Source-Free Domain Adaptation via Perceptual Semantic Decoupling for WiFi Gesture Recognition

Yelin Wei*, Xiang Zhang*, Bin Liu*, Songming Jia*, Jinyang Huang†, Zhi Liu§, Huan Yan‡
*University of Science & Technology of China, Hefei, China
†Hefei University of Technology, Hefei, China; ‡Guizhou Normal University, Guiyang, China
§The University of Electro-Communications, Tokyo, Japan
Email: {yhweiyelin, songming0612}@mail.ustc.edu.cn; {zhangxiang, liu}@ieee.org; flowice@ustc.edu.cn;
hjy@hfut.edu.cn; yh1995.cs@gmail.com

***Abstract*—Generalizable WiFi gesture recognition has gained increasing attention for its contactless operation, ubiquitous infrastructure and enhanced robustness. Among existing methods, source-free domain adaptation (SFDA) stands out by preserving privacy and reducing computational demands without relying on source data. Current methods typically process low-level WiFi signals and their high-level semantic representations from a unified perspective, making temporal semantic learning highly susceptible to low-level signal noise and lacking consistent semantic guidance for cross domain alignment, thereby limiting the effectiveness. In this paper, we propose ViFi, a novel SFDA framework specifically designed for cross-domain WiFi gesture recognition. Unlike prior work, ViFi introduces a viewpoint-hierarchical strategy that explicitly process cross-domain sensing from two perspectives: the perceptual (signal-level) and the semantic (gesture-level). This separation mitigates the impact of signal noise on high-level semantics while preventing semantic space drift during domain alignment. ViFi operates in two key stages. First, it anchors the perceptual encoder and employs masked signal semantic reconstruction to learn robust high-level temporal semantics. Then, it freezes the semantic encoder and aligns the perceptual encoder across domains, again leveraging masked reconstruction to ensure alignment under a unified and meaningful semantic space. We evaluate ViFi on a public dataset, and experimental results show that our viewpoint-hierarchical method achieves over 15% improvement compared to the baseline and significantly outperforms state-of-the-art approaches.**

***Index Terms*—WiFi sensing, channel state information, source-free domain adaptation, gesture recognition.**

I. Introduction

Gesture recognition underpins a wide range of human-computer interaction (HCI) applications [1]. Among various sensing modalities, WiFi Channel State Information [2] (CSI)-based gesture recognition [3] has attracted increasing attention due to its ubiquity, privacy-preserving nature, and contactless sensing capabilities. The adoption of machine learning and deep learning techniques has significantly improved the performance of WiFi-based gesture recognition systems [4], [5]. However, these systems often struggle with generalization across domains, primarily due to distribution shifts caused by differences in physical environments and sensing perspectives between training and deployment. To bridge this gap and advance the practical deployment of WiFi sensing, cross-domain WiFi gesture recognition [6] has seen notable progress in recent years.

Current cross-domain WiFi gesture recognition approaches can be broadly categorized into two types: domain generalization (DG) and domain adaptation (DA) methods. The core idea behind DG is to leverage existing data to learn domain-invariant features. Early studies mainly use handcrafted design to derive domain-invariant features from WiFi CSI, based on expert knowledge. For example, Widar 3.0 [7] utilizes the Body-coordinate Velocity Profile (BVP) to achieve domain-independent gesture representation. However, such handcrafted approaches may lead to the loss of critical information from the original signal. Thus, recent studies have shifted their focus toward automated generalization feature learning. These methods typically employ adversarial learning [8] and complementary information [9] to automatically refine robust features. Nevertheless, relying solely on domain-invariant features limits performance by overlooking domain-specific cues that remain valuable for gesture recognition.

On the other hand, DA methods aim to explicitly reduce the distribution gap between source and target domains. Therefore, DA offers a more effective way to leverage all the available information [10]. These approaches typically assume access to labeled source-domain data and unlabeled (or partially labeled) target-domain data during training. Through techniques such as adversarial training, feature alignment, or self-supervised learning, they align representations across domains to enhance cross-domain generalization [11]. Due to privacy concerns and high computational costs, most recent studies avoid using raw WiFi source data (Source Free) for DA, relying instead only on pre-trained source models. Some adopt few-shot strategies with limited labeled target data [5], [12], while others like Wi-SFDAGR [13] use self-supervised clustering on unlabeled data. These frameworks jointly learn low-level perception and high-level semantic features from CSI during both representation learning and DA. However, CSI is a highly noisy and environment-sensitive signal. This joint learning strategy may lead the model to overemphasize denoising at the perceptual level, rather than effectively capturing meaningful temporal semantics. Furthermore, during the DA phase, these methods may tend to focus more on finetune semantic encoder, while neglecting the more fundamental yet challenging task of unifying perceptual viewpoints across domains.

To address these limitations, we propose ViFi, a novel Source-Free Domain Adaptation (SFDA) framework tailored for WiFi-based cross-domain gesture recognition. Unlike prior methods, ViFi adopts a viewpoint-hierarchical strategy that separates WiFi sensing into two levels: a perceptual (signal-level) and a semantic (gesture-level) perspective. The perceptual level captures low-level signal variations, while the semantic level focuses on temporal relations in CSI that underlie gesture representation. Specifically, in source domain training stage, ViFi first trains and freezes a perceptual encoder on the source domain to maintain a stable perceptual viewpoint, then a semantic encoder is subsequently trained by reconstructing masked temporal segments of the CSI. During target domain adaptation, the semantic encoder is fixed, and the same reconstruction strategy guides the target perceptual encoder to align with the source, enabling perceptual domain adaptation under a unified semantic view. Experiments show ViFi achieves over 15% improvement over the baseline and significantly outperforms current state-of-the-art (SOTA) methods.

II. Preliminaries

A. Channel State Information

CSI describes the fine-grained channel characteristics during the propagation of WiFi signals from the transmitter (Tx) to the receiver (Rx), and it can be decomposed into static components $H_s(f,t)$ and dynamic components $H_d(f,t)$ as [14], [15]:

$$H(f,t) = H_s(f,t) + H_d(f,t) \tag{1}$$

where f, t, $H_s(f,t)$ and $H_d(f,t)$ represent the signal frequency, the timestamp, the CSI from Line-of-Sight (LoS) and static reflections, and the CSI caused by moving objects, respectively.

$H_d(f,t)$ can be further represented as [16]:

$$H_d(f,t) = \sum_{n \in P_d} a_n(f,t) e^{-j2\pi \frac{d_n(t)}{\lambda}} \tag{2}$$

where P_d denotes the collection of dynamic paths, and $a_n(f,t)$, $e^{-j2\pi \frac{d_n(t)}{\lambda}}$ and $d_n(t)$ denote the complex attenuation, phase shift and path length of the n-th path, respectively.

During gesture execution, the movement of hands and arms alters the propagation path lengths of dynamic components and make two dominant features changes: amplitude fluctuations $|H(f,t)|$ from multipath interference patterns, and phase modulation $e^{-j2\pi \frac{d_n(t)}{\lambda}}$ directly proportional to motion displacement. By analyzing these variations, different gestures can be effectively distinguished.

B. Problem Definition

ViFi is designed to address the challenge of SFDA for WiFi-based gesture recognition. We first define the notations used throughout this work. Let $\mathcal{D}_S$ denote the source dataset, consisting of pairs $\{X_S, Y_S\}$, where $X_S \in \mathcal{X}_S$ represents the preprocessed CSI data and $Y_S \in \mathcal{Y}_S$ denotes the corresponding ground truth gesture labels. The target dataset, $\mathcal{D}_T$, comprises only unlabeled CSI measurements $\{X_T\}$, where $X_T \in \mathcal{X}_T$. The corresponding gesture labels $Y_T \in \mathcal{Y}_T$ are reserved solely for evaluation purposes.

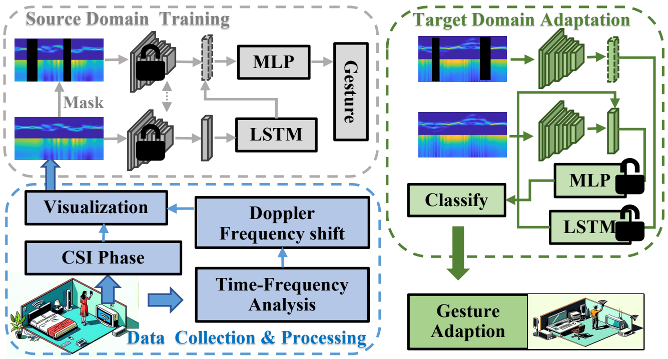

Fig. 1. The system framework of ViFi.

CSI data typically exhibits a large volume and may contain sensitive information. Therefore, when deploying a WiFi-based gesture recognition system in a new environment, using $\mathcal{D}_S$ to facilitate the adaptation process would not only raise privacy concerns but also impose significant computational and storage demands. ViFi aims to enable a WiFi gesture recognition system to adapt to new domains and be deployed using only $\{X_T\}$, without requiring access to the $\mathcal{D}_S$.

III. Methods

A. System Overview

The overview of SFDA is illustrated in Fig. 1, which comprises the following components: data collection and processing, source domain training and target domain adaptation.

Data Collection and Processing: The transmitter continuously emits WiFi signals while the receiver recive it and extract the CSI. The acquired CSI data inevitably contains substantial noise, which degrades recognition accuracy. Thus, we first employ the CSI ratio method [17] to denoise the signals and extract clean amplitude and phase information. After obtaining relatively clean CSI signals, we apply the Short-Time Fourier Transform (STFT) to extract Doppler Frequency Shift (DFS) in the frequency domain while retaining phase information in the time domain. By vertically concatenating the visual representations of DFS and phase, we generate time-frequency composite images that simultaneously captures spectral and temporal characteristics. These images are then used for subsequent modules.

Source Domain Training: In the source domain training phase, ViFi aims to establish a hierarchical sensing structure that separates low-level perceptual features from high-level semantic representations. First, a perceptual encoder combined with a classifier is pre-trained to extract signal-level features from inputs. Then, the perceptual encoder is frozen to preserve a consistent sensing viewpoint, and a LSTM-based semantic encoder is optimized through a masked reconstruction task, which forces the semantic network to capture meaningful gesture-related temporal dependencies. This two-stage training ensures robust feature learning while maintaining a clean separation between perception and semantics.

Target Domain Adaptation: During deployment in a new environment, ViFi performs source-free domain adaptation without accessing the source data. The semantic encoder and classifier, pre-trained on the source domain, remain fixed to anchor the learned gesture semantics. Only the perceptual encoder is fine-tuned using unlabeled target domain CSI data. To guide the adaptation, masked reconstruction is again employed, enforcing consistency between the adapted perceptual features and the frozen semantic representations. This design enables the model to align domain-specific sensing variations while preserving the temporal semantic structure, thus achieving robust cross-domain adaptation.

B. Data Collection and Processing

In ideal scenarios, uncontaminated signal measurements could be utilized directly without additional preprocessing. However, practical implementations must account for inherent signal distortions caused by carrier frequency offset, sampling frequency offset, and packet detection delays. These impairments inevitably introduce additive noise $e^{-j\varphi}$ to the acquired CSI, significantly degrading its reliability. Formally, the observed noisy CSI measurement can be modeled as [18]:

$$\begin{aligned} H(f,t) &= e^{-j\varphi}(H_s(f,t) + H_d(f,t)) \\ &= e^{-j\varphi}\left(H_s(f,t) + \sum_{n\in P_d} a_n(f,t)e^{-j2\pi\frac{d_n(t)}{\lambda}}\right) \end{aligned} \tag{3}$$

Random noise significantly corrupts CSI phase extraction accuracy. Since all antennas on the network interface card share the same RF oscillator, the phase offset $e^{-j\varphi}$ remains consistent across all subcarriers. To effectively suppress this noise, we adopt the CSI-ratio method, formulated as [19]:

$$\begin{aligned} H_q(f,t) &= \frac{H_1(f,t)}{H_2(f,t)} \\ &= \frac{e^{-j\varphi}\left(H_{s,1} + \sum_{n\in P_d} a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda}}\right)}{e^{-j\varphi}\left(H_{s,2} + \sum_{n\in P_d} a_2(f,t)e^{-j2\pi\frac{d_2(t)}{\lambda}}\right)} \\ &= \frac{\sum_{n\in P_d} a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda}} + H_{s,1}}{\sum_{n\in P_d} a_2(f,t)e^{-j2\pi\frac{d_1(t)+\Delta d}{\lambda}} + H_{s,2}} \end{aligned} \tag{4}$$

where $H_1(f,t)$ and $H_2(f,t)$ are the CSI of two nearby antennas. And $\triangle d$ can be regarded as a constant because $H_1(f,t)$ and $H_2(f,t)$ are close to each other. The transformation in (4), comprising scaling and rotation operations on the phase shift $e^{-j2\pi\frac{d_1(t)}{\lambda}}$, is trend-preserving in the complex domain. This property enables simultaneous noise suppression and gesture information retention.

To further mitigate error propagation caused by parameter Δd, we introduce a proportional coefficient ρ_k for optimal antenna pair selection, formulated as:

$$\rho_k = \frac{1}{I}\sum_{i=1}^{I}\frac{\mathrm{var}\left(|H_k(f_i,t)|\right)}{\mathrm{mean}\left(|H_k(f_i,t)|\right)}, \quad k \in [1,3] \tag{5}$$

where var and mean denote the variance and mean value of amplitude readings for the k-th antenna of the i-th subcarrier. Since CSI with larger variance typically exhibit higher sensitivity to motion, while those with greater amplitude often contain stronger static path components, we strategically select the antenna with the highest and lowest ρ_k values as $H_1(f,t)$ and $H_2(f,t)$. This selection criterion effectively minimizes the impact of Δd on $H_1(f,t)$ while maintaining optimal motion detection capability.

By applying the CSI-ratio method combined with our antenna selection strategy, we first obtain denoised phase information. This processed data is then transformed into DFS through STFT. To enhance visual clarity and removing background information, we employ the CSI visualization techniques in [3] to convert the DFS into heatmap images. Finally, we vertically concatenate DFS and denoised phase images to generate a composite DFS-phase diagram sharing a unified timeline, which serves as the input X for subsequent analysis.

C. Source Domain Training

In the source domain training phase, ViFi adopts a two-stage hierarchical learning strategy to separate low-level perceptual features from high-level semantic representations. Specifically, ViFi first anchor the perceptual encoder to extract reliable low-level representations and then train the semantic encoder to model temporal relations through a masked signal reconstruction task. We first illustrate our masking approach here as the preliminary knowledge. Given an input $X \in \mathbb{R}^{H\times W}$, where H and W represent the time and frequency dimensions respectively, we first divide the image into N vertical blocks of equal width along the time dimension.

The masking process is formally defined as:

$$X'(i,j) = \begin{cases} 0, & \text{if } j \in \mathcal{M} \\ X(i,j), & \text{otherwise} \end{cases} \tag{6}$$

where $\mathcal{M}$ denotes the set of randomly selected vertical blocks to be masked, and (i,j) represents the pixel coordinates, and we use X' denote the masked input.

In the pre-training phase, a perceptual encoder $p_\phi : \mathcal{X}_S \mapsto \mathcal{F}_S$ combined with a gesture classifier $c_\theta : \mathcal{F}_S \mapsto \mathcal{Y}_S$ is trained to extract stable and discriminative signal-level features from X by using cross-entropy loss:

$$\mathcal{L}_{cls}^{S} = -\mathbb{E}_{(X_S,Y_S)\sim\mathcal{D}_S}\left[\sum_{m=1}^{M}\mathbb{I}[Y_S = m]\log\sigma(c_\theta(p_\phi(X_S)))\right] \tag{7}$$

where M denotes the number of gesture classes, $\sigma(\cdot)$ represents the softmax function, and $\mathbb{I}[\cdot]$ is the indicator function.

After achieving satisfactory classification performance, the perceptual encoder is frozen to preserve the learned stable sensing viewpoint. Subsequently, a semantic encoder s_ψ, based on a LSTM network, is trained to model the temporal relationships embedded within the feature sequences by using the masked reconstruction task. This task aims to reconstruct $p_\phi(X_S)$ form $p_\phi(X'_S)$ to encourages the s_ψ to capture robust, gesture-relevant temporal dependencies, enhancing its generalization capability across domains.

The s_ψ is optimized by minimizing the feature-space mean squared error:

$$\mathcal{L}^S_{recon} = \mathbb{E}_{X_S \sim \mathcal{X}_S} \left[\| p_\phi(X_S) - s_\psi(p_\phi(X'_S)) \|_2^2 \right] \quad (8)$$

Notably, the reconstruction loss $\mathcal{L}^S_{recon}$ exclusively constrains the semantic network s_ψ, while the classification loss $\mathcal{L}^S_{cls}$ only applies to the perceptual network p_ϕ and classifier c_θ. This decoupled optimization strategy yields two key advantages: (1) During domain adaptation, the feature alignment can be performed separately at different levels, preventing simultaneous alignment of high-level and low-level features that may cause semantic deviation; (2) The perceptual network maintains stable feature extraction capabilities unaffected by reconstruction errors, whereas the semantic network focuses solely on temporal relationship modeling. Finally, This two-stage training procedure establishes a clean separation between perception and semantics, ensuring that the subsequent domain adaptation can focus on aligning only the perceptual sensing variations without disrupting the gesture semantics.

D. Target Domain Adaptation

During deployment in unseen target environments, ViFi performs SFDA, relying solely on the pre-trained source model in Section III-C without accessing the source dataset. In this phase, the s_ψ and c_θ are kept frozen, preserving the learned temporal semantic structures from the source domain. Only the p_ϕ is fine-tuned to adapt to domain-specific variations present in the target CSI data.

Specifically, masked $\{X'_T\}$ are passed through the p_ϕ and frozened s_ψ, and a feature reconstruction loss is computed between the reconstructed features and the original perceptual outputs:

$$\mathcal{L}^T_{adapt} = \mathbb{E}_{X_T \sim \mathcal{X}_T} \left[\| p_\phi(X_T) - s_\psi(p_\phi(X'_T)) \|_2^2 \right] \quad (9)$$

The frozen semantic network s_ψ serves as an anchor to maintain temporal sementaic constraints, while only the perceptual network p_ϕ undergoes optimization. This selective adaptation forces target features to conform to the source domain's reconstruction criteria, achieving implicit perceptual viewpoint matching, and ensures stable perceptual transformation by preventing semantic deviation caused by simultaneous alignment of multiple levels. In addition, a pseudo-labeling mechanism is adopted to provide a soft supervision signal for the classifier on unlabeled target data. The overall adaptation objective combines temporal reconstruction consistency and classification alignment:

$$\mathcal{L}_{total} = \mathcal{L}^T_{adapt} + \lambda \mathbb{E}_{X_T \sim \mathcal{D}_T} \left[-\log c_\theta(\hat{y}_T | p_\phi(X_T)) \right] \quad (10)$$

where λ regulates the relative importance weight between perceptual-level temporal consistency and temporal semantic-level alignment, and $\hat{y}_T$ denotes pseudo-labels generated by the fixed source classifier.

Notably, the second term in (10) offers flexible implementation options: while we employ pseudo-labeling here, any established SFDA technique (e.g., entropy minimization or prototype matching) can be alternatively applied, provided it operates solely on the perceptual network's output features. This approach ensures stable feature transformation while preventing semantic deviation caused by simultaneous alignment of multiple levels.

IV. Evaluations

A. Datasets

The performance evaluation employs the Widar3.0 [7] containing CSI measurements collected from commercial Wi-Fi devices in multiple indoor environments. Our experiments primarily utilize data from Environment 1 (a classroom) involving nine users performing six distinct gestures, with each gesture repeated five times across five orientations and five spatial positions, and additionally include one user's data from Environment 2 (a hall) for cross-environment evaluation. The data collection employed six transceiver pairs, while single-antenna experiments consistently used Receiver 2 as the default configuration similar to Wi-learner [12].

To systematically evaluate cross-domain performance, we examine five key domain factors: location changes, orientation differences, environmental shifts, user variations and antenna pair configurations. The evaluation follows a structured protocol where for each domain factor being tested, we employ 80% of the data from one domain instance for adaptation and reserve the remaining 20% for testing, while using data from all other domain instances for training. This evaluation methodology is consistently applied across all domain factors, with results reported as average accuracy across all possible held-out domain configurations.

B. Implementation Details

The experimental implementation involves MATLAB for CSI data preprocessing and PyTorch for network construction. We employ a ResNet18 pretrained on ImageNet as the perceptual network and a LSTM as semantic network. For unified cross-domain evaluation, all experiments share identical hyperparameters: the source domain training uses an initial learning rate of 1×10^{-3} with decay factor 0.1 every 10 epochs over 40 total epochs, while target domain adaptation employs a lower initial rate of 1×10^{-4} for 10 epochs. Both phases utilize a batch size of 64 and default $\lambda = 1$ for loss balancing. This configuration ensures consistent comparison of cross-domain performance while maintaining training stability.

C. Overall Performance

The overall performance of ViFi is presented in Table I and Table II. These tables provide comparative results between the ViFi framework and existing DA and DG methods on the Widar3.0 dataset for both in-domain and cross-domain recognition tasks. Evidently, our proposed ViFi framework outperforms most baseline methods. Remarkably, ViFi consistently achieves over 90% accuracy in all single-antenna cross-domain experiments, demonstrating exceptional robustness

TABLE I
PERFORMANCE COMPARISON ON WIDAR3 DATASET USING 1 ANTENNA PAIR

Methods	Accuracy Across Different Scenarios (%)					
	In-Domain	Cross-Loc	Cross-Ori	Cross-Env	Cross-User	Cross-Ant
SelfReg [20]	–	76.71	86.67	39.11	53.10	–
WiSGP [21]	–	78.49	88.46	43.17	56.77	–
WiSR [9]	–	77.51	88.80	42.52	55.18	–
Wi-Learner [12]	93.20	91.40	86.50	74.20	89.40	**94.40**
WiOpen [4]	–	86.40	77.67	84.44	82.71	–
UniFi* [6]	97.50	92.50	92.00	87.50	–	–
ViFi	**98.67**	**97.78**	**93.74**	**98.00**	**92.67**	93.11

*Approximated from graphical data. Bold indicates best performance.

TABLE II
PERFORMANCE COMPARISON ON WIDAR3 DATASET USING 6 ANTENNA PAIRS

Methods	Accuracy Across Different Scenarios (%)				
	In-Domain	Cross-Loc	Cross-Ori	Cross-Env	Cross-User
EI [8]	97.40	73.33	79.70	63.50	–
Widar3.0 [7]	79.25	76.22	78.07	62.24	–
WiHF [22]	97.25	89.11	87.55	82.25	–
WIGRUNT [3]	98.66	92.07	91.92	85.19	–
PAC-CSI [23]	99.46	98.77	98.90	96.47	97.54
UniFi [6]	99.40	99.18	99.40	97.73	96.27
Wi-SFDAGR [13]	–	97.30	97.17	95.52	–
MetaFormer [5]	**100.00**	99.00	**100.00**	81.00	94.67
ViFi	**100.00**	**100.00**	99.51	**100.00**	**98.67**

Bold indicates best performance.

to domain shifts. Specifically, when utilizing only a single antenna pair for gesture recognition, ViFi achieves SOTA performance in all configurations except for the cross-antenna pair scenarios, where it trails Wi-Learner [12] by merely 1%. And in cross-environment experiments, ViFi demonstrates more than twofold accuracy improvement over some competing methods.

In the six-antenna-pair configuration, where most methods achieve high accuracy due to abundant available information, our approach remains competitive with existing methods, attaining 100% accuracy in some cross-domain experiments. It is worth noting that several methods adopt similar baseline architectures: WiSGP [21], WiSR [9] and Wiopen [4] all employ ResNet as their backbone, while MetaFormer [5] utilizes a more complex transformer architecture. However, except for the cross-orientation case, none surpass ViFi in overall performance. This may due to that jointly learning low-level perceptual features and high-level semantic features from CSI data through representation learning and data analysis may inevitably suffer from environmental noise and CSI sensitivity limitations, preventing networks from achieving optimal performance.

D. Ablation Study

To further validate the impact of the viewpoint-hierarchical strategy across various domains, we conducted ablation studies by removing the vertical masking method while maintaining identical data processing pipelines, backbone, SFDA methods and hyperparameters. As illustrated in Fig. 2 and Fig. 3, the experimental results demonstrate that the ViFi framework achieves an average accuracy improvement exceeding 15% compared to the baseline without viewpoint-hierarchical strategy in single-antenna configurations. Notably, the accuracy gains reach nearly 30% for cross-orientation and cross-user tasks. This significant performance gap stems from two inherent challenges: severe projection distortion of identical gestures when viewed from different orientations, and user-specific motion patterns for the same gesture category. These factors make these cross-domain tasks particularly challenging for perceptual networks alone, underscoring the necessity of incorporating temporal semantic information. Although the performance difference diminishes in multi-antenna scenarios

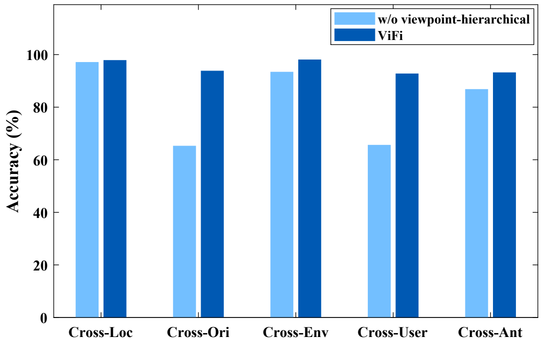

Fig. 2. The performance with/without the viewpoint-hierarchical strategy using 1 antenna pair

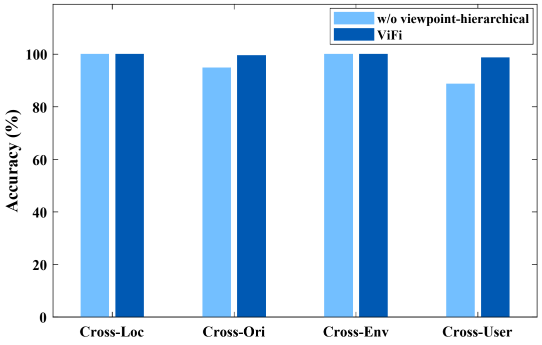

Fig. 3. The performance with/without the viewpoint-hierarchical strategy using 6 antenna pair

(with similar trends observed), the results collectively indicate that losing semantic understanding critically degrades gesture recognition performance in complex tasks. The ablation studies confirm that explicit modeling of viewpoint hierarchies is essential for robust cross-domain adaptation.

V. Conclusion and Future Work

This paper presents ViFi, a novel viewpoint-hierarchical source-free domain adaptation (SFDA) framework designed for cross-domain WiFi-based gesture recognition. ViFi decouples perceptual sensing and semantic modeling by anchoring a stable perceptual encoder and learning robust temporal semantics through masked signal reconstruction, thereby addressing the challenges of signal noise sensitivity and semantic drift inherent in conventional methods. During deployment, ViFi enables efficient adaptation by fine-tuning only the perceptual encoder on unlabeled target data while preserving the learned semantic structures, achieving robust performance without accessing source domain data. Extensive experiments on the Widar3.0 dataset demonstrate that ViFi achieves consistent improvements across multiple domain shifts, surpassing state-of-the-art methods by over 15% on average and up to 30% in challenging cross-orientation and cross-user scenarios. Future work will explore extending ViFi to dynamic and continuously evolving environments, incorporating online adaptation mechanisms, and further reducing the adaptation overhead to enhance the practicality and scalability of WiFi-based human sensing systems.

References

[1] J. Huang, J.-X. Bai, X. Zhang, Z. Liu, Y. Feng, J. Liu, X. Sun, M. Dong, and M. Li, "Keystrokesniffer: An off-the-shelf smartphone can eavesdrop on your privacy from anywhere," *IEEE Transactions on Information Forensics and Security*, 2024.

[2] E.-Z. Yi, K. Niu, F.-S. Zhang, R.-Y. Gao, J. Luo, and D.-Q. Zhang, "Multi-person respiration monitoring leveraging commodity wi-fi devices," *Journal of Computer Science and Technology*, vol. 40, no. 1, pp. 229–251, 2025.

[3] Y. Gu, X. Zhang, Y. Wang, M. Wang, H. Yan, Y. Ji, Z. Liu, J. Li, and M. Dong, "Wigrunt: Wifi-enabled gesture recognition using dual-attention network," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 736–746, 2022.

[4] X. Zhang, J. Huang, H. Yan, Y. Feng, P. Zhao, G. Zhuang, Z. Liu, and B. Liu, "Wiopen: A robust wi-fi-based open-set gesture recognition framework," *IEEE Transactions on Human-Machine Systems*, vol. 55, no. 2, pp. 234–245, 2025.

[5] B. Sheng, R. Han, F. Xiao, Z. Guo, and L. Gui, "Metaformer: Domain-adaptive wifi sensing with only one labelled target sample," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–27, 2024.

[6] Y. Liu, A. Yu, L. Wang, B. Guo, Y. Li, E. Yi, and D. Zhang, "Unifi: A unified framework for generalizable gesture recognition with wi-fi signals using consistency-guided multi-view networks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 4, pp. 1–29, 2024.

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

[8] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th annual international conference on mobile computing and networking*, 2018, pp. 289–304.

[9] S. Liu, Z. Chen, M. Wu, C. Liu, and L. Chen, "Wisr: Wireless domain generalization based on style randomization," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 4520–4532, 2023.

[10] C. Chen, G. Zhou, and Y. Lin, "Cross-domain wifi sensing with channel state information: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–37, 2023.

[11] B.-B. Zhang, D. Zhang, Y. Li, Y. Hu, and Y. Chen, "Unsupervised domain adaptation for rf-based gesture recognition," *IEEE Internet of Things Journal*, vol. 10, no. 23, pp. 21 026–21 038, 2023.

[12] C. Feng, N. Wang, Y. Jiang, X. Zheng, K. Li, Z. Wang, and X. Chen, "Wi-learner: Towards one-shot learning for cross-domain wi-fi based gesture recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–27, 2022.

[13] H. Yan, X. Zhang, J. Huang, Y. Feng, M. Li, A. Wang, W. Ou, H. Wang, and Z. Liu, "Wi-sfdagr: Wifi-based cross-domain gesture recognition via source-free domain adaptation," *IEEE Internet of Things Journal*, 2025.

[14] J. Huang, B. Liu, C. Miao, X. Zhang, J. Liu, L. Su, Z. Liu, and Y. Gu, "Phyfinatt: An undetectable attack framework against phy layer fingerprint-based wifi authentication," *IEEE Transactions on Mobile Computing*, 2023.

[15] H. Wang, X. Li, J. Li, H. Zhu, and J. Luo, "Vr-fi: Positioning and recognizing hand gestures via vr-embedded wi-fi sensing," *IEEE Transactions on Mobile Computing*, 2025.

[16] X. Zhang, Y. Gu, H. Yan, Y. Wang, M. Dong, K. Ota, F. Ren, and Y. Ji, "Wital: A cots wifi devices based vital signs monitoring system using nlos sensing model," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 3, pp. 629–641, 2023.

[17] Y. Zeng, D. Wu, J. Xiong, E. Yi, R. Gao, and D. Zhang, "Farsense: Pushing the range limit of wifi-based respiration sensing with csi ratio of two antennas," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–26, 2019.

[18] X. Zhang, J. Zhang, Z. Ma, J. Huang, M. Li, H. Yan, P. Zhao, Z. Zhang, B. Liu, Q. Guo, T. Zhang, and N. Yu, "Camlopa: A hidden wireless camera localization framework via signal propagation path analysis," in *2025 IEEE symposium on security and privacy (SP)*. IEEE, 2025.

[19] D. Wu, R. Gao, Y. Zeng, J. Liu, L. Wang, T. Gu, and D. Zhang, "Fingerdraw: Sub-wavelength level finger motion tracking with wifi signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–27, 2020.

[20] D. Kim, Y. Yoo, S. Park, J. Kim, and J. Lee, "Selfreg: Self-supervised contrastive regularization for domain generalization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9619–9628.

[21] S. Liu, Z. Chen, M. Wu, H. Wang, B. Xing, and L. Chen, "Generalizing wireless cross-multiple-factor gesture recognition to unseen domains," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 5083–5096, 2023.

[22] C. Li, M. Liu, and Z. Cao, "Wihf: Gesture and user recognition with wifi," *IEEE Transactions on Mobile Computing*, vol. 21, no. 2, pp. 757–768, 2020.

[23] J. Su, Q. Mao, Z. Liao, Z. Sheng, C. Huang, and X. Zhang, "A real-time cross-domain wi-fi-based gesture recognition system for digital twins," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 11, pp. 3690–3701, 2023.