IMPROVING PERFORMANCE WITH FEATURE ENHANCEMENT AND RANKING CONSTRAINTS FOR RADAR-BASED HUMAN ACTIVITY RECOGNITION

Yi Zhou, [1,2,3,5]***, Miguel López-Benítez*** [5]***, Limin Yu*** [2*]***, Yutao Yue*** [134*]

[1]Institute of Deep Perception Technology, JITRI, Wuxi 214000, China
[2] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[3]XJTLU-JITRI Academy of Industrial Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[4]Department of Mathematical Sciences, University of Liverpool, Liverpool L69 7ZX, UK
[5]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 7ZX, UK
**Email: limin.yu@xjtlu.edu.cn; yueyutao@idpt.org*

Keywords: HUMAN ACTIVITY CLASSIFICATION, DEEP LEARNING, MICRO-DOPPLER

Abstract

Radar-based human activity recognition has been extensively studied using deep learning models. To better suit embedded devices, it is essential to design models that are small in size and computationally efficient. This research paper presents a lightweight model architecture specifically tailored for processing radar micro-Doppler spectrograms. The key contributions of this work include the introduction of a lightweight front-end that employs symmetric depthwise convolution and spectral pooling to enhance features and perform temporal downsampling. In our encoder design, we treat the spectrogram as a multidimensional temporal sequence and exclusively utilize 1D convolution in our model. Additionally, the paper introduces a two-branch temporal modeling module: one branch utilizes attentional LSTM to extract and summarize temporal features, while the other employs a 1D FCN to aggregate features along the temporal dimension. Furthermore, the paper applies ranking constraints to the output of the attentional LSTM, effectively leveraging temporal order information. Experimental results conducted on three HAR datasets validate the effectiveness of the proposed model, with significant performance improvements attributed to the spectral pooling layer. This highlights the importance of addressing temporal redundancy in spectrograms for improved recognition accuracy. The code is available at https://github.com/ZHOUYI1023/ConvLSTM-for-RadarHAR.

1 Introduction

Human activity recognition (HAR) involves the task of identifying and categorizing human actions and behaviors using data from various sources, such as camera, LiDAR, radar and Wi-Fi. As the development of CMOS techniques and antenna-in-package (AiP) technology, radar sensors achieve low-cost and highly integrated, making them widely utilized in IoT applications. However, low-cost radar sensors with small apertures often face limitations in angular resolution, which can restrict their ability to accurately reconstruct human motion. Nonetheless, radar sensors can measure the superimposed Doppler velocity of different parts of the human body. This spatial-temporal variation in the Doppler pattern serves as a unique motion signature for the activities of interest. To extract this motion signature, the time-frequency analysis techniques, such as short time Fourier transform (STFT), are applied. The output of this analysis is referred to as a micro-Doppler spectrogram.

The micro-Doppler spectrogram of human activities exhibits specific characteristics. For example, the moving torso typically exhibits a narrow low-frequency band, while the limbs in motion produce wider Doppler spreads and higher frequencies. Limbs with smaller scattering areas tend to have weaker high-frequency components compared to the torso. Furthermore, the motion is time-dependent, and oscillating limbs introduce periodic motion patterns. Traditional methods involve extracting hand-crafted features, such as Doppler bandwidth and temporal period, to train machine learning classifiers. With the advent of deep learning, neural network models have become a more effective approach for classifying these spectrograms [1].

There are two distinct approaches to handling the radar spectrogram. The first approach treats the spectrogram as an M × N image and utilizes CNNs for processing. However, due to the small dataset size and the sparsity of the micro-Doppler spectrogram, 2D convolutional networks often rely excessively on local patterns and are prone to overfitting background cells. The second approach views the spectrogram as a multi-dimensional time series with M frames of N-dimensional Doppler features. This approach involves two main components: the feature encoder and temporal modeling. The feature encoder typically employs 1D convolutions that slide along the temporal dimension, and then frame-wise feature vectors are aggregated in the temporal modeling module The temporal summarization approaches include LSTM-FCN [2] and temporal feature pooling [3]. However, the performance of this second paradigm is comparatively

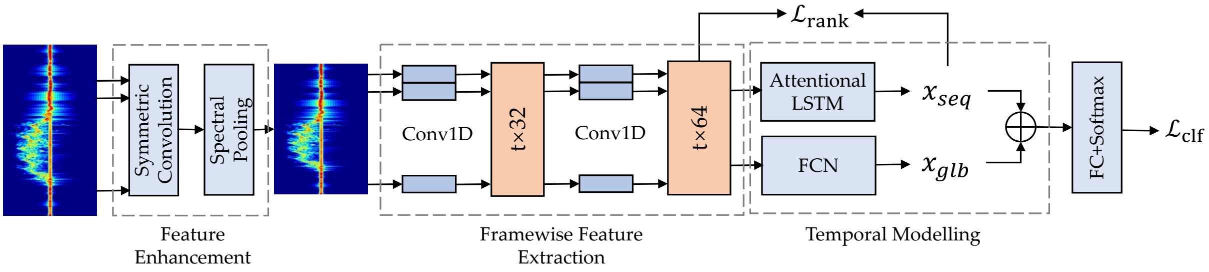


Figure 3.1 Our proposed model architecture

weaker. This can be attributed to two factors: the limited capacity of the model, which makes it less robust to noisy input, and the challenge of modeling temporal dynamics over long sequences.

To address these issues, we improve the classical Conv1D-LSTM framework in two key ways. First, we propose a lightweight feature enhancement front end to effectively improve the Signal-to-Noise Ratio (SNR) of the input and reduce temporal redundancy. In the temporal modeling module, we introduce a rank loss to encourage the LSTM to better capture motion dynamics rather than overfitting to local patterns. The contributions can be summarized as follows:

- We propose a light-weight front-end comprising symmetric depthwise convolution and spectral pooling for feature enhancement and temporal downsampling.
- We design a two-branch temporal modelling module. One branch utilizes an attentional LSTM to extract and summarize the temporal features, and the other employs 1D FCN to aggregate features over the temporal dimension. We further apply ranking constraints to the representation extracted by the attentional LSTM to effectively leverage temporal order information.
- Our experimental results on three HAR datasets confirm the effectiveness of our proposed model. The performance improvements resulting from the spectral pooling layer underscore the significance of addressing temporal redundancy in the spectrogram.

The remaining sections of this article are organized as follows: Section 2 provides an introduction to related works. Section 3 provides a detailed explanation of the proposed network. Section 4 offers a description of the datasets for benchmarking. The experimental results and analysis are presented in Section 5. Finally, Section 6 serves as the conclusion of this paper.

2 Related Works

As radar sensors find applications in IoT devices, there is a growing emphasis on developing lightweight model architectures due to the constraints of power consumption and cost sensitivity. These approaches explore various data representations, focusing on innovations such as attention mechanisms and feature enhancement modules. Zhu *et al.*[4] utilize the micro-Doppler spectrogram as input and use depthwise convolution to extract channel-wise features, along with pointwise convolutions to summarize features from each Doppler channel. Lai *et al.* [5] construct a two-branch module with 1D Convolution and attention mechanisms. This module is designed to extract two 1D features, one along the time axis and the other along the frequency axis of the spectrogram. The extracted features are then expanded to recover the input size, added together, and subjected to a softmax operation to generate an attention map. Ding *et al.* [6] take a different paradigm, extracting sparse point clouds from the micro-Doppler spectrogram as input and employing the PointNet model for classification. Ding *et al.* [7] consider the use of the 3D range-Doppler-time tensor as input data. Their method incorporates spatial attention mechanisms to enhance features in the range-Doppler dimension, utilizes conv-LSTM to extract frame-wise features, and employs temporal attention mechanisms to enhance temporal features. Yang *et al.* [8] extract two kinds of Doppler information from the Range-Time map: microscopic Doppler, estimated using Gaussian kernel estimation, and macroscopic Doppler, consisting of trajectories of peaks fitted using Lagrangian trajectory estimation. They construct a lightweight model using a combination of MobileNet and MobileViT modules for classification.

3 Methodology

Our proposed model, depicted in Figure 3.1, comprises three main components: the feature enhancement front-end, the frame-wise feature extractor, and the temporal aggregation module. The feature enhancement front-end is divided into two parts. Firstly, it includes a lightweight convolution layer that utilizes 1D depthwise convolution and symmetrically shared weights to enhance features in a learnable manner. Secondly, there is a spectral pooling layer responsible for downsampling the input with minimal information loss. Subsequently, the feature extractor employs two layers of 1D convolution to extract frame-wise features while preserving temporal order information. Finally, the frame-wise features are sent to the temporal aggregation module, which consists of two branches. The first branch employs a 1D FCN to summarize the temporal features, while the second branch utilizes an attentional LSTM to model temporal relationships and summarize temporal features. The features from both branches are concatenated and passed to the classification head. It's worth noting that we impose a ranking constraint to ensure the final representation encodes temporal order information effectively.

3.1 Feature Enhancement Front-End

Large-capacity neural networks are prone to overfitting noisy patterns, making feature enhancement crucial in radar HAR tasks. The core idea involves mitigating redundant or noisy information while highlighting critical patterns within the signal. Instead of employing hand-crafted front ends, we propose a lightweight front-end with minimal parameters. Our feature enhancement front-end serves two primary functions: firstly, we utilize a symmetric depthwise convolution layer to enhance important features, and secondly, we incorporate a spectral pooling layer to reduce temporal redundancy while preserving essential information. The result of this front-end processing is an enhanced input data with downsampled temporal dimension.

3.1.1 Symmetric Convolution: To enhance critical patterns and avoid overfitting, we design a lightweight front-end with very few parameters. When considering an input size of D×T, where D represents the size of the Doppler channel, and T represents the temporal length, popular feature enhancement methods such as the self-attention module can be computationally intensive for long sequences, with a complexity of $O(T^2D)$. However, for front-end modules, heavy computations should be avoided. In this work, we propose a light-weight module based on symmetric depthwise convolution. We firstly use 1D depthwise convolution to efficiently extract Doppler-channel-wise information. As shown in Figure 3.2 (a), depthwise convolution uses a $1 \times c$ filter sliding along temporal dimension for each Doppler channel to extract channel-wise features. For output cells at frequency index f and temporal index t, the operation can be represented as

$$X'(t,f) = w_f * X(t,f) \qquad (3.1)$$

where w_f is a normalized weight with size of $1 \ \times c$, where in our case, c is set to 3. These weights are trainable and undergo normalization across the temporal dimension through a softmax operation.

Given the symmetric characteristics of the micro-Doppler spectrogram in the Doppler dimension, we further propose a symmetric weight-sharing mechanism to reduce the model size. As depicted in Figure 3.2 (b), we group every two symmetric Doppler channels around the zero Doppler axis and assign them the same filter. This grouping strategy effectively reduces the number of parameters by a factor of $D/2$, contributing to a more compact layer.

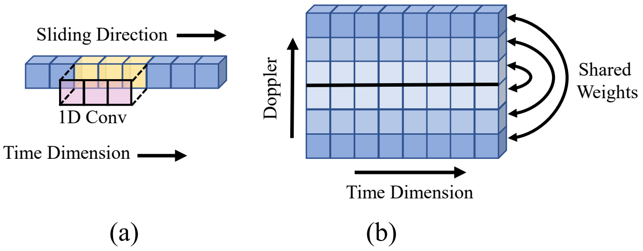


Figure 3.2 Symmetric depthwise convolution

3.1.2 Spectral Pooling: For the audio spectrogram classification task, recent findings [9] suggests that downsampling along the temporal dimension has a negligible impact on performance. Similarly, to address the issue of temporal redundancy within radar spectrograms and reduce computational complexity, we employ spectral pooling [10] as a method of adaptive downsampling along the temporal dimension.

The approach begins by applying the Discrete Hartley Transform (DHT) [11] along the temporal dimension, resulting in a frequency representation denoted as

$$Y = \text{DHT}(X) \in C^{T\times F} \qquad (3.2)$$

The DHT is a Fourier-related transform of discrete, periodic data, similar to the discrete Fourier transform (DFT). The difference is DHT transforms real inputs to real outputs without involving complex numbers, simplifying gradient computation in pytorch framework. Subsequently, a low-pass filter is applied to remove high-frequency components, yielding $Y_{crop} \in C^{kT\times F}$, where k is the compression factor. Finally, we perform the inverse DHT to transform the data back to the temporal domain, resulting in the downsampled input as

$$X_{down} = \text{IDHT}(Y_{crop}) \in R^{kT\times F} \qquad (3.3)$$

Since the DHT can be considered as a linear mapping, the backpropagation process can be directly derived. Compared to other pooling methods, spectral pooling has advantages of information preservation and interpretability. The degree of information loss can be quantified by the Parseval's Theorem, which establishes a connection between the energy loss in spatial and frequency domains. Furthermore, spectral pooling exhibits the capability to accommodate inputs of arbitrary sizes while generating fixed-sized outputs, effectively decoupling input size from the model architecture. In our model design, we place the spectral pooling layer after the symmetric depthwise convolution layer. This configuration ensures that feature enhancement operates at a finer resolution, with downsampling carried out afterward to effectively eliminate unimportant information along the temporal dimension.

3.2 Frame-wise Feature Extraction

In the next step, we employ 1D convolution to extract frame-wise features from the enhanced spectrogram. Unlike 2D convolution, which slides over both time and frequency dimensions, 1D convolution utilizes a set of convolutional kernels with a size of $D \times W$, where D is the length of Doppler channel, and W is the kernel width. These kernels slide along the temporal dimension, effectively preserving the temporal resolution. Several reasons support the choice of 1D convolution for the spectrogram data. Firstly, 2D convolution breaks down the temporal relationship by extracting translational-invariant features, potentially leading to overfitting to local patterns regardless of their position in the spectrogram. However, for micro-Doppler spectrograms, the temporal order of the motion pattern can be important for classifying an activity. Secondly, given the sparsity of radar data, 2D convolutions often convolve with background cells, resulting in unnecessary computational costs. Moreover, the receptive field of 2D convolution is usually limited in size, requiring a deep hierarchical design to enhance the receptive field, consequently increasing the model's size and complexity.

In our model, we design a two-layer convolutional architecture for feature extraction. Each layer comprises 1D convolution, followed by batch normalization and ReLU activation. The first convolutional layer involves 32 kernels with a kernel size of 5 and a stride of 1, followed by max pooling with a kernel size of 2 to halve the size. The subsequent convolutional layer increases the number of channels from 32 to 64 using kernels with a size of 3 and a stride of 1. Our experiments indicate that this two-layer architecture is sufficient to achieve strong performance.

3.3 Temporal Modelling

In the feature extractor, we obtain frame-wise information using 1D convolution with a small fixed temporal context. For the activity classification task, it's essential to aggregate these frame-wise features to summarize the sequential information effectively. In our model design, we adopt a two-branch architecture for temporal feature aggregation.

3.3.1 Attentional LSTM Branch: The LSTM serves as an effective way for modelling the temporal relationship between frames. The meaningful sequential representations are extracted as hidden features while the unimportant information is forgotten. In contrast to directly using the last hidden vector to represent the sequence, we employ an attention mechanism to summarize all the hidden features. This attention module dynamically assigns weights to different hidden states and then sum all the weighted features. This mechanism allows the model to selectively focus on the most relevant elements, effectively enhancing its ability to handle long sequences and capture complex motion patterns. Given an input feature sequence $F = (f_1, f_2, \ldots, f_T)$, where T is the length of the input sequence. The attentional LSTM computes the sequence feature f_{seq} as following:

$$e_t = \text{softmax}(W_a \tanh(W_x F + W_h h_{t-1})) \qquad (3.4)$$

$$f_{seq} = \sum_{i=1}^{T} e_{t,i} h_i \qquad (3.5)$$

where h_i is the hidden state, W_a, W_x, and W_h are learnable weight matrices and $e_{t,i}$ represents the attention weights.

3.3.2 Ranking Constraints: In the context of attentional-LSTM, a potential issue arises from the loss of temporal ordering information in the final sequential representation due to the weighted sum operations. This loss of temporal information can be problematic when classifying complex motion patterns like swimming styles. To address this challenge, we draw inspiration from previous work and apply ranking constraints to the sequence feature. The notation of ranking constraints comes from the convex optimization [12] and is adopted in video understanding [13,14]. As shown in Figure 3.3, the intuition is to find a direction in which the frame-wise features exhibit an ordered structure. Constraining the rank of a feasible solution can be thought of as introducing a linear objective function whose normal opposes the direction of search. In our cases, we relax the condition by regulating the angle between feature vectors with a margin. Since in high dimensional space the angle between vectors is hard to determine, we employ feature smoothing by calculating the cumulative sum of vector elements along the time dimension and dividing it by the corresponding accumulated time steps. Given the smoothed features, we introduce an additional ranked loss as

$$\text{ranked loss} = \min_\theta \sum_t \text{softplus}(\zeta_t) \qquad (3.6)$$

$$\zeta_t = \langle f_{seq}, \tilde{f}_{t-1} \rangle + \beta - \langle f_{seq}, \tilde{f}_t \rangle \qquad (3.7)$$

where f_{seq} denotes the sequence feature output from the attentional-LSTM and $\tilde{f}_t$ is the smoothed hidden feature at frame t, β indicates the margin, ⟨ ⟩ denotes the inner product operator. During the training, this ranked loss is added to the cross entropy loss with a predetermined weight.

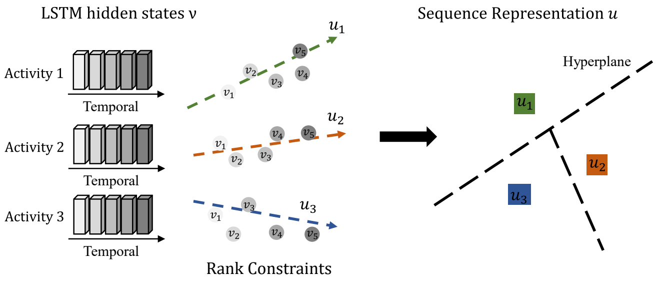


Figure 3.3 Ranking constraint (adapted from [12])

3.3.3 FCN Branch: In additional to LSTM for modelling the sequential dependencies, similar to [2], we adopt an FCN branch to summarize the temporal features over the channel dimension. Firstly, we transpose the temporal features so that the 1D convolution can slide along the channel dimension. The FCN module consists of two 1D convolution layers with filter sizes of 8 and 64. Subsequently, the feature map is converted into a global feature vector through global average max pooling. This global feature vector is then concatenated with the sequence feature extracted from the attentional LSTM. Finally, the concatenated feature is fed into a fully connected layer for classification.

4 Datasets and Experiment Settings

A significant challenge in evaluating these methods lies in the absence of publicly available benchmarks for comparing algorithms. Each method is assessed and compared using self-collected datasets with varying levels of difficulty. Additionally, the quality of the spectrogram is influenced by both the radar sensor used and the signal processing pipeline employed. In this work, we aim to address this challenge by benchmarking models on three different public datasets that provide raw data. We also unify the signal processing pipeline to minimize the impact of signal processing on the evaluation.

4.1 Datasets

Three datasets are chosen for benchmarking. The first is the Glasgow indoor HAR dataset [15], 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.

The second dataset [16] also aims at indoor HAR. The datasets provide three sensors, including UWB radar, 24 GHz radar and 77 GHz radar. In this project, we only use the data recorded by the 77GHz radar. The radar is a 77GHz radar with 0.7675 GHz bandwidth and 0.3125 ms chirp duration. Compared to the first one. This dataset covers 11 classes of activities, including walking towards radar, walking away from radar, picking up an object, bending, sitting, kneelling, crawling, walking on both ties, limping with stiff, short steps and scissors gait. The walking direction and some similar activities are considered. Each participant conducted 10 repetitions of each activity, resulting in a total of 60 samples per class per sensor.

The third dataset [17] focuses on aquatic human activity recognition. It employs a 77 GHz FMCW radar with a 1.7 GHz bandwidth and 0.33 ms chirp duration. Nine-class aquatic human activities: struggle, drowning, float with buoys, wave for help, pull buoys, swim with buoys, backstroke, breaststroke and 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 discriminating different activities.

Table 1 Radar Configurations

Model	Glasgow	CI4R	Aquatic
Operating Frequency	5.8 GHz	77 Ghz	77 Ghz
Bandwidth	0.4 GHz	0.77 GHz	1.7 GHz
Chirp Time	1 ms	0.3125 ms	0.33 ms
# ADC per Chirp	128	256	256
# Chirp per Frame	128	256	128

4.2 Unified Signal Processing

To mitigate the effects of the signal processing pipeline, we standardize the pipeline to process these datasets. We retain only one channel from the raw ADC data, without considering the angle information. We conduct a 256-point FFT along the range dimension, followed by a 4th order Butterworth filter for MTI processing. Finally, we conduct STFT with a 128-point window size, spanning 128 samples, and an overlap ratio of 0.95 along the Doppler dimension to extract the micro-Doppler spectrogram. Some studies save the micro-Doppler as a color image by applying a cutoff filter to the log-spectrogram and mapping the intensity to RGB space using a color map. However, we argue that saving the spectrogram as an RGB image is an unnecessary process that leads to information loss. Instead, we directly encode the log-spectrogram into a normalized map. The color map is only used for visualization.

4.3 Experiment Settings

Our approach utilizes raw micro-Doppler spectrograms as inputs and performs input normalization using precomputed mean and standard deviation values derived from all training samples. In addition to the proposed model, our evaluation includes four additional architectures: two CNN models, including a light-weight VGG [18] and ResNet-18 [19]; two LSTM-based models, including CRNN [20] and plain Conv1D-LSTM [21]. The optimization process employs the Adam optimizer with a learning rate of either 1e-3 or 1e-4. To adaptively adjust the learning rate when the validation loss plateaus, we utilize the ReduceLROnPlateau scheduler to enhance convergence. We split the dataset into 80% training and 20% testing sets. Due to the small size of the first two datasets, we apply ten-fold cross-validation to better utilize the available training data. The selection of the best model is determined by monitoring the validation loss, and we incorporate early stopping with a patience of 5 epochs, meaning that training halts if the validation loss does not improve within this specified time window.

5 Result Analysis

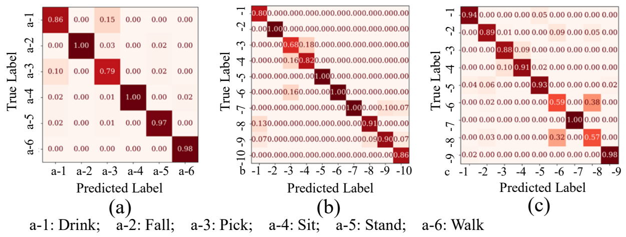
5.1 Classification Performance

In Table 2, we present the classification results for three selected datasets. The table reveals that our model achieves the best performance in the aquatic HAR dataset and the second-best result in the CI4R dataset. All the models perform comparably in the Glasgow HAR dataset, suggesting that this dataset may be too simple to provide a fair evaluation of model performance. Notably, the CNN models exhibit significantly lower accuracy in the CI4R dataset, which is a small dataset with many classes. This suggests that CNN models are prone to overfitting in such cases, while the LSTM models demonstrate markedly improved performance. In contrast, for the aquatic HAR dataset with its complex motion patterns, the CNN models outperform the LSTM models. Importantly, our proposed models offer significantly enhanced efficiency while maintaining comparable accuracy in this dataset compared to the CNN models, all while having fewer parameters and FLOPs.

Table 2 Classification Performance

Model	Accuracy (%)			Params	FLOPs
	GLA	CI4R	AQUA	(G)	(M)
CRNN	94.53	90.28	82.46	0.415	0.896
VGG7	93.49	76.84	85.56	2.095	0.298
ResNet18	95.83	78.25	88.73	1.824	11.180
ConvLSTM	94.53	86.33	80.31	0.008	0.111
Our Model	93.49	89.57	88.67	0.010	0.127

In Figure 5.1, we further inspect the class-wise performance using confusion matrices. In the first plot, we observe that although "drink" and "sit" activities can be occasionally confused, our model successfully classifies all the other activities with high accuracy. In the second plot, we notice that distinguishing between "picking up" and "bending" poses a challenge, and different types of walking activities, such as "walking towards radar" and "short steps," are prone to being misclassified into other walking types. Finally, in the third plot for the aquatic dataset, we find that "pull buoy" and "swim with buoy" activities are challenging to discriminate, while the model performs well with the other activities. Notably, different swimming styles are successfully distinguished from each other, showcasing the model's effectiveness in handling intricate patterns in aquatic activities.



a-1: Drink; a-2: Fall; a-3: Pick; a-4: Sit; a-5: Stand; a-6: Walk
b-1: Walking Towards Radar; b-2:Walking Away From Radar; b-3:Picking Up An Object;
b-4: Bending; b-5: Sitting; b-6: Kneeling; b-7: Crawling; b-8: Walking On Both Toes;
b-9: Limping With Stiff; b-10: Short Steps .
c-1: Backstroke; c-2: Breaststroke; c-3: Float; c-4: Float with Buoy; c-5: Freestyle;
c-6: Pull Buoy; c-7: Struggle; c-8: Swim with Buoy; c-9: Wave with Buoy

Figure 5.1 Confusion matrices

5.2 Ablation Study

In Table 3, we present the results of our ablation study. The symmetric depthwise convolution layer exhibits slight improvements on the latter two datasets while minimizing the increase in model size and computation. Since the design of encoder layers has been extensively discussed in [20], our focus here is on investigating the impact of the temporal modeling module. From the results, it becomes evident that removing the LSTM module and relying solely on the FCN module for sequence summarization leads to significant performance degradation. Specifically, we observe a 5.99% drop on the Glasgow dataset, an 8.67% drop on the CI4R dataset, and an 8.59% drop on the aquatic dataset, underscoring the importance of temporal modelling. If we remove the FCN branch, we witness a 5.21% performance drop on the CI4R dataset and a 2.73% drop on the aquatic dataset. Replacing the FCN with a nonparametric max-pooling layer results in a 2.73% performance decrease on the aquatic dataset, with minimal impact on the other two datasets. This reaffirms the crucial role of our two-branch design in aggregating temporal features. The FCN module appears to be particularly effective for complex motion patterns due to its increased model capacity. When we remove the attention mechanism and use the last hidden feature as the output, we observe a substantial 10.27% performance drop on the CI4R dataset, highlighting the significance of the attention mechanism for long sequence classification. Furthermore, the inclusion of ranking constraints contributes to performance improvement without increasing computation, as these constraints modify the training loss.

Table 3 Ablation Study

Model	Accuracy (%)			Params	FLOPs
	GLA	CI4R	AQUA	(M)	(K)
Full	93.49	89.57	88.67	10.390	126.65
rmSymConv	95.53	88.83	87.11	10.390	126.65
rm FCN	93.49	83.36	85.74	9.933	117.93
FCN2Pool	96.35	89.50	85.94	9.933	118.25
rm LSTM	87.50	80.90	80.08	5.470	51.01
rm Attention	94.53	79.30	86.13	10.386	126.59
rmRankLoss	95.31	87.23	85.94	10.390	126.65

5.3 Effect of the Feature Enhancement Module

Since the ablation study does not fully reflect the importance of the feature enhancement module, we conduct further analysis through feature map visualization. In Figure 5.2(a), we observe that the trained weights of the depthwise kernel exhibit patterns such as descending and periodical motion. To illustrate this, we provide an example using a spectrogram of freestyle and display the difference map between the test image and the enhanced image in Figure 5.2(c). This visualization highlights that critical motion patterns are effectively enhanced by the feature enhancement module.

Moreover, the feature enhancement module implicitly contributes to noise filtering due to the use of softmax. To assess this effect, we introduced Gaussian white noise with a variance of 0.1 to the test data, as depicted in Figure 5.2(d). The results in Figure 5.2(f) show that the enhanced image is less noisy, indicating the noise filtering capabilities of the module.

These findings underscore the importance of the feature enhancement module in improving the model's ability to highlight critical motion patterns and reduce the impact of noise in radar micro-Doppler spectrograms.

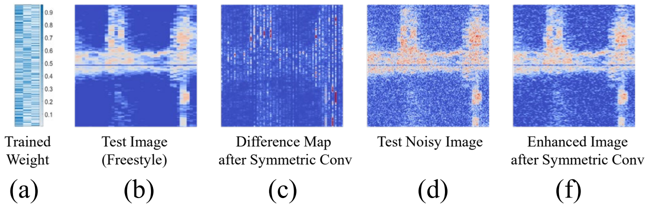


Figure 5.2 Effect of feature enhancement

In Figure 5.3, we present visualizations of the feature map after spectral pooling using different scaling factors. A consistent colormap is applied to emphasize the effect of SNR enhancement achieved by the spectral pooling layer. Upon visual inspection, we observe significant temporal redundancy in the spectrogram.

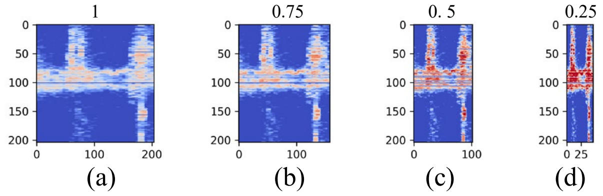


Figure 5.3 Visualization of spectral pooling

Figure 5.4 displays the accuracy changes as we modify the temporal scaling factor. The figure suggests that increasing temporal compression can improve performance, possibly due to SNR improvements and enhanced temporal modelling capabilities of the LSTM with shorter sequences. If we replace spectral pooling with adaptive max pooling or adaptive average pooling, we observe a drop in performance due to resolution loss. These findings can inspire future research on more efficient temporal information compression techniques for HAR tasks.

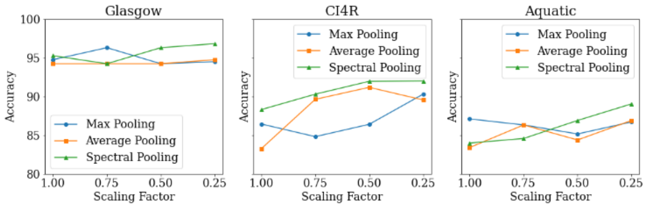


Figure 5.4 Effect of spectral pooling

5.3 Effect of the Ranking Constraints

Table 3 has already demonstrated the benefits of rank constraints on performance. To gain a deeper understanding of how these constraints work, we present the learning curve in Figure 5.5 and employ t-SNE for feature visualization in Figure 5.6, using the aquatic dataset as our example. In order to compare the learning dynamics with and without ranking loss optimization, we conducted a 50-epoch training session, setting the weight for ranking loss to 1 and 0, respectively. From the learning curve, we can find the ranking loss tends to be consistent if left unoptimized. When optimization is applied, the loss experiences a constantly slight decrease. Examining the cross-entropy loss, we observe that the ranking loss does not have a significant impact during the initial epochs. However, as the cross-entropy loss reaches a plateau, we notice that optimization of the ranking loss leads to a more pronounced decrease in cross-entropy loss compared to the unoptimized scenario.

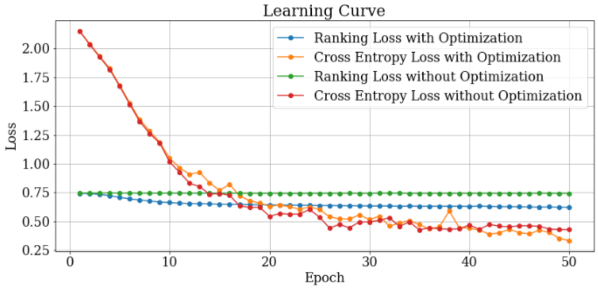


Figure 5.5 Learning curve for the loss optimization

In Figure 5.6, the t-SNE results reveal that different swimming styles are more effectively separated in the feature space after the application of ranking constraints. Given that different swimming styles often exhibit similar periodical energy distributions, the temporal order of features appears to play a more crucial role in classification.

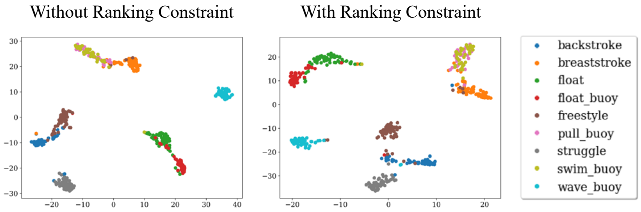


Figure 5.6 Feature visualization

6 Conclusion

In conclusion, this work introduces a lightweight model architecture tailored for radar micro-Doppler spectrogram processing. The model comprises three main components: a lightweight feature enhancement module, a 1D convolution based temporal feature extractor, and a temporal modelling module. The feature enhancement module consists of a lightweight symmetric depthwise convolution layer to emphasize important features and a spectral pooling layer to reduce temporal redundancy. Furthermore, a two-branch temporal modelling module is devised, utilizing an attentional LSTM for temporal relationship modelling and summarization, alongside a 1D FCN for feature aggregation over the temporal dimension. The incorporation of ranking constraints on the attentional LSTM's output effectively leverages temporal order information. Experimental results on three HAR datasets validate the model's effectiveness.

The experimental results particularly emphasize the significance of addressing temporal redundancy through the spectral pooling layer. Future research should focus on finding ways to better harness this temporal redundancy to further reduce model size and computational costs.

7 Acknowledgements

This work received financial support from Jiangsu Industrial Technology Research Institute (JITRI) and Wuxi National Hi-Tech District (WND).

8 References

[1] Zhou, Y., Liu, L., Zhao, H., et al.:'Towards deep radar perception for autonomous driving: Datasets, methods, and challenges', Sensors, 2022, 22(11), p.4208

[2] Karim, F., Majumdar, S., Darabi., et al.: 'LSTM fully convolutional networks for time series classification', IEEE Access, 2017, 6, pp.1662-1669

[3] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., et al.: ' Beyond short snippets: Deep networks for video classification', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Boston, USA, June 2015, pp. 4694-4702

[4] Zhu, J., Lou, X., Ye, W.: 'Lightweight deep learning model in mobile-edge computing for radar-based human activity recognition', IEEE Internet of Things Journal, 2021, 8(15), pp.12350-12359

[5] Lai, G., Lou, X., Ye, W.: 'Radar-based human activity recognition with 1-D dense attention network', IEEE Geoscience and Remote Sensing Letters, 2021, 19, pp.1-5

[6] Ding, C., Zhang, L., Chen, H.: 'Sparsity-based Human Activity Recognition with PointNet using a Portable FMCW Radar', IEEE Internet of Things Journal, 2023, 10(11), pp. 10024-10037

[7] Ding, C., Zhang, L., Chen, H., et al.: 'Human motion recognition with spatial-temporal-convLSTM network using dynamic range-doppler frames based on portable FMCW

radar', IEEE Transactions on Microwave Theory and Techniques, 2022, 70(11), pp.5029-5038

[8] Yang, X., Gao, W., Qu, X., et al.: 'A lightweight multi-scale neural network for indoor human activity recognition based on macro and micro-Doppler features', IEEE Internet of Things Journal, 2023

[9] Liu, X., Liu, H., Kong, Q., et al.: 'Simple pooling front-ends for efficient audio classification', IEEE Int. Conf. Acoustics, Speech and Signal Processing, Rhodes Island, Greece, June 2023, pp. 1-5

[10] Rippel, O., Snoek, J., Adams, R.P., et al.: 'Spectral representations for convolutional neural networks', Advances in Neural Information Processing Systems, 2015, 28

[11] Bracewell, R.N.: 'Discrete Hartley transform', JOSA, 1983, 73(12), pp.1832-1835.

[12] Boyd, S.P.,Vandenberghe, L. : 'Convex optimization', (Cambridge University Press, 2004)

[13] Fernando, B., Gavves, E., Oramas, J.M., et al.: 'Modeling video evolution for action recognition', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Boston, USA, June 2015, pp. 5378-5387

[14] Cherian, A., Wang, J., Hori, C., et al.: 'Spatio-temporal ranked-attention networks for video captioning. ', Proc. IEEE Winter Conf. Applications of Computer Vision, Snowmass, USA, March 2020, pp. 1617-1626

[15] Fioranelli, F., Shah, S. A., Li, H., et al.: 'Radar signatures of human activities', 2019, University of Glasgow

[16] Gurbuz, S.Z., Rahman, M.M., Kurtoglu, E., et al.: 'Cross-frequency training with adversarial learning for radar micro-Doppler signature classification', Proc. SPIE Radar Sensor Technology XXIV, 11408, May 2020, pp. 58-68

[17] Yu, X., Cao, Z., Wu, Z., et al.: 'A novel potential drowning detection system based on millimeter-wave radar', Int. Conf. Control, Automation, Robotics and Vision, Singapore, December 2022, pp. 659-664

[18] Fonseca, E., Favory, X., Pons, J., et al.: 'Fsd50k: an open dataset of human-labeled sound events', IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30, pp.829-852

[19] He, K., Zhang, X., Ren, S., et al.: 'Deep residual learning for image recognition', Proc. IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, USA, June 2016, pp. 770-778

[20] Shi, B., Bai, X., Yao, C.: '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, 2016, 39(11), pp.2298-2304

[21] Zhu, J., Chen, H., Ye, W.: 'A hybrid CNN–LSTM network for the classification of human activities based on micro-Doppler radar', IEEE Access, 2022, 8, pp.24713-24720