This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3557561

VR-Fi: Positioning and Recognizing Hand Gestures via VR-embedded Wi-Fi Sensing

Hongbo Wang, *Graduate Student Member, IEEE*, Xin Li, *Member, IEEE*, Jiachun Li, *Graduate Student Member, IEEE*, Haojin Zhu, *Fellow, IEEE* and Jun Luo, *Fellow, IEEE*

***Abstract*—Accurate gesture-based interactions are crucial for enhancing the immersive experience in VR (*virtual reality*) systems; they in turn necessitate gesture positioning and recognition in *physical world*. However, existing VR gesture recognition methods are predominantly vision-based, incurring high computational demands and raising privacy concerns. Meanwhile, Wi-Fi-based gesture recognition methods, deemed as promising complement to vision-based ones, typically lack gesture positioning capabilities. To this end, we propose VR-Fi, a gesture positioning and recognition system leveraging VR(-headset)-embedded Wi-Fi. To position gestures across different areas, VR-Fi innovates in a *frequency-hopping bandwidth expansion* (FHBE) technique to improve spatial resolution for locating a target. Additionally, VR-Fi innovates in neural models to process the FHBE-enhanced Wi-Fi CSI (channel state information) and enable the multi-task requirements of the joint positioning and recognition of hand gestures. Extensive experimental results demonstrate that VR-Fi achieves a positioning accuracy of 94.47%, a recognition accuracy of 92.13%, and a joint accuracy of 89.47%.**

***Index Terms*—Wi-Fi human sensing, ISAC, localization, gesture recognition, virtual reality.**

I. Introduction

VR (*virtual reality*) systems employ gesture-based interactions to enrich immersive experiences, enabling users to intuitively manipulate digital content and interact with virtual interfaces. These interactions are pivotal for tasks such as selecting options, navigating menus, and handling virtual objects, thereby improving the usability and accessibility of VR [1]. Traditional controller-based manipulation, which constrains the natural fluidity of user movements, is increasingly being supplanted by bare-hand control, offering a more genuine interactive experience [2], [3]. This shift necessitates the development of robust, device-free systems that can precisely track and recognize gestures within designated areas and convert them into meaningful VR actions. Most VR gesture recognition methods currently in use are vision-based [4], [5], utilizing various cameras to capture gestures. This procedure typically involves gesture segmentation, tracking hand feature points, estimating hand direction, and recognizing gestures. However, these methods may be resource-intensive, risking hardware overload and degrading real-time performance [6]. They also present privacy concerns, as they depend on cameras to capture and analyze user movements [7], [8]. Additionally, these systems may encounter difficulties in achieving accurate gesture recognition under varied lighting conditions [9], [10] or in the presence of obstructions [11], [12].

In response to the limitations of optical vision technologies, researchers have explored alternative non-optical approaches for gesture recognition [13]–[15]. Among these, Wi-Fi-based methods [16]–[19] stand out thanks to the wide deployment of Wi-Fi infrastructure. These methods are particularly well-suited for VR scenarios, as they eliminate the need for any hardware modifications to existing commercial VR headsets. Specifically, these methods monitor changes in Wi-Fi signal characteristics to detect hand movements, with recent influential studies often employing commercial network interface cards (NICs) to capture CSI (channel state information) for gesture recognition. Such Wi-Fi-based systems are particularly valued for their minimal computational requirements, attributed to the sparsity of electromagnetic signals, which facilitates the real-time responsiveness essential for sustained immersion and seamless interaction in VR applications. Furthermore, these systems operate without the need for video recording, significantly enhancing user privacy. Building on these strengths, Wi-Fi-based gesture recognition serves as a valuable supplement to vision-based systems, particularly in scenarios of limited hardware capabilities, high privacy

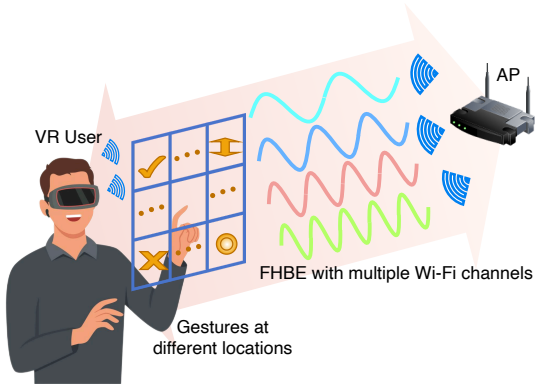


Fig. 1: Vision of VR-Fi: By utilizing the FHBE technique to enhance spatial resolutions, VR-Fi obtains necessary frequency diversity for subsequent joint gesture positioning and recognition, enabling accurate gesture recognition in a 3×3 grid of areas of a VR environment.

Hongbo Wang is with the Collaborative Initiative, Interdisciplinary Graduate Programme, Nanyang Technological University (NTU), Singapore 639798, and also with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore 639798 (e-mail: hongbo001@ntu.edu.sg).

Xin Li and Jun Luo are with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore 639798 (e-mail: l.xin@ntu.edu.sg; junluo@ntu.edu.sg).

Jiachun Li, and Haojin Zhu are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, 200240. (Email: jiachunli@sjtu.edu.cn; zhu-hj@cs.sjtu.edu.cn)

requirements, or insufficient visual conditions.

However, the adaptation of Wi-Fi-based technologies to VR gesture recognition presents its own set of challenges. Unlike traditional gestures in non-VR environments, VR settings typically allow users to perform varied gestures in different areas to fulfill various operation needs, thus necessitating an initial determination of the area of hand's position. Moreover, owning to Wi-Fi's limited capability for device-free tracking of moving objects, particularly for fine-grained hand movements,[1] we reform continuous tracking into a discrete classification problem. Specifically, considering users' familiarity with vertical and horizontal orientations [21], [22], organizing the reachable gesture space into a 3×3 grid should enhance intuitive and precise interactions, as shown in Fig. 1. Despite these adaptations, whether existing Wi-Fi sensing techniques may effectively handle joint positioning (albeit coarse-grained) and recognition remains an open issue. This inability largely stems from the inherent hardware constraints of Wi-Fi sensing: the limited bandwidth and number of antennas typically available on commercial NICs confine the spatial resolution (in both range and bearing) achievable. Consequently, these limitations prevent the Wi-Fi sensing system from separating gesture-induced paths from the background, thereby hindering accurate identification of gesture positions.

To address the challenges outlined above, we propose VR-Fi, the first gesture positioning and recognition system in VR environments, leveraging VR(-headset)-embedded Wi-Fi, as depicted in Fig. 1[2]. To handle positioning, VR-Fi innovates a novel *frequency-hopping bandwidth expansion* (FHBE) technique, which captures greater frequency diversity of CSI to enhance spatial resolution. FHBE specifically involves a channel selection algorithm to choose the optimal channels for accurate positioning and recognition of gestures, incorporating precise frequency hopping control to ensure reliable data collection via commercial Wi-Fi NICs embedded in VR devices. Given that traditional signal processing techniques are incapable of synthesizing FHBE-enhanced CSI samples (with frequency diversity) into gestures and positions, VR-Fi adopts a neural model combined with a multi-task learning strategy. Specifically, VR-Fi employs an adaptive gating mechanism to regulate the contributions of various expert sub-networks to different tasks, thus achieving joint positioning and recognition of gestures in VR environments. In summary, our major contributions are:

- We propose VR-Fi as the first VR embedded Wi-Fi sensing system that simultaneously implements gesture positioning and recognition.
- To achieve accurate gesture positioning, we design an FHBE technique for VR-Fi to captures greater frequency diversity using commercial Wi-Fi NICs with a limited number of antennas.
- We develop novel deep learning models with a multi-task learning strategy to address the challenge of synthesizing FHBE-enhanced channel samples, ultimately enabling the joint positioning and recognition of hand gestures.
- We implement the first prototype of VR-Fi on commercial VR headset, and conduct extensive evaluations on it to demonstrate VR-Fi's excellent capabilities in gesture positioning and recognition.

The rest of our paper is structured as follows. Sec. II introduces the background and motivation of VR-Fi. Sec. III elaborates on the system design of VR-Fi. Sec. IV and Sec. V explain VR-Fi's implementation and report the extensive evaluations. Related works and discussion of VR-Fi are briefly discussed in Sec. VI, followed by the conclusion of our paper in Sec. VII.

II. Background and Motivation

In this section, we first establish a basic model for Wi-Fi sensing and analyze the relationship between CSI with range and bearing under ideal antenna hardware conditions. Subsequently, we derive a simplified sensing model tailored to the constraints of commercial NICs to elucidate why existing gesture recognition methodologies fall short in achieving gesture positioning. Finally, we compare the gesture positioning results across various bandwidths to demonstrate the potential benefits of bandwidth expansion for gesture positioning.

A. Wi-Fi Sensing Basic

Assuming a Wi-Fi sensing system with a transmitter-receiver (Tx-Rx) pair, we start by introducing a CSI model to establish the foundation for gesture positioning. Fig. 2 illustrates the Rx as an ideal antenna array with M elements arranged vertically and N elements arranged horizontally. The spacing between adjacent antennas is dz and dy, respectively. The CSI model involves parameters (τ, θ, ϕ), representing the time of flight (ToF), azimuth direction of the angle of arrival ($\mathrm{AoA_A}$), and elevation direction of the angle of arrival ($\mathrm{AoA_E}$) respectively, as determined by the range and bearing of the sensing subjects. Considering L distinct propagation paths of the Wi-Fi orthogonal frequency division multiplexing (OFDM) signal with K subcarriers, the received CSI $\boldsymbol{H} = [h_{m,n,k}]$ can be expressed as follows:

$$h_{m,n,k} = \textstyle\sum_{l=1}^{L} \alpha_{m,n,k,l} \cdot h_{k,l}^{\mathrm{ToF}} \cdot h_{n,l}^{\mathrm{AoA_A}} \cdot h_{m,l}^{\mathrm{AoA_E}}, \quad (1)$$

where $h_{k,l}^{\mathrm{ToF}}$, $h_{n,l}^{\mathrm{AoA_A}}$, and $h_{m,l}^{\mathrm{AoA_E}}$ are given by

$$\begin{aligned} h_{k,l}^{\mathrm{ToF}} &= \mathrm{e}^{-j2\pi(f_c + k\Delta f)\tau_l}, \\ h_{n,l}^{\mathrm{AoA_A}} &= \mathrm{e}^{j2\pi(n-1)d_y \cos\theta_l \sin\phi_l (f_c + k\Delta f)/c}, \\ h_{m,l}^{\mathrm{AoA_E}} &= \mathrm{e}^{j2\pi(m-1)d_z \sin\theta_l (f_c + k\Delta f)/c}. \end{aligned}$$

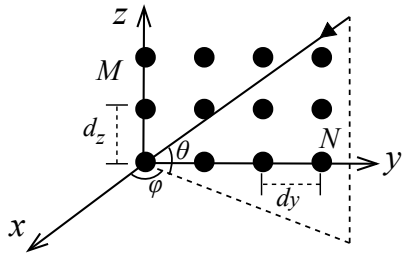
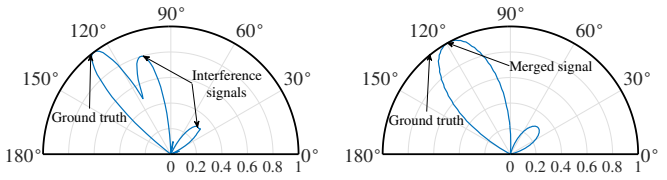


Fig. 2: Rx antenna array

[1]For example, Widar2.0 [20] (arguably the best to our knowledge) has 20% errors exceeding 1 m, rendering it hardly be usable for VR applications.

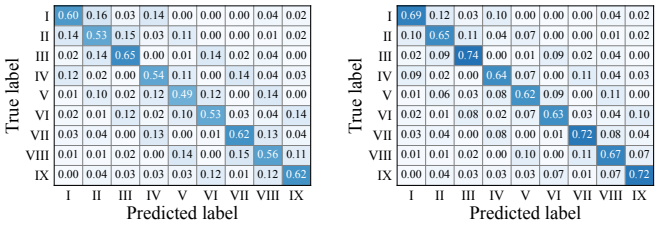
[2]Although shown in single-user scenario, VR-Fi can be readily adapted to multi-user environments by leveraging the near-field domain effect, as explored in the multi-user-focused study, MUSE-Fi [23]

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3557561



(a) Eight antennas. (b) Four antennas.

Fig. 3: AoA spectrum with different number of antennas.



(a) Existing proposal of 20 MHz. (b) Maximum of 160 MHz.

Fig. 4: Gesture positioning under different Wi-Fi bandwidth.

where m, n, k, and l respectively index the vertical antenna, horizontal antenna, subcarrier, and path, α represents channel gain, f_c and Δf respectively denote channel centre frequency and subcarrier bandwidth, and c is the speed of light.

B. Infeasible Gesture Positioning of Prior Art

Existing gesture recognition works mainly used commercial NICs to capture dynamic features such as Doppler Frequency Shift (DFS) and phase variations to recognize gestures [17], [24], [25]. However, gesture positioning requires accurate acquisition of static features that characterize the position (e.g., ToF and AoA). In this subsection, we further explore the capability of existing Wi-Fi gesture solutions to estimate ToF, AoA_A, and AoA_E. Sec. II-A describes an antenna array under the assumption of ideal hardware conditions. However, typical real-life commercial NICs are equipped with only 2 to 4 linearly distributed antennas. In this scenario, with such N antennas, Eqn. (1) simplifies to:

$$h_{n,k} = \sum_{l=1}^{L} \alpha_{n,k,l} \cdot h_{k,l}^{\text{ToF}} \cdot h_{n,l}^{\text{AoA}_\text{A}}. \tag{2}$$

It is evident that in this configuration, the capability to estimate the elevation angle of arrival (i.e., AoA_E) is absent. This limitation significantly impairs the system's ability to accurately determine the vertical position of gestures.

For the resolution of the azimuth direction of AoA (i.e., AoA_A), existing AoA-based localization systems typically employ multiple antennas to find the intersection of AoAs [26]. The resolution of AoA is primarily determined by the number of antennas. Increasing the number of antennas results in narrower angle beams, thereby enhancing the ability to distinguish between two adjacent angle signals, as illustrated in Fig. 3. Fig. 3(a) illustrates that with eight antennas, it is possible to separate the target signal from interference signals, accurately capturing the angle matching the ground truth. However, in Fig. 3(b), using only four antennas—the maximum number available on commercial NICs—results in signals merging. Consequently, due to the limited number of antennas in commercial NICs, existing Wi-Fi gesture recognition methods can only provide limited bearing resolution[3].

Next, we discuss the capability of the aforementioned Wi-Fi system to estimate ToF. According to [27], the temporal resolution of a radio system follows the formula $\Delta\tau = 1/B$, where B represents the total sensing bandwidth. This implies that the range resolution of human gestures, $\Delta R = c\Delta\tau = c/B$, is directly proportional to the bandwidth B. To distinguish the 3×3 grid of areas depicted in Fig. 1 to achieve gesture positioning, decimeter-level resolution is generally required, given the typical range of hand movements. Therefore, the required bandwidth B needs to be approximately 1 GHz. However, existing Wi-Fi gesture recognition works often rely on the Wi-Fi 5 protocol, which offers a bandwidth of only 20 MHz, significantly below the bandwidth requirement for accurate ToF estimation. As a result, these systems struggle to accurately estimate the ToF of gesture-induced signal paths.

In conclusion, the *range resolution* and *bearing resolution* derived from the ToF and AoA terms in Eqn. (2) are insufficient for accurate hand positioning. To further illustrate this point, following the experimental setup in Sec. IV-2, we collect CSI data across 3×3 grid of gesture positions using the 20 MHz sensing bandwidth of existing proposals. We then apply the feature extraction model from Widar3.0 [17], followed by a fully connected layer for gesture positioning. As shown in Fig. 4(a), the average accuracy across various positions is only 57.1%, highlighting the significant limitations of current methods in precise gesture positioning. Therefore, it is imperative to develop innovative approaches to effectively overcome such limitation for VR scenarios.

C. Bandwidth Expansion for Gesture Positioning

Considering that gesture positioning requires an accurate estimation of the hand's range and bearing, enhancing its resolution becomes a critical task. As previously noted, the resolution of ToF is linearly related to the sensing bandwidth; thus, a larger bandwidth enables a better differentiation of ToFs along different paths. Moreover, while the bearing resolution is dependent on the number of antennas, the estimation of AoA_A can still benefit from a wider bandwidth. This is because improving the resolution of $d_y \cos\theta_l \sin\phi_l / c$ of $h_{n,l}^{\text{AoA}_\text{A}}$ (which also is a temporal component) results in higher precision in AoA_A estimation. Additionally, in commercial NIC configurations with linear antennas, although AoA_E cannot be estimated at the physical layer, gestures at different positions within the 3×3 grid can result in unique combinations of ToF and AoA_A in the received signals. Therefore, increasing the sensing bandwidth in Wi-Fi systems can significantly enhance the spatial resolution available for accurately estimating the range and bearing of hand gestures, potentially establishing a viable solution for positioning gestures in various areas.

To further validate this point, we present the accuracy of gesture positioning under the maximum directly accessible bandwidth of 160 MHz, as depicted in Fig. 4(b). The average

[3]Even attempts to estimate AoA_E by rearranging the antenna array (e.g., 2×2 configuration), the limited number of antennas in the elevation direction still constrains resolution.

accuracy across various locations is 67.5%, which underscores the effectiveness of increased bandwidth in enhancing the accuracy of gesture positioning, compared to the 20 MHz results shown in Fig. 4(a). However, a sensing bandwidth of 160 MHz represents the current upper limit achievable on mature commercial Wi-Fi hardware and still falls short for highly accurate identification of gesture positions. Therefore, it is crucial to explore methods to further expand the bandwidth to achieve precise gesture positioning.

III. VR-Fi System Design

Our VR-Fi is specifically designed for accurate gesture positioning and recognition in VR scenarios with extended Wi-Fi sensing bandwidth. VR-Fi consists of two major components, as shown in Fig. 5:

- Frequency-hopping Bandwidth Expansion: This technique is meticulously designed to expand the Wi-Fi sensing bandwidth. It integrates an optimal channel selection algorithm and employs precise frequency-hopping control to ensure effective and reliable data collection.
- MTPG-Net with Multi-task Learning: The model performs feature extraction, eliminates unknown interference, and employs various expert sub-networks to different tasks, thus achieving joint positioning and recognition of gestures in VR scenarios.

A. Frequency-hopping Bandwidth Expansion (FHBE)

Intuitively, expanding the sensing bandwidth B can be understood as increasing the number of subcarriers in Eqn.(1) and Eqn.(2), thereby enhancing frequency diversity for algorithms estimating range and bearing. A straightforward approach to achieving this is continuous channel stitching [28], [29]. However, this method is impractical due to several limitations: (1) the unavailability of a large number of channels at all times, (2) the time-consuming sampling process, which exceeds the coherence time budget, and (3) the computational complexity of the stitching procedure [30]. Fortunately, according to compressed sensing principles [31] and considering the sparse nature of many physical phenomena, it is unnecessary to acquire complete and redundant information—that is, continuous, full bandwidth—for the sensing of a single physical phenomenon. This insight motivates us to explore the feasibility of using a limited number of channel samples to reconstruct information that traditionally requires full-bandwidth sensing for accurate localization, thereby overcoming the aforementioned challenges of continuous channel stitching.

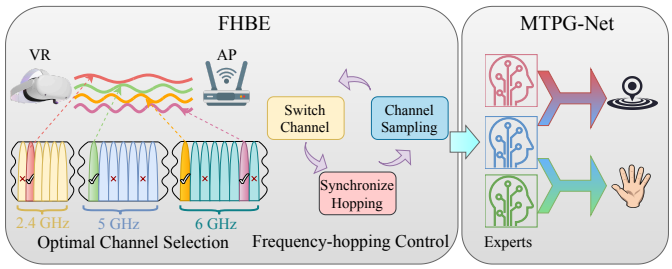


Fig. 5: VR-Fi system overview.

Specifically, FHBE of VR-Fi implements a discrete hopping strategy across all standard channels specified by the IEEE 802.11ax Wi-Fi protocol to extend the sensing bandwidth, thereby enhancing the spatial resolution of gesture positioning. In terms of channel selection, this strategy theoretically allows for the free choice of any channel, contingent on their availability at runtime. Despite this flexibility, VR-Fi has meticulously devised an optimal channel selection algorithm, engineered to ensure that the selected channels maximize frequency diversity for the subsequent MTPG-Net tasked with performing joint gesture positioning and recognition, thereby achieving high accuracy in both tasks. Moreover, this technique necessitates sampling only a few discrete channels, in contrast to utilizing continuous full bandwidth. Despite the minimal channel requirements, given the constraints of channel coherence time budgets [32], we have developed a precise hopping control mechanism to ensure that the commercial Wi-Fi NIC embedded in VR devices can rapidly acquire the necessary number of channel CSIs.

1) Optimal channel selection algorithm: The up-to-date Wi-Fi protocols (such as 802.11ax) accommodate the 2.4 GHz, 5 GHz, and 6 GHz bands. Considering the diverse usage across different countries and regions, we have selected 97 commonly used channels, each with a bandwidth of 20 MHz, as frequency-hopping candidates. They include 13 channels between 2412-2472 MHz, 8 channels from 5180-5320 MHz, 12 channels spanning 5500-5720 MHz, 5 channels within 5745-5825 MHz, and 59 channels from 5955-7115 MHz. Initially, the VR headset as Rx initiates gesture positioning and recognition commands, and the Tx access point (AP) and the Rx perform a sweep of all candidate channels to identify those available in the current environment (not occupied by other devices or reserved for special purposes). Additionally, the CSI of each candidate is recorded, providing critical input for the subsequent channel selection algorithm.

To make efficient use of the limited selection of available channels, we develop an optimal channel selection algorithm, as shown in **Algorithm 1**. Considering that closely spaced channels exhibit similar multipath propagation characteristics and environmental reflection and diffraction properties, the differences in CSI are minimal [33], [34]. Consequently, the frequency diversity they can provide is limited. The channel selection algorithm aims to prioritize channels with larger frequency separation and more distinct CSI differences, thereby avoiding information redundancy. The effectiveness of this selection strategy is further validated in Sec. V-C2. The algorithm begins by generating an initial set of candidate channel combinations $\mathcal{P}$, where each combination $\mathcal{C}$ consists of N_c randomly selected channels from the available pool $\mathcal{C}_{\text{avail}}$. For each combination $\mathcal{C} \in \mathcal{P}$, a score $D(\mathcal{C})$ is computed, representing the total pairwise Euclidean distance among the selected channels. This score can reflect the degree of distinctiveness in CSI values among the channels [32] and serves as a measure of their frequency diversity. The algorithm then selects the top k combinations with the highest scores, denoted as $\mathcal{P}_{\text{retain}}$, for further refinement.

To further enhance the selected combinations, the algorithm applies an adjustment process. For each $\mathcal{C} \in \mathcal{P}_{\text{retain}}$, the algo-

Algorithm 1: Optimal channel selection algorithm

Input: Available channels $\mathcal{C}_{\text{avail}}$; CSI data CSI_c; Number of channels to select N_{c}.
Output: Optimal channel set $\mathcal{C}_{\text{opt}}$.

Initialize: $\mathcal{P} \leftarrow \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m\}$, where $\mathcal{C}_i$ are randomly selected sets of size N_{c} from $\mathcal{C}_{\text{avail}}$.
for *each iteration* **do**
  foreach $\mathcal{C} \in \mathcal{P}$ **do**
    $D(\mathcal{C}) \leftarrow \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|\text{CSI}_{c_i} - \text{CSI}_{c_j}\|_2$;
  end
  $\mathcal{P}_{\text{retain}} \leftarrow \text{Top-k}(\mathcal{P}, D)$;
  $\mathcal{P}_{\text{adjusted}} \leftarrow \text{Adjust}(\mathcal{P}_{\text{retain}}, \mathcal{C}_{\text{avail}})$;
  $\mathcal{P}_{\text{new}} \leftarrow \mathcal{P}_{\text{retain}} \cup \mathcal{P}_{\text{adjusted}}$; $\mathcal{P} \leftarrow \mathcal{P}_{\text{new}}$;
end
$\mathcal{C}_{\text{opt}} \leftarrow \arg\max_{\mathcal{C} \in \mathcal{P}} D(\mathcal{C})$;
Function $\text{Adjust}(\mathcal{P}_{\text{retain}}, \mathcal{C}_{\text{avail}})$**:**
  $\mathcal{P}_{\text{adjusted}} \leftarrow \emptyset$;
  foreach $\mathcal{C} \in \mathcal{P}_{retain}$ **do**
    $c_{\text{swap}} \leftarrow \arg\min_{c \in \mathcal{C}} \sum_{j \in \mathcal{C} \setminus \{c\}} \|\text{CSI}_c - \text{CSI}_j\|_2$;
    $c' \leftarrow \arg\max_{c' \in \mathcal{C}_{\text{avail}} \setminus \mathcal{C}} \sum_{j \in \mathcal{C} \setminus \{c_{\text{swap}}\}} \|\text{CSI}_{c'} - \text{CSI}_j\|_2$;
    $\mathcal{C} \leftarrow (\mathcal{C} \setminus \{c_{\text{swap}}\}) \cup \{c'\}$;
    $\mathcal{P}_{\text{adjusted}} \leftarrow \mathcal{P}_{\text{adjusted}} \cup \{\mathcal{C}\}$;
  end
  return $\mathcal{P}_{\text{adjusted}}$;
end

rithm identifies the channel c_{swap} with the smallest contribution to $D(\mathcal{C})$, which is determined by the sum of its pairwise differences in CSI with other channels in $\mathcal{C}$. This least-contributing channel is replaced with a new channel c' from $\mathcal{C}_{\text{avail}} \setminus \mathcal{C}$, where $\setminus$ represents set subtraction. The replacement channel is selected to maximize the updated score $D(\mathcal{C})$. The refined combinations are then merged with the retained ones to form an updated set $\mathcal{P}_{\text{new}}$, which undergoes further evaluation and refinement in subsequent iterations. This iterative process continues until a termination condition is met, such as a predefined number of iterations or the stabilization of $D(\mathcal{C})$. Finally, the channel combination $\mathcal{C}_{\text{opt}}$ with the highest score is selected as the optimal solution. By systematically evaluating and refining channel combinations, this algorithm strategically places channels near boundaries while maintaining sufficient spacing to mitigate potential signal overlap. This ensures that the selected channels provide maximum frequency diversity and minimal information redundancy.

2) Precise frequency-hopping control: Following the optimal channel selection, FHBE of VR-Fi progresses to the frequency hopping stage. Upon receiving a hopping command from the VR headset on the Rx side, the Tx (AP) initiates frequency hopping by sending a probe frame on the first optimal channel to synchronize the hopping with the Rx. The Rx responds immediately upon signal reception and samples the CSI of the current channel. Following the Rx's response, the Tx promptly transitions to the next optimal channel. VR-Fi cyclically hops through optimal channels, repeating this behavior and conducting channel sampling[4]. Subsequently, according to the duration of gestures, FHBE compiles the CSI samples from N_{s} cycles across N_{c} optimal channels into $N_{\text{s}} \times N_{\text{c}}$ CSI samples, which are utilized as training data for the subsequent MTPG-Net. Notably, each channel hopping occurs within several milliseconds, allowing the required samples to be collected with appropriate delay. If the Rx fails to respond within the allocated time, the Tx logs the failure and resends the signal on the current channel; if the failure persists, the Tx reverts to the initial optimal channel for a reset. Considering that the channels may suddenly become unavailable (possibly occupied by other devices), FHBE will, after several resets, rescan and re-execute the optimal channel selection algorithm to effectively restore frequency-hopping operations.

B. MTPG-Net with Multi-task Learning

After collecting optimal channel samples, achieving accurate gesture positioning and recognition remains a significant challenge. Traditional signal processing methods, inherently designed for continuous frequency signals within a single channel [37], are unsuitable for FHBE-enhanced CSI samples, which span a broad frequency range with discrete frequency diversity [30]. Even if processed independently, no existing approach effectively synthesizes such processed signals from irregularly distributed discrete channels into meaningful gestures. In addition, beyond common issues such as noise and channel fading, the collected samples also contain numerous random and unknown parameters. A critical factor is carrier phase offset (CPO), which may vary randomly with each channel hop [38]. For channel samples acquired through multiple samplings, the cumulative randomness can obscure the sensing information embedded in these samples [39].

Fortunately, the universal approximation theorem [40] allows us to employ a well-trained neural network to approximate such mapping functions, thus offering an effective tool for addressing the challenges posed by various frequency diversities and unknown parameters. In the context of VR scenarios, VR systems need to simultaneously capture the user's gestures and hand positions to ensure prompt and precise responses. This necessitates the implementation of a joint network output for both gesture positioning and recognition. Therefore, we implement MTPG-Net with Multi-task Learning scheme. This model consists of four main modules: spatial feature extraction, temporal modeling, expert weighting, and task-specific components, as shown in Fig. 6. MTPG-Net operates at the millisecond level during the inference phase, combined with the millisecond-level channel hopping time discussed in Sec. III-A2, thus significantly enhancing the real-time sensing capabilities of VR-Fi.

1) Spatial feature extraction module: This module, functioning as both a feature extractor and a trainable matching filter, is designed to mitigate hardware-related disturbances or biases within the extracted features. It leverages DenseNet [41]

[4]Notably, since conventional Wi-Fi sensing typically utilizes a single frame for sensing while also carrying data traffic [35], [36], VR-Fi ensures seamless operation by segmenting the same data payload across all hopping frames, preventing interruptions to default Wi-Fi communications.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3557561

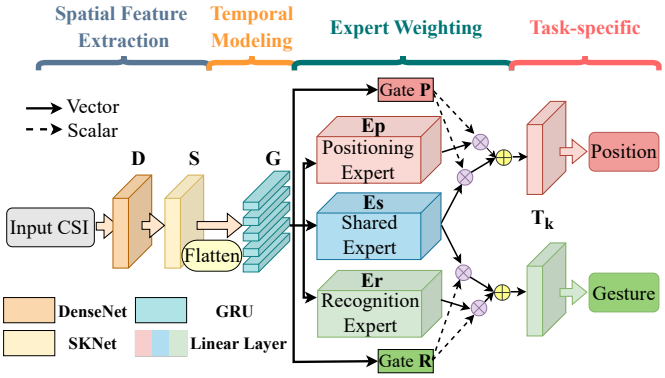


Fig. 6: MTPG-Net: One-stop model for joint gesture positioning and recognition with Multi-task Learning.

for its superior feature extraction capabilities and SKNet [42], noted for its attention-dense architecture, to serve as effective matching filters that minimize noise and eliminate interference. Additionally, the module combines these filtered results with the input data to concentrate key channel information for computational efficiency while preserving the availability of model gradients. To further mitigate variance drift during the feature extraction and filtering processes, batch normalization is specifically employed for regularization. Denoting DenseNet by $\mathbf{D}(\cdot)$, SKNet by $\mathbf{S}(\cdot)$, the spatial feature vector can be represented as:

$$\mathbf{x}^{\mathrm{sf}} = [\mathbf{x}, \mathbf{S}(\mathbf{D}(\mathbf{x})], \tag{3}$$

where x indicates the input of $N_{\mathrm{s}} \times N_{\mathrm{c}}$ CSI sequence.

2) Temporal modeling module: Beyond local spatial features, the input CSI sequences also encapsulate the temporal dynamics of gestures. Therefore, the temporal modeling module employs Gated Recurrent Units (GRUs) [43], which are adept at learning long-term dependencies with fewer parameters, thus optimizing training on limited datasets. This module is specifically designed to extract dynamic features of gestures across multiple N_{s} temporal cycles, enhancing the model's ability to discern temporal patterns within gesture data. Denoting GRU by $\mathbf{G}(\cdot)$, then the temporal feature vector can be shown as:

$$\mathbf{x}^{\mathrm{tf}} = \mathbf{G}(\mathsf{Flatten}(\mathbf{x}^{\mathrm{sf}})), \tag{4}$$

where Flatten refers to the process of reshaping the multi-dimensional vector after extracting the spatial feature into a one-dimensional vector.

3) Expert weighting module: This module adopts a multi-task learning scheme to enhance learning outcomes and generalization across the related tasks of gesture positioning and recognition. By concurrently training these tasks, it leverages shared model parameters and feature representations, facilitating the transfer of relevant information between tasks and reducing the risk of overfitting [44], [45]. Drawing from the Mixture-of-Experts (MoE) model [46], this module includes three specialized networks: Positioning Expert, Shared Expert, and Recognition Expert. Rather than using outputs directly for task prediction, an adaptive gating mechanism adjusts the outputs' combination and weighting, dynamically refining the representations for each task based on the input characteristics and task requirements, thus optimizing processing efficiency.

Specifically, the adaptive gating networks are linear transformations of the input of the module with a softmax layer:

$$\mathbf{S}_k = \mathrm{Softmax}(\mathbf{W}_k \mathbf{x}^{\mathrm{tf}}), \tag{5}$$

where $\mathbf{W}_k \in \mathbb{R}^{2\times d}$ is a trainable matrix, k represents the task index (1 for positioning, 2 for recognition), and d is the dimension of the feature $\mathbf{x}^{\mathrm{f}}$. The output of the gating mechanism is expressed as:

$$\mathbf{x}_k^{\mathrm{g}} = \mathrm{S}_k^1 \mathbf{E}_s(\mathbf{x}^{\mathrm{tf}}) + \mathrm{S}_k^2 \mathbf{E}_t(\mathbf{x}^{\mathrm{tf}}), \tag{6}$$

where $\mathbf{E}_t$ is the respective task-specific experts ($\mathbf{E}_p$ for $k = 1$ and $\mathbf{E}_r$ for $k = 2$). By adaptively assigning weights to different experts, this module allows each expert to focus on learning distinct knowledge efficiently without interference.

4) Task-specific module: Finally, the task-specific module processes the gated, weighted outputs from various experts through multiple fully connected layers to generate the final predictions for their respective tasks. The prediction for task k is formulated as:

$$\mathbf{y}_k = \mathbf{T}_k \mathbf{x}_k^{\mathrm{g}}, \tag{7}$$

where $\mathbf{T}_k$ denotes the task-specific network of task k.

5) Joint loss optimization: During the multi-task learning training phase, we utilize a joint loss function for back-propagation to update the model parameters. We denote the parameters of the models in the Spatial Feature Extraction, Temporal Modeling, and Expert Weighting modules as θ_{s}. θ_k is task-specific parameters of task k. The training procedure is summarized as follows:

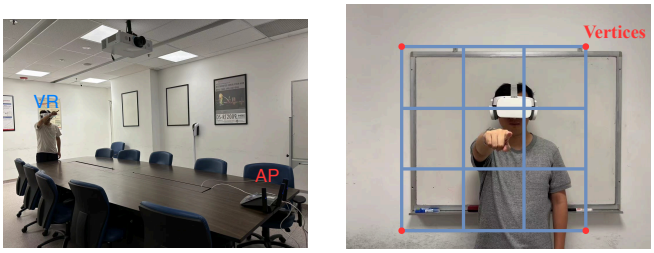
$$(\hat{\theta}_{\mathrm{s}}, \hat{\theta}_1, \ldots, \hat{\theta}_{\mathrm{K}}) = \arg\min_{\theta_{\mathrm{s}}, \theta_1, \ldots, \theta_{\mathrm{k}}} \sum_{k=1}^{K} \omega_k \mathcal{L}(\theta_{\mathrm{s}}, \theta_{\mathrm{k}}), \tag{8}$$

where ω_k is the task-specific parameter of task k and $\mathcal{L}(\theta_{\mathrm{s}}, \theta_{\mathrm{k}})$ is the cross-entropy losses between $\mathbf{y}_k$ and ground truth.

IV. Prototype and Experiment Setup

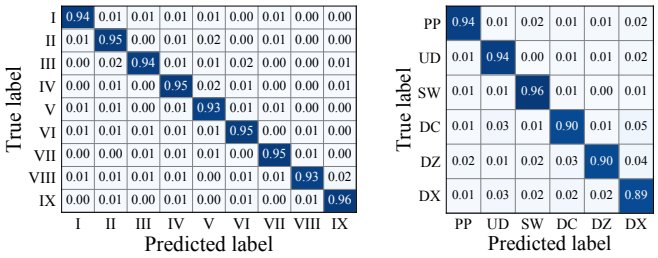
In this section, we provide a detailed introduction to the implementation of the VR-Fi prototype, explaining its key components and deployment configuration. Additionally, we outline the experimental setup, including the environments, participants, and specific procedures designed to evaluate VR-Fi's performance comprehensively. To ensure rigorous analysis, we introduce two state-of-the-art baseline methods in Wi-Fi sensing as reference points to benchmark VR-Fi's gesture recognition and positioning capabilities.

1) Prototype of VR-Fi: VR-Fi is deployed across a VR headset and an AP, each equipped with an Intel AX210 NIC [47]. VR headset, functions as the Rx, while the AP serves as the Tx. This Wi-Fi NIC has two antennas and supports all channels mentioned in Section III-A1. VR-Fi utilizes a single transmission and dual reception configuration, where only one transmitting antenna is used at AP, reducing dependency on physical hardware and minimizing costs. In the optimal channel selection algorithm, we set $m = 30$, $k = 10$, and the number of iterations to 200. We utilize the PicoScenes platform [48], and a custom plugin to implement the FHBE technique, to collect FHBE-enhanced channel samples. We set



(a) Training scene in the MR. (b) 3×3 grid in the SR.

Fig. 7: Experiment setting examples: (a) training environment (b) a layout of 9 gesture areas.



(a) Gesture positioning. (b) Gesture recognition.

Fig. 8: Overall performance of VR-Fi.

$N_c = 4$ and $N_s = 15$ according to the frequency-hopping time and gesture duration. These CSI channel samples are then parsed into their real and imaginary parts and concatenated as input to MTPG-Net. MTPG-Net is implemented within the PyTorch 1.7.1 environment and contains about 4.3 million model parameters, which is compatible with common commercial mobile processors in VR-headset [49].

2) Experiment Setup: We recruit 8 subjects, consisting of five males and three females, to conduct experiments across six different environments: meeting room (MR), auditorium (AD), classroom (CR), library (LB), office (OF), and study room (SR). For training purposes, we utilize data collected exclusively in the MR, and conduct tests in the remaining environments. The subject wears the VR headset, and the AP is placed on the opposite side of the room, as depicted in Fig. 7(a). We define the 3×3 grid of areas based on the subject's shoulder joint as the center. When the arm is raised, the four outermost vertices are positioned at $\pm 45°$ in the azimuth direction (left and right) and simultaneously at $\pm 45°$ in the elevation direction (upward and downward) respectively. The distance between two outermost points is equally divided into three segments, resulting in nine regions as follows: top-left (i), top-center (ii), top-right (iii), middle-left (iv), center (v), middle-right (vi), bottom-left (vii), bottom-center (viii), and bottom-right (ix), as illustrated in Fig. 7(b). Subjects perform six distinct gestures, including push-pull (PP), up-down (UD), sweeping (SW), drawing a circle (DC), drawing a zig-zag (DZ), and drawing a cross (DX), 300 times with both the left and right hand at each of the nine designated areas. These experiments have strictly followed our IRB.

3) Baseline: Given the lack of existing Wi-Fi sensing research that realizes gesture positioning and recognition concurrently, we refine two types of gesture recognition methods as baselines: the BVP-based method and the CSI-based method. The representative and influential frameworks for these methods are Widar3.0 [17] and WiGesID [50] respectively. Specifically, Widar3.0 extracts a domain-independent feature, known as body coordinate velocity (BVP), from the Doppler frequency shift (DFS) spectrum of raw CSI measurements for gesture recognition. Concurrently, we adopt the gesture feature extraction model from WiGesID as our baseline for another choice to process raw CSI data directly. We collect the 20MHz CSI data as utilized in both Widar3.0 and WiGesID, adhering to the experimental setup outlined in Section IV-2. For gesture positioning, we modify the final fully connected layer of these models to output gesture positions. We strictly adhere to the model processes established in each framework to train separately for gesture recognition and positioning, thereby obtaining distinct results for each task.

V. Evaluation

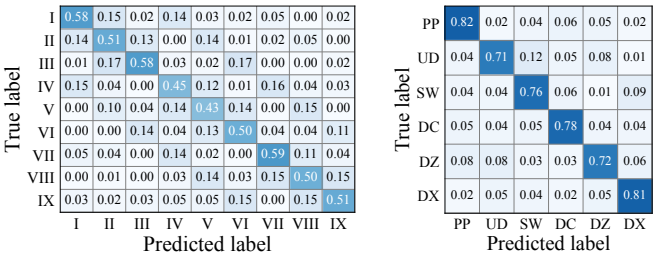
In this section, we conduct a comprehensive evaluation of VR-Fi's capabilities, focusing primarily on its essential functions of joint gesture positioning and gesture recognition. We start with an overall investigation into VR-Fi's performance for the two tasks. Subsequently, VR-Fi is compared with two baselines to highlight its superior gesture positioning capability. Furthermore, we examine the influence of various practical factors on VR-Fi's performance. An ablation study is then conducted to analyze the contributions of key components, to the overall system performance. Finally, we extend VR-Fi to support simultaneous gesture positioning and recognition, with an in-depth assessment of its recognition accuracy across 10 gestures performed in different directions.

A. Overall Performance

In this section, we present the overall accuracy of VR-Fi, detailing both gesture positioning and gesture recognition accuracies. Fig. 8(a) illustrates the confusion matrix for VR-Fi identifying different gesture positions, showing an average recognition accuracy of 94.47%. Moreover, VR-Fi maintains a consistently high accuracy, exceeding 93% across all positions. Notably, the fifth position—corresponding to the center (v)—exhibits slightly lower accuracy. This reduction in accuracy is intuitive, as each position tends to be most frequently confused with its adjacent positions, and the fifth position is centrally located. Despite this inherent confusion, the high average accuracy and minimal variance demonstrate VR-Fi's exceptional capability in gesture positioning.

Subsequently, Fig. 8(b) presents the confusion matrix for VR-Fi in recognizing six types of gestures, indicating an average recognition accuracy of 92.13% with consistently high accuracies above 89%. It is observed that gestures such as "drawing a circle", "drawing a zig-zag", and "drawing a cross" exhibit slightly lower accuracies. This disparity is mainly attributable to the complex nature of these gestures, which involve movements in both horizontal and vertical two directions, contrasting with the other three simpler unidirectional gestures. Despite some fluctuations, VR-Fi effectively recognizes a diverse range of gestures.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3557561



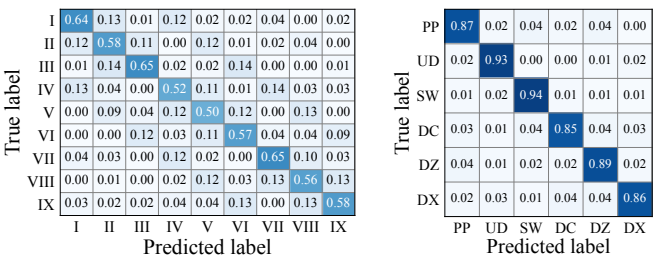
(a) Gesture positioning. (b) Gesture recognition.

Fig. 9: Performance of BVP-based method.

In contrast, the performance of the two baselines of BVP-based and CSI-based in gesture positioning is unsatisfactory, as shown in Fig. 9(a) and Fig. 10(a), with accuracies of only 51.76% and 58.31%, respectively, and displaying significant variance. The superior performance of VR-Fi compared to the baselines can be attributed to the implementation of our unique FHBE technology, which substantially increases the Wi-Fi sensing bandwidth. This enhancement provides more frequency diversity of range and bearing on gesture positions. Moreover, the higher positioning accuracy of the CSI-based method over the BVP-based method stems from the fact that when BVP is extracted from raw CSI, essential domain features, including gesture position features, are abstracted away, making it more difficult to identify gesture positions.

Next, Fig. 9(b) and Fig. 10(b) illustrate the accuracies of the two baselines in gesture recognition. The BVP-based method achieves an accuracy of 76.68%, exhibiting relatively large variance among different gestures. Notably, the performance is considerably lower than that previously reported in Widar3.0 [17]. This discrepancy is largely attributed to the VR-Fi setup, which is deployed on a single Wi-Fi communication link to meet VR scenarios and significantly reduce hardware dependencies, whereas Widar3.0 requires at least three communication links to achieve its reported performance.

In contrast, the CSI-based method shows a gesture recognition accuracy of 89.12%, with higher consistency across different gestures. The discrepancy between the two baselines is due to the BVP extraction process, which not only removes position information but also inadvertently reduces gesture features. Moreover, compared to the CSI-based baseline, VR-Fi shows a certain degree of improvement in gesture recognition. This enhancement is primarily due to VR-Fi's implementation of FHBE technology, which expands the sensing bandwidth and thereby increases the frequency diversity available for gesture recognition. Furthermore, VR-Fi's MTPG-Net employs multi-task learning scheme, effectively leveraging the potential connections between the two tasks to enhance the accuracy of both gesture positioning and recognition.



(a) Gesture positioning. (b) Gesture recognition.

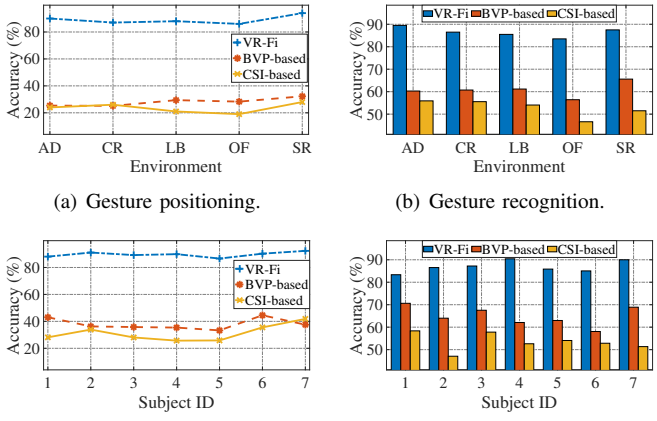
Fig. 10: Performance of CSI-based method.

B. Impact Factors

1) Environments and Subjects: To demonstrate the cross-environment and cross-subject generalizability of VR-Fi, we assess its gesture positioning and recognition accuracy on unseen subjects across five different unseen environments. Notably, in each environment, we apply the optimal channel selection algorithm separately to identify the optimal channels. As illustrated in Fig. 11, the results reveal that the average accuracies for gesture positioning and recognition with VR-Fi are 89.05% and 86.37%, respectively, across various environments, and 89.33% and 86.94% across different subjects. Concurrently, these figures highlight high consistency in performance, adequately demonstrating the robust generalizability of VR-Fi across diverse environments and subjects.

This broad generalizability is partly attributable to the enhanced resolution stemming from expanded bandwidth, which more effectively distinguishes gesture-induced paths from the background. Simultanously, the predominance of near-field channel variations, induced by gestures within the proximity of the VR headset, significantly mitigates the impact of distant interferences [23]. Additionally, we observe that while the training and testing channels may differ across environments, the impact on accuracy remains minimal. This is due to the optimal channel selection algorithm, which ensures a balance between proximity to boundaries and inter-channel isolation. As a result, the CSI characteristics of training and testing channels remain relatively similar, minimizing discrepancies and preserving positioning and recognition accuracy.

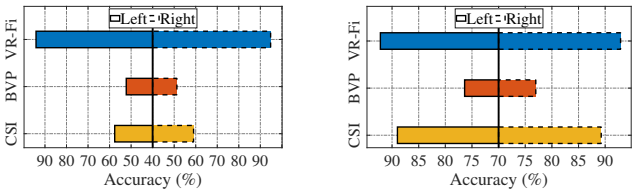
In contrast, the two baselines all display gesture positioning accuracies below 43% in varying environments and subjects, as depicted in Fig. 11(a) and Fig. 11(c), rendering them nearly ineffective for cross-environment and cross-subject applications. Furthermore, we evaluate the original gesture recognition capabilities of the baselines under diverse conditions of environments and subjects. Fig. 11(b) and Fig. 11(d) show



(a) Gesture positioning. (b) Gesture recognition.
(c) Gesture positioning. (d) Gesture recognition.

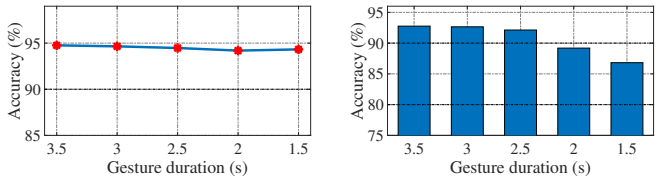
Fig. 11: Impacts of environment and subject.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3557561



(a) Gesture positioning. (b) Gesture recognition.

Fig. 12: Impacts of usage of the left or right hand.



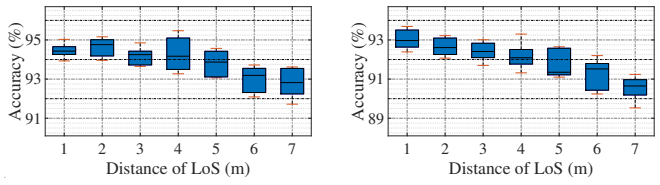
(a) Gesture positioning. (b) Gesture recognition.

Fig. 14: Impacts of gesture duration.

that the BVP-based method achieves an accuracy near 60%, while the CSI-based method's accuracy shows even lower, both of which are unsuitable for cross-domain applications. The poor performance of the former is attributed to the limited communication links in VR settings, while the latter suffers due to its model lacking generalization capability.

2) Left or right hand: We further evaluated the adaptability of VR-Fi in accommodating left-hand versus right-hand usage. Fig. 15 illustrates that the average accuracies for gesture positioning with VR-Fi are 94.15% for the left hand and 94.78% for the right hand. The average accuracies for gesture recognition are 91.94% for the left hand and 92.33% for the right hand. The two baselines similarly exhibit comparable accuracy performances, as previously mentioned in Sec. V-A. These findings confirm that VR-Fi effectively accommodates either hand for gesture positioning and recognition, significantly enhancing the flexible control and immersive experience for users in VR environments. Given the consistently poor performance of the baseline, we will exclude it from subsequent evaluations for comparison.

3) Distance of LoS: We further evaluate the performance of VR-Fi when the user's location varies at different distances from the AP (i.e., distances of Line-of-Sight (LoS)). In this evaluation, subjects are asked to stand at five different locations without fixed objects (e.g., tables or chairs) at distances ranging from 1 m to 10 m from the AP. The results, as shown in Fig. 13(a) and Fig. 13(b), indicate that under consistent LoS conditions, the accuracy of gesture positioning and recognition remains relatively stable, exhibiting only minor fluctuations of less than 2% across different locations. Simultaneously, as the LoS distance increases, the performance of VR-Fi shows a slight decline due to greater signal attenuation and increased interference over longer propagation paths. Despite this, even at the upper limit of common indoor distances (10 m), VR-Fi continues to deliver outstanding performance, maintaining gesture localization and recognition accuracy close to 90%. These findings demonstrate VR-Fi's ability to adapt effectively to users at varying locations within an indoor environment.



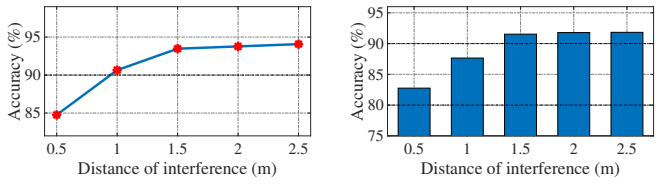
(a) Gesture positioning. (b) Gesture recognition.

Fig. 13: Impacts of distance of LoS.

4) Gesture duration: We next investigate the impact of varying gesture durations (i.e., gesture speeds) on the performance of VR-Fi. Fig. 14(a) and Fig. 14(b) illustrate VR-Fi's gesture positioning and recognition accuracy, respectively, for durations ranging from 1.5 s to 3.5 s. As observed, gesture positioning accuracy exhibits only a slight decrease as gesture speed increases. This minor decline can be attributed to the reduced sampling data caused by faster gestures, which introduces more uncertainty. On the other hand, gesture recognition accuracy experiences a more noticeable drop with faster gestures. This is because gesture recognition relies heavily on detecting dynamic changes in motion, and faster gestures are more likely to cause confusion. Nonetheless, it is worth noting that even at the fastest gesture speed of 1.5 s—representing the upper limit for most users—VR-Fi achieves an accuracy exceeding 86%. These findings clearly demonstrate VR-Fi's robust adaptability to varying gesture speeds, ensuring reliable performance across users' changing interaction habits.

5) Distance of Interference: We next investigate the impact of interference at varying distances on the performance of VR-Fi. Specifically, A fan measuring 40 cm $\times$ 40 cm and 1.2 m in height is used as the interference source. The distance between the fan and the subject is varied from 0.5 m to 2.5 m. The results for gesture positioning and recognition accuracy are presented in Fig.15(a) and Fig.15(b), respectively. The results indicate that as the interference source moves closer, both gesture positioning and recognition accuracy experience a certain degree of decline. However, even at the closest interference distance of 0.5 m, VR-Fi maintains accuracy levels exceeding 84.7% for gesture positioning and 82.1% for gesture recognition. This robustness is primarily attributed to VR-Fi's enhanced resolution, derived from its expanded bandwidth, which allows it to effectively differentiate between gesture-induced paths and interference. These findings underscore VR-Fi's resilience to interference, even in challenging scenarios with proximal interference sources.

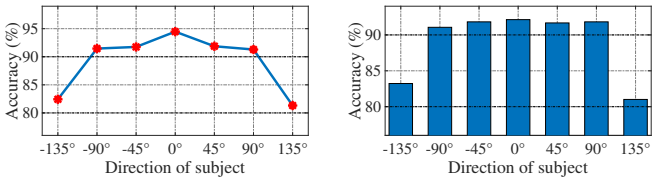
6) Direction of subject: We next investigate the impact of the subject's orientation on the performance of VR-Fi. Specifically, participants were asked to rotate from directly



(a) Gesture positioning. (b) Gesture recognition.

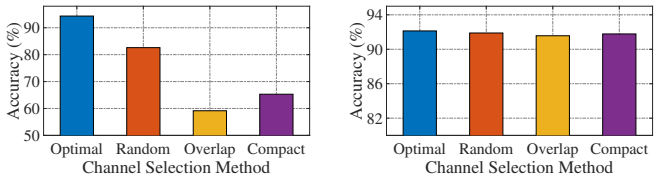
Fig. 15: Impacts of distance of interference.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3557561



(a) Gesture positioning. (b) Gesture recognition.

Fig. 16: Impacts of direction of subject.



(a) Gesture positioning. (b) Gesture recognition.

Fig. 18: Impacts of channel selection.

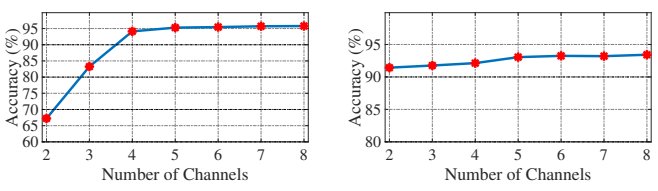
facing the AP to the left and right at angles of 45°, 90°, and 135°. The results for gesture positioning and recognition accuracy are presented in Fig. 16(a) and Fig. 16(b), where "-" denotes left rotation and "+" denotes right rotation. In Fig. 16(a), slight variations in gesture positioning accuracy are observed at angles of 45° and 90° compared to 0°. This can be attributed to the reduced differences in the AoA of gestures at these orientations, which makes it more challenging to distinguish between different gesture positions compared to the directly facing AP condition. Conversely, in Fig. 16(b), the same 45° and 90° rotations have minimal impact on gesture recognition accuracy, indicating that gesture recognition is less sensitive to moderate angular changes. However, at an angle of 135°, both gesture positioning and recognition accuracy experience a noticeable decline. This degradation is due to the subject facing away from the AP, where the body obstructs the gestures, leading to significant attenuation of the LoS sensing signal strength. These findings emphasize that while VR-Fi maintains robust performance under moderate angular deviations, its accuracy diminishes when substantial signal blockage occurs due to the subject's orientation.

C. Ablation study

1) Number of Channels: We initially conduct a detailed evaluation of the impact of varying channel numbers on VR-Fi, using **Algorithm 1** of optimal channel selection algorithm to select between 2 to 8 available channels. The results, illustrated in Fig. 17, indicate that as the number of channels increases, there is a notable improvement in the gesture positioning accuracy of VR-Fi. However, the accuracy improvement becomes less pronounced after four channels, while the gesture recognition accuracy only exhibits a marginal increase consistently. Considering the concern to minimize channel occupation and thereby avoid interference with the normal use of Wi-Fi communication in VR environments, we opt for a configuration of four channels.

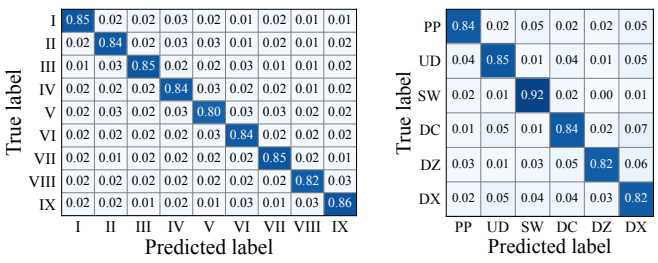
2) Channel Selection: Next, we comprehensively assess the significant contribution of our optimal channel selection algorithm to VR-Fi. For comparison, we selected three control groups: random channel selection from all available channels, overlap channel selection from continuous channels in the 2.4 GHz band, and compact channel selection from continuous channels in the 5 GHz band, as illustrated in Fig. 18. Fig. 18(a) shows that while VR-Fi with the optimal channel selection algorithm sustains high accuracy in gesture positioning, all other three control groups experienced notable reductions in accuracy. Specifically, the accuracy rates are 82.85% for random channel selection, 59.73% for overlap channel selection, and 65.29% for compact channel selection. These declines are attributed to the insufficient frequency diversity provided in these selections, which largely constrained the actual expansion capabilities of the bandwidth. Fig. 18(b) shows the impact of four channel selection methods on gesture recognition. The results show that the optimal channel selection algorithm has a slight advantage because the expansion of the frequency band is mainly to improve the positioning resolution. The results clearly demonstrate that the optimal channel selection algorithm can effectively enhance VR-Fi's sensing bandwidth, providing maximum frequency diversity for gesture positioning and recognition.

3) MTPG-Net network: We further assess the critical contribution of the MTPG-Net of VR-Fi to joint gesture positioning and gesture recognition. Specifically, we evaluate the expert weighting module, a pivotal component of our multi-task learning scheme. For comparative analysis, we append multiple fully connected layers to the gesture feature vectors extracted by the temporal modeling module and directly produce an output that concatenates gesture positioning and recognition label dimensions. Fig. 19 displays the confusion matrices for gesture positioning and recognition with the modified MTPG-Net of VR-Fi, recording average accuracies of 83.81% and 84.80%, respectively. These figures are notably lower than those achieved by the complete VR-Fi with MTPG-Net. This decrease in performance results from the modified model treating gesture positioning and recognition together as



(a) Gesture positioning. (b) Gesture recognition.

Fig. 17: Impacts of the number of channels.



(a) Gesture positioning. (b) Gesture recognition.

Fig. 19: Performance of VR-Fi without MTPG-Net.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2025.3557561

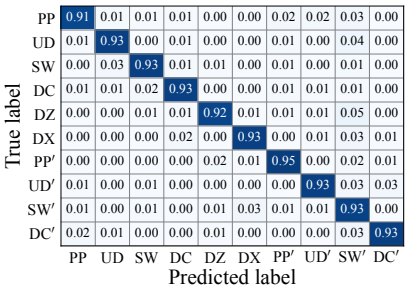


Fig. 20: Extended gesture recognition.

a whole. In this setup, the model attempts to address both two tasks simultaneously but struggles to prioritize or distinguish between the tasks of gesture positioning and recognition. In contrast, the MTPG-Net with multi-task learning allocates distinct expert network weights to different tasks during training, effectively capturing and utilizing latent interconnections between tasks, while minimizing mutual task interference. The ablation study of the model clearly demonstrates the efficacy of MTPG-Net as an integrated solution for joint gesture positioning and gesture recognition.

D. Extension of VR-Fi

Considering VR-Fi's application scenarios, we further explore its ability to simultaneously position and recognize gestures in VR environments. The results show that VR-Fi achieved an impressive accuracy of 89.47%. Note that this result is not a direct mathematical product of gesture positioning and recognition accuracy, as there is an overlap in error data samples between the two metrics. Meanwhile, the performance of the two baselines of BVP-based and CSI-based in gesture positioning is unsatisfactory, with accuracies of only 45.33% and 55.58%, respectively. These outcomes further highlight VR-Fi's exceptional performance in addressing both gesture positioning and recognition tasks within VR scenarios.

In addition, to further extend VR-Fi's ability to differentiate gestures of different directions, we include four intuitive reverse gestures in our experiments, derived from the original six gestures. Specifically, the gesture "PP", originally "push-then-pull", is expanded to "pull-then-push" (PP′). Similarly, "UD", initially "Up-then-Down", is extended to "Down-then-Up" (UD′); "SW", which is originally "right-then-left", is extended to "left-then-right" (SW′); and "DC", originally drawing a circle clockwise, is extended to drawing a circle counterclockwise (DC′). The gesture recognition results of VR-Fi are shown in Fig. 20, indicating an average accuracy rate of 92.45% across these 10 gestures. We also observe that VR-Fi exhibits excellent differentiation capability for gestures of the same category but in opposite directions. This outcome closely matches the results for the six gestures shown in Fig. 8(b), highlighting VR-Fi's ability to recognize various gestures, including those with directional variations.

VI. Related Work and Discussion

In this section, we introduce existing gesture recognition proposals related to VR-Fi, categorizing them into two main types: Wi-Fi-based methods and non-Wi-Fi-based methods. Additionally, we incorporate a discussion of VR-Fi for its comparison with other solutions and future works.

a) Wi-Fi-based Gesture Recognition: Early attempts tried to use histograms of Wi-Fi signal amplitudes [51] and Received Signal Strength Indicator (RSSI) [52] for gesture recognition. In contrast, CSI, measured across multiple subcarriers, offers a greater diversity of information, rendering it more suitable for activity recognition. For instance, WiG [53] utilized features derived from CSI amplitude variations to train a Support Vector Machine (SVM) classifier, enabling the recognition of four common gestures: right, left, push, and pull. Similarly, WiFinger [54] employed principal component analysis (PCA) to extract CSI amplitude patterns and used dynamic time warping (DTW) alongside k-nearest neighbor (KNN) algorithms to compare waveform shapes, facilitating the recognition of nine distinct gestures. However, most current Wi-Fi gesture recognition research typically lacks the capability to position gestures. Some studies crudely consider the gesture's position and the person's location as factors affecting recognition accuracy, often referring to these as "domain" factors. These are typically only eliminated rather than explicitly recognized, in contrast to VR-Fi which aims to identify and utilize position factors.

WiAG [16] developed a transformation function to generate virtual samples that capture the positional relationship of a person's hand relative to the Wi-Fi transmitter and receiver, thereby enabling cross-domain gesture recognition. To circumvent the need for additional training in new domains, Widar3.0 [17] extracts the Doppler Frequency Shift (DFS) on at least three links to derive domain-independent features and has developed a deep learning model suitable for cross-domain gesture recognition. Unlike manually designed domain-independent features, [55] employed adversarial learning to construct a domain-independent feature space. Recent research has expanded beyond using CSI as the sole data source for gesture recognition; WiKI-Eve [18] and MuKI-Fi [56] utilized beamforming feedback information—a compressed digital version of CSI—to recognize finger typing motions.

b) Non-Wi-Fi-based Gesture Recognition: For wearable sensors, [57] achieved gesture recognition of sign language using a data glove. Smartwatches [58] and wearable rings [59] enabled text input recognition through hand movements. These methods all required users to wear additional physical sensors. Pioneering work by [60] established a foundational framework for gesture recognition using dedicated cameras. Advances in imaging technology enabled the use of depth and infrared cameras [61]. Zhang et al. [62] further advanced the field by employing 3D convolutional neural networks (3DCNN) and bidirectional convolutional long short-term memory (ConvLSTM) networks to encode both global temporal and local spatial information, enabling the extraction of sophisticated spatiotemporal features for gesture recognition.

Radar-based gesture recognition systems offer fine-grained sensing capabilities due to their large bandwidth but require specialized and often expensive hardware. Hazra et al. [63] utilized a compact 60-GHz millimeter-wave (mmWave) radar sensor to extract and process a sequence of range-Doppler

images, employing a long-recurrent fully convolutional neural network (L-RFCN) for real-time dynamic gesture recognition. Similarly, Ahmed et al. [64] used a Frequency-Modulated Continuous Wave (FMCW) radar with two receiving channels to design a novel three-stream convolutional neural network (CNN). Their approach incorporated range-time, Doppler-time, and angle-time spectrograms as inputs, fusing these features in the later stages for gesture recognition. OCHID-Fi [15], on the other hand, employed wideband RF sensors integrated into smart devices to detect 3D human hand poses. This system used a cross-modality and cross-domain training process to extract skeletal structures, even in the presence of obstacles.

c) Discussion: For non-Wi-Fi-based solutions, while wearable sensors require additional contact devices, vision-based solutions face issues such as being resource-intensive, privacy concerns, and requirements for specific lighting conditions, as discussed in Sec. I. Although radar-based solutions provide ample bandwidth for fine-grained ranging, integrating radar chips into VR headsets presents practical challenges. These challenges include hardware compatibility, power consumption limitations, and cost considerations associated with the widespread adoption of radar chips in consumer-grade VR devices. In contrast, Wi-Fi is already embedded in VR headsets, making it a more accessible and cost-effective sensing solution for VR-Fi. Considering that the AX210 NIC is the only contemporary hardware capable of extracting CSI in the 802.11ax format within 6 GHz channels, VR-Fi has not yet tested its cross-hardware capabilities. We will explore this further as Wi-Fi hardware evolves in the future.

VII. Conclusion

We have developed VR-Fi as a pioneering VR-embedded Wi-Fi sensing system, designed to implement joint gesture positioning and recognition within a VR environment. VR-Fi initially incorporates the FHBE technique to facilitate bandwidth expansion, thereby capturing greater frequency diversity of CSI. This technique is particularly beneficial for commercial Wi-Fi NICs equipped with a limited number of antennas to enhance spatial resolution. Given the lack of suitable signal processing tools for handling FHBE-enhanced channel samples with unknown interference, we have innovated in the MTPG-Net model. Simultaneously, this model employs multi-task learning to achieve joint gesture positioning and recognition by utilizing an adaptive gate scheme to dynamically assign weights to different task experts. Through extensive experiments with our VR-Fi prototype, we have demonstrated its capability to accurately position and recognize user gestures in various indoor VR scenarios.

References

[1] L. Yang, J. Huang, T. Feng, W. Hong-An, and D. Guo-Zhong, "Gesture Interaction in Virtual Reality," *Virtual Reality & Intelligent Hardware*, vol. 1, no. 1, pp. 84–112, 2019.

[2] M. Lee, R. Green, and M. Billinghurst, "3D Natural Hand Interaction for AR Applications," in *Proc. of 23rd IVCNZ*. IEEE, 2008, pp. 1–6.

[3] Y. Shen, S.-K. Ong, and A. Y. Nee, "Vision-Based Hand Interaction in Augmented Reality Environment," *Intl. Journal of Human–Computer Interaction*, vol. 27, no. 6, pp. 523–544, 2011.

[4] A. Vaitkevičius, M. Taroza, T. Blažauskas, R. Damaševičius, R. Maskeliūnas, and M. Woźniak, "Recognition of American Sign Language Gestures in a Virtual Reality Using Leap Motion," *Applied Sciences*, vol. 9, no. 3, p. 445, 2019.

[5] J. J. LaViola Jr, "Context aware 3d gesture recognition for games and virtual reality," in *ACM SIGGRAPH 2015 Courses*, 2015, pp. 1–61.

[6] S. Poularakis and I. Katsavounidis, "Low-Complexity Hand Gesture Recognition System for Continuous Streams of Digits and Letters," *IEEE Transactions on Cybernetics*, vol. 46, no. 9, pp. 2094–2108, 2015.

[7] "Are Security Cameras Legal?" https://www.security.org/security-cameras/legality/, accessed: 2022-05-19.

[8] "Security Camera Laws, Rights, and Rules," https://www.safewise.com/security-camera-laws/, accessed: 2022-05-19.

[9] T. Arici, S. Dikbas, and Y. Altunbasak, "A Histogram Modification Framework and Its Application for Image Contrast Enhancement," *IEEE Transactions on Image Processing*, pp. 1921–1935, 2009.

[10] K. Xu, X. Yang, B. Yin, and R. W. Lau, "Learning to Restore Low-Light Images via Decomposition-and-Enhancement," in *Proc. of the IEEE/CVF CVPR*, June 2020.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A Large-scale Hierarchical Image Database," in *Proc. of the IEEE/CVF CVPR*, 2009, pp. 248–255.

[12] A. Kosaka, A. Saito, Y. Furuhashi, and T. Shibasaki, "Augmented Reality System for Surgical Navigation using Robust Target Vision," in *Proc. of the IEEE/CVF CVPR*, vol. 2, 2000, pp. 187–194.

[13] X. Cai, J. Ma, W. Liu, H. Han, and L. Ma, "Efficient Convolutional Neural Network for FMCW Radar based Hand Gesture Recognition," in *Adjunct Proc. of the UbiComp/ISWC*, 2019, pp. 17–20.

[14] Z. Chen, C. Cai, T. Zheng, J. Luo, J. Xiong, and X. Wang, "RF-based Human Activity Recognition using Signal Adapted Convolutional Neural Network," *IEEE Transactions on Mobile Computing*, 2021.

[15] S. Zhang, T. Zheng, Z. Chen, J. Hu, A. Khamis, J. Liu, and J. Luo, "Ochid-Fi: Occlusion-Robust Hand Pose Estimation in 3D via RF-Vision," in *Proceedings of the IEEE/CVF ICCV*, 2023, pp. 15 112–15 121.

[16] A. Virmani and M. Shahzad, "Position and Orientation Agnostic Gesture Recognition Using WiFi," in *Proc. of the 15th ACM MobiSys*.

[17] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort Cross-domain Gesture Recognition with Wi-Fi," in *Proc. of the 17th ACM MobiSys*, 2019, pp. 313–325.

[18] J. Hu, H. Wang, T. Zheng, J. Hu, Z. Chen, H. Jiang, and J. Luo, "Password-Stealing without Hacking: Wi-Fi Enabled Practical Keystroke Eavesdropping," in *Proc. of the 2023 ACM CCS*, 2023, pp. 239–252.

[19] R. Xiao, J. Liu, J. Han, and K. Ren, "Onefi: One-Shot Recognition for Unseen Gesture via COTS WiFi," in *Proc. of the 19th ACM Sensys*, 2021, pp. 206–219.

[20] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, "Widar2.0: Passive Human Tracking with a Single Wi-Fi Link," in *Proc. of the 16th ACM MobiSys*, 2018, p. 350–361.

[21] H. Emsley, "Irregular Astigmatism of the Eye: Effect of Correcting Lenses," *Transactions of the Optical Society*, vol. 27, no. 1, p. 28, 1925.

[22] S. Appelle, "Perception and Discrimination as a Function of Stimulus Orientation: The" Oblique Effect" in Man and Animals." *Psychological bulletin*, vol. 78, no. 4, p. 266, 1972.

[23] J. Hu, T. Zheng, Z. Chen, H. Wang, and J. Luo, "MUSE-Fi: Contactless MUti-person SEnsing Exploiting Near-field Wi-Fi Channel Variation," in *Proc. of the 29th ACM MobiCom*, 2023, pp. 1–15.

[24] R. H. Venkatnarayan, G. Page, and M. Shahzad, "Multi-user gesture recognition using wifi," in *Proc. of the 16th ACM MobiSys*, 2018, pp. 401–413.

[25] J. Liu, C. Xiao, K. Cui, J. Han, X. Xu, and K. Ren, "Behavior Privacy Preserving in RF Sensing," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 1, pp. 784–796, 2022.

[26] Z. Chen, Z. Li, Z. Xu, G. Z. Zhu, Y. Xu, J. Xiong, and X. Wang, "AWL: Turning Spatial Aliasing From Foe to Friend for Accurate WiFi Localization," in *Proc. of the 13th ACM CoNEXT*, 2017, pp. 238–250.

[27] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3D Tracking via Body Radio Reflections," in *Proc. of the 11th USENIX NSDI*, 2014, pp. 317–329.

[28] Y. Xie, Z. Li, and M. Li, "Precise Power Delay Profiling with Commodity Wi-Fi," in *Proc. of the 21st ACM MobiCom*, 2015, pp. 53–64.

[29] J. Xiong, K. Sundaresan, and K. Jamieson, "ToneTrack: Leveraging Frequency-Agile Radios for Time-Based Indoor Wireless Localization," in *Proc. of the 21st ACM MobiCom*, 2015, pp. 537–549.

[30] X. Li, H. Wang, Z. Chen, Z. Jiang, and J. Luo, "UWB-Fi: Pushing Wi-Fi towards Ultra-wideband for Fine-Granularity Sensing," in *Proc. of the 22nd ACM MobiSys*, 2024, pp. 42–55.

[31] J. Wright and Y. Ma, "Dense Error Correction Via ℓ^1-Minimization," *IEEE Trans. on Information Theory*, vol. 56, no. 7, pp. 3540–3560, 2010.

[32] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge University Press, 2005.

[33] D. Vasisht, S. Kumar, and D. Katabi, "Decimeter-Level Localization with a Single WiFi Access Point," in *Proc. of the 13th USENIX NSDI*, 2016, p. 165–178.

[34] X. Li, H. Wang, J. Hu, Z. Chen, Z. Jiang, and J. Luo, "CCS-Fi: Widening Wi-Fi Sensing Bandwidth via Compressive Channel Sampling," in *Proc. of the 44th IEEE INFOCOM*, 05 2025.

[35] Y. He, G. Yu, Y. Cai, and H. Luo, "Integrated sensing, computation, and communication: System framework and performance optimization," *IEEE Trans. Wireless Commun.*, vol. 23, no. 2, pp. 1114–1128, Feb. 2024.

[36] Y. He, J. Liu, M. Li, G. Yu, J. Han, and K. Ren, "SenCom: Integrated Sensing and Communication with Practical WiFi," in *Proc. of the 29th ACM MobiCom*, 2023, pp. 1–16.

[37] R. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[38] Y. Ma, G. Zhou, and S. Wang, "WIFI Sensing With Channel State Information: A Survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.

[39] Kotaru, Manikanta and Joshi, Kiran and Bharadia, Dinesh and Katti, Sachin, "SpotFi: Decimeter Level Localization Using WiFi," in *Proc. of 29th ACM SIGCOMM*, 2015, p. 269–282.

[40] D.-X. Zhou, "Universality of Deep Convolutional Neural Networks," *Applied and Computational Harmonic Analysis*, vol. 48, no. 2, pp. 787–794, 2020.

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proc. of the 30th IEEE/CVF CVPR*, 2017, pp. 4700–4708.

[42] X. Li, W. Wang, X. Hu, and J. Yang, "Selective Kernel Networks," in *Proc. of the 32nd IEEE/CVF CVPR*, 2019, pp. 510–519.

[43] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[44] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "Modeling Task Relationships in Multi-Task Learning With Multi-Gate Mixture-Of-Experts," in *Proc. of the 24th ACM KDD*, 2018, pp. 1930–1939.

[45] H. Tang, J. Liu, M. Zhao, and X. Gong, "Progressive Layered Extraction (Ple): A Novel Multi-Task Learning (Mtl) Model for Personalized Recommendations," in *Proc. of the 14th ACM RecSys*, 2020, pp. 269–278.

[46] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-Of-Experts Layer," *arXiv preprint arXiv:1701.06538*, 2017.

[47] I. Corporation, "Intel® Wi-Fi 6E AX210," https://www.intel.com/content/www/us/en/products/sku/204836/intel-wifi-6e-ax210-gig/specifications.html, 2024, online; accessed 25 Jun 2024.

[48] Z. Jiang, T. H. Luan, X. Ren, D. Lv, H. Hao, J. Wang, K. Zhao, W. Xi, Y. Xu, and R. Li, "Eliminating the Barriers: Demystifying Wi-Fi Baseband Design and Introducing the PicoScenes Wi-Fi Sensing Platform," *IEEE Internet of Things Journal*, pp. 1–21, 2021.

[49] A. G. Howard, "Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.

[50] R. Zhang, C. Jiang, S. Wu, Q. Zhou, X. Jing, and J. Mu, "Wi-Fi Sensing for Joint Gesture Recognition and Human Identification From Few Samples in Human-Computer Interaction," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 7, pp. 2193–2205, 2022.

[51] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-Eyes: Device-Free Location-Oriented Activity Identification Using Fine-Grained WiFi Signatures," in *Proc. of the 20th ACM MobiCom*, 2014, pp. 617–628.

[52] P. Melgarejo, X. Zhang, P. Ramanathan, and D. Chu, "Leveraging Directional Antenna Capabilities for Fine-Grained Gesture Recognition," in *Proc. of the 16nd UbiComp*, 2014, pp. 541–551.

[53] W. He, K. Wu, Y. Zou, and Z. Ming, "Wig: Wifi-Based Gesture Recognition System," in *Proc. of 24th ICCCN*. IEEE, 2015, pp. 1–7.

[54] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, "Wifinger: Talk to Your Smart Devices With Finger-Grained Gesture," in *Proc. of the 18th UbiComp*, 2016, pp. 250–261.

[55] H. Zou, J. Yang, Y. Zhou, and C. J. Spanos, "Joint Adversarial Domain Adaptation for Resilient Wifi-Enabled Device-Free Gesture Recognition," in *Proc. of 17th IEEE ICMLA*. IEEE, 2018, pp. 202–207.

[56] H. Wang, J. Hu, T. Zheng, J. Hu, Z. Chen, H. Jiang, Y. Zheng, and J. Luo, "MuKI-Fi: Multi-person Keystroke Inference With BFI-enabled Wi-Fi Sensing," *IEEE Transactions on Mobile Computing*, 2024.

[57] R.-H. Liang and M. Ouhyoung, "A Real-Time Continuous Gesture Recognition System for Sign Language," in *Proc. of 3rd IEEE FG*. IEEE, 1998, pp. 558–567.

[58] C. Xu, P. H. Pathak, and P. Mohapatra, "Finger-Writing With Smartwatch: A Case for Finger and Hand Gesture Recognition Using Smartwatch," in *Proc. of the 16th HotMobile*, 2015, pp. 9–14.

[59] J. Gummeson, B. Priyantha, and J. Liu, "An Energy Harvesting Wearable Ring Platform for Gestureinput on Surfaces," in *Proc. of the 12th ACM MobiSys*, 2014, pp. 162–175.

[60] J. M. Rehg and T. Kanade, "Visual Tracking of High Dof Articulated Structures: An Application to Human Hand Tracking," in *Computer Vision—ECCV'94*. Springer, 1994, pp. 35–46.

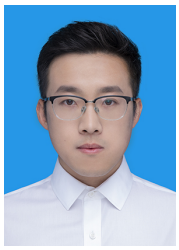
[61] G. Marin, F. Dominio, and P. Zanuttigh, "Hand Gesture Recognition With Leap Motion and Kinect Devices," in *2014 IEEE ICIP*. IEEE, 2014, pp. 1565–1569.

[62] L. Zhang, G. Zhu, P. Shen, J. Song, S. Afaq Shah, and M. Bennamoun, "Learning Spatiotemporal Features Using 3DCNN and Convolutional LSTM for Gesture Recognition," in *Proceedings of the IEEE ICCV*, Oct 2017.

[63] S. Hazra and A. Santra, "Robust Gesture Recognition Using Millimetric-Wave Radar System," *IEEE sensors letters*, vol. 2, no. 4, pp. 1–4, 2018.

[64] S. Ahmed, W. Kim, J. Park, and S. H. Cho, "Radar-Based Air-Writing Gesture Recognition Using a Novel Multistream CNN Approach," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23 869–23 880, 2022.

Hongbo Wang (Graduate Student Member, IEEE) is a currently pursuiting Ph.D. student with the College of Computing and Data Science, Nanyang Technological University, Singapore. He received the MS degree in Comunications Engineering from Nanyang Technological University in 2021 and the BS degree in Electrical Engineering from University of Electronic Science and Technology of China in 2020. He has published papers in ACM Sensys, ACM MobiCom, ACM CCS, etc. His research interests include Integrated Sensing and Communication (ISAC) and deep learning.



Xin Li (Member, IEEE) received his Ph.D. degree in Power Machinery and Engineering from Tianjin University, China, in 2022. From 2022 to 2023, he has worked as a senior software engineer in Huawei Technologies Co., Ltd., Intelligent Automotive Solution Business Unit. He now is a research fellow at College of Computing and Data Science, Nanyang Technological University. His research interests include machine learning, signal processing, and wireless sensing.

Jiachun Li (Graduate Student Member, IEEE) is a Ph.D. candidate in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. He received B.S. degree in Communication Engineering from Huazhong University of Science and Technology in 2020. He has published paper in USENIX Security, IEEE INFOCOM, IEEE TIFS, IEEE TDSC, etc. His research interests include smart home security and the metaverse security.

Haojin Zhu (Fellow, IEEE) received his B.Sc. degree (2002) from Wuhan University (China), his M.Sc.(2005) degree from Shanghai Jiao Tong University (China), both in computer science and the Ph.D. in Electrical and Computer Engineering from the University of Waterloo (Canada), in 2009. He is currently a professor in the Computer Science department at Shanghai Jiao Tong University. He has published more than 160 paper in top-tier conferences IEEE S&P, ACM CCS, USENIX Security, NDSS, and top-tier transaction JSAC, TDSC, TPDS, TMC, TIFS. He received a number of awards including SIGSOFT Distinguished Paper of ESEC/FSE (2023), ACM CCS Best Paper Runner-Ups Award (2021), USENIX Security Distinguished Paper (2024), etc. His current research interests include network security and privacy enhancing technologies. More information can be found at https://nsec.sjtu.edu.cn/ hjzhu/.

Jun Luo (Fellow, IEEE) received his BS and MS degrees in Electrical Engineering from Tsinghua University, China, and the Ph.D. degree in Computer Science from EPFL (Swiss Federal Institute of Technology in Lausanne), Lausanne, Switzerland. From 2006 to 2008, he has worked as a postdoctoral research fellow in the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. In 2008, he joined the faculty of Nanyang Technological University in Singapore, where he is currently an Associate Professor. His research interests include mobile and pervasive computing, wireless networking, machine learning and computer vision, applied operations research, as well as security. More information can be found at http://www.ntu.edu.sg/home/junluo.