2025 IEEE 26th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM) | 979-8-3315-3832-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/WoWMoM65615.2025.00030

MAGIC: Meta-Learning Adaptive Gesture Recognition with mmWave MIMO CSI

Khandaker Foysal Haque*, K M Rumman*, Arman Elyasi*, Francesca Meneghello‡, Francesco Restuccia*
* Institute for the Wireless Internet of Things, Northeastern University, United States
‡ Department of Information Engineering, University of Padova, Italy

***Abstract*—In this paper, we present MAGIC, a novel approach to gesture recognition utilizing mmWave multiple-input multiple-output (MIMO) Channel State Information (CSI). Unlike existing mmWave gesture recognition methods that often rely on radar signals, MAGIC leverages CSI extracted from mmWave MIMO integrated sensing and communication (ISAC) systems. While advanced radar systems, such as those operating in frequency-modulated continuous wave (FMCW) mode, can achieve high frequency and spatial resolution, they typically require dedicated sensing infrastructure, which increases system complexity. In contrast, MAGIC utilizes high-granular CSI from orthogonal frequency-division multiplexing (OFDM) systems, enabling fine spatial, temporal, and frequency-domain information for robust gesture recognition. This eliminates the need for dedicated radar transceivers, simplifying the system and reducing transmission overhead. MAGIC employs a learning-based architecture, integrating a temporal convolutional network (TCN) to classify gestures by capturing long-range temporal dependencies. To address the critical challenge of domain adaptation in gesture recognition, we propose adaptive temporal embedding network (ATEN), a meta-learning framework that combines the temporal modeling capabilities of TCN with task-specific adaptation mechanisms. We evaluate MAGIC through a comprehensive data collection campaign involving two subjects performing 10 micro gestures across three different environments, with synchronized video streams providing the ground truth. The proposed system achieves a baseline accuracy of 99.24% using TCN. The system continues to perform well – achieving up to 98.82% accuracy – when adapting to new domains using ATEN, outperforming other state-of-the-art domain adaptation methods by 14% on average.**

I. Introduction

Next-generation wireless systems must support intelligent, autonomous, and immersive applications, driving the need for significantly enhanced network capabilities [1]. Emerging applications such as augmented reality (AR) and virtual reality (VR) demand data rates exceeding 40 Gbps to enable seamless user experiences and efficient neural network processing [2]. In contrast, current wireless technologies, including Wi-Fi, typically support data rates up to 1.2 Gbps [3], [4]. Thus, to meet these stringent requirements necessitates moving beyond traditional sub-6 GHz bands and leverage the larger bands available in the millimeter-wave (mmWave) spectrum.

The mmWave spectrum offers a substantial bandwidth advantage, enabling high-data-rate communications. Moreover, mmWave radio transmissions offer precise wireless sensing capabilities beyond communication due to super high-resolution channel information. The intuition behind wireless sensing is that any object in the physical world acts as an obstacle to the propagation of radio signals that undergo reflections, diffractions, and scattering, making the signals collected at the receiver differ from the transmitted ones.

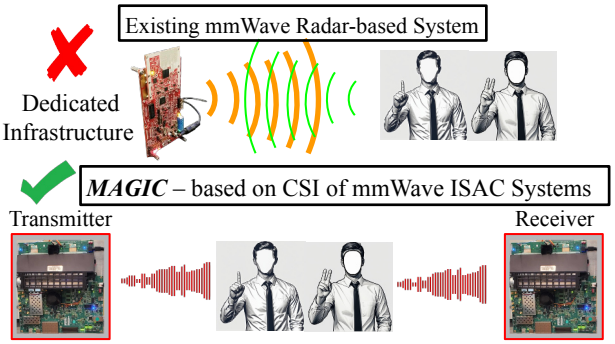


Fig. 1: MAGIC vs existing radar-based mmWave gesture recognition systems.

Wireless sensing aims to detect changes in radio signals and associate them with the way any object or human stays/moves in the environment, thus allowing device-free monitoring solutions. This dual capability positions mmWave systems as a promising foundation for applications requiring fine-grained sensing, such as human activity classification [5], [6], gesture recognition [7], [8], intrusion detection [9], and patient monitoring [10].

While radar-based systems are widely used for gesture recognition in the mmWave band, they require dedicated sensing infrastructure, which adds complexity to the system. Advanced radar systems, such as those operating in frequency-modulated continuous wave (FMCW) mode, achieve high spatial and frequency resolution, but their reliance on hardware designed solely for sensing tasks can limit their integration with communication systems.

In contrast, mmWave Channel State Information (CSI) offers a transformative approach by leveraging the spatial diversity of multiple-input multiple-output (MIMO) systems (e.g., 8x8 antennas) and the high-frequency granularity of orthogonal frequency-division multiplexing (OFDM) subchannels across a wide bandwidth. This provides rich spatial, temporal, and frequency-domain data, enabling detailed characterization of gesture patterns. Furthermore, CSI-based sensing seamlessly integrates into integrated sensing and communication (ISAC) systems, eliminating the need for dedicated radar hardware, simplifying system design, and reducing overall complexity and sensing overhead.

In stark contrast to the traditional radar-based systems, we propose MAGIC, a domain adaptive gesture recognition system that integrates sensing functionalities directly into communication systems using mmWave MIMO CSI. As depicted in Figure 1, by leveraging rich spatial, temporal, and

2770-0542/25/$31.00 ©2025 IEEE
DOI 10.1109/WoWMoM65615.2025.00030

frequency-domain data offered by CSI, MAGIC significantly improves sensing precision while reducing system complexity and deployment overhead.

The core of MAGIC is a learning-based architecture that incorporates a temporal convolutional network (TCN) to effectively capture long-range temporal dependencies from CSI for gesture classification. Moreover, domain generalization being one of the crucial challenges in wireless sensing, we also propose adaptive temporal embedding network (ATEN) – a novel meta-learning approach that combines the temporal features extracted from TCN with the task-specific meta-learning framework for domain adaptation with only a few data samples from the new environment or subject.

Summary of Contributions:

• We propose MAGIC, a novel gesture recognition system that leverages CSI to seamlessly integrate sensing functionalities into mmWave MIMO ISAC systems. By utilizing the high-resolution wideband CSI estimated during communication, MAGIC provides fine spatial, temporal, and frequency-domain granularity for robust gesture recognition. Unlike radar-based systems, which require dedicated sensing hardware, MAGIC eliminates this need, reducing system complexity, and minimizing sensing overhead while maintaining high performance in diverse environments;

• We develop ATEN, an innovative domain generalization algorithm that integrates TCN for efficient long-range temporal feature extraction with a task-specific meta-learning framework to adaptively align query features with domain-specific contexts, enabling adaptation to unseen environments and subjects. To enhance robustness, we introduce the domain-adaptive preprocessing pipeline (DAPP) for systematic mitigation of domain-specific artifacts, incorporating path loss compensation, spectral alignment, and entropy-guided filtering. This preprocessing pipeline transforms raw mmWave CSI into a domain-invariant representation, ensuring reliable gesture recognition across diverse environments;

• We evaluate MAGIC through extensive data collection campaigns involving three environments and ten micro gestures, achieving up to 99.24% accuracy for gesture recognition with baseline TCN and maintaining 98.82% accuracy with ATEN under domain-adaptive conditions. To foster reproducibility, we open-source all the datasets, and the codes at **https://github.com/kfoysalhaque/MAGIC**.

II. MAGIC System Workflow

The channel estimates, expressed as channel frequency responses (CFRs)[1], form the basis for the MAGIC's functionality and are obtained via a MIMO procedure known as *channel sounding*, illustrated in **Step 1** of Figure 2. During channel sounding, the beamformer broadcasts null data packets (NDPs), allowing the channel properties between every pair of m transmitter antennas and n receiver antennas to be measured across the entire operational bandwidth. The beamformer r estimates the CFR, $\mathbf{H}_r^{m,n}$ on each OFDM subchannel $k \in \{1, \ldots, K\}$ by comparing the received signal $\mathbf{Y}_r^{m,n}$ with the known transmitted signal $\mathbf{X}_r^{m,n}$. The CFR is calculated as $\mathbf{H}_r^{m,n} = \mathbf{Y}_r^{m,n} / \mathbf{X}_r^{m,n}$, resulting in a matrix of dimensions $M \times N$, where M and N represent the number of transmit and receive antenna, respectively. Indicating with $h_{k,s}^{m,n}$ the CFR of the S-th sample at the K-th subchannel from transmit antenna m to receive antenna n, the CFR matrix can be expressed as:

$$H_r^{m,n} = \begin{bmatrix} h_{r,1,1}^{m,n} & \cdots & h_{r,1,s}^{m,n} & \cdots & h_{r,1,K}^{m,n} \\ \vdots & & \vdots & & \vdots \\ h_{r,s,1}^{m,n} & \cdots & h_{r,s,k}^{m,n} & \cdots & h_{r,s,K}^{m,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ h_{r,S,1}^{m,n} & \cdots & h_{r,S,k}^{m,n} & \cdots & h_{r,S,K}^{m,n} \end{bmatrix}. \quad (1)$$

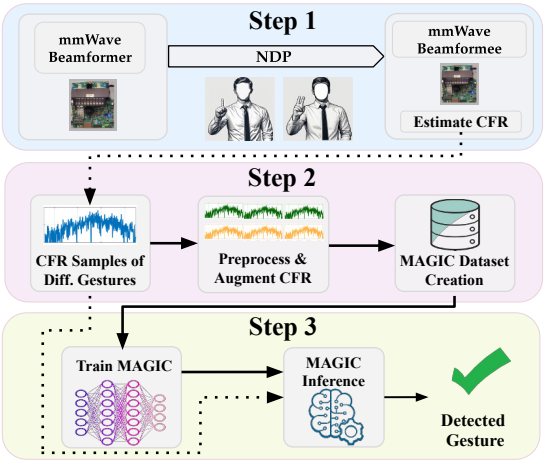


Fig. 2: The overview of MAGIC gesture recognition system.

We leverage MIMO transmissions to estimate and extract the CFR during the execution of different gestures. The CFR is then forwarded through a rigorous preprocessing pipeline, as depicted in **Step 2** of Figure 2. This preprocessing stage leverages the DAPP framework, detailed in Section II-A, to transform the raw CFR data into a domain-independent representation. By mitigating environment-specific artifacts and highlighting gesture-relevant features, DAPP ensures the data is robust for model training. Subsequently, the processed data undergoes augmentation, as described in Section II-B, to further enhance the generalization capabilities of the learning architecture. The combination of preprocessed and augmented CFR data forms a comprehensive and diverse dataset, ready for training the MAGIC learning architecture. Hence, the MAGIC learning architecture ATEN is trained using this dataset (**Step 3** of Figure 2). ATEN integrates the temporal modeling capabilities of TCN with a task-adaptive meta-learning framework, enabling the system to generalize effectively across domains. This dual approach – elaborated in Section II-C – ensures robust gesture recognition performance in diverse and dynamic environments.

[1]We use the terms CSI and CFR interchangeably for the rest of the paper.

A. Domain-adaptive Preprocessing Pipeline (DAPP)

The domain-adaptive preprocessing pipeline (DAPP) is designed to systematically address the inherent variability and domain-specific artifacts present in mmWave CFR data. Environmental factors such as spatial layout, device placement, and dynamic movements introduce significant challenges that can impair the generalization performance of gesture recognition models when deployed across diverse environments. DAPP helps mitigate these challenges by transforming raw CFR data into a domain-invariant and gesture-specific representation through a set of carefully designed processing steps as follows.

Path Loss Compensation. Variations in the propagation distance between the transmitter and receiver introduce substantial variability in CFR magnitudes due to path loss. Without correction, these variations embed domain-specific biases into the data, hindering model generalization. Path loss $P_L(d)$ is modeled as $P_L(d) \propto 10n\log_{10}(d) + \chi_\sigma$, where d is the propagation distance, n is the path loss exponent, and $\chi_\sigma \sim \mathcal{N}(0, \sigma^2)$ represents the shadow fading component with σ as the standard deviation. An attenuation factor $A_{PL}(d) = 1/d^n$ is derived to compensate for the path loss. Thus, the CFR matrix $\mathbf{H}_r^{m,n}$ is scaled using $\mathbf{C}_{\text{comp}}(f) = \mathbf{H}_r^{m,n} \cdot A_{PL}(d)$. This compensation standardizes the CFR magnitudes across different distances, allowing gesture-specific features to be more prominent while attenuating domain-induced biases.

Frequency Normalization. Next, to ensure uniform scaling, frequency normalization is applied to the compensated CFR:

$$\mathbf{C}_{\text{norm}}(f) = \frac{\mathbf{C}_{\text{comp}}(f)}{\|\mathbf{C}_{\text{comp}}\|}, \tag{2}$$

where $\|\mathbf{C}_{\text{comp}}\|$ is the Frobenius norm [11] expressed as

$$\|\mathbf{C}_{\text{comp}}\| = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K}|\mathbf{C}_{ijk}|^2}. \tag{3}$$

This normalization removes amplitude inconsistencies while preserving the relative variations critical for distinguishing different gestures.

Entropy-Guided Filtering. The stability of the CFR over the sub-channels varies due to environmental factors like multi-path and motion dynamics, affecting gesture recognition accuracy. Entropy is used to quantify this variability, identifying stable sub-channels with consistent energy distribution. For sub-channel k, entropy $E(k)$ is computed as:

$$E(k) = -\sum_{i=1}^{M\times N} p_i \log p_i, \quad p_i = \frac{|\mathbf{C}_{\text{norm }k,i}|^2}{\sum_{j=1}^{M\times N}|\mathbf{C}_{\text{norm }k,j}|^2}, \tag{4}$$

where p_i is the normalized power of the i-th transmit-receive antenna pair in the normalized CFR, $\mathbf{C}_{\text{norm}}$. Low $E(k)$ indicates stable subchannels likely to capture gesture-relevant features, while high $E(k)$ suggests noisy subchannels. Weights $w(k) = \exp(-E(k))$ are applied, producing the filtered CFR as $\mathbf{C}_{\text{filtered}} = \mathbf{C}_{\text{norm}} \cdot \text{diag}(w(k))$. This emphasizes stable, gesture-relevant subchannels while suppressing noise and enhancing the signal-to-noise ratio. Figure 3 presents examples of both the raw and preprocessed CFR for the received streams 1-4 of the first transmit antenna.

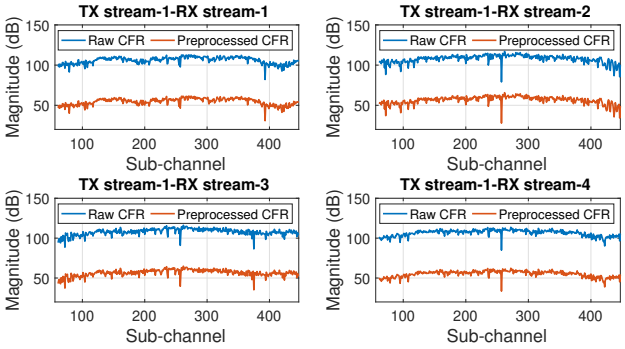


Fig. 3: An example of a CFR sample before and after preprocessing through the DAPP framework. The four plots depict the CFR for received streams 1-4 and the first transmit antenna.

B. Data Augmentation

To further enhance the generalization capabilities of the gesture recognition model, data augmentation is integrated into the preprocessing pipeline. This step simulates a variety of environmental conditions and channel dynamics that the learning model may encounter in real-world scenarios, thus improving its robustness to domain shifts.

Noise Addition. To emulate additive white Gaussian noise (AWGN) commonly present in wireless channels, Gaussian noise is added to the filtered CFR matrix as follows:

$$C_{\text{aug1}} = C_{\text{filtered}} + N, \quad N \sim \mathcal{CN}(0, \sigma^2), \tag{5}$$

where N represents complex Gaussian noise with variance σ^2. This augmentation helps the model become resilient to noise-induced fluctuations.

Scaling. Variations in transmit power or propagation distances are simulated by scaling the CFR magnitudes with a random factor:

$$C_{\text{aug2}} = \alpha \cdot C_{\text{filtered}}, \quad \alpha \in [\alpha_{\min}, \alpha_{\max}]. \tag{6}$$

The scaling factor α is sampled from a uniform distribution within the range $[\alpha_{\min}, \alpha_{\max}]$, specifically $\alpha \in [0.7, 1.3]$ with an interval of 0.1 for data augmentation. This range reflects realistic variations in signal strength, accounting for path loss effects due to varying distances between the transmitter and the receiver in typical indoor environments. By simulating such variations, this augmentation helps the model to generalize across different signal strengths without overfitting to specific power levels.

Phase Perturbation. To replicate the phase distortions caused by multipath effects due to the variations in the environment dynamics, random phase shifts are applied:

$$C_{\text{aug3}}(f) = C_{\text{filtered}}(f) \cdot e^{j\phi}, \quad \phi \sim U(-\phi_{\max}, \phi_{\max}), \tag{7}$$

where ϕ is a random phase shift sampled from a uniform distribution. This augmentation makes the model robust to phase variations that occur due to the variation in the environment dynamics.

By generating diverse and realistic training samples through these augmentation techniques, the model is better equipped to handle variations encountered in real-world deployments.

C. Adaptive Temporal Embedding Network (ATEN)

Gesture recognition using mmWave CFR data is challenging due to the dynamic nature of temporal signals and variations across domains, with traditional machine learning models often failing to generalize across environments. To address these issues, we propose the ATEN framework, which integrates four main components: (i) *feature encoding with TCNs*, (ii) *task embedding*, (iii) *query-specific adaptation and classification*, and (iv) *meta-learning-driven adaptation*. The first component – feature encoding with TCNs – extracts gesture-relevant features from high-dimensional CFR data by modeling both short and long-term temporal dependencies through causal and dilated convolutions. This ensures causality, efficiently expands the receptive field, and minimizes noise and domain-specific artifacts. The second component – task embedding – generates a task-specific representation by aggregating feature embeddings of data samples with their labels, creating a global descriptor that reduces domain-specific influences while bridging task-specific and global characteristics. The third component – query-specific adaptation and classification – refines query sample features by integrating them with task embedding using a learnable transformation. This process dynamically aligns query features with task-specific attributes, enabling robust recognition even in unseen environments. The refined features are passed through a task-adaptive classifier that adjusts decision boundaries based on the task embedding. Finally, meta-learning-driven adaptation optimizes the entire framework by training across a distribution of tasks. This objective ensures the framework generalizes effectively by capturing task-agnostic patterns and adapting to task-specific details with minimal fine-tuning, making it highly resilient to domain shifts.

1) Feature Encoding with TCNs: This section describes the input representation and the architectural details of the TCN used in our framework, as illustrated in Figure 4.

Input Representation. The raw CFR data from the mmWave system is inherently multi-dimensional, comprising information across time steps, subchannels, and multiple streams. This data can be represented as a tensor $\mathbf{X} \in \mathbb{R}^{T \times K \times N_s}$, where T is the number of CFR samples in the temporal sequence, reflecting the temporal resolution of the CFR signal, K is the number of subchannels, capturing the frequency domain granularity, and $N_s = M \times N$ is the total number of streams where M and N are the numbers of transmit and receive antennas, respectively. To make the data suitable for temporal processing using a TCN, the tensor

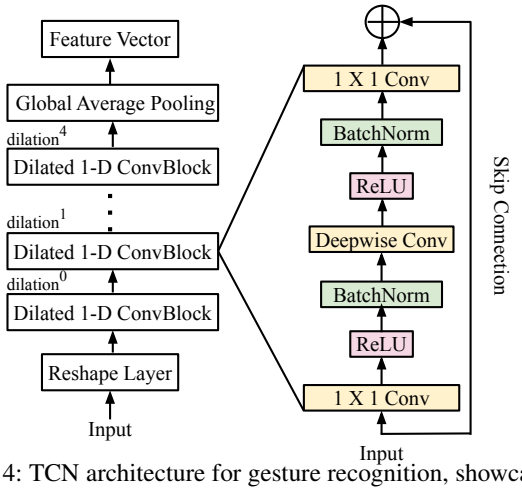


Fig. 4: TCN architecture for gesture recognition, showcasing dilated convolutions, deepwise separable convolutions, and residual connections.

is reshaped into a 2D matrix, $\mathbf{X}_{\text{flat}} \in \mathbb{R}^{T \times (N_k \cdot N_s)}$. This transformation preserves the temporal structure (T) while flattening the spatial and frequency dimensions, enabling the TCN to process temporal dependencies while embedding the combined spatial-frequency information in each time step.

TCN Architecture. The proposed TCN architecture, illustrated in Figure 4, consists of five dilated 1-D convolutional layers with dilation rates $\{0, 1, 2, 3, 4\}$, leveraging both causal and dilated convolutions to efficiently capture short and long-term temporal dependencies while preserving the temporal order of the input data. Each convolutional layer uses a kernel size of 3, balancing local feature extraction and computational efficiency. Causal convolutions ensure that the output at any time step depends only on the current and past inputs, which is crucial for time-series modeling. Residual connections are integrated to stabilize training and prevent vanishing gradients, while deepwise separable convolutions in the skip paths reduce parameter overhead. The temporal feature map generated after the final layer undergoes global average pooling to produce a compact feature vector for downstream tasks. This design ensures effective temporal modeling, computational efficiency, and robust gesture recognition from multi-dimensional mmWave CFR.

Causal and Dilated Convolutions: The mathematical definition of causal convolutions is defined as:

$$\mathbf{Z}^{(l)}[t] = \sum_{i=0}^{q-1} \mathbf{W}^{(l)}[i] \cdot \mathbf{Z}^{(l-1)}[t-i], \tag{8}$$

where $\mathbf{W}^{(l)}[i]$ represents the convolutional filters, and q is the kernel size. To capture long-term dependencies, dilated convolutions expand the receptive field by skipping input values, determined by the dilation rate r. This operation is expressed as:

$$\mathbf{Z}^{(l)}[t] = \sum_{i=0}^{q-1} \mathbf{W}^{(l)}[i] \cdot \mathbf{Z}^{(l-1)}[t-r \cdot i], \tag{9}$$

with r increasing exponentially across layers ($r = 2^l$). Together, causal and dilated convolutions enable the TCN to capture both local and global temporal patterns while maintaining the sequential integrity of the input.

Residual Connections: Residual connections, illustrated in Figure 4, are incorporated to stabilize training and mitigate vanishing gradient issues. Specifically, for a given layer l, the output is computed as $\mathbf{Z}^{(l)} = \text{ReLU}\left(\mathbf{Z}^{(l)} + \mathbf{Z}^{(l-1)}\right)$. These connections enable the network to learn perturbations to the input rather than entirely new transformations, leading to faster convergence and improved generalization.

Global Pooling for Feature Representation: After processing through L layers, the TCN outputs a temporal feature map, denoted by $\mathbf{Z}^{(L)}$, where $\mathbf{Z}^{(L)} \in \mathbb{R}^{T\times p}$ represents a matrix with T CFR time step samples and p as the feature dimension per time step. To generate a compact representation for each input sample x_i, global average pooling is applied across the temporal dimension, resulting in a feature vector $z_i = \text{GlobalAveragePooling}(\mathbf{Z}^{(L)})$, where $z_i \in \mathbb{R}^p$ represents the final feature vector, encapsulating the temporal, spatial, and frequency characteristics of the input sample x_i. This feature vector z_i is then utilized for downstream tasks such as task embedding and classification.

2) Task Embedding: The concept of task embedding serves as the backbone of meta-learning by distilling the information from the support set into a compact representation. This representation encapsulates the distinctive characteristics of a given task while preserving features that are invariant across domains. By bridging task-specific nuances with domain-independent traits, task embeddings lay the foundation for robust model generalization across diverse tasks and environments.

Let the support set for a task $\mathcal{T}$ be defined as $\mathcal{S}_i = \{(x_i, y_i)\}_{i=1}^n$, where x_i and y_i denote the input and corresponding label of the i-th sample. The task embedding $z_\mathcal{T} \in \mathbb{R}^q$ is designed to aggregate this information, producing a high-level representation of the task. Formally, the task embedding is computed as:

$$z_\mathcal{T} = g_\psi\left(\{z_i\}_{i=1}^n, \{y_i\}_{i=1}^n\right), \tag{10}$$

where g_ψ is a neural network parameterized by ψ, and z_i represents the latent embedding of the i-th input x_i, learned through the TCN feature extractor as presented in 4. The function g_ψ maps the support set to a single embedding vector $z_\mathcal{T}$ that summarizes the task efficiently. A straightforward yet effective mechanism for computing $z_\mathcal{T}$ involves the following aggregation strategy:

$$z_\mathcal{T} = \frac{1}{n}\sum_{i=1}^{n} \rho\left(z_i, y_i\right), \tag{11}$$

where ρ is a task-specific embedding function that combines the latent representation z_i of the input with its label y_i. This combination can be implemented through simple operations such as concatenation or element-wise multiplication, ensuring that label information is seamlessly integrated into the task embedding. The neural network g_ψ, is designed to refine the aggregated embeddings by learning the relationship between the features unique to a specific task and broader patterns or characteristics that are consistent across multiple tasks. By capturing the essence of the task, the embedding $z_\mathcal{T}$ not only facilitates rapid adaptation to new tasks but also ensures that the model remains resilient to domain shifts.

This approach for task embedding offers multiple advantages. First, it achieves a balance between task specificity and domain generality by encoding diverse task attributes without overfitting to any single task. Second, the aggregation mechanism is computationally efficient, making it scalable to large support sets. Finally, the inclusion of labels in the embedding process ensures that the representation aligns closely with the task objective, enhancing its discriminative power.

3) Query-Specific Adaptation and Classification: The query-specific adaptation mechanism tailors the representation of each query sample to the characteristics of the given task, leveraging the task embedding for improved alignment with task-specific nuances. This adaptation ensures that the model effectively incorporates contextual information from the support set when processing each query sample.

Given a query sample $x_i \in \mathcal{Q}_i$, where x_i is the input query sample and $\mathcal{Q}_i$ denotes the set of all query samples for a task, its feature embedding $z_i \in \mathbb{R}^p$ is adjusted using the task embedding $z_\mathcal{T} \in \mathbb{R}^q$. Here, p represents the dimension of the query embedding, corresponding to the number of features extracted for each query sample, while q represents the dimension of the task embedding, capturing the compressed information extracted from the task-specific support set. The adjusted or adapted representation $z_i^{\text{adapted}} \in \mathbb{R}^p$ is computed as $z_i^{\text{adapted}} = \eta\left(z_i, z_\mathcal{T}\right)$, where η denotes an adaptation function. A commonly used adaptation mechanism involves an affine transformation formulated as $z_i^{\text{adapted}} = Wz_i + Vz_\mathcal{T} + b$, where $W \in \mathbb{R}^{p\times p}$ is a learnable matrix that transforms the query embedding z_i, maintaining its dimensionality p. The matrix $V \in \mathbb{R}^{q\times p}$ maps the task embedding $z_\mathcal{T}$ from its q-dimensional space into the p-dimensional query embedding space, while $b \in \mathbb{R}^p$ is a learnable bias vector that ensures flexibility in the transformation. This affine transformation combines task-level and query-specific features, enabling z_i^{adapted} to reflect both global and task-specific contexts. Once adapted, the embeddings are passed to a classifier h_θ, where h_θ is a neural network parameterized by θ. The classifier operates on the adapted embedding z_i^{adapted} to produce gesture label predictions. The predicted label $\hat{y}_i$ for the query x_i is computed as

$$\hat{y}_i = \arg\max_{c\in\mathcal{Y}} h_\theta\left(z_i^{\text{adapted}}, c\right), \tag{12}$$

where $\mathcal{Y}$ is the set of all possible gesture labels, and h_θ outputs a probabilistic distribution over the labels in $\mathcal{Y}$. The final prediction is determined by the class c with the highest probability. The classifier in our framework consists of a single fully connected layer that maps the adapted embeddings z_i^{adapted} to the number of gesture classes $|\mathcal{Y}|$. This is

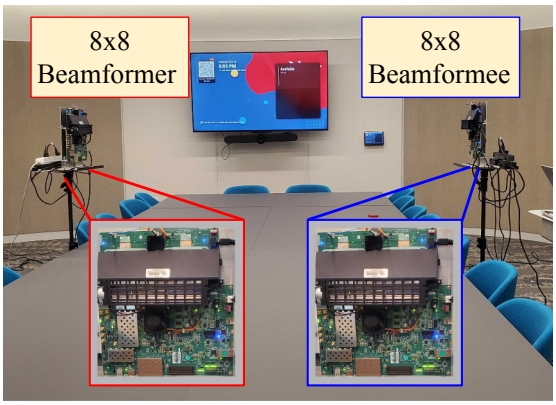


Fig. 5: m^{3}MIMO ISAC system for data collection in MAGIC.

followed by a softmax activation function, which outputs a probabilistic distribution over the gesture labels.

4) Meta-learning-driven adaptation: The meta-learning adaptation serves as the cornerstone for training models that generalize across diverse tasks. By leveraging task-specific losses, the model is iteratively refined to enhance its ability to adapt to new tasks while retaining generalization capabilities.

For each task $\mathcal{T}_i$, the query loss measures the discrepancy between the model's predictions and the true labels in the query set $\mathcal{Q}_i$ where the task-specific loss is defined by

$$\mathcal{L}_{\text{task}}(\mathcal{T}_i) = \frac{1}{m} \sum_{(x_i, y_i) \in \mathcal{Q}_i} \mathcal{L}_{\text{CE}} \left(h_\theta \left(z_i^{\text{adapted}} \right), y_i \right), \tag{13}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss function, h_θ is the classifier parameterized by θ, and m represents the number of query samples in the task. The use of cross-entropy loss ensures a probabilistic interpretation of the model's output, aligning the predictions $h_\theta(z_i^{\text{adapted}})$ with the ground truth labels y_i. The overall meta-loss aggregates the task-specific losses across all tasks in a meta-batch as

$$\mathcal{L}_{\text{meta}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{task}}(\mathcal{T}_i), \tag{14}$$

where N represents the number of tasks in the meta-batch. This aggregation ensures that the model learns a shared representation that generalizes across tasks, balancing task-specific performance and domain invariance. To optimize the meta-loss, the parameters (ϕ, ψ, θ) of the feature extractor, task embedding module, and classifier are updated using gradient descent, $\phi, \psi, \theta \leftarrow \phi, \psi, \theta - \eta \nabla_{(\phi,\psi,\theta)} \mathcal{L}_{\text{meta}}$, where η is the learning rate. This update step enables the model to iteratively refine its parameters by minimizing the meta-loss, ensuring that it captures both global patterns and task-specific nuances. The meta-learning adaptation plays a pivotal role in enabling the model to adapt to new tasks efficiently.

III. Experimental Setup and Data Collection

We evaluate MAGIC through extensive experimentation using the m^{3}MIMO [12] testbed—a fully digital 8×8 mmWave MIMO system. The testbed features two Zynq Ultra-Scale+

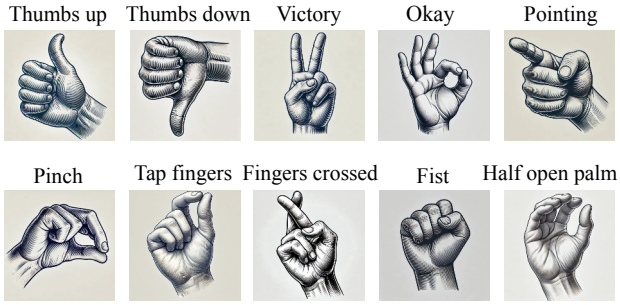


Fig. 6: The different micro-gestures in our evaluation: *thumbs up, thumbs down, victory, okay, pointing, pinch, tap two fingers, fingers crossed, fist, half open palm.*

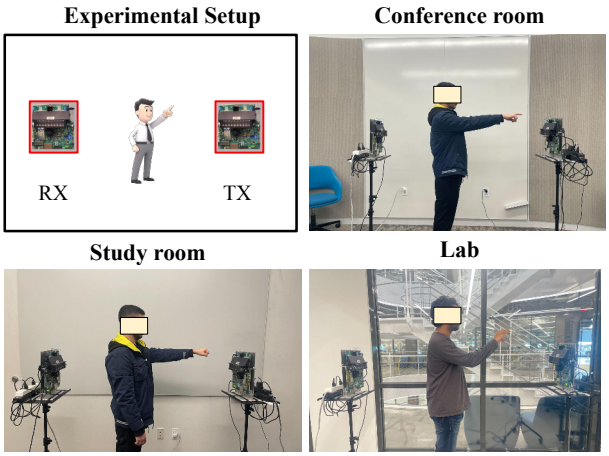


Fig. 7: Experimental setup and sample frames captured in three environments with synchronized video streams for ground truth.

RFSoC Software Defined Radios (SDRs) equipped with Pi-radio transceivers, functioning as beamformer and beamformee. Each SDR supports up to 8×8 MIMO, with eight independent transmitter and receiver RF chains and antennas. An example of the experimental setup is illustrated in Figure 5. Operating in the 57–64 GHz mmWave band, m^{3}MIMO supports up to 1 GHz of bandwidth and functions in a fully digital mode, enabling 8×8 MIMO with OFDM transmissions. While m^{3}MIMO is capable of both single-user multi-input multi-output (SU-MIMO) and multi-user multi-input multi-output (MU-MIMO) modes, MAGIC leverages only the SU-MIMO mode for extracting the CFR, which serves as the key sensing primitive in our system. Further technical details of the testbed are available in [12].

A. Data Collection Campaigns

We conducted data collection campaigns in three different environments – a Conference room, a Study room, and a Lab space with two different subjects performing ten different micro-gestures. The considered gestures are: thumbs up, thumbs down, victory, okay, pointing, pinch, tap fingers, fingers crossed, fist, half open palm as depicted in Figure 6. We chose this set of micro-gestures as they represent fine,

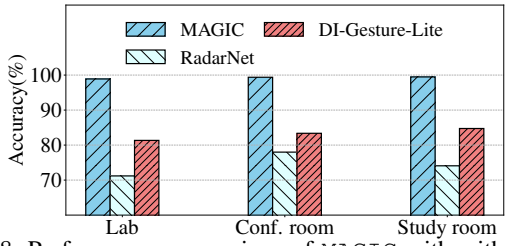


Fig. 8: Performance comparison of MAGIC with with state-of-the-art (SOTA) approaches.

nuanced movements that are common in everyday communication and interactions. These gestures are often subtle and exhibit minimal variation in physical movement, making them particularly challenging for gesture recognition systems. This selection is motivated by the need to evaluate the robustness of our system in recognizing gestures that require high sensitivity to spatial and temporal variations. Unlike broader, more distinct gestures, micro-gestures like "fist" or "half open palm" demand advanced processing capabilities to accurately identify subtle differences in motion patterns and orientations.

In each of the environments, we perform data collection by placing the transmitter and the receiver at a distance of 2 m from each other while the subjects perform different gestures facing toward the beamformer as presented in Figure 7. The transmitter broadcasts the NDP at a rate of 166 packets/s while we collect the estimated CFR at the receiver during the execution of different gestures by different subjects. We synchronously record the video stream from a fixed camera location to create the vision-based ground truth. Every subject performs all the different gestures 50 times in each of the environments while each of the gestures lasts for 3 seconds. The CFR corresponding to each gesture is captured and labeled synchronously to create the training dataset. The captured dataset is then processed and fed to the learning model as presented in Section II-C.

IV. Performance Evaluation

In this section, we evaluate the performance of MAGIC by comparing it with two state-of-the-art mmWave radar-based gesture recognition approaches – RadarNet, and DI-Gesture-Lite. We also analyze the performance variations of MAGIC with different subchannel resolutions and variations in subject. Next, we perform a detailed ablation study to understand the impact of TCN in overall performances and the impact of DAPP preprocessing pipeline and augmentation in generalization. Finally, we compare the generalization performance of the proposed learning approach ATEN with two other SOTA few-shot learning approaches – ReWiS and FREL.

A. MAGIC Vs SOTA Approaches

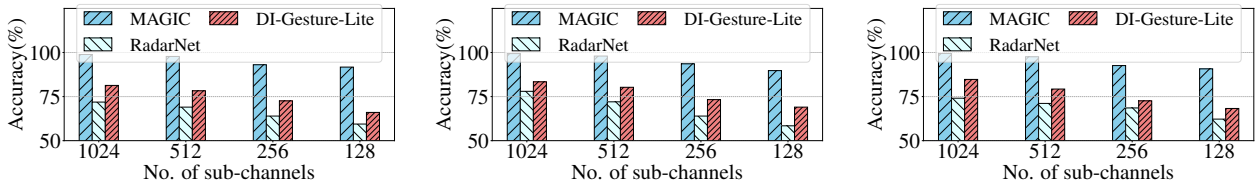
Figure 8 presents the performance comparison of MAGIC with other SOTA approaches. The results demonstrate the clear superiority of MAGIC compared to SOTA models – RadarNet and DI-Gesture-Lite, across various environments. In the Lab setting, MAGIC achieves an accuracy of 98.89%, significantly outperforming the SOTA models by margins of 27.7% and 17.2%, respectively. The trend is consistent in the Study and Conference room, where MAGIC achieves over 98.8% accuracy, with an average improvement of approximately 24.97% over other SOTA models. Notably, MAGIC's performance remains remarkably stable across different environments, with a variation of less than 0.6% accuracy among the Lab, Conference, and Study environments. This stability contrasts sharply with the SOTA models, which exhibit significantly larger performance fluctuations, reflecting their susceptibility to environmental changes. The results emphasize MAGIC's robustness and adaptability, showcasing its ability to maintain high accuracy regardless of the operating conditions.

B. MAGIC performance with different subchannel resolution

We evaluate the performance of MAGIC as a function of different subchannel resolutions by comparing it with SOTA approaches as presented in Figure 9. The results show that the MAGIC consistently outperforms the state-of-the-art approaches, DI-Gesture-Lite and RadarNet, across all subchannel resolutions and environments. Notably, MAGIC demonstrates remarkable robustness as the number of subchannels decreases, maintaining high accuracy levels even at lower resolutions. For instance, in the Lab environment, while DI-Gesture-Lite and RadarNet experience significant drops in accuracy – 59.43% and 66.03%, respectively, at 128 subchannels, MAGIC sustains an accuracy of 91.73%. This trend is similarly observed in the Conference and Study room scenarios, where MAGIC achieves approximately 89.73% and 90.73% accuracy at 128 subchannels, respectively. In contrast, the competing models, DI-Gesture-Lite and RadarNet, show a more pronounced sensitivity to the reduction in subchannel resolution, suggesting that their feature extraction or adaptation mechanisms are less robust to such changes. Additionally, MAGIC's accuracy remains above 90% in all scenarios when the subchannel resolution is 256 or higher, reflecting its reliability in high-resolution settings.

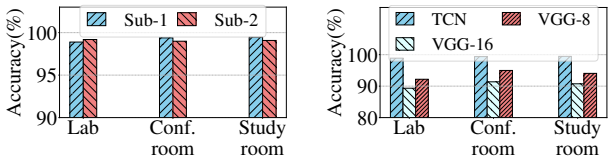
C. MAGIC performance with different subjects

Figure 10a presents the performance of MAGIC with different subjects in different environments. MAGIC demonstrates remarkable consistency and robustness across different subjects in all environments, with minimal variation in accuracy between subjects. In the Lab environment, while both subjects achieve near-perfect performance, the slight improvement for Sub-2 (99.19%) over Sub-1 (98.89%) might indicate marginal differences in gesture execution clarity or environmental factors. Interestingly, the Study room shows the highest accuracy for Sub-1 (99.47%) across all environments, suggesting that Sub-1's gestures in this environment were exceptionally well captured, possibly due to favorable environmental dynamics or subject performance. Overall, the small differences in performance indicate that MAGIC



(a) Lab environment (b) Conference room (c) Study room

Fig. 9: Comparative analysis of MAGIC with SOTA approaches as a function of number of subchannels



(a) MAGIC with different subjects (b) MAGIC performance with TCN vs VGG variants

Fig. 10: MAGIC performance with (a) different subjects and (b) impact of TCN vs VGG-8 and VGG-16.

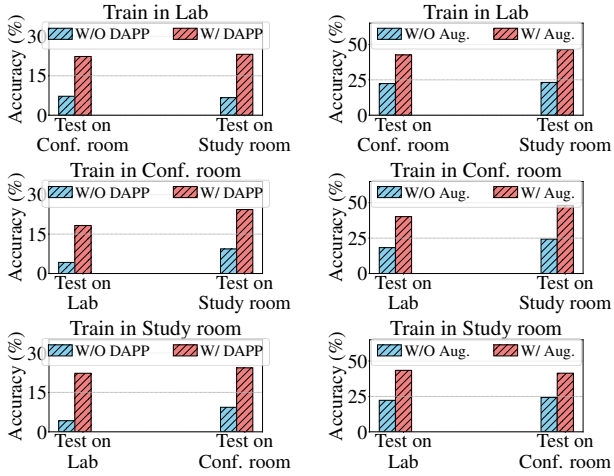
effectively generalizes across subjects while maintaining consistently high accuracy, underscoring its robustness and suitability for real-world multi-user applications.

D. Evaluating the role of TCN in MAGIC

We evaluate the impact of adopting the TCN architecture over traditional convolutional neural network (CNN) models, such as VGG-8 and VGG-16, within MAGIC. The results are presented in Figure 10b. The results reveal interesting insights into MAGIC's performance with TCN compared to VGG-8 and VGG-16 across environments. TCN consistently outperforms both VGG variants, highlighting its ability to better capture temporal dependencies in gesture recognition tasks. While VGG-16 slightly outperforms VGG-8 due to its deeper architecture, the performance gap between TCN and VGG-16 suggests that increasing network depth alone is insufficient for these tasks. The consistent performance advantage of TCN across all environments, demonstrates its robustness and suitability for dynamic, real-world scenarios. These findings underscore the importance of leveraging architectures like TCN, which are better aligned with the temporal nature of gesture data, over purely depth-focused designs like VGG.

E. Impact of the DAPP Preprocess Pipeline and Data Augmentation in Generalization

We now highlight the impact of the proposed data preprocessing timeline, DAPP, on domain generalization, as illustrated in Figure 11a. The results show that DAPP significantly enhances cross-environment accuracy, with notable improvements over models without preprocessing. For example, when trained with the data from the Lab environment and tested with the data from the Conference room, accuracy increases from 7.21% without DAPP to 22.34% with DAPP, and in the Study room, it improves from 6.68% to 23.17%. These improvements are consistent across all scenarios. Interestingly,



(a) Impact of preprocessing pipeline – DAPP (b) Impact of data augmentation

Fig. 11: Impact of preprocess pipeline – DAPP and data augmentation in generalization

DAPP is particularly effective in handling challenging domain shifts, as seen in the Conference room to Lab scenario, where the improvement is more than four times the baseline – from 4.23% to 18.28%. This shows that DAPP excels at extracting robust features, enabling the model to better adapt to unseen domains.

As depicted in Figure 11, data augmentation provides even greater improvements, amplifying the effects of DAPP. When trained in the Lab and tested in the Conference room, accuracy increases from 22.34% (with only DAPP) to 42.56% with combined DAPP and augmentation, while testing in the Study room shows a similar boost from 23.17% to 46.32%. These results demonstrate augmentation's ability to enhance generalization across diverse environments by introducing variability during training. The consistent performance gains across all settings suggest that augmentation addresses domain shifts more comprehensively than preprocessing alone.

DAPP and data augmentation complement each other, with DAPP providing foundational improvements and augmentation introducing the necessary variability to handle unseen conditions. For example, in the Lab-to-Conference scenario, the combined improvements take accuracy from a mere 7.21% (without DAPP or augmentation) to 42.56% by applying both DAPP and augmentation strategies. This combined approach proves critical for handling domain shifts, ensuring robust generalization in gesture recognition tasks

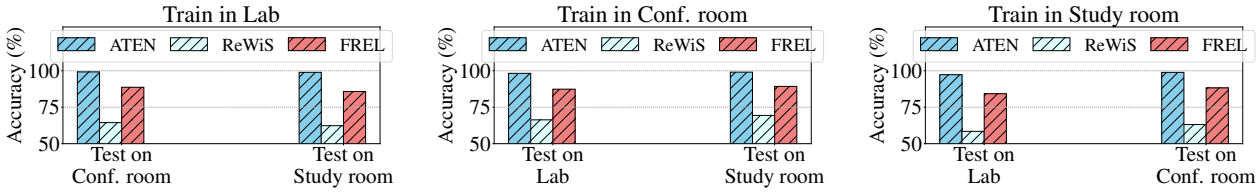


Fig. 12: Environment generalization performance of ATEN – the domain generalization algorithm of MAGIC

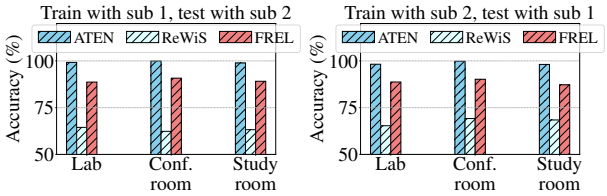


Fig. 13: Subject generalization performance of ATEN – the generalization algorithm of MAGIC

across diverse environments.

F. ATEN in Generalizing Environments

We present the generalization performance of ATEN – the domain generalization algorithm of MAGIC in Figure 12. ATEN demonstrates exceptional generalization performance across environments, significantly outperforming the SOTA wireless sensing generalization algorithm – ReWiS and FREL. For instance, when trained in the Lab and tested in the Conference room, ATEN achieves 99.21% accuracy with only 1000 new samples from new environment which is worth of only 6 Seconds of data collection. On the contrary, ReWiS and FREL achieve only 64.46% and 88.68% repectively, while ATEN shows improvements of 54% and 12%, respectively. A similar trend is observed when trained in the Study room and tested in the Lab, where ATEN achieves 97.28%, outperforming ReWiS (58.47%) and FREL (84.24%) by 66% and 15%, respectively. On average, ATEN improves accuracy by approximately 60% over ReWiS and 14% over FREL across all environments. These results highlight ATEN's ability to handle domain shifts effectively, achieving consistently high performance where other methods falter. ATEN's ability to achieve such consistent improvements underscores its robustness and makes it a superior choice for cross-domain wireless sensing.

G. ATEN in Generalizing Subjects

As presented in Figure 13, ATEN demonstrates outstanding cross-subject generalization performance, significantly outperforming ReWiS and FREL. For instance, when trained with Sub-1 and tested on Sub-2, ATEN achieves 99.21% accuracy in the Lab, compared to 64.46% for ReWiS and 88.68% for FREL. On average, ATEN improves accuracy by over 55% compared to ReWiS and approximately by 10% compared to FREL, showcasing its effectiveness in handling subject variability. These findings confirm ATEN's robustness and adaptability for cross-subject gesture recognition, consistently maintaining high performance across diverse scenarios.

V. RELATED WORK

Gesture recognition has become a critical area of research due to its applications in Human-Computer Interaction (HCI), healthcare, and security. To address the limitations of traditional vision-based systems, researchers have explored alternative sensing methods such as Wi-Fi signals [13]–[15] and mmWave signals [16], [17]. For a detailed overview, readers can refer to [18] and [9], which provide insights into performance metrics, applications, and machine learning techniques used with mmWave radar.

Significant progress has been made in mmWave gesture recognition. For instance, Yu et al. [7] developed a mmWave MIMO radar-based gesture recognition system using CNN and LSTM, achieving over 90% accuracy for 12 gestures. Liu et al. [19] proposed mHomeGesUser, a lightweight CNN-based framework for real-time arm gesture recognition in smart home scenarios. Identification has also emerged as an important factor in mmWave gesture systems [8], [20]. Xu et al. [8] introduced GesturePrint, combining attention-based mechanisms for gesture recognition and user identification, achieving over 98% accuracy for 15 gestures and identifying 17 participants. Liu et al. [21] presented M-Gesture, a person-independent, real-time gesture recognition system. Addressing data scarcity, Yan et al. [22] proposed mmGesture, a semi-supervised system leveraging data augmentation to minimize labeling costs. Various hardware architectures have also been explored in recent works [23], [24]. Mao et al. [23] introduced a multiple Frequency Modulation Continuous Wave (FMCW) radar-based system using LSTM, achieving 98% accuracy for eight gestures. Yu et al. [24] employed continuous wave (CW) radar for detecting valid frames in real-time gesture recognition. For generalization across environments, Liu et al. [25] proposed mTransSee, a transfer learning framework that adapts to new environments while preserving recognition accuracy. Compact and efficient frameworks have been designed for constrained devices, such as RadarNet [26], which combines CNN and LSTM for computational efficiency, and Gesture-mmWAVE [27], a system using multilevel feature fusion and transformers for embedded deployment. DI-Gesture [28] and its variant DI-Gesture-Lite incorporate Dynamic Range Angle Images (DRAI) and basic data augmentation to enhance domain independence and robustness. RadarNet [26] processes range-Doppler maps for spatial-temporal motion details, summarizing radar data into compact representations for practical applications.

While these approaches achieve impressive results, they rely on specialized radar hardware, increasing system com-

plexity and cost. In contrast, MAGIC eliminates the need for dedicated radars by leveraging mmWave MIMO CSI, integrating a domain-adaptive preprocessing pipeline (DAPP), robust data augmentation, and the ATEN meta-learning framework. This enables MAGIC to achieve high accuracy, robust generalization, and adaptability across diverse environments and subjects, establishing itself as a practical and cost-effective solution for real-world gesture recognition.

VI. Conclusion

In this paper, we introduced MAGIC, a novel gesture recognition framework leveraging mmWave MIMO CSI, eliminating the reliance on dedicated radar hardware and significantly reducing system complexity and data overhead. By incorporating the domain-adaptive preprocessing pipeline (DAPP), robust data augmentation, and the ATEN meta-learning framework, MAGIC achieves exceptional adaptability, maintaining up to 99% accuracy and demonstrating strong generalization across diverse environments and subjects. Compared to SOTA wireless sensing generalization approaches like ReWiS and FREL, MAGIC improves accuracy by 60% and 14% on average, respectively, highlighting its robustness under domain shifts and subject variability. Furthermore, its ability to operate with reduced subchannel resolutions and minimal training data emphasizes its practicality for resource-constrained scenarios. These results show that MAGIC is a scalable, efficient, and robust solution for real-world gesture recognition applications, setting a new benchmark for mmWave-based sensing systems.

Acknowledgments

This work is funded in part by the National Science Foundation (NSF) grant CNS-2134973, ECCS-2229472, and ECCS-2329013, by the Air Force Office of Scientific Research under contract number FA9550-23-1-0261, by the Office of Naval Research under award number N00014-23-1-2221.

References

[1] Declan McGlynn, Rolling Stones., "Music and the Metaverse: Are we on the brink of a virtual artist revolution?" https://tinyurl.com/39ed34dd, 2022.

[2] J. Wang, R. Shi, W. Zheng, W. Xie, D. Kao, and H.-N. Liang, "Effect of frame rate on user experience, performance, and simulator sickness in virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2478–2488, 2023.

[3] K. F. Haque, F. Meneghello, and F. Restuccia, "Integrated Sensing and Communication forv Efficient Edge Computing," in *WiMob*, 2024.

[4] I. of Electrical and E. E. (IEEE), "Ieee standard for information technology– part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications," *IEEE Std 802.11ax-2021 (Amendment to IEEE Std 802.11-2020)*, pp. 1–767, 2021.

[5] K. F. Haque, M. Zhang, and F. Restuccia, "Simwisense: Simultaneous multi-subject activity classification through wi-fi signals," in *2023 IEEE 24th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2023, pp. 46–55.

[6] K. F. Haque, M. Zhang, F. Meneghello, and F. Restuccia, "BeamSense: Rethinking Wireless Sensing with MU-MIMO Wi-Fi Beamforming Feedback," *Computer Networks*, vol. 258, p. 111020, 2025.

[7] J.-T. Yu, Y.-H. Tseng, and P.-H. Tseng, "A mmwave mimo radar-based gesture recognition using fusion of range, velocity, and angular information," *IEEE Sensors Journal*, 2024.

[8] L. Xu, K. Wang, C. Gu, X. Guo, S. He, and J. Chen, "Gestureprint: Enabling user identification for mmwave-based gesture recognition systems," in *2024 IEEE ICDCS*. IEEE, 2024, pp. 1074–1085.

[9] A. Soumya, C. Krishna Mohan, and L. R. Cenkeramaddi, "Recent advances in mmwave-radar-based sensing, its applications, and machine learning techniques: A review," *Sensors*, vol. 23, no. 21, p. 8901, 2023.

[10] P. Pan, F. Zhang, A. Zhou, H. Ma, and H. Jia, "mmcare: A nursing care activity monitoring system via mmwave sensing," in *Proceedings of the ACM Turing Award Celebration Conference-China 2024*, 2024, pp. 18–22.

[11] A. Böttcher and D. Wenzel, "The frobenius norm and the commutator," *Linear algebra and its applications*, vol. 429, no. 8-9, pp. 1864–1885, 2008.

[12] K. F. Haque, F. Meneghello, K. M. Rumman, and F. Restuccia, "m3MIMO: An 8×8 mmWave Multi-User MIMO Testbed for Wireless Research," in *Proceedings of the MobiCom*, 2024, p. 1922–1929.

[13] J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, and L. Xie, "Efficientfi: Towards large-scale lightweight wifi sensing via csi compression," *IEEE Internet of Things Journal*, 2022.

[14] Q. Bu, X. Ming, J. Hu, T. Zhang, J. Feng, and J. Zhang, "Transfersense: towards environment independent and one-shot wifi sensing," *Personal and Ubiquitous Computing*, vol. 26, no. 3, pp. 555–573, 2022.

[15] X. Zheng, K. Yang, J. Xiong, L. Liu, and H. Ma, "Pushing the limits of wifi sensing with low transmission rates," *IEEE Transactions on Mobile Computing*, 2024.

[16] T. Gu, Z. Fang, Z. Yang, P. Hu, and P. Mohapatra, "Mmsense: Multi-person detection and identification via mmwave sensing," in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, 2019, pp. 45–50.

[17] K. Liang, A. Zhou, Z. Zhang, H. Zhou, H. Ma, and C. Wu, "mmstress: Distilling human stress from daily activities via contact-less millimeter-wave sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–36, 2023.

[18] J. Zhang, R. Xi, Y. He, Y. Sun, X. Guo, W. Wang, X. Na, Y. Liu, Z. Shi, and T. Gu, "A survey of mmwave-based human sensing: Technology, platforms and applications," *IEEE Communications Surveys & Tutorials*, 2023.

[19] H. Liu, Y. Wang, A. Zhou, H. He, W. Wang, K. Wang, P. Pan, Y. Lu, L. Liu, and H. Ma, "Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 4, pp. 1–28, 2020.

[20] Y. Wang, T. Gu, T. H. Luan, and Y. Yu, "Your breath doesn't lie: Multi-user authentication by sensing respiration using mmwave radar," in *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2022, pp. 64–72.

[21] H. Liu, A. Zhou, Z. Dong, Y. Sun, J. Zhang, L. Liu, H. Ma, J. Liu, and N. Yang, "M-gesture: Person-independent Real-time In-air Gesture Recognition using Commodity Millimeter Wave Radar," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3397–3415, 2021.

[22] B. Yan, P. Wang, L. Du, X. Chen, Z. Fang, and Y. Wu, "mmGesture: Semi-Supervised Gesture Recognition System using mmWave Radar," *Expert Systems with Applications*, vol. 213, p. 119042, 2023.

[23] Y. Mao, L. Zhao, C. Liu, and M. Ling, "A low-complexity hand gesture recognition framework via dual mmwave fmcw radar system," *Sensors*, vol. 23, no. 20, p. 8551, 2023.

[24] M. Yu, N. Kim, Y. Jung, and S. Lee, "A frame detection method for real-time hand gesture recognition systems using cw-radar," *Sensors*, vol. 20, no. 8, p. 2321, 2020.

[25] H. Liu, K. Cui, K. Hu, Y. Wang, A. Zhou, L. Liu, and H. Ma, "mtranssee: Enabling environment-independent mmwave sensing based gesture recognition via transfer learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–28, 2022.

[26] E. Hayashi, J. Lien, N. Gillian, L. Giusti, D. Weber, J. Yamanaka, L. Bedal, and I. Poupyrev, "RadarNet: Efficient Gesture Recognition Technique Utilizing a Miniature Radar Sensor," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[27] B. Jin, X. Ma, B. Hu, Z. Zhang, Z. Lian, and B. Wang, "Gesture-mmwave: Compact and accurate millimeter-wave radar-based dynamic gesture recognition for embedded devices," *IEEE Transactions on Human-Machine Systems*, 2024.

[28] Y. Li, D. Zhang, J. Chen, J. Wan, D. Zhang, Y. Hu, Q. Sun, and Y. Chen, "Di-Gesture: Domain-Independent and Real-time Gesture Recognition with Millimeter-Wave Signals," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 5007–5012.