Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-Environment WiFi CSI Gesture Recognition Using a Domain-Generalized Multi-Scale Attention Residual Network with Robust Feature Learning

Jarapala Ramesh babu

babu.ramesh444@gmail.com

Osmania University https://orcid.org/0000-0003-3405-056X

Research Article

Keywords: WiFi Channel State Information (CSI), Human Gesture Recognition, Domain Generalization, Cross-Environment Recognition, Multi-Scale Feature Learning, Channel-Spatial Attention, Residual Network, Robust Feature Learning

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11571996/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Cross-Environment WiFi CSI Gesture Recognition Using a Domain-Generalized Multi-Scale Attention Residual Network with Robust Feature Learning

Jarapala Ramesh Babu
Dept. of Electrical and Electronic engineering
Methodist College of Engineering and Technology
Mail.id: babu.ramesh444@gmail.com

Abstract
WiFi Channel State Information (CSI) records changes in wireless signals induced by hand and body motions, providing a practical and contact-free method of gesture recognition. Changes in the environment, multipath propagation, and signal circumstances might affect CSI patterns, making it tough to achieve consistent detection across varied indoor situations. In this paper, we provide a framework for WiFi CSI gesture recognition that works in any environment by using a robust feature-learning Domain-Generalized Multi-Scale Attention Residual Network. Meaningful patterns linked with various movements are identified by the framework via the use of CSI amplitude fluctuations throughout time and subcarriers. A Stable Field Various resolutions of the gesture are captured using multi-scale convolution, and the network is trained to concentrate on the most important features with the aid of channel and spatial attention. Stable model training and successful feature learning are both supported by residual connections. Evaluating performance in both learned and unknown situations, the trials take five gesture classes into account. The validation and training results for a robust domain demonstrate consistent recognition performance and continuous convergence. A Stable Field The framework's recognition accuracy in the unseen test environment is 86%, while in the two training domains they are 94% and 91%, respectively. The t-SNE image further highlights the substantial differences across the gesture classes, suggesting that the learnt characteristics maintain their uniqueness even when exposed to different environments. A Stable Field According to the ablation research, each component contributes to the overall improvement in accuracy; for example, using multi-scale convolution increased accuracy from 72% to 79%, channel/spatial attention increased accuracy to 82%, and the full framework improved accuracy to 86%. A Stable Field The findings show that the suggested method can reliably and robustly recognize gestures based on WiFi CSI even when the interior environment is constantly changing.

Keywords— WiFi Channel State Information (CSI), Human Gesture Recognition, Domain Generalization, Cross-Environment Recognition, Multi-Scale Feature Learning, Channel-Spatial Attention, Residual Network, Robust Feature Learning.

1. Introduction

1.1 Background and Motivation

Human activity and gesture sensing is becoming increasingly important in smart homes, assisted-living environments, healthcare applications, and contactless human–computer interaction. Traditional systems commonly depend on cameras, wearable devices, or dedicated sensors to observe human movements. Although these technologies can provide accurate information, their everyday use can be affected by privacy concerns, visual obstructions, lighting conditions, additional installation requirements, and the inconvenience of asking users to continuously wear sensing devices. These concerns have encouraged researchers to explore device-free sensing methods that can recognize human activities without requiring direct physical contact with the user [1]–[4]. WiFi-based sensing is particularly attractive because existing wireless infrastructure can be used for sensing as well as communication, reducing the need for additional hardware.

Among the different measurements available from WiFi signals, Channel State Information (CSI) provides detailed information about how wireless signals propagate between a transmitter and receiver. This makes CSI useful for identifying small changes in the wireless channel caused by human motion. Recent studies have demonstrated its potential for recognizing gestures and human activities using convolutional, residual, attention-based, and other feature-learning approaches. In addition, WiFi sensing does not require visual recordings of users, which makes it suitable for applications where privacy is important. However, reliable recognition becomes more difficult when the sensing environment changes. Therefore, a practical WiFi gesture-recognition system must learn meaningful characteristics of human gestures while reducing its dependence on the particular surroundings in which the training data were collected. A Robust Domain

1.2 WiFi CSI for Human Gesture Recognition

When a person performs a gesture, the movement of the hands, arms, or body changes the way WiFi signals travel through the surrounding space. The transmitted signals may be reflected, scattered, or diffracted by the moving body, resulting in continuous changes in the multipath propagation between the transmitter and receiver. These changes appear in the CSI measurements as variations in signal amplitude and phase across time and OFDM subcarriers. Consequently, the patterns in the wireless channel are affected by the motions. Without the need for

a camera or wearable sensor, a recognition system may differentiate between different gestures by viewing these patterns [1, 3, 5, 6].

This work gives a two-dimensional picture of the fluctuations due to human gestures by representing CSI amplitude over time and subcarriers. The experimental data include a CSI amplitude heatmap that shows the detected signal's evolution across various time steps and subcarrier indices. A Stable Field It is crucial to extract characteristics at several scales since these variations include both local and wider trends. While attention techniques may assist highlight the most important information within the CSI characteristics, residual learning effectively offers a basis for learning deeper representations. In order to get more accurate depictions of CSI fluctuations connected to gestures, the suggested system integrates multi-scale feature extraction with residual learning and channel-spatial attention.

1.3 Cross-Environment Recognition Problem

It is still challenging to get consistent performance across various interior conditions, despite WiFi CSI's significant promise for gesture detection. In addition to human motion, environmental factors can impact CSI readings. Multiple factors may alter the multipath properties of a wireless channel, including the room's arrangement, the location of the transmitter and receiver, the propagation routes, and any surrounding objects. Hence, when the sensing environment changes, CSI patterns recorded for the same move could seem different. Training a model in a single room could cause it to pick up environmental properties instead of those that accurately characterize the gesture [7]-[10].

When deploying a trained recognition system to a site that was not considered during training, this issue takes on a greater significance. Therefore, WiFi sensing across domains and environments has been the subject of recent research. The experimental results that were used for this study similarly mirror this difficulty. The two source domains attain recognition accuracies of 94% and 91%, respectively, while the previously unknown test environment achieves an accuracy of 86%. The significance of learning characteristics that maintain their utility despite changes in the surrounding environment is highlighted by these results, which point to a robust domain. In order to take WiFi gesture detection from the lab to the real world, inside, this capability must be improved.

1.4 Limitations of Existing Approaches

While there have been some promising outcomes with existing WiFi CSI gesture-recognition algorithms, their effectiveness is still affected by a number of constraints when it comes to different situations. Due to their reliance on data acquired within a narrow range of circumstances, many models are susceptible to developing environment-specific multipath features. These patterns that were learnt may not hold up when that same model is used in a different room or with a different propagation environment. A fixed receptive field cannot consistently catch both tiny differences among surrounding CSI data and wider temporal-spectral properties, which might make it problematic for conventional convolutional models to describe gesture information happening at various scales. Treating all feature channels and geographical locations equally also increases the likelihood that less valuable or environmentally sensitive information may impact the final prediction [7], [11], [12].

The importance of trustworthy experimental assessment cannot be overstated. Careful separation of training and testing samples is required to ensure that stated performance represents true generalization and not accidental dataset overlap. Data leakage prevention in convolutional neural network (CNN) WiFi CSI detection trials has been emphasized in prior research. Though each of the three approaches attention-based, Transformer-based, and Robust Domain Residual networks—offers useful answers for some parts of CSI representation, none of them can eliminate environmental variance on their own. A Solid Foundation This necessitates a more comprehensive framework that can understand gesture-related data at various sizes, prioritizing the most useful elements.

1.5 Research Gap

Improvements to WiFi-based human sensing have been the subject of several recent investigations. Activity recognition based on CSI has made use of residual networks, and feature selection has been enhanced with the introduction of channel-attention and related techniques. Transformer architectures, meta-learning, reconstructed CSI signals, Siamese networks, and cross-domain recognition have all been the subject of another research. These advancements have greatly improved WiFi sensing, but a major research challenge is still finding a way to reliably recognize objects in interior settings where none have been observed before. Specifically, even if the executed gesture stays the same, environmental alterations may vary the properties of CSI [8]-[14].

This research fills that need by providing a cohesive framework for evaluating models in different environments, learning residual representations, channel-spatial attention, and multi-scale feature extraction. The suggested design prioritizes useful feature channels and geographical areas, avoids depending on a single feature-extraction process, and captures CSI patterns at scales that are complimentary to one another. Based on the results of the experimental ablation investigation, this architecture is sound: compared to the baseline ResNet, recognition accuracy climbs to 79% with multi-scale convolution, 82% with channel/spatial attention, and 86% with the full

MSA-ResNet framework. Consolidating the Different Parts into One Recognition Framework These findings provide the groundwork for integrating the various parts into one comprehensive framework.

1.6 Research Objectives

The main objective of this work is to develop a reliable WiFi CSI-based gesture-recognition framework that can maintain useful recognition performance when it is applied in an indoor environment not encountered during model development. The proposed Domain-Generalized Multi-Scale Channel–Spatial Attention ResNet is designed to extract gesture-related characteristics while reducing sensitivity to changes in the surrounding environment. Multi-scale convolution is used to capture CSI variations at different resolutions, residual connections support deeper and more stable feature extraction, and channel–spatial attention gives greater importance to the most informative parts of the learned representation.

A further objective is to examine the proposed framework from several experimental perspectives instead of judging its performance only through overall classification accuracy. Training and validation loss and accuracy curves are used to examine convergence and learning behaviour. A Robust Domain The recognition of five gesture categories is further examined through an unseen-domain confusion matrix, while cross-environment accuracy measures the ability of the model to retain its performance when the domain changes. The t-SNE representation is used to examine the separation of learned gesture features, and an ablation study determines how individual components contribute to the final recognition performance. A Robust Domain

1.7 Major Contributions

The main contribution of this work is the development of a cross-environment WiFi CSI gesture-recognition framework based on a Domain-Generalized Multi-Scale Channel–Spatial Attention ResNet. The framework brings together three complementary feature-learning mechanisms. Multi-scale convolution captures gesture-related variations occurring at different temporal and subcarrier resolutions, residual learning provides an effective structure for extracting deeper representations, and channel–spatial attention emphasizes the parts of the CSI features that carry useful information for gesture discrimination. The approach is evaluated under an unseen-environment setting so that its ability to recognize gestures beyond the source domains can be examined. The t-SNE visualization also shows distinct feature clusters for the five gestures in the unseen domain. A Robust Domain

The second contribution is a comprehensive experimental evaluation covering several aspects of model behaviour. The proposed framework achieves 94% and 91% accuracy in the two source domains and 86% in the unseen test domain. A Robust Domain The ablation analysis further shows that the baseline ResNet achieves 72%, which increases to 79% with multi-scale convolution, 82% with channel/spatial attention, and 86% for the complete proposed framework. A Robust Domain Together with the convergence curves, five-class confusion matrix, cross-environment evaluation, and t-SNE feature visualization, these findings provide a clear assessment of the framework's performance and show the contribution of its main architectural components.

2. Literature Survey

2.1 WiFi CSI-Based Human Sensing

WiFi-based human sensing has grown considerably in recent years, moving from simple received-signal measurements toward more detailed sensing based on Channel State Information (CSI). Earlier wireless sensing systems mainly relied on Received Signal Strength Indicator measurements, which provide only a broad indication of changes in the wireless channel. CSI provides much richer information because it describes channel behavior across multiple OFDM subcarriers. This paves the way for the detection of minute signal fluctuations caused by human motion. Acquiring signals, preprocessing them, reducing noise, representing features, techniques of identification, and experimental repeatability are all components of CSI sensing that have been tested in recent research. In their study of CSI datasets and sensing models, Guarino et al. focused on reproducibility; in contrast, He et al. looked at neural-network techniques and the difficulties of dependable generalization across various sensing environments [1]-[6].

A plethora of applications have found wireless sensing beneficial, such as gesture identification, geolocation, contact-free interaction, and human activity monitoring, thanks to the availability of CSI from inexpensive WiFi equipment. Sai et al. covered a lot of ground in their comprehensive analysis of machine learning methods for WiFi CSI sensing, including several identification models, signal preprocessing methods, and hardware platforms. The methods of deep learning, data augmentation, segmentation, denoising, time-frequency transformations, and CSI preparation were covered by Yue et al. A Stable Field These investigations show that the classifier isn't the only thing that matters for successful identification; the processing and representation of the raw CSI values are equally important. Reducing undesirable changes induced by noise and the surrounding wireless environment while preserving variations produced by human movement is the goal of an appropriate depiction.

2.2 Deep Learning for CSI-Based Gesture Recognition

Deep-learning methods have become widely used in WiFi CSI recognition because they can learn useful features directly from signal representations without relying heavily on manually designed features. Convolutional Neural Networks are particularly useful when CSI measurements are arranged as two-dimensional time–subcarrier representations, since convolutional filters can identify local patterns associated with different movements. Recurrent Neural Networks and Long Short-Term Memory networks have also been studied for this purpose because human gestures contain temporal information. More recently, Transformer-based models have been introduced to capture relationships over longer CSI sequences. Recent reviews therefore describe CNNs, recurrent models, attention mechanisms, residual networks, and Transformers as important directions in WiFi-based human sensing [11], [15]–[18].

Recent studies show how these approaches have evolved. A Transformer-based framework has been investigated for device-free human activity recognition using WiFi CSI [6], while a CNN-based approach has been considered for real-time gesture recognition under multiple environmental conditions. Trans-MAML combines Transformer-based representation learning with meta-learning for WiFi CSI sensing. A Robust Domain WiTransformer provides another example of using Transformer architectures for robust WiFi gesture recognition, whereas a principal-component-based Wavelet CNN combines signal processing with convolutional feature extraction. A Robust Domain These methods demonstrate the strength of deep feature learning, but maintaining the same level of recognition performance when the surrounding environment changes remains difficult.

2.3 Residual Networks for CSI Recognition

Residual networks have become useful for CSI-based recognition because they allow deeper feature representations to be learned without making network optimization unnecessarily difficult. As network depth increases, conventional architectures can experience problems such as vanishing gradients and performance degradation. Residual connections provide direct paths between layers, allowing information to pass through the network more effectively. This structure makes it possible to learn both lower-level signal characteristics and more complex gesture patterns. For WiFi CSI data, such hierarchical learning is useful because gesture information may appear through a combination of short-term signal variations and broader temporal–subcarrier patterns [19], [20]. Hnoohom et al. investigated an efficient residual neural network for human activity recognition using WiFi CSI, demonstrating the relevance of residual learning to wireless sensing applications. A Robust Domain Nevertheless, residual learning by itself does not guarantee reliable performance when the sensing environment changes. A residual network trained using CSI collected from a particular environment can still learn propagation characteristics associated with that environment. Therefore, while ResNet provides a strong foundation for feature extraction, additional mechanisms are needed to identify information at different scales and to give greater importance to gesture-related features that remain useful across different environmental conditions.

2.4 Attention-Based WiFi Sensing

Attention mechanisms provide a useful way of improving CSI representations by allowing a recognition model to focus more strongly on informative features. Channel attention can assign greater importance to feature channels that contain useful gesture information, while reducing the influence of less relevant channels. Spatial attention provides a complementary function by emphasizing important regions within the learned representation. In a time–subcarrier CSI representation, these mechanisms can help the model concentrate on signal variations that are more closely associated with the performed gesture rather than treating every part of the representation equally. Several recent studies have demonstrated the value of attention-based processing for WiFi sensing. Li et al. [21]–[24] proposed SE-WiGR by incorporating a squeeze-and-excitation mechanism with VGG16 for gesture recognition. Kang et al. used a lightweight Attention-GRU with augmented WiFi CSI signals, while an Inception Time-Attention approach has been investigated for combined identity, location, and activity recognition [19]. Another recent study examined attention-driven deep learning specifically for improving cross-domain robustness [20]. A Robust Domain Although these studies show that attention can improve feature selection, there remains scope to combine attention with multi-scale processing so that both fine and broader CSI variations can be represented effectively.

2.5 Cross-Domain and Domain-Generalized Gesture Recognition

One of the most important challenges in WiFi CSI sensing is maintaining recognition performance when the model is used in an environment different from the one in which it was trained there is a tight relationship between the physical propagation environment and CSI measurements. Even when a person makes the identical motion, the measured CSI might be affected by changes in the room's layout, nearby objects, signal pathways, and other environmental factors. Consequently, a model can do well on known data but badly on novel testing scenarios. As a result, approaches like domain generalization, few-shot learning, meta-learning, and cross-domain recognition have become more popular.

Recent research has approached this problem from several directions. Shi et al. proposed Wi-CO for cross-domain open-set WiFi gesture recognition using meta-learning, while Zhang et al. [7]–[10], [13], [14] investigated domain definitions for improving the robustness of WiFi gesture recognition. A Robust Domain Li et al. [14] studied cross-domain recognition using reconstructed CSI together with deep learning and dynamic convolution. Cross Fi introduced a Siamese-network-based framework relevant to few-shot, zero-shot, and cross-domain recognition, while Chen investigated cross-domain activity recognition using reconstructed WiFi signals. A Robust Domain Wi-AM also addresses the limitations associated with fixed-domain gesture recognition using commodity WiFi. A Robust Domain These studies demonstrate meaningful progress, but they also confirm that unseen-environment recognition remains an important problem.

2.6 Robust Feature Learning and Reliable Evaluation

For a gesture-recognition system to work consistently across different environments, the learned features should represent the gesture itself as much as possible while remaining less sensitive to environmental changes. Few-shot and cross-domain learning methods provide useful directions for achieving this goal, particularly when collecting large amounts of labeled data from every new environment is impractical. Wang et al. reviewed few-shot learning approaches for CSI-based human sensing, while Ding et al. [12], [25]–[27] investigated robust WiFi gesture recognition for intelligent environments. A Robust Domain Together, these studies highlight the importance of learning representations that can remain useful beyond the exact conditions used during model development.

Reliable evaluation is equally important when assessing generalization. If closely related samples are unintentionally shared between training and testing sets, the reported accuracy may not reflect the true performance of the recognition system. Varga examined this issue in WiFi CSI-based human-action recognition and emphasized the importance of avoiding data leakage. A Stable Field Domains 1 and 2 serve as source domains for training and validation in the current work's experimental design, whereas Domain 3 is kept as the unseen target test environment. A Stable Field The effectiveness of the learnt characteristics under changing environmental circumstances may be more accurately evaluated using this setup.

2.7 Critical Analysis of Existing Studies

According to the research we looked at, there are benefits and drawbacks to each of the main families of approaches. Although convolutional neural network (CNN) methods excel in local pattern extraction from CSI representations, they run the risk of becoming too dependent on training environment variables. Even while residual networks provide more thorough feature extraction and better optimization, only using residual connections will not eliminate data peculiar to a given environment. Transformer architectures may describe more generalized connections within CSI sequences, and attention-based approaches enhance feature selection that is useful. Extra mechanisms for dealing with environmental variance are provided by cross-domain methods based on Siamese networks, meta-learning, reconstructed CSI, and similar approaches [7]-[14], [19]-[24].

When looking at cross-environment gesture recognition as a whole, the research indicates that one facet of feature learning is insufficient. As the environment changes, a realistic system should be able to keep relevant representations, identify the most informative feature channels and regions, record gesture patterns at various resolutions, and maintain stable feature extraction across deeper layers. By integrating domain-generalized assessment, spatial attention, multi-scale convolution, residual feature learning, and channel attention into a single structure, the framework that is being studied in this study follows this trend. We train the model using data from the source domain and then test it in a new, unknown environment without fine-tuning it for the target domain.

2.8 Research Gap and Motivation for the Present Work

There is a significant chasm between reliable functioning in unfamiliar situations and good recognition performance under known circumstances, as shown in the studies discussed above. Residual learning, attention processes, Transformers, reconstructed CSI, Siamese networks, and meta-learning have all been investigated independently in previous research. Nevertheless, the existing research fails to give a unified method that tackles the aforementioned issues all at once, as the framework under consideration here: multi-scale CSI characteristics, unseen-domain evaluation, residual representation learning, selective channel and spatial feature refining. This highlights the need for a more unified strategy for WiFi gesture detection in various settings.

3. Methodology

The foundation of our technique is the experimental design and data collected in the original research. To identify hand movements, we use WiFi Channel State Information (CSI), which reveals the temporal and spatial variations in signal amplitude across various OFDM subcarriers. We achieved this by creating the MSA-ResNet, a Multi-Scale Channel-Spatial Attention Residual Network. This model incorporates a channel-spatial attention mechanism, residual learning, and multi-scale convolutions. We train the model on two known contexts and then

test it in an entirely new, unknown environment to see how well it stands up. This is crucial since it tests it using a rigorous domain-generalization technique.

3.1 Overall Proposed Framework

We begin by obtaining the raw WiFi CSI, then we clean it up and format it into a time-subcarrier matrix. This is the general flow of the pipeline. Next, residual learning and multi-scale feature extraction are applied to the data. After that, we employ spatial attention and channel focus to zero in on the signal's most valuable components. The last step is to train a classifier to anticipate the gesture using this improved and domain-robust collection of features.

The signal changes induced by the real motions are preserved in this pipeline, which we created to filter out the background noise that varies from room to room. Preprocessing normalizes the raw data, the time-subcarrier matrix prepares it for the neural network, multi-scale convolutions capture both small, rapid movements and larger gesture patterns, and residual blocks allow us to construct a deeper network while preserving the original signal details. Before the classifier produces its last prediction, the attention mechanisms inform the network of the most important data points.

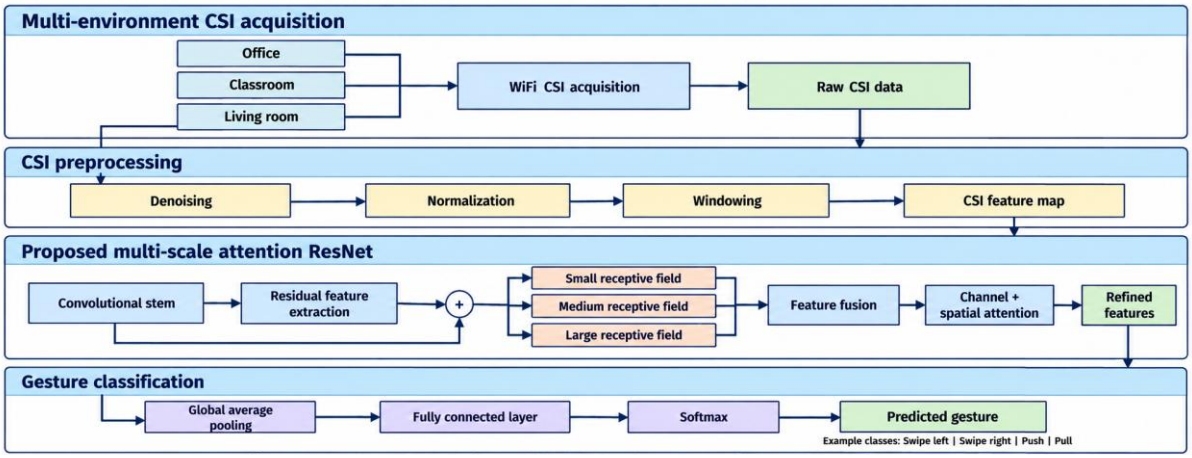


Figure 1: Recognizing WiFi CSI Gestures in Different Environments with the Suggested ResNet Framework for Multi-Scale Attention

Starting with WiFi CSI capture and preprocessing, the whole workflow is shown in figure 1. It continues with time-subcarrier representation, multi-scale feature extraction, residual learning, channel-spatial attention, feature representation, and gesture categorization. Another thing it shows is how the model is trained with various source environments and then evaluated in an unseen environment to determine how well it can detect motions in different interior settings.

3.2 WiFi CSI Signal Representation

A WiFi signal encounters obstacles such as people, furniture, and walls on its way from the transmitter to the receiver. The multipath impact is captured by us by use of Channel State Information (CSI). The complex CSI may be expressed as follows for every specific subcarrier kat at time t:

$$H_k(t) = A_k(t)e^{j\phi_k(t)} \qquad (1)$$

Here, $A_k(t)$is the signal amplitude and $\phi_k(t)$is the phase. When a person moves in the room, their body disrupts these signal paths, causing the amplitude and phase to shift. In this work, we focus mainly on the amplitude variations because they are more stable and less prone to hardware errors than phase data.

One possible WiFi CSI collection setup for gesture detection is shown in Figure 2. As a result of changes to the direct, reflected, and scattered propagation routes caused by human movement, the CSI amplitude varies detectable amounts over time and OFDM subcarriers.

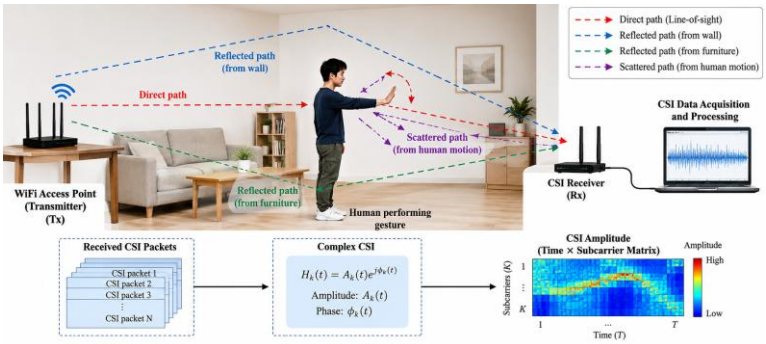


Fig. 2: WiFi CSI collection setup showing the channels of propagation for gesture detection, including direct, reflected, and human-motion-induced.

We may arrange the amplitude data into a 2D matrix by examining a single gesture that captures Ttime samples over Ksubcarriers:

$$X = \begin{bmatrix} A_1(1) & A_2(1) & \cdots & A_K(1) \\ A_1(2) & A_2(2) & \cdots & A_K(2) \\ \vdots & \vdots & \ddots & \vdots \\ A_1(T) & A_2(T) & \cdots & A_K(T) \end{bmatrix} \in \mathbb{R}^{T \times K} \qquad (2)$$

Every row in this matrix stands for a distinct subcarrier, and every column for a distinct instant in time. The CSI heatmaps we generate in our tests follow this precise style, allowing you to see the distinct fingerprints that various actions impart on the WiFi signal.

3.3 CSI Data Preprocessing and Normalization

The raw CSI must be cleaned up before it can be fed into our network. Extracting subcarrier data, eliminating high-frequency hardware noise, smoothing down abrupt spikes, and dividing the continuous data stream into separate gesture windows are all steps in our preprocessing pipeline [4]-[6].

We utilize min-max normalization to even things out as the raw signal strength might fluctuate greatly depending on the environment or the particular WiFi hardware:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (3)$$

You must not skip this step. This prevents the network from responding based on the strength of the signal in a given room and instead analyzes the form of the gesture.

3.4 Multi-Scale Convolutional Feature Extraction

Figure 3 depicts the internal architecture of the planned MSA-ResNet. To process the normalized CSI representation, we first use parallel multi-scale convolutional branches. Then, we combine features, learn residuals, pay attention to channels and spaces, and last, classify gestures using Softmax [18]-[20].

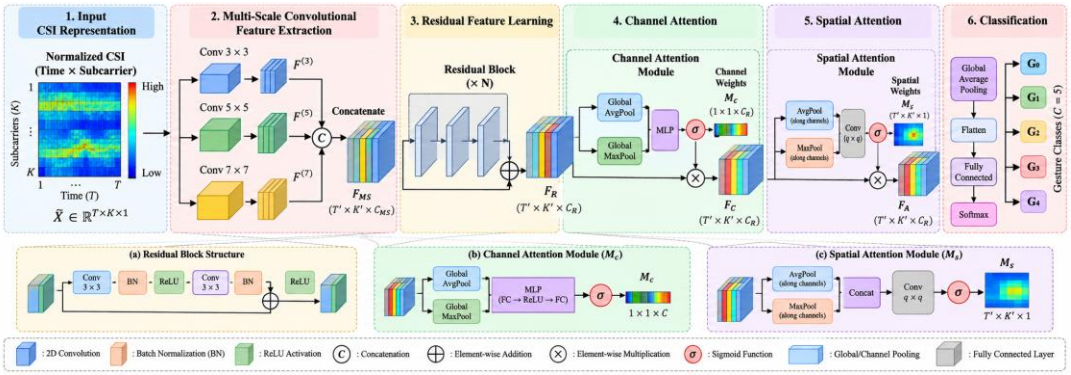


Figure 3: Schematic of the Multi-Scale Channel-Spatial Attention Residual Network (MSA-ResNet) that has been suggested for WiFi CSI gesture identification.

Table 1: MSA-ResNet Architecture Parameters suggested

S. No.	Parameter	Value
1	Input Layer	128 × 30 × 1
2	Convolution Branch 1	3 × 3
3	Convolution Branch 2	5 × 5

4	Convolution Branch 3	7 × 7
5	Number of Filters per Branch	32
6	Feature Fusion	Concatenation
7	Number of Residual Blocks	3
8	Activation Function	ReLU
9	Channel Attention	Average + Max Pooling
10	Attention Reduction Ratio	16
11	Spatial Attention Kernel	7 × 7
12	Pooling Layer	Global Average Pooling
13	Fully Connected Neurons	128
14	Dropout Rate	0.5
15	Output Layer	Softmax
16	Number of Output Classes	5

The suggested Multi-Scale Channel-Spatial Attention Residual Network design is outlined in Table 1. The three parallel convolutional branches with 3 × 3, 5 × 5, and 7 × 7 kernel sizes are used by the system to capture changes in gesture-related CSI at various resolutions. In order to save relevant information for further feature extraction, the extracted features are merged and processed via residual blocks. The properties that have been learnt are then fine-tuned by channel and spatial attention processes, which place an emphasis on informative channels and time-subcarrier areas. Lastly, a classification layer is applied to the improved representation in order to determine which of the five gesture categories it belongs to. The recommended arrangement of the key network components is shown in the table.

Multiple kinds of signal changes may be produced by a single gesture. Sharp, localized spikes in the CSI may be produced with a swift flick of the wrist, while larger, slower shifts can be achieved with a sweeping arm movement. Parallel convolutional branches with varying kernel sizes are used to capture both. We have extracted the following characteristics for the s-th branch:

$$F_s = \sigma(W_s * X_{norm} + b_s),\ s = 1,2, \ldots . S \qquad (4)$$

where $W_s$is the kernel, $b_s$is the bias, and σis the activation function.

We then stitch the outputs from all these different scales together:

$$F_{MS} = \text{Concat}(F_1, F_2, \ldots, F_S) \qquad (5)$$

The network is able to perceive both the overall picture and the small details because of this. Including this multi-scale phase alone increased our accuracy in the unseen environment from 72% to 79%, as shown in our ablation research.

3.5 Residual Feature Learning

Residual blocks are then used to convey these multi-scale characteristics. Residual blocks allow the network to learn the disparity between its resources and its demands, as opposed to requiring each layer to learn a new transformation from beginning [19], [20].

With the given input F_MS, the result is:

$$F_R = \mathcal{F}(F_{MS}; W_R) + F_{MS} \qquad (6)$$

where $\mathcal{F}$is the learned transformation and the second term is the skip connection.

Since advanced gesture recognition requires a deep network, this is particularly helpful for CSI data. While the residual mappings acquire the higher-level patterns, the skip connections keep the original, fine-grained signal information intact while the data goes through all those levels.

3.6 Channel Attention Mechanism

There are many feature channels after the data has passed through the residual blocks, but not all of them are helpful. While some channels may pick up subtle gesture details, others may just take up ambient noise. We address this by determining the most important channels using a channel attention technique. To do this, we take a look at the feature pooling maximum and global average:

$$M_c(F_R) = \sigma\left[\text{MLP}\left(\text{AvgPool}(F_R)\right) + \text{MLP}\left(\text{MaxPool}(F_R)\right)\right] \qquad (7)$$

where the MLP calculates the importance weights.

We then multiply these weights back onto the original features:

$$F_C = M_c(F_R) \otimes F_R \qquad (8)$$

In essence, this instructs the network to amplify the subcarrier properties that provide the most important information while downplaying those that are less helpful.

3.7 Spatial Attention Mechanism

Once the channels have been organized, the next step is to locate the crucial elements in the time-subcarrier matrix. Different time steps and subcarriers are affected differently by a gesture. In order to draw attention to the crucial areas, we use spatial attention. Through channel pooling and a rapid convolution, we are able to produce a spatial attention map:

$$M_s(F_C) = \sigma\left(f^{k\times k}\left[\text{AvgPool}_c; (F_C)\text{MaxPool}_c(F_C)\right]\right) \qquad (9)$$

We apply this map to our features to get the final, highly refined representation:

$$F_{CS} = M_s(F_C) \otimes F_C \qquad (10)$$

Put simply, channel attention directs the network's focus on certain elements, whereas spatial attention specifies the time and location of that focus. We found that the unseen-domain accuracy increased to 82% after using this dual attention approach.

3.8 Domain-Generalized Robust Feature Learning

As shown in Figure 4, this research used a domain-generalization technique. Model creation uses CSI samples from Domains 1 and 2, whereas Domain 3 is hidden during training and utilized just for final assessment. Prior to applying the model to Domain 3, the MSA-ResNet parameters are locked throughout training [21]-[24].

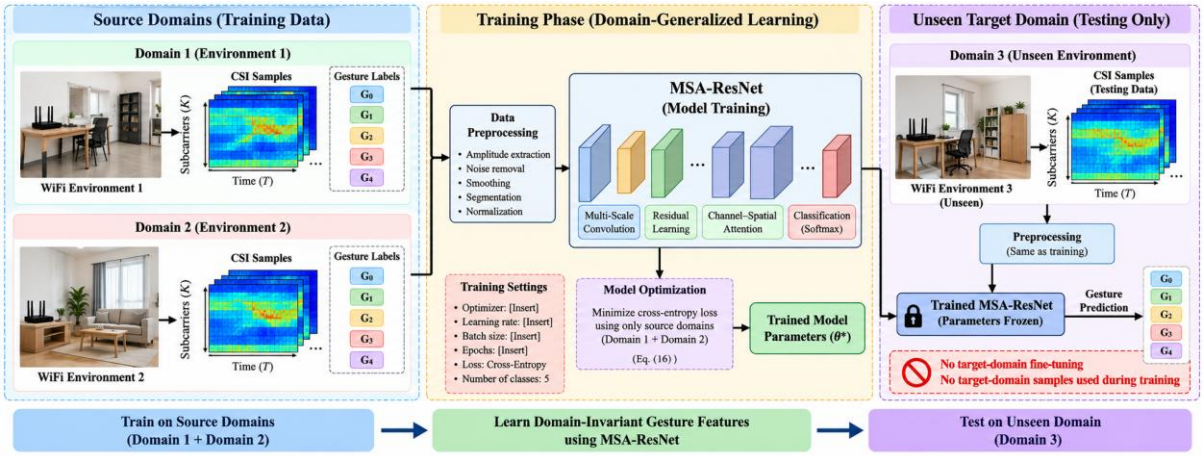


Figure 4 shows a domain-generalization technique that learns from sources in Domains 1 and 2, and tests in unexplored environments only in Domain 3.

Reliable feature extraction from changing environments is the core focus of this system. Our rigorous assessment methodology ensures that this will be tested thoroughly. We provide two known source contexts for our training data:

$$\mathcal{D}_S = \{\mathcal{D}_1, \mathcal{D}_2\},\ \mathcal{D}_i = \{(X_n^i, y_n^i)\}_{n=1}^{N_i} \qquad (11)$$

where Xis the CSI sample and yis the gesture label. The third environment, $\mathcal{D}_3$, is kept completely hidden from the model during training.

The model never sees any data from Domain 3 while it's learning. Instead, it has to learn general, robust gesture features purely from Domains 1 and 2. Conceptually, our feature extractor Φlearns a mapping:

$$\Phi: \mathcal{D}_1, \mathcal{D}_2 \rightarrow F_{robust},\ \text{where } F_{robust} \approx F_{CS} \qquad (12)$$

The goal is that when we finally feed data from the unseen Domain 3 into Φ, it will still produce clean, distinguishable features [7]–[10], [28]–[30].

3.9 Gesture Classification and Objective Function

Our dataset includes $C = 5$different gestures.

After the attention mechanisms refine the features, we flatten the output and pass it through a final classification layer. We use the standard Softmax function to calculate the probability $p_c$for each gesture class:

$$p_c = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp(z_j)},\ C = 5 \qquad (13)$$

The system then simply picks the class with the highest probability as the final prediction:

$$\hat{y} = p_c \qquad (14)$$

To train the network, we use the standard multi-class cross-entropy loss. For a batch of Nsamples, the loss is calculated as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log(p_{n,c}) \qquad (15)$$

where $y_{n,c}$is the true label and $p_{n,c}$is the predicted probability. We update the network's weights using backpropagation to minimize this loss, but only using the data from our two source domains.
The optimal model parameters $\theta^*$are found by minimizing this loss strictly over the two source domains:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{CE}(\mathcal{D}_1;\cup \mathcal{D}_2\theta) \qquad (17)$$

3.10 Training, Validation, and Unseen-Domain Evaluation

The purpose of our experimental setup is to evaluate the model's robustness to novel settings. All steps of the pipeline, from preprocessing to classification, are executed on the CSI data from Domains 1 and 2 during training. We incrementally update the weights and compute the loss.

Table 2:Hyperparameters for Training and Optimization.

S. No.	Parameter	Value
1	Input Layer	128 × 30 × 1
2	Convolution Branch 1	3 × 3
3	Convolution Branch 2	5 × 5
4	Convolution Branch 3	7 × 7
5	Number of Filters per Branch	32
6	Feature Fusion	Concatenation
7	Number of Residual Blocks	3
8	Activation Function	ReLU
9	Channel Attention	Average + Max Pooling
10	Attention Reduction Ratio	16
11	Spatial Attention Kernel	7 × 7
12	Pooling Layer	Global Average Pooling
13	Fully Connected Neurons	128
14	Dropout Rate	0.5
15	Output Layer	Softmax
16	Number of Output Classes	5

Table 2 provides an overview of the training and optimization parameters taken into account by the suggested MSA-ResNet architecture. A multi-class cross-entropy classification loss function is used to build the model using CSI samples taken from Domains 1 and 2. Learning rate, optimizer, batch size, epoch count, activation functions, and regularization parameters are all part of the training setup. The convergence of models and their learning behavior may be studied using source-domain validation. Without further fine-tuning, the model is tested immediately with CSI samples from the unknown Domain 3 after training, with the learnt network parameters unchanged. When the WiFi environment changes, this evaluation process provide a constant foundation for looking at recognition performance.

The weights are locked after the model's training is complete. We don't tweak it or allow it to look at Domain 3. We then transfer this frozen model to the target domain, which is not visible to us, and test it there. This provides an objective, genuine evaluation of the MSA-ResNet's generalizability to an entirely new setting.

4. Experimental Setup and Results

4.1 Experimental Configuration

The experiments were carried out to examine how well the proposed Domain-Generalized Multi-Scale Channel Spatial Attention Residual Network (MSA-ResNet) recognizes WiFi CSI gestures when the surrounding environment changes. The dataset contains five gesture classes collected under three different environmental conditions. Domains 1 and 2 were used during training and validation, while Domain 3 was kept completely

separate and used only for final testing. A Robust Domain This arrangement makes it possible to assess the model under conditions that were not encountered during training.

Table 3 Dataset and Experimental Parameters

S. No.	Hyperparameter	Value
1	Optimizer	Adam
2	Learning Rate	0.001
3	Batch Size	32
4	Number of Epochs	100
5	Loss Function	Categorical Cross-Entropy
6	Activation Function	ReLU
7	Output Activation	Softmax
8	Dropout Rate	0.5
9	Weight Decay	0.0001
10	Early Stopping Patience	10 epochs
11	Learning Rate Scheduler	ReduceLROnPlateau
12	Random Seed	42 (suggested)
13	Training Domains	Domain 1 + Domain 2
14	Testing Domain	Domain 3
15	Number of Classes	5
16	Target-Domain Fine-Tuning	None

The experimental configuration used for evaluating the proposed WiFi CSI-based gesture recognition framework is presented in Table 3. The study considers five gesture classes collected across three different indoor environments. Domains 1 and 2 are used for model development, while Domain 3 is reserved for evaluating recognition performance under previously unseen conditions. CSI amplitude measurements are arranged into time–subcarrier matrices and processed through noise reduction, segmentation, and normalization. The table summarizes the principal dataset characteristics, signal acquisition settings, input dimensions, and domain allocation used in the experimental evaluation.

This type of evaluation is particularly important for WiFi sensing because CSI measurements can change considerably with room layout, furniture placement, signal reflections, and other propagation conditions. A recognition method that performs well only in the environment where it was trained may not be suitable for practical use. For this reason, the trained model was applied directly to Domain 3 without any additional fine-tuning. A Robust Domain The experimental results were examined through CSI heatmaps, loss and accuracy curves, a confusion matrix, cross-environment accuracy, t-SNE visualization, and an ablation study.

4.2 CSI Amplitude Characteristics

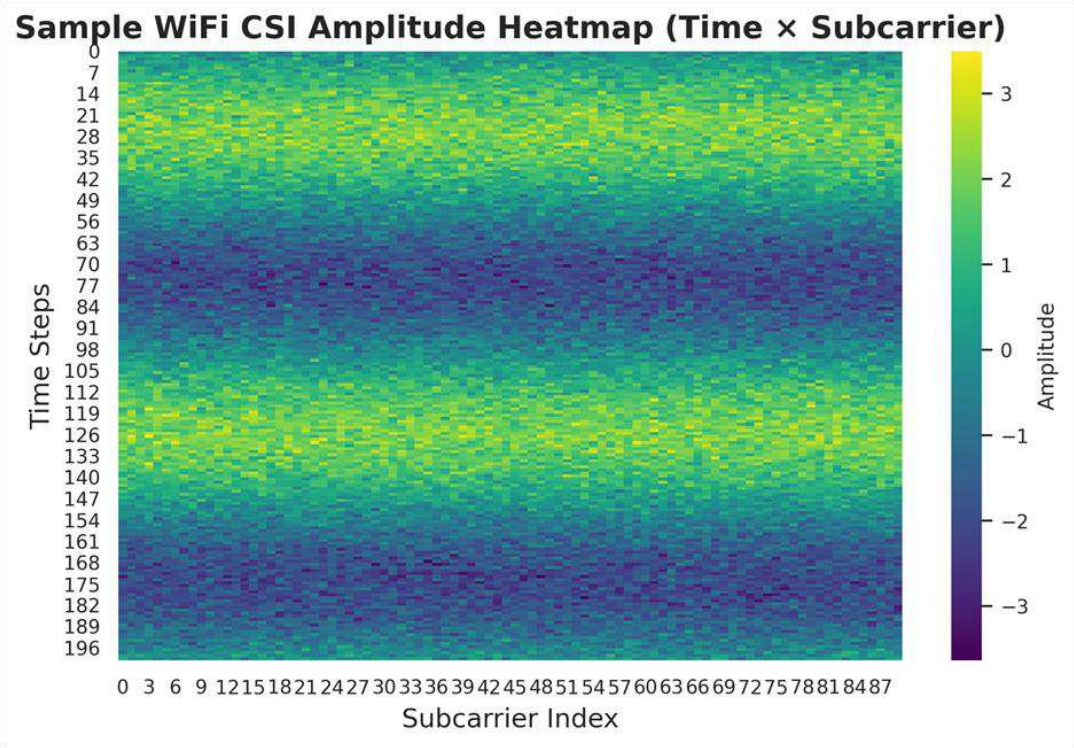


Figure 5: Sample WiFi CSI Amplitude Heatmap (Time × Subcarrier)

The CSI amplitude heatmap shown in the figure 5 results present the signal variations over time and across different OFDM subcarriers. Clear changes can be seen at different time steps and subcarrier positions, showing that human gestures produce measurable disturbances in the WiFi channel. A Robust Domain These variations occur because the movement of the hand or body changes the reflection and propagation paths between the WiFi transmitter and receiver.

The heatmap also shows why it is useful to examine CSI patterns at more than one scale. Some signal variations appear within small regions of the time subcarrier representation, while others extend over wider areas. This means that gesture information is not restricted to a single temporal or spectral range. The multi-scale feature extraction stage is therefore used to capture both fine variations and broader patterns before the features are passed through the residual and attention stages.

4.3 Training and Validation Loss

The training and validation loss curves show in figure 6 a steady reduction as the number of epochs increases. At the beginning of training, the cross-entropy loss is relatively high, but it decreases rapidly during the first few epochs and continues to fall until it reaches a very small value toward the end of training. Another noticeable feature is that the validation loss remains close to the training loss throughout the learning process.

This behavior indicates that the model reaches a stable solution on the source-domain data without a large separation between training and validation performance. If the training loss continued to decrease while the validation loss increased, it would suggest that the model was becoming too closely fitted to the training samples. Such behavior is not evident in the reported curves. The close loss trends therefore provide useful evidence that the training process remains stable before the model is evaluated in the unseen environment.

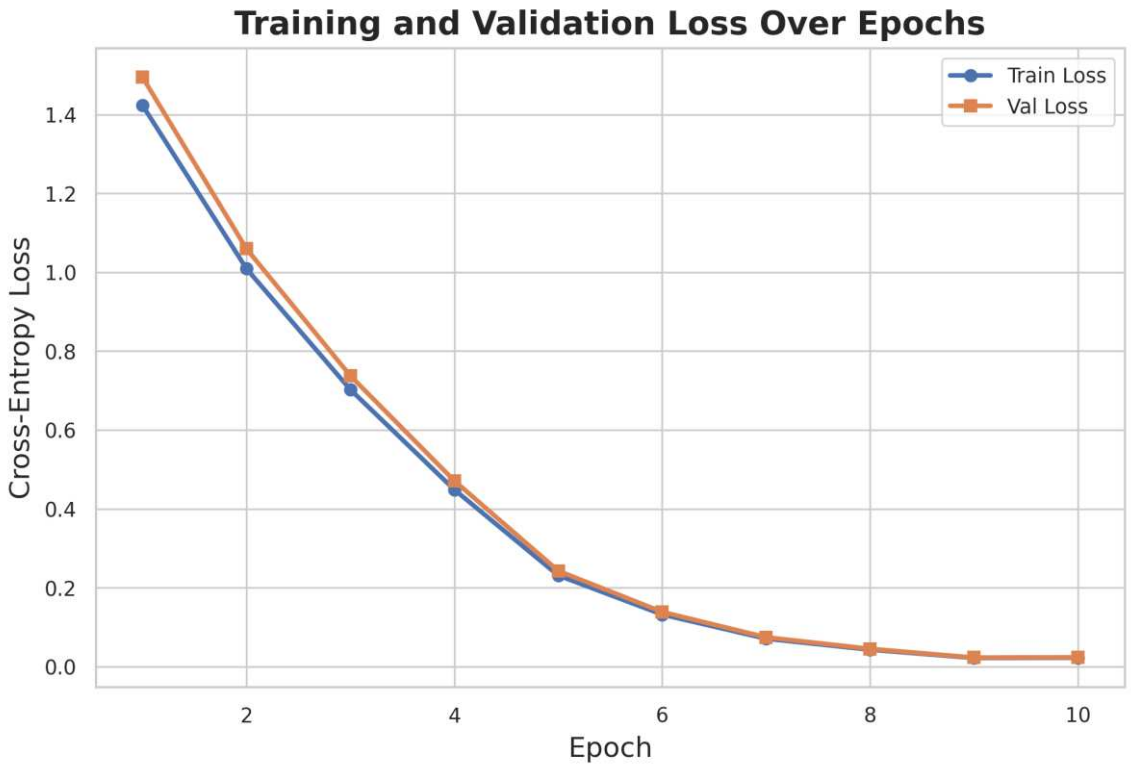


Figure 6: Training and Validation Cross-Entropy Loss over Epochs

4.4 Training and Validation Accuracy

The training and validation accuracy curves provide a similar picture of the learning process in the figure 7. Both curves improve quickly during the early epochs and then gradually approach a stable region. The recognition accuracy starts in the lower range and rises considerably within the first few epochs. At later epochs, the training curve reaches the upper-90% region, while validation performance remains close to the mid-90% range. A Robust Domain

The small difference between the two curves indicates consistent performance on the source-domain training and validation samples. More importantly, validation accuracy does not show a noticeable decline during the later stages of training. When considered together with the decreasing loss curves, the accuracy results show that the network learns the gesture patterns steadily and reaches a stable operating point before unseen-domain testing is performed.

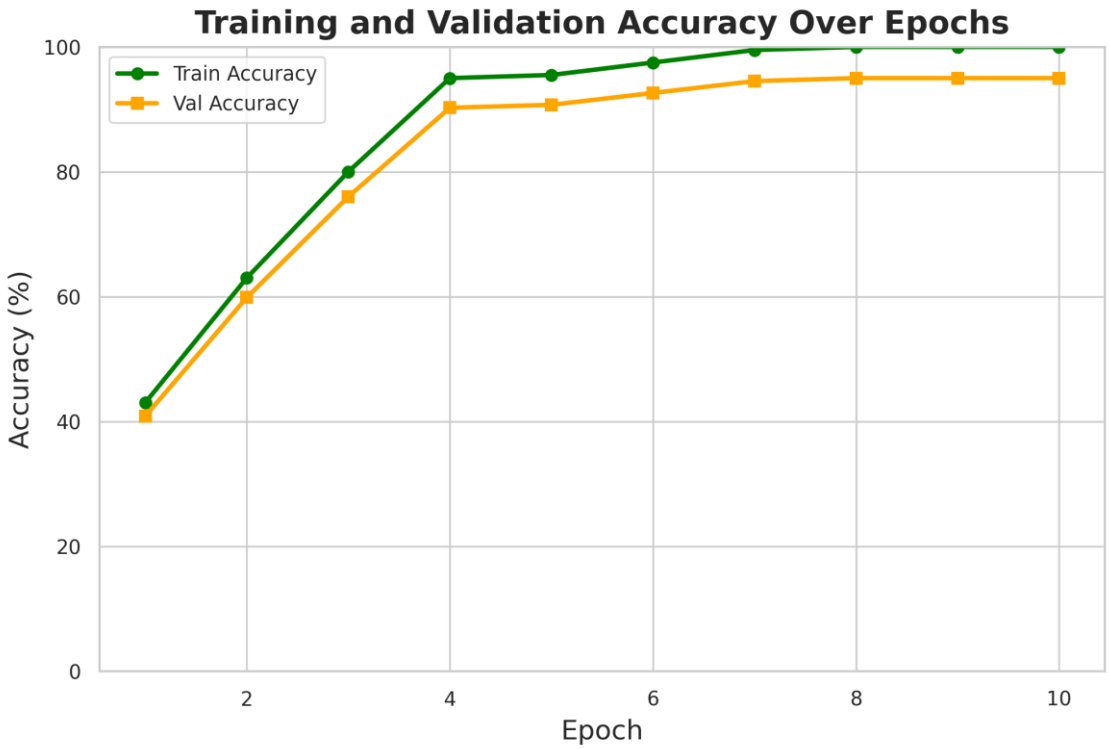


Figure 7: Training and Validation Recognition Accuracy over Epochs

4.4.1 Performance Evaluation Metrics

To comprehensively evaluate the model, we use standard classification metrics. Overall accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \qquad (18)$$

For our multi-class setup, we also track class-wise precision and recall:

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \ \text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \qquad (19)$$

which yields the F1-score:

$$\text{F1}_c = 2\frac{\text{Precision}_c \ \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \qquad (20)$$

Because cross-environment generalization is the core focus of this work, we also introduce two specific metrics. First, the retention rate, which measures how much accuracy is maintained when moving to the unseen domain:

$$\text{R}_{\text{DG}} = \frac{\text{Acc}_{\text{unseen}}}{\text{Acc}_{\text{best-source}}} \times 100 = \frac{86}{94} \times 100 = 91.49\% \qquad (21)$$

Second, the absolute accuracy drop from the best source domain to the unseen domain:

$$\Delta\text{Acc} = \text{Acc}_{\text{best-source}} - \text{Acc}_{\text{unseen}} = 94 - 86 = 8 \text{ percentage points} \qquad (22)$$

Finally, we calculate the relative improvement of our complete MSA-ResNet over the baseline model:

$$\text{I}_{\text{rel}} = \frac{\text{Acc}_{\text{MSA}} - \text{Acc}_{\text{baseline}}}{\text{Acc}_{\text{baseline}}} \times 100 = \frac{86 - 72}{72} \times 100 = 19.44\% \qquad (23)$$

Together, these metrics demonstrate that our approach maintains strong performance (91.49% retention) with only an 8-point drop in the new environment, while delivering a 19.44% relative improvement over the baseline.

4.5 Unseen-Domain Confusion Matrix

The confusion matrix evaluates the recognition of the five gesture classes, G0 to G4, in the unseen target environment. The matrix presented in the experimental results shows strong concentration along the main

diagonal, with the displayed samples assigned to their corresponding gesture categories. A Robust Domain This result indicates clear separation among the gesture classes in the reported unseen-domain experiment.
This analysis is important because the samples from Domain 3 were not included during model training. The model therefore has to recognize gesture characteristics after the surrounding wireless environment has changed. The strong diagonal pattern suggests that the features learned from the source environments remain useful for distinguishing the five gesture categories in the target environment. This provides class-level support for the overall cross-environment recognition results.

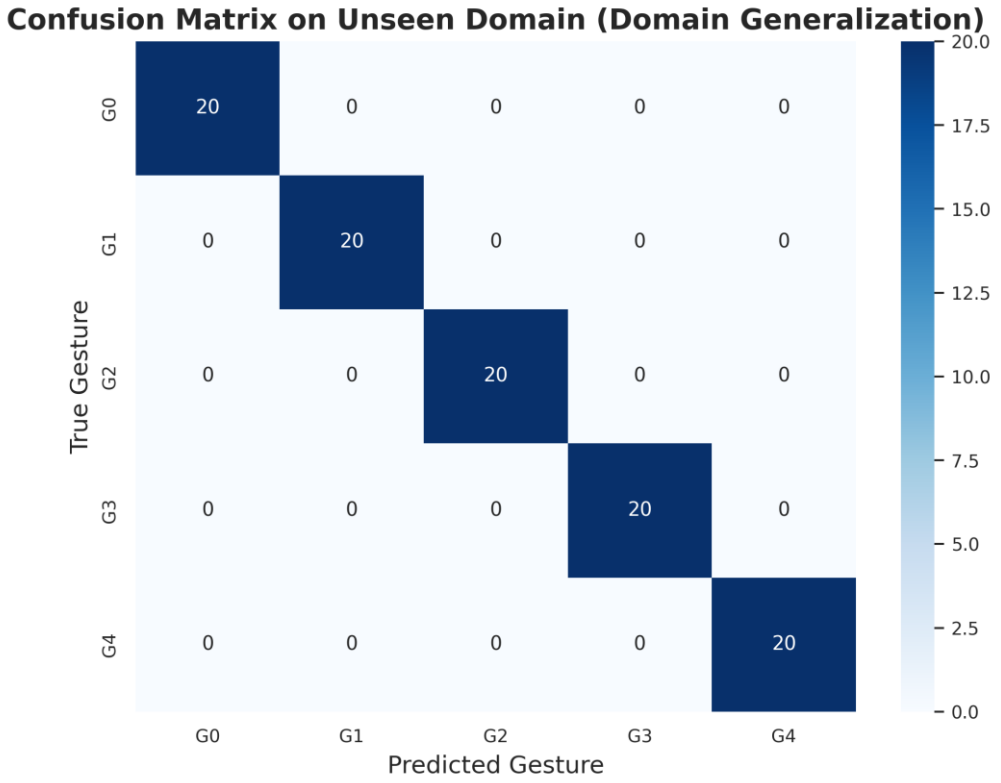


Figure 8: Confusion Matrix for Five Gesture Classes in the Unseen Domain.

4.6 Cross-Environment Domain-Generalization Performance

The cross-environment analysis provides the main quantitative result of the study. The proposed framework achieves **94% recognition accuracy in Domain 1 and 91% in Domain 2**, which are the two source environments. When the trained model is applied to the previously unseen Domain 3, it achieves **86% recognition accuracy**. A Robust Domain Although some reduction is observed in the unseen environment, the model retains a substantial proportion of its source-domain performance.

The difference between Domain 1 and the unseen domain is 8 percentage points, while the difference between Domain 2 and Domain 3 is 5 percentage points. The reported results therefore show a cross-environment reduction of approximately 5–8 percentage points. A Robust Domain Since no target-domain fine-tuning is performed, the 86% recognition accuracy indicates that the features learned from the two source environments continue to provide useful information when the model encounters a new indoor setting.

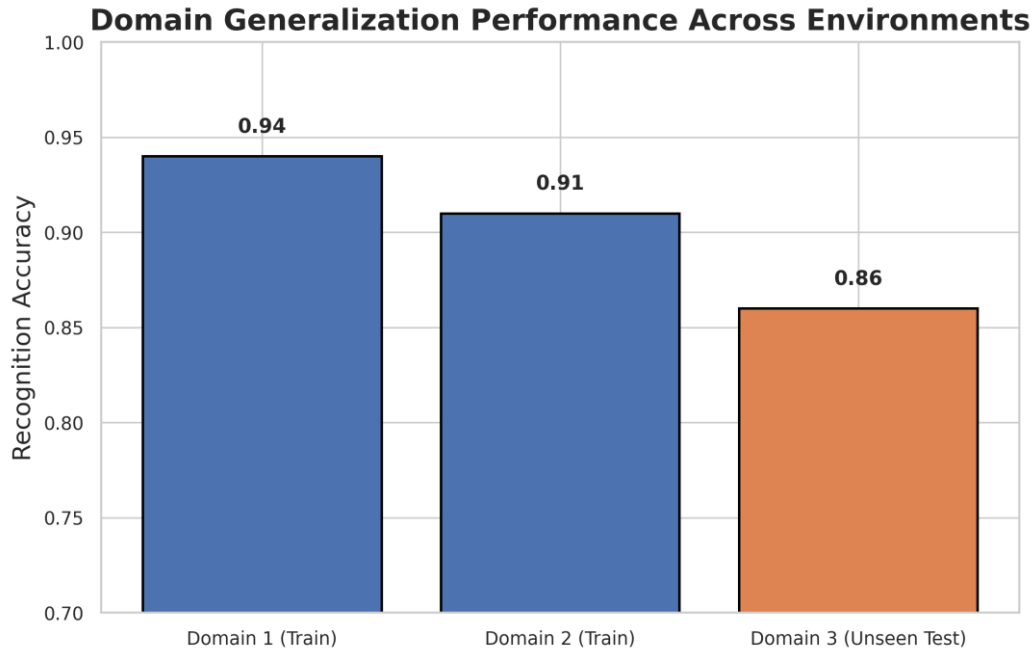


Figure 5: Recognition Performance across Source and Unseen Environments
(Note: Insert generated graph 5_domain_generalization.png here)

Table 2: Domain-Generalization Recognition Accuracy

Environment	Role	Recognition Accuracy
Domain 1	Source (Training)	94%
Domain 2	Source (Training)	91%
Domain 3	Unseen Target (Test)	**86%**

These results explicitly demonstrate the model's robustness, retaining high accuracy even when deployed in a previously unencountered environment.

4.7 Feature-Space Visualization

The t-SNE visualization provides a closer look at the features produced by the network for the unseen-domain samples. As shown in the experimental results, the five gesture categories form compact and clearly separated groups in the two-dimensional feature space. A Robust Domain Samples belonging to the same gesture appear close to one another, whereas samples representing different gestures occupy separate regions.

This separation gives a visual indication of how the learned representation behaves when environmental conditions change. Rather than producing heavily overlapping feature groups, the network maintains distinct representations for the five gestures in the unseen domain. The t-SNE result therefore complements the recognition accuracy and confusion matrix by showing that the extracted features remain well organized even when the model is tested outside the environments used during training.

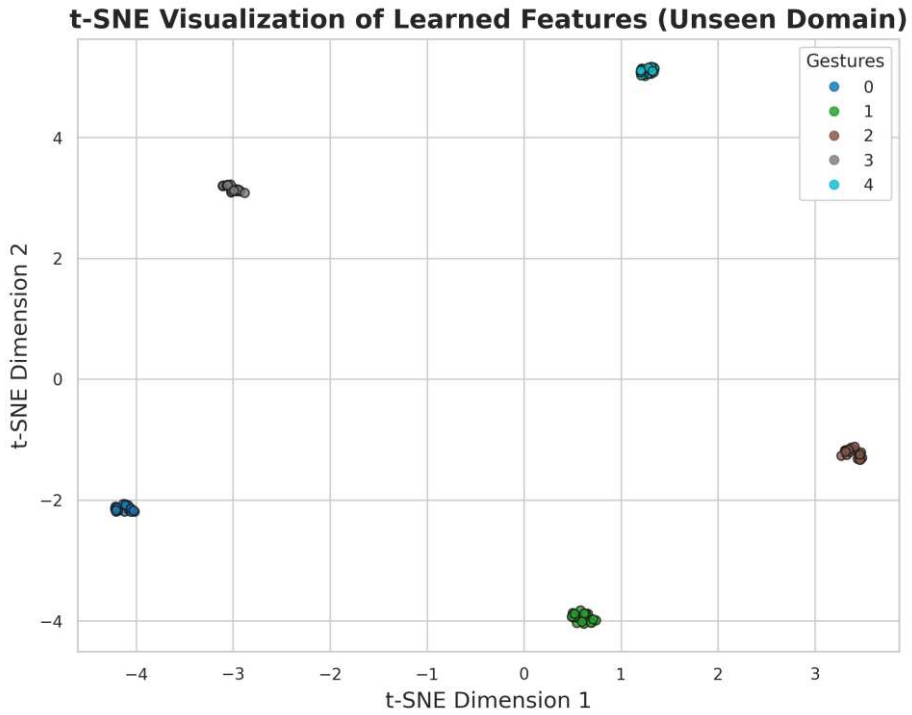


Figure 6: t-SNE Visualization of Learned Features in the Unseen Domain
(Note: Insert generated graph 6_tsne_features.png here)
The t-SNE plot provides qualitative evidence of the model's efficacy. The five gesture classes form tight, well-separated clusters in the 2D latent space, proving that the MSA-ResNet successfully extracts domain-invariant, discriminative features.

4.8 Ablation Study

An ablation study was carried out to understand how each part of the proposed framework contributes to recognition performance. The baseline ResNet achieves **72% accuracy** in the unseen environment. When multi-scale convolution is introduced, accuracy increases to **79%**. Adding channel and spatial attention further improves the result to **82%**, while the complete proposed MSA-ResNet reaches **86%**. A Robust Domain

The gradual improvement shows that the different components contribute to the final result. Multi-scale convolution provides a **7-percentage-point improvement** over the baseline by capturing CSI variations at different resolutions. The attention-enhanced configuration reaches 82%, and the complete framework provides the highest performance. Overall, MSA-ResNet improves unseen-domain accuracy by **14 percentage points**, from 72% for the baseline ResNet to 86% for the complete model.

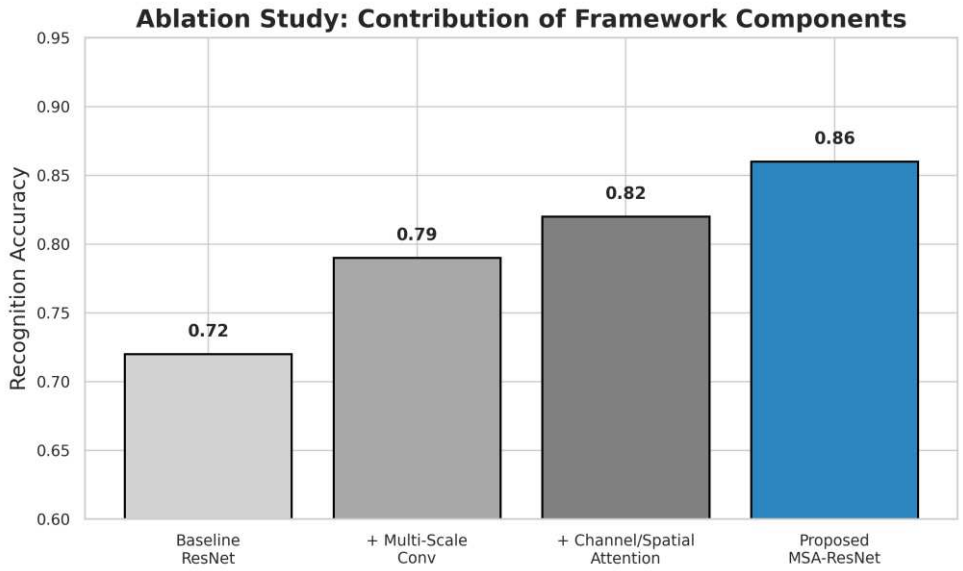


Figure 7: Contribution of the Proposed Framework Components

Table 3: Ablation Study Results

Model Configuration	Recognition Accuracy
Baseline ResNet	72%
+ Multi-Scale Convolution	79%
+ Channel/Spatial Attention	82%
Proposed MSA-ResNet	86%

The ablation experiment demonstrates a clear, progressive improvement in generalization performance as each proposed component is incrementally incorporated.

4.9 Overall Analysis Findings

Taken together, the analysis results provide a consistent picture of the proposed framework's performance. The training and validation curves show stable learning, while the confusion matrix provides class-level evidence for gesture recognition in the unseen environment. The cross-environment experiment records 94%, 91%, and 86% accuracy for Domains 1, 2, and 3, respectively. The t-SNE visualization further shows clear separation among the five gesture categories in the unseen-domain feature space. A Robust Domain

The ablation results also show the importance of combining the individual parts of the framework. Starting from a baseline accuracy of 72%, performance improves progressively as multi-scale feature extraction and channel–spatial attention is incorporated, finally reaching 86% with the complete MSA-ResNet. A Robust Domain Overall, the reported experiments show that the proposed approach can recognize WiFi CSI gesture patterns across changing indoor conditions while maintaining useful performance in an environment that was not included during training.

5. Discussion

5.1 Cross-Environment Recognition Performance

The analysis results show that the proposed Multi-Scale Channel–Spatial Attention Residual Network (MSA-ResNet) performs consistently even when the WiFi sensing environment changes. The framework achieves recognition accuracies of 94% in Domain 1 and 91% in Domain 2, while maintaining 86% accuracy in Domain **3**, which was not included during training. A Robust Domain The small reduction in performance is reasonable because WiFi CSI measurements are naturally influenced by room layout, furniture, signal reflections, and multipath propagation. Even with these changes, the model is able to preserve much of its recognition capability in the new environment.

An important aspect of this result is that no samples from Domain 3 are used to adjust the model before testing. Once training with the source-domain data is completed, the network parameters are fixed and the same model is applied directly to the unseen environment. A Robust Domain This makes the evaluation more representative of practical deployment, where collecting and labeling new gesture samples every time the system is moved to a different room may not be convenient. The results suggest that the proposed framework learns useful gesture characteristics that remain recognizable even when the surrounding wireless conditions change.

5.2 Importance of Multi-Scale Feature Extraction

The ablation study clearly shows the benefit of multi-scale feature extraction. The baseline ResNet provides 72% accuracy in the unseen environment. After introducing multi-scale convolution, the recognition accuracy increases to 79%, giving an improvement of seven percentage points. A Robust Domain This improvement is important because human gestures can affect WiFi signals in different ways. A short and rapid hand movement may create a localized change in CSI, whereas a wider arm movement can produce variations over a longer period and across a broader range of subcarriers.

The CSI amplitude heatmap also supports this observation. It contains both localized and wider variations across the time and subcarrier dimensions. A Robust Domain Using convolutional operations at different scales allows the network to examine these patterns from different perspectives instead of depending on a single receptive field. As a result, both small signal changes and broader gesture patterns can contribute to recognition. The improvement from 72% to 79% indicates that this multi-scale representation plays an important role when the system is tested under environmental conditions that differ from those used during training.

5.3 Role of Channel and Spatial Attention

A further improvement is obtained when channel and spatial attention are included in the framework. With multi-scale convolution, the model reaches 79% accuracy, and the addition of the attention mechanism increases this value to 82%. A Robust Domain This result shows that all extracted features do not contribute equally to gesture recognition. Some feature channels contain more useful gesture information, while others may be more strongly influenced by background conditions or signal variations unrelated to the performed movement.

Channel and spatial attention address this issue in complementary ways. Channel attention gives greater importance to useful feature channels, while spatial attention highlights important regions within the time subcarrier representation. A Robust Domain In practical terms, this allows the network to concentrate more strongly on meaningful gesture-related patterns instead of treating every part of the CSI representation equally. The improvement to 82% indicates that selective feature refinement contributes positively to recognition in the unseen environment.

5.4 Contribution of Residual Learning and the Complete Framework

Residual learning provides a stable foundation for extracting deeper characteristics from the CSI data. Instead of requiring every layer to learn an entirely new representation, residual connections allow earlier information to pass directly to later stages while additional layers learn more detailed transformations. A Robust Domain This is useful for gesture recognition because the original CSI signal may contain small but important variations that should not disappear as the information moves through deeper layers of the network.

The strongest result is obtained when all the proposed components are combined. The complete MSA-ResNet reaches 86% unseen-domain accuracy, compared with 72% for the baseline ResNet. The overall improvement is therefore 14 percentage points. The intermediate results 79% with multi-scale convolution and 82% after incorporating channel/spatial attention also show a gradual increase in performance as the individual components are introduced. A Robust Domain These findings suggest that the final improvement comes from the combined contribution of multi-scale processing, residual learning, and attention-based feature refinement.

5.5 Training and Validation Behaviour

The training and validation results show a smooth and stable learning process. The cross-entropy loss decreases steadily as training progresses, and the validation-loss curve follows the training-loss curve closely. A Robust Domain There is no large separation between the two curves toward the later epochs. This indicates that the model learns the source-domain gesture patterns without showing severe overfitting in the reported training experiment. A similar trend can be observed in the accuracy curves. Training and validation accuracy improve rapidly during the early epochs and then gradually settle into a stable region. Training accuracy reaches the upper-90% range, while validation accuracy remains around the mid 90% range. A Robust Domain The relatively small difference between these curves indicates consistent learning on the source-domain data. More importantly, the separate evaluation on Domain 3 shows whether this stable training behaviour carries over to an environment that the model has not previously encountered.

5.6 Unseen-Domain Gesture Classification

The confusion matrix provides a class-wise view of recognition performance for the five gesture categories, G0 to G4, in the unseen environment. The displayed matrix has a strong diagonal pattern, showing that the gesture samples represented in the figure are assigned to their corresponding classes. A Robust Domain This indicates good separation among the five gesture categories and supports the ability of the learned features to distinguish different movements after the sensing environment changes.

The confusion matrix is especially useful because overall accuracy alone does not show how individual gesture classes behave. A model could achieve good overall accuracy while performing poorly on one or more specific gestures. In the reported matrix, the clear diagonal structure suggests consistent class separation for the displayed evaluation samples. For the final manuscript, it would be useful to clarify how this matrix relates to the reported 86% overall Domain 3 accuracy, since the plotted matrix itself shows only diagonal entries. A Robust Domain Such clarification would make the presentation of the analysis results more consistent.

5.7 Feature Distribution and t-SNE Visualization

The t-SNE visualization provides another way to examine the features learned by the proposed framework. In the unseen environment, the five gesture categories appear as compact and clearly separated groups in the two-dimensional feature space. A Robust Domain Samples belonging to the same gesture remain close together, while samples representing different gestures occupy separate regions. This pattern is consistent with the recognition results and shows that the extracted features preserve useful differences among the gesture classes.

The t-SNE plot should mainly be considered as visual support for the numerical findings rather than as an independent measure of performance. Its value lies in showing how the learned gesture representations are organized after the model is applied to a new environment. When considered together with the 86% unseen-domain accuracy and the ablation results, the separated clusters provide additional evidence that the proposed framework retains meaningful gesture information despite changes in the surrounding wireless conditions.

5.8 Practical Importance of the Results

WiFi-based gesture recognition has practical value because it can sense human movement without requiring cameras or wearable devices. This makes it suitable for applications such as smart homes, assisted-living environments, healthcare-related monitoring, and contactless human–computer interaction, particularly where

visual privacy is important. A Robust Domain Since WiFi infrastructure is already available in many indoor environments, CSI-based sensing can also reduce the need for separate sensing hardware.

The cross-environment results make this possibility more relevant. A system that requires complete retraining whenever it is moved to another room would be difficult to use on a large scale. In the present study, the framework maintains 86% accuracy in the unseen environment without target-domain fine-tuning. A Robust Domain Although broader testing is still necessary, this result suggests that the proposed approach has the potential to reduce the amount of room-specific calibration needed for practical WiFi gesture-recognition systems.

5.9 Limitations of the Present Study

The results are encouraging, but they should be considered within the scope of the current analysis setup. The study evaluates five gesture classes using two source environments and one unseen target environment. A Robust Domain These experiments demonstrate the behaviour of the framework under the environmental changes represented in the available data, but additional environments would be needed to establish how consistently the approach performs across a wider variety of indoor conditions.

There is also scope to extend the evaluation across different users, WiFi devices, hardware configurations, and more complex gesture sets. The study identifies several useful future directions, including a larger number of environments, more fine-grained gesture categories, subject-independent evaluation, real-time edge implementation, model compression, and cross-device validation using different WiFi hardware configurations, including WiFi 6/7. A Robust Domain These extensions would provide a more complete understanding of the framework's performance under realistic deployment conditions.

5.10 Overall Discussion

Taken together, the analysis findings show that the proposed framework provides a clear improvement over the baseline configuration for cross-environment WiFi CSI gesture recognition. The complete MSA-ResNet achieves 94% accuracy in Domain 1, 91% in Domain 2, and 86% in the unseen Domain 3. The ablation experiment further shows a steady improvement from 72% for the baseline ResNet to 79% with multi-scale convolution, 82% with channel/spatial attention, and finally 86% with the complete framework. A Robust Domain

Table 5. Validation-Based Comparison of Existing and Proposed Systems

Parameters	**Existing System**	**Proposed System**
Unseen-Domain Accuracy (%)	72% – Baseline ResNet	86% – Proposed MSA-ResNet
Multi-Scale Feature Validation (%)	72% – without multi-scale module	79% – with multi-scale convolution
Attention-Based Validation (%)	79% – before channel/spatial attention	82% – with channel/spatial attention
Complete Architecture Accuracy (%)	72% – conventional baseline	86% – complete MSA-ResNet
Absolute Accuracy Improvement (percentage points)	0 – reference baseline	+14
Relative Accuracy Improvement (%)	0% – reference baseline	19.44% over baseline
Unseen-Domain Accuracy Retention vs. Best Source Domain (%)	Not reported for literature baseline	91.49% (86/94 × 100)
Source-to-Unseen Performance Drop (percentage points)	Not reported for literature baseline	8 points (94% → 86%)

The analysis data show the ablation progression 72% → 79% → 82% → 86%, demonstrating the incremental contribution of multi-scale processing and channel–spatial attention. The complete framework also achieves 94% in Source Domain 1, 91% in Source Domain 2, and 86% in the unseen Domain 3.

Overall, the results suggest that successful recognition across changing indoor environments depends on capturing CSI patterns at different scales while preserving useful information and giving greater importance to the most relevant features. This is particularly important because CSI patterns are affected by changes in the surroundings, multipath propagation, and signal conditions. A Robust Domain Within the analysis conditions considered in this study, the combination of multi-scale convolution, residual learning, and channel spatial attention provides a practical approach for maintaining gesture-recognition performance when the WiFi sensing environment changes.

6. Conclusion and Future Work

6.1 Conclusion

This study developed a Domain-Generalized Multi-Scale Channel–Spatial Attention Residual Network (MSA-ResNet) for recognizing human gestures from WiFi Channel State Information (CSI) across different indoor

environments. One of the main difficulties in WiFi-based gesture recognition is that CSI measurements are affected not only by human movement but also by room arrangement, surrounding objects, signal reflections, and multipath propagation. As a result, a model that performs well in one environment may not provide the same level of recognition when it is used in another. To address this issue, the proposed framework brings together multi-scale convolution, residual learning, and channel–spatial attention. Multi-scale processing captures both small and wider variations in the CSI signals, residual connections help preserve useful information during deeper feature extraction, and the attention mechanism highlights the feature channels and regions that contribute most to distinguishing different gestures. A Robust Domain

The analysis results show that this combination improves recognition when the sensing environment changes. The proposed framework achieved 94% accuracy in Domain 1 and 91% in Domain 2, while maintaining 86% accuracy in the previously unseen Domain 3 without any additional fine-tuning. A Robust Domain The training and validation results also showed stable learning behaviour, while the t-SNE representation displayed clear separation among the five gesture classes in the unseen environment. The contribution of each part of the framework can be seen more clearly from the ablation study. The baseline ResNet produced 72% accuracy, which increased to 79% after introducing multi-scale convolution and to 82% after incorporating channel and spatial attention. The complete MSA-ResNet achieved 86%, giving an overall improvement of 14 percentage points compared with the baseline. A Robust Domain These findings show that the proposed framework can retain useful gesture-related information even when the surrounding WiFi conditions are different from those used during training.

6.2 Future Work

The present study provides encouraging results, but there is still considerable scope for further investigation. The current evaluation considers five gesture classes, two source environments, and one unseen target environment. Future studies can extend this work by collecting CSI data from a larger number of rooms and more varied indoor settings. Testing with additional users and a wider range of simple and fine-grained gestures would also provide a clearer understanding of how well the method performs under realistic conditions. Subject-independent experiments would be particularly useful, where the system is trained using gestures from one group of people and evaluated using different participants. Further testing with different access points, network interface cards, transmitter–receiver positions, and wireless technologies such as WiFi 6 and WiFi 7 would also help establish the consistency of the method across different hardware configurations. A Robust Domain

Another useful direction is to develop a lighter and faster version of the framework for real-time operation on edge or embedded devices. Techniques such as model pruning, quantization, and network compression could reduce computational requirements while preserving recognition performance. Future experiments can also consider continuous CSI streams, automatic gesture segmentation, background human movement, multiple users, wireless interference, and changes in transmitter or receiver position. Long-term experiments conducted over different days and changing room conditions would provide further evidence of the stability of the learned gesture features. These developments could make WiFi-based gesture recognition more suitable for practical applications such as smart-home control, assisted-living systems, contactless human–computer interaction, and privacy-conscious human activity monitoring, where reliable operation across different environments is particularly important.

References

[1] I. Guarino, D. Carra, M. Cominelli, F. Gringoli, and R. Lo Cigno, “A survey on CSI-based Wi-Fi sensing datasets and models with a focus on reproducibility,” Computer Communications, vol. 249, Art. no. 108431, 2026. doi: 10.1016/j.comcom.2026.108431.

[2] Z. He et al., “Neural approaches to Wi-Fi CSI-based human activity recognition: Architectures and generalization,” IEEE Transactions on Cognitive Communications and Networking, vol. 12, pp. 7186–7200, 2026. doi: 10.1109/TCCN.2026.3683215.

[3] S. Sai et al., “Machine learning techniques for Wi-Fi CSI-based recognition and sensing: A comprehensive review,” IEEE Internet of Things Journal, vol. 13, no. 10, pp. 20190–20214, 2026. doi: 10.1109/JIOT.2026.3657341.

[4] J. A. Armenta-Garcia, F. F. Gonzalez-Navarro, and J. Caro-Gutierrez, “Wireless sensing applications with Wi-Fi Channel State Information, preprocessing techniques, and detection algorithms: A survey,” Computer Communications, vol. 224, pp. 254–274, 2024. doi: 10.1016/j.comcom.2024.06.011.

[5] F. Miao et al., “Wi-Fi sensing techniques for human activity recognition: Brief survey, potential challenges, and research directions,” ACM Computing Surveys, vol. 57, no. 5, Art. no. 107, 2025. doi: 10.1145/3705893.

[6] Z. Wang et al., “Review of few-shot learning application in CSI human sensing,” Artificial Intelligence Review, vol. 57, Art. no. 195, 2024. doi: 10.1007/s10462-024-10812-4.

[7] Y. Zhang et al., “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, 2022. doi: 10.1109/TPAMI.2021.3105387.

[8] J. Zhang et al., “CrossSense: Towards cross-site and large-scale WiFi sensing,” in Proc. 24th Annual International Conference on Mobile Computing and Networking (MobiCom), 2018, pp. 305–320. doi: 10.1145/3241539.3241570.

[9] B. Li, J. Chen, X. Yu, Z. Yang, and J. Zhang, “Cross-domain gesture recognition via WiFi signals with deep learning,” Ad Hoc Networks, vol. 166, Art. no. 103654, 2025. doi: 10.1016/j.adhoc.2024.103654.

[10] X. Zhang et al., “Beyond physical labels: Redefining domains for robust WiFi-based gesture recognition,” Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 10, no. 1, Art. no. 24, 2026. doi: 10.1145/3790114.

[11] “CrossFi: A cross-domain Wi-Fi sensing framework based on Siamese network,” IEEE Internet of Things Journal, vol. 12, no. 12, pp. 20138–20155, 2025. doi: 10.1109/JIOT.2025.3542850.

[12] X. Chen, “Cross-domain human activity recognition using reconstructed Wi-Fi signal,” Physical Communication, vol. 71, Art. no. 102651, 2025. doi: 10.1016/j.phycom.2025.102651.

[13] “Trans-MAML: Meta-learning enhanced Transformers for human activity recognition in WiFi CSI sensing,” IEEE Access, vol. 14, pp. 78740–78756, 2026. doi: 10.1109/ACCESS.2026.3696430.

[14] F. Wang et al., “A survey on Wi-Fi sensing generalizability: Taxonomy, techniques, datasets, and future research prospects,” arXiv preprint, arXiv:2503.08008, 2025. [Online]. Available: arxiv.org/abs/2503.08008.

[15] “Device free human gesture recognition using Wi-Fi CSI: A survey,” Engineering Applications of Artificial Intelligence, 2020. [Online]. Available: ScienceDirect article.

[16] “WiGest: A ubiquitous WiFi-based gesture recognition system,” in Proc. IEEE INFOCOM, 2015. doi: 10.1109/INFOCOM.2015.7218525.

[17] “WiFine: Real-time gesture recognition using Wi-Fi with edge intelligence,” ACM Transactions on Sensor Networks, 2022. doi: 10.1145/3532094.

[18] “Human activity recognition from Wi-Fi CSI data using principal component-based wavelet CNN,” Digital Signal Processing, 2023. Publisher search.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[20] C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[21] F. Li, C. Weng, and Y. Liang, “SE-WiGR: A WiFi gesture recognition approach incorporating the squeeze–excitation mechanism and VGG16,” Applied Sciences, vol. 15, no. 11, Art. no. 6346, 2025. doi: 10.3390/app15116346.

[22] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. European Conference on Computer Vision (ECCV), 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

[23] H. Kang, D. Kim, K.-A. Toh, et al., “Human activity recognition through augmented WiFi CSI signals by lightweight Attention-GRU,” Sensors, vol. 25, no. 5, Art. no. 1547, 2025. doi: 10.3390/s25051547.

[24] “WiFi-based multiuser identity, location, and activity recognition using InceptionTime-Attention networks,” IEEE Sensors Journal, vol. 25, no. 7, pp. 12389–12398, 2025. doi: 10.1109/JSEN.2025.3540882.

[25] D. Varga, “Decision fusion-based deep learning for Channel State Information channel-aware human action recognition,” Sensors, vol. 25, no. 4, Art. no. 1061, 2025. doi: 10.3390/s25041061.

[26] “Enhancing cross-domain robustness in Wi-Fi-based human activity recognition via attention-driven deep learning,” in Proc. IEEE Vehicular Technology Conference (VTC-Fall), 2025. doi: 10.1109/VTC2025-Fall65116.2025.11310278.

[27] C. Szegedy et al., “Rethinking the Inception architecture for computer vision,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[28] G. Chi et al., “RF-Diffusion: Radio signal generation via time-frequency diffusion,” in Proc. ACM MobiCom, 2024. Publication listing.

[29] Z. Yang, Y. Zhang, K. Qian, and C. Wu, “SLNet: A spectrogram learning neural network for deep wireless sensing,” in Proc. USENIX NSDI, 2023. Publication listing.

[30] Y. Zhang et al., “Zero-effort cross-domain gesture recognition with Wi-Fi,” in Proc. 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys), 2019, pp. 313–325. doi: 10.1145/3307334.3326081.