Corruption Robustness Analysis of Radar Micro-Doppler Classification for Human Activity Recognition

Yi Zhou, Xuliang Yu, Miguel López-Benítez, Limin Yu*, and Yutao Yue*

***Abstract*—Radar-based human activity recognition (HAR) is a popular research field. Despite claims of high accuracy on self-collected datasets, the ability of these models to handle unexpected scenarios has been largely overlooked. This work introduces a framework for analyzing corruption robustness of radar micro-Doppler spectrogram classification. A set of corruptions are categorized, applied, and systematically tested on common model architectures. Diverse training methods, including adversarial training, cadence velocity diagram (CVD) transformation and data augmentation, are explored. The performance is evaluated on two tasks: indoor HAR and continuous aquatic HAR. Our study unveils several insights. Firstly, relying solely on accuracy may not adequately assess model performance due to dataset limitations. All well-trained models exhibit sensitivity to corruptions. Secondly, deeper convolutional neural network (CNN) models excel in both accuracy and robustness, but confront the problem of overfitting to background. Thirdly, adversarial training enhances robustness against corruptions, albeit at the cost of a slight decrease in accuracy. Lastly, combining data augmentation and adversarial training achieves a balance between accuracy and robustness. In essence, our study contributes to a more profound understanding of the complex interplay between model architecture, classification accuracy, and corruption robustness in radar HAR tasks.**

***Index Terms*—human activity recognition, micro-Doppler, robustness**

I. Introduction

The issue of robustness is a prominent research topic within the field of machine learning. Despite the impressive accuracy achieved by deep neural networks in various classification tasks, researchers have discovered that these networks are highly sensitive to small perturbations. The categorization presented in [1] establishes three kinds of robustness: adversarial noise robustness, natural noise robustness, and system noise robustness. Adversarial robustness, often considered the worst-case scenario, involves deliberately crafted small perturbations that lead neural networks to make incorrect predictions. Ongoing debates revolve around the feasibility of achieving both adversarial robustness and generalization simultaneously [2], [3]. The second type is natural noise robustness, also referred to as robustness under common corruptions [4]. This type of noise originates from observed objects or environmental factors. While the ideal approach is to capture this noise through data collection, real-world challenges often make it difficult to gather these unusual cases. As a result, natural noise is often overlooked in many datasets. The third type is system noise, which refers to noise that occurs in sensor models. Examples include noise introduced in the signal processing pipeline or the data transmission module. While much of the research on radar-based human activity recognition (HAR) focuses on adversarial robustness, implementing adversarial attacks at the data level can be tough due to the complex signal processing pipeline in real radar applications. In contrast, the last two types of robustness are more commonly observed but are not as well-studied. In this study, our main focus is on the last two categories: specifically, the robustness against natural noise and system noise in radar HAR tasks. In the subsequent sections, we will use the term 'corruption robustness' to collectively refer to both forms of robustness.

The corruption robustness has drawn significant research attention across various modalities. In the case of visual images, which are high-dimensional and rich in semantic information, the design space for image corruptions is substantial, making it an extensively studied area. Corruptions tailored to images typically involve noise and blur effects induced by the camera sensor [4]. Additionally, motion blur introduced by observed objects and environmental factors such as illumination [5], [6] can also contribute to the natural corruptions. In the audio domain, similar to radar, time-frequency analysis plays an important role in signal classification. Compared to radar signals, audio signals are well separated across different frequency domains, showcasing distinct patterns in both temporal and frequency dimensions. This distinctiveness underscores the need for perturbations to adhere to strict constraints, ensuring the preservation of the inherent natural characteristics of the audio signal. Therefore, the perturbations involves only additive noise and room reverberation. The robustness also

Manuscript received August 29, 2023; revised . This work received financial support from Jiangsu Industrial Technology Research Institute (JITRI) and Wuxi National Hi-Tech District (WND). (Corresponding author: Yutao Yue.)

Yi Zhou is with the Institute of Deep Perception Technology, JITRI, Wuxi 214000, China, and also with the XJTLU-JITRI Academy of Industrial Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: zhouyi1023@tju.edu.cn)

Xuliang Yu is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 12331100@zju.edu.cn)

Miguel López-Benítez is with the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3GJ, UK, and also with the ARIES Research Centre, Antonio de Nebrija University, 28040 Madrid, Spain (email: m.lopez-benitez@liverpool.ac.uk)

Limin Yu is with the Department of Electrical and Electronic Engineering, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (email: limin.yu@xjtlu.edu.cn)

Yutao Yue is with the Institute of Deep Perception Technology, JITRI, Wuxi 214000, China, and also with the XJTLU-JITRI Academy of Industrial Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China and Department of Mathematical Sciences, University of Liverpool, Liverpool L69 7ZX, UK (email: ytyue@ustc.edu)

captures attention in the field of point clouds [7], [8], [9]. Subtle local perturbations are introduced to spatial dimensions of points without undermining the overall structure. Examples include point jittering, alongside the addition or removal of local points.

In comparison to other fields, radar HAR stands out as a highly promising task for the study of robustness, owing to several factors. Firstly, the micro-Doppler spectrogram serves as the motion signature containing the superimposed reflections from different parts of the human body. Perturbations in human motion can exhibit substantial variations while retaining a sense of inherent naturalness. Secondly, the quality of radar spectrograms is closely linked to the underlying signal processing procedures. System noise introduced during signal processing can be viewed as a form of corruption. Thirdly, the majority of radar datasets lack diversity. Data collection for radar HAR often takes place in controlled laboratory environments, characterized by simplified conditions. Despite the extended time duration of the collected data, the diversity within these datasets remains limited. Consequently, models trained on these datasets are at risk of overfitting to features that may not be representative, such as background noise patterns. This concern is especially noteworthy given the prevalence of deep learning models with high capacity.Therefore, it is beneficial to include robustness as an additional metric to assess the overall generalization of the trained model.

In this paper, we introduce a robustness analysis framework for micro-Doppler spectrogram classification tasks. Our contributions can be summarized as follows:

- Taxonomy and Design of Corruptions: We categorize diverse corruptions affecting human activities and radar signals into three classes: temporal, Doppler, and intensity domains, outlining strategies to manage their severity for balanced evaluation.
- Benchmarking Model Architectures: Our evaluation involves various model architectures for the radar-based HAR task. These architectures include Conv1D-LSTM, Convolutional RNN (CRNN) [10], and different CNN variants [11], [12], [13]. We investigate their performance across two HAR scenarios: indoor HAR and continuous aquatic HAR. Our findings indicate that CNNs with higher capacity yield superior results in terms of both classification accuracy and robustness, but risk overfitting.
- Enhancing CNN Robustness: Our efforts to improve CNN robustness involve three approaches: using cadence velocity diagram (CVD) [14] as input, adversarial training [15], and data augmentation. While CVD has limited impact, adversarial training and data augmentation improve robustness and mitigate overfitting in deeper CNNs.

The remainder of this article is organized as follows. Section II introduces related works from both the perspectives of radar HAR and robustness. Section III describes the signal processing pipeline for micro-Doppler spectrogram extraction. Section IV elaborates on the definition and taxonomy of corruptions specifically designed for radar HAR tasks. Section V describes the model architectures and training methods for benchmark. Section VI specifies the datasets and experimental settings. Section VII summarizes and discusses the results with respect to classification accuracy and robustness. Finally, section VIII concludes this article.

II. Related Works

A. Radar-based HAR

Recently, the field of radar perception has witnessed significant advancements thanks to the progress in deep learning techniques [16]. Yang *et al*. [17] summarize model performance on public radar HAR datasets, showing significant improvements in this field. Despite achieving high accuracy, radar datasets often feature simplistic activity designs, laboratory environments, and ideal conditions. These limitations encourage the introduction of new challenges, such as continuous activity recognition, which involves identifying specific activities within sequential activities with unknown duration [18]. The continuous characteristics encapsulates the temporal diversity inherent in human activities. Furthermore, researchers also address variations associated with motion patterns, signal processing, and environments. Abdulatif *et al*. [19] use a generative adversarial network (GAN) to denoise micro-Doppler spectrograms, improving noise robustness. Yang *et al*. [20] propose a neural network module to generate super-resolution spectrograms, overcoming the limitations of the time-frequency uncertainty present in temporal-frequency analysis. Patel *et al*. [21] highlight the sensitivity of neural networks to domain shifts, corruptions and unknown objects, and underscore the problem of high-confidence incorrect predictions. A related work is the study by Czerkawski *et al*. [22], which shows CNN's sensitivity to subtle temporal shifts and adversarial examples in micro-Doppler classification. Training on adversarial examples and augmented samples improves robustness. Models operating on CVD representations also show adversarial robustness. Nevertheless, their evaluation of robustness is conducted within a classification task framework, rather than the robustness analysis framework employed in our study.

B. Robustness to Corruptions

Research on robustness spans across various modalities, including images [4], videos [5], [6], and point clouds [7], [8], [9]. The consistent findings underscore the vulnerability of neural networks to corruptions in the context of supervised learning. Data augmentation techniques have proven effective in enhancing robustness. AugMix, proposed by Hendrycks *et al*. [23], combines augmented views of images while preserving both semantics and local statistics, proving its effectiveness in enhancing robustness. Rusak *et al*. [24] demonstrate improved robustness with simple augmentations like Gaussian and Speckle noise. Modas *et al*. [25] prioritize semantically-preserving augmentation, and propose a strategy that samples transformations from a max-entropy distribution to preserve naturalness. FourierMix, introduced by Sun *et al*. [26], utilizes Fourier-based transformations to expand spectral coverage. Zhang *et al*. [27] design FourierShuffle, which shuffles high-frequency components to mitigate their impact. Furthermore,

the connection between data augmentation and domain adaptation is gaining attention. Schneider *et al*. [28] leverages the unlabeled corruption data to adapt batch normalization statistics. Xie *et al*. [29] discover that augmented data improves corruption robustness by aligning perceptual similarities between training and test data.

In addtion to data augmentation, the interplay between adversarial and natural robustness is also investigated. Tang *et al*. [1] and Kireev *et al*. [30] both demonstrate the capability of L_p adversarial training against common image corruptions with proper perturbation radius. The impact of model architecture on robustness is also a topic of discussion. Among models of comparable size, Tang *et al*. [1] highlight CNNs' superior robustness against natural and system noises over other model architectures like Transformers. Hooker *et al*. [31] observe robustness degradation during the compression of models. Timpl *et al*. [32] clarify that capacity reduction, rather than sparsity, drives robustness loss due to network compression.

III. Signal Processing Pipeline for HAR

Radar sensors are capable of detecting Doppler frequency shifts resulting from relative motion. In scenarios involving moving human limbs, subtle micro-motions introduce frequency modulations as sidebands around the primary Doppler-shifted frequency [14]. This dynamic Doppler pattern, known as the micro-Doppler signature, is examined through time-frequency analysis. In this section, we will elaborate on the signal processing pipeline for extracting micro-Doppler spectrograms and subsequently analyze their characteristics in the context of HAR tasks.

A. Micro-Doppler Spectrogram Extraction

Commercial radar systems transmit frames of frequency-modulated continuous wave (FMCW) chirp sequences to illuminate scenes. The received signals are mixed with the transmitted signals before being down-converted to an Intermediate Frequency (IF) at which an Analogue to Digital Converter (ADC) samples the mixed signal. An Analogue to Digital Converter (ADC) samples the mixed signal, with sampling index dimension referred to as the 'fast time.' The chirp dimension corresponds to the 'slow time.' For simplicity, we use only one received channel to extract the micro-Doppler spectrogram. As illustrated in fig. 1, the data acquired by the ADC is organized into a three-dimensional tensor, with dimensions corresponding to chirp (x-axis), ADC samples (y-axis), and frames (z-axis). The micro-Doppler spectrogram can be computed in several ways. For scene-based activity classification, the tensor is concatenated along the slow time, forming a lengthy 2D multidimensional sequence. Then, a range Fast Fourier Transform (FFT) is conducted along the fast time dimension, followed by Moving Target Indication (MTI) [33] to mitigate static clutters. Short-time Fourier transform (STFT) with a window size spanning the entire time duration is used to extract the micro-Doppler spectrogram, denoted as:

$$\mathrm{S}(t, f_d) = \mathrm{STFT}_{f \in w}(\mathrm{FFT}_{range}(R_{ADC})) \tag{1}$$

For continuous activity recognition, the storage and concatenation of multiple frames along the temporal dimension is infeasible. Alternatively, a simplified approach is employed to extract the micro-Doppler spectrogram in real-time. Following the conventional signal processing pipeline, a Range FFT, MTI, and Doppler FFT are applied within a single frame. This yields the Range-Doppler (RD) map, represented as:

$$\mathrm{RD}(f_r, f_d, t) = \mathrm{FFT}_{Doppler}(\mathrm{FFT}_{range}(R_{ADC})) \tag{2}$$

The RD map is then summed over the region of interest (ROI) along the range dimension, producing a Time-Doppler (TD) vector at time step t, expressed as:

$$\mathrm{TD}(t, f_d) = \sum\nolimits_{f_r \in \mathrm{ROI}} \mathrm{RD}(f_r, f_d, t) \tag{3}$$

Stacking TD vectors along temporal dimension results in the micro-Doppler spectrogram, also known as TD map. This procedure can be regarded as an STFT with a window size equal to one frame duration. Smaller window sizes yield decreased frequency resolution but increased temporal resolution, making them well-suited for continuous activities characterized by frequent motion changes.

B. HAR Task

For the HAR task, the micro-Doppler spectrogram encodes motion signature which can be utilized for classification of different motion patterns. The human body can be thought of non-rigid and modelled as multiple scattering points from torso and limbs. Suppose their corresponding scattering areas are A_T and A_{L_k}, respectively, and the instantaneous velocities at time t occupy several Doppler frequency bands f_T and f_{L_k}, respectively. The motion signal components can be expressed as

$$|\mathrm{S}(t, f_d)|^2 = |\xi_{\mathrm{T}} A_{\mathrm{T}} S\left(t, f_{\mathrm{T}}\right)| + \sum_{k=1}^{K} |\xi_{L_k} A_{L_k} S\left(t, f_{L_k}\right)| \tag{4}$$

where ξ is the attenuation coefficient of different body parts, which is a function of viewing angle and range. The overall spectrogram is the composed of motion signal components, background environment components and noise components.

The micro-Doppler spectrogram of human activity exhibits the following characteristics:

- The frequency of the moving torso is usually located at a narrow low frequency band. Moving limbs, on the other hand, can generate high frequencies and have a wider Doppler spread.
- The intensity of high-frequency components from limbs with smaller scattering areas is weaker than that of the torso.
- The motion is a function of time, and oscillatory limbs introduce a periodic motion pattern.

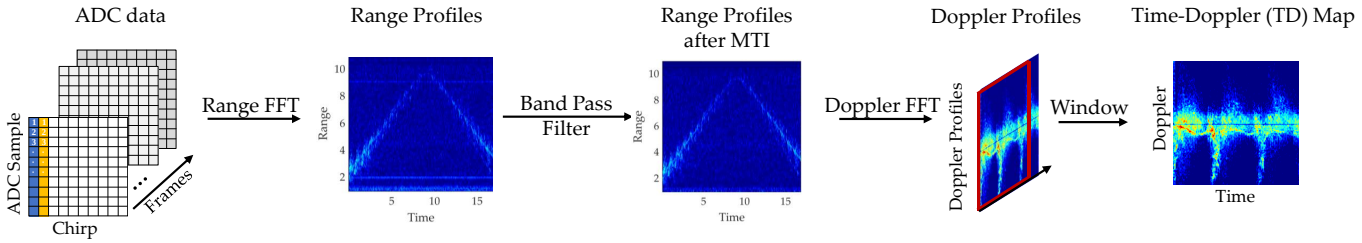

Fig. 1. Signal Processing Pipeline

IV. Corruptions for Micro-Doppler Spectrogram

A. Definition of Corruptions

Classifiers often encounter scenarios where they must classify low-quality or corrupted inputs. Essentially, corruption robustness measures the average performance of the classifier when faced with various corruptions from the set C, while adversarial robustness assesses the classifier's worst-case performance when dealing with small additive perturbations.

It is important to acknowledge that we cannot really measure the exact occurrence probability of specific corruptions. Therefore, we often make the assumption that $\mathbb{P}_C(c)$ is equal for all corruptions. As a result, the primary focus shifts towards designing an appropriate set of corruption functions C. To effectively design the set of corruption functions, several important factors should be taken into account. First and foremost, the choice of corruptions should be aligned with real world scenarios. These corruptions should maintain physical fidelity to ensure that the evaluation remains meaningful and realistic. Secondly, the definition of what constitutes a 'corruption' is often dataset-dependent. The infrequent occurrence of a corruption is defined based on the perspective of the target dataset. For instance, if a specific type of corrupted data is well-captured in the dataset, it may not be considered as a corruption. Conversely, for datasets with limited data diversity, some common data augmentations might also be considered as corruptions.

B. Corruption Taxonomy for Radar HAR

We establish a taxonomy for categorizing a diverse range of corruptions for radar HAR task, capturing variations in both human activities and radar signal processing complexities. These corruptions can be organized into three primary classes: Doppler domain, temporal domain, and intensity domain.

1) Corruptions in Doppler Domain: The frequency domain contains valuable information about the distribution of energy in terms of Doppler velocity. While the primary velocity of the human body is the main component, the limbs of a human introduce notable spread along the Doppler dimension. Instead of suppressing this spread, we can leverage it for distinguishing between various human activities. To this end, we have devised three types of corruptions that are applied in the frequency domain as depicted in fig. 2.

The first type involves scaling along the Doppler dimension, which can be interpreted as variations in the orientation of objects. Given that Doppler velocity corresponds to radial velocity, projecting the full velocity onto the observing direction is a nonlinear function of orientation. Therefore, recovering the full velocity and reprojecting it to different orientations is challenging due to the absence of spatial information. Kern *et al.* [34] address this problem by data augmentation through simulation, but we take a simpler approach by slightly scaling the Doppler dimension. This approximation is sufficient to capture the variance and serves as a robustness test. The second type is Doppler jittering, where the energy in neighboring cells is randomly shuffled. This captures the randomness in the scattering point locations. By disrupting the local structure, it serves as a test of whether the neural network can effectively utilize global information for classification. The final type involves changing the FFT resolution by using fewer FFT points. This manipulation explores the effect of reducing frequency resolution on classification performance.

2) Corruptions in Temporal Domain: We design four types of corruptions that target the temporal dimension, corresponding to variations along the x-axis. The first type involves temporal scaling, where the temporal frames are stretched or compressed to model variations in motion change or periodicity. The second type, temporal drop, simulates the occurrence of frame drops that could happen due to unreliable connections during data transfer. The third type, temporal STFT, is achieved by utilizing a larger window size in the temporal dimension. Lastly, the fourth type, temporal masking, randomly masks out consecutive temporal sequences within the spectrogram. This type of corruption mimics scenarios of occlusion, where parts of the activity are not captured.

3) Corruptions in Intensity Domain: We implement two kinds of corruptions in intensity. The first type involves enhancing the intensity of background cells to intentionally reduce the SNR. This process begins with the application of the Cell Average Constant False Alarm Rate (CA-CFAR) detector [33] to distinguish between foreground and background cells. Then, we increase the intensity of background cells to lower the SNR. The second type aims to modify the SNR by introducing local random noise, specifically Gaussian white noise, across the entire spectrogram. This particular corruption simulates real-world scenarios where signals can be corrupted by various sources of interference, clutter, or distortion.

C. Controlled Corruption Severity

Following the conventions in studying robustness across other modalities, we also apply each type of corruption at different severity levels. However, controlling the severity of corruptions demands a systematic and numerical approach

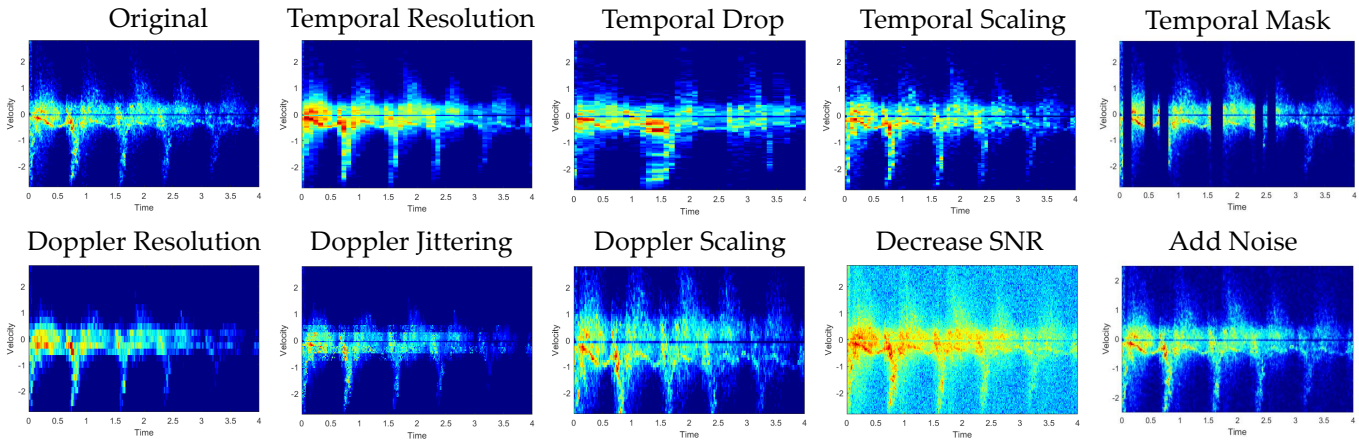

Fig. 2. Examples of Corruptions for Radar HAR

rather than arbitrarily assigning severity values. This task is inherently complex due to the presence of distinct physical constraints associated with each type of corruption. To overcome this challenge, we categorize the corruptions into three types based on their underlying physical interpretations.

The first type encompasses geometric transformations, such as scaling and downsampling. In implementing these transformations, our primary consideration is on preserving physical realism and adhering to the constraints imposed by the underlying physical processes. For instance, when applying temporal scaling to activities like walking, we ensure that the scaled data remain within the typical human walking speed range of 1.0 to 1.6 meters per second.

The second category involves structural corruptions, including drop and jittering. These corruptions introduce disruptions to the local structure of the original data. To assess the impact of local perturbations on global information, we leverage the structural similarity index (SSIM) [35] as a metric to control the severity. The SSIM is a widely accepted measure in image quality assessment that jointly takes into account the similarity in terms of luminance, contrast, and structure.

The third type of corruptions applies to intensity. Given that the intensity of motion can exhibit fluctuations contingent on the type of motion, motion scale, and the object's distance, we adopt an adaptive-threshold-based SNR measurement to quantify the extent of signal corruptions. In the context of spectrograms, the adaptive threshold is computed by applying a CA-CFAR detector to the windowed spectrogram. This process separates the reflections into two categories: human activities denoted as Ω_s and background noise signals denoted as Ω_n. Then, the SNR is calculated according to

$$\mathrm{SNR}_{\mathrm{tf}} = 10 \log_{10} \left(\frac{\mathrm{mean} \left| S\left(t, f_d\right) \right|^2_{(t,f_d)\in\Omega_s}}{\mathrm{mean} \left| S\left(t, f_d\right) \right|^2_{(t,f_d)\in\Omega_n}} \right) \tag{5}$$

V. Model Architecture and Training Methods

A. Model Architectures

As depicted in fig. 3, three paradigms for processing radar spectrograms using neural networks are showcased. The first paradigm views the spectrogram as a multi-dimensional temporal sequence. Each time step is processed using lightweight 1D convolutions, capturing frame-wise features. LSTM modules are then employed to model temporal relationships between frames. The second paradigm interprets the spectrogram as an image and leverages CNN architectures. 2D convolutional kernels slide across the spectrogram in both frequency and temporal dimensions to capture local patterns. A hierarchical structure expands the receptive field, enabling extraction of global information. The third paradigm combines elements of both previous paradigms by using 2D convolutional layers for feature extraction and LSTM to model the temporal dimension. In order to reduce the 2D feature maps into a 1D temporal sequence, the kernel size of pooling layers is specifically designed to condense the channels along the frequency dimension while preserving temporal resolution along the temporal dimension.

Given the focus on benchmarking different paradigms rather than introducing novel architectural designs, complex models are avoided. Four representative model architectures are evaluated: Conv1D LSTM, CRNN [10], a shallow CNN named VGG-7 [11], and a deeper CNN termed ResNet-18 [12]. Additionally, we explores the integration of Convolutional Block Attention Module (CBAM) [13] attention into ResNet to investigate its effects.

1) Conv1D LSTM: For Conv1D LSTM, the input spectrogram is treated as a multidimensional time series. In our implementation, the initial layer of this architecture consists of a stack of two 1D convolutional layers. These convolutional layers use kernel sizes of 4 and 3, and they have channel sizes of 32 and 64, respectively. The 1D convolutions are applied along the Doppler dimension to extract frame-wise features from the spectrogram. Subsequently, the extracted features are fed into bidirectional LSTM cells, which capture temporal relationships between features. Finally, the output of the LSTM cells at the last time step is sent to a fully connected (FC) layer with softmax for classification.

2) CRNN: CRNN employs 2D convolution for feature extraction and progressively compresses the Doppler dimension through pooling layers. The initial convolutional layer

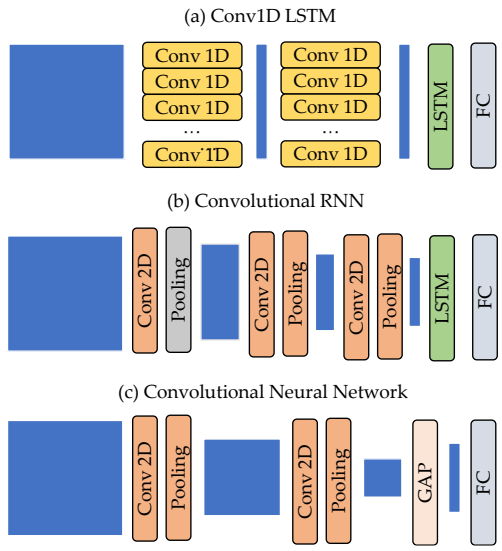

Fig. 3. Typical Architectures to Process Spectrogram

comprises a stack of three 2D convolutional layers with kernel sizes of 5, and channel sizes of 128, followed by maxpooling and ReLu. The max pooling layer is specially designed so that the feature map is downsampled in the frequency dimension with a large step, and in the final layer the frequency dimension is reduced to one and can be squeezed. Then, the feature map becomes a 2D temporal sequence, which is sent to a bi-directional LSTM for modelling the sequential information, followed by an FC layer with softmax for classification.

3) VGG: VGG is a CNN architecture utilizing small kernels to gradually increase receptive fields. To assess shallow CNNs, we construct a light-weight VGG-7. In our implementation, we reduce convolutions to 7 layers and replace FC layers with global average pooling. This consists of 7 convolutional layers, all 3×3. The first 3 layers have 32 channels, then 2 layers each with 64 channels, and 1 layer with 128 channels. Max-pooling of size 2×2 is applied after each convolutional blocks. Finally, global average pooling reduces feature maps to 1D vector, which are fed to a softmax FC layer for classification.

4) ResNet: ResNet, built by stacking multiple residual blocks, stands as the most prominent CNN architecture in addressing the vanishing gradient problem and enabling the efficient training of very deep networks. Each residual block is comprised of two sequential 3×3 convolutional layers and a residual connection. This residual connection performs identity mapping, directly connecting the input of a residual block to its output. The final output of the block is determined by adding the residual with the output from the convolutional layers. Specifically, ResNet-18 comprises five modules, with the first module featuring a single convolutional layer and the subsequent four modules employing two stacked residual blocks each.

5) ResNet with CBAM Attention: CBAM [13] is an attention mechanism enhancing CNN's representational power by modeling interdependencies among channels and spatial locations. It has two sub-modules: channel attention emphasizing 'what,' and spatial attention emphasizing 'where.'

Channel attention aggregates spatial information using average-pooling and max-pooling. Shared Multi-Layer Perceptron (MLP) processes this information to learn channel importance. The channel attention weight W_c is computed as follows:

$$W_c = \sigma(\text{MLP}(f_{avg}(X)) + \text{MLP}(f_{max}(X))) \tag{6}$$

where $f_{avg}(X)$and $f_{max}(X)$ represent the average pooled and max pooled feature maps. MLP denotes a shared MLP with one hidden layer and σ is the sigmoid activation function.

Spatial attention compresses feature maps along channel dimensions through max-pooling and average-pooling. Concatenating the resulting feature maps, they are passed through a convolutional layer to summarize spatial information. Spatial attention weight W_s is computed as:

$$W_s = \sigma(Conv(f_{pool}(X); f_{max}(X))) \tag{7}$$

where $f_{pool}(X)$ and $f_{max}(X)$ are the average-pooled features and max-pooled features across the channels. Conv denotes a 7×7 convolutional layer and σ is the sigmoid activation function.

The channel attention and spatial attention are sequentially applied to reweight the original feature maps. In our implementation, CBAM is introduced to the final feature maps in each basic module of the ResNet-18 architecture.

B. Training Methods

1) Adversarial Training: Adversarial training is reported to be beneficial for corruption robustness [1], [30]. This technique involves augmenting the training data with carefully crafted adversarial examples. By exposing the model to these adversarial examples during training, it learns to better generalize and improve its resistance to perturbations at test time. In this study, adversarial examples are constructed using a multi-step first-order method known as projected gradient descent (PGD) [15]. The PGD perturbation for a given input x can be expressed as follows:

$$\delta^{t+1} = \delta^t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f(x + \delta^t), y_{\text{true}})) \tag{8}$$

where δ^{t+1} denotes the adversarial perturbation at iteration $t+1$, α is a hyper paratmeter that adjusts the learning rate, $f(x + \delta^t)$ is the model's prediction on the perturbed input, y_{true} is the true label of input x, and $\mathcal{L}$ represents the loss function. The perturbation is updated over multiple steps with a small learning rate α until the maximum iteration is reached. After each update, the current perturbation δ^{t+1} is projected onto a set of constraints. For adversarial examples computed on normalized data, this constraint is typically $x + \delta^t \in [0, 1]$. Additionally, adversarial examples are constrained to an L_∞ ball of size ϵ around x. In the training process, the original training dataset is dynamically augmented with these adversarial examples at each epoch.

2) Cadence Velocity Diagram: The CVD representation can be computed by applying another FFT along the temporal dimension of the micro-Doppler spectrogram:

$$\mathrm{CVD}(f_t, f_d) = \mathrm{FFT}_t(\mathrm{S}(t, f_d)) \tag{9}$$

The reversible nature of the FFT operation ensures that the CVD operates as a transformed representation without any loss of information. The notable advantage of the CVD lies in its ability to capture information related to periodic features.

3) Data Augmentation for Image: Visual classification tasks often require invariance to color variations and geometric transformations. As such invariance cannot be intrinsically ensured by network architecture, these inductive biases are introduced through data augmentation. Common image data augmentation techniques are conveniently integrated into deep learning frameworks like PyTorch and TensorFlow, and many works directly apply these techniques to radar tasks. However, it is important to note that radar spectrogram have different interpretation compared to images. In this work, we consider three kinds of data augmentation and examines their applicability to radar spectrograms.

The first type of augmentation involves color jittering, which adjusts the brightness and contrast of the image. Notably, in an image, changes in brightness correspond to variations in intensity in radar spectrograms, while alterations in contrast can approximate changes in SNR in radar spectrograms.

For geometric transformations, we employ the random resized crop augmentation, which randomly crops a section of the image and resizes it to the original size. For radar spectrograms, the transformation scale must be chosen carefully to avoid unrealistic data.

Lastly, Gaussian blur is randomly applied to the image. It involves convolving the original image with a Gaussian kernel, which approximates the reduction in resolution that can occur in radar spectrograms.

4) Spec Augmentation for Audio: Data augmentation techniques designed for audio spectrograms take into account the distinct temporal and frequency dimensions. These methods aim to enhance the robustness of the features against variations in the time direction, partial loss of frequency information, and partial loss of small segments of speech. One prominent approach, known as SpecAug [36], employs the following three types of augmentation.

The first augmentation method is time warping. It involves cropping consecutive frames and moving them along the temporal dimension by a distance of w. The entire sequence is then interpolated using spline-based grid interpolation. Warping augmentation does not alter the data length, thereby avoiding information loss that can occur with scaling and cropping.

The second augmentation method is frequency masking, which involves masking consecutive frequency channels in the spectrogram. Specifically, f consecutive frequency channels are masked, where f is chosen from a uniform distribution between 0 and the frequency mask parameter F.

The third augmentation method is time masking. Similarly, time masking is applied so that t consecutive time steps are masked, where t is first chosen from a uniform distribution between 0 and the time mask parameter T.

VI. Datasets and Experiment Setting

A. Radar HAR Datasets

As shown in table I, two datasets are chosen for benchmarking. The first is the Glasgow indoor HAR dataset [37], which employs a 5.8 GHz FMCW radar featuring a 400 MHz bandwidth and 1 ms chirp duration. The dataset consists of recordings from 20 volunteers performing six different activities: walking, sitting down, standing up, picking up an object, drinking water, and falling. Each activity class comprises 300 data instances, and each data instance lasts for 10 seconds. During the signal processing stage, the entire sequence is treated as a single sample, and a STFT with a window size equal to the entire sequence duration is applied to extract the spectrogram. The second dataset [38] focuses on aquatic human activity recognition. For this dataset, a 77 GHz FMCW radar with 1.7 GHz bandwidth, 128 chirps, and 0.33 ms chirp duration is used. Five activities including floating, struggling and three swimming styles (backstroke, breaststroke, freestyle) are recorded for a consecutive 20 or 40 seconds for each recorded sequence. A 128-point Doppler FFT along slow time is applied in signal processing to obtain TD maps. Utilizing a small frame length of 20 with a 0.5 overlap, each class of activity consists of approximately 600 data instances. Notably, swimming activities exhibit variable and prolonged periods, making it highly probable to encompass incomplete motion patterns within a data instance. Also, submerged bodies lead to weaker reflected energy, further complicating the task of accurately discerning different activities.

TABLE I
Radar Configuration

	Glasgow Indoor HAR	ZJU Aquatic HAR
Operating Frequency	5.8 GHz	77 GHz
Bandwidth	0.4 GHz	1.7 GHz
Chirp Time	1 ms	0.33 ms
Number of ADC per Chirp	128	256
Number of Chirp per Frame	128	128

B. Corruption Implementation

As explained in section IV, we have categorized the corruptions into three distinct classes and a total of nine possible types. In the case of the aquatic dataset, these corruptions are applied to entire sequences before dividing them into individual instances, ensuring consistent neighboring frames. The detailed settings are shown in table II. For each class of corruption, a base value is chosen and subsequently scaled according to different levels of severity.

C. Experiment Setting

Our approach directly employs raw micro-Doppler spectrograms as inputs. These spectrograms are resized to dimensions of (224, 224) for compatibility with common image-based network architectures. Input normalization is performed using precomputed mean and standard deviation values from all training samples. Our evaluation covers five distinct architectures: Conv1D LSTM, CRNN, VGG-7, ResNet-18, and

TABLE II
CORRUPTION CONFIGURATION

Corruption Type	Base	Scaling Factor
Temporal Resolution	Window Size 128	[2, 0.5, 0.25, 0.125]
Temporal Drop Step	Time Step 10	[1, 2, 3, 4, 5]
Temporal Scaling Ratio	X-Axis Length 128	[0.5, 0.8, 1.4, 1.7]
Temporal Mask Ratio	Base Percentage*	[0.9, 0.8, 0.7, 0.6, 0.5]
Doppler Resolution	FFT Point 128	[0.5, 0.25, 0.125]
Doppler Jittering Ratio	Kernel Size (1×1)	[2 ,3 ,4 ,5 ,6]
Doppler Scaling Ratio	Y-Axis Length 128	[2, 0.5, 0.25, 0.125]
Background Increase	Relative SNR (dB)	[-5, -10, -15, -20, -30]
Gaussian White Noise**	Average SNR (dBW)	[5, 7, 10, 12, 15]

* Determined class-wise by maintaining a consistent SSIM value of 0.1.
** Introduced as Additive White Gaussian Noise (AWGN) assuming that the input signal power is 0 dBW.

ResNet-18 with CBAM attention. We explore various training techniques, including adversarial training, where we generate PGD perturbations with parameters $\epsilon = 0.1$, $\alpha = 0.01$, and perform 20 iterations. For real-time CVD transformation, we compute a 256-point FFT along the time dimension of the spectrogram during preprocessing. Our image augmentation techniques encompass Color Jitter for contrast and brightness adjustments, Random Resized Crop to (224, 224) following random cropping (200, 200), and Random Gaussian Blur with a kernel size of 3 and a 20% chance of application. In the case of spec augmentation, we utilize Temporal Warping with a maximum warping distance of 40 time steps, along with Frequency Masking and Temporal Masking with maximum masked ratios of 0.15 and 0.5, respectively.

The optimization objective is based on Cross Entropy Loss. The optimization process employs the Adam optimizer with a learning rate of either 10^{-3} or 10^{-4}. To adaptively adjust the learning rate when the validation losses plateau, we utilize the ReduceLROnPlateau scheduler to enhance convergence. The selection of the best model is determined by monitoring the validation loss. Early stopping is incorporated with a patience of 5 epochs, meaning that training halts if the validation loss does not improve within this specified window.

D. Evaluation Metrics

Considering the balanced nature of the dataset, accuracy serves as the primary evaluation metric for the original dataset. For a comprehensive assessment of robustness, we adopt the Corruption Error (CE) metric [4]. The formula for CE is as follows:

$$\mathrm{CE}_i = \frac{\sum_{l=1}^{s} (1 - \mathrm{Acc}_{i,l})}{\sum_{l=1}^{s} \left(1 - \mathrm{Acc}_{i,l}^{\mathrm{base}}\right)} \tag{10}$$

where $\mathrm{Acc}_{i,l}$ represents the accuracy of a corrupted test set i at a specific severity level $l \in [1, s]$, and $\mathrm{Acc}_{i,l}^{\mathrm{base}}$ denotes the baseline model's accuracy. In essence, CE quantifies the relative performance degradation with respect to the base model for a given test corruption.

To evaluate the model's overall robustness across different corruptions, we employ the mean Corruption Error (mCE) metric: $\mathrm{mCE} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CE}_i$, where N represents the number of corruptions. This metric aligns with the definition of corruption robustness outlined in section IV-A. While mCE is capable of evaluating the model's performance on corrupted sets, it may not fully consider the original performance. In situations where a certain level of accuracy trade-off is acceptable in order to ensure consistent performance across both clean and corrupted datasets, we adopt the concept of Relative mCE (RmCE):

$$\mathrm{RmCE} = \frac{1}{N} \frac{\sum_{l=1}^{s} (\mathrm{Acc}_{\mathrm{clean}} - \mathrm{Acc}_{i,l})}{\sum_{l=1}^{s} \left(\mathrm{Acc}_{\mathrm{clean}}^{\mathrm{base}} - \mathrm{Acc}_{i,l}^{\mathrm{base}}\right)}, \tag{11}$$

where $\mathrm{Acc}_{\mathrm{clean}}$ is the accuracy on the clean test set. RmCE provides a more nuanced perspective by quantifying the degree of performance drop compared to the accuracy on the clean test set.

VII. RESULT ANALYSIS

A. Classification Accuracy of Models

Table III provides a performance comparison of various models for indoor and continuous aquatic HAR. All models demonstrate high accuracy on the indoor HAR dataset, with minor variations due to the relatively limited dataset size. However, the aquatic HAR dataset presents more challenges, resulting in lower accuracy across models. The model performance ranking indicates that ResNet outperforms VGG, which in turn surpasses CRNN, followed by Conv1D LSTM. This ranking underscores the significance of capturing local motion patterns for accurately classifying continuous activities. Deeper convolutional architectures, like ResNet-18, are more suitable for handling complex recognition tasks but demand greater computational resources. Conversely, the lightweight Conv1D LSTM is better suited for simpler tasks.

The confusion matrix, illustrated in fig. 4, provides insights into class-wise accuracy. In indoor HAR, high-precision classification is achieved for most classes, with occasional confusion between activities like 'drink' and 'pick.' Higher-capacity models improve all classes. For aquatic HAR, distinguishing between activities such as 'freestyle' and 'breaststroke' proves challenging. Temporal models yield enhanced differentiation for 'backstroke' and 'breaststroke,' although they struggle with 'float.' VGG-7 performs well in classifying 'float' but faces difficulties with swimming styles. ResNet-18 significantly improves the classification of dynamic activities like 'struggle.'

Saliency maps, depicted in fig. 5, provide further insights into the models' predictions. Temporal models emphasize the temporal dimension, with Conv1D LSTM showing relatively weaker emphasis on the Doppler dimension. The CRNN architecture enhances Doppler dimension emphasis through the utilization of 2D convolutions. VGG places greater focus on local patterns, while ResNet-18 exhibits a tendency to attend to background areas. Similar observations apply to aquatic HAR, where the overfitting of ResNet-18 becomes more pronounced. Notably, extreme scenarios like the 'float' activity reveal predictions based on background occupancy rather than specific activities.

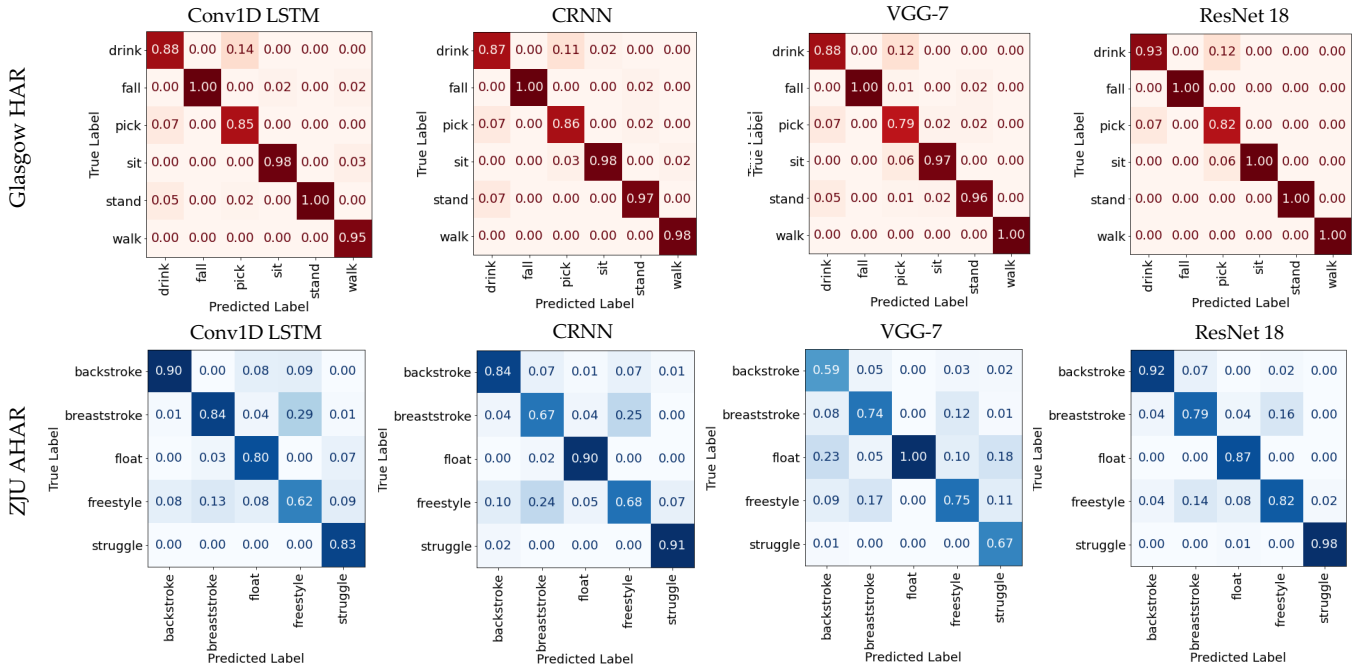

Fig. 4. Confusion Matrix

TABLE III
CLASSIFICATION PERFORMANCE

Model	Accuracy(%)		Params(G)	FLOPs(M)	Time(ms)
	Indoor	Aquatic			
Conv1D LSTM	94.53	80.31	0.008	0.111	0.609
CRNN	94.53	82.46	0.415	0.895	0.279
VGG-7	93.49	85.56	2.095	0.298	0.226
ResNet-18	95.83	88.73	1.824	11.180	4.576
ResNet-18+CBAM	95.31	90.89	1.825	11.269	11.048
ResNet-18+CVD	93.49	78.48	1.824	11.180	5.083
ResNet-18+ImageAug	96.88	87.50	1.824	11.180	4.576
ResNet-18+SpecAug	95.83	90.94	1.824	11.180	4.576
ResNet-18+Adv	94.79	87.17	1.824	11.180	4.576
ResNet-18+All	93.75	88.44	1.824	11.180	4.576

* The inference time was measured using an Nvidia RTX 2080 Ti.

B. Robusntess Analysis of Models

The baseline model selected for further analysis is ResNet-18 due to its high accuracy on both datasets. Before proceeding to compare the methods using the robustness metric, we firstly inspect the performance drop for different datasets and corruption types. Figure 6 shows the accuracy with respect to different corruptions. Notably, scene-based HAR exhibits lower sensitivity to noise and temporal corruptions. Doppler and intensity corruptions appear to have a more consistent impact across datasets.

Table IV provides a comparative analysis of robustness to different corruption types across methods. All models exhibit lower robustness compared to the ResNet-18 baseline, as indicated by higher mCE values. Analyzing the relative mCE values, we observe that CRNN showcases a smaller performance drop (RmCE 0.88) relative to the clean dataset compared to ResNet. This trend is aligned with the observations from saliency maps, where CRNN demonstrates better attention to the temporal and Doppler dimensions, while ResNet-18 relies more on features from the background.

Inspecting corruption-wise CE, temporal models outperform ResNet-18 in noise corruptions on both datasets since the homegeneous noise has less impact on the temporal relationships. Conversely, ResNet-18 demonstrates superior performance in handling temporal corruptions and variations in SNR due to its ability to identify local patterns. VGG's emphasis on activity patterns rather than background details, explains its reduced sensitivity to increased background noise levels. Furthermore, VGG, featuring a smaller receptive field, is extreme sensitive to scaling and dropping corruptions, whereas ResNet's superior utilization of global information results in improved performance.

Dataset-specific differences are highlighted in yellow in table IV. Temporal models excel against frequency corruptions in the first dataset and struggle in the second. The underlying reason can be attributed to the temporal model's tendency to treat each frame as a holistic entity. As a result, it demonstrates resilience to Doppler corruptions in the simpler dataset with clear energy fluctuations over time. In the aquatic dataset, where swimming activities manifest similar fluctuations, doppler corruptions disrupt the energy distribution, leading to the model's struggle in distinguishing activities. Regarding temporal masking, all three models show lower performance compared to ResNet in the first dataset, but they demonstrate superior performance in the second dataset. This phenomenon can be explained by considering that the ResNet model trained on the aquatic dataset overfit to the background. The information loss introduced by temporal masking makes the weak activity signals challenging to identify, but this has a comparatively lesser impact on the background. Interestingly,

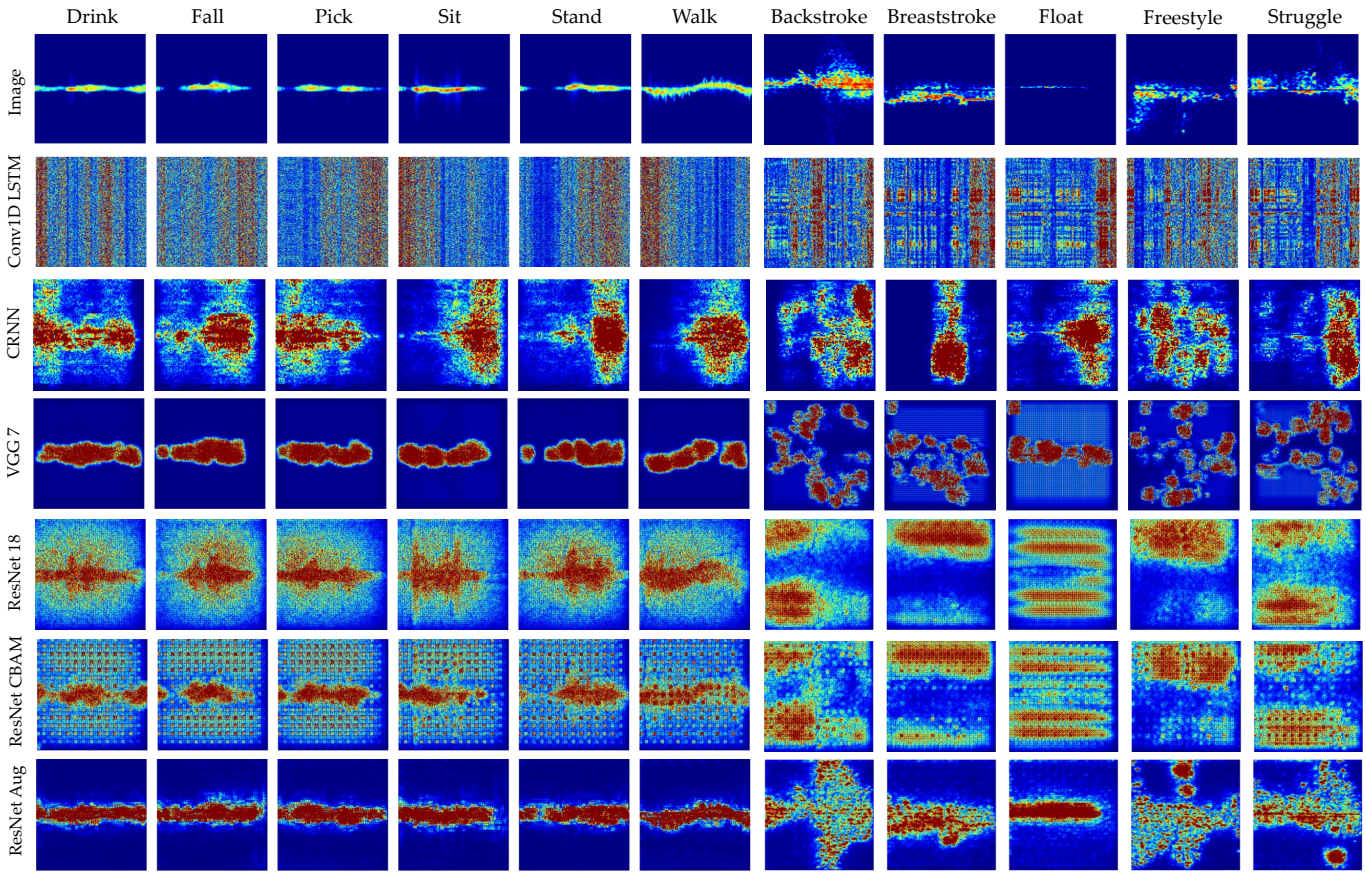

Fig. 5. Saliency Maps for Model Architectures

VGG performs better than ResNet when subjected to local random noise in the first dataset but worse in the second dataset. The saliency map analysis suggests that the first VGG model effectively attends to activity patterns, making it robust to noise, while the second model relies more on local patterns, making it more sensitive to random noise. In summary, from a robustness perspective, the primary distinction between these two datasets lies in the temporal characteristics of the captured activities.

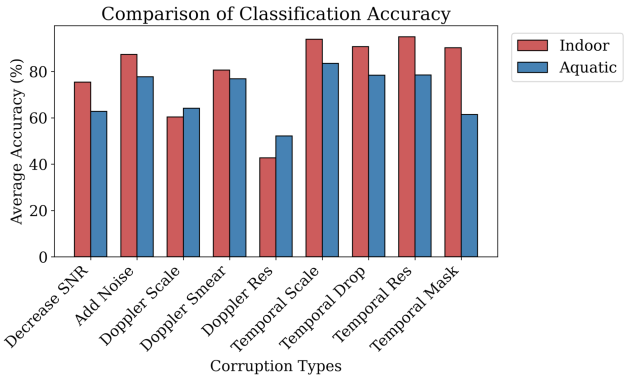

Fig. 6. Accuracy for Corruption Types

C. Analysis of Improved Models

Notably, the baseline model, ResNet-18, demonstrates high accuracy in both datasets. However, upon closer inspection of the saliency map and mCE, we observe that ResNet tends to overfit to background cells, particularly evident in the aquatic dataset. This overfitting phenomenon compromises the model's robustness and raises concerns about its generalization to activities not explicitly represented in the dataset. Thus, it becomes imperative to explore potential modifications that can alleviate the issue of overfitting and improve robustness. We tested a range of methods, including changes in model architecture, data augmentation, and adversarial training. The results of both classification and robustness evaluations are presented in the lower sections of table III and table IV, respectively.

From table III, several key observations can be made. Firstly, due to the small size of the indoor HAR dataset, slight performance improvements or degradation are observed. However, more noticeable performance enhancements are observed on the aquatic dataset. The integration of CBAM attention modules or the application of spec augmentation contribute to improved performance. It is important to note that the performance boost from attention modules comes at the expense of increased inference time. In contrast, data augmentation techniques lead to performance improvements without additional computational cost. On the other hand, the

TABLE IV
ROBUSNTESS TO CORRUPTIONS

Indoor HAR Dataset											
Method	mCE	RmCE	SNR ↓	Add Noise	Freq Scale	Freq Jitter	Freq STFT	Temp Scale	Temp Drop	Temp STFT	Temp Mask
ConvLSTM	1.29	1.23	2.84	0.53	1.07	0.88	0.77	1.11	1.30	1.25	1.90
CRNN	1.06	0.88	2.24	0.43	0.90	0.68	0.69	1.00	1.16	1.24	1.26
VGG-7	1.61	2.10	0.50	0.60	1.28	2.47	1.03	1.73	2.29	2.25	2.34
ResNet-18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ResNet-18+CBAM	1.33	1.40	2.04	2.29	1.09	1.04	0.94	1.18	1.24	1.09	1.03
ResNet-18+CVD	1.30	1.42	0.28	0.49	1.31	1.71	0.95	1.62	2.05	1.87	1.43
ResNet-18+Adv	0.73	0.29	0.75	0.37	0.71	0.42	0.69	0.67	0.78	0.98	1.17
ResNet-18+ImageAug	0.80	1.19	0.20	0.96	0.16	0.52	0.79	0.94	1.20	1.26	1.19
ResNet-18+SpecAug	0.70	0.37	0.49	0.90	0.97	0.63	0.93	0.52	0.56	0.81	0.51
ResNet-18+All	0.79	0.51	0.29	0.50	0.36	0.49	0.70	1.23	1.22	1.49	0.87
Aquatic HAR Dataset											
Method	mCE	RmCE	SNR↓	Add Noise	Freq Scale	Freq Smear	Freq STFT	Temp Scale	Temp Drop	Temp STFT	Temp Mask
ConvLSTM	1.22	2.31	1.54	0.82	1.25	1.47	1.05	1.47	1.34	1.26	0.79
CRNN	1.11	0.80	1.51	0.72	0.93	1.26	1.00	1.35	1.27	1.46	0.49
VGG-7	1.19	1.24	0.62	1.37	1.26	1.69	1.13	1.47	1.36	1.30	0.54
ResNet-18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ResNet-18+CBAM	1.03	1.20	1.73	0.61	0.98	0.97	1.05	1.05	1.00	1.10	0.81
ResNet-18+CVD	1.71	1.95	0.63	1.06	1.33	1.50	1.12	3.31	2.72	2.60	1.12
ResNet-18+Adv	0.78	0.50	0.76	0.56	0.83	0.81	0.78	0.89	0.92	0.87	0.62
ResNet-18+ImageAug	0.99	1.08	0.32	0.52	0.59	1.01	0.81	1.75	1.56	1.51	0.83
ResNet-18+SpecAug	0.73	0.65	1.27	0.43	0.81	0.83	0.83	0.74	0.75	0.69	0.18
ResNet-18+All	0.68	0.43	0.54	0.49	0.86	0.80	0.82	0.84	0.82	0.76	0.23

introduction of adversarial training results in a slight negative impact on accuracy for both datasets, which aligns with the literature indicating that adversarial robustness can conflict with accuracy. The CVD transformation shows comparable performance on the indoor HAR dataset and worse performance on the aquatic dataset. Although the CVD transformation can yield better representations for periodic motion patterns, it is less effective for continuous HAR tasks. In fact, applying the CVD transformation on partially observed motion could hinder the identification of local motion patterns, thereby posing a challenge in discerning relevant features.

Since classification accuracy alone may not fully reveal model superiority, a comprehensive analysis of robustness to corruptions, as depicted in table IV, is conducted for further insights. The best and worst performances are highlighted in green and red, respectively, to facilitate visualization. Notably, similar trends emerge across datasets, suggesting a dataset-independent characteristics of the evaluation. Firstly, the integration of CBAM attention modules appears to compromise the robustness of ResNet, particularly evident as background noise intensity increases. A potential explanation lies in the tendency of the ResNet model equipped with CBAM attention to excessively focus on background cells, as inferred from the saliency map in fig. 5.

As a data augmentation technique, image augmentation significantly enhances robustness against SNR corruptions, albeit at the cost of performance drop in temporal corruptions. The CVD transformation displays high sensitivity to corruptions, especially for the temporal corruptions. The reason is well explained when we analyze the performance drop of CVD in accuracy. Spec augmentation substantially improves resistance to temporal corruptions, thanks to the inclusion of temporal warping. Adversarial training consistently reduces errors across various corruption types at the cost of accuracy drop. When combining spec augmentation, image augmentation, and adversarial training, robustness improves across all categories, resulting in the lowest mCE (0.68). The final row in fig. 5 highlights that the implementation of data augmentations and adversarial training enables the model to effectively focus on motion patterns, thereby partially explaining the improved robustness.

VIII. CONCLUSION

This study has presented a framework for analyzing the corruption robustness of radar micro-Doppler spectrogram classification. Corruptions capturing human activity variability and radar signal processing complexities has been applied to indoor and aquatic HAR tasks. CNNs have demonstrated superior accuracy and robustness among the test models, but overfitting is a concern with deep CNNs. Strategies like CVD transform, adversarial training, and data augmentation have been explored. The obtained results have demonstrated that CVD representation provides marginal gains for indoor HAR and decreases the robustness for aquatic HAR. Adversarial training slightly reduces accuracy but improves robustness. Data augmentation effectively enhances robustness and mitigates overfitting in deep CNNs. Combining adversarial training and data augmentation has been shown to achieve a balanced trade-off between accuracy and corruption robustness. Future work will focus on refining data augmentation and model architecture for improved robustness in radar HAR tasks.

REFERENCES

[1] S. Tang, R. Gong, Y. Wang, A. Liu, J. Wang, X. Chen, F. Yu, X. Liu, D. Song, A. Yuille, *et al.*, "Robustart: Benchmarking robust-

ness on architecture design and training techniques," *arXiv preprint arXiv:2109.05211*, 2021.
[2] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," *arXiv preprint arXiv:1805.12152*, 2018.
[3] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang, "Understanding and mitigating the tradeoff between robustness and accuracy," *arXiv preprint arXiv:2002.10716*, 2020.
[4] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.
[5] C. Yi, S. Yang, H. Li, Y.-p. Tan, and A. Kot, "Benchmarking the robustness of spatial-temporal models against corruptions," *arXiv preprint arXiv:2110.06513*, 2021.
[6] O. F. Kar, T. Yeo, A. Atanov, and A. Zamir, "3d common corruptions and data augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18963–18974, 2022.
[7] J. Ren, L. Pan, and Z. Liu, "Benchmarking and analyzing point cloud classification under corruptions," *arXiv preprint arXiv:2202.03377*, 2022.
[8] S. Li, Z. Wang, F. Juefei-Xu, Q. Guo, X. Li, and L. Ma, "Common corruption robustness of point cloud detectors: Benchmark and enhancement," *arXiv preprint arXiv:2210.05896*, 2022.
[9] J. Sun, Q. Zhang, B. Kailkhura, Z. Yu, C. Xiao, and Z. M. Mao, "Benchmarking robustness of 3d point cloud recognition against common corruptions," *arXiv preprint arXiv:2201.12296*, 2022.
[10] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.
[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.
[14] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-doppler effect in radar: phenomenon, model, and simulation study," *IEEE Transactions on Aerospace and electronic systems*, vol. 42, no. 1, pp. 2–21, 2006.
[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
[16] Y. Zhou, L. Liu, H. Zhao, M. López-Benítez, L. Yu, and Y. Yue, "Towards deep radar perception for autonomous driving: Datasets, methods, and challenges," *Sensors*, vol. 22, no. 11, p. 4208, 2022.
[17] S. Yang, J. Le Kernec, O. Romain, F. Fioranelli, P. Cadart, J. Fix, C. Ren, G. Manfredi, T. Letertre, I. D. H. Sáenz, *et al.*, "The human activity radar challenge: Benchmarking based on the 'radar signatures of human activities' dataset from glasgow university," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 1813–1824, 2023.
[18] I. Ullmann, R. G. Guendel, N. C. Kruse, F. Fioranelli, and A. Yarovoy, "A survey on radar-based continuous human activity recognition," *IEEE Journal of Microwaves*, 2023.
[19] S. Abdulatif, K. Armanious, F. Aziz, U. Schneider, and B. Yang, "Towards adversarial denoising of radar micro-doppler signatures," in *2019 International Radar Conference (RADAR)*, pp. 1–6, IEEE, 2019.
[20] Z. Yang, Y. Zhang, K. Qian, and C. Wu, "Slnet: A spectrogram learning neural network for deep wireless sensing," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pp. 1221–1236, 2023.
[21] K. Patel, W. Beluch, K. Rambach, A.-E. Cozma, M. Pfeiffer, and B. Yang, "Investigation of uncertainty of deep learning-based object classification on radar spectra," in *2021 IEEE Radar Conference (RadarConf21)*, pp. 1–6, IEEE, 2021.
[22] M. Czerkawski, C. Clemente, C. Michie, I. Andonovic, and C. Tachtatzis, "Robustness of deep neural networks for micro-doppler radar classification," in *2022 23rd International Radar Symposium (IRS)*, pp. 480–485, IEEE, 2022.
[23] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019.
[24] E. Rusak, L. Schott, R. S. Zimmermann, J. Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel, "A simple way to make neural networks robust against diverse image corruptions," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 53–69, Springer, 2020.
[25] A. Modas, R. Rade, G. Ortiz-Jiménez, S.-M. Moosavi-Dezfooli, and P. Frossard, "Prime: A few primitives can boost robustness to common corruptions," in *European Conference on Computer Vision*, pp. 623–640, Springer, 2022.
[26] J. Sun, A. Mehra, B. Kailkhura, P.-Y. Chen, D. Hendrycks, J. Hamm, and Z. M. Mao, "A spectral view of randomized smoothing under common corruptions: Benchmarking and improving certified robustness," in *European Conference on Computer Vision*, pp. 654–671, Springer, 2022.
[27] Z. Zhang, D. Meng, L. Zhang, W. Xiao, and W. Tian, "The range of harmful frequency for dnn corruption robustness," *Neurocomputing*, vol. 481, pp. 294–309, 2022.
[28] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge, "Improving robustness against common corruptions by covariate shift adaptation," *Advances in neural information processing systems*, vol. 33, pp. 11539–11551, 2020.
[29] E. Mintun, A. Kirillov, and S. Xie, "On interaction between augmentations and corruptions in natural corruption robustness," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3571–3583, 2021.
[30] K. Kireev, M. Andriushchenko, and N. Flammarion, "On the effectiveness of adversarial training against common corruptions," in *Uncertainty in Artificial Intelligence*, pp. 1012–1021, PMLR, 2022.
[31] S. Hooker, A. Courville, G. Clark, Y. Dauphin, and A. Frome, "What do compressed deep neural networks forget?," *arXiv preprint arXiv:1911.05248*, 2019.
[32] L. Timpl, R. Entezari, H. Sedghi, B. Neyshabur, and O. Saukh, "Understanding the effect of sparsity on neural networks robustness," *arXiv preprint arXiv:2206.10915*, 2022.
[33] M. A. Richards, *Fundamentals of radar signal processing*. McGraw-Hill Education, 2022.
[34] N. Kern, J. Aguilar, P. Schoeder, and C. Waldschmidt, "Improving the robustness of automotive gesture recognition by diversified simulation datasets," in *2023 IEEE Radar Conference (RadarConf23)*, pp. 1–6, IEEE, 2023.
[35] J. Nilsson and T. Akenine-Möller, "Understanding ssim," *arXiv preprint arXiv:2006.13846*, 2020.
[36] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
[37] F. Fioranelli, S. A. Shah, H. Li, A. Shrestha, S. Yang, and J. L. Kernec, "Radar signatures of human activities," 2019.
[38] X. Yu, Z. Cao, Z. Wu, C. Song, J. Zhu, and Z. Xu, "A novel potential drowning detection system based on millimeter-wave radar," in *2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 659–664, IEEE, 2022.

Yi Zhou received the B.Eng. (Hons.) degree in engineering mechanics from Tianjin University, Tianjin, China, in 2016, and the M.Sc. degree in advanced control and system engineering from the University of Manchester, UK, in 2017, and the M.Sc. degree in computer vision, robotics and machine learning from University of Surrey, UK, in 2020. He is currently with the the Institute of Deep Perception Technology, JITRI, China. His research interests include millimeter-wave radar signal processing and deep learning.

Xuliang Yu received the B.S. degree in electronic and information engineering from Taiyuan University of Technology, Taiyuan, China, in 2021. He is currently pursuing the Ph.D. degree in electronic Information, Zhejiang University, Hangzhou. His research interests include millimeter-wave radar signal processing and deep learning.

Miguel López-Benítez (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees (Hons.) in telecommunication engineering from Miguel Hernández University, Elche, Spain, in 2003 and 2006, respectively, and the Ph.D. degree (summa cum laude) in telecommunication engineering from the Technical University of Catalonia, Barcelona, Spain, in 2011. From 2011 to 2013, he was a Research Fellow with the Centre for Communication Systems Research, University of Surrey, Guildford, U.K. In 2013, he became a Lecturer (Assistant Professor) with the Department of Electrical Engineering and Electronics, University of Liverpool, U.K., where he has been a Senior Lecturer (Associate Professor), since 2018. His research interests include wireless communications and networking, with special emphasis on mobile communications, dynamic spectrum access, and the Internet of Things.

Limin Yu (Member, IEEE) received the B.Eng. degree in telecommunications engineering and the M.Sc. degree in radio physics/underwater acoustic communications from Xiamen University, China, in 1999 and 2002, respectively, and the Ph.D. degree in telecommunications engineering from The University of Adelaide, Australia, in 2007. She worked with ZTE Telecommunications Company, Shenzhen, China, as a Software Engineer. She also worked with South Australia University and The University of Adelaide as a Research Fellow and a Research Associate. She has been with Xi'an Jiaotong-Liverpool University (XJTLU), since 2012; and is currently an Associate Professor. Her research interests include sonar detection, wavelet analysis, sensor networks, coordinated multi-AGV systems design, and medical image analysis.

Yutao Yue (Member, IEEE) received the B.S. degree in applied physics from the University of Science and Technology of China, in 2004, and the M.S. and Ph.D. degrees in computational physics from Purdue University, USA, in 2006 and 2010, respectively. From 2011 to 2017, he worked as a Senior Scientist with the Shenzhen Kuang-Chi In-stitute of Advanced Technology and a Team Leader of the Guangdong "Zhujiang Plan" 3rd Introduced Innovation Scientific Research Team. From 2017 to 2018, he was a Research Associate Professor with the Southern University of Science and Technology, China. Since 2018, he has been the Founder and the Director of the Institute of Deep Perception Technology, JITRI, Jiangsu, China. Since 2020, he has been working as an Honorary Recognized Ph.D. Advisor of the University of Liverpool, U.K., and Xi'an Jiaotong-Liverpool University, China. He is the co-inventor of over 300 granted patents of China, USA, and Europe. He is also the author of over 20 journals and conference papers. His research interests include computational modeling, radar vision fusion, perception and cognition cooperation, artificial intelligence theory, and electromagnetic field modulation. Dr. Yue was a recipient of the Wu WenJun Artificial Intelligence Science and Technology Award in 2020.