remote sensing

MDPI

Article

Vulnerable Road User Skeletal Pose Estimation Using mmWave Radars

Zhiyuan Zeng [1,2,*], Xingdong Liang [1,2], Yanlei Li [1,2] and Xiangwei Dang [1]

[1] National Key Laboratory of Microwave Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: zengzhiyuan19@mails.ucas.ac.cn

Abstract: A skeletal pose estimation method, named RVRU-Pose, is proposed to estimate the skeletal pose of vulnerable road users based on distributed non-coherent mmWave radar. In view of the limitation that existing methods for skeletal pose estimation are only applicable to small scenes, this paper proposes a strategy that combines radar intensity heatmaps and coordinate heatmaps as input to a deep learning network. In addition, we design a multi-resolution data augmentation and training method suitable for radar to achieve target pose estimation for remote and multi-target application scenarios. Experimental results show that RVRU-Pose can achieve better than 2 cm average localization accuracy for different subjects in different scenarios, which is superior in terms of accuracy and time compared to existing state-of-the-art methods for human skeletal pose estimation with radar. As an essential performance parameter of radar, the impact of angular resolution on the estimation accuracy of a skeletal pose is quantitatively analyzed and evaluated in this paper. Finally, RVRU-Pose has also been extended to the task of estimating the skeletal pose of a cyclist, reflecting the strong scalability of the proposed method.

Keywords: mmWave radar; skeletal pose estimation; radar signal processing; convolutional neural network

check for updates

Citation: Zeng, Z.;Liang, X.; Li, Y.; Dang, X. Vulnerable Road User Skeletal Pose Estimation Using mmWave Radars. *Remote Sens.* **2024**, *16*, 633. https://doi.org/10.3390/rs16040633

Academic Editor: Massimiliano Pieraccini

Received: 24 October 2023
Revised: 30 January 2024
Accepted: 31 January 2024
Published: 8 February 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Skeletal pose estimation, especially for pedestrians and cyclists, is important for autonomous driving, road monitoring, and other applications. Estimating the current posture of vulnerable road users, such as walking, running, jumping, waving, cycling, etc., can aid in preventative decision making [1]. In addition to the transportation domain, skeletal pose estimation has attracted much attention in home behavior monitoring, medical patient monitoring, human–computer interaction, and other applications.

Skeletal pose estimation was first studied in computer vision and artificial intelligence. Recently, a large number of solutions have been proposed for 2D and 3D human pose estimation [2–4]. In existing studies on human pose estimation, the sensors used include vision sensors, radar, RFID [5,6], tactile sensors [7], etc. The tactile sensor-based skeletal pose recognition uses the force exerted by the target to obtain data from the tactile sensor. This is a scheme suitable for small indoor applications. RFID-based skeletal pose recognition requires the installation of multiple radio tags on the surface of the test target, and the system is complex and not scalable. Skeletal pose recognition with vision sensors is the prevailing scheme, but existing vision schemes do not work effectively in dark environments, line-of-sight occlusion, and distant targets, which makes it difficult to directly apply vision sensor-based skeletal pose recognition to automatic driving and other road scenarios. Radar is robust to illumination and weather compared to vision sensors, and millimeter-Wave (mmWave) Multiple-Input Multiple-Output (MIMO) radar has a higher resolution than conventional radar, which has been widely used in the autonomous driving of vehicles [8].

Therefore, the radar sensor can be used as a high-quality complementary solution to the visual sensor for the estimation of the user's skeletal pose in the real road environment.

The direct recovery of fine skeletal poses from radar images is difficult due to the low resolution of radar imaging. By applying deep learning techniques, the automatic estimation of skeletal poses can be achieved without hand-crafted features. However, training deep learning models is data greedy and requires a large amount of manually labeled data to guarantee performance [9–11]. On the scattering mechanism of an electromagnetic wave, optical scattering mainly occurs when visible and infrared wavelengths (350–1000 nm) perceived by visual sensors interact with targets. However, the interaction between the microwave frequency band (1 mm–1 m) perceived by the radar sensor and targets mainly takes place with Rayleigh scattering and Mie scattering. The difference in scattering mechanisms makes radar images more abstract than optical images, so it is not possible to manually label radar image datasets as it is for optical images. RF-Pose [12] solves the radar data-labeling problem by introducing cross-modal supervised learning techniques for the first time, synchronously collecting visual and radar data, and using optical data as supervision.

By applying cross-modal supervised learning techniques, the radar-based skeletal pose estimation schemes can be divided into two categories based on the radar system, namely, the planar array radar-based skeletal pose estimation scheme and the linear array radar-based skeletal pose estimation scheme. The planar array radar-based skeletal pose estimation scheme can obtain a 3D point cloud of the target in range–azimuth–pitch space and estimate the skeletal pose based on it. Under certain requirements of estimation accuracy and robustness, the signal processing complexity of this scheme is low, but the radar system is highly complex. Skeletal pose estimation based on a single linear array radar can only obtain 2D spatial information of the target due to the observation dimension limitation. Under certain estimation accuracy and robustness requirements, this scheme has low radar system complexity but high signal processing complexity. In road applications such as automatic driving, automotive radar sensors have stringent requirements on volume, power consumption and cost. Therefore, it is difficult for the scheme based on planar array radar to meet the application requirements. For the scheme based on a single linear array radar, although it meets the requirements at the system level, the absence of the system observation dimension may lead to errors in some cases, which is unacceptable in terms of road traffic safety.

To balance the performance and complexity of radar systems, a distributed non-coherent radar system is proposed in this paper. It consists of two linear radars placed horizontally and vertically. By combining the range–azimuth and range–pitch dimension of the non-coherent signal, the lack of observation dimension of a single linear radar can be overcome, and the system cost and complexity can be greatly reduced compared to a planar array radar. We study the quantitative relationship between the resolution of a radar system and the accuracy of pose estimation based on a distributed non-coherent radar system. Meanwhile, we propose a deep learning model training method based on multi-resolution radar images to achieve remote object pose estimation that cannot be detected by visual sensors. Moreover, the object of pose estimation is extended from pedestrians to cyclists, which makes the results of pose estimation more consistent with the real world. Finally, skeletal pose estimation for vulnerable road users such as pedestrians and cyclists is implemented in this paper. The proposed method is also applicable to long-distance and multi-target environments.

The main contributions of the paper are briefly summarized as follows.

- In terms of radar system, a non-coherent radar system is proposed for the detection of vulnerable road users, and the relationship between the radar resolution and the estimation accuracy of skeletal posture is analyzed.
- In terms of modeling, a deep learning network, called RVRU-Pose, which is applicable to non-coherent radar for skeletal posture estimation, is proposed to achieve an accu-

rate estimation of human posture. Furthermore, based on this model, the estimation of the skeletal pose of the cyclist is also implemented.
- In terms of model generalization, a multi-resolution data augmentation and training method is proposed for radar to achieve pose estimation for long-range and multi-target application scenarios.

The rest of the paper is structured as follows. Section 2 gives a brief overview of related works on fine-grained human sensing tasks. Section 3 discusses the signal collection and preprocessing process with some analysis. Section 4 introduces the proposed RVRU-Pose deep learning architecture and the multi-resolution data augmentation method. Section 5 describes the experimental results. Finally, Section 6 concludes this article.

2. Related Work

This section will intrude the methods for human skeletal posture estimation; they can be divided into two categories, depending on whether their radar system is linear or planar. The planar array radar can obtain azimuthal pitch two-dimensional images similar to optical images, providing a favorable understanding of the human eye and easier extraction of bone pose information. However, the complexity of the system limits its practical application. Linear array radar has low volume and weight cost but few radar observation dimensions and poor reliability in object recognition and bone pose estimation. In order to find a compromise between the performance and complexity of these two radar systems, the fusion of multiple linear array radar systems provides an opportunity.

2.1. Human Skeletal Posture Estimation with Planar Array Radar

The planar array radar antenna elements are arranged in the planar array mode. It can provide 3D spatial information about the distance, azimuth, and pitch of the target, which can then be used to obtain information about the body structure and behavioral attitude of the target, which is similar to optical images.

Existing studies on human skeletal pose estimation based on planar array radar are based on the application of Through-the-Wall Radar Imaging (TWRI), which uses a relatively low electromagnetic frequency band. Ahmad et al. [13] were the first to investigate the 3D ultra-wideband radar imaging technology and implemented near-field 3D imaging of the human body using a beamforming method. Fadel et al. [14] proposed the RF-capture method based on planar array radar. This approach studied the technique of scattering point alignment based on a human motion model and obtained a clear image of the human profile, but only a few types of motion could be distinguished. Song et al. [15] proposed a 3D convolutional network and implemented it to estimate the 3D skeletal pose of a wall-penetrating human body with a 10Tx $\times$ 10Rx array radar in the 1.2 to 2.2 GHz band. However, manual annotation of the radar data is required, and a dataset of 10 human poses for classification is formed based on this [16]. Zheng et al. [17] used a 4Tx $\times$ 8Rx array radar at 0.5 to 1.5 GHz to achieve 2D estimation of a human skeletal pose based on acquired azimuthal 2D radar images by cross-modal supervised learning. Furthermore, Zhang et al. [18] implemented human pose and shape recovery from azimuth–pitch radar images based on multi-task learning.

However, planar array-based methods for human skeletal pose estimation are limited by the large volume and high cost of radar systems, which cannot meet the needs of vehicle applications such as autonomous driving and road monitoring.

2.2. Human Skeletal Posture Estimation with Linear Array Radar

The transceiver antenna elements of the linear array radar are arranged in linear array mode and provide two-dimensional information about the distance and angle of the target. Ding et al. [19] utilized the range–Doppler maps acquired by mmWave radar, combined with kinematic constraints on human joints, to achieve skeletal pose estimation for a single human in indoor scenes. Cao et al. [20] provided a joint global–local human pose estimation network using a mmWave radar Range–Doppler Matrix (RDM) sequence, which

achieves high estimation accuracy compared to using global RDM only. Xue et al. [21] proposed to use the range–azimuth point cloud obtained from mmWave radar and a combination of convolutional neural network (CNN) and Long Short-Term Memory (LSTM) networks to achieve 3D human mesh model reconstruction. Zhong et al. [22] provided a skeletal pose estimation method that combines point convolution to extract the local information and density from the point cloud, and they found that the accuracy of the pose estimation increased.

Zhao et al. [12] proposed RF-Pose for the first time. By integrating the horizontal and vertical radar heatmaps and using cross-modal learning, Zhao obtained 2D results for human skeletal pose estimation. The results are also compared with OpenPose. Based on RF-Pose, Sengupta et al. [23] proposed mm-Pose, which uses regression instead of heatmap-based methods to achieve 3D human skeletal pose estimation. Li et al. [24] applied the RF-Pose method to mmWave radars to achieve an estimation of individual human skeletal poses in open environments. To estimate the skeletal pose of the human body at a short range, radar images from multiple perspectives also need to be fused at this time due to the limited field-of-view angle of radar. Sengupta et al. [25,26] proposed to use two mmWave radars aligned along the height to obtain data, thus obtaining complete echo data of the human body in the height direction. The layout of two radars along the height also appears in ref. [27]. After obtaining the deep learning results, the position of the skeletal joints were adjusted and optimized through the spatial constraints of the joints so that the estimation result of human skeletal pose is closer to the actual situation.

So far, the scenario of skeletal pose estimation with mmWave radar is limited to close-range and wide-open scenarios, and there are no related studies on long-distance road scenes. At the same time, only human subjects have been studied, and there is a lack of research on other road targets. To achieve this goal, we take a step forward in this article.

3. Signal Model and Data Analysis

In this section, the signal collection system is first introduced. Then, the signal processing chain is discussed. Finally, the vulnerable road user detection and extraction method is described before the skeletal posture estimation.

3.1. Signal Collection System

To balance the performance and complexity of the radar system, a distributed non-coherent radar system, consisting of two linear arrays of mmWave radars placed horizontally and vertically, is adopted in this paper. In addition, a depth camera is installed in the signal collection system, as the deep learning training process for skeletal pose estimation relies on cross-modal supervision by cameras. To ensure synchronization of the signal collection in the time domain, the trigger signals of the two radars and the depth camera are synchronized within a negligible error by designed synchronization module. To ensure synchronization of the signal collection in the spatial domain, the geometrical centers of two radars and depth camera are precisely measured, and the viewing angles of each sensor are set to ensure maximum overlap.

To ensure that the two radars can operate simultaneously without interference, a frequency division multiplexing mode is used with one radar operating at 77 to 79 GHz and the other at 79 to 81 GHz. The scanning rate of the radar is set to 10 frames per second. The Intel RealSense D455 depth camera is used to capture color images and depth images for the targets in the environment. The depth camera captures color frames and depth frames at 30 frame per second (FPS) and downsamples the result three times to obtain the RGB-D images consistent with the radar images.

3.2. Signal Processing Chain

The radar signal process pipeline of the raw radar echo signals is shown in Figure 1. The signal processing procedure used by both radars is the same.

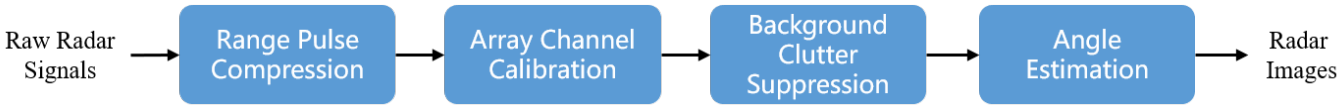


Figure 1. Radar signal process pipeline of the raw radar echo signals.

The echo data from each channel of the Frequency-Modulated Continuous Wave (FMCW) mmWave radar first dechirp to perform range pulse compression for the echo complex signal [28].

$$s(t_r, t_a) = \sigma_0 \exp\left[-\frac{4\pi K_r}{c} R(t_a) \cdot t_r - j\frac{4\pi}{\lambda} R(t_a)\right] + w(t_r, t_a) \tag{1}$$

where t_r is fast time, t_a is slow time, σ_0 is the Radar Cross-Section (RCS) of the target, $R(t_a)$ is the distance from the radar to the target at the slow-time t_a, K_r is the frequency slope, λ is wavelength, and $w(t_r, t_a)$ is noise. The range pulse compression can be completed by FFT of the complex echo signal along the range direction.

$$s(f_r, t_a) = \sigma_0 \text{sinc}\left[f_r + \frac{2K_r}{c} R(t_a)\right] \exp\left[-j\frac{4\pi}{\lambda} R(t_a)\right] + w(t_r, t_a) \tag{2}$$

After the range pulse compression in each channel is completed, the amplitude and phase errors in each channel should be corrected to reduce the bias in the target angle estimation due to the amplitude and phase errors between channels. Some strong reflection targets, such as corner reflectors, can be arranged in the scene before the experiment. The radar system collects the echo signal from the corner reflector and performs pulse compression. After that, the complex information, including the amplitude and phase of the range bin corresponding to the target of the corner reflector in each radar channel is extracted, and a certain radar channel is selected as the reference channel to solve the complex calibration of the other channels.

$$C_{am,ph} = \frac{C_{ref}}{C_i} \tag{3}$$

where C_{ref} is a complex value of the reference channel, C_i is the complex value of the $i-th$ channel to be calibrated, $C_{am,ph}$ is the calculated calibration complex value, including amplitude and phase, $C_{am,ph} = 1$. $C_{am,ph}$ is multiplied by the compressed pulse data of each channel to achieve amplitude and phase error correction between channels.

For skeletal pose estimation applications, where we focus on moving targets, clutter suppression is required to remove the effect of static clutter. Moving Target Indication (MTI) is a common clutter cancellation technique, which uses the structure of a delay canceller to construct a high-pass filter to filter out low-frequency clutter.

Range–azimuth 2D radar images can be obtained by combining multi-channel data to estimate the target angle. For a multi-channel radar system, after each channel is compressed in the range direction pulse, a data matrix of size can be obtained, where each row of the matrix represents a range bin and each column represents a radar channel. For signals within a range bin, the target angle is determined using the Direction of Arrival (DoA) estimation technique. For an array with N elements, the space can be partitioned into K parts [29].

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} \alpha(\theta_1) & \alpha(\theta_2) & \dots & \alpha(\theta_K) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & \dots & 1 \\ \exp\left(-j2\pi\frac{d\sin\theta_1}{\lambda}\right) & \exp\left(-j2\pi\frac{d\sin\theta_2}{\lambda}\right) & \dots & \exp\left(-j2\pi\frac{d\sin\theta_K}{\lambda}\right) \\ \dots & \dots & \dots & \dots \\ \exp\left(-j2\pi\frac{(N-1)d\sin\theta_1}{\lambda}\right) & \exp\left(-j2\pi\frac{(N-1)d\sin\theta_2}{\lambda}\right) & \dots & \exp\left(-j2\pi\frac{(N-1)d\sin\theta_K}{\lambda}\right) \end{bmatrix}\end{aligned} \tag{4}$$

For each channel data vector $\mathbf{x}_{N\times1}$, of some range bin, the mathematical model for DoA estimation in one shot can be formulated as follows.

$$\mathbf{x}_{N\times1} = \mathbf{A}_{N\times K} \cdot \mathbf{s}_{K\times1} + \mathbf{n} \tag{5}$$

where $\mathbf{x}_{N\times1}$ is the rang bin signal after range pulse compress, $\mathbf{A}_{N\times K}$ is an array manifold matrix, $\mathbf{s}_{K\times1}$ is the azimuth signal for a range bin, and $\mathbf{n}$ is noise. Digital Beam Forming (DBF) is one of the most common and effective algorithms to solve the DoA estimation problem with the minimum L2 norm [30].

$$\mathbf{s} = \mathbf{A}^{\dagger}\mathbf{x} \tag{6}$$

where $\mathbf{A}^{\dagger}$ is the Moore–Penrose pseudoinverse of matrix $\mathbf{A}$.

3.3. Vulnerable Road User Detection and Extraction

In existing work on skeletal pose estimation in mmWave radar, the most common approach is to directly use the entire radar image of the monitoring area as input to the deep learning network. This strategy is simple to operate and performs well in small areas of the monitoring area where only a single target exists. However, when the monitoring area is large, the amount of data for the neural network is greatly increased. When there are multiple targets in the monitoring area, directly feeding the entire radar image to the neural network for single target detection will cause serious errors.

To achieve multi-target detection in distant and wide surveillance areas, this paper proposes a method that first detects the interested targets from radar heatmaps using a radar target detection algorithm. Then, skeleton pose estimation is performed individually on each target. The schematic diagram is shown in Figure 2. There are two main approaches for radar target detection. The first approach is based on radar signal detection theory, which selects several target points from the radar heatmaps based on their intensity and performs clustering to distinguish multiple targets. However, this approach fails to differentiate between target types, whereas we are interested only in pedestrian targets. The second approach is based on deep learning, where models like YOLO [31] are trained on radar heatmaps to achieve accurate detection and recognition of pedestrian targets. However, this method has higher computational complexity. In this paper, a combination of these two approaches is adopted. When there are only moving pedestrian targets in the scene, the first approach is used for target detection. When there are multiple types of moving targets in the scene, the second approach is employed.

After target detection, the centroid position of each target is selected, and a region of interest with dimensions of 2 m $\times$ 2 m $\times$ 2 m is extracted around the centroid. The spatial size of pedestrians is generally not larger than 2 m, so a size of 2 m is chosen as the dimension for the region of interest. The skeleton pose estimation is then performed on the radar sub-image of the region of interest instead of directly detecting the entire radar heatmap. However, the cropped radar sub-image loses the global localization information of the target. Taking inspiration from range–Doppler (RD) positioning in Synthetic Aperture Radar (SAR) imaging [32], this paper also preserves the pixel position of the radar sub-image in the radar heatmap when cropping. Therefore, the input to the skeleton pose estimation network includes not only the cropped intensity heatmap but also the encoded heatmap of the cropping region coordinates.

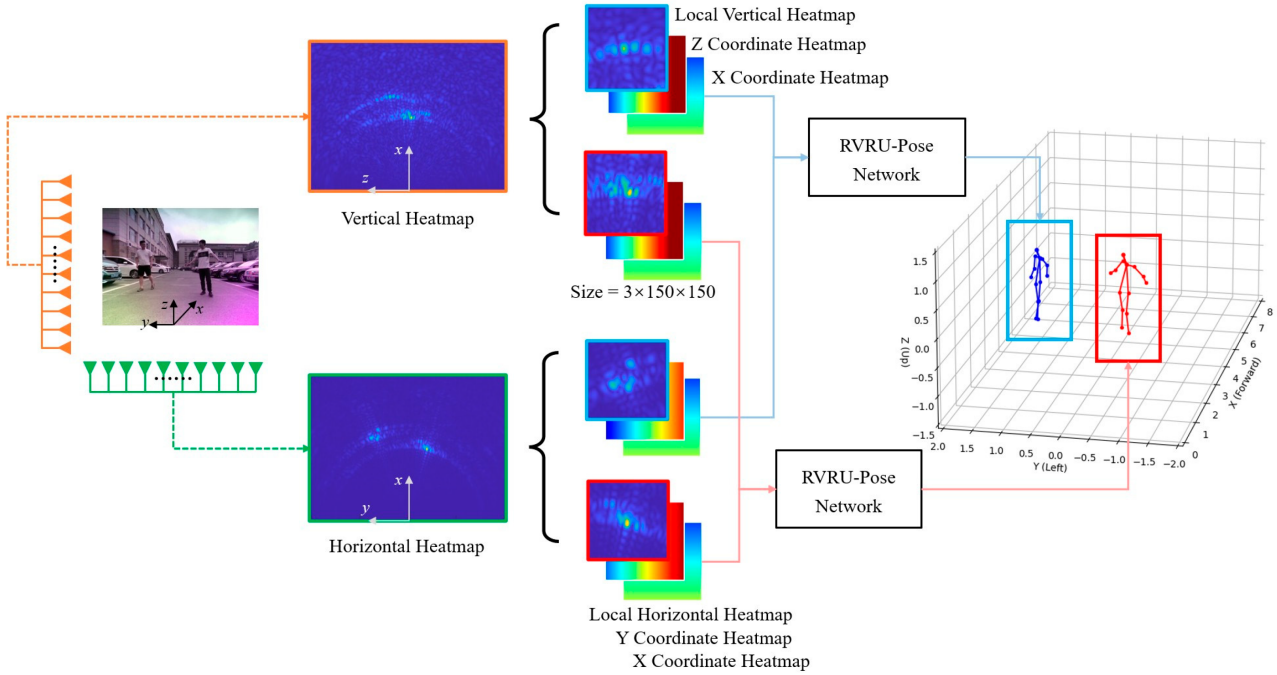


Figure 2. Radar multi-target extraction method and intensity heatmaps–coordinate heatmaps generation method.

4. Methodology

In this section, the pipeline of the proposed method is first given. Then, the cross-modal supervised network RVRU-Pose is illustrated. Finally, a multi-resolution data enhancement and training method for radar is proposed.

4.1. Method Pipeline

The complete process of the proposed method is shown in Figure 3. It mainly includes four parts: Radar Signal Process, Targets Detection and Extraction, Teacher Network (OpenPose), and Student Network (RVRU-Pose).

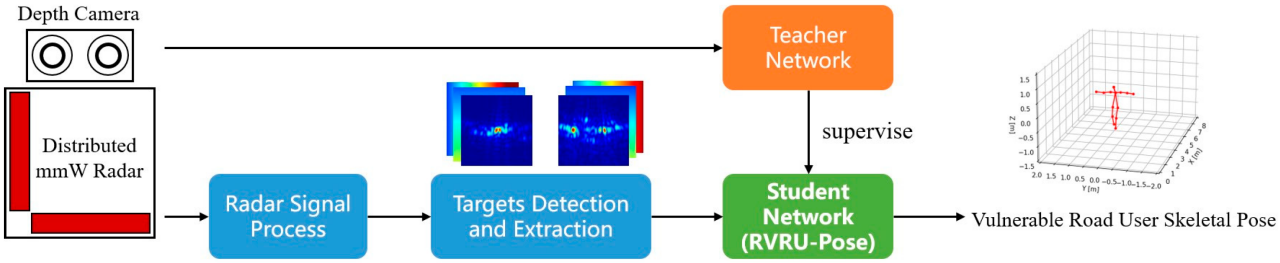


Figure 3. Block diagram of skeletal pose estimation method for vulnerable road user.

In the training stage, the method of cross-modal supervised learning is used. In the open scenario, vulnerable road users such as humans are assigned as subjects. Then, the distributed radar and depth camera synchronously collect experimental data to obtain paired 3D optical image set $S_{visual3D} = \left\{ s_{visual3D}^{(n)} \right\}_{n=1}^{N}$, horizontal radar image set $S_{radarH} = \left\{ s_{radarH}^{(n)} \right\}_{n=1}^{N}$, and vertical radar image set $S_{radarV} = \left\{ s_{radarV}^{(n)} \right\}_{n=1}^{N}$. For the 3D optical image dataset, the computer vision-based deep learning network OpenPose [4] is first trained using 2D color images as the teacher network $TN(\cdot)$ to obtain heatmaps of the target skeletal joints. Then, the pixel coordinates of the skeletal joints in the color image are projected onto the depth image to obtain the 3D spatial coordinates $Pose_{visual} = TN(S_{visual3D})$

of the skeletal joints. By training the teacher network on the massive dataset, $Pose_{visual}$ has sufficient accuracy and robustness to automatically generate supervised labels for the student network. The student network $SN(\cdot)$ has a two-headed input, which is used to input the H radar image and coordinating heatmap sequence and V radar image and coordinating heatmap sequence, respectively, to obtain the 3D spatial coordinate $Pose_{radar} = SN(S_{radarH}, S_{radarV})$ of the skeletal pose joints via regression. Then, the difference between the outputs of the teacher network and the student network is computed using the loss function $Loss(Pose_{visual}, Pose_{radar})$, and the model parameters of the student network are updated by stochastic optimization using the adaptive momentum method. During training, limited by the performance of the depth camera, the distance of the collected data generally does not exceed 5 m in order to ensure the accuracy of the results of the teacher network. However, practical road applications often require long-distance skeletal pose estimation. Therefore, we propose a multi-resolution radar image training method. Based on the relation between the radar angular resolution and the radar array aperture, the radar image of an equivalent distant target is intercepted by the aperture of the radar data, and then the coordinate heatmap is adjusted accordingly to achieve dataset augmentation. After training, accurate 3D skeletal pose estimation results can only be obtained by feeding the H and V radar image sequences into the student network.

4.2. Cross-Modal Supervised Network

In this section, a cross-modal supervised network named RVRU-Pose (Radar Vulnerable Road User Pose) is proposed for skeletal pose estimation. RVRU-Pose uses the radar heat map sequence from the horizontal and vertical radars as input through forked CNN architecture, and it uses data-driven regression to return the three-dimensional spatial coordinates of the skeletal postural joints.

The horizontal and vertical radar heatmaps obtained from the radar are three-channel images, where the three channels are the radar heatmaps of target, the radar forward coordinate heatmaps, and the radar transverse coordinate heatmaps. The size of radar heatmap is set to 128×128. The horizontal and vertical radar heatmap sequences are, respectively, fed into a multi-layer CNN channel with the same structure. The tensors after feature extraction are fused in the channel dimension, and then the combined features of the fused tensors are extracted by the multi-layer CNN channel. Finally, the tensor is expanded into vectors by a fully connected layer to form a mapping relation between the radar heatmap sequence and the 3D coordinates of the skeletal pose. The basic structure of the proposed RVRU-Pose network is illustrated in Figure 4, while Figure 5 exhibits the specific parameters of the RVRU-Pose network.

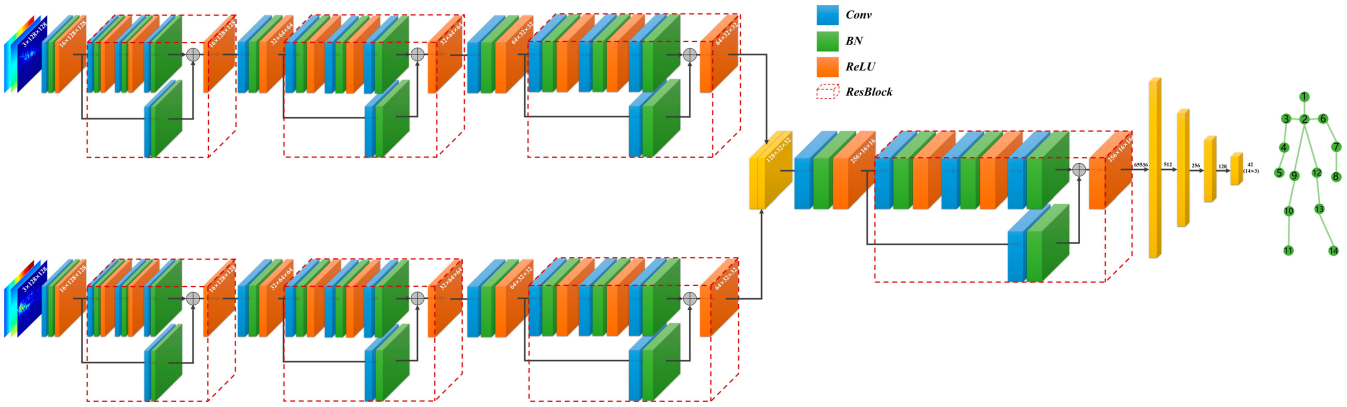


Figure 4. The cross-modal supervised network structure diagram of RVRU-Pose.

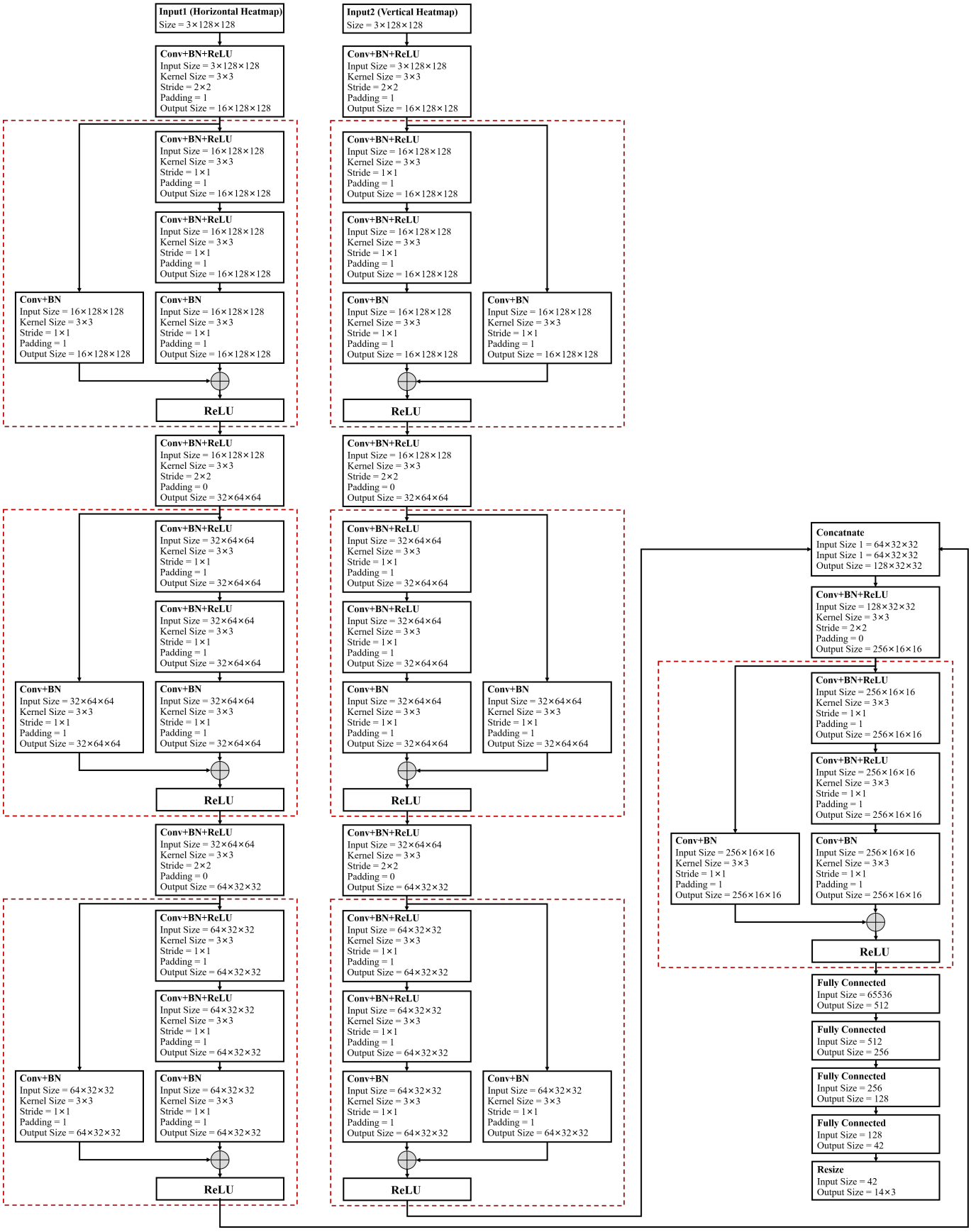


Figure 5. The cross-modal supervised network parameters of RVRU-Pose.

In the deep learning task of optical images, the deep features of an image can be extracted by multi-layer convolution. However, mmWave radar images suffer from issues such as poor angular resolution, sidelobes, and noise interference, which makes the design of mmWave radar deep learning networks different from traditional optical deep learning networks. For the design of deep learning networks, this paper analyzes and designs the convolutional receptive field and the number of convolutional layers by fully considering the characteristics of radar images.

The receptive field in a convolutional neural network represents the region of the input map corresponding to a pixel in the output feature map [18]. If denotes the size of the convolution kernel in the i-th layer, is the step size, and is the receptive field size, then

$$r_i = r_{i-1} + (f_i - 1) \prod_{i=1}^{n-1} stride_i \tag{7}$$

For mmWave radar images, the relationship between sensitivity field and radar resolution must be balanced so that the extracted feature maps contain multiple radar resolution units and fully exploit the resolution information of the radar images. For the range or azimuthal dimension of the radar image, let the physical size corresponding to the receptive field of the radar image be L, and let the radar resolution be ρ; then, $L > \rho$ should be satisfied. Specifically, for the azimuthal resolution $\rho_a = R\frac{\lambda}{D}$, where R is the range, λ is the wavelength, and D is the azimuthal aperture of the radar antenna. When the convolution kernel size is 3 and the step size is 1, the sensitivity field size of the n-th convolution layer is $r_i = 2n - 1$, and $r_i > \frac{\rho_a}{A_{bin}}$ can be obtained according to the constraint $L > \rho$, where ρ_a is the radar azimuth resolution and A_{bin} is the actual length corresponding to an azimuth pixel in the radar image. It follows that the minimum depth n of the convolutional layer should satisfy $n > \frac{\rho_a + A_{bin}}{2A_{bin}}$. Based on the above principles and the adopted mmWave radar system, the convolutional layer number of RVRU-Pose is reasonably designed in this paper.

According to the above analysis, the low resolution of radar images prevents the number of convolutional layers from being too small. However, multi-layer convolution has two problems: namely, the gradient vanishing often faced by deep convolutional networks and the loss of raw information in radar images in deep feature extraction. Raw information such as the intensity contained in a radar image has a reference meaning for distinguishing sidelobes from noise interference. To combine deep and shallow features of radar images, the ResBlock module [33] is introduced in this paper, as shown in Figure 4.

4.3. Multi-Resolution Radar Imaging Training

The training of deep learning networks cannot be separated from large-scale datasets. The strategy of mmWave radar deep learning by using visual cross-modal supervised learning has been widely used in existing works. However, since visual sensors are mainly used for angular resolution, human skeletal pose detection based on visual sensors is generally used for relatively short distances. When the target is far away, the number of pixel units occupied by the target in the image decreases, making skeletal pose detection difficult. As a result, visual sensors can only automatically label datasets of close-range targets, and existing related works mainly focus on indoor close-range application scenarios, which do not leverage the advantages of mmWave radar for long-range detection.

In this paper, a dataset enhancement method based on multi-resolution radar images is proposed for deep learning training. Radar images of targets at different distances are generated by adjusting the radar angular resolution, and the optical detection of the human skeletal pose is shifted synchronously to the target distance for data augmentation.

Based on radar azimuthal resolution

$$\rho_a = R\rho_\theta = R\frac{k\lambda}{D} = R\frac{k\lambda}{(n-1)d} \tag{8}$$

where ρ_a is the azimuth resolution, R is the radar detection range, k is the constant 0.886, λ is the wavelength, D is the radar aperture of the array, and n is the array element number of the uniform array with interval d. It can be seen that increasing the distance R and decreasing the number of elements n can achieve an equivalent effect on reducing the azimuth angle resolution.

In the data acquisition step, the complete radar array is used to collect data within a certain close range, e.g., 3–8 m, where the optical sensor can work normally, and the optical image is automatically labeled. In the data processing step, different target distances can be simulated by selecting different radar apertures for multi-resolution imaging based on the equivalence between the m-fold reduction in array elements and the m-fold increase in range, as shown in Figure 6. Human skeletal pose detection data suitable for remote targets can be constructed by jointly training on multiple resolution radar images.

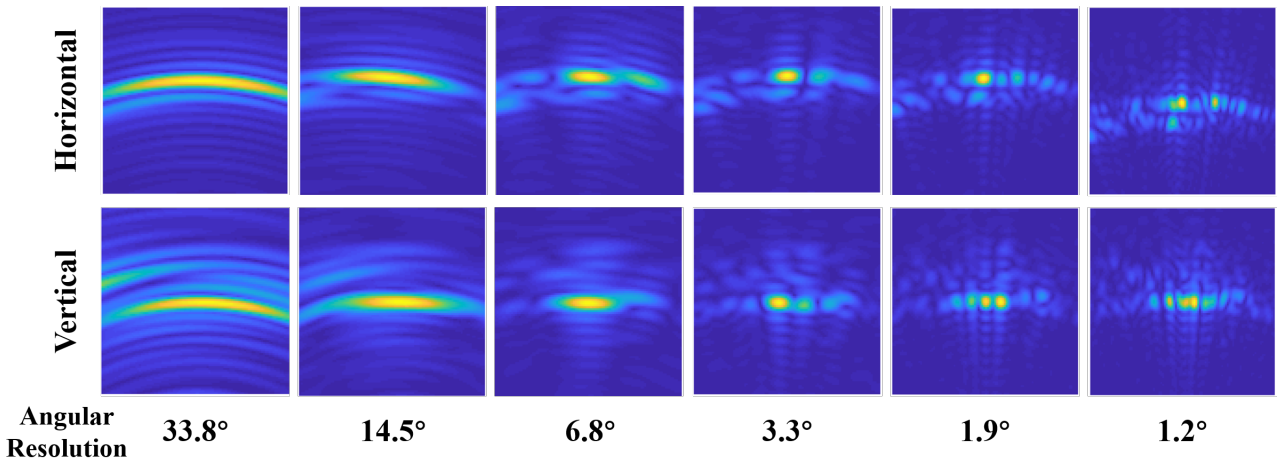


Figure 6. Radar images of the same target under different angular resolutions; the target is a human with hands up.

5. Experiments and Results

5.1. Experimental Setup

In this study, we use two Texas Instruments MWCAS mmWave radars working at 77 to 79 GHz and 79 to 81 GHz, respectively, to ensure non-interference. The MIMO system can form a virtual linear array with 86 elements and the highest azimuthal angle resolution up to 1 degree. To obtain ground truth data for cross-modal supervised learning, the Intel RealSense D455 depth camera (Intel company, Santa Clara, CA, USA) was used to detect human skeletal poses through the OpenPose-Python project. To achieve the synchronization of multiple sensors, we design a data acquisition module to achieve the real-time acquisition of mmWave radar frames and optical image frames with a frame rate of 10 Hz and a temporal synchronization error of better than 1 ms.

The experimental data collection scenarios mainly include outdoor open space, outdoor parking lot scenes, and indoor scenes. Six human subjects with varying sizes were used, one at a time, to collect the skeletal pose data. We acquired about 50,000 samples of training data and about 9000 samples for a validation/development dataset to be used for training the model. About 3000 samples of test data were also collected with the human subject.

RVRU-Pose, as described in Section 4, was used as our learning algorithm. We implement the model based on Pytorch running on a NVIDA Quadro RTX 4000 GPU (NVIDA company, Santa Clara, CA, USA). We use 2D convolutional layers to extract the spatial features of radar heatmaps, in which the kernel size is 5. The stride is set to 2. Each convolutional layer is followed by a BatchNorm layer and a ReLU nonlinear activation layer. The Conv-BN-ReLU group is followed by the ResBlock module, which is used to extract deeper features and improve model generalization. After passing through multiple

Conv-BN-Relu-ResBlocks, the feature tensors extracted from the forked CNN structure are fused in the channel dimension, and finally, the 3D mapping relation of the skeletal pose is formed by the fully connected network. In the experimental part, the number of joints is set to 14, so the output dimension of the skeleton is (14, 3). The modal was trained using an Adam optimizer with the objective of minimizing the mean squared error (MSE) of the output with respect to the ground truth data. The training and validation loss curves for RVRU-Pose are shown in Figure 7.

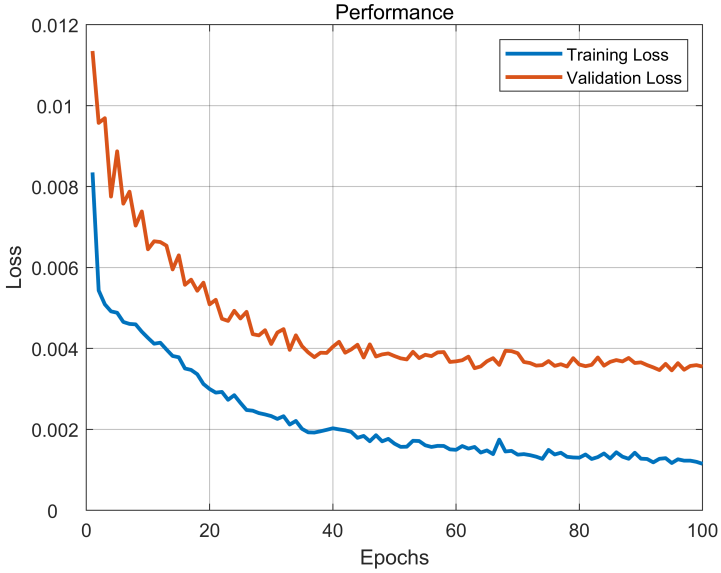


Figure 7. The training and validation loss curves for RVRU-Pose.

5.2. Skeletal Pose Results

The RVRU-Pose network is tested using our dataset, and the results are analyzed in detail in this section. Figure 8 shows the results of estimating the human skeletal pose from mmWave radar and synchrotron optical images when standing, stepping, moving, falling and other different poses. The results of human skeletal pose estimation show that RVRU-Pose can better adapt to different poses in different scenes, such as indoor and outdoor.

When the scene size is set to 6 m $\times$ 6 m, compare RF-Pose [12], mmPose [23], RVRU-Pose algorithm, RVRU-Pose with resblock ablation, RVRU-Pose with horizontal heatmap ablation, and RVRU-Pose with vertical heatmap ablation in estimating the accuracy and time for the completion of skeletal poses, as shown in Table 1. The proposed method can achieve an average localization accuracy of 2 cm skeletal pose joints. At the same time, the proposed model provides an average inference time of 4 ms per frame, which make it suited for real-time implementation. Below, we analyze several methods employed in comparative experiments. The RF-Pose utilizes complete-sized radar heatmaps as input, resulting in an exceedingly large network scale. While it achieves relatively good performance, it entails significant computational resource consumption. Thus, RF-Pose is more suitable for application in small-scale indoor scenarios. On the other hand, mmPose drastically reduces the data volume by encoding the input heatmaps, albeit at the cost of a certain level of estimation accuracy loss. Ablation experiments confirm the crucial role of resblock in the estimation accuracy of RVRU-Pose. Furthermore, the removal of either the horizontal or vertical heatmap leads to a decline in the estimation accuracy of RVRU-Pose along the corresponding dimension. The calculation method for evaluating the accuracy of skeletal joint localization in Table 1 is as follows.

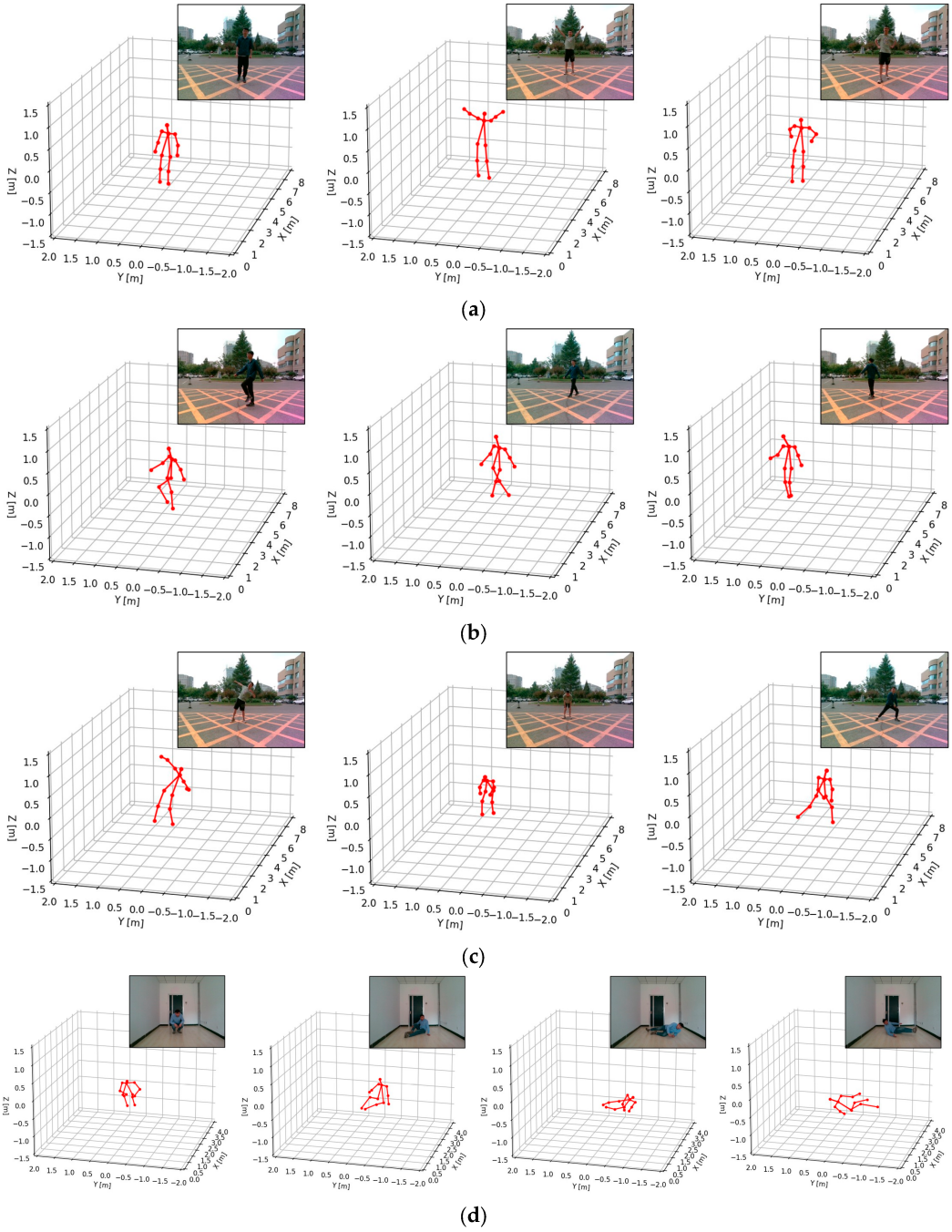


Figure 8. Skeletal pose estimation results at different postures. (**a**) Skeletal pose estimation when the subject is standing; (**b**) skeletal pose estimation when the subject is stepping; (**c**) skeletal pose estimation when the subject is taking exercise; (**d**) skeletal pose estimation when the subject is falling down indoors.

$$MSE_x = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(p_{x,i} - p_{x,i,ref}\right)^2} \tag{9}$$

where the number of skeletal pose joints $M = 14$ is considered. The x-axis coordinate value of the i-th skeletal pose joint is represented as $p_{x,i}$, while $p_{x,i,ref}$ refers to the reference ground truth coordinate of the corresponding joint. The reference ground truth of the joint coordinates is obtained from OpenPose running on a depth camera, which has been

tested to achieve a positioning accuracy of less than 1 cm within an effective range of <6 m. Similarly, MSE_y and MSE_z can be obtained.

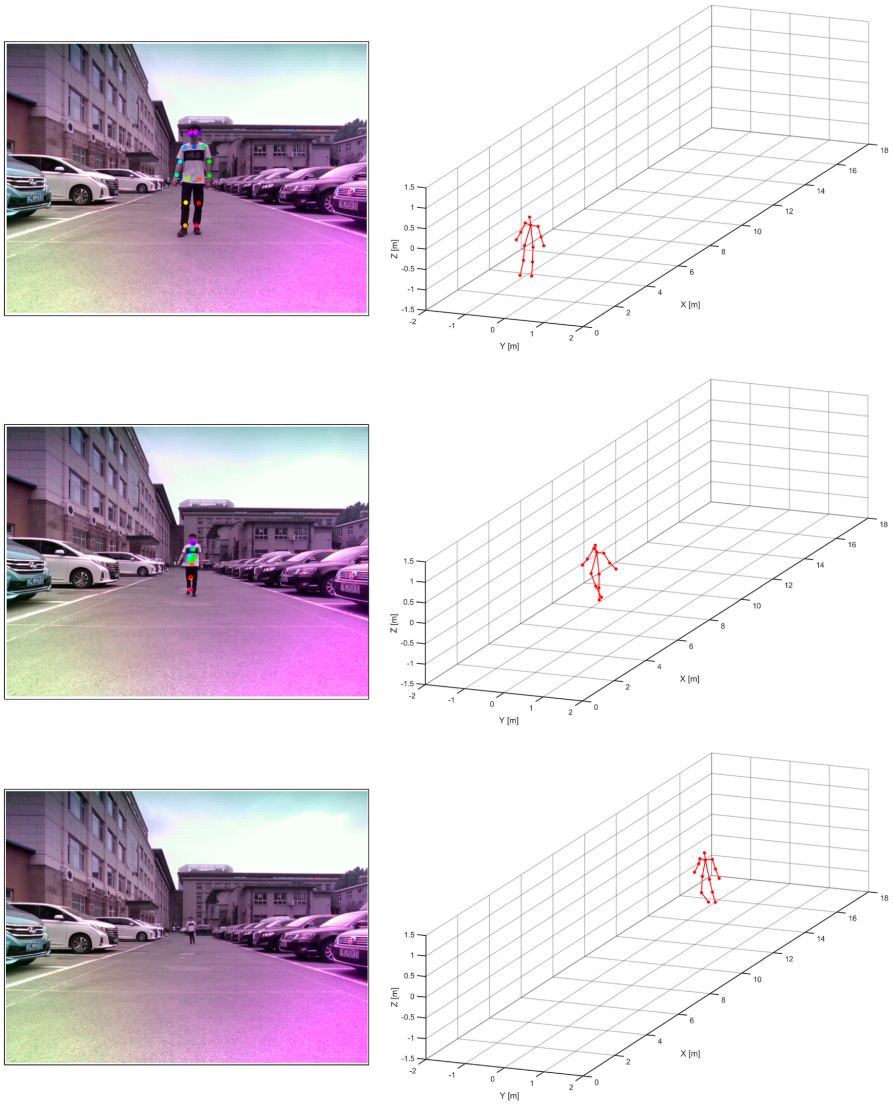


Figure 9. Skeletal pose estimation results at far distance. The distances from the target to the radar are 3 m, 9 m, and 16 m from the left to right subfigures. The color dots in the left column represent the results of vision-based human skeletal pose detection.

Table 1. Localization accuracy and time consuming comparison.

	X (Depth) Accuracy	Y (Azimuth) Accuracy	Z (Elevation) Accuracy	Time Consuming
RF-Pose	1.5 cm	2.5 cm	1.7 cm	276 ms
mmPose	1.6 cm	3.1 cm	3.1 cm	2 ms
RVRU-Pose	1.2 cm	2.3 cm	1.4 cm	4 ms
Ablation1 [1]	3.6 cm	3.5 cm	2.8 cm	2 ms
Ablation2 [2]	1.7 cm	24.8 cm	1.7 cm	3 ms
Ablation3 [3]	1.5 cm	3.0 cm	7.9 cm	3 ms

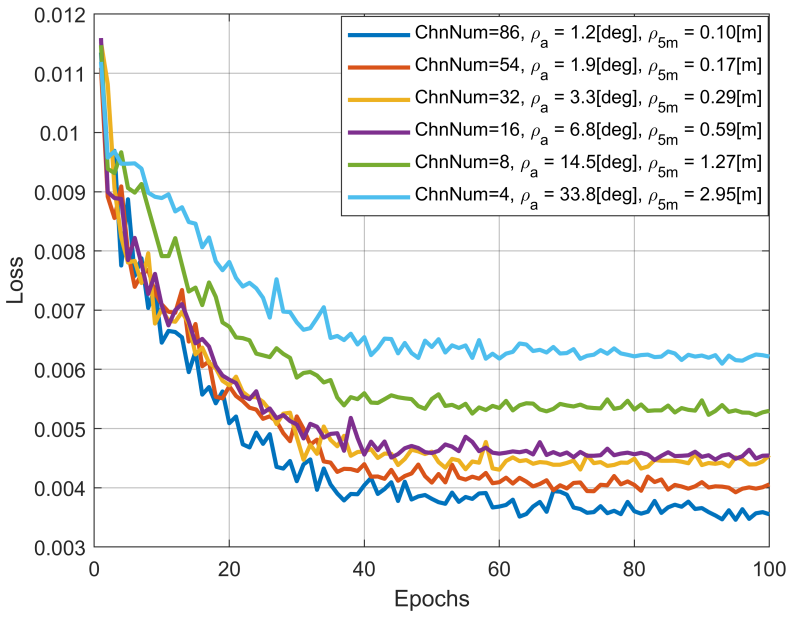
[1] RVRU-Pose with ResBlock Ablation; [2] RVRU-Pose with Horizontal Heatmap Ablation; [3] RVRU-Pose with Vertical Heatmap Ablation.

An obvious advantage of the proposed method over existing mmWave radar-based human skeletal pose estimation schemes is that it is not easily limited by the scene size.

When the scene size increases, the proposed method can still maintain high accuracy and a high frame rate. As shown in Figure 9, pedestrians in the parking lot are 3 m, 9 m and 16 m away from the sensor. Vision-based human skeletal pose detection can successfully detect all skeletal joints at a pedestrian distance of 3 m, some missing joints at a pedestrian distance of 9 m, and the presence of pedestrians at a pedestrian distance of 16 m. The proposed method is able to achieve full human skeletal shutdown detection for long-distance pedestrians, which benefits from the organized form of intensity-coordinate heatmap data and the multi-resolution data augmentation method proposed in this paper. In addition, the proposed method can be directly extended to multi-objective estimation applications, as shown in Figure 2.

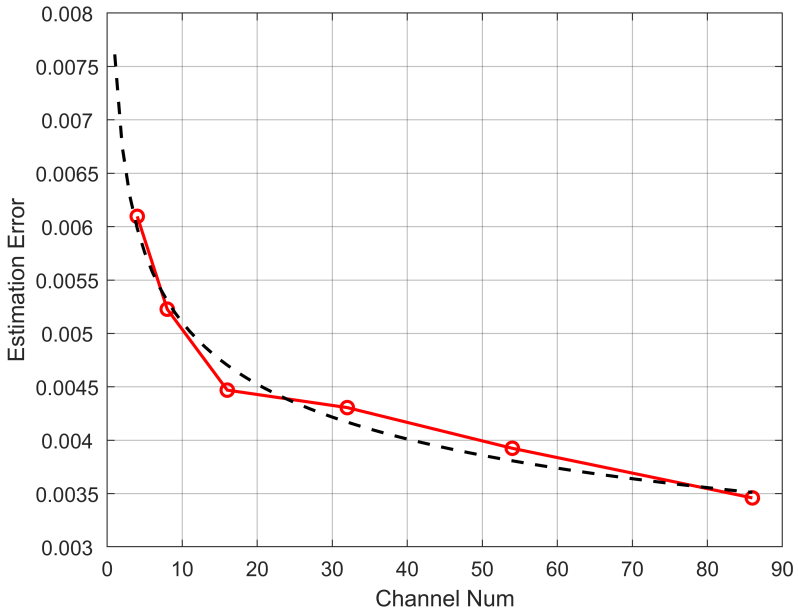
5.3. Influence of Radar Resolution on Localization Accuracy

In this section, the impact of the angular resolution of mmWave radar on the accuracy of human skeletal pose detection will be studied. Depending on the dependence of the radar angular resolution on the number of array channels, the radar angular resolution can be reduced by sampling the number of radar channels. Figure 10a shows the loss convergence of RVRU-Pose in the verification set when the number of array channels is 86, 54, 32, 16, 8, 4, respectively, and the corresponding angular resolutions are 1.2 deg, 1.9 deg, 3.3 deg, 6.8 deg, 14.5 deg, and 33.8 deg, respectively. The red curve in Figure 10b shows the variation of skeletal pose estimation error with the number of radar channels, and its variation trend can be fitted as a black dotted line. It can be seen that the error increases rapidly when the number of channels is lower than 16: that is, when the angular resolution is lower than 6.8 degrees. It can be obtained by combining the average distance of the 5 m dataset, which is the boundary of the human skeleton detection effect when the radar resolution is around 0.5 m. When the resolution is further reduced, the accuracy of the estimation drops dramatically. The analysis shows that the azimuthal resolution should be at least comparable to the bulk width along the azimuth.



(**a**)

Figure 10. *Cont.*



(**b**)

Figure 10. Effect of radar angle resolution on the estimation accuracy of skeletal pose. (**a**) Loss convergence of validation sets with different radar angular resolutions. (**b**) The estimation error of skeletal pose varies with the number of radar channels or angular resolution. The black dotted line means variation trend.

5.4. Model Extension

Currently, published work on skeletal pose detection focuses on pedestrian targets. However, cyclists are also vulnerable road users. They are more maneuverable than pedestrians. Fortunately, the human skeletal pose estimation method presented in this paper can be directly extended to estimate the skeletal pose of cyclists. We built an eight-point skeletal model of the cyclist and collected 4000 mmWave datasets for training, validation, and testing of RVRU-Pose. Unlike human skeletal pose estimation, there are no open-source datasets and optical methods for estimating the skeletal pose of cyclists, and it is not possible to annotate the data independently through cross-modal supervised learning. Data annotation can only be performed manually. Figure 11 shows the results of skeletal pose estimation for cyclists using the proposed method, which reflects the scalability and generality of the proposed radar signal processing method and the proposed RVRU-Pose deep learning model.

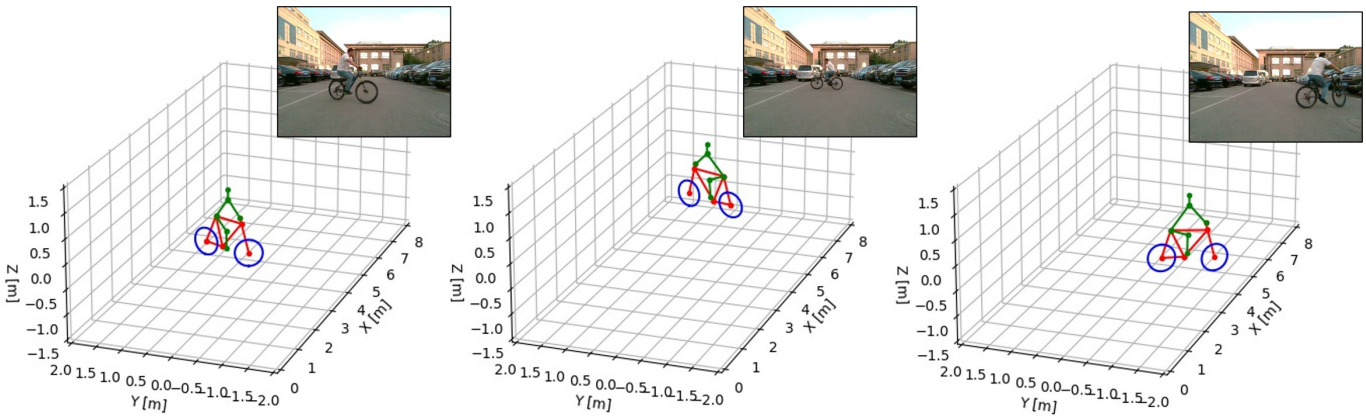


Figure 11. The skeletal pose estimation result of a cyclist.

6. Conclusions

In this paper, we propose a skeletal pose estimation method called RVRU-Pose based on distributed non-coherent mmWave radar for estimating the skeletal pose of vulnerable road user such as pedestrians and cyclists. Then, a multi-resolution radar data augmentation and training method is proposed. Experiments show that the proposed method can achieve an average accuracy of 2 cm in estimating skeletal poses, and it can still work properly in multi-object scenes as well as long-distance scenes with distances greater than 16 m. The effect of radar angular resolution on the estimation accuracy of skeletal poses was also evaluated, and it was concluded that the radar resolution level should be at least comparable to the target size. Estimation of various types of target skeletal poses and the prediction of target behavior through skeletal poses will be future research work.

Author Contributions: Conceptualization, Z.Z. and X.L.; Methodology, Z.Z. and X.D.; Software, Z.Z.; Validation, Y.L.; Writing—original draft, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

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

References

1. Fang, Z.; López, A.M. Intention recognition of pedestrians and cyclists by 2d pose estimation. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 4773–4783. [CrossRef]
2. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep Learning-Based Human Pose Estimation: A Survey. *Acm Comput. Surv.* **2023**, *56*, 1–37. [CrossRef]
3. Munea, T.L.; Jembre, Y.Z.; Weldegebriel, H.T.; Chen, L.; Huang, C.; Yang, C. The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access* **2020**, *8*, 133330–133348. [CrossRef]
4. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef]
5. 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]
6. Yang, C.; Wang, L.; Wang, X.; Mao, S. Environment adaptive RFID based 3D human pose tracking with a meta-learning approach. *IEEE J. Radio Freq. Identif.* **2022**, *6*, 413–425. [CrossRef]
7. Luo, Y.; Li, Y.; Foshey, M.; Shou, W.; Sharma, P.; Palacios, T.; Torralba, A.; Matusik, W. Intelligent carpet: Inferring 3d human pose from tactile signals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
8. Chen, A.; Wang, X.; Zhu, S.; Li, Y.; Chen, J.; Ye, Q. Rfid-pose: MmBody Benchmark: 3D Body Reconstruction Dataset and Analysis for Millimeter Wave Radar. *arXiv* **2022**, arXiv:2209.05070.
9. Chen, J.; Cao, X. SAR motion parameter estimation based on deep learning. In Proceedings of the 2022 3rd China International SAR Symposium (CISS), Shanghai, China, 2–4 November 2022; pp. 1–4.
10. Pu, W. Deep SAR Imaging and Motion Compensation. *IEEE Trans. Image Process.* **2021**, *30*, 2232–2247. [CrossRef]
11. Liu, C.; Wang, Y.; Ding, Z.; Wei, Y.; Huang, J.; Cai, Y. Analysis of Deep Learning 3-D Imaging Methods Based on UAV SAR. In Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 2951–2954.
12. 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, 18–23 June 2018.
13. Ahmad, F.; Zhang, Y.; Amin, M.G. Three-Dimensional Wideband Beamforming for Imaging Through a Single Wall. *Geosci. Remote Sens. Lett.* **2008**, *5*, 176–179. [CrossRef]
14. Adib, F.; Hsu, C.Y.; Mao, H.; Katabi, D.; Durand, F. Capturing the human figure through a wall. *Acm Trans. Graph.* **2015**, *34*, 1–13. [CrossRef]
15. 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]
16. Tian, J.; Yongkun, S.; Yongpeng, D.; Xikun, H.; Yongping, S.; Xiaolong, Z.; Zhifeng, Q. UWB-HA4D-1.0: An Ultra-wideband Radar Human Activity 4D Imaging Dataset. *J. Radars* **2022**, *11*, 27–39.
17. 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]

18. 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*, 1–15. [CrossRef]
19. Ding, W.; Cao, Z.; Zhang, J.; Chen, R.; Guo, X.; Wang, G. Radar-Based 3D Human Skeleton Estimation by Kinematic Constrained Learning. *IEEE Sensors J.* **2021**, *21*, 23174–23184. [CrossRef]
20. Cao, Z.; Ding, W.; Chen, R.; Zhang, J.; Guo, X.; Wang, G. A Joint Global-Local Network for Human Pose Estimation with Millimeter Wave Radar. *IEEE Internet Things J.* **2022**, *10*, 434–446. [CrossRef]
21. Xue, H.; Ju, Y.; Miao, C.; Wang, Y.; Wang, S.; Zhang, A.; Su, L. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services, Virtual Event Wisconsin, 24 June–2 July 2021.
22. Zhong, J.; Jin, L.; Wang, R. Point convolution based human skeletal pose estimation on millimetre wave frequency modulated continuous wave multiple input multiple output radar. *Iet Biom.* **2022**, *11*, 333–342. [CrossRef]
23. Sengupta, A.; Jin, F.; Zhang, R.; Cao, S. mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sensors J.* **2020**, *20*, 10032–10044. [CrossRef]
24. Li, G.; Zhang, Z.; Yang, H.; Pan, J.; Chen, D.; Zhang, J. Capturing human pose using mmWave radar. In Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops, Austin, TX, USA, 23–27 March 2020.
25. Sengupta, A.; Jin, F.; Cao, S. NLP based skeletal pose estimation using mmWave radar point-cloud: A simulation approach. In Proceedings of the IEEE Radar Conference, Florence, Italy, 21–25 September 2020.
26. 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]
27. Cui, H.; Dahnoun, N. Real-Time Short-Range Human Posture Estimation Using mmWave Radars and Neural Networks. *IEEE Sensors J.* **2021**, *22*, 535–543. [CrossRef]
28. Patole, S.M.; Torlak, M.; Wang, D.; Ali, M. Automotive radars: A review of signal processing techniques. *IEEE Signal Process. Mag.* **2017**, *34*, 22–35. [CrossRef]
29. Xu, S.; Wang, J.; Yarovoy, A. Super Resolution DOA for FMCW Automotive Radar Imaging. In Proceedings of the IEEE Conference on Antenna Measurements and Applications, Västerås, Sweden, 3–6 September 2018.
30. Zeng, Z.; Dang, X.; Li, Y.; Bu, X.; Liang, X. Angular Super-Resolution Radar SLAM. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, Czech Republic, 27 September–1 October 2021.
31. Kim, W.; Cho, H.; Kim, J.; Kim, B.; Lee, S. YOLO-Based Simultaneous Target Detection and Classification in Automotive FMCW Radar Systems. *Sensors* **2020**, *20*, 2897. [CrossRef] [PubMed]
32. Bennett, J.R.; Cumming, I.G.; Deane, R.A.; Widmer, P.; Fielding, R.; McConnell, P. SEASAT Imagery. *Aviat. Week Space Technol.* **1979**, *19*.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.

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.