electronics

MDPI

Article

A Novel Spatial–Temporal Network for Gait Recognition Using Millimeter-Wave Radar Point Cloud Videos

Chongrun Ma and Zhenyu Liu *

School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China; machongrun@mail2.gdut.edu.cn
* Correspondence: zhenyuliu@gdut.edu.cn

Abstract: Gait recognition is a behavioral biometric technology that aims to identify individuals through their manner of walking. Compared with vision and wearable solutions, millimeter-wave (mmWave)-radar-based gait recognition has drawn attention because radar sensing is privacy-preserving and non-contact. However, it is challenging to capture the motion dynamics of walking people from mmWave radar signals, which is crucial for robust gait recognition. In this study, a novel spatial–temporal gait recognition network based on mmWave radar is proposed to address this problem. First, a four-dimensional (4D) radar point cloud video (RPCV) was introduced to characterize human walking patterns. Then, a PointNet block was utilized to extract spatial features from the radar point clouds in each frame. Finally, a Transformer layer was applied for the spatial–temporal modeling of the 4D RPCVs, capturing walking motion information, followed by fully connected layers to output the identification results. The experimental results demonstrated the superiority of the proposed network over mainstream networks, which achieved the best human identification performance on a dataset of 15 volunteers.

Keywords: millimeter-wave radar; gait recognition; point clouds; Transformer

Citation: Ma, C.; Liu, Z. A Novel Spatial–Temporal Network for Gait Recognition Using Millimeter-Wave Radar Point Cloud Videos. *Electronics* **2023**, *12*, 4785. https://doi.org/10.3390/electronics12234785

Academic Editors: Yin Zhang, Yulin Huang, Yachao Li and Deqing Mao

Received: 7 November 2023
Revised: 21 November 2023
Accepted: 22 November 2023
Published: 26 November 2023

1. Introduction

Human gait, which is defined as the manner of walking, is a behavioral biometric trait that is unique for each person and can be used to authenticate individuals [1]. Compared with other biometrics, such as faces, fingerprints, DNA, and irises, gait signatures can be captured from a distance and without cooperation from individuals, and it is hard to conceal and disguise gait characteristics [2,3]. These advantages make gait recognition a promising human identification technology for diverse applications, including public security, forensics, and healthcare [4,5]. Vision- and wearable-sensor-based methods are the two main categories of gait recognition techniques used in the community [6,7]. However, vision sensors are limited by the illumination conditions, and people may feel constrained by wearable devices and find them inconvenient. More importantly, vision-based devices may raise privacy concerns in non-public scenarios, such as home and office, which can result in the leakage of private or confidential information (e.g., human habits, relationships, visited places, and so on).

mmWave-radar-based human sensing has attracted extensive attention in recent years for its high sensing sensitivity [8]. Radar sensors are non-contact devices, can work under any lighting conditions, and do not infringe on the privacy of monitored individuals [9]. Therefore, mmWave-radar-based gait recognition can be a promising option for insensitive human identification compared to vision and wearable solutions.

To achieve accurate gait recognition, it is crucial to capture walking dynamics that produce inter-personal differences in walking patterns [7]. Most existing radar-based gait-recognition methods exploit micro-Doppler signatures from radar echoes to characterize the micro-motion patterns of human gait, combined with machine learning or deep learning

technologies for human identification [10–12]. Micro-Doppler signatures are induced by the motions of different body parts (e.g., the torso and limbs), embodying the unique kinematic patterns of differently walking individuals [13]. However, micro-Doppler signatures are not robust against viewpoint changes [14], which is limited when people walk in a wide field of view.

Radar point clouds are collections of reflection points representing the target surface, which are generated by performing a series of target-detection algorithms on multiple-input multiple-output (MIMO) radar echoes, containing spatial coordinates and velocity information [8]. Radar point clouds across consecutive frames can be regarded as a radar point cloud video (RPCV). Time-varying radar point clouds reveal motion dynamics, as well as the physical shape of a walking human [15], which are also more resilient to changes in viewpoints than micro-Doppler signatures. Therefore, RPCV with spatial–temporal signatures have been taken into consideration for gait recognition in some related studies [16–18].

Nonetheless, 4D-radar-point-cloud-based solutions still pose challenges for high-performance gait recognition. First, the limited numbers of antennas on commercial mmWave radars result in sparse radar point clouds that exhibit a lack of appearance or geometric information [19]. Second, due to the specular reflection phenomenon of mmWave signals, only parts of human body reflections propagate back to the received antennas [20,21]. Consequently, radar point clouds emerge inconsistently across different frames, resulting in difficulty in modeling the spatial–temporal signatures and motion dynamics of human gait.

To address the problem mentioned above, a 4D-RPCV-based spatial–temporal network for gait recognition is proposed in this study. In our proposed network, PointNet [22] was adopted to extract the spatial features from sparse radar point clouds in each frame. In PointNet, shared multi-layer perceptrons (MLPs) are utilized to extract high-level representations from point clouds. Furthermore, a max pooling operation is applied to process an unordered set of point features. After being processed via the PointNet block, 4D RPVs are transformed into point feature sequences. To capture the motion dynamics of human gait, a Transformer layer is deployed to perform multi-head attention on point feature sequences. Transformers [23] have dominated the field of natural language processing in recent years and have been extended to the computer vision community for their capacity to capture global correlations [24]. Inspired by Transformers and self-attention mechanisms, a Transformer layer is employed in the proposed network to further exploit the spatial–temporal correlation across the 4D RPCVs, thereby capturing the motion dynamics of human gait. Finally, fully connected layers are utilized to output the identification results.

The gait-recognition method in this study can be formulated as follows: (1) First, mmWave radar signals reflected from walking human subjects are transformed into 4D RPCVs through a series signal-processing algorithms, which can be used to characterize walking patterns. (2) Second, a PointNet block was adopted to extract spatial features from the radar point clouds in each frame of the 4D RPCVs, followed by a Transformer layer for the temporal modeling of features from consecutive frames, capturing the motion dynamics of walking individuals. (3) Third, the class token in the output of the Transformer layer is fed into fully connected layers to predict the identities. After training and evaluation, the experimental results demonstrated the effectiveness and robustness of the proposed spatial–temporal gait recognition network, which achieved the best performance in the case of identifying 10 and 15 subjects.

The main contributions of this study are summarized as follows:

- A 4D-RPCV-based spatial–temporal network is proposed to better capture the motion dynamics of gait for accurate human identification, which is capable of modeling walking motion from time-varying sparse radar point clouds.
- A Transformer encoder architecture is introduced in the proposed network to learn radar point features' sequences, capturing the spatial–temporal dependencies that contribute to accurate gait recognition.

- A 4D RPCV human gait dataset was built on real mmWave MIMO frequency-modulated continuous wave (FMCW) radar measurements, which involved 15 volunteers walking along different paths. Furthermore, experiments on this dataset showed that the proposed spatial–temporal network effectively improved the accuracy of gait recognition.

The remainder of this paper is organized as follows. Section 2 reviews the related works on radar-based human sensing and gait recognition research. Section 3 introduces the 77 GHz radar system used in this study. Section 4 describes the proposed radar-based gait-recognition method, including the generation of 4D RPCVs and the spatial–temporal network. In Section 5, the performance of the proposed gait recognition network is evaluated. The conclusion and future works are provided in Section 6.

2. Related Works

With the rapid development of mmWave radar technology, radar sensors have been widely explored for various human sensing tasks. In 2016, Google designed the mmWave radar sensing module Soli, which supports gesture recognition [25]. Various radar-based human activity recognition methods have been studied in recent years, involving many kinds of radar representations such as range–time maps, range–Doppler maps (RDM), micro-Doppler signatures, and radar point clouds [26]. Radar-based human recovery is also a field that attracts attention. Ref. [27] estimated 25 human skeletal joints from radar point clouds, and Ref. [19] reconstructed a three-dimensional human mesh by combining mmWave sensing and the SMPL model. In summary, mmWave radar is suitable for human sensing tasks, including gait recognition.

In [28], a micro-Doppler signature-based multi-branch convolutional neural network (CNN) for human gait recognition was proposed. However, micro-Doppler signatures are limited due to their poor robustness to viewpoint changes, which is also computationally demanding in a multi-person scenario. Ref. [17] designed a sequence radar point network combining PointNet and bidirectional long short-term memory (Bi-LSTM) to learn on 4D radar point cloud sequences. Similarly, the researchers in [29] proposed a gait recognition network combining PointNet and a temporal convolution network (TCN). However, the networks of these methods are based on the recurrent neural network (RNN) architecture, which does not consider global dependencies related to walking motion dynamics with time-varying sparse radar point clouds.

3. Frequency-Modulated Continuous Wave-MIMO Radar System

This study utilized a mmWave FMCW-MIMO radar that transmits linear *chirp* sequences. The frequency of transmission is linearly increased over time through the transmit (TX) antenna. A single *chirp* with the carrier frequency f_c can be expressed as [30]

$$S(t) = e^{j2\pi\left(f_c+\frac{1}{2}\frac{B}{T_c}t\right)t},\ 0 \le t \le T_c. \tag{1}$$

where B is the bandwidth and T_c is the chirp duration.

Denoting c as the speed of light and R and v as the range and velocity of the target, the time delay of the received signal can be expressed as

$$\tau = \frac{2(R+vt)}{c}. \tag{2}$$

The received (RX) signal is mixed with the TX signal and, subsequently, filtered by a low-pass filter, generating the intermediate frequency (IF) signal.

A *radar frame* is a sequence of consecutive chirps that can be structured as a two-dimensional waveform across two temporal dimensions. In a frame with M chirps, each chirp is sampled with sampling rate f_s to obtain N points (*fast time* dimension), while M samples, corresponding to the number of chirps, are obtained with sampling period T_{rep}

(*slow time* dimension). Thus, the IF signal of the target in a frame across these two time dimensions can be approximately expressed as [30]

$$d(n,m) \approx \exp\left\{ j2\pi \left[(f_b + f_d)\frac{n}{f_s} + f_d m T_{rep} + \frac{2f_c R}{c} \right] \right\}. \quad (3)$$

where n indicates the index of fast time samples within each chirp and m is the index of slow time samples across successive chirps. The beat frequency $f_b = 2BR/cT_c$ and Doppler frequency $f_d = 2f_c v/c$ reveal the range and velocity of the target, respectively. The information can be extracted by a two-dimensional fast Fourier transform (FFT) along the fast and slow time dimensions.

Because each antenna's received signal has a different phase, a radar with a linear antenna array can be used to estimate a target's azimuth. Denoting by d the distance between two adjacent antennas and $\lambda = c/f_c$ the base wavelength of the transmitted chirp, the phase shift between the received signals from these two antennas is [31]

$$\Delta\phi = 2\pi \frac{dsin\theta}{\lambda}. \quad (4)$$

where θ denotes the azimuth of the target. For Q number of targets, the three-dimensional (3D) FMCW-MIMO radar IF signal can be represented as [30]

$$d(n,m,l) \approx \sum_{q=1}^{Q} \alpha_q \exp\left\{ j2\pi \left[\left(f_{bq} + f_{dq} \right) \frac{n}{f_s} + \frac{ld \sin\theta_q}{\lambda} + f_{dq} m T_{rep} + \frac{2f_c R_q}{c} \right] \right\}. \quad (5)$$

where l indicates the index of the receiving antenna and α_q is the complex amplitude of the qth target. The samples of the IF signal can be arranged into a 3D matrix across fast time, slow time, and channel dimensions, forming the *Raw Data Cube*. Range, velocity, and angle estimation can be achieved by applying FFT along these three dimensions, respectively.

4. Method

4.1. Four-Dimensional Radar Point Cloud Videos

Four-dimensional RPCVs are introduced to characterize human walking patterns. In comparison to 3D point clouds obtained from Lidar or depth cameras, 4D radar point clouds include velocity information, providing benefits for modeling human walking patterns. The generation of 4D radar point clouds involves using a frequency-modulated continuous-wave (FMCW) multiple-input multiple-output (MIMO) radar with antennas placed both horizontally and vertically. This configuration enables the estimation of the azimuth and elevation of scatter points from walking human targets.

The steps for generating 4D radar point clouds from IF signals are shown in Figure 1. First, a 2D-FFT is applied to the raw radar data to obtain a range–Doppler matrix (RDM). Here, the 2D-FFT involves applying the FFT along the fast time and the slow time dimensions sequentially. Subsequently, a moving target indication (MTI) filter is utilized to remove static clutter caused by the environment. Following this, a two-dimensional constant false alarm rate (2D-CFAR) is applied on the RDM to select prominent range–Doppler pixels as potential scattering points, using a threshold that varies according to the noise level. For each potential scatter point in the range–Doppler domain, the signal along channel dimension is arranged into a 2D matrix based on the antenna array positions. The spatial spectrum in the horizontal and vertical directions can be obtained by performing the 2D-FFT on this 2D matrix. The angle of arrival (AoA) in the horizontal and vertical directions of each scattering point can be estimated by applying peak searching to the spatial spectrum. A 4D detected scattering point can be expressed by $p = [r, v, \theta, \phi]$, where

θ and ϕ are the azimuth and elevation angles, respectively. After coordinate transformation, $p = [x, y, z, vs.]$, and the transformation is

$$\begin{aligned} x &= rsin\theta cos\phi, \\ y &= rcos\theta cos\phi, \\ z &= rsin\phi. \end{aligned} \tag{6}$$

where x, y, and z are the 3D coordinates in the Cartesian coordinate system and v is the radial velocity.

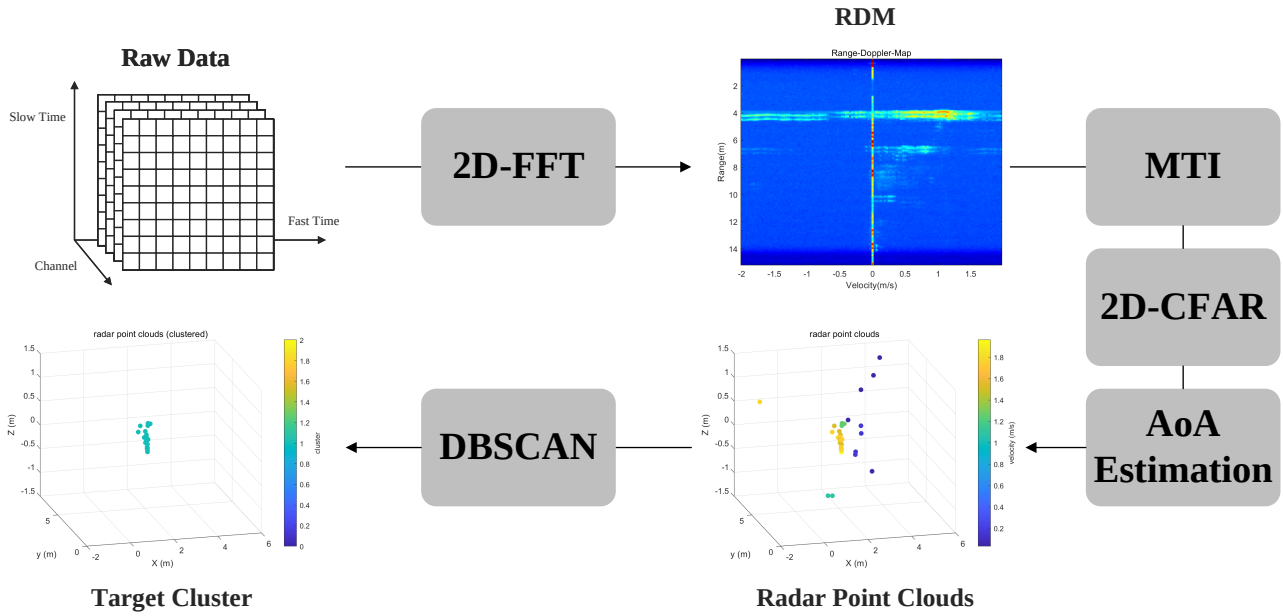


Figure 1. Flowchart of generating radar point clouds.

Finally, to remove clutter points and cluster the scattering points belonging to the same target, the density-based spatial clustering of applications with noise (DBSCAN) algorithm [32] is employed. A cluster consisting of multiple scattering points in frame k can be expressed as

$$C_k = \left\{ p_i = [x_i, y_i, z_i, v_i] \middle| i = 1, \ldots\ldots, I \right\}. \tag{7}$$

Furthermore, a 4D RPCV can be constructed by combining clusters belonging to the same target from consecutive frames, which can be expressed as $C_{1:L} = [C_1, C_2, \ldots, C_L]^{\mathrm{T}}$. A short 4D RPCV sample is shown in Figure 2, and the cluster with the most scattering points is regarded as the human target. Radar point clouds emerge inconsistently across multiple frames due to the specular reflection phenomenon of mmWave signals. The 4D RPCVs were taken as the input of the spatial–temporal network introduced in Section 4.2 for human identification.

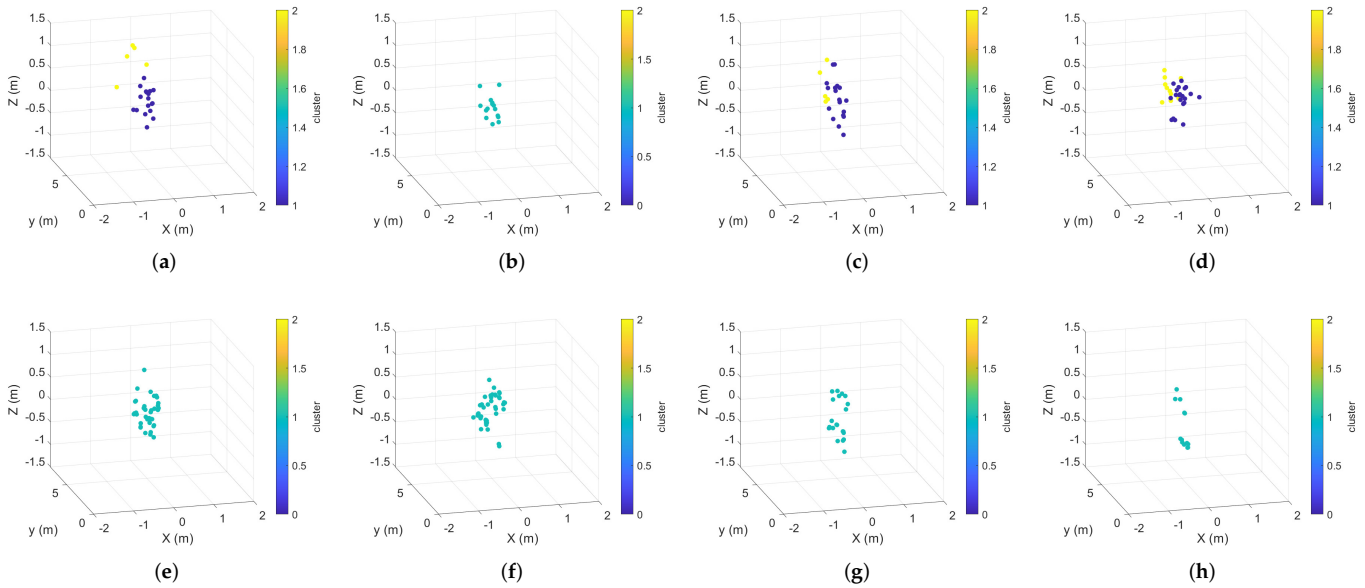


Figure 2. A 4D RPCV sample lasting 8 frames. (**a**) Frame 1; (**b**) Frame 2; (**c**) Frame 3; (**d**) Frame 4; (**e**) Frame 5; (**f**) Frame 6; (**g**) Frame 7; (**h**) Frame 8.

4.2. Spatial–Temporal Network

The key factor for the recognition network is extracting person-specific gait features, which are related to both spatial and motion patterns. The proposed network was designed to exploit time-varying sparse 4D radar point clouds and capture unique spatial information and walking motion dynamics. It consists of three modules, termed the *PointNet block*, *Transformer layer,* and *Output layer*, as shown in Figure 3.

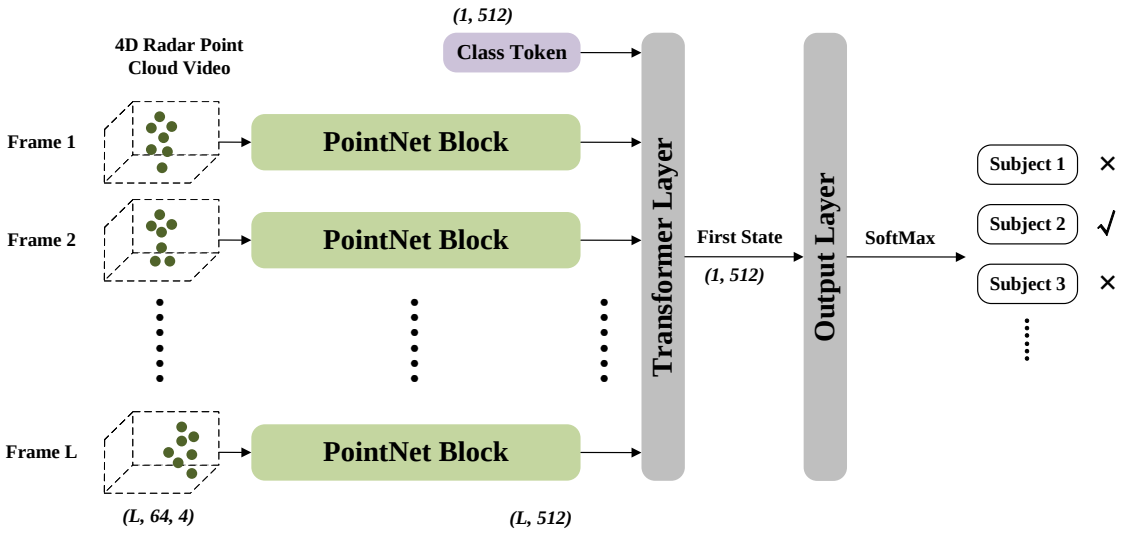


Figure 3. Architecture of the spatial–temporal network.

4.2.1. PointNet Block

L identical PointNet encoders are applied to process the input of 4D RPCVs, which consists of radar point clouds in *L* frames. As shown in Figure 4, for radar point clouds in each frame, the PointNet block implements MLPs in parallel to extract pointwise features, and all the MLPs in parallel share the same weights. In the MLP layer, each MLP extracts high-dimensional features from a single 4D radar point through a linear transformation.

Compared to the original PointNet, the T-Nets for the input and feature transform are removed for better consistency of the point clouds from consecutive frames, as well as an easier training phase.

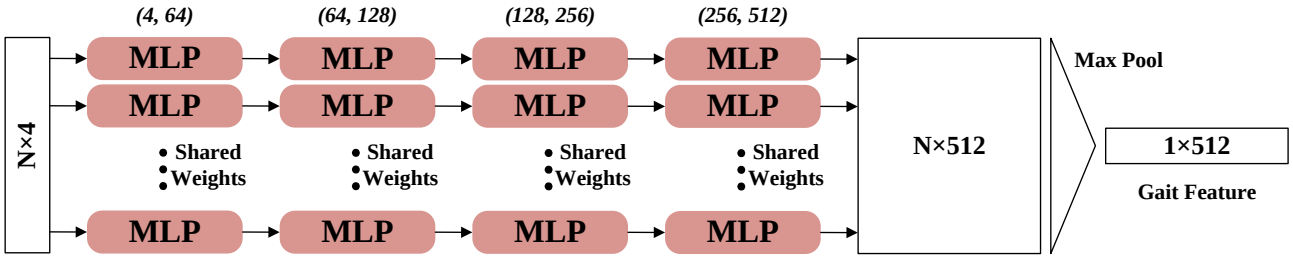


Figure 4. Architecture of the PointNet block.

After extracting high-level human walking motion and appearance-related representations from radar point clouds, a max pooling operation is applied to the unordered set of pointwise features to obtain a global spatial feature in a single frame. Specifically, the 4D RPCV $C_{1:L} \in \mathbb{R}^{L \times Num \times 4}$ is transformed to a feature sequence $F_{1:L} = [f_1, f_2, \ldots, f_L]^{\mathrm{T}}, F_{1:L} \in \mathbb{R}^{L \times Dim}$ by the PointNet block, where L is the length, Num is the number of points in each frame, and Dim is the dimension of the global spatial feature.

4.2.2. Transformer Layer

To model person-specific walking motion dynamics with feature sequences obtained from the PointNet block, a Transformer layer with a multi-head attention mechanism is applied in the spatial–temporal network.

The feature sequences are regarded as gait embeddings in this Transformer layer. A learnable vector, termed as the *class token*, is initialized and concatenated with the gait embeddings, as shown in Figure 3. The class token interacts with the features in all states, avoiding preference for motion information in specific states. Compared to simply pooling features from all states, using the class token for further classification is a better way to aggregate gait information across the entire RPCV. The input of the Transformer layer can be expressed as

$$F = [f_{cls}, f_1, f_2, \ldots, f_L]^{\mathrm{T}}, F \in \mathbb{R}^{(L+1) \times Dim} \tag{8}$$

where f_{cls} is the class token.

The architecture of the Transformer layer is shown in Figure 5a, consisting of a multi-head attention block and a positionwise feed-forward block with layer normalization applied and residual connections used. First, F is projected to the *Query*, *Key*, and *Value* by linear transformation, which can be expressed as

$$\begin{aligned} Q &= W^q(F); \\ K &= W^k(F); \\ V &= W^v(F). \end{aligned} \tag{9}$$

where W^q, W^k, and W^v are the weights of the linear transformation.

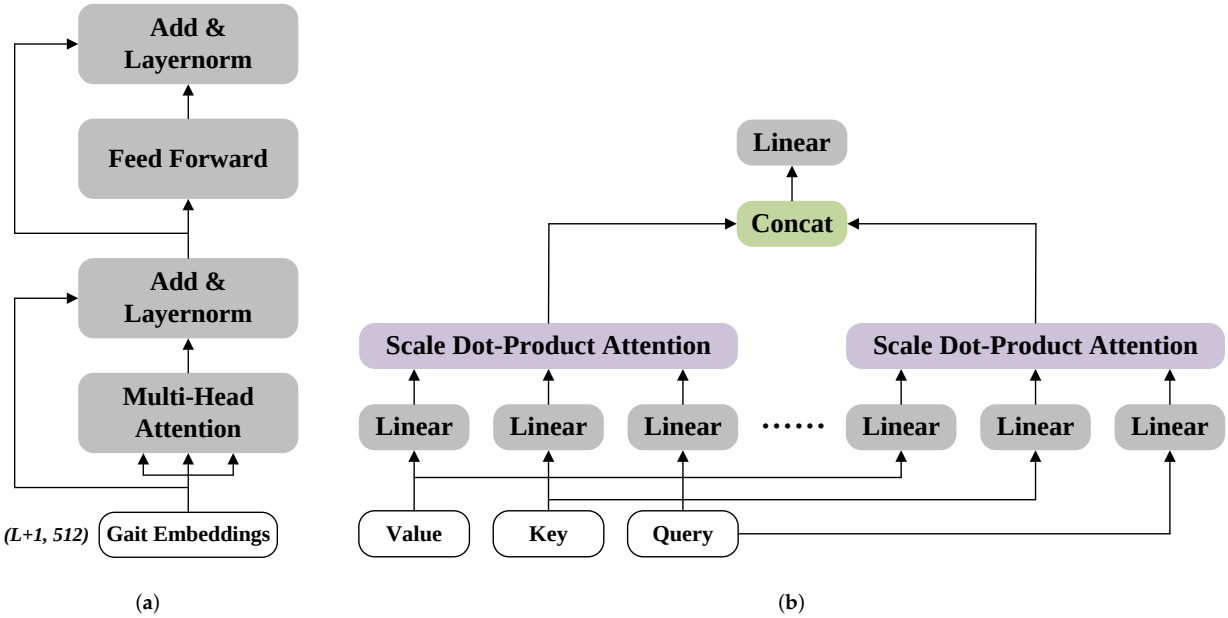


Figure 5. Architecture of Transformer layer and multi-head attention. (**a**) Transformer layer; (**b**) multi-head attention.

Multi-head attention then divides Q, K, and V into different representation subspaces by linear projection and aggregates features from all the representation subspaces to capture various dependencies within the 4D RPCVs, as shown in Figure 5b. The process of this transformation can be expressed as

$$\begin{aligned} Q_h &= W^{q_h}(Q) = [q_{cls}, q_1, \ldots, q_L]^{\mathrm{T}}; \\ K_h &= W^{k_h}(K) = [k_{cls}, k_1, \ldots, k_L]^{\mathrm{T}}; \\ V_h &= W^{v_h}(V) = [v_{cls}, v_1, \ldots, v_L]^{\mathrm{T}}. \end{aligned} \tag{10}$$

where h represents the index of the representation subspace (termed as the *head*) and Q_h, K_h, $V_h \in \mathbb{R}^{(L+1)\times\frac{Dim}{H}}$. Furthermore, H is the number of heads. The scaled dot-product attention for each head is calculated as

$$Att_h = softmax(\frac{Q_h K_h^{\mathrm{T}}}{\sqrt{\frac{Dim}{H}}})V_h \in \mathbb{R}^{(L+1)\times\frac{Dim}{H}}. \tag{11}$$

Attention from all heads is concatenated and processed by linear projection. Finally, the positionwise feed-forward block applies an identical MLP to each state for further feature extraction. In addition, the use of layer normalization and residual connections facilitates building a deeper architecture.

4.2.3. Output Layer

In the output layer, the feature vector corresponding to the class token in attention is extracted, and FC layers are applied to reduce the dimension of the gait feature. A dropout layer is used to prevent overfitting problems. Afterward, a softmax layer is applied to predict the human identity $\hat{y}$. The categorical cross-entropy loss function compares $\hat{y}$ with the ground-truth label y of the walking human, instructing the optimization of the spatial–temporal network. The loss function can be expressed as

$$Loss = -\sum_{p=1}^{P} y_p log(\hat{y_p}) \tag{12}$$

where p is the number of people registered in the gait-recognition system.

5. Experimental Results and Analysis

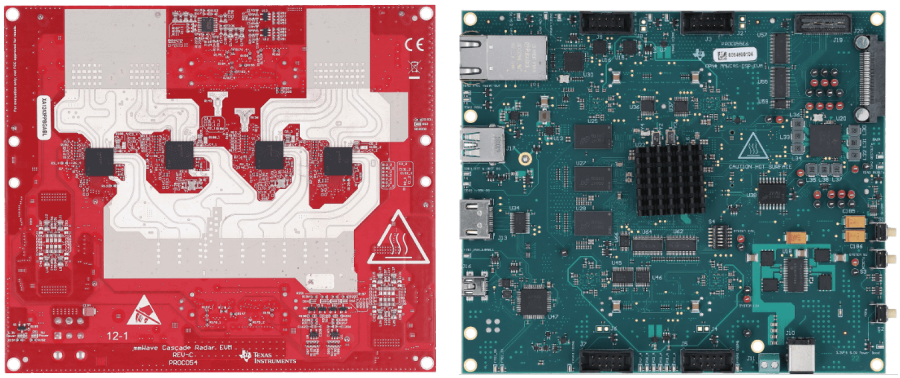
5.1. Data Collection

A mmWave FMCW-MIMO radar platform developed by Texas Instruments was utilized for evaluation in this study. The radar platform comprises an RF module and a DSP module, implementing a four-device cascaded array of AWR1243 chips, as shown in Figure 6. The radar, equipped with a two-dimensional antenna array, can be used for azimuth and elevation estimation. It employs the time-division multiplexing (TDM) technique to achieve waveform orthogonality. The virtual RX antenna array is shown in Figure 7. The detailed parameters of the radar system can be found in Table 1.

Table 1. Parameters of the radar FMCW-MIMO system.

Parameters	Value
Start frequency	77 GHz
Chirp bandwidth	2529 MHz
Chirp duration	32 µs
Frame duration	62 ms
Number of samples per chirp	256
Number of chirps per frame	128
Number of TX antennas	12
Number of RX antennas	16
Range resolution	5.93 cm
Velocity resolution	0.0311 m/s
Azimuth resolution	1.4°
Elevation resolution	18°

The data were collected in an open area, as shown in Figure 8, with a sensing area measuring 10 m × 15 m. The radar platform was placed 3 m away from the sensing area and mounted on a tripod stand at a height of 1 m. We recruited 15 volunteers for the experiment, with heights ranging from 160 cm to 183 cm and weights from 51 kg to 75 kg, as detailed in Table 2. Each volunteer was instructed to walk within the sensing area from six different viewpoints relative to the radar platform. For each viewpoint, we collected five sequences, each lasting 100 frames. Four of these sequences were allocated for training, while the remaining one was used for testing. In total, we collected 45,000 radar frames.



(**a**) (**b**)

Figure 6. Radar platform for evaluation. (**a**) Radar RF module; (**b**) radar DSP module.

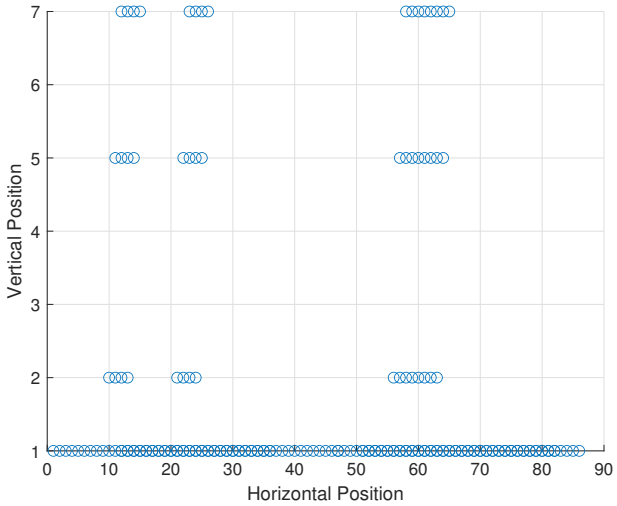


Figure 7. Virtual RX antenna array.

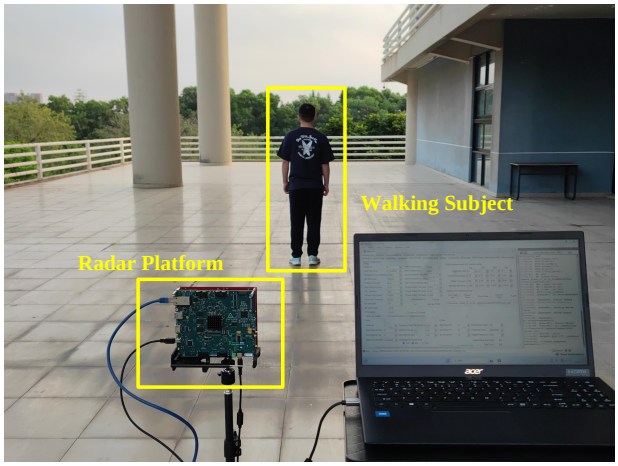


Figure 8. Experimental scene.

Table 2. Information about the subjects.

	Height (cm)	Weights (kg)	Gender	Ages		Height (cm)	Weights (kg)	Gender	Ages
Person 1	172	65	male	41	Person 9	171	66	male	22
Person 2	177	66	male	25	Person 10	166	52	female	23
Person 3	165	56	female	26	Person 11	177	83	male	22
Person 4	172	52	male	23	Person 12	160	51	female	26
Person 5	168	64	male	24	Person 13	169	66	male	35
Person 6	183	61	male	23	Person 14	171	69	male	23
Person 7	170	62	female	22	Person 15	180	74	male	21
Person 8	178	70	male	24					

5.2. Implementation Details

The size of the 4D RPCVs was set to $50 \times 64 \times 4$. For the implementation of DBSCAN, the radius and minimal number of points in the neighborhood were set to 0.8 m and 10, respectively. The hidden dimension *Dim* was set to 512, and the number of heads H in the multi-head attention block was set to four. The data of 10 volunteers were used for the basic evaluation, and the data of all 15 volunteers were used for further stability analysis of the networks. In the training phase, Adam was chosen as the optimizer, with a batch size of 32 and a learning rate of 0.0001. We trained the networks for 250 epochs. All the

training and evaluating processes of the networks were implemented using PyTorch with an NVIDIA A40 GPU.

5.3. Performance Analysis

5.3.1. Comparison of Performance

To verify the effectiveness of the proposed spatial–temporal gait-recognition network, we compared it with several gait-recognition benchmarks based on radar point clouds or micro-Doppler signatures. A combination of PointNet and an RNN-based temporal block is the mainstream for radar-point-cloud-video-based gait recognition. In this experiment, we compared the proposed network with "PointNet + BiLSTM" and "PointNet + TCN", both of which use PointNet to extract radar point cloud features and employed an RNN-based temporal block to capture the time-varying characteristics. mmGaitNet [16] uses 2D convolutional kernel to extract spatial–temporal features from the RPCVs. The Multi-Channel CNN [28] captures gait Doppler features from the micro-Doppler signature using an Inception and residual-connection-based network.

As shown in Table 3, the proposed spatial–temporal network achieved the best identification performance with an accuracy of 94.44% on the test set, showcasing the capacity of our model in capturing human walking dynamics. The spatial–temporal network was 8.33% more accurate than "PointNet + BiLSTM" and 6.94% more accurate than "PointNet + TCN", respectively, demonstrating the effectiveness of using the Transformer layer to model the temporal correlation from the 4D RPCVs. In the case of 4D-RPCV-based gait recognition, mmGaitNet achieved a higher accuracy than that of "PointNet + BiLSTM" and "PointNet + TCN", showcasing the potential of the 2D CNN in capturing spatial–temporal human gait features from time-varying radar point clouds. Since the samples in our dataset were collected from different viewpoints relative to the radar, the micro-Doppler signatures were affected, resulting in the lowest accuracy among all the compared networks for the Multi-Channel CNN. The confusion matrices of the proposed network, mmGaitNet, "PointNet + TCN", and Multi-Channel CNN are shown in Figure 9. All the networks in this experiment exhibited poor recognition accuracy for certain users. However, the spatial–temporal network achieved more than 92% accuracy for all subjects except 'Subject 1', an outcome that other networks could not achieve.

Table 3. Comparison of the proposed spatial–temporal network with different gait recognition networks.

Input	Network	Accuracy
4D Radar Point Cloud Videos	Spatial–Temporal Network (Ours)	**94.44**%
	PointNet + BiLSTM	86.11%
	PointNet + TCN	87.50%
	mmGaitNet	90.88%
Micro-Doppler Signatures	Multi-Channel CNN	83.33%

As shown in Figure 10, as the number of subjects increased, the performance of all networks degraded. However, the proposed spatial–temporal network still achieved the highest accuracy, demonstrating the robustness and stability of our network in the gait recognition task. The proposed network captured the motion dynamics of the human gait by combining PointNet and the Transformer encoder, fully exploiting the spatial–temporal structure of the 4D RPCVs.

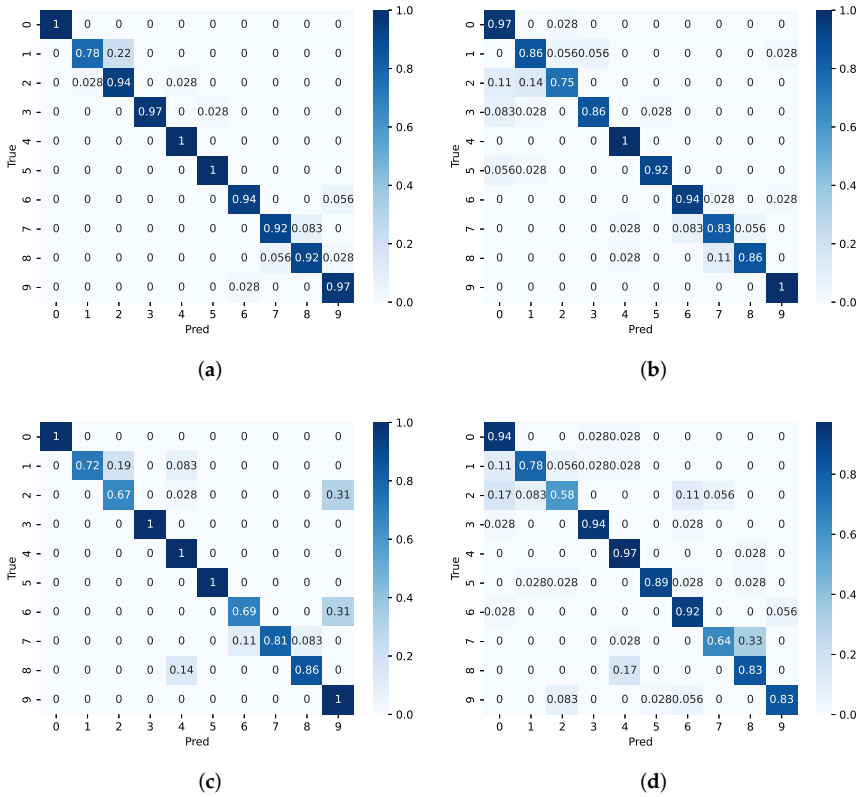


Figure 9. Confusion matrices of different gait recognition networks. (**a**) Spatial–temporal network; (**b**) mmGaitNet; (**c**) PointNet + TCN; (**d**) Multi-Channel CNN.

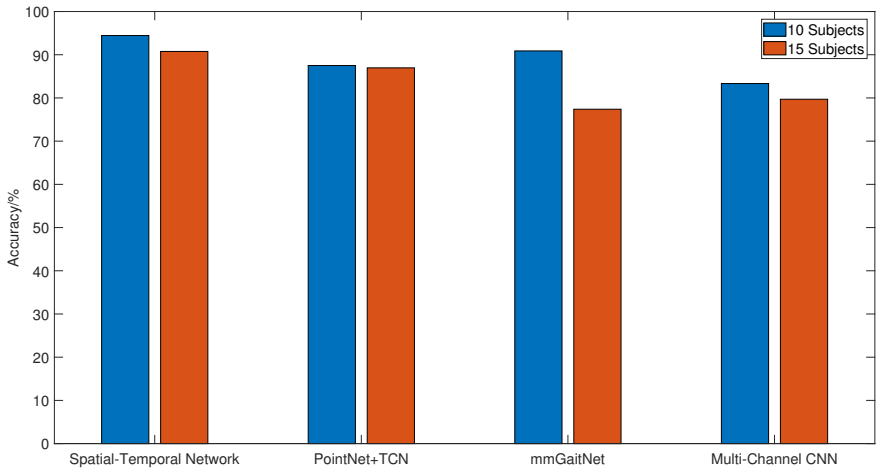


Figure 10. Comparison of different networks with the increase of the number of subjects.

5.3.2. Impact of Hidden Dimension

The hidden dimension *Dim* is correlated with the size and performance of the network, and we compared different sizes of a hidden dimension in this experiment. As shown in Table 4, with the increase of the hidden dimension, the size of the network increased, which was not conducive to the deployment of the network. It is worth noting that the performance of the network with a hidden dimension of 1024 slightly degraded compared to that with 512, potentially due to overfitting. After considering both the performance and computational complexity of the network, the hidden dimension was set to 512.

Table 4. Performance of network with different hidden dimensions.

Hidden dimension	256	512	1024
Accuracy	92.71%	94.44%	94.17%
Parameters	0.540 m	1.656 m	6.447 m

5.4. Discussion

This sub-section discusses the potential applications in real-time scenarios, as well as the limitations of the proposed radar-based gait-recognition method. The designed spatial–temporal gait recognition network, with 1.656 million parameters and 1.158 billion FLOPs, can be deployed on many commercial AI edge computing devices for real-time processing. The gait-recognition method in this study can be applied in many real-world scenarios. It has great potential application over traditional vision solutions in personalized surveillance systems such as smart homes and enterprise settings, where the number of individuals involved is a few tens. The sensing device used in the proposed method is a single radar module, making it easy to deploy with edge computing devices in most practical scenarios without requiring extensive additional hardware deployment in the environment.

Although radar-based gait recognition offers a non-invasive way of human identification, it is limited in certain cases. As a soft biometric, gait cannot be used to identify subjects within very large groups, since it is hard to separate each subject's gait representation from the radar echoes of a large crowd. In addition, abnormal walking patterns due to injuries may lead to poor gait recognition performance.

6. Conclusions

In this article, a 4D-RPCV-based spatial–temporal network for gait recognition was proposed. The 4D RPCV was introduced to characterize human gait. In the proposed network, PointNet was adopted to extract spatial features from sparse radar point clouds in each frame. Furthermore, a Transformer layer was employed to further exploit the spatial–temporal correlation across the 4D RPCVs, enabling the capture of motion dynamics in human gait. The experimental results demonstrated the effectiveness and robustness of the proposed spatial–temporal network, achieving an accuracy of 94.44% in identifying 10 subjects and 90.76% for 15 subjects.

In the future, related research will be continued from two aspects. On the one hand, to make the proposed network more general and robust, we will increase the number and diversity of the subjects in the experiment, as well as evaluate the network with various environmental settings. On the other hand, we will study the radar-based gait recognition network that is robust to environment changes. Radar sensing is easily affected by interference associated with the surroundings, for which it is important to enhance the environment adaptivity of the radar-based gait-recognition model. The potential of meta-learning methods in radar-based gait recognition will be explored, with the goal of rapid adaptation to new environments with minimal observations.

Author Contributions: Conceptualization, C.M.; methodology, C.M.; software, C.M.; validation, C.M.; formal analysis, C.M.; investigation, C.M.; resources, Z.L.; data curation, C.M.; writing—original draft preparation, C.M.; writing—review and editing, Z.L.; visualization, C.M.; supervision, Z.L.; project administration, Z.L.; funding acquisition, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Guangdong Provincial Science and Technology Plan Project under Grant 2021A0505080014, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515012873, and in part by the Guangzhou Key Research and Development Project under Grant 2023B01J0011.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Wan, C.; Wang, L.; Phoha, V.V. A survey on gait recognition. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–35. [CrossRef]
2. Fan, C.; Liang, J.; Shen, C.; Hou, S.; Huang, Y.; Yu, S. OpenGait: Revisiting Gait Recognition Towards Better Practicality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 9707–9716.
3. Singh, J.P.; Jain, S.; Arora, S.; Singh, U.P. Vision-based gait recognition: A survey. *IEEE Access* **2018**, *6*, 70497–70527. [CrossRef]
4. Shen, C.; Yu, S.; Wang, J.; Huang, G.Q.; Wang, L. A comprehensive survey on deep gait recognition: Algorithms, datasets and challenges. *arXiv* **2022**, arXiv:2206.13732.
5. Sepas-Moghaddam, A.; Etemad, A. Deep gait recognition: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 264–284. [CrossRef] [PubMed]
6. Muro-De-La-Herran, A.; Garcia-Zapirain, B.; Mendez-Zorrilla, A. Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors* **2014**, *14*, 3362–3394. [CrossRef] [PubMed]
7. Marsico, M.D.; Mecca, A. A survey on gait recognition via wearable sensors. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–39. [CrossRef]
8. Zhang, J.; Xi, R.; He, Y.; Sun, Y.; Guo, X.; Wang, W.; Na, X.; Liu, Y.; Shi, Z.; Gu, T. A Survey of mmWave-Based Human Sensing: Technology, Platforms and Applications. *IEEE Commun. Surv. Tutor.* **2023** , *52*, 2052–2087. [CrossRef]
9. Gurbuz, S.Z.; Amin, M.G. Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring. *IEEE Signal Process. Mag.* **2019**, *36*, 16–28. [CrossRef]
10. Le, H.T.; Phung, S.L.; Bouzerdoum, A. Human gait recognition with micro-Doppler radar and deep autoencoder. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 3347–3352.
11. Addabbo, P.; Bernardi, M.L.; Biondi, F.; Cimitile, M.; Clemente, C.; Orlando, D. Temporal convolutional neural networks for radar micro-Doppler based gait recognition. *Sensors* **2021**, *21*, 381. [CrossRef]
12. Yang, Y.; Ge, Y.; Li, B.; Wang, Q.; Lang, Y.; Li, K. Multiscenario Open-Set Gait Recognition Based on Radar Micro-Doppler Signatures. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–13. [CrossRef]
13. Chen, V.C.; Li, F.; Ho, S.S.; Wechsler, H. Micro-Doppler effect in radar: Phenomenon, model, and simulation study. *IEEE Trans. Aerosp. Electron. Syst.* **2006**, *42*, 2–21. [CrossRef]
14. Chen, Z.; Li, G.; Fioranelli, F.; Griffiths, H. Personnel recognition and gait classification based on multistatic micro-Doppler signatures using deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 669–673. [CrossRef]
15. Kim, Y.; Alnujaim, I.; Oh, D. Human activity classification based on point clouds measured by millimeter wave MIMO radar with deep recurrent neural networks. *IEEE Sens. J.* **2021**, *21*, 13522–13529. [CrossRef]
16. Meng, Z.; Fu, S.; Yan, J.; Liang, H.; Zhou, A.; Zhu, S.; Ma, H.; Liu, J.; Yang, N. Gait recognition for co-existing multiple people using millimeter wave sensing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 849–856.
17. Cheng, Y.; Liu, Y. Person reidentification based on automotive radar point clouds. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
18. Canil, M.; Pegoraro, J.; Rossi, M. MilliTRACE-IR: Contact tracing and temperature screening via mmWave and infrared sensing. *IEEE J. Sel. Top. Signal Process.* **2021**, *16*, 208–223. [CrossRef]
19. 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, 24 June–2 July 2021; pp. 269–282.
20. Guan, J.; Madani, S.; Jog, S.; Gupta, S.; Hassanieh, H. Through fog high-resolution imaging using millimeter wave radar. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11464–11473.
21. Cao, D.; Liu, R.; Li, H.; Wang, S.; Jiang, W.; Lu, C.X. Cross vision-rf gait re-identification with low-cost rgb-d cameras and mmwave radars. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2022**, *6*, 1–25. [CrossRef]
22. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
25. Lien, J.; Gillian, N.; Karagozler, M.E.; Amihood, P.; Schwesig, C.; Olson, E.; Raja, H.; Poupyrev, I. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graph. (TOG)* **2016**, *35*, 1–19. [CrossRef]
26. Li, X.; He, Y.; Jing, X. A survey of deep learning-based human activity recognition in radar. *Remote Sens.* **2019**, *11*, 1068. [CrossRef]
27. 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]

28. Xia, Z.; Ding, G.; Wang, H.; Xu, F. Person identification with millimeter-wave radar in realistic smart home scenarios. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
29. Pegoraro, J.; Rossi, M. Real-time people tracking and identification from sparse mm-wave radar point-clouds. *IEEE Access* **2021**, *9*, 78504–78520. [CrossRef]
30. 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]
31. Soumekh, M. Array imaging with beam-steered data. *IEEE Trans. Image Process.* **1992**, *1*, 379–390. [CrossRef] [PubMed]
32. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. Kdd* **1996**, *96*, 226–231.