remote sensing

MDPI

Article

SCRP-Radar: Space-Aware Coordinate Representation for Human Pose Estimation Based on SISO UWB Radar

Xiaolong Zhou , Tian Jin *, Yongpeng Dai , Yongping Song and Kemeng Li

College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China; zhouxiaolong@nudt.edu.cn (X.Z.); dai_yongpeng@nudt.edu.cn (Y.D.); songyongping08@nudt.edu.cn (Y.S.); likemeng18@nudt.edu.cn (K.L.)
* Correspondence: tianjin@nudt.edu.cn

Abstract: Human pose estimation (HPE) is an integral component of numerous applications ranging from healthcare monitoring to human-computer interaction, traditionally relying on vision-based systems. These systems, however, face challenges such as privacy concerns and dependency on lighting conditions. As an alternative, short-range radar technology offers a non-invasive, lighting-insensitive solution that preserves user privacy. This paper presents a novel radar-based framework for HPE, SCRP-Radar (space-aware coordinate representation for human pose estimation using single-input single-output (SISO) ultra-wideband (UWB) radar). The methodology begins with clutter suppression and denoising techniques to enhance the quality of radar echo signals, followed by the construction of a micro-Doppler (MD) matrix from these refined signals. This matrix is segmented into bins to extract distinctive features that are critical for pose estimation. The SCRP-Radar leverages the Hrnet and LiteHrnet networks, incorporating space-aware coordinate representation to reconstruct 2D human poses with high precision. Our method redefines HPE as dual classification tasks for vertical and horizontal coordinates, which is a significant departure from existing methods such as RF-Pose, RF-Pose 3D, UWB-Pose, and RadarFormer. Extensive experimental evaluations demonstrate that SCRP-Radar significantly surpasses these methods in accuracy and robustness, consistently exhibiting lower average error rates, achieving less than 40 mm across 17 skeletal keypoints. This innovative approach not only enhances the precision of radar-based HPE but also sets a new benchmark for future research and application, particularly in sectors that benefit from accurate and privacy-preserving monitoring technologies.

Keywords: short-range radar; human pose estimation; micro-Doppler; coordinate representation

check for updates

Citation: Zhou, X.; Jin, T.; Dai, Y.; Song, Y.; Li, K. SCRP-Radar: Space-Aware Coordinate Representation for Human Pose Estimation Based on SISO UWB Radar. *Remote Sens.* **2024**, *16*, 1572. https://doi.org/10.3390/rs16091572

Academic Editor: Yanhua Wang

Received: 21 February 2024
Revised: 24 April 2024
Accepted: 25 April 2024
Published: 28 April 2024

1. Introduction

Recent advances in wireless technologies have significantly enhanced the capabilities of short-range radar systems [1], particularly in remote sensing for human target detection and perception. These systems are increasingly sought after for sophisticated surveillance and monitoring applications that require operation in diverse environments without compromising privacy or safety [2]. Human pose estimation, a critical component of urban environmental perception, involves deducing the posture of the human body by recognizing and locating different body parts, such as ankles, shoulders, and wrists [3]. This technology is pivotal in applications ranging from security to healthcare monitoring, where understanding human intentions and actions through pose estimation is essential.

Traditional methods for human pose estimation have predominantly relied on optical sensors, such as cameras, which have achieved significant success in accurately capturing human movements [4]. These camera-based systems are adept at providing high-resolution data and have been instrumental in advancing the field. However, the use of continuous surveillance raises substantial privacy concerns. The intrusive nature of constant video monitoring and the vulnerability of wireless security cameras to hacking are significant drawbacks that have prompted researchers to explore less invasive methods [5]. To address these privacy issues, researchers have explored alternative non-visual technologies such as WiFi for human pose estimation. WiFi-based systems utilize Channel State Information (CSI) from commodity WiFi devices to deduce human poses. Pioneering work by Wang et al. [6] utilized deep learning to process 1D WiFi signals for this purpose, proving that non-visual methods could effectively estimate human postures. However, the coarse resolution of common WiFi frequencies, such as 2.4 GHz and 5 GHz with limited bandwidths of 20 MHz and 40 MHz, restricts their ability to capture fine-grained movements, which is crucial for accurate pose estimation. The use of Radio Frequency Identification (RFID) technology has been proposed to overcome some of these limitations. RFID systems, such as the RFID-Pose by Mao et al. [7], employ wearable tags and commodity RFID readers for 3D pose estimation, effectively monitoring multiple human joints in real time. Although RFID technology offers a small form factor and the ability to track multiple points, it struggles with the low data rates that are typical of RFID systems, making it challenging to generate a joint confidence map for all joints with the precision required in pose estimation.

Despite advancements in WiFi and RFID-based systems, both approaches face inherent resolution and data rate limitations that affect their practicality and effectiveness in pose estimation. This sets the stage for considering alternative technologies offering more accurate, reliable, and non-intrusive methods. Radar technology emerges as a superior alternative in this context. Unlike cameras, WiFi, and RFID systems, radar-based human pose estimation is not hindered by lighting conditions or line-of-sight restrictions, allowing continuous operation in diverse environments. However, accurately interpreting radar reflections to estimate human poses presents significant challenges. The complexity arises from the need to distinguish between signals reflected from the body and those from other objects or backgrounds and the dynamic nature of human movement. Recent studies have focused on developing advanced signal processing algorithms and machine learning models to improve the accuracy and reliability of pose estimation. These approaches often use deep learning techniques to classify radar signals and predict body positions. For instance, technologies such as RF-Pose [8], RF-Pose 3D [9], and RF-Capture [10] utilize Frequency Modulated Continuous Wave (FMCW) technology and require a complex assembly of a 16 + 4 T-shaped antenna array with extensive bandwidth (1.78 GHz) to produce depth maps.

However, a significant challenge MIMO radar imaging-based approaches encounter is their susceptibility to environmental variations. Factors such as changes in the surrounding environment and the relative distance between the human target and the radar can drastically affect the quality of radar imaging, thereby impacting the pose estimation accuracy. To address these challenges, this paper introduces a novel approach by employing SISO FMCW UWB radar for human pose estimation. Unlike traditional methods that rely heavily on radar imaging quality, our technique capitalizes on the micro-Doppler signature, a feature inherently less affected by environmental variations and the target's distance from the radar, thereby offering a more robust solution for accurate pose estimation [11]. Here are some of the key challenges:

(1) Dataset Limitations in Radar-Based Human Pose Estimation: Acquiring accurately labeled radar data for human poses is time-consuming and labor-intensive. Large, diverse, and accurately labeled datasets are necessary for training and evaluating machine learning models. Determining which part of the reflected signal corresponds to the human target is not straightforward, making manual annotation of radar signals with key-points an impractical task.

(2) Challenges with Limited Channel Radar Systems: Current research in radar-based human pose estimation predominantly relies on massive MIMO radar systems. Short-range radar systems often have limited spatial resolution due to the radar antennas' physical size and the signals' wavelengths. This limitation can make distinguishing between closely spaced body parts difficult, leading to less precise pose estimation.

(3) Subject Variability: People come in various shapes and sizes, and their clothing can also affect radar signal reflection. A system trained on a specific dataset might struggle with generalization across different subjects. Human movements are complex and

dynamic, with various possible poses and actions. Capturing the full extent of this variability poses a significant challenge for pose estimation systems.

The micro-Doppler signatures of a human target strongly indicate the target's pose and motion. For instance, swinging arms or legs during walking generates characteristic patterns in the micro-Doppler spectrum that are distinct from those generated by other actions such as running or waving [12]. This distinction arises because each type of movement has a unique velocity profile over time, captured by the micro-Doppler effect [13]. Consequently, analyzing these signatures allows for the inference of specific postures and motions, making it possible to identify and classify human activities based on radar signals alone.

In radar-based human pose estimation approaches, the head network predominantly employs heatmap-based methodologies to predict human poses. This technique involves marking the probable locations of various human joints on predictive heatmaps, offering an intuitive representation of the spatial distribution and confidence levels of key-points. However, heatmap-based methodologies are constrained by heatmap resolution, which affects the precision of joint localization. Higher resolutions improve accuracy but increase computational demands, which is particularly challenging for real-time applications. Limitations in resolution may also obscure fine movements, necessitating a balance between precision and performance.

To address these challenges, our paper introduces a novel approach, SCRP-Radar (space-aware coordinate representation for human pose estimation based on SISO UWB radar), which utilizes the Simcc method [14] for human pose estimation from the micro-Doppler signature of human motion. By incorporating the Simcc method, which focuses on transforming the regression problem of human key-point detection into separate classification tasks for the x and y axes, we introduce the innovative SCRP-Radar approach to radar-based human pose estimation. This pioneering methodology redefines pinpointing human key-points by categorizing them into two distinct classification tasks: one for vertical coordinates and the other for horizontal coordinates. This refined approach significantly enhances accuracy and provides a more nuanced perspective on human pose estimation within radar systems.

The main contributions of our work can be summarized as follows.

(1) We introduce a novel benchmark for human pose estimation using UWB radar, named HPSUR. This benchmark, recorded using a SISO UWB radar system and N3 motion capture system, encompasses a comprehensive dataset comprising three rooms and two halls in a living room configuration. The dataset includes 311,963 frames, featuring five subjects of diverse heights and weights, each performing four different types of actions.

(2) We propose a new approach to coordinate representation for human pose estimation utilizing SISO UWB radar. This approach separates the representation of key-point x and y coordinates into individual 1D vectors, allowing us to treat key-point localization as separate classification tasks in the vertical and horizontal directions.

(3) We introduce the Hrnet and LiteHrnet models as the foundational backbones for the SCRP-Radar framework. This approach begins with a high-resolution subnetwork at the initial stage, progressively incorporating parallel subnetworks of lower resolutions. Such a configuration allows for the nuanced processing of micro-Doppler features across different scales and velocities, adeptly representing human motion's dynamic and intricate patterns, as captured by radar signals.

The remainder of this paper is structured as follows: Section 2 delves into the existing literature on human pose estimation and coordinate representation. Section 3 outlines the theoretical framework, encompassing the geometric modeling of human targets and radar systems, the micro-Doppler characteristics of human posture, and the structural information inherent in human models. Section 4 is dedicated to introducing the architecture of the proposed SCRP-Radar network. Section 5 presents quantitative and qualitative assessments of the proposed method, utilizing the HPSUR dataset. Finally, Section 6 offers conclusions from the research and outlines potential future directions.

2. Related Works

2.1. Human Pose Estimation

Human pose estimation (HPE) is a fundamental task in computer vision with critical implications across various applications. Accurately estimating human posture is essential in inferring specific behaviors, especially in remote monitoring. However, this task is complex due to the intricate mechanics of the human body and accompanying constraints. Research in human pose estimation can be classified into two primary approaches: camera-based and wireless sensing-based.

Camera-based HPE. Deep learning has revolutionized the camera-based HPE approach, setting new benchmarks for accuracy and efficiency in this domain. A notable example is DeepPose [15], which pioneered the application of deep learning techniques in this field. DeepPose employs an iterative architecture, extracting image features using cascaded convolutional neural networks and regressing the joint coordinates with fully connected layers. Xiao et al. [16] proposed a baseline method that predicts the heatmap by adding several deconvolutional layers to a backbone network. Building upon this, Sun et al. [17] introduced the Hrnet model, which maintains high-resolution representations throughout the heatmap estimation process.

Wireless sensing-based HPE. The exploration of wireless sensing technologies marks a significant departure from traditional camera-based methods, offering novel approaches that leverage the omnipresence of wireless signals. This research segment capitalizes on the ability of wireless signals, such as WiFi, RFID, and other radio frequencies, to penetrate occlusions and operate in non-line-of-sight conditions, thus overcoming some of the intrinsic limitations of visual sensors. Wang et al. [18] developed a deep learning methodology that utilizes annotations on 2D images, processes received 1D WiFi signals, and achieves end-to-end HPE. However, these solutions are typically limited to capturing poses from a single perspective or constructing poses of individuals at a stationary location, which hinders their broader applicability in everyday scenarios. Lu et al. introduced Wi-Pose [19] and Wi-Mose [20]. This system derives skeletons from synchronized video frames as supervision for WiFi signals and employs a novel neural network to obtain detailed human skeleton images. Mao et al. introduced the RFID-Pose system [7], marking the first instance of 3D HPE using standard RFID readers and tags. This system is adept at monitoring multiple human joints concurrently in real time. However, the bandwidth range of WiFi results in a too-coarse resolution for capturing fine-grained human poses, and RFID systems have a low data rate, which makes generating a comprehensive joint confidence map for all joints as arduous as in other RF-based systems.

In light of this challenge, skeletal estimation utilizing radar devices represents a burgeoning area of research. Radar-based devices can be broadly categorized into two groups: high-frequency radars, such as millimeter-wave (mmWave) or terahertz radars [21,22], and lower-frequency radars, operating around a few GHz. Studies [23–25] have leveraged mmWave radar's reflection signals, combined with convolutional neural networks, to estimate the positions of distinct joints in the human body. Chen et al. [26] innovated a domain discriminator that filters user-specific characteristics from mmWave signals, enabling robust skeleton reconstruction across users with minimal training effort. Naim Dahnoun et al. [27] designed a novel neural network model for HPE based on point cloud data, comprising a part detector for initial key-point positioning and a spatial model that refines these estimates by learning joint relationships.

Conversely, low-frequency radar offers several benefits: it can penetrate walls and obstructions, function effectively in both daylight and darkness, and is inherently more privacy-preserving due to its non-interpretability by humans. Pioneering work by MIT researchers [8,9,28] introduced a neural network system that interprets radar signals for 2D human pose and dynamic 3D human mesh estimations. Tian Jin et al. [29] developed a novel through-wall 3D pose reconstruction method using UWB MIMO radar and 3D CNNs. Guangyong Fang et al. [30] proposed a cross-modal CNN-based method for postural reconstruction through wall radar imaging (TWRI). Choi et al. [31] introduced the 3D-TransPose algorithm for 3D HPE, leveraging an attention mechanism to focus on relevant periods in time-domain IR-UWB radar signals. Nevertheless, these approaches rely on MIMO radar imaging, and the variances between the human target and the surrounding environment can significantly impact the quality of radar imaging. Our work employs SISO FMCW UWB radar for HPE, capitalizing on the micro-Doppler signature, which represents an innovative convergence of radar signal processing and pattern recognition techniques, targeting the dynamic and nuanced task of discerning human postures and movements.

2.2. Coordinate Representation

Accurately modeling and predicting human joints and limb positions depends on the representation of coordinates [32]. The method used to encode the spatial locations of crucial body parts in the input data is called coordinate representation. This foundational aspect of pose estimation significantly impacts the effectiveness of the estimation process, influencing both the precision of the pose inference and the efficiency of the computational models. Traditionally, three main types of coordinate representation have been prevalent: heatmap-based, regression-based, and Simcc-based.

Heatmap-based. Heatmap-based representations involve creating a 2D probability map for each joint, which indicates the likelihood of each pixel being the location of that joint. This approach provides a more detailed representation of uncertainty. It is commonly used with convolutional neural networks (CNNs) owing to its compatibility with the spatial processing strengths of CNNs. In [33,34], Gaussian-smoothed heatmaps are constructed by assigning higher values to the pixels closer to the dot annotation than those farther away.

Regression-based. Regression-based representations aim to predict the numerical values of joint coordinates directly. This method often requires less computational overhead than heatmap-based methods and can be more straightforward. The authors of [35] proposed a novel regression paradigm called residual log-likelihood estimation (RLE) to capture the underlying output distribution. In [14], a new Simcc-based coordinate coding scheme is introduced, which represents the coordinate estimation task as two classification tasks of horizontal and vertical coordinates.

3. Theory

3.1. Geometry Model of the Human Target and Radar

The geometric relationship between the transmitting and receiving antennas of the SISO UWB radar is shown in Figure 1. The geometric motion relationship between radar and moving human target is shown in Figure 2. The coordinate system (U, V, W) is the global coordinate system, Tx is the position of the radar transmitting antenna, and Rx is the position of the radar receiving antenna, where Tx $= (0, 0, 0)^T$, Rx $=(u_1, v_1, w_1)^T$. The reference coordinate system is (X, Y, Z) parallel to the global coordinate system, and the origin of the coordinate is Tx. The target coordinate system is (x, y, z), and the origin is O, as is the reference coordinate system. The initial position vector of the origin O in the global coordinate system is $\mathbf{R}_o = (U_o, V_o, W_o)^T$, and the initial azimuth angle and elevation angle are defined as α, β, respectively. Furthermore, the radial unit vector extending from the radar toward the target is defined as

$$\mathbf{n} = \mathbf{R_o} / \|\mathbf{R_o}\| = (\cos\alpha\cos\beta, \sin\alpha\cos\beta, \sin\beta)^{\mathrm{T}} \tag{1}$$

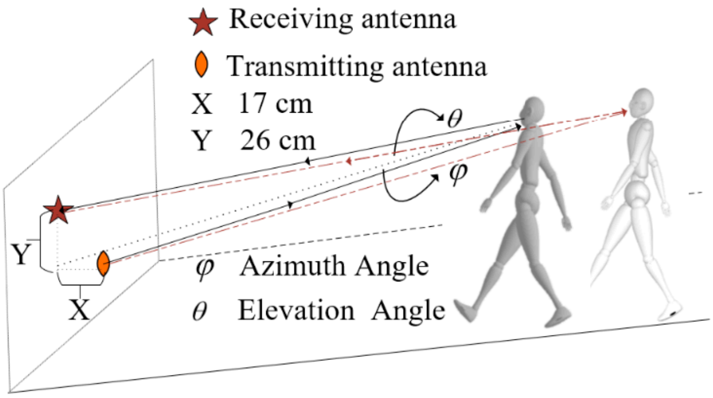


Figure 1. Configuration of SISO UWB radar for human pose estimation showing the relative positioning of the transmitting and receiving antennas, and the azimuth and elevation angles to the moving human target.

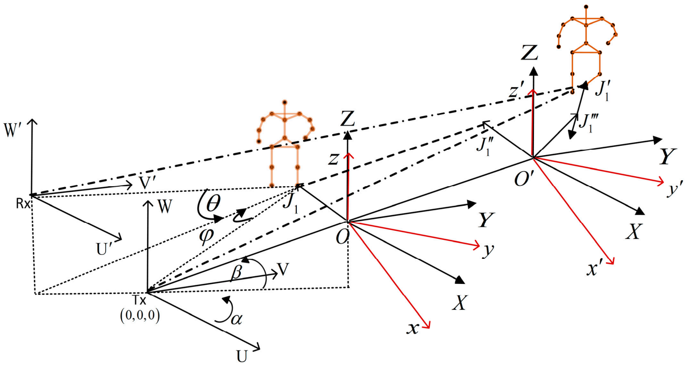


Figure 2. The geometric relationship between human motion model and radar.

Assume that the position of the left foot bone of the moving human target at the initial time $t = 0$ is designed as J_1, and the position vector in the global coordinate system is $\mathbf{r_o} = (X_o, Y_o, Z_o)^{\mathrm{T}}$. During the observed period, point J_1 undergoes four simultaneous movements characterized by their distinct kinematic properties.

1. The skeleton translates with speed $\mathbf{V}$ in the radar coordinate system;
2. The skeleton accelerates with acceleration $\mathbf{a}$;
3. The skeleton vibrates sinusoidally with frequency $\mathbf{f}_v$ and amplitude $\mathbf{D}_v$. The azimuth angle and pitch angle are $\boldsymbol{\alpha}_p$, $\boldsymbol{\beta}_p$, respectively, and the unit vector of the vibration direction is $\mathbf{n}_v = (cos\alpha_p cos\beta_p,\ sin\alpha_p cos\beta_p,\ sin\beta_p)^{T\cdot}$;
4. The skeleton rotates in the reference coordinate system with an angular velocity of $\boldsymbol{\omega} = (\omega_X, \omega_Y, \omega_z)^T$. At time t, the J_1 skeleton point moves to the new position J_1''.

Then, the distance from the radar transmitting antenna to joint J_1''' at time t is:

$$\begin{aligned} \mathbf{R_{tx}}(t) = \overline{\mathbf{TxJ_1}} &= \mathbf{R}_o + \mathbf{r}_o + \overline{\mathbf{J_1J_1''}} + \overline{\mathbf{J_1''J_1'''}} + \overline{\mathbf{J_1'''J_1'}} \\ &= \mathbf{R_o} + \mathbf{r_o} + \mathbf{V}t + 1/2\mathbf{a}t^2 + \mathbf{Rot}(t) \cdot \overline{\mathbf{O'J_1''}} + D_v \sin(2\pi f_v t) \cdot \mathbf{n}_v \\ &= \mathbf{R_o} + \mathbf{r_o} + \mathbf{V}t + 1/2\mathbf{a}t^2 + \mathbf{Rot}(t) \cdot \mathbf{r_o} + D_v \sin(2\pi f_v t) \cdot \mathbf{n}_v \end{aligned} \quad (2)$$

Then, the distance from the radar receiving antenna to joint J_1''' at time t is:

$$\begin{aligned} \mathbf{R_{Rx}}(t) = \overline{\mathbf{RxJ_1}} &= \mathbf{R}_x + \mathbf{R}_o + \mathbf{r}_o + \overline{\mathbf{J_1J_1''}} + \overline{\mathbf{J_1''J_1'''}} + \overline{\mathbf{J_1'''J_1'}} \\ &= \mathbf{R}_x + \mathbf{R_o} + \mathbf{r_o} + \mathbf{V}t + 1/2\mathbf{a}t^2 + \mathbf{Rot}(t) \cdot \overline{\mathbf{O'J_1''}} + D_v \sin(2\pi f_v t) \cdot \mathbf{n}_v \\ &= \mathbf{R}_x + \mathbf{R_o} + \mathbf{r_o} + \mathbf{V}t + 1/2\mathbf{a}t^2 + \mathbf{Rot}(t) \cdot \mathbf{r_o} + D_v \sin(2\pi f_v t) \cdot \mathbf{n}_v \end{aligned} \quad (3)$$

The sum of the distances from the joint point J_1 to the transmitting antenna and the receiving antenna at moment t is:

$$\mathbf{R}(t) = \mathbf{R_{tx}}(t) + \mathbf{R_{Rx}}(t) \tag{4}$$

Then, the distance from the radar to the J_1' joint at moment t is:

$$\begin{aligned} R(t) &= \|\mathbf{R}(t)\| = \|\mathbf{R_{tx}}(t)\| + \|\mathbf{R_{Rx}}(t)\| \\ &= \|\mathbf{R_o} + \mathbf{r_o} + \mathbf{V}t + 1/2\mathbf{a}t^2 + \mathbf{Rot}(t)\cdot\mathbf{r_o} + D_v\sin(2\pi f_v t)\cdot\mathbf{n}_v\| \\ &+ \|\mathbf{R}_x + \mathbf{R_o} + \mathbf{r_o} + \mathbf{V}t + 1/2\mathbf{a}t^2 + \mathbf{Rot}(t)\cdot\mathbf{r_o} + D_v\sin(2\pi f_v t)\cdot\mathbf{n}_v\| \end{aligned} \tag{5}$$

where $\boldsymbol{\omega}' = \frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|} = \left(\omega_X', \omega_Y', \omega_Z'\right)^{\mathrm{T}}$, $\Omega = \|\omega\|$, $\hat{\boldsymbol{\omega}} = \begin{bmatrix} 0 & -\omega_Z & \omega_Y \\ \omega_Z & 0 & -\omega_X \\ -\omega_Y & \omega_X & 0 \end{bmatrix}$, $\hat{\boldsymbol{\omega}}' = \begin{bmatrix} 0 & -\omega_Z' & \omega_Y' \\ \omega_Z' & 0 & -\omega_X' \\ -\omega_Y' & \omega_X' & 0 \end{bmatrix}$, and the rotation matrix $\mathbf{Rot}(t)$ can be expressed as:

$$\mathbf{Rot}(t) = I + \hat{\boldsymbol{\omega}}'\sin(\Omega t) + \hat{\boldsymbol{\omega}}'^2(1-\cos(\Omega t)) = \exp(\hat{\boldsymbol{\omega}}t) \tag{6}$$

The baseband signal of the radar echo can be expressed as:

$$s(t) = \rho(x,y,z)\exp\left\{\mathrm{j}2\pi f\frac{R(t)}{c}\right\} = \rho(x,y,z)\exp\{\mathrm{j}\Phi(R(t))\} \tag{7}$$

where $\Phi(R(t)) = \frac{2\pi f R(t)}{c}$.

Derivation of the phase function $\Phi(R(t))$ yields the Doppler frequency of the echo f_d.

$$\begin{aligned} f_d &= \tfrac{1}{2\pi}\tfrac{\mathrm{d}\Phi(R(t))}{\mathrm{d}t} = \tfrac{f}{c}\tfrac{\mathrm{d}R(t)}{\mathrm{d}t} \\ &= \tfrac{f}{c}\tfrac{\mathrm{d}(R_{tx}(t)+R_{Rx}(t))}{\mathrm{d}t} \\ &= \tfrac{2f}{c}\mathbf{V}^{\mathrm{T}}\cdot\mathbf{n}_{p'} + \tfrac{2f}{c}\left(\mathbf{a}^{\mathrm{T}}\cdot\mathbf{n}_{p'}\right)t + \tfrac{2f}{c}\tfrac{\mathrm{d}}{\mathrm{d}t}(\mathbf{Rot}(t)\cdot\mathbf{r_0})^{\mathrm{T}}\cdot\mathbf{n}_{p'} + \tfrac{4f}{c}\pi f_v D_v\cos(2\pi f_v t)\cdot\mathbf{n}_v^{\mathrm{T}}\cdot\mathbf{n}_{p'} \end{aligned} \tag{8}$$

Noting $\mathbf{r} = \mathrm{Rot}(t)\cdot\mathbf{r_0}$, combining $\omega \times r = \hat{\omega}\cdot r$ and $\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{Rot}(t)) = \frac{\mathrm{d}}{\mathrm{d}t}(\exp(\hat{\boldsymbol{\omega}}t)) = \hat{\boldsymbol{\omega}}\cdot\exp(\hat{\boldsymbol{\omega}}t)$, the above equation can be expressed as:

$$f_d = \frac{2f}{c}\mathbf{V}^{\mathrm{T}}\cdot\mathbf{n}_{p'} + \frac{2f}{c}\left(\mathbf{a}^{\mathrm{T}}\cdot\mathbf{n}_{p'}\right)t + \frac{2f}{c}(\boldsymbol{\omega}\times\mathbf{r})^{\mathrm{T}}\cdot\mathbf{n}_{p'} + \frac{4f}{c}\pi f_v D_v\cos(2\pi f_v t)\cdot\mathbf{n}_v^{\mathrm{T}}\cdot\mathbf{n}_{p'} \tag{9}$$

When $n = R_0/\|R_0\|$ is used as an approximation instead of $\mathbf{n}_{p'}$, the above equation can be written in the following form:

$$f_d = \frac{2f}{c}\mathbf{V}^{\mathrm{T}}\cdot\mathbf{n} + \frac{2f}{c}\left(\mathbf{a}^{\mathrm{T}}\cdot\mathbf{n}\right)t + \frac{2f}{c}(\boldsymbol{\omega}\times\mathbf{r})^{\mathrm{T}}\cdot\mathbf{n} + \frac{4f}{c}\pi f_v D_v\cos(2\pi f_v t)\cdot\mathbf{n}_v^{\mathrm{T}}\cdot\mathbf{n} \tag{10}$$

The human left ankle joint's micro-Doppler is:

$$f_{m-d} = \frac{2f}{c}\left(\mathbf{a}^{\mathrm{T}}\cdot\mathbf{n}\right)t + \frac{2f}{c}(\boldsymbol{\omega}\times\mathbf{r})^{\mathrm{T}}\cdot\mathbf{n} + \frac{4f}{c}\pi f_v D_v\cos(2\pi f_v t)\cdot\mathbf{n}_v^{\mathrm{T}}\cdot\mathbf{n} \tag{11}$$

However, only the modulation characteristics of human motion frequency caused by acceleration and vibration can be obtained from the above formula. In order to better understand the modulation characteristics of rotating motion on frequency, the relevant parameters of the moving human target are set in the target coordinate system. Suppose at time t $= 0$, the position vector of the joint point J_1 of the human target in the target coordinate system is $\mathbf{r_0} = (x_0, y_0, z_0)^{\mathrm{T}}$, and then rotates in the target coordinate system with the angular velocity $\boldsymbol{\omega}_l = \left(\omega_x, \omega_y, \omega_z\right)^{\mathrm{T}}$, and (ϕ, θ, ψ) represents the initial Euler angles. The initial rotation matrix is represented by $\mathbf{R}_{\text{init}}$:

$$\mathbf{R}_{\text{init}} = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{12}$$

Noting $\boldsymbol{\omega}^l_{\prime} = \frac{\mathbf{R}_{\text{init}} \cdot \boldsymbol{\omega}_l}{\|\boldsymbol{\omega}_l\|} = \left(\omega'_x, \omega'_y, \omega'_z\right)^{\mathrm{T}}$, $\hat{\boldsymbol{\omega}}_l = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}$, $\hat{\boldsymbol{\omega}}'_l = \begin{bmatrix} 0 & -\omega'_x & \omega'_y \\ \omega'_z & 0 & -\omega'_x \\ -\omega'_y & \omega'_x & 0 \end{bmatrix}$, and $\Omega_l = \|\omega_l\|$, the rotation matrix is still represented by $\mathbf{Rot}(t)$:

$$\mathbf{Rot}(t) = I + \hat{\boldsymbol{\omega}}'_l \sin(\Omega_l t) + \hat{\boldsymbol{\omega}}'^2_l (1 - \cos(\Omega_l t)) = \exp(\hat{\boldsymbol{\omega}}_l t) \tag{13}$$

Then, the distance from the radar to the joint J'_1 at moment t is:

$$\begin{aligned} R(t) &= \|\mathbf{R}(t)\| = \|\mathbf{R_{tx}}(t)\| + \|\mathbf{R_{Rx}}(t)\| \\ &= \|\mathbf{R_o} + \mathbf{r_o} + \mathbf{V}t + 1/2\mathbf{a}t^2 + \mathbf{Rot}(t) \cdot \mathbf{R}_{init} \cdot \mathbf{r_o} + D_v \sin(2\pi f_v t) \cdot \mathbf{n}_v\| \\ &+ \|\mathbf{R}_x + \mathbf{R_o} + \mathbf{r_o} + \mathbf{V}t + 1/2\mathbf{a}t^2 + \mathbf{Rot}(t) \cdot \mathbf{R}_{init} \cdot \mathbf{r_o} + D_v \sin(2\pi f_v t) \cdot \mathbf{n}_v\| \end{aligned} \tag{14}$$

Derivation of the phase function yields the Doppler frequency f_d of the echo:

$$\begin{aligned} f_d &= \frac{1}{2\pi} \frac{\mathrm{d}\Phi(R(t))}{\mathrm{d}t} = \frac{f}{c} \frac{\mathrm{d}R(t)}{\mathrm{d}t} \\ &= \frac{f}{c} \frac{\mathrm{d}(R_{tx}(t) + R_{Rx}(t))}{\mathrm{d}t} \\ &= \frac{2f}{c} \mathbf{V}^{\mathrm{T}} \cdot \mathbf{n}_{p'} + \frac{2f}{c} \left(\mathbf{a}^{\mathrm{T}} \cdot \mathbf{n}_{p'}\right) t + \frac{2f}{c} \frac{\mathrm{d}}{\mathrm{d}t} (\mathbf{Rot}(t) \cdot \mathbf{R}_{\text{init}} \cdot \mathbf{r}_0)^{\mathrm{T}} \cdot \mathbf{n}_{p'} \\ &+ \frac{4f}{c} \pi f_v D_v \cos(2\pi f_v t) \cdot \mathbf{n}_v{}^{\mathrm{T}} \cdot \mathbf{n}_{p'} \end{aligned} \tag{15}$$

Noting $\mathbf{r} = \mathbf{Rot}(t) \cdot \mathbf{R}_{\text{init}} \cdot \mathbf{r_0}$, with $n = R_0 / \|R_0\|$ used as an approximation instead of $\mathbf{n}_{p'}$, at the time of human movement, the target joint of the micro-Doppler f_{m-d} is as follows:

$$\begin{aligned} f_{m-d} &= \frac{2f}{c} (\mathbf{a}^{\mathrm{T}} \cdot \mathbf{n}) t + \frac{2f}{c} (\Omega_l \boldsymbol{\omega}'_l \times \mathbf{r})^{\mathrm{T}} \cdot \mathbf{n} + \frac{4f\pi f_v D_v}{c} \cos(2\pi f_v t) \cdot \mathbf{n}_v{}^{\mathrm{T}} \cdot \mathbf{n} \\ &= \frac{2f}{c} (\mathbf{a}^{\mathrm{T}} \cdot \mathbf{n}) t + \frac{2f}{c} (\Omega_l \hat{\boldsymbol{\omega}}'_l \cdot \mathbf{Rot}(t) \cdot \mathbf{R}_{init} \cdot \mathbf{r_0})^{\mathrm{T}} \cdot \mathbf{n} + \frac{4f\pi f_v D_v}{c} \cos(2\pi f_v t) \cdot \mathbf{n}_v{}^{\mathrm{T}} \cdot \mathbf{n} \\ &= \frac{2f}{c} (\mathbf{a}^{\mathrm{T}} \cdot \mathbf{n}) t + \frac{2f}{c} \left(\Omega_l \left[\hat{\boldsymbol{\omega}}'^2_l \sin(\Omega_l t) - \hat{\boldsymbol{\omega}}'^3_l \cos(\Omega_l t) + \hat{\boldsymbol{\omega}}'_l (I + \hat{\boldsymbol{\omega}}'^2_l)\right] \mathbf{R}_{init} \cdot \mathbf{r_0}\right)^{\mathrm{T}} \cdot \mathbf{n} \\ &+ \frac{4f\pi f_v D_v}{c} \cos(2\pi f_v t) \cdot \mathbf{n}_v^{\mathrm{T}} \cdot \mathbf{n} \end{aligned} \tag{16}$$

The formula presented above indicates that when the target simultaneously exhibits translation, acceleration, vibration, and rotation characteristics, the parameter f_{m-d} will undergo linear modulation. This modulation in frequency is directly proportional to the acceleration of the target. It exhibits a periodic variation over time, with the cycle period influenced by both the vibration and rotation periods. Furthermore, the amplitude of these changes depends on the vibration frequency, vibration amplitude, and rotational angular velocity.

3.2. The Micro-Doppler of the Human Posture

Researchers commonly model the subject as a combination of interconnected rigid segments when analyzing radar scattering from nonrigid body motion. This approach simplifies nonrigid body motion by breaking it down into the movements of several rigid bodies. Human motion, noted for its high degree of articulation and flexibility, presents a complex and intriguing case of micro-motion. The micro-Doppler (MD) signatures, presented in a combined time-frequency domain, introduce an invaluable additional time dimension, which facilitates the examination of evolving MD signatures linked to targets' rotating or vibrating components. These signatures, which illustrate the kinematics of a target's motion, act as unique identifiers and provide deeper insights into the target's movements. Fahad Jibrin Abdu et al. proposed an efficient CCA-based feature fusion algorithm that effectively combines multi-deep CNN features of radar MD spectrograms [36]. Shahid Hassan et al. crafted a method for classifying human activity based on micro-Doppler and interferometric micro-motion signatures using a DCNN classifier [37].

The generation of MD frequency is a result of the echo frequency modulation caused by various movements, including vibration, rotation, and other motions of moving objects. The traditional Fourier transform analysis, which is often used for frequency analysis, requires improvements to capture this dynamic frequency information adequately. Yuan He and Francesco Fioranelli et al. suggested an instance-based transfer learning approach for recognizing human motion with radar using limited training data [38]. Peng Li et al. developed a novel spatiotemporal weighted MD spectrum that accounts for the primary and secondary importance information in the reduced dimension space and the temporal information of the sequences [39].

The radar data preprocessing chain chart is shown in Figure 3, which outlines the process from raw radar data to the final output. The first steps in this process involve applying clutter suppression and noise reduction techniques to boost the signal-to-noise ratio and reduce unwanted interference. Next, the fast Fourier transform (FFT) is used along the fast-time dimension of the raw radar data to produce a range bin. Then, the FFT is applied along the slow-time dimension to create multiple Range-Doppler maps, as illustrated in Figure 3, referred to as Range-FFT and Doppler-FFT. The Range-FFT calculates the target distance for each chirp in the original data matrix, while the Doppler-FFT helps determine target velocities for each distance unit. Each element in the resulting Range-Doppler maps is known as a "Range Bin" and is represented in the frequency domain and expressed in decibels. The next step involves summing the Range Bins along the range axis for each Range-Doppler map, generating a vector **e** comprising L Doppler Bins. Finally, concatenating n consecutive frames produces a time-length n-frame micro-Doppler signature map, which is also known as a time-Doppler spectrogram.

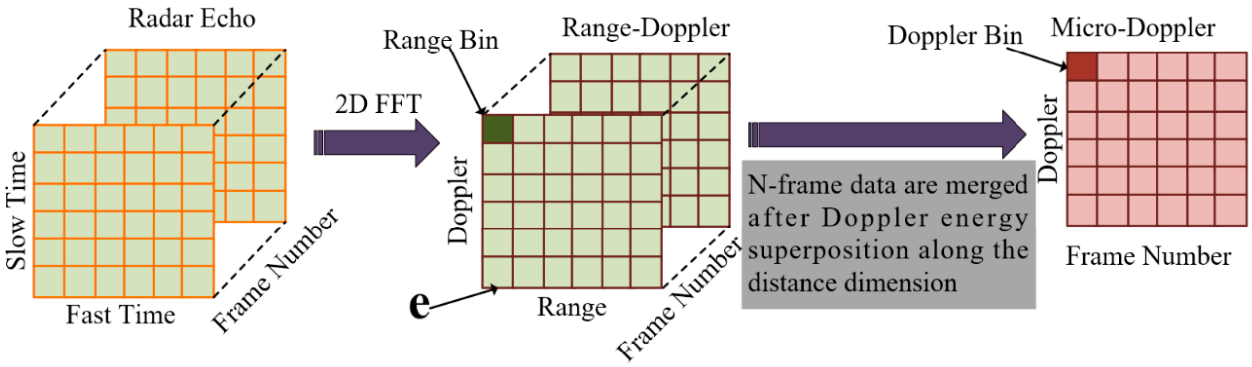
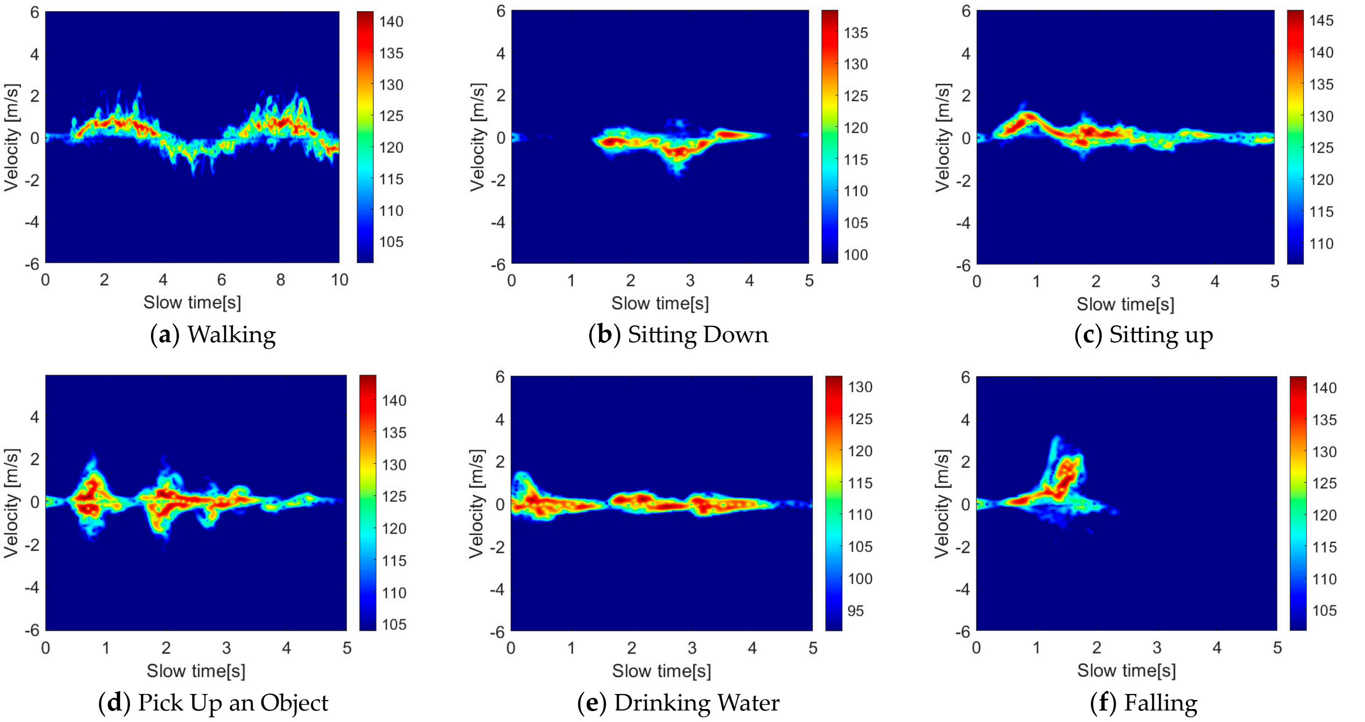


Figure 3. Radar data preprocessing chain chart.

Additionally, window functions are employed during signal processing to mitigate spectral leakage and related issues. Our analysis transformed the radar signals of human activities into the micro-Doppler spectrum using the UOG dataset [12], which comprises over 1700 radar signatures from six types of human activities. Figure 4 displays the micro-Doppler spectrum for these activities, where the intense yellow and red zones indicate the Doppler frequency range associated with the human torso. Meanwhile, the peripheral pale-yellow regions represent the micro-Doppler signals produced by the movement of human limbs.



(**a**) Walking (**b**) Sitting Down (**c**) Sitting up

(**d**) Pick Up an Object (**e**) Drinking Water (**f**) Falling

Figure 4. Micro-Doppler spectra for six different human activities.

Our paper employed an FFT size of 512, a frame number of 500 frames, and a Hamming window for radar signal processing. Notably, our use of ultra-wideband radar necessitated a pulse repetition frequency (PRF) of 960 frames per second, and we specifically selected 500 frames for analysis. This choice equates to each micro-Doppler representation encapsulating data spanning 0.52 s. The rationale behind this decision lies in the context of human pose estimation, where we aim to estimate the coordinates of skeletal points corresponding to a specific moment in time. Given that human movement cycles, such as walking or other dynamic activities, typically exhibit periods of 2 to 5 s, our selection of 500 frames per micro-Doppler instance ensures better extraction of micro-Doppler features from radar echoes, providing a comprehensive representation of the motion characteristics associated with the human body at that particular moment in time. This parameterization aligns with our objective of capturing meaningful and temporally relevant information for accurate human pose estimation using radar signals. The mapped micro-Doppler spectrum is used as input to the subsequent network to estimate human poses owing to the distinct features of micro-Doppler signals and leveraging insights from deep learning.

We simplify the intricate Boulic human body model in this paper, initially featuring 62 degrees of freedom across 32 joints, into a more manageable framework that includes 13 standard rigid bodies and 17 nodes, as illustrated in Figure 5. These 13 segments effectively capture the human body's complexity, representing the head, shoulders, upper arms, forearms, thighs, lower legs, and upper and lower torso. The 17 nodes identified within this configuration are pinpointed at crucial anatomical points such as the hips, upper and lower segments of both legs, feet, spine, head, shoulders, upper arms, forearms, and hands, ensuring comprehensive skeletal mapping. Moreover, the interaction between the foot and the floor is accurately depicted through a rigid contact model, which considers the shape of the foot's bottom and its orientation relative to the ground.

Based on experimental data, the proportionate lengths of different human body parts were calculated and are presented in Figure 5. According to the data, the segment from the top of the head to the bottom of the neck makes up 18.2% of the body's height, the shoulders constitute 25.9%, and the torso accounts for 28.8%. Additionally, the upper arms represent 18.8% of the length, the lower arms 14.5%, the thighs 24.5%, the calves 28.5%, and the hips 19.1%.

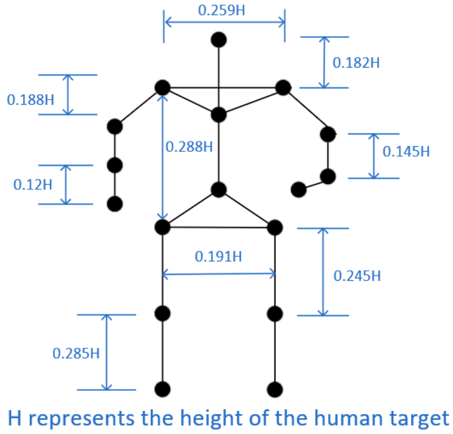


Figure 5. The structure information of the human model.

In order to examine how different parts of the human body interact during movement, we analyzed the trajectories of body segments from MOCAP data of a human subject at Carnegie Mellon University. Figure 5 displays the structural information of the human target segment, while Figure 6 shows the experimental setup of the simulation. In this setup, a human target, approximately 175cm, starts to fall face down in place at around 2 s, and the total data duration is roughly 5.5 s. We then extracted the corresponding MD spectrum from the radar echo data of the moving human body. Figure 7 reveals the Doppler frequency variations of different parts of the moving human body. The zero-frequency line in this figure represents the human body's torso. The human target was stationary in a standing position for 2 s and after 4.5 s. However, between 2 and 4.5 s, the MD frequency changes in the figure correspond to the movement dynamics of various parts of the moving human target. A higher MD frequency indicates a larger movement amplitude in that particular body part.

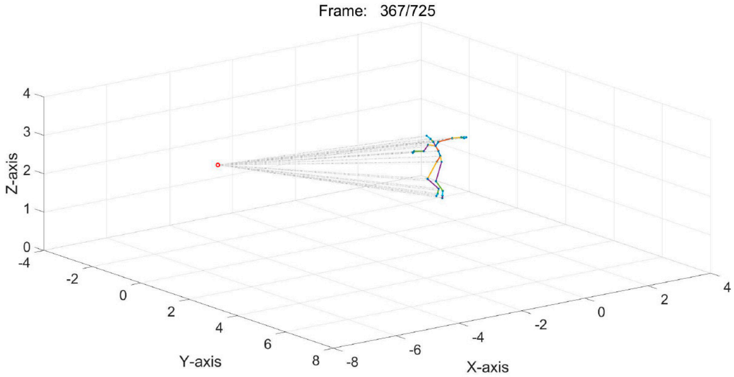


Figure 6. Simulated experimental scene.

The human body is an asymmetric, non-rigid structure with bilateral symmetry, and during human movement, it follows specific patterns caused by the MD effects. In Figure 8, the left and right sides of the human body are represented in the first and third rows and second and fourth rows, respectively, to facilitate a comparative analysis of the MD effects resulting from micro-movements in the left and right structures of the moving human subject. For instance, Figure 8c,g showcase the left and right arms resembling a left–right symmetrical structure. To maintain balance during most movements, the arms often exhibit symmetrical or reverse symmetrical movements centered around the trunk.

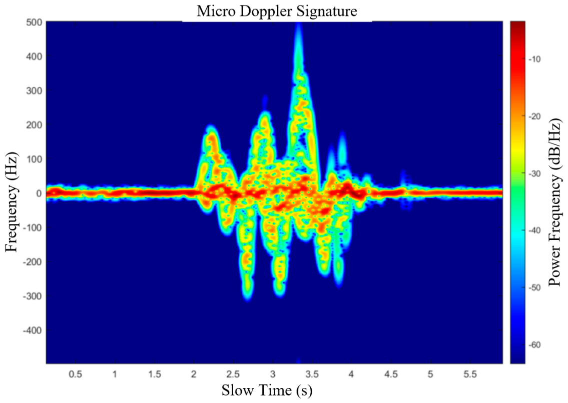
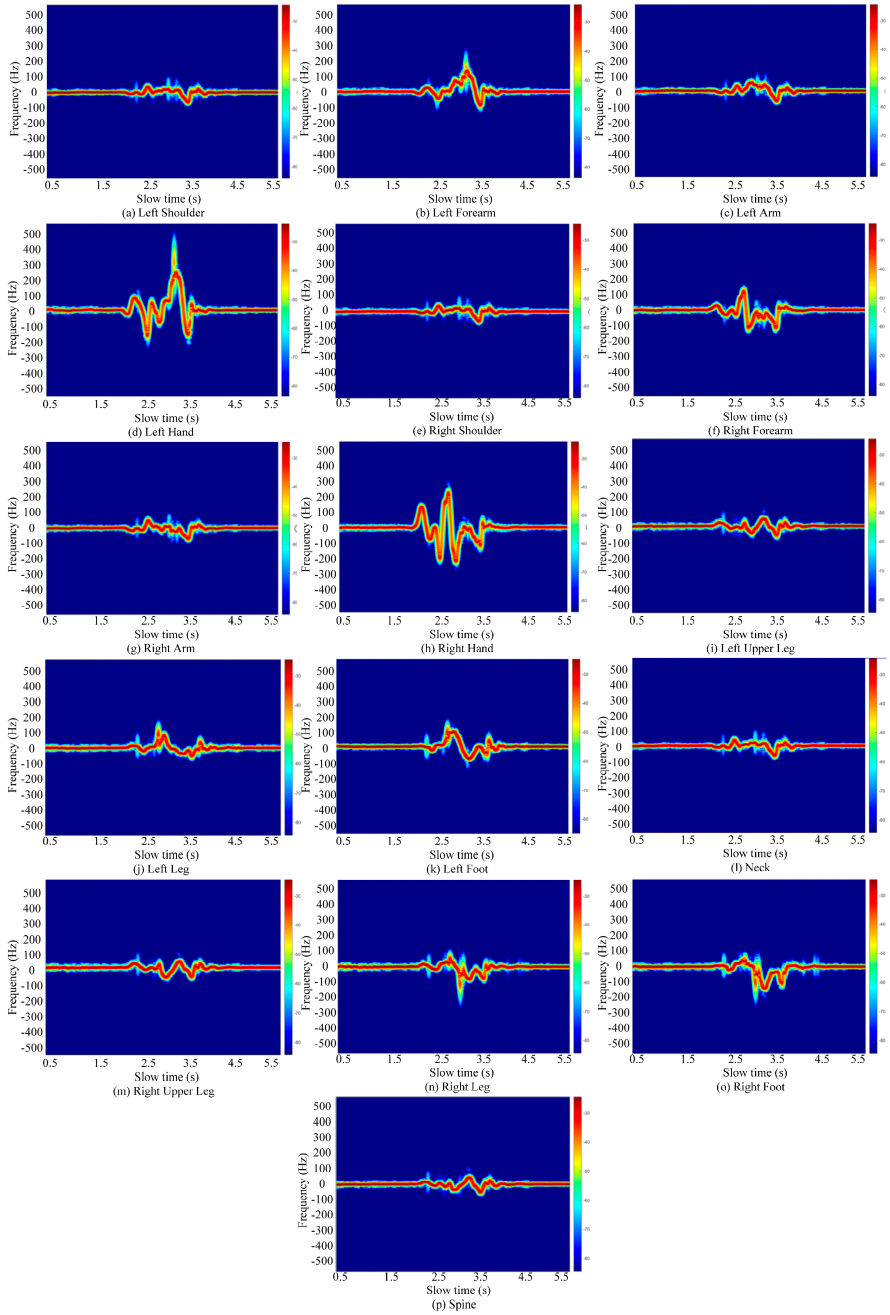


Figure 7. MD spectrum of the human body.



(a) Left Shoulder
(b) Left Forearm
(c) Left Arm
(d) Left Hand
(e) Right Shoulder
(f) Right Forearm
(g) Right Arm
(h) Right Hand
(i) Left Upper Leg
(j) Left Leg
(k) Left Foot
(l) Neck
(m) Right Upper Leg
(n) Right Leg
(o) Right Foot
(p) Spine

Figure 8. MD spectra for individual human body parts during movement analysis.

Both sides' upper and lower arms, thighs, and calves exhibit large motion amplitudes during human movement, resulting in higher MD frequencies. The corresponding MD spectrum mirrors the human body's inherent symmetrical structural characteristics. By simulating the MD effect differences caused by micro-motions of different human body parts during movement, we can more effectively demonstrate that MD spectra accurately reflect the characteristics inherent in various postural states of the human body. Therefore, the MD features of moving human subjects can be instrumental in addressing the challenge of human pose reconstruction.

4. Method

4.1. Experimental Setup and Data Collection

Our UWB radar system utilizes a single transmitter and receiver channel, as shown in Figure 9. This system is co-located with a inter realsense camera and the Noitom Perception Neuron 3 (N3) system, simultaneously capturing the subjects' radar returns and velocity information. To collect data comprehensively, we set up the UWB radar, a camera, the N3 device, and a personal computer (PC). The UWB radar acquires radar data, whereas the camera captures scene imagery. The N3 device employs a 2.4G wireless method to collect data on skeletal key-points of the human body. The PC synchronizes data across these three sensors, ensuring cohesive and aligned data capture for subsequent analysis.

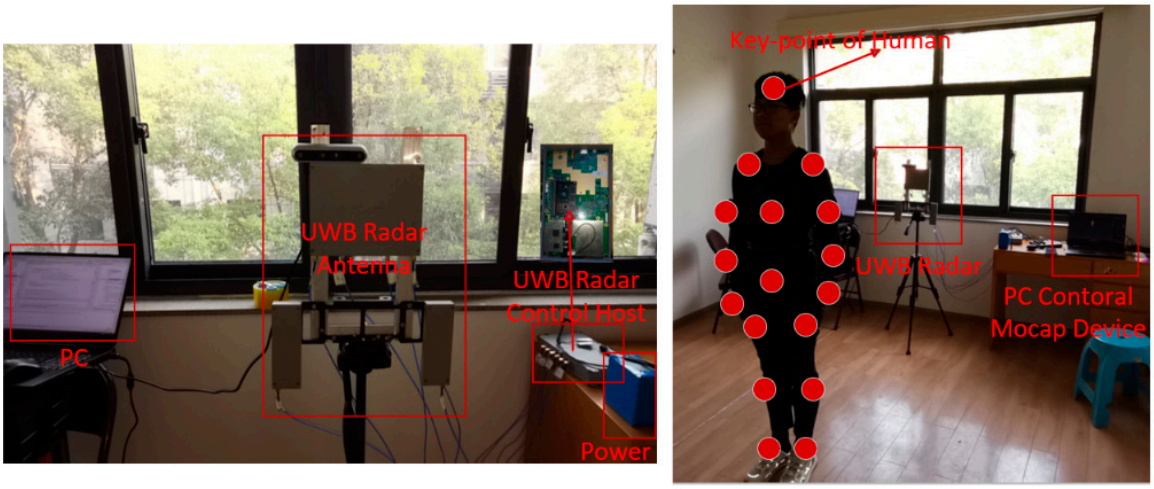


Figure 9. Experimental setup for UWB Radar Data acquisition.

The research involved collecting experimental data in a living room with four indoor movement scenarios. The radar was mounted on a tripod about 1.2 m above the ground and paired with a camera placed above the radar antenna. Ground truth (GT) data were obtained using an N3 device, which is an inertial sensor-based motion capture system. The 17 skeletal key-points captured by the N3 were timestamped using universal time coordinates (UTC) to make associating them with radar frames and compiling the dataset easier.

This paper utilized a SISO UWB radar, specifically an FMCW radar that operates in the 2.7 to 3.2 GHz range and has a bandwidth of 500 MHz. Table 1 provides more details about the specific parameters of the FMCW radar used. The chosen frequency band for the radar provides a degree of penetration and high resolution, which is crucial for accurately determining the posture of the human target. Estimating human poses in indoor environments filled with desks, chairs, and debris can be challenging. High-frequency radars and cameras can have difficulty estimating poses through obstacles, leading to aliasing and loss of target pose visibility. A practical approach to independently estimate each joint of the human body in indoor settings is to use lower-frequency UWB radar. For this purpose, the compact SISO UWB radar is the ideal choice due to its suitability in indoor scenarios. As Table 1 shows, the SISO UWB radar has a bandwidth of 500 MHz and a range resolution of 0.33m. However, the limited data acquired from the single transmitter and receiver antenna pose a significant challenge in estimating the extended posture of the human target.

Table 1. The specification of SISO UWB radar system parameters.

Parameters	Values
Frequency	2.7 GHz~3.2 GHz
Bandwidth	500 MHz
Lambda	1.017×10^{-1}
Pulse width	4.4×10^{-4}
FM slope	1.136×10^{12}
Pulse repetition frequency (PRF)	1923
Sampling frequency	4 MHz
Sampling points	1460

4.2. Overview Architecture

The SCRP-Radar approach is a novel technique for human pose estimation that treats it as a dual classification task for vertical and horizontal positions. The aim is to minimize quantization errors by segmenting each micro-Doppler signature patch into several bins. This strategy is inspired by the techniques used in Simcc [14]. Traditional radar-based human pose estimation research relies on heatmaps or regression to determine joint coordinates. However, these processes demand substantial computational power owing to the larger size of radar echo data than conventional images. To address this issue, we introduce the SCRP-Radar method, which marks its debut as a more efficient alternative that significantly reduces the need for computational resources while enhancing the accuracy of human joint estimation.

The SCRP-Radar's overview architecture is demonstrated in Figure 10. The human pose estimation model processes the radar echo with clutter suppression, producing a micro-doppler signature, and then extracts features using a backbone network, which can be either a CNN-based or a Transformer-based network. Through the full convolutional layer, n key-point representations are extracted. The obtained key-point representation is flattened from (n, H', W') to $(n, H' \times W')$ for subsequent classification. The SCRP-Radar performs coordinate classification independently for the horizontal and vertical axes based on the n key-point representations to generate the final predictions. For the i-th key-point representation, the horizontal and vertical coordinate classifiers generate the i-th key-point predictions O_x^i and O_y^i, respectively, using only one linear layer for each classifier. The head and decoding networks can restore the feature map's resolution and learn from the loss between the predicted coordinates and the ground truth coordinate. Our model mainly focuses on converting the regression problem in the human pose estimation task into a classification problem so that the loss function can incorporate a classification loss with better properties than L2 (MSE) loss.

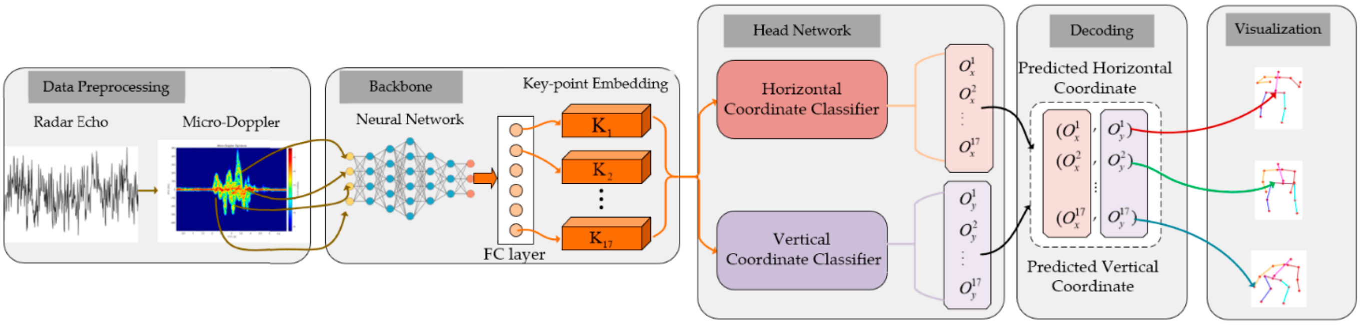


Figure 10. Overview of the SCRP-Radar framework for Human Pose Estimation (HPE) using SISO UWB Radar Data.

4.3. Efficient Feature Extraction Using Hrnet and LiteHrnet

The SCRP-Radar framework utilizes a High-Resolution Network (Hrnet) model as its backbone to process micro-Doppler features from radar echoes. The Hrnet model maintains high resolution throughout the process, allowing for detailed capture of micro-Doppler features that are essential for human pose estimation accuracy. The network starts with a high-resolution subnetwork and gradually adds lower-resolution subnetworks, enabling cross-resolution feature fusion. This design enables information to flow across different resolution pathways, ensuring that the network can process micro-Doppler features with varying scales and velocities, effectively capturing the dynamic and complex nature of human movements represented by radar signals. Using the Hrnet model as the backbone, the SCRP-Radar network can take advantage of the high-resolution representations of micro-Doppler features, resulting in precise human pose estimation.

The LiteHrnet model is a streamlined version of the Hrnet model that efficiently processes radar-derived micro-Doppler features for human pose estimation. It preserves high-resolution pathways like the Hrnet model but reduces computational costs and memory usage, making it ideal for resource-constrained settings. The LiteHrnet model combines lightweight design with competitive accuracy and a simplified feature fusion strategy, ensuring low computational overhead and suitability for real-time applications on limited-capacity devices. In subsequent experimental comparisons detailed in our paper, we rigorously evaluate the LiteHrnet model against the Hrnet model in the SCRP-Radar network architecture context.

4.4. Advancing Human Pose Estimation with Space-Aware Coordination

In SCRP-Radar, the x and y coordinates are represented as separate one-dimensional vectors instead of being encoded into a single representation. This disentanglement allows for independent manipulation of each coordinate, providing flexibility in handling spatial information. To encode the input's micro-Doppler features, we generate supervision signals in a space-aware way.

Coordinate encoding. Given a human motion's micro-Doppler signature of size $H \times W$, we denote the ground-truth coordinate for the p-th type of key-point as (x^p, y^p). To improve localization precision, we incorporate a splitting factor, denoted as k (where $k \geq 1$) and subsequently scale the ground-truth coordinates to obtain a refined coordinate system.

$$\mathbf{p}' = (x', y') = (\text{round}(x^p \cdot k), \text{round}(y^p \cdot k)) \tag{17}$$

where round(.) is a round function, the utilization of this splitting factor has the capability to elevate localization precision to sub-pixel levels. Additionally, the supervision signals are defined as follows:

$$\mathbf{p}'_{\mathbf{x\_sa}} = [x_0, x_1, \ldots, x_{W \cdot k-1}] \in \mathbb{R}^{W \cdot k}, x_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(i - x')^2}{2\sigma^2}\right) \tag{18}$$

$$\mathbf{p}'_{\mathbf{y\_sa}} = [y_0, y_1, \ldots, y_{H \cdot k-1}] \in \mathbb{R}^{H \cdot k}, y_j = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(j - y')^2}{2\sigma^2}\right) \tag{19}$$

where $i \in \{0, 1, \ldots, W \cdot k - 1\}, j \in \{0, 1, \ldots, H \cdot k - 1\}$, and σ is the standard deviation. Both p'_x and p'_y are one-dimensional vectors.

Coordinate decoding. Assuming the model generates two one-dimensional vectors O_x and O_y, corresponding to a specific type of human key-point, the predicted joint position $(\hat{O}_x, \hat{O}_y)$ is computed as follows:

$$\hat{o}_x = \frac{\text{argmax}_i(\mathbf{o}_\mathbf{x}(i))}{k}, \hat{o}_y = \frac{\text{argmax}_j(\mathbf{o}_\mathbf{y}(j))}{k} \tag{20}$$

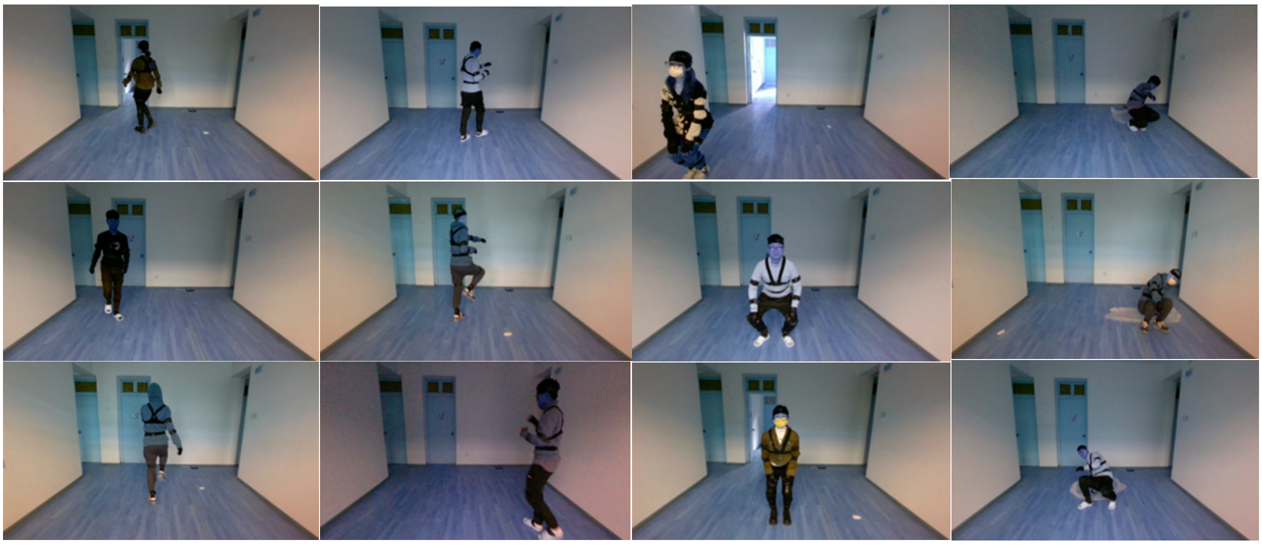
5. Results

5.1. Dataset and Annotations and Evaluation Metrics

Dataset and Annotations. We evaluated our model using the HPSUR dataset, which was created using a SISO UWB radar and N3 systems. The dataset comprises a comprehensive indoor environment, specifically a living room, featuring four distinct indoor movement scenarios, as detailed in Table 2. We collected 311,963 data frames from five subjects of varying height and weight. Each subject performed four types of actions, as shown in Figure 11, within a controlled visual environment. We divided the dataset into training and testing subsets to conduct our experiments. The training set included data from three subjects, totaling 189,462 frames. The testing set included data from the remaining two subjects, which totaled 122,401 frames. Ground truth (GT) for human pose key-points was acquired using the N3 system, which accurately captures 17 key-points of the human skeleton.

Table 2. Detailed descriptions of different human postures in the HPSUR dataset.

ID	Type of Posture	Specific Description
1001	Walking	Subjects walked back and forth along the radial path of the radar, including linear movements at both 45 degrees and 135 degrees to the radar's central axis.
1002	Punch forward	Subjects performed walking exercises along the radar's radial path and diagonally at 45 and 135 degrees, incorporating fist movements during the walk.
1003	Sit and stand	Subjects assumed sitting and standing postures at designated positions relative to the radar, specifically at (0 m, 2 m), (0 m, 3 m), (−1 m, 2 m), and (1 m, 3m), using the radar as the origin point.
1004	Fall	Subjects performed fall motion at the same coordinates as the sitting and standing postures, namely at (0 m, 2 m), (0 m, 3 m), (−1 m, 2 m), and (1 m, 3 m).



(**a**) Walking (**b**) Punch forward (**c**) Sit and stand (**d**) Fall

Figure 11. Various postures captured for the HPSUR dataset in a controlled environment.

Evaluation metric. We use the MPJPE metric to evaluate the accuracy of our estimated 2D human poses against the GT under the HPSUR dataset. The MPJPE calculates the average Euclidean distance between the estimated joint positions and their corresponding

GT counterparts. Since our research primarily focuses on human pose estimation in single-person scenarios, we also use the Percentage of Correct Key-points (PCK) as a metric to assess the effectiveness of our training process. The PCK measures the ratio of accurately estimated human key points, where a key-point is considered correct if the normalized distance between the estimated key-point and the GT is less than a predefined threshold. We set the threshold T_k to 0.7, as outlined in our experimental setup.

$$PCK_{\text{mean}}^{k} = \sum_{i} \delta\left(\frac{\|\mathbf{p}_i - \mathbf{g}_i\|_2}{h} \leq T_k\right) / \sum_{i} 1 \tag{21}$$

where i represents the i-th key-point, T_k represents the k-th threshold, $\mathbf{p}_i$ represents the predicted value of the i-th key point, $\mathbf{g}_i$ represents the ground truth of the i-th key-point, h represents the length of body, and T_k represents the manually set threshold; we choose 0.7. This means that the Euclidean distance between the predicted value of the key-point and the ground truth is less than or equal to T_k* (length of body), it is judged that the prediction of the key-point is correct. During the training process, the accuracy of all joint points of the batch size is calculated as an evaluation index, which is used to guide the network training.

5.2. Implementations Details

For our study on human pose estimation, we used a dataset called HPSUR to train and test our networks. This dataset contains UWB radar data from five volunteers: S1 (Male), S2 (Female), S3 (Male), S4 (Male), and S5 (Female). We used data from volunteers S1, S2, and S3 for the training phase, while data from volunteers S4 and S5 were reserved for the testing phase. We experimented with four network models: Hrnet-w32, Hrnet-w64, LiteHrnet-18, and LiteHrnet-30.

During the training phase, we optimized our networks using Adaptive Moment Estimation (Adam), which combines the benefits of the AdaGrad and RMSProp algorithms. We set the optimizer with an initial learning rate of 1×10^{-2}. We decreased the learning rate by magnitude after the 20th and 40th epochs to prevent overfitting and facilitate convergence. We used a batch size of 24 and trained our networks for 60 epochs to ensure sufficient learning. We implemented our networks using the PyTorch framework and accelerated all experiments using an NVIDIA RTX3090 GPU.

5.3. Quantitative Evaluation of the HPSUR Dataset

The average precision (AP) curves of the four backbone models throughout their training iterations are displayed in Figure 12. All the networks demonstrate an increasing trend in AP, with the Hrnet-w32 and Hrnet-w64 models reaching a plateau earlier than the LiteHrnet variants. After about 50,000 iterations, the LiteHrnet-30 model exhibits the highest AP, indicating a more refined learning capability. Hrnet-w64 shows marginally better performance than Hrnet-w32.

The loss curves for different networks show how the optimization process went during training, as shown in Figure 13. All networks initially had a quick drop in loss, followed by a gradual convergence toward a minimum value. The LiteHrnet-30 model had the lowest loss, which means it had better generalization ability on the training data from the three volunteers. Although the LiteHrnet-18 model did not perform better than the LiteHrnet-30, it had a lower loss than both the Hrnet models, indicating the effectiveness of the LiteHrnet architecture in capturing pose features with fewer parameters. Despite its lightweight design, the AP and loss curves show that the LiteHrnet-30 model can effectively learn complex human poses from UWB radar data. These results suggest that LiteHrnet models have potential in scenarios where model efficiency is crucial without significantly compromising performance.

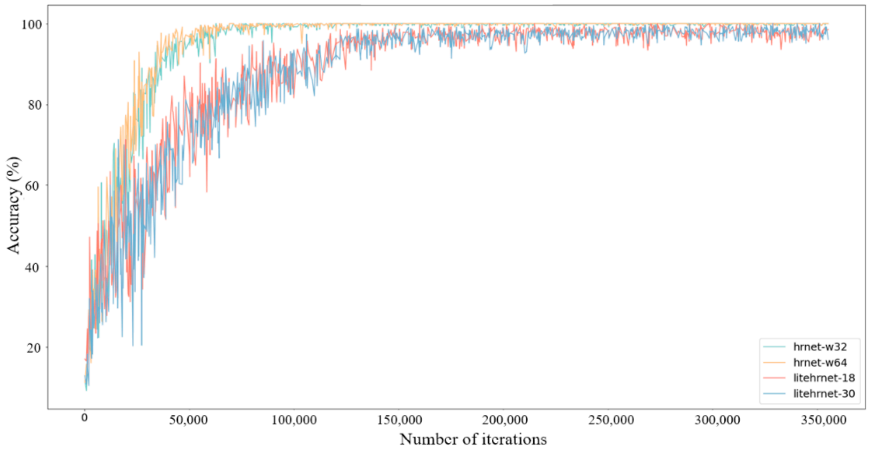


Figure 12. The average precision performance comparison of different neural networks.

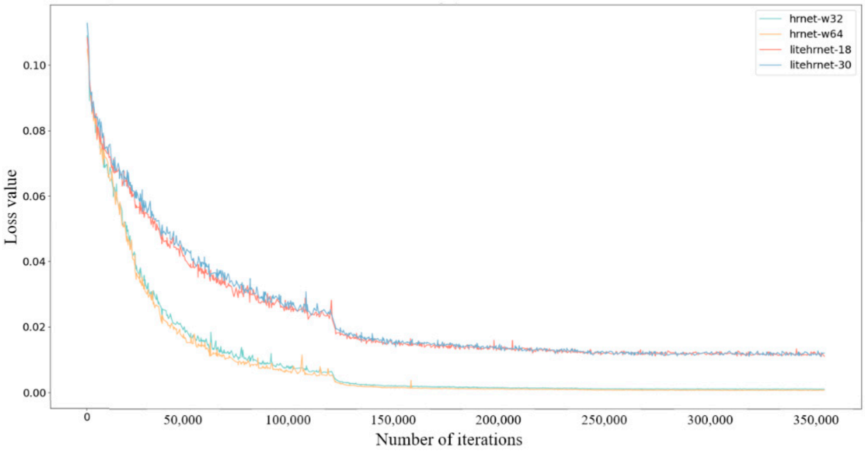


Figure 13. Training loss curves for different neural network architectures.

Table 3 and Figure 14 present a comparison of human pose estimation errors across various network models. Established models such as RF-Pose and RF-Pose 3D exhibit mean per-joint position errors (MPJPE) of 62.4 mm and 43.6 mm, respectively, with RF-Pose 3D being assessed over a substantial dataset comprising more than 1.6 million samples. In contrast, UWB-Pose and RadarFormer demonstrate pose estimation errors of 37.87 mm and 33.5 mm, respectively, indicating their higher accuracy in estimating human poses. RadarFormer's evaluation involved a dataset of 162,280 samples, highlighting its effectiveness across a considerable number of data points.

Table 3. Comparison of human pose estimation error based on different network models (unit: mm).

Model	Dataset Size	MPJPE
RF-Pose [8]	—	62.4
RF-Pose 3D [9]	1,693,440	43.6
UWB-Pose [29]	120,000	37.87
RadarFormer [40]	162,280	33.5
Hrnet-w32 (ours)	311,963	34.43
Hrnet-w64 (ours)	311,963	33.48
LiteHrnet-18 (ours)	311,963	32.09
LiteHrnet-30 (ours)	311,963	34.28

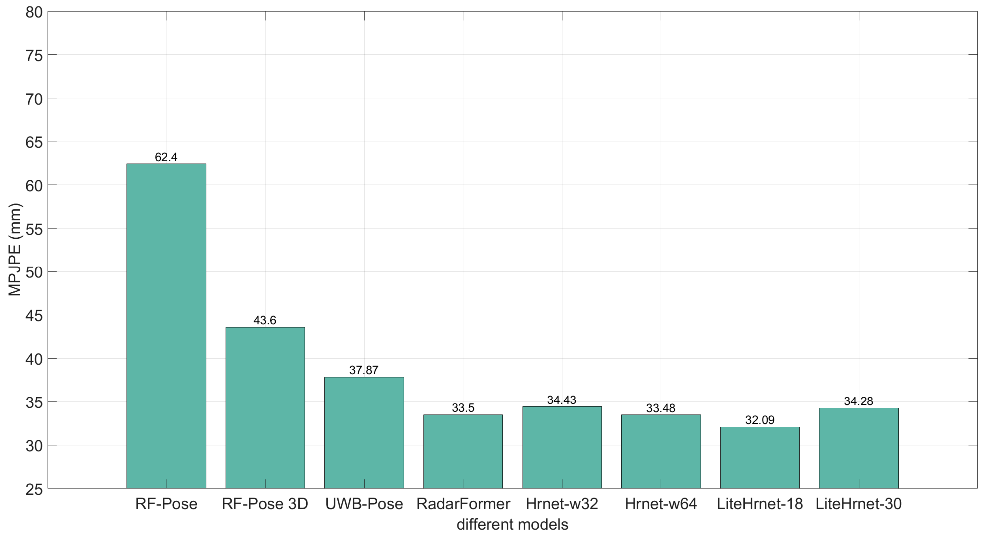


Figure 14. Comparison of our proposed method with several state-of-the-art methods of human pose estimation.

Our paper proposes Hrnet-w32, Hrnet-w64, LiteHrnet-18, and LiteHrnet-30 models. These methods demonstrate MPJPEs of 34.43 mm, 33.48 mm, 32.09 mm, and 34.28 mm, respectively. We analyzed each model over a dataset of 311,963 samples. The LiteHrnet-18 model outperforms all models with the lowest error rate, signifying a significant advancement in radar-based human pose estimation.

Table 4 and Figure 15 provide the performance details of various backbone models for human pose estimation, as measured using the HPSUR dataset. The focus is on the methods proposed in this paper. Among the backbones, the Hrformer-base and Hrformer-small models exhibit higher mean pose estimation errors of 59.16 mm and 49.94 mm, respectively. Other models, such as theCPM, Mobilenetv2, Resnet 50, Resnet 101, and Vipnas models, show a range of errors, with CPM having the lowest mean error of 39.75 mm, indicating a trend toward improvement with more recent architectures.

Table 4. Overall evaluation of the performance of the HPSUR dataset (unit: mm).

Backbone	Mean	Variance	Maximum	Minimum
Hrformer-base [41]	59.16	8.13	65.06	10.77
Hrformer-small [41]	49.94	8.53	60.24	6.76
CPM [42]	39.75	4.40	46.45	9.36
Mobilenetv2 [43]	37.85	3.10	39.77	9.36
Resnet 50	35.18	2.62	37.53	8.75
Resnet 101	36.93	3.37	39.36	8.52
Vipnas [44]	36.99	1.74	36.24	11.53
Hrnet-w32 (ours)	34.43	3.42	39.65	7.40
Hrnet-w64 (ours)	33.48	2.85	38.89	7.48
LiteHrnet-18 (ours)	32.09	2.40	36.01	8.26
LiteHrnet-30 (ours)	34.28	3.07	38.22	7.24

The final four backbones, part of this study's contribution, show significant advancements in estimation accuracy. The Hrnet-w32 backbone model achieved a mean error of 34.43 mm, showcasing a relatively higher variation in performance with a variance of 3.42 mm. Its maximum and minimum errors were recorded at 39.65 mm and 7.40 mm, respectively. The Hrnet-w64 model demonstrated a slightly lower mean error of 33.48 mm and a reduced variance of 2.85 mm, indicating a more consistent performance. The maximum error was slightly lower at 38.89 mm, with a minimum error close to the Hrnet-w32 model at 7.48 mm. The LiteHrnet-18 model exhibited a mean error of 32.09 mm, the lowest among the four models, and had the most minor variance at 2.40 mm, suggesting a more stable prediction capability. However, it registered a higher maximum error of 36.01 mm but also showed a better minimum error of 8.26 mm, implying that while its peaks were higher, its overall performance tended to be more reliable. Lastly, the LiteHrnet-30 model reported a mean error of 34.28 mm with a variance of 3.07 mm. Its maximum error was recorded at 38.22 mm, and the minimum error was 7.24 mm, which was the best among all models.

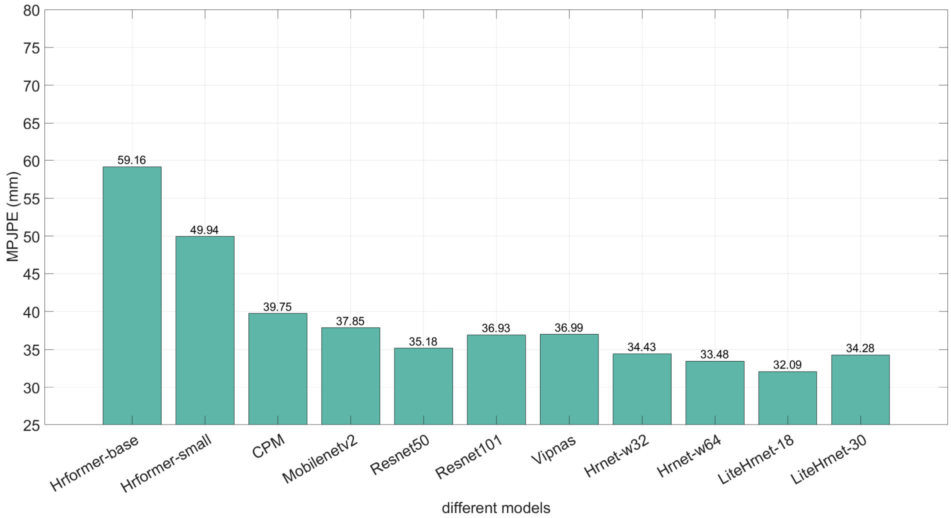


Figure 15. Comparison of the human pose estimation error based on our proposed method with several backbone models.

To summarize the comparison between the human pose estimation models, LiteHrnet-18 had the lowest mean error, which indicates that it has better accuracy on average. LiteHrnet-30 had the best minimum error, indicating its potential to yield highly accurate predictions in the best scenarios. Although the Hrnet variants were less consistent than the LiteHrnet models, they still maintained a competitive range of error metrics. These insights into the models' performance highlight the trade-offs between mean and variance and between the maximum and minimum errors. These trade-offs are crucial for practical applications of human pose estimation technology.

Tables 5 and 6 provide data on the performance metrics of four different neural network models evaluated on subjects S4 and S5, respectively. These evaluations were conducted as part of our study on how well these models perform when estimating human poses across different subjects. Table 5 shows that, for subject S4, the mean errors ranged from 36.16 mm for the LiteHrnet-18 model to 40.28 mm for the LiteHrnet-30 model, which indicates that these models vary in their ability to generalize across different subjects. However, the Hrnet-w64 and LiteHrnet-18 models showed lower variances of 3.10 mm and 2.73 mm, respectively, suggesting that they consistently performed well across various poses of the subject. The Hrnet-w64 model had the lowest maximum error of 41.73 mm, while the LiteHrnet-18 model exhibited the highest minimum error of 10.85 mm, indicating that it performed well in best-case scenarios. Table 6 shows that, for subject S5, all models had a reduced average error rate when compared with their performance on subject S4. The LiteHrnet-18 model had the lowest average error rate of 30.09 mm, followed closely by the LiteHrnet-30 model at 31.33 mm. The error variances were minimal across all models, indicating consistent performance. The Hrnet-W64 model registered the lowest variance at 2.72 mm. The LiteHrnet-18 model had the advantage in maximum error, recording a significantly lower figure of 34.34 mm, thereby demonstrating its superior capability in accurately estimating the poses of subject S5. Additionally, the LiteHrnet-30 model recorded the smallest minimum error at 5.93 mm, highlighting its outstanding accuracy under optimal conditions.

Table 5. Overall evaluation of the performance of S4 of the HPSUR dataset (unit: mm).

Backbone	Mean	Variance	Maximum	Minimum
Hrformer-base [41]	61.99	9.27	68.19	11.33
Hrformer-small [41]	52.89	9.53	63.47	7.93
CPM [42]	46.41	4.82	51.49	11.73
Mobilenetv2 [43]	44.93	4.06	46.60	10.91
Resnet 50	41.25	3.05	42.23	11.19
Resnet 101	43.73	4.34	46.14	10.04
Vipnas [44]	43.72	2.16	42.23	13.79
Hrnet-w32 (ours)	38.81	3.85	42.86	9.70
Hrnet-w64 (ours)	36.96	3.10	41.73	10.24
LiteHrnet-18 (ours)	36.16	2.73	39.38	10.85
LiteHrnet-30 (ours)	40.28	3.73	41.77	9.88

Table 6. Overall evaluation of the performance of S5 of the HPSUR dataset (unit: mm).

Backbone	Mean	Variance	Maximum	Minimum
Hrformer-base [41]	57.76	7.56	63.52	10.49
Hrformer-small [41]	48.49	8.03	58.65	6.17
CPM [42]	36.46	4.20	43.97	8.20
Mobilenetv2 [43]	34.36	2.63	36.40	8.65
Resnet 50	32.19	2.41	35.20	7.55
Resnet 101	33.57	2.90	36.02	7.77
Vipnas [44]	33.66	1.53	33.28	10.43
Hrnet-w32 (ours)	32.27	3.20	38.06	6.26
Hrnet-w64 (ours)	31.76	2.72	37.49	6.12
LiteHrnet-18 (ours)	30.09	2.24	34.34	6.98
LiteHrnet-30 (ours)	31.33	2.74	36.47	5.93

The analysis that compares the effectiveness of models across different subjects highlights the unique capabilities of each architectural design. The Hrnet model versions demonstrate consistent performance, while the LiteHrnet models excel in specific metrics and have average or minimal error rates. This evaluation emphasizes the importance of considering individual differences when creating and evaluating human pose estimation models. Gaining such an understanding is crucial for developing pose estimation technologies that are robust and reliable across different individuals.

Tables 7 and 8 show the performance of the S4 subject performing two different postures, labeled 1001 and 1003. These tables give us insights into the models' performance across diverse subjects and postures. For posture 1001, as shown in Table 7, the mean errors for the models range from 36.33 mm for the LiteHrnet-30 model to 39.91 mm for the Hrnet-w32 model, indicating a modest range of mean prediction accuracy across the network architectures. The LiteHrnet-18 and LiteHrnet-30 models show the lowest variances of 3.09 and 2.69 mm, respectively, indicating consistent performance across multiple instances of posture 1001. The maximum error is smallest for the LiteHrnet-30 model at 39.72 mm, whereas the minimum error does not vary significantly across the models, with the Hrnet-w32 model showing slightly better performance at 8.64 mm.

Table 7. Overall evaluation of the performance on S4 1001 of the HPSUR dataset (unit: mm).

Backbone	Mean	Variance	Maximum	Minimum
Hrnet-w32	39.91	3.96	44.96	8.64
Hrnet-w64	38.31	3.47	43.20	8.77
LiteHrnet-18	36.37	3.09	41.85	9.27
LiteHrnet-30	36.33	2.69	39.72	9.78

Table 8. Overall evaluation of the performance of S4 1003 of the HPSUR dataset (unit: mm).

Backbone	Mean	Variance	Maximum	Minimum
Hrnet-w32	37.97	3.77	41.29	10.50
Hrnet-w64	35.94	2.83	40.62	11.34
LiteHrnet-18	35.70	2.46	37.52	12.04
LiteHrnet-30	43.24	4.51	43.32	9.96

In the case of posture 1003, as illustrated in Table 8, the mean errors are relatively lower, with the Hrnet-w64 and LiteHrnet-18 models showing similar performance at around 35.94 mm and 35.70 mm, respectively. However, the LiteHrnet-30 model shows a notable increase in mean error to 43.24 mm. The variance metrics are consistent with the previous posture, but the LiteHrnet-18 model shows a marginally better variance of 2.46 mm. The maximum error for the LiteHrnet-18 model is significantly lower at 37.52 mm, highlighting its effectiveness in handling posture 1003. Conversely, the LiteHrnet-30 model has the least favorable minimum error at 9.96 mm.

It is important to choose an appropriate model for estimating posture based on the specific posture to be estimated. The LiteHrnet models offer more consistent performance with less variance, while the Hrnet models fluctuate more accurately. The difference in performance between postures 1001 and 1003 for the same subject also highlights the models' varying degrees of adaptability to different postural dynamics. This comprehensive evaluation is necessary for developing nuanced pose estimation models that can adapt to the subtleties of individual subject postures.

Tables 9–12 present a comprehensive analysis of the performance of four distinct neural network backbones when applied to subject S5 across four different postures (1001, 1002, 1003, and 1004). This analysis aims to gain a granular understanding of how well these models capture the nuanced differences in human movements. For posture 1001, as shown in Table 9, the LiteHrnet-30 model outperforms the other models with the lowest mean error of 29.52 mm and a reasonable variance of 2.24 mm, indicating its robust performance. Additionally, this model has the lowest maximum error, demonstrating its effectiveness in dealing with posture 1001. However, the LiteHrnet-18 model shows the highest minimum error among all the models at 7.66 mm.

Table 9. Overall evaluation of the performance of S5 1001 of the HPSUR dataset (unit: mm).

Backbone	Mean	Variance	Maximum	Minimum
Hrnet-w32	32.47	2.93	37.35	6.92
Hrnet-w64	31.93	2.60	37.07	7.12
LiteHrnet-18	30.14	2.04	33.73	7.66
LiteHrnet-30	29.52	2.24	34.02	6.61

Table 10. Overall evaluation of the performance of S5 1002 of the HPSUR dataset (unit: mm).

Backbone	Mean	Variance	Maximum	Minimum
Hrnet-w32	26.84	2.26	33.52	5.62
Hrnet-w64	26.91	2.20	33.74	5.66
LiteHrnet-18	26.11	1.82	32.90	6.38
LiteHrnet-30	26.58	2.03	33.04	5.46

Table 11. Overall evaluation of the performance of S5 1003 of the HPSUR dataset (unit: mm).

Backbone	Mean	Variance	Maximum	Minimum
Hrnet-w32	35.57	4.43	42.34	6.07
Hrnet-w64	34.55	3.23	40.89	5.48
LiteHrnet-18	33.43	2.92	36.92	7.26
LiteHrnet-30	36.45	3.96	42.10	5.75

Table 12. Overall evaluation of the performance of S5 1004 of the HPSUR dataset (unit: mm).

Backbone	Mean	Variance	Maximum	Minimum
Hrnet-w32	33.31	2.83	38.08	5.92
Hrnet-w64	33.06	2.77	37.34	5.54
LiteHrnet-18	29.23	2.04	32.90	5.62
LiteHrnet-30	32.77	2.61	36.47	5.29

For posture 1002, Table 10 reveals a closer range of mean errors among the models, with the LiteHrnet-18 model achieving the lowest mean error of 26.11 mm. Furthermore, it exhibits the lowest variance of 1.82 mm, illustrating its consistent performance across different instances of posture 1002. Once again, the LiteHrnet-30 model demonstrates a robust minimum error at 5.46 mm, which is the best among all models.

Table 11 shows that for posture 1003, the LiteHrnet-18 model records the lowest mean error of 33.43 mm, with a variance of 2.92 mm. This model also has the lowest maximum error of 36.92 mm, signifying a favorable performance. Although the LiteHrnet-30 model does not have the lowest mean error, it does maintain a competitive minimum error of 5.75 mm. Lastly, Table 12 shows the most significant difference in performance for posture 1004. Here, the LiteHrnet-18 model achieves a substantially lower mean error of 29.23 mm and the lowest variance of 2.04 mm, indicating an exceptional ability to predict this posture accurately. The LiteHrnet-30 model, however, has a slightly higher mean error of 32.77 mm but maintains a relatively low minimum error of 5.29 mm.

These performance metrics across diverse postures for subject S5 illustrate the distinct capabilities and limitations of the different models. LiteHrnet models, particularly the LiteHrnet-18 model, consistently show lower mean and variance errors, implying better overall performance in diverse posture estimation. These findings are crucial for developing advanced human pose estimation models that can adapt to various human postures and movements, ensuring high accuracy and reliability in real-world applications.

Table 13 compares the performance of four neural network backbone models when tested on two subjects, S4 and S5. The analysis shows how each model performs on different subjects in the same task domain. The Hrnet-w32 model significantly reduces the mean error from 38.81 mm for subject S4 to 32.27 mm for subject S5. This reduction in mean error is coupled with a decrease in variance, suggesting a better fit for subject S5. Similarly, the Hrnet-w64 model performs better on subject S5 with a decrease in mean error and variance. The mean errors are 36.96 mm for S4 and 31.76 mm for S5. The LiteHrnet-18 model also shows a decrease in mean error from 36.16 mm for S4 to 30.09 mm for S5, with a notable reduction in variance, indicating a more stable performance on subject S5. Although the LiteHrnet-30 model shows an increased mean error for subject S4 at 40.28 mm, it presents a significantly lower mean error of 31.33 mm for subject S5, again with reduced variance. Maximum errors are consistently lower for subject S5 across all models, with the most considerable improvement seen in the Hrnet-w32 model, from 42.86 mm down to 38.06 mm. Minimum errors follow a similar trend, with all models achieving lower errors on subject S5, indicating that the models are better at capturing the least complex poses of subject S5 than S4.

Table 13. Comparison of the performance of S4 and S5 of the HPSUR dataset (unit: mm).

Backbone	**Mean**		**Variance**		**Maximum**		**Minimum**	
	S4	**S5**	**S4**	**S5**	**S4**	**S5**	**S4**	**S5**
Hrnet-w32	38.81	32.27	3.85	3.20	42.86	38.06	9.70	6.26
Hrnet-w64	36.96	31.76	3.10	2.72	41.73	37.49	10.24	6.12
LiteHrnet-18	36.16	30.09	2.73	2.24	39.38	34.34	10.85	6.98
LiteHrnet-30	40.28	31.33	3.73	2.74	41.77	36.47	9.88	5.93

This comparison highlights the significance of evaluating models for specific subjects in human pose estimation research. Although all models show improved performance metrics for subject S5, the reduction in variance and error indicates that either the models are inherently more adaptable or the poses of subject S5 are less challenging for the models to estimate. These findings are crucial in developing customized pose estimation solutions accommodating inter-subject variability.

Table 14 uses four neural network backbone models to compare human pose estimation results for four different postures labeled 1001, 1002, 1003, and 1004. The models' performance is measured in terms of mean error and variance, which gives a comprehensive view of each model's capabilities in handling different human movements. In postures 1001 and 1002, the LiteHrnet models outperform the Hrnet variant models, with the LiteHrnet-30 model achieving the lowest mean error at 32.05 mm in posture 1001 and the LiteHrnet-18 model demonstrating the lowest variance at 2.41 mm in posture 1001. The LiteHrnet models are better suited to capture posture 1001 and 1002 more consistently.

Table 14. Comparison of the performance for 1001, 1002, 1003, and 1004 of the HPSUR dataset (unit: mm).

Backbone	**1001**		**1002**		**1003**		**1004**	
	Mean	**Var**	**Mean**	**Var**	**Mean**	**Var**	**Mean**	**Var**
Hrnet-w32	35.23	3.31	26.84	2.26	36.76	4.10	33.31	2.83
Hrnet-w64	34.30	2.92	36.91	2.20	35.24	3.03	33.06	2.77
LiteHrnet-18	32.45	2.43	26.11	1.82	34.56	2.69	29.23	2.04
LiteHrnet-30	32.05	2.41	26.58	2.03	39.82	4.23	32.77	2.61

Conversely, in posture 1003, the LiteHrnet-30 model exhibits a considerable increase in mean error to 39.82 mm, the highest among all models for this posture, which might indicate a reduced ability to estimate this particular pose accurately. However, the Hrnet-w64 and LiteHrnet-18 models maintain lower mean errors and variances, with the Hrnet-w64 model achieving the lowest variance, suggesting it is less sensitive to the variations within posture 1003. Lastly, posture 1004 showcases LiteHrnet-18's dominance, with the lowest mean error and variance across all models, suggesting its strong adaptability and reliability for this specific posture. Across all postures, the LiteHrnet-18 model consistently maintains low mean errors and variances, indicating its robustness and efficiency in human pose estimation tasks. The Hrnet models, while generally exhibiting higher mean errors and variances, still maintain competitive performance, especially the Hrnet-w64 model, which has the lowest variance for posture 1003. The comparison of these models provides critical insights into their posture-specific performance, underlining the importance of model selection based on the specific requirements of the pose estimation task at hand. This detailed assessment aids in discerning the strengths and limitations of each model, facilitating more informed decisions in the development and application of human pose estimation technologies.

Table 15 compares the performance of four backbone models in estimating poses 1001 and 1003 for subjects S4 and S5. This comparison helps to demonstrate how effectively the models identify different poses for different subjects. For pose 1001, all models showed a reduced mean error when evaluating subject S5 compared with S4. The LiteHrnet-30 model had the most significant reduction, indicating its heightened sensitivity to the subtleties of subject S5's posture. The variances for S5 were consistently lower, implying a more stable estimation across different instances of pose 1001. However, for pose 1003, the mean errors were generally lower for subject S5 across all models except for the LiteHrnet-30 model, which showed an increase. The variances for S5 were higher in the Hrnet-w32 and LiteHrnet-30 models, suggesting that pose 1003 presents more complexity or diversity in this subject's movements than in S4.

Table 15. Comparison of performance for 1001 and 1003 of S4 and S5 of the HPSUR dataset (unit: mm).

Backbone	**1001**				**1003**			
	S4		**S5**		**S4**		**S5**	
	Mean	**Var**	**Mean**	**Var**	**Mean**	**Var**	**Mean**	**Var**
Hrnet-w32	39.91	3.96	32.47	2.93	37.97	3.77	35.57	4.43
Hrnet-w64	38.31	3.47	31.93	2.60	35.94	2.83	34.55	3.23
LiteHrnet-18	36.37	3.09	30.14	2.04	35.70	2.46	33.43	2.92
LiteHrnet-30	36.33	2.69	29.52	2.24	43.24	4.51	36.45	3.96

The Hrnet-w64 and LiteHrnet-18 models showed notable consistency across subjects and poses, with competitive and stable mean errors and variances. The LiteHrnet-18 model showed the lowest variance in both poses for S5, reinforcing its robustness in pose estimation across different subjects. This comparison demonstrates the importance of considering subject variability and poses difficulty when developing human pose estimation models. The results indicate that although LiteHrnet architectures generally offer superior accuracy and stability, the choice of model may depend on the specific subject and pose combination, requiring a tailored approach for optimal performance in practical applications.

A boxplot visualization compares the mean per-joint position error (MPJPE) for four backbone models, Hrnet32, Hrnet64, LiteHrnet18, and LiteHrnet30, as shown in Figure 16. The bule and red lines of the figure represent the interquartile range (IQR) and mean value, respectively. The comparison is made across all test data and subsets for subjects S4 and S5. The boxplot gives an overview of the estimation errors for the models. The horizontal line within each box shows the median MPJPE, which helps quickly compare the central tendencies of the models. The LiteHrnet models consistently demonstrate a lower median error across all datasets, suggesting better pose estimation performance.

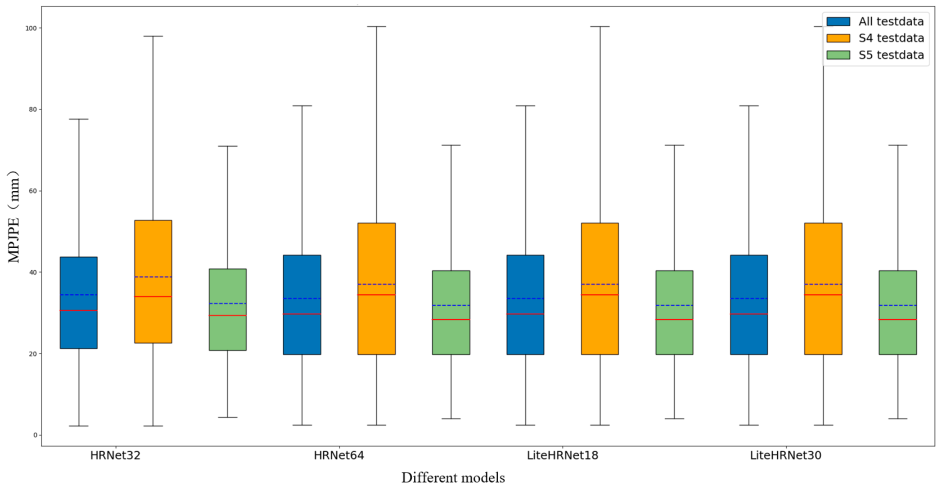


Figure 16. Comparative boxplot analysis of MPJPE across the Hrnet32, Hrnet64, LiteHrnet18, and LiteHrnet30 models using S4, S5, and all test data in the HPSUR dataset.

Each model’s interquartile range (IQR) is represented by the height of a box, which shows the middle 50% of the data with red lines in Figure 16. A smaller IQR indicates less variability in the model’s performance. The LiteHrnet-18 model consistently displays a compact IQR, especially for S5 test data, indicating robust performance with fewer outliers and less dispersion. The whiskers extending from the boxes illustrate the range of the data. At the same time, outliers, depicted as individual points, represent data points that fall beyond the whiskers and indicate pose estimates that significantly deviate from the typical error range. All models have outliers, indicating challenges in estimating certain poses. Comparing the performance of the S4 and S5 test data shows that the models’ performances are subject-specific. The LiteHrnet-30 model shows a notable increase in the median MPJPE for S4 compared with S5. This could mean that the LiteHrnet30 model is more attuned to the characteristics of subject S5’s data or that subject S4 presents more challenging poses for this model. The boxplot in Figure 16 succinctly encapsulates the performance distributions of the tested models, providing insights into their reliability and precision. The LiteHrnet models, particularly the LiteHrnet-18 model, exhibit consistently high performance across different subjects, making them promising candidates for real-world applications where pose estimation accuracy is critical.

A boxplot analysis of the mean per-joint position error (MPJPE) for four human pose estimation models, Hrnet-32, Hrnet-64, LiteHrnet-18, and LiteHrnet-30, is shown in Figure 17. The bule and red lines of the figure represent the interquartile range (IQR) and mean value, respectively. The analysis is conducted over various test datasets, including a collective test dataset and specific actions (1001 and 1003) for subject S4 and actions 1001, 1002, 1003, and 1004 for subject S5. The dashed line within each box represents the median MPJPE, which suggests that the LiteHrnet models, especially the LiteHrnet-18 model, offer a lower median error across most actions and subjects than the Hrnet models, which indicates higher accuracy in pose estimation for the LiteHrnet models. The interquartile range (IQR) for each action of both subjects is compact for LiteHrnet-18, signifying consistent estimation across different poses. The other models exhibit slightly wider IQRs, indicating more variability in their pose estimations. Whiskers extending from the boxes demonstrate the range of data, excluding potential outliers, and reflect the variability in estimation accuracy for more challenging poses. The presence of outliers, as indicated by points above and below the whiskers, is observed across all models and actions, highlighting instances where pose estimation deviates from typical error ranges.

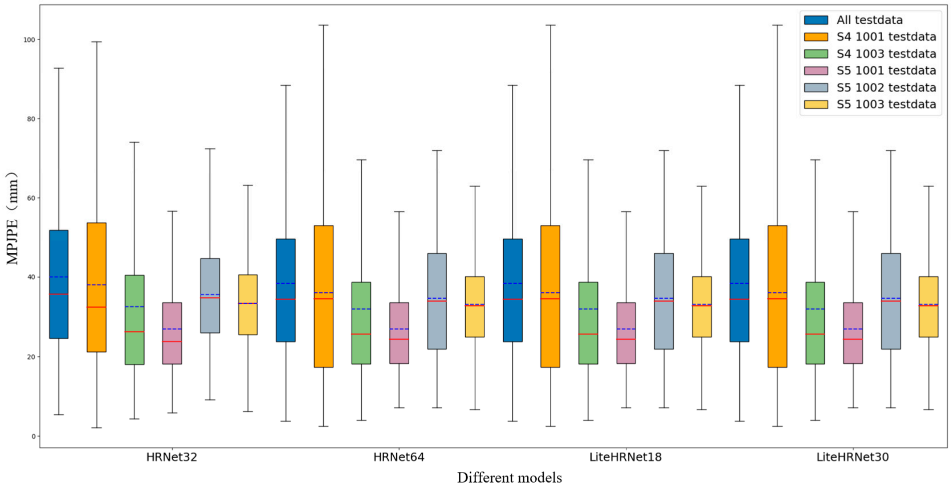


Figure 17. Comparative boxplot analysis of MPJPE across the Hrnet32, Hrnet64, LiteHrnet18, and LiteHrnet30 models on S4 1001, S4 1003, S5 1001, S5 1002, S5 1003, and S5 1004 data from the HPSUR dataset.

A comparison of S4 and S5 data for actions 1001 and 1003 reveals that model performance is not only model-specific but also action-specific. Some models handle specific actions better than others. For instance, the LiteHrnet-30 model tends to have a higher median error for S4’s action 1003, which suggests a potential model–subject–action interaction effect. Figure 10 effectively illustrates the performance distribution of human pose estimation models across different subjects and actions, providing valuable insights into model precision and reliability. The LiteHrnet models, particularly the LiteHrnet-18 model, demonstrate a lower median MPJPE, signifying their potential as robust solutions for accurate human pose estimation in diverse scenarios. Such detailed analysis is essential for advancing pose estimation technology and its application in real-world settings where accuracy and consistency are paramount.

5.4. Qualitative Evaluation of the HPSUR Dataset

Figure 18 visualizes the performance of four models in estimating human poses across various predefined motions using the mean per-joint position error (MPJPE) metric, which is commonly used in human pose estimation tasks to measure a model’s accuracy. This metric measures the accuracy of a model by calculating the average distance between the predicted and true joint locations across all tested poses. The four models compared are two versions of the Hrnet models (Hrnet-w32 and Hrnet-w64) and two versions of the LiteHrnet models (LiteHrnet-18 and LiteHrnet-30) that are designed to capture the spatial hierarchies in human poses by connecting high-to-low-resolution convolutions in parallel and exchanging information across resolutions.

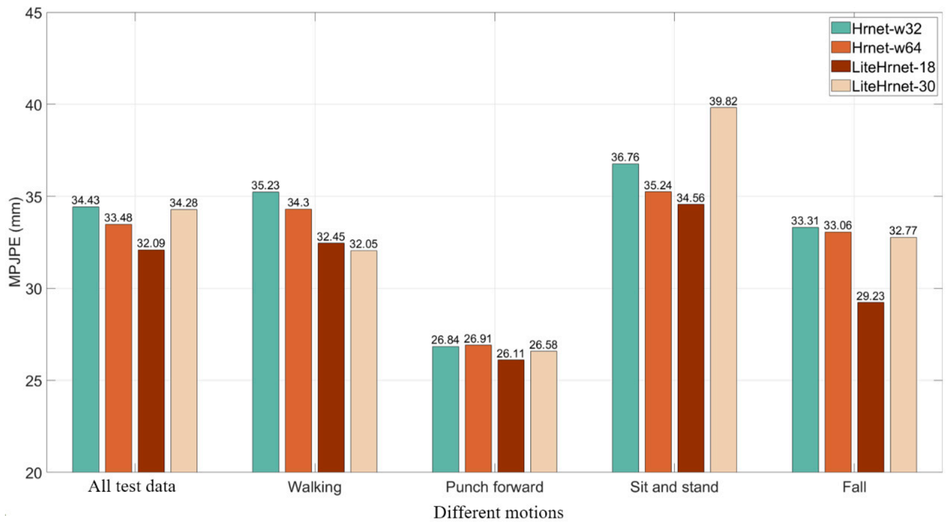


Figure 18. Bar graph comparing the MPJPE of two variants of Hrnet (Hrnet-w32 and Hrnet-w64) and two variants of LiteHrnet (LiteHrnet-18 and LiteHrnet-30) models across different motions.

Overall, each model has a distinct MPJPE across all motions, with the LiteHrnet30 and Hrnet64 models performing better than the other models, suggesting that these models have a better overall ability to capture and accurately capture human poses. However, the performance varies across motions when examining individual motions (labeled 1001 to 1004). This could be due to the varying complexity of the poses, where some poses may be easier or more complicated for the models to estimate accurately. There is a slight difference in performance between the Hrnet32 and Hrnet64 models, with the Hrnet64 model usually outperforming the Hrnet32 model, which could be attributed to the larger capacity and potentially more powerful feature extraction in the Hrnet64 model. Similarly, the LiteHrnet30 model tends to have a lower MPJPE across most motions than the LiteHrnet18 model, suggesting that the increased complexity in the LiteHrnet30 model offers an advantage in capturing the nuances of human poses.

Figure 19 compares different backbone models for human pose estimation using radar-based technology. The figure displays micro-Doppler radar signatures, the ground truth pose label, and the estimated poses by four backbone models: Hrnet-w32, Hrnet-w64, LiteHrnet-18, and LiteHrnet-30. The data include four motions, numbered 1001 to 1004. Each row corresponds to a separate motion that was captured by the system.

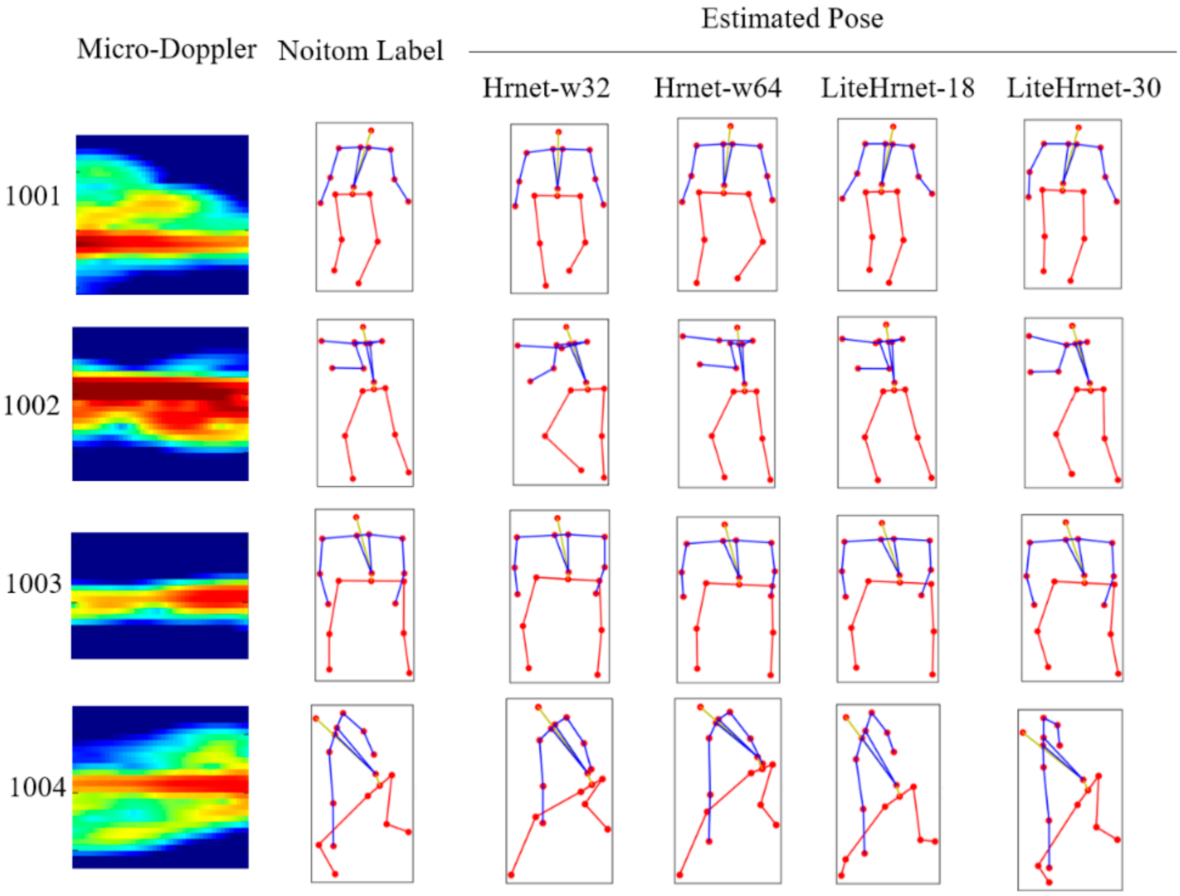


Figure 19. Visualization of pose estimation comparisons of the Hrnet and LiteHrnet models based on SISO UWB radar data for various human motions.

The first column of the figure displays the micro-Doppler signatures, which are radar-generated representations that capture the dynamic movement of different body parts. The second column shows the ground truth pose, represented as the "Noitom Label", against which the estimated poses are compared. Each model's estimated poses are represented by a skeletal diagram, with joints and limbs aligned according to the model's interpretation of the radar data. For motion 1001, the estimated poses closely match the ground truth, indicating effective model performance for this motion. However, for motion 1002, slight variations exist among the models' estimations, suggesting differences in model sensitivities or methodologies. Motion 1003 has significant discrepancies between the models' estimations and the ground truth, indicating potential challenges are inherent in this motion's complexity. Finally, motion 1004 also exhibits variations in pose estimation, with some models aligning more closely with the ground truth than others.

6. Conclusions

Our research and experimental analyses demonstrate the effectiveness of SISO UWB radar technology for human pose estimation (HPE). The innovative SCRP-Radar framework, which uses the Hrnet and LiteHrnet networks as backbone models, utilizes a unique space-aware coordinate representation and an up-sampling module. Our approach was extensively evaluated using the HPSUR dataset, which includes a wide range of actions and subjects, and the results provide robust empirical evidence supporting its accuracy.

The experimental results from the HPSUR dataset show that our methods are robust across various indoor scenarios and provide high precision in pose estimation. We quantified the performance of multiple backbone architectures, including Hrnet and LiteHrnet variants, and the latter showed impressive adaptability and accuracy, as evidenced by statistical analysis for subjects S4 and S5. The differences in performance metrics across the tested models highlight the importance of selecting the appropriate model for specific environmental contexts and pose estimation tasks. Specifically, the LiteHrnet-30 model demonstrates an impressive balance between accuracy and processing speed, making it suitable for real-time applications.

This investigation confirms the SCRP-Radar method is an important advancement in non-visual HPE and emphasizes the broader applicability of radar-based systems. The implications of this research are significant and indicate transformative prospects for UWB radar technology in ambient assisted living environments, interactive systems, and potentially in the burgeoning field of privacy-preserving surveillance. The positive outcomes of this study provide a strong foundation for subsequent innovation and practical deployment of radar-based human pose estimation systems.

Author Contributions: Conceptualization, X.Z., T.J. and Y.D.; methodology, X.Z. and Y.D.; formal analysis, X.Z.; investigation, X.Z. and K.L.; resources, X.Z. and T.J.; writing—original draft preparation, X.Z. and Y.S.; writing—review and editing, X.Z., Y.D., Y.S. and K.L.; supervision, X.Z. and T.J.; project administration, T.J.; funding acquisition, T.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China, grant number 61971430 entitled, "Study on ultra-wideband radar gait recognition technique".

Data Availability Statement: Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Acknowledgments: The authors thank anonymous reviewers and academic editors for their valuable comments and helpful suggestions. The authors are also grateful to the assistant editor for her meticulous work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ge, S.; Feng, D.; Song, S.; Wang, J.; Huang, X. Sparse Logistic Regression-Based One-Bit SAR Imaging. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5217915. [CrossRef]
2. Wang, Y.; Han, C.; Zhang, L.; Liu, J.; An, Q.; Yang, F. Millimeter-wave radar object classification using knowledge-assisted neural network. *Front. Neurosci.* **2022**, *16*, 1075538. [CrossRef] [PubMed]
3. Gamra, M.; Akhloufi, M. A review of deep learning techniques for 2D and 3D human pose estimation. *Image Vis. Comput.* **2021**, *114*, 104282. [CrossRef]
4. Ning, G.; Zhang, Z.; He, Z. Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation. *IEEE Trans. Multimed.* **2018**, *20*, 1246–1259. [CrossRef]
5. Jiang, W.; Xue, H.; Miao, C.; Wang, S.; Lin, S.; Tian, C.; Murali, S.; Hu, H.; Sun, Z.; Su, L. Towards 3d human pose construction using wifi. In Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, New York, NY, USA, 21–25 September 2020; pp. 1–14.
6. Wang, F.; Zhou, S.; Panev, S.; Han, J.; Huang, D. Person-in-WiFi: Fine-grained person perception using WiFi. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5452–5461.
7. Yang, C.; Wang, X.; Mao, S. RFID-pose: Vision-aided three-dimensional human pose estimation with radio-frequency identification. *IEEE Trans. Reliab.* **2020**, *70*, 1218–1231. [CrossRef]
8. Zhao, M.; Li, T.; Abu Alsheikh, M.; Tian, Y.; Zhao, H.; Torralba, A.; Katabi, D. Through-wall human pose estimation using radio signals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7356–7365.
9. Zhao, M.; Tian, Y.; Zhao, H.; Alsheikh, M.A.; Li, T.; Hristov, R.; Kabelac, Z.; Katabi, D.; Torralba, A. RF-based 3D skeletons. In Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication, Budapest, Hungary, 20–25 August 2018; pp. 267–281.
10. Adib, F.; Hsu, C.Y.; Mao, H.; Katabi, D.; Durand, F. Capturing the human figure through a wall. *ACM Trans. Graph. (TOG)* **2015**, *6*, 1–3. [CrossRef]

11. Zhou, X.; Jin, T.; Dai, Y.; Song, Y.; Qiu, Z. MD-Pose: Human Pose Estimation for Single-Channel UWB Radar. *IEEE Trans. Biom. Behav. Identity Sci.* **2023**, *5*, 449–463. [CrossRef]
12. Zhou, X.; Jin, T.; Du, H. A lightweight network model for human activity classifiction based on pre-trained mobilenetv2. In Proceedings of the IET International Radar Conference (IET IRC 2020), Chongqing, China, 4–6 November 2020; pp. 1483–1487.
13. Qi, F.; Lv, H.; Liang, F.; Li, Z.; Yu, X.; Wang, J. MHHT-based method for analysis of micro-Doppler signatures for human finer-grained activity using through-wall SFCW radar. *Remote Sens.* **2017**, *9*, 260. [CrossRef]
14. Li, Y.; Yang, S.; Liu, P.; Zhang, S.; Wang, Y.; Wang, Z.; Yang, W.; Xia, S.T. Simcc: A simple coordinate classification perspective for human pose estimation. In Proceedings of the 2022 European Conference on Computer Vision, Tel Aviv, Israel, 23–24 October 2022; pp. 89–106.
15. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
16. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
17. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
18. Wang, F.; Panev, S.; Dai, Z.; Han, J.; Huang, D. Can WiFi estimate person pose? *arXiv* **2019**, arXiv:1904.00277.
19. Guo, L.; Lu, Z.; Wen, X.; Zhou, S.; Han, Z. From signal to image: Capturing fine-grained human poses with commodity Wi-Fi. *IEEE Commun. Lett.* **2019**, *24*, 802–806. [CrossRef]
20. Wang, Y.; Guo, L.; Lu, Z.; Wen, X.; Zhou, S.; Meng, W. From point to space: 3D moving human pose estimation using commodity WiFi. *IEEE Commun. Lett.* **2021**, *25*, 2235–2239. [CrossRef]
21. Wang, K.; Wang, Q.; Xue, F.; Chen, W. 3D-skeleton estimation based on commodity millimeter wave radar. In Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020; pp. 1339–1343.
22. Sengupta, A.; Cao, S. mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 8418–8429. [CrossRef] [PubMed]
23. Sengupta, A.; Jin, F.; Zhang, R.; Cao, S. mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sens. J.* **2020**, *20*, 10032–10044. [CrossRef]
24. Shi, C.; Lu, L.; Liu, J.; Wang, Y.; Chen, Y.; Yu, J. mPose: Environment-and subject-agnostic 3D skeleton posture reconstruction leveraging a single mmWave device. *Smart Health* **2022**, *23*, 100228. [CrossRef]
25. Sengupta, A.; Jin, F.; Cao, S. NLP based skeletal pose estimation using mmWave radar point-cloud: A simulation approach. In Proceedings of the 2020 IEEE Radar Conference (RadarConf20), Florence, Italy, 21–25 September 2020; pp. 1–6.
26. Ding, W.; Cao, Z.; Zhang, J.; Chen, R.; Guo, X.; Wang, G. Radar-based 3D human skeleton estimation by kinematic constrained learning. *IEEE Sens. J.* **2021**, *21*, 23174–23184. [CrossRef]
27. Cui, H.; Dahnoun, N. real-time short-range human posture estimation using mmWave radars and neural networks. *IEEE Sens. J.* **2021**, *22*, 535–543. [CrossRef]
28. Li, T.; Fan, L.; Yuan, Y.; Katabi, D. Unsupervised learning for human sensing using radio signals. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 3288–3297.
29. Song, Y.; Jin, T.; Dai, Y.; Song, Y.; Zhou, X. Through-wall human pose reconstruction via UWB MIMO radar and 3D CNN. *Remote Sens.* **2022**, *13*, 241. [CrossRef]
30. Zheng, Z.; Pan, J.; Ni, Z.; Shi, C.; Ye, S.; Fang, G. Human posture reconstruction for through-the-wall radar imaging using convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 3505205. [CrossRef]
31. Kim, G.W.; Lee, S.W.; Son, H.Y.; Choi, K.W. A Study on 3D Human Pose Estimation Using Through-Wall IR-UWB Radar and Transformer. *IEEE Access* **2023**, *11*, 15082–15095. [CrossRef]
32. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-aware coordinate representation for human pose estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7093–7102.
33. Yang, S.; Quan, Z.; Nie, M.; Yang, W. Transpose: Towards explainable human pose estimation by transformer. *arXiv* **2020**, arXiv:2012.14214.
34. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. Tokenpose: Learning keypoint tokens for human pose estimation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11313–11322.
35. Li, J.; Bian, S.; Zeng, A.; Wang, C.; Pang, B.; Liu, W.; Lu, C. Human pose regression with residual log-likelihood estimation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11025–11034.
36. Abdu, F.J.; Zhang, Y.; Deng, Z. Activity classification based on feature fusion of FMCW radar human motion micro-Doppler signatures. *IEEE Sens. J.* **2022**, *22*, 8648–8662. [CrossRef]
37. Hassan, S.; Wang, X.; Ishtiaq, S.; Ullah, N.; Mohammad, A.; Noorwali, A. Human Activity Classification Based on Dual Micro-Motion Signatures Using Interferometric Radar. *Remote Sens.* **2023**, *15*, 1752. [CrossRef]

38. Li, X.; He, Y.; Fioranelli, F.; Jing, X. Semisupervised human activity recognition with radar micro-Doppler signatures. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5103112. [CrossRef]
39. Li, P.; Wang, T.; He, Z.; Gao, M.; Yang, Y.; Huang, J. Spatiotemporal Weighted Micro-Doppler Spectrum Design for Soft Synchronization FMCW Radar. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–14. [CrossRef]
40. Zheng, Z.; Zhang, D.; Liang, X. RadarFormer: End-to-End Human Perception with Through-Wall Radar and Transformers. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–15. [CrossRef] [PubMed]
41. Yuan, Y.; Fu, R.; Huang, L. Hrformer: High-resolution transformer for dense prediction. In Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems, Online Conference, 6–14 December 2021; pp. 1–15.
42. Wei, S.E.; Ramakrishna, V.; Kanade, T. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
43. Sandler, M.; Howard, A.; Zhu, M. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 28–23 June 2018; pp. 4510–4520.
44. Xu, L.; Guan, Y.; Jin, S. Vipnas: Efficient video pose estimation via neural architecture search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16072–16081.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.