

# **POINT TRANSFORMER-BASED HUMAN ACTIVITY RECOGNITION USING HIGH-DIMENSIONAL RADAR POINT CLOUDS**

MASTER THESIS REPORT

**Zhongyuan Guo**



# **POINT TRANSFORMER-BASED HUMAN ACTIVITY RECOGNITION USING HIGH-DIMENSIONAL RADAR POINT CLOUDS**

## **Thesis**

to obtain the degree of Master of Science  
in Electrical Engineering  
at Delft University of Technology  
to be defended publicly on August 31st

by

**Zhongyuan GUO**

Born in Shaanxi Province, China

This thesis has been approved by  
promotor: Prof. DSc. A. Yarovoy

Thesis committee:

Dr. F. Fioranelli,  
Dr. J. Dauwels,  
R. Guendel,

Technische Universiteit Delft  
Technische Universiteit Delft  
Technische Universiteit Delft



*Keywords: Deep Learning, Imaging Radar, Human Activity Recognition,  
Point Transformer*

An electronic version of this dissertation is available at  
<http://repository.tudelft.nl/>.

# ACKNOWLEDGEMENTS

My master's journey is coming to an end. When I look back on the two years of study, it has been a fantastic experience: pleasure with new friends, curiosity for new knowledge, and achievement upon finishing my master's project. Numerous things deeply impressed me. Despite all the restrictions due to the pandemic in the first year of study, I will always adore my study and life abroad in the Netherlands. Many thanks to everyone around me for making things better.

First, I would like to express my gratitude to my supervisors: Dr. Francesco Fioranelli and Mr. Ronny Guendel. At the beginning of my thesis, thanks for providing a creative and comfortable atmosphere during weekly meetings, which allows me to explore the topic freely. When I got stuck, they always offered holistic and professional advice to help me push further. Also, they are enthusiastic about giving me solutions when my laptop goes down. The most impressive thing is the hands-on teaching of academic skills, such as guidance for my academic writing and suggestions for academic presentation.

I am also grateful to Prof. Alexander Yarovoy. The first lesson I attended was the electromagnetics course taught by Prof. Yarovoy. His in-depth explanation of electromagnetic phenomena and cutting-edge technology attracted me deeply, which guided me to the MS3 group. I would also like to thank all the professors for their lectures and for expanding my horizon.

My sincere thanks extend to Ph.D. candidates and MSc students in the MS3 group. I would like to show my gratitude to Simin, Sen, Yue, and Max for attending the midterm presentation rehearsal and giving me suggestions and encouragement. I am also grateful for deep and mindful discussions with Liyuan, Xingzhuo, Yue and Qinyu when I had questions about radars. In addition, I would like to thank Yubin, Yongdian, Xubin, and Ignacio for providing me with technical support. In addition, I would like to thank the secretary of MS3 for arranging my defense.

Special thanks to my family. Without your support and effort, I would not have had the opportunity to study abroad, and your meticulous care fills my heart.

The end of my master's journey is also the end of my student status. Still, my curiosity to explore nature and society will be like the fire of Prometheus, always shining.

# ABSTRACT

Today, with increasingly aging population, healthcare systems in many countries need to improve their effectiveness, and the automatic Human Activity Recognition (HAR) technology can be beneficial. This can provide early diagnosis of changes in behavioral patterns in the home environment, without hospitalization, and detect critical events such as falls in a timely manner. In this area, radar-based HAR solutions are attracting the researchers' attention because no optical images are captured by radars, and thus respect of privacy and functionality in darkness can be guaranteed. Furthermore, no sensors need to be worn by the person being monitored.

Most previous work related to radar-based HAR employs image-like data representation such as spectrograms, range profiles, and snapshots of point clouds, and the information contained in these data representation is limited. For instance, spectrograms and range profiles cannot reflect the body shape of the subjects, while snapshots of point cloud do not contain Doppler or intensity information.

To overcome the limitation of these data representations, we propose to utilize the data from a mm-wave FMCW MIMO radar to create a novel data representation of point cloud with Doppler and intensity values, plus temporal information to achieve accurate HAR. Specifically, this thesis work focuses on the high dimensional radar point clouds and on a pipeline to generate and process this novel data representation. The proposed method combines the spatial information of point clouds with other features like Doppler, intensity/SNR, and time, expanding each point from 3D coordinates to a 6D vector. Hence, the movement of every part of the body can be expressed by those points. A module consisting of adaptive noise cancelation, frame selection, and resampling is proposed to process point clouds to match the input of subsequent classifiers.

Considering that the core of the self-attention concept matches well the information within point clouds, we investigate three self-attention based models as classifiers. These models can learn the spatial distribution of point clouds with their extra features with self-attention mechanism. The best combination of different input features, the positive contribution of the proposed adaptive noise cancellation method, and the performance of these three models are studied with experimental data from the MMActivity dataset and a purposely collected TU Delft (TUD) dataset.

# CONTENTS

<b>Acknowledgements</b>	<b>v</b>
<b>Summary</b>	<b>vi</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Problem formulation . . . . .	2
1.3 Thesis contributions . . . . .	3
1.4 Thesis structure . . . . .	4
<b>2 Literature Review</b>	<b>5</b>
2.1 Deep learning approaches for human activity recognition in radar . . . . .	5
2.1.1 HAR using CNN and its variants . . . . .	5
2.1.2 HAR using RNN and its variants . . . . .	8
2.1.3 HAR using transformer and its variants . . . . .	11
2.2 Summary and gaps . . . . .	12
2.2.1 Data generation . . . . .	12
2.2.2 Radar data representation . . . . .	13
2.2.3 Feature extraction network and classifier . . . . .	13
<b>3 Methodology</b>	<b>15</b>
3.1 Deep learning models . . . . .	15
3.1.1 Convolutional neural network . . . . .	15
3.1.2 Recurrent neural network . . . . .	16
3.1.3 Transformer and its variants . . . . .	17
3.2 Data pre-processing . . . . .	29
3.2.1 2D-FFT . . . . .	29
3.2.2 CFAR . . . . .	30
3.2.3 DOA estimation . . . . .	31
3.3 Proposed HAR pipeline . . . . .	31
<b>4 Dataset Preparation and Relevant Preprocessing</b>	<b>33</b>
4.1 MMAActivity dataset . . . . .	33
4.1.1 Radar information . . . . .	33
4.1.2 Measurement . . . . .	34
4.1.3 Dataset for this thesis . . . . .	36

4.2	TUD Dataset . . . . .	36
4.2.1	Radar information . . . . .	36
4.2.2	Measurement . . . . .	39
4.2.3	Basic dataset for this thesis. . . . .	40
4.2.4	Dataset with adaptive clutter cancellation . . . . .	41
4.2.5	Datasets of selected frames . . . . .	43
<b>5</b>	<b>Results</b>	<b>46</b>
5.1	Feasibility results of Using Point Transformer on MMA dataset . . . . .	46
5.1.1	Comparison between proposed pipeline and RadHAR pipeline . . . . .	46
5.1.2	Results with different input . . . . .	47
5.2	Results of using PT and PCT on TUD's dataset . . . . .	49
5.2.1	Results of different features as inputs . . . . .	49
5.2.2	Results with adaptive clutter cancellation . . . . .	51
5.2.3	Comparison among three attention-based models. . . . .	51
5.2.4	Leave-one-subject-out test. . . . .	53
5.2.5	Person recognition . . . . .	54
<b>6</b>	<b>Conclusion and Future Work</b>	<b>57</b>
	References . . . . .	60



# LIST OF FIGURES

3.1	Example of CNN architecture from the literature [1] . . . . .	15
3.2	Model of RNN architecture . . . . .	16
3.3	Illustration of (a) LSTM and (b) GRU. (a) $i$ , $f$ and $o$ are the input, forget and output gates, respectively. $c$ and $\tilde{c}$ denote the memory cell and the new memory cell content. (b) $r$ and $z$ are the reset and update gates, and $h$ and $\tilde{h}$ are the activation and the candidate activation [2] . . . . .	17
3.4	An example of self-attention analyzing a sequence of text [3] . . . . .	18
3.5	Graphical illustration of scaled dot-product attention calculation [4] . . . .	19
3.6	(left) Scaled Dot-Product Attention. (right) Multi-Head Attention, which consists of several attention layers running in parallel. . . . .	20
3.7	Example of the Transformer - model architecture. . . . .	21
3.8	Details of the point transformer layer . . . . .	22
3.9	Example of point transformer networks for semantic segmentation (top) and classification (bottom) [5] . . . . .	23
3.10	Point transformer block (left) and transition down block (right) . . . . .	23
3.11	Example of the PCT architecture: the encoder mainly comprises an Input Embedding module and four stacked Attention modules. The decoder mainly comprises multiple Linear layers. The numbers above each module indicate its output channels. MA-Pool concatenates Max-Pool and Average-Pool. LBR combines Linear, Batch Norm, and ReLU layers. LBRD means LBR followed by a Dropout layer [6] . . . . .	25
3.12	Overview of the Point Transformer architecture, which consists of two branches to generate local and global features. SortNet produces an ordered set of local features against the global structure of the input point cloud. Depending on the task, classification or part segmentation heads are employed. Red Boxes denote sorted sets, only for the segmentation part [7] . . . . .	26
3.13	Overview of the SortNet: a score is learned from a latent feature representation to extract important points from the input. Local features are aggregated from neighboring points. SortNet outputs a permutation invariant and sorted feature set. Red boxes denote sorted sets [7] . . . . .	27
3.14	Overview of the radar data preprocessing from radar cube to point cloud .	29
3.15	Principle of OS-CFAR algorithm: orange blocks represent the training cells, shallow yellow blocks stand for the guard band cells and green block is the CUT [8] . . . . .	30
3.16	Overview of the proposed pipeline to address HAR problem . . . . .	31
4.1	Picture of IWR1443 radar board [9] . . . . .	34
4.2	MMAActivity dataset collection setup [10]. . . . .	35

4.3	The storage format of one point in the MMAActivity dataset.[11]	35
4.4	Picture of the four-device cascaded AWR2243 radar board	37
4.5	Data measurement setup in an office-like room at TU Delft [12, 13]	39
4.6	Classes and data distribution of the TUD dataset [12]	40
4.7	Example of data from a 6D point cloud file: from left to right are the coordinates of x, y, z, Doppler, SNR, and frame index, separated by commas	41
4.8	Visualization of point cloud of a standing human subject. The red dot represents the centroid of the point cloud.	42
4.9	Comparison of point cloud without/with proposed adaptive clutter cancellation. The red sphere shows the threshold to determine whether a point is considered clutter or not.	43
4.10	Visualization of point cloud of the 'bending' activity for 20 frames. Each plot shows the point cloud of a single frame, and the color bar indicates the Doppler value. From left top to right bottom corner, the point clouds from frame 0 to frame 19 are shown	44
5.1	Normalized confusion matrix for Point Transformer to classify the five motions in the MMAActivity dataset, with 1360 as the number of input points	47
5.2	Classification accuracy and F1 score values with different input features from the point clouds. x, y, and z indicate the spatial coordinates, and D, I, and T indicate Doppler, intensity, and time, respectively	48
5.3	Classification accuracy and F1 scores with different number of input points in the point cloud, from 128 to 1024	49
5.4	Normalized confusion matrix for Point Transformer to classify six motions and postures with (a) only coordinates, (b) coordinates and Doppler, (c) coordinates and intensity, (d) coordinates and time, (e) coordinates, Doppler, intensity and time, (f) coordinates, Doppler and time	50
5.5	Classification F1 score for 3 different models with decreasing number of frames as input, where the horizontal axis represents the number of frames. The F1 scores are the average of 5-fold cross validation	52
5.6	Classification F1 score in leave-one-subject-out test when using 20 frames, where the horizontal axis represents the index of the left-out subject, and S represents specific subjects	54
5.7	Normalized confusion matrix for Point Cloud Transformer to classify seven human subjects through their 'bending' motions.	55
5.8	Normalized confusion matrix for Point Cloud Transformer to classify seven human subjects through their 'standing' postures.	56

# LIST OF TABLES

1.1	Summary of pros and cons of different types of sensors for HAR [14] . . . . .	2
2.1	Summary of relevant radar-based HAR studies using CNN, where # classes represents the number of classes. . . . .	6
2.2	Summary of relevant radar-based HAR studies using RNN, where # classes represents the number of classes . . . . .	9
2.3	Summary of radar-based HAR studies using Transformer, where # classes represents the number of classes and - means that the paper did not indicate the number of classes. . . . .	11
3.1	Summary of the numbers of parameters in the three models considered in this work. . . . .	29
4.1	Classes distribution of the MMActivity dataset . . . . .	34
4.2	Relation among the number of points, frames and duration in adjusted MMActivity dataset . . . . .	36
4.3	Waveform parameters and derived features of the radar. . . . .	38
4.4	Mapping relation between number of frames and size of resulting input data	45
5.1	RadHAR results from [10]: test accuracy of different activity recognition classifiers trained on the MMActivity Dataset. . . . .	47
5.2	F1 score of Point Transformer with and without the proposed adaptive clutter cancellation . . . . .	51

# ABBREVIATIONS

- ACC** Adaptive Clutter Cancellation. 41, 42, 51, 58
- CA** Cell-Average. 30
- CFAR** Constant False Alarm Rate. 7, 15, 30–32, 35, 40, 44, 58
- CNN** Convolutional Neural Network. 2, 5–8, 10–13, 15–18, 47
- CTC** Connectionist Temporal Classification. 10
- CUT** Cell Under Test. 30
- CV** Computer Vision. 3, 5, 11, 12
- CVD** Cadence Velocity Diagram. 12, 14
- DCNN** Deep Convolutional Neural Network. 7–9
- DL** Deep Learning. 2–5, 10–15, 17, 32, 47, 59
- DOA** Direction Of Arrival. 35, 40, 58
- DSP** Digital Signal Processor. 35
- FFT** Fast Fourier Transform. 15, 29–32, 40, 58
- FMCW** Frequency-Modulated Continuous-Wave. 31, 33
- FPS** Farthest Point Sampling. 28, 45
- GAN** Generative Adversarial Network. 13, 59
- GRU** Gated Recurrent Unit. 8–10, 16, 17
- HAR** Human Activity Recognition. vi, 1–10, 12–14, 16, 31, 38, 46, 57–59
- HRRP** High Resolution Range Profiles. 7, 9, 16
- LSTM** Long-Short Time Memory. 8–11, 16, 17
- MIMO** Multiple-Input and Multiple-Output. 2, 7, 29, 31, 33
- MLP** Multilayer Perception. 10, 12, 17

**MUSIC** Multiple Signal Classification. 31

**NLP** Natural Language Processing. 3, 5, 8, 11, 12

**OS** Ordered-Statistic. 30

**RGB** Red Green Blue. 15

**RNN** Recurrent Neural Network. 2, 8–11, 13, 15–18, 47

**ROS** Robot Operating System. 8, 10, 35

**RX** Receiver. 36, 38

**SNR** Signal-to-Noise Ratio. 12, 31, 32, 40–42

**ST-GCN** Spatial-Temporal Graph Convolutional Network. 7

**SVM** Support Vector Machine. 10

**TDMA** Time Division Multiple Access. 36

**TX** Transmitter. 36, 38

**ViT** Vision Transformer. 11–13



# 1

## INTRODUCTION

*This chapter describes the background on the radar-based HAR in section 1.1, the problem formulation according to the gaps of existing research studies in section 1.2, the contributions of this work and finally the structure of this thesis in sections 1.3 and 1.4, respectively.*

### 1.1. BACKGROUND

Today, most people can expect to live into their sixties and beyond. Every country in the world is experiencing growth in both the size and the proportion of older people in the population [15]. This challenge of aging population is pushing toward novel healthcare provision that evolves from the traditional hospital-based system. For example, for a person-centric healthcare, patients can be monitored in their home remotely via modern technology. Home-centric healthcare devices can recognize the situation of aged people and patients and thus enhance the life quality of these aged people and patients by minimizing the disruption to their usual routine and lifestyle [16].

In addition to being used for eldercare and healthcare as part of assistive technologies, in isolation or in ensemble with other technologies such as Internet of Things (IoT) [17], HAR also plays a significant role in many applications such as gaming, smart screen interaction, sign language interpretation (gesture recognition), remote monitoring, and human-computer interaction. Therefore, we can say that HAR occupies an important place in improving our life quality and reducing living cost [18].

From a source-domain perspective, HAR can be categorized into using external and wearable sensors. For external sensors, cameras are used for HAR very commonly because of their high accuracy and robustness for different backgrounds. However, processing video sequences shows great disrespect to users' privacy, high computational cost and poor performance without light. Radars have also been typical external sensors to capture physical information from human targets compared to cameras. For instance, radars can directly estimate the range and Doppler/velocity information from human

Table 1.1: Summary of pros and cons of different types of sensors for HAR [14]

Source domain	Sensor type	Pros	Cons
External	Cameras	a) Robustness against different backgrounds b) Storage of records	a) Privacy issues b) High computational cost c) Dependency on light
	Radars	a) Privacy is ensured b) Functionality in darkness	a) Sensitive to the direction of arrival of motions b) Required installation and calibration c) Directional functionality
Wearable	Embedded sensors	a) High velocity accuracy b) Privacy is ensured	a) Uncomfortable b) Expensive
	Smartphones	a) Privacy is ensured b) No extra devices	a) Low accuracy

subjects. Specifically, radars collect and analyze the reflected electromagnetic waves to obtain the features to recognize the situation of the monitored subjects. Unlike cameras, radars do not capture optical images or videos from the monitored subjects, which cause less problems in terms of privacy and guarantee functionality in darkness.

For wearable sensors, the majority of sensor types are gyroscopes and accelerometers embedded in clothes or built in smartphones. These sensors are generally not considered to be privacy sensitive and have high velocity accuracy [14], but may cause some discomfort due to these embedded sensors. Unlike wearable sensors, radar systems do not require users to wear, carry, or interact with any additional electronic device or modify their daily routine and behavior [19]. Pros and cons of these mainstream sensors for HAR are listed in table 1.1.

## 1.2. PROBLEM FORMULATION

After reviewing the literature of radar-based HAR, it is realized that most of the features and data representations extracted from radar raw data to solve radar-based HAR problems are spectrograms, range-Doppler heatmaps and range profiles, which are stored as an image. Therefore, with the help of those Deep Learning (DL) networks for processing images such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), it is successful to utilize DL classifiers to recognize human activities with images as input. For RNN and its variants, the image data are cut into slices along the time axis so that the inputs are sequential signals.

With the popularity of the mm wave Multiple-Input and Multiple-Output (MIMO) radar, radar point clouds are available, and this data representation can be used to distinguish different human activities since point clouds can better reflect the posture of the human body and its shape in the 3D space. However, in most current work of using point clouds to solve HAR, they either take pictures of point clouds [20] or embed point clouds into a cube [10]. Essentially, this is equivalent to 'flatten' the point clouds into lower-dimension representations, sometimes just still images. Moreover, several studies only used the coordinates of the points while ignoring the other features of points such



as Doppler and intensity. Therefore, the potential of using point clouds with more features as data representation to solve HAR is considerable.

From DL model perspective, Transformer has dominated Natural Language Processing (NLP) since it was proposed, and its variants like Vision Transformer also perform well in the Computer Vision (CV) field, showing the good performances of this family of models for many tasks. However, the combination of Transformer-based DL models and radar data is so far rare, and most related works are listed in Table 2.3. All research indicates that the use of a Transformer-based model can result in better performance, thus opening up an opportunity for research in the radar-based HAR field.

In summary, the research gaps from the literature can be listed as the following points:

- The research for HAR using **point clouds with high dimensional information** is lacking.
- Transformer and its variants showed superiority compared to other DL models, but **the applications of Transformer in radar-based HAR are so far limited**.

Based on these research gaps, the problem to be studied in this thesis can be formulated as:

**"Can we combine attention-based DL models and point clouds with high dimension features such as Doppler, intensity, and time to achieve accurate HAR, and how can these added information of point clouds affect the classification results?"**

### 1.3. THESIS CONTRIBUTIONS

The main contributions of this thesis are summarized in the following aspects:

- This work developed and tested a pipeline that solely utilized point clouds as input data to train attention-based DL models, and this pipeline can classify both motions and static postures using only one classifier.
- An adaptive clutter cancellation method is developed to remove the clutter in point cloud data. This increases the performance by 2%-5%, varying for different input features. The effect of selecting different input features was also investigated and quantified.
- This work investigates performance of three different attention-based networks with only point clouds as input. It shows that the classification results of Hengshuang's model is slighter better than Menghao, but F1 score of them are obviously higher than Nico's model. With fewer frames of input data, the F1 score of three models decrease, but the decline is more significant in Nico's model.
- Part of this thesis is being written as a paper to be submitted to the IEEE Radar Conference, which will take place in San Antonio, USA, in May 2023.

## 1.4. THESIS STRUCTURE

The following chapters in this thesis are structured as follows. Chapter 2 reviews previous researches related to addressing HAR problems, and this chapter is divided based on different types of DL classifiers. Chapter 3 introduces the architectures of three attention-based networks investigated in this thesis and other DL models related to the literature review. Then, there is an overview of the data preprocessing algorithms that convert complex radar signals into point clouds. Moreover, the pipeline proposed in this thesis is also described. Chapter 4 gives an introduction to the two datasets used to test the proposed pipeline, and then describes how to adjust the data format of these two datasets to match the pipeline. Chapter 5 shows the classification results using two datasets and three networks with different input sizes and input features. In Chapter 6, the conclusion and the future work are presented.

# 2

## LITERATURE REVIEW

*This chapter gives an account of related work on the topic of deep learning-based human activity classification from the literature. In section 2.1, relevant previous research studies are introduced based on the various DL models used, from CNN to Transformer. Section 2.2 provides a summary from the aspects of data generation, data representation, and classifier choice, and indicates the gaps between previous works and this research.*

### 2.1. DEEP LEARNING APPROACHES FOR HUMAN ACTIVITY RECOGNITION IN RADAR

Thanks to the rapid development in CV and NLP field, numerous well-performing DL algorithms for extracting features and classifying radar data have been presented in the literature. A deep learning model has multiple processing layers to learn high-level representations automatically[21], so that difficult and complicate classification and recognition tasks could be solved. Therefore, radar-based HAR could also learn from those models. In this section, we investigate several classical and novel DL models that can be utilized for HAR.

#### 2.1.1. HAR USING CNN AND ITS VARIANTS

CNN is currently the state-of-art network for almost any computer vision task. As an end to end classification approach, it can extract features from an image automatically, compared to the conventional handcrafted feature-based classifiers that were used in the initial stages of research into radar-based HAR [22]-[23]. For radar-based HAR, most data representations after processing the I/Q raw data include range-Doppler heat maps, spectrograms, and time-range heat maps. These can be forwarded to a classifier as image matrices. Therefore, it is theoretically feasible to apply CNN for radar-based HAR by taking the radar feature maps as inputs and actually CNN is a popular DL model in radar-based HAR field. More specific cases are reviewed and sorted by the different radar data representations. The table 2.1 summarizes papers reviewed for HAR using CNN, with additional discussion provided for each group of papers clustered with respect to the

data representation used as input of the network.

Table 2.1: Summary of relevant radar-based HAR studies using CNN, where # classes represents the number of classes.

Paper	#classes	Data format	Network architecture	Scenario	Radar type
[22] Lang	7	spectrogram	CNN	MOCAP simulated	UWB radar 3-5GHz
[24] Bresnahan	8	spectrogram	CNN	in-vehicle	UWB radar 3-5GHz
[25] Zhang	4	spectrogram scalogram	CNN	lab	Doppler radar 5.8GHz
[26] Le	3	spectrogram	Bayesian optimized CNN	lab	Doppler radar 24GHz
[27] Shao	7	time-range	DCNN	lab	UWB radar 0.7-7GHz
[28] Alujaim	7	point cloud image	CNN	lab	FMCW radar 77GHz
[29] Kim	7	spectrogram	DCNN	indoor	Doppler radar 2.4GHz
[30] Park	5	spectrogram	CNN	Aquatic	Doppler radar 7.25GHz
[31] Shao	8	spectrogram	CNN with inception module	indoor 8 subjects	FMCW radar 24GHz
[32] Trommel	6	spectrogram	DCNN	indoor 29 subjects	CW radar X-band
[20] Lee	10 5	point clouds	ST-GCN	MARS MMActivity	FMCW radar 77GHz

## A Spectrograms

Currently, spectrogram is the most common feature extracted from radar data using time-frequency analysis. A spectrogram reveals the instantaneous spectral content of the time-domain signal and the variations of the spectral content over time [33]. Therefore, in many HAR works, spectrograms are fed into CNN to extract features of different human activities such as [22, 24, 25, 29].

To begin with, in [22], Yue Lang et al. employed a CNN to classify seven human activities based on spectrograms generated from MOtion CAPture database (MOCAP). Since the data were simulated, varying levels of noise were added and all the average accuracy was above 90%, which showed the robustness of the CNN utilized for radar based HAR. Besides, the spectrograms in grayscale were also fed, and the results showed the eventual accuracy remained unchanged but its coverage time increased. Hoang Thanh Le et al. [26] used Bayesian optimization model for optimizing hyperparameters of CNN such as the number of convolution layers, the learning rate, the momentum and the L2 regularization coefficient. Meanwhile, the spectrogram-based classifications of human aquatic activities and driver head motions were also investigated in [30] and [24] respectively, where data were generated from real measurements. Worth to mention, micro-Doppler signatures can also be used for personnel recognition as a personal identification [31]. Additionally, spectrograms were utilized for human gait and gesture classification

in [32] and in [25] respectively.

### B Range information

The above studies address only the radial velocity information of human subjects, but neglect the range information during the activities. For the majority of human activities, errors in spectrograms are easily caused by low speed movements that generate little Doppler, that can be confused with static clutter in the scene. However, High Resolution Range Profiles (HRRP) can differentiate the motions when the difference of spectrograms are not obvious [27]. To obtain HRRP, a radar with a wide bandwidth is needed because the resolution is proportional to the inverse of the bandwidth. UWB radar can provide HRRP due to its small pulse duration or its chirp bandwidth [34]. Otherwise, the human subject will look like a large point. Therefore, in HAR, range information is usually bound to UWB radar or FMCW radar with wide bandwidth [27, 35, 36]. Furthermore, the Doppler information is dependent on the aspect angle between the line of sight of the radar and the movement trajectory, while range information is not.

In [37], range information was introduced for fall detection, and it decreased the probability of false alarms. For human motion classification, in [27], Yuming Shao et al. investigated the method of recognizing human motions by HRRP with a Deep Convolutional Neural Network (DCNN), the results of which showed more robustness than the micro-Doppler signatures, especially for unknown radial velocity conditions.

### C Point clouds

Radar point cloud represents the possible point scattered by an object. Usually, it contains 3D coordinates information plus the intensity and Doppler features, so it is easy to describe the shape and size of an object, including a human subject. MIMO millimeter-wave radar can generate 3D point cloud with some target detection algorithms such as Constant False Alarm Rate (CFAR) and estimation algorithms to measure the range, angle, Doppler and intensity from every cell selected by the CFAR.

Radar point cloud is relatively sparse compared to Lidar due to the wavelength and hardware noise. To overcome the sparsity, in [13] the author aggregates several frames over time to generate a relatively dense point cloud, but it comes at a price that multiple activities can be merged, and it is impossible to classify the symmetric motions such as sitting down and standing up. In [28], Ibrahim Alujaim et al. measured 7 human motions using a 77GHz FMCW radar and calculated the direction of arrival angle to obtain the 3D point clouds. Then they classified the motions by taking the point cloud images as the input of a CNN.

In [20], Gawon Lee et al. designed a Spatial-Temporal Graph Convolutional Network (ST-GCN) based model to process point cloud data from MARS[38] dataset and MMActivity dataset[10]. MARS contains the point cloud data from 10 different indoor activities from 19 human subjects, collected using a MIMO 77GHz FMCW

radar. The MMActivity dataset directly includes the point cloud data of 5 indoor motions, collected using TI's IWR1443BOOST radar with Robot Operating System (ROS).

### 2.1.2. HAR USING RNN AND ITS VARIANTS

With the successful implementation of RNN in speech recognition and NLP, the performance of RNN processing temporal sequences was proved to be excellent and thus, RNN caught the researchers' attention in HAR. RNN is famous for its ability to remember the previous information, that is to say, the current outputs are influenced by previous inputs. But gradients will exponentially shrink down as we back propagate because the gradient is calculated at each step, and this is the so-called 'vanishing gradient' problem. This means it is too difficult for the RNN to learn over many timesteps.

To solve this problem, there are two variants of RNN proposed, Gated Recurrent Unit (GRU) [39] and Long-Short Time Memory (LSTM)[40]. In GRU, update gate and reset gate are introduced to decide whether the memory cell should be updated and if the previous cell state is important, respectively. In LSTM, the forget gate decides what information should be kept and what should be forgotten, and the output gate determines what the next hidden state will be. Thus, the problem of vanishing gradient is overcome. Therefore, in HAR, many researchers opt for these two network architectures rather than original RNN.

The radar signals of human motions have high temporal sequentiality: human motions are continuous in time and not separated snapshots of individual activities [41]. However, CNN cannot utilize such sequential information to learn more relevant features. RNN can learn more information through the time-varying radar signals and take advantage of the sequentiality of human motions.

The input of RNN can be different data representation of radar data in sequence format, but can also be the features extracted by a CNN. So the following sub-sections will be organised in terms of papers using RNN only, and using hybrid model, i.e. models that combine CNN and RNN together. The summary of papers reviewed for HAR using RNN can be seen in table 2.2.

#### A RNN only

RNN can learn the temporal information of the radar data by treating them as a sequential signal rather than individual images. For spectrograms and time-range maps, these heatmaps will be cut into slices along the time axis to be temporal sequences as the input of a RNN. Usually, multiple RNN layers can be stacked to extract more generalized sequential features like the structure in [41], where two layers of LSTM are stacked and connected to an output layer to generate the probabilities of six different human motions.

One of the advantages of using RNN is that low computational power is needed, especially for embedded system [42]. In [41], the authors also compared the performance of DCNN and the network they proposed. The result showed the RNN

Table 2.2: Summary of relevant radar-based HAR studies using RNN, where # classes represents the number of classes

Paper	#classes	Data format	Network architecture	Scenario	Radar type
[35] Li	6	HRRP	LSTM	MOCAP	UWB radar 4GHz
[42] Jiang	6	spectrogram	LSTM & GRU	lab 72 subjects	FMCW radar 5.8GHz
[41] Wang	6	spectrogram	stacked LSTM	lab	CW radar 25GHz
[43] Zhang	8	stacked time-range	3D-CNN+LSTM+CTC	lab	FMCW radar 24GHz
[44] Park	5	amplitude spectrum	GRU	lab(indoor)	UWB radar 6.0-8.5GHz
[45] Tang	4	spectrogram	attention-based LSTM	outdoor	pulse radar L-band
[46] Wang	11	range-Doppler	CNN+RNN	lab	solis sensor
[36] Jian	4	HRRP	attention-augmented GRU	indoor	UWB radar 1.6-2.2GHz
[47] Guo	6	3D image frame	CNN+RNN	indoor	FMCW radar
[48] Klarenbeek	4	spectrogram	LSTM	indoor 29 subjects	CW radar X-band
[49] Sadreazami	2	range-Doppler	stacked LSTM	indoor	UWB radar 5.9-10.3GHz
[50] Sun	7	range-angle	LSTM	indoor	FMCW radar 77GHz
[10] Singh	5	point clouds	CNN+LSTM	lab	FMCW radar 77GHz
[51] Li	6	micro-Doppler signatures	bi-LSTM	lab	FMCW radar 5.8GHz
[52] Li	12	spectrogram	bi-LSTM	lab	FMCW radar 24GHz UWB radar X-band

network had better accuracy even if the number of parameters is one eighth of the DCNN, which means employing RNN can significantly decrease the computational load.

The comparison of GRU and LSTM was made in [42]. There are six target classes in this classification task, and the accuracy of GRU is slightly higher than LSTM. Moreover, a data augmentation algorithm was proposed based on the RNN property of being able to handle any length of input data. In this augmentation algorithm, the original samples are cut into different lengths to increase the number of samples. Hence, the overfitting phenomenon was suppressed.

Utilizing range information as the input of RNN was studied in [35] and [36] with the architecture of LSTM and GRU, respectively. In the investigation of Yuan He et al., simulated one-dimensional HRRPs were fed into bi-directional LSTM and uni-directional LSTM for comparison [35]. The results demonstrated that using HRRP for HAR can still obtain comparable results, though HRRPs are not as intuitive as

spectrograms and bi-directional LSTM has higher accuracy and shorter computation time. In [36], an attention-augmented sequential classification method was proposed by Qiang Jian et al. After using GRU to extract semantic features of various human activities, the attention mechanism is deployed to enhance the correlation of temporal output sequences of GRU. That is, the attention reinforcement replaces the stacking of multiple RNN layers to acquire high accuracy.

Apart from these, RNN can also be utilized for radar classification tasks in other scenarios such as fall detection [49][50], hand gesture recognition citepark2021hand, target recognition [45] and human gait classification [48]. The novel part in [50] is that, to reduce the redundancy in spatial domain, a radar low-dimension embedding (RLDE) algorithm is employed to preprocess the range-angle reflection heatmap sequence. To be more specific, RLDE concatenates horizontal plane and vertical plane as the input of LSTM. This network architecture not only reduces the computational complexity and memory consumption as an alternative of CNN, but also outperforms the state-of-the-art with 3% increases on F1\_score.

## B Hybrid models

Each DL model has its own property and advantages, so it is not possible for a single model to be proficient in all the classification tasks. Hybrid models can take advantages of each model's own strength by integrating several network architectures to acquire better performance. In HAR, it is common to see RNN and CNN are combined because CNN does well in extracting abstract and hierarchical features from spectrograms and other data representations, while RNN and its variants are good at exploiting the temporal information. How these two models can be exactly combined is for example described in [43][47][46][10]. The pipelines in these papers are similar: data collection, preprocessing, CNN, RNN and final prediction.

Saiwen Wang et al.[46] generated range-Doppler images from radar echos and then leveraged a CNN to extract the spatial features and a LSTM to capture the temporal information to predict the probabilities of hand gestures with softmax. Compared with [46], Zhenyuan Zhang et al. [43], introduced a 3D-CNN for short spatial-temporal modeling to process stacked spectrograms. At the end of the network, to recognize hand gestures continuously with unsegmented data, a Connectionist Temporal Classification (CTC) layer was introduced to process the results of LSTM after softmax function. The results of ablation experiment demonstrated the CTC and the hybrid models can indeed bring increases on accuracy. A significant difference between [10] and the other three papers is that Singh et al. utilized the point cloud data generated directly from a millimeter-wave radar through ROS without any preprocessing. They compared the performance of various classifiers such as Support Vector Machine (SVM), Multilayer Perception (MLP), bidirectional LSTM and CNN+LSTM and the result demonstrated that the hybrid model could obtain the highest accuracy.



### 2.1.3. HAR USING TRANSFORMER AND ITS VARIANTS

Transformer is a novel DL model using only attention mechanism and was originally designed to resolve the problem that RNN cannot be parallelized and thus improve the computational efficiency [53]. Since transformer was first proposed in 2017, it has dominated in NLP. Vision Transformer (ViT)[54], as a variant of Transformer, also outperforms many state-of-the-art models in CV field. However, for objects classification with radar, there are only a few studies using transformer and its variants. Three related papers are summarized in table 2.3.

Table 2.3: Summary of radar-based HAR studies using Transformer, where # classes represents the number of classes and - means that the paper did not indicate the number of classes.

Paper	#classes	Data format	Network architecture	Scenario	Radar type
[55] Zheng	8	range-Doppler range-azimuth range-elevation	CNN+Transformer	in-vehicle	FMCW radar 77GHz
[56] Chen	-	CVD & spectrogram	ViT	indoor 98 subjects	FMCW radar 24GHz
[57] Bai	5	point cloud	radar Transformer	outdoor	FMCW radar 77GHz

With the limited number of papers, the usage of transformer in radar can be divided into following three categories:

#### A Using CNN+Transformer [55]

Usually, a sentence contains up to dozens of words, so in the Transformer model, the upper limit for the input sequence length is 512[53]. However, the size of an image for DL model is 224 by 224, and it is not feasible to feed spectrograms and other image-like data representations of radar data into the Transformer, since the pixels in these feature maps are obviously far more than the given limit. Therefore, in [55], CNNs are employed to extract spatial features and reduce the length of input data without changing the architecture of the Transformer. Range-Doppler maps, range-azimuth maps and range-elevation maps were generated from the radar echoes of different hand gestures. The input data of the Transformer were the spatial features extracted and concatenated from these three maps by three independent convolutional layers and a fusion block. After the Transformer, a fully connected layer and softmax function were deployed to predict the probability distribution among 8 different gestures. Besides, the most suitable number of Transformer encoder modules was studied to be 3 and a comparison with other mainstream models demonstrated this model has higher accuracy than 3D-CNN and CNN+LSTM.

#### B Using ViT

ViT is modified for images input by dividing an image into  $16 \times 16$  patches and taking each patch as a word in a sentence [54]. Therefore, it is feasible to directly utilize spectrograms and other radar images as input. In [56] the authors

proposed a dual-stream ViT pipeline for radar gait recognition. Shiliang Chen et al. represented the radar signal as spectrograms and Cadence Velocity Diagram (CVD) and then used the ViT to split the spectrogram and CVD into patches and embed their positions separately, so that the information embedded in different frequency bands could be effectively extracted. Then an attention-based fusion block integrated the discriminant features from these two representations and a MLP was connected to complete the classification task. The experimental results showed that the accuracy of this proposed network is higher than other CNN network such as VGG, AlexNet, and ResNet.

### C Using radar Transformer

Radar Transformer takes 5-dimensional point clouds as input and each object point contains 3D coordinates, Doppler velocity information, and Signal-to-Noise Ratio (SNR) or a related intensity quantity. The design of radar Transformer [57] refers to the attention mechanism in [53] and hierarchical feature extraction as well as fusion of global and local features in [58]. In the feature encoder part of the radar Transformer, the radar point clouds are first mapped into a feature vector and are then divided into two branches. One passes through three stacked set abstractions and vector attention modules to gradually extract deep local features, while the other one goes to three stacked attention modules to extract global features. The local features from each hierarchy are concatenated with global feature of the same hierarchy as new global features. A scalar attention module integrates the final global features and a MLP is deployed to predict the class of each object.

## 2.2. SUMMARY AND GAPS

In this section, summary and gaps are indicated from aspects of data generation, data representation and feature extraction network.

### 2.2.1. DATA GENERATION

Unlike CV and NLP, only a few open source datasets for radar-based HAR are available. Therefore, for each research project, the researchers have often to generate radar data by themselves. Based on the papers reviewed above, methods to generate radar data can be divided into three main categories:

- Experiments

Radar data from experiments are the most common in HAR researches and the data are realistic due to multi-path effect, noise, and clutters captured. However, collecting data from experiments is time-consuming and expensive, especially for DL models which are data-driven: 98 human subjects spent two weeks accomplishing a data set [56].

- Simulation

To avoid complex and time-consuming measurements, simulation is sometimes adopted due to its convenience [22][35] and most of the simulated data are from the CMU MOCAP dataset. This type of data are generated based on kinematic

models where human skeleton motions are simulated to analyze the scattering behavior and model the radar signal [59][60]. Besides, Generative Adversarial Network (GAN), as a DL model, is also used to generate radar data [61][62].

- **Transfer learning**  
Transfer learning refers to transferring data from other source domains such as speech signal and image to radar like such as spectrogram [63].

A big challenge for solving HAR problems with DL methods is that these models require numerous data. For instance, the dataset used for training Transformer is the standard WMT 2014 English-German dataset which consists of about 4.5 million sentence pairs [53]. However, it is extremely time-consuming and expensive to set up such a dataset with radar data.

### 2.2.2. RADAR DATA REPRESENTATION

There are several types of data representations in radar, but the most significant point is to find the optimal data representation for a specific HAR problem. The most common radar data representation for DL is 2D image-like format including spectrogram [22], range-Doppler [49], time-range [35], amplitude spectrum [44], range-angle [55] and point cloud image [28]. For CNN these representations are treated as images while for RNN, these representations are cut into temporal sequences as input. However, by stacking these 2D representations, we can obtain 3D data representations and utilize 3D-CNN to extract features. For instance, in [43], many time-range maps are stacked and in [64], Erol et al. created a tensor structure, called radar data cube, by stacking consecutive range-Doppler frames.

Apart from these conventional data representations, a point cloud is a set of many points with associated values. Except the coordinate information, each point can also contain Doppler and intensity information. The number of researches using point cloud to solve HAR problem are so far limited. In [28], point clouds are treated as images while in [10] and [20], only the coordinates of points are used for HAR. Although five dimensional features of point clouds are utilized in [57], the objects are automotive objects such as car, bus and cyclist. Therefore, the research for radar-based HAR using point clouds with high dimensional information is lacking.

### 2.2.3. FEATURE EXTRACTION NETWORK AND CLASSIFIER

There have been numerous researches using CNN, RNN, variants of RNN, and hybrid models but only a few works using Transformer and its variants to solve HAR problems. The hyperparameters of DL models can be optimized by machine learning methods such as Bayesian optimizer [26]. CNN treats the input radar data as images mainly considering pixel-related features, while RNN and its variants treat radar data as temporal sequences focusing on the temporal relations. The hybrid models can take advantages of these two networks. For CNN and RNN, classifier and data representation are mutually dependent. For Transformer, according to the existent papers, Zheng et al. employed Transformer to process the features of range-angle maps extracted by CNN [55], ViT is

deployed to extract feature from spectrograms and CVD directly [56], and radar Transformer are leveraged to process the radar point clouds from automotive targets [57].

All of these three papers showed superiority of Transformer and its variants compared to other DL models and in [5], in a shape classification task, point Transformer performed better than all the other DL models on the ModelNet40 dataset.

Therefore, it is promising to apply point Transformer to solve radar based HAR problems, although to the best of our knowledge this has not been done in the related literature.

# 3

## METHODOLOGY

*This chapter describes the different aspects of the methodology related to this research, including DL models and data pre-processing algorithms. In section 3.1, some classical and typical DL models such as CNN and RNN are briefly introduced, and then there is an elaboration on the architectures of Transformer and its variants. Section 3.2 focuses on the most relevant algorithms, such as CFAR, Fast Fourier Transform (FFT), DOA estimation that can process the radar raw data into point clouds.*

### 3.1. DEEP LEARNING MODELS

#### 3.1.1. CONVOLUTIONAL NEURAL NETWORK

Convolutional neural network (CNN) was originally proposed for image processing [65]. To be specific, CNN is designed to process the data in multiple arrays format. For instance, a typical image consists of three two-dimensional arrays containing the pixel luminance in the Red Green Blue (RGB) channels. Many realistic data formats conform to the multiple arrays model: 1D for signal; 2D for spectrogram and images; 3D for video or volumetric images. There are four key ideas behind CNN exploiting the properties of natural signals: local connections, shared weights, pooling, and the use of multiple layers [66].

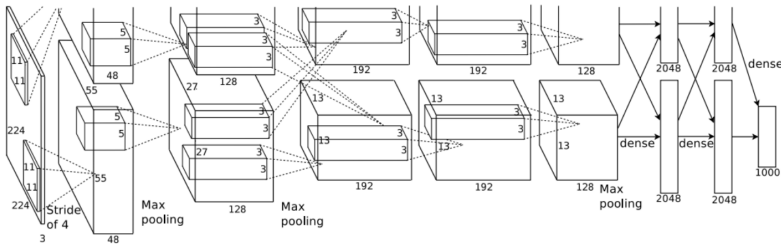


Figure 3.1: Example of CNN architecture from the literature [1]

A typical architecture of CNN is illustrated in Figure 3.1. CNN adopts a series of layers as a basic block including convolutional layers, pooling layers and activation function. Many blocks can be stacked to make the network deeper. The role of the convolutional layer is to extract the features and detect local conjunctions from the previous layer via sliding a filter over the feature map. The filter is also called a convolution kernel containing trainable weights, followed by a pooling layer to merge semantically similar features into one to decrease overfit and expand the perception field. A typical pooling unit computes the maximum of a local patch of units in one feature map. The activation function is often used to introduce more non-linearity to the network. Therefore, the features of an input image can be well represented as a vector after passing through the above blocks. Lastly, fully connected layers are connected to the last block and via a softmax function the output vector contains the probability for each class.

For radar-based HAR, CNNs are deployed when the inputs are image-like radar data representation such as spectrograms, HRRP arranged into a matrix, range-time plots, and so on. Here, the radar data are processed and classified like normal images.

### 3.1.2. RECURRENT NEURAL NETWORK

RNN is also a classical artificial neural network, where connections between nodes form a directed or undirected graph along a temporal sequence, so this network is able to exhibit temporal features. As the topology of RNN shown in figure 3.2, this type of network can utilize the internal state (hidden layer) to store memory from previous inputs. To be more specific, each input is corresponding to an output, and each output is determined by current input and previous hidden states jointly.

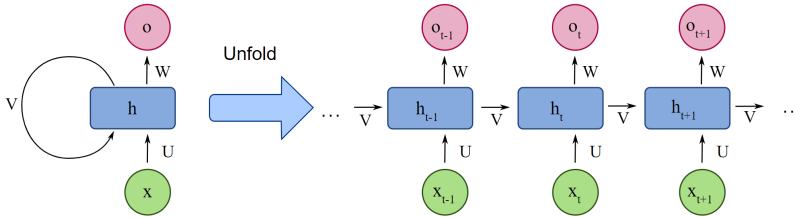


Figure 3.2: Model of RNN architecture

Where,  $o$  is the output,  $h$  is hidden layer,  $x$  is the input and  $U$ ,  $V$ , and  $W$  are three matrices (linear layers in network), respectively, so that the output of RNN is decided by previous hidden layers and current input. Compared with CNN which can only process certain input size, RNN can process arbitrary length of sequences of input, at least theoretically. However, one of the main problems in the original RNN is that it cannot remember the past very well because gradients will exponentially shrink down as we back propagate, when the gradient is calculated at each step. To solve this problem ('vanishing gradient'), two variants of RNN with some modifications to the network architecture are proposed: LSTM and GRU.

- LSTM was first proposed by Hochreiter et al. in 1997, and he aimed to improve the storage of information over extended time intervals [40]. Its network architecture

can be seen in figure 3.3. In LSTM three gates are introduced to overcome the gradient vanishing or exploding. These gates are input, forget and output gate, respectively. The input gate is responsible for adding new memory, and the forget gate decides which memory should be forgotten. The output gate modulates the amount of memory content exposure. LSTM unit is able to decide whether to keep the existing memory via the introduced gates. Intuitively, if the LSTM unit detects an important feature from an input sequence at an early stage, it easily carries this information (the existence of the feature) over a long distance, hence, capturing potential long-distance dependencies.

- GRU was designed by Cho et al. in 2014 to have the capability to capture dependencies of different time scales adaptively [67]. Figure 3.3 shows the graphical illustration of GRU. Compared with LSTM, GRU can make sure that features will not be lost during long-term transmission by introducing only 2 gates (reset gate and update gate), which makes the network architecture relatively simple, thus having less computational requirement. According to the experiment results of [68], GRU can save a lot of time due to its simpler architecture while not sacrificing performance in long text scenarios.

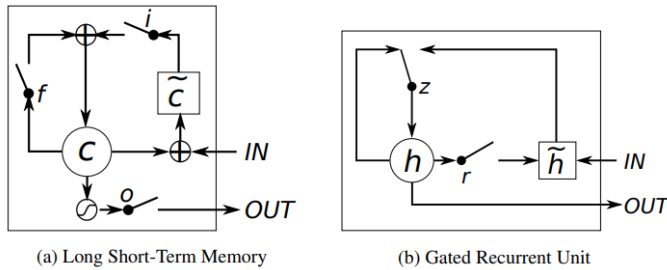


Figure 3.3: Illustration of (a) LSTM and (b) GRU. (a)  $i$ ,  $f$  and  $o$  are the input, forget and output gates, respectively.  $c$  and  $\tilde{c}$  denote the memory cell and the new memory cell content. (b)  $r$  and  $z$  are the reset and update gates, and  $h$  and  $\tilde{h}$  are the activation and the candidate activation [2]

### 3.1.3. TRANSFORMER AND ITS VARIANTS

Before the Transformer was proposed, the dominant sequences translation models were dependent on complex recurrent or convolutional neural network. Transformer is a novel and simple DL model using solely attention mechanism as its core to implement sequences translation. It has dominated Natural Language Processing (NLP) since it was proposed in 2017: it achieved 28.4 BLEUs (bilingual evaluation understudy) on the WMT 2014 English-to-German translation [53], and it can be regarded as initiating the fourth general DL model after MLP, CNN, RNN.

### SELF-ATTENTION AND TRANSFORMER

Self-attention mechanism is the core of Transformer. As the name of the paper that proposes Transformer says, "Attention is all you need". In previous research, attention mechanism are always combined with CNN and RNN to enhance the performance of models, while Transformer use solely self-attention mechanism without any convolutional layers to solve sequence to sequence problems.

Self-attention, as known as intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence.

3

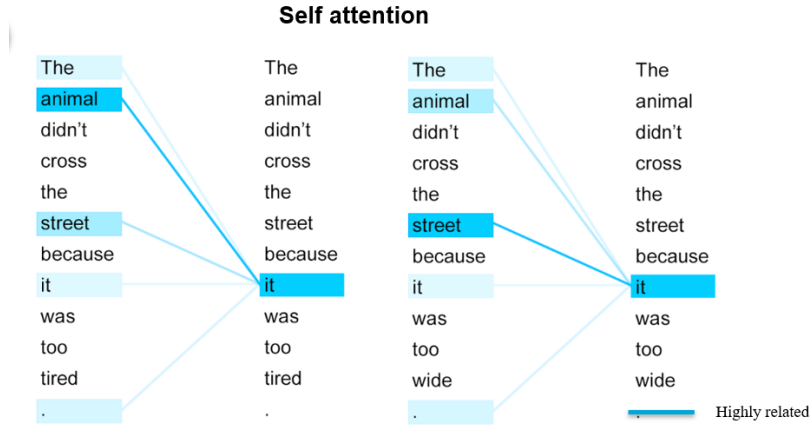


Figure 3.4: An example of self-attention analyzing a sequence of text [3]

Figure 3.4 shows an example to describe what the self-attention focus is in order to find the pairwise relations between a sequence based on its context. In this example, RNN will process the sentence word by word, and it just considers the relation around the word 'it' so, RNN is likely to associate the left 'it' with street since they are near. However, when 'self-attention' processes these two sentences, the model can associate the left 'it' with animal and the right 'it' with street by calculating the attention scores based on the pairwise relations. When processing long text, self-attention can capture the semantic relation even if across long intervals. Moreover, self-attention can be computed in parallel, thus training more effectively.

How exactly the self-attention calculates the relations between its key inputs and outputs is described in equation 3.1 and figure 3.5.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.1)$$



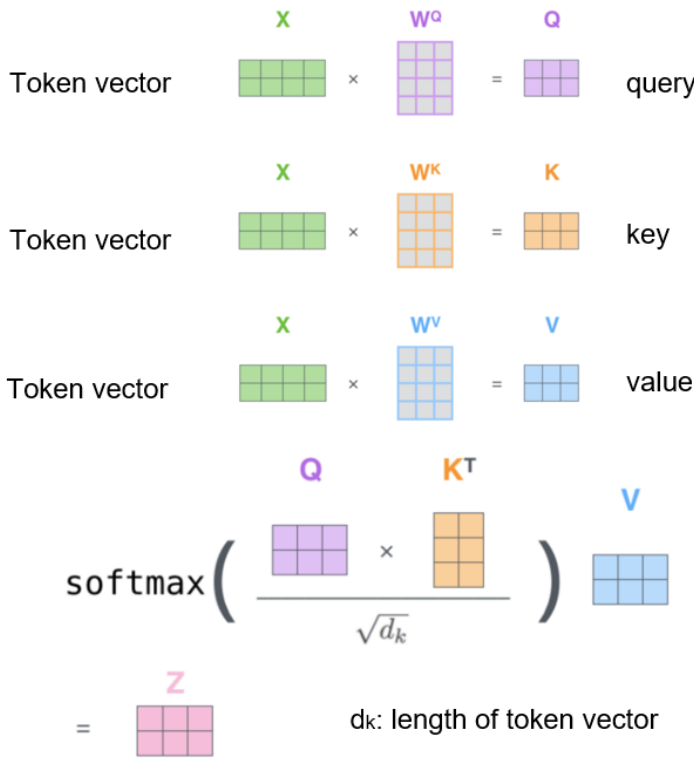


Figure 3.5: Graphical illustration of scaled dot-product attention calculation [4]

The mapping of a query and a collection of key-value pairs to an output, where the query, keys, values, and output are all vectors, is known as the *attention function*. The result is calculated as a weighted sum of the values, with the weights assigned to each value determined by how well the query matches the key in question.

To be specific, in Figure 3.5,  $x$  is a token vector and  $W^Q$ ,  $W^K$ ,  $W^V$  are three weight matrices, fully connected layers in a network. After multiplying the token vector with three weight matrices, we get the query key and value matrices, respectively, and for self-attention, the three matrices  $Q$ (Query),  $K$ (Key),  $V$ (Value) all come from the same input. Now, we need to calculate the dot product between  $Q$  and  $K$ , and then in order to prevent the result from being too large, it is divided by a scale  $\sqrt{d_k}$ , where  $d_k$  is the dimension of a query and key vector. Then we use the Softmax operation to normalize the result to a probability distribution, and then multiply by the matrix  $V$  to get the representation of the weight sum.

Actually, in [53], the authors who first proposed the Transformer found it beneficial to repeat the single self-attention function  $h$  times in parallel, and concatenate and once again project these outputs, resulting in the final values, as depicted in Figure 3.6.

With the introductions of self attention and multi-head, it is easier to get an understanding of the whole architecture of Transformer, as depicted in Figure 3.7.

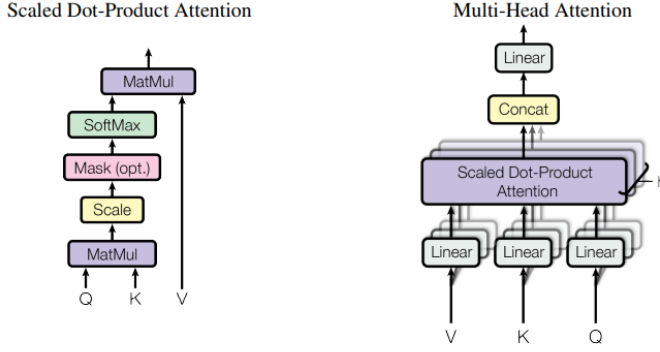


Figure 3.6: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention, which consists of several attention layers running in parallel.

### A Embedding

Inputs of the Transformer are usually words, but they should be represented as vectors so that the network can process them, so these tokens are converted to vectors of dimension  $d_{model}$  in the input embedding layer.

### B Positional Encoding

Since the Transformer has no recurrence and no convolution, in order for the model to utilize the sequence's order, some information about the relative or absolute position of the tokens in the sequence should be injected. The positional encodings and embeddings both have the same dimension  $d_{model}$ , allowing the two to be added together. There are many choices of positional encodings, in Transformer the authors use sine and cosine functions of different frequencies [53].

$$\begin{aligned} PE_{(pos, 2i)} &= \sin\left(pos/10000^{2i/d_{model}}\right) \\ PE_{(pos, 2i+1)} &= \cos\left(pos/10000^{2i/d_{model}}\right) \end{aligned} \quad (3.2)$$

Where pos is the position and i is the dimension. In other words, a sinusoid corresponds to each dimension of the positional encoding. From  $2\pi$  to  $20000\pi$ , the wavelengths follow a geometric progression. We selected this function because we believed it would make it simple for the model to pick up on relative positioning.

### C Encoder

The encoder is shown in the left part of Figure 3.7, which consists of a stack of  $N = 6$  identical layers. For each layer, there are two sub-layers. The first sub-layer is a multi-head self-attention mechanism, and the second sub-layer is just a positionwise fully connected feed-forward network. The authors employ layer normalization after using a residual connection around each of the two sub-layers. The model's embedding layers and all sub-layers generate outputs of dimension  $d_{model} = 512$  to enhance these residual connections.

### D Decoder

Similar to the encoder, the decoder is also composed of a stack of  $N = 6$  identical layers, but the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. Besides, there is also a residual connection around each of the sub-layers followed by layer normalization. In order to stop positions from paying attention to succeeding positions, the authors additionally modify the self-attention sub-layer in the decoder stack.

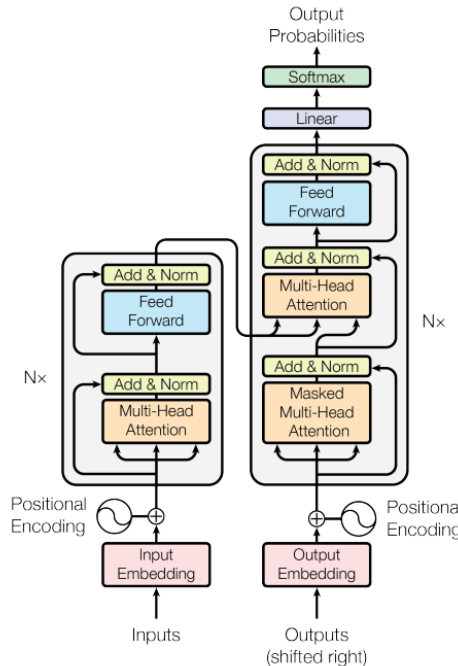


Figure 3.7: Example of the Transformer - model architecture.

### POINT TRANSFORMER - HENGSHUANG ZHAO'S MODEL

Since radar point clouds are simply sets that are sporadically embedded in a metric space and self-attention can find the relation among inputs in parallel, self-attention fits naturally with point clouds. Here the first self-attention model used in this thesis is described.

#### A Point Transformer Layer

In Hengshuang's point Transformer model, the foundation of the point transformer layer is vector self-attention [5]. The authors employ the subtraction relation and add a position encoding  $\delta$  to both the attention vector  $\gamma$  and the transformed features  $\alpha$ , which is shown in equation 3.3

$$\mathbf{y}_i = \sum_{\mathbf{x}_j \in \mathcal{X}(i)} \rho(\gamma(\varphi(\mathbf{x}_i) - \psi(\mathbf{x}_j) + \delta)) \odot (\alpha(\mathbf{x}_j) + \delta) \quad (3.3)$$

Where,  $\mathcal{X}(i) \subseteq \mathcal{X}$  and it is a local neighborhood set containing the  $k$  nearest points of point  $\mathcal{X}(i)$ .  $\mathbf{y}_i$  is the output feature,  $\varphi$ ,  $\psi$ , and  $\alpha$  are pointwise feature transformations, like linear projections or MLPs.  $\delta$  is a position encoding function and  $\rho$  is a normalization function such as softmax.  $\odot$  stands for the Hadamard product, that is, element-wise product.  $\gamma$  is a mapping function with two linear layers and one ReLU nonlinearity. These tokens have the same meanings in Hengshuang's model.

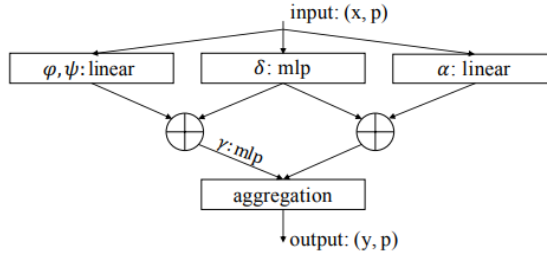


Figure 3.8: Details of the point transformer layer

The input of point transformer layer is  $(x, p)$  where  $p$  stands for the coordinate (position) of the input point and  $x$  represents the features of the input point such as the normalized vector of the coordinate or the Doppler information of the point plus the coordinate. The output of point transformer layer contains  $(y, p)$  where  $p$  remains unchanged, and  $y$  represents the new features after self-attention mechanism.

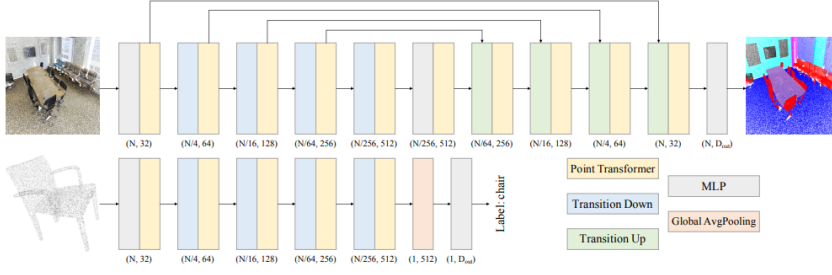


Figure 3.9: Example of point transformer networks for semantic segmentation (top) and classification (bottom) [5]

### B Position Encoding

Position encoding is crucial for self-attention because it enables the operator to adjust to local data structure. The 3D point coordinates themselves are a suitable option for position encoding in 3D point cloud processing. In Hengshuang's point transformer model, the position encoding function is described as follows:

$$\delta = \theta(\mathbf{p}_i - \mathbf{p}_j) \quad (3.4)$$

Where,  $\mathbf{p}_i$  and  $\mathbf{p}_j$  are the 3 dimensions coordinates of points  $i$  and  $j$ . An MLP with two linear layers and one ReLU nonlinearity makes up the encoding function  $\theta$ . Worth to mention, it is beneficial to apply position encoding to both the attention generation branch and the feature transformation branch. So it can be seen in Equation 3.3 that the trainable position encoding  $\theta$  is added in both branches.

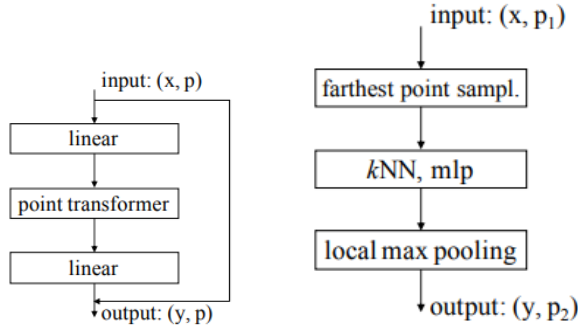


Figure 3.10: Point transformer block (left) and transition down block (right)

### C Point Transformer Block

The structure of the point transformer block is illustrated in Figure 3.10. As shown, there is a residual connection in the block. Besides, the self-attention layer and

linear projections, which can reduce dimensionality and speed up processing, are integrated into the transformer block. The input is a collection of feature vectors  $x$  and their corresponding 3D coordinates. The point transformer block permits information interchange between these localized feature vectors, creating new feature vectors for all data points as its output.

#### D Transition Down

Reducing the cardinality of the point set when necessary is the main function of the transition down module, for instance reducing the number of points from  $N$  to  $N/4$ . Denote  $P_1$  and  $P_2$  as the input and output point set of the transition down block. First, we apply the farthest point sampling [69] in  $P_1$  to acquire a well-spread subset  $P_2 \subset P_1$  with the requisite cardinality. Then, we employ  $kNN$  on  $P_1$  to pool feature vectors from  $P_1$  onto  $P_2$  (empirically  $k = 16$  can have the best results). Each input features are forwarded to a linear transformation, followed by batch normalization, ReLU and a max pooling onto each point in  $P_2$  from its  $k$  neighbors in  $P_1$ .

#### E Network Architecture

The complete 3D point cloud process network is depicted in Figure 3.9. For classification, the pipeline is based on several repeated blocks, such as point transformer block, transition down block, MLP and global average pooling. The network's main feature aggregation operator is the point transformer, and convolutions are not used for preprocessing: the architecture of the network is entirely built on point transformer layers, pointwise transformations, and pooling.

$N$  points inputs first pass through an MLP and a point transformer block to be resized to  $(N, 32)$ , where  $N$  is the number of input points. They are forwarded to a couple of repeated transition down and point transformer blocks with down sample rate  $[4 \ 4 \ 4 \ 4]$ , and thus the cardinality of the point set for each stage is  $[N/4, N/16, N/64, N/256]$ . Notably, the number of stage and the downsampling rate can be varied based on different tasks. At the end of the architecture, a global average pooling layer and an MLP are connected to output the final classification result.

#### POINT CLOUD TRANSFORMER - MENGHAO GUO'S MODEL

The complete architecture of the point cloud transformer (PCT) is shown in Figure 3.11. This is the second self-attention based model considered in this thesis. The purpose of PCT is to convert the input points into a new, higher-dimensional feature space that may describe the semantic affinities between points as a foundation for other point cloud processing tasks.

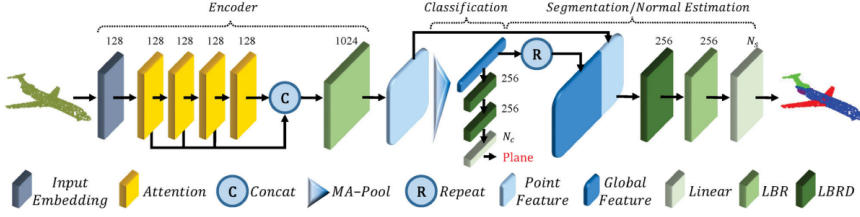


Figure 3.11: Example of the PCT architecture: the encoder mainly comprises an Input Embedding module and four stacked Attention modules. The decoder mainly comprises multiple Linear layers. The numbers above each module indicate its output channels. MA-Pool concatenates Max-Pool and Average-Pool. LBR combines Linear, Batch Norm, and ReLU layers. LBRD means LBR followed by a Dropout layer [6]

### A Encoder

The input coordinates are first embedded into a new feature space by the PCT encoder. The embedded features are later forwarded into 4 stacked attention modules to learn a semantically rich and discriminative representation for each point, then a linear layer to produce the output feature. Overall, the PCT encoder has a nearly identical construction as the original Transformer, except that the position embedding is removed, since the point's coordinates already provide positional information.

Formally, with an input point cloud  $P \in \mathbb{R}^{N \times d}$ , where  $N$  is the number of points and each point has  $d$ -dimensional features, a  $d_e$ -dimensional embedded feature  $F_e \in \mathbb{R}^{N \times d_e}$  is first transformed in the input embedding module. The attention output of each attention layer is then concatenated through the feature dimension, and followed by a linear transformation to generate the PCT output: point-wise  $d_o$ -dimensional feature  $F_o \in \mathbb{R}^{N \times d_o}$ . This process can be briefly expressed as Equation 3.5

$$\begin{aligned} F_1 &= AT^1(F_e) \\ F_i &= AT^i(F_{i-1}), \quad i = 2, 3, 4 \\ F_o &= \text{concat}(F_1, F_2, F_3, F_4) \cdot W_o \end{aligned} \quad (3.5)$$

where  $F_e$  denotes the feature vectors after input embedding,  $W_o$  denotes the linear layer's weights and  $AT^i$  denotes the  $i$ th attention layer, each of which has the same output dimension as its input.

At the end, the outputs from two pooling operators: a max-pooling and an average-pooling on the learned pointwise feature representation are concatenated so that an effective global feature vector  $F_g$  can be extracted to represent the point cloud.

### B Classification

For classification task, we can focus on the right part of the Figure 3.11. After the 4 stacked attention layers and a linear transformation, we can obtain the global

feature vector  $F_g$ . To classify the input point cloud, the  $F_g$  is fed into a classification decoder which is composed of 2 cascaded LBRDs (Linear, Batch Norm, ReLU, Dropout layer), each with a dropout rate of 0.5, followed by a linear layer to predict the final classification probability of each class. The predicted class of the input point cloud is determined as the class with maximal probability.

### POINT TRANSFORMER - NICO ENGEL'S MODEL

The overview of Nico's point transformer architecture is shown in Figure 3.12. This is the third self-attention based model considered in this thesis. The objective of this model is to relate local and global input properties in order to investigate the shape information of the point set.

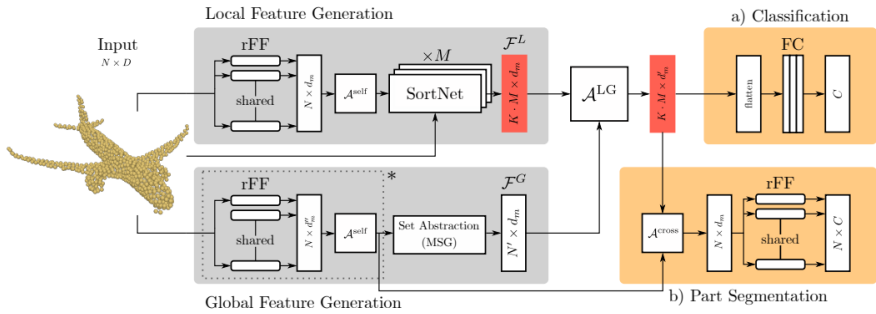


Figure 3.12: Overview of the Point Transformer architecture, which consists of two branches to generate local and global features. SortNet produces an ordered set of local features against the global structure of the input point cloud. Depending on the task, classification or part segmentation heads are employed. Red Boxes denote sorted sets, only for the segmentation part [7]

This pipeline can be divided into three parts:

- 1) **SortNet** that extracts ordered local feature sets from different subspaces.
- 2) **Global feature generation** of the whole point set.
- 3) **Local-Global attention**, which relates local and global features.

The input point set is considered as  $P = \{p_i \in \mathbb{R}^D, i = 1, \dots, N\}$ , where  $D$  is the dimension of each point.  $D = 3$  when only coordinates are given, and it is possible to let  $D = 5$  when points are added additional features such as Doppler and time step. Overall, Nico's point transformer model is composed of two independent branches: a local feature generation module, and a global feature.

For the local feature generation branch, the input point set is first projected to latent space with dimension  $d_m$  via a row wise feed-forward network. To link the points to one



another, the authors then use self-multi-head attention on the latent features. Eventually, the SortNet outputs a sorted set of fixed length. For the global feature generation branch, set abstraction and multi-scale grouping are deployed here to extract global features. After collecting local and global features, there is a local-global attention to combine and aggregate these features from input point cloud. Finally, the outputs of local-global attention are flatten and forwarded to fully connected layers to predict the class.

### A Local Feature Generation

In this module, SortNet is the crucial part which is illustrated in Figure 3.13. First and foremost, SortNet receives the original point cloud and the projected latent features from row-wise feed forward networks, followed by a multi self-attention layer. This additional self multi-head attention layer is to extract the spatial and higher-order relation between each input point  $p_i \in P$ .

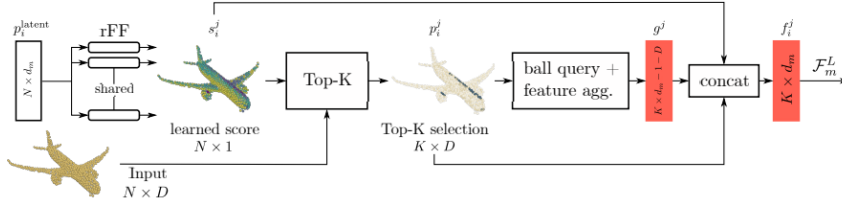


Figure 3.13: Overview of the SortNet: a score is learned from a latent feature representation to extract important points from the input. Local features are aggregated from neighboring points. SortNet outputs a permutation invariant and sorted feature set. Red boxes denote sorted sets [7]

Subsequently, a row-wise feedforward network is employed to reduce the feature dimension to one, thus generating a learnable scalar score  $s_i \in \mathbb{R}$  for each input point, which integrates spatial relationships as a result of the layer of multi-heads attention. Pair  $\langle p_i, s_i \rangle_{i=1}^N$  assigns the score corresponding to each point. Let  $\mathcal{Q}$  a completely ordered set and based on  $\langle p_i, s_i \rangle_{i=1}^N$ , we select  $K \leq N$  points from the original point cloud with the highest score :

$$\mathcal{Q} = \{q_j, j = 1, \dots, K\}$$

where,  $q_j = \langle p_i^j, s_i^j \rangle_{j=1}^K$ ,  $p_i^j \in P$  and  $s_i^1 \geq \dots \geq s_i^K$ . That is, the author deploy the top-K operation to look for the  $K$  points associated with the highest scores  $s_i$ . After choosing  $k$  points based on the learnable scores, a ball query search is applied to group the points in the original points within the Euclidean distance  $r$  to the chosen  $k$  points. Grouped points are denoted by  $g^j \in \mathbb{R}^{d_m-1-D}$ ,  $j = 1, \dots, K$ . The local features  $g^j$  and the scores  $s_i$  are concatenated with the corresponding input

points  $p_i$ . Thus, the local feature vector is acquired:

$$f_i^j = p_i^j \oplus s_i^j \oplus g^j, \quad f_i^j \in \mathbb{R}^{d_m}. \quad (3.6)$$

where,  $\oplus$  operation denotes matrix concatenation. Therefore, the SortNet output makes up one local feature set:

$$\mathcal{F}_m^L = \{f_i^j, j = 1, \dots, K\} \quad (3.7)$$

$\mathcal{F}_m^L$  is also an ordered set since  $\mathcal{Q}$  is ordered. To collect dependencies and regional features from several subspaces, the  $M$  feature sets are concatenated to create an ordered local feature set of fixed size:

$$\mathcal{F}^L = \mathcal{F}_1^L \cup \dots \cup \mathcal{F}_M^L, \quad \mathcal{F}^L \in \mathbb{R}^{K \cdot M \times d_m} \quad (3.8)$$

where  $\cup$  denotes the union of two sets.

### B Global Feature Generation

The purpose of this branch is to extract the global features from the original point cloud. The authors employ the set abstraction multiscale grouping layer (MSG) [69] so that computational time and memory can be reduced with decreased input points.  $N$  input points are down sampled to  $N'$  through Farthest Point Sampling (FPS) and acquire neighboring points to aggregate features of dimension  $d_m$  to obtain a global feature  $N' \times d_m$  with dimension of  $d_m$ . In particular, the global features are not ordered since no sorting operation is performed.

### C Local-Global Attention

This part is responsible for relating global and local feature sets,  $\mathcal{F}^L$  and  $\mathcal{F}^G$ , respectively, to extract shape and context information from the input point cloud. Therefore, the authors employ self-attention to both local feature  $\mathcal{F}^L$  and global characteristic  $\mathcal{F}^G$ , followed by multihead cross attention:

$$A^{LG} := A^{\text{cross}} \left( A^{\text{self}}(F^L), A^{\text{self}}(F^G) \right) \quad (3.9)$$

$$A^{\text{cross}}(P, Q) := A^{\text{MH}}(P, Q) \quad (3.10)$$

$$A^{\text{MH}}(X, Y) = \text{LayerNorm}(S + \text{rFF}(S)) \quad (3.11)$$

$$S = \text{LayerNorm}(X + \text{Multihead}(X, Y, Y)) \quad (3.12)$$

$$\text{Multihead}(Q, K, V) = (\text{head}_1 \oplus \dots \oplus \text{head}_h) W^O \quad (3.13)$$

$$\text{head}_i = A \left( QW_i^Q, KW_i^K, VW_i^V \right) \quad (3.14)$$

where,  $(F^L$  and  $F^G$  are matrix representation of  $\mathcal{F}^L$  and  $\mathcal{F}_1^G$  respectively and  $A(Q, K, V) = \text{score}(Q, K)V$  which is described in Equation 3.1. The last row-wise feed forward network in cross multi-head attention changes the feature dimension to  $d'_m < d_m$  so that the computational complexity can be reduced.

Finally, after the description of each model used in this thesis, the number of trainable parameters, the total number of operations, and the network size for each of them are listed in Table 3.1.

Table 3.1: Summary of the numbers of parameters in the three models considered in this work.

	Henghuang[5]	Menghao[6]	Nico[7]
Total parameters	1,431,813	2,932,805	21,994,673
Trainable parameters	1,431,813	2,932,805	21,994,673
Total multi-add (M)	4.17	550.09	48.43
Parameters size (MB)	5.46	11.19	83.90

### 3.2. DATA PRE-PROCESSING

Since the representation of radar data in this thesis is radar point cloud, this section will focus on the data pre-processing algorithms that transform radar complex signals into point clouds with Doppler, intensity and time step features. Also, we assume that the radar type is a MIMO millimeter wave, so that multiple channels provide angle information, and that the millimeter wave is short enough to detect human subjects as extended targets with multiple, distributed reflections. The overview of the radar data pre-processing pipeline from radar cube to point cloud is displayed in Figure 3.14, with some of the key steps detailed in the following sub-sections.

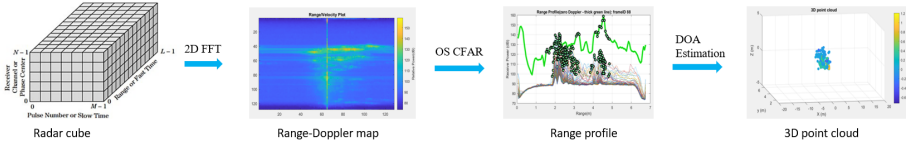


Figure 3.14: Overview of the radar data preprocessing from radar cube to point cloud

#### 3.2.1. 2D-FFT

As shown in Figure 3.14, the received radar cube has three axes: fast time ( $M$  bins), slow time ( $L$  bins), and channels ( $N$  bins), respectively; so, for each channel, the radar data can be regarded as a 2D discrete signal. Therefore, in order to estimate accurate Doppler and range information of the target, 2D FFT has been applied on this 2D discrete matrix with an assumption that the beat signal is stationary in each original chirp. The expression of 2D FFT (range-Doppler) is listed in Equation 3.15.

$$\begin{aligned}
 S_{Rx}(k, l) &= \sum_{m=0}^{M-1} S_{Rx}^{(m)}(k) e^{-j2\pi l m / M} \\
 &= \sum_{m=0}^{M-1} \sum_{n=0}^{L-1} s_{Rx}^{(m)}(n) e^{-j2\pi k n / N} e^{-j2\pi l m / M}
 \end{aligned} \tag{3.15}$$

where,  $l = 0, \dots, M-1$ , and  $k = 0, \dots, N-1$ ,  $M$  and  $L$  is the number of bins for fast time and slow time, respectively.

### 3.2.2. CFAR

After applying the 2D FFT for each channel, the Doppler range matrix is obtained, where the fast-time dimension is converted to range and the slow-time dimension is converted to Doppler. To further detect which bins are occupied by the extended target, CFAR is applied.

CFAR is a popular objection detection algorithm to adaptively suppress homogeneous background clutter so that undesired targets can be filtered. The principle of CFAR is that for a given Doppler-range cell, namely Cell Under Test (CUT), a threshold can be calculated using the training cells and based on the false alarm rate, and this threshold is compared with CUT to determine whether there is an object in CUT. That is, we estimate the noise level from the training cells and determine whether CUT is occupied by an object or not. There are two common CFAR algorithms: Cell-Average (CA)-CFAR and Ordered-Statistic (OS)-CFAR. The reason we choose OS-CFAR is that it shows superior performance over Cell-Averaging (CA) CFAR with non-uniform clutter [70]. The principle of OS-CFAR is illustrated in Figure 3.15.

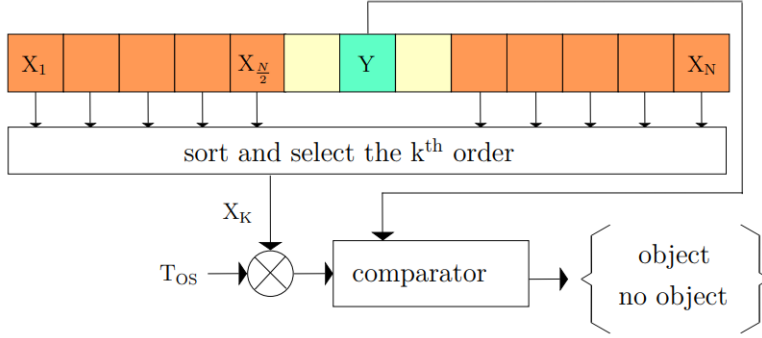


Figure 3.15: Principle of OS-CFAR algorithm: orange blocks represent the training cells, shallow yellow blocks stand for the guard band cells and green block is the CUT [8]

In contrast to CA-CFAR, which utilizes all the training cells to calculate the threshold, OS-CFAR choose a single amplitude. The general idea is that noise estimation is based on the  $k^{th}$  values of the training cells sorted in ascending order, as in Equation 3.16. That is, the arithmetic mean used in CA-CFAR is replaced by a single rank of the ordered statistic  $\leq X_k$ .

$$X_1 \leq X_2 \leq \dots \leq X_k \leq \dots \leq X_{N-1} \leq X_N \quad (3.16)$$

Therefore, if there is another object present in the training cells, its value will not affect the peak detection in the CUT. For the threshold, the noise level also need to be multiplied by a scaling factor  $T_{OS}$ .

$$P_{FA} = k \binom{N}{k} \frac{(k-1)!(T_{OS} + N - k)!}{(T_{OS} + N)!} \quad (3.17)$$

According to [71], the suitable value of  $k$  is  $k = 3/4N$ . The scaling factor  $T_{OS}$  can be derived by solving Equation 3.17.

### 3.2.3. DOA ESTIMATION

After 2D FFT and CFAR detection, the extended targets' occupied cells are determined. That is, the range, Doppler and intensity features of a point are already obtained. Based on the arrangement of the antenna array on the MIMO radar, we can estimate the elevation angle and azimuth angle by applying a beamforming algorithm, such as FFT and Multiple Signal Classification (MUSIC) to the multiple channels. With the range, azimuth, and elevation angles, we can convert the spherical coordinate to the Cartesian coordinate. In this thesis, the simple FFT-based beamforming was used, leaving any super-resolution technique such as MUSIC for future work.

## 3.3. PROPOSED HAR PIPELINE

The used MIMO mm-wave Frequency-Modulated Continuous-Wave (FMCW) radar can generate six intrinsic features of the target: range, azimuth angle, elevation angle, Doppler, SNR, temporal relation by continuously transmitting and receiving modulated millimeter wave. For the first three features, they can be represented in Cartesian coordinates and form the spatial aspect of the point clouds. For the last three features, the most common data representations are range-Doppler heatmaps and spectrograms, as thoroughly analyzed in Chapter 2. In the majority of the previous works, these are separately used to recognize human activities. The pipeline proposed in this thesis aims to try to integrate Doppler, SNR, temporal relation information features on point clouds to address HAR via this more information-rich representation. This is done in conjunction with the self-attention models described in the previous sections, that will be used as classifiers. The overview of the proposed pipeline is given in Figure 3.16.

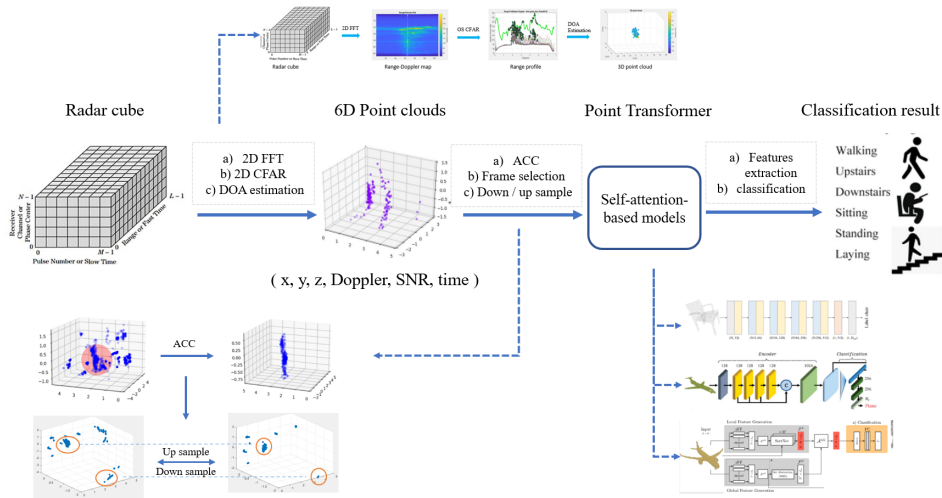


Figure 3.16: Overview of the proposed pipeline to address HAR problem

Specifically, this pipeline is made up of three main modules, which are indicated by three dashed lines in Figure 3.16, and the descriptions of the three modules are listed as follows:

- The *first module* is responsible for converting the radar raw data cubes containing complex signals to 6D point clouds including 3D coordinates, Doppler (velocity), intensity (SNR), time. In this module, 2D FFT is first applied to 2D discrete signals to generate the range-Doppler map. Then, we employ 2D OS-CFAR on the map to detect which bins are occupied by subjects, and the coordinates of the detected bins are the range and Doppler information of the point, while the values of the detected bins correspond to SNR. Last but not least, an FFT is applied along the channel axis to estimate the azimuth angle and elevation angle. With the angle and range information, we can thus derive the 3D Cartesian coordinates.
- Actually, a lot of points in the point clouds are clutter, and the number of points after the first module does not match the input of the network. For these two reasons, we need the second block of data pre-processing. In the *second module*, a method of removing the clutter is proposed: 1) calculate the centroid of the point cloud with SNR as weights, 2) filter out the points with the distance to the centroid higher than 1 *m*, assumed to be a reasonable number for an average human body size. After removing clutter, a certain number of frames with the highest Doppler are selected because points in these frames can better represent the features of the motion. Lastly, we apply down sampling or up sampling on the point clouds to match the input size of the network. The specific resampling algorithms are described in detail in Chapter 4.
- The *third module* is the DL classifier. In this thesis, we investigate three different attention-based networks for processing point clouds. They are the point transformer of Hengshuang Zhao et al. [5], the point cloud transformer of Menghao Guo et al. [6], and the point transformer of Nico Engel et al. [7]. For the first two models, the authors employ a hierarchical architecture to extract the features of the input point cloud with an attention mechanism, and use fully connected layers to present the classification results. For Nico's model, local and global features are related by cross multi-head attention after being extracted separately, and similar to the other two models, fully connected layers are deployed to provide the classification results.

# 4

## DATASET PREPARATION AND RELEVANT PREPROCESSING

*This thesis utilizes two datasets to prove its proposed pipeline. One is the MMActivity dataset [10] and the other one is TUD dataset also used in [13]. In this chapter, the radar types and measurements of two datasets are depicted in detail. Section 4.1 and section 4.2 introduce the MMActivity dataset and the TUD dataset, respectively.*

### 4.1. MMACTIVITY DATASET

The dataset includes five continuous motions of only two human subjects, including boxing, jumping jack, jumping, squats, and walking. Given its small size, this dataset was initially intended to help with just a feasibility study before moving to a larger one.

#### 4.1.1. RADAR INFORMATION

The radar used to collect this dataset is TI's IWR1443BOOST [72]. The waveform of this radar is FMCW using a chirp signal. The antenna layout of the radar can be seen in Figure 4.1. As shown, the array arrangement of antennas enables the estimation of both elevation and azimuth angle. It is a MIMO radar with three transmitters and four receivers so that the radar can detect the angle information of targets. This radar and the radar used in the TUD data set are both FMCW radar produced by the TI, so the working principle is similar even if the number of channels is different. The operating frequency band ranges from 76 to 81 GHz, according to Equation 4.1, since the maximum bandwidth that can be used is 4 GHz, the best range resolution is 4 cm.

$$\Delta R = \frac{c}{2B} \quad (4.1)$$

where,  $B$  is the bandwidth in Hz swept by the chirp of the radar,  $c$  is the speed of light, and  $\Delta R$  is the range resolution.

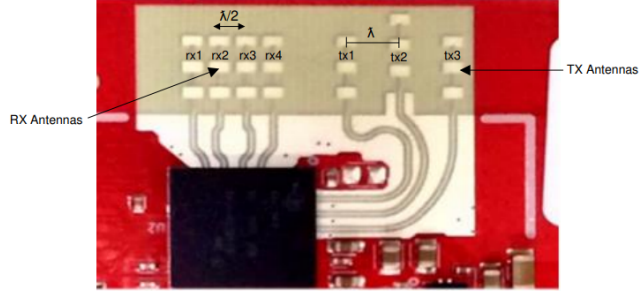


Figure 4.1: Picture of IWR1443 radar board [9]

4

For angle resolution, we can refer to the Equation 4.2.

$$\theta_{BW} \approx \frac{\lambda}{Md \cos \theta} \quad (4.2)$$

Where,  $M$  is the number of channels,  $\theta$  is azimuth compared to boresight,  $d$  is the separation between antenna elements and  $\lambda$  is wavelength.

According to the antenna pattern in [9] that at 78 GHz, based on the 3-dB drop in the gain, the horizontal 3dB-beamwidth is approximately  $\pm 28$  degrees and elevation 3dB-beamwidth is approximately  $\pm 14$  degrees. Additionally, the specific usage of IWR1443 can also be found in [9].

#### 4.1.2. MEASUREMENT

For the measurement of human motion data, the radar is mounted on a tripod stand at a height of 1.3 m and only two human subjects participated in the data collection. They performed five different activities including walking, jumping, jumping jacks, squatting, and boxing. The experimental scene can be seen in Figure 4.2, and human subjects kept performing one activity for a period of 20 seconds in front of the radar to collect the data. Totally, they collected 93 minutes of radar point cloud data and the distribution of the data can be seen in Table 4.1, with a sampling rate of 30 frames per second.

Table 4.1: Classes distribution of the MMActivity dataset

Activity	Number of data files	Total duration (seconds)
Boxing	39	1115
Jumping	38	1062
Jumping jack	37	1045
Squats	39	1090
Walking	47	1269





Figure 4.2: MMACTIVITY dataset collection setup [10].

The data collected are then transferred to a laptop using ROS over USB. ROS can interact with the TI radar development board where raw data preprocessing is implemented in a Digital Signal Processor (DSP), and thus point clouds can be acquired directly. The storage format of point cloud data in ROS is rosbag file [10], and the data was converted into .txt files. Unlike the storage format of TUD dataset in Matlab, where each frame is a single matrix, the data storage format in MMACTIVITY is .txt files and each file contains all points detected in a period of time. Figure 4.3 shows the data format to save the information of a detected point. These .txt files store the 3D coordinates, Doppler, intensity, and time information of each point. These data are directly received from ROS and data preprocessing such as Direction Of Arrival (DOA) estimation and CFAR detection are automatically performed in the radar development board. If radar point clouds are collected using ROS, we cannot access complex signal data, since the board has only a few megabytes of memory, not enough to store big complex signals.

```
header:
  seq: 6264
  stamp:
    secs: 1538888235
    nsecs: 712113897
  frame_id: "ti_mmwave" # Frame ID used for multi-sensor scenarios
point_id: 17 # Point ID of the detecting frame (Every frame starts with 0)
x: 8.650390625 # Point x coordinates in m (front from antenna)
y: 6.92578125 # Point y coordinates in m (left/right from antenna, right positive)
z: 0.0 # Point z coordinates in m (up/down from antenna, up positive)
range: 11.067276001 # Radar measured range in m
velocity: 0.0 # Radar measured range rate in m/s
doppler_bin: 8 # Doppler bin location of the point (total bins = num of chirps)
bearing: 38.6818885803 # Radar measured angle in degrees (right positive)
intensity: 13.6172780991 # Radar measured intensity in dB
```

Figure 4.3: The storage format of one point in the MMACTIVITY dataset.[11]

### 4.1.3. DATASET FOR THIS THESIS

Obviously, data in the format shown in Figure 4.3 cannot be fed into networks directly, so a Python script is utilized to convert this data format into a matrix as in Figure 4.7. It shows the .txt file containing the features such as Doppler, intensity and point\_id as features of the point.

In order to explore the variation of recognition results as the number of points decreases, the continuous data are segmented into multiple samples with a certain number of points. The samples have no overlap part with each other. The number of points and its corresponding duration and frames are showed in Table 4.2.

Table 4.2: Relation among the number of points, frames and duration in adjusted MMActivity dataset

number of points	Frames	Duration
1360	60	2 s
1024	40	1.5 s
512	20	0.8 s
256	10	0.4 s
128	5	0.2 s

The standard number of input points is 1024, so the reason to choose 1360 points as the input size is that in [10] the author of MMActivity dataset use the data of 60 frames to train networks and the number of points is approximately 1360 in that case. Therefore, it is fair to compare the pipeline of this thesis and pipeline in [10] with the same number of input points.

## 4.2. TUD DATASET

The TUD dataset [13] includes radar data for 4 motions such as sitting down to a chair, standing up from a chair, bending and standing up after bending, and 2 static postures such as standing still and sitting still on a chair. There are 8 male human subjects participating in the measurement, in the age between 20 and 30 years.

### 4.2.1. RADAR INFORMATION

The radar board used to collect data is composed of four cascaded AWR2443 chips, each of which has 4 transmitters and 3 receivers. Therefore, there are 16 Transmitter (TX)s and 12 Receiver (RX)s embedded in the board. While the radar board is working, all RX receive the echo signals simultaneously, but the TXs are working with Time Division Multiple Access (TDMA). Overlapping channels are discarded, resulting in 86 equivalent azimuth channels and 7 equivalent elevation channels. According to the TI manual [73], the field of view in azimuth is approximately -60 to +60 degrees and -20 to +20 degrees in elevation, for attenuation of less than 10dB in the radiation pattern. A picture of the actual radar board is shown in Figure 4.4.

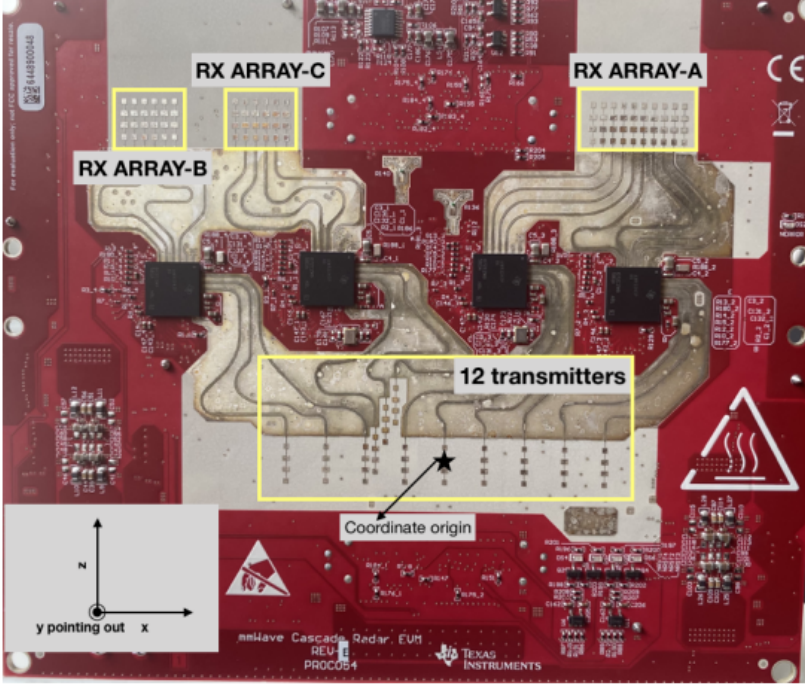


Figure 4.4: Picture of the four-device cascaded AWR2243 radar board

The derived radar parameters, such as the maximum detection range and velocity, the range and velocity resolution can be seen in Equation 4.3, respectively. Range resolution is the ability of a radar to discriminate between two objects that are extremely close to each other. The velocity resolution means the smallest different of velocity a radar can measure.

$$\begin{aligned}
 R_{\max} &= \frac{cf_{ADC}}{2r_{\text{chirp}}}, \\
 v_{\max} &= \pm \frac{\lambda_{\text{chirp}}}{4N_{TX}T_{\text{chirp},\text{total}}}, \\
 \Delta R &= \frac{c}{2B} = \frac{c}{2T_{\text{chirp}}r_{\text{chirp}}}, \\
 \Delta v &= \frac{\lambda_{\text{chirp}}}{2T_{CPI}}
 \end{aligned} \tag{4.3}$$

Where,  $c$  is the light speed,  $f_{ADC}$  is the sample frequency,  $r_{\text{chirp}}$  is the slope of the chirp,  $\lambda_{\text{chirp}}$  is the wavelength,  $T_{\text{chirp}} = N_{ADC}/f_{ADC}$  is the ADC sampled interval.  $2T_{CPI}$  is the coherent processing interval. For angle resolution, it is not dependent on the bandwidth but relies on the proportion between the chirp center wavelength and the MIMO aperture size (i.e. the number of channels), since it is estimated from the phase shift of multi-

channel. The specific expressions of azimuth and elevation angle resolution are in Equation 4.4.

$$\begin{aligned}\Delta\phi &= \frac{\lambda_{\text{chirp}}}{L_{\phi}}, \\ \Delta\theta &= \frac{\lambda_{\text{chirp}}}{L_{\theta}},\end{aligned}\tag{4.4}$$

Here,  $L_{\phi}$  and  $L_{\theta}$  represent the azimuth and elevation apertures, respectively.  $\Delta\phi$  and  $\Delta\theta$  are azimuth resolution and elevation resolution, respectively. According to [13], with the above factors considered, the radar parameters for HAR are empirically configured as shown in Table 4.3.

4

Table 4.3: Waveform parameters and derived features of the radar. The definition of parameter and feature is subject to whether it is directly configurable, the directly configurable term is referred to as parameter, the other as derived feature [13]

Parameter	Symbol	Value
Antenna design wavelength	$\lambda_{\text{antenna}}$	3.90mm
Number of TXs	$N_{TX}$	12
Number of RXs	$N_{RX}$	16
Total number of virtual channels	$N_{\text{channel}}$	192
MIMO aperture on azimuth	$L_{\phi}$	$42.5\lambda_{\text{antenna}}$
MIMO aperture on elevation	$L_{\theta}$	$3\lambda_{\text{antenna}}$
ADC Sampling Rate	$f_{ADC}$	2.7MHz
Chirp Ramp Interval	$T_{\text{chirp}}$	60μs
Total Chirp Interval	$T_{\text{chirp},\text{total}}$	63μs
Number of chirps per sub-frame	$N_{\text{chirp}}$	1536
Start Frequency	$f_{\text{start}}$	77GHz
Chirp Ramp Slop	$r_{\text{chirp}}$	60MHz/μs
Sub-frame Periodicity	$T_{\text{sub-frame}}$	100ms
Field of view on azimuth	$FOV_{\phi}$	[−40deg, 40deg]
Field of view on elevation	$FOV_{\theta}$	[−20deg, 20deg]
Coherent processing interval	$T_{CPI}$	0.1s
Derived Features	Symbol	Value
Equivalent number of channels on x-axis	$N_{\phi}$	86
Equivalent number of channels on z-axis	$N_{\theta}$	7
Transmitted Chirp Bandwidth	$B_{Tx}$	3.6GHz
Received Chirp Bandwidth	$B_{Rx}$	2.84GHz
Valid chirp center wavelength	$\lambda_{\text{chirp}}$	3.82mm
Maximum Measurement Range	$R_{\text{max}}$	6.75m
Maximum Unambiguous Velocity	$v_{\text{max}}$	±1.26m/s
Range Resolution	$\Delta R$	5.28cm
Velocity Resolution	$\Delta v$	0.0286m/s
Azimuth Angle Resolution (broadside)	$\Delta\phi$	1.4deg
Elevation Angle Resolution (broadside)	$\Delta\theta$	18deg

### 4.2.2. MEASUREMENT

The experiments were carried out at the Lage Hallen of the EWI building in TU Delft. The specific setups for the measurement are as follows:

- The radar is placed at a height of 0.75 m so that all body parts of human subjects can be covered within the field of view of the radar.
- Figure 4.5 shows the layout of the measurement room, describing the positions of the radar and other items in the room. As shown, the distance between the radar and the chair is 2.7 m.

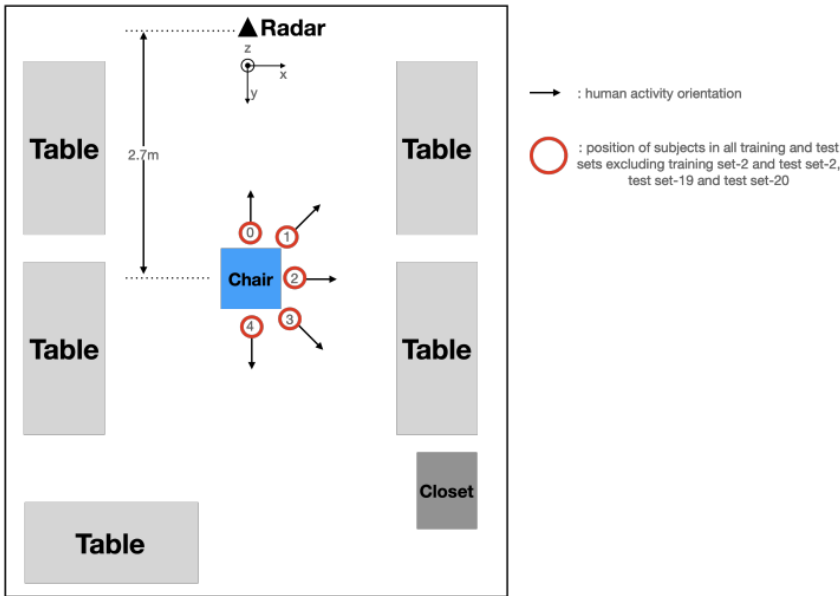


Figure 4.5: Data measurement setup in an office-like room at TU Delft [12, 13]

In this measurement, motions are recorded in pairs. For example, standing up from a chair and sitting down are paired motions, and the subjects were asked to sit down and stand up successively for 2 minutes with a regular period of 2 seconds for each motion. For static postures such as standing and sitting, the human subjects just stood or sat in front of the radar for 2 minutes. Therefore, it is easy to label the radar data following the above procedure. The data distribution of six different activities are displayed in the pie chart of 4.6 showing a balanced distribution.

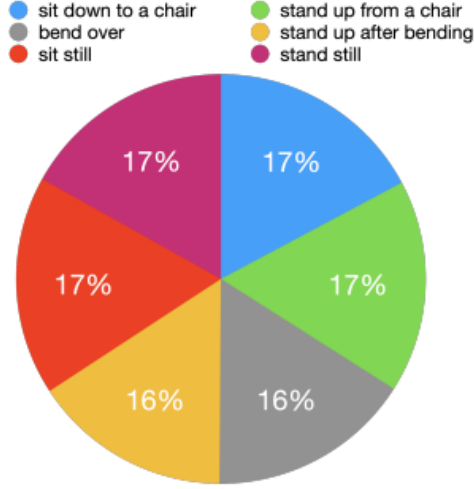


Figure 4.6: Classes and data distribution of the TUD dataset [12]

In this thesis, only data of 7 human subjects with zero aspect angle are used out of the whole dataset. Each activity was recorded for 2 minutes and each motion lasts for 2 seconds, so there are  $420 = 120/2 \times 7$  samples for each activity and  $2520 = 420 \times 6$  samples totally.

For cross validation test, we apply 5-fold validation, 80 percentage for training and 20 percentage for testing.

#### 4.2.3. BASIC DATASET FOR THIS THESIS

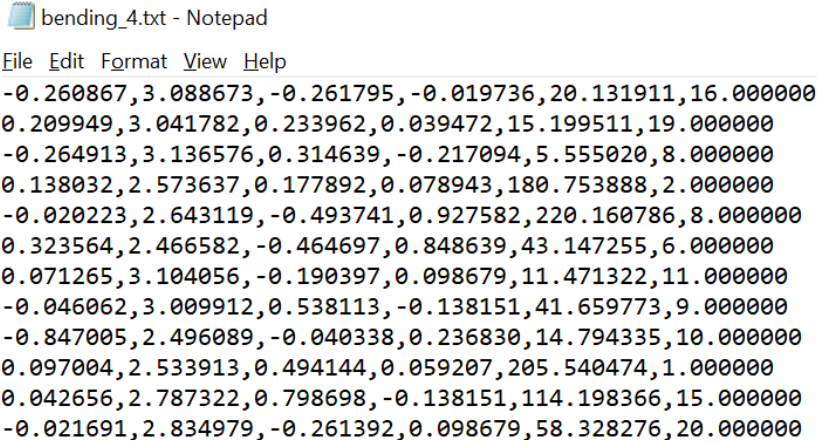
The data can be directly acquired from the TI radar board, stored in bin files. These bin files can be parsed into complex signals in Matlab. Then, after applying 2D-FFT, range-Doppler OS-CFAR and DOA estimation, the point clouds with Doppler and intensity information are thus obtained [74].

The data structure to store the processed data in Matlab is cell with size of  $1 \times 1197$ , where this length means that there are 1197 matrices in this cell and each matrix store the coordinates and corresponding features such as Doppler, SNR and angle information in a frame. The coordinates already contain the range and angle information, so in a single frame, the features selected for each point are 3D coordinates, Doppler, and SNR. Besides, for a complete activity, the sequential indices of 20 frames are added as the sixth feature to a point. Time information is added by considering that the time step describes the temporal order of points, which is relevant to capture kinematic information. It may for example help the network to recognize the pair motions.

For a specific example, if we look at the snapshot of standing up and sitting down, the spatial distributions of points from 20 frames are almost identical, but if we take the temporal information into consideration, it is interesting to find some points appearing earlier in certain areas in the point cloud of sitting down and some points appear later

in these areas in the point cloud of standing up. Theoretically, this can be used by the model to distinguish pair motions without using spectrograms.

With prior knowledge that each motion occupies 20 frames (2 seconds) and human subjects successively repeated paired motions, we can label the data of each 20 frames with the corresponding activities. To match the Python environment where we train our networks, while labeling these processed data are transferred into 6D point clouds (3D coordinates, Doppler, SNR, time step) and saved as a .txt file with 6 decimal places, as shown in Figure 4.7. Finally, the input data can be regarded as a 2D matrix.



```

bending_4.txt - Notepad
File Edit Format View Help
-0.260867,3.088673,-0.261795,-0.019736,20.131911,16.000000
0.209949,3.041782,0.233962,0.039472,15.199511,19.000000
-0.264913,3.136576,0.314639,-0.217094,5.555020,8.000000
0.138032,2.573637,0.177892,0.078943,180.753888,2.000000
-0.020223,2.643119,-0.493741,0.927582,220.160786,8.000000
0.323564,2.466582,-0.464697,0.848639,43.147255,6.000000
0.071265,3.104056,-0.190397,0.098679,11.471322,11.000000
-0.046062,3.009912,0.538113,-0.138151,41.659773,9.000000
-0.847005,2.496089,-0.040338,0.236830,14.794335,10.000000
0.097004,2.533913,0.494144,0.059207,205.540474,1.000000
0.042656,2.787322,0.798698,-0.138151,114.198366,15.000000
-0.021691,2.834979,-0.261392,0.098679,58.328276,20.000000

```

Figure 4.7: Example of data from a 6D point cloud file: from left to right are the coordinates of x, y, z, Doppler, SNR, and frame index, separated by commas

#### 4.2.4. DATASET WITH ADAPTIVE CLUTTER CANCELLATION

As shown in Figure 4.5, around the human subject there are a couple of tables, a closet and walls which can cause a lot of clutter reflections. As shown in Figure 4.8, the point cloud looks disorganized, and it is hard to find the exact position of a human intuitively since the human subject is surrounded by clutter. In order to remove these clutters adaptively, an algorithm called Adaptive Clutter Cancellation (ACC) is proposed in this thesis. The general idea is to find the centroid of a human subject and keep only the points with a certain distance from the centroid.

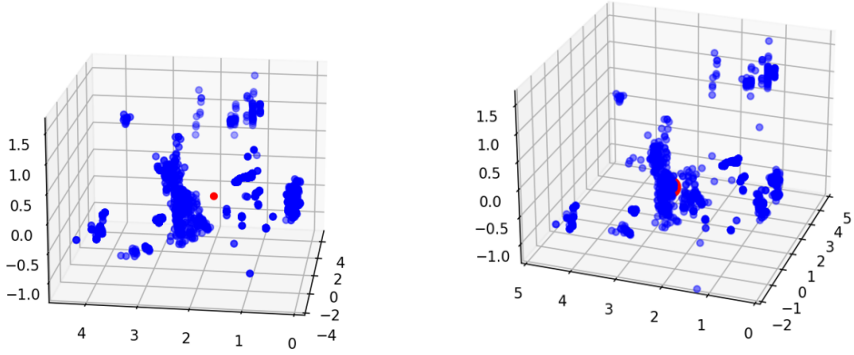
The first step is to determine the centroid. If the coordinates of the centroid is simply chosen as the mean value of coordinates of all points, there will be an offset, as shown in Figure 4.8a. The reason is that clutter contributions are not symmetric and there are more clutter elements near to the radar. After observing that the point clouds data like in Figure 4.7, it is noticeable that points in the position of human subject have significantly higher SNR values than clutter, so SNR values can be used as weights to calculate the

precise position of centroid as per Equation 4.5.

$$\begin{aligned}
 weight &= snr / mean(snr) \\
 x_c &= \langle x, weight \rangle / n \\
 y_c &= \langle y, weight \rangle / n \\
 z_c &= \langle z, weight \rangle / n
 \end{aligned} \tag{4.5}$$

where,  $snr$   $x$ ,  $y$ , and  $z$  are 1D vectors containing SNR coordinates of  $x$ ,  $y$ ,  $z$  of all points, respectively.  $n$  and  $\langle \rangle$  represents the number of points and inner product, respectively.

4



(a) calculating the centroid of the point cloud without weights.

(b) calculating the centroid of the point cloud using SNR values as weights.

Figure 4.8: Visualization of point cloud of a standing human subject. The red dot represents the centroid of the point cloud.

After using SNR as weights to revise the centroid coordinate against offset, we can calculate the precise centroid. As Figure 4.8b shows, the red dot is almost embedded in the body of the human subject.

Subsequently, with the assumption that the average height and body size of all human subjects are less than 2 m, we can consider that all points within a sphere are generated by human activities. The center of the sphere is the calculated centroid, and the radius is 1 m. Finally, we just need to retain the points within the sphere and points outside the sphere can be regarded as clutter and discarded. Figure 4.9 shows the comparison of before and after using the proposed ACC and it is obvious that the remaining points look like the shape of a human body. Intuitively, this algorithm is capable to remove most of the clutter contributions.



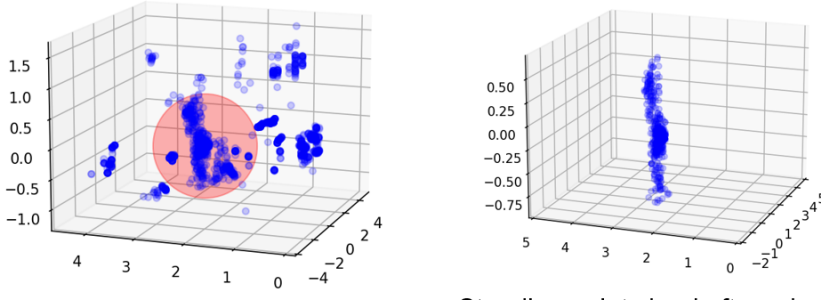


Figure 4.9: Comparison of point cloud without/with proposed adaptive clutter cancellation. The red sphere shows the threshold to determine whether a point is considered clutter or not.

#### 4.2.5. DATASETS OF SELECTED FRAMES

As shown in Table 4.3,  $T_{sub-frame}$  is 100 ms, and each motion is recorded for 2 seconds, so there are 20 frames in each sample. However, in our common sense, the period of an actual movement such as sitting down and standing up is far less than 2 seconds. In the motion data, some frames record movement and the other frames record posture. Therefore, much information contained in the data of 2 seconds is redundant to recognize a single motion, and indeed not necessary to utilize 20 frames to classify the activity. In order to explore how many numbers of frames are necessary, 3 new datasets with fewer frames are created based on the dataset with applied adaptive clutter cancellation. As introduced in section 4.2.3, each sample in the basic dataset contains the radar data of 20 frames. To filter the most valuable frames among the 20 available, Doppler information is used. Figure 4.10 displays the visualization of point clouds from 20 frames, and the color of the points can indicate the corresponding velocity. As shown, not all frames contain the information on how a human subject moves, so there is great potential to eliminate redundant data. This means that we can utilize only part of 20 frames to recognize activities so that fewer points and less computational power are needed, and the classification can happen at a faster rate. Since movements can generate more points in contrast to static postures, it is relatively fair to choose the average Doppler value as a suitable indicator to filter frames.

More specific details on this process of selecting frames and modifying where needed the number of input points are given in the following.

- **Strategy for selecting frames**

Considering that the actual motion is continuous, the selected frames should also be continuous, so we deploy a sliding window on the 20 frames with a step of 1 frame to calculate the average Doppler. The window with the highest Doppler value is selected as new data to represent the sample. The length of the sliding window can be varying, and in this thesis the length is set as 3, 5 and 10, and 3 new

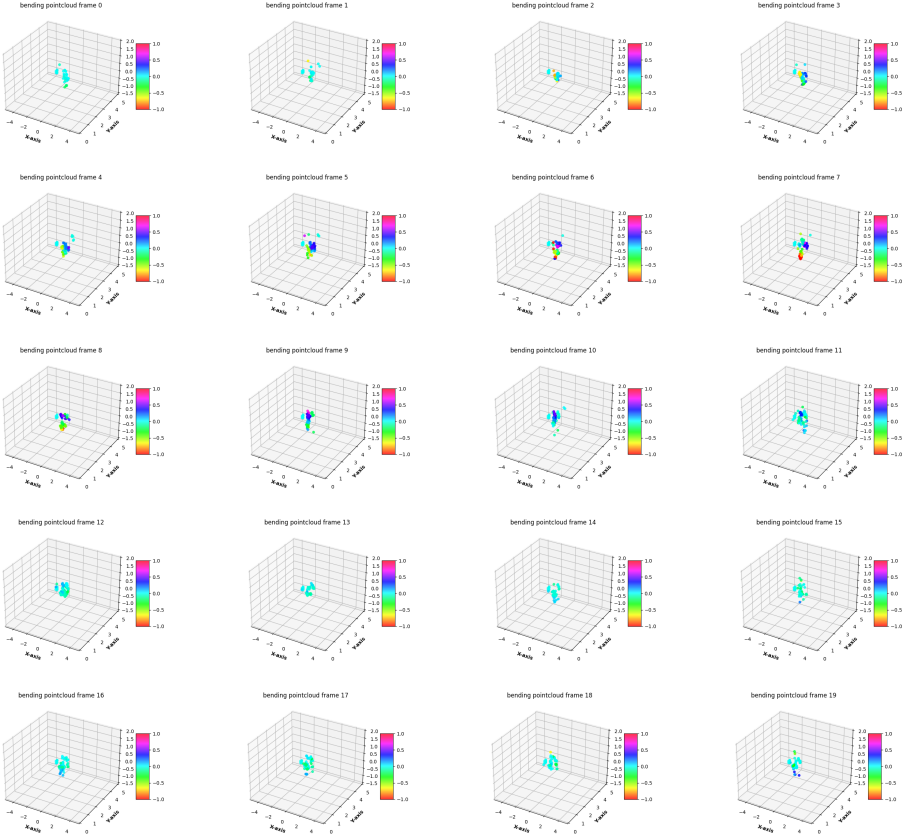


Figure 4.10: Visualization of point cloud of the 'bending' activity for 20 frames. Each plot shows the point cloud of a single frame, and the color bar indicates the Doppler value. From left top to right bottom corner, the point clouds from frame 0 to frame 19 are shown

datasets of different frames contained in one sample are created. As introduced in Chapter 3, the network architecture is made up of a few hierarchical blocks, where the input size is decreasing proportionally with a ratio of 4, so the number of points contained in each sample of new datasets is shown in Table 4.4.

- **Padding strategy**

Since the detection algorithm applied on range-Doppler map is OS-CFAR, we cannot make sure that the number of points per frame is a constant. After applying adaptive clutter cancellation, the number of remained points per frame fluctuates around 100 with a standard deviation of 10, depending on how many movements are captured by radar. If the human subjects perform more movements, there will be more points generated. In order to ensure that the total number of points is

Table 4.4: Mapping relation between number of frames and size of resulting input data

Number of frames	Size of input per sample
20	(1024,6)
10	(512, 6)
5	(256, 6)
3	(256, 6)

constant, we apply up sample and down sample if the number of points is fewer or more than the preset constant, respectively.

For the *up sampling*, let us consider that the preset number of points is  $N_0$  and the actual number of points after adaptive noise cancellation is  $N_1$  ( $N_0 > N_1$ ), so the number of points we need to fill is  $N_2 = N_0 - N_1$ . The most common method for padding is to fill the data with  $N_2$  zero vectors, but this operation may add too homogenized values with little information to all activities, and thus limit the performance of the networks. In order to keep the original information in the data, the idea is to repeat some representative points and make the features of the selected points unchanged. Therefore, we use the Farthest Point Sampling (FPS) algorithm on the original points to sample  $N_2$  points and then concatenate the new points with the original data to generate a matrix with a size of  $(N_0, 6)$  and the features of the added points remained unchanged. For these repeated points, they will be removed in the down sample block of networks, so this up sample operation can avoid adding homogenized information artificially, while filling the points to the needed preset number. If  $N_1$  is far less than  $N_0$  such as  $N_0 > 2N_1$ , the solution is to double the data first and then apply FPS to sample the remaining number of points. If the  $N_1$  is even smaller, we can assume that this sample does not contain enough information to be correctly recognized and discard it.

For the *down sampling*, consider  $N_0$  and  $N_1$  ( $N_0 < N_1$ ) as the preset number and the actual number of points in the input point cloud, respectively. The simplest method is to use a random sample, that is, to randomly select  $N_0$  points as input data. This method may work well when the point cloud is dense, but for a small number of point set as we often have with radar-based sensing, such as 256 points, random selection is likely to change the spatial distribution of points and make the features of points a lot different. To avoid this, we also apply FPS on the original points to sample  $N_0$  points. This operation can ensure that the spatial distribution is unchanged in great extent, but the price is that it will increase the computational complexity.

# 5

## RESULTS

*This chapter analyzes the performance of the proposed method through different datasets and input features. Section 5.1 shows the feasibility results of using point transformer (Hengshuang’s version) on MMActivity dataset. In section 4.2, there are thorough investigations of different input features derived from point cloud, three attention-based models, and their leave-one-subject-out test.*

### 5.1. FEASIBILITY RESULTS OF USING POINT TRANSFORMER ON MMA DATASET

In this thesis, training and testing of the networks are done in an Alienware laptop with an NVIDIA GeForce RTX 3070 Laptop GPU, and the GPU memory is 8 GB.

Since the MMActivity dataset can directly provide the point cloud data and the activities are continuous motions, it is possible to label the data and convert the data into the 6D point cloud format as described in Chapter 4. Thus, this dataset is utilized to study first the feasibility of using point transformer to address HAR problems.

#### 5.1.1. COMPARISON BETWEEN PROPOSED PIPELINE AND RADHAR PIPELINE

This subsection compares the classification results of the proposed pipeline and the pipeline in RadHAR [10], where the authors voxelized points from 60 frames and embedded these points into a cube. The input data format to classifiers is a 3D matrix. The classification results from the literature when using different classifiers is listed in Table 5.1.

Table 5.1: RadHAR results from [10]: test accuracy of different activity recognition classifiers trained on the MMActivity Dataset.

S.No	Classifier	Accuracy
1	SVM	63.74
2	MLP	80.34
3	Bi-directional LSTM	88.42
4	Time-distributed CNN+ Bi-directional LSTM	90.47

In order to have a fair comparison with the results above, the number of input points for the proposed pipeline of this thesis is selected as 1360, which is approximately the number of points in 60 frames. The normalized confusion matrix of the proposed pipeline is shown in Figure 5.1.

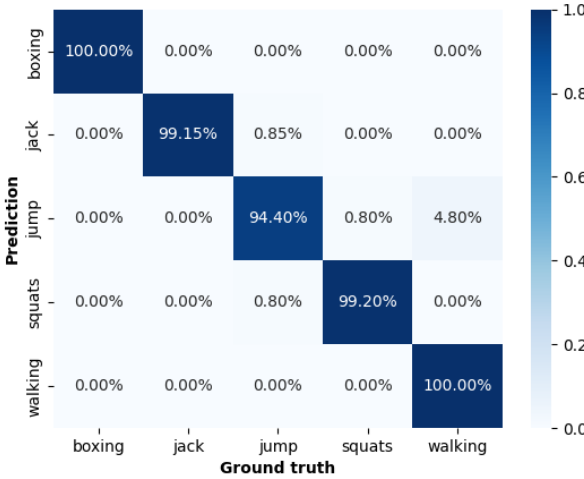


Figure 5.1: Normalized confusion matrix for Point Transformer to classify the five motions in the MMActivity dataset, with 1360 as the number of input points

The numbers in the diagonal show the accuracy for each activity. All test data of walking and boxing are classified correctly. Only a few test data of the other three classes are given a wrong prediction. The overall accuracy is 98.43%, which is 8% more than the best results with the RadHAR pipeline proposed in [10]. This great enhancement shows the potential of point transformer architectures to process radar point cloud data in contrast with the traditional DL models such as CNN and RNN.

### 5.1.2. RESULTS WITH DIFFERENT INPUT

For the purpose of exploring how the input will affect the classification result, extra tests are performed with different input features of points and varying number of in-

put points. First, we feed only the 3D spatial coordinates of the points to the network and use this result as a baseline. Then we associate the points with different features as the inputs. Worth to mention is that we make input data with all features as 0 to study the influences brought by input shape. Figure 5.2 shows the accuracy and F1 score of different input features.

In the case of MMActivity dataset, input data with Doppler and intensity features can benefit the classification most. For single extra feature, the improvement brought by Doppler and intensity is slightly greater than temporal information. The possible reason for this is that the activities are continuous motion, so the order of how limbs move can vary from sample to sample. Thus, the temporal distribution of points are not important. For the group with 0 as extra features, the accuracy and F1 score decline slightly instead, so we can consider that from the point transformer perspective the homogenized information may reduce the variance of different activities in radar point cloud, thus declining the overall classification accuracy.

## 5

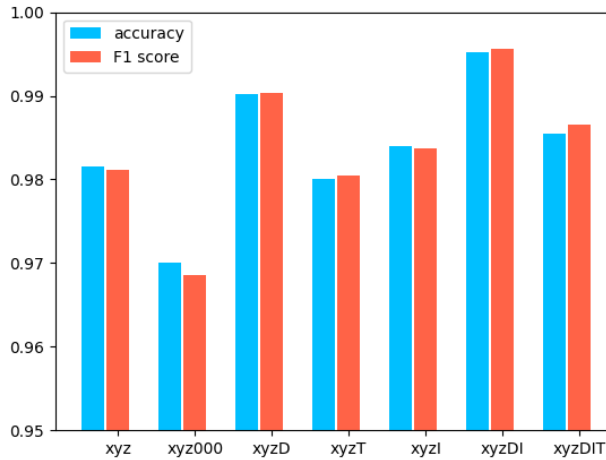


Figure 5.2: Classification accuracy and F1 score values with different input features from the point clouds. x, y, and z indicate the spatial coordinates, and D, I, and T indicate Doppler, intensity, and time, respectively

Furthermore, the number of points needed for the networks to recognize the activities is also important, since in the real scenario many human motions happen in a short period and the number of generated points is related to how many frames a radar records. If the number of needed points is much more than the actual motions can generate, it can be hard to utilize radar in a timely manner to recognize these activities.

Since the point transformer has a hierarchical architecture to extract features from

point cloud as described in Chapter 3, the input number should be chosen as power of 2 in the proposed pipeline. Figure 5.3 displays the accuracy and F1 score of feeding different number of points to the point transformer. As shown, with decreasing number of input points, accuracy and F1 score are declining at the same time. However, even only 128 points are fed into the network, the classification result is still better than the best result in the RadHAR pipeline with non transformer-based networks and approaches. This means that 128 points and their added information are enough for the proposed pipeline to distinguish the different motion in MMActivity dataset.

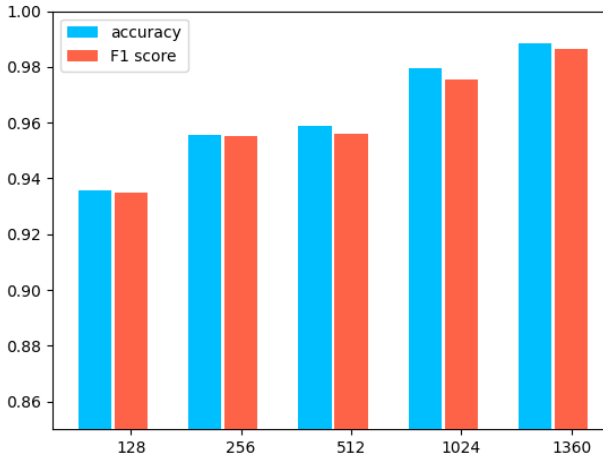


Figure 5.3: Classification accuracy and F1 scores with different number of input points in the point cloud, from 128 to 1024

## 5.2. RESULTS OF USING PT AND PCT ON TUD'S DATASET

In this section, we thoroughly investigate the characteristics of the three attention-based models chosen in this thesis with various input data. We also have investigated person-based recognition based on their activities.

### 5.2.1. RESULTS OF DIFFERENT FEATURES AS INPUTS

There are many combinations of 3D coordinates and extra features. To find the best combination and explore how these different features contribute to the classification results, we first train point transformer (Hengshuang's model) to see the results with different combinations of features. The reason why we select the point transformer by Hengshuang is that the numbers of parameters and operations in the model are the fewest, as shown in Figure 3.1, and thus we can experiment effectively by running many tests with limited computational burden.

The experiment results with different input features are displayed in the confusion matrices of Figure 5.4.

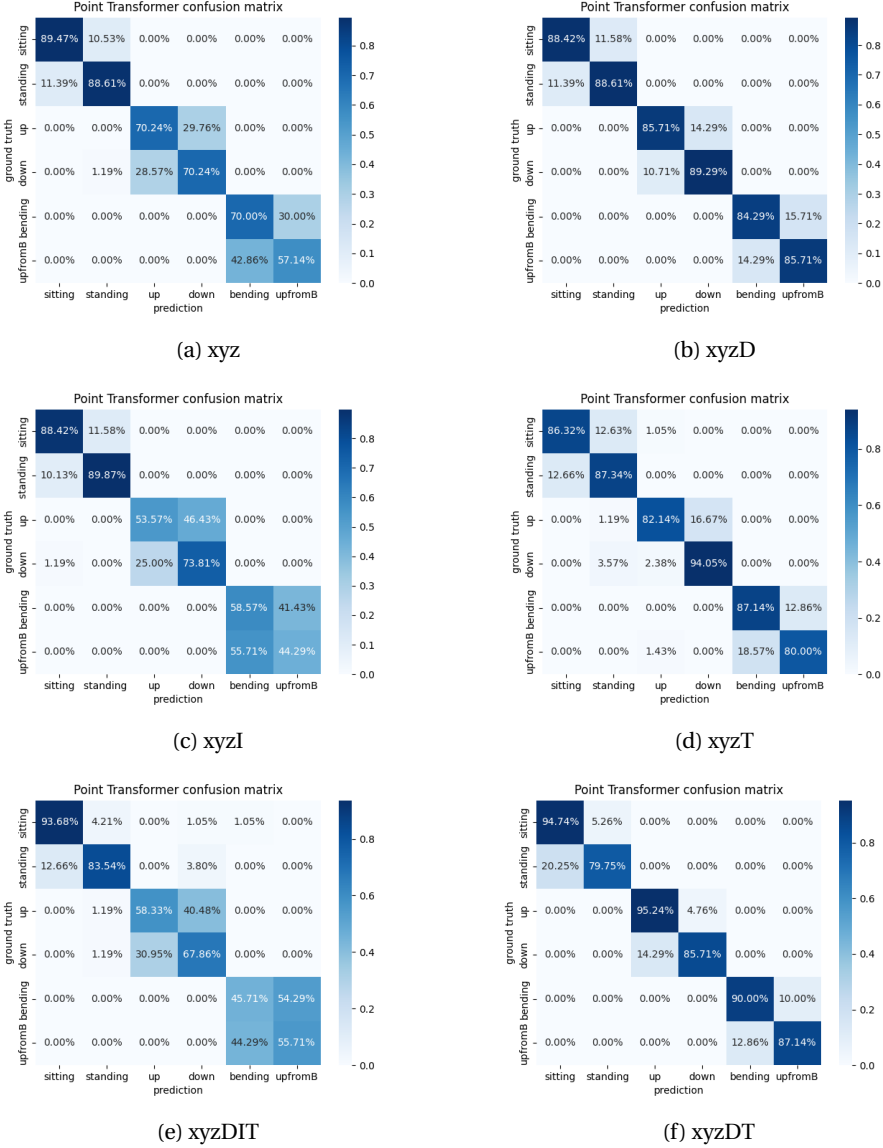


Figure 5.4: Normalized confusion matrix for Point Transformer to classify six motions and postures with (a) only coordinates, (b) coordinates and Doppler, (c) coordinates and intensity, (d) coordinates and time, (e) coordinates, Doppler, intensity and time, (f) coordinates, Doppler and time



Figure 5.4a shows the results of using only 3D coordinates, and it is used to compare with other results as the baseline. Here, we can notice that the classification results are almost in pair, e.g. it is possible for the network to be confused with standing up and sitting down, but the network is unlikely to mix standing up with bending down.

Only with the coordinates of the point cloud, the network is able to classify six classes roughly, and it performs better when distinguishing static posture. For the motions, the results are not good enough, since the spatial distribution for these paired motions are almost the same, and they are symmetrical in time. That is, standing up can be regarded as the inverse motion of sitting down. In the previous work in [12], the solution to distinguish the paired motions is to introduce spectrograms, which can reflect the difference of Doppler over time for each motion. For the proposed pipeline, as Figure 5.4b shows, after adding the Doppler information to each point, the classification results for the motions improve and the results for all classes are balanced. Moreover, we can achieve similar results if we add time information as the extra feature as shown in Figure 5.4c. However, after adding intensity information to input points, the classification accuracy for motions declines greatly while the results for postures remain unchanged, as Figure 5.4c and Figure 5.4e show. When the Doppler and time information are added to points, we can obtain the best classification results, that is, the highest accuracy as shown in Figure 5.4f.

### 5.2.2. RESULTS WITH ADAPTIVE CLUTTER CANCELLATION

As the description in Chapter 4, many points in the point clouds are actual clutter contributions due to the items around the human subjects, so the proposed ACC is applied to remove the clutter in the experiment scene. The improvements brought by ACC for different features as input are listed in Table 5.2.

Table 5.2: F1 score of Point Transformer with and without the proposed adaptive clutter cancellation

Input features	With ACC	Without ACC	Difference
xyz	0.792	0.741	+0.051 (5%)
xyzD	0.893	0.869	+0.024 (2.4%)
xyzT	0.880	0.861	+0.021 (2.1%)
xyzDT	0.928	0.888	+0.040 (4%)

In this table, it is noticeable that after removing the clutter points outside the human movement area, the improvement for the F1 score is significant. For the best case, the F1 score improves by about +4% and reaches 92.8%. For other cases, ACC can also bring improvements, and it benefits most the group with only coordinates as input.

### 5.2.3. COMPARISON AMONG THREE ATTENTION-BASED MODELS

As introduced in Chapter 3, we have three attention-based models to investigate, and they are the point transformer from Hengshuang [5], the point cloud transformer from Menghao [6], and the point transformer from Nico [7], respectively. Figure 5.5 displays

the F1 scores of these three attention-based models with decreasing number of frames as inputs. Since F1 score combines the precision and recall of a classifier into a single metric by taking their harmonic mean, it is a more complete indicator to compare the performance of classifiers rather than recall and precision. In this section, F1 score is used to analyze and compare the performance of three models.

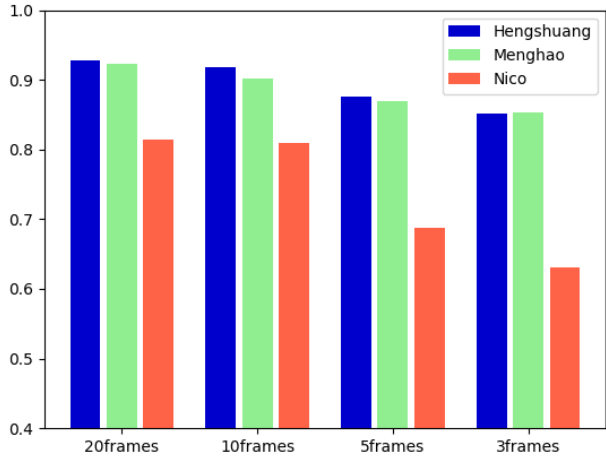


Figure 5.5: Classification F1 score for 3 different models with decreasing number of frames as input, where the horizontal axis represents the number of frames. The F1 scores are the average of 5-fold cross validation

Overall, the F1 score of Hengshuang’s model is slightly higher than Menghao’s model, but both perform clearly better than Nico’s model. Even the worst case of Hengshuang and Menghao’s model perform better than the best case of Nico’s model.

Additionally, a noticeable trend reflected from this figure is that as the number of frames is decreasing, the F1 score of the three models are also declining. For Hengshuang and Menghao’s models, the F1 scores decline slowly from about 0.92 to 0.85 over 4 different number of frames as input. The classification results of Hengshuang and Menghao’s model are still reliable even if only 3 frames of radar point cloud data are used. This trend is more evident for Nico’s model, and the F1 score shrinks from around 0.8 to 0.68 as the number of frames decreases from 10 to 5.

As for the reason for these differences, it can be traced back to the architecture of the three models. In the architectures of Hengshuang and Menghao’s models, the self attention mechanism is deployed in each hierarchical blocks, and that is to say, the features from the point clouds are extracted by self-attention mechanism. However, in Nico’s model, the features are extracted by SortNet, and the self-attention is just utilized to re-

late the local and global feature. Moreover, there is no hierarchical structure in Nico's network.

Accordingly, we can conclude from these initial results that the self-attention mechanism and hierarchical structure is very suitable for extracting the features from data representation of radar point cloud.

#### 5.2.4. LEAVE-ONE-SUBJECT-OUT TEST

It is expected that each person has its own specific kinematic patterns when performing certain activities. According to the interview with the subjects, some performed the motions with swinging arms, while other holding the arms static.

In order to train a generally applicable pipeline, it is important to test the generalization of the pipeline. That is to say, how the various kinematic patterns from each different human individual can affect the classification results. Leave-one-subject-out test means that the data of one subject are used for testing and the data of all the other subjects are used for training. Essentially, in every iteration the pipeline is trained to classify unseen kinematic patterns of data.

As shown in Figure 5.6, the average F1 score is around 0.82, which is approximately 0.1 lower than for the cross validation results.

- The drop in average classification F1 score in the leave-one-subject-out test fits the expectation that different people have their own kinematic patterns, making this a more complicated problem than using simpler cross-validation.
- Although there is a drop in average F1 score, it is of interest to find that the F1 score of subject 3 is 0.958 which is even 0.3 more than the original results. It can be inferred that the kinematic patterns of subject 3 contain somehow more general features of the activities.
- The F1 score of subject 1 and subject 4 are relatively low, and this shows that either the kinematic patterns of them are extremely different from the others, or that they performed the activities without following the 2 seconds interval approach (thus making invalid some of the labels used for the data).

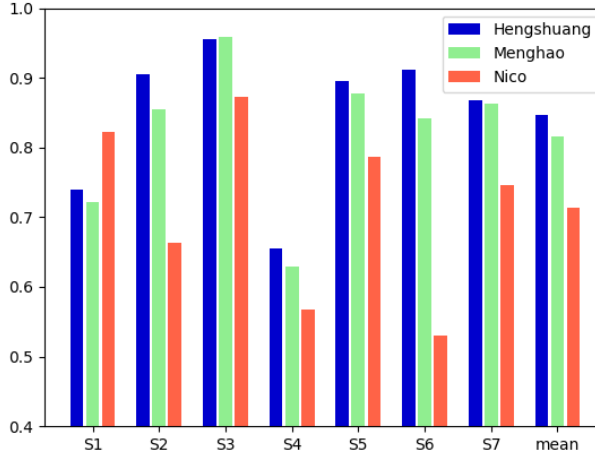


Figure 5.6: Classification F1 score in leave-one-subject-out test when using 20 frames, where the horizontal axis represents the index of the left-out subject, and S represents specific subjects

### 5.2.5. PERSON RECOGNITION

The drop in the average F1 score of leave-one-subject-out test means that specific individual differences in kinematic patterns can be encoded by the radar in the point cloud data. These differences can be captured by the proposed pipeline. Accordingly, it is possible to utilize the proposed pipeline to recognize specific individual based on the different kinematic patterns. As an example of the potential for this approach, the bending motion is selected as the data from which infer the identity of a specific individual. The reason we choose data of bending to recognize people is that this action contains the most movements, and thus it can reflect more differences in kinematic patterns.

In this test, we only use the bending data and relabel these data with an ID of the participants. For testing, there are only 12 samples for one class. As shown in Figure 5.7, it is surprising that only 2 subjects accuracy is 91.67 %, and the other accuracy are all 100 %, which means only 2 data are given wrong predictions.

Apart from the different kinematic patterns, the body characteristics such as weights and heights of different human subjects can also contribute to the people recognition via the shape of their bodies. In order to exclude the influence by different body characteristics, the radar point clouds for the static 'standing' posture with different human subjects as classes are fed into the proposed pipeline to see the variant of classification results.

As can be seen in Figure 5.8, the confusion matrix is disordered and the results are

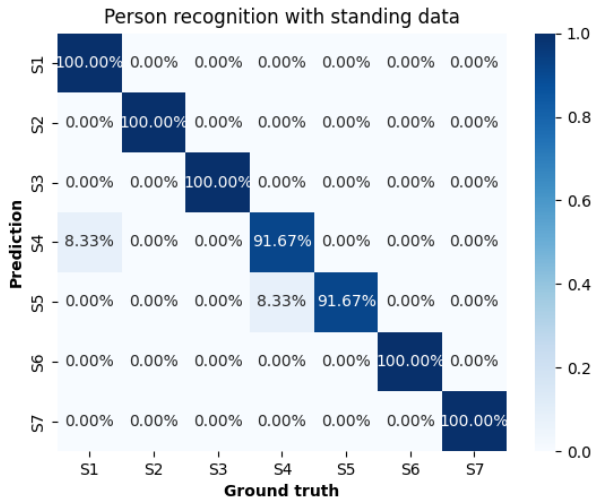


Figure 5.7: Normalized confusion matrix for Point Cloud Transformer to classify seven human subjects through their 'bending' motions.

just slightly better than random predictions. Hence, it is realized that the variety of body characteristics have little contribution to the people recognition. Though this promising results might be caused by the small amount of data, it is enough to demonstrate that the proposed pipeline has the potential to both recognize activities of various people and recognize people based on their unique kinematic patterns.

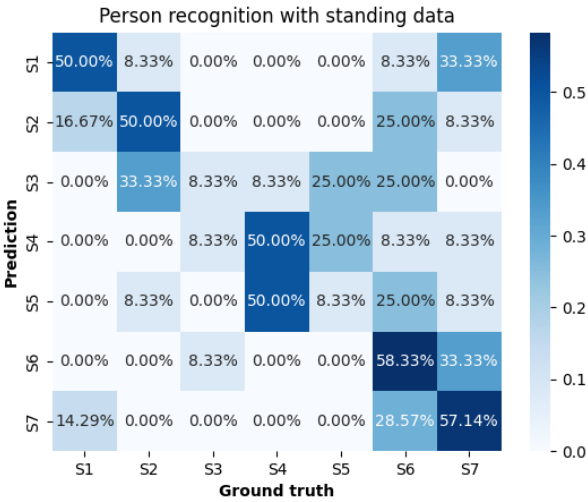


Figure 5.8: Normalized confusion matrix for Point Cloud Transformer to classify seven human subjects through their 'standing' postures.

# 6

## CONCLUSION AND FUTURE WORK

Automatic indoor HAR is a crucial technology to address the significant scarcity of healthcare workers caused by the aging population issue, and can improve the healthcare condition both in hospital and at home. With the technology of radar-based HAR, critical events such as seizure of epileptic patients and falling of frail people can be remotely detected in a timely manner from home, while protecting privacy.

Through literature review, for the most used HAR pipelines, the limitation of previous research on radar-based HAR is excessive dependency on image-like data representations such as spectrograms and range profiles. These are hardly sensitive to static activity because the features of static postures cannot be reflected by the spectrogram (no range or shape information) and the range profiles (no direct movement over time represented). Although the point cloud is a good data representation to contain the shape information of the target, for most researches, point clouds are processed as pictures to be forwarded to classifiers rather than the coordinates of points. This cannot make effective use of the whole information point clouds can provide.

Therefore, this MSc thesis work focuses on the limitation of common radar data representations and utilizes the data from MIMO imaging radar to test the proposed pipeline. The main reason to choose imaging radar rather than conventional radar is that the provided angle information allows more informative depiction of the shape of a human body combined with Doppler and temporal information. In principle, this combination can reflect different features for motions as well as postures. Hence, it is promising to solve HAR with this combination. Correspondingly, some challenges are presented as follows:

- Currently, no research explores systematically advantages and disadvantages of combining the coordinates of the point cloud with additional features such as Doppler and intensity.
- It is not yet established what classifier can be used to extract features from radar

point clouds to solve HAR problems.

To address these two challenges, a combination of features is proposed in Chapter 4, and an overall processing pipeline is proposed in Chapter 3. Data representation adds additional features to the point so that the 3D coordinates are extended to a 6D vector of  $(x, y, z, \text{Doppler}, \text{SNR}, \text{time})$ . The point cloud is therefore represented as a 2D matrix of the size of  $(N, 6)$ , where  $N$  is the number of points in a point cloud.

There are three modules in the proposed pipeline, the first of which aims to convert complex signals to point clouds by applying 2D FFT in range and Doppler, 2D-CFAR detector, and DOA estimation. In the second module, we proposed the ACC to remove the clutter and a sample strategy to match the point clouds with the input of the network, counteracting the fact that the number of points of the radar point cloud can be very unstable even for simple measurements. The third module contains the classifier to process the point cloud. Considering that point clouds are essentially sets embedded irregularly and disorderly in a metric space, it is suitable to deploy a self-attention based model to process point clouds, since the core of self-attention is to relate the different positions of inputs. Three self-attention-based models are investigated in this thesis. Furthermore, some datasets with fewer frames in a sample are also generated to investigate the classification results with a shorter period of available radar data. The main findings from the results are as follows:

- For the MMActivity dataset, the proposed pipeline can obtain the best classification using 6D  $(x, y, z, \text{Doppler}, \text{SNR}, \text{time})$  input data. It achieves an overall accuracy of 98.43%, 8% more than the best results of the RadHAR pipeline [10].
- For the MMActivity dataset, even when the number of input points is 128, corresponding to just 5 frames and 0.2 seconds of radar data, the proposed pipeline can still achieve promising accuracy (93.57%).
- For the TUD dataset, the proposed pipeline with proposed data representation achieves the accuracy of 92.8% for the problem of classifying 4 motion and 2 postures, bringing +5.8% improvement compared with the previous work on the same TUD dataset that did not use attention-based models [12].
- The proposed clutter removal method ACC is proved to be a crucial contribution of the pipeline, and it can improve the accuracy by 2% to 5%, depending on the different input features.
- The comparison of the three models shows that the point transformer from Hengshuang performs best: it obtains the highest classification F1 score while consuming the least training time.
- The leave-one-out test demonstrates that each human subject has unique kinematic patterns. From the aspect of classifiers, the point cloud features of the same activity but performed by different human subjects can be huge, and this needs to be taken into account.



- The person recognition test confirms the above findings and shows that the proposed pipeline can not only solve HAR issues, but can be also potentially used for people recognition problems.

### **Future work:**

Apart from improving the performance of the proposed pipeline with approaches such as parameter tuning, there are many aspects of future work worth to follow up. Referring back to the challenges of radar-based HAR, future work can also be divided into such two categories. Some general ideas are given as follows:

1. More realistic dataset needs to be collected. In the TUD dataset, human subjects are asked to perform the activities with a controlled 2-second interval repeatedly. However, in real scenes, people have the freedom to perform daily activities casually, but how to label these realistic data can be a new challenge, apart from the effort for the actual data collection. In addition to performing experiments to obtain radar data, simulation can also be a crucial method to generate additional, diverse radar point cloud data. For the data representation used in the proposed pipeline in the point cloud and in [75], a GAN based point cloud generation method has already been explored. With an increase in the amount of data, the performance of many DL models can also improve.
2. The work in this thesis only explores the condition that human subjects perform the activity from the line-of-sight direction of the radar. However, in the actual scene, the aspect angle can vary from 0 to 180, thus affecting Doppler information. How the proposed pipeline performs when processing the radar data of human activities from the non-line-of-sight orientation is worthy exploring.
3. Triggered by the opportunity of high-dimensional point clouds as the chosen data representation, many advanced classifiers of processing point clouds can be investigated for solving radar-based HAR. For instance, ModelNet40 is a point cloud dataset containing 40 classes, and in [76] so far there are more than 60 models of point clouds processing, and their results are updated. Although not directly related to HAR radar data, these could be used in transfer learning schemes to improve the results obtainable for radar problems.

## REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>
- [2] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.
- [3] “How do transformers work in nlp a guide to the latest state-of-the-art models,” <https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models>, accessed: 2020-08-12.
- [4] “The illustrated transformer,” <https://jalammar.github.io/illustrated-transformer/>, accessed: 2020-08-12.
- [5] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, “Point transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 259–16 268.
- [6] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, “Pct: Point cloud transformer,” *Computational Visual Media*, vol. 7, no. 2, pp. 187–199, 2021.
- [7] N. Engel, V. Belagiannis, and K. Dietmayer, “Point transformer,” *IEEE Access*, vol. 9, pp. 134 826–134 840, 2021.
- [8] C. Katzlberger and D. Gerstmair, “Object detection with automotive radar sensors using cfar algorithms,” Ph.D. dissertation, Johannes Kepler University Linz Linz, Austria, 2018.
- [9] “Iwr1443boost evaluation module mmwave sensing solution user guide,” <https://www.ti.com/lit/ug/swru518d/swru518d.pdf?ts=1660644884514>, accessed: 2020-08-15.
- [10] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, “Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar,” in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 2019, pp. 51–56.
- [11] “Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar,” <https://github.com/nesl/RadHAR>, accessed: 2020-08-15.
- [12] Y. Zhao, A. Yarovoy, and F. Fioranelli, “Angle-insensitive human motion and posture recognition based on 4d imaging radar and deep learning classifiers,” *IEEE Sensors Journal*, vol. 22, no. 12, pp. 12 173–12 182, 2022.

- [13] Y. Zhao, "Angle-insensitive human motion and posture recognition based on 2d fmcw mimo radar and deep learning classifiers," 2022. [Online]. Available: <http://resolver.tudelft.nl/uuid:059faecd-1713-459b-b089-c5ee3aae8ee0>
- [14] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.
- [15] "Ageing and health," <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>, accessed: 2020-08-23.
- [16] S. A. Shah and F. Fioranelli, "Rf sensing technologies for assisted daily living in healthcare: A comprehensive review," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 11, pp. 26–44, 2019.
- [17] C. Jobanputra, J. Bavishi, and N. Doshi, "Human activity recognition: A survey," *Procedia Computer Science*, vol. 155, pp. 698–703, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050919310166>
- [18] Z. Hussain, Q. Z. Sheng, and W. E. Zhang, "A review and categorization of techniques on device-free human activity recognition," *Journal of Network and Computer Applications*, vol. 167, p. 102738, oct 2020. [Online]. Available: <https://doi.org/10.1016%2Fj.jnca.2020.102738>
- [19] F. Fioranelli, J. Le Kernec, and S. A. Shah, "Radar for health care: Recognizing human activities and monitoring vital signs," *IEEE Potentials*, vol. 38, no. 4, pp. 16–23, 2019.
- [20] G. Lee and J. Kim, "Improving human activity recognition for sparse radar point clouds: A graph neural network model with pre-trained 3d human-joint coordinates," *Applied Sciences*, vol. 12, no. 4, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/4/2168>
- [21] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sensing*, vol. 11, no. 9, p. 1068, 2019.
- [22] Y. Lang, C. Hou, Y. Yang, D. Huang, and Y. He, "Convolutional neural network for human micro-Doppler classification," in *Proc. Eur. Microw. Conf.*, 2017, pp. 1–4.
- [23] Y. Kim and H. Ling, "Human activity classification based on micro-doppler signatures using a support vector machine," *IEEE transactions on geoscience and remote sensing*, vol. 47, no. 5, pp. 1328–1337, 2009.
- [24] D. G. Bresnahan and Y. Li, "Classification of driver head motions using a mm-wave FMCW radar and deep convolutional neural network," *IEEE Access*, vol. 9, pp. 100 472–100 479, 2021.
- [25] J. Zhang, J. Tao, and Z. Shi, "Doppler-radar based hand gesture recognition system using convolutional neural networks," in *International Conference in Communications, Signal Processing, and Systems*. Springer, 2017, pp. 1096–1113.

- [26] H. T. Le, S. L. Phung, A. Bouzerdoum, and F. H. C. Tivive, "Human motion classification with micro-Doppler radar and bayesian-optimized convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2961–2965.
- [27] Y. Shao, S. Guo, L. Sun, and W. Chen, "Human motion classification based on range information with deep convolutional neural network," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, 2017, pp. 1519–1523.
- [28] I. Alujaim, I. Park, and Y. Kim, "Human motion detection using planar array FMCW radar through 3d point clouds," in *2020 14th European Conference on Antennas and Propagation (EuCAP)*, 2020, pp. 1–3.
- [29] Y. Kim and T. Moon, "Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 8–12, 2016.
- [30] J. Park, R. Javier, T. Moon, and Y. Kim, "Micro-Doppler based classification of human aquatic activities via transfer learning of convolutional neural networks," *Sensors*, vol. 16, p. 1990, 11 2016.
- [31] Y. Shao, Y. Dai, L. Yuan, and W. Chen, "Deep learning methods for personnel recognition based on micro-Doppler features," 11 2017, pp. 94–98.
- [32] R. Trommel, R. Harmanny, L. Cifola, and J. Driessen, "Multi-target human gait classification using deep convolutional neural networks on micro-Doppler spectrograms," in *2016 European Radar Conference (EuRAD)*, 2016, pp. 81–84.
- [33] R. I. A. Harmanny, J. J. M. de Wit, and G. P. Cabic, "Radar micro-Doppler feature extraction using the spectrogram and the cepstrogram," in *2014 11th European Radar Conference*, 2014, pp. 165–168.
- [34] A. Yarovoy, L. Ligthart, J. Matuzas, and B. Levitas, "Uwb radar for human being detection," *IEEE Aerospace and Electronic Systems Magazine*, vol. 21, no. 3, pp. 10–14, 2006.
- [35] X. Li, Y. He, Y. Yang, Y. Hong, and X. Jing, "Lstm based human activity classification on radar range profile," in *2019 IEEE International Conference on Computational Electromagnetics (ICCEM)*. IEEE, 2019, pp. 1–2.
- [36] Q. Jian, S. Guo, P. Chen, P. Wu, and G. Cui, "A robust real-time human activity recognition method based on attention-augmented gru," in *2021 IEEE Radar Conference (RadarConf21)*. IEEE, 2021, pp. 1–5.
- [37] B. Erol, M. Amin, Z. Zhou, and J. Zhang, "Range information for reducing fall false alarms in assisted living," in *2016 IEEE Radar Conference (RadarConf)*, 2016, pp. 1–6.

- [38] S. An and U. Y. Ogras, “Mars: mmwave-based assistive rehabilitation system for smart healthcare,” *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 20, no. 5s, pp. 1–22, 2021.
- [39] R. Dey and F. M. Salem, “Gate-variants of gated recurrent unit (gru) neural networks,” in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. IEEE, 2017, pp. 1597–1600.
- [40] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [41] M. Wang, Y. D. Zhang, and G. Cui, “Human motion recognition exploiting radar with stacked recurrent neural network,” *Digital Signal Processing*, vol. 87, pp. 125–131, 2019.
- [42] H. Jiang, F. Fioranelli, S. Yang, O. Romain, and J. Le Kernec, “Human activity classification using radar signal and rnn networks,” 2021.
- [43] Z. Zhang, Z. Tian, and M. Zhou, “Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor,” *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3278–3289, 2018.
- [44] G. Park, V. K. Chandrasegar, and J. Koh, “Hand gesture recognition using deep learning method,” in *2021 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (APS/URSI)*. IEEE, 2021, pp. 1347–1348.
- [45] T. Tang, C. Wang, and M. Gao, “Radar target recognition based on micro-Doppler signatures using recurrent neural network,” in *2021 IEEE 4th International Conference on Electronics Technology (ICET)*, 2021, pp. 189–194.
- [46] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, “Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum,” in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, ser. UIST ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 851–860. [Online]. Available: <https://doi.org/10.1145/2984511.2984565>
- [47] H. Guo, N. Zhang, S. Wu, and Q. Yang, “Deep learning driven wireless real-time human activity recognition,” in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.
- [48] G. Klarenbeek, R. I. A. Harmanny, and L. Cifola, “Multi-target human gait classification using lstm recurrent neural networks applied to micro-doppler,” in *2017 European Radar Conference (EURAD)*, 2017, pp. 167–170.
- [49] H. Sadreazami, M. Bolic, and S. Rajan, “On the use of ultra wideband radar and stacked lstm-rnn for at home fall detection,” in *2018 IEEE Life Sciences Conference (LSC)*, 2018, pp. 255–258.

- [50] Y. Sun, R. Hang, Z. Li, M. Jin, and K. Xu, "Privacy-preserving fall detection with deep learning on mmwave radar signal," in *2019 IEEE Visual Communications and Image Processing (VCIP)*, 2019, pp. 1–4.
- [51] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-lstm network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2019.
- [52] H. Li, A. Mehul, J. Le Kernec, S. Z. Gurbuz, and F. Fioranelli, "Sequential human gait classification with distributed radar sensor fusion," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7590–7603, 2020.
- [53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [55] L. Zheng, J. Bai, X. Zhu, L. Huang, C. Shan, Q. Wu, and L. Zhang, "Dynamic hand gesture recognition in in-vehicle environment based on fmcw radar and transformer," *Sensors*, vol. 21, no. 19, p. 6368, 2021.
- [56] S. Chen, W. He, J. Ren, and X. Jiang, "Attention-based dual-stream vision transformer for radar gait recognition," 2021.
- [57] J. Bai, L. Zheng, S. Li, B. Tan, S. Chen, and L. Huang, "Radar transformer: An object classification network based on 4d mmw imaging radar," *Sensors*, vol. 21, no. 11, p. 3854, 2021.
- [58] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.
- [59] B. Erol and S. Z. Gurbuz, "A kinect-based human micro-doppler simulator," *IEEE Aerospace and Electronic Systems Magazine*, vol. 30, no. 5, pp. 6–17, 2015.
- [60] B. Erol, C. Karabacak, S. Z. Gürbüz, and A. C. Gürbüz, "Radar simulation of different human activities via kinect," in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2014, pp. 1015–1018.
- [61] I. Alnujaim, D. Oh, and Y. Kim, "Generative adversarial networks for classification of micro-doppler signatures of human activity," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 3, pp. 396–400, 2020.
- [62] S. Vishwakarma, C. Tang, W. Li, K. Woodbridge, R. Adve, and K. Chetty, "Gan based noise generation to aid activity recognition when augmenting measured wifi radar data with simulations," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–6.

- [63] Y. Li, K. He, D. Xu, and D. Luo, "A transfer learning method using speech data as the source domain for micro-doppler classification tasks," *Knowledge-Based Systems*, vol. 209, p. 106449, 2020.
- [64] B. Erol and M. G. Amin, "Radar data cube processing for human activity recognition using multisubspace learning," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 6, pp. 3617–3628, 2019.
- [65] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.
- [66] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [67] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.
- [68] S. Yang, X. Yu, and Y. Zhou, "Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example," in *2020 International workshop on electronic communication and artificial intelligence (IWECAI)*. IEEE, 2020, pp. 98–101.
- [69] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.
- [70] H. Rohling, "Radar cfar thresholding in clutter and multiple target situations," *IEEE transactions on aerospace and electronic systems*, no. 4, pp. 608–621, 1983.
- [71] M. A. Habib, M. Barkat, B. Aissa, and T. Denidni, "Ca-cfar detection performance of radar targets embedded in "non centered chi-2 gamma" clutter," *Progress In Electromagnetics Research*, vol. 88, pp. 135–148, 2008.
- [72] "Iwr1443single-chip76-to8-ghzmmwavesensor," <https://www.ti.com/lit/ds/symmlink/iwr1443.pdf?ts=1659677302151>, accessed: 2020-08-05.
- [73] "Design guide: Tidep-01012 imaging radar using cascaded mmwave sensor reference design," Texas Instrumentation, online; accessed 19 August 2021.
- [74] "Mmwave\_studio 03\_00\_00\_14 user guide," [https://software-dl.ti.com/ra-processors/esd/MMWAVE-STUDIO-2G/latest/index\\_FDS.html](https://software-dl.ti.com/ra-processors/esd/MMWAVE-STUDIO-2G/latest/index_FDS.html), accessed: 2020-08-13.
- [75] D. W. Shu, S. W. Park, and J. Kwon, "3d point cloud generative adversarial network based on tree structured graph convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [76] "Princeton modelnet project," <https://modelnet.cs.princeton.edu/>, accessed: 2020-08-21.