Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO25263

DSA-Net: A Dual-Path Spatial-Temporal Attention Network for WiFi-Based Human Activity Recognition

Shengxiong Xiao

Formerly with Department of Computer Science, Bishop's University, Sherbrooke, Canada
s.xiao.cs@gmail.com

Abstract. WiFi-based Human Activity Recognition (HAR) enables privacy-preserving, device-free motion detection using Channel State Information (CSI) from commodity devices. However, CSI's low resolution and noise complicate spatiotemporal feature extraction. We propose DSA-Net, a Dual-Path Spatial-Temporal Attention Network tailored to CSI-based HAR. It combines a slow-fast temporal pathway with cross-spatial attention to capture fine-grained and long-range dependencies, while a Transformer-based fusion module adaptively integrates spatiotemporal features. Evaluated on the Widar3.0 dataset, DSA-Net surpasses Vision Transformer (ViT) baselines by 18.91 percentage points, achieving superior accuracy with low computational overhead. Our results demonstrate DSA-Net's potential for scalable, real-time activity recognition in IoT and smart environments.

Keywords: WiFi sensing, Channel State Information, Human Activity Recognition, spatiotemporal feature fusion.

1. Introduction

Human Activity Recognition (HAR) is fundamental to the development of intelligent environments, enabling machines to interpret and respond to human behavior in real time. Applications range from healthcare monitoring and smart homes to gesture-based control and security systems. While traditional activity recognition methods—such as wearable sensors and vision-based systems—have made significant progress, they often suffer from drawbacks like user discomfort, ongoing maintenance requirements, and serious privacy concerns, especially in residential or sensitive settings [1-4].

To address these limitations, WiFi-based sensing has emerged as a compelling alternative. By leveraging Channel State Information (CSI) from commercial WiFi devices, it enables passive, device-free activity recognition without requiring cameras or wearable hardware. This approach offers a scalable, privacy-preserving alternative to wearable and camera-based systems [5-10].

Human movements alter WiFi signal propagation, introducing subtle variations in strength, phase, and path that can be captured by CSI for inferring activities. Unlike vision systems, WiFi sensing operates robustly in occluded or dark environments and does not rely on user cooperation [11]. Recent advances in device-free sensing and deep learning have further improved the accuracy and feasibility of CSI-based HAR in real-world conditions [12-17].

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO25263

However, despite these promising directions, current models face significant challenges. CSI's noisy, low-resolution nature hinders effective spatial-temporal pattern extraction. Conventional models capture local features but struggle with long-term dependencies. Moreover, they often conflate spatial and temporal information, hindering their ability to learn discriminative features.

Transformer-based attention mechanisms have recently been introduced to address temporal dependencies, but their effectiveness on CSI data remains limited. Due to the coarse spatial resolution and domain-specific noise in CSI feature maps, standard attention models struggle to capture meaningful interactions across time and space. As noted by Lu et al. [18-20], CSI contains richer multidimensional features than traditional representations, but fully leveraging this potential requires architectures that are explicitly tailored to the spatial-temporal structure of wireless signals.

As a result, advancing CSI-based HAR demands architectures that can not only extract but also intelligently fuse spatial, temporal, and channel-specific features—ideally without requiring handcrafted inputs or pre-processing heuristics. Recent work by Zhang et al. [21] emphasized the importance of attention mechanisms for capturing long-range dependencies but also noted the lack of architectures tailored to CSI's multidimensional and low-resolution nature. Furthermore, most current models do not specifically separate or jointly optimize spatial and temporal learning, resulting in limited generalization and weak performance in edge cases. These gaps motivate the need for a more holistic and adaptive architecture that treats CSI's spatial-temporal structure as a first-class design priority.

Our contributions are as follows:

• We propose DSA-Net, a dual-path architecture tailored to CSI data, decoupling spatial and temporal feature learning.

• We design Cross-Spatial Attention and Self-Attention Fusion modules to enhance feature discriminability and noise robustness.

• Extensive experiments on Widar3.0 demonstrate a significant accuracy improvement over baselines, with lightweight computational overhead suitable for IoT deployment.

2. Related works

2.1. CNN-based deep learning for WiFi HAR

Early research on deep learning for WiFi HAR primarily utilized Convolutional Neural Networks (CNNs) to extract spatial features from CSI matrices. CNNs apply local receptive fields and parameter sharing to model the spatial patterns embedded in CSI amplitude and phase measurements. Studies and [22,23] demonstrated that neural networks could effectively capture fine-grained motion cues in WiFi signals by treating CSI matrices analogously to images. In particular, Lu et al. [18] proposed CE-HAR, which introduced channel exchange operations within deep neural network architectures to better exploit spatial dependencies across subcarriers. How ever, standard CNNs are limited in their ability to capture long range temporal dependencies across sequential CSI frames. Although extensions like 3D convolutional layers attempt to jointly model spatial and short-term temporal features, they remain computationally expensive and often underperform when faced with noisy, low-resolution wireless data. Consequently, methods that separately or effectively mitigate temporal dynamics became necessary to better exploit the sequential nature of human activities captured in CSI streams.

DOI: 10.54254/2755-2721/2025.PO25263

2.2. Related studies using multipath architectures

Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks and Gated Re current Units (GRUs), were introduced to address the temporal modeling gap left by CNNs. These architectures excel at capturing sequential dependencies and have been applied to CSI based tasks. Several studies reviewed in [24] have successfully employed LSTM and GRU models to process signal sequences for HAR, highlighting the potential of recurrent architectures to capture temporal patterns despite challenges in spatial feature extraction. However, a major limitation of RNN-based models lies in their weak spatial feature extraction capabilities. CSI signals are multi-dimensional, requiring models that can capture both spatial and temporal aspects effectively [9].

To overcome these challenges, multipath architectures, such as SlowFast networks [25], were proposed in the video recognition domain. SlowFast networks employ a dual-path design: a slow path processes low frame rate inputs to capture semantic context, while a fast path processes high frame rate inputs to preserve motion detail. This architecture decouples semantic and dynamic temporal information, offering a powerful way to model activities across different timescales. Several improvements have been proposed to refine multipath designs. For example, reference [26] fused SlowFast paths with residual connections to enhance spatiotemporal feature learning, while reference [27,28] introduced additional modules to better capture multi scale temporal dynamics. Although those networks are effective for such tasks, their assumptions of spatial continuity and dense input make them suboptimal for WiFi signal matrices, which are characterized by coarse resolution, multipath fading, and significant domain-specific noise. Consequently, adapting multipath architectures to CSI-based activity recognition re quires customized designs tailored to the irregular wireless signals.

2.3. Enhance feature learning with attention mechanisms

In parallel with the development of multipath architectures, attention mechanisms have emerged as a powerful tool for feature enhancement in sequential data processing. Transformers [29] introduced self-attention to model long range dependencies without relying on recurrent structures, revolutionizing natural language processing and, later, vision tasks [30]. Recent works have explored the application of attention to HAR. For example, Mazzia et al. [31] proposed an Action Transformer that applies self-attention to 2D human pose sequences for short-term activity recognition, while Xu et al. [32] introduced Evo-ViT, which evolves slow and fast tokens dynamically to balance temporal resolution and computational cost in video recognition.

However, most of these attention-based methods are designed for dense, high-resolution, and structured input modalities, such as RGB frames or pose key points. Applying them directly to WiFi CSI signals is non-trivial. CSI data is inherently sparse, noisy, and irregular, posing significant challenges in capturing meaningful spatial-temporal dependencies. Yang et al. [9] highlight these challenges, emphasizing the low signal-to-noise ratio (SNR), domain-specific variability, and sparsity of CSI measurements, which fundamentally differ from conventional visual data.

To address these challenges, we propose a tailored approach that adapts attention mechanisms. Specifically, we design two critical components:

• Cross-Spatial Attention Module — selectively emphasizes channel-specific features during spatial path extraction, enhancing the model's ability to discern informative patterns within noisy and low-resolution CSI maps.

DOI: 10.54254/2755-2721/2025.PO25263

• Self-Attention Fusion Module — dynamically integrates spatial-temporal representations after the slow and fast temporal paths, ensuring effective feature fusion across multiple scales and reducing the impact of irrelevant noise.

By simultaneously modeling both spatial channel dependencies and temporal dynamics, our attention modules enable the network to focus on salient signal patterns while suppressing noise and redundancy. This design bridges the gap between generic attention mechanisms and the specialized requirements in current experimental settings, leading to improved recognition performance in noisy and sparse environments.

3. Method

Unlike conventional image or video data, CSI signals pose significant challenges for effective feature extraction and temporal modeling [9]. Building upon the SlowFast dual-path paradigm [33], our architecture addresses these challenges by isolating spatial and temporal learning processes. We intro duce two key components: a Cross-Spatial Attention Module, designed to isolate and emphasize channel-specific spatial features, and a Temporal Slow-Fast Pathway, which captures both fine-grained and long-term temporal dependencies.

To effectively fuse these heterogeneous features, we further employ a Self-Attention Fusion Module that adaptively prioritizes salient patterns while suppressing irrelevant noise. This design enables robust activity recognition even in low-SNR, high-variability CSI environments, aligning with our objective of achieving accurate, generalizable model performance without handcrafted features or heavy pre-processing.

The overall architecture of DSA-Net is illustrated in Figure 1, which highlights its dual-path structure and attention-driven fusion strategy.

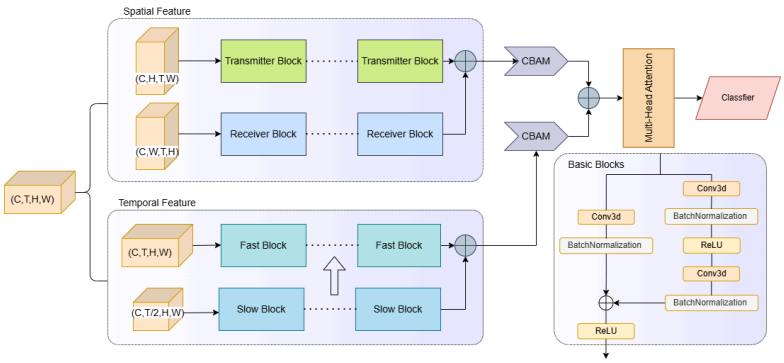

Figure 1. The proposed DSA-Net framework for WiFi-based human activity recognition

The model comprises multiple convolutional layers followed by pooling layers for feature extraction, with additional self-attention layers for feature fusion. This unique architecture integrates temporal and spatial information from CSI, enhancing human activity recognition performance.

3.1. Widar dataset utilizing CSI

The Widar3.0 dataset [34] provides a comprehensive plat form for WiFi-based HAR, specifically designed to enable cross-domain recognition of complex human actions. Unlike wearable or vision-based systems, Widar3.0 leverages the passive sensing capability of WiFi signals, capturing human movements by analyzing how these actions affect wireless signal propagation between transmitters and receivers [35].

DOI: 10.54254/2755-2721/2025.PO25263

At the core of this approach is CSI, which offers fine grained measurements of wireless channels at the physical layer, including amplitude and phase information for each subcarrier [24]. CSI serves as a rich, non-intrusive source of behavioral data, immune to privacy concerns and environmental lighting conditions.

For each WiFi packet received, raw CSI is extracted into a three-dimensional matrix of size $N_t \times N_r \times N_s$, where N_t and N_r denote the number of transmit and receive antennas, and N_s represents the number of subcarriers. However, to facilitate robust activity recognition, Widar3.0 further processes these CSI values into Body Velocity Profile (BVP) representations. These BVP sequences encapsulate subtle signal variations caused by human movements, providing a more structured input for learning-based models. This BVP representation effectively transforms raw CSI measurements into a structured spatiotemporal sequence, analogous to video frames, which is critical for enabling SlowFast temporal modeling in DSA-Net.

The resulting dataset comprises 3D BVP sequences with dimensions $H \times W$ over T consecutive frames, forming a spatiotemporal tensor input represented as $X_{in} \in R^{C \times T \times H \times W}$. This structured format enables models to simultaneously exploit spatial and temporal signal characteristics

3.2. Main architecture of DSA-Net

The design of DSA-Net is motivated by the need to accurately recognize human activities from WiFi CSI data, which is characterized by low spatial resolution, temporal sparsity, and signal noise. Traditional CNN and RNN-based architectures often conflate spatial and temporal information, limiting their ability to extract discriminative features from such data [21]. Furthermore, attention-based models, while effective for dense visual inputs, struggle to adapt to the irregular and noisy structure of CSI signals [9].

To overcome these challenges, DSA-Net decouples spatial and temporal feature extraction through a dual-path architecture, followed by a self-attention fusion mechanism to adaptively integrate multi-scale information. The network is composed of three key modules:

• Cross-Spatial Attention Module: Enhances the model's ability to isolate channel-specific spatial patterns, addressing the lack of strong spatial locality in CSI data.

• Temporal Slow-Fast Pathways: Capture both fine-grained and long-range temporal dependencies by processing input sequences at multiple temporal resolutions, inspired by SlowFast networks [33].

• Self-Attention Fusion Module: Dynamically integrates spatial and temporal representations, prioritizing salient features while suppressing noise, ensuring robust activity recognition in low-SNR environments.

This modular design allows DSA-Net to effectively model the complex spatiotemporal relationships inherent in CSI based HAR tasks, achieving high recognition accuracy without relying on handcrafted features or extensive preprocessing.

3.2.1. Cross learning spatial path

Unlike visual data with well-defined spatial locality (e.g., edges, contours), WiFi CSI data presents a unique challenge: its spatial features are distributed across antenna channels and subcarrier dimensions, lacking consistent spatial patterns. Consequently, traditional convolutional approaches struggle to isolate meaningful spatial dependencies in such low-resolution, noisy inputs.

DOI: 10.54254/2755-2721/2025.PO25263

To address this, we design a Cross-Spatial Attention Mod ule that selectively emphasizes channel-specific spatial pat terns. The key idea is to transform the input tensor $X_{in} \in R^{C\times T\times H\times W}$ into alternate spatial orientations, enabling the network to focus on directional variations in signal propaga tion.

Specifically, we reshape the input features along two axes:

$$x_h = X_{in} \in R^{C\times T\times H\times W} \tag{1}$$

$$x_v = X_{in} \in R^{C\times T\times H\times W} \tag{2}$$

This transformation allows independent feature extraction along the $T \times W$ and $H \times T$ spaces, capturing distinct spatial dependencies introduced by human activities interacting with WiFi signals. By processing these two feature spaces through separate convolutional paths, the module effectively models subtle channel-specific variations, which are critical for distinguishing human movements in CSI data.

The Cross-Spatial Attention mechanism computes attention maps over these transformed features, dynamically prioritizing salient regions of interest while suppressing irrelevant noise. This approach draws inspiration from multi-path CNN architectures, yet is specifically tailored to the irregular and sparse nature of CSI measurements.

Through this design, our network achieves a more nuanced understanding of spatial signal patterns, enhancing its ability to recognize fine-grained activity cues embedded within noisy wireless environments.

3.2.2. Temporal slow-fast path

Human activities captured through WiFi CSI signals exhibit diverse temporal characteristics: some actions involve brief, transient motions (e.g., gestures), while others represent sustained postures (e.g., standing still). Effectively recognizing these activities requires modeling both fine-grained temporal changes and long-term dependencies.

To address this, we adopt a Temporal Slow-Fast Pathway inspired by the SlowFast network paradigm [33]. This dual path architecture processes input sequences at two distinct temporal resolutions:

• Slow Path: Focuses on capturing long-range temporal dependencies by sampling input frames with a larger time stride (τ) . In our implementation, the slow path processes $T = 11$ frames, enabling the network to model broader activity patterns over extended durations.

• Fast Path: Emphasizes fine-grained temporal details by sampling input frames at a higher rate, using a reduced time stride (τ/α) . We set $\alpha = 2$ considering the short-lived nature of most activities in the Widar dataset, resulting in the fast path processing $\alpha T = 22$ frames.

Both paths operate on input tensors of shape $X_{in} \in R^{C\times T\times H\times W}$, but differ in their sampling rates and channel capacities. The fast path maintains high temporal fidelity with reduced channel width, while the slow path captures global temporal context with richer feature representations.

By decoupling temporal resolutions, the Temporal Slow Fast Pathway enables DSA-Net to simultaneously learn rapid signal fluctuations and sustained activity patterns, addressing the temporal sparsity and variability inherent in CSI-based HAR tasks.

DOI: 10.54254/2755-2721/2025.PO25263

3.2.3. Feature fusion in self-attention encoder

After extracting spatial and temporal features through the Cross-Spatial Attention Module and Temporal Slow-Fast Pathways, it is essential to refine and effectively fuse these heterogeneous representations. Given the inherent defects of CSI data, naive feature concatenation risks diluting informative patterns amidst irrelevant signals.

To mitigate this, we first apply a Convolutional Block Attention Module (CBAM) [36] independently to the output of each path. CBAM sequentially infers channel attention and spatial attention maps, allowing the network to adaptively emphasize meaningful features and suppress noise at both the channel and spatial levels. This lightweight attention refinement ensures that each path outputs a more discriminative representation before fusion.

Let x_s and x_f denote the refined outputs of the slow and fast temporal paths after CBAM, respectively. Similarly, spatial path features processed through CBAM are denoted as xspatial. These features are then concatenated to form a unified representation:

$$X_a = x_{spatial} \oplus x_s \oplus x_f \in R^{C\times T\times H\times W} \tag{3}$$

where $\oplus$ denotes concatenation along the channel dimension.

To adaptively model inter-feature dependencies and prioritize salient information, we employ a Self-Attention Fusion Module on X_a. Unlike fixed-weight fusion methods, self attention allows the network to dynamically adjust the con tribution of each feature based on the context, enhancing the model's ability to focus on subtle signal variations critical for activity recognition. Following the standard attention mechanism, we compute the query (Q), key (K), and value (V) projections:

$$Q = \mathrm{X_a W_Q},\ K = \mathrm{X_a W_K},\ V = \mathrm{X_a W_V} \tag{4}$$

where W_Q, W_K, $and\, W_V$ are learnable projection matrices.

The attention weights are computed as:

$$A = softmax\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d_k}}}\right) \tag{5}$$

with dk denoting the key dimension. The fused representation is then obtained as:

$$Attention(\mathrm{X_a}) = AV \tag{6}$$

This fusion process ensures that the final representation selectively integrates spatial and temporal features, robustly handling the noisy and low-SNR nature of CSI data. The result is a discriminative and noise-resilient feature embedding, optimized for accurate human activity recognition.

DOI: 10.54254/2755-2721/2025.PO25263

3.3. Learning objective

Given the multi-class nature of activity recognition tasks, we adopt the categorical cross-entropy loss to supervise model training. This loss function measures the divergence between the predicted class probabilities and the ground-truth activity labels, providing an effective objective for optimizing the network's classification performance.

Formally, the cross-entropy loss is defined as:

$$\mathscr{L}_{\mathrm{CE}} = \sum_{i=1}^{N} \sum_{c=1}^{c} y_{i,c} \bullet \log(p_{i,c}) \tag{7}$$

where N denotes the number of training samples, C is the total number of activity classes, $y_{i,c}$ represents the ground truth indicator (1 if sample i belongs to class c, otherwise 0), and $p_{i,c}$ is the predicted probability of sample i being classified as class c after the final SoftMax layer.

The choice of cross-entropy loss aligns with our objective of achieving robust activity classification in the presence of noisy, sparse CSI data. Its probabilistic formulation allows the model to calibrate its confidence, effectively handling ambiguous signal patterns and domain variability. Additionally, cross entropy seamlessly integrates with gradient-based optimizers, facilitating efficient model convergence during training.

By minimizing this loss, DSA-Net learns to differentiate between subtle activity-induced variations in wireless signals, enabling accurate recognition across diverse activity classes and deployment scenarios.

4. Experiments

4.1. WiFi dataset

For our experimental evaluation, we utilize the Widar3.0 dataset [34], a large-scale benchmark specifically designed for device-free, WiFi-based HAR. Unlike vision or wearable sensor datasets, Widar captures human movements by analyzing CSI — complex-valued measurements reflecting the propagation of WiFi signals through an environment.

Human activities cause distinct variations in signal re f lection, scattering, and Doppler shifts, which are embedded within the CSI measurements. To facilitate effective learning from these subtle signal perturbations, Widar3.0 introduces the concept of BVPs. These BVPs represent the velocity distributions of different human body parts over time, derived from processing the raw CSI amplitudes and phases.

The dataset poses several unique challenges aligned with real-world deployment scenarios:

• Sparse and Noisy Inputs: Due to limited antenna arrays and environmental multipath effects, CSI signals are intrinsically low-resolution and noisy.

• Domain Variability: Data spans multiple subjects, positions, and domains, demanding models with strong generalization capabilities.

• Fine-Grained Activity Distinction: Differentiating between subtle activities (e.g., hand gestures vs. postural changes) requires precise spatiotemporal modeling.

These characteristics make Widar3.0 an ideal testbed to evaluate the robustness and effectiveness of our proposed DSA-Net. By leveraging the BVP sequences structured as spatiotemporal tensors

DOI: 10.54254/2755-2721/2025.PO25263

$X_{in} \in R^{C\times T\times H\times W}$, DSA-Net is specifically designed to address the dataset's challenges through its dual-path temporal processing, cross-spatial attention, and adaptive fusion mechanisms.

4.2. Results and analysis

We benchmark the performance of our proposed DSA-Net against several representative architectures, including Multilayer Perceptron (MLP), CNN, RNN, GRU, LSTM, and ViT. These baselines reflect common paradigms for spatial and temporal feature extraction in sequential data. Table 1 summarizes the comparative results on the Widar3.0 dataset in terms of classification accuracy, computational complexity (FLOPs), and parameter count.

Table 1. Widar dataset – model performance comparison

Method	Accuracy $(\%)$	FLOPs (M)	Params (M)
MLP	67.24	9.15	9.150
CNN-5	70.19	3.38	0.299
RNN	46.77	0.66	0.031
GRU	62.50	1.98	0.091
LSTM	63.35	2.64	0.121
ViT	64.85	9.28	0.106
DSA-Net	83.76	6.08	1.078

Performance Insights: DSA-Net achieves an accuracy of 83.76%, outperforming all baselines by a significant margin.

Traditional CNNs, while effective in extracting local spatial features, struggle with modeling long-term temporal dependencies and are sensitive to the coarse spatial granularity of CSI data. Recurrent models (RNN, GRU, LSTM) are sequential and face difficulties handling sparse and noisy inputs, leading to suboptimal performance.

ViT introduces self-attention to model global dependencies but assumes dense, high-resolution inputs, making it less effective on irregular CSI matrices. Additionally, its computational overhead is not justified by the modest accuracy gain observed.

In contrast, DSA-Net's architecture explicitly addresses these limitations:

• The Slow-Fast Temporal Pathways capture both fine grained signal fluctuations and long-range dependencies, crucial for distinguishing diverse activity patterns in CSI.

• The Cross-Spatial Attention Module enhances channel specific feature extraction, compensating for CSI's lack of spatial coherence.

• The Self-Attention Fusion adaptively prioritizes salient spatial-temporal representations, effectively suppressing domain-specific noise.

Efficiency Considerations: Beyond accuracy, DSA-Net maintains a lightweight profile with only 1.08M parameters and 6.08M FLOPs, striking a balance between performance and computational efficiency. This makes it suitable for deployment on resource-constrained edge devices in IoT and smart home environments.

These results validate the effectiveness of DSA-Net's specialized design for CSI-based HAR, confirming its superiority over both conventional sequential models and generic attention-based architectures.

DOI: 10.54254/2755-2721/2025.PO25263

4.3. Ablation study

To assess the individual contributions of DSA-Net's core components, we conduct an ablation study focusing on three key modules: the Cross-Spatial Path, responsible for channel specific spatial feature extraction; the Temporal Slow-Fast Pathways, designed to capture multi-scale temporal dependencies; and the Self-Attention Fusion Module, which dynamically integrates spatiotemporal representations. Table 2 summarizes the classification accuracy achieved on the Widar3.0 dataset with various component combinations.

Table 2. Ablation study: component-wise performance on Widar3.0

Spatial Path	Temporal Path	Self-Attention	Accuracy (%)
√	×	×	73.19
√	√	×	78.61
√	×	√	75.43
√	√	√	83.76

When utilizing only the Cross-Spatial Path, the model achieves an accuracy of 73.19%, highlighting its effectiveness in capturing spatial patterns from CSI data, which inherently lacks the structured spatial hierarchies found in visual modalities. Introducing the Temporal Slow-Fast Pathways elevates accuracy to 78.61%, underscoring the importance of modeling both fine-grained signal fluctuations and long-range activity dependencies. This decoupled temporal design proves particularly effective for sparse, low-resolution CSI sequences where transient and sustained motions coexist.

Incorporating the Self-Attention Fusion Module, without temporal pathways, further improves performance to 75.43%, demonstrating its utility in filtering irrelevant noise and adaptively prioritizing informative spatiotemporal patterns. The complete DSA-Net configuration, combining spatial, temporal, and attention-driven fusion, achieves the highest accuracy of 83.76%, confirming the synergistic contribution of these modules in addressing CSI-specific challenges.

These results empirically validate our architectural choices: dedicated Cross-Spatial learning captures channel dependencies effectively; the Slow-Fast dual-pathway addresses temporal sparsity and multi-scale dynamics; and the Self-Attention Fusion enhances robustness against domain variability. Collectively, these components enable DSA-Net to achieve superior recognition performance aligned with the intrinsic properties of CSI-based HAR tasks.

4.4. Conclusion of experimental results

The experimental evaluation on the Widar3.0 dataset high lights the effectiveness of DSA-Net in addressing the key challenges of WiFi-based HAR. With an accuracy of 83.76%, DSA-Net consistently outperforms baseline models such as CNNs, RNNs, and Vision Transformers, confirming the value of its tailored design for CSI signals. Unlike conventional models that falter under sparse, noisy, and low-resolution CSI conditions, DSA-Net integrates temporal multi-scale modeling, channel-aware spatial attention, and adaptive feature fu sion to extract robust and discriminative features. This results in improved recognition accuracy and stronger resilience to domain-specific variability.

Ablation studies further confirm that each component—Cross-Spatial Attention, Slow-Fast Temporal Pathways, and Self-Attention Fusion—contributes significantly to overall performance,

DOI: 10.54254/2755-2721/2025.PO25263

with the full configuration achieving the highest accuracy. Crucially, DSA-Net achieves these gains with only 1.08M parameters and 6.08M FLOPs, ensuring compatibility with resource-constrained edge devices. These findings substantiate the architectural advantages of DSA-Net and establish its practicality for real-world HAR applications in IoT environments.

5. Conclusion

In this work, we introduced DSA-Net, a Dual-Path Spatial Temporal Attention Network specifically designed for CSI based HAR. Unlike conventional CNN, RNN, or Transformer models, DSA-Net is architected to address the core challenges of CSI signals—namely, their sparsity, low spatial resolution, and susceptibility to environmental noise. By integrating Temporal Slow-Fast Pathways, Cross-Spatial Attention, and a Self-Attention Fusion Module, DSA-Net effectively models multi-scale temporal dependencies, isolates informative spatial features, and adaptively prioritizes relevant patterns. This unified design eliminates the need for handcrafted inputs or extensive preprocessing. Extensive experiments on the Widar3.0 dataset demonstrate that DSA-Net achieves high recognition accuracy while maintaining a lightweight computational footprint—making it ideal for edge deployment in privacy-sensitive and resource-constrained environments. Our results underscore the importance of architecture-level adaptations tailored to the unique structure of wireless CSI data. Looking forward, future work will explore improving DSA-Net's domain generalization capabilities and optimizing it for real-time, on-device inference in dynamic environments.

References

[1] Beddiar, D. R., Nini, B., Sabokrou, M., & Hadid, A. (2020). Vision-based human activity recognition: A survey. Multimedia Tools and Applications, 79, 30509–30555. https: //doi.org/10.1007/s11042-019-08336-9

[2] Yang, J., Huang, H., Zhou, Y., Chen, X., Xu, Y., Yuan, S., Zou, H., Lu, C. X., & Xie, L. (2023). Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. NeurIPS Datasets and Benchmarks Track. https: //doi.org/10.48550/arXiv.2305.10345

[3] Chen, X., & Yang, J. (2024). X-fi: A modality-invariant foundation model for multimodal human sensing. International Conference on Learning Representations (ICLR-25). https: //doi.org/10.48550/arXiv.2410.10167

[4] Zhou, C., & Yang, J. (2025). Hololllm: Multisensory foundation model for language-grounded human sensing and reasoning. arXiv preprint arXiv: 2505.17645. https: //arxiv.org/abs/2505.17645

[5] Yang, J., Zou, H., Jiang, H., & Xie, L. (2018). Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes. IEEE Internet of Things Journal, 5(5), 3991–4002. https: //doi.org/10.1109/JIOT.2018.2835520

[6] Yang, J., Zou, H., Jiang, H., & Xie, L. (2018). Carefi: Sedentary behavior monitoring system via commodity WiFi infrastructures. IEEE Transactions on Vehicular Technology, 67(8), 7620–7629. https: //doi.org/10.1109/TVT.2018.2830452

[7] Yang, J., Zou, H., Zhou, Y., & Xie, L. (2019). Learning gestures from WiFi: A Siamese recurrent convolutional architecture. IEEE Internet of Things Journal, 6(6), 10763–10772. https: //doi.org/10.1109/JIOT.2019.2938065

[8] Yang, J., Chen, X., Zou, H., Wang, D., & Xie, L. (2022). Autofi: Towards automatic WiFi human sensing via geometric self-supervised learning. IEEE Internet of Things Journal. https: //doi.org/10.1109/JIOT.2022.3174760

[9] Yang, J., Chen, X., Wang, D., Zou, H., Lu, C. X., Sun, S., & Xie, L. (2023). SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing. Patterns, 4(3). https: //doi.org/10.1016/j.patter.2023.100677

[10] Yang, J., Zou, H., & Xie, L. (2022). SecureSense: Defending adversarial attack for secure device-free human activity recognition. IEEE Transactions on Mobile Computing. https: //doi.org/10.1109/TMC.2022.3160981

[11] Kumar, P., Chauhan, S., & Awasthi, L. K. (2024). Human activity recognition (HAR) using deep learning: Review, methodologies, progress and future research directions. Archives of Computational Methods in Engineering, 31(1), 179–219. https: //doi.org/10.1007/s11831-023-09964-w

[12] Al-Qaness, M. A. A., Li, F., Ma, X., Zhang, Y., & Liu, G. (2016). Device-free indoor activity recognition system. Applied Sciences, 6(11), 329. https: //doi.org/10.3390/app6110329

DOI: 10.54254/2755-2721/2025.PO25263

[13] Yang, J., Chen, X., Zou, H., Wang, D., Xu, Q., & Xie, L. (2022). EfficientFi: Toward large-scale lightweight WiFi sensing via CSI compression. IEEE Internet of Things Journal, 9(15), 13086–13095. https: //doi.org/10.1109/JIOT.2022.3157755
[14] Zou, H., Yang, J., Zhou, Y., Xie, L., & Spanos, C. J. (2018). Robust WiFi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation. In 2018 27th International Conference on Computer Communication and Networks (ICCCN) (pp. 1–8). IEEE. https: //doi.org/10.1109/ICCCN.2018.8487320
[15] Zou, H., Zhou, Y., Yang, J., Gu, W., Xie, L., & Spanos, C. (2018). WiFi-based human identification via convex tensor shapelet learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https: //doi.org/10.1609/aaai.v32i1.11678
[16] Zou, H., Zhou, Y., Yang, J., Jiang, H., Xie, L., & Spanos, C. J. (2018). DeepSense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network. In 2018 IEEE International Conference on Communications (ICC) (pp. 1–6). IEEE. https: //doi.org/10.1109/ICC.2018.8422243
[17] Zou, H., Yang, J., Zhou, Y., & Spanos, C. J. (2018). Joint adversarial domain adaptation for resilient WiFi-enabled device-free gesture recognition. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 202–207). IEEE. https: //doi.org/10.1109/ICMLA.2018.00038
[18] Lu, X., Li, Y., Cui, W., & Wang, H. (2022). CeHAR: CSI-based channel-exchanging human activity recognition. IEEE Internet of Things Journal, 10(7), 5953–5961. https: //doi.org/10.1109/JIOT.2022.3153332
[19] Zhou, Y., Huang, H., Yuan, S., Zou, H., Xie, L., & Yang, J. (2023). MetaFi++: WiFi-enabled transformer-based human pose estimation for metaverse avatar simulation. IEEE Internet of Things Journal, 10(16), 14128–14136. https: //doi.org/10.1109/JIOT.2023.3269519
[20] Yang, J., Zhou, Y., Huang, H., Zou, H., & Xie, L. (2022). MetaFi: Device-free pose estimation via commodity WiFi for metaverse avatar simulation. In IEEE World Forum on Internet of Things 2022 (pp. 1–6). IEEE. https: //doi.org/10.1109/WF-IoT49695.2022.00022
[21] Zhang, Y., Chen, Y., Wang, Y., Liu, Q., & Cheng, A. (2021). CSI-based human activity recognition with graph few-shot learning. IEEE Internet of Things Journal, 9(6), 4139–4151. https: //doi.org/10.1109/JIOT.2021.3052457
[22] Moshiri, P. F., Shahbazian, R., Nabati, M., & Ghorashi, S. A. (2021). A CSI-based human activity recognition using deep learning. Sensors, 21(21), 7225. https: //doi.org/10.3390/s21217225
[23] Zhang, J., Wu, F., Wei, B., Zhang, Q., Huang, H., Shah, S. W., & Cheng, J. (2020). Data augmentation and dense-LSTM for human activity recognition using WiFi signal. IEEE Internet of Things Journal, 8(6), 4628–4641. https: //doi.org/10.1109/JIOT.2020.3002864
[24] Gu, F., Chung, M.-H., Chignell, M., Valaee, S., Zhou, B., & Liu, X. (2021). A survey on deep learning for human activity recognition. ACM Computing Surveys (CSUR), 54(8), 1–34. https: //doi.org/10.1145/3453160
[25] Xiao, F., Lee, Y. J., Grauman, K., Malik, J., & Feichtenhofer, C. (2020). Audiovisual SlowFast networks for video recognition. arXiv preprint arXiv: 2001.08740. https: //arxiv.org/abs/2001.08740
[26] Feichtenhofer, C., Pinz, A., & Wildes, R. P. (2017). Spatiotemporal multiplier networks for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4768–4777). https: //doi.org/10.1109/CVPR.2017.506
[27] Zeng, W., Huang, J., Zhang, W., Nan, H., & Fu, Z. (2022). SlowFast action recognition algorithm based on faster and more accurate detectors. Electronics, 11(22), 3770. https: //doi.org/10.3390/electronics11223770
[28] Liu, Y., Yuan, J., & Tu, Z. (2022). Motion-driven visual tempo learning for video-based action recognition. IEEE Transactions on Image Processing, 31, 4104–4116. https: //doi.org/10.1109/TIP.2022.3172605
[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https: //proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
[30] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. et al. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929. https: //arxiv.org/abs/2010.11929
[31] Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., & Chiaberge, M. (2022). Action Transformer: A self-attention model for short-time pose-based human action recognition. Pattern Recognition, 124, 108487. https: //doi.org/10.1016/j.patcog.2021.108487
[32] Xu, Y., Zhang, Z., Zhang, M., Sheng, K., Li, K., Dong, W., Zhang, L., Xu, C., & Sun, X. (2022). Evo-ViT: Slow-fast token evolution for dynamic vision transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), 2964–2972. https: //doi.org/10.1609/aaai.v36i3.20115
[33] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6202–6211). https: //doi.org/10.1109/ICCV.2019.00630

DOI: 10.54254/2755-2721/2025.PO25263

[34] Zhang, Y., Zheng, Y., Qian, K., Zhang, G., Liu, Y., Wu, C., & Yang, Z. (2021). Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11), 8671–8688. https: //doi.org/10.1109/TPAMI.2021.3088747

[35] Wang, W., Liu, A. X., Shahzad, M., Ling, K., & Lu, S. (2017). Device-free human activity recognition using commercial WiFi devices. IEEE Journal on Selected Areas in Communications, 35(5), 1118–1131. https: //doi.org/10.1109/JSAC.2017.2680903

[36] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3–19). https: //doi.org/10.1007/978-3-030-01234-2_1