remote sensing

MDPI

Article

Three-Dimensional Human Pose Estimation from Micro-Doppler Signature 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: In this paper, we propose an innovative approach for transforming 2D human pose estimation into 3D models using Single Input–Single Output (SISO) Ultra-Wideband (UWB) radar technology. This method addresses the significant challenge of reconstructing 3D human poses from 1D radar signals, a task traditionally hindered by low spatial resolution and complex inverse problems. The difficulty is further exacerbated by the ambiguity in 3D pose reconstruction, as multiple 3D poses may correspond to similar 2D projections. Our solution, termed the Radar PoseLifter network, leverages the micro-Doppler signatures inherent in 1D radar echoes to effectively convert 2D pose information into 3D structures. The network is specifically designed to handle the long-range dependencies present in sequences of 2D poses. It employs a fully convolutional architecture, enhanced with a dilated temporal convolutions network, for efficient data processing. We rigorously evaluated the Radar PoseLifter network using the HPSUR dataset, which includes a diverse range of human movements. This dataset comprises data from five individuals with varying physical characteristics, performing a variety of actions. Our experimental results demonstrate the method's robustness and accuracy in estimating complex human poses, highlighting its effectiveness. This research contributes significantly to the advancement of human motion capture using radar technology. It presents a viable solution for applications where precision and reliability in motion capture are paramount. The study not only enhances the understanding of 3D pose estimation from radar data but also opens new avenues for practical applications in various fields.

Keywords: 3D human pose estimation; Micro Doppler; Radar PoseLifter; dilated temporal convolutions network; SISO UWB radar

Citation: Zhou, X.; Jin, T.; Dai, Y.; Song, Y.; Li, K. Three-Dimensional Human Pose Estimation from Micro-Doppler Signature Based on SISO UWB Radar. *Remote Sens.* **2024**, *16*, 1295. https://doi.org/10.3390/rs16071295

Academic Editors: Dusan Gleich, Nebojsa Doncov and Venceslav Kafedziski

Received: 4 March 2024
Revised: 27 March 2024
Accepted: 2 April 2024
Published: 6 April 2024

1. Introduction

With the rapidly evolving field of urban wireless sensing, significant strides have been made, particularly in complex cityscapes. These intelligent systems are designed to interpret human behavior using pervasive wireless signals, playing a crucial role in understanding the pedestrian dynamics essential for autonomous and semi-autonomous vehicle operations. These advancements are not only pivotal in vehicular contexts but also hold immense potential in healthcare applications, notably in aiding the disabled and elderly [1].

Within the urban sensing domain, estimating human poses is critical for discerning intentions and actions, an essential aspect of environmental perception in urban settings [2]. This area is becoming increasingly relevant in indoor, human-focused environments, where the goal is to determine human postures through various sensor inputs. Human pose capture, a cornerstone of human–computer interaction, has been challenging [3]. The emphasis is primarily on identifying and classifying different body parts, such as ankles, shoulders, and wrists. While camera-based systems have seen success in human pose estimation [4–7], privacy concerns are a significant hurdle. The omnipresence of video

surveillance can be intrusive, and the vulnerability of millions of wireless security cameras to hacking globally is a concern. In response, wireless sensing systems emerge as a privacy-preserving alternative, showing resilience against factors like clothing, background, lighting, and occlusion [8].

WiFi-based human sensing presents a promising solution to privacy concerns. Commercial WiFi devices, functioning as RF sensors in the 2.4 GHz and 5 GHz bands [9,10], offer a less intrusive means of monitoring. By utilizing WiFi signals, this technology bypasses the need for visual surveillance, thereby protecting individual privacy. In ref. [9], deep learning techniques applied to WiFi signals have shown potential for end-to-end human pose estimation. Following this, Wi-Mose [10] introduced a method to extract pose-related features from WiFi signals, translating them into human poses.

Despite these advances, WiFi-based sensing systems have limitations, primarily due to the coarse resolution offered by the bandwidths (20 MHz and 40 MHz) used in standard WiFi protocols. This limitation hampers the ability to accurately capture fine-grained human poses. Moreover, WiFi signals are prone to interference from environmental factors, which can significantly affect the reliability of pose estimations in urban settings.

In light of these challenges, the focus has increasingly shifted towards radar-based intelligent wireless sensing systems. Radar technology, with its ability to penetrate through obstacles and low sensitivity to environmental variables, offers a robust alternative for urban sensing. These systems can detect human pose, body shape, and activities even through walls and in poorly lit settings. 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 [1,3,11–17], and lower frequency radars, operating around a few GHz [18–24]. High-frequency radar signals, with their shorter wavelengths, provide greater precision in posture capture but lack the ability to penetrate walls and furniture. Studies [1,12,13] 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. [11] innovated a domain discriminator that filters user-specific characteristics from mmWave signals, enabling robust skeleton reconstruction across different users with minimal training effort. Dahnoun et al. [16] designed a novel neural network model for human posture estimation based on point cloud data, comprising a part detector for initial keypoint 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 [18–20] introduced a neural network system that interprets radar signals for 2D human pose and dynamic 3D human mesh estimation. Jin et al. [21] developed a novel through-wall 3D pose reconstruction framework using UWB MIMO radar and 3D CNNs for concealed target detection. Fang et al. [22] proposed a cross-modal CNN-based method for postural reconstruction in Through the Wall Radar Imaging (TWRI). Then, they proposed a pose estimation framework (Hourglass) and a semantic segmentation framework (UNet) to serve as the teacher network to convert the RGB images into the pose keypoints and the shape masks [23]. Choi et al. [24] introduced the 3D-TransPose algorithm for 3D human pose estimation, leveraging an attention mechanism to focus on relevant time periods in time-domain IR-UWB radar signals. Nevertheless, these approaches rely on MIMO radar imaging, and the quality of radar imaging can be significantly impacted by the changes in the surrounding environment and the relative distance between the human target and the radar. Therefore, we use SISO UWB radar to capture human poses based on the micro-Doppler signature, which is not susceptible to the human target and environment. However, reconstructing fine-grained human skeletal spatial information from the 1D radar echo with low spatial resolution is a severely ill-posed problem.

Numerous studies have demonstrated that the Micro-Doppler (MD) signatures are resilient to variations in the human target and environment, offering subject-independent

and environment-independent features. He et al. [25] propose a multiscale residual attention network (MRA-Net) for joint activity recognition and person identification with radar micro-Doppler signatures. Kim et al. [26] apply deep convolutional neural networks directly to a raw micro-Doppler spectrogram for both human detection and activity classification problems. The frequency shifts in different body parts, as captured in MD spectrograms, provide a comprehensive perception of human movements. The dynamic motion of human body parts, such as the torso, arms, legs, hands, and feet, results in distinct MD signatures that can be visually differentiated from one another.

To address these challenges, we formulate a two-step approach to realize 3D human pose estimation. Initially, we employ the Swin Transformer (MDST) network to estimate 2D human poses based on micro-Doppler signatures. Subsequently, we introduce the innovative Radar PoseLifter network, designed to elevate 2D human poses to 3D using SISO UWB radar. In summary, our contributions can be succinctly summarized as follows.

(1) We propose the Radar PoseLifter network, a fully convolutional-based architecture with dilated temporal convolutions for 3D human pose estimation based on SISO UWB radar, which is simple and efficient, to lift 2D human joints to 3D poses.

(2) To learn inherently enforces long-range dependencies, and the external knowledge information of the human target is injected into the Radar PoseLifter network. This addition enhances the network's capability to discern and accurately estimate intricate human poses.

(3) Numerous experiments are carried out to verify the effectiveness and robustness of the proposed method, which is conducted across four distinct human motions, demonstrating our approach's broad applicability and reliability in accurately estimating human poses.

The remainder of this paper is structured as follows: Section 2 outlines the theoretical framework, encompassing the geometric modeling of human targets and radar systems, the structural information inherent in human models, and the Micro Doppler characteristics of human posture. Section 3 is dedicated to introducing the architecture of the proposed radar poselifter network. Section 4 presents both quantitative and qualitative assessments of the proposed method, utilizing the HPSUR dataset. Finally, a discussion and conclusion are presented in Section 5 and Section 6, respectively.

2. Theory

The geometric relationship between the transmitting and receiving antennas of the SISO UWB radar is shown in Figure 1. The geometric motion relationship between the radar and moving human targets 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 $= (0, 0, 0)^T$ and 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 $\boldsymbol{R_o} = (U_o, V_o, W_o)^T$ and the initial azimuth angle and elevation angle are defined as α, β, respectively. Furthermore, the radial unit vector extending from the radar towards the target is defined as:

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

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 $\boldsymbol{r_o} = (X_o, Y_o, Z_o)^{\mathrm{T}}$. During the observed period, point J_1 undergoes four simultaneous movements characterized by their distinct kinematic properties.

(1) The skeleton translates with speed $\boldsymbol{v}$ in the radar coordinate system;

(2) The skeleton accelerates with acceleration $\boldsymbol{a}$;

(3) The skeleton vibrates sinusoidally with frequency f_v and amplitude D_v. The azimuth angle and pitch angle are α_p, β_p, respectively, and the unit vector of the vibration direction is $n_v = (\cos\alpha_p \cos\beta_p, \sin\alpha_p \cos\beta_p, \sin\beta_p)^T$;

(4) The skeleton rotates in the reference coordinate system with an angular velocity of $\omega = (\omega_X, \omega_V, \omega_Z)^T$. At time t, the J_1 skeleton point moves to the new position J_1''.

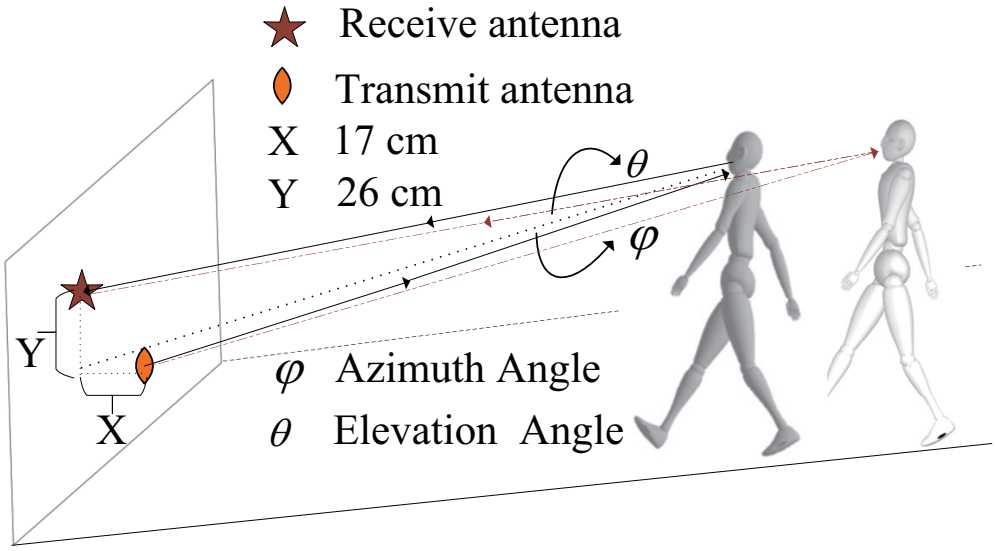

Figure 1. Configuration of SISO UWB bistatic radar for human pose estimation showing the relative positioning of the transmit and receive antennas, along with the azimuth and elevation angles to the moving human target.

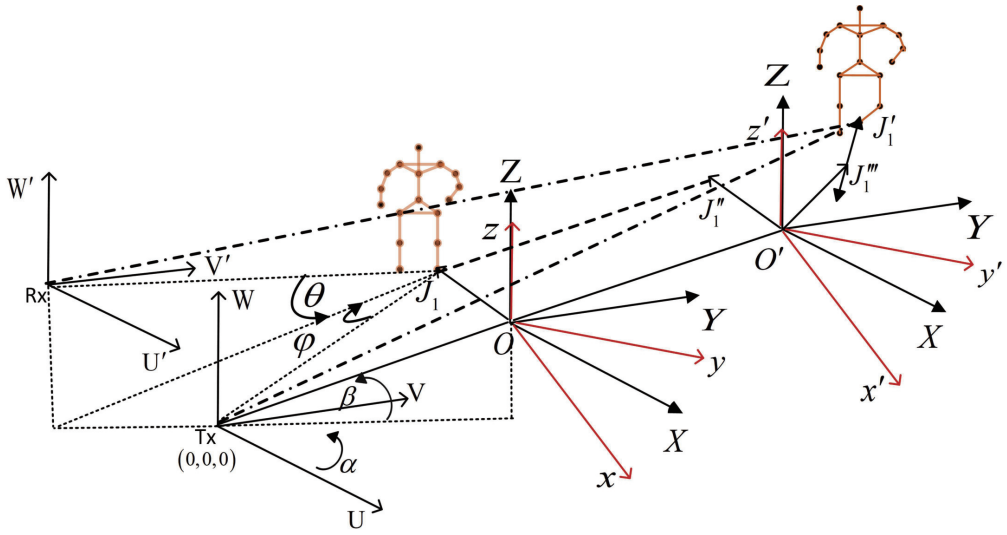

Figure 2. The geometric relationship between human motion model and radar.

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

$$\begin{aligned} \boldsymbol{R_{tx}}(t) &= \overline{\boldsymbol{TxJ_1}} = \boldsymbol{R_o} + \boldsymbol{r_o} + \overline{\boldsymbol{J_1J_1''}} + \overline{\boldsymbol{J_1''J_1'''}} + \overline{\boldsymbol{J_1'''J_1'}} \\ &= \boldsymbol{R_0} + \boldsymbol{r_0} + \boldsymbol{V}t + 1/2\boldsymbol{a}t^2 + \boldsymbol{Rot}(t) \cdot \overline{\boldsymbol{O'J_1''}} + D_v \sin(2\pi f_v t) \cdot \boldsymbol{n_v} \\ &= \boldsymbol{R_0} + \boldsymbol{r_0} + \boldsymbol{V}t + 1/2\boldsymbol{a}t^2 + \boldsymbol{Rot}(t) \cdot \boldsymbol{r_0} + D_v \sin(2\pi f_v t) \cdot \boldsymbol{n_v} \end{aligned} \quad (2)$$

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

$$\begin{aligned}\boldsymbol{R}_{\boldsymbol{Rx}}(t) &= \overline{\boldsymbol{RxJ}_1} = \boldsymbol{R}_x + \boldsymbol{R}_o + \boldsymbol{r}_o + \overline{\boldsymbol{J_1J_1''}} + \overline{\boldsymbol{J_1''J_1'''}} + \overline{\boldsymbol{J_1'''J_1'}} \\ &= \boldsymbol{R}_x + \boldsymbol{R_o} + \boldsymbol{r_o} + \boldsymbol{V}t + 1/2\boldsymbol{a}t^2 + \boldsymbol{Rot}(t) \cdot \overline{\boldsymbol{O'J_1''}} + D_v \sin(2\pi f_v t) \cdot \boldsymbol{n}_v \\ &= \boldsymbol{R}_x + \boldsymbol{R_0} + \boldsymbol{r_0} + \boldsymbol{V}t + 1/2\boldsymbol{a}t^2 + \mathrm{Rot}(t) \cdot \boldsymbol{r_0} + D_v \sin(2\pi f_v t) \cdot \boldsymbol{n}_v\end{aligned} \tag{3}$$

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

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

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

$$\begin{aligned}\boldsymbol{R}(t) &= \|\boldsymbol{R}(t)\| = \|\boldsymbol{R_{tx}}(t)\| + \|\boldsymbol{R_R}(t)\| \\ &= \left\|\boldsymbol{R_o} + \boldsymbol{r_o} + \boldsymbol{V}t + 1/2\boldsymbol{a}t^2 + \boldsymbol{Rot}(t) \cdot \boldsymbol{r_o} + D_v \sin(2\pi f_v t) \cdot \boldsymbol{n}_v\right\| \\ &+ \left\|\boldsymbol{R}_x + \boldsymbol{R_0} + \boldsymbol{r_0} + \boldsymbol{V}t + 1/2\boldsymbol{a}t^2 + \mathrm{Rot}(t) \cdot \boldsymbol{r_0} + D_v \sin(2\pi f_v t) \cdot \boldsymbol{n}_v\right\|\end{aligned} \tag{5}$$

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

$$\boldsymbol{Rot}(t) = I + \hat{\boldsymbol{\omega}} \sin(\Omega t) + \hat{\boldsymbol{\omega}}^2(1 - \cos(\Omega t)) = \exp(\hat{\boldsymbol{\omega}} t) \tag{6}$$

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.

$$\begin{aligned}f_d &= \frac{1}{2\pi}\frac{\mathrm{d}\Phi(R(t))}{\mathrm{d}t} = \frac{f}{c}\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \frac{f}{c}\frac{\mathrm{d}(R_{tx}(t) + R_{Rx}(t))}{\mathrm{d}t} \\ &= \frac{2f}{c}\boldsymbol{V}^{\mathrm{T}} \cdot \boldsymbol{n}_{p'} + \frac{2f}{c}\left(\boldsymbol{a}^{\mathrm{T}} \cdot \boldsymbol{n}_{p'}\right)t + \frac{2f}{c}\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{Rot}(t) \cdot \boldsymbol{r}_0)^{\mathrm{T}} \cdot \boldsymbol{n}_{p'} + \frac{4f}{c}\pi f_v D_v \cos(2\pi f_v t) \cdot \boldsymbol{n}_v^{\mathrm{T}} \cdot \boldsymbol{n}_{p'}\end{aligned} \tag{8}$$

Noting $\boldsymbol{r} = \mathrm{Rot}(t) \cdot \boldsymbol{r}_0$, 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 in the following form:

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

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

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

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

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

However, only the modulation characteristics of human motion frequency caused by acceleration and vibration can be seen 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 $\boldsymbol{r}_0 = (x_0, y_0, z_0)^{\mathrm{T}}$ and rotates in the target coordinate system with the angular velocity $\omega_l = \left(\omega_x, \omega_y, \omega_z\right)^{\mathrm{T}}$; (ϕ, θ, ψ) represents the initial Euler angles.The initial rotation matrix is represented by $\boldsymbol{R}_{init}$:

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

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

$$\begin{aligned} R(t) &= \|\boldsymbol{R}(t)\| = \|\boldsymbol{R_{tx}}(t)\| + \|\boldsymbol{R_{Rx}}(t)\| \\ &= \|\boldsymbol{R_o} + \boldsymbol{r_o} + \boldsymbol{V}t + 1/2\boldsymbol{a}t^2 + \boldsymbol{Rot}(t) \cdot \boldsymbol{R}_{\text{init}} \cdot \boldsymbol{r_o} + D_v \sin(2\pi f_v t) \cdot \boldsymbol{n}_v\| \\ &+ \|\boldsymbol{R}_x + \boldsymbol{R_o} + \boldsymbol{r_o} + \boldsymbol{V}t + 1/2\boldsymbol{a}t^2 + \mathrm{Rot}(t) \cdot \boldsymbol{R}_{\text{init}} \cdot \boldsymbol{r}_{\mathrm{o}} + D_v \sin(2\pi f_v t) \cdot \boldsymbol{n}_v\| \end{aligned} \tag{13}$$

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

$$\begin{aligned} f_d &= \frac{1}{2\pi}\frac{\mathrm{d}\Phi(R(t))}{\mathrm{d}t} = \frac{f}{c}\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \frac{f}{c}\frac{\mathrm{d}(R_{tx}(t) + R_{Rx}(t))}{\mathrm{d}t} \\ &= \frac{2f}{c}\boldsymbol{V}^{\mathrm{T}} \cdot \boldsymbol{n}_{p'} + \frac{2f}{c}\left(\boldsymbol{a}^{\mathrm{T}} \cdot \boldsymbol{n}_{p'}\right)t + \frac{2f}{c}\frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{Rot}(t) \cdot \boldsymbol{R}_{\text{init}} \cdot \boldsymbol{r}_0)^{\mathrm{T}} \\ &\cdot \boldsymbol{n}_{p'} + \frac{4f}{c}\pi f_v D_v \cos(2\pi f_v t) \cdot \boldsymbol{n_v}^{\mathrm{T}} \cdot \boldsymbol{n}_{p'} \end{aligned} \tag{14}$$

Noting $\boldsymbol{r} = \boldsymbol{Rot}(t) \cdot \boldsymbol{R} \cdot \boldsymbol{r}_0$, with $\boldsymbol{n} = \boldsymbol{R_0}/\|\boldsymbol{R_0}\|$ to approximate instead of $\boldsymbol{n}_{p'}$, at this time of human movement target joint of micro-Doppler f_{m-d} as follows:

$$\begin{aligned} f_{m-d} &= \frac{2f}{c}\left(\boldsymbol{a}^{\mathrm{T}} \cdot \boldsymbol{n}\right)t + \frac{2f}{c}\left(\Omega_l \boldsymbol{\omega}_l' \times \boldsymbol{r}\right)^{\mathrm{T}} \cdot \boldsymbol{n} + \frac{4f\pi f_v D_v}{c}\cos(2\pi f_v t) \cdot \boldsymbol{n}_v^{\mathrm{T}} \cdot \boldsymbol{n} \\ &= \frac{2f}{c}\left(\boldsymbol{a}^{\mathrm{T}} \cdot \boldsymbol{n}\right)t + \frac{2f}{c}(\Omega_l \hat{\boldsymbol{\omega}}_l \cdot \boldsymbol{R} \circ \mathrm{t}(t) \cdot \boldsymbol{R}_{\text{init}} \cdot \boldsymbol{r}_0)^{\mathrm{T}} \cdot \boldsymbol{n} + \frac{4f\pi f_v D_v}{c}\cos(2\pi f_v t) \cdot \boldsymbol{n}_v^{\mathrm{T}} \cdot \boldsymbol{n} \\ &= \frac{2f}{c}\left(\boldsymbol{a}^{\mathrm{T}} \cdot \boldsymbol{n}\right)t + \frac{2f}{c}\left(\Omega_l\left[\hat{\boldsymbol{\omega}}_l^2 \sin(\Omega_l t) - \hat{\boldsymbol{\omega}}_l^3 \cos(\Omega_l t) + \hat{\boldsymbol{\omega}}_l\left(I + \hat{\boldsymbol{\omega}}_l^2\right)\right]\boldsymbol{R}_{\text{init}} \cdot \boldsymbol{r}_0\right)^{\mathrm{T}} \\ &\cdot \boldsymbol{n} + \frac{4f\pi f_v D_v}{r}\cos(2\pi f_v t) \cdot \boldsymbol{n}_v^{\mathrm{T}} \cdot \boldsymbol{n} \end{aligned} \tag{15}$$

The formula presented above indicates that when the target simultaneously exhibits translation, acceleration, vibration, and rotation characteristics, the parameter f_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 the rotation periods. Furthermore, the amplitude of these changes depends on the vibration frequency, vibration amplitude, and rotational angular velocity.

In this paper, the Boulic human body model, characterized by its 62 degrees of freedom and 32 joints, is abstracted into 13 standardized rigid bodies and 17 nodal points for simplification. The rigid bodies represent various parts of the human anatomy: the head,

the shoulders (left and right), the arms (left and right), the forearms (left and right), the thighs (left and right), the calves (left and right), and the torso (upper and lower parts). The nodal points identified are the hips, the upper legs (right and left), the legs (right and left), the feet (right and left), the spine, the head, the shoulders (right and left), the arms (right and left), the forearms (right and left), and the left hand. This study models the interaction between the foot and the ground as a rigid contact, with the contact point determined by the geometry of the foot's plantar surface and the foot's orientation.

The diagram in Figure 3 offers a detailed proportional representation of the human skeletal structure, indicating the relative orientations of each segment of the body. These orientations are denoted by specific angles and are not fully independent, particularly during the double support phase of movement, due to the interconnected nature of the model, akin to a closed-loop system. The figure illustrates how the body's proportions are segmented in relation to overall height (denoted as 'H'). According to the empirical data captured, the segment from the top of the head down to the lower neck represents 18.2% of a person's total height. The shoulders are measured to be 25.9% of the height, while the torso contributes to 28.8% of the height. In terms of limb proportions, the upper arms are 18.8% of the height, the lower arms make up 14.5%, the thighs account for 24.5%, the calves for 28.5%, and the hips for 19.1% of the total height. These measurements provide a quantified overview of the human form, which is essential for the study of biomechanics and related fields.

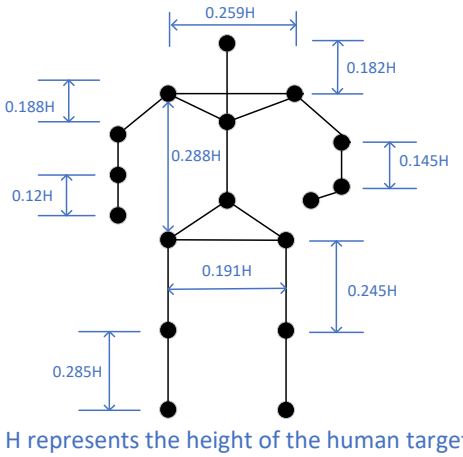

Figure 3. Schematic representation of the proportional human skeletal structure.

In investigating the interactions between various human body segments during motion, body segment trajectories were analyzed using data from the Carnegie Mellon University Motion Capture (MOCAP) database for a human subject. Figure 3 provides structural information on the targeted human body segments, while Figure 4 outlines the experimental setup for the simulation. This setup features a human target, with a height of approximately 175 cm, beginning a face-down fall at approximately 2 s into the simulation, with the total duration of the data collection being approximately 5.5 s. Subsequent processing of radar echo data from the moving human body facilitated the Micro-Doppler (MD) spectrum extraction. As depicted in Figure 5, specific movements were executed by the target between 2 and 4.5 s, after which the target remained stationary for the rest of the observation period.

Figure 5 also clearly delineates the variations in Doppler frequency attributed to the movement of different human body segments. In this depiction, the zero-frequency line is indicative of the torso of the human body. The figure shows that, prior to 2 s and subsequent to 4.5 s, the human target maintains a stationary stance. Conversely, the period between 2 and 4.5 s is characterized by changes in the Micro-Doppler (MD) frequency, reflecting the movement dynamics of the human target's various segments. It is observed that an

increase in MD frequency corresponds to a greater amplitude of movement in the respective body part.

Notably, the human body is an asymmetric, non-rigid structure with bilateral symmetry. Figure 6 demonstrates that the MD effects caused by this symmetrical structure during human movement follow specific patterns. In the figure, the first and third rows represent the left side of the human body, while the second and fourth rows depict the right side. This arrangement facilitates 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 6c,g showcase the left and right arms, respectively, 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.

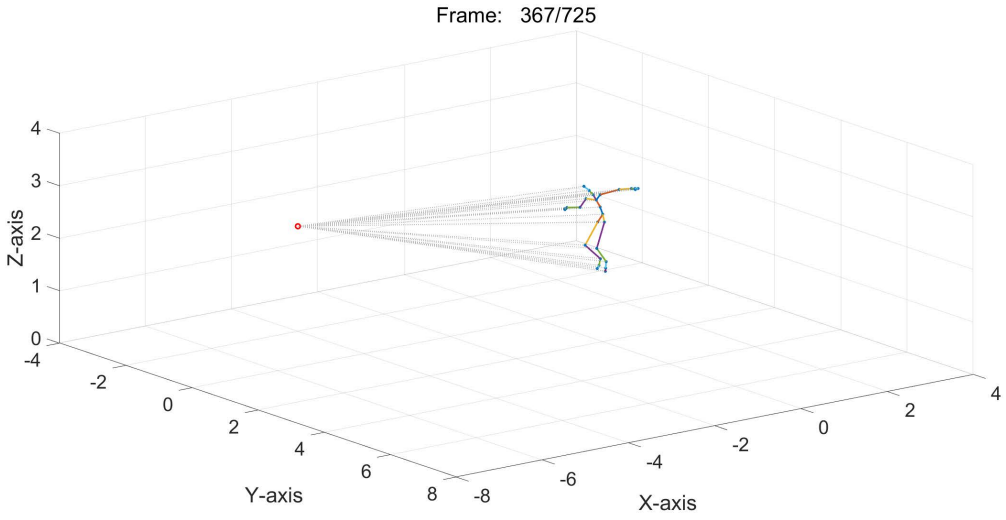

Figure 4. Simulated experimental scene from MOCAP data.

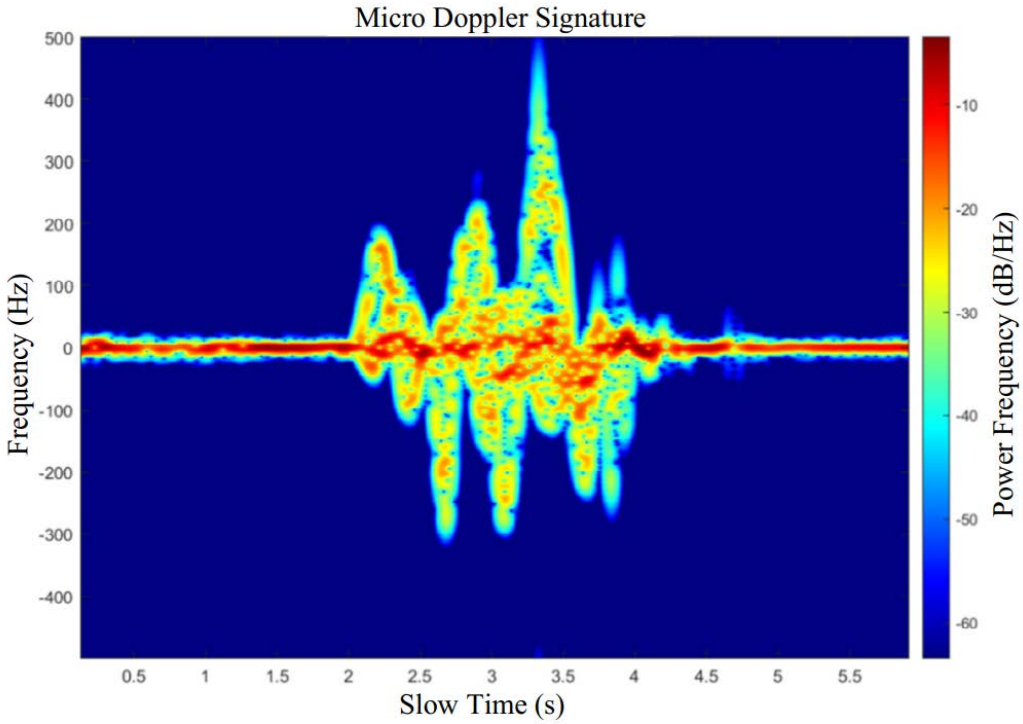

Figure 5. The micro-Doppler spectrum of human motion.

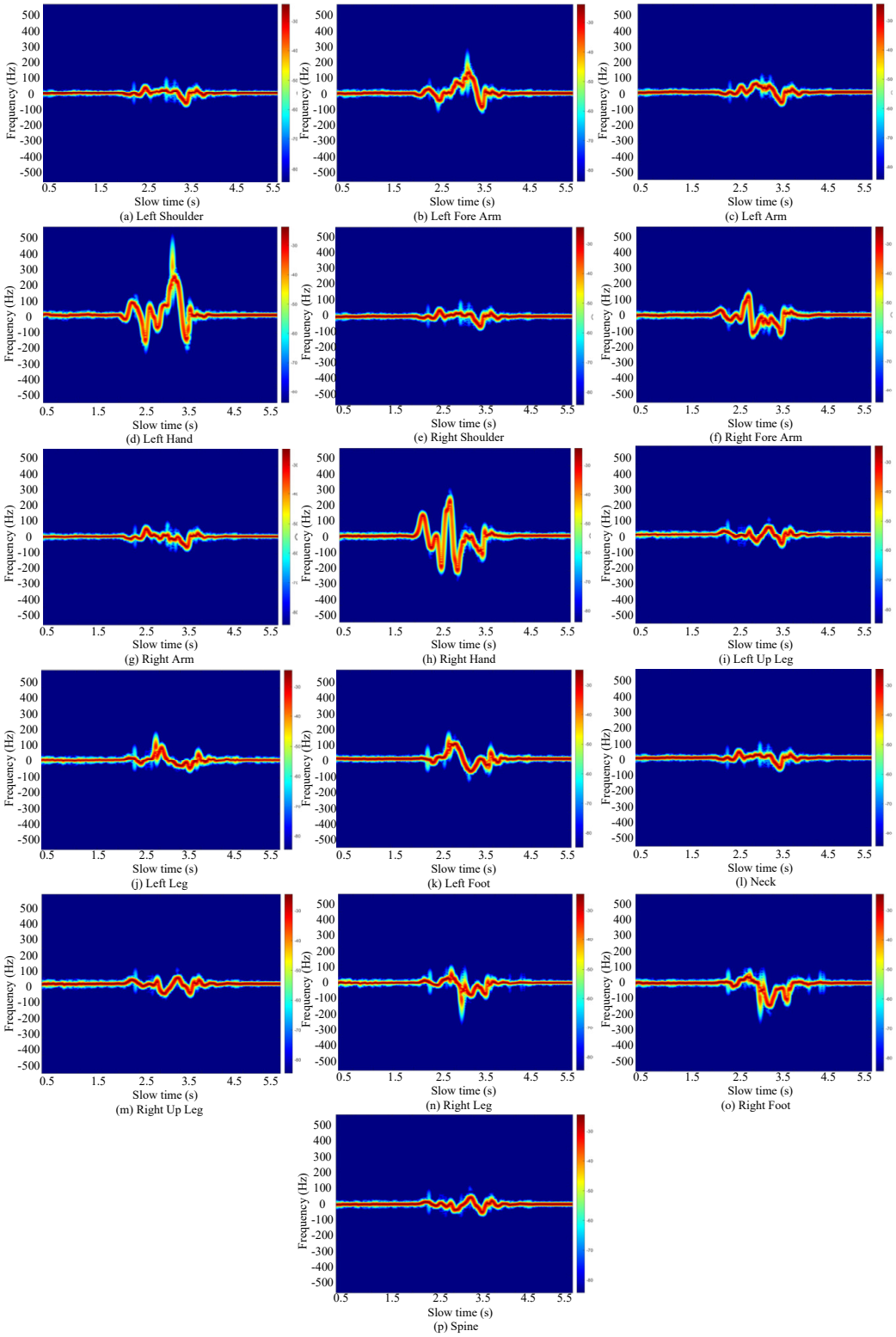

Figure 6. The micro-Doppler spectrum of individual human body parts for movement analysis.

It is also evident that the upper and lower arms, thighs, and calves on both sides exhibit larger motion amplitudes during human movement, resulting in higher MD frequencies. The human body's inherent symmetrical structural characteristics are also mirrored in the corresponding MD spectra. 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. Thus, the MD features of moving human subjects can be instrumental in addressing the challenge of human pose reconstruction.

3. Method and Approach

This section is dedicated to the estimation of human keypoint locations in three-dimensional space utilizing a radar PoseLifter that operates on SISO UWB radar technology. The approach involves taking a sequence of two-dimensional points, $\boldsymbol{x} \in \mathbb{R}^{2n}$, derived from a two-dimensional human pose estimation detector as input, to generate an output sequence of three-dimensional space points $\boldsymbol{y} \in \mathbb{R}^{3n}$. The primary aim is to devise a mapping function $f^* : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{3n}$ to minimize the prediction error across a dataset comprising N poses.

$$f^* = \min_f \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(\boldsymbol{x}_i) - \boldsymbol{y}_i) \tag{16}$$

where x_i is obtained from the detector of 2D human pose estimates. The f^* is the architecture of the radar-based human PoseLifter through the micro-Doppler spectrum.

3.1. Signal Preprocessing of Radar Human Posture Echo

Figure 7 shows the radar data preprocessing chain chart. The processing chain starts from raw radar data, which then undergoes clutter suppression and noise reduction techniques. These preprocessing steps are essential to enhance the signal-to-noise ratio and mitigate the impact of unwanted interference. The signal processing pipeline initiates with the application of Fast Fourier Transform (FFT) along the fast-time dimension to the raw radar data, yielding a Range bin. Subsequently, FFT is employed along the slow-time dimension to derive multiple Range–Doppler maps, as shown in Figure 7, termed Range-FFT and Doppler-FFT. The Range-FFT, processed along each chirp in the original data matrix, facilitates the computation of target distance, while the Doppler-FFT, executed along each distance unit, is instrumental in determining target velocities. Each element in the resulting Range–Doppler maps, referred to as "Range Bin," is represented in the frequency domain and expressed in decibels. Following this, the summation of Range Bins along the range axis for each Range–Doppler map is performed, yielding a vector comprising L Doppler Bins. Subsequently, the concatenation of n consecutive frames forms a time-length n frames micro-Doppler signature map (time-Doppler spectrogram).

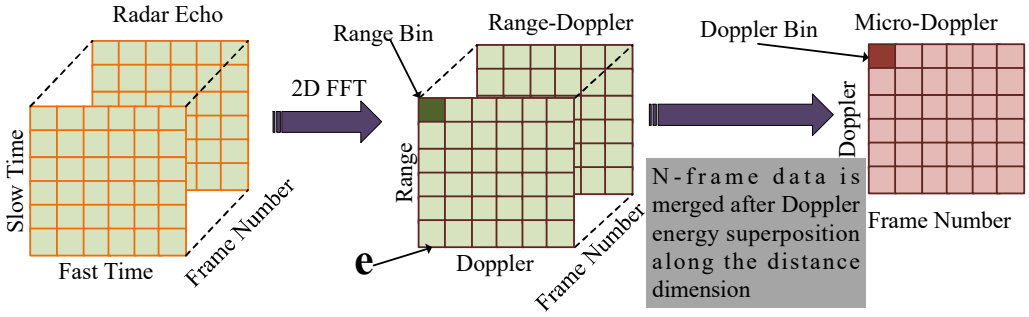

Figure 7. Radar data preprocessing chain chart.

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 HPSUR dataset, which comprises over 311,963 frame radar signatures from four types of human activities. Figure 8 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.

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 due to the distinct features of micro-Doppler signals and leveraging insights from deep learning.

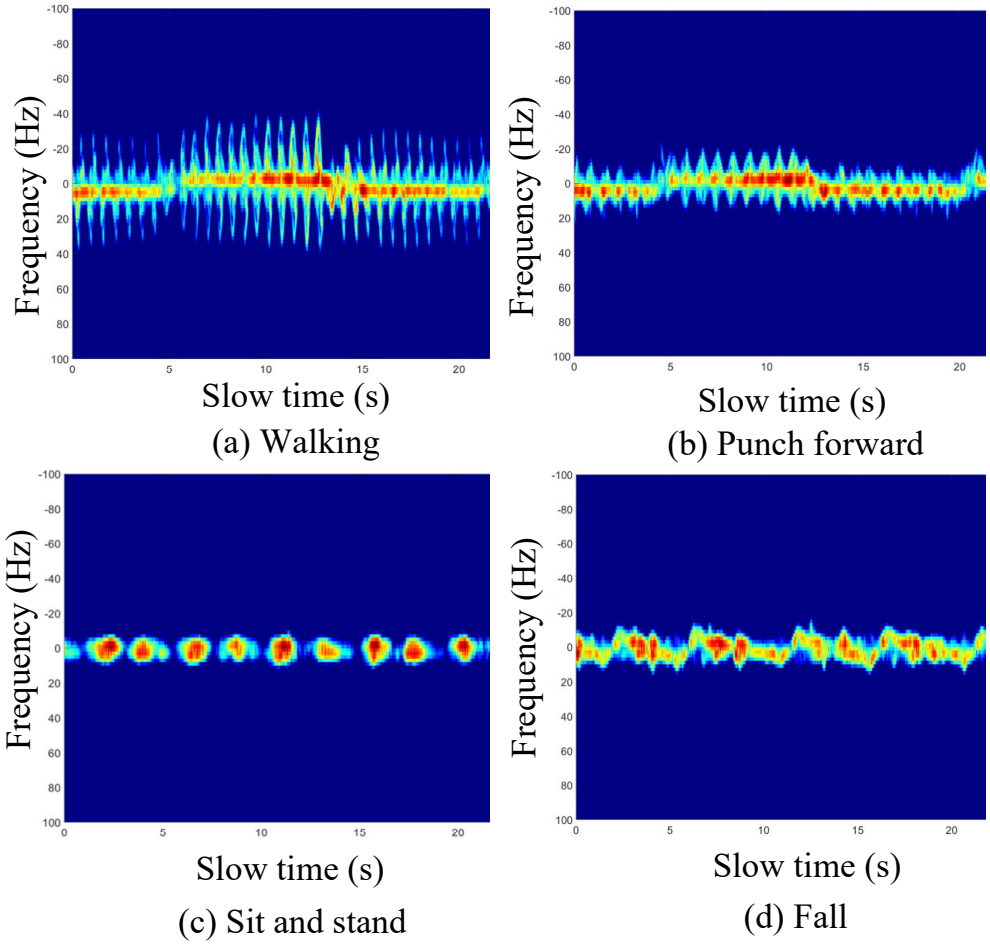

Figure 8. The micro-Doppler spectrum of four different human activities of the HPSUR dataset.

3.2. 2D Key-Points Estimator

The Swin Transformer network, designated as the Micro-Doppler Swin Transformer (MDST), is utilized to estimate 2D human keypoints by analyzing micro-Doppler signatures. The MDST framework integrates both window-based and shift window-based multi-head self-attention mechanisms, facilitating the comprehensive capture of micro-Doppler signatures' inner-frame and intra-frame dynamics for radar-based human pose estimation. These signatures are initially encoded as paths and subsequently segmented into discrete path blocks to improve the network's ability to learn and accurately characterize micro-Doppler signatures.

Within this framework, each path is conceptualized as a 'Token', which forms the foundational data structure for the transformer's input. The MDST architecture is structured into four sequential stages, which collectively form the backbone of the human pose estimation network. The initial stage, or 'Stage 1', amalgamates patch and position embedding outputs while maintaining a consistent number of tokens (H/4, W/4) through linear embedding. This is followed by 'Stage 2', which executes patch merging and feature transformation. This process is duplicated in 'Stage 3' and 'Stage 4', yielding output resolutions of H/16 × H/16 and H/32 × H/32, respectively.

To ensure an equitable comparison with other vision Transformers under analogous conditions, we adhere to the stage, block, and channel configurations of the original Swin Transformer. This approach is applied to two distinct configurations of the MDPST and MDCST, maintaining consistency with the established Swin Transformer structure while exploring its application in micro-Doppler analysis.

- MDPST-T & MDCST-T (Tiny): $C = 96$, layer numbers = $\{2, 2, 6, 2\}$, number heads = $\{3, 6, 12, 24\}$.
- MDPST-B & MDCST-B (Base): $C = 128$, layer numbers = $\{2, 2, 18, 2\}$, number heads = $\{4, 8, 16, 32\}$.

where C denotes the number of channels in the hidden layers of the first stage and the layer numbers refers to the count of blocks within each stage.

3.3. Radar-Based Human PoseLifter Network

The radar-based Human PoseLifter network is a fully convolutional architecture with residual connections designed to process a sequence of 2D poses derived from a 2D keypoint estimator. It leverages a Temporal Convolutional Network (TCN) to facilitate parallelization across both batch and temporal dimensions, capitalizing on the limitations of Recurrent Neural Networks (RNNs), which lack such temporal parallelization due to their inherent sequential processing nature. The input layer concatenates the (x, y) coordinates of J joints for each frame, initiating a temporal convolution with a kernel size of W and yielding C output channels. Subsequently, four ResNet blocks, augmented by skip connections, are employed. Each block conducts a 1D convolution with kernel size W and dilation factor $D = W^B$, followed by convolution, normalization, rectified linear unit activation, and dropout, as delineated in Figure 9. The exponential expansion of the receptive field by a factor of W is achieved within each block while maintaining a linear growth in parameter count. Hyperparameters W and D are judiciously chosen to ensure that the receptive field for any output frame delineates a tree structure encompassing all input frames. Finally, a 1D fully convolutional network is deployed to comprehensively forecast 3D poses for all frames in the input sequence, integrating temporal context.

During the training process, the external knowledge is injected into the PoseLifter network to satisfy the human body's skeletal structure. This incorporation of knowledge-guided learning effectively imposes constraints on the joints' long-range dependencies and spatial arrangements. Consequently, a Mean Squared Error (MSE) loss function augmented with a bone length ratio—termed as Bone Loss—is employed, which is MSE with weights across different bone.

$$\text{BoneLoss} = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N}\left\|\alpha_n Y_n^{(t)} - \hat{Y}_n^{(t)}\right\|_2 \tag{17}$$

where t is the batch size ($T = 24$) during training, N is the total number of joints ($N = 17$), $Y_n^{(t)}$ donates the 3D ground truth coordinate of the $n-$joint, $\hat{Y}_n^{(t)}$ donates the 3D predicted coordinate of the $n-$joints, and α_n is weights across different bone.

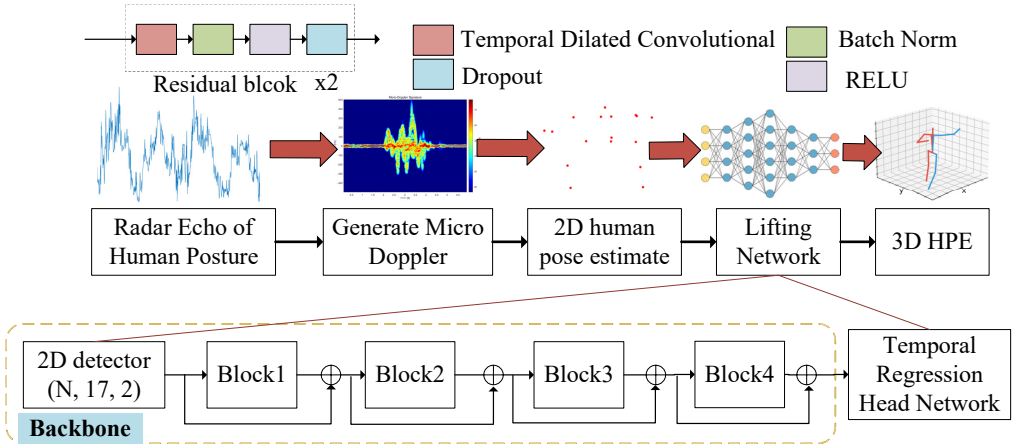

Figure 9. Schematic overview of the methodological framework for 3D human pose estimation.

4. Experimental Description and Results

4.1. Datasets, Annotation, and Evaluation Metrics

We evaluated our model using the HPSUR dataset, collaboratively captured with a SISO UWB radar system and an N3 system, with specific radar parameters detailed in Table 1. The dataset encompasses a comprehensive indoor environment, specifically a living room with three rooms and two halls, featuring four distinct indoor movement scenarios, as detailed in Table 2. We amassed 311,963 data frames, contributed by five subjects varying in height and weight. Each subject executed four types of actions, as enumerated in Table 2 and Figure 10 , within a controlled visual environment. The dataset was partitioned into training and testing subsets for our experimental design. The training set, comprising data from three subjects, totaled 189,462 frames, while the testing set, encompassing the remaining two subjects, consisted of 122,401 frames. Ground Truth (GT) for human pose keypoints was acquired using the N3 system, which precisely captures 17 keypoints of the human skeleton.

Table 1. Specification of SISO UWB radar system parameters.

Parameters	Values
Frequency	2.7∼3.2 GHz
Bandwidth	500 MHz
Pulse width	4.4×10^{-4}
Transmitted signal	FMCW
Pulse repetition frequency (PRF)	1923

We employ the Mean Per Joint Position Error (MPJPE), Procrustes-aligned Mean Per Joint Position Error (P-MPJPE), and Normalized Mean Per Joint Position Error (N-MPJPE) metrics to evaluate the accuracy of our estimated human 2D poses against the GT under the HPSUR dataset. The MPJPE is a fundamental metric for assessing the accuracy of 3D human pose estimation. It is computed by measuring the Euclidean distance between the predicted and true 3D joint positions and then averaging these distances across all joints. MPJPE reflects the average error in joint positioning, encompassing both global and local accuracy. The P-MPJPE is an enhanced variant of MPJPE that incorporates a Procrustes analysis alignment step before error calculation. This process mitigates the impact of global positional, rotational, or scaling inaccuracies, which primarily evaluates the relative accuracy of predicted joint positions independent of global pose accuracy. N-MPJPE is another variant of MPJPE and predicted that the pose undergoes a different form of normalization (such as scaling to a fixed size or normalization according to a certain

standard) before error calculation. This metric provides insights into the accuracy of pose scaling and global positioning but does not account for rotational or translational errors.

Table 2. Detailed description of different human postures of HPSUR dataset.

ID	Type of Posture	Specific Description
1001	Walking	Walk back and forth along the radar radially, walk back and forth along the radar diagonally at 45 degrees, and walk back and forth along the radar diagonally at 135 degress.
1002	punch forward	Walk radially along the radar and back and forth with fists, walk diagonally 45 degrees along the radar and walk back and forth with fists, and walk diagonally 135 degrees along the radar and walk back and forth with fists.
1003	Sit and stand	Take the radar as the origin and make siting and standing posture at (0 m, 2 m), (0 m, 3 m), (−1 m, 2 m), (1 m, 3 m).
1004	Fall	Take the radar as the origin and perform falling motion at (0 m, 2 m), (0 m, 3 m), (−1 m, 2 m), (1 m, 3 m).

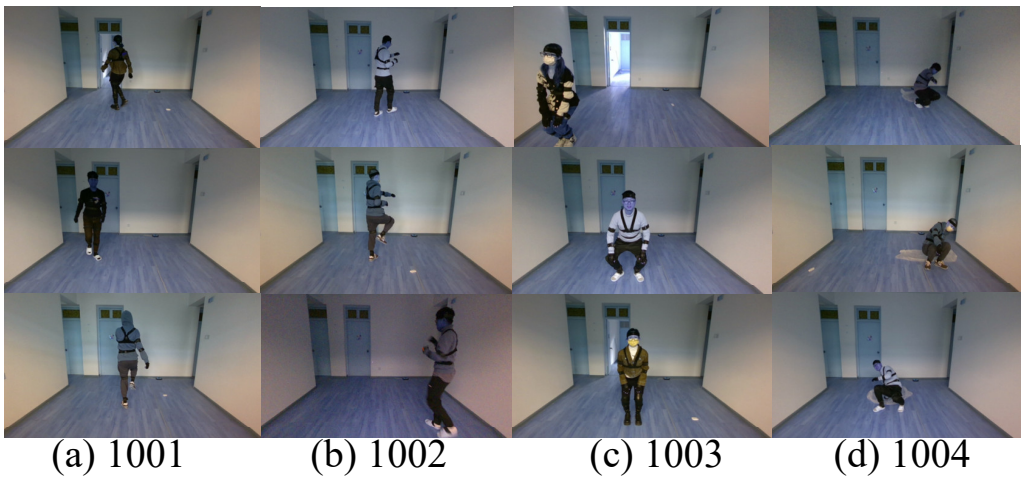

Figure 10. Illustration of data collection scenarios of HPSUR dataset.

4.2. Quantitative Results

Our method, based on direct regression from 2D joint coordinates, naturally depends on the quality of the output of a 2D pose estimator, named the MDST network, which achieves human pose estimation errors within 40mm, as shown in Table 3. The MD-CST-T model shows a mean error of 37.49 mm and the least minimum error at 10.98 mm, suggesting strong accuracy. The MD-CST-B is comparable, with a slightly lower mean error of 36.37 mm and a minimum error of 10.11 mm. The MDPST-T and MDPST-B models have mean errors of 37.82 mm and 37.62 mm, respectively, and demonstrate slightly higher minimum errors of 12.45 mm and 12.18 mm. Therefore, the method adopted for 2D human keypoint estimation is the MDCST-B network in this paper.

Table 4 present a comparative evaluation of radar-based human pose estimation methods. The RadarFormer method outperforms others, with the smallest error at 33.5 mm, while the RF-Pose has the highest error at 62.4 mm. Other methods, like RF-Pose 3D, mm-Pose, and UWB-Pose, show varying degrees of accuracy with errors of 43.6 mm, 44.67 mm, and 37.87 mm, respectively.

The last three rows of Table 4 show the results of the proposed method for lifting 2D human poses to 3D using micro-Doppler signatures based on Single Input–Single Output (SISO) Ultra-Wideband (UWB) radar, with varying numbers of blocks. The results indicate a decrease in MPJPE from 40.26 mm for a single block configuration to 38.62 mm when utilizing two blocks. A marginal increase to 39.86 mm was observed with the integration of four blocks. Additionally, the P-MPJPE, which may represent a variant of the error measurement adjusted for certain conditions or normalized in some way, shows a decreasing trend with more blocks initially, from 31.69 mm (one block) to 31.04 mm

(two blocks), but then slightly increases to 32.31 mm (four blocks). The N-MJPE, possibly a normalized version of MPJPE, presents similar performance improvements with two blocks (38.17 mm) over one block (39.70 mm) and a slight increase for the four-block configuration (39.64 mm).

Table 3. Two-dimensional radar-based human pose estimation based on MDCST and MDPST models (unit: mm).

Method	Mean	Variance	Maximum	Minimum
MDCST-T	37.49	6.32	70.63	10.98
MDCST-B	36.37	6.31	70.42	10.11
MDPST-T	37.82	5.82	68.71	12.45
MDPST-B	37.62	5.89	69.18	12.18

Table 4. Comparative evaluation of radar-based human pose estimation methods (unit: mm).

Method	MPJPE	P-MPJPE	N-MPJP
RF-Pose [18]	62.4	-	-
RF-Pose 3D [19]	43.6	-	-
mm-Pose [1]	44.67	-	-
UWB-Pose [21]	37.87	-	-
RadarFormer [26]	33.5	-	-
Ours (1 block)	40.26	31.69	39.70
Ours (2 blocks)	38.62	31.04	38.17
Ours (4 blocks)	39.86	32.31	39.64

Overall, our method demonstrates an improved performance over traditional methods when using two blocks, which suggests that this configuration strikes a good balance between complexity and accuracy in pose estimation using micro-Doppler signatures from SISO UWB radar systems.

Table 5 presents a comparative analysis of the 3D human pose reconstruction accuracy for four different postures, employing MPJPE, P-MPJPE, and N-MPJPE as the evaluation metrics. The reconstruction performance is quantified in millimeters (mm) and is dissected across three experimental configurations: 1 block, 2 blocks, and 4 blocks.

The results show that the introduction of additional blocks generally improves the pose estimation accuracy. For instance, the 1001 posture demonstrates a progressive decrease in MPJPE from 34.13 mm for 1 block to 33.95 mm for 4 blocks, indicating an enhancement in the precision of pose reconstruction with increased complexity. Notably, the 1004 posture exhibits the highest error reduction when the configuration is shifted from 1 block to 4 blocks, with MPJPE decreasing from 77.77 mm to 77.31 mm, P-MPJPE from 57.33 mm to 58.51 mm, and N-MPJPE from 77.17 mm to 76.52 mm. This suggests a significant dependency of reconstruction accuracy on the complexity of the employed blocks, especially for more challenging postures.

In contrast, the 1003 posture shows minimal variation in MPJPE, decreasing from 33.85 mm to 33.35 mm as the block configuration increases, indicating a lower sensitivity to the number of blocks used. The P-MPJPE and N-MPJPE metrics follow a similar trend, with modest improvements from 25.94 mm to 26.34 mm and from 33.64 mm to 33.37 mm, respectively, for 1 block versus 4 blocks. These findings underscore the necessity of optimizing the block configuration according to the specific posture to achieve the most accurate 3D pose reconstruction.

Table 5. Comparative analysis of 3D human pose reconstruction for four different postures (unit: mm).

Metrics (mm)	Postures	1 Block	2 Blocks	4 Blocks
MPJPE	1001	34.13	32.75	33.95
	1002	48.86	47.57	48.13
	1003	33.85	32.19	33.35
	1004	77.77	74.64	77.31
P-MPJPE	1001	28.15	27.59	29.07
	1002	39.44	38.65	39.49
	1003	25.94	25.21	26.34
	1004	57.33	56.87	58.51
N-MPJPE	1001	33.89	32.65	34.07
	1002	46.52	45.23	46.77
	1003	33.64	32.12	33.37
	1004	77.17	73.99	76.52

4.3. Quantitative Results

Figure 11 illustrates the visualization of the 3D human pose reconstruction process derived from radar-based estimations. Each row corresponds to one of the four postures investigated, labeled as 1001, 1002, 1003, and 1004, respectively. The first column displays the micro-Doppler signatures, which encode the motion dynamics of the human subject. These signatures exhibit distinct patterns characteristic of the different motion captured.

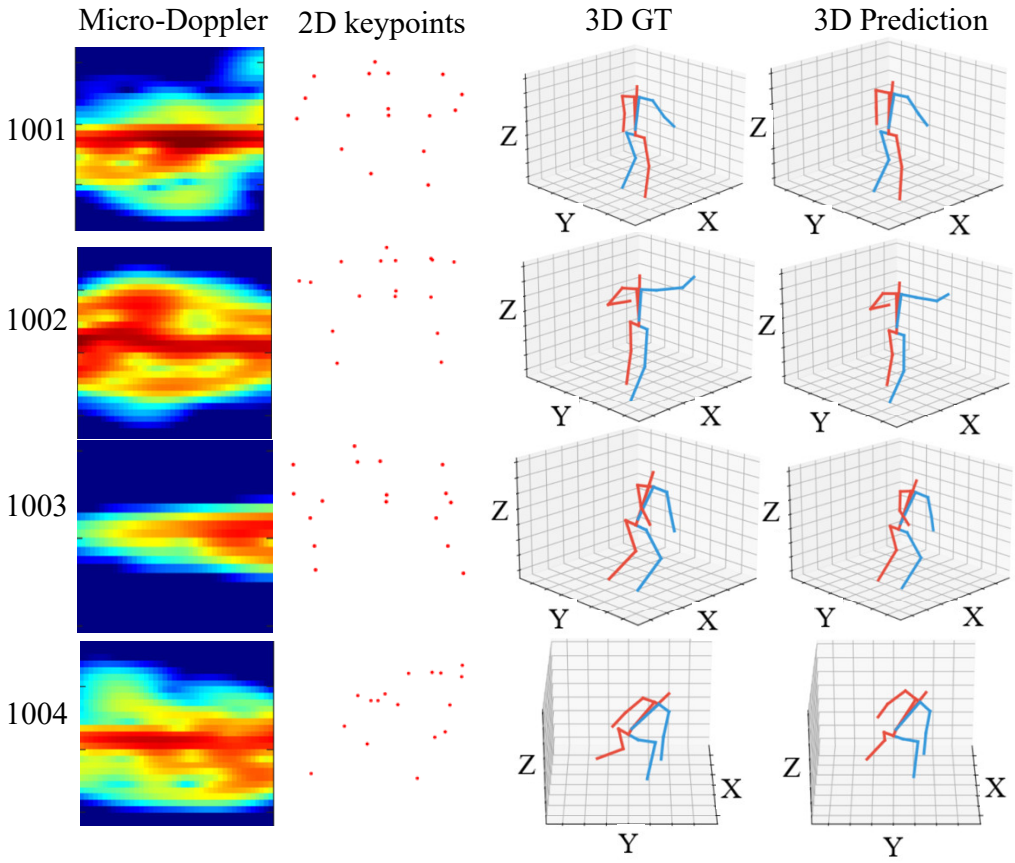

Figure 11. Comparative visualization of 3D human pose reconstruction from radar-derived 2D keypoints.

The second column represents 2D keypoints, extracted features from the micro-Doppler data serving as the foundation for constructing the 3D pose. Although these points are scattered in the 2D space, the underlying spatial relationships indicate the pose's

structural framework. Columns three and four present the ground truth (3D GT) and the predicted 3D poses, respectively. The 3D GT models, depicted with red lines, serve as benchmarks for evaluating the accuracy of the 3D pose predictions, shown with blue lines. A visual inspection reveals a close resemblance between the predicted poses and their corresponding ground truths, suggesting high accuracy in the pose estimation process. However, the subtle discrepancies observed, particularly in the complex postures of 1002 and 1004, highlight the challenges inherent in radar-based human pose estimation.

The grid structure imposed on the 3D plots provides a reference for depth perception, allowing a more apparent appreciation of the position and orientation of limbs in space. This visual analysis validates the proposed method's effectiveness and demonstrates the potential of radar technology in capturing and reconstructing complex human movements in three-dimensional space.

4.4. Ablation Study

Table 6 evaluates the contribution of various components within our proposed radar-based pose estimation method. The ablation study is designed to quantify the impact of individual components on the performance, measured in terms of MPJP, P-MPJP, and N-MPJPE, with all values reported in millimeters (mm).

The results demonstrate that our complete method achieves an MPJPE of 38.62 mm, P-MPJPE of 31.04 mm, and N-MPJPE of 38.17 mm. Removing batch normalization (w/o batch norm) significantly degrades performance, increasing MPJPE to 101.69 mm, P-MPJPE to 47.39 mm, and N-MPJPE to 60.67 mm. This underlines the batch normalization's critical role in model regularization and training stability. The exclusion of the residual connections (w/o residual) also results in performance deterioration, with an increase in MPJPE and N-MPJPE to 55.23 mm and 53.81 mm, respectively, emphasizing the importance of residuals in learning complex functions and enabling deeper architectures.

A further combined removal of both batch normalization and residual connections (w/o residual w/o batch norm) leads to the most pronounced decrease in pose estimation accuracy, with MPJPE soaring to 94.91 mm, P-MPJPE to 78.28 mm, and N-MPJPE to 94.25 mm. Comparatively, the simplified versions of our method with only 1 block yield an MPJPE of 40.26 mm, slightly higher than our full model with 2 blocks (ours), at 38.62 mm, suggesting that a more complex model structure does not necessarily compromise efficiency.

Table 6. Assessment of component impact on pose estimation accuracy in the proposed method (unit: mm).

	MPJPE	**P-MPJPE**	**N-MPJPE**
Our method	38.62	31.04	38.17
w/o batch norm	101.69	47.39	60.67
w/o residual	55.23	44.18	53.81
w/o residual w/o batch norm	94.91	78.28	94.25
1 block	40.26	31.69	39.70
2 blocks (ours)	38.62	31.04	38.17

The results of this study highlight the critical importance of each component in realizing the enhanced accuracy of the proposed approach. It is confirmed that the integrated effect of batch normalization, residual connections, and an optimally determined number of blocks plays a crucial role in the exceptional performance of our pose estimation framework.

Table 7 details the ablation study outcomes, elucidating the influence of distinct model components on radar-based human pose estimation accuracy across four different postures. The 'Our method (2 blocks)' is the baseline, with the overall MPJPE recorded at 38.62 mm. Upon removal of batch normalization, a substantial increase in error is observed, with the overall MPJPE surging to 101.69 mm, signifying the vital role of this component in the model's ability to generalize across different postures. The omission of residual connections

results in an overall MPJPE of 55.23 mm, indicating their importance in capturing the hierarchical structure of human poses. The combined absence of both batch normalization and residual connections further exacerbates the error, inflating the overall MPJPE to 94.91 mm, which underscores the compounded benefits of these components.

Table 7. Model performance across varying postural archetypes using different architectural components (unit: mm).

Metrics (mm)	Motions	Our (2 Blocks)	w/o Batch Norm	w/o Residual	w/o Residual w/o Batch Norm	1 Block	4 Blocks
MPJPE	Overall	38.62	101.69	55.23	94.91	40.26	39.86
	1001	32.75	106.95	49.52	97.28	34.13	33.95
	1002	47.57	102.59	68.88	112.38	48.86	48.13
	1003	32.19	90.67	45.93	72.76	33.85	33.35
	1004	74.64	123.39	94.99	147.82	77.77	77.31
P-MPJPE	Overall	31.04	47.39	44.18	78.28	31.69	32.31
	1001	27.59	46.21	40.69	84.99	28.15	29.07
	1002	38.65	61.06	53.74	98.46	39.44	39.49
	1003	25.21	34.94	36.85	56.75	25.94	26.34
N-MPJPE	Overall	38.17	60.67	53.81	94.25	39.70	39.64
	1001	32.65	59.66	47.75	96.88	33.89	34.07
	1002	45.23	72.03	66.25	111.71	46.52	46.77
	1003	32.12	45.87	45.18	72.31	33.64	33.37
	1004	73.99	106.91	94.13	145.20	77.17	76.52

The performance variations across different posture are also notable. For instance, the 1004 posture exhibits the most significant increase in MPJPE when batch normalization is removed, rising from 74.64 mm to 123.39 mm. This highlights the component's criticality in complex pose estimations. The impact of reducing the model to '1 block' is relatively less severe, with an overall MPJPE increase to 40.26 mm, whereas employing '4 blocks' slightly improves the baseline to 39.86 mm. Similar trends in P-MPJPE and N-MPJPE metrics are observed, with 'Our method (2 blocks)' consistently outperforming the configurations where critical components are omitted. This indicates that the architecture of our method is optimally balanced for the diversity of postures encountered in radar-based pose estimation.

5. Discussion

The Radar PoseLifter network introduced in this paper embodies a novel approach that marries the precision of micro-Doppler signatures with the computational prowess of fully convolutional neural networks. The network's ability to handle long-range dependencies and its design tailored for dilated temporal convolutions set it apart from conventional pose estimation techniques. When compared with WiFi-based sensing systems and traditional camera surveillance, our method surmounts privacy concerns and overcomes limitations of spatial resolution.

High-frequency radar systems, such as millimeter-wave (mmWave) and terahertz radars, offer greater precision in capturing postures due to their shorter wavelengths. However, they lack the ability to penetrate walls and furniture, which is a critical drawback for applications in urban environments where obstructions are common. The proposed method overcomes this by using micro-Doppler signatures to accurately estimate human poses, demonstrating improved performance in these challenging environments, as evidenced by the comparative results with mm-Pose found in Table 4. In contrast, 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. Moreover, these methods depend on MIMO radar imaging,

and the quality of the radar images can be greatly affected by environmental changes and the varying distance between the radar and the human subject. Therefore, we employ SISO UWB radar technology to detect human poses through micro-Doppler signatures, which are less influenced by such environmental and subject-related variabilities. We also conducted a comprehensive comparison between our proposed micro-Doppler method and other existing approaches, including RF-Pose, RF-Pose 3D, and UWB-Pose, as illustrated in Table 4. This comparison effectively validates the efficacy of our method and provides valuable insights and feasible strategies for research in human pose estimation using micro-Doppler features, advancing radar-based sensing for complex human pose estimation in diverse settings.

The implications of our research extend beyond academic interest and into practical applications. For example, in the realm of smart homes and healthcare, our technology could offer non-invasive monitoring of patients or elderly individuals, preserving their privacy while providing critical data for their care. In urban environments, the precision and reliability of our UWB radar-based pose estimation could enhance the safety and efficiency of autonomous vehicle navigation by providing accurate pedestrian dynamics.

6. Conclusions

This paper has presented a novel Radar PoseLifter network, harnessing the capabilities of SISO UWB radar technology with micro-Doppler signature to elevate 2D human pose estimation to 3D reconstructions. This work overcomes traditional challenges in radar-based human motion capture, such as low spatial resolution and the ambiguity of pose reconstruction from 1D radar signals. The network's employment of a fully convolutional architecture with dilated temporal convolutions caters to the efficient processing of long-range dependencies in pose sequences. Our empirical validation on the HPSUR dataset illustrates the method's efficacy in handling diverse human movements and its superiority over existing techniques.

Notably, our approach demonstrates that a two-block configuration in the network achieves an optimal balance between system complexity and estimation accuracy. The ablation studies reinforce the importance of network components like batch normalization and residual connections in minimizing pose estimation errors. Our method's adaptability makes it suitable for a variety of applications where accurate and reliable motion capture is critical, such as autonomous vehicle guidance, urban planning, and healthcare monitoring, while also maintaining privacy. The insights gained from this paper pave the way for future enhancements in radar-based human pose estimation, with the potential for broader application and integration into smart environments.

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. 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]
2. Dai, Y.; Lin, Y.; Lin, X.; Wen, C.; Xu, L. SLOPER4D: A Scene-Aware Dataset for Global 4D Human Pose Estimation in Urban Environments. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 682–692.
3. Li, G.; Zhang, Z.; Yang, H.; Pan, J.; Chen, D.; Zhang, J. Capturing human pose using mmwave radar. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 23–27 March 2020; pp. 1–6.
4. Ning, G.; Zhang, Z.; He, Z. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Trans. Multimed.* **2017**, *20*, 1246–1259. [CrossRef]
5. Xiong, Z.; Wang, C.; Li, Y.; Luo, Y.; Cao, Y. Swin-pose: Swin transformer based human pose estimation. In Proceedings of the 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR), Virtual, 2–4 August 2022; pp. 228–233.
6. 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 (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5686–5696.
7. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23 June 2014; pp. 1653–1660.
8. 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, London, UK, 21–25 September 2022 ; pp. 1–14.
9. Wang, F.; Zhou, S.; Panev, S.; Han, J.; Huang, D. Person-in-WiFi: Fine-Grained Person Perception Using WiFi. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5451–5460.
10. 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]
11. Ding, W.; Cao, Z.; Zhang, J.; Chen, R.; Guo, X.; Wang, G. Radar-based 3D human skeleton estimation by kinematic constrained learning. *IEEE Syst. J.* **2022**, *16*, 3036–3047. [CrossRef]
12. 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]
13. 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), Washington, DC, USA, 28–30 April 2020; pp. 1–6.
14. Wang, K.; Wang, Q.; Xue, F.; Chen, W. 3d-skeleton estimation based on commodity millimeter wave radar. In Proceedings of 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020; pp. 1339–1343.
15. 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.* **2020**, *34*, 8418–8429. [CrossRef] [PubMed]
16. 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]
17. Zeng, Z.; Liang, X.; Li, Y.; Dang, X. Vulnerable Road User Skeletal Pose Estimation Using mmWave Radars. *Remote Sens.* **2024**, *16*, 633. [CrossRef]
18. 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, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7356–7365.
19. Zhao, M.; Tian, Y.; Zhao, H.; Abu Alsheikh, M.; 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.
20. Li, T.; Fan, L.; Yuan, Y.; Katabi, D. Unsupervised learning for human sensing using radio signals. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 3288–3297.
21. Song, Y.; Jin, T.; Dai, Y.; Song, Y.; Zhou, X. Through-Wall Human Pose Reconstruction via UWB MIMO Radar and 3D CNN. *Remote Sens.* **2021**, *13*, 241. [CrossRef]
22. 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*, 1–5. [CrossRef]
23. Zheng, Z.; Pan, J.; Ni, Z.; Shi, C.; Zhang, D.; Liu, X.; Fang, G. Recovering Human Pose and Shape From Through-the-Wall Radar Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–5. [CrossRef]
24. 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]

25. He, Y.; Li, X.; Jing, X. A Mutiscale Residual Attention Network for Multitask Learning of Human Activity Using Radar Micro-Doppler Signatures. *Remote Sens.* **2019**, *11*, 2584. [CrossRef]
26. Kim, Y.W.; Moon, T. Human Detection and Activity Classification Based on Micro-Doppler Signatures Using Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 8–12. [CrossRef]