Toward Cross-Environment Continuous Gesture User Authentication With Commercial Wi-Fi

Lei Zhang, *Member, IEEE*, Yazhou Ma, Mingzi Zuo, Zhen Ling, *Member, IEEE*, Changyu Dong, *Member, IEEE*, Guangquan Xu, *Member, IEEE*, Lin Shu, *Senior Member, IEEE*, Xiaochen Fan, *Member, IEEE*, and Qian Zhang, *Fellow, IEEE*

***Abstract*—Behavior biometrics-based user authentication with Wi-Fi gains significant attention due to its ubiquitous and contact-free manners. An individual's identity can be verified by analyzing activities induced signal variances, excellently balancing the security demands and user experience. However, the inherent complexity of Wi-Fi signals presents significant challenges for behavior biometrics-based user authentication. The susceptibility of Wi-Fi signals results in a poor cross-environment generalization capability, which is overlooked by the existing research. In addition, most existing works of behavior-based user authentication are based on one-off activity. This makes them vulnerable to zero-effort attacks and imitation attacks. To address these issues, we propose a cross-environment continuous gesture-based user authentication framework with Wi-Fi, dubbed Wi-CGAuth. Specifically, the cross-environment generalization capability is enhanced by the cross-layer joint optimization approach. At the lowest signal layer, the signals' time, spatial, and frequency diversity are extended maximally, by a novel, subcarrier-level, cost-effective signal optimization strategy. At the middle layer, the multi-view fusion method, i.e., multi-transfer component analysis (TCA), is applied to refine the signals from transceiver pairs after signal preprocessing. The continuous gesture segmentation problem is modeled as the classification problem, which is solved by CNN. At the upper layer, a Convolutional Neural Network-Transformer (CNN-Transformer) model is employed to achieve the dual task of effective user authentication and accurate gesture recognition. After extensive experiments in three typical indoor scenarios, Wi-CGAuth can achieve an average authentication accuracy of 92.7%, demonstrating its robustness and effectiveness.**

***Index Terms*—Gesture-based, user authentication, Wi-Fi, cross-environment.**

Received 7 February 2024; revised 24 June 2024 and 21 December 2024; accepted 28 February 2025; approved by IEEE TRANSACTIONS ON NETWORKING Editor S. Ioannidis. This work was supported in part by the Natural Science Foundation of Tianjin under Grant 22JCYBJC00120, in part by the Opening Fund of Key Laboratory of Computing Power Network and Information Security under Grant 2023ZD036, in part by the Opening Fund of Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems under Grant TJKL-AITS-20241001, in part by the National Key Research and Development Program of China under Grant 2022YFB3102100, in part by the National Science Foundation of China under Grant U22B2027 and Grant 62172297, in part by Hainan Province Science and Technology Special Fund under Grant ZDYF2024GXJS008, in part by Guangxi Science and Technology Plan Project (Guangxi Science and Technology Base and Talent Special Project) under Grant AD23026096 and Grant 2022AC20001, in part by the Xinjiang Corps "Tianchi Talent" Introduction Program, in part by Xinjiang Production and Construction Corps Key Laboratory of Computing Intelligence and Network Information Security, in part by Research Grants Council (RGC) under Contract CERG 16204523 and Contract AoE/E-601/22-R, in part by the National Natural Science Foundation of China under Grant 92467205, in part by Jiangsu Provincial Key Laboratory of Network and Information Security under Grant BM2003201, in part by the Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grant 93K-9, in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization, in part by The Taihu Lake Innovation Fund for the School of Future Technology of South China University of Technology under Grant 2024B105611003, and in part by Guangzhou Basic and Applied Basic Research Project under Grant 2024A03J0324. *(Yazhou Ma and Mingzi Zuo equally contributed to this work.) (Corresponding author: Qian Zhang.)*

Lei Zhang is with the School of Cyber Security, Tianjin University, Tianjin 300050, China, also with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250202, China, also with the Key Laboratory of Data and Intelligent System Security, Ministry of Education, Nankai University, Tianjin 300072, China, and also with Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tianjin 300072, China (e-mail: lzhang@tju.edu.cn).

Yazhou Ma, Mingzi Zuo, and Guangquan Xu are with the School of Cyber Security, Tianjin University, Tianjin 300050, China (e-mail: yazhouma@tju.edu.cn; zuomingzi@tju.edu.cn; losin@tju.edu.cn).

Zhen Ling is with the School of Computer Science and Engineering, Southeast University, Nanjing 211189, China (e-mail: zhenling@seu.edu.cn).

Changyu Dong is with the Institute of Artificial Intelligence, Guangzhou University, Guangzhou 510000, China (e-mail: changyu.dong@gzhu.edu.cn).

Lin Shu is with the School of Future Technology, South China University of Technology, Guangzhou, Guangdong 510641, China (e-mail: shul@scut.edu.cn).

Xiaochen Fan is with the Institute for Electronics and Information Technology at Tianjin, Tianjin 300467, China, and also with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: fanxiaochen33@gmail.com).

Qian Zhang is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China (e-mail: qianzh@ust.hk).

Digital Object Identifier 10.1109/TON.2025.3548464

I. INTRODUCTION

IN RECENT years, user authentication has widely penetrated various application fields to protect user privacy [1], [2], [3], [4], [5], including finance, healthcare, e-commerce, social media, government services, etc. Existing user authentication methods can be divided into two categories: password-based authentication and user biometrics (e.g., fingerprint [1], voiceprint [2], and facial information [3]) based authentication. They both have their own limitations. The issue of password-based authentication is password forgetting or leakage. The issue of biometrics-based authentication is the user biometrics forgery or imitation [1], [2], [3]. Recently, behavior biometrics-based user authentication [6], [7], [8], [9], [10], [11], [12] have piqued significant attention. User authentication is achieved by verifying an individual's identity through daily activities or gestures, excellently balancing the security demands and better user experience. Different users have their unique behavior biometrics and physiological characters when performing specific gestures, but these unique

identifiers remain consistent over time for the same user [11], [12], [13], [14], [15], [16]. This provides the feasibility of the behavior biometrics-based user authentication.

Behavior biometrics-based user authentication can be achieved by wearing wearable devices [6], [7]. However, there are issues such as intrusiveness, high cost, and unfriendly user experience. The video-based user authentication [8], [16] is a non-intrusive behavior-based user authentication. However, this kind of user authentication has issues of privacy violation, lighting, and occlusion sensitivity. With the quick development of Wi-Fi technology, there is the paradigm shift of Wi-Fi from the communication method to being integrated with the sensing capability. Wi-Fi sensing has the predominant advantages of low cost, ubiquitousness, privacy protection, occlusion insensitivity, non-intrusiveness, and friendly user experience. This expedites the development of the behavior biometrics-based user authentication with Wi-Fi [9], [10], [11], [12].

Behavior biometrics-based user authentication with Wi-Fi has lots of potential applications. In a smart home, behavior biometrics-based user authentication with Wi-Fi revolutionizes user interaction with household appliances. The appliances are set according to the authorized user's preferences. The user is authenticated by his biometric gestures. This enables a user-friendly experience, allowing the authorized user to trigger his personalized settings. For instance, when a subject walks close to the air conditioner and performs the gestures, he is identified as an authorized user through his gestures, and the temperature is adjusted to his favorite one. Another example is that the smart TV tries to identify an authorized user by her gesture and switches to her favorite channel. The intelligent appliances are simultaneously set to the user's preference if he is identified as an authorized user. This seamless integration of convenience and functionality greatly improves the security and user experience.

Firstly, it is challenging to build a "one-fit-all" model and improve the cross-environment generalization ability of the model. Due to the low spatial resolution of Wi-Fi signals, the features derived from primitive signals definitely carry adverse information specific to the environment unrelated to the user identity [10], [17], [18]. Unfortunately, the existing behavior biometrics-based user authentication research with Wi-Fi [11], [12], [13] do not take this issue into consideration. Previously, the cross-environment sensing solutions could be classified into the model-based and the learning-based. Previous studies on behavior sensing through Wi-Fi signals mainly utilize two types of approaches for cross-environment sensing: model-based approaches and learning-based approaches. However, for model-based sensing, accurately modelling the relationship between Wi-Fi signal variations and behaviors in complex scenarios solely through observation and experience remains challenging [10], [17], [18]. One of the most effective learning-based cross-environment sensing solutions is transfer learning [19], [20], [21]. The knowledge learned from a source domain with an abundant data set is transferred so that a classifier in the target domain can be efficiently trained with very limited labeled data. Originally being applied to image processing, the learning-based solutions can hardly be directly applied to Wi-Fi signal processing. How to seamlessly integrate the model-based into the learning-based solutions and make the cross-environment sensing more effective is challenging.

Secondly, effectively segmenting continuous activities into a series of single ones remains a significant challenge. Previous work [9], [10], [11], [12], [13], [22] of the behavior biometrics-based user authentication with commercial Wi-Fi rely chiefly on one-off activity for user authentication. This makes it vulnerable to the zero-effort attacks and the imitation attacks. In contrast, the continuous activities offer more sufficient temporal information of activities [23], [24], exhibit more substantial spatial and temporal dynamic relations, compensate for the inherent defect of Wi-Fi, and can be utilized for a more precise portrayal of various gesture patterns. Therefore, leveraging continuous gestures can definitely contribute to behavior biometrics-based user authentication. Currently, most segmentation research uses threshold-based segmentation methods. However, the threshold is an empirical value and can not adapt to dynamic changes. Therefore, it is challenging to segment the continuous gestures accurately into a sequence of atomic ones.

Finally, it is challenging to derive a high-quality activity-induced Wi-Fi signal variation, which is the fundamental guarantee for effective activity recognition and user authentication. The reasons are the following. Firstly, the original sensing signal is weak. As illustrated in [25], Wi-Fi sensing captures information only from the weak reflection signals and the subtle movement-induced signal variations that can be easily buried in noise. Secondly, CSI (Channel State Information) dynamics fluctuate even in a static environment without human movements since Wi-Fi signals are susceptible to surrounding electromagnetic interferences. Thirdly, due to the imperfections of the commercial wireless network card, the Wi-Fi signal contains a lot of noise, such as the impulse noise in CSI amplitude, the random offset in CSI phase, and the measurement noise, etc. Therefore, it is challenging to derive the effective activity-induced CSI dynamics without an appropriate physical model between the signal fluctuation and the mobile subjects. However, this critical issue is overlooked by current behavior biometrics-based authentication systems.

To address these challenges, we propose a cross-layer user authentication framework called Wi-CGAuth. Wi-CGAuth achieves effective user authentication from the bottom signal layer up to the top user authentication layer. Compared with learning-based approaches, such as the transfer learning-based approach, Wi-CGAuth can seamlessly integrate the model-based approach into the learning-based approach and enhance the movement-induced signal quality effectively by the layering framework. At the lowest signal layer, the best signal quality of the wireless Wi-Fi is fully exploited by a subcarrier-level joint optimization to leverage the signal's time, space, and frequency diversity. Specifically, the conjugate multiplication is applied to the raw Wi-Fi signals to eliminate the phase and amplitude errors led by the network card imperfections. The highest sampling rate and maximum number of antennas are adopted to maximize the signal's time and spatial diversity. The maximum frequency diversity gain is achieved by combining received signals from all the antennas as well as the subcarriers and conducting the corresponding sub-carrier level alignment. At the middle layer, the sensing generalization capability and signal quality are significantly enhanced through cross-layer signal optimization. Wi-Fi signals after the processing at the lowest layer are transferred to the middle layer for the further enhancement. Specifically,the multi-view fusion methodology, implemented as multi-transfer component analysis (Multi-TCA), aligns features from multiple transceivers to eliminate perception discrepancies across different views.The signals after the multi-view fusion are

partitioned. The problem of continuous gesture segmentation is formulated as a classification task implemented with CNN. These refined and partitioned signals are further transferred to the upper layer. Finally, at the upper layer, a Convolutional Neural Network-Transformer (CNN-Transformer) dual-task model is adopted to achieve precise gesture recognition and reliable user authentication. Accurate user authentication is achieved by extracting the unique behavior biometric features when gestures are performed. CNN effectively captures spatial dependencies in spectrograms through relevant convolution layers and pooling layers. The transformer model can effectively capture long-range temporal dependencies and accurately discern complex patterns as well as features by leveraging self-attention mechanisms. Therefore, the transformer is employed to segment the feature maps generated by CNN for effectively handling sequential relationships and extracting the feature maps of the behavioral characteristics.

In summary, through cross-layer collaboration, sensing generalization capabilities of Wi-CGAuth are enhanced from various aspects. The comprehensive evaluation results indicate that Wi-CGAuth has achieved significant improvements over state-of-the-art, the most effective system WiHF (6% average authentication accuracy improvement in three typical indoor scenarios). This validates the effectiveness and robustness of Wi-CGAuth, demonstrating its potential for practical.

The contribution of this research can be summarized as following:

- The cross-environment user authentication framework is proposed, achieved from the lower signal to the upper signal level. The signal's time, frequency, and space diversity are maximized at the lower signal level. Then, the multi-perspective information fusion is applied to the calibrated signals to enhance signal quality at the middle layer. Finally, the gesture recognition and user authentication dual-task are executed at the upper level. To the best of our knowledge, this research is the first cross-layer framework to realize cross-environment user authentication.
- A user authentication method and gesture recognition method based on continuous gestures is proposed, and effective gesture segmentation is implemented by modeling the gesture segmentation to the classification problem.
- Wi-CGAuth is evaluated by extensive experiments. Its superiority in various indoor environments demonstrates its effectiveness and robustness.

To better illustrate the effectiveness of Wi-CGAuth, remaining sections of the paper are as followings: Sec. II introduces the preliminary of Wi-CGAuth. Sec. III presents the Wi-CGAuth, an identity authentication system based on continuous gestures with Wi-Fi. Sec. IV introduces the prototyping of Wi-CGAuth and details the extensive experiments conducted. Sec. V surveys currently proposed Wi-Fi-based human behavior recognition and identity authentication methods. The paper concludes in Sec. VI.

II. Related Work

A. Wi-Fi-Based Human Recognition

In contrast to cameras [26], [27], [28], wearable sensors [29], [30], [31], and ambient floor sensors [32], [33], Wi-Fi has emerged as a promising sensing modality due to its pervasiveness, non-invasiveness, extensive coverage, minimal deployment requirements, and absence of privacy violations. Camera-based human behavior sensing often faces privacy, lighting, and obstruction issues, limiting its applications in certain scenarios [26], [27]. Human behavior sensing using ambient floor sensors is constrained by limited sensing range, susceptibility to environmental and obstruction effects, and high maintenance and deployment costs [29], [31]. Similarly, wearable sensors, while effective in direct measurements, present challenges such as user discomfort, privacy concerns, short battery life, and data accuracy limitations [32], [33].

With advancements in Wi-Fi technology, various prospective applications have been explored, including respiration monitoring [34], [35], [36], gesture recognition [10], [17], [18], gait recognition [22], [37], [38], [39], indoor localization [40], and identity authentication [10], [13], [41], [42]. Among these, Wi-Fi-based identity authentication holds significant promise. By analyzing Wi-Fi signal variations caused by user behavior, automatic identity verification can be achieved without the need for passwords or physical tokens, offering both convenience and enhanced security.

B. Wi-Fi-Based User Identity Authentication

Recent studies on Wi-Fi-based user authentication rely on human behavior features such as gestures [10], [17], [43], gait [38], [39], and respiration [44]. These methods often attempt to address performance degradation caused by environmental changes by extracting domain-independent features [10], [17], [45]. The recently proposed user authentication solution uses fingerprints, voices, and other personal identifiers, requiring wearable devices [46] or specialized sensors [7], [47], incurring additional costs and inconvenience. With the rapid development of Wi-Fi infrastructure, researchers are beginning to explore Wi-Fi-based user authentication.

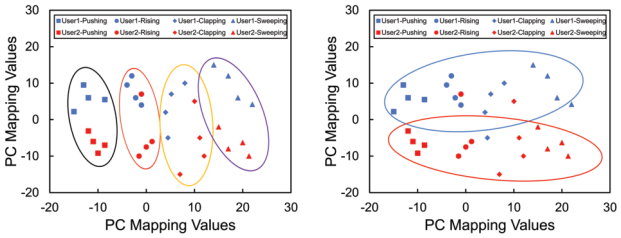
Several Wi-Fi-based identity authentication systems have been proposed in recent years. In WiID [12], predefined gestures are utilized for user authentication and gesture recognition. In [13], adversarial learning is employed to recognize individuals without relying on specific gestures. In [9], an adversarial network is used to remove environmental factors and achieve behavior-based authentication. WiHF [10] leverages motion change patterns and a deep neural network (DNN) to perform both identity authentication and gesture recognition. Similarly, FingerPass [11] continuously authenticates users through finger gestures using Wi-Fi signals' channel state information (CSI).

Despite these advancements, existing Wi-Fi-based identity authentication systems face three major limitations. First, they often rely on a single predefined activity for authentication, which limits flexibility. Second, they fail to adequately address cross-environment sensing generalization, leading to degraded performance in varying environments. Third, many systems conflate user authentication with activity recognition, resulting in suboptimal identity verification methods.

To address these limitations, we introduce Wi-CGAuth, a cross-environment, continuous gesture-based user authentication framework. By integrating multi-layer optimizations, including signal preprocessing, multi-view fusion, and dual-task modeling, Wi-CGAuth achieves robust and accurate authentication across diverse environments while overcoming the limitations of existing systems.

III. Preliminary

The existence of the behavioral uniqueness are verified in research work [6], [7], [8], [9], [10], [11], [12], [13], [14],



(a) Distinguish the gesture (b) Distinguish the user

Fig. 1. Distribution of four gestures performed by two users.

[15]. As demonstrated in these research, there are extrinsic and intrinsic behavioral uniqueness. A behavior performed by limbs and torso suits the person's physiology. Therefore, behaviors are always constrained by extrinsic human physiological characteristics (e.g., the length of limbs, the power generated by limb movements). These extrinsic physiological characteristics induce the behavioral uniqueness for different people. For example, people with different muscle masses perform behaviors in different accelerations and velocities, resulting in their behavioral uniquenesses. Hence, the behavioral uniqueness is determined by the extrinsic physiological characteristics of each person. Different from the extrinsic behavioral characteristics, the intrinsic physiological characteristics are behavior-independent, i.e., such features remain static for a specific individual no matter what kind of the behavior performed. The intrinsic physiological characteristics relate more to the inborn physical and biochemical functions of people, so they hardly change in different behaviors, which induce the invariant intrinsic behavior uniqueness. Experiments are conducted to ascertain the distinctive physiological characteristics of different users by investigating their respective underlying statistical distributions. Two participants are asked to perform four basic gestures: pushing, rising, sweeping, and clapping.

The principal component analysis (PCA) technique is utilized to characterize the underlying statistical patterns to check whether each gesture has unique and user-related signal characteristics. Fig. 1 illustrates the distribution of two principle components from four gestures performed by two participants. The x-axis and y-axis represent the first principal component and the specifically filtered PCA component, respectively. As shown in Fig. 1(a), various users' gesture-induced CSI data have a similar statistical distribution when performing the same gesture. This indicates that the various gestures are separable through Wi-Fi sensing. As presented in Fig. 1(b), a user's gesture-induced CSI data have a similar statistical distribution when performing different gestures. This indicates that the various users exhibit significant divergence in their physiological characteristics, and various users are distinguishable through Wi-Fi sensing. This inspires us to build a user authentication framework based on user gestures.

IV. System Design

Wi-CGAuth is a system to utilize continuous gestures for user authentication with commercial Wi-Fi, as depicted in Fig. 2. Wi-CGAuth comprises five modules. The continuous gestures data collection module collects continuous CSI measurements. The data noise reduction and enhancement module makes efforts to fully extend the Wi-Fi's sensing capability and improve the signal quality. The Multi-view data fusion module conducts joint learning, fuses the features from various views, and eliminates the difference between different receiver views. The continuous gesture segmentation problem is transferred to a classification problem by the continuous gesture segmentation module to achieve accurate segmentation. In the user authentication & gesture recognition module, a CNN-Transformer dual-task model is adopted for user identity authentication and gesture recognition.

A. Noise Reduction and Signal Enhancement

In the raw CSI data, there are amplitude impulse noise, phase offset, and environmental noise, which influences sensing results [39], [40]. To address these issues, conjugate multiplication is applied to the data collected between every two pairs of transmitter-receiver antennas. Therefore, the random phase offset is eliminated while the characteristics of the original signal are eliminated. To enhance the dynamic signal strength, we propose a novel, subcarrier-level, cost-effective optimization strategy, which involves a series of processing steps for the data from each antenna pair after conjugate multiplication. These processes include removing static vectors, normalization, identifying the optimal rotation angle, and aligning the initial phases of other subcarriers accordingly. By conducting multidimensional signal joint optimization at the subcarrier level, we effectively improve the overall signal quality from the aspect of the underlying signal.

1) Denoising the Original Signal: The CSI values at time t are represented in APPENDIX A. A Wi-Fi network interface card (NIC) is typically equipped with several antennas (such as the AX210 Wi-Fi card, each with two antennas). The time-variant phase offsets remain consistent across various antennas since all antennas use an identical RF oscillator [48], [49]. Conjugate multiplication [40] is applied to the CSI streams of two antennas on the same NIC to remove the random phase offsets. The details are in APPENDIX B.

With increased distance between the subject and transceivers, sensing ability decreases. Therefore, a novel, microscopic, cost-effective, subcarrier-level optimization strategy is introduced to enhance overall sensing capability based on the previous study [50].

The initial idea comes from the Khinchin theorem of large numbers. Assuming that $X_n, n = 1, 2, 3, \ldots$ are independently and identically distributed random samples. With an increasing sample size n, the mean approaches the expected value $E[X]$. $E[X] : \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{P} E[X]$. The frequency diversity is improved as 57 subcarriers in 20MHz is leveraged. The spatial diversity is enhanced through the utilization of the available antenna arrays. The time diversity is improved as the maximal sampling rate is reached. The environmental noise is independent random, so are the dynamic components. Therefore, the Khinchin theorem of large numbers is utilized to enhance overall signal quality.

2) Single Subcarrier Noise: As indicated in [51], environmental noise of CSI follows a Gaussian distribution with a mean of zero. CSI consists of a real part (I) and an imaginary part (Q). Following conjugate multiplication, environmental noise from the CSI signal is separated to check the normality of I and Q. The data in a static environment without dynamic vectors are collected. Furthermore, static vectors are constants, subtracted to obtain the environmental noise. In a static scenario, sufficient CSI signals are collected (we empirically collect it for 30 seconds at a frequency of 1000 Hz). The static component is obtained by computing the

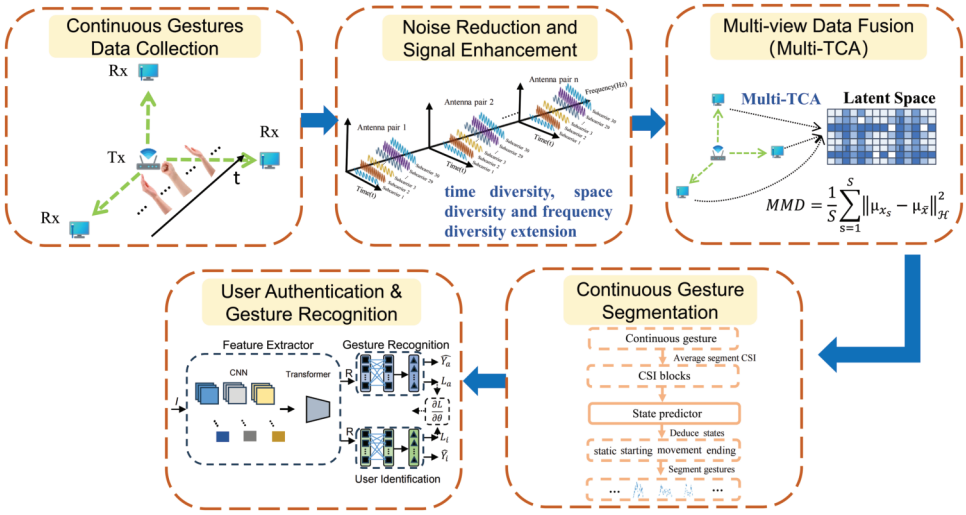


Fig. 2. An Overview of Wi-CGAuth. It comprises five components: continuous gestures data collection, data noise reduction and enhancement, Multi-view data fusion, continuous gesture segmentation, and user authentication & gesture recognition.

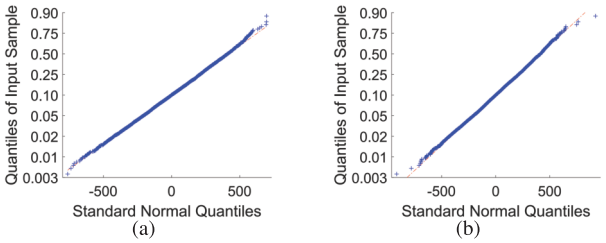


Fig. 3. The quantile-quantile (Q-Q) plot of I/Q components.

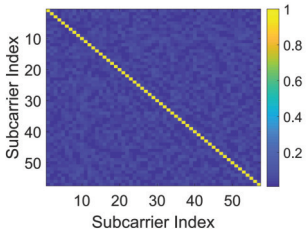


Fig. 4. Cross-correlation test shows environmental noise is independent across subcarriers.

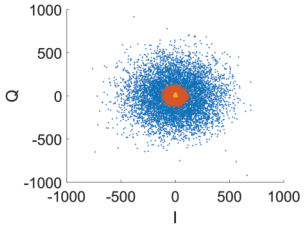


Fig. 5. The variance of environmental noise can be reduced with more samples averaged.

average, and it is subtracted to derive residual environmental noise. The normality of both components of environmental noise is validated. The quantile-quantile (Q-Q) is plotted for a given subcarrier's normalized and standard normal distributions in a static environment. As depicted in Fig. 3, I/Q components of environmental noise closely approximates a normal distribution. Next, how to utilize Wi-Fi signals' time, spatial, and frequency diversity to eliminate environmental noise is introduced.

3) Mitigating CSI Environmental Noise Through Time Diversity: Once confirming the normality of the environmental noise about I/Q components, environmental noise is mitigated by overlaying successive samples in the temporal domain. Given that the environmental noise in I/Q components of the CSI after conjugate multiplication has the mean 0 and the variance $\sigma^2(f)$, the variance can be reduced by gathering more samples. In this way, environmental noise can be suppressed.

4) Mitigating CSI Environmental Noise Through Space and Frequency Diversity: As discussed in above section, the noise level can be decreased by combining the environmental noise of one subcarrier in various instants. While the environmental noise of various subcarriers has identical expectations of 0, the variances might vary. Hence, the distribution of environmental noise across various subcarriers is not entirely identical. Based on Kolmogorov's strong law of large numbers, let X_1, X_2,...be independent with means μ_1, μ_2,...and variances σ_1^2, σ_2^2,...under such conditions $\sum_{k=1}^{\infty}\frac{\sigma_k^2}{k^2} < \infty$. Then $\frac{X_1+X_2+\cdots+X_n-(\mu_1+\mu_2+\cdots+\mu_n)}{n} \xrightarrow{\text{a.s.}} 0$. Therefore, when X_i satisfies the following three conditions: (1) X_i are mutually independent, (2) expectations of X_i exist, and (3) the variance of X_i is bounded by a finite value, all samples' average converges towards the expectations average.

In a static scenario, conjugate multiplication is performed with one transmitting antenna and two receiving antennas crossing 57 subcarriers, resulting in 2×57 results (where $A_2^1 = 2$). Following the previous section, environmental noise is derived by various subcarriers. For satisfying condition (1), the input signals must be mutually independent. Fig. 15 presents the correlation output of 57 subcarriers, demonstrating the independence of environmental noise across various subcarriers. Furthermore, a distinct converging trend is found in both the noise mean and variance, satisfying condition (2) and condition (3). Therefore, the Kolmogorov's strong law of large numbers can be applied to to mitigate noise. To prove the method's validity above, Fig. 5 depicts the noise distributions before (represented by blue point) and after the combination (represented by red point). The noise level can be greatly reduced with the signals combination of the various subcarriers in various antennas. In addition, the temporal combination in

the previous section can also be used additionally. To demonstrate this, time diversity with the derived combined noise data is utilized to compute the average over the contiguous 50 samples. As illustrated by bright color points in Fig. 5, it is obvious that noise is significantly depressed.

In sum, the analysis demonstrates noise levels can effectively be mitigated by signals' time, space, and frequency diversity extension. A novel, subcarrier-level, cost-effective signal optimization strategy is proposed.

5) Aligning Dynamic CSI: In the previous section, when the environment is static, the noise is reduced by combining multiple CSI data at numerous subcarriers, antennas, and higher sampling rates. In a real environment with a motion subject, the CSI after conjugate multiplication comprises static components, dynamic components, and environmental noises. The CSI data after conjugate multiplication are combined from antenna pairs and subcarriers, static components (① in Eq. 10) remain unchanged, the noises (③ in Eq. 10) is mitigated according to the study above. However, it's not guaranteed the dynamic components (② in Eq. 10) can achieve maximum simultaneously. Due to different wavelengths, combining the dynamic components simply from various subcarriers may interfere mutually, resulting in a destructive dynamics signal.The primary factor causing phase differences among dynamic vectors from different subcarriers is the wavelength.The detailed derivation of the phase differences among subcarriers and antennas is in the APPENDIX C.

To effectively enhance the dynamic signal strength and depress environmental noise, the CSI initial phases are rotated and aligned after the conjugate multiplication of each subcarrier from all antenna pairs after conjugate multiplication. When two CSI signals after conjugate multiplication are identical, they are in coincidence with each other in the I-Q plane. Motivated by this insight, the minimum distance sum of two CSI signals after conjugate multiplication is utilized to find the optimal alignment angles. Consequently, numerous data must be utilized to synchronize the phase of the dynamic components, allowing the distances between noise contributions from two signals to converge to the minimum. The rotation angles of the dynamic components are obtained.

The alignment algorithm after CSI conjugate multiplication consists of three steps: (1) Elimination: Eliminate the static components' impact. (2) Normalization: Normalize the CSI signals for every subcarrier on all antennas. (3) Alignment: Select one subcarrier's CSI signal after conjugate multiplication as a reference and align the remaining to it, ensuring identical initial phases.

a) Elimination: The static components are depressed since it has no relationship to the subject gesture. Given a sufficient number of samples *T*, the mean of a CSI signal after conjugate multiplication is represented in APPENDIX D.

b) Normalization: A moving mean method by the window length of T' is applied to the CSI data after conjugate multiplication to compute the results. It is formulate in APPENDIX E.

c) Alignment: One CSI signal after conjugate multiplication is chosen as the reference, and others are aligned towards it. The optimization function is developed to calculate the distances among various subcarriers and iterates through every possible angle to find the finest one.It is formulated in APPENDIX F. Since overall subcarriers share an identical initial phase and contribute closely equally, the signals sampled from each subcarrier contribute to the overall signal enhancement. Regarding noise, its power declines as the number of subcarrier samples rises. In the way, the environmental noise is depressed and the movement induced CSI strength is increased without changing the phase difference.

6) Combine Subcarriers, Antennas, and Temporal Together: Two efforts are made to improve signal quality: 1) Synchronizing phases of dynamic components across all antennas and subcarriers; 2) combining as many samples as possible. An experiment is conducted to demonstrate the validation of utilizing the temporal, space, and frequency diversity.

An experiment is conducted using computers with AX210 NICs in an empty hall, as depicted in Fig. 13(a). The sending device is equipped with one omnidirectional antenna, and the receiving device is equipped with two omnidirectional antennas. The distance between the transceivers measures 3 meters. A subject makes a circle gesture. There are $A_2^1 = 2$ kinds of combinations to calculate conjugate multiplication from the two antennas. For multiple subcarriers, the CSI signals after conjugate multiplication can be utilized. The CSI sampling rate increases to $2000Hz$.

Fig. 6 shows the CSI signal variation after conjugate multiplication by time, space, and frequency diversity. Five kinds of variation processing are presented: (a) the CSI dynamic after conjugate multiplication and phase alignment, (b) signal combinations with various antennas, (c) signal combinations with various subcarriers, (d) signal combinations with various antennas and subcarriers, and (e) signal combinations with various antennas and subcarriers at a higher sample rate (2000Hz). As the sample numbers for combinations increase, the noise is significantly reduced, and the signal quality is greatly improved.

In [40], conjugate multiplication is primarily employed to eliminate the phase shift error $e^{-j2\pi\frac{d(t)}{\lambda_f}}$ caused by the imperfection of the Wi-Fi hardware. In [50], the method is proposed to reduce measurement noise and enhance the quality of the Wi-Fi signal by fully exploiting the Wi-Fi signal diversity. A denoising method is proposed to reduce the phase shift error and measurement noise as well as enhance signal quality by combining the approaches proposed in [40] and [50]. The sole goal of the signal processing method in the paper [40] and [50] is to enhance movement-induced signal quality at only the lowest layer. In other words, the model-based approaches try to enhance the movement-induced signal quality by establishing the physical models at the lowest layer. However, this research proposes a cross-environment continuous gesture-based user authentication framework with Wi-Fi called Wi-CGAuth. By layering framework design, Wi-CGAuth seamlessly integrates the model-based into the learning-based solutions and makes the cross-environment sensing more effective. It has four layers. As an indispensable component of the framework, the model-based signal processing method works at the lowest layer and plays a significant role in serving its upper layer. The outputs of the lowest layer are the essential inputs for its upper layer. The behavioral based user authentication is achieved by layer-by-layer cooperation and cross-layer joint optimization.

After the subcarrier-level signals' time, spacial, and frequency diversity optimization, the various noises of raw Wi-Fi signals are mitigated effectively, and a high-quality activity-induced Wi-Fi signal variation is derived. In order to further enhance the cross-environment sensing generalization capa-

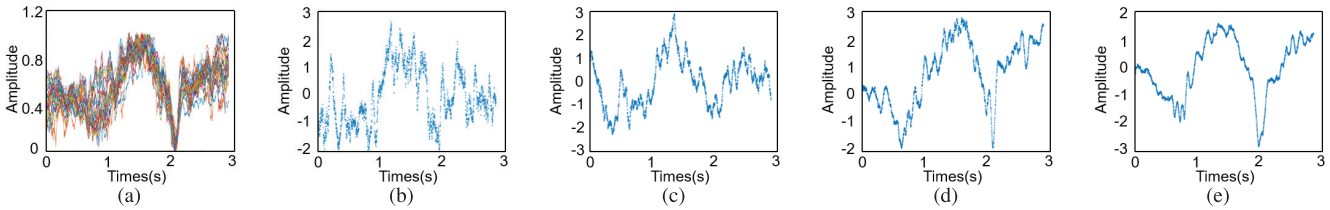


Fig. 6. The CSI signal variation by time, space, and frequency diversity. (a) the CSI dynamic after conjugate multiplication and phase alignment, (b) signal combinations with various antennas, (c) signal combinations with various subcarriers, (d) signal combinations with various antennas and subcarriers, and (e) signal combinations with various antennas and subcarriers at a higher sample rate (2000Hz).

bility and signal quality, the multi-view fusion method, i.e., multi-transfer component analysis (multi-TCA), is leveraged to refine the signals after the preprocessing.

B. Multi-View Data Fusion (Multi-TCA)

The TCA (Transfer Component Analysis) algorithm [52] is used for domain adaptation of both source and target domains to improve the model's generalization. The core of the TCA algorithm is to find a transfer matrix that maps data of both domains to a distribution and learn a feature representation with better generalization by minimizing the distribution differences between domains.

Traditional TCA algorithm [52] is typically designed for source and target domains. Different sensing device pairs sense the same activity in various aspects. The existing Wi-Fi based Multi-view framework [53], focuses on static target through Wi-Fi imaging, which is not in practical real-world scenario. In this work, the continuous gestures are sensed from different perspectives with various Wi-Fi device pairs. This is modeled as a multi-view fusion problem. Each transceiver pair serves as an independent perspective, and the Multi-TCA algorithm is employed for multi-view fusion. The Multi-TCA approach extends the traditional TCA algorithm from two domains to multiple domains.By applying multi-TCA, the sensing data from multiple receivers are projected to a shared subspace. The movement pattern dynamics across various data sources are preserved maximally. At the same time, the irrelevances across various data sources are minimized.

The application condition for Multi-TCA is when $P\left(X_s\right) \neq P\left(X_t\right), 1 \leq s < t \leq U$, where X_s, X_t represents the perspective dataset, $P(X_s)$ denotes the probability distributions of X_s, and U indicates the numbers of datasets for all perspectives. In Multi-TCA, the objective is to discover the features mapping satisfying $P(\phi(X_s)) \approx P(\phi(X_t))$. Assuming ϕ represents a feature mapping generated by the universal kernel. The Maximum Mean Discrepancy (MMD) is a statistical metric calculating differences in both two probability distributions of two perspectives in the Reproducing Kernel Hilbert Space (RKHS). Two sensing perspectives are extended to multiple perspectives as followings.

$$\mathrm{MMD} = \frac{1}{S}\sum_{s=1}^{S}\left\|\mu_{x_s} - \mu_{\bar{x}}\right\|_{\mathcal{H}}^2 \tag{1}$$

where $\mu_{x_s} = \frac{1}{n_s}\sum_{i=1}^{n_s}\phi\left(x_{si}\right)$, $\mu_{\bar{x}} = \frac{1}{S}\sum_{s=1}^{S}\mu_{x_s}$, n_s represents the sample numbers in X_s, x_{s_i} denotes the i-th sample in X_s, S denotes the total numbers of all perspectives, $\|\cdot\|_{\mathcal{H}}^2$ indicates RKHS norm. Assuming K is a Gram matrix [54] that integrates cross-domain data from all perspectives $X_1, X_2, \ldots, X_S$.

$$K = \begin{bmatrix} K_{X_1,X_1} & K_{X_1,X_2} & \cdots & K_{X_1,X_S} \\ K_{X_2,X_1} & K_{X_2,X_2} & \cdots & K_{X_2,X_S} \\ \vdots & \vdots & \ddots & \vdots \\ K_{X_S,X_1} & K_{X_S,X_2} & \cdots & K_{X_S,X_S} \end{bmatrix} \in \mathbb{R}^{N\times N} \tag{2}$$

where $N = \sum_{s=1}^{S} n_s$, $K_{i,j}$ is calculated by $\phi(x_i)^T\phi(x_j)$. Based on the derivation, *MMD* in Eq. 1 is reformulated as $tr(KL)$. $L_{i,j}$ is expressed as follows:

$$L_{i,j} = \begin{cases} \dfrac{S-1}{N^2 n_s^2} & x_i, x_j \in X_s \\ -\dfrac{1}{N^2 n_s n_u} & x_i \in X_s, x_j \in X_u \text{ and } s \neq t \end{cases} \tag{3}$$

where $s, t \in \{1, 2, \ldots, S\}$. A parameterized kernel mapping $K = (KK^{-1/2})(K^{-1/2}K)$ is utilized to help solve computationally intensive semidefinite programming. The kernel matrix $\tilde{K} = KWW^TK$, $W \in \mathbb{R}^{N\times m}, m \ll N$ is identified a transformation matrix in TCA [52]. Here, the *MMD* distance in Eq. 1 is transformed as: $MMD = tr((KWW^TK)L) = tr(W^TKLKW)$.

Another goal of Multi-TCA is to maximize sensing data variance to ensure sufficient diversity among different sensing perspectives to capture the feature differences better. In other words, it helps to improve the generalization, enabling the model to update more effectively to different data distributions in the target perspective. By maximizing data variance, Multi-TCA can better address feature alignment and domain adaptation challenges in multiple perspectives, providing enhanced feature representation for sensing. The variance of the projected samples in transformed feature space is denoted as W^TKHKW, where $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ and represents centering matrix. Here, $\mathbf{I} \in \mathbb{R}^{N\times N}$ denotes an identity matrix, and $\mathbf{1} \in \mathbb{R}^N$ represents a column vector consisting of all ones. The intrinsic characteristics and differences between various perspectives are captured by maximizing the variance.

The objective function of Multi-TCA is represented as follows: $\min_W \mathrm{tr}\left(\mathrm{W^TKLKW}\right) + \mu\mathrm{tr}\left(\mathrm{W^TW}\right)$, s.t. $\mathrm{W^TKHKW = I}$ by introducing adjustment terms $tr(W^TW)$ and weighting parameters μ. W^TK represents the embedding of data in the latent space, and W is determined by $m \ll N$ dominant eigenvectors of $(KLK + \mu I)^{-1}KHK$. Here, $\mu > 0$ is served as a weighting parameter that controls the complexity of W.

To model the statistical distributions of each perspective's CSI induced gestures before and after multi-TCA, the Principle Component Analysis (PCA) is applied derives the correlations between different CSI sequences and exhibits the principal

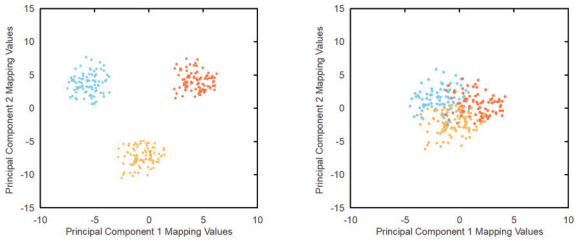


Fig. 7. The statistical distributions of sensing data from three perspectives(top, left, right) before and after using Multi-TCA. (a) represents the statistical distributions from the top, left, and right perspectives before using Multi-TCA. (b) represents the statistical distributions from the top, left, and right perspectives after using Multi-TCA.

components with minimum redundancy. This can exhibit the statistical distributions clearly under CSI sequences of each perspective. As presented in Fig. 7, before applying multi-TCA, the movement patterns statistical distributions at different perspectives are various significantly. After applying multi-TCA, the movement patterns statistical distributions at various perspective have some consistency. By applying multi-TCA, the sensing data from multiple receivers are projected to a shared subspace. The movement pattern dynamics across various data sources are preserved maximally. At the same time, the irrelevances across various data sources are minimized and the sensing data variance is maximized.

Overall, the multi-TCA further enhances the cross-environment sensing generalization capability and signal quality. As illustrated in [23] and [24], the continuous activities offer more substantial spatial and temporal dynamic relations of activities and compensate for the inherent defect of Wi-Fi. Therefore, continuous gestures instead of one-off activity can definitely benefit the behavior biometrics-based user authentication. The problem of continuous gesture segmentation is modeled as a classification task implemented with CNN.

C. Continuous Gesture Segmentation

Continuous gestures contain more temporal and spatial relation of motions, improving the accuracy of the authentication process. Currently, most Wi-Fi-based researches on continuous action segmentation rely on threshold-based segmentation methods, where the significant fluctuation in the CSI amplitude beyond a certain threshold is considered as an activity appearance. Although threshold-based segmentation methods can be effective in specific scenarios, they have a practical issue. Threshold is empirical value and can not adapt to the dynamic changes.

A CNN-based activity segmentation approach is introduced to address these issues based on previous work [55]. Specifically, the gesture states are regarded as classes and the continuous gesture segmentation task is transformed to a classification problem. Initially, the continuous CSI stream is divided into equally sized blocks. Then, a trained state predictor classifies these blocks into four distinct states: static, starting, movement, and ending. Finally, all block states are utilized to identify the gestures' starting and ending points.

1) State Predictor: The state predictor is the classifier, which is trained with the labeled state blocks from single activity data. The received continuous CSI series is divided into equally sized blocks, defined by four state labels, static state, starting state, movement state, and ending state. The static/movement state signifies whether the CSI data indicates the absence or presence of a gesture within this block. The starting/ending state means the CSI data contains an activity's starting/ending point in this block.

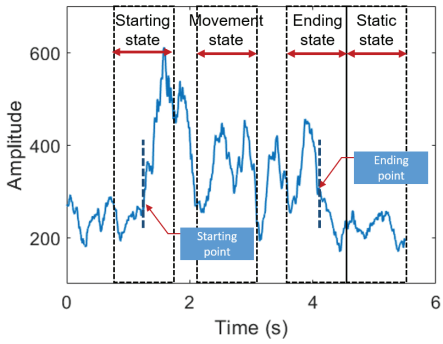


Fig. 8. Four kinds of states of continuous gesture.

Fig. 8 plots the waveform of the consecutive gestures. In both the starting and ending states, one half is the non-gesture segment, while the other half corresponds to the gesture segment. Conversely, the static state solely encompasses the non-gesture segment, while the motion state solely encompasses the gesture segment. For the static, starting, movement, and ending states, the starting points are $t_{end} + w/2$, $t_{start} - w/2$, $t_{start} + t_{end} - w/2$, and $t_{end} - w/2$, respectively, and the ending points are $t_{end} + w/2 + w$, $t_{start} + w/2$, $t_{start} + t_{end} + w/2$, and $t_{end} + w/2$, respectively. t_{start} and t_{end} denote the actual starting and ending points of gesture, while w denotes the window size. Following the definition, each CSI series about gesture can produce four blocks for training the state predictor.

The state predictor takes a CNN model, which is characterized as follows:

$$\mathrm{Y} = \mathrm{CNN}(\mathrm{X}) \tag{4}$$

where X represents the CSI series after the data caliberation, while Y represents fully connected layer output. The CNN model consists of the convolutional layer, dropout layer, and max-pooling layer. W_f and b_f denotes convolutional parameters. The prediction probability $p(r \mid x; \Phi)$ can be presented as followings:

$$p(r \mid x; \Phi) = \mathrm{softmax}(\mathrm{W}_r * Y + b_r) \tag{5}$$

where W_r and b_r represent fully connected layers parameters, $\Phi = \{\mathrm{W}_f, b_f, \mathrm{W}_r, b_r\}$, and objective loss with cross-entropy loss can be expressed as followings:

$$\mathcal{J}(\Phi) = -\frac{1}{|X|}\sum_{i=1}^{|X|} \log p\left(r_i \mid x_i; \Phi\right) \tag{6}$$

The Adam optimizer with default hyperparameters is employed as the optimization method.

2) Determining Gestures' Starting and Ending Points: Upon completing the training of state predictor, gesture motions can be separated from continuous CSI series by the following stages:

(1) Divide the CSI series into blocks employing a sliding window of size w and a sliding steps of 50.
(2) Label the blocks with the trained state predictor.
(3) Determine the starting and ending points of all gesture motions by considering the block states and the mode change, where the mode refers to the most frequently occurring numbers in the label list.

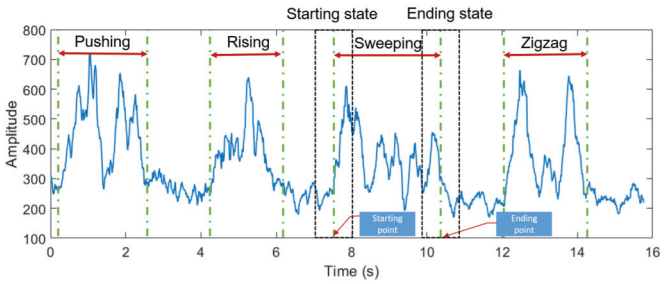


Fig. 9. Continuous gesture segmentation (pushing, rising, sweeping, zigzag) is implemented by a CNN model.

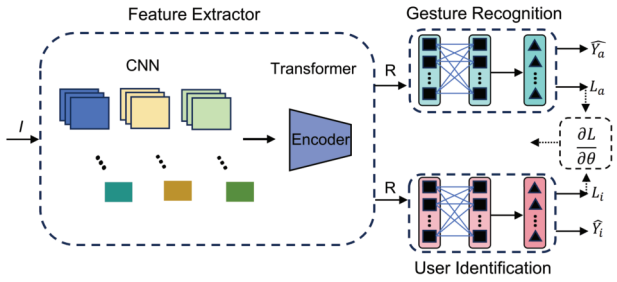


Fig. 10. The dual-task model based on CNN and transformer.

If there is a transition from the static state to the starting state in terms of the mode, the block is recognized as the initiation of a gesture. Conversely, once the transition is from the ending state to the static, the block is considered the ending of a gesture. The training samples are the segmented calibrated CSI series with size w. The algorithm begins by segmenting the input CSI series into blocks of size w. Subsequently, the trained state predictor is utilized to label these blocks, which are saved in the label sequence as $inferred\_label$. The algorithm traverses the entire $inferred\_label$ to identify all starting and ending points of all gestures. A window of length m is utilized to traverse the $inferred\_label$, and m denotes the size of the block label window used for mode calculation. The window's state is determined by the mode of all block labels within the window. When the number of appearing state are equal in a window, the state of the newly traversed block is taken as the window's state. The criterion to decide the starting of a gesture activity during the traversal is as follows: When the detection status for the starting point is undetected, and the mode transitions from static to the starting, the algorithm enables i-m/2+1 as the starting point of a gesture. Here, i denotes the index of the currently traversed $inferred\_label$. The criterion for determining the ending of a gesture is as follows: When the starting point of a gesture has been detected, and the mode transitions from the ending to the static, the algorithm enables i-m/2+1-w as the ending point of a gesture and updates the detection status for the starting point as undetected. The reason for subtracting w is that when the state mode transitions to static state, the current block is filled with static segments.

Fig. 9 displays the segmentation results of the CNN model for four consecutive gestures (pushing, rising, sweeping, zigzag). The starting and ending points of each gesture are obtained through mode transition. The CNN-based gesture segmentation algorithm avoids reliance on threshold determination, and makes the segmentation adopt to the dynamics. After segmenting the continuous gestures accurately into a sequence of atomic ones, these atomic gestures are input into the CNN-Transformer model to achieve user identity authentication and gesture recognition.

D. User Authentication & Gesture Recognition

The gesture-based identity authentication comprises two core tasks: user identity authentication and gesture recognition. Therefore, a dual-task model (CNN-Transformer) is developed.

Fig. 10 presents the CNN-Transformer architecture, including a shared feature extractor and two separate fully connected layers specialized for gesture recognition and user authentication. The feature extractor consists of three CNNs and one Transformer. The CNN component consists of convolution layer and pooling layer, where convolution layer makes input images into compressed representations while pooling layer reduces the dimensions of these compressed representations. The input consists of spectrograms of the gestures from various users, capturing fine-grained behavioral features from the pixel level. Transformer is employed to segment the feature maps generated by CNN for handling sequential relationships and extract feature maps *R* embedding behavioral characteristics of various users.

The user authentication and gesture recognition networks share a similar structure, which includes two fully connected layers and a softmax layer. With feature maps *R* generated by a feature extractor as input, both networks focus on distinct scales to extract higher features, facilitating both identity authentication and gesture recognition.

The identity authentication network generates user labels $\hat{Y}_i$ and user loss L_i as indicators of authentication errors. In parallel, the gesture recognition network generates gesture labels $\hat{Y}_a$ and activity loss L_a to indicate errors in gesture recognition. The combination of these two losses for the joint model training is as follows:

$$L = \alpha\left(L_a + m\right) + \beta e^{(L_i + n)} \tag{7}$$

where α and β represent the weights for gesture and user loss, *m* and *n* denote the biases of the both, respectively. Considering that user authentication demands more deep features compared to gesture recognition, an overall loss function is devised as an exponent function with a higher convergence priority for identity loss over gesture loss. By continually backpropagation gradients of the loss $\frac{\partial L}{\partial \theta}$, the model can derive more representative features for identity authentication and gesture recognition.

Wi-CGAuth not only authenticates legitimate users but also detects potential illegal users. Both subjective factors and objective physiological traits determine human behavioral characteristics. Therefore, even when illegal users attempt to imitate the extrinsic behaviors of legitimate users, their behavioral features exhibit noticeable differences. This makes it possible for the Wi-CGAuth to identify the illegal users. Specifically, Wi-CGAuth distinguishes each user by comparing every class of the identity probability Y_i^k to a predefined threshold Θ. If $Y_i^k < \theta$ holds for all $\forall k \in [1, n]$, Wi-CGAuth identifies the user as an illegal one.

V. Experimental Evaluation

A. Experimental Setup and Dataset

1) Experimental Setups: The PicoScense platform [56] with Ubuntu 20.04 LTS operating system runs on four Lenovo

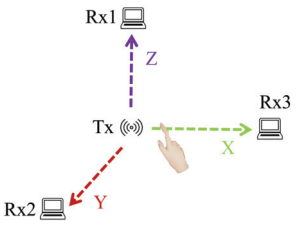


Fig. 11. 3D Transceiver Setup.

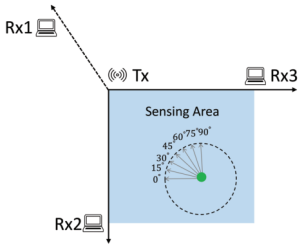
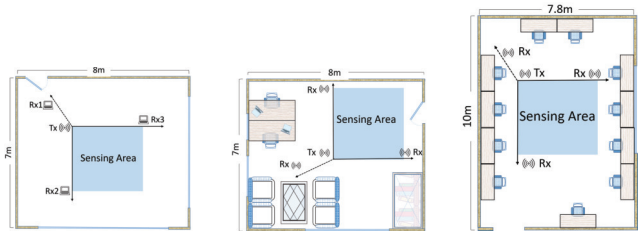


Fig. 12. User orientations.



(a) Empty hall. (b) Office. (c) Lab.

Fig. 13. Overviews of three experimental scenarios.

E73S desktop computers. Each computer is equipped with one AX210 NICs, each of which has two antennas.As illustrated in Fig. 11, one computer works as a transmitter, while others three serve as receivers. The gesture is performed in a 3D environment. Specifically, in a 3D environment, there is one pair of sensing devices for each mutual perpendicular direction X, Y, and Z.Rx1 is positioned 3m above ground. Tx, Rx2, and Rx3 are all placed 0.8m above the ground. The distance between Tx and Rx2 and the distance between Tx and Rx3 are 3m. This deployment can provide more spatial information and help to achieve cross-layer joint optimization more effectively. The Wi-Fi channel frequency is configured at 5 GHz on channel 165 with a bandwidth of 20 MHz. Each receiving antenna has 57 subcarriers, a total of 114 for each data stream.The default packet transmission rate is 2000 packets per second. Wi-CGAuth is tested in three indoor experimental scenarios, as depicted in Fig. 13 including an empty room with the size of $7m \times 8\ m$, an office with the size of $7m \times 8m$, and a lab with the size of $7.8\ m \times 10\ m$. The Kinect 2.0 camera is utilized to record the ground truth.

2) Experimental Dataset: 32 volunteers (20 males and 12 females) are recruited and do the experiment in three scenarios. The age of all volunteers varies from 18 to 55 years, heights from $155cm$ to $185cm$, and weights between $45kg$ and $80kg$. We randomly select 24 volunteers (15 males and 9 females) as legitimate users, the others as illegal ones. Both legal and illegal user groups maintain consistent distributions of age, height and weight. Six gestures are performed, including pushing, rising, sweeping, clapping, zigzag, and circle. During the experiment, each subject performs the gestures continuously within the sensing area, facing the Tx-Rx1 transceiver pair. The hand performing the gesture is in front of the user's face. All volunteers perform a set of continuous gestures(six gestures) 20 times in three typical experimental scenarios. The data from empty hall are divided into two sets: 80% for training and the remaining 20% for testing the in-domain accuracy.The data from office and lab are for testing the cross-domain accuracy. Legitimate user gesture data is employed to train user authentication, whereas illegal users only participate in testing. The final dual-task model, which achieves both user authentication and gesture recognition, is trained with the data collected from the empty hall and tested with data from the other two scenarios.

TABLE I

THE OVERALL PERFORMANCE OF IDENTITY AUTHENTICATION OF WI-CGAUTH IN THREE SCENARIOS

Scenario	Accuracy
Empty hall	93.56%
Office	92.51%
Lab	92.33%

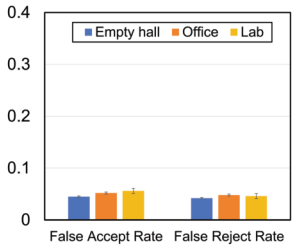


Fig. 14. False accept rate and false reject rate in three scenarios.

B. Overall Performance of Identity Authentication

Table I presents the overall authentication results of Wi-CGAuth in three typical experimental scenarios (Empty hall, Office, Lab), with average recognition accuracies of 93.56%, 92.51%, and 92.33%, respectively. These results demonstrate the excellent performance of Wi-CGAuth in various indoor environments. The continuous gestures makes it possible to improve authentication accuracy since the final authentication result is based on the probability. Wi-CGAuth maintains a high authentication accuracy across various scenarios, which verifies the effectiveness of the cross-environment continuous gusture based user authentication framework.

Fig. 14 presents the false accept rate (FAR) and false reject rate (FRR) of Wi-CGAuth in three experimental scenarios, with average accuracies of 5.1% and 4.5%, respectively. The false accept rate represents the probability that an illegal user is authenticated as a legal user, and the false reject rate represents the probability that a legal user is authenticated as an illegal user. Based on the experiment results, Wi-CGAuth demonstrates its capability in accurately identifying authorized and unauthorized users in various indoor settings.

C. Overall Performance of Gesture Recognition

Besides user authentication, Wi-CGAuth is capable of identifying the gestures conducted by the users. As shown in Fig. 15, the recognition performance of six gestures is conducted in three experimental scenarios(Empty hall, Office, Lab), with average recognition accuracies of 93.44%, 91.25%,

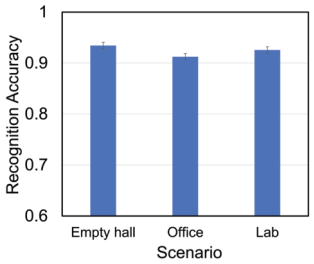


Fig. 15. The overall performance of gesture recognition of Wi-CGAuth.

TABLE II
ABLATION STUDY

Room	*Empty hall*	*Office*	*Lab*
Without Filter, With Multi-TCA	88.25%	85.32%	86.21%
With Filter, Without Multi-TCA	90.23%	81.68%	82.62%
Filter + Multi-TCA	93.56%	92.51%	92.33%

and 92.56%, respectively. Wi-CGAuth exhibits the best gesture recognition performance in the empty hall due to its simpler layout and less multipath effect. However, in the office and lab scenarios with relatively complex layouts, gesture recognition accuracy decreases but still maintains excellent recognition performance. This indicates that the CSI signal quality is enhanced, and cross-environment influence is mitigated by mitigating environmental noise through temporal, spatial, and frequency diversity and employing Multi-TCA. Additionally, the application of Multi-TCA significantly improves the system's robustness against cross-environment influences. To further validate this, we conducted ablation studies on Wi-CGAuth's core modules.

D. Ablation Study

The identity authentication results with and without the method of noise reduction and signal enhancement are shown in Table II, where the Filter replaces the noise reduction and signal enhancement. The method of noise reduction and signal enhancement contributes to improved average authentication accuracies, with 93.56%, 92.51%, and 92.33% achieved in the empty hall, office, and lab, respectively, compared to average accuracies of 88.25%, 85.32%, and 86.21% obtained without one. Experiment results confirm the effectiveness of the noise reduction and signal enhancement.

An ablation study about multi-TCA is conducted. The authentication performance with and without multi-TCA is evaluated in three scenarios, as presented in Table II. Multi-TCA contributes to improved average authentication accuracies, with 93.56%, 92.51%, and 92.33% achieved in the empty hall, office, and lab, respectively, compared to average accuracies of 90.23%, 81.68%, and 82.62% obtained without multi-TCA. Experiment results confirm the effectiveness of the multi-TCA method.

E. Performance Evaluation on Various Impacts

1) Comparison With Existing Approaches: The user authentication is achieved by the joint layer optimization. The user authentication is improved in various scenarios. In three typical scenarios, three existing gesture-based user authentication systems, i.e.,WiHF [10], FingerPass [11], and

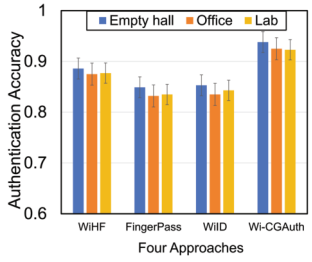


Fig. 16. The comparisons with others user authentication systems.

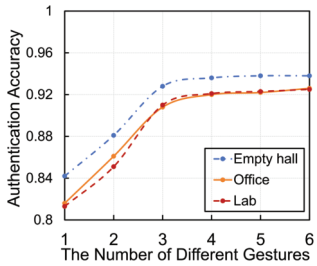


Fig. 17. The impact of distinct number of continuous gesture.

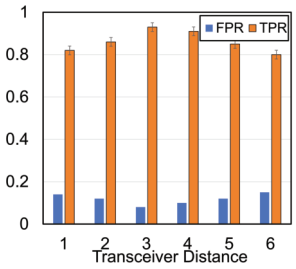


Fig. 18. The impact of various Line of-Sight (LoS) lengths(1m ∽ 6m).

WiID [12] are compares with Wi-CGAuth. The results present the average accuracy across scenarios. In Fig. 16, it is evident that Wi-CGAuth outperforms the other three systems.

2) Impact of Distinct Number of Continuous Gesture: Identity authentication is implemented using continuous gestures, and the number of various gestures during testing may impact the experiment results. During a test, each subject performs a group of gestures continuously, and this is conducted ten times in three typical scenarios. The number of gestures performed in a group varies randomly from one to six. Fig. 17 illustrates the identity authentication accuracy at various numbers of continuous gestures in three scenarios. The results indicate that Wi-CGAuth's authentication accuracy initially rises and stabilizes as the number of continuous gestures increases. Authentication results tend to be stable when there are more than three continuous gestures. Since user identity authentications are achieved by segmenting continuous gestures and selecting the category with the highest frequency, the impact on the system diminishes as the number of gestures exceeds three. Therefore three is the optimal number of gestures need to be performed.

3) Impact of Various Line-of-Sight (LoS) Lengths: At first, the default LoS length is set to 3m. Then, the LoS lengths varies from 1m to 6m, increasing by 1m each time. Fig. 18 illustrates the effects for various LoS lengths with a classifier threshold of 0.5. The results represent averages under various environments. The TPR initially rises, peaks at 3m, and then gradually declines as LoS length grows. Conversely, the FPR exhibits an opposite trend. FPR decreases as the LoS length

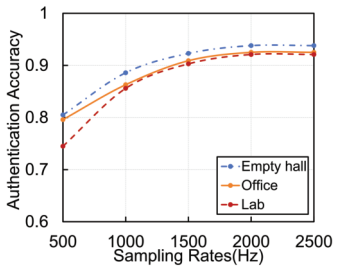


Fig. 19. The impact of various sampling rate(300Hz ∽ 2000Hz).

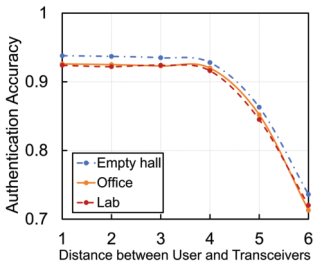


Fig. 20. The impact of user-toreceiving antennas distance.

increases, reaches its lowest value at 3m, and subsequently increases. In other words, Wi-CGAuth performs best at a distance of 3m. These findings validate the experimental results presented in [57].

4) Impact of Various Sampling Rate: Authentication performance is evaluated at various sampling rates to determine the optimal sampling rate for Wi-CGAuth. Fig. 19 displays the identity authentication accuracy under the various sampling rates in three scenarios. The results indicate that Wi-CGAuth's authentication accuracy initially rises and then stabilizes with an increasing sampling rates. Wi-CGAuth achieves over 92% authentication accuracy in all three environments as sampling rates near 2000Hz.

5) Impact of User-to-Receiving Antennas Distance: Since users perform gestures toward the transceivers for identity authentication, the performance of Wi-CGAuth at various distances between users and the receiving antennas is evaluated. The distance ranges from 1m to 6m, increasing by 1m each time. Fig. 20 presents the identity authentication accuracy at various distances between users and receiving antennas in three scenarios. When the distance changes from 1m to 4m, Wi-CGAuth maintains stable performance. However, authentication performance starts to decline as the distance continues to increase. When a user is are farther from the transceivers, the signal propagation distance increases, leading to reduced signal transmission power, elevated noise levels, and increased packet loss. The authentication performance can achieve up to 85%, even when the distance between users and receiving antennas is 6 meters. This satisfies the requirements of various applications.

6) Impact of Environmental Interferer: Additional experiments are conducted in the empty hall scenario to investigate the impact of the subject in the sensing environment. In these experiments, the subjects perform gestures within the sensing area while others are asked to randomly walk within a range of 2m to 6m from the transceivers. Fig. 21 indicates that the impact on authentication performance decreases as the interferers move farther from the transceivers. For instance, the authentication performance declines seriously when interferers move around the subjects (within 2m). When the interferers

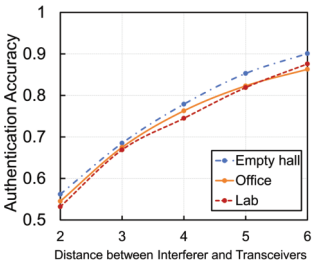
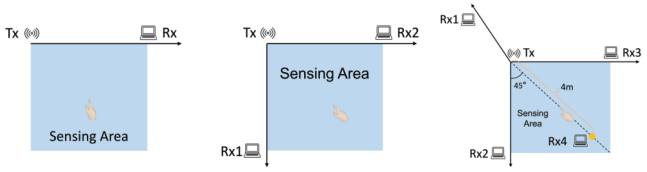


Fig. 21. The impact of environmental interferer.

TABLE III
IMPACT OF DIFFERENT NUMBERS OF TRANSCEIVERS

Number of transceivers	*Empty hall*	*Office*	*Lab*
Four	93.85%	93.01%	92.65%
Three	93.56%	92.51%	92.33%
Two	89.56%	85.12%	84.35%
One	86.33%	73.26%	72.14%



(a) One pair of transceiver devices. (b) Two pair of transceiver devices. (c) Four pair of transceiver devices.

Fig. 22. The experimental setup for data collection with various pairs of transceivers.

are 5m from the transceivers, the system achieves an authentication accuracy of 87.32%. Although there is a slight decrease, it still meets application requirements. Therefore, Wi-CGAuth exhibits robustness when there are environment interferers, given the interferers are far from transceivers.

7) Impact of Different Numbers of Transceivers: In this section, the impact of the different numbers of transceivers on user authentication is presented. In the experiment, the number of transceivers varies from four to one in three typical environments. Fig. 22 shows the experimental setup with one pair, two pairs, and four pairs of transceiver devices. As shown in the Table III, the authentication accuracy decreases as the number of transceivers changes. Compared with two pairs of transceivers, three pairs of transceivers receive more Wi-Fi data from more receivers. This will be more beneficial for gesture recognition and user authentication. In addition, as shown in Fig. 22(c), the fourth receiver, Rx4, is placed on the same height as R2 and R3. The angle between Rx2-Tx-Rx4 and Rx3-Tx-Rx4 is forty-five degrees. Rx4 is four meters away from Tx. Collecting data from four pairs of transceivers does not significantly improve the accuracy compared to three pairs of transceivers. Adding more transceivers will introduce more interference. Therefore, three pairs of transceivers are the best choice for Wi-CGAuth.

8) Impact of the Special Scenario: To evaluate the robustness of Wi-CGAuth, the tests are conducted in the library. As shown in Fig. 23, the layout of the library is complicated. There are some occlusions in the sensing area, such as tables, chairs, and bookshelves. All transceivers are deployed in 3D. In the experiments, there are six subjects with similar height and weight. They perform continuous gestures ten times in

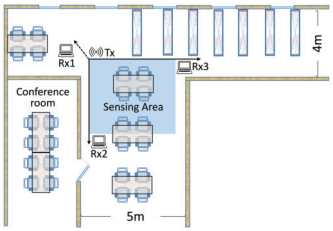


Fig. 23. Library layout.

TABLE IV
THE AUTHENTICATION ACCURACY OF FOUR SCENARIOS

Scenario	Accuracy
Empty hall	91.35%
Office	89.42%
Lab	89.36%
Library	87.25%

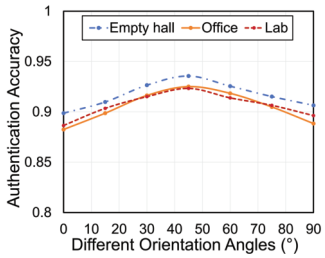


Fig. 24. The impact of user orientations on authentication accuracy.

each of four scenarios: empty hall, office, lab, and library. From Table IV, it is obvious that Wi-CGAuth demonstrates excellent performance across all four scenarios with the aid of its cross-layer framework. In the library, the authentication accuracy is slightly lower than in other scenarios because of its more complicated layout.

9) The Impact of User Orientation: As shown in Fig. 12, in the 3D deployment, user orientation is defined as the direction of the user's face when performing gestures. The user orientation angle is denoted by the angle between the user's face direction and Tx-Rx3. The experiments are done to test the impact of user orientation on authentication accuracy. In 3D deployment, there is one transmitter, Tx, and three receivers, Rx1, Rx2, and Rx3. For 3D deployment, Tx-Rx1, Tx-Rx2, and Tx-Rx3 are mutually perpendicular. When a user faces Tx-Rx2, the user orientation degree is regarded as zero degree. When a user faces Tx-Rx3, the user orientation degree is regarded as ninety degrees. During the experiment, the user varies his orientation from zero to ninety degrees with the step of fifteen degrees. For each user orientation, a user stands at the point with a three-meter distance to Tx. Each time, the user performs a group of gestures continuously ten times.

Fig. 24 shows the impact of different user orientations on user authentication accuracy in three typical environments. When the user orientation angle is 45°, the user authentication accuracy is the highest one. As the orientation changes from 45° to 0° and from 45° to 90°, the user authentication accuracy declines slightly. The different user orientations can significantly influence the Fresnel zone cutting in hand movement. This leads to significant changes in signal dynamics. However, the framework proposed mitigates the environment dependence. Therefore, Wi-CGAuth demonstrates excellent cross-orientation performances.

VI. DISCUSSION AND LIMITATION

A. Impact of Surrounding People's Activities

One subject's gesture-induced CSI dynamics can be distorted by the activities of surrounding people when they are within the sensing area. According to the analysis in [58], the interference from surrounding people's movements is negligible if they are outside the sensing region. However, when multiple individuals are present within the sensing area, separating each individual's gesture-induced CSI dynamics becomes essential to maintain satisfactory user authentication accuracy.

Recent advancements in data-driven techniques provide promising solutions to mitigate the influence led by the surrounding people' activities. The Spectrogram Learning Network (SLNet) has shown effectiveness in enhancing CSI resolution by leveraging deep learning method, thereby minimizing the impact of the interference [59].This kind of data-driven approach could be integrated into the future work of Wi-CGAuth.

B. Sensing Area Limitations

The bandwidth constrain of Wi-Fi signals inherently limits the effective sensing area of Wi-CGAuth. This constraint poses challenges to Wi-CGAuth application, particularly in dynamic environments such as labs or offices. While increasing the number of transceivers can provide partial improvements, it also leads to higher deployment complexity and cost. This makes it an impractical solution for certain scenarios.

To address this issue, the potential of multimodal sensing approaches has been demonstrated in recent research [60]. GaitFi effectively integrates Wi-Fi CSI with video data and makes them complementary of both modalities. Wi-Fi provides robustness in scenarios with weak lighting, while vision offers precise spatial resolution. This fusion significantly enhances the overall performance and broadens the sensing coverage by overcoming individual modality constraints.This work motivates us to integrate other modality to Wi-Fi and trackle the coverage issue of Wi-Fi in the future.

VII. CONCLUSION

In this paper, we propose Wi-CGAuth, a cross-environment continuous gesture-based user authentication system with Wi-Fi. A novel and cross-layer optimization strategy is implemented from the bottom layer signal's time, space, and frequency diversity extension up to the middle layer TCA-based multi-view signal fusion and classification-based continuous gesture segmentation to the upper layer dual-task accurate gesture recognition and user authentication. Through cross-layer collaboration, cross-environment sensing generalization capability is extended to the maximum extent. Extensive experiments in three typical indoor scenarios demonstrate Wi-CGAuth's feasibility and effectiveness. Wi-CGAuth represents a promising stride toward developing a practical user authentication prototype, laying the groundwork for novel insights in the realm of future wireless sensing applications.

APPENDIX A
CSI VALUES

The CSI values at time t are represented as:

$$H(f,t) = A_{\text{noise}}\ (f,t)e^{-j\theta_{\text{offset}}\ (f,t)}(H_s(f,t) + a(f,t)e^{-j2\pi\frac{d(t)}{\lambda_f}}) + \epsilon(f,t) \quad (8)$$

where $H_s(f,t)$, $a(f,t)e^{-j2\pi\frac{d(t)}{\lambda_f}}$ represent the static and dynamic component in CSI, separately, $\epsilon(f,t)$ represents measurement noise. $a(f,t)$ represents amplitude, while $e^{-j2\pi\frac{d(t)}{\lambda_f}}$ represents phase variation induced by dynamic objects. $A_{\text{noise}}\ (f,t)$ and $e^{-j\theta_{\text{offset}}\ (f,t)}$ denote amplitude impulse noise and random phase offsets, separately.

APPENDIX B
CONJUGATE MULTIPLICATION

$$\begin{aligned}
&H_{cm}(f,t) = H_1(f,t) * \overline{H_2}(f,t)\\
&= A^2_{\text{noise}}\ (f,t)\left(H_{s,1}(f) + a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda_f}} + \varepsilon_1(f,t)\right)\\
&\left(\overline{H_{s,2}}(f) + a_2(f,t)e^{j2\pi\frac{d_2(t)}{\lambda_f}} + \overline{\varepsilon}_2(f,t)\right)\\
&= A^2_{\text{noise}}\ (f,t)(H_{s,1}(f)\overline{H_{s,2}}(f) + a_1(f,t)a_2(f,t)\\
&e^{-j2\pi\frac{d_1(t)-d_2(t)}{\lambda_f}} + (H_{s,1}(f) - \alpha) * a_2(f,t)e^{j2\pi\frac{d_2(t)}{\lambda_f}}\\
&+ (\overline{H_{s,2}}(f) + \beta) * a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda_f}})\\
&+ \varepsilon_1(f,t)\left(\overline{H_{s,2}}(f) + a_2(f,t)e^{j2\pi\frac{d_2(t)}{\lambda_f}} + \overline{\varepsilon_2}(f,t)\right)\\
&+ \overline{\varepsilon_2}(f,t)\left(H_{s,1}(f) + a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda_f}} + \varepsilon_1(f,t)\right) \quad (9)
\end{aligned}$$

where $H_{cm}(f,t)$ represents the result of conjugate multiplication. $H_1(f,t)$ denotes the CSI value of one antenna, while $\overline{H_2}(f,t)$ denotes the conjugate of the CSI value of the other antenna. $H_{s,1}(f) * \overline{H_{s,2}}(f)$ represents the product of the static path Channel Frequency Response (CFR) from two receiving antennas. It is regarded as a constant value during a short period. The dynamic path Channel Frequency Response (CFR) $a_1(f,t)a_2(f,t)e^{-j2\pi\frac{d_1(t)-d_2(t)}{\lambda_f}}$ can be ignored since it is very small compared with the value of $H_{s,1}(f) * \overline{H_{s,2}}(f)$. $H_{s,1}(f) * a_2(f,t)e^{j2\pi\frac{d_2(t)}{\lambda_f}}$ and $\overline{H_{s,2}}(f) * a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda_f}}$ are the combination of one antenna's static path CFR and the other's dynamic path CFR. Both of them include CSI dynamics induced by movements. Due to the similar multipath effects of the two nearby antennas, the Doppler velocity information reflected in the dynamic path CFRs has similar values but opposite directions. The term $\overline{H_{s,2}}(f) * a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda_f}}$ can be amplified by increasing the weight α on one antenna and decreasing the weight β on the other antenna, reducing the static path attenuation on the first antenna while increasing it on the second one.[1] Similarly, the noise terms $\varepsilon_1(f,t)\left(a_2(f,t)e^{j2\pi\frac{d_2(t)}{\lambda_f}} + \overline{\varepsilon_2}(f,t)\right)$ and $\overline{\varepsilon_2}(f,t)\left(a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda_f}} + \varepsilon_1(f,t)\right)$ are so small that they can be ignored. Consequently, conjugate multiplication can eliminate the random phase offsets in CSI readings from different antennas caused by hardware imperfections. The equation is simplied as:

$$\begin{aligned}
H_{cm}(f,t) = A^2_{\text{noise}}\ (f,t)(&\overbrace{H_{s,1}(f) * \overline{H_{s,2}}(f)}^{①}\\
&\overbrace{+\overline{H_{s,2}}(f) * a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda_f}}}^{②}\\
&\overbrace{+\varepsilon_1(f,t)\overline{H_{s,2}}(f) + \overline{\varepsilon_2}(f,t)H_{s,1}(f)}^{④} \quad (10)
\end{aligned}$$

①, ② and ③ correspond to the static component, dynamic component, and environmental noise of the CSI, respectively. In this case, there is still environmental noise, i.e., ③, even after the conjugate multiplication process.

APPENDIX C
DERIVATION OF PHASE DIFFERENCES AMONG SUBCARRIERS AND ANTENNAS

When a subject is in motion at a distant location, phase variations $\Delta\theta$ of various subcarriers resulting from subtle movements $\Delta d(t)$ exhibit similarity. Similarly, when a subject is in motion at a distant location, the phase variations of various antennas resulting from subtle movements will be consistent. The difference in path lengths between the two antennas' reflected paths can be regarded as similar [25]. Therefore, each subcarrier's dynamic reflection path length ($d(t)$ for all antenna pairs in Eq. 10) can be divided into two parts: 1) $d(1)$, denotes the initial dynamic reflection path length, 2) $\Delta d(t)$, is the dynamic reflection path length change over $d(1)$. In this case, the ② of Eq. 10 is formulated as:

$$\begin{aligned}
&\overline{H_{s,2}}(f) * a_1(f,t)e^{-j2\pi\frac{d_1(t)}{\lambda_f}}\\
&= \overline{H_{s,2}}(f) * a_1(f,t)e^{-j2\pi\frac{d(1)+\Delta d(t)}{\lambda_f}}\\
&= \overline{H_{s,2}}(f) * a_1(f,t)e^{-j2\pi\frac{\Delta d(t)}{\lambda_f}}e^{-j2\pi\frac{d(1)}{\lambda_f}}\\
&= \overline{H_{s,2}}(f) * A_1(f,t)e^{-j2\pi\frac{\Delta d(t)}{\lambda_f}}e^{-j\theta_{Ini,f}} \quad (11)
\end{aligned}$$

where $\theta_{Ini,f}$ comprises $2\pi\frac{d(1)}{\lambda_f}$ and phase of $a_1(f,t)$, $A_1(f,t) = |a_1(f,t)|$. The initial phases of distinct subcarriers may vary, and the phase differences of identical subcarriers across various antennas depend on the initial phases as well.

APPENDIX D
ELIMINATION

Given a sufficient number of samples T, the mean of a CSI signal after conjugate multiplication is represented as followings:

$$\begin{aligned}
E(f,t) &= \frac{1}{T}\sum_{t=1}^{T} H_1(f,t) * \overline{H_2}(f,t)\\
&= \frac{1}{T}\sum_{t=1}^{T}\left(\delta^2(t)\overline{H_{s,2}}(f) * A(f,t)e^{-j2\pi\frac{\Delta d(t)}{\lambda_f}}e^{-j\theta_{\text{Ini,f}}}\right)
\end{aligned}$$

[1] In the implementation, in each estimation window, we choose α so the minimum amplitude of CSI across all the samples within the window at the first antenna is reduced to zero, and we set β as 1000α.

$$
\begin{aligned}
&+\frac{1}{T}\sum_{t=1}^{T}\delta^2(t)H_{s,1}(f)*\overline{H_{s,2}}(f)+\frac{1}{T}\sum_{t=1}^{T}\varepsilon(f,t)H_s(f)\\
&=\frac{1}{T}\sum_{t=1}^{T}\left(\delta^2(t)\overline{H_{s,2}}(f)*A(f,t)e^{-j2\pi\frac{\Delta d(t)}{\lambda_f}}e^{-j\theta_{\text{Ini, f}}}\right)\\
&+\delta^2(t)H_{s,1}(f)*\overline{H_{s,2}}(f)\\
&=\delta^2(t)\overline{H_{s,2}}(f)A(f,t)e^{-j\theta_{\text{Ini,f}}}\cdot K+\delta^2(t)H_{s,1}(f)*\overline{H_{s,2}}(f)
\end{aligned}
\tag{12}
$$

where $K=\frac{1}{T}\sum_{t=1}^{T}e^{-j2\pi\frac{\Delta d(t)}{\lambda_f}}$. It is demonstrated that ③ in Eq. 10 follows a Gaussian distribution with zero mean. Here, $\varepsilon(f,t)H_s(f)$ is substituted for the ③. By Eq. 10, the noise is eliminated because its mean is 0. K is consistent across various subcarriers and antennas. Subsequently, $E(f,t)$ is subtracted from $H_{cm}(f,t)$, which is expressed as followings:

$$
\begin{aligned}
&H_{cm}(f,t)-E(f,t)=\delta^2(t)\overline{H_{s,2}}(f)A(f,t)*\\
&e^{-j\theta_{\textbf{Ini, f}}}\left(e^{-j2\pi\frac{\Delta d(t)}{\lambda_f}}-K\right)+\delta^2(t)\varepsilon(f,t)H_s(f)
\end{aligned}
\tag{13}
$$

APPENDIX E
NORMALIZATION

$$
\begin{aligned}
S(f,t)&=|\frac{1}{T'}\sum_{t}^{t+T'}(H_{cm}(f,t)-E(f,t))|\\
&=|\frac{1}{T'}\sum_{t}^{t+T'}\delta^2(t)\overline{H_{s,2}}(f)A(f,t)e^{-j\theta_{\text{Ini,f}}}\left(e^{-j2\pi\frac{\Delta d(t)}{\lambda_f}}-K\right)|\\
&+\frac{1}{T'}\sum_{t}^{t+T'}\delta^2(t)\epsilon(f,t)H_s(f)\\
&=\left|\delta^2(t)\overline{H_{s,2}}(f)A(f,t)e^{-j\theta_{\text{Ini,f}}}\left(K'(f,t)-K\right)\right|
\end{aligned}
\tag{14}
$$

where $K'=\frac{1}{T'}\sum_{t}^{t+T'}e^{-j2\pi\frac{\Delta d(t)}{\lambda_f}}$. As $K'(f,t)$ is consistent among various subcarriers/antennas. The various CSI signals after conjugate multiplication reaches max $K'(f,t)$ at the same t_0. The maximum value of $K'(f,t_0)-K$ is marked as S, and $H_{cm}(f,t)-E(f,t)$ is normalized using $S(f,t_0)=\frac{A(f,t)S}{\overline{H_{s,2}}(f)}$ to obtain $R(f,t)$.

APPENDIX F
ALIGNMENT

$$
\begin{aligned}
\arg\min_{\theta_j}\text{Dis(i,j)}&=\sum_{t=1}^{T}\left|R(i,t)-R(j,t)e^{j\theta_j}\right|^2\\
&=\sum_{t=1}^{T}|\frac{(e^{-j\theta_{\text{Ini, }i}}-e^{j\theta_j}e^{-j\theta_{\text{Ini,j}}})\left(e^{-j2\pi\frac{\Delta d(t)}{\lambda_j}}-K\right)}{S}\\
&+\frac{\varepsilon(i,t)H_s(f)}{\overline{H_{s,2}}(f,t)A(f,t)S}+\frac{\varepsilon(j,t)H_s(f)e^{j\theta_j}}{\overline{H_{s,2}}(f,t)A(f,t)S}|^2\\
&=\sum_{t=1}^{T}|\frac{(e^{-j\theta_{\text{Ini,i}}}-e^{j\theta_j}e^{-j\theta_{\text{Ini,j}}})\left(e^{-j2\pi\frac{\Delta d(t)}{\lambda_j}}-K\right)}{S}\\
&+\frac{\varepsilon'(i,j,t,\theta_j)}{S}|^2
\end{aligned}
\tag{15}
$$

where $Dis(i,j)$ denotes the norm sums of the subcarrier i and j. $\frac{\varepsilon'(i,j,t,\theta_j)}{S}$ represents the differences in environmental noises across various subcarriers and fits into a normal distribution. The reference subcarrier is determined by the first one in an antenna pair, and the remaining are adjusted by it to achieve phase alignment. Once θ_j has been obtained for all subcarriers j ($j>1$), the rotated subcarriers are combined to compute the mean.

$$
\begin{aligned}
C(t)&=\frac{1}{N}\sum_{i=1}^{N}R(i,t)e^{-j\theta_i}\\
&=\frac{1}{N}\sum_{i=1}^{N}\frac{\left(e^{-j2\pi\frac{\Delta d(t)}{\lambda_i}}-K\right)e^{-j\theta_{\text{Ini},i}}}{S}+\frac{1}{N}\sum_{i=1}^{N}\frac{\epsilon''(i,t)}{S}
\end{aligned}
\tag{16}
$$

where N denotes the total numbers of the subcarriers, $\frac{\epsilon''(i,t)}{S}$ represents final outcome of elimination, normalization, aligning $\delta^2(t)\varepsilon(f,t)H_s(f)$ by θ_i, which also follows a normal distribution with a zero mean. $C(t)$ represents the CSI signal after enhancement, consisting of two components: the dynamic components and noise.

REFERENCES

[1] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva, "Local contrast phase descriptor for fingerprint liveness detection," *Pattern Recognit.*, vol. 48, no. 4, pp. 1050–1058, Apr. 2015.

[2] L. Zhang, S. Tan, J. Yang, and Y. Chen, "VoiceLive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 1080–1091.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[4] J. Li et al., "Rhythmic RFID authentication," *IEEE/ACM Trans. Netw.*, vol. 31, no. 2, pp. 877–890, Apr. 2023.

[5] J. Liu, K. Cui, X. Zou, J. Han, F. Lin, and K. Ren, "Reliable multi-factor user authentication with one single finger swipe," *IEEE/ACM Trans. Netw.*, vol. 31, no. 3, pp. 1117–1131, Jun. 2023.

[6] Y. Li and M. Xie, "Understanding secure and usable gestures for realtime motion based authentication," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 13–20.

[7] J. Ranjan and K. Whitehouse, "Object hallmarks: Identifying object users using wearable wrist sensors," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2015, pp. 51–61.

[8] J. Wu, J. Konrad, and P. Ishwar, "Dynamic time warping for gesture-based user identification and authentication with Kinect," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 2371–2375.

[9] C. Shi, J. Liu, N. Borodinov, B. Leao, and Y. Chen, "Towards environment-independent behavior-based user authentication using WiFi," in *Proc. IEEE 17th Int. Conf. Mobile Ad Hoc Sensor Syst. (MASS)*, Dec. 2020, pp. 666–674.

[10] C. L. Li, M. Liu, and Z. Cao, "WiHF: Gesture and user recognition with WiFi," *IEEE Trans. Mobile Comput.*, vol. 21, no. 2, pp. 757–768, Feb. 2022.

[11] H. Kong, L. Lu, J. Yu, Y. Chen, L. Kong, and M. Li, "FingerPass: Finger gesture-based continuous user authentication for smart homes using commodity WiFi," in *Proc. 20th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jul. 2019, pp. 201–210.

[12] M. Shahzad and S. Zhang, "Augmenting user identification with WiFi based gesture recognition," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–27, Sep. 2018.

[13] H. Kong et al., "Push the limit of WiFi-based user authentication towards undefined gestures," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2022, pp. 410–419.

[14] C. Bo, L. Zhang, X.-Y. Li, Q. Huang, and Y. Wang, "SilentSense: Silent user identification via touch and movement behavioral biometrics," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, 2013, pp. 187–190.

[15] M. Abuhamad, A. Abusnaina, D. Nyang, and D. Mohaisen, "Sensor-based continuous authentication of smartphones' users using behavioral biometrics: A contemporary survey," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 65–84, Jan. 2021.

[16] S. Keykhaie and S. Pierre, "Lightweight and secure face-based active authentication for mobile users," *IEEE Trans. Mobile Comput.*, vol. 22, no. 3, pp. 1551–1565, Mar. 2023.

[17] Y. Zheng et al., "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proc. 17th Annu. Int. Conf. Mobile Syst. Appl. Services*, 2019, pp. 313–325.

[18] R. Gao et al., "Towards position-independent sensing for gesture recognition with Wi-Fi," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–28, Jun. 2021.

[19] S. J. Pan, V. W. Zheng, Q. Yang, and D. H. Hu, "Transfer learning for WiFi-based indoor localization," in *Proc. Assoc. Advancement Artif. Intell. (AAAI) Workshop*, 2008, pp. 1–6.

[20] S. Arshad, C. Feng, R. Yu, and Y. Liu, "Leveraging transfer learning in multiple human activity recognition using WiFi signal," in *Proc. IEEE 20th Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Jun. 2019, pp. 1–10.

[21] H. Liu et al., "MTransSee: Enabling environment-independent mmWave sensing based gesture recognition via transfer learning," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 1, pp. 1–28, Mar. 2022.

[22] Y. Zeng, P. H. Pathak, and P. Mohapatra, "WiWho: WiFi-based person identification in smart spaces," in *Proc. 15th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2016, pp. 1–12.

[23] J. Chauhan, H. Jameel Asghar, M. Ali Kaafar, and A. Mahanti, "Gesture-based continuous authentication for wearable devices: The Google glass case," 2014, *arXiv:1412.2855*.

[24] I. Ahmed et al., "Checksum gestures: Continuous gestures as an out-of-band channel for secure pairing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2015, pp. 391–401.

[25] Y. Zeng, J. Liu, J. Xiong, Z. Liu, D. Wu, and D. Zhang, "Exploring multiple antennas for long-range WiFi sensing," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 4, pp. 1–30, Dec. 2021.

[26] A. Bandini and J. Zariffa, "Analysis of the hands in egocentric vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6846–6866, Jun. 2023.

[27] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2019.

[28] C. Wang, Z. Liu, and S.-C. Chan, "Superpixel-based hand gesture recognition with Kinect depth camera," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 29–39, Jan. 2015.

[29] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, and L. Yuan, "Hand gesture recognition using deep feature fusion network based on wearable sensors," *IEEE Sensors J.*, vol. 21, no. 1, pp. 539–547, Jan. 2021.

[30] J. Wu and R. Jafari, "Orientation independent activity/gesture recognition using wearable motion sensors," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1427–1437, Apr. 2019.

[31] L. Zhang et al., "Montage: Combine frames with movement continuity for realtime multi-user tracking," *IEEE Trans. Mobile Comput.*, vol. 16, no. 4, pp. 1019–1031, Apr. 2017.

[32] H. Ma and W.-H. Liao, "Human gait modeling and analysis using a semi-Markov process with ground reaction forces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 597–607, Jun. 2017.

[33] S. Pan, N. Wang, Y. Qian, I. Velibeyoglu, H. Y. Noh, and P. Zhang, "Indoor person identification through footstep induced structural vibration," in *Proc. 16th Int. Workshop Mobile Comput. Syst. Appl.*, Feb. 2015, pp. 81–86.

[34] F. Zhang et al., "From Fresnel diffraction model to fine-grained human respiration sensing with commodity Wi-Fi devices," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–23, Mar. 2018.

[35] 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," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–26, Sep. 2019.

[36] Y. Yang, J. Cao, X. Liu, and K. Xing, "Multi-person sleeping respiration monitoring with COTS WiFi devices," in *Proc. IEEE 15th Int. Conf. Mobile Ad Hoc Sens. Syst. (MASS)*, Oct. 2018, pp. 37–45.

[37] L. Zhang et al., "Wi-diag: Robust multisubject abnormal gait diagnosis with commodity Wi-Fi," *IEEE Internet Things J.*, vol. 11, no. 3, pp. 4362–4376, Feb. 2024.

[38] L. Zhang, C. Wang, and D. Zhang, "Wi-PIGR: Path independent gait recognition with commodity Wi-Fi," *IEEE Trans. Mobile Comput.*, vol. 21, no. 9, pp. 3414–3427, Sep. 2022.

[39] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2015, pp. 65–76.

[40] X. Li et al., "IndoTrack: Device-free indoor human tracking with commodity Wi-Fi," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–22, 2017.

[41] J. Zhang, Z. Chen, C. Luo, B. Wei, S. S. Kanhere, and J. Li, "MetaGanFi: Cross-domain unseen individual identification using WiFi signals," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 3, pp. 1–21, Sep. 2022.

[42] C. Shi, J. Liu, H. Liu, and Y. Chen, "Smart user authentication through actuation of daily activities leveraging WiFi-enabled IoT," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jul. 2017, pp. 1–10.

[43] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign language recognition using WiFi," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–21, Mar. 2018.

[44] F. Zhang et al., "From Fresnel diffraction model to fine-grained human respiration sensing with commodity Wi-Fi devices," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–23, Mar. 2018.

[45] Z. Wang, S. Chen, W. Yang, and Y. Xu, "Environment-independent Wi-Fi human activity recognition with adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Feb. 2021, pp. 3330–3334.

[46] S. Roy, U. Roy, and D. Sinha, "Identifying soft biometric traits through typing pattern on touchscreen phone," *Commun. Comput. Inf. Sci.*, vol. 836, pp. 546–561, Jan. 2018.

[47] S. Yi, Z. Qin, E. Novak, Y. Yin, and Q. Li, "GlassGesture: Exploring head gesture interface of smart glasses," in *Proc. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.

[48] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "SpotFi: Decimeter level localization using WiFi," in *Proc. ACM Conf. Special Interest Group Data Commun.*, Aug. 2015, pp. 269–282.

[49] X. Li, S. Li, D. Zhang, J. Xiong, Y. Wang, and H. Mei, "Dynamic-MUSIC: Accurate device-free indoor localization," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2016, pp. 196–207.

[50] Y. Li et al., "DiverSense: Maximizing Wi-Fi sensing range leveraging signal diversity," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 6, no. 2, pp. 1–28, Jul. 2022.

[51] F. Zhang, C. Chen, B. Wang, and K. J. Ray Liu, "WiSpeed: A statistical electromagnetic approach for device-free indoor speed estimation," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 2163–2177, Jun. 2018.

[52] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2010.

[53] M. H. Kefayati, V. Pourahmadi, and H. Aghaeinia, "Multi-view WiFi imaging," *Signal Process.*, vol. 197, Aug. 2022, Art. no. 108552.

[54] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 1997, pp. 583–588.

[55] C. Xiao, Y. Lei, Y. Ma, F. Zhou, and Z. Qin, "DeepSeg: Deep-learning-based activity segmentation framework for activity recognition using WiFi," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5669–5681, Apr. 2021.

[56] Z. Jiang et al., "Eliminating the barriers: Demystifying Wi-Fi baseband design and introducing the PicoScenes Wi-Fi sensing platform," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4476–4496, Mar. 2022.

[57] X. Wang et al., "Placement matters: Understanding the effects of device placement for WiFi sensing," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 1, pp. 1–25, 2022.

[58] X. Guo, J. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, "Device-free personalized fitness assistant using WiFi," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 4, pp. 1–23, 2018.

[59] Z. Yang, Y. Zhang, K. Qian, and C. Wu, "*SLNet: A spectrogram learning neural network for deep wireless sensing*," pp. 1221–1236, Apr. 2023.

[60] L. Deng, J. Yang, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "GaitFi: Robust device-free human identification via WiFi and vision multimodal learning," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 625–636, Jan. 2023.