






Article

# Dual-Scale Doppler Attention for Human Identification

Sunjae Yoon <sup>1</sup>, Dahyun Kim <sup>1</sup>, Ji Woo Hong <sup>1</sup>, Junyeong Kim <sup>2</sup> and Chang D. Yoo <sup>1,\*</sup><sup>1</sup> School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea<sup>2</sup> Department of AI, Chung-Ang University, Seoul 06974, Korea

\* Correspondence: cd\_yoo@kaist.ac.kr; Tel.: +82-10-3774-1007

**Abstract:** This paper considers a Deep Convolutional Neural Network (DCNN) with an attention mechanism referred to as Dual-Scale Doppler Attention (DSDA) for human identification given a micro-Doppler (MD) signature induced as input. The MD signature includes unique gait characteristics by different sized body parts moving, as arms and legs move rapidly, while the torso moves slowly. Each person is identified based on his/her unique gait characteristic in the MD signature. DSDA provides attention at different time-frequency resolutions to cater to different MD components composed of both fast-varying and steady. Through this, DSDA can capture the unique gait characteristic of each person used for human identification. We demonstrate the validity of DSDA on a recently published benchmark dataset, IDRad. The empirical results show that the proposed DSDA outperforms previous methods, using a qualitative analysis interpretability on MD signatures.

**Keywords:** deep learning; human identification; micro-Doppler radar; fine-grained feature analysis



**Citation:** Yoon, S.; Kim, D.; Hong, J.W.; Kim, J.; Yoo, C.D. Dual-Scale Doppler Attention for Human Identification. *Sensors* **2022**, *22*, 6363. <https://doi.org/10.3390/s22176363>

Academic Editors: Moulay A. Akhloufi and Mozhddeh Shahbazi

Received: 26 July 2022

Accepted: 20 August 2022

Published: 24 August 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



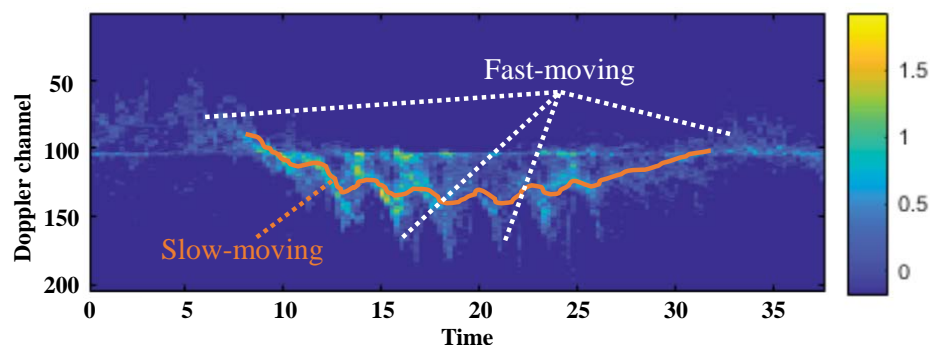
**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Human Identification (HI) serves as an essential building block for many personal identification services including surveillance, security and identification systems. In general, HI adopts visual video as it has information that can be easily understood. However, HI through visual video can be problematic under low lighting conditions and with the privacy infringement issues [1]. As an alternative to the use of cameras, radar devices bypass these problems. Radar is a detection system that measures the distance, direction, angle, and speed of a target by emitting electromagnetic waves from a device and analyzing electromagnetic waves that are reflected and returned from an object. Radar can operate under low light conditions, and its tendency to bend around obstacles makes it suitable for identification in obscured environments [2–4]. Moreover, radar information is relatively safe regarding privacy, as people cannot directly interpret information obtained by radar. Furthermore, radar also has the advantage of observing far-distance targets. Overall, radar is a more robust sensor for HI than visual video. However, since it is difficult to observe the iris, voice, and face using radar, this paper adopts gait instead of aforementioned biometrics. Gait can be observed from a distance, and unlike the biometrics (iris, voice, and face), it is behavior over a certain period of time, so the security is relatively high. These advantages are consistent with the reason for using radar-based HI instead of video-based HI.

For this radar-based HI, the micro-Doppler (MD) signature has been a popular choice. As shown in Figure 1, it sequentially records Doppler effects in electromagnetic signals of moving targets [5], and the superposition of these Doppler signals is summarized to make MD signatures, where it holds granularity to specify their information. In the methodology, conventional machine learning (ML) algorithms have attempted to analyze the statistics of the MD signatures [6–9]. Unfortunately, ML has several drawbacks as it is based on heuristic feature extraction and a low capacity model. Recent Deep Convolutional Neural Network (DCNN), via capturing spatial relationships and comprehending high-level spatial features, overcomes these limitations and has revolutionized many applications including radar-based human identification and motion recognition. Motion recognition is easily

conducted by training DCNN to recognize the common patterns among identical motions. However, human identification [10,11] should identify the unique characteristics of each person in an uncontrolled scenario, which requires more fine-grained understanding than motion recognition.



**Figure 1.** Micro-Doppler signature of human walking. When a person walks, the arms and legs generate fast-moving signatures represented by a white line, and the torso generates slow-moving signatures indicated by an orange line.

For these fine-grained understandings, we exploit that MD signatures from different body parts such as arms, legs, and torso display different signal characteristics. Each of us has a unique gait characteristic distinguished by different sized body parts moving and swinging in a distinctive pattern, as the arms and legs move rapidly, while the torso moves slowly. Our empirical analysis validates that more than 95% of MD signatures on the radar dataset (Validation is performed on the IDRad dataset) include this distinguishable gait characteristic. Therefore, we utilize these unique gait characteristics from the MD signatures to identify humans. For the detailed explanation of the gait characteristics of MD signatures, as shown in Figure 1, MD signatures have fast and slow-moving components. The MD signatures recorded the Doppler effects along the time, where the orange curve shows a small amplitude wave denoting slow-moving features, and the signal that spreads around the orange curve makes a high-amplitude wave representing fast-moving features.

Under this observation, our proposed Dual-Scale Doppler Attention (DSDA) is composed of Temporal Window Attention (TWA), which attends fast-moving MD signatures via building temporal windows for MD signatures, and Holistic Window Attention (HWA), which attends slow-moving MD signatures via building holistic windows. To perform TWA and HWA, we define a common attention module defined as Window Attention (WA), which conducts self-attention among windows by calculating their similarities. In the overall pipeline, TWA extracts fast-moving features by generating multiple temporal windows and applying WA among them. HWA generates slow-moving features by subtracting the attended fast-moving features from original signals, and WA again performs self-attention between subtracted and original signals. With these attended fast and slow-moving features, DSDA extracts unique characteristics of each person in fast and slow moving, and identifies human identity. Using the IDRad [10] dataset, we validate state-of-the-art performances on human identification tasks and show the interpretability of the proposed DSDA.

## 2. Related Works

### 2.1. Doppler Radar Systems for Human Identification

Doppler radar, which uses a single-tone radio wave, has been frequently used for human identification [6,10,12,13]. By the Doppler effect, the received frequency  $f_r$  of the

moving target is shifted away from the transmitted frequency  $f_t$ , and the Doppler frequency is defined by subtracting  $f_t$  from  $f_r$  as:

$$f_r = f_t(1 + v/c)/(1 - v/c), \quad (1)$$

$$f_d = f_r - f_t = 2vf_t/(c - v), \quad (2)$$

where  $c$  is the speed of light, and  $v$  is the radial speed of the moving target. By capturing the Doppler shift, it is able to detect the human motions [5,14,15]. Moreover, frequency-modulated continuous-wave (FMWC) radar is commonly used for short-range multiple targets detection by generating a Doppler map within a certain range [10,16,17]. Vandermissen et al. [10] utilized low-power 77 GHz FMCW radar for person identification and constructed the IDentification with Radar (IDRad) data set. IDRad is a micro-Doppler map received from several people walking around spontaneously in any possible direction. Our proposed DSDA is experimented on IDRad and validates its interpretability on MD signatures.

### 2.2. Deep Learning for Micro-Doppler Signatures

As the use of radar-based systems increases, several applications of MD signatures using deep learning have emerged [18–21]. Kim et al. [22] first applied a neural network for human motion recognition on MD signatures and showed the applicability of deep learning on radar signals. After that, several deep learning techniques have been applied for the radar-based motion recognition, including large-scale pre-training [23] and recurrent neural network [24]. Lin et al. [25] proposed iterative CNN followed with random forests in MD signatures which showed performance boost. Park et al. [23] utilized DCNN pre-trained with a large-scale image classification dataset, ImageNet [26], which presented the connectivity between radar and computer vision. Furthermore, Wang et al. [24] used a recurrent neural network (RNN) to detect dynamic gestures with a short-range radar-based sensor, Google's Soli. Recent studies have been conducted via performing human identification (HI) on an MD signature, which requires the understanding of unique characteristics in a single person. In detail, MD signatures from heartbeat signals are utilized for HI [11,27]. Henceforth, MD signatures on gait characteristics of humans [10] are used for this HI, which is more challenging as they are performed in an uncontrolled scenario where a target is allowed to walk around in a free and spontaneous way. Cao et al. [27] primarily applied DCNN to MD signatures for HI. Vandermissen et al. [10] also used the DCNN and constructed public dataset IDRad for HI, which contributed to subsequent research. Although several methods have been proposed for the aforementioned tasks, they do not fully utilize the details of MD signatures induced from moving human body parts. Therefore, we propose DSDA, which can exclusively recognize unique signals generated by human walking.

### 2.3. Recent Radar-Based Human Identification Analysis

Radar systems have mainly been applied on the radar-based human identifications [10,11,27–29]. We also compare these previous works to our research in terms of the differences, advantages and disadvantages. The work [28] is performed on detecting humans in specific conditions (i.e., short-range through-wall and long-range foliage penetration). The difference between this study and ours is that it was carried out to find people under specific conditions, but our proposed DSDAs are more contributing on human identification from general human behaviors including human walking motions, arm and body movements. In terms of the method, they applied SVM for human detection. However, our study also contain SVM methods and makes more experimental contributions (SVM, CNN, RNN, and Attention model). The advantage of this study, we think, is that it defines specific tasks well and suggests their solutions. However, the disadvantage is that it is too task-specific and reduces its applicability for other research.

The work [29] holds the commonality with our work in that it performs human identification through the recognition of gait characteristics in MD signatures. However, this work mainly focused on open-set feature analysis, which means how the model can do better when the ‘unknown class’ exists in the inference, but our research is mainly about sequential feature pattern analysis; thus, our experimental contributions are more relying on sequential pattern analysis and its solutions. The advantage of the work [29] is novel problem definition for establishing the generality of human identification, but the disadvantage is that the feature analysis is insufficient in that the MD signatures contain sequential information.

The works in [11,27] are also holding commonality with our work in that they perform feature pattern analysis (limbs, torso, heart beat). However, the difference is that our model introduces an attention method to better recognize sequential feature patterns. As shown in Table 1, in the paper, our initial attempts also include the CNN models such as [10,11,27], but we confirmed that the RNN-based model should perform better. Therefore, our further studies are focused on designing sequential feature processing model (RNN and DSDA). The advantages of works [11,27] are specific feature pattern analysis, and we speculate that the disadvantages are a lack of concern for a model that can recognize the feature pattern well because of reliance on the popular CNN model.

**Table 1.** Window Encoder Specifications for Temporal Window Attention.

Layer	Kernel Size	Stride	# of Filters	Data Shape
INPUT				$(150 \times 205 \times 1)$
Temporal Window				$(3 \times 50 \times 205 \times 1)$
Conv 1	(3,3)	(1,1)	8	$(3 \times 50 \times 205 \times 8)$
ELU 1				$(3 \times 50 \times 205 \times 8)$
MAXPool 1	(2,3)	(2,3)		$(3 \times 25 \times 68 \times 8)$
Conv 2	(3,3)	(1,1)	16	$(3 \times 25 \times 68 \times 16)$
ELU 2				$(3 \times 25 \times 68 \times 16)$
MAXPool 2	(2,3)	(2,3)		$(3 \times 12 \times 22 \times 16)$
Conv 3	(3,3)	(1,1)	32	$(3 \times 12 \times 22 \times 32)$
ELU 3				$(3 \times 12 \times 22 \times 32)$
MAXPool 3	(3,1)	(3,1)		$(3 \times 4 \times 22 \times 32)$
Conv 4	(3,3)	(1,1)	64	$(3 \times 4 \times 22 \times 64)$
ELU 4				$(3 \times 4 \times 22 \times 64)$
MAXPool 4	(1,5)	(1,5)		$(3 \times 4 \times 4 \times 64)$
Pooling				$(3 \times 1024)$

For the work [10], our proposed DSDA is validated on the same dataset (IDRad) released in [10]. However, the simple convolutional neural network used in [10] is limited in understanding sequential patterns in the radar continuity; thus, we more focused on a method that can accurately recognize sequential information of the radar sequence. In this respect, we have performed several experimental contributions including Recurrent Neural Network models and an Attention model. Finally, we proposed a Dual-Scale Doppler Attention technique for recognizing fast and slow-moving patterns that are prominently present in continuous signals, where we speculate that this makes the methodological differences comparing to the work [10]. The advantages in the work [10] exist in the contributions from dataset release and task proposal, but for the disadvantages, we guess that there is a lack of contribution on how to better understand the features pattern in sequential MD signatures.

### 3. Method

#### 3.1. Generating Micro-Doppler Signatures

Our proposed Dual-Scale Doppler Attention (DSDA) is validated on the Micro-Doppler signatures (MD signatures); we first explain how to generate the MD signatures. To make 45 time stamps of MD signatures, we first build a single time stamp MD signature and connect 45 stamps in a row. As shown in Figure 2, the single time stamp MD signature is obtained from a single range-Doppler map. We integrate the range-Doppler map (e.g., the dimension for the range-Doppler map is  $256 \times R$ , where 256 is the Doppler channel and  $R$  is the number of discrete range bins) along the range axis, which constructs the single MD signature (e.g.,  $256 \times 1$ ).

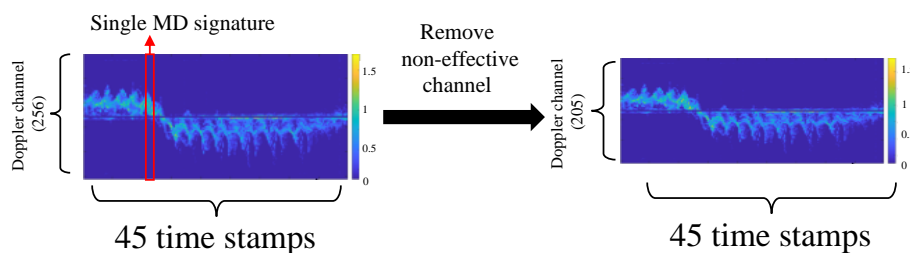


Figure 2. Illustration of Generating Micro-Doppler signature for Dual-Scale Doppler Attention.

This single MD signature contains the Doppler-shift information, where this Doppler-shift value denotes whether the target is moving closer or farther away (e.g., 129–256 channels contain the Doppler shift by the target moving closer and 1–128 channels contain the Doppler shift by the target moving farther away). Following the [10], some channels (i.e., 127–129 channels are zero-Doppler effective, and Doppler channels at both ends are too noisy) are not helpful to identify the human. We also remove them and finally can generate  $205 \times 45$  time stamps MD signatures.

#### 3.2. Model Overview

MD signatures are provided for identifying the human identity [10,11,27–31]. DSDA takes MD signature  $R \in \mathbb{R}^{C \times T}$  consisting of the number of channels  $C$  and the number of time stamps  $T$  ( $C = 205, T = 45$ ). The time stamp can be modulated according to the size of MD signatures. As shown in Figure 3, the MD signature is a time-evolving sequence representing the micro-Doppler effect of a person moving.

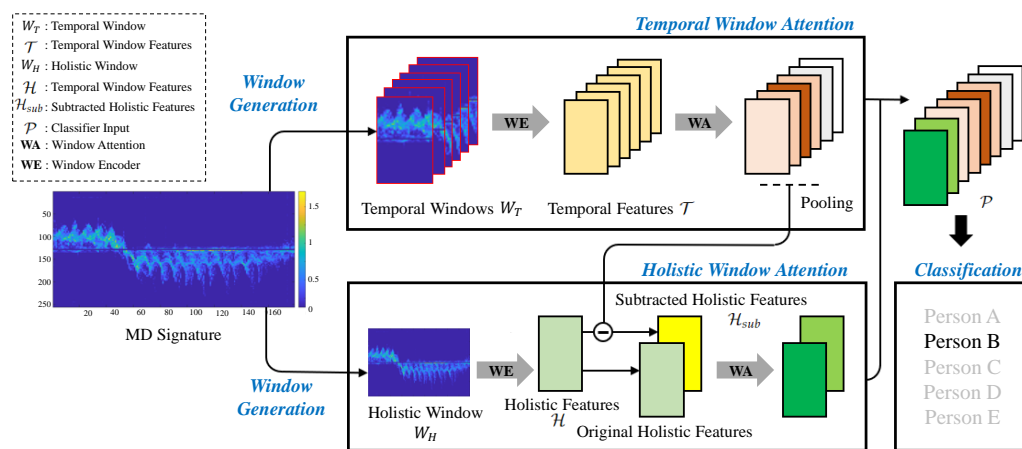


Figure 3. Illustration of Dual-Scale Doppler Attention (DSDA) for human identification. DSDA is composed of the following components: (1) Window Generation for designing temporal window and holistic window, (2) Temporal Window Attention for observing fast-moving features of arms and legs, (3) Holistic Window Attention for recognizing slow-moving features of the torso, and (4) Classification for classifying final targets.

Figure 3 gives a schematic of DSDA consisting of Window Generation, Temporal Window Attention (TWA) and Holistic Window Attention (HWA). Window Generation provides dual-scale windows: a temporal window and holistic window. The temporal window  $W_T$  is focused on recognizing fast-moving features extracted by MD signatures so that it is uniformly divided with a temporal sliding window of stride  $S$ . The temporal window size is  $C \times L$ ; as such, the total number of temporal windows is  $N$  ( $S = 5$ ,  $L = 25$  and  $N = 5$ ). The holistic window  $W_H$  is focused on recognizing the slow-moving signal and extracted by an original MD signature. For the Window Encoder (WE), it embeds temporal windows and holistic window into  $d$ -dimensional feature representation through a series of Conv-ELU-MaxPooling layers. The encoded temporal features  $\mathcal{T}$  and holistic features  $\mathcal{H}$  are defined as:

$$\mathcal{T} = \text{LN}(\text{MaxPool}(\text{ELU}(\text{Conv}(W_T)))) \in \mathbb{R}^{N \times d}, \quad (3)$$

$$\mathcal{H} = \text{LN}(\text{MaxPool}(\text{ELU}(\text{Conv}(W_H)))) \in \mathbb{R}^{1 \times d}, \quad (4)$$

where LN denotes the layer normalization [32] and the Exponential Linear Unit (ELU) [33] is nonlinearity operation. We present detailed specifications of the Window Encoder used in TWA and HWA in Tables 1 and 2. The number of temporal windows is considered as  $N = 3$  in Table 1 (i.e., this is designed to help understand the Window Encoder's structure used in the Temporal Window Attention). The following sub-sections will explain the details of the remaining model components.

**Table 2.** Window Encoder Specifications for Holistic Window Attention.

Layer	Kernel Size	Stride	# of Filters	Data Shape
INPUT				$(150 \times 205 \times 1)$
Holistic Window				$(150 \times 205 \times 1)$
Conv 1	(3,3)	(1,1)	8	$(150 \times 205 \times 8)$
ELU 1				$(150 \times 205 \times 8)$
MAXPool 1	(6,3)	(6,3)		$(25 \times 68 \times 8)$
Conv 2	(3,3)	(1,1)	16	$(25 \times 68 \times 16)$
ELU 2				$(25 \times 68 \times 16)$
MAXPool 2	(2,3)	(2,3)		$(12 \times 22 \times 16)$
Conv 3	(3,3)	(1,1)	32	$(12 \times 22 \times 32)$
ELU 3				$(12 \times 22 \times 32)$
MAXPool 3	(3,1)	(3,1)		$(4 \times 22 \times 32)$
Conv 4	(3,3)	(1,1)	64	$(4 \times 22 \times 64)$
ELU 4				$(4 \times 22 \times 64)$
MAXPool 4	(1,5)	(1,5)		$(4 \times 4 \times 64)$
Pooling				$(1 \times 1024)$

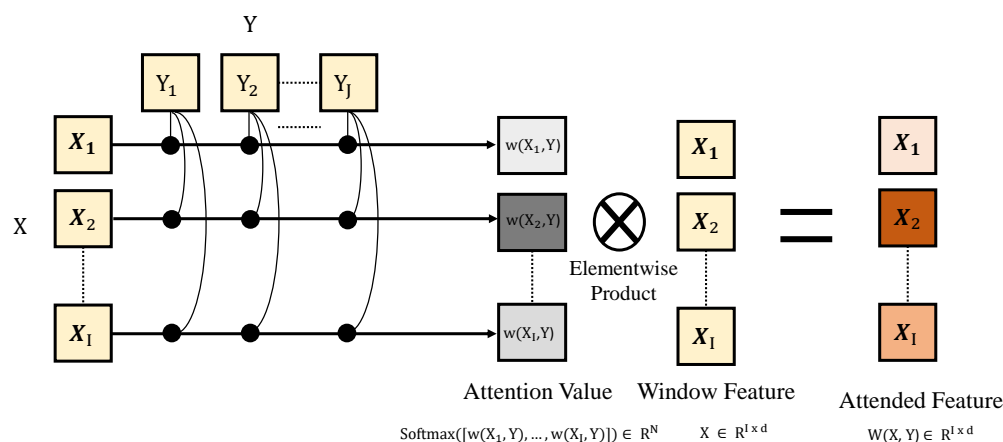
### 3.3. Window Attention

Our proposed basic attention unit of the TWA and HWA is referred to as Window Attention (WA), which calculates the attention weights among multi-element features. Given  $X = [\mathbf{x}_1 \dots \mathbf{x}_I]^T \in \mathbb{R}^{I \times d}$  and  $Y = [\mathbf{y}_1 \dots \mathbf{y}_J]^T \in \mathbb{R}^{J \times d}$ , the WA  $W(X, Y) : \mathbb{R}^{I \times d} \times \mathbb{R}^{J \times d} \rightarrow \mathbb{R}^{I \times d}$  attends  $X$  using  $Y$ , which is defined as follows:

$$w(\mathbf{x}_i, Y) = \sum_{j=1}^J (\text{ELU}(W_1 \mathbf{x}_i) \text{ELU}(W_2 \mathbf{y}_j)^T), \quad (5)$$

$$W(X, Y) = \text{Softmax}([w(\mathbf{x}_1, Y) \dots w(\mathbf{x}_I, Y)])X, \quad (6)$$

where  $w(x_i, Y) : \mathbb{R}^d \times \mathbb{R}^{J \times d} \rightarrow \mathbb{R}$  and  $W_1, W_2 \in \mathbb{R}^{d \times d}$  are learnable parameters. (Our experiments also validate the attention method in the Section 4.4) Using the ELU nonlinearity, WA tries to preserve common signatures between input  $X$  and  $Y$  windows. Based on the WA, the following TWA and HWA are defined and attend their moving signatures. We also provide the process of Window Attention in Figure 4, where WA is performed on two window features. For the TWA, the two types of window are temporal windows and, for the HWA, the two types are holistic windows.



**Figure 4.** Illustration of Window Attention. The WA is performed on TWA and HWA according to the window features.

### 3.4. Temporal Window Attention

We observe that the fast-moving features of arms and legs are unique depending on the person’s height and walking pattern. To preserve and highlight these features, the Temporal Window Attention (TWA) takes multiple temporal windows as inputs and applies WA as follows:

$$\mathcal{T} \leftarrow \text{LN}(W(\mathcal{T}, \mathcal{T}) + \mathcal{T}), \tag{7}$$

$$\mathcal{T}_{avg} = \text{AvgPool}(\mathcal{T}) \in \mathbb{R}^{1 \times d}. \tag{8}$$

where AvgPool() is the average pooling function. TWA produces two types of features. One is the original TWA output  $\mathcal{T}$  in Equation (7), which contributes to classifying the person identification, and the other is the average of  $\mathcal{T}$  over  $N$  temporal windows  $\mathcal{T}_{avg}$  in Equation (8). This average of temporal window attended features  $\mathcal{T}_{avg}$  is treated to have the characteristic of the overall fast-moving features. Furthermore,  $\mathcal{T}_{avg}$  is used to remove fast-moving information from the original holistic features in the following Holistic Window Attention.

### 3.5. Holistic Window Attention

The Holistic Window Attention (HWA) is designed to recognize slow-moving signatures such as the torso. Different from TWA, HWA takes two types of features as input. One is original holistic features  $\mathcal{H}$ , and the other is subtracted holistic features  $\mathcal{H}_{sub}$  obtained by subtracting the  $\mathcal{T}_{avg}$  from  $\mathcal{H}$  in Equation (9). Semantically, the  $\mathcal{H}$  can include slow and fast-moving features, and the  $\mathcal{H}_{sub}$  includes only the slow-moving features by removing characteristics of overall fast-moving features. These two holistic features are concatenated, and HWA attends the common slow-moving signals using the WA in Equations (10) and (11):

$$\mathcal{H}_{sub} = \mathcal{H} - \mathcal{T}_{avg}, \tag{9}$$

$$\mathcal{H}^* = [\mathcal{H} \mathcal{H}_{sub}] \in \mathbb{R}^{2 \times d}, \tag{10}$$

$$\mathcal{H}^* \leftarrow \text{LN}(W(\mathcal{H}^*, \mathcal{H}^*) + \mathcal{H}^*). \tag{11}$$

where  $[\cdot]$  is the concatenation operation. Holistic features  $\mathcal{H}^*$  through HWA represent the slow-moving feature in the MD signature and are utilized to classify the targets.

### 3.6. Classification

In classification, we use two aforementioned temporal and holistic features for classifying targets. These two attended temporal and holistic features are concatenated and transmitted to the final classifier as follows:

$$\mathcal{P} = [\mathcal{T}\mathcal{H}^*] \in \mathbb{R}^{(N+2) \times d}, \quad (12)$$

$$\mathcal{P}^* = \text{Flat}(\mathcal{P}) \in \mathbb{R}^{((N+2) \times d)}, \quad (13)$$

$$\mathcal{C} = W_c(\text{Dropout}(\text{ELU}(W_p\mathcal{P}^*))) \in \mathbb{R}^{C_T}, \quad (14)$$

$$p = (\text{Softmax}(\mathcal{C})) \quad (15)$$

where  $\text{Flat}()$  is a function that converts 2D data into 1D in order to apply the CNN data type to the fully connected neural network, Equation (14) is the last Dropout–ELU–Linear block in the classifier,  $W_p \in \mathbb{R}^{128 \times ((N+2) \times d)}$  and  $W_c \in \mathbb{R}^{C_T \times 128}$  are the learnable parameters,  $C_T = 5$  (IDRad dataset contain five different targets made by five different people.) is the number of targets and  $p = \{p[1], p[2], p[3], p[4], p[5]\}$  is inferred prediction distribution values. In the inference, prediction is performed using argmax function on the  $p$ . We also add some specifications of the classifier network in Table 3. The  $(3 + 2)$  in Table 3 denotes concatenation between TWA and HWA, where HWA includes  $(2 \times 1024)$  features from the holistic feature  $(1 \times 1024)$  in Table 2 and the subtracted feature  $(1 \times 1024)$ . TWA includes  $(3 \times 1024)$  features in Table 3.

**Table 3.** Classifier detailed specifications.

Layer	Kernel SIZE	Stride	# of Filters	Data Shape
INPUT				$((3 + 2) \times 1024)$
Pooling	(3,3)	(1,1)	8	(5120)
Linear				(128)
ELU	(3,3)	(1,1)	16	
Dropout 4	(1,5)	(1,5)		
Linear				5

### 3.7. Training Loss

The entire model is trained in an end-to-end manner using cross-entropy loss  $\mathcal{L}_{HI}(y, p)$ , where the  $y$  is the ground-truth target information and the  $p$  is the prediction of human identification from DSDA, as shown below:

$$\mathcal{L}_{HI}(y, p) = -\log(p[y]) \quad (16)$$

## 4. Experiments

### 4.1. IDRad Dataset

As the MD signature dataset for HI, we use IDRad [10] using FMCW radar, which records range-Doppler maps with a speed of around 15 FPS. The IDRad dataset contains 95,650 frames of 20 min for a training set and 22,535 frames of 5 min for a test set. One frame contains a Doppler frequency channel along the range axis. To construct micro-Doppler signatures, the IDRad dataset integrates a range-Doppler map along the range axis and connects the integrated Doppler signals in temporal order to make Doppler-time maps as previously stated [10]. Here, one time stamp includes 256 Doppler channels, and one MD signature is composed of 45 time stamps. For a fair comparison, we also performed the same preprocessing as [10] and used the default input of  $205 \times 45$  MD signatures,



which translates to 205 Doppler channels and 45 time stamps. Regarding the details of preprocessing, among 256 Doppler channels, 127~129 Doppler channels representing static objects are removed and 24 Doppler channels in the top and bottom of the Doppler axis in MD signatures are also removed, because they are a too high speed range for humans to demonstrate. The IDRad dataset films the movements of five subjects whose age ranges are from 23 to 32 years old, weight ranges are from 60 to 90 kg and their height ranges are from 178 to 185 cm. Five subjects are able to move freely within a certain range of filmed room, and they are able to freely walk, run, or stop in various directions.

#### 4.2. Experimental Details

The dimension of the hidden layer is set to  $d = 1024$ . For DSDA, we use three blocks of the Conv-ELU-MaxPooling layer in the Window Encoder and four blocks of the Dropout [34]-ELU-Linear layer in the classifier. Our model can be easily implemented with six layers of convolutional neural networks and trained on NVIDIA TITAN V (12 GB of memory) GPU with an Adam optimizer [35] with  $\beta_1 = 0.9, \beta_2 = 0.98$  and  $\epsilon = 10^{-9}$ . For all experiments, we select the batch size of 64, a dropout rate of 0.2, and the model is trained up to 15 epochs. From this condition, we do not perform any hyperparameter fine tuning.

The overall evaluation of DSDA is performed using error rate, where the error rate is calculated as: 'error rate' =  $100 \times (\text{number of incorrect predictions}) / (\text{total samples})$ . The total sample can be composed of a validation set and test set.

#### 4.3. Experimental Results

Table 4 summarizes the experimental results on the IDRad Dataset where we compare DSDA with several recent methods. Considering the temporal properties, we use 10 s MD signatures as an input (150 time stamps), which is different from the input of 3 s MD signatures (45 time stamps) in [10]. For the fair comparison, we also measured the baseline with the same sizes of the input MD signatures, where the baseline with 10 s MD signatures is reproduced using their public code and reported as 'Baseline (150 time stamp)' in Table 4. The extended input size of the baseline also shows slight effectiveness, but it still remained in the variation of original performances reported in the paper [10]. To validate the deep learning-based model on this task, we first built a PCA model with SVM classifier, which shows the classical classifying performance with a machine learning algorithm. Comparing to the performances of the PCA model, the deep learning-based models (i.e., baseline, LSTM-based model, DSDA) are giving superior performances. Although there are sequential data in the IDRad dataset, there have not been sequential models to perform human identification. Thus, for these sequential radar image data, we also consider a sequential RNN (Recurrent Neural Network) model on this task. We devised a Long Short-Term Memory (LSTM) [36,37] based model. The LSTM model shows better performances compared to the baseline built with a CNN structure, which explains the necessity of sequential understanding for radar data. The recent success of Transformer [38–40], our proposed DSDA, is based on the attention mechanism and also utilizes the specifications in radar (fast-moving and slow-moving features). DSDA achieves state-of-the-art performance against all methods with a large margin. The results indicate that extracting slow and fast-moving features and attending them improve the interpretability of MD signatures.

**Table 4.** Error rate (%) on the IDRad Dataset.

Methods	Validation	Test
Baseline [10]	24.70	21.54
Baseline [10] (reproduced on 150 time stamp)	22.42	20.37
PCA SVM	37.67	32.91
LSTM based model	22.83	18.42
DSDA	<b>10.87</b>	<b>8.65</b>

#### 4.4. Ablation Study

We experiment with several variants of DSDA to measure the effectiveness of the proposed key components. The first block of Table 5 is full DSDA with three multi-scale windows composed of strides  $s = \{5, 10, 20\}$  and windows  $L = \{35, 25, 25\}$ . The second block of Table 5 provides ablation results of TWA and HWA. Since TWA and HWA have influenced performance improvements, both fast-moving features and slow-moving features imply target features. Especially as TWA boosts performance significantly, we can see that unique information, which identifies a person, is inherent in fast-moving features. The third block of Table 5 provides ablation results on the scale of temporal windows in TWA. Window size and stride between windows influence the extraction of the proper characteristic features. We have confirmed that the best performance is achieved when the stride  $s$  is 5 and window size  $L$  is 35 for single fixed window. The fourth block of Table 5 provides ablation results with multi-scale windows. Here, we use multi-scale temporal windows of various sizes for TWA via selecting several windows in the third block of Table 5. The  $M$  is the number of different types of windows. When  $M = 3$ , we select the several combinations of three windows, where the best performances are shown with the windows composed of strides  $s = \{5, 10, 20\}$  and windows  $L = \{35, 25, 25\}$ . We also validate for  $M = 4$  via adding one more window on top of the best-performance condition in  $M = 3$ . If all the results are not improved, then  $M = 3$  is the best-performance condition, so we report the averaging results on  $M = 4$ . To adopt multi-scale temporal windows, TWA is applied to each window scale, and HWA uses several subtracted holistic features obtained by multi-scale temporal windows. We consider that several moving features extracted by multi-scale temporal windows make it suitable to capture a person's unique characteristics.

**Table 5.** Ablation study on model variants of DSDA on the validation split of IDRAd.

Model Variants	Error Rate (%)
Full DSDA	10.87
w/o TWA	24.32
w/o HWA	23.94
w/ stride $s = 10$ , window $L = 25$	12.99
w/ stride $s = 20$ , window $L = 25$	13.71
w/ stride $s = 5$ , window $L = 15$	14.73
w/ stride $s = 10$ , window $L = 15$	14.80
w/ stride $s = 5$ , window $L = 35$	11.93
w/ stride $s = 10$ , window $L = 35$	13.45
Multi-Scale ( $M = 3$ )	10.87
Multi-Scale ( $M = 4$ )	12.79

Table 6 presents the ablation on the attention methods for  $\text{ELU}(W_1 x_i) \text{ELU}(W_2 y_j)^T$  in Equation (5). We experimented with three attention methods that highlight common features. Based on the results, we selected  $A * B$  for the final DSDA. Here, the operation  $*$  denotes concatenation with a  $d$ -dimensional axis. To follow the dimensional condition, we add more of an embedding matrix  $W_{add} \in \mathbb{R}^{d \times 1}$  for  $A + B$  and embedding matrix  $W_{cat} \in \mathbb{R}^{2d \times 1}$  for  $A; B$ . Our empirical experiments for more attention methods are in the variance of  $A + B$  and  $A; B$ , and they do not give further performance gain.

Table 7 represents the Kappa Index Analysis of our proposed DSDA with Baseline [10]. 'TRUE' means 'correct' on the target, and 'FALSE' means 'incorrect' on the target. To calculate the kappa value  $K = (1 - Pa) / (Pa - Pc)$ , where  $Pa$  is the observational probability of agreement and  $Pc$  is the hypothetical expected probability of agreement.  $Pa$  is obtained as  $Pa = (1082 + 122) / 1490 = 0.808$  and  $Pc$  is obtained as  $Pc = (1328 / 1490) \times (1122 / 1490) + (162 / 1490) \times (368 / 1490) = 0.698$ . Thus, the kappa value is obtained as  $K = 0.36$ . Therefore, according to the kappa value analysis, our pro-

posed DSDA decision performance has ‘fair’ strength of agreement with the baseline [10]. We consider this is because DSDA includes improved performance (10.8% error rate) compared to the baseline (24.7% error rate), which contributes to the disagreement with baseline in the case the baseline performs an incorrect decision on the target.

**Table 6.** Comparison of attention methods in WA on the validation split of IDRAd.

Attention Method for Calculating Similarity		Error Rate (%)
$A * B$	$w(x_i, Y) = \sum_{j=1}^J (\text{ELU}(W_1 x_i) \text{ELU}(W_2 y_j)^T)$	10.87
$A + B$	$w(x_i, Y) = \sum_{j=1}^J ([\text{ELU}(W_1 x_i) + \text{ELU}(W_2 y_j)^T] W_{add})$	11.28
$A; B$	$w(x_i, Y) = \sum_{j=1}^J ([\text{ELU}(W_1 x_i); \text{ELU}(W_2 y_j)^T] W_{cat})$	11.02

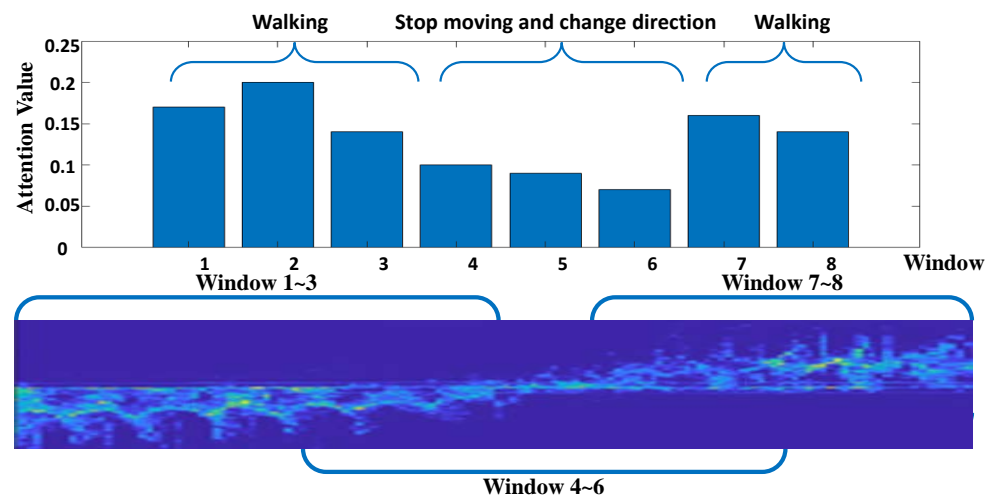
**Table 7.** Kappa Index Analysis of DSDA on the validation split of IDRAd.

Human Identification		Baseline [10]		Total
		True	False	
DSDA	True	1082	246	1328
	False	40	122	162
Total		1122	368	1490

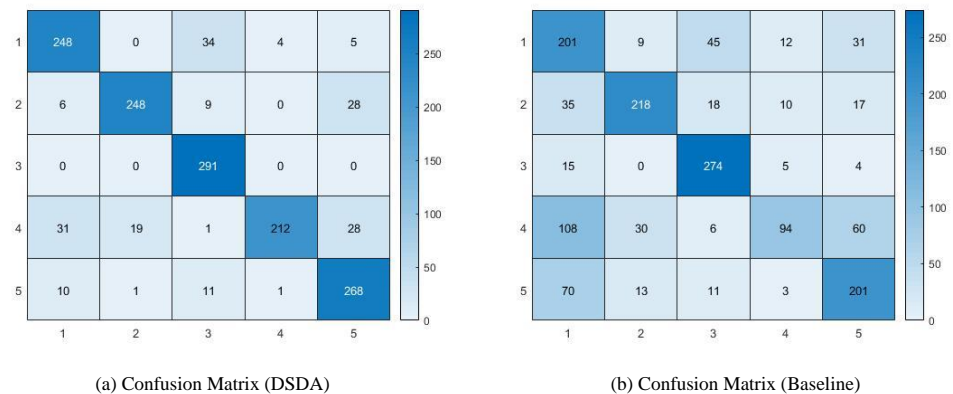
#### 4.5. Qualitative Results

Figure 5 visualizes attention weights for temporal windows. For the input MD signature of 150 time stamps, attention weights are represented as a bar chart in the Figure 5. Here, temporal windows adopt 45 time stamps and 15 strides. Although the multi-scale windows performed well in this human identification task, we select the single window that can help easily understand how the attention weights of TWA are formed and these weights identify the fast-moving information of the MD signatures. Thus, we can find out which windows have been highlighted through the attention weight. The MD signatures corresponding to three windows (i.e., 1, 2, 3) from the left contain fast-moving features, and also, the signatures corresponding to two windows (i.e., 7, 8) from the right contain fast-moving features. Our analysis on synchronized video identifies that these fast-moving features are from the human walking. The other signatures (i.e., MD signatures from windows 4, 5, 6) are formed when the humans stop moving and change the direction of walking. The attention weights in TWA are highlighted on these windows containing fast-moving features such as walking; however, the small weights are given on slow-moving features such as changing direction or non-moving. Therefore, it is confirmed that the TWA is properly trained to recognize the fast-moving information from the MD signatures.

In Figure 6, we build a confusion matrix of DSDA. The vertical axis denotes the ground-truth category and the horizontal axis denotes the predicted category. For all targets, DSDA perform 100% accuracy on target 3 with the highest accuracy and performs 72% accuracy on target 4. We consider the reason why DSDA’s prediction is the lowest on target 4, where the target 4, target 1 and target 5 show similar movements and have a relatively similar body shape. This gives the challenge in identifying their characteristics. To qualitatively compare the confusion matrix with the baseline [10], we give the confusion matrix predicted from the baseline (We obtain confusion matrices from fully trained baseline and DSDA.) in the (b) of Figure 6. Comparing to the baseline, the performance improvement can be confirmed in all targets for our proposed DSDA, and it was confirmed that both the baseline and DSDA showed excellent performance in target 3. However, it is also notable that our DSDA is more improved in target 4, where this is because DSDA is more robust to distinguish fine-grained information in the MD signatures.



**Figure 5.** Attention weights according to temporal windows in TWA. The 8 windows are generated via traversing MD signatures with a sliding window composed of 45 time stamps and 15 strides.



**Figure 6.** (a) Confusion matrix ( $5 \times 5$ ) of human identification from DSDA and (b) Confusion matrix ( $5 \times 5$ ) of human identification from baseline [10] on IDRAd validation split. Vertical axis denotes ground-truth type and horizontal axis denotes predicted type.

In Figure 7, we also perform a target-wise confusion matrix to calculate the sensitivity and specificity according to the targets. The average sensitivity is 0.88, and the average specificity is 0.97. Our DSDA is effective in the specificity, which means our model is highly sensible on the negative targets. It is also notable that DSDA performs 100% on target 3, which implies that the model finds suitable feature space that can classify target 3, and also, the model localizes other target features on the proper feature spaces.

In Figure 8, we perform efficiency analysis. The attention mechanism in DSDA does not require many memory resources (The resources that are required are joint space embedding matrices.) but also performs early saturation of training error rate. As shown in orange curve of Figure 8, the error rate on the IDRAd training dataset converges more early than the blue curve of the baseline, which denotes that the window attention mechanism promotes weights in the network to be sensible on this identification task. After epoch 15, two curves from the baseline and DSDA are converged and become saturated.

In Figure 9, we also perform dataset sensitivity analysis according to the ratio of training dataset usage. Our proposed DSDA gives robustness until 60% usage of the training dataset by keeping the error rate below 30%, and then, the error rate increases. However, the baseline shows the robustness until 80% usage of the training dataset, which shows that our model learns more efficiently on the dataset and is less sensitive to dataset scarcity.

DSDA Target-wise Confusion Matrix																			
Target 1	GT		Target 2	GT		Target 3	GT		Target 4	GT		Target 5	GT						
	T	F		T	F		T	F		T	F		T	F					
DSDA	T	248	43	DSDA	T	248	43	DSDA	T	291	0	DSDA	T	212	79	DSDA	T	268	23
	F	47	1117		F	20	1144		F	55	1109		F	5	1159		F	61	1103
SE	0.841	SP	0.963	SE	0.925	SP	0.964	SE	0.841	SP	1	SE	0.977	SP	0.936	SE	0.815	SP	0.98

Figure 7. Target-wise Confusion Matrix to calculate sensitivity (SE) and specificity (SP) between DSDA and the ground-truth (GT) on the validation split of the IDRad Dataset (T: True, F: False).

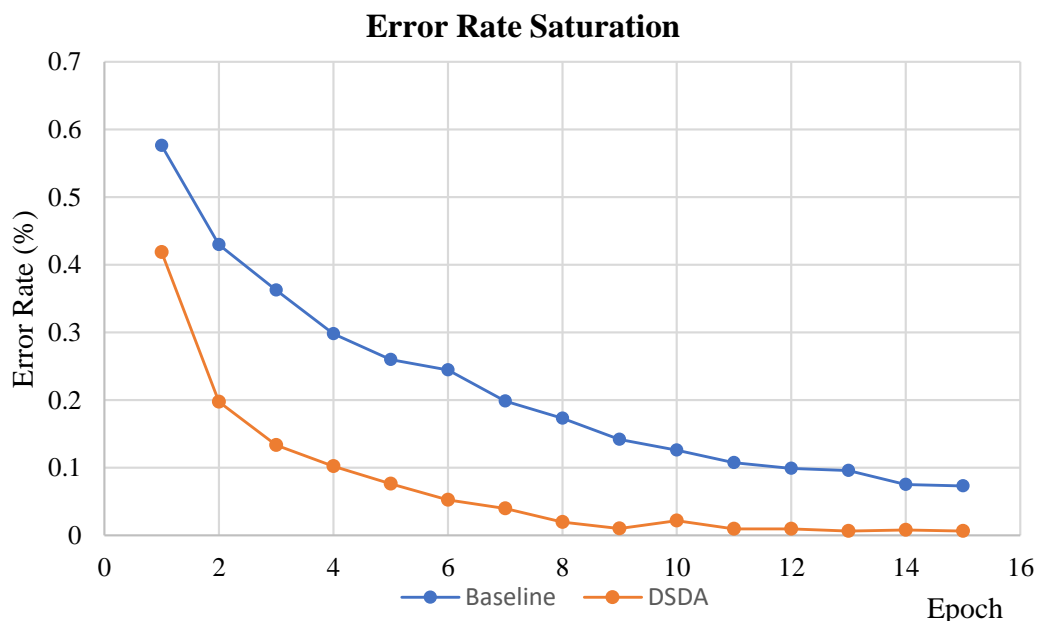


Figure 8. Error rate saturation along the epoch on the training dataset. The blue curve shows the saturation of baseline [10] and the orange curve shows the saturation of the proposed DSDA.

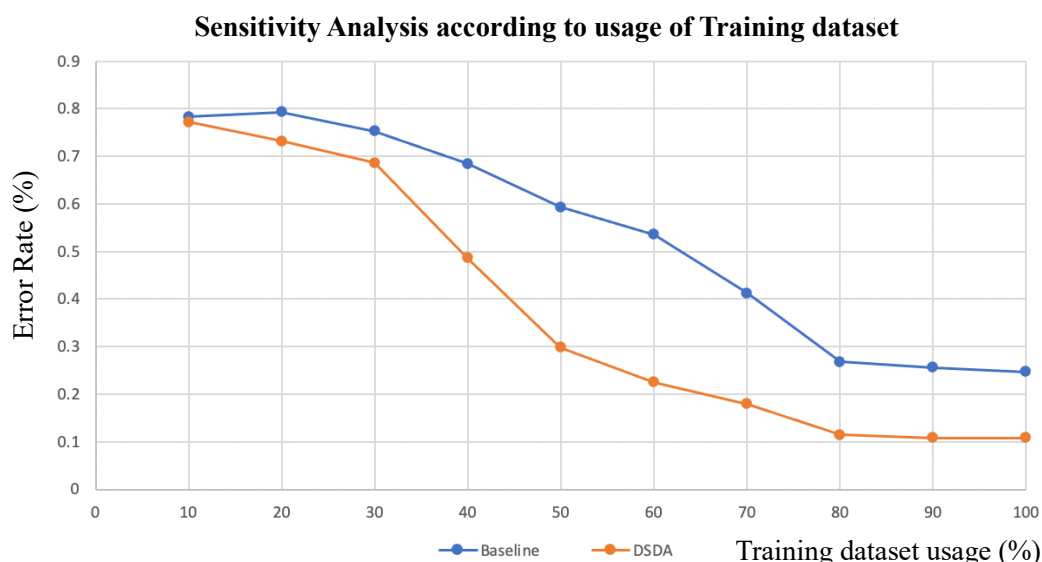
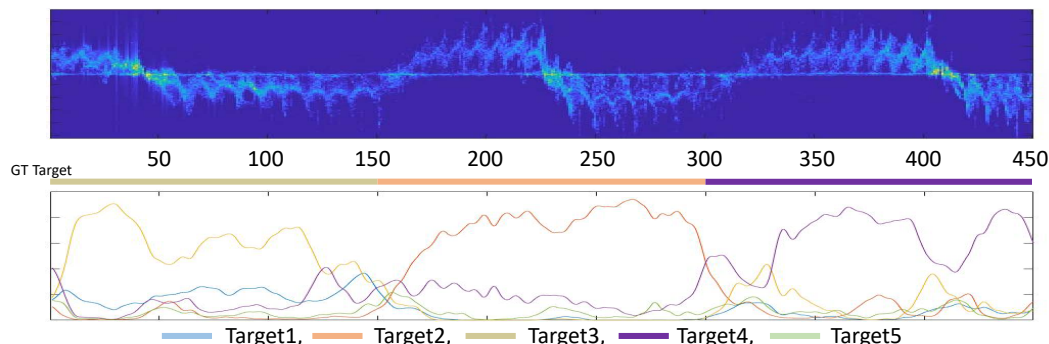


Figure 9. Sensitivity analysis of model performances according to the training dataset usage. The blue curve shows baseline [10], and the orange curve shows the proposed DSDA.

In Figure 10, to verify the practicality of the proposed model, we extended the HI task into a human localization task, where the MD signatures are extended to the longer size (i.e., 450 time stamps), and we perform time stamp wise human identifications. This can be thought of as human localization in MD signatures, which is more challenging and requires

a fine-grained understanding of MD signatures. MD signatures are randomly selected in the validation set. The prediction of each stamp is performed on the center of 150 time stamp windows. DSDA traverses all the MD signatures and predicts every time stamp, which shows its predictions in the below Figure 10, where the y-axis denotes the probability of each target. Thus, five distributions corresponding to five targets are generated, and the above bar shows the ground-truth target for every signature. It is clearly confirmed that the distributions matched with the ground-truth target are highlighted for the given MD signatures. This denotes that DSDA is available to perform fine-grained human identification and is properly extended to the human localization task in radar signals.



**Figure 10.** Extended experiment on MD signatures. The below distribution denotes time stamp wise target predictions, where this represents the availability of the human localization task of DSDA.

## 5. Limitation

We would like to present two limitations from two different perspectives. The first limitation exists in terms of task. Our current experiments are mainly performed under a human identification task. However, as shown in Figure 10 of the paper, we found that this task can be extended up to the human detection task in the radar sequence by stitching MD signatures and localizing the human in them. Our further studies will include this extended analysis and possibilities of human detection. The second limitation is the lack of datasets for this study. We validate DSDA on the IDRAd dataset. To build a general radar-based model, we should also validate more different radar datasets. In this respect, our further research will focus on how to establish model generality on radar signals.

## 6. Future Work

Our future works are three-fold. The first is that we will annotate the radar dataset to be suitable localization tasks. As shown in Figure 10, we further evaluate our DSDA in terms of human identification and localization tasks on the radar dataset, where DSDA shows enough applications on these tasks. The second is that we will modify the window. Currently, our window operates on the sliding window mechanism, but we will consider more diverse windowing methods. The third is that we will validate our DSDA on another radar dataset to be a general attention module for radar understanding.

## 7. Conclusions

We propose Dual-Scale Doppler Attention (DSDA) for human identification. DSDA adopts Temporal Window Attention (TWA) that attends fast-moving MD signatures of arms and legs and Holistic Window Attention (HWA) that attends slow-moving MD signatures of the torso. Our experimental results on the IDRAd dataset show state-of-the-art performance and the qualitative results validate the efficiency of our proposed module. The model analysis including a sensitivity and confusion matrix shows DSDA's robustness. Extended experiments incorporating radar-based human detection tasks show the flexibility of DSDA. From these experiments, our future works are more obvious and substantial.

**Author Contributions:** Conceptualization, S.Y.; methodology, S.Y., D.K. and J.K.; software, S.Y. and D.K.; validation, S.Y. and D.K.; formal analysis, S.Y. and C.D.Y.; writing S.Y., D.K., J.W.H., J.K. and C.D.Y.; supervision, C.D.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partly supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-01381, Development of Causal AI through Video Understanding) and partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00951, Development of Uncertainty-Aware Agents Learning by Asking Questions).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank the experimental contributions from researcher Junehyune Park and Jaewook Park in LIG Nex1, where they perform additional experiments in this paper to validate model generality including target-wise accuracy analysis via confusion matrix and model efficiency analysis.

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

**Sample Availability:** Samples of the compounds ... are available from the authors.

## Abbreviations

The following abbreviations are used in this manuscript:

DCNN	Deep Convolution Neural Network
DSDA	Dual-Scale Doppler Attention
HWA	Holistic Window Attention
IDRad	IDentification with Radar
MD	Micro-Doppler
ML	Machine Learning
TWA	Temporal Window Attention
WE	Window Encoder
WA	Window Attention

## References

1. Cunningham, S.J.; Masoodian, M.; Adams, A. Privacy issues for online personal photograph collections. *J. Theor. Appl. Electron. Commer. Res.* **2010**, *5*, 26–40. [[CrossRef](#)]
2. Kang, D.; Kum, D. Camera and radar sensor fusion for robust vehicle localization via vehicle part localization. *IEEE Access* **2020**, *8*, 75223–75236. [[CrossRef](#)]
3. Bai, J.; Li, S.; Huang, L.; Chen, H. Robust detection and tracking method for moving object based on radar and camera data fusion. *IEEE Sens. J.* **2021**, *21*, 10761–10774. [[CrossRef](#)]
4. Yang, B.; Guo, R.; Liang, M.; Casas, S.; Urtasun, R. Radarnet: Exploiting radar for robust perception of dynamic objects. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020.
5. Chen, V.C.; Li, F.; Ho, S.S.; Wechsler, H. Micro-Doppler effect in radar: Phenomenon, model, and simulation study. *IEEE Trans. Aerosp. Electron. Syst.* **2006**, *42*, 2–21. [[CrossRef](#)]
6. Kim, Y.; Ling, H. Human activity classification based on micro-Doppler signatures using a support vector machine. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1328–1337.
7. Zhou, Z.; Cao, Z.; Pi, Y. Dynamic gesture recognition with a terahertz radar based on range profile sequences and Doppler signatures. *Sensors* **2017**, *18*, 10. [[CrossRef](#)] [[PubMed](#)]
8. Fioranelli, F.; Ritchie, M.; Griffiths, H. Classification of unarmed/armed personnel using the NetRAD multistatic radar for micro-Doppler and singular value decomposition features. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1933–1937. [[CrossRef](#)]
9. Kim, Y.; Ha, S.; Kwon, J. Human detection using Doppler radar based on physical characteristics of targets. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 289–293. [[CrossRef](#)]
10. Versmissen, B.; Knudde, N.; Jalalv, A.; Couckuyt, I.; Bourdoux, A.; De Neve, W.; Dhaene, T. Indoor person identification using a low-power FMCW radar. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3941–3952. [[CrossRef](#)]

11. Cao, P.; Xia, W.; Li, Y. Heart id: Human identification based on radar micro-doppler signatures of the heart using deep Learning. *Remote Sens.* **2019**, *11*, 1220. [CrossRef]
12. Kim, Y.; Moon, T. Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2015**, *13*, 8–12. [CrossRef]
13. Le H.T.; Phung, S.L.; Bouzerdoum, A.; Tivive, F.H. Human motion classification with micro-Doppler radar and Bayesian-optimized convolutional neural networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2961–2965.
14. Gong, T.; Cheng, Y.; Li, X.; Chen, D. Micromotion detection of moving and spinning object based on rotational Doppler shift. *IEEE Microw. Wirel. Compon. Lett.* **2018**, *28*, 843–845. [CrossRef]
15. Seddon, N.; Bearpark, T. Observation of the inverse Doppler effect. *Science* **2003**, *302*, 1537–1540. [CrossRef] [PubMed]
16. Winkler, V. Range Doppler detection for automotive FMCW radars. In Proceedings of the 2007 European Radar Conference, Waltham, MA, USA, 17–20 April 2007; pp. 166–169.
17. Lin, J., Jr.; Li, Y.P.; Hsu, W.C.; Lee, T.S. Design of an FMCW radar baseband signal processing system for automotive application. *SpringerPlus* **2016**, *5*, 42. [CrossRef] [PubMed]
18. Li, X.; He, Y.; Jing, X. A survey of deep learning-based human activity recognition in radar. *Remote Sens.* **2019**, *11*, 1068. [CrossRef]
19. Lee, D.; Park, H.; Moon, T.; Kim, Y. Continual learning of micro-Doppler signature-based human activity classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
20. Martinez, J.; Vossiek, M. Deep learning-based segmentation for the extraction of micro-doppler signatures. In Proceedings of the 2018 15th European Radar Conference (EuRAD), Madrid, Spain, 26–28 September 2018; pp. 190–193. [CrossRef]
21. Abdulatif, S.; Wei, Q.; Aziz, F.; Kleiner, B.; Schneider, U. Micro-doppler based human-robot classification using ensemble and deep learning approaches. In Proceedings of the 2018 IEEE Radar Conference (RadarConf18), Oklahoma City, OK, USA, 23–27 April 2018.
22. Kim, Y.; Ling, H. Human activity classification based on micro-Doppler signatures using an artificial neural network. In Proceedings of the 2008 IEEE Antennas and Propagation Society International Symposium, San Diego, CA, USA, 5–11 July 2008; pp. 1–4.
23. Park, J.; Javier, R.J.; Moon, T.; Kim, Y. Micro-Doppler based classification of human aquatic activities via transfer learning of convolutional neural networks. *Sensors* **2016**, *16*, 1990. [CrossRef]
24. Wang, S.; Song, J.; Lien, J.; Poupirev, I.; Hilliges, O. 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, Tokyo, Japan, 16–19 October 2016; pp. 851–860.
25. Lin, Y.; Le Kernec, J.; Yang, S.; Fioranelli, F.; Romain, O.; Zhao, Z. Human activity classification with radar: Optimization and noise robustness with iterative convolutional neural networks followed with random forests. *IEEE Sens. J.* **2018**, *18*, 9669–9681. [CrossRef]
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
27. Cao, P.; Xia, W.; Ye, M.; Zhang, J.; Zhou, J. Radar-ID: Human identification based on radar micro-Doppler signatures using deep convolutional neural networks. *IET Radar Sonar Navig.* **2018**, *12*, 729–734. [CrossRef]
28. Narayanan, R.M.; Smith, S.; Gallagher, K.A. A multifrequency radar system for detecting humans and characterizing human activities for short-range through-wall and long-range foliage penetration applications. *Int. J. Microw. Sci. Technol.* **2014**, *2014*. [CrossRef]
29. Ni, Z.; Huang, B. Open-set human identification based on gait radar micro-Doppler signatures. *IEEE Sens. J.* **2021**, *21*, 8226–8233. [CrossRef]
30. Kwon, J.; Kwak, N. Human detection by neural networks using a low-cost short-range Doppler radar sensor. In Proceedings of the 2017 IEEE Radar Conference (RadarConf), Seattle, WA, USA, 8–12 May 2017.
31. Qiao, X.; Shan, T.; Tao, R. Human identification based on radar micro-Doppler signatures separation. *Electron. Lett.* **2020**, *56*, 195–196. [CrossRef]
32. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
33. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* **2015**, arXiv:1511.07289.
34. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Zhang, Z.; Tian, Z.; Zhou, M. Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor. *IEEE Sens. J.* **2018**, *18*, 3278–3289. [CrossRef]
37. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. Available online: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf> (accessed on 10 August 2022).



39. Ma, M.; Yoon, S.; Kim, J.; Lee, Y.; Kang, S.; Yoo, C.D. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020.
40. Kim, J.; Ma, M.; Pham, T.; Kim, K.; Yoo, C.D. Modality shifting attention network for multi-modal video question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.