

Transparent and Resilient Activity Recognition via Attention-Based Distributed Radar Sensing

Mina Shahbazifar , Zolfa Zeinalpour-Yazdi , Matthias Hollick , Arash Asadi , and Vahid Jamali 

Abstract—Distributed radar sensors enable robust human activity recognition. However, scaling the number of coordinated nodes introduces challenges in feature extraction from large datasets, and transparent data fusion. We propose an end-to-end framework that operates directly on raw radar data. Each radar node employs a lightweight 2D Convolutional Neural Network (CNN) to extract local features. A self-attention fusion block then models inter-node relationships and performs adaptive information fusion. Local feature extraction reduces the input dimensionality by up to 480×. This significantly lowers communication overhead and latency. The attention mechanism provides inherent interpretability by quantifying the contribution of each radar node. A hybrid supervised contrastive loss further improves feature separability, especially for fine-grained and imbalanced activity classes. Experiments on real-world distributed Ultra Wide Band (UWB) radar data demonstrate that the proposed method reduces model complexity by 70.8%, while achieving higher average accuracy than baseline approaches. Overall, the framework enables transparent, efficient, and low-overhead distributed radar sensing.

Index Terms—Distributed radar sensors, activity recognition, interpretability, attention mechanism, sensor fusion.

I. INTRODUCTION

The increasing demand for intelligent ambient systems has driven significant research interest in wireless activity recognition. Various signal representation and neural network architectures have been investigated to extract meaningful and actionable insights from wireless signals [1], [2]. In particular, deep neural networks have received significant attention due to their capability to automatically learn discriminative features from intricate data [3]. Several signal modalities (e.g., WiFi and mmWave) have been investigated, with UWB radar offering superior temporal and spatial resolution due to its extremely short pulses [4]. For real-world deployment of activity recognition, especially in sensitive scenarios such

as high-risk operational zones in industrial environments, continuous sensing is essential to ensure that user activities can be reliably recognized regardless of location or orientation [5], [6]. In this regard, distributed sensing has emerged as a compelling solution for resilient activity recognition [7]. However, scaling up the number of coordinated sensing nodes introduces two key challenges: *First*, additional radars produce substantially larger data volumes and higher preprocessing costs, particularly when forming Doppler signatures or point clouds. *Second*, achieving robust and interpretable fusion becomes difficult because each radar node experiences geometry-dependent variations in signal quality due to changes in target position, orientation, and propagation conditions.

Prior work. Most distributed radar systems still depend on heavy preprocessing before learning or classification, including Range-Doppler (RD) maps [6], [8], Doppler spectra [5], [7], and point clouds [9], [10]. These Discrete Fourier Transform (DFT)-based pipelines require additional operations such as side lobe suppression [11], [12], which introduce extra computational overhead. Furthermore, such representations may discard useful raw information that hinders the ability to model geometry-dependent variability [11]. A second body of work focuses on fusion. Existing approaches typically rely on early, intermediate, or late fusion mechanisms, each with inherent limitations. Early fusion requires transmitting raw measurements from all nodes to a central processor, resulting in significant communication bandwidth and making the system highly sensitive to noisy or low-quality nodes [11], [13], [14]. Late fusion reduces bandwidth by combining high-level decisions, but restricts interaction across viewpoints and shifts computational load to individual sensors, limiting the ability to exploit complementary information [6], [7], [10]. Intermediate fusion offers a more balanced trade-off by allowing each radar to extract local features before joint integration, yet the resulting representations often lack interpretability because the contribution of individual nodes becomes entangled [5], [7], [9].

Research gap. Existing works either focuses on recognition accuracy without considering communication constraints [7], [13], or on communication efficiency for simpler tasks such as localization [14], [15]. Consequently, unified and practical distributed human activity recognition remains largely under-explored.

Our solution. To address this gap, we propose an end-to-end framework that jointly considers communication constraints (e.g., bandwidth and latency) and system-level requirements (e.g., accuracy and transparency) for practical continuous distributed activity recognition. The proposed framework

Shahbazifar, Hollick, and Jamali's work has been co-funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center [LOEWE/1/12/519/03/05.001(0016)/72]. Zeinalpour-Yazdi's work has been supported in part by Iran National Science Foundation (INSF) under Grant 4025264.

M. Shahbazifar and Z. Zeinalpour-Yazdi are with the Electrical Engineering Department, Yazd University, Yazd 8915818411, Iran (e-mail: minashahbazifar@stu.yazd.ac.ir; zeinalpour@yazd.ac.ir).

V. Jamali and M. Shahbazifar are with the Resilient Communication Systems Group, Technical University of Darmstadt, 64283 Darmstadt, Germany (e-mail: vahid.jamali@tu-darmstadt.de; m.shahbazifar@rcs.tu-darmstadt.de).

M. Hollick is with the Secure Mobile Networking Lab, Technical University of Darmstadt, 64289 Darmstadt, Germany (e-mail: mhollick@seemoo.tu-darmstadt.de) as well as with the IMDEA Networks Institute, Madrid, Spain (e-mail: matthias.hollick@networks.imdea.org).

A. Asadi is with the Wireless Communication and Sensing Lab, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: A.Asadi@tudelft.nl).

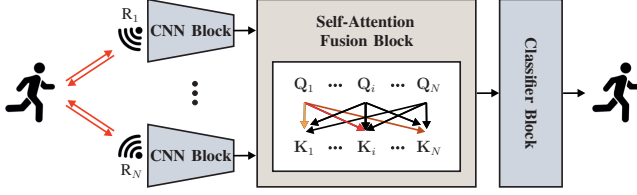


Fig. 1: The proposed distributed radar recognition model consists of a 2D CNN block for local feature extraction, a self-attention block for adaptive and interpretable fusion, and a final classification block.

improves efficiency through four key design choices: (i) eliminating the need for explicit preprocessing by operating directly on raw data, (ii) integrating feature compression into the model architecture to reduce communication cost, (iii) shifting representation learning complexity to the training stage via a hybrid loss function, and (iv) employing an attention-based intermediate fusion strategy that enables hierarchical feature aggregation while preserving interpretability through explicit node-level contribution modeling. Our main contributions are:

- **End-to-end raw-data framework.** We propose an end-to-end framework for distributed UWB radar-based continuous activity recognition that operates directly on raw radar data, eliminating costly preprocessing like Doppler signatures and point clouds. The architecture includes a 2D CNN block for local feature extraction, a self-attention block for adaptive inter-node fusion, and a fully connected block for classification (Fig. 1). Training uses a hybrid supervised contrastive loss to enhance feature separability for fine-grained, imbalanced activity classes.
- **Interpretable attention-based multi-node fusion.** We introduce an attention-based fusion mechanism for distributed radar sensing. A self-attention block explicitly models inter-node relationships and produces a weight matrix that quantifies the contribution of each radar node. Ablation experiment and comparison with GradientSHAP attribution method show that the attention-derived importance scores reflect each node's true impact on the final decision. This provides inherent interpretability that is not available in early, intermediate, or late fusion schemes.
- **Communication efficiency.** By deploying the trained lightweight 2D CNN encoder at each node, semantically rich compressed features are transmitted to the central processor instead of raw data which significantly reduces communication overhead and latency. We demonstrate the effectiveness of this approach compared to conventional downsampling schemes.

II. SYSTEM AND SIGNAL MODEL

System model. A network of N monostatic UWB radar nodes is deployed around an area of interest to perform activity recognition regardless of the location and orientation of the target. The transmission of radar signals are performed in an orthogonal manner that allows simultaneous operation. Each radar transmits pulses with the same Pulse Repetition Interval (PRI) and in synchronization with other nodes. The

radars collect the backscattered echoes, which contain motion-induced modulation characteristics of target movements from different viewpoints. Radar nodes forward their data to a fusion processor for the recognition task.

Signal model. UWB signal is typically organized as a 2D fast-time $\times$ slow-time matrix, where fast time represents range bins and slow time captures pulse-to-pulse variations, such as target motion. Let $y[n, m]$ denote the (n, m) -th element of this matrix, corresponding to the n -th fast-time sample of the m -th pulse. Following [4], the m -th transmitted pulse, with a carrier angular frequency ω_c , can be modeled as:

$$x(t, mT) = p(t, mT) \sin(\omega_c t), \quad (1)$$

where $p(t, mT)$ is the short-duration pulse and T is the PRI. The received echo at this sample, acquired with a fast-time sampling period T_s , is expressed as:

$$y[n, m] = \alpha(n, m) p(nT_s - t_D, mT) e^{j\omega_c t_D(n, m)}, \quad (2)$$

where $\alpha(n, m)$ and $t_D(n, m)$ denote the attenuation factor and propagation delay, respectively. Both quantities are influenced by the target's orientation, location, and the type of activity being performed.

III. PROPOSED FRAMEWORK

In this section, we outline the proposed end-to-end framework. We first describe the input structure, followed by the model architecture composed of three components: (i) a 2D CNN block for per-radar feature extraction, (ii) a self-attention fusion block for modeling inter-node dependencies and emphasizing informative nodes, and (iii) fully connected layers for final classification. We then present the hybrid supervised contrastive learning objective and, finally, the procedure for estimating node importance.

Model input representation. For each fixed-length slow-time window, the complex samples are converted to polar form, $(|y[n, m]|, \angle y[n, m])$, and arranged into a three-dimensional tensor with dimensions corresponding to range bins (fast time) $\times$ window length (slow time) $\times$ 2 (magnitude/phase) which enables the model to capture both spatial and temporal variations in the scene.

Model architecture. The proposed architecture models temporal dependencies hierarchically by extracting local temporal patterns using a 2D CNN from each node and capturing higher-level dependencies via a multi-head self-attention fusion mechanism across radar streams.

Feature extraction block. The fast-time and slow-time dimensions are treated as height and width, respectively, with the polar representation as the channel dimension. Three 2D CNN layers are employed to capture spatial-temporal patterns at each radar. Each layer is followed by batch normalization (BN) and ReLU activation to improve optimization stability and nonlinear representational capability. A 1×1 convolutional layer is added to enhance channel-wise feature interactions. Finally, an average pooling is employed to condense depth and spatial information into compact yet discriminative representations to enhance both computational efficiency and practical usability. Each radar node is processed through the same

block using a weight-sharing strategy [7]. This contributes to reduced model complexity and enhanced model scalability. The extracted features are first flattened into an array of size d_{model} and then concatenated across the nodes to form the output $\mathbf{S}$ with shape $N \times d_{\text{model}}$ compatible with the self-attention input, and forwarded to the fusion block. In this setting, self-attention is computationally efficient since N is small relative to d_{model} in typical indoor sensing.

Fusion block. Each radar observes the activity from a distinct viewpoint, depending on the target's location and orientation, which affects the resolution and the information it acquires. Treating the nodes as a sequence, we employ a self-attention mechanism to dynamically aggregate information across them and assign higher weights to more informative ones. At the core of this mechanism lies connecting all the nodes, which allows the model to find the hidden inter-node relation. This is achieved through three learnable linear projections that map the input features into the Query ($\mathbf{Q}$), Key ($\mathbf{K}$) and Value ($\mathbf{V}$). In this manner, the information that nodes lack can be obtained from other nodes and they complete their information from different viewpoints. By leveraging attention weights, each node focuses on the nodes that provide the most complementary information and combines their values according to the weights. In multi-head attention with H heads, rather than a single set of $\mathbf{Q}, \mathbf{K}, \mathbf{V}$, each head maintains its own set:

$$\mathbf{Q}_h = \mathbf{S}\mathbf{W}_h^Q, \quad \mathbf{K}_h = \mathbf{S}\mathbf{W}_h^K, \quad \mathbf{V}_h = \mathbf{S}\mathbf{W}_h^V. \quad (3)$$

The attention weights and outputs of each head are given by:

$$\alpha_h = \text{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{d_k}}\right), \quad \mathbf{S}_h = \alpha_h \mathbf{V}_h. \quad (4)$$

The outputs of all heads are then concatenated and linearly projected to form the fused feature matrix, $\mathbf{S}_a$. The attention weights across all heads are averaged to produce a matrix of shape $N \times N$ that represents the overall inter-node attention, α . A residual connection is added to facilitate gradient flow. The fused features are flattened into an array $\mathbf{s}_a$ to be prepared for the next block.

Classifier block. The classifier consists of two fully connected layers with a dropout layer to mitigate overfitting. A softmax function produces class probabilities, and the final output $\hat{y}$ corresponds to the class with the highest predicted probability.

Learning and optimization. Inspired by [16], we adopt a hybrid loss to improve robustness for fine-grained activity classes under imbalanced distributions:

$$\min_{\theta_c, \theta_a, \theta_d} \mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{CE}}(y, \hat{y}) + \gamma \cdot \mathcal{L}_{\text{SCL}}(y, \mathbf{s}_a), \quad (5)$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, and $\mathcal{L}_{\text{SCL}}$ is the supervised contrastive loss over subspace $\mathbf{s}_a$, to enhance feature separability by pulling embeddings of the same class together and pushing apart those of different classes. For a batch of B samples:

$$\mathcal{L}_{\text{SC}} = \frac{1}{B} \sum_{i=1}^B \frac{-1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}, \quad (6)$$

where $\mathcal{P}(i) = \{j \mid y_j = y_i, j \neq i\}$ is the set of positive samples sharing the same class as anchor i , $\mathbf{z}_i$ is the L2-normalized embedding of sample i , and τ is a temperature parameter controlling the distribution sharpness. The numerator $\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)$ encourages *intra-class cohesion*, while the denominator $\sum_{k \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)$ enforces *inter-class separation* by penalizing similarity with negative samples [17].

Attention weights and node importance. The weight matrix α encodes pairwise node interactions [18], [19]. Each row, $\alpha_{i,:}$, represents a probability distribution of the attention weights assigned by node i to other nodes, while each column, $\alpha_{:,j}$, quantifies the total attention received by node j , serving as a measure of its importance or attention centrality. The global importance of node j is defined as:

$$I(j) = \sum_{i=1}^N \alpha_{i,j}. \quad (7)$$

A higher $I(j)$ indicates greater influence on model decision making, a concept used in explainability analyses [19], [20].

IV. EXPERIMENTS AND RESULTS

In this section, we evaluate the proposed architecture. We begin by summarizing the dataset and implementation details, including network and training parameters. We then compare the model's performance against the baseline, followed by t-SNE visualizations and an ablation study on node-attribution interpretability. Finally, we examine the model's compression capability for real-time deployment.

Dataset description. We use the public dataset from [21], which includes nine human activities performed by 14 participants at varying locations and orientations, recorded using five UWB radars with 480 fast time bins. The dataset exhibits class imbalance and activity similarities, with the following distribution: 1) walking 29.7%, 2) stationary 14.8%, 3) sitting down 5.1%, 4) standing up from sitting 4.7%, 5) bending from sitting 11.8%, 6) bending from standing 12.9%, 7) falling while walking 3.3%, 8) standing up from the ground 11.6%, and 9) falling while standing 5.3%.

Training and network parameters. Different input window sizes were evaluated, and the best results were achieved with a window length of 30 which results a final input shape of $480 \times 30 \times 2$. Model parameters are detailed in in Table I.

TABLE I: Network Architecture

Block	Layer Structure
Feature Extraction (FE)	Conv2D 6, kernel (7,3), padding (3,1), BN, ReLU
	Conv2D 8, kernel (3,3), padding 1, BN, ReLU
	Conv2D 6, kernel (3,3), padding 1, BN, ReLU
	Conv2D 6, kernel 1
Fusion	AdaptiveAvgPool (5,4) $\rightarrow$ Flatten
	Concatenate ($\mathbf{FE}_1, \dots, \mathbf{FE}_N$)
	Multi-head Attention 4 heads, $d_k = d_v = 24$ Residual Connection $\rightarrow$ Flatten
Classifier	Dense layer 64, ReLU
	Dropout rate=0.3 Dense layer 9 $\rightarrow$ Softmax

Training is performed using the Adam optimizer at a learning rate of 3×10^{-3} for up to 100 epochs. Early stopping with a patience of 10 epochs is applied to prevent overfitting.

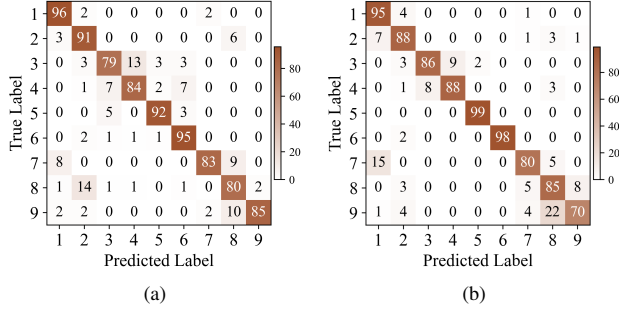


Fig. 2: Confusion matrices (in percentage) for (a) the proposed model and (b) the baseline CNN-RNN [7].

TABLE II: Performance comparison with baseline model

Model	Max. Test	Ave. Test	Input type	Params
CNN-RNN [7]	90.8%	85.1%	Doppler Spectrum	72K
Proposed method	92.53%	88.54%	Raw data	21K

The learning rate is halved if the loss doesn't improve for 10 consecutive epochs. Parameters γ and τ are set to 1 and 0.5, respectively, based on empirical observations. A leave-one-person-out (LOPO) evaluation over all participants is adopted with an 4:1 train-validation split across samples that holds one participant out as unseen test data. The best test result is used for visualization. To mitigate class imbalance when all participants are included, data augmentation is applied using additive Gaussian noise (scale = 0.1).

System performance benchmarking. We compare the proposed framework with the CNN-RNN model introduced in [7]. As shown in the confusion matrices (Fig. 2), both models achieve higher accuracy for classes with larger sample distributions. Our model performs better in several classes (e.g., 1, 2, 7, and 9), while the baseline performs better in others (e.g., 3, 4, 5, 6, and 8). Importantly, classes with opposite movements (e.g., 3 vs. 4 and 8 vs. 9) show the highest mutual misclassifications in both models. Overall, the proposed framework shows a more favorable performance trend. As reported in Table II, both the maximum and average test accuracies are improved. Notably, the proposed method reduces model complexity and eliminates costly preprocessing. Although training is more resource-intensive due to the larger dataset and hybrid optimization, this overhead is limited to offline training. In contrast, inference latency, critical for deployment, is significantly reduced: our 3.4 $\times$ smaller model achieves 0.013 s per sample versus 0.049 s for the baseline on the same Intel 11th Gen Core i5-11400H CPU.

Feature embedding visualization. We use t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the feature embeddings. As shown in Fig. 3, plot (a) indicates substantial class overlap in the raw input data; plot (b) demonstrates clear class separation in the fused feature space that highlights effective feature discrimination. In plot (c), the classifier outputs form well-defined clusters corresponding to individual classes. Consistent with Fig. 2(a), improved class separation in the t-SNE space (i.e., less overlap) corresponds to higher classification accuracy in confusion matrix. For classes

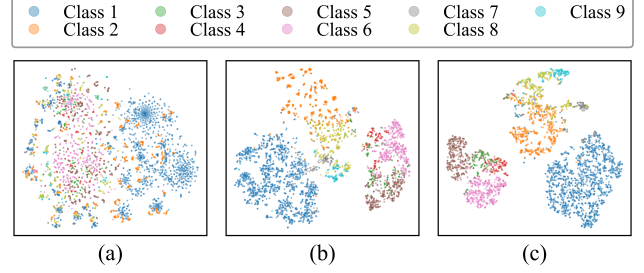


Fig. 3: t-SNE visualization of (a) raw input features, (b) fused representation, and (c) output of the final dense layer.

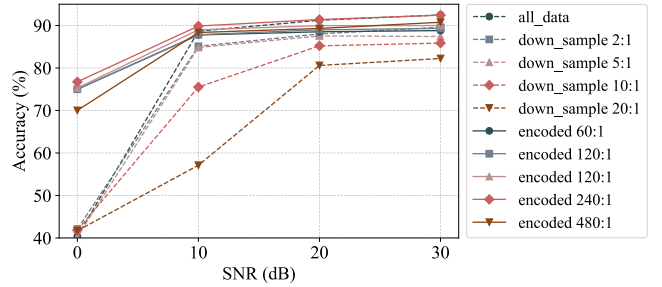


Fig. 4: Encoder-based feature compression versus conventional downsampling, evaluated under varying SNR conditions.

with overlapping movements (e.g., 7 and 9), finer segmentation could further improve both per-class and overall accuracy.

Data communication overhead. In real deployment, communication bottlenecks such as channel noise, latency, and limited bandwidth can substantially degrade inference performance. To address this, each radar node transmits compact features extracted by the lightweight 2D CNN block, rather than raw measurements. We benchmark this encoder-based compression scheme against conventional downsampling. An AWGN channel is considered, where white Gaussian noise is added to the transmitted data, which corresponds to either raw radar samples or compressed features depending on the experimental setup. Feature map sizes of (5 $\times$ 2), (5 $\times$ 4), (5 $\times$ 8), (10 $\times$ 4), and (10 $\times$ 8) with a fixed channel depth of 6 yield compression factors of 480 $\times$, 240 $\times$, 120 $\times$, 120 $\times$, and 60 $\times$, respectively. For comparison, fast-time downsampling ratios of 2 $\times$, 5 $\times$, 10 $\times$, and 20 $\times$ are evaluated. As shown in Fig. 4, the encoder-based transmission achieves substantially higher compression and significantly better robustness to channel noise due to its extraction of noise-resilient features. Among them, the (5 $\times$ 4) feature map performs best. Notably, even without any compression (i.e., transmitting raw measurements), the classification performance is lower than 480 $\times$ compression. This highlights the advantage of the proposed intermediate fusion, which facilitates noise-resilient classification while reducing communication overhead. Moreover, performing part of the inference locally without additional dimensionality reduction (e.g., PCA) reduces overall latency.

Attention weights interpretability. We perform an ablation study in which portions of the input are systematically removed to evaluate the model's dependency on these inputs. Individual nodes are ablated with zeros or random Gaussian

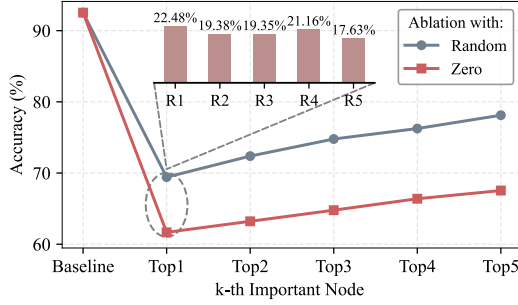


Fig. 5: Ablation study of node importance, with the bar plot showing the distribution of the most important nodes across the radars.

noise according to their computed importance from attention weights (Eq. 7). As shown in Fig. 5, ablating nodes with higher importance leads to larger drops in prediction accuracy. The bar plot illustrates the distribution of radars with the highest importance which indicates that the model adaptively prioritizes information from all sources. Such node-level attribution is crucial to prevent overfitting or bias during training and to ensure correct operation during the online phase (e.g., under radar failure or changes in radar orientation or placement). We further compare node importance with GradientSHAP, a Shapley-value-based XAI method [22]. Node attributions are computed by averaging absolute attribution maps. Spearman and Pearson correlations show strong agreement, with values of 0.875 and 0.9375, respectively.

V. CONCLUSION

We propose an end-to-end framework for continuous activity recognition using distributed radar that operates directly on raw data, eliminating costly DFT preprocessing. A lightweight 2D CNN extracts compact semantic features locally at each radar node, reducing communication overhead to the fusion processor. Multi-head attention-based intermediate fusion improves robustness to target position and orientation while providing interpretable node-level importance. The proposed method achieves higher accuracy with lower model complexity than baseline approaches. Ablation studies confirm the contribution of individual architectural components. The proposed feature compression strategy outperforms conventional downsampling, demonstrating its suitability for practical, low-latency distributed radar sensing.

ACKNOWLEDGMENT

The authors acknowledge the compute resources provided by the high-performance computer Lichtenberg II at TU Darmstadt, funded by the German Federal Ministry of Education and Research (BMBF) and the State of Hesse.

REFERENCES

[1] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1629–1645, 2019.

[2] J. Zhang, R. Xi, Y. He, Y. Sun, X. Guo, W. Wang, X. Na, Y. Liu, Z. Shi, and T. Gu, "A survey of mmwave-based human sensing: Technology, platforms and applications," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2052–2087, 2023.

[3] R. Zhang, X. Jing, S. Wu, C. Jiang, J. Mu, and F. R. Yu, "Device-free wireless sensing for human detection: The deep learning perspective," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2517–2539, 2020.

[4] L. Ren, Y. S. Koo, Y. Wang, and A. E. Fathy, "Noncontact heartbeat detection using UWB impulse doppler radar," in *2015 IEEE Topical Conference on Biomedical Wireless Technologies, Networks, and Sensing Systems (BioWireless)*. IEEE, 2015, pp. 1–3.

[5] S. Waqar, M. Muaaz, and M. Pätzold, "Direction-independent human activity recognition using a distributed MIMO radar system and deep learning," *IEEE Sensors Journal*, vol. 23, no. 20, pp. 24916–24929, 2023.

[6] R. G. Guendel, M. Unterhorst, E. Gambi, F. Fioranelli, and A. Yarovoy, "Continuous human activity recognition for arbitrary directions with distributed radars," in *2021 IEEE Radar Conference (RadarConf21)*. IEEE, 2021, pp. 1–6.

[7] S. Zhu, R. G. Guendel, A. Yarovoy, and F. Fioranelli, "Continuous human activity recognition with distributed radar sensor networks and CNN-RNN architectures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[8] R. G. Guendel, N. C. Kruse, F. Fioranelli, and A. Yarovoy, "Multipath exploitation for human activity recognition using a radar network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[9] P. Zheng, A. Zhang, J. Chen, J. Zhang, and F. Qi, "Direction-independent human behavior recognition using distributed radar sensor system and hybrid neural network with dual-view attention," *IEEE Sensors Journal*, 2024.

[10] N. C. Kruse, F. Fioranelli, and A. Yarovoy, "Radar point cloud processing methods for human activity classification with point transformer networks," *IEEE Transactions on Radar Systems*, vol. 2, pp. 1–12, 2023.

[11] Y. Liu, F. Wang, N. Wang, and Z.-X. Zhang, "Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53964–53982, 2023.

[12] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3. 0: Zero-effort cross-domain gesture recognition with Wi-Fi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8671–8688, 2021.

[13] R. G. Guendel, F. Fioranelli, and A. Yarovoy, "Distributed radar fusion and recurrent networks for classification of continuous human activities," *IET Radar, Sonar & Navigation*, vol. 16, no. 7, pp. 1144–1161, 2022.

[14] C. Eckrich, A. M. Zoubir, and V. Jamali, "Fronthaul-constrained distributed radar sensing," *arXiv preprint arXiv:2409.17753*, 2024.

[15] R. Thomä, C. Andrich, M. Döbereiner, R. Faramarzhangari, J. Ged-schold, M. F. C. Miranda, S. J. Myint, S. Schieler, C. Schneider, S. Semper *et al.*, "Distributed multisensor isac," *npj Wireless Technology*, vol. 2, no. 1, p. 22, 2026.

[16] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 943–952.

[17] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.

[18] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.

[19] L. Hu, X. Wang, Y. Liu, N. Liu, M. Huai, L. Sun, and D. Wang, "Towards stable and explainable attention mechanisms," *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[20] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[21] R. Guendel, M. Unterhorst, E. Gambi, F. Fioranelli, and A. Yarovoy, "Dataset of continuous human activities performed in arbitrary directions, collected with a distributed radar network of five nodes," 2021, 4TU.ResearchData. [Online]. Available: <https://doi.org/10.4121/16691500.v1>

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.