SHARP: Environment and Person Independent Activity Recognition with Commodity IEEE 802.11 Access Points

Francesca Meneghello⬤, Domenico Garlisi⬤, Nicolò Dal Fabbro⬤, Ilenia Tinnirello⬤, Michele Rossi⬤

Abstract—In this article we present SHARP, an original approach for obtaining human activity recognition (HAR) through the use of commercial IEEE 802.11 (Wi-Fi) devices. SHARP grants the possibility to discern the activities of different persons, across different time-spans and environments. To achieve this, we devise a new technique to clean and process the channel frequency response (CFR) phase of the Wi-Fi channel, obtaining an estimate of the Doppler shift at a radio monitor device. The Doppler shift reveals the presence of moving scatterers in the environment, while not being affected by (environment-specific) static objects. SHARP is trained on data collected as a person performs seven different activities in a single environment. It is then tested on different setups, to assess its performance as the person, the day and/or the environment change with respect to those considered at training time. In the worst-case scenario, it reaches an average accuracy higher than 95%, validating the effectiveness of the extracted Doppler information, used in conjunction with a learning algorithm based on a neural network, in recognizing human activities in a subject and environment independent way. The collected CFR dataset and the code are publicly available for replicability and benchmarking purposes [1].

Index Terms—Wi-Fi sensing, contactless indoor monitoring, human activity recognition, neural networks, CSI, CFR, IEEE 802.11ac.

✦

1 INTRODUCTION

Human activity recognition (HAR) systems are key elements of emerging applications devised for implementation in smart buildings. Among their other functionalities, they can be profitably exploited to optimize the buildings' energy consumption, implement alert systems, provide platforms for smart entertainment, etc. They can also be used to complement IT solutions for elderly care, to promptly detect critical situations or, for assisted-living purposes, to support physically-impaired people in their daily activities [2]. HAR can be performed in indoor environments through ad-hoc devices – such as radars and cameras – or by leveraging already deployed systems without the burden of installing additional instruments. This second possibility is being investigated by several researchers and has recently attracted the interest of the industrial community. In this context, in September 2020, the IEEE established a working group for the definition of a new version of the IEEE 802.11 (Wi-Fi) standard, named IEEE 802.11bf. The new specifications will include modifications at the physical and medium access control layers of the protocol stack to enable the joint provisioning of the communication and the human/environmental sensing services [3]. Devices implementing the new standard are expected to be commercialized by 2024. However, the standard is focused on innovative and cooperative control mechanisms for triggering and exchanging sensing measurements. The definition of algorithms for processing these measurements and detect

features of a target, under unknown environmental conditions, is still an open issue.

In the last decade, since the pioneering work of Adib and Katabi [4], wireless signals of opportunity have been extensively researched as a means to perform device-free localization and activity recognition tasks, including pattern and gesture recognition [5], and even for estimating biological signals at distance, such as the respiration rate or the heartbeat [6]. In this respect, Wi-Fi signals are particularly appealing, due to the high availability of Wi-Fi-enabled devices in most residential and working spaces [7]. While active sensing techniques allow localizing and following users as they carry along a Wi-Fi-connected device, passive sensing approaches obtain such information by monitoring the changes in the channel, where users act as scatterers that modify the channel frequency response (CFR), a frequency estimate of the radio channel describing how the signal changes as it propagates from the transmitter to the receiver [8]. The CFR – also referred to as the channel state information (CSI) – is continuously estimated by Wi-Fi routers for communication purposes. Recently, specific tools have allowed gathering such information from commercial off-the-shelf (COTS) devices, enabling its use for environmental sensing applications.

In this work, we present and validate SHARP (*sensing human activities through Wi-Fi radio propagation*), a device-free HAR system for the automatic detection and classification of human activities in indoor spaces. SHARP is a novel approach to HAR that leverages standard-compliant Wi-Fi devices and achieves unobtrusive human sensing through the analysis of the CFR gathered by COTS IEEE 802.11 routers. We emphasise that, although similar sensing technology has already been proposed as part of prior work, e.g., [4], [5], [6], [7], [8], existing algorithms are unable, unlike SHARP, to generalize across unknown environments and people, and

• F. Meneghello, N. Dal Fabbro and M. Rossi are with the Department of Information Engineering (DEI), University of Padova, via Gradenigo 6/b, 35131, Padova, Italy. M. Rossi is also with the Department of Mathematics "Tullio Levi-Civita", via Trieste 63, 35131, Padova, Italy. F. Meneghello is the corresponding author. E-mail: {name.surname}@dei.unipd.it
• D. Garlisi and I. Tinnirello are with the Department of Engineering, University of Palermo, Viale delle Scienze, edificio 7, 90128, Palermo, Italy and also with CNIT, 43124 Parma. E-mail: {name.surname}@unipa.it

situations for which they were not trained. That is, they *fall short in generalizing across persons and environments*. The proposed system sharply differs from these previous solutions, and improves upon them, in that it reliably works across different indoor environments and people, without having to be re-trained each time the physical wireless propagation setup, or the monitored person, change. To make this possible, SHARP features an integrated procedure combining some preliminary processing steps and a learning-based activity classification algorithm. At first, *an original phase-cleaning method* permits the extraction of *micro-Doppler* traces from the CFR data. Without this procedure, the phase offsets that affect any Wi-Fi transmission would be so strong as to hide finer phase variations and, in turn, would prevent the extraction of the micro-Doppler information. The proposed phase sanitization approach is a major step forward in the HAR via Wi-Fi channels: as we show in Section 6.3, prior sanitization algorithms fail to provide a sufficiently accurate phase signal and are thus ineffective to our purpose. Once available, the Doppler shift reveals the velocities of the scattering points during the transmission events and is not affected by static objects (furniture, walls, etc.), allowing the gauging of dynamic HAR features. As quantified in Section 6, we found it to be a robust representation for environment- and person-independent HAR. To do that, the micro-Doppler traces are used as input for a neural network architecture that is trained to recognize the activities of interest. The effective combination of the preliminary processing steps and the learning algorithm ensures high performance also in changing environmental conditions. This, of course, is a much-desirable feature in off-the-shelf implementation for smart-home applications.

Our work focuses on the recognition of dynamic activities performed by a human in an indoor environment. The identification of static activities – such as relax poses – is out of the scope of the present work. For the recognition of static activities, a paradigm change in both the monitoring technology and the data processing would be needed. Using Wi-Fi devices operating in the sub-6 GHz bands, static users can be hardly distinguished from the empty room situation as the features that can be acquired or computed, e.g., amplitude, phase, Doppler, can not reliably reveal the presence of a static human. In case the amplitude information is considered, the static person only introduces an offset that is constant over time and which would be hardly distinguishable from small changes in the positions of objects inside the room (as discussed at the beginning of Section 3.2). Moreover, the presence of a static person has no impact on the Doppler traces as no micro-Doppler effect is introduced. The recognition of static poses through radio-wave technologies would be possible by leveraging devices operating at higher frequencies, e.g., millimeter-wave radars [9], [10]. However, the detection of static poses is less relevant than dynamic activities in the development of smart-home applications, e.g., involving the contactless monitoring of people. In this context, the main objective is the detection of dynamic movements that may reveal normal or critical situations. SHARP provides proper support for these use cases through standard-compliant IEEE 802.11ac devices that nowadays are almost ubiquitous in residential and working spaces.

The present work sharply departs from the literature on HAR with Wi-Fi devices, as follows.
● We propose an original method for CSI phase correction (also referred to as *sanitization*). The algorithm is independently used at each receiving antenna without the need for a reference stream. Hence, the space diversity provided by the antennas can be fully exploited for sensing purposes.
● We present and validate SHARP, an environment- and person-independent learning-based framework for HAR, which leverages the Doppler effect caused by human motion. The information collected at the monitoring antennas is combined to identify the activity, regardless of the user's position in the monitored area.
● We assess SHARP robustness to changing environmental conditions through CSI data collected with commercial off-the-shelf IEEE 802.11ac devices operating on an 80 MHz frequency band. It reaches an average accuracy of 96% in the identification of four activities – namely walking, running, jumping and sitting – in the most challenging scenario, i.e., when the person, the day and the indoor environment change with respect to those provided in the training set. SHARP is validated against prior work – DeepSense [11], EI [12] and MatNet-eCSI [13] – achieving superior generalization capabilities across subjects and environments. SHARP can be extended to recognize more activities in an environment-independent manner. The numerical results show that it allows reaching an accuracy higher than 95% in the recognition of seven activities, being standing, doing arm gymnastics and sitting down/standing up in addition to the former ones. Note that we validate SHARP on devices implementing the most commonly used Wi-Fi standard at the time of writing, namely the IEEE 802.11ac. However, our approach can also be integrated – without modification – with older and upcoming IEEE standards adopting the orthogonal frequency division multiplexing (OFDM) modulation scheme, such as IEEE 802.11n/ax/be/bf. We released the dataset of the collected CFR traces and the code implementing the devised algorithm to allow replicability and provide a performance benchmark [1].

The rest of the paper is organized as follows. The related work is reviewed in Section 2. In Section 3 and Section 4, we respectively detail the processing on the CSI data and the learning architecture for HAR, which represent the SHARP building blocks. The experimental setup is presented in Section 5, while the performance of the proposed approach is evaluated in Section 6. Conclusions are drawn in Section 7. For completeness, we include Appendix A where an overview of the OFDM Wi-Fi channel model is provided.

2 RELATED WORK

In this section, to put our contribution into context, we review the existing literature on Wi-Fi-based HAR. We specifically focus on works tackling the robustness and generalization of HAR algorithms to different days, environments, and people. Hence, we review the contributions regarding the sanitization of the CFR phase information, the extraction and use of micro-Doppler features.

2.1 CSI based human activity recognition

HAR through CSI data from COTS Wi-Fi devices was first studied in, e.g., E-eyes [14], CARM [15]. More recently,

several articles showed the effectiveness of machine learning techniques in building algorithms that distinguish human activities based on CSI features [11], [16], [17], [18]. However, these works do not focus on the robustness to environmental changes and on the generalization capability to previously unseen environments and subjects, which are key enablers for the successful development of Wi-Fi-based sensing systems [8]. Only a few works try to address these weaknesses. In [12], the authors use a neural network-based approach to extract environment-independent features from the CSI amplitude to recognize human movements. The performance of their algorithm is promising, but remains below 80% in the best scenario. In [19], transfer learning is shown to be effective to adapt the Wi-Fi-based HAR algorithm to different persons and days for the same environment. The algorithm presented in [20] leverages generative adversarial networks to generalize on new persons, while in [13] the matching network one-shot learning approach [21] is proposed to bridge the gap between previously seen environments and new ones. A recent work [22] addresses the problem of location and subject independent HAR through a learning architecture consisting of three deep neural networks. The algorithm is trained on the CSI amplitude collected by monitor routers placed in different positions inside a room, and is tested in the same room by changing the location of a single router. The approach presents a significant performance degradation when evaluated on other datasets (going from 99% to 80% of accuracy).

To the best of our knowledge, no work in the literature proposes a system that generalizes well on unknown environments, days and subjects without any re-training step and using COTS devices. In the present work, we propose an effective solution to this problem, and we consider as benchmarks for comparison, three recent approaches from the literature, i.e., DeepSense [11], EI [12], MatNet-eCSI [13].

2.2 Exploitation of the CFR phase

Some works, e.g, [23], [24], [25], consider a reference antenna and use the phase differences at the remaining ones to perform sensing. By construction, such phase difference is not affected by systematic offsets (same value across the antennas), but it is unable to cope with offsets that are antenna-specific. Exploiting a similar idea, in [26], the authors correct the rotation errors using a reference signal, obtained by connecting with a cable one of the available monitoring antennas with the transmitter. However, the need for a reference reduces the spatial diversity that can be exploited for sensing purposes at the monitor station. The possibility to remove the unwanted phase offsets without exploiting any reference is less addressed in the literature. In [27], the authors propose to mitigate the errors by considering the average signal over a number of time instants, while in [28] a two-step approach is presented to estimate and remove some of the antenna-independent offsets. A different strategy is adopted in [13], where the authors estimate the systematic phase shift at each CFR acquisition. With their approach they remove the static component, that is used as a reference for the algorithm, and retain the dynamic one, using it as the input for the subsequent HAR framework.

In this work, we correct the phase offsets by devising an optimization approach that finds the optimal CFR parameters from the raw channel data. For each CFR sample, we extract the contributions of the different radio paths and exploit the component related to the strongest path as a reference to correct the signal phase. This allows the removal of the non-systematic offsets that affect the signal. As discussed above, this would result infeasible if considering as a reference a signal extracted form a different physical measurement (either in the time or in the spatial dimension).

2.3 Doppler based applications

The micro-Doppler effect extracted from ad-hoc RADAR devices has been extensively used in the literature to perform HAR. Several studies emphasise its effectiveness in highlighting relevant features for the correct identification of the dynamics in the scene [29]. However, only a few works in the literature exploit the Doppler shift computed from commercial Wi-Fi devices, without recurring to ad-hoc infrastructures. In [30], [31], [32], the Doppler information is used to track the movement of a user inside an environment. Two different monitoring devices, placed in strategic positions, are used to obtain effective Doppler shift estimates. In [33] the authors leverage the Doppler effect for breathing detection and people counting. The sensing system consists of one Wi-Fi access point and two universal software radio peripherals (USRPs) monitoring the target channel and a reference one that needs to remain stable. A sensing system consisting of a single monitoring station is presented in [34], to recognize three different bodyweight exercises. In [34], the transmitter and the monitor stations must be placed in specific positions with respect to the user, so that the activity is performed perpendicularly to the radio link. Moreover, the detection algorithm is based on metrics that are manually extracted from the raw CSI data and the Doppler trace. This makes the proposed approach highly activity-dependent, as the features are specifically designed to separate the three considered physical exercises. HAR through Doppler information has been recently addressed in [35]. The latter algorithm uses phase sanitization based on a reference antenna, followed by a support vector machine (SVM) classifier. This strategy allows achieving a good recognition accuracy (96%) on test data collected in the same environment where the algorithm is trained, proving the effectiveness of the Doppler information for HAR purposes. Also, the system performance is improved considering additional data from a wearable inertial measurement unit (IMU), increasing the accuracy to nearly 100%. However, no performance assessment is carried out for scenarios that are not seen at training time. SHARP outperforms [35] by exploiting an advanced phase sanitization approach and a more powerful learning architecture. As a result, it reaches accuracies close to 100% by only exploiting CSI data (no IMU). Moreover, SHARP generalizes to unknown environments, by correctly recognizing the activity in at least 95.99% of the cases.

Our intuition, behind the design of SHARP, is that the Doppler shift naturally lends itself to the separation of the radio reflections coming from static objects or structures (e.g., room walls) from those generated by moving targets (e.g., humans). If available, it can be used as a reference signal domain to perform environment independent recognition tasks, assuming that the environment is mostly static.

Our work descends from this line of reasoning and our novelty resides in the fact that we were, for the first time, able to extract the Doppler signal from 802.11ac COTS devices. After that, such Doppler signal must be used in combination with dedicated processing and learning architectures, which are the second main contribution of our present work. In the remainder of this work, the extraction of the Doppler is presented in detail, along with its use with inception-based neural networks.

3 CSI DATA PROCESSING

Wi-Fi systems adopt orthogonal frequency division multiplexing (OFDM), by transmitting the user information over K partially overlapping and orthogonal sub-channels, with K even. At the OFDM receiver, the channel parameters (amplitudes and phases) are continuously estimated for all the sub-channels. This information is computed for each received packet based on known preamble symbols, and is collected by the CSI, a large (environment-dependent) complex matrix describing the CFR for each sub-channel along every receiving antenna.[1]

For each pair of transmit and receive antennas, the CFR is a set of complex numbers $A_k e^{j\phi_k}$ specifying the attenuation A_k and the phase shift ϕ_k for each sub-channel $k \in \{-K/2, \dots, K/2 - 1\}$. Considering a multi-path propagation channel, and using index $p \in \{0, \dots, P - 1\}$ to indicate the P copies of the transmitted signal that are collected at the receiver, the CFR estimated on packet n is written as

$$
\begin{aligned}
H_k(n) &= A_k(n)e^{j\phi_k(n)} \\
&= \sum_{p=0}^{P-1} A_p(n)e^{-j2\pi(f_c+k/T)\tau_p(n)},
\end{aligned} \quad (1)
$$

where f_c is the main carrier frequency, $T = 1/\Delta f$ is the OFDM symbol time, with Δf being the sub-channel spacing (see Appendix A), while $A_p(n)$ and $\tau_p(n)$ respectively represent the attenuation and the delay associated with path p. The complete OFDM model for the Wi-Fi channel, along with a derivation of Eq. (1) are given in Appendix A.

In this work, the CFR matrix is estimated by a monitor Wi-Fi device running the Nexmon CSI tool [36], as detailed in Section 5. We remark that the collected CFR slightly deviates from the theoretical model in Eq. (1) due to hardware artifacts, which introduce an undesired phase offset $\phi_{\text{offs},k}$, i.e.,

$$
\bar{H}_k(n) = H_k(n)e^{\phi_{\text{offs},k}}. \quad (2)
$$

Next, we present the steps that we implemented to clean the CFR matrix that is extracted by the Nexmon CSI tool. Note that a new CFR matrix is retrieved for each received packet. Thus, the interval between subsequent acquisitions of the CFR is variable but, for the sake of exposition, in the following analysis we assume that a new sample is available every T_c seconds, with T_c fixed. We remark that, in a real Wi-Fi network, the traffic exchanged between already deployed access points and user devices (e.g., streaming services), can be captured and exploited for HAR recognition purposes. When this is not possible, two Wi-Fi access points are required to recreate the setup, using a simple

1. In the rest of the paper, the terms CSI and CFR will be used interchangeably.

network controller to transmit packets at regular intervals and a second monitor device to infer the CFR from these.

3.1 Phase sanitization

The undesired phase offset $\phi_{\text{offs},k}$ in Eq. (2) contains different contributions (see Appendix A for their detailed explanation) [26], [37]. Some of them, namely, the channel frequency offset (CFO), the phase-locked loop (PPO) and the phase ambiguity (PA), although changing in time have the same value across the sub-channels in each receiving antenna. The sampling frequency offset (SFO) and the packet detection delay (PDD) are instead sub-channel dependent. As shown in Appendix A, the offset $\phi_{\text{offs},k}$ experienced at one receiving antenna in sub-channel k can be expressed as

$$
\phi_{\text{offs},k} = -2\pi k(\tau_{\text{SFO}} + \tau_{\text{PDD}})/T + \phi_{\text{CFO}} + \phi_{\text{PPO}} + \phi_{\text{PA}}. \quad (3)
$$

The main idea behind our approach to phase sanitization is that the contribution of each path in Eq. (2) is affected by the same phase shift $\phi_{\text{offs},k}$. Hence, if we were able to separate the different paths, we could use (any) one of them as a reference to remove the phase offset from the CFR. In our method, the strongest path will be used to this end, as this is the path whose parameters are more reliably estimated at the receiver. The details are given shortly below. (Note that in the following analysis the time index n is omitted in the interest of readability.)

Let $\mathbf{h}$ be the K-dimensional vector collecting the CFR information for the K sub-channels,

$$
\mathbf{h} = [\bar{H}_{-K/2}, \dots, \bar{H}_0, \dots \bar{H}_{K/2-1}]^T. \quad (4)
$$

To separate the P multi-path contributions, we define a grid of P' possible paths with $P' > P$ and solve a minimization problem to select the P components out of the P' ones that contribute to the CFR. We consider the following decomposition

$$
\mathbf{h} = \mathbf{Tr}, \quad (5)
$$

where $\mathbf{T} = [\mathbf{T}_{-K/2} \dots \mathbf{T}_{K/2-1}]^T$ is an $(K \times P')$-dimensional matrix collecting the contributions to the CFR that depend on the sub-channel index k, while $\mathbf{r}$ is a P'-dimensional column vector representing the sub-channel independent terms. The P'-dimensional row vectors $\mathbf{T}_k$ are defined through a dictionary of candidate total delays $\tau_{p,\text{tot}} = \tau_p + \tau_{\text{SFO}} + \tau_{\text{PDD}}$, with $p \in \{0, \dots, P' - 1\}$ as

$$
\mathbf{T}_k = \left[e^{-j2\pi k \tau_{0,\text{tot}}/T} \quad \dots \quad e^{-j2\pi k \tau_{P'-1,\text{tot}}/T} \right], \quad (6)
$$

The P'-dimensional column vector $\mathbf{r}$ can be modeled as

$$
\mathbf{r} = e^{j(\phi_{\text{CFO}} + \phi_{\text{PPO}} + \phi_{\text{PA}})} \begin{bmatrix} A_0 e^{-j2\pi f_c \tau_0} \\ \vdots \\ A_{P'-1} e^{-j2\pi f_c \tau_{P'-1}} \end{bmatrix}, \quad (7)
$$

and is obtained by solving the minimization problem

$$
\mathbf{P1}: \quad \mathbf{r} = \underset{\tilde{\mathbf{r}}}{\arg\min} \ ||\mathbf{h} - \mathbf{T}\tilde{\mathbf{r}}||_2^2 + \lambda ||\tilde{\mathbf{r}}||_1, \quad (8)
$$

where $\lambda > 0$ is the weighing parameter for the ℓ_1 regularization term in the Lasso regression. As $\mathbf{r}$ reflects the sparse channel impulse response, $\mathbf{P1}$ is a compressive sensing reconstruction problem and can be solved through quadratic optimization using, e.g., [38]. Note that $\mathbf{h}$, $\mathbf{T}$ and, in turn,

r, are complex-valued vectors/matrices while the majority of quadratic optimization tools, e.g., [38], work with real numbers. In turn, to solve **P1**, we devised a strategy to convert our problem into a real-valued minimization task as follows. The K-dimensional complex-valued **h** vector is first converted into a $2K$-dimensional real-valued vector – $\mathbf{h}_{\text{ext}}$ – by concatenating the real and the additive inverse of the imaginary parts, i.e.,

$$\mathbf{h}_{\text{ext}} = \begin{bmatrix} \text{Re}(\mathbf{h}) & -\text{Im}(\mathbf{h}) \end{bmatrix}. \tag{9}$$

The $K \times P'$-dimensional complex-valued T matrix is converted into a $2K \times 2P'$-dimensional real-valued matrix – $\mathbf{T}_{\text{ext}}$ – by combining the imaginary and real parts as

$$\mathbf{T}_{\text{ext}} = \begin{bmatrix} \text{Re}(\mathbf{T}) & -\text{Im}(\mathbf{T}) \\ \text{Im}(\mathbf{T}) & \text{Re}(\mathbf{T}) \end{bmatrix}. \tag{10}$$

Using these conversions, **P1** in Eq. (8) is solved through [38] obtaining a $2P'$-dimensional real-valued $\mathbf{r}_{\text{ext}}$ vector. To obtain the P'-dimensional complex-valued **r** vector, the real and imaginary parts of $\mathbf{r}_{\text{ext}}$ are recombined by applying the reverse of the processing explained above for **h**, i.e.,

$$\mathbf{r} = \mathbf{r}_{\text{ext}}[0 : P' - 1] - j\mathbf{r}_{\text{ext}}[P' : 2P' - 1], \tag{11}$$

where the indices inside the squared brackets represent the first and the last considered entries of the $\mathbf{r}_{\text{ext}}$ vector, starting from entry zero.

The non-zero entries in the solution **r** reveal the presence of a path p' with corresponding total delay $\tau_{p',\text{tot}}$ with $p' \in \{0, \dots, P' - 1\}$. Note that it would not be possible to separate the contributions in **r** as encoded in Eq. (7). However, from Eq. (7) we know that the phase offset is the same across all the paths and we will leverage this fact in the following to sanitize the CFR phase.

Using the decomposition in Eq. (5), the vector **r** found solving problem **P1**, and Eq. (6), we are able to separate the contributions of the different paths in the CFR for each sub-channel k by applying the following Hadamard product

$$\mathbf{X}_k = \mathbf{T}_k^T \circ \mathbf{r}. \tag{12}$$

$\mathbf{X}_k$ can be rewritten by replacing the terms in Eq. (12), as

$$\mathbf{X}_k = e^{-j2\pi k(\tau_{\text{SFO}} + \tau_{\text{PDD}})/T} e^{j(\phi_{\text{CFO}} + \phi_{\text{PPO}} + \phi_{\text{PA}})} \times$$
$$\times \begin{bmatrix} A_0 e^{-j2\pi(f_c + k/T)\tau_0} \\ \vdots \\ A_{P'-1} e^{-j2\pi(f_c + k/T)\tau_{P'-1}} \end{bmatrix}. \tag{13}$$

At this point, we define $p^* \in \{0, \dots, P' - 1\}$ as the position where **r** has the highest amplitude, which, in turn, is associated with the strongest path. Let X_{k,p^*} be the p^*-th entry of $\mathbf{X}_k$. By multiplying $\mathbf{X}_k$ by the complex conjugate of X_{k,p^*}, we remove the phase components that are constant across all the paths, including the offset, obtaining

$$\bar{\mathbf{X}}_k = A_{p^*} e^{j2\pi(f_c + k/T)\tau_{p^*}} \begin{bmatrix} A_0 e^{-j2\pi(f_c + k/T)\tau_0} \\ \vdots \\ A_{P'-1} e^{-j2\pi(f_c + k/T)\tau_{P'-1}} \end{bmatrix}. \tag{14}$$

Hence, summing up the elements of $\bar{\mathbf{X}}_k$ (associated with the paths from $p' = 0$ to $p' = P' - 1$), we attain a CFR

estimate for each of the sub-channels, where the phase offset is mitigated, as follows,

$$\hat{H}_k = \hat{A}_k(n)e^{j\hat{\phi}_k(n)}$$
$$= A_{p^*} \sum_{p'=0}^{P'-1} A_{p'} e^{-j2\pi(f_c + k/T)\hat{\tau}_{p'}} \tag{15}$$
$$\simeq A_{p^*} e^{j2\pi(f_c + k/T)\tau_{p^*}} H_k,$$

where $\hat{\tau}_{p'} = \tau_{p'} - \tau_{p^*}$. The last relation – that links $\hat{H}_k$ with H_k – requires an approximate equality sign as, in general, $P' \neq P$ and the paths $p' \in \{0, \dots, P' - 1\}$ obtained by solving the compressive sensing reconstruction problem in **P1** may deviate from the actual $p \in \{0, \dots, P - 1\}$ ones. Note that $\hat{H}_k$ of Eq. (15) represents the CFR estimate for sub-channel k, where the phase offset has been removed and the contribution of each path is modulated according to the amplitude and phase of path p^*.

3.2 Doppler trace computation

Fig. 1 shows how the amplitude and the phase change during a time interval of 3 seconds in two different scenarios, i.e., an empty room (on the left) and a room with a moving person (right plots). The presence of a human induces changes in the extracted channel parameters, which can be exploited by HAR algorithms. However, the CFR is affected by the whole indoor multi-path propagation environment and hence also accounts for the reflections from static objects. The reflected signals combine differently at each sub-channel, and the delays induce sub-channel specific phase shifts (see Eq. (1)). Such a behavior is *environment-specific* and is clearly identifiable in the amplitude plots of Fig.1 (see the horizontal patterns in the figure). This fact does not allow developing robust algorithms for HAR that generalize across different environments, as the CFR is strongly affected by the room configuration itself (static objects, including walls, and the room shape). Also, even considering the same indoor space, slight changes in the position of the objects therein have a non-negligible effect on the measured CFR, thus making the HAR task more challenging. As an example, Fig. 2 shows the CFR amplitude collected in two different days within the same empty room. Note that the presence of a person that remains still during the acquisition time modifies the CFR in a way that would be indistinguishable from changing the CFR of the same empty room by moving around static objects. We found that the micro-mobility of the chest and other body parts and, in turn, static poses, are hardly detectable using radio frequencies at sub-6 GHz and off-the-shelf IEEE 802.11ac technology, independently of the type of features extracted. Instead, static poses can be successfully identified and classified by leveraging the propagation behavior of radio waves at higher frequencies, e.g., in the millimeter-wave spectrum.

Due to these facts, in our framework we focus on the detection of dynamic activities and we exploit the Doppler effect to obtain effective features for environment-independent HAR. The Doppler effect corresponds to a shift in the signal phase measured at the receiver as the geometry of the multi-path propagation changes during a transmission event. The movements of the scatterers cause variations in the time needed by the signal to reach the
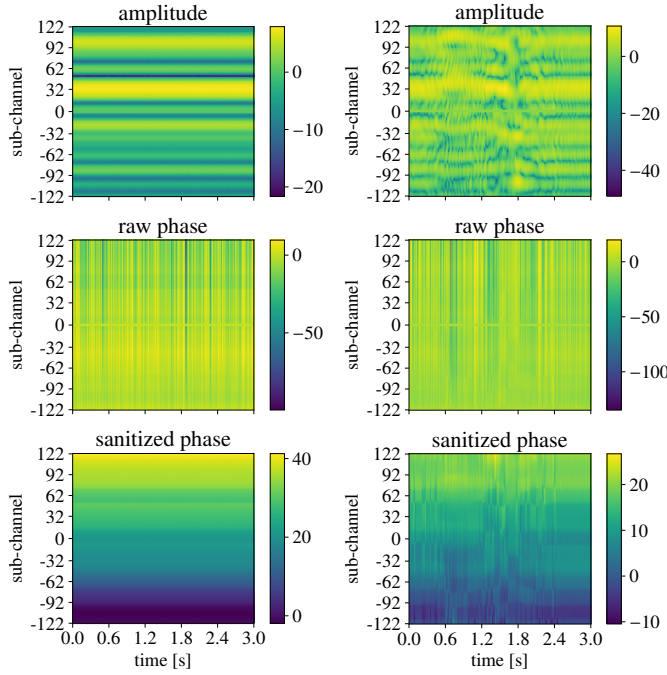
Fig. 1: Amplitude (in dB scale), raw and sanitized phases (unwrapped) of CFR data for an empty room (left plots) and with a person running (right plots). Each trace is three seconds long and shows the behavior on each of the monitored sub-channels (y-axis). Note that CSI is not available on the three central sub-channels, see Section 5 for details.
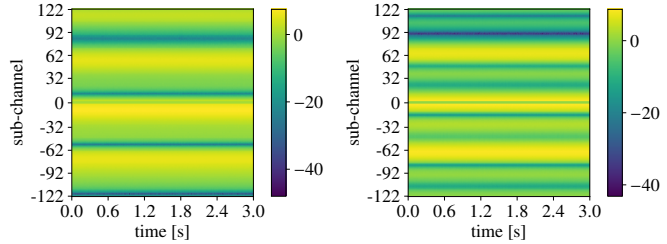

Fig. 2: CFR amplitude (dB scale) collected in the same empty room but in two different days. Environmental variations from one day to another change the multi-path CFR.

receiver through each of the propagation paths. This reflects in a phase shift in the received OFDM signal and, in turn, in the CFR samples defined in Eq. (1) (see Appendix A for a description of the OFDM model). Specifically, considering path p in $H_k(n)$, its associated delay, $\tau_p(n)$, can be expressed as the sum of two contributions. Let ℓ_p be the path length related to the initial position of the scattering point and $\Delta_p(n)$ be the delta caused by the movement of the point during the transmission period nT_c. We have

$$\tau_p(n) = \frac{\ell_p + \Delta_p(n)}{c}, \qquad (16)$$

with

$$\Delta_p(n) = -\int_0^{nT_c} v_p(x) \cos \alpha_p(x) dx, \qquad (17)$$

where c is the speed of light, v_p indicates the speed of scatterer p, and $\cos \alpha_p$ results from the combination of sinusoidal functions related to the angles of motion of the scatterer, and the angles of arrival/departure of the signal.

The activity-related movements of a human cause complex variations in the phase, as each body part acts as a

scatterer moving at a specific velocity $v_p(x)$. This is revealed by the Doppler vector computed from the CFR samples through a short-time Fourier transform over N subsequent channel estimates, i.e., during a *channel observation window*. These N estimates are respectively acquired by extracting the CFR samples at the Wi-Fi monitor for N subsequent packets, collected with sampling period T_c. The value of N is selected so that the attenuation, the velocities, and the angles can be considered constant during the i-th observation window $[iNT_c, (i+1)NT_c]$, with $i \geq 0$. This allows us to rewrite $\Delta_p(n)$ (Eq. (17)) as

$$\Delta_p(n) = -v_p \cos \alpha_p nT_c. \qquad (18)$$

Our aim is to estimate the Doppler shift $v_p \cos \alpha_p$ from Eq. (15) using Eq. (16) and Eq. (18). However, note that, the sanitized phase expressed by Eq. (15) depends on $\hat{\tau}_p = \tau_p - \tau_{p^*}$, where τ_{p^*} is the propagation delay associated with the strongest path. Fortunately, we can reasonably assume that the strongest path, p^*, refers to a static component, thus $\Delta_{p^*}(n) = 0$ and $\tau_{p^*} = \ell_{p^*}/c$, which is constant. It follows that the term depending on p^* in Eq. (15) does not alter the dynamic component of the signal, which is entirely contained in H_k and, in turn, reveals the subject's movements. The reference static component is either a line-of-sight (LOS) or a non-line-of-sight (NLOS) path. Typically, in an indoor environment the signals propagate through several paths where at least walls contribute with static components even if the direct path between the transmitter and the receiver is blocked. It descends that we are entitled to use Eq. (15) to track the dynamic component in the Doppler domain. We do so by defining the $K \times N$ dimensional CFR matrix $\mathbf{H}_i$ associated with the i-th observation window, with entry (k, n) referring to the CFR estimate from Eq. (15) obtained for OFDM sub-channel k and packet n within the current window i, formally,

$$\mathbf{H}_i = \begin{bmatrix} \hat{H}_{-K/2}(iN) & \dots & \hat{H}_{-K/2}((i+1)N-1) \\ & \vdots & \\ \hat{H}_{K/2-1}(iN) & \dots & \hat{H}_{K/2-1}((i+1)N-1) \end{bmatrix}. \qquad (19)$$

A parallel can be built between matrix Eq. (19) and the signal obtained by a frequency-modulated continuous-wave (FMCW) radar, where the *subcarrier* axis (row index) encodes the *fast-time* component, while the *time* axis (column index, i.e., subsequent time estimates) serves as the *slow-time* component [29]. This analogy guides us on the transformation that is to be applied to the matrix to extract the desired Doppler information. Specifically, each element of the N_D-dimensional Doppler vector $\mathbf{D}_i = [d_i(-N_D/2), \dots, d_i(N_D/2-1)]^T$ is obtained as

$$d_i(u) = \sum_{k=-K/2}^{K/2-1} |\mathscr{F}\{\mathbf{H}_i\}(k, u)|^2, \qquad (20)$$

where $u \in \{-N_D/2, \dots, N_D/2 - 1\}$ is the Doppler index and $\mathscr{F}\{\cdot\}$ indicates the Fourier transform.

Proof. The mathematical expression for $\mathscr{F}\{\mathbf{H}_i\}(k, u)$ is derived as follows. According to the formulation in Section 3, element k, n of $\mathbf{H}_i$ is $\hat{H}_k(n)$. The models for the channel frequency response and the Doppler effect used in the computation are given in Eq. (15), Eq. (16) and Eq. (18), while
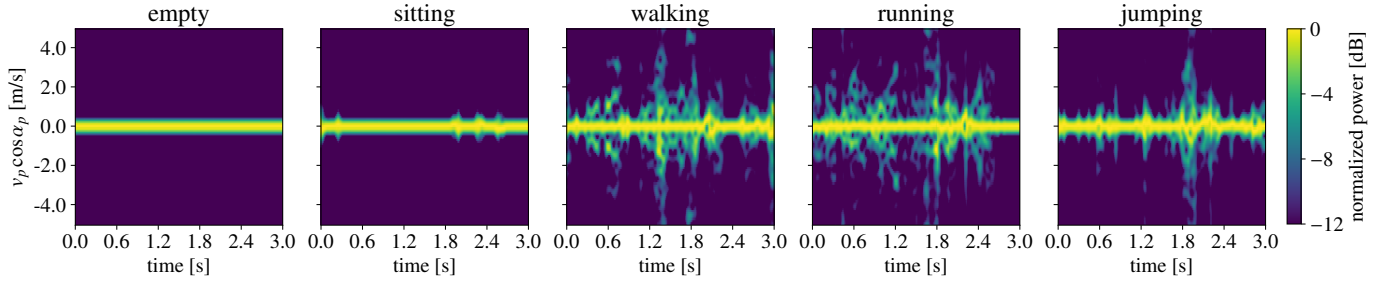
Fig. 3: Doppler spectrogram expressed as $v_p \cos \alpha_p$ for five different setups (from left to right: empty room, sitting, walking, running and jumping). The quantity can be positive or negative depending on the way the scattering points move, changing the geometry of the multi-path propagation channel.

W_k is the Hanning function, selected for the windowing operation. The term $e^{j2\pi(k\hat{v}_p \cos \hat{\alpha}_p nT_c/T)/c}$ is negligible and omitted in the final expression. Note also that, to increase the velocity resolution, the signal can be zero-padded out to N_D samples before applying the Fourier transform.

$$\mathscr{F}\{\mathbf{H}_i\}(k,u) = \sum_{n=0}^{N_D-1} \hat{H}_k(n) W_k(n) e^{-j2\pi nu/N_D}$$

$$\simeq A_{p^*} \sum_{p=0}^{P'-1} A_p e^{-j2\pi(f_c+k/T)\hat{\ell}_p/c} \times$$

$$\times \sum_{n=0}^{N_D-1} W_k(n) e^{j2\pi n(f_c \hat{v}_p \cos \hat{\alpha}_p T_c/c - u/N_D)}. \tag{21}$$

$\square$

By summing over the subcarrier axis k, in Eq. (20), we preclude the possibility of retrieving the length ℓ_p of path p taking the maximum of the Fourier transform over k. However, by definition, ℓ_p is constant during a transmission slot and only depends on the position of the corresponding scattering point inside the room. In this work, we are interested in capturing the path variations ($\Delta_p(n)$) caused by a moving subject, thus the constant value ℓ_p is irrelevant to our HAR task.

The non-zero entries u in the Doppler vector reveal the presence of a scatterer with associated velocity

$$v_p \cos \alpha_p = \frac{uc}{f_c T_c N_D}. \tag{22}$$

Note that, we do not need to estimate v_p and $\cos \alpha_p$ separately. In fact, we are interested in finding a representative feature for the activity-related movements, i.e., we aim at revealing the dynamic component $\Delta_p(n)$ (Eq. (18)) of the paths in the environment. It follows that the quantity in Eq. (22) is a good proxy to such dynamic components and is a good input feature for SHARP. We recall that this feature is obtained considering time windows consisting of N subsequent channel estimates each. In the following, we refer to Doppler trace, or Doppler spectrogram, as the matrix obtained by stacking a number of subsequent Doppler vectors, computed for consecutive observation windows.

As an example, in Fig. 3 we plot the Doppler spectrograms as a person performs each of the four considered activities inside a room, compared with an empty room case. The spectrograms show the evolution of the quantity in Eq. (22) over three-seconds long observation windows and

the colormap refers to the normalized amplitudes in dB. As expected, the Doppler trace related to the empty room only presents non-negligible power at the zero-velocity bin, revealing no movement in the environment. Instead, in the presence of a moving person the power is spread across different bins, reflecting the human-related multi-path changes.

We stress that the main idea behind SHARP is to perform environment and person independent activity recognition. To this end, we exploit the proprieties of the Doppler spectrogram discussed above combined with the learning architecture detailed in the next section. The framework is trained on a single scenario (i.e., same environment and person) and its robustness is assessed considering different test cases. Note that the framework is used at testing time without performing any retraining, to assess its performance when the day, the person and/or the environment change with respect to those used at training time.

4 LEARNING ARCHITECTURE FOR HAR

SHARP is conceived to fully exploit the data gathered from a commercial Wi-Fi access point, i.e., using amplitude and phase information from the CFR, and combining this data for the available N_{ant} receiving antennas. Activity recognition is performed using N_w subsequent channel estimates at a time, which amounts to monitoring the channel for $N_w T_c$ seconds. Note that this can be implemented in a sliding window fashion, thus updating the activity label at every new CSI sample.

The HAR algorithm consists of two steps.

1) First, we compute the Doppler traces from the data collected at all the receiving antennas and we use them to obtain activity estimates though a neural-network based algorithm (Section 4.1). Specifically, the machine learning based prediction chain that we build is *independently applied* to the data stream coming from each of the available antennas.

2) As a result of the previous step, we obtain N_{ant} independent predictions, one per antenna, which are combined, in a second step, through a decision fusion method that leads to the final activity estimate (Section 4.2).

We remark that an alternative approach that combines the data from the antennas (at the input of the above decision chains), would also be possible. Such combined input data (the Doppler traces coming from the antennas) would then be fed to a neural network to classify the activity. We experimentally verified that this is not a viable method to
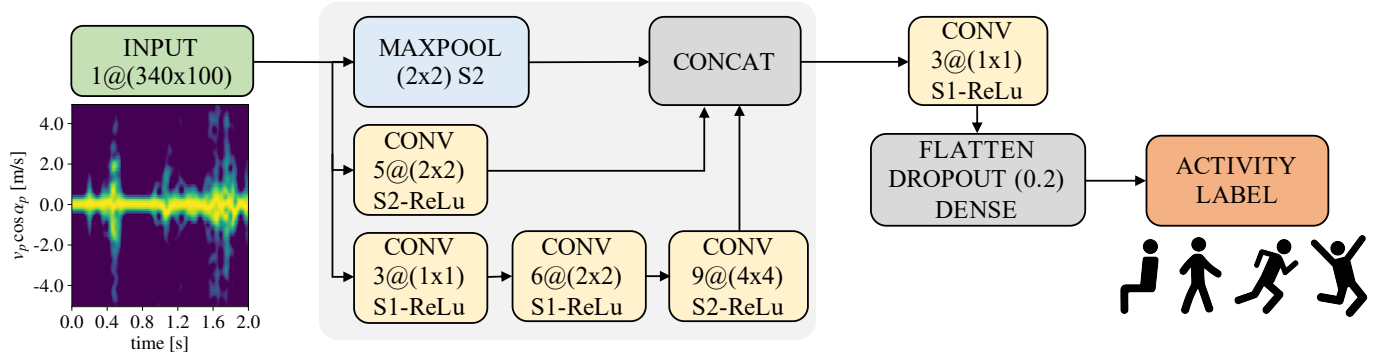
Fig. 4: SHARP classification architecture for a single antenna system. The input of the network consists of 340 (about two seconds) long Doppler traces with 100 velocity bins. [CONV maps@(kernel)] blocks are convolutional filters with stride S#, followed by ReLu activation functions.

obtain an environment-invariant classifier, as the classification result depends on the antenna ordering and, in turn, on the person's location inside the room.

4.1 Activity classifier for the single antenna system

We framed the problem as a multi-class classification task with five classes (four user activities plus empty room), tackling it through a learning-based approach. The designed neural-network classifier takes as input the Doppler trace from a single antenna element, i.e., an $N_w \times N_D$-dimensional matrix obtained by stacking N_w subsequent Doppler vectors, and returns the label of the activity being performed. The classification architecture for the single-antenna system is shown in Fig. 4.

Firstly, the input is fed to a simplified Inception module with dimensional reduction. This block extracts significant features from the Doppler trace at several scales by using layers with different kernel sizes in a parallel fashion. The structure is inspired by the *reduction block* proposed for the Inception-v4 neural network in [39], and consists of three branches combining max-pooling (MAXPOOL) and convolutional (CONV) layers. Overall, the proposed neural network has 128, 535 parameters. In the interest of obtaining a lightweight model, we do not use the full 17 blocks Inception-v4 network which consists of 43 million parameters, but leads to negligible performance improvements. The $N_w/2 \times N_D/2$ dimensional feature maps obtained at the output of each branch of the Inception module are concatenated and passed to the following convolutional filter with 1×1 kernel, used to reduce the number of feature maps from 15 to 3. Hence, the output of such filter is flattened and passed to a fully connected (dense) layer with five output neurons, one for each activity class. The Dropout technique is used as a regularization strategy, randomly zeroing 20% of the elements in the flatted vector preceding the dense layer. The classifier is trained in a supervised manner on data collected from a single indoor environment, using the cross-entropy loss function. The Doppler traces collected from the different antennas are used without any distinction among them, i.e., they are all added to a unique training set, without keeping track of the antenna that generated them. Using the trained architecture of Fig. 4, each input trace is associated with the activity having the highest score in the output five-dimensional vector, referred to as *activity vector*.

4.1.1 On the selection of the learning architecture

The Inception module was selected among other candidate learning architectures as the subject's movements introduce both small and large scale variations in the Doppler spectrogram (see, e.g., Fig. 3). As humans, we are able to detect and visually extract patterns at different scales, but this is not straightforward for a neural network, where each kernel captures features at a specific scale, depending on its size. The Inception module tries to mimic such a human-specific capability by simultaneously processing the input through different-sized kernels. As a consequence, in our case, the combination of the features extracted by an Inception module is expected to better represent the input than using a single kernel.

An alternative approach to obtain such multi-scale features is to implement a deep neural network by stacking several convolutional layers. In this way, fine-grained features are extracted in the first steps and coarser characteristics are derived subsequently [40]. However, this approach leads to an architecture with a considerable number of trainable parameters which entails a longer training time and, more importantly, an increased memory to store the model. This limits the applicability of the method in the case of devices with scarce computing resources. As an alternative approach for resource constrained devices, the authors of [41] propose the multi-scale dense-net (MSDNet). MSDNet consists of a convolutional architecture where layers of parallel sub-networks working at different scales are combined to reach scale invariance. In this way, fine- and coarse-grained features are extracted at each layer of the combined network. In turn, each layer can output a reliable label for the classification problem and, at run-time, the processing can be stopped as soon as computing resources are exhausted. A similar approach is presented in [42] with the introduction of the resolution adaptive network (RANet). As for MSDNet, several parallel sub-networks operating at different scales are used. However, while in MSDNet the classification is always performed at the lowest scale, the RANet classifiers are placed both in low and high-resolution layers. The rationale behind this choice is that some "easy" input samples can be recognized directly at a higher scale without extracting coarse-grained features, thus reducing the computation burden. Another scale-invariance approach is presented in [43], where the authors introduce ELASTIC, a strategy that can be applied to any convolutional archi-

tecture. Specifically, down- and up-sampling functions are added as parallel branches at each layer of the network to change the spatial resolution of the extracted features. This allows obtaining a dynamic resolution structure where the network can activate branches with different resolutions at each layer, adapting the network to each input signal.

Overall, the approaches in [41], [42], [43] are very useful in complex computer vision tasks. However, for our recognition problem, they would not lead to significant improvements at the cost of an increased model complexity. As our aim is to propose a lightweight approach that can be implemented in low-cost devices, we decided not to consider these more complex solutions.

4.2 Decision fusion for the multiple antenna system

At runtime, the trained classification engine from the single antenna system is independently applied to the Doppler stream gathered by each antenna. This returns N_{ant} independent classification outcomes that are combined as we now explain. In detail, for each antenna we obtain a five-dimensional *activity vector* (the classifier output) and an *activity label*, corresponding to the largest element in the activity vector. When at least $N_{\mathrm{ant}} - 1$ activity labels agree on a certain activity, there is a clear winner, and the Doppler trace is associated with that activity. Otherwise, an overall *decision vector* is computed by summing, element-wise, the N_{ant} activity vectors. The trace is then associated with the activity having the highest score in this decision vector.

5 EXPERIMENTAL SETUP

In this section we present the experimental setup designed to train and validate SHARP. We first introduce the CSI extraction method (Section 5.1), and then present the measurement scenarios and campaigns selected for building the dataset (Section 5.2).

5.1 Nexmon extraction tool

Although most commercial Wi-Fi chipsets can potentially generate CSI data, few manufacturers make this data available to developers and researchers, especially for modern chipsets. Hence, the majority of the Wi-Fi-based HAR works in the literature have used outdated chipsets, for which some CSI extraction tools have been developed over the years: widely used ones are [44], [45], that target network interface cards implementing the IEEE 802.11n protocol. Recently, as part of the Nexmon project [36], [46], [47], a firmware patch allowing the extraction of CSI from specific Broadcom/Cypress Wi-Fi chipsets has been released.

In the present work, we use the Nexmon CSI extraction tool presented in [48] to obtain CSI data from an Asus RT-AC86U IEEE 802.11ac Wi-Fi router. The extraction tool is compatible with the very-high-throughput mode, defined by IEEE 802.11ac, working with a total bandwidth of 80 MHz. Each CSI sample results in complex-valued channel information from 242 data sub-channels for each transmit-receive antennas pair. In our experiments, with one transmitter antenna and four at the monitoring device, each CSI sample corresponds to four vectors of 242 complex values. Although the total number of sub-channels at 80 MHz is 256, each antenna vector has 242 components as the CFR is only provided for data sub-channels,

namely sub-channels whose indexes are $\{-122, \ldots, -2\}$ and $\{2, \ldots, 122\}$, i.e., no CFR value is provided for the control sub-channels. Moreover, the values returned by the tool on the sub-channels from -63 to 122 need an inversion on the sign, probably due to hardware artifacts.

In the next sections, the Asus router that estimates the CFR is referred to as *monitor* device. The name reflects the status of the wireless interface of the router, that should be set in monitor mode to capture the packets sent over-the-air by other devices transmitting in its proximity.

5.2 Dataset acquisition and organization

In order to analyze the effects of different room geometries and static obstacles, we collected CSI samples in three different environments, i.e., a bedroom (Fig. 5-a), a living room (Fig. 5-b) and a University laboratory (Fig. 6), where one person moves within the area. Specifically, we obtained data from three volunteers (a male, P1, and two females, P2, P3) while they were walking or running around, jumping in place, or sitting somewhere in the room. We recall that, during a Wi-Fi communication, the CSI samples are computed based on known packet preambles that allow estimating the channel conditions. CSI samples can be estimated by any device having access to the wireless channel where the communication is ongoing. In our experimental setup, we considered a system made of two active terminals – i.e., that exchange Wi-Fi traffic – and one passive device configured in monitor mode. Such system reflects a real-world scenario where terminals (e.g., laptops and smartphones) transmit Wi-Fi packets that are leveraged by the monitor node as "signals of opportunity" to sense the surrounding environment through the computation of the CFR. The monitor node is implemented on an Asus router equipped with $N_{\mathrm{ant}} = 4$ antennas and running the Nexmon firmware (see Section 5.1). To generate the signals of opportunity, we set up a Wi-Fi transmission link using two Netgear X4S AC2600 Wi-Fi routers, equipped with Qualcomm Atheros chipsets and hereafter referred to as transmitter (Tx) and receiver (Rx) devices. Note that we used routers for experimental convenience: in a real setup Wi-Fi signals of opportunity can be generated by any other Wi-Fi-enabled device as it uses the spectrum for communications. The packets are transmitted through a single antenna, by using a fixed modulation and coding scheme – namely, MSC 4 – by disabling frame aggregation, and by setting the source rate to 173 packets per second. Since the Nexmon tool is configured for reading CSI samples on data packets only (i.e., by neglecting the acknowledgement frames sent by the receiver), a new channel estimate is generated every $T_c \simeq 6 \times 10^{-3}$ s.

The proposed setup reflects a real-life situation where Wi-Fi traffic is already present in the environment, so we only need a monitor node to obtain the CSI traces from the signals of opportunity. This is in line with the envisioned scenario for next-generation Wi-Fi routers implementing the upcoming IEEE 802.11bf standard [3]. The IEEE 802.11bf project started in September 2020 with the aim to empower Wi-Fi routers with sensing capabilities. The new version of the Wi-Fi standard will enable the joint provisioning of the communication and the sensing services to the users by introducing some modifications at the physical and medium access control layers of the protocol stack to fulfill the slightly different requirements of the two services. In
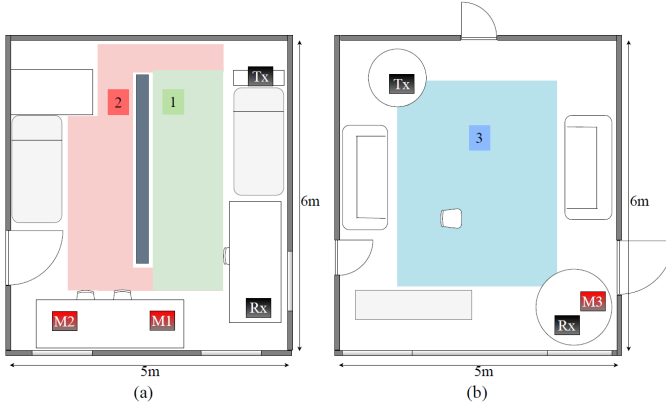
Fig. 5: Monitored environment for (a) sets $S1$ - $S5$, bedroom with a bookcase on the middle; (b) set $S6$, living room. `Tx` and `Rx` denote the transmitter and the receiver, respectively. `Mj`, $j \in \{1, 2, 3\}$, denotes the monitor station.
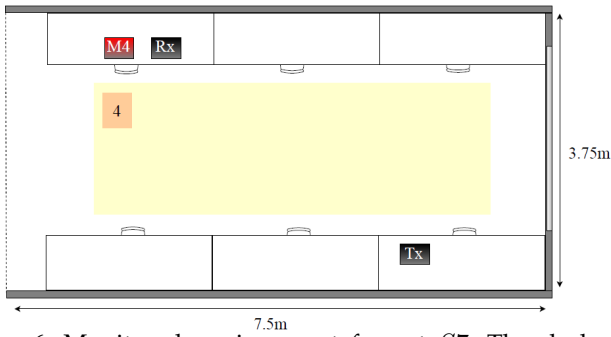

Fig. 6: Monitored environment for set $S7$. The desks are fully of computers and monitors. `Tx` and `Rx` denote the transmitter and the receiver, respectively. `M4` denotes the monitor station.

this scenario, a Wi-Fi router will serve both as an access point for communication purposes and a monitor device to provide the sensing service by leveraging the ongoing Wi-Fi transmissions. In our experimental setup, the monitor router provides the sensing functionality only as with the current physical and medium access control layer protocols the communication and the sensing services cannot be simultaneously provided to the users. However, our proposed sensing strategy will remain valid and effective for human activity recognition also once the two functionalities will be jointly supported by IEEE 802.11bf devices.

Fig. 5 and Fig. 6 display the positions of transmitter, receiver and monitor routers in the different environments. The black boxes represent the transmitter (`Tx`) and the receiver (`Rx`), while the red boxes labeled with `Mj` (with $j \in \{1, \ldots, 4\}$) indicate the position of the monitor station in different measurement sets. The activities are performed in the areas identified by a color and a number. Note that in the bedroom (Fig. 5-a), the direct path between `Tx` and `M2` is occluded by the bookcase in the middle of the room (grey rectangle in the figure).

We performed several measurement campaigns and we grouped them into seven sets, Sj with $j \in \{1, ..., 7\}$ each corresponding to a different triplet of environment-day-person. For each set, Table 1 provides the position of the monitoring station (`Mj`), the area where the activities take place (identified by the color), and the person performing them (`Pi`). We also include an indication of whether there

set	monitor position, `Mj`	person, `Pi`	direct path	type
$S1$	M1	P1	yes	train-test
$S2$	M1	P1	yes	test
$S3$	M1	P2	yes	test
$S4$	M2	P1	no	test
$S5$	M2	P2	no	test
$S6$	M3	P1	yes	test
$S7$	M4	P3	yes	test

TABLE 1: Measurement conditions. For each set Sj we specify the position of the monitor station (`Mj`) (see Fig. 5 and Fig. 6), the person (`Pi`) performing the activity, and the presence of a direct path between the transmitter and the monitor. The last column indicates whether the set is used for training SHARP or testing its performance.

communication and processing parameters	
monitored channel	IEEE 802.11ac ch no. 42
OFDM sample duration, T	3.2×10^{-6} s
modulation and coding scheme	MCS 4
channel estimates interval, T_c	$\simeq 6 \times 10^{-3}$ s
no. OFDM sub-channels, M	256 (245 used)
no. ch. estimates for Doppler computation, N	31
no. bins in a Doppler vector, N_D	100
no. Doppler vectors per window, N_w	340
no. monitoring antennas, N_{ant}	4

TABLE 2: Summary of parameters related to the communication setup, and the Doppler traces computation.

is a direct path between the transmitter and the monitor stations. The configuration for sets $S1$-$S2$ is the same apart from the day of measurement. Set $S1$ is used to train the model, while set $S2$ is used to test the generalization over different days. In $S3$ we monitor the same environment as the previous sets, but in a different day and with a different person performing the activities. For sets $S4$ and $S5$, the person is required to move in both area 1 and 2 of the bedroom depicted in Fig. 5-a. In these configurations, the direct path between the transmitter and the monitor is disturbed by the bookcase. Set $S6$-$S7$ are collected in two different days and environments. $S7$ represents the more challenging situation, in which no element (room, day, person) is in common with the training set scenario.

Each measurement campaign involves 120 seconds of data for each activity, plus an additional trace of 120 seconds of data collected when the room is empty. Activities are repeated continuously by the volunteers during the trace acquisition time. The campaigns have been performed in different days through several months (April-December, 2020 and January 2022), resulting in a considerable time diversity. Overall, we collected nearly 120 minutes of CSI data, consisting of the CFRs estimated at all the four antennas of the monitor station.

6 EXPERIMENTAL RESULTS

SHARP has been tested in the scenarios detailed in Table 1. The adopted communication and processing parameters are summarized in Table 2 and are presented in Section 6.1. The performance is discussed in Section 6.2. Our implementation of SHARP, along with the collected activity dataset, are available at [1].

6.1 Pre-processing steps

We considered some pre-processing operations for extracting the features of interest from the dataset. For each CSI

		empty	sitting	walking	running	jumping	**mean**
$S1$	accuracy (%)	100	100	100	100	97.93	**99.59**
	F1-score	1	1	0.997	0.996	0.990	**0.996**
$S2$	accuracy (%)	100	100	100	99.28	99.64	**99.78**
	F1-score	1	0.998	0.999	0.996	0.998	**0.998**
$S3$	accuracy (%)	100	100	96.64	100	100	**99.33**
	F1-score	1	1	0.983	0.984	1	**0.993**
$S4$	accuracy (%)	100	100	99.84	93.27	94.63	**97.55**
	F1-score	1	1	0.965	0.941	0.972	**0.976**
$S5$	accuracy (%)	100	100	98.58	84.93	99.45	**96.59**
	F1-score	1	1	0.923	0.911	0.997	**0.966**
$S6$	accuracy (%)	99.76	100	100	100	99.20	**99.79**
	F1-score	0.999	0.999	0.999	0.997	0.996	**0.998**
$S7$	accuracy (%)	100	100	81.24	98.71	100	**95.99**
	F1-score	1	1	0.89	0.909	1	**0.960**

TABLE 3: SHARP performance (decision fusion). The accuracy and the F1-score are reported for each of the five classes, along with their average value.

Fig. 7: SHARP normalized confusion matrix for test set $S7$ (worst case scenario). Environment, day and person change with respect to the training.

sample, we divided the CFR values by the mean amplitude over the 242 monitored sub-channels to remove unwanted amplifications. Then, the phase sanitization algorithm presented in Section 3.1 was applied, fixing $\lambda = 10^{-1}$. We reconstructed the CFR values on the three central sub-channels together with the other 242 using Eq. (14), thus obtaining a CFR complex-valued vector (amplitudes and phases) of 245 components. The Doppler vectors have been computed considering $N = 31$ subsequent sanitized CFR samples. The velocity resolution was increased by zero-padding the signal out to $N_D = 100$ points before applying the Fourier transform. A threshold was used on the resulting Doppler vectors to remove noisy contributions with power smaller than 12 dB. Finally, the Doppler trace acting as input feature for the HAR system was built by stacking $N_w = 340$ consecutive Doppler vectors. The Doppler vectors were generated for each CSI acquisition by using a sliding window mechanism. Therefore, the complete Doppler trace lasted roughly $340 \cdot T_c = 2$ s of measurements.

6.2 SHARP training and performance assessment

The HAR learning algorithm presented in Section 4 has been trained by using the features extracted on set $S1$. Specifically, 60% of the data in $S1$ makes up the training set, while the remaining 40% is evenly split between the validation and the test sets. The other measurement sets, i.e., $S2$-$S7$, are only considered in the test phase. The classification accuracy and the F1-score obtained on each of the test sets are reported in Table 3. Note that, for $S1$, the evaluation is performed on the test set only, i.e., on new data never seen during training. Overall, the mean recognition accuracy is higher than 95%, reaching almost 100% when the environment and location of the monitor node ($M1$) remain the same of the training data, regardless of day of measurement and the person performing the activity ($P1$ or $P2$). We still obtain a good accuracy, around 97%, when the direct path between the transmitter and the monitor stations is blocked (sets $S4$ and $S5$). In this situation, the running activity is sometimes confused with the walking one, as revealed by the F1-score associated with these classes. However, this is a reasonable limit of our approach, as the activities are recognized by considering the velocity of different parts of the human body (i.e., the
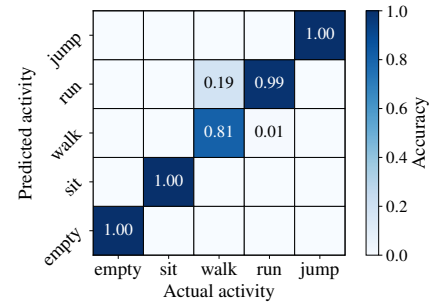
scattering sources), which can be very similar when people walk or run in indoor spaces. We also believe that a small confusion between the running and walking activities is admissible in almost all the fields of application of a Wi-Fi sensing system. For example, in a residential scenario, it is more relevant to correctly distinguish between static and dynamic activities, rather than providing details about the type of movement.

For testing the generality of our algorithm under different environments, we consider the measurement sets $S6$ and $S7$. When the person is the same of the training data, the average HAR accuracy approaches 100% while it decreases to 95.99% when both the person and the room change. Again, the lower accuracy is achieved for the walking activity, which is wrongly classified as the running one, as displayed in the normalized confusion matrix in Fig. 7.

6.3 On the selection of the phase sanitization approach

The removal of the phase offsets is necessary to obtain meaningful Doppler traces. The unwanted offsets, and in particular the CFO, disrupt the Doppler shift information contained in the CFR data, making its extraction impossible. Prior to developing our own phase sanitization method (see Section 3.1), we investigated existing phase sanitization techniques from the literature. The comparison between these strategies is not straightforward due to two main reasons. First, the target phase (i.e., the one without offsets) can be hardly obtained through a mathematical model, as the multi-path propagation is an environment-specific process that depends on many factors. One may think of evaluating the effectiveness of a phase sanitization algorithm by developing a simulator of the Wi-Fi channel, including the phase offsets introduced by the hardware imperfections. However, even with such simulator a fair comparison among the different strategies would be difficult, as they use different reference systems for the sanitized output signal. Specifically, the extensively adopted conjugate-multiplication between antennas (see, e.g., [5], [23], [24], [25]), uses one of the antennas as the reference for the processed signal. The method in [13], instead, refers the signal to its static component, while our proposed approach considers the strongest channel path as the reference.

As our main objective is to develop an environment-independent algorithm for HAR, we compare these phase sanitization approaches based on the quality of their features, i.e., by analyzing their respective Doppler spectrograms. Fig. 8 shows the Doppler spectrogram computed
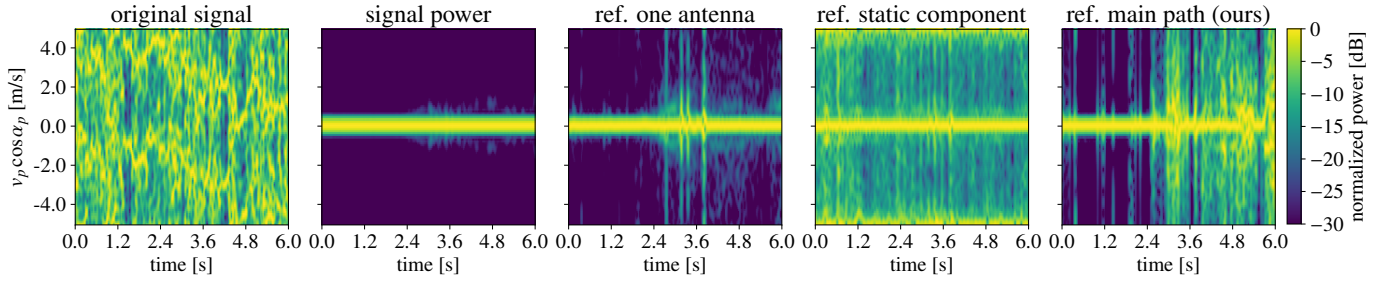
Fig. 8: Doppler spectrograms computed from a six seconds long CFR signal, where the activity performed are sitting ($[0, 3]$ s) and running ($[3, 6]$ s). The first two subfigures show the Doppler extracted from the original signal (containing offsets) and the signal power. The following three plots represent the Doppler computed on the sanited signal – i.e., obtained with three phase sanitization techniques – labeled according to the reference (ref.) used.

from a 6 second long CFR trace and its sanitized versions using the just mentioned techniques as a preprocessing step on the signal. In the considered time window, a person sits for the first 3 seconds and then starts running inside the room. The "original signal" is clearly too noisy to be useful. A straightforward strategy, proposed in [15] to solve this issue, consists in computing the spectrogram from the power of the original signal that, naturally, is not affected by phase offsets. The result depicted in the second subfigure – entitled "signal power" – indicates that the sole amplitude information is not sufficient to obtain a clear spectrogram. The Doppler trace obtained with our sanitization algorithm (see Section 3.1) is referred to as "ref. main path", while the "ref. one antenna" and "ref. static component" respectively refer to the approaches in [5] and [13]. The figure reveals that our proposed methodology allows obtaining a clearer Doppler spectrogram where the segments corresponding to the two different situations are more distinguishable. The use of one antenna as the reference allows removing most of the phase offset contributions in the CFR, but it also removes some relevant information related to the environmental changes. This is due to the conjugate multiplication operation, which can lead to destructive interference between the signal to be cleaned and the reference. On the other hand, when the reference is the static component, the processed signal contains some noisy contributions which are probably associated with energy leakage between the static and the dynamic part. This prevents the correct estimation of the reference used for phase sanitization. We experimentally verified that these previous algorithms do not lead to good HAR performance. These drawbacks led us to the development of our approach, detailed in the earlier Section 3.1. A quantitative evaluation of the impact of the phase sanitization strategy is reported in Table 4. Specifically, we assessed the accuracy and the F1-score of the classification algorithm (see Section 4) using the Doppler spectrograms computed from the traces sanitized with the "ref. one antenna" strategy, and we compared the results with those obtained through our approach in Table 3 ("ref. main path"). We selected the "ref. one antenna" as a benchmark for comparison as it returns the cleanest Doppler spectrograms among the other candidate approaches in the literature (see Fig. 8). The results confirm the importance of the phase sanitization step for the extraction of informative Doppler traces that can serve as feature maps for HAR. Note that the "ref. one antenna" is the most used approach in the literature. However, existing works on HAR only focus on

		$S1$	$S2$	$S3$	$S4$	$S5$	$S6$	$S7$
ref. one ant.	accuracy (%)	99.23	59.77	75.99	43.45	39.20	59.44	60.31
	F1-score	0.992	0.563	0.698	0.323	0.273	0.556	0.559
ref. main path	accuracy (%)	99.59	99.78	99.33	97.55	96.59	99.79	95.99
	F1-score	0.996	0.998	0.993	0.976	0.966	0.998	0.960

TABLE 4: Impact of using different phase sanitization methods for the estimation of the Doppler traces. Our approach ("ref. main path") is here compared with the most adopted approach in the literature ("ref. one antenna").

the scenario here identified as $S1$. In $S1$, the performance obtained through the two sanitization methods are similar as the person and the environment remain the same at training and testing time. In this case, coarser Doppler spectrograms suffice to perform a correct classification of the activities. However, once the environment, the person or both change at testing time ($S2$-$S7$), the performance of the "ref. one antenna" sanitization method degrades, while the accuracy of our proposed strategy ("ref. main path") remains above 95%.

6.4 On the selection of the fusion approach

SHARP relies on the decision fusion strategy detailed in Section 4.2. As an alternative, one may implement a data fusion approach. This consists in merging the data from the four antennas at the input of the neural network instead of combining the outputs of the classification performed on the single antennas. Hence, the $N_w \times N_D \times N_{ant}$-dimensional input is forwarded through the neural network that returns the most probable human activity class. We implemented this approach by applying some slight modifications to the SHARP classification architecture of Section 4.1. Specifically, we doubled the feature maps of the convolutional layer in the second branch of the simplified Inception block, while we used respectively 6, 9 and 12 maps for the convolutional layers in the third branch. These adaptations are needed because of the higher dimensionality of the input matrix and lead to a total of $129, 751$ trainable parameters. The network is trained by randomly swapping the order of the antennas in the $N_w \times N_D \times N_{ant}$-dimensional input matrices to make the model independent of the antennas' positions along the channels dimension. Without the use of this strategy, the network would learn which antenna is the most important in the training scenario and will mostly base its classification on that one. However, the most significant antenna is likely to change when changing the environmental conditions thus leading to poorer classification performance. The accuracy and F1-score of the data-fusion strategy averaged over the

		$S1$	$S2$	$S3$	$S4$	$S5$	$S6$	$S7$
data fus.	accuracy (%)	99.80	97.45	89.43	88.51	81.68	98.18	88.02
	F1-score	0.998	0.966	0.890	0.884	0.796	0.982	0.881
decision fus.	accuracy (%)	99.59	99.78	99.33	97.55	96.59	99.79	95.99
	F1-score	0.996	0.998	0.993	0.976	0.966	0.998	0.960

TABLE 5: Performance of the data fusion (fus.) approach compared to the decision fusion strategy in Section 4.2.

five considered activities are reported in Table 5. The results show that the data fusion approach is less effective than the decision fusion one in recognizing the activity being performed by the subject, especially when the LOS is blocked – $S4$ and $S5$ – or the environment and subject both differ than in the training phase – $S7$. This because by merging the four monitoring antennas' data at the input of the neural network architecture, we force the algorithm to learn useful information from all the monitoring antennas. The shuffling of the antennas' position at the input is required to grant independency from the antennas' physical positioning. However, as a result, the signal captured by each monitor antenna has almost the same weight (importance) in the online classification outcome. This entails the fact that antennas that lead to a misclassifying of the activity have a considerable impact on the final activity classification as well. This drawback is bypassed by our proposed decision fusion approach of Section 4.2: with it, we still take into account all of the antennas' data when forwarding them through the neural network, one at a time. However, at the output, the decisions based on the single antennas are combined in a way that an antenna likely to lead to a misclassification – as it is in discordance with the others – is not considered for the final result.

6.5 On the impact of the number of monitor antennas

Here we evaluate the performance of SHARP when increasing the number of monitor antennas from 1 to 4. The neural network is trained as detailed in Section 4.1. The classification output is obtained considering subsets of the four available antennas by applying the decision fusion algorithm of Section 4.2. The results are reported in Fig. 9(a). The activity recognition accuracy for $N_{\mathrm{sub}} \in \{1, \dots, 4\}$ antennas is obtained by averaging the metric over all activities and combinations of N_{sub} antennas. The results confirm that the higher the number of antennas, the higher the accuracy. This is associated with the amount of information that is collected for the classification. The accuracy is always above 80% even with a single antenna. However, as confirmed by the results, the spatial diversity provided by the antenna array is beneficial for increasing the robustness of the system to changing environmental conditions. Having more antennas increases the possibility to sense environmental changes as the signals are collected at slightly different spatial positions.

6.6 HAR algorithms performance comparison

In Figs. 9(b)-9(c), SHARP is compared against three HAR systems from the literature: DeepSense [11], EI [12] and MatNet-eCSI [13]. *DeepSense* and *EI* are used in place of the single antenna classifier described in Section 4.1 (i.e., before the decision fusion step) and rely on the sole CSI amplitude. *MatNet-eCSI*, instead, considers phase and amplitude information, and takes as input their combination at all the antennas. The three approaches use different neural network

architectures. In the interest of a fair comparison, we trained them with the same portion of the set $S1$ used to train SHARP. Note that MatNet-eCSI would require two training steps: the first performed on the training scenario, whereas the second on the test scenario, to fine tune the classification architecture to the new setup. For a fair comparison, in our evaluation we only considered the first step (using $S1$ data), i.e., without benefiting from additional data collected in the test scenarios. Specifically, we respectively used the training and the validation portions of $S1$ as the *reference* and the *target data* for the MatNet-eCSI neural network. For the EI algorithm, we considered the three different days containing the measurements for the training set $S1$ as the different domains for its adversarial learning framework. Then, for DeepSense and EI, the decision fusion approach of Section 4.2 is applied at runtime to combine the output of the classifiers operating on the single antenna data and obtain the final activity label. As MatNet-eCSI already takes the combination of features from all the antennas as input to its neural network classifier, the decision fusion is not performed for this scheme.

From Figs. 9(b)-9(c), we observe that the performance of all the considered competing algorithms is in line with that of SHARP when tested on set $S1$, while it substantially degrades in all the other cases, even when data are collected in the same environment used for training, but on a different day (e.g., set $S2$). These results demonstrate that the sole amplitude information (DeepSense and EI) or its combination with the phase (MatNet-eCSI) are insufficient for the classifiers to generalize well across different scenarios. Indeed, as shown and discussed in Section 3, the CFR amplitude and its phase are environment-dependent and are affected by small scene variations over days, such as changes in the positioning of the monitor node, room obstacles, and objects. Conversely, SHARP generalizes across different days, environments and persons, by reliably distinguishing between static and dynamic situations. This is achieved by: i) exploiting the Doppler shift as the input feature for HAR, and ii) using the Inception module as the core of the learning-based classifier, to automatically extract coarse- and fine-grained details from the Doppler spectrograms.

6.7 SHARP for smart-living applications

In this section, we assess the performance of SHARP in recognizing a larger set of activities within a smart-living scenario. In addition to those already investigated, we added three additional human movements to the dataset, namely, standing still, sitting down/standing up and doing arm gymnastics, by retraining SHARP to distinguish them. The number of neurons in the last layer SHARP neural network classifier was changed to reflect the new number of classes. We experimentally verified that the standing-up and sitting-down movements produce an almost indistinguishable effect on the Doppler traces extracted from the Wi-Fi router. This is linked with the physical limitations given by the operational parameters (bandwidth and carrier frequency) of the used off-the-shelf Wi-Fi-based system. The obtained accuracy and F1-score are reported in Table 6 considering a single subject in LOS conditions. These results show that SHARP recognizes all the eight considered activities with high accuracy.

(a) Impact of the number of monitor antennas combined by the decision fusion algorithm on the average recognition accuracy.

(b) Average accuracy of SHARP compared against the three considered approaches from the literature.

(c) Average F1-score of SHARP compared against the three considered approaches from the literature.
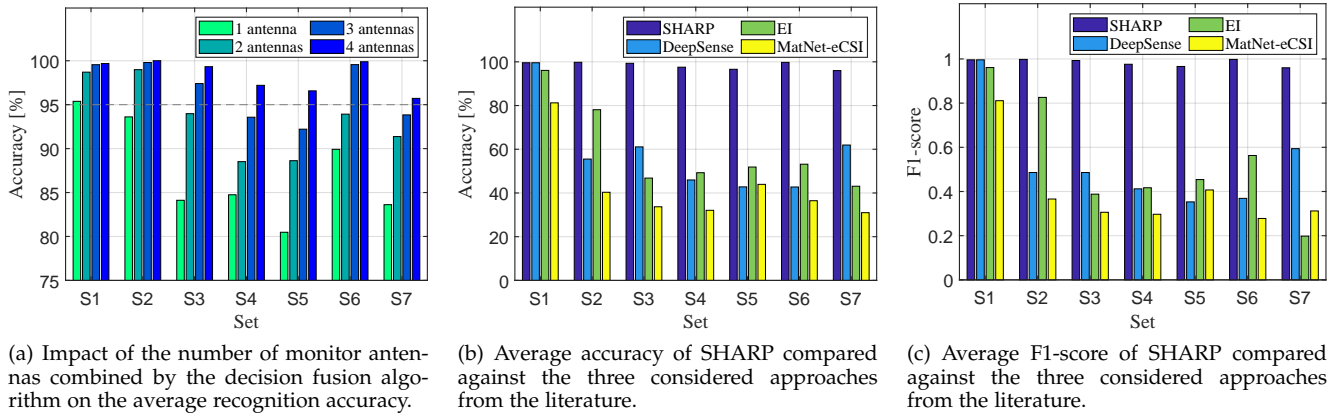
Fig. 9: SHARP performance when changing the number of monitor antennas and comparison against three HAR algorithms from the literature – namely DeepSense [11], EI [12], and MatNet-eCSI [13] – for all the configurations in Table 1.

		empty	sitting	standing	sit down stand up	walking	running	jumping	arm gym	mean
$S1$	acc. (%)	100	100	100	100	100	100	100	96.02	**99.50**
	F1-score	1	1	1	0.994	1	1	0.978	0.980	**0.994**
$S2$	acc. (%)	100	100	100	100	100	99.84	100	94.56	**99.30**
	F1-score	1	1	1	0.994	1	0.999	0.976	0.972	**0.993**
$S6$	acc. (%)	100	100	100	100	100	73.16	99.82	90.22	**95.40**
	F1-score	1	1	0.997	0.960	0.880	0.844	0.995	0.949	**0.953**

TABLE 6: SHARP performance with eight activities for scenarios $S1$, $S2$ and $S6$. The accuracy (acc.) and the F1-score are shown for each, along with their average value.

A remark is in order: The five activities considered in Table 3 are reliably distinguished by SHARP across subjects, days and environments. This is due to the fact that the corresponding Doppler traces contain sufficiently separable dynamics, which can be reliably isolated. Instead, the three new activities that are considered in this section generate Doppler traces that are similar to those from the previous five. As such, their reliable separation requires looking at finer details, i.e., at subject-specific traits. This means that SHARP can reliably distinguish them across different days and environments, via a subject-specific training.

7 CONCLUDING REMARKS

In this work, SHARP, a novel low-cost system for human activity recognition (HAR) in indoor spaces has been presented and experimentally validated. This system analyzes Wi-Fi signals scattered into the environment and runs on COTS Wi-Fi routers, which are usually available in indoor spaces for communication purposes, thus making the deployment and maintenance of an ad-hoc sensing infrastructure unnecessary. The use of Wi-Fi routers as environmental sensors is enabled by the possibility of reliably estimating the wireless channel frequency response while receiving background data traffic.

In contrast with previous solutions, we implemented a *robust* HAR system that does not require complex and periodic calibrations, i.e., once trained, it can be used at run-time for a completely unseen setup (new environment and individuals). SHARP revolves around the idea of processing the CSI data gathered by a monitor access point to quantify the Doppler effect due to the presence of moving objects over time. Indeed, the Doppler shift describes the velocity of the scattering points (i.e., the human body parts), whose temporal trace depends on the specific activity and is neither affected by the environment geometry, nor by static objects. A learning based algorithm, working on the Doppler trace, has been proposed to distinguish among different activities.

Four coarse-grained human activities, i.e., sitting, walking, running and jumping are considered, together with the empty space recognition. The robustness of SHARP has been challenged considering seven combinations of environments, access point locations, individuals and days. In each scenario, several measurement campaigns have been conducted, using a four-antennas IEEE 802.11ac router by Asus, operating on a frequency band of 80 MHz. Experimental results show that our system achieves a very high accuracy in all the considered scenarios. In the most challenging situation, i.e., when the person and the environment change with respect to those used at training time, the average accuracy is 96%. Moreover, the system outperforms state-of-art methods based on the analysis of the CSI amplitude and phase. Small errors are observed only when working with unknown persons and when distinguishing between walking and running. These errors somehow represent the fundamental limits of the approach, because they are due to the behavioral differences among different persons, or to the movement similarities in walking and running activities. We experimentally verified that SHARP is effective even when considering additional activities with similar dynamics, namely standing, sitting down / standing up and doing arm gymnastics. The accuracy remains above 95% when the performance is assessed in an environment that is different from the training one.

Future research avenues include the investigation of algorithms that work across differing hardware, at both the transmitter and the monitor access point, as well as the design of a new learning based algorithm that identifies multiple persons moving concurrently in the environment. Both extensions have important practical implications for real deployments and applications.

opinions expressed in this work are those of the authors and do not necessarily reflect those of the funding institutions.

APPENDIX A
OFDM MODEL FOR THE WI-FI CHANNEL

Next, we detail the main blocks of a Wi-Fi transmission chain, deriving the expression for the CFR in Eq. (2).

A.1 Transmitted signal model

In this subsection, we summarize how orthogonal frequency-division multiplexing (OFDM) is implemented by an IEEE 802.11 transmitter. OFDM uses a large number of closely spaced orthogonal sub-channels, transmitted in parallel. Each sub-channel is modulated with a conventional digital modulation scheme (such as QPSK, 16QAM, etc.).

The input bits are grouped and mapped onto source data symbols, which are complex numbers representing the modulation constellation points. Such constellation points are further grouped into K elements each, i.e., the OFDM symbols. OFDM symbols are then fed to an IFFT block that transforms the data prior to transmitting it over the wireless channel, in parallel, over K sub-channels with carriers spaced apart by $\Delta f = 1/T$ Hz (T is the OFDM symbol time).

The duration of an OFDM symbol is $\bar{T} = T + T_{\mathrm{CP}}$, where T_{CP} is the duration of the cyclic prefix, added to mitigate inter-symbol interference. Specifically, for IEEE 802.11ac the transmission bandwidth is 80 MHz, the samples are clocked out at 80 Msps, the number of sub-channels is $K = 256$, $T = 1/\Delta f = 3.2$ μs (i.e., $\Delta f = 312.5$ kHz), $T_{\mathrm{CP}} = 0.8$ μs, and, in turn, $\bar{T} = 4$ μs.

Let $\mathbf{a}_m = [a_{m,-K/2}, \dots, a_{m,K/2-1}]$ be m-th OFDM symbol, where $a_{m,k}$ is the k-th OFDM sample. After digital to analog conversion, the baseband OFDM signal for the m-th symbol is

$$x_m(t) = \sum_{k=-K/2}^{K/2-1} a_{m,k} e^{j2\pi kt/T}, \qquad (23)$$

where $k \in \{-K/2, \dots, K/2-1\}$ is the sub-channel index. Considering M subsequent blocks, the baseband signal is

$$x(t) = \sum_{m=0}^{M-1} x_m(t)\xi(t - m\bar{T}), \qquad (24)$$

with

$$\xi(t) = \begin{cases} 1 \text{ if } t \in [-T_{\mathrm{CP}} - TK/2, TK/2] \\ 0 \text{ otherwise} \end{cases}, \qquad (25)$$

and the signal transmitted over the Wi-Fi channel is obtained by upconverting $x(t)$ to the carrier frequency f_c,

$$s_{\mathrm{tx}}(t) = e^{j2\pi f_c t} x(t). \qquad (26)$$

A.2 Received signal model

At each receiver antenna, P signal copies are collected, due to the scatterers that the signal $s_{\mathrm{tx}}(t)$ encounters (multi-path propagation). Each path p is characterized by an attenuation

$A_p(t)$ and a delay $\tau_p(t)$. Neglecting the additive white Gaussian noise, the received signal $s_{\mathrm{rx}}(t)$ is written as

$$s_{\mathrm{rx}}(t) = \sum_{p=0}^{P-1} A_p(t) s_{\mathrm{tx}}(t - \tau_p(t))$$
$$= e^{j2\pi f_c t} \sum_{p=0}^{P-1} A_p(t) e^{-j2\pi f_c \tau_p(t)} x(t - \tau_p(t)), \qquad (27)$$

and its baseband representation $y(t)$ is expressed as,

$$y(t) = s_{\mathrm{rx}}(t) e^{-j2\pi f_c t}. \qquad (28)$$

A rectangular window $[m\bar{T}, m\bar{T} + T]$ is used at the receiver to collect and decode the information carried by an OFDM symbol at a time. Without loss of generality, we assume $m = 0$ and hence we omit such index from the following equations. The transmitted symbol a_k is recovered by computing the Fourier transform of the signal in the received window:

$$\hat{a}_k = \int_{\bar{T}}^{\bar{T}+T} y(t) e^{-j2\pi kt/T} dt$$
$$= \sum_{p=0}^{P-1} A_p e^{-j2\pi f_c \tau_p} \sum_{b=-K/2}^{K/2} a_b e^{-j2\pi b\tau_p/T} \times$$
$$\times \int_{\bar{T}}^{\bar{T}+T} e^{j2\pi(b-k)t/T} dt \qquad (29)$$
$$= a_k T \sum_{p=0}^{P-1} A_p e^{-j2\pi(f_c+k/T)\tau_p},$$

where we consider A_p and τ_p constant over an integration interval T. The sum in the last line of Eq. (29) corresponds to the frequency response of the Wi-Fi channel,

$$H_k = \sum_{p=0}^{P-1} A_p e^{-j2\pi(f_c+k/T)\tau_p}, \qquad (30)$$

that is estimated based on the known preamble symbols. In Eq. (29), we consider that the path attenuation and delay remain constant over each window, i.e., $A_p(t) = A_p$ and $\tau_p(t) = \tau_p$. Also, exchanging the order of integration and summation is legitimate as we deal with finite quantities, and we used $\int_{\bar{T}}^{\bar{T}+T} e^{j2\pi(b-k)t/T} dt = 0$ if $k \neq b$.

A.3 Phase offsets in the Wi-Fi channel estimates

Hardware artifacts make the CFR gathered from Wi-Fi devices slightly deviate from the model in Eq. (30). These artifacts introduce offsets (rotation errors) in the phase information, among which the most significant are [26], [37]:

- *carrier frequency offset (CFO)*, due to the difference between the carrier frequency of the transmitted signal and the one measured at the receiver. The CFO is only partially compensated for at the receiver [28].
- *sampling frequency offset (SFO)*, due to the imperfect synchronization of the clocks between transmitter and receiver.
- *packet detection delay (PDD)*, due to the time required to recover the transmitted modulated symbols from the received signal [28].

- *phase-locked loop phase offset (PPO)*, due to the phase-locked loop (PLL), the entity responsible for randomly generating the initial phase at the transmitter.
- *phase ambiguity (PA)*, due to the phase difference (multiples of π) between the antennas, that in static conditions should remain constant.

Considering these contributions, the complete expression for the phase of the p-th path in the received signal is

$$
\begin{aligned}
\bar{\phi}_{p,k} = & -2\pi(f_c + k/T)\tau_p + \phi_{\mathrm{CFO}} \\
& -2\pi k(\tau_{\mathrm{SFO}} + \tau_{\mathrm{PDD}})/T \\
& + \phi_{\mathrm{PPO}} + \phi_{\mathrm{PA}}.
\end{aligned}
\tag{31}
$$

Note that, while the CFO, SFO and PDD contributions take the same value across different antennas, the initial PLL phase (PPO) and PA are antenna specific [49], [50].

REFERENCES

[1] F. Meneghello, D. Garlisi, N. Dal Fabbro, I. Tinnirello, and M. Rossi, "SHARP – code and dataset," https://github.com/signetlabdei/SHARP, 2021.

[2] B. Dong, V. Prakash, F. Feng, and Z. O'Neill, "A review of smart building sensing system for better indoor environment control," *Energy and Buildings*, vol. 199, pp. 29 – 46, 2019.

[3] F. Restuccia, "IEEE 802.11bf: Toward ubiquitous Wi-Fi sensing," *arXiv preprint arXiv:2103.14918*, 2021.

[4] F. Adib and D. Katabi, "See through walls with WiFi!" *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 75–86, Aug. 2013.

[5] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, Seoul, Republic of Korea, 2019.

[6] B. Wang, Q. Xu, C. Chen, F. Zhang, and K. J. R. Liu, "The promise of radio analytics: A future paradigm of wireless positioning, tracking, and sensing," *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 59–80, 2018.

[7] Y. He, Y. Chen, Y. Hu, and B. Zeng, "WiFi vision: Sensing, recognition, and detection with commodity MIMO-OFDM WiFi," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8296–8317, 2020.

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

[9] H. Cui and N. Dahnoun, "Real-time short-range human posture estimation using mmWave radars and neural networks," *IEEE Sensors Journal*, vol. 22, no. 1, pp. 535–543, 2022.

[10] A. Sengupta and S. Cao, "mmPose-NLP: A natural language processing approach to precise skeletal pose estimation using mmwave radars," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022.

[11] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, and C. J. Spanos, "DeepSense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network," in *Proceedings of the IEEE International Conference on Communications (ICC)*, Kansas City, MO, USA, 2018.

[12] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2018.

[13] Z. Shi, J. A. Zhang, Y. D. R. Xu, and Q. Cheng, "Environment-robust device-free human activity recognition with channel-state-information enhancement and one-shot learning," *IEEE Transactions on Mobile Computing*, 2020.

[14] 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 the 20th Annual International Conference on Mobile Computing and Networking*, Maui, Hawaii, USA, 2014.

[15] 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 the 21st Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2015.

[16] J. Ding and Y. Wang, "WiFi CSI-based human activity recognition using deep recurrent neural network," *IEEE Access*, vol. 7, pp. 174 257–174 269, 2019.

[17] N. Damodaran, E. Haruni, M. Kokhkharova, and J. Schäfer, "Device free human activity and fall recognition using WiFi channel state information (CSI)," *CCF Transaction on Pervasive Computing Interaction*, vol. 2, pp. 1–17, 2020.

[18] T. Tegou, A. Papadopoulos, I. Kalamaras, K. Votis, and D. Tzovaras, "Using auditory features for WiFi channel state information activity recognition," *SN computer science*, vol. 1, no. 3, 2020.

[19] J. K. Brinke and N. Meratnia, "Scaling activity recognition using channel state information through convolutional neural networks and transfer learning," in *Proceedings of the First International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things*, ser. AIChallengeIoT'19. New York, NY, USA: Association for Computing Machinery, 2019.

[20] C. Xiao, D. Han, Y. Ma, and Z. Qin, "CsiGAN: Robust channel state information-based activity recognition with GANs," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 191–10 204, 2019.

[21] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proceedings of the Advances in Neural Information Processing Systems*, Barcelona, Spain, 2016.

[22] Y. Ma, S. Arshad, S. Muniraju, E. Torkildson, E. Rantala, K. Doppler, and G. Zhou, "Location- and person-independent activity recognition with WiFi, deep neural networks, and reinforcement learning," *ACM Trans. Internet Things*, vol. 2, no. 1, 2021.

[23] X. Wang, C. Yang, and S. Mao, "On CSI-based vital sign monitoring using commodity WiFi," *ACM Transactions on Computing for Healthcare*, vol. 1, no. 3, 2020.

[24] Y. Zeng, D. Wu, R. Gao, T. Gu, and D. Zhang, "FullBreathe: Full human respiration detection exploiting complementary of CSI phase and amplitude of WiFi signals," *Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, 2018.

[25] D. Zhang, Y. Hu, Y. Chen, and B. Zeng, "BreathTrack: Tracking indoor human breath status via commodity WiFi," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3899–3911, 2019.

[26] N. Keerativoranan, A. Haniz, K. Saito, and J. I. Takada, "Mitigation of CSI temporal phase rotation with B2B calibration method for fine-grained motion detection analysis on commodity Wi-Fi devices," *Sensors (Basel)*, vol. 18, no. 11, p. 3795, 2018.

[27] A. Borhani and M. Pätzold, "RF-based human activity recognition: A non-stationary channel model incorporating the impact of phase distortions," in *Proceedings of the International Work-Conference on Artificial Neural Networks*, Gran Canaria, Spain, 2019.

[28] Y. Chen, X. Su, Y. Hu, and B. Zeng, "Residual carrier frequency offset estimation and compensation for commodity WiFi," *IEEE Transactions on Mobile Computing*, vol. 19, no. 12, pp. 2891 – 2902, 2019.

[29] J. Pegoraro, F. Meneghello, and M. Rossi, "Multiperson continuous tracking and identification from mm-Wave micro-Doppler signatures," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2020.

[30] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, "IndoTrack: Device-free indoor human tracking with commodity Wi-Fi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–22, 2017.

[31] K. Qian, C. Wu, Z. Zhou, Y. Zheng, Z. Yang, and Y. Liu, "Inferring motion direction using commodity Wi-Fi for interactive exergames," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado, USA, 2017.

[32] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, "Widar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi," in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Chennai, India, 2017.

[33] W. Li, R. J. Piechocki, K. Woodbridge, C. Tang, and K. Chetty, "Passive WiFi radar for human sensing using a stand-alone access point," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 1986–1998, 2021.

[34] S. Li, X. Li, Q. Lv, G. Tian, and D. Zhang, "WiFit: Ubiquitous bodyweight exercise monitoring with commodity Wi-Fi device," in *Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, Guangzhou, China, 2018.

[35] M. Muaaz, A. Chelli, A. A. Abdelgawwad, A. C. Mallofré, and M. Pätzold, "WiWeHAR: Multimodal human activity recognition using Wi-Fi and wearable sensing modalities," *IEEE Access*, vol. 8, pp. 164 453–164 470, 2020.

[36] M. Schulz, D. Wegemer, and M. Hollick. (2017) Nexmon: The C-based firmware patching framework. [Online]. Available: https://nexmon.org

[37] H. Zhu, Y. Zhuo, Q. Liu, and S. Chang, "π-splicer: Perceiving accurate CSI phases with commodity WiFi devices," *IEEE Transactions on Mobile Computing*, vol. 17, no. 9, pp. 2155–2165, 2018.

[38] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, "OSQP: an operator splitting solver for quadratic programs," *Mathematical Programming Computation*, vol. 12, no. 4, pp. 637–672, 2020.

[39] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1602.07261

[40] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. MIT press Cambridge, 2016, vol. 1.

[41] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," in *Proceedings of the International Conference on Learning Representations*, 2017.

[42] L. Yang, Y. Han, X. Chen, S. Song, J. Dai, and G. Huang, "Resolution adaptive networks for efficient inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[43] H. Wang, A. Kembhavi, A. Farhadi, A. L. Yuille, and M. Rastegari, "Elastic: Improving CNNs with dynamic scaling policies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2258–2267.

[44] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 41, no. 1, p. 53, 2011.

[45] Y. Xie, Z. Li, and M. Li, "Precise power delay profiling with commodity WiFi," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2015.

[46] M. Schulz, D. Wegemer, and M. Hollick, "NexMon: A cookbook for firmware modifications on smartphones to enable monitor mode," Dec. 2015, arXiv:1601.07077.

[47] M. Schulz, J. Link, F. Gringoli, and M. Hollick, "Shadow Wi-Fi: Teaching smartphones to transmit raw signals and to extract channel state information to implement practical covert channels over Wi-Fi," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, New York, NY, USA, 2018.

[48] F. Gringoli, M. Schulz, J. Link, and M. Hollick, "Free your CSI: A channel state information extraction platform for modern Wi-Fi chipsets," in *Proceedings of the 13th International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization (WiNTECH)*, Los Cabos, Mexico, Oct. 2019.

[49] A. Zubow, P. Gawlowicz, and F. Dressler, "On phase offsets of 802.11ac commodity WiFi," *arXiv preprint arXiv:2005.03755*, 2020.

[50] D. Zhang, Y. Hu, Y. Chen, and B. Zeng, "Calibrating phase offsets for commodity WiFi," *IEEE Systems Journal*, vol. 14, no. 1, pp. 661–664, 2020.