applied sciences

Article

Human Activity Recognition Based on Point Clouds from Millimeter-Wave Radar

Seungchan Lim [1], **Chaewoon Park** [1], **Seongjoo Lee** [2,3] and **Yunho Jung** [1,4,*]

[1] School of Electronics and Information Engineering, Korea Aerospace University, Goyang 10540, Republic of Korea; lsc2007@kau.kr (S.L.); pcw0201@kau.kr (C.P.)
[2] Department of Electrical Engineering, Sejong University, Seoul 05006, Republic of Korea; seongjoo@sejong.ac.kr
[3] Department of Convergence Engineering of Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea
[4] Department of Smart Air Mobility, Korea Aerospace University, Goyang 10540, Republic of Korea
* Correspondence: yjung@kau.ac.kr; Tel.: +82-2-300-0133

Abstract: Human activity recognition (HAR) technology is related to human safety and convenience, making it crucial for it to infer human activity accurately. Furthermore, it must consume low power at all times when detecting human activity and be inexpensive to operate. For this purpose, a low-power and lightweight design of the HAR system is essential. In this paper, we propose a low-power and lightweight HAR system using point-cloud data collected by radar. The proposed HAR system uses a pillar feature encoder that converts 3D point-cloud data into a 2D image and a classification network based on depth-wise separable convolution for lightweighting. The proposed classification network achieved an accuracy of 95.54%, with 25.77 M multiply–accumulate operations and 22.28 K network parameters implemented in a 32 bit floating-point format. This network achieved 94.79% accuracy with 4 bit quantization, which reduced memory usage to 12.5% compared to existing 32 bit format networks. In addition, we implemented a lightweight HAR system optimized for low-power design on a heterogeneous computing platform, a Zynq UltraScale+ ZCU104 device, through hardware–software implementation. It took 2.43 ms of execution time to perform one frame of HAR on the device and the system consumed 3.479 W of power when running.

Keywords: millimeter-wave radar; 3D point cloud; human activity recognition; field-programmable gate array

check for updates

Citation: Lim, S.; Park, C.; Lee, S.; Jung, Y. Human Activity Recognition Based on Point Clouds from Millimeter-Wave Radar. *Appl. Sci.* **2024**, *14*, 10764. https://doi.org/10.3390/app142210764

Academic Editor: Grigorios Beligiannis

Received: 23 October 2024
Revised: 15 November 2024
Accepted: 19 November 2024
Published: 20 November 2024

1. Introduction

The increasing number of single-person and elderly households has drawn attention to the necessity for continuous monitoring to enable the prevention of and fast response to accidents at home. In the event of an accident, individuals often struggle to receive immediate assistance from those nearby; therefore, human activity recognition (HAR) systems should be deployed to accurately infer human activity in case of emergency. In addition, the power consumption for detecting human activity must be low at all times, and the cost of operating the HAR system must be low in order for it to be available for as many people as possible. Therefore, it is essential to research low-power and lightweight HAR systems that achieve high accuracy in human activity inference. In response to this need, HAR systems using wearable sensors with built-in accelerometers and gyroscopes for indoor accident detection [1–3], and sensors that observe the surrounding environment, such as cameras [4–6], light detection and ranging (LiDAR) [7], ultrasonic sensors [8,9], and radar [10–15], have been actively researched.

A limitation of wearable sensors is that the device must be worn continuously to infer human activity [16]. Ultrasonic sensors have a narrow sensing range, and their performance varies significantly depending on the material or shape of the object being detected [17]. Camera-based sensors can cause portrait-rights infringement when collecting image data [17]. LiDAR has the disadvantages of high cost, narrow detection range, and sensors that are highly affected by the surrounding environment [17]. In comparison, radar has the advantages of having a lower cost than LiDAR and a wider range of detection and is less affected by the surrounding environment [16,17].

Several types of data can be collected using radar; however, two main types of data can be used when performing HAR based on radar sensors—range-Doppler map and point clouds. Point-cloud data have a sufficiently high resolution to distinguish the shape of objects, and a higher inference accuracy can be achieved while maintaining a level of network complexity similar to that when using range-Doppler maps as input data [18].

Singh et al. [11] constructed an HAR system using point-cloud data collected by millimeter-wave radar. They collected an MMActivity dataset consisting of five classes and used a model consisting of a time-distributed, convolutional neural network (CNN) and a bi-directional long short-term memory (LSTM). The classification network of this HAR system achieved 90.47% accuracy using 291 K parameters. Kim et al. [12] conducted human motion classification using a point-cloud dataset consisting of seven classes, which they collected themselves. The authors used a model consisting of 2D-DCNN and DRNN, and they achieved 96% accuracy with 1510 K parameters using 2D-DCNN alone, and 98% accuracy using a combination of 2D-DCNN and DRNN. Huang et al. [13] constructed an HAR system using their own point-cloud dataset and a range-Doppler dataset consisting of six classes. The point-cloud data pass through a 3D CNN and LSTM, and the range-Doppler data pass through a 3D CNN. The two types of data are concatenated to classify human activity. An accuracy of 97% was achieved using a fusion network that placed two networks in parallel. Ding et al. [14] used six classes of point-cloud data collected by themselves to classify human activity. The method using time-Doppler (TD) achieved 95% accuracy in combination with 3D-PointNet using 1610 K parameters, and the method using range-Doppler (RD) achieved 98% accuracy in combination with 4D-PointNet. Gu et al. [15] augmented five classes of self-collected point-cloud data with segment-wise point-cloud augmentation (SPCA) to organize a dataset and infer human activity. In this study, Lite-PointNet and a bidirectional lightweight LSTM (BiLiLSTM) were used to achieve 95% accuracy. Lite-PointNet achieved high accuracy with a lightweight network using 79.7 K parameters, and the HAR system was ported to Raspberry Pi to implement the edge device.

To effectively utilize HAR technology in indoor environments, edge devices capable of performing HAR must be deployed to infer the activity of people, wherever they are. Multiple edge devices are required to build such an environment, whose cost is closely related to the cost of the edge devices. To reduce the cost of edge devices, a lightweight HAR system aimed at low memory usage is essential, and a low-power design should be used to reduce maintenance costs. Although high HAR accuracy is important, it is also important to design an appropriate HAR model that considers network complexity and therefore memory usage.

In this study, a pillar feature encoder (PFE) was used as an encoder for 3D point-cloud data for the purpose of lightweighting the network [19]. The PFE clusters 3D point clouds and converts them into 2D images. The advantage of using PFE is that 2D convolution can be applied on 2D images instead of 3D images in the classification network, thereby reducing the network complexity. The classification network was optimized based on depth-wise separable convolution and has the advantage of low complexity. Depth-wise separable convolution consists of depth-wise convolution and point-wise convolution, in which fewer parameters are required for the operation compared to a general convolution [20].

Existing radar point-cloud HAR studies have achieved high accuracy but are limited by the lack of discussion on lightweight HAR systems and low-power designs. Considering the practicality of HAR technology, it is crucial to develop a lightweight HAR system that operates efficiently at low power on edge devices. Therefore, in this study we used Xilinx's FINN [21] to design a hardware–software implementation of an HAR system on an edge device, aiming to achieve a low-power, lightweight solution while maintaining an inference accuracy comparable to existing studies.

The remainder of this paper is organized as follows: In Section 2, we introduce the dataset collection. In Section 3, we describe the structure of the proposed HAR system in relation to data pre-processing, the encoding of point-cloud data, and a classification network. In Section 4, we evaluate the performance of the hardware–software implementation of the HAR system using FINN. Finally, in Section 5, we conclude the study.

2. Data Collection Using Frequency Modulated Continuous Wave (FMCW) Radar

The dataset used in this study was collected using the RETINA-4SN radar (Smart Radar System, Gyeonggi, Republic of Korea) of the Smart Radar system [22]. The RETINA-4SN radar is an FMCW and multi-input–multi-output (MIMO) millimeter-wave radar that can obtain (x, y, z) coordinates and p (power) for each point in a point cloud at a rate of 20 frames per second. The detailed specifications of the radar are listed in Table 1.

Table 1. Radar specifications.

Parameter	Quantity
Start frequency	77 GHz
Stop frequency	81 GHz
Bandwidth	4 GHz
Azimuth angle FoV	90°
Elevation angle FoV	90°
Detection range	12 m
Number of transmitter antennas	12
Number of receiver antennas	16
Number of frames per second	20

The data collection process was carried out as shown in Figure 1, where the radar was positioned 0.8 m above ground. Data were collected by having a person perform the actions of each class in the center of a 5 m long and 4 m wide area. A total of three volunteers (subjects), two men and one woman, participated in the experiment. Their heights ranged from 163 to 180 cm, and their weights ranged from 58 to 80 kg. The collected dataset was organized into 11 classes based on activities that often occur in daily life. An example of point-cloud data for each class is shown in Figure 2. Frames with fewer than 10 points included in one frame were excluded from the dataset. Table 2 shows details of the dataset, where each class consists of approximately 1000 to 1100 frames, the maximum number of points in a frame is 1280, and there is an average of 343 points per frame across all classes.

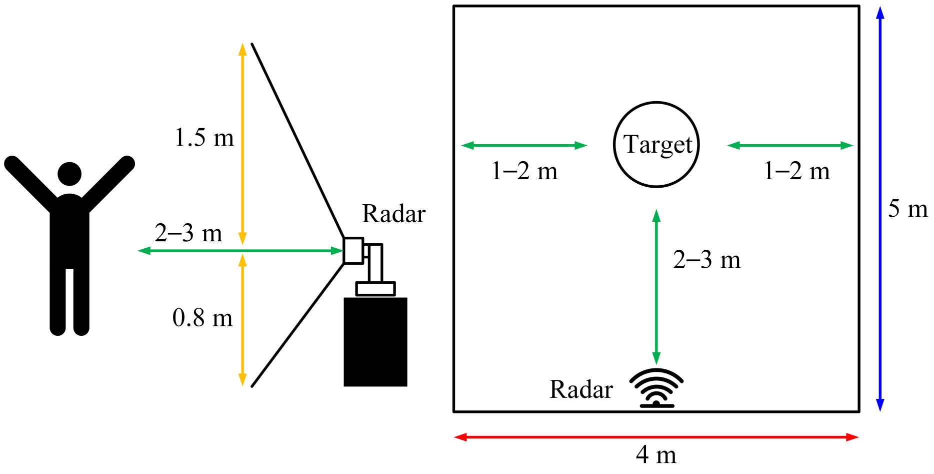
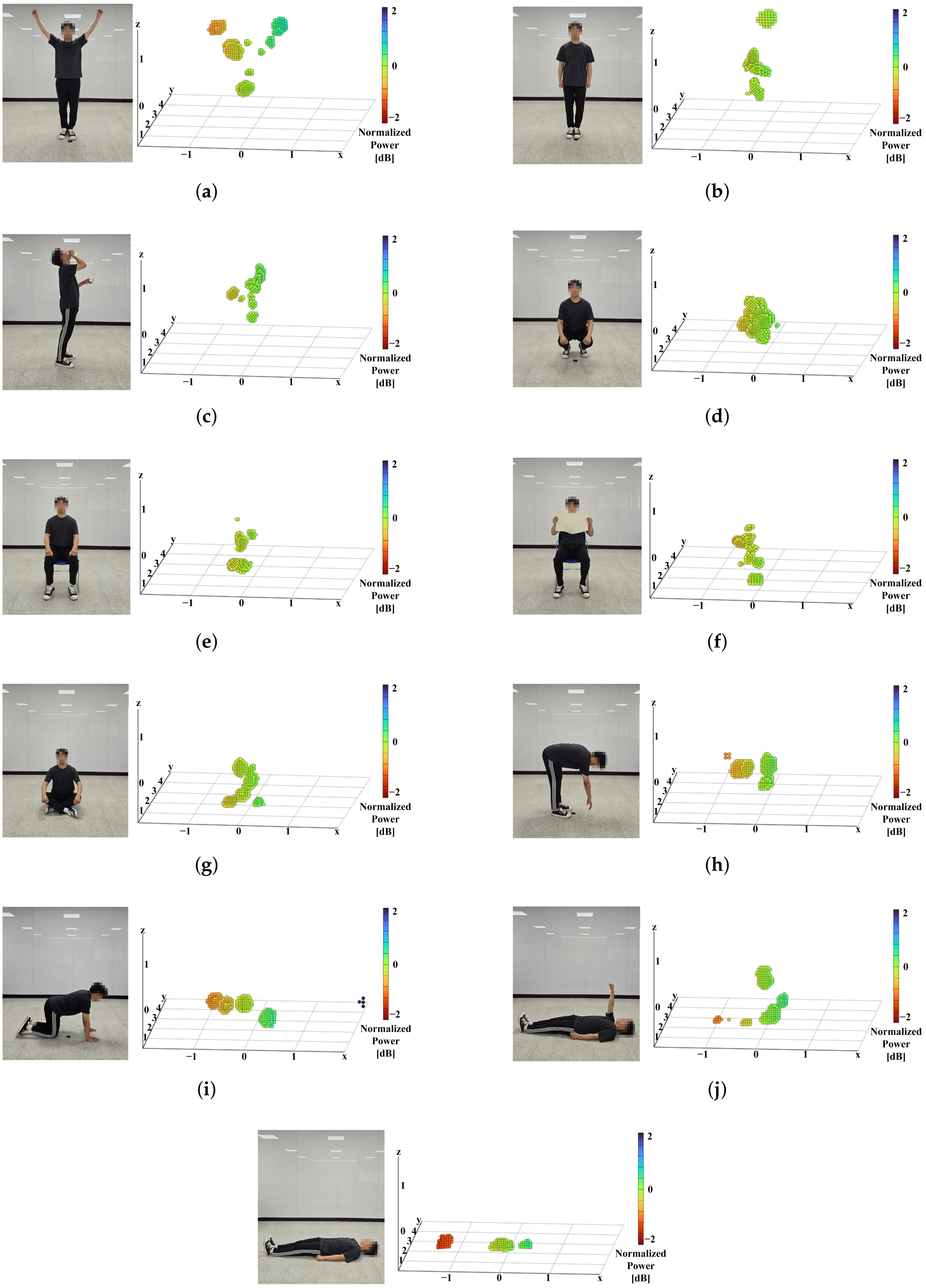


Figure 1. Data collection setup.



(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

(**f**)

(**g**)

(**h**)

(**i**)

(**j**)

(**k**)

Figure 2. Configuration of dataset classes and their corresponding point clouds: (**a**) Stretching; (**b**) Standing; (**c**) Taking medicine; (**d**) Squatting; (**e**) Sitting chair; (**f**) Reading news; (**g**) Sitting floor; (**h**) Picking; (**i**) Crawl; (**j**) Lying wave hands; (**k**) Lying.

Table 2. Eleven classes of activity type in the dataset.

Class	Number of Frames	Average Number of Points per Frame
Crawl	1069	321
Lying	1048	172
Lying wave hands	1075	336
Picking	1075	382
Reading news	1098	422
Sitting chair	1060	374
Sitting floor	1063	280
Squatting	1071	389
Standing	1073	340
Stretching	1062	402
Taking medicine	1075	345

3. Proposed HAR System

To design a low-power and lightweight HAR system, first, we designed the HAR system with a 32 bit floating-point format in a GPU environment. Subsequently, the classification network was quantized to a 4 bit fixed-point format for lightweight, and the memory usage was reduced to 12.5% of the original 32 bit floating-point format network. Finally, for a low-power design, the HAR system was hardware–software designed on a heterogeneous computing platform, ZCU104 [23]. Figure 3 shows an overview of the proposed HAR system. The point-cloud data obtained by the radar are input into the microprocessor, and pre-processing is performed on the input point-cloud data. The pre-processed point-cloud data are input to the PFE and converted into a 2D pseudo image by grouping 3D point-cloud data, creating handcraft features and scattering. The converted 2D image is input to the classification network IP implemented on a field-programmable gate array (FPGA), and the activity class is inferred and output through deep-learning operations based on depth-wise separable convolution.

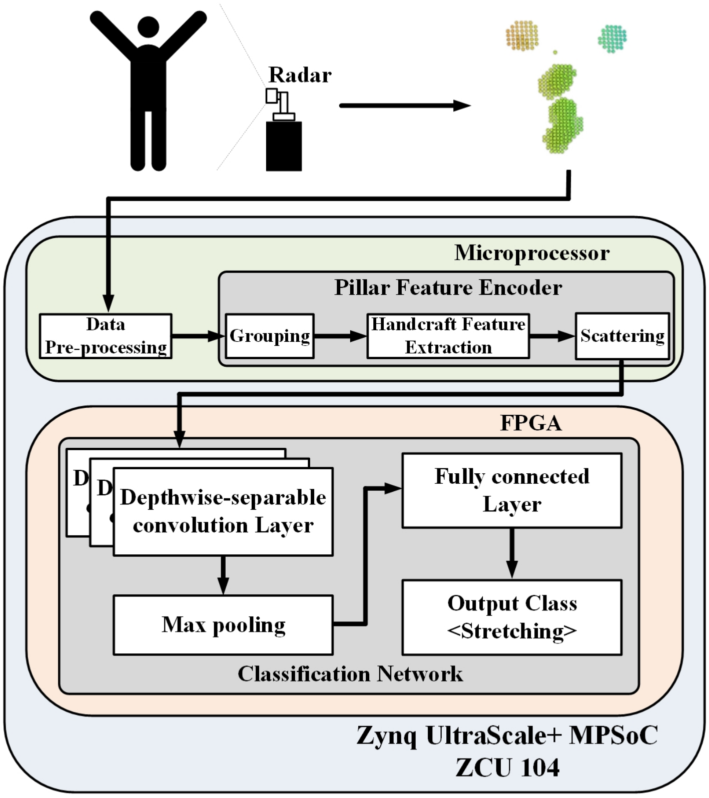


Figure 3. Overview of the proposed HAR system.

3.1. Data Pre-Processing

The data pre-processing performed in the microprocessor included both parallel translation and normalization processes. Normalizing point-cloud data ensures that all points are distributed within a certain range without information loss [24]. This facilitates gradient descent learning and prevents bias toward specific values. In addition, by performing a parallel translation before normalization—to move the center of the point cloud to the origin—the point cloud was evenly distributed around the origin when normalization was performed.

In this study, parallel translation was performed to move the center of the maximum and minimum values along the (x, y, z) axis to the origin for the points included in each frame. Subsequently, data pre-processing was performed to normalize the (x, y, z) coordinates of all points by dividing the (x, y, z) coordinates of all points by the largest absolute value.

3.2. Pillar Feature Encoder

There are several methods to encode point-cloud data. The voxel feature encoding (VFE) layer of VoxelNet [25] uses voxels to encode in three dimensions; this has the disadvantage of increasing the network complexity using 3D convolution. In addition, the multi-view (MV)CNN [26] uses as its input the 2D images that are projected from multiple directions of 3D point-cloud data; this has the disadvantage of overlapping points occurring during the process of projecting points, resulting in information loss. However, the PFE uses pillars to convert 3D point-cloud data into 2D pseudo images to enable image-based inference. Therefore, the PFE can be used to convert 3D point-cloud data into 2D images without loss of information, and as it outputs a 2D image it has the advantage of lowering the complexity of the network by utilizing 2D convolutions instead of 3D convolutions in the classifier.

The PFE consists of three major steps: grouping, handcraft feature extraction, and scattering. Grouping is the process of grouping points into pillars with an infinitely extended z-axis. The xy plane was divided into uniformly spaced regions, and each region was treated as a pillar. Points within a uniform area were grouped into the internal set of pillars of the region; if no points were inside a pillar, the pillar was excluded from the grouping. Finally, the remaining pillar contained the pillar index, which indicated its position on the xy plane, the number of points inside the pillar, and the (x, y, z, p) channel information of the internal point. In the handcraft feature extraction process, the point data consisting of (x, y, z, p) 4 channels were expanded to a total of $(x, y, z, p, x_c, y_c, z_c, x_p, y_p)$ 9 channels by adding the center point coordinates (x_c, y_c, z_c) of the points inside the pillar and the center coordinates (x_p, y_p) of the pillar. Subsequently, the point data of the 9 channels were expanded to 64 channels by passing them through point-wise convolution, batch normalization, and the ReLU activation function. In the scattering process, the pillar index information saved in the grouping process was used to place each pillar and the point data of the 64 channels within the pillar at the existing location on the xy plane.

3.3. Classification Network

In this study, we optimized and used a depth-wise separable convolution-based network from various existing image inference networks. A depth-wise separable convolution consists of point-wise convolution and depth-wise convolution. Point-wise convolution adjusts the number of output channels using a 1×1 filter, whereas depth-wise convolution is a convolution in which one filter is applied to only one channel. When the input has M channels, depth-wise convolution requires only M filters; therefore, the network can be constructed using fewer filters, or in other words, fewer parameters, compared to general convolution.

4. Results

4.1. Experiment

In this paper, we compared the performance of the different network configurations in Table 3 by setting the following as parameters: the 2D pseudo image size encoded in PFE, the number of depth-wise separable convolution layers (#ds-conv.), and the channel configuration of the network. Table 4 lists the numbers of output channels per convolution layer for each network (A, B, C, D, E). The Network A configuration has 64 output channels for all depth-wise separable convolution layers, and Network B has 128 output channels for the last depth-wise separable convolution layer. Network C has the last two depth-wise separable convolution layers, each with 128 output channels. Network D has the last depth-wise separable convolution layer with 256 output channels and the previous layer with 128 output channels. Finally, Network E has 256 output channels in the last depth-wise separable convolutional layer and 128 output channels in each of the two previous layers. In this way, a total of 42 networks were configured, and an NVIDIA RTX A6000 GPU [27] (NVIDIA, Santa Clara, CA, USA) was used for network training and verification. In addition, all the networks used the same training settings (300 epochs, learning rate of 0.001, batch size of 24, Adam optimizer, and negative log likelihood loss (NLL loss)). For training and verification, we used a dataset of 11,769 frames collected by ourselves, which were divided into training data of 10,594 frames and test data of 1175 frames. In addition, to investigate the generality of the proposed HAR system, we performed stratified 10-fold cross validation. "Accuracy" was calculated to evaluate the classification performance in terms of true positives (*TPs*), true negatives (*TNs*), false positive (*FP*), and false negatives (*FNs*).

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

Table 3. Accuracy of models by various parameters and networks.

Parameter		Network				
Image Size	**#ds-conv.**	**A**	**B**	**C**	**D**	**E**
32 × 32	2	91.22%	92.89%	93.25%	93.96%	-
	3	92.76%	93.55%	94.09%	94.13%	94.38%
	4	92.52%	92.97%	93.31%	94.08%	94.40%
64 × 64	2	92.36%	94.66%	94.72%	95.36%	-
	3	94.65%	95.54%	95.36%	96.41%	96.65%
	4	95.38%	95.81%	95.76%	96.53%	96.56%
128 × 128	2	92.82%	94.72%	94.95%	95.74%	-
	3	94.22%	95.21%	95.60%	96.16%	96.47%
	4	95.30%	95.52%	95.78%	96.62%	96.75%

Table 4. Output channel configuration for network (A,B,C,D,E).

Network	#ds-conv.		
	2	**3**	**4**
A	64, 64	64, 64, 64	64, 64, 64, 64
B	64, 128	64, 64, 128	64, 64, 64, 128
C	128, 128	64, 128, 128	64, 64, 128, 128
D	128, 256	64, 128, 256	64, 64, 128, 256
E	-	128, 128, 256	64, 128, 128, 256

Finally, the model with the smallest number of network parameters (64 × 64 image size, 3 depth-wise separable convolutions, and B network configuration) was selected as the final image classification network among those models with an HAR accuracy of 95% or higher (Table 3) to configure the HAR system.

The backbone network structure of the proposed HAR system is shown in Figure 4. When a 2D image of size 64 × 64 with 64 channels was input, it was converted to an 8 × 8 image with 128 channels using three depth-wise separable convolution layers. It was then passed through max pooling and the fully connected layer to classify the 11 classes of human activity.

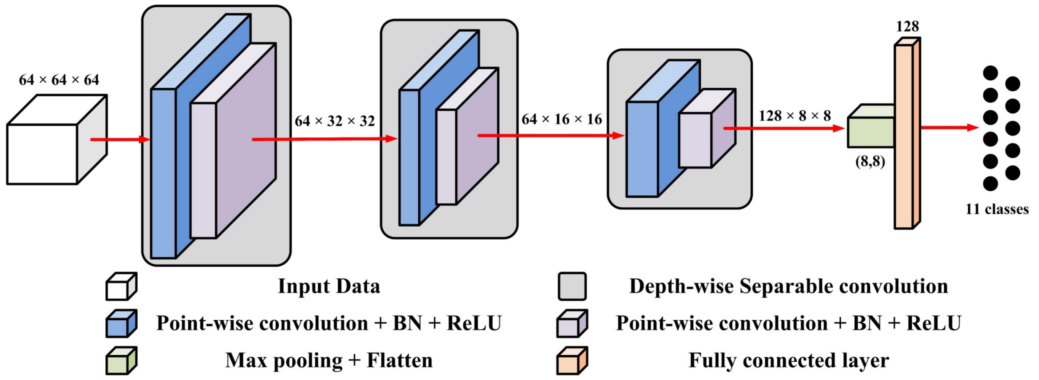


Figure 4. Proposed classification network.

We compared the performance of the optimized depth-wise separable convolution-based network with the existing, representative, image classification networks LeNet5 [28], VGGNet [29], ResNet [30], and MobileNet [31] as the backbone network of the HAR system. Table 5 shows the results of comparing the classification accuracy and network complexity between the networks, confirming that the network proposed in this study is superior in terms of accuracy, number of multiply–accumulate (MAC) operations, and number of network parameters.

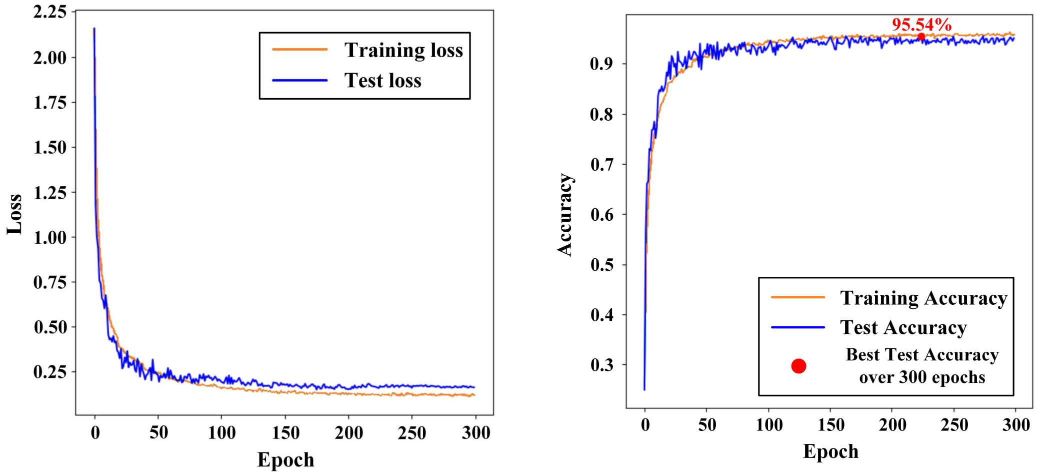
Table 5. Comparison between image classification networks.

Network	Accuracy	#MACs	#Parameters
LeNet5 [28]	94.33%	190.55 M	1.94 M
VGG11 [29]	94.01%	759.10 M	9.35 M
Resnet18 [30]	93.98%	2.37 G	11.23 M
MobileNetV1 [31]	90.57%	66.00 M	3.24 M
Ours	95.54%	25.77 M	22.28 K

4.2. Evaluation and Analysis

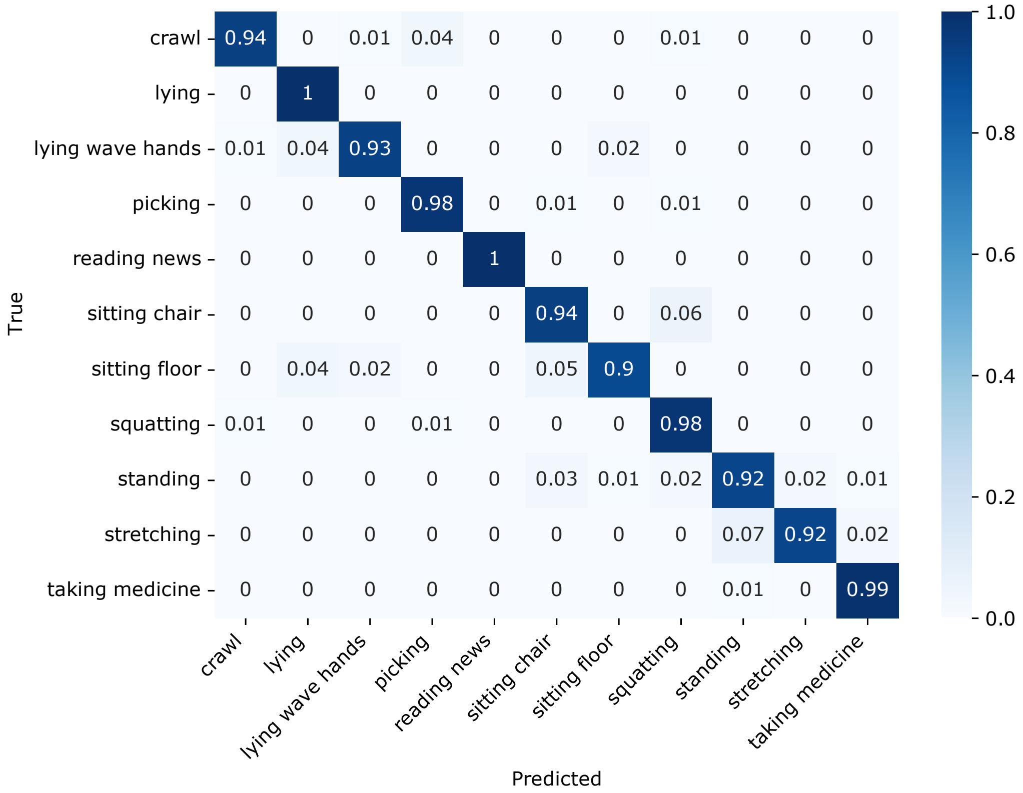
The loss curve and accuracy curve for training and testing after 300 epochs of training are shown in Figure 5, where NLL loss is used as the loss function. The loss decreased significantly in the early epochs of training and tended to decrease moderately after approximately 50 epochs. Similarly, the accuracy increased significantly in the early epochs of training and tended to increase and maintain a moderate rate after approximately 50 epochs.

The confusion matrix for the HAR system using our dataset is shown in Figure 6. The sitting floor class tended to be confused with the lying, lying wave hands, and sitting chair classes. This is because the sitting floor class is similar to the sitting chair class in terms of sitting motion and is similar to the lying class in terms of being close to the ground.



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

Figure 5. Training and test loss curve and accuracy curve: (**a**) Training and test loss curve; (**b**) Training and test accuracy curve.



	crawl	lying	lying wave hands	picking	reading news	sitting chair	sitting floor	squatting	standing	stretching	taking medicine
crawl	0.94	0	0.01	0.04	0	0	0	0.01	0	0	0
lying	0	1	0	0	0	0	0	0	0	0	0
lying wave hands	0.01	0.04	0.93	0	0	0	0.02	0	0	0	0
picking	0	0	0	0.98	0	0.01	0	0.01	0	0	0
reading news	0	0	0	0	1	0	0	0	0	0	0
sitting chair	0	0	0	0	0	0.94	0	0.06	0	0	0
sitting floor	0	0.04	0.02	0	0	0.05	0.9	0	0	0	0
squatting	0.01	0	0	0.01	0	0	0	0.98	0	0	0
standing	0	0	0	0	0	0.03	0.01	0.02	0.92	0.02	0.01
stretching	0	0	0	0	0	0	0	0	0.07	0.92	0.02
taking medicine	0	0	0	0	0	0	0	0	0.01	0	0.99

Figure 6. Confusion matrix.

4.3. Performance Comparison by Quantization Bit Formats

A lightweight model design and a low-power design are essential for the practical use of HAR systems. For this purpose, the proposed HAR system was implemented on the Xilinx Zynq UltraScale+ ZCU104 (Xilinx, Santa Clara, CA, USA) for a low-power design. To implement a classification network on the FPGA part of the ZCU104 platform, a register transfer level (RTL) design is required, which includes the process of converting the system implemented in a floating-point format to a fixed-point format. In RTL design, the more bits that express a number, the larger the range of numbers that can be expressed; this improves accuracy but requires a large memory for system storage and operation. To find a tradeoff between accuracy and memory usage, we compared the HAR accuracy based on the bit formats of the classification network, as shown in Table 6. The Brevitas library provided by Xilinx was used to quantize the input data, weight, bias, and activation functions [32]. As shown in Table 6, the accuracy decreased by less than 1% from 32 bits to 4 bits; however, at 2 bits the accuracy degradation was approximately 13% compared with other bit formats. Therefore, 4 bit quantization was selected as the best tradeoff between accuracy and memory usage for the classification network and was implemented on the FPGA.

Table 6. Comparison of the classification network accuracy by bit format.

Classification Network Bit Format	Accuracy
32 bits	95.54%
16 bits	95.48%
8 bits	95.46%
4 bits	94.79%
2 bits	82.21%

4.4. Hardware–Software Implementation

FINN is a deep-learning framework developed by the Integrated Communications and AI Lab of AMD Research & Advanced Development. Using FINN, we optimized a classification network configuration based on the Xilinx library and generated the RTL code and IP of the optimized network. Our implementation of the HAR hardware–software performs data pre-processing and PFE on a microprocessor and deploys a classification network IP created using FINN on an FPGA. Figure 7 shows the environment in which the classification network was implemented on an FPGA using FINN, and where the HAR system was validated with a hardware–software implementation.

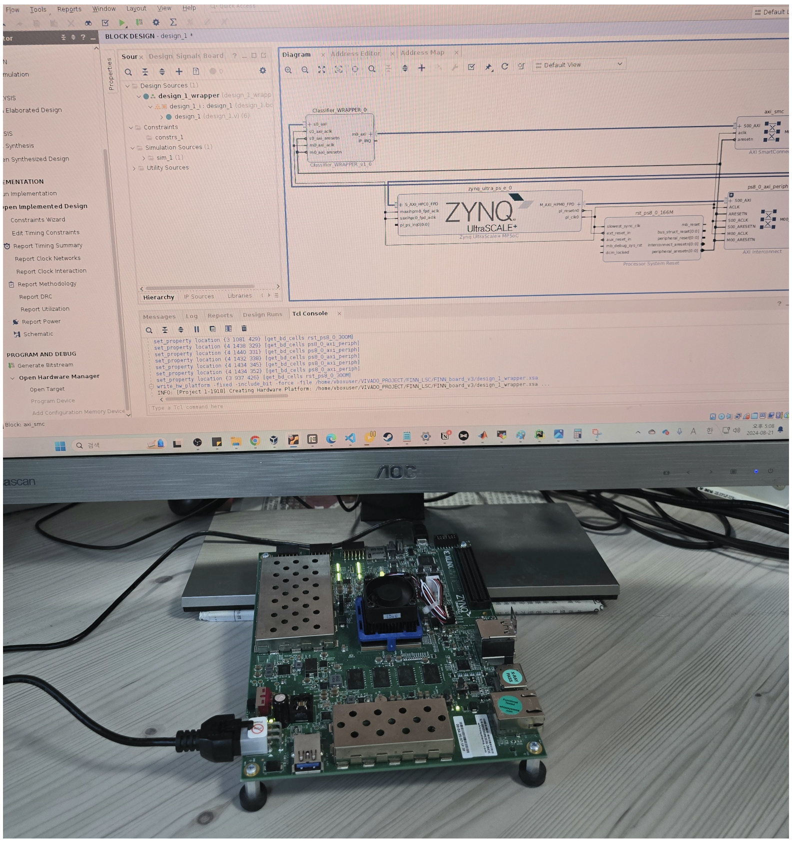


Figure 7. Environment used for FPGA implementation and verification.

The specifications of the HAR system with the hardware–software implementation is shown in Table 7. The platform used was an Xilinx Zynq UltraScale+ ZCU104 with an ARM Cortex A53 processor (ARM, Cambridge, England, UK) operating at 1.2 GHz, and the execution time of the HAR system in this environment was 2.43 ms. The proposed HAR system utilized 29,720 configurable logic block (CLB) look-up tables (LUTs), 22,893 CLB registers, 72 digital signal processors (DSPs), and 25.5 block RAMs. The maximum operating frequency was 300 MHz, and the power consumption was 3.479 W, making it a lightweight HAR system that runs at low power on edge devices.

Table 7. Specifications of the HAR system implementation on the Zynq UltraScale+ ZCU104.

Parameter	Proposed System
Platform	ZCU104
Execution time	2.43 ms
CLB LUTs	29,720
CLB Registers	22,893
DSPs	72
Block RAMs	25.5
Frequency	300 MHz
Power	3.479 W

4.5. Performance Comparison

Table 8 shows the results of the performance comparison of the HAR system proposed in this paper with those of other radar point-cloud-based HAR systems. The comparison was conducted based on the input feature size of the classification network, number of data classes, accuracy, number of network parameters, the device that implemented the system, and the power consumption. The proposed HAR system distinguished 11 classes, the most, and achieved 94.79% accuracy by performing 4 bit quantization. It also used 22.28 K parameters in the network, the smallest number of parameters. In terms of power consumption, the RTX3080 consumed 320 W, the RTX2060 consumed 175 W, and the Raspberry Pi consumed 3.6 W. However, the proposed HAR system using the ZCU104 platform consumed 3.479 W of power, which is up to 92 times less than using RTX3080. Compared to other HAR systems, the proposed HAR system achieved the lowest power consumption, indicating that we have achieved a low-power design.

Table 8. Performance of the reference HAR system versus the proposed HAR system.

Method	Feature Size	#Data Classes	Accuracy	#Parameters	Platform	Power
[11]	10 × 32 × 32	5	90.47%	291 K	-	-
[12]	3 × 224 × 224	7	96.10%	1510 K	-	-
[13]	10 × 32 × 32	6	90.20%	131.62 K	GTX3080	320 W
[14]	9 × 400 × 3	6	94.53%	1610 K	RTX2060	175 W
[15]	2 × 25 × 3	5	95.12%	79.7 K	Raspberry Pi	3.6 W
Ours	64 × 64 × 64	11	94.79%	22.28 K	ZCU104	3.479 W

5. Conclusions

In this study, we proposed a low-power and lightweight HAR system using radar point-cloud data. The proposed HAR system consists of pre-processing 3D point data with four channels and then inputting the converted 2D image through a PFE to the image classification network. For the HAR technology to be practical, a lightweight and low-power design must be used. In this paper, PFE is used as an encoder for point-cloud data, and an image classification network based on depth-wise separable convolution is used to realize a lightweight HAR system. Also, the 4 bit quantized classification network was implemented on an FPGA to achieve a low-power design. The 32 bit format classification network was quantized to a 4 bit one, resulting in a lightweight design that used 12.5% of the memory of the original 32 bit format network. As a result of the implementation, in terms of network complexity, 25.77 M MAC operations were performed and 22.28 K parameters were used, which is the smallest among the compared HAR systems. An accuracy of 95.54% was achieved with a 32 bit floating-point data format in a GPU environment and 94.79% with a 4 bit fixed-point data format in a hardware–software implementation environment for 11 classes of datasets collected using the FMCW MIMO millimeter-wave radar. In terms of power consumption, the proposed HAR system consumed 3.479 W, which is the lowest power consumption compared to other HAR systems. The results show that the proposed HAR system has a level of accuracy similar to other HAR systems, has relatively low complexity, and can operate at low power. In future work, we plan to implement a low-power, lightweight HAR system on a very-large-scale integrated circuit (VLSI) to achieve a better low-power, lightweight design while maintaining current accuracy.

Author Contributions: S.L. (Seungchan Lim) designed and implemented the proposed HAR system, performed the experiments and evaluation, and wrote the paper. C.P. evaluated the proposed HAR system and revised the manuscript. S.L. (Seongjoo Lee) evaluated the proposed HAR system and revised the manuscript. Y.J. conceived of and led the research, analyzed the experimental results, and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Technology Innovation Program (Nos. 00144290 and 00433615) and funded by the Ministry of Trade, Industry, and Energy (MOTIE, Korea); the CAD tools were supported by IDEC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

References

1. Dirgová Luptáková, I.; Kubovčík, M.; Pospíchal, J. Wearable sensor-based human activity recognition with transformer model. *Sensors* **2022**, *22*, 1911. [CrossRef]
2. Zhang, S.; Li, Y.; Zhang, S.; Shahabi, F.; Xia, S.; Deng, Y.; Alshurafa, N. Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors* **2022**, *22*, 1476. [CrossRef]
3. Prasad, A.; Tyagi, A.K.; Althobaiti, M.M.; Almulihi, A.; Mansour, R.F.; Mahmoud, A.M. Human activity recognition using cell phone-based accelerometer and convolutional neural network. *Appl. Sci.* **2021**, *11*, 12099. [CrossRef]
4. Alrashdi, I.; Siddiqi, M.H.; Alhwaiti, Y.; Alruwaili, M.; Azad, M. Maximum entropy Markov model for human activity recognition using depth camera. *IEEE Access* **2021**, *9*, 160635–160645. [CrossRef]
5. Song, K.T.; Chen, W.J. Human activity recognition using a mobile camera. In Proceedings of the 2011 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Incheon, Republic of Korea, 23–26 November 2011; IEEE: New York, NY, USA, 2011; pp. 3–8.
6. Jalal, A.; Kamal, S.; Kim, D. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensors* **2014**, *14*, 11735–11759. [CrossRef] [PubMed]
7. Roche, J.; De-Silva, V.; Hook, J.; Moencks, M.; Kondoz, A. A multimodal data processing system for LiDAR-based human activity recognition. *IEEE Trans. Cybern.* **2021**, *52*, 10027–10040. [CrossRef] [PubMed]
8. Ghosh, A.; Chakraborty, A.; Chakraborty, D.; Saha, M.; Saha, S. UltraSense: A non-intrusive approach for human activity identification using heterogeneous ultrasonic sensor grid for smart home environment. *J. Ambient. Intell. Humaniz. Comput.* **2023**, pp. 1–22. [CrossRef]
9. Wang, Z.; Hou, Y.; Jiang, K.; Zhang, C.; Dou, W.; Huang, Z.; Guo, Y. A survey on human behavior recognition using smartphone-based ultrasonic signal. *IEEE Access* **2019**, *7*, 100581–100604. [CrossRef]
10. Papadopoulos, K.; Jelali, M. A Comparative Study on Recent Progress of Machine Learning-Based Human Activity Recognition with Radar. *Appl. Sci.* **2023**, *13*, 12728. [CrossRef]
11. Singh, A.D.; Sandha, S.S.; Garcia, L.; Srivastava, M. 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, Los Cabos, Mexico, 25 October 2019; pp. 51–56.
12. Kim, Y.; Alnujaim, I.; Oh, D. Human activity classification based on point clouds measured by millimeter wave MIMO radar with deep recurrent neural networks. *IEEE Sens. J.* **2021**, *21*, 13522–13529. [CrossRef]

13. Huang, Y.; Li, W.; Dou, Z.; Zou, W.; Zhang, A.; Li, Z. Activity recognition based on millimeter-wave radar by fusing point cloud and range–doppler information. *Signals* **2022**, *3*, 266–283. [CrossRef]
14. Ding, C.; Zhang, L.; Chen, H.; Hong, H.; Zhu, X.; Fioranelli, F. Sparsity-based human activity recognition with PointNet using a portable FMCW radar. *IEEE Internet Things J.* **2023**, *10*, 10024–10037. [CrossRef]
15. Gu, Z.; He, X.; Fang, G.; Xu, C.; Xia, F.; Jia, W. Millimeter Wave Radar-based Human Activity Recognition for Healthcare Monitoring Robot. *arXiv* **2024**, arXiv:2405.01882.
16. Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; Liu, Y. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–40. [CrossRef]
17. Ayala, R.; Mohd, T.K. Sensors in autonomous vehicles: A survey. *J. Auton. Veh. Syst.* **2021**, *1*, 031003. [CrossRef]
18. Cha, D.; Jeong, S.; Yoo, M.; Oh, J.; Han, D. Multi-input deep learning based FMCW radar signal classification. *Electronics* **2021**, *10*, 1144. [CrossRef]
19. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12697–12705.
20. Li, Y.; Han, Z.; Xu, H.; Liu, L.; Li, X.; Zhang, K. YOLOv3-lite: A lightweight crack detection network for aircraft structure based on depthwise separable convolutions. *Appl. Sci.* **2019**, *9*, 3781. [CrossRef]
21. Xilinx. Xilinx FINN. Available online: https://xilinx.github.io/finn (accessed on 13 September 2024).
22. Smart Radar System. Smart Radar System RETINA-4SN. Available online: https://www.smartradarsystem.com/en/products/retina_4s.html (accessed on 13 September 2024).
23. Xilinx. UltraSclae+ ZCU104. Available online: https://www.xilinx.com/products/boards-and-kits/zcu104.html#overview (accessed on 13 September 2024).
24. Sola, J.; Sevilla, J. Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Trans. Nucl. Sci.* **1997**, *44*, 1464–1468. [CrossRef]
25. Zhou, Y.; Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3D object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.
26. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3D shape recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 945–953.
27. NVIDIA. Jetson AGX Xavier Developer Kit | NVIDIA Developer. Available online: https://www.nvidia.com/en-us/design-visualization/rtx-6000 (accessed on 13 September 2024).
28. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
32. Pappalardo, A. Xilinx Brevitas. Available online: https://xilinx.github.io/brevitas (accessed on 13 September 2024).

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