang, and Lu Su. 2022. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys)*.

[67] Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. 2023. Towards Generalized mmWave-based Human Pose Estimation through Signal Augmentation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking (MobiCom)*.

[68] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*.

[69] Huanqi Yang, Mingda Han, Xinyue Li, Di Duan, Tianxing Li, and Weitao Xu. 2024. iRadar: Synthesizing Millimeter-Waves from Wearable Inertial Inputs for Human Gesture Sensing. *arXiv preprint arXiv:2412.15980 (arXiv)* (2024).

[70] Jiarui Yang, Songpengcheng Xia, Yifan Song, Qi Wu, and Ling Pei. 2024. mmBaT: A Multi-Task Framework for Mmwave-Based Human Body Reconstruction and Translation Prediction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[71] yenanjing. 2024. mmEgo. https://github.com/yenanjing/mmEgo Accessed: 2024-07-10.

[72] Tianhong Catherine Yu, Peter He, Chi-Jung Lee, Cassidy Cheesman, Saif Mahmud, Ruidong Zhang, François Guimbretière, Cheng Zhang, et al. 2024. SeamPose: Repurposing Seams as Capacitive Sensors in a Shirt for Upper-Body Pose Tracking. *arXiv preprint arXiv:2406.11645 (arXiv)* (2024).

[73] Bo Zhang, Zimeng Zhou, Boyu Jiang, and Rong Zheng. 2024. SUPER: Seated Upper Body Pose Estimation using mmWave Radars. In *Proceedings of the 9th International Conference on Internet-of-Things Design and Implementation (IoTDI)*.

[74] Jia Zhang, Rui Xi, Yuan He, Yimiao Sun, Xiuzhen Guo, Weiguo Wang, Xin Na, Yunhao Liu, Zhenguo Shi, and Tao Gu. 2023. A survey of mmWave-based human

sensing: Technology, platforms and applications. *IEEE Communications Surveys & Tutorials* (2023).

[75] Xiaotong Zhang, Zhenjiang Li, and Jin Zhang. 2022. Synthesized Millimeter-Waves for Human Motion Sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys)*.

[76] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, and Dina Katabi. 2019. Through-wall human mesh recovery using radio signals. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[77] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*.

[78] Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. 2023. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal (IoTJ)* (2023).