WiHF: Enable User Identified Gesture Recognition with WiFi

Chenning Li, Manni Liu, Zhichao Cao
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
{lichenni, liumanni, caozc}@msu.edu

***Abstract*—User identified gesture recognition is a fundamental step towards ubiquitous device-free sensing. We propose WiHF, which first simultaneously enables cross-domain gesture recognition and user identification using WiFi in a real-time manner. The basic idea of WiHF is to derive a cross-domain motion change pattern of arm gestures from WiFi signals, rendering both unique gesture characteristics and the personalized user performing styles. To extract the motion change pattern in realtime, we develop an efficient method based on the seam carving algorithm. Moreover, taking as input the motion change pattern, a Deep Neural Network (DNN) is adopted for both gesture recognition and user identification tasks. In DNN, we apply splitting and splicing schemes to optimize collaborative learning for dual tasks. We implement WiHF and extensively evaluate its performance on a public dataset including 6 users and 6 gestures performed across 5 locations and 5 orientations in 3 environments. Experimental results show that WiHF achieves 97.65% and 96.74% for in-domain gesture recognition and user identification accuracy, respectively. The cross-domain gesture recognition accuracy is comparable with the state-of-the-art methods, but the processing time is reduced by 30×.**

***Index Terms*—Gesture Recognition; User Identification; WiFi Channel State Information, Cross-domain**

I. Introduction

Every time I lift my arm, it distorts a small electromagnetic field that is maintained continuously across the room. Slightly different positions of my hand and fingers produce different distortions and my robots can interpret these distortions as orders. I only use it for simple orders: Come here! Bring tea! and so on. Such an amazing scenario was described in the science fiction ”The Robots of Dawn” [1] by Isaac Asimov in 1983. Nowadays, WiFi based sensing is making it happen and researchers have proposed several WiFi based systems for gesture recognition [2]–[8], which can improve the efficiency and quality of human living in a smart home. The fundamental principle that enables humans to naturally interact with and control smart devices or even robots via WiFi is that WiFi channel metrics get distorted by arm or hand gestures as shown in Figure 1.

Besides the semantic meaning of diverse gestures, many applications usually require user identities for access control and content customization. For example, smart home devices can be controlled by only family members, but not guests or strangers. Moreover, when the identity of children or parents can be known, it can recommend different contents when they are watching TV or listening to music. As shown in Figure 1, the true potential of WiFi based gesture recognition can be unleashed only when it can associate the performed gesture with a specific user simultaneously, just like HuFu used in ancient Chinese military, which can simultaneously provide authentication (e.g., match a pair of HuFu pieces) and semantic meaning (e.g., deploy military force). Hence, user identified gesture recognition becomes an emerging research topic.

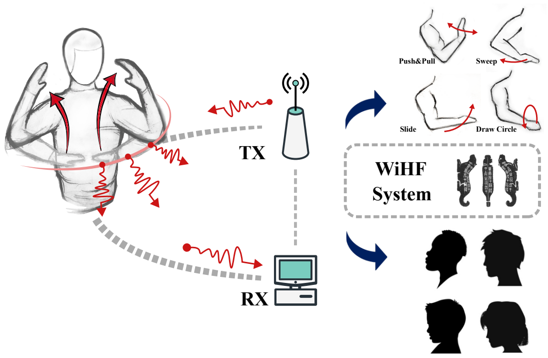

Fig. 1. The illustration of WiHF system inspired by China HuFu, which is designed for both authentication and application purposes.

Nevertheless, user identified gesture recognition encounters three fundamental challenges in practice. First, since the WiFi signals are usually noisy, it is difficult to derive an effective feature, which represents both unique gesture characteristics and personalized user performing styles, to naturally support gesture recognition and user identification simultaneously. Second, the computation complexity of feature extraction should be low enough so that we can apply user identified gesture recognition in a real-time manner. Third, since a user may perform the same gesture in diverse locations, orientations, and environments, the recorded WiFi signals are no longer the same at all. It is challenging to keep user identified gesture recognition accurate across domains (e.g., locations, orientations, environments) [7] with low extra overhead.

State-of-the-art methods fail to resolve all three challenges. For example, Widar3.0 [7] enables a zero-effort cross-domain gesture recognition, which means extra efforts are unnecessary in either data collection or model retraining when gestures are performed in new domains. The feature extracted however cannot support user identification and the computation complexity of feature extraction is too high to achieve real-time. Moreover, WiID [9] achieves user identification for gestures while it must take as input the classification of the gesture performed, which can restrict the user identification accuracy significantly. Consequently, real-time efficiency and cross-domain ability are degraded.

In this paper, we propose WiHF (**Wi**Fi **HuF**u), a pioneering attempt to achieve cross-domain user identified gesture recognition with WiFi in a real-time manner. It aims to capture the personalized motion change pattern caused by arm gestures, which includes rhythmic velocity fluctuation and characteristic pause distribution. Moreover, the pattern keeps consistent across domains. Then an efficient method based on the seam carving algorithm [10] is developed to carve the motion change pattern for computation efficiency. Finally, we design a dual-task DNN model to simultaneously achieve accurate user identified gesture recognition, which further applies the splitting and splicing scheme to bootstrap the cross-domain ability and collaborative learning.

We implement WiHF and evaluate it extensively on a public dataset. Results demonstrate WiHF achieves 97.65% and 96.73% for in-domain gesture recognition and user identification, respectively. Moreover, WiHF demonstrates zero-effort cross-domain characteristics for gesture recognition comparable with the state-of-the-art methods [7], but the processing time is reduced by 30× and thus can be running in real-time. In a nutshell, the contributions of this paper are as follows:

- We design a novel motion change pattern of arm gestures and a dual-task network that can recognize gestures and identify users with WiFi simultaneously.
- We develop several efficient methods to enable real-time processing and ensure the cross-domain recognition accuracy with zero-effort.
- We implement WiHF and conduct extensive experiments to evaluate its performance. The results demonstrate the feasibility and effectiveness of our system.

The rest of this paper is organized as follows. Section II demonstrates the preliminary and observation. The WiHF design is detailed in Section III before the performance evaluation in Section IV. Related works are reviewed in Section V followed by the conclusion Section VI.

II. Preliminary and Observation

In this section, we first discuss the preliminary principles behind the user identification based on arm gestures across domains (§II-A). Then, we demonstrate the feasibility to utilize the motion change pattern of arm gestures for user identified gesture recognition with WiFi (§II-B).

A. Arm Gesture based User Identification across Domains

We survey the existing works to verify two fundamental questions as follows:

a) Are the characteristics of arm gestures representative enough for user identification?: The answer is true due to the following observations. During performing an arm gesture, the potential biometric feature [9], [11]–[14], which associates with the shape of arm and hand, is tightly coupled with the corresponding movements and brings an opportunity for user identification. Moreover, arm gestures, including arm sweep [13] and gesture vocabulary (e.g., drawing the line, circle, rectangle) [11], are diversely performed from user to user due to their different habits, then have been used for user authentication. Several works have shown these unique characteristics of arm gestures can be captured by either WiFi signals or inertial sensors. Specifically, WiID [9] conducts a comprehensive measurement study to validate that the time-series of the frequencies appearing in WiFi Channel State Information (CSI) measurements while performing a given gesture is different from that of the same gesture performed by different users but similar to that performed by the same user in a long period. Moreover, wrist acceleration samples during performing arm gestures are collected with inertial sensors and able to provide personalized characteristics with long term stability over a month [12], [14]. Hence, the characteristics of arm gestures are indeed representative enough for user identification.

b) Can we achieve accurate user identification while avoiding extra re-training efforts across domains?: The key challenge whether a user identification system can adapt the various inputs of the same arm gesture performed by the same user across domains is to extract a domain-independent feature from the induced variances. For example, Widar3.0 [7] derives a domain-independent feature called Body-coordinate Velocity Profile (BVP) from the Doppler Frequency Shift (DFS) spectrogram of raw CSI measurements. It reflects the relative velocity of arm motion in the user's body coordinate system which is irrelevant to the user's orientations, locations, and surrounding environments. With BVP, Widar3.0 achieves cross-domain gesture recognition without the extra cost of data collection and model retraining. BVP shows the possibility of the across-domain feature design, but it is still a challenging problem for user identification because BVP cannot preserve the personalized characteristics while performing arm gestures. Besides, BVP is too computation-intensive to be running in real-time. Hence, to solve the problem, a possible way is to derive a new domain-independent feature for both gesture recognition and user identification.

Overall, to enable user identified gesture recognition in real-time, we need to derive a new feature of arm gestures, which is domain-independent and supports both gesture recognition and user identification with efficient computation complexity.

B. Empirical Study of WiFi Motion Feature

We intend to investigate several feasible personalized features preserved by WiFi CSI measurements while performing arm gestures. Upon receiving the raw CSI measurements, we can derive the DFS spectrogram using Short-Term Fourier Transform (STFT). By observing the DFS spectrogram, we notice that both power and temporal features (called *carving paths*) have the potential to indicate the personalized arm motion from velocity and rhythm aspects. Specifically, DFS spectrograms [5] can separate the movement of different arm

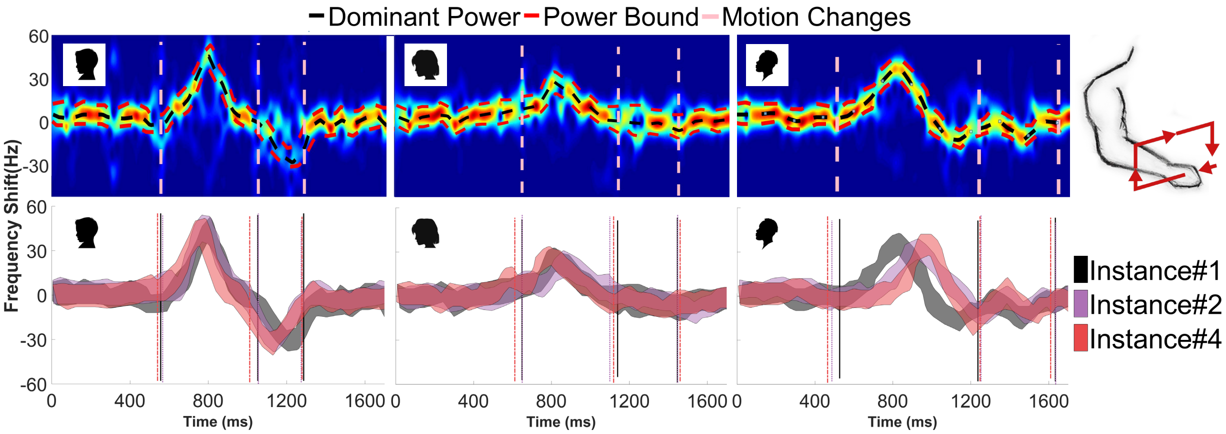

Fig. 2. When three different users perform a gesture (e.g., drawing rectangle), three carving paths (e.g., 3 different color dash lines) of their denoised DFS spectrograms are shown in the top three sub-figures, respectively. For each user, given the same gesture, the corresponding power bound and motion changes of three different instances (e.g., 3 color bold and dash lines) are shown in each bottom sub-figure.

parts when they move at different speeds since spectrogram power varies as the reflection areas change for the certain velocity component over time. Based on the power of DFS spectrograms, two carving paths can be derived. One is *Dominant Power* which reflects the most dominant power in the DFS spectrograms. The other is *Power Bound* which profiles the dominant power area and velocity bounds [15]. Note that both are power based features. Moreover, an arm gesture can be usually divided into some atomic motions (e.g., drawing the line, arc) in the temporal order. For example, drawing a rectangle contains four lines towards four different directions. The switches between two adjacent atomic motions are called *Motion Changes*, which indicate motion pause/restart. We extract a carving path, called *Motion Change Pattern*, to represent the timing of motion changes, namely the temporal rhythmic motion during gesture drawing.

To validate whether the three carving paths are distinguishable among different users and stable for any single user across domains, we further conduct some empirical experiments. With the collected CSI measurements of each gesture instance, we manually extract and label the three carving paths from DFS spectrograms. We have three observations as follows.

a) The motion change pattern of an individual user performing the same gesture stays consistent over time while different users manifest various motion change patterns for performing the same gesture, but it does not hold for power based carving paths all the time: The three top sub-figures of Figure 2 show the three carving paths of three different users while performing the same gesture (e.g., drawing the rectangle). Specifically, the black and red dashed lines denote the carving paths of dominant power and power bound. The pink dashed lines which are parallel with the axis of frequency shift indicate the motion changes. With the visualization of the carving paths, we can observe both power based features and motion change pattern varies among different users. Intuitively, different users perform the same gesture with personalized action understanding and performing style. Among different users, their diverse arm shapes and sizes reinforce the impacts on WiFi signals, leading diverse power based features and the motion change pattern.

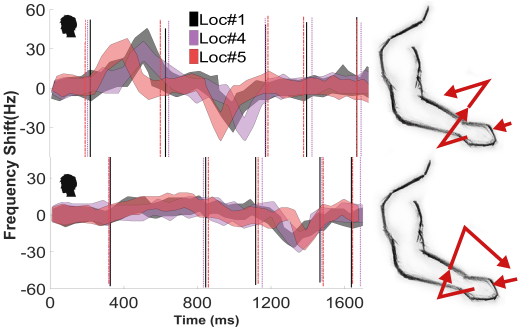

Fig. 3. The distributions of the DFS power bound and motion changes while a user performs two different gestures (e.g., drawing zigzag and triangle) in three different locations (e.g., 3 different color lines).

Further, we collect three instances of the same gesture from all three users and superpose their carving paths for each user in the bottom three sub-figures of Figure 2. We can see the power based carving paths of different instances may shift along the time axis, especially for the third user. In contrast, for all users, their motion changes can be grouped into three clusters which correspond to the three pauses during drawing a rectangle gesture. In a cluster, the largest period between two motion changes across different instances is less than 70ms (the first cluster of the third user appeared at about 500ms), which demonstrates its consistency of motion change pattern for each user across instances. The reason behind this is the inevitable noise (e.g. multi-path, body motion) has a significant influence on the power based carving paths, but motion changes are less affected.

b) An individual user introduces similar motion changes when performing gesture across domains, but the power based carving paths vary significantly: To verify the stability of different carving paths across domains, a user performs two gestures (e.g., drawing zigzag and triangle) at three different locations. For each gesture, the distribution of the derived carving paths is shown in Figure 3. We can see the motion change pattern is consistent across locations for both gestures. The largest period between two motion changes in a cluster is 50ms (the second cluster in the top sub-figure for drawing zigzag). Meanwhile, the power based carving paths exhibit obvious dynamics when a gesture is performed in different locations. The reason is the motion change pattern reflects the rhythm of the arm motion and the temporal characteristic is only related to user drawing habit and gesture composition, but not where the user performs the gesture. In contrast, in the view of WiFi transceiver pairs, the velocity of the detected arm motion varies with the location changes so that incurs DFS power dynamic since the human body is best modeled as a quasi-specular reflector [16].

c) The motion change pattern does not work if we treat gesture recognition and user identification as separate tasks, but works when the two tasks are operated collaboratively.: As Figure 2 shows, we observe among different users, their motion change patterns demonstrate noticeable variances while performing the same gesture. The inconsistency makes it difficult to be utilized for gesture recognition. But if we know the user identity, as shown in Figure 3, we can see the motion change pattern between two different arm gestures performed by the same user are quite different since the timing distribution of noticeable pause/restart is usually different, which inspires us to design both classification tasks in a collaborative manner. The corresponding components contained in the motion change pattern can be extracted for respective task and bootstrap each other, leading to the design of dual-task module for collaborative learning (§III-C).

Overall, in comparison with power based features, the motion change pattern is a better feature to achieve user identified gesture recognition across domains.

III. System Design

Based on the empirical observations, we propose WiHF, which leverages the motion change pattern and designs a collaborative dual-task module to recognize gestures and identify users simultaneously. Figure 4 provides an overview of the architecture of WiHF. It consists of three modules, from bottom to top, including data acquisition, pattern extraction, and collaborative dual-task, respectively.

Data Acquisition Module: Upon receiving raw CSI measurements from WiFi transceiver pairs, WiHF first sanitizes the CSI series using band-pass filters and conjugate multiplication [17], [18]. Then dominant DFS spectrogram components reflected by different body parts (e.g. hand, elbow, arm) are collected using Principal Component Analysis (PCA). Thus we can remove the interference while retaining the unique gesture characteristic and the personalized user performing styles (§III-A).

Pattern Extraction Module: We extract the domain-independent motion change pattern(§III-B). To derive the DFS spectrogram, we first operate the time-frequency analysis by adopting Short-Time Fourier Transform (STFT) on the denoised CSI. Then, WiHF develops an efficient method based on the seam carving algorithm [10] to capture the motion change pattern readily, which is fed into the collaborative dual-task Module.

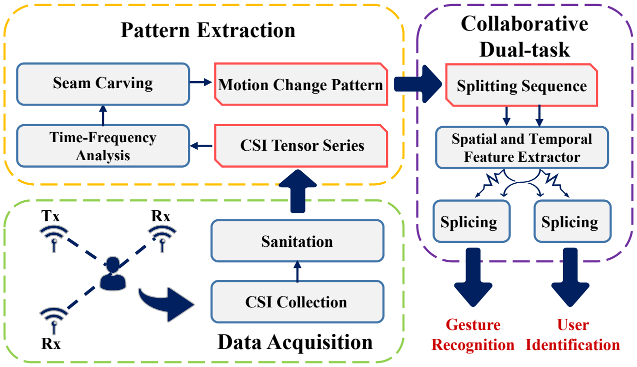

Fig. 4. The system architecture of WiHF.

Collaborative Dual-task Module: This module works for collaborative classification tasks including gesture recognition and user identification at runtime (§III-C). First, the motion change pattern is filtered and split to the corresponding dual inputs for the dual tasks. Then convolution-based Recurrent Neural Network extracts the spatial (e.g., different body parts) and temporal (e.g., motion change pattern) features of gesture motion. Next, the gradient block layer is integrated for splicing respective features while ensuring that they do not affect each other during back-propagation with the loss function. Finally, it outputs the predictions for gesture recognition and user identification simultaneously.

A. CSI Acquisition

A time series of CSI matrices characterizes MIMO channel variations from different dimensions including time (packet), frequency (subcarrier), space (transceivers). For a MIMO-OFDM channel with M transmit antennas, N receiver antennas, K subcarriers, and T packets, the CSI matrix sequence is a 4-D matrix $H \in C^{N\times M\times K\times T}$ representing amplitude attenuation and phase shift of multi-path channels [19]. Thus it has the following measurement at time t, frequency f and antenna a:

$$H(f,t,a) = (H_s(f,t,a) + H_d(f,t,a) + N(f,t,a))e^{j\epsilon(f,t)} \tag{1}$$

where $\epsilon(f,t)$ is the phase offset caused by cyclic shift diversity, sampling time offset, sampling frequency offset [19]. N is the complex white Gaussian noise capturing the background noise [20]. H_s is the static component with zero DFS while the dynamic component H_d is a superposition of vectors with time-varying phases and amplitudes [5].

Generally, we remove the phase offset $\epsilon(f,t,a)$ by calculating conjugate multiplication with raw CSI of two antennas on the same WiFi receiver [7]. Then H_s and N are filtered out using band-pass filter [21]. The left H_d is affected by the motion of multiple body parts. We further obtain the dominant components using PCA. Empirically, the first three components are selected for motion change pattern profiling while balancing the processing time.

B. Motion Change Pattern Extraction

Upon receiving the dominant CSI tensor series, WiHF first applies STFT and obtains the doppler shift frequency f_D. And it's incurred by the movement of the arm and associated with the velocity of different body parts in Equation. (2), where P_d is the set of dynamic paths ($f_D \neq 0$) [5], [7], [21]. Thus we derive the DFS spectrogram with the dimension as $R \times P \times F \times T$, where R and P are the numbers of transceiver links and PCA components. And F and T denote the sampling points in frequency and time domain, respectively.

$$H_d(f,t,a) = \sum_{k \in P_d} \alpha_l(f,t,a) e^{j2\pi \int_{-\infty}^{t} f_{D_k}(u)du} \tag{2}$$

As mentioned in (§II-B), users express unique personalized styles while performing the same gesture, resulting in rhythmic increase, drop or even pause at certain instances. Thus signals reflected by diverse body parts generate consistent motion change patterns and form the corresponding DFS spectrogram sequence. The remaining challenge is how to extract the motion change pattern efficiently. Intuitively, rhythmic increase, drop or even pause usually induce noticeable moving velocity fluctuation detected by the DFS spectrogram. And it occurs at certain moments representing the velocity change peaks in the time domain. But intensive computation for the derivative operation of velocity sacrifices the real-time characteristic. To retain personalized features while balancing the computation cost, the motion change pattern is derived out of carving DFS spectrograms.

The basic idea is to extract the motion change pattern comparable with acceleration as biometrics for dominant body parts, such as wrist, elbow, arm. However, we are facing three challenges. First, DFS only demonstrates the power value for the specific velocity component over time. It cannot provide accurate fine-grained acceleration corresponding body parts due to the superimposition of velocity components at the receiver. Second, the motion change pattern requires the derivative calculation of high dimension data, which is computation-intensive and cannot be running in realtime. Third, the DFS spectrogram contains excessive irrelevant interference, resulting in unnecessary computation and memory.

For the velocity components superimposition challenge, we propose a model to fill the gap between body part acceleration sequence and power distribution of the DFS spectrogram, which outputs the motion change pattern. As CARM [5] elaborates, the power distribution of spectrograms changes as reflection areas S vary for specific doppler shift frequency at instance t. Thus the power P_{ds} for the DFS Spectrogram can be defined with the scaling factor c due to propagation loss as the Equation. (3a). Assuming K body parts are dominant to define the gesture, the relation between P_{ds} and the superimposition of body parts can be modeled as Equation. (3b):

$$P_{ds}(f_D,t) = c \cdot S(f_D,t) \tag{3a}$$

$$= c \cdot \sum_{k=1}^{K} Ref(k,t) \cdot \mathbb{1}(f_{dfs}(k,t) = f_D) \tag{3b}$$

where $Ref(\bullet)$ and $f_{dfs}(\bullet;\bullet)$ denote the individual reflection area and doppler shift frequency at time t for the k_{th} body part.

Nevertheless, accurate P_{ds} is non-reachable due to the resolution of the WiFi signals and the grid estimation for body part K. Thus we can get the experimental approximation $\hat{P}_{ds}$ with acceptable computing error ε and attenuation factor a_{at} as (4a). For accessibility and derivability, Gaussian distributions is applied to model the the superimposition for movements of body parts as (4b):

$$\hat{P}_{ds}(f_D,t) \approx c \sum_{k=1}^{K} Ref(k,t) a_{at} Rect(\frac{f_{dfs}(k,t) - f_D}{2\varepsilon}) \tag{4a}$$

$$\approx c \sum_{k=1}^{K} Ref(k,t) \cdot e^{-\frac{(f_{dfs}(k,t)-f_D)^2}{2(\varepsilon/3)^2}} \tag{4b}$$

Since f_{dfs} demonstrates the moving velocity v with $v = f_{dfs} \times \frac{\lambda}{2}$ [20], [21], corresponding acceleration a_k of each body part can be denoted as $\frac{\partial}{\partial t}|f_{dfs}(k,t) - f_{D_0}|$ for a fixed f_{D_0}. Assuming the $Ref(\bullet)$ variance can be omitted compared with the superimposition effects on $\hat{P}_{ds}$ between consecutive DFS spectrogram, power change rate can de derived as Equation. (5a). With the limitation of rigid body part of human and acceleration continuation, we can properly decimate the derivative of power as Equation. (5b) to simplify the relationship between the DFS power change and acceleration a_k:

$$|\frac{\partial \hat{P}_{ds}(f_{D_0},t)}{\partial t}| \approx |c \sum_{k=1}^{N} Ref(k,t) \cdot -\frac{9\Delta_k}{\varepsilon^2} e^{-\frac{9\Delta_k^2}{2\varepsilon^2}} \cdot a_k| \tag{5a}$$

$$\approx c \sum_{k=1}^{N} Ref(k,t) \cdot \frac{9\Delta_k}{\varepsilon^2} e^{-\frac{9\Delta_k^2}{2\varepsilon^2}} \cdot |a_k| \tag{5b}$$

$$(\Delta_k = |f_{dfs}(k,t) - f_{D_0}|, a_k = \frac{\partial \Delta_k}{\partial t})$$

Thus we find the power change rate increases as a_k rises for all K body parts. The personalized acceleration sequence as biometrics over time [11]–[14] can be detected when users perform gestures by computing the derivative of DFS power spectrograms.

As for the remaining challenges on derivative calculation and redundant data, we are inspired by the seam carving problem in computer graphics for content-aware image resizing [10]. First, we filter the redundant interference with edge detection methods while optimizing derivative calculation using the difference scheme with the convolution operator. Then we develop an efficient method based on the seam carving algorithm to generate multiple dominant carving paths mentioned in (§II-B) as the motion change pattern on each

Algorithm 1 Motion Change Pattern Extraction

```
Input : $[M_{Pow_{ds}}]_{F_D \times T}$, $W_{F_D \times T_s}$, $ker_{sobel}$, $K_{path}$, $T_s$
Output : $MCP_{velocity\_bins \times T_s}$
$[M_{Pow_{ds}}]_{F_D \times T_s} = meanCompressing([M_{Pow_{ds}}]_{F_D \times T}, T_s)$;
for $n = 1$ to $K_{path}$ do
    initialize weight matrix $W_{F_D \times T_s}$;
    $P_{ds} = W \circ M_{Pow_{ds}}$; $M_{gradient} = Conv(P_{ds}, ker_{sobel})$;
    $M_{sum}(1, 1 : T_s) = P_{ds}(1, 1 : T_s)$; $M_{index}(1, 1 : T_s) = 1$;
    for $i = 2$ to $F_D$ do
        for $j = 1$ to $T_s$ do
            $[Val, Index] = max\{M_{sum}(i-1, max(j-1, 1) : min(j+1, T_s))\}$;
            $M_{index}(i, j) = (Index - min(2, j)) + M_{index}(i, j)$;
            $M_{sum}(i, j) = Val + M_{sum}(i, j)$;
        end for
    end for
    $[\sim; Index_{tail}] = max\{M_{sum}(F_D, 1 : T_s)\}$;
    $[M_{path}(n, 1 : F_D), Index] = BackTrack(M_{index}, Index_{tail})$;
    $W = UpdateWeight(W, Index)$;
end for
$MCP_{velocity\_bins \times T_s} = VelocityMapping(M_{path})$;
```

power spectrogram of PCA components. Suppose estimations of K dominant carving paths are considered and each path demonstrates the most significant motion change pattern over time. We use $w_{i,j} \in (0, 1]$ to denote the weight of the $|\frac{\partial}{\partial t}\hat{P}_{ds}(f_D, t)|_{i,j}$ for the i_{th} frequency bin at j_{th} packet. Thus, the optimal Motion Change Pattern (MCP) along the frequency axis, as the function of indices of timestamps, can be defined as:

$$MCP_{opt} = MCP(\underset{t_1,\cdots,t_{F_D}}{\operatorname{argmax}} \sum_{i=1}^{F_D} w_{i,t_i} |\frac{\partial \hat{P}_{ds}(f_D, t)}{\partial t}|_{i,t_i}) \quad (6)$$
$$(s.t.|t_i - t_{i-1}| < 2; i = 1, \cdots, F_D)$$

where F_D denotes the numbers of frequency bins with STFT.

For computation efficiency, the Sobel operator for the time axis [22] is applied on DFS spectrogram $P_{ds}(f_D, t)$ thus we can get the temporal gradient matrix for each power spectrogram.

Algorithm 1 elaborates on the motion change pattern extraction. Note that assigning an appropriate value to the segmentation number T_s is crucial since if T_s is too large, the segmentation may be instantaneous and cannot guarantee the robust feature extracted with a small sliding window. In contrast, with a small T_s, the unique motion change pattern of individual users get too averaged out to be distinguished with a large sliding window. Besides, an adaptive algorithm is too computation-intensive to be real-time. We set T_s with a constant value 60. Therefore, the duration for each segmentation is restricted between $35ms \sim 70ms$ with the total sample length, which is demonstrated as an appropriate segmentation guideline range [9] for DFS spectrograms. Besides, the empirical study of motion change pattern (§II-B) also shows that the most evident difference between adjacent carving paths is less than 70ms. We set the 0.16 m/s resolution within $[-1.6, 1.6]$ velocity range to achieve 20 velocity bins. $[MCP_{VB \times T_s}]_{R \times P}$ is derived out of each sample and fed into the Dual-task Module, where VB denotes the number of Velocity Bins, the segmentation number T_s, the number of receivers R as well as the number PCA components P.

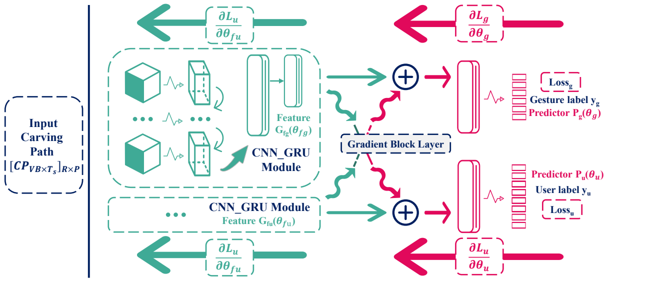
$\frac{\partial L_u}{\partial \theta_{fu}}$
$\frac{\partial L_g}{\partial \theta_g}$
Input Carving Path $[CP_{VB \times T_s}]_{R \times P}$
Feature $G_{fg}(\theta_{fg})$
CNN_GRU Module
Gradient Block Layer
Loss$_g$
Gesture label y_g
Predictor $P_g(\theta_g)$
Predictor $P_u(\theta_u)$
User label y_u
Loss$_u$
CNN_GRU Module
Feature $G_{fu}(\theta_{fu})$
$\frac{\partial L_u}{\partial \theta_{fu}}$
$\frac{\partial L_u}{\partial \theta_u}$

Fig. 5. The structure of the dual-task neural network for gesture recognition and user identification.

C. Dual-task Module

Data Adaptation and Feature Extraction: Upon receiving the motion change pattern, we reshape the motion change pattern as dimensions $[MCP_{(VB \times R) \times P}]_{T_s}$. Thus it is similar to a digital image with the spatial resolution of $VB \times R$ and P as color channels. The underlying principle of the modification is that signals across receivers convey the Angle-of-Arrival (AoA) information [19], [20] while velocity bins contain body part movements [7], [15]. In contrast, PCA components are set as color channels with different scales of signals.

Inspired by Widar3.0 [7], we first extract spatial features from the individual motion change pattern and then profile the temporal dependencies of the whole feature sequence. To do this, the network of Convolutional Neural Network (CNN) based Gated Recurrent Units (GRU) is adopted with the input tensor $[MCP_{(VB \times R) \times P}]_{T_s}$. For each t_s th sampling component, the matrix $[MCP_{(VB \times R) \times P}]$ is fed into a CNN, which contains 16 3×3 filters and two 64-unit dense layers sequentially. ReLU function and the flatten layer are applied for non-linear feature mapping and dimension reshape, resulting in the final output for CNN characterizing the t_s th sampling component. Next, the spatial feature series is fed into the following GRU for temporal profiling. Empirically, we adopt the 128-unit single-layer GRUs to profile the temporal relationships. A dropout layer is further added for avoiding over-fitting followed by the Softmax layer with cross-entropy loss for dual-task prediction. Note that the early stopping scheme is utilized to halt the training at the right time with the patience epochs 30 for value loss [23].

Splitting and Splicing Scheme: As illustrated in Figure 5, the dual-task module requires to splice individual unique features for bootstrapping performance of dual tasks collaboratively.

First, we split $[MCP_{(VB \times R) \times P}]_{T_s}$ evenly along the time axis. Thus we can avoid the issue of vanishing gradients with too long time series ($T_s \sim 60$). On the other hand, splitting the input generates dual inputs for the module and the correlation between them can enhance the performance of the user identification. Then the gradient block layer is tailored for feature splicing inspired by Multi-Task Learning (MTL).

However, different from traditional MTL, the dual tasks here are expected to only slice respective features for collaborative learning while avoiding impacts of loss from each other during the back-propagation process. Take the gesture recognition task as an example, we utilize the output of its own CNN-based GRU module as the superior feature while the feature extracted by the CNN_GRU module for user identification is taken as an inferior one. Further, both features are spliced together and fed into the final layer to predict the gesture. The key point is we do not back-propagate the gesture prediction loss to the CNN-based GRU for user identification. In other words, we keep the CNN-based GRU of user identification from being influenced by gesture predictions. Thus we introduce the Gradient Reversal Layer (GRL) [24] and adapt it into the gradient block layer by setting the splicing factor with zero, which can be used for the generative adversarial network with a positive factor while MTL with a negative one as normal back-propagation.

The underlying rationale comes from both theoretical analysis and experimental validation. First, the dual tasks are defined as a sub-type (user-defined gesture vs gesture-defined user) task instead of a main-type task(gesture vs user), required by the preliminary (§II-B). That means the features for the main-type task can assist the sub-type task as a superior indicator while it should not be influenced by the loss of the inferior feature for the sub-type tasks. On the other hand, it has been validated that the user-specific feature is noise to gesture recognition as domain information [7]. Previous work [24]–[26] applies GRL with a negative factor to eliminate domain noise and extract a cross-domain feature while sacrificing the performance of the predictor. Since the motion change pattern already contains cross-domain knowledge, we no longer need to apply GRL to eliminate noise at the expense of predictor performance.

IV. Implementation and Evaluation

We implement WiHF and evaluate its performance through extensive experiments. The detailed settings are illustrated as follows:

Dataset: We leverage a public dataset from Widar3.0 [7], which contains 9 gestures of 16 users collected across 75 domains (5 positions × 5 orientations × 3 environments). A WiFi receiver consists of 3 antennas that record the raw CSI measurements when the packet generation rate is 1,000Hz at the WiFi transmitter side. And we mainly leverage 6 gestures and 6 users for overall performance distribution due to the non-uniform distribution of other gestures or users across 75 domains. And it's enough for a normal smart home scenario. Besides, we conduct an extensive parameter analysis on the dataset including 9 gestures and 9 users (§IV-D). Considering the sampling length for each gesture, We make three feature datasets using the Pattern-Carving Module. First, HuFuM (**HuFu M**ini) feature dataset is collected using the original dataset for comparison with Widar3.0. Then we find that the original dataset is too short with the average duration of $1.619s$ for each gesture to generate the characteristic carving path representing the personalized motion change pattern. Thus we concatenate raw CSI data across instances and make the dataset HuFuE (**HuFu E**xtend) with the doubled average sample length of $3.238s$. Each extended gesture shows various performance improvement for user identification, which demonstrates that the motion change pattern represented by the carving path is impacted by the duration and complexity of gestures. For example, drawing the rectangle is more sophisticated than sweeping since the former has a much larger influential coverage area and more explicit rhythmic velocity fluctuation. Therefore, the final **HuFu** feature dataset is composed by concatenating raw CSI data across gestures and instances for the demands on gesture duration and complexity. And it contains 10 instances of 6 gestures from 6 users across 75 domains.

Metrics: To characterize the WiHF's performance, ACCuracy (ACC) and latency are the two main metrics for both of gesture recognition and user identification. And the former is the measure of the confidence of prediction for each instance. Besides, False Authorized Rate (FAR) and False Unauthorized Rate (FUR) are adopted for user identification as [27]. And FAR measures the probability that WiHF incorrectly accepts an unauthorized user while FUR evaluates the likelihood that WiHF incorrectly refuses the target user. All the metrics can be calculated with the confusion matrix as:

$$ACC = \frac{TruePositive}{TruePositive + FalsePositive} \tag{7}$$

$$FAR = \frac{FalsePositive}{FalsePositive + TrueNegative} \tag{8}$$

$$FUR = \frac{FalseNegative}{FalseNegative + TruePositive} \tag{9}$$

A. Overall Accuracy

We first evaluate the performance of distinguishing gestures and users on HuFu dataset. Taking all domain factors into consideration, WiHF achieves an overall accuracy of 97.65% and 96.74% with 80 and 20 percentage data used for training and testing. Figure 6 shows the overall metrics including accuracy for the gesture and user as well as FAR and FUR for user identification. WiHF achieves consistently high accuracy of over 95% for all 6 gestures and 6 users in domains. Meanwhile, the FAR rates are even less than 1%, meaning that WiHF can almost correctly identify all unauthorized users 100%. However, we notice that the FURs slightly increase while they are still less than 5%. The reason is that some concatenated gestures are still simple and have limited motion details for users to express their consistent performing styles. This inspires the possibility to further optimize the gesture design.

B. Cross-Domain Evaluation

As for cross-domain characteristics of motion change pattern, we first compare it with Widar3.0 [7] using BVP and HuFuM feature Dataset. Then HuFuE Dataset is adopted for exploring the impact of the gesture duration and complexity on the motion change pattern. Averaged accuracy across domains

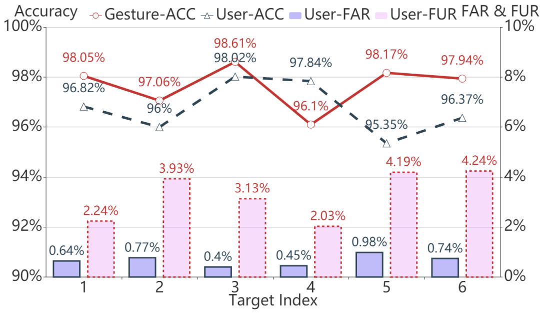

Fig. 6. Overall performance for gesture recognition and user identification on HuFu Dataset. Target Index denotes the specific gesture and user target for dual-task.

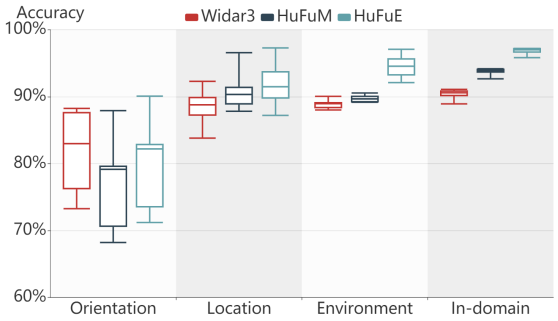

Fig. 7. Performance comparison between WiHF and Widar3.0 in terms of gesture recognition across domains.

is calculated by using one out of all domain instances for testing while others for training. Note that the other domain components keep unchanged for evaluation on the specific domain component. By doing this, we can evaluate its zero-effort cross-domain ability as Widar3.0 does. Figure 7 plots the accuracy distribution for each domain component.

We can find WiHF achieves almost 95% in-domain recognition accuracy using HuFuM while it reaches 92% as reported for Widar3.0 [7]. Compared with Widar3.0 on cross-domain gesture recognition, the average accuracy values of WiHF on HuFuM Dataset slightly increases across locations and environments but drops for orientation. The worst instance performance is only 68.19% with edge orientations (orientation#1 and #5) [7] as the target orientation. Such an accuracy decrease can also be observed for Widar3.0, which is 73.26%. And the reason behind this is that gestures might be shadowed by human body parts in edge orientations, resulting in unrecoverable signal loss. WiHF drops more since such shadow behaviors can cause significant detail missing for the fine-grained motion change pattern. And gestures with appropriate duration and complexity can supply motion details. The performance of HuFuE demonstrates the longer gesture instances can improve the overall cross-domain recognition and compensate for the pattern missing due to body shadow.

Further, we evaluate the performance of cross-domain dual tasks on HuFu Feature Dataset. Table I shows that the cross-domain performance for gesture recognition remains above 85% for orientation 2,3,4 and all the locations while declines by over 10% at edge orientation 1,5, which demonstrates consistent and better accuracy with Widar3.0, HuFuM and HuFuE. It also results from the body part shadow which decreases effective wireless signal sources for the motion change pattern, such as transceivers and PCA components.

For user identification, Table I shows an apparent decrease compared with gesture recognition and even descends to the worst 57.17% when testing at orientation 1. Generally, WiHF achieves 75.31% and 69.52% across locations and orientations, respectively. The reason for the performance decrease is that Dual-task Module identifies users with finer-grained information than gestures. It is weaker due to the cross-domain noise, especially considering the short and simple gestures.

TABLE I
ACCURACY FOR DUAL TASKS ON HUFU FEATURE DATASET

Target Label[a]	1	2	3	4	5
Gesture	In-domain	97.65%			
Location	90.39%	91.33%	92.00%	90.89%	95.72%
Orientation	70.17%	87.11%	86.44%	90.06%	75.67%
User	In-domain	96.74%			
Location	73.83%	66.69%	73.44%	76.22%	86.33%
Orientation	57.17%	77.33%	73.17%	77.39%	59.00%

[a] The target label denotes the data for testing while others for training.

We can alleviate the decrease across domains by extracting more PCA components and carving paths. On the other hand, the performance loss can be minimized by designing more sophisticated gestures, which can minimize the body shadow effect.

To conclude, WiHF achieves averaged performance of 92.07% and 82.38% for gesture across domains. It shows 75.31% for all locations and 75.96% for centering orientations in user identification simultaneously.

C. Latency

In practice, the time consumption of WiHF mainly comes from feature extraction for Pattern-carving Module and recognition as well as identification for Dual-task Module. Table II shows the time consumption distribution. Note that results are all computed assuming it runs in parallel across receivers since the feature $[MCP_{VB\times R\times P}]_{T_s}$ can be derived from various receivers individually and concatenated together. We can find WiHF spends more time on signal processing due to STFT operation for more PCA components. Widar3.0 demands on averaged $70.61s$ for feature extraction while WiHF takes $69.932s$ less than Widar3.0. The total time consumption for HuFu feature extraction is $2.488s$ with the average gesture duration $3.669s$. We believe it is sufficient for most user identified gesture recognition application scenarios. The remarkable improvement on time consumption lies in two folds. First, Widar3.0 derives the BVP in body part coordinate system using the l_0 optimization problem with respect to Earth Mover's Distance metric. And it estimates the high dimensional BVP as the square of velocity bins resolution (20^2 variables) and becomes computation-intensive although

TABLE II
TIME CONSUMPTION COMPARISON IN PARALLEL

	Widar3	HuFuM	HuFuE	HuFu
Signal Processing	0.162s	0.992s	1.312s	1.557s
Feature Extraction	70.29s	0.194s	0.358s	0.379s
Total Time Consumption[a]	70.61s	1.462s	2.162s	2.488s
Gesture Duration	1.619s	1.619s	3.238s	3.669s

[a] It includes procedures for loading data, signal processing, feature extraction and recognition & identification.

Widar3.0 controls the estimated variables using sparsity coefficients. Second, WiHF designs the optimization problems individually for each receiver to estimate the low dimensional carving path as the resolution of velocity bins (20 variables). Moreover, it leverages the efficient method based on the seam carving algorithm instead of the constrained nonlinear multivariable function as Widar3.0.

In a nutshell, WiHF demonstrates comparable cross-domain characteristics using the motion change pattern with Widar3.0 for gesture recognition but the processing time is reduced by 30×.

D. Extensive Study

To study the impact of the number of gestures and user types, various sets of the HuFu feature Dataset are used for different total type numbers (the default type number is 6 for both tasks). Figure III shows that the in-domain accuracy remains above 92% though the number of gestures or users increases to 9. As far as the accuracy of WiHF is concerned, it's not significantly affected by the number of gestures and users. Moreover, we can find that the respective number of subjects for each dual task cannot influence each other. That means performance for gesture recognition stays consistent as the number of users varies even across domains, resulting from the effect of the gradient block layer.

V. RELATED WORK

Passive Human Identification: Most prior work performs with location-oriented activities and leverages intrinsic physiological distinctions or behavioral features. Existing studies capture human gait [28]–[33] or daily activity patterns [34]–[37] in well-defined locations. WiWho [28], WifiU [33], WiFi-ID [30] represent the variations with cascaded statistical signal features to extract user-specific gait features. WFID [29] characterizes the uniqueness of subcarrier-amplitude frequency when users are standing or walking. As for the activity based identification, E-eyes [34] utilizes statistical distribution and time series characteristics for walking and in-place activities recognition. It requires user input to be adaptively updated for domain changes and profile calibration. Duet [35] combines probabilistic inference with the first-order logic to reason about the users' locations and identification but it relies on special devices and assumes the identity of the person is associated with her cell phone. All these passive human identification methods aim to identify users with location-oriented activities and cannot convey extra information. They are vulnerable to domain or activity style changes.

TABLE III
IMPACT OF GESTURE AND USER TYPE NUMBERS

#Gesture		6	7	8	9
Gesture Recognition	In-domain	97.65%	96.14%	95.33%	93.11%
	Location	92.07%	85.81%	84.92%	83.81%
	Orientation	82.38%	74.46%	72.72%	74.55%
User Identification	In-domain	96.74%	97.19%	97.29%	95.33%
	Location	75.31%	68.00%	70.65%	71.36%
	Orientation	69.52%	66.43%	68.34%	70.59%
#User		6	7	8	9
Gesture Recognition	In-domain	97.65%	96.17%	96.99%	97.21%
	Location	92.07%	90.94%	91.62%	91.22%
	Orientation	82.38%	83.81%	79.62%	80.64%
User Identification	In-domain	96.74%	92.56%	93.76%	94.43%
	Location	75.31%	66.98%	64.70%	65.26%
	Orientation	69.52%	63.26%	55.86%	57.26%

[a] The target label denotes the test dataset.

Gesture Recognition across domains: WiFi based gesture recognition systems need to adaptively updated to new data domains. Researchers have proposed to either translate features between domains [3] or generate domain-independent features [7]. Widar3.0 [7] extracts the domain-independent feature BVP from CSI but requires the accurate location of transceivers, otherwise, it may suffer due to the noise and outlier [38]. None of the aforementioned solutions can be real-time or retain the user-specific component. Domain adversarial training approaches [24] are proposed to learn a representation that is predictive about learning tasks on the source domain, but uninformative to the domain of the input. Zhao et al. [25] propose a conditional adversarial architecture, which retains the sleep-specific domain-invariant features. EI [26] incorporates the unlabeled data into conditional adversarial architecture.

VI. CONCLUSIONS

In this paper, we propose WiHF to enable real-time gesture recognition and user identification with WiFi simultaneously. WiHF proposes to derive a cross-domain motion change pattern of arm gestures from WiFi signals, rendering both unique gesture characteristics and the personalized user performing styles. To be real-time, we carve the dominant motion change pattern and develop an efficient method based on the seam carving algorithm. Taking the carving path for motion change pattern as input, a dual-task DNN with splitting and splicing schemes are adopted. We implement WiHF and evaluate its performance on a public dataset. Experimental results show that WiHF achieves 97.65% and 96.74% for in-domain gesture recognition and user identification accuracy, respectively. And the cross-domain gesture recognition performance is comparable with the state-of-the-art methods, but the processing time is reduced by 30×.

ACKNOWLEDGMENT

We thank the authors of Widar3.0 for publicizing the valuable gesture dataset.

References

[1] I. Asimov, *The Robots of Dawn (Robot, 4)*. Spectra, 1983.

[2] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proceedings of MobiCom*, 2013.

[3] A. Virmani and M. Shahzad, "Position and orientation agnostic gesture recognition using WiFi," in *Proceedings of MobiSys*, 2017.

[4] R. H. Venkatnarayan, G. Page, and M. Shahzad, "Multi-user gesture recognition using WiFi," in *Proceedings of MobiSys*, 2018.

[5] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in *Proceedings of MobiCom*, 2015.

[6] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, "Keystroke recognition using WiFi signals," in *Proceedings of MobiCom*, 2015.

[7] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with WiFi," in *Proceedings of MobiSys*, 2019.

[8] Z. Zhou, C. Wu, Z. Yang, and Y. Liu, "Sensorless sensing with WiFi," *Tsinghua Science and Technology*, vol. 20, no. 1, pp. 1–6, Feb 2015.

[9] M. Shahzad and S. Zhang, "Augmenting user identification with WiFi based gesture recognition," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 134, 2018.

[10] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," in *Proceedings of SIGGRAPH*, 2007.

[11] J. Liu, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "uwave: Accelerometer-based personalized gesture recognition and its applications," 2009.

[12] J. Guerra-Casanova, C. Sánchez-Ávila, G. Bailador, and A. de Santos Sierra, "Authentication in mobile devices through hand gesture recognition," *International Journal of Information Security*, vol. 11, no. 2, pp. 65–83, 2012.

[13] F. Okumura, A. Kubota, Y. Hatori, K. Matsuo, M. Hashimoto, and A. Koike, "A study on biometric authentication based on arm sweep action with acceleration sensor," in *Proceedings of ISPACS*, 2006.

[14] K. Matsuo, F. Okumura, M. Hashimoto, S. Sakazawa, and Y. Hatori, "Arm swing identification method with template update for long term stability," in *Proceedings of ICB*, 2007.

[15] P. Van Dorp and F. Groen, "Feature-based human motion parameter estimation with radar," *IET Radar, Sonar & Navigation*, vol. 2, no. 2, pp. 135–145, 2008.

[16] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, "Capturing the human figure through a wall," *ACM Transactions on Graphics*, vol. 34, no. 6, 2015.

[17] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, "Indotrack: Device-free indoor human tracking with commodity WiFi," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 72, 2017.

[18] K. Qian, C. Wu, Z. Zhou, Y. Zheng, Z. Yang, and Y. Liu, "Inferring motion direction using commodity WiFi for interactive exergames," in *Proceedings of CHI*, 2017, pp. 1961–1972.

[19] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," *ACM Computing Surveys*, vol. 52, no. 3, p. 46, 2019.

[20] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu, "Widar2.0: Passive human tracking with a single WiFi link," in *Proceedings of MobiSys*, 2018.

[21] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, "Widar: Decimeter-level passive tracking via velocity monitoring with commodity WiFi," in *Proceedings of MobiHoc*, 2017.

[22] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," *presented at a talk at the Stanford Artificial Project*, 1968.

[23] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[24] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[25] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: A conditional adversarial architecture," in *Proceedings of ICML*, 2017.

[26] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of MobiCom*, 2018.

[27] C. Zhao, Z. Li, T. Liu, H. Ding, J. Han, W. Xi, and R. Gui, "Rf-mehndi: A fingertip profiled rf identifier," in *Proceedings of INFOCOM*, 2019.

[28] Y. Zeng, P. H. Pathak, and P. Mohapatra, "Wiwho: WiFi-based person identification in smart spaces," in *Proceedings of IPSN*, 2016.

[29] F. Hong, X. Wang, Y. Yang, Y. Zong, Y. Zhang, and Z. Guo, "Wfid: Passive device-free human identification using WiFi signal," in *Proceedings of Mobiquitous*, 2016.

[30] J. Zhang, B. Wei, W. Hu, and S. S. Kanhere, "Wifi-id: Human identification using WiFi signal," in *Proceedings of DCOSS*, 2016.

[31] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, and C. J. Spanos, "WiFi-based human identification via convex tensor shapelet learning," in *Proceedings of AAAI*, 2018.

[32] T. Xin, B. Guo, Z. Wang, M. Li, Z. Yu, and X. Zhou, "Freesense: Indoor human identification with WiFi signals," in *Proceedings of GLOBECOM)*, 2016.

[33] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using WiFi signals," in *Proceedings of UbiComp*, 2016.

[34] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: Device-free location-oriented activity identification using fine-grained WiFi signatures," in *Proceedings of MobiCom*, 2014.

[35] D. Vasisht, A. Jain, C.-Y. Hsu, Z. Kabelac, and D. Katabi, "Duet: Estimating user position and identity in smart homes using intermittent and incomplete rf-data," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, p. 84, 2018.

[36] C. Shi, J. Liu, H. Liu, and Y. Chen, "Smart user authentication through actuation of daily activities leveraging WiFi-enabled iot," in *ACM on International Symposium on Mobile Ad Hoc Networking and Computing*, 2017.

[37] B. Wei, W. Hu, M. Yang, and C. T. Chou, "From real to complex: Enhancing radio-based activity recognition using complex-valued csi," *ACM Transactions on Sensor Networks*, vol. 15, no. 3, Aug. 2019.

[38] Z. Yang, L. Jian, C. Wu, and Y. Liu, "Beyond triangle inequality: Sifting noisy and outlier distance measurements for localization," *ACM Transactions on Sensor Networks*, vol. 9, no. 2, 2013.