This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3408041

Attention-Refined Unrolling for Sparse Sequential micro-Doppler Reconstruction

Riccardo Mazzieri[‡*], *Graduate Student Member, IEEE,* Jacopo Pegoraro[‡], *Member, IEEE,* and Michele Rossi[†‡], *Senior Member, IEEE*

***Abstract*— The reconstruction of micro-Doppler signatures of human movements is a key enabler for fine-grained activity recognition wireless sensing. In Joint Communication and Sensing (JCS) systems, unlike in dedicated radar sensing systems, a suitable trade-off between sensing accuracy and communication overhead has to be attained. It follows that the micro-Doppler has to be reconstructed from *incomplete* windows of channel estimates obtained from communication packets.**

Existing approaches exploit compressed sensing, but produce very poor reconstructions when only a few channel measurements are available, which is often the case with real communication patterns. In addition, the large number of iterations they need to converge hinders their use in real-time systems.

In this work, we propose and validate STAR, a neural network that reconstructs micro-Doppler sequences of human movement even from highly incomplete channel measurements. STAR is based upon a new architectural design that combines a *single* unrolled iterative hard-thresholding layer with an attention mechanism, applied at its output. This results in an interpretable and lightweight architecture that reaps the benefits of both model-based and data driven solutions.

STAR is evaluated on a public JCS dataset of 60 GHz channel measurements of human activity traces. Experimental results show that it substantially outperforms state-of-the-art techniques in terms of the reconstructed micro-Doppler quality. Remarkably, STAR enables human activity recognition with satisfactory accuracy even with 90% of missing channel measurements, for which existing techniques fail.

***Index Terms*—Joint Communication and Sensing, Micro-Doppler signatures, Sparse Reconstruction, Algorithm Unrolling, Attention, gHuman Activity Recognition.**

I. Introduction

NEXT-generation wireless networks are expected to gain the capability of sensing their surroundings via Radio Frequency (RF) signals, in addition to their primary communication functionality [1]. The vast number of applications of such context-aware networks spans domains like remote healthcare [2], safety [3], vehicle and crowd monitoring [4], and touchless human-computer interaction [5], which all require real-time processing of a huge amount of raw sensing data. Moreover, advanced JCS systems that analyze the movement of complex targets (e.g., humans) often involve computation-heavy Deep Learning (DL) architectures [6], [7]. However, while research on JCS is rapidly growing, there is an increasing concern that endowing communication systems with radar-like capabilities is bound to increase the network overhead and the channel occupation in time and frequency [8].

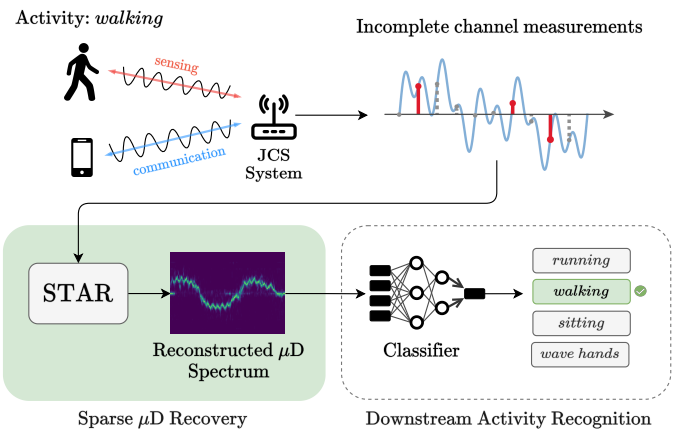

Fig. 1: Processing chain for human activity recognition: our model (STAR) is capable of recovering the micro-Doppler (μD) spectrum from very few CIR measurements.

In this work, we tackle the design of a *lightweight* and *ultra-low overhead* JCS method for human movement analysis based on the *sparse* reconstruction of the subject's μD signature. μD refers to the frequency modulation of the reflected radio signal caused by the motion of multiple target parts [9]. μD is a widespread method for complex target recognition and motion analysis in radar [10] and, more recently, in JCS [11], [12]. Its accurate computation usually requires regular and dense channel estimates, for the extraction of the Doppler spectrum. The main challenge of obtaining μD signatures in JCS is that the overhead introduced by frequent channel estimations is excessive. Therefore, the Doppler estimation must be carried out relying on few and irregular channel estimates obtained from the communication packets that are naturally exchanged by the network terminals, which prevents the use of standard Time-Frequency (TF) analysis methods.

To the best of our knowledge, only our previous work [12] has addressed this problem, devising a sparse reconstruction technique for the μD based on the Iterative Hard-Thresholding (IHT) compressed sensing algorithm [13]. However, this former approach has two main limitations: (i) when faced with high amounts of missing measurements, e.g., 90%, it produces low-quality reconstructions of the μD that can lead to movement recognition errors, and (ii) it takes many iterations

[‡]These authors are with the Department of Information Engineering at the University of Padova. [†]These authors are with the Department of Mathematics "Tullio Levi-Civita" at the University of Padova. *Corresponding author email: riccardo.mazzieri@phd.unipd.it.

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE0000001 - program "RESTART").

to converge, which translate into a high computation cost and large processing delays. Note that both (i) and (ii) go against the critical requirements of an efficient JCS system, which should operate with low overhead, performing very few channel measurements (i.e., avoiding the transmission of dummy packets for the sole purpose of sensing), and should guarantee fast reconstruction of the μD, enabling real-time sensing applications.

Additionally, classical compressed sensing approaches are not specifically tailored for μD spectrogram reconstruction and do not account for the temporal correlations of this type of data. Indeed, μD spectrograms exhibit specific temporal features, which may vary across different application domains. For example, in the case of human sensing, μD spectrograms are characterized by specific energy patterns, which might heavily differ from those of drones, cars, or other target types. Therefore, the key technical challenge tackled by this work is to design a system capable of improving upon classical compressed sensing methodologies, by leveraging domain-specific features of the signals to improve the quality of the reconstructions. To do so, while meeting the requirements of a low-overhead JCS system, we propose and validate Single Thresholding with Attention Refinement (STAR). STAR is an interpretable Neural Network (NN) architecture that accurately reconstructs μD signatures from highly incomplete time-domain channel measurements, with a computational complexity comparable to that of *a single* IHT iteration.

The considered processing chain for human activity recognition via wireless sensing is illustrated in Fig. 1: A wireless device samples the channel at irregular intervals, i.e., when data packets are transmitted over the wireless medium. This leads to the collection of incomplete vectors of channel measurements. Those are then translated by STAR into the corresponding μD spectrum, which succeeds in this task even in the presence of a high percentage of missing measurements. A classifier is finally used to assess the user activity.

STAR solves the drawbacks of prior techniques [12] by effectively combining: (1) a model-based learning block, that obtains a candidate reconstruction by *unrolling* a single IHT iteration into a NN layer, (2) a dot product attention layer with no learnable parameters, which exploits the sequential nature of the μD to learn its (temporal) correlation properties and to provide context features, and (3) a solution refinement block, which improves the unrolled solution from step (1) using the context information from step (2), in an interpretable fashion, producing the final (refined) μD spectrum at its output.

A thorough evaluation of the proposed architecture on experimental data is carried out, using the publicly available DISC dataset containing IEEE 802.11ay Channel Impulse Response (CIR) measurements at 60 GHz [14]. We focus on the task of reconstructing μD signatures of human movement for different activities, computing the reconstructed μD error with respect to the ground truth, and the resulting activity classification accuracy of a NN classifier. The considered activities are walking, running, waving hands, and sitting down/standing up [14]. STAR shows superior performance to both the original IHT and state-of-the-art NN models from the literature, providing accurate reconstructions even when only 10% of the input channel measurements are available.

STAR paves the way for the utilization of lightweight interpretable NN models to drastically lower the overhead of μD reconstruction in JCS. Note that human activity recognition is just one of the possible applications of STAR, which may be trained on any application involving μD recognition, e.g., drone and vehicle classification [10], [15] and gait analysis [16], among others.

The main contributions of this work are:

1) We develop and validate STAR, the first NN model to reconstruct μD spectrograms from (very) few time-domain estimates of the JCS channel.
2) STAR features a new way of exploiting the sequential nature of the μD, by refining an unrolled IHT solution with context information extracted through an attention mechanism. The resulting μD spectrum quality is excellent, even with over 90% missing channel measurements.
3) STAR is lightweight, having the computational complexity of *just one* IHT iteration. This makes it amenable to real-time operation in JCS systems.
4) We test STAR on a publicly available JCS dataset [14], on the μD of human activities. When the μD signatures reconstructed from 90%-incomplete measurements are used for the classification of challenging activities, existing approaches completely fail, yielding F1-scores that approach zero. Conversely, STAR provides F1-scores in the range $0.5 - 0.8$, showing a huge performance gain.

The paper is organized as follows. In Section II we discuss the related work, while Section III introduces the necessary background on compressed sensing and deep unrolling. Section IV presents our CIR model, focusing on the JCS aspect and introducing sparse μD reconstruction. STAR is presented in Section V, along with a detailed explanation of each processing block. In Section VI, STAR is evaluated on a publicly available experimental dataset, showing its superior performance with respect to state-of-the-art solutions. Concluding remarks are provided in Section VII.

II. Related Work

A. micro-Doppler spectrogram applications

μD analysis was originally introduced in the radar signal processing field [9], [17], as an enabler for advanced target recognition and motion estimation applications [6], [10], [15]. In the past few years, the high sensitivity of Millimeter-Waves (mmWaves) to μD shifts, together with DL methods for spectrogram analysis and classification, has led to the successful utilization of μD analysis for human activity recognition [18], [19], person identification [20], [21] and bio-mechanical gait analysis [16]. Due to its wide applicability to unobtrusive human motion analysis, μD processing has been identified as a prominent technology in remote healthcare and continuous monitoring of hospitals, homes, and public spaces.

Although we use a JCS human activity recognition dataset to evaluate STAR, our contribution is completely agnostic to the specific application scenario. As shown in Fig. 1, STAR enables μD reconstruction in cases where very few, irregular channel estimates are available. As such, it is an algorithm

to *obtain* the μD, rather than to analyze it, thus it applies to any type of target and motion after an appropriate training process.

B. STFT-based micro-Doppler extraction.

The standard way of obtaining the μD signature of a target is to perform Short Time Fourier Transform (STFT) on an estimate of the propagation channel [6], [15], [22]. In the context of JCS, [11] has shown that similar processing can be performed using standard-compliant IEEE 802.11ay channels estimates, achieving comparable performance to radar devices in terms of activity recognition and person identification.

The drawback of these previous works is the need for *regular* and *dense* transmission of probing signals, to retrieve channel estimates with the required fine-grained Doppler resolution. This has a twofold negative impact on JCS systems: (i) it requires a considerable amount of power, time and frequency resources, and (ii) it causes a significant overhead to the communication process, as dedicated sensing packets or waveforms have to be transmitted at the required rate, even though no data packets need to be transmitted by the users/applications. In the present work, instead, we focus on the less explored scenario where the above techniques fail because the channel estimates are obtained from incomplete sampling windows. Indeed, directly applying STFT on irregularly sampled data causes significant artifacts in the resulting spectrum. This poses the challenging research problem we tackle in this paper, involving the design of reconstruction algorithms that can recover the spectrum from a very few and irregularly sampled measurements. We do so by exploiting the channel sparsity in the frequency domain and the temporal correlation of μD spectra, stemming from the continuity of the underlying target movement.

C. Sparse micro-Doppler reconstruction.

In the radar context, sparse μD reconstruction with compressed sensing has been addressed in [23]–[26]. In the JCS literature instead, most works have focused on sparse sensing parameters estimation [27]–[29] but none dealt with the challenging problem of μD reconstruction, where all the Doppler frequencies of the different body parts have to be retrieved. To the best of our knowledge, only our previous work in [12] has directly tackled sparse μD reconstruction from IEEE 802.11ay channel estimates collected using irregular Wi-Fi traffic patterns. In the present work, we tackle this latter scenario due to its relevance to JCS, where the irregularity or the channel measurements is a direct consequence of the underlying communication traffic patterns. In addition, all the radar-based methods in [23]–[26] and [12] adopt standard iterative compressed sensing strategies, which suffer from two main limitations: (i) the reconstruction accuracy is significantly degraded when the number of available measurements becomes very low, and (ii) they are iterative methods requiring several iterations to converge. This leads to a consequent increase in processing delay and energy consumption, which hinders their real-time use in practical systems. As a solution to the latter issues, STAR enhances compressed sensing with data-driven feature learning, which exploits the sequential nature of μD to maintain a high reconstruction quality even when just a few measurements are available. This aspect is critical for JCS, since communication traffic may present long idle periods and every additional transmission causes possible interference and overhead. Moreover, STAR combines the concepts of attention [30] and deep algorithm unrolling [31] to reduce the number of iterations to just a single one. This drastic reduction in computational complexity makes our approach much more suited to real-time or near-real-time sensing applications.

D. Deep algorithm unrolling for sequential sparse recovery.

Deep *unrolling* (or *unfolding*) is a framework for enhancing traditional model-based iterative optimization algorithms with DL. Each iteration of the algorithm is implemented by a neural network layer, whose parameters are learned from data through backpropagation [31]. While most of the related research has focused on providing unrolled versions of popular iterative optimization algorithms, a few works have tackled the extension of deep unrolling to sparse temporal data [32]–[34]. These works dealt with the video recovery problems, modeling it as either an ℓ_1-ℓ_1 or an ℓ_1-ℓ_2 sparse reconstruction. The solver algorithm is unrolled using a Recurrent Neural Network (RNN), [32], [33], or a Transformer network, [34]. In previous unrolling formulations, the temporal correlation structure of the data was exploited to *initialize* the sparse recovery solution at the algorithm onset, in [32], [33], or at the beginning of every iteration, in [34], in an attempt of improving the recovery performance. In this work, we show that these techniques fail when the μD has to be reconstructed from very few measurements. In fact, initializing the IHT solution using past information simply provides an initial estimate of the support of the solution, which does not guarantee a good reconstruction. Conversely, STAR makes full use of the signal temporal correlation to *directly* suppress or enhance spectral components at the *output* of the unrolled IHT, thus strengthening its reconstruction capability even when very few measurements are available. This grants it superior performance with respect to state-of-the-art approaches, as demonstrated by our experimental results in Section VI.

E. Deep learning for sinusoid estimation.

μD reconstruction is a sequential spectral estimation problem on time-varying sinusoidal signals. A few works have used deep learning methods to estimate the sinusoid parameters in noisy signals [35], [36], and in low-resolution quantized signals [37]. These approaches rely on standard, *blackbox* deep neural networks with little interpretability. STAR instead combines an unrolled version of IHT, whose processing steps have a well-known interpretation of thresholded fixed point iterations [38], and an attention layer that finds temporal correlations among μD spectra. This makes STAR easily interpretable using signal processing domain knowledge. Moreover, existing approaches significantly differ from our proposed technique in that: (i) they only work for uniformly sampled signals, i.e., they can not handle incomplete sampling patterns

and hence they are not suited for JCS; (ii) they are developed for single-shot sinusoidal spectral estimation, thus they can not exploit the sequentiality of μD data. Conversely, STAR is specifically designed to deal with incomplete sampling of the CIR, and it does so by jointly mimicking the structure of robust sparse reconstruction algorithms and exploiting the powerful feature extraction capabilities of attention mechanisms. This endows it with enhanced robustness to severe sparsity levels and a much-improved convergence speed.

III. Background

Next, we provide an overview of compressed sensing methods for sparse reconstruction and deep algorithm unrolling.

A. Notation

A continuous-time signal s is denoted by $s(\cdot)$, whereas square brackets are used for discrete-time signals, e.g., $s[\cdot]$. $\Re(x)$ and $\Im(x)$ denote the real and imaginary part of $x \in \mathbb{C}$, while $|x|$ is its magnitude. $\mathbf{X}^T$, $\mathbf{X}^H$, and $\mathbf{X}^*$ denote the transpose, Hermitian, and complex conjugate of matrix $\mathbf{X}$. $||\mathbf{x}||_p$ refers to the ℓ_p-norm of vector $\mathbf{x}$, with $p = 0, 1, 2$. The *soft-thresholding* and *hard-thresholding* operators applied to vector $\mathbf{x}$ are denoted by $\mathcal{S}_\omega(\mathbf{x})$ and $\mathcal{H}_\omega(\mathbf{x})$, respectively. Soft-thresholding is defined as $\mathcal{S}_\omega(\mathbf{x}) = \text{sign}(\mathbf{x}) \cdot \max(|\mathbf{x}| - \omega, 0)$, where operations on vector $\mathbf{x}$ are applied elementwise. $\mathcal{H}_\omega(\mathbf{x})$ sets to 0 all the components of $\mathbf{x}$ but the ω largest ones.

When operating with NNs, complex-valued vectors $\mathbf{x} \in \mathbb{C}^N$ are transformed into real-valued vectors using the mapping $\mathbf{x}' = \mathcal{R}(\mathbf{x}) = [\Re(\mathbf{x}) \ -\Im(\mathbf{x})]^T \in \mathbb{R}^{2N}$. When applied to matrices, the transformation computes $\mathbf{X}' \in \mathbb{R}^{2M \times 2N}$ as

$$\mathbf{X}' = \mathcal{R}(\mathbf{X}) = \begin{bmatrix} \Re(\mathbf{X}) & -\Im(\mathbf{X}) \\ \Im(\mathbf{X}) & \Re(\mathbf{X}) \end{bmatrix}. \tag{1}$$

We denote by $\mathbf{F}_N$ the inverse N-point Fourier matrix, whose elements are $F_{n,m} = (1/\sqrt{N}) \exp(j2\pi nm/N)$, $n, m = 0, \ldots, N-1$. $\mathbf{I}_N$ is the N-dimensional identity matrix. Finally, $\mathcal{U}([a, b])$ denotes the continuous uniform distribution in the interval $[a, b]$.

B. Compressed sensing primer

Compressed Sensing (CS) provides a framework to solve *underdetermined* linear systems of the form

$$\mathbf{x} = \mathbf{\Phi}\mathbf{z} + \mathbf{n}, \tag{2}$$

where $\mathbf{x} \in \mathbb{C}^M$ is a vector of noisy measurements, $\mathbf{z} \in \mathbb{C}^K$ is an unknown signal vector, to be reconstructed, $\mathbf{\Phi} \in \mathbb{C}^{M \times K}$ is the sensing matrix, and $\mathbf{n} \in \mathbb{C}^M$ is a noise vector. Under the assumption that the signal vector is *sparse*, meaning that $||\mathbf{z}||_0 \ll K$, i.e., it has only a few non-zero components, the core CS result is that $\mathbf{z}$ can be reconstructed *exactly* even when $M < K$ [13], [38]. This is subject to the requirement of having a sufficient number of measurements, that scales as the logarithm of K. The reconstruction is performed by solving an optimization problem of the form

$$\arg\min_{\mathbf{z}} \left[f(\mathbf{\Phi}\mathbf{z}, \mathbf{x}) + g(\mathbf{z}) \right], \tag{3}$$

where f is a measure of the reconstruction error, e.g., the ℓ_2-norm $||\mathbf{\Phi}\mathbf{z} - \mathbf{x}||_2$, and $g(\mathbf{z})$ is a regularization term that enforces the sparsity of the solution, e.g., $||\mathbf{z}||_0$ or $\lambda||\mathbf{z}||_1$, where $\lambda > 0$ is used to tune the importance of the regularization. Popular, fast iterative algorithms to solve Eq. (3) are: (i) the Iterative Shrinkage-Thresholding Algorithm (ISTA), that uses $g(\mathbf{z}) = \lambda||\mathbf{z}||_1$, (ii) IHT, that uses $g(\mathbf{z}) = ||\mathbf{z}||_0$, and (iii) Orthogonal Matching Pursuit (OMP), which also uses the ℓ_0 regularizer [13]. ISTA and IHT belong to the category of *iterative thresholding* methods, whose $(i+1)$-th iteration computes

$$\mathbf{z}^{(i+1)} \leftarrow \mathcal{T}\left(\frac{1}{\mu}\mathbf{\Phi}^H\mathbf{x} + \left(\mathbf{I} - \frac{1}{\mu}\mathbf{\Phi}^H\mathbf{\Phi} \right) \mathbf{z}^{(i)} \right), \tag{4}$$

where μ is the inverse of the learning step size. $\mathcal{T}$ is a suitable (algorithmic-dependent) thresholding operator. ISTA uses soft-thresholding, $\mathcal{T}(\mathbf{z}) = \mathcal{S}_{\frac{\lambda}{\mu}}(\mathbf{z})$, while IHT uses hard-thresholding, $\mathcal{T}(\mathbf{z}) = \mathcal{H}_\Omega(\mathbf{z})$, with Ω being a pre-defined sparsity level. Typical stopping criteria for Eq. (4) involve setting a maximum number of iterations or stopping the algorithm upon convergence of the difference $||\mathbf{z}^{(i+1)} - \mathbf{z}^{(i)}||_2$.

C. Sequential compressed sensing

CS recovery has been extended to the case where the underlying signal is time-varying, and the processing relies upon a moving window approach [39]. Denote by $\mathbf{x}[t]$ and $\mathbf{z}[t]$ the time-dependent vectors of measurements and unknown signal, respectively. In the sequential formulation, the signal vectors in previous windows, $\mathbf{z}[t-1], \ldots, \mathbf{z}[0]$, are used as *side information* on the locations of the non-zero entries of $\mathbf{z}[t]$. Eq. (3) can be adapted to the sequential setting as follows [39]

$$\arg\min_{\mathbf{z}[t]} \left[f(\mathbf{\Phi}_t\mathbf{z}[t], \mathbf{x}[t]) + d(\mathbf{z}[t], \ldots, \mathbf{z}[0]) + g(\mathbf{z}[t]) \right], \tag{5}$$

where $\mathbf{\Phi}_t$ is the sensing matrix in the t-th processing window, and function $d(\cdot)$ serves to regularize the optimization using past information. Typically, $d(\cdot)$ is simplified to only capture the one-step correlation between $\mathbf{z}[t]$ and $\mathbf{z}[t-1]$, neglecting the potential dependency of $\mathbf{z}[t]$ on the full μD sequence. Previous works have then used ℓ_1 or ℓ_2 norm-based formulations for $d(\cdot)$ to aid the reconstruction of $\mathbf{z}[t]$, forcing it to share a similar support to $\mathbf{z}[t-1]$ [34], [39]. Using the ℓ_1-norm for $d(\cdot)$ and $g(\cdot)$, and the ℓ_2-norm for $f(\cdot)$, the ℓ_1-ℓ_1 CS problem is obtained [39], i.e.,

$$\arg\min_{\mathbf{z}[t]} \left[||\mathbf{\Phi}_t\mathbf{z}[t] - \mathbf{x}[t]||_2^2 + \eta||\mathbf{z}[t] - \mathbf{D}\mathbf{z}[t-1]||_1 + \lambda||\mathbf{z}[t]||_1 \right], \tag{6}$$

where $\eta > 0$ weighs the importance of the past information, and matrix $\mathbf{D}$ represents a linear evolution model for the reconstructed signal across subsequent processing windows. The motivation behind modeling $d(\cdot)$ as in Eq. (6) stems from a Laplacian approximation of the distribution of the error $\mathbf{z}[t] - \mathbf{D}\mathbf{z}[t-1]$, which holds well in video processing and other domains. Eq. (6) is typically solved using the algorithm in [39] or an unrolled proximal gradient method [33] (see Section III-D below).

D. Deep unrolling for compressed sensing

Although iterative algorithms have found widespread application thanks to their accuracy and ease of implementation, in practical settings they may require a large number of iterations to converge [31]. The idea behind algorithm unrolling is to construct a NN architecture in which each layer corresponds to one iteration of Eq. (4). This is done by exploiting the structural similarity between Eq. (4) and a recurrent NN layer with state vector $\mathbf{z}^{(i)}$ and a fixed input $\mathbf{x}$, where the thresholding operator plays the role of a non-linear activation function [40]. In the unrolling implementation, Eq. (4) is usually rewritten by using the complex-to-real transformation, $\mathcal{R}(\cdot)$, introduced in Section III-A. The input-output equation of an unrolled NN layer can be written as

$$\mathbf{z}^{(i+1)} = \mathcal{T}\left(\frac{1}{\mu}\mathbf{W}^T\mathbf{x} + \mathbf{S}\mathbf{z}^{(i)}\right), \tag{7}$$

where $\mathbf{W}$ is a set of learnable weights and $\mathbf{S}$ can be selected as in the original algorithm, i.e., $\mathbf{S} = \mathbf{I} - \frac{1}{\mu}\mathbf{W}^T\mathbf{W}$, or as an additional set of learnable weights [31]. The network is constructed by stacking a fixed number of layers of the type in Eq. (7), with weights $\mathbf{W}, \mathbf{S}$ being *shared* among the different layers. Training is then performed with standard backpropagation on a dataset of D input-output pairs, $\mathcal{X} = \{\mathbf{x}_m, \mathbf{z}_m^*\}_{m=1}^D$, where $\mathbf{z}_m^*$ is the solution obtained with the original iterative algorithm with input $\mathbf{x}_m$, which can be pre-computed. In the literature, unrolled versions of several popular algorithms have been proposed [41], [42]. The unrolled ISTA and IHT algorithms are commonly referred to as Learned Iterative Shrinkage-Thresholding Algorithm (LISTA) and Learned Iterative Hard-Thresholding (LIHT) [43], [44].

Reference [33] has shown that deep unrolling is well suited to address sequential CS problems tackled via, e.g., $\ell_1 - \ell_1$ minimization. There, an RNN is presented that unrolls a proximal gradient method solving Eq. (6). The key difference between this approach and STAR lies in the way in which the prior information, represented by past solutions $\mathbf{z}[t-1], \ldots, \mathbf{z}[0]$, is exploited. In detail, instead of specifying a model for $d(\cdot)$, we let the NN learn it directly from data through an attention layer. This better exploits the previous reconstructions, which ultimately leads to enhanced results with very few available measurements.

IV. Channel model and micro-Doppler

In JCS systems, CIR estimates are reused (besides using them to decode data) to obtain information about the sensing targets. In the interest of better conveying our framework, we keep the experimental setup simple, by assuming to be in a *monostatic* scenario, i.e., with co-located transmitter and receiver. We underline that this is not a restrictive assumption for STAR. In this setup, the channel can be estimated whenever the reflections of transmitted packets are collected back at the transmitter. Therefore, CIR estimates are obtained at irregular (and random) time instants which depend on the transmission pattern at the transmitter side, and coincide with the reception of packets. As shown in [12], CIR can sampled at *regular* intervals, obtaining a grid of channel samples evenly spaced by T_c seconds, where T_c is the (arbitrary) channel sampling period. However, in communication networks, it is very unlikely that data packets are available for transmission every T_c seconds (leading to channel estimates at this granularity), and this very much depends on the application data pattern. Thus, it follows that the resulting CIR measurements grid is likely to be incomplete, i.e., some of the measurements on the grid are missing. To compensate for this, one may think of transmitting dummy packets (with no data) to attain channel estimates at regular intervals, but this is undesirable as it implies a large communication overhead.

A. Channel impulse response model

In the following, we use a Single Carrier (SC) CIR model similar to that in [12], since the dataset used for our experimental validation is based on a SC IEEE 802.11ay JCS implementation [14]. However, the presented approach is equally applicable to Orthogonal Frequency Division Multiplexing (OFDM) waveforms.

For the channel model, we consider discrete timesteps $nT_c, n \in \mathbb{Z}^+$, with time granularity T_c. Although our system operates on incomplete windows of channel samples, the CIR is mathematically modeled as if all the samples were available. Shortly below, in Section IV-B, we show how the missing samples can be added to the formulation. At time nT_c, the CIR is a function of the propagation delay τ, expressed as a sum of L Dirac delta components which correspond to the *resolvable* signal propagation paths [45]. The finite delay resolution of the system is $\Delta\tau = 1/B$, where B is the transmitted signal bandwidth. Denoting by τ_l the propagation delay of the l-th scatterer, and by $h_l(nT_c)$ the l-th complex CIR component, we have

$$h(\tau, nT_c) = \sum_{l=1}^{L} h_l(nT_c)\delta(\tau - \tau_l(nT_c)). \tag{8}$$

The propagation delay of path l is associated with a specific distance from the JCS transceiver, according to the relation $d_\ell = c\tau_l/2$, where c is the speed of light. Hence, the signal bandwidth determines a minimum distance threshold, equal to $\Delta d = c\Delta\tau/2$, under which two targets produce reflections that overlap in the same CIR peak at the receiver. For complex targets with many moving parts, such as the human body, the system bandwidth in common communication systems (even in the mmWave range) is typically insufficient to entirely resolve the resulting reflections [12].

Denote by: $Q_l(nT_c)$ the number of reflections due to the unresolvable scatterers composing the sensing target, $f_{l,q}$ the Doppler shift of the q-th of such reflections and $\alpha_{l,q}$ a complex coefficient that accounts for the scatterer's Radar Cross-Section (RCS), the propagation loss, and the beamforming gains. The l-th CIR component at time nT_c is

$$h_l(nT_c) = \sum_{q=1}^{Q_l(nT_c)} \alpha_{l,q}(nT_c)e^{j2\pi f_{l,q}(nT_c)nT_c}. \tag{9}$$

The expression of the Doppler shift depends on the movement speed of the scatterer, denoted by $v_{l,q}$, on the angle of the

direction of motion with respect to the incident wave, $\theta_{l,q}$, and on the carrier frequency f_c, as

$$f_{l,q}(nT_c) = \frac{v_{l,q}(nT_c)\cos\theta_{l,q}(nT_c)}{c} f_c, \tag{10}$$

where c is the speed of light.

Finally, we make the two following assumptions to further simplify Eq. (9).

Assumption 1 - Window-based processing: The CIR samples are processed in time windows spanning K subsequent timesteps, where KT_c is sufficiently short, so that the parameters of the scatterers, $\alpha_{l,q}, f_{l,q}, Q_l$, can be considered constant within each window. Timesteps within each processing window are indexed by variable $k = 1, \ldots, K$. We allow instead the parameters to change across different processing windows (indexed by t). This assumption is ubiquitous in radar signal processing and JCS, and has been empirically verified for human sensing [4], [6], [45].

Assumption 2 - Tracking multipath reflections: We assume we can track the signal propagation paths corresponding to the sensing targets of interest, following the evolution of their delay parameter τ_l across time. Tracking τ_l amounts to tracking the distance of the target from the transmitter, termed d_l, across time, since it holds that $d_l = c\tau_l/2$. This is a standard processing step in wireless human sensing applications, which can be done using one of the many existing multiple target tracking techniques based on multiple Kalman filters [45], probabilistic data association filters [46], or multiple hypotheses tracking [47]. The performance of these algorithms depends on the system bandwidth, B, which determines the delay resolution, and on the extension of the target. Tracking of several human targets in challenging indoor scenarios has been successfully demonstrated in many works, e.g., [11], [48]. Tracking τ_l allows extracting the CIR component due to the sensing target, separating it from those of the other scatterers. As such, it is a necessary processing step for any wireless sensing system that obtains target-specific μD signatures: If no separation of the different target reflections were performed, the μD contributions of different targets would overlap in the Doppler domain, interfering with one another. It follows that Assumption 2 is not due to a limitation of the proposed system but to an intrinsic requirement for multitarget wireless sensing. Note that in single-target scenarios Assumption 2 becomes unnecessary and it can be removed.

Combining assumptions 1 and 2 allows rewriting Eq. (9) by replacing the time dependency of $\alpha_{l,q}, f_{l,q}, Q_l$ with a coarser dependency on the processing window, t, and dropping index l as we consider each target to be consistently tracked and separated from the others. Hence, denoting by $\alpha_q[t], f_q[t], Q[t]$ the sensing parameters of target q in the t-th window, Eq. (9) becomes

$$h[t,k] = \sum_{q=1}^{Q[t]} \alpha_q[t] e^{j2\pi f_q[t](tK+kT_c)}, \tag{11}$$

where we use the compact, discrete-time notation $h[t,k]$ to refer to the k-th CIR sample in the t-th processing window. Eq. (11) forms the basis for the formulation of the sparse sequential μD recovery problem, as it expresses each CIR component as a superposition of $Q[t]$ complex sinusoids with frequencies $f_q[t]$. Each sinusoid is associated with one of the scatterers overlapping in the same CIR component. The CIR can be considered sparse in the frequency domain if $Q[t] \ll K$. This follows from the fact that each complex exponential has a single active frequency, hence the Fourier transform of Eq. (11) has few non-zero elements. In [12], it has been shown that for targets involving few moving parts (such as the human body), channels observed by wideband, directional communication systems lead to few scattering components $Q[t]$, hence justifying the use of CS methods.

We stress that Eq. (11) refers to a single component in the CIR, but contains information about multiple unresolvable paths that all overlap in the same CIR peak. This demonstrates the importance of μD analysis in recognizing the movement of complex targets, as it allows resolving multiple moving parts in the Doppler domain, leveraging their different moving speeds. Moreover, the evolution of the sensing parameters, $\alpha_q[t], f_q[t], Q[t]$, across different windows is *correlated*. This stems from the complex dynamics of the underlying movement of the target and how these affect the channel depending on the observation angle, distance, and propagation characteristics. As a result, such sequential correlation is extremely challenging to model, as further discussed in Section IV-B2.

B. Sparse sequential micro-Doppler reconstruction

The μD signature of a target is typically obtained by using the STFT [6], [9], [11], which applies a Discrete Fourier Transform (DFT) to each window of $h[t,k]$. The resulting spectrum peaks at values $f_q[t] = v_q[t]\cos\theta_q[t] f_c/c$. Hence, spectral analysis of the CIR reveals important features of the underlying physical movement of the target, especially when considering its temporal evolution across subsequent windows. The frequency resolution (Δf) and the maximum resolvable frequency (f_m) of the STFT depend on T_c as $\Delta f = 1/(KT_c)$, and $f_\mathrm{m} = 1/(2T_c)$, respectively. These can then be mapped onto the corresponding velocity of the scatterer as $\Delta v = c/(2f_cKT_c)$ and $v_\mathrm{m} = c/(4f_cT_c)$. When CIR measurements are incomplete, STFT is not directly applicable as the resulting spectrum would be degraded by the lack of a fixed sampling interval. Hence, alternative approaches have to be sought relying on *sparse reconstruction* [12].

1) Sparse μD reconstruction: Due to the missing CIR values, out of K timesteps in window t we only have M_t available samples, whose indices in the window are denoted by $m_1, \ldots, m_{M_t}$. We define vector $\mathbf{h}[t] = [h[t,m_1], \ldots, h[t,m_{M_t}]]^T$, containing the available CIR samples in the window, and matrix $\mathbf{M}_t = \left[\mathbf{e}_{m_1}, \ldots, \mathbf{e}_{m_{M_t}}\right]^T$, where $\mathbf{e}_i$ is the i-th vector of the canonical basis. Left multiplication of a matrix by $\mathbf{M}_t$ has the effect of selecting the rows of such matrix whose indices correspond to the available samples. Moreover, denote by $\mathbf{z}[t] \in \mathbb{C}^K$ the DFT of the (unknown) complete window of CIR samples $h[tK], \ldots, h[(t+1)K-1]$. The following compressed sensing model can be formulated relating $\mathbf{h}[t]$ and $\mathbf{z}[t]$,

$$\mathbf{h}[t] = \mathbf{M}_t\mathbf{F}_K\mathbf{z}[t] + \mathbf{n}, \tag{12}$$

where $\mathbf{n}$ is a M_t-dimensional complex noise vector, and $\mathbf{F}_K$ is the inverse Fourier matrix defined in Section III-A. By setting $\mathbf{\Phi}_t = \mathbf{M}_t\mathbf{F}_K$, Eq. (12) is in the form expressed by Eq. (2), hence we can tackle it by solving the optimization problem in Eq. (3). Our aim is to recover $\mathbf{z}[t]$ from the incomplete measurement vector $\mathbf{h}[t]$, so that we can obtain the μD spectrum of the t-th CIR window as $\mathbf{y}[t] = |\mathbf{z}[t]|^2$. To this end, we solve the following compressed sensing problem, which finds a vector $\mathbf{z}[t]$ which is a solution to Eq. (12), while being at least Ω-sparse

$$\underset{\mathbf{z}[t]}{\arg\min} ||\mathbf{h}[t] - \mathbf{\Phi}_t\mathbf{z}[t]||_2^2 \text{ s.t. } ||\mathbf{z}[t]||_0 \leq \Omega. \tag{13}$$

The constant $\Omega \in \mathbb{N}$ is a pre-defined sparsity level that is closely related to the value of Q in Eq. (11). We set Ω as an upper bound to Q, as the latter is unknown in practice. In this way, the sparse reconstruction retrieves a solution that is Ω-sparse in the frequency domain, ensuring that all the Q frequency components of the CIR can be reconstructed. To solve Eq. (13), previous work has adopted IHT [12]. However, this solution has significant limitations in terms of computational speed, as it involves an iterative process that may take several iterations to converge, and suffers from low reconstruction quality when very few measurements per window are available [12].

2) Sequential μD reconstruction: Solving Eq. (13) independently for each processing window does not fully exploit the temporal features of the μD spectrogram. Indeed, the evolution of the μD in time exhibits particular patterns that depend on the type of target and movement. In our model, this can be expressed as the availability of *side information* on $\mathbf{z}[t]$, which depends on the previous μD spectrum windows. This can be modeled as a sequential CS problem, as discussed in Section III-C. Eq. (13) is modified to account for the correlation among μD windows as follows [39]

$$\begin{aligned} &\underset{\mathbf{z}[t]}{\arg\min} \left[||\mathbf{h}[t] - \mathbf{\Phi}_t\mathbf{z}[t]||_2^2 + d(\mathbf{z}[t], \ldots, \mathbf{z}[0])\right] \\ &\text{s.t. } ||\mathbf{z}[t]||_0 \leq \Omega. \end{aligned} \tag{14}$$

Using the formulation in Eq. (6) to solve the above problem fails to capture complex and long-term correlations among the sequence of μD spectra. Moreover, as we discuss in Section VI-C2, previous approaches that adopted unrolling to address sequential CS did not fully exploit the information from previous reconstructions, as such information is used to obtain a better *initialization* for the reconstruction algorithm (at the input of the IHT processing block).

In the next section, we describe our approach, which is significantly less computationally complex than IHT and fully exploits the μD temporal correlation structure. To reduce the computational complexity of standard IHT, we unroll a *single iteration* of IHT into a neural network. Then, we enhance the resulting solution by learning sequential features of the μD using an attention mechanism, which acts as a refinement step applied to the output of the IHT block. This avoids having to specify a model for the function $d(\cdot)$ in Eq. (14) and rather lets the neural network learn μD correlations directly from data.

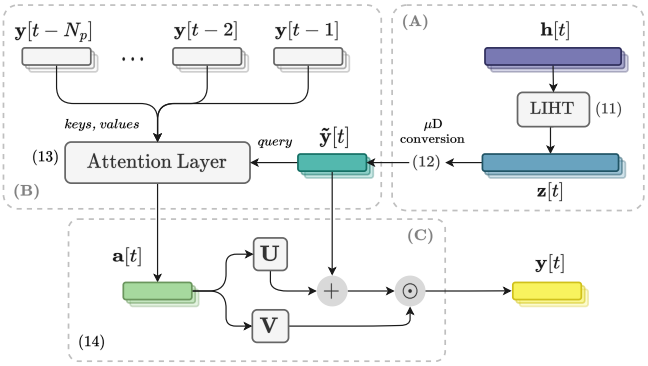

Fig. 2: Block diagram of STAR.

V. The STAR architecture

The block diagram of STAR is shown in Fig. 2, while the step-by-step computations are reported in Alg. 1. STAR processes the current input measurements by unrolling the IHT algorithm, and subsequently improves the resulting reconstruction by leveraging the sequential structure of the μD spectrogram. The computations performed by STAR can be subdivided into three distinct blocks:

(A) Single-layer LIHT: With the first block, we retrieve an approximate solution for the reconstruction of the micro-Doppler window at time t (of size K), denoted by $\tilde{\mathbf{y}}[t]$, with minimal computations. To do so, the incomplete CIR input window is processed by a single LIHT layer, which unrolls one IHT iteration, as discussed in-depth in Section V-A. The LIHT output is then refined in steps (2) and (3).

(B) Attention mechanism: In this block, temporal features of the μD are exploited to provide a *context* for the approximate solution found in (1). μD signatures of human movements exhibit a strong temporal correlation, which we take into account to refine the quality of the reconstruction. To this end, we introduce a scaled dot product attention mechanism. This computes the correlation of the current approximate reconstruction with past recovered μD spectra, to obtain a context feature vector $\mathbf{a}[t]$. Details regarding the attention mechanism are reported in Section V-B.

(C) Solution refinement: Finally, the outputs from the previous two blocks are combined to retrieve the final sparse μD spectrum. The solution refinement block learns a transformation of the current approximate solution, $\tilde{\mathbf{y}}[t]$, based on the context feature vector $\mathbf{a}[t]$. This is achieved by stacking two parallel feedforward NN layers (with weights $\mathbf{U}$ and $\mathbf{V}$) that are fed with $\mathbf{a}[t]$, and whose output is used to transform $\tilde{\mathbf{y}}[t]$ additively and multiplicatively. The resulting output is the current μD reconstruction. Block **(C)** is thoroughly discussed in Section V-C.

A. LIHT Module

The LIHT block operates on each incomplete CIR measurement window independently, as summarized in lines $3-5$ of Alg. 1. We use a single LIHT layer to obtain a baseline sparse reconstruction of $\mathbf{z}[t]$. Calling $\mathcal{H}_\Omega$ the hard-thresholding operator described in Section III-B, μ the reciprocal of

the learning step size, and initializing the reconstruction as $\mathbf{z}^{(0)}[t] = \mathcal{H}_\Omega\left(\frac{1}{\mu}\mathbf{W}^T\mathbf{h}[t]\right)$, the LIHT block computes

$$\mathbf{z}[t] = \mathcal{H}_\Omega\left(\left(\mathbf{I} - \frac{1}{\mu}\mathbf{W}^T\mathbf{W}\right)\mathbf{z}^{(0)}[t] + \frac{1}{\mu}\mathbf{W}^T\mathbf{h}[t]\right). \quad (15)$$

The specific choice of the initialization follows [43]. All vectors and matrices in Eq. (15) are subjected to the complex-to-real transformation defined in Section III-A. Therefore, we have $\mathbf{h}[t], \mathbf{z}[t] \in \mathbb{R}^{2K}$. Matrix $\mathbf{W}$ is initialized as $\mathbf{W} = \mathcal{R}(\mathbf{M}_t\mathbf{F}_K) \in \mathbb{R}^{2M_t \times 2K}$ and then learnt during training.

We remark that $\mathbf{z}[t]$ is a reconstruction of the DFT of $\mathbf{h}[t]$, not of its μD spectrum $\mathbf{y}[t] \in \mathbb{R}^K$. To obtain the latter, we compute

$$\tilde{\mathbf{y}}[t] = \begin{bmatrix}\mathbf{I}_K & \mathbf{I}_K\end{bmatrix}\mathbf{z}[t]^2, \quad (16)$$

where the square operation is applied elementwise to $\mathbf{z}[t] \in \mathbb{R}^{2K}$. In detail, with Eq. (16) we obtain the squared magnitude (μD spectrum) of the DFT of the t-th CIR window. The two identity matrices are required to combine the real and imaginary part of such (complex-valued) DFT, that respectively appear in the first and second K elements of vector $\mathbf{z}[t]$ (due to our definition of the transform $\mathcal{R}$).

B. Attention mechanism

This is a key (and novel) step of the proposed technique: As it will be discussed in Section VI-C2, positioning the attention layer as we did, i.e., towards the end of the processing chain, is non-standard and we found it to be highly effective. To exploit the strong temporal correlations in the μD spectrograms due to the sequentiality of human movement, STAR looks for correlations between $\tilde{\mathbf{y}}[t]$, and the reconstructions obtained at previous time steps, $\mathbf{y}[t-1], \mathbf{y}[t-2], \ldots, \mathbf{y}[t-N_p]$. N_p is the number of past μD windows considered in the computation of the context features, which is set as a hyperparameter of our model. The key idea behind this attention layer is to obtain context features to be used by block **(C)** to improve the current approximate solution $\tilde{\mathbf{y}}[t]$. Specifically, the attention layer learns the temporal correlation model that best represents the evolution of a spectrogram window.[1] Such model is then utilized to refine $\tilde{\mathbf{y}}[t]$, improving the reconstruction of those window elements that are heavily corrupted due to undersampling. Note that our approach does not explicitly require specifying function $d(\cdot)$ in Eq. (14), but rather uses an attention mechanism to learn complex sequential features of the μD. The computations performed by the attention mechanism are described next, and reported in lines $6-7$ of Alg. 1.

Following the original terminology introduced in [30], attention compares a set of *query* vectors to some *key* vectors, according to a specific alignment function, to obtain a set of *attention weights* for each query vector. Such weights are then mapped within $[0, 1]$ by a $\mathrm{Softmax}$ function and used to perform a weighted convex sum of some *value* vectors, which represents the final output of the attention layer. In STAR, we define matrix $\mathbf{Y}[t] = [\mathbf{y}[t-1], \ldots, \mathbf{y}[t-N_p]]^T$, which contains the reconstructions obtained at past time steps. The columns of $\mathbf{Y}[t]$ serve as both the keys and the values of our attention mechanism. Queries are instead represented by the current approximate reconstruction $\tilde{\mathbf{y}}$, which we want to contextualize with respect to the sequence of μD spectra. As the alignment function, we use the dot product, which measures the level of correlation between queries and the key while maintaining the whole layer extremely efficient and lightweight. The computations performed by the attention mechanism are summarized in the following equation

$$\begin{aligned}\mathbf{a}[t] &= \sum_{i=1}^{N_p} \frac{e^{\frac{1}{\sqrt{K}}\mathbf{y}[t-i]^T\tilde{\mathbf{y}}[t]}}{\sum_{u=1}^{N_p} e^{\frac{1}{\sqrt{K}}\mathbf{y}[t-u]^T\tilde{\mathbf{y}}[t]}}\mathbf{y}[t-i] \\ &= \mathbf{Y}[t]^T\mathrm{Softmax}\left(\frac{1}{\sqrt{K}}\mathbf{Y}[t]\tilde{\mathbf{y}}[t]\right). \quad (17)\end{aligned}$$

In Eq. (17), the correlations are computed directly in the frequency spectrum domain, without introducing any additional learnable parameter. This is in contrast with the typical approach used in DL models of projecting queries, keys, and values to high dimensional feature spaces before the computation of the attention weights [30]. For the considered human-sensing application, experimental results showed no performance improvement when learning such projections.

C. Solution refinement Module

In this final step, the initial reconstruction $\mathbf{z}[t]$ is refined using the context provided by the attention layer. The key idea behind this step is to leverage the typical structure of μD spectra, which show *localized* spectral components around the Doppler frequencies of the different body parts.

When applying the standard IHT with few input measurements we observe two main shortcomings (see Fig. 6 and Fig. 9): (i) the resulting reconstruction shows unstructured background noise with no physical meaning, and (ii) the frequency components due to the body parts are inaccurately reconstructed. Therefore, we apply two different kinds of operations to $\tilde{\mathbf{y}}[t]$, in order to mitigate these two issues separately:

1) First, an *additive* transformation is applied to $\tilde{\mathbf{y}}[t]$. This is done to give the model the ability to learn how to improve the solution in those frequency bins that have been poorly reconstructed. The additive term is parametrized as a dense neural network layer with a Rectified Linear Unit (ReLU) activation function, to enforce only positive-valued changes to the approximate known solution. The ReLU is defined as $\mathrm{ReLU}(x) = \max(0, x)$;
2) Next, STAR applies a *multiplicative* masking to the output of step 1). This step is designed to let the model learn how to denoise the reconstruction. The multiplicative term is parametrized by a dense neural network layer with a sigmoid activation function, denoted by $\sigma(\cdot)$, to constrain the output in the interval $[0, 1]$.

The computations performed by the refinement block are (line 8 in Alg. 1)

$$\mathbf{y}[t] = \left(\mathbf{z}[t] + \mathrm{ReLU}\left(\mathbf{U}\mathbf{a}[t] + \mathbf{b}\right)\right) \odot \sigma\left(\mathbf{V}\mathbf{a}[t]\right), \quad (18)$$

with $\mathbf{y}[t]$ being the final output μD spectrum.

[1] This model is strongly connected with the underlying physical movement that is being tracked, e.g., human-related.

Algorithm 1 Computational steps of STAR.

Input: Sequence of incomplete CIR windows $\mathbf{h}[t], t = 1, 2, \dots$

Output: μD spectrogram $\mathbf{y}[t], t = 1, 2, \dots$.

// Initialize learnable weights

$\mathbf{W} \leftarrow \mathbf{F}_K$; $\mathbf{U}, \mathbf{V}, \mathbf{b} \sim \mathcal{U}(\frac{1}{\sqrt{K}}, \frac{1}{\sqrt{K}})$

for $t = 1, 2, \dots$ **do**

// Single-layer LIHT

$\mathbf{z}^{(0)} \leftarrow \mathcal{H}_\Omega \left(\frac{1}{\mu} \mathbf{W}^T \mathbf{h}[t] \right)$

$\mathbf{z}[t] \leftarrow \mathcal{H}_\Omega \left\{ \left(\mathbf{I} - \frac{1}{\mu} \mathbf{W}^T \mathbf{W} \right) \mathbf{z}^{(0)}[t] + \frac{1}{\mu} \mathbf{W}^T \mathbf{h}[t] \right\}$

$\tilde{\mathbf{y}}[t] \leftarrow \begin{bmatrix} \mathbf{I}_K & \mathbf{I}_K \end{bmatrix} \mathbf{z}[t]^2$

// Attention mechanism

$\mathbf{Y}[t]_i \leftarrow \left[\mathbf{y}[t-i]^T \right]$

$\mathbf{a}[t] \leftarrow \mathbf{Y}[t]^T \mathrm{Softmax} \left(\frac{1}{\sqrt{K}} \mathbf{Y}[t] \tilde{\mathbf{y}}[t] \right)$

// Solution refinement

$\mathbf{y}[t] \leftarrow (\tilde{\mathbf{y}}[t] + \mathrm{ReLU}\left(\mathbf{U}\mathbf{a}[t] + \mathbf{b}\right)) \odot \sigma\left(\mathbf{V}\mathbf{a}[t]\right)$

end for

D. Final remarks

In this section, we provide some final remarks to link STAR to the mathematical formulation of the problem defined in Section IV-B2, and to other sequential CS algorithms.

STAR solves the sequential CS problem in Eq. (14) taking a different approach compared to existing works, such as Sequential LISTA (SLISTA) [32], the $\ell_1 - \ell_1$ RNN in [33], and Deep Unfolding Sparse Transformer mode (DUST) [34]. These approaches propose NNs that maintain the same structure of the original iterative optimization algorithm and exploit the temporal correlation of the data to initialize the solver. Specifically, the unrolling formulation in [32], [33] uses past information to initialize the solution once, before the first iteration of the unrolled algorithm. [34] instead uses it to initialize the solution at the beginning of every iteration. In Section VI, we show that with highly incomplete measurements (i.e., 90% missing samples) these initialization strategies fail, as the information carried by the current measurement vector can be very low, making the unrolled iterative algorithm converge to a poor solution despite the good initialization.

To cope with this, with STAR we take an opposite approach with respect to existing works. We use a single deep unrolling iteration to find an initial solution, which is then *refined* by exploiting the past evolution of the μD using temporal features obtained from an attention layer. This avoids a direct optimization of Eq. (14), which requires specifying function $d(\cdot)$ and, in turn, making restrictive assumptions on the μD dynamics. Thus, the attention layer is used to *learn* how to enhance the output of the sparse recovery problem solved through the LIHT block. It follows that STAR splits the problem into a CS part, implemented via deep unrolling, and a sequential feature learning part, implemented through the attention mechanism. The features obtained from the attention layer are then utilized to refine the solution *at the output* of the unrolling block, rather than to initialize it. Note that, although the attention layer is not part of the model-based unrolling framework (block **(A)**), its effect on the current solution $\tilde{\mathbf{y}}[t]$, by exploiting the correlation structure learned from past outputs, is still clearly interpretable.

TABLE 1: CIR parameters of the DISC dataset.

B [GHz]	$\Delta\tau$ [ns]	T_c [ms]	f_c [GHz]	Δv [m/s]	v_m [m/s]
1.76	0.568	0.27	60	0.14	$\pm$4.48

VI. Experimental Results

In this section, we present experimental results to validate STAR and compare it against state-of-the-art algorithms. The code implementation of our model, which was carried out in PyTorch [49], will be made available on GitHub[2] to facilitate reproducibility.

A. Dataset description and system parameters

We tested STAR on the publicly available DISC dataset[3] [14]. DISC contains, among other data, 416 IEEE 802.11ay CIR sequences at 60 GHz, obtained at a fixed sampling rate of $T_c = 0.27$ ms using a monostatic JCS Software Defined Radio (SDR) platform. CIR estimates are obtained in a standard-compliant fashion, using so-called TRN fields of pilot symbols appended as trailers to IEEE 802.11ay packets. The CIR sequences contain signal reflections on humans performing four different activities: *walking*, *running*, *waving hands* and *sitting down/standing up*. The dataset includes data from 7 different subjects. A summary of the parameters of the DISC CIR data is provided in Tab. 1, while for additional details regarding the data collection and the experimental testbed we refer to [11], [14]. We remark that sequences have different durations, ranging from 0.52 to 9.22 seconds, and that the dataset is *unbalanced*, with more samples belonging to *walking* compared to the other activities.

B. Training details and model parameters

1) Dataset splitting and preparation: We process the CIR data in windows of $K = 64$ steps in the temporal grid with spacing T_c, where each new window is shifted to the right by $\delta = 32$ samples with respect to the previous one, leading to $K - \delta$ overlapping samples between every two adjacent windows, as illustrated in Fig. 3. Therefore, each of the 416 sequences in DISC provides tens to hundreds of input CIR windows, where the exact number depends on the duration of the sequence. The whole set of μD sequences is split into non-overlapping training, validation, and test sets, with ratios $0.8, 0.01, 0.19$. The choice of these ratios is motivated later in this section. We split the dataset at the sequence level, before extracting the CIR windows, as: (i) we want to ensure a sufficient level of diversity between the training and validation/test sets, and (ii) the training process for STAR is *sequential*, i.e., the CIR windows must be processed in their original temporal ordering. In the training and evaluation of STAR, we obtain incomplete CIR measurement patterns by randomly removing samples from the DISC CIR sequences. During training, we provide the network with a wide range of diverse sampling patterns, with different sparsity levels. For this, we dynamically augment our training set by generating

[2] https://github.com/rmazzier/STAR

[3] https://dx.doi.org/10.21227/2gm7-9z72

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3408041

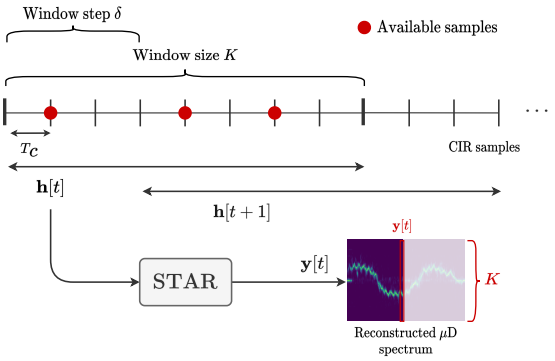

Fig. 3: Framing CIR samples into subsequent windows to be processed by STAR.

random binary masks applied to each CIR input window. The masks are generated by first sampling a mask probability $p \sim \mathcal{U}(0, 0.9)$, and then setting each element of the mask to 0 with probability p. This augmentation procedure makes the STAR training process extremely robust to overfitting, as the probability that two identical input sequences are presented to the network is negligible. This is the reason why we select a smaller validation set compared to the test set: We give more importance to obtaining a reliable estimate of the final metrics (guaranteed by a larger sample) than to evaluating the validation performance, as we observe no overfitting in our experiments. Moreover, despite being a 0.01 fraction of the whole dataset, our validation set contains 2837 CIR windows. We found this to be a sufficient sample size to tune the hyperparameters of STAR and to monitor its generalization capabilities during training.

Furthermore, to address the unbalanced nature of the dataset, we performed random oversampling of the μD spectrograms related to the *running* activity. This was done to balance the amount of *walking* and *running* sequences in the training set. This oversampling procedure was crucial: we empirically observed that the dataset being unbalanced, combined with the presence of higher Doppler frequency components in *running* compared to the other activities (due to the higher movement speed of the limbs) led to poor reconstruction of the *running* μD sequences. The oversampling procedure provides an effective solution to this issue, and thanks to the random mask generation there is no repetition of the same input CIR samples during training, which could cause overfitting.

2) Ground truth and loss function: To train the model, we first defined the ground truth as the μD spectrum reconstructed by the IHT algorithm at convergence, *when provided with a complete window of CIR samples*, i.e., when all the measurements are available. This served as reference data to train STAR. We denote the ground truth μD and the reconstructed DFT at time step t as $\mathbf{y}_{\text{gt}}[t]$ and $\mathbf{z}_{\text{gt}}[t]$, respectively. Our model was trained to faithfully reconstruct both the sparse reconstruction and the final μD spectrum, by minimizing the Mean Squared Error (MSE) between (i) the ground truth μD, $\mathbf{y}_{\text{gt}}[t]$, and its reconstruction, $\mathbf{y}[t]$, and (ii) the ground truth DFT $\mathbf{z}_{\text{gt}}[t]$ and $\mathbf{z}[t]$. We denote these two loss terms as $\mathcal{L}_{\mu\text{D}}[t]$ and $\mathcal{L}_{\text{IHT}}[t]$, respectively. The final training loss of our model is then

$$\mathcal{L}[t] = \alpha \mathcal{L}_{\mu\text{D}}[t] + \beta \mathcal{L}_{\text{IHT}}[t], \tag{19}$$

where $\alpha, \beta > 0$ are used to tune the relative importance of the two losses. Experimental results showed that putting more emphasis on the reconstruction of the μD spectrum ($\mathcal{L}_{\mu\text{D}}[t]$) yielded better performance, but $\mathcal{L}_{\text{IHT}}[t]$ is still useful to provide additional training feedback on the LIHT block, so we set $\alpha = 0.9, \beta = 0.1$, as this combination of weights led to the best results.

STAR hyperparameters		
Window length	K	64
Window shift	δ	32
Sparsity parameter	Ω	5
Inverse LIHT step	μ	20
No. of past windows	N_p	6
μD loss weight	α	0.9
IHT loss weight	β	0.1

TABLE 2: STAR hyperparameters used in the implementation.

In the LIHT module, we set the reciprocal of the step size $\mu = 20$, while in the attention block we consider $N_p = 6$ past μD windows. The model was trained for 5 epochs, using the Adam optimizer [50] with a learning rate of $2 \cdot 10^{-4}$, on a NVIDIA RTX3080 GPU. In Tab. 2, we summarize all the relevant hyperparameters that were used in our experiments. Note that, in the deep unrolling framework, parameters of the standard IHT algorithm (e.g., the sparsity parameters Ω and the inverse LIHT step size μ) can be seen as hyperparameters of the unrolled neural network, which are optimized using a validation set and kept *constant* during training. This distinction is done to set such parameters apart from the neural network weights and learnable parameters, which are instead optimized during training using backpropagation.

C. μD reconstruction results

1) Reconstruction quality: As a first evaluation step, we compare the reconstructed μD spectrum to the ground truth using two different metrics: the Root MSE (RMSE) and the Structural Similarity Index Metric (SSIM) [51]. Note that in all our results the ground truth μD spectra have been normalized in the interval $[0, 1]$, using min-max normalization as in [11]. The RMSE treats all the components of the two signals equally, without considering any specific features of the μD data. However, this is not always the best way of assessing the quality of reconstruction: When the goal is to evaluate the human-perceived signal quality, the RMSE has often proven to be inadequate [52]. Therefore, we also compute the SSIM, which is based on a combination of several visual aspects like lightness, contrast, and structural information of the reconstructed μD spectrum, and provides a more reliable measure of perceptual fidelity.

2) Comparison with state-of-the-art solutions: In our numerical analysis, we provide a comparison between our model, IHT [12], and DUST [34], which is the most recent approach for sequential sparse recovery based on deep unrolling and attention in the context of video processing. STAR and DUST

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3408041

differ under several aspects. First, DUST applies the self-attention operation across the whole input frames sequence, thus looking for correlations between the current frame and both past and future inputs. While this approach provides richer modeling of the correlations present in the sequence, it is impractical for real-time μD reconstruction, since it is *non-causal* and requires knowing the future input frames. Therefore, for the sake of a fair comparison, we constrain the attention steps of DUST to only operate on the same past windows considered by STAR.

Furthermore, another distinction lies in the domain in which the attention operation is applied. Unlike STAR, which directly applies the attention operation between μD spectra, DUST reconstructs the original signal and computes the correlations in the time domain. Specifically, the DUST attention step is described by the following equation

$$\begin{aligned}\mathbf{z}[t] &= \xi \sum_{i=1}^{N_p} \frac{e^{\mathbf{z}[t-i]^T \mathbf{W}^T \mathbf{W}\mathbf{z}[t]}}{\sum_{u=1}^{N_p} e^{\mathbf{z}[t-u]^T \mathbf{W}^T \mathbf{W}\mathbf{z}[t]}} \mathbf{z}[t-i] \\ &= \xi \mathbf{Z}[t]^T \mathrm{Softmax}\left(\mathbf{Z}[t]^T \mathbf{W}^T \mathbf{W}\mathbf{z}[t]\right), \end{aligned} \tag{20}$$

where $\mathbf{W}\mathbf{z}[t]$ is the reconstructed signal in the time domain, $\mathbf{Z}[t] = [\mathbf{z}[t-1], \ldots, \mathbf{z}[t-N_p]]^T$, and ξ is a trainable parameter. In the original version of DUST, the output of the attention layer is used to initialize a LISTA layer, which is then trained to solve the sparse reconstruction problem. We underline that this is a key distinction between DUST and STAR, as in the former the attention is applied at the input section, whereas in the latter it is employed at the end of the processing chain. Now, to understand the impact of this architectural choice, for DUST, we replace the LISTA module with a LIHT module. In this case, DUST and STAR reconstruct the micro-Doppler spectrum using a similar approach and their main difference resides in the positioning of the attention mechanism. Moreover, we also consider a DUST variant, which we denote by DUST-V2, which still uses the attention at the input, but the corresponding operations are executed in the frequency domain (as we do in STAR).

The experimental results show almost identical performance for the two DUST variants, which proves that the gain brought by STAR is not due to the domain (time versus frequency) in which the attention is performed. It is also important to observe that DUST requires *at least two* iterations of LIHT to be performed to successfully compute its attention weights (see Section 3.1 of [34]). This has to be taken into account when comparing it to STAR, which instead entails a single LIHT iteration.

To provide a broader benchmark of STAR's performance, we also compare it to established compressed sensing algorithms, namely OMP and Least Absolute Shrinkage and Selection Operator (LASSO) (or $\ell 1$-norm minimization). Note that we do not evaluate OMP and LASSO against IHT's ground truth, as for STAR and DUST, but against their own ground truth reconstructions obtained with a full measurement window. This choice is motivated by the fact that, unlike STAR and DUST which are based on IHT, OMP and LASSO may yield intrinsically different reconstructions compared to IHT since they are based on different optimization approaches. Our comparison method removes this mismatch and only considers the degradation due to the fewer CIR samples.

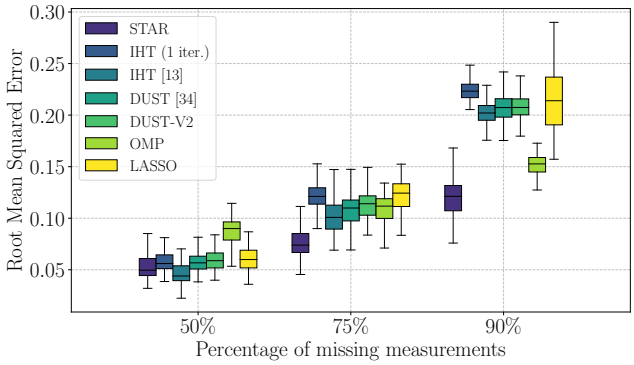

Fig. 4: Evolution of the RMSE over the test set for increasing percentages of missing measurements, achieved by STAR, in comparison with IHT at convergence, IHT stopped after one iteration, DUST, DUST-V2, OMP, and LASSO algorithms.

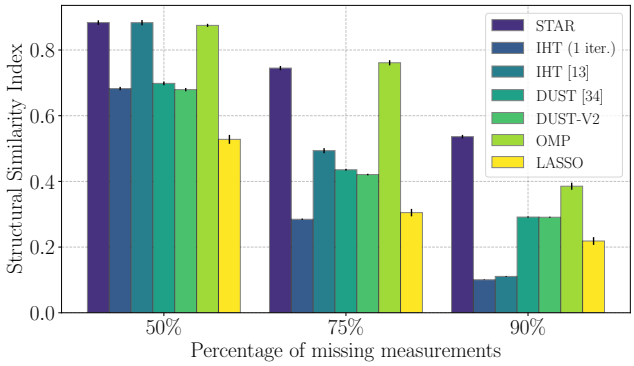

Fig. 5: Evolution of the SSIM over the test set for increasing percentages of missing measurements, achieved by STAR, in comparison with IHT at convergence, IHT stopped after one iteration, DUST, DUST-V2, OMP, and LASSO algorithms.

Fig. 4 and Fig. 5 show the μD reconstruction errors in terms of RMSE and SSIM, respectively. Error bars in Fig. 5 represent the standard error dispersion measure, i.e., $\sigma/\sqrt{n_{\text{test}}}$, where σ and n_{test} are the standard deviation and number of samples in our test set, respectively.

These results show that STAR performs close to IHT when the channel is densely sampled (low sparsity), which is the best-case scenario for IHT, and where STAR output gets very close to the ground truth reconstructions. However, as the number of available samples decreases, STAR consistently outperforms IHT stopped after one iteration (IHT 1 iter.), and IHT at convergence, both in terms of RMSE and SSIM. STAR also steadily outperforms both DUST variants. To visually compare the quality of the reconstructions for a highly incomplete window (90% missing measurement), in Fig. 6 we show the μD spectrograms reconstructed using STAR, IHT at convergence, DUST, OMP and LASSO.

Two key observations are in order.

(i) Placement of the attention layer: In the DL literature, attention layers are typically placed at the very beginning of the processing chain. For example, the classical transformer

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3408041

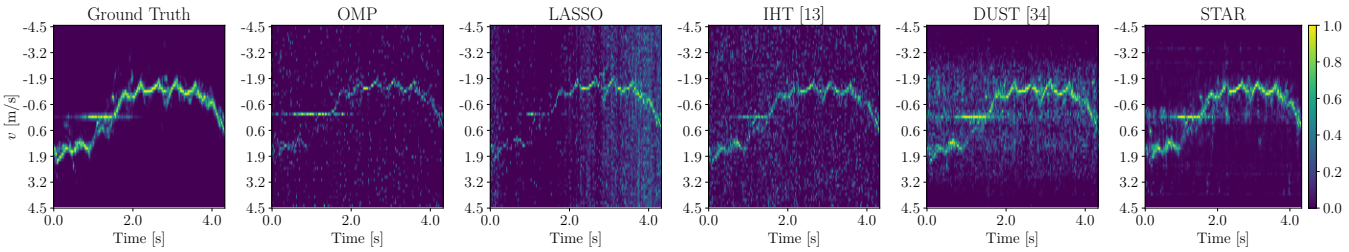

Fig. 6: Comparison of different sparse reconstructions provided by OMP, LASSO, IHT, DUST and STAR, when provided with CIR windows with 90% missing channel measurements.

architecture applies multiple self-attention layers in its input encoder module to learn contextualized features that are exploited by the subsequent layers. Similarly, DUST applies the self-attention mechanism at the input, to provide good initialization for the following LISTA algorithm. This is the key difference that sets STAR apart from DUST: Our model is architecturally different, as the attention mechanism is utilized towards the output of the processing chain rather than at its input, to refine the coarse-grained reconstruction obtained by the first unrolled layer. This choice was found to be highly effective for the reconstruction of highly sparse μD spectra.

(ii) Role of the unrolled layer: Existing approaches primarily rely on the predictions from the unrolled algorithms. This is not the case for STAR, where the primary role of the unrolled LIHT layer is to provide a first approximation of the solution. For the considered sensing application, experimental results revealed that just providing a good initialization to the LIHT layer (i.e., using attention at its input) is ineffective, and leads to poor-quality solutions.

3) Improved activity recognition capabilities: In this section, we evaluate the quality of the reconstructed μD signatures by using them as the input to a standard JCS application that performs human activity recognition. In the literature, this is typically implemented by applying a Convolutional Neural Network (CNN) classifier on the μD spectrograms [6], [11], [12]. We show that STAR provides a substantial improvement in the activity recognition task when only a few channel measurements are available. To this end, the following methodology was adopted:

i. We built a set of training data pairs $\mathcal{X}_{\text{train}} = \{(\mathbf{X}_i, \ell_i)\}_{i=1}^{N_{\text{train}}}$. Each $\mathbf{X}_i$ is a crop of a ground truth μD spectrogram, obtained by applying IHT until convergence on a sequence of $\Gamma = 200$ consecutive CIR windows. As a result, all the spectrogram crops in $\mathcal{X}_{\text{train}}$ have a time duration of $((\Gamma-1)\delta+K)T_c \approx 1.7$ seconds of measurements, see Fig. 3. ℓ_i is the activity label of the corresponding crop $\mathbf{X}_i$, which can be one among *walking*, *running*, *waving hands* or *sitting/standing up*.

ii. We built a set of test data pairs $\hat{\mathcal{X}}(p) = \{(\hat{\mathbf{X}}_i, \hat{\ell}_i)\}_{i=1}^{N_{\text{test}}}$, where each $\hat{\mathbf{X}}_i$ is a crop of Γ consecutive μD windows reconstructed by our model from an incomplete sequence of CIR samples, with a level of sparsity p. We stress that this set is built using the *test* CIR sequences, hence there is no overlap between the $\hat{\mathbf{X}}_i$ and the training set of STAR.

iii. We then trained a CNN classifier on the task of activity recognition on set $\mathcal{X}_{\text{train}}$. The structure of this neural network classifier is summarized in Tab. 3. We trained this CNN for 100 iterations, using the Adam optimizer with a learning rate of 10^{-3}.

iv. Finally, we obtained the CNN output classifications on the test set $\hat{\mathcal{X}}(p)$, and computed the global and per-class F1-scores as the evaluation metrics.

Layer	In Channels	Out Channels
`Conv_1 (ReLU)`	1	8
`Conv_2 (ReLU)`	8	16
`Conv_3 (ReLU)`	16	32
`Conv_4 (ReLU)`	32	64
`Conv_5 (ReLU)`	64	128
`Conv_6 (ReLU)`	128	128
`Flatten`	-	-
`Linear (ReLU)`	512	64
`Dropout (p=0.2)`	-	-
`Linear`	64	4

TABLE 3: Baseline CNN Architecture used in the evaluation. All convolutional layers use 3×3 kernels with a stride of 2.

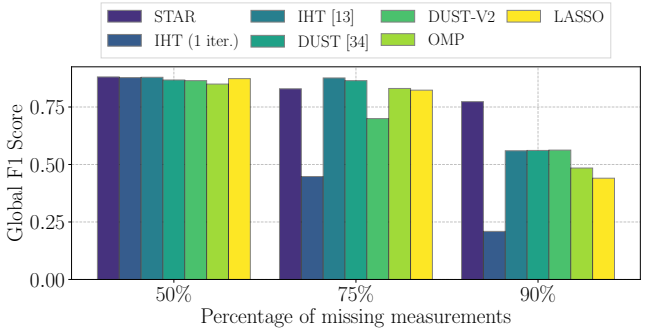

Fig. 7: Global F1-score for different sparsity levels.

In Fig. 7 and Fig. 8 we report the F1-scores achieved by STAR, IHT at convergence, IHT stopped after one iteration, DUST, and DUST-V2. In terms of global F1-score, results show that our method performs similarly to IHT and DUST at sparsity levels of 50% and 75%, but it greatly outperforms them with a 90% measurement sparsity. The comparable performance when many measurements are available is due to the robustness of CNNs to the presence of noise in the input data. Indeed, the reconstructions of all the evaluated models with low sparsity are only slightly noisier than those obtained from a full measurement window. However, at extreme sparsity levels, the quality of the reconstructions provided by the other

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3408041

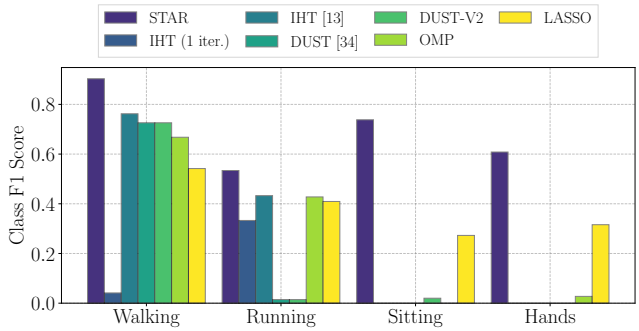

Fig. 8: Per-class F1-score with 90% incomplete channel measurements.

models drops significantly. Conversely, the reconstructions provided by STAR preserve the specific features of human movements, reaching an F1-score close to 0.8. A deeper insight into this can be gained by observing the values of the class-specific F1-scores (see Fig. 8). The class-specific F1-score evaluates the capability of the classifier to recognize each class individually, whereas the global F1-score is a measure of the aggregate performance. The class-specific F1-score is useful to understand which classes are more affected by the high number of missing measurements. With 90% missing measurements, STAR largely outperforms its competitors on each activity. The gain is especially evident for those activities involving fine-grained movements, i.e., *sitting/standing up* and *waving hands*. DUST, IHT, and OMP instead fail to provide discriminative reconstructions, obtaining an F1-score of 0. This is due to the insufficient reconstruction quality: As shown in Fig. 9, in μD recovered by DUST, IHT, and OMP the movement features are hidden by severe reconstruction artifacts. Conversely, LASSO has lower (not exceeding 55%) but more consistent F1-scores across activities.

STAR achieves good performance with over 0.6 F1-score for all classes. This shows that the gain brought by STAR is much more significant than the 0.2 gain shown in Fig. 7. STAR is the only model that can recover the μD features of small movements, which is a key enabler for gesture recognition applications. Fig. 9 provides a qualitative comparison of μD reconstructions relative to the *waving hands* activity.

D. Ablation studies

We perform an ablation study by evaluating several variations of our model, obtained by modifying or removing different components of the architecture, to assess their impact on the final performance. The considered variations are:

- STAR with $N_p = 1, 3, 9$;
- STAR where we replace matrix $\mathbf{I} - \mathbf{W}^T\mathbf{W}$ in Eq. (15) with matrix $\mathbf{S} \in \mathbb{R}^{2K \times 2K}$, containing $4K^2$ additional learnable weights. We refer to this variant as "Learn $\mathbf{S}$".
- a single LIHT block, i.e., STAR without the attention and solution refinement blocks. We call this variation "No Attention";
- STAR where the solution refinement block is modified by removing the multiplicative branch, only retaining the additive modification of the LIHT solution ("Only Add");

In Tab. 4, we report the RMSE and SSIM metrics, as well as the global F1-score achieved by the CNN classifier applied to the reconstructed μD signatures, for each model variation.

Our results reveal three main insights regarding our model. Firstly, increasing N_p above 6 does not yield a consistent improvement for all metrics. Indeed, the best CNN F1-score with 90% sparsity, as well as the best RMSE values are obtained with $N_p = 6$. Moreover, larger values of N_p require storing a longer sequence of past outputs and increase the complexity of the attention mechanism. This is not necessarily beneficial, since the autocorrelation of μD spectra quickly goes to zero as the time lag increases. Therefore, we choose $N_p = 6$ as the default configuration for STAR.

Secondly, learning matrix $\mathbf{S}$ provides a performance improvement in terms of SSIM and CNN F1-score at lower sparsity levels, but it is not beneficial with a sparsity of 90%. The additional cost of learning $4K^2$ additional weights, and the fact that learning $\mathbf{S}$ makes the model less interpretable has to be considered as well. Hence, in STAR we decided not to learn $\mathbf{S}$, and to use matrix $\mathbf{I} - \mathbf{W}^T\mathbf{W}$ in Eq. (15) instead.

Finally, the results for "No Attention" and "Only Add" show that the proposed additive and multiplicative refinement block, based on a context vector provided by attention, brings a substantial improvement. This is especially evident for high percentages of missing input measurements. The reason for this improvement lies in the additional information provided by attention, which is particularly important in the presence of highly incomplete input measurements.

E. Test in a different environment

To assess the ability of our model to generalize across different environments, we evaluate STAR on CIR sequences collected in a different room (called *test room*) than the one it was trained on (*training room*). The data from the test room is also provided in the DISC dataset [14] and contains measurements of a single human subject performing the same set of actions of the training set. Following the same methodology of Section VI-C, we evaluate the model in terms of μD spectrogram reconstruction error and F1-score on the activity recognition task using the CNN of Tab. 3. In Tab. 5, we compare the values of RMSE, SSIM, and CNN F1-scores obtained on reconstructions coming from the training room and the test room. The results are in line with those presented in Section VI-C, showing the same reconstruction quality, and similar degradation on the classification task when providing the CNN model with reconstructions from 90% missing measurements. This demonstrates that STAR generalizes beyond the specific environment used during training.

F. Computational complexity and model size

In this section, we show that STAR is a lightweight model both in terms of computational complexity and model size (intended as the number of learnable parameters).

1) Computational complexity: The complexity of STAR is in the same order of just one IHT iteration, making it well suited for real-time applications. We start by analyzing the complexity of IHT. From Eq. (15), one can see that the number

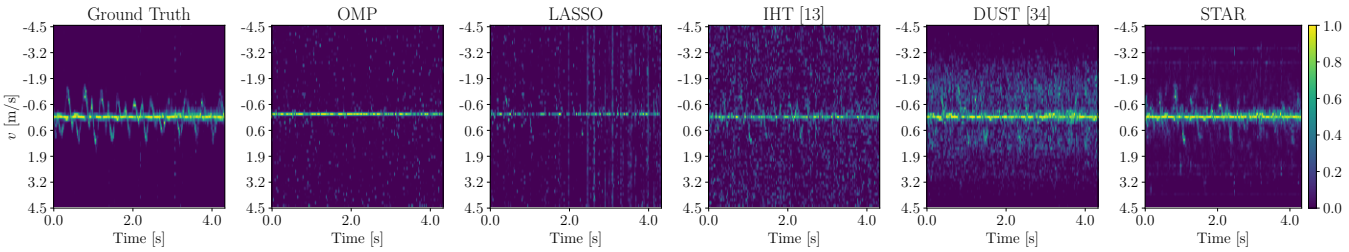

Fig. 9: Comparison of different sparse reconstructions for the *Waving hands* activity, with 90% missing input channel measurements. STAR is capable of retaining the characteristic shape of the activity while removing the majority of the background noise.

	RMSE			SSIM			CNN F1-score			No. parameters
Missing measurements [%]	50%	75%	90%	50%	75%	90%	50%	75%	90%	
$N_p = 1$	0.0564	0.0828	0.1379	0.806	0.608	0.374	0.875	0.814	0.764	24,640
$N_p = 3$	0.0546	0.0784	0.1252	0.859	0.687	0.433	0.873	0.831	0.766	24,640
$N_p = 9$	0.0548	0.0784	0.1214	0.887	0.760	**0.581**	**0.882**	0.829	0.757	24,640
Learn $\mathbf{S}$	0.0578	0.0846	0.1412	**0.897**	**0.771**	0.488	0.876	**0.859**	0.330	41,024
No Attention	0.0635	0.1128	0.1999	0.724	0.492	0.319	0.845	0.701	0.582	16,384
Only Add	0.0636	0.1128	0.1998	0.714	0.479	0.230	0.841	0.700	0.586	20,544
STAR	**0.0545**	**0.0779**	**0.1213**	0.884	0.745	0.536	0.881	0.829	**0.773**	24,640

TABLE 4: Results of the ablation studies in which we compare different variations of STAR. We highlight the best values using a **bold** font.

TABLE 5: STAR performance in a different environment. We report μD reconstruction quality metrics (RMSE and SSIM) in the training room and the (unseen) test room.

	RMSE	SSIM
Training room	0.121 ± 0.002	0.536 ± 0.005
Test room	0.115 ± 0.004	0.479 ± 0.006

TABLE 6: CNN F1-score on ground truth (IHT full) and STAR reconstructions obtained in the test room.

	Walking	Running	Sitting	Hands	Global
IHT (full)	0.896	0.902	0.607	0.588	0.750
STAR (90%)	0.829	0.750	0.557	0.432	0.652

of operations required at each iteration is asymptotically dominated by the matrix product $\mathbf{W}^T\mathbf{W}$. As $\mathbf{W} \in \mathbb{C}^{M_t \times K}$, the number of operations involved in a single iteration is in the order of $K^2 M_t$. Overall, considering N IHT iterations, the complexity is $\mathcal{O}(NK^2M_t)$.

We can now analyze the complexity of STAR, by deriving the number of computations for each module:

1) The LIHT block has the same complexity as a single IHT iteration, namely, $\mathcal{O}(K^2M_t)$.
2) The attention mechanism performs N_p dot products between vectors of dimension K, resulting in a complexity of $\mathcal{O}(KN_p)$. Linearly combining the resulting correlations does not increase the complexity order.
3) The computational complexity of the solution refinement module is determined by the vector-matrix multiplications performed by the two feedforward layers (Eq. (18)). Recalling that matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{2K \times 2K}$, and $\mathbf{a}[t] \in \mathbb{R}^K$, the complexity is $\mathcal{O}(K^2)$.

Combining the above steps we obtain a total complexity of $\mathcal{O}(K^2M_t)$, which represents an N-fold gain with respect to applying IHT to convergence. In practice, this provides a huge speedup in reconstructing the μD sequences, as IHT easily takes $N \approx 15$ to 20 iterations to converge.

A similar analysis can be performed for DUST, yielding a complexity of $\mathcal{O}(NK^2M_t)$, which is identical to that of IHT. However, DUST takes much fewer iterations to provide acceptable results, although a minimum of 2 is required, as explained in Section VI-C. This shows that DUST is not faster than our model, even in the best case with $N = 2$.

2) Model size: Our approach is also lightweight in terms of number of learnable parameters. In the LIHT module, which performs the computations described in Eq. (15), the only learnable parameters are the entries of matrix $\mathbf{W}$, resulting in $4K^2 = 16384$ parameters. As discussed in Section VI-D, learning additional weights in this module does not significantly enhance the performance. The attention module does not contain learnable parameters. Finally, the solution refinement block in Eq. (18) learns the weight matrices $\mathbf{U}$, $\mathbf{V}$, and the bias vector $\mathbf{b}$, resulting in a total of $2K^2 + K = 8256$ parameters. Hence, STAR consists of a total of 24640 learnable parameters (32 bit floating point), amounting to 98 kB memory space. As a comparison, DUST has 32769 parameters, as it learns an additional weight matrix in its unrolling layer.

G. Communication overhead reduction

Next, we evaluate the impact of applying STAR to a JCS system to reduce the overhead introduced by the sensing process on communication. To this end, we consider an IEEE 802.11ay system as a reference. We consider the system parameters of Tab. 1 along with processing windows of $W = 64$ time slots, in which communication payloads with a Physical layer Service Data Unit (PSDU) size of 4 kB are

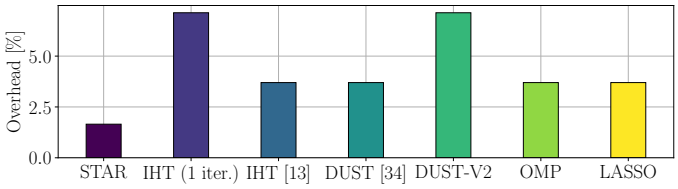

Fig. 10: Overhead on communication introduced by each model to obtain a global F1-score over 0.75.

transmitted in standard-compliant packets using Modulation and Coding Scheme (MCS) 9 (similar but rescaled results are obtained for other MCSs). Each packet includes preamble fields containing 4352 $\pi/2$-Binary Phase-Shift Keying (BPSK)-modulated symbols [53]. To perform sensing, we follow [11], [54], assuming that a TRN field with 768 symbols is appended to the packets.

For each method, we derive the number of CIR samples needed to achieve a global F1-score threshold of at least 0.75 on the activity recognition task, by inspecting Fig. 7. As an example, STAR can reach this threshold even with just 10% available measurements per window, hence it requires 7 CIR samples with $W = 64$. Conversely, DUST requires 16 CIR samples (25% of the window). We assume that 10 communication packets are transmitted in the window, serving as the reference number of information bits transmitted by the user. The overhead is obtained as the ratio between the length of the *added* PHY layer symbols in the TRN units (for sensing purposes) and the total number of symbols in the packet, considering the number of information bits plus PHY and MAC headers [53].

Fig. 10 shows the overhead introduced by each μD reconstruction method based on how many measurements per window it needs to reach the F1-score threshold (0.75). STAR adds less than half the overhead of the other methods, thus demonstrating its usefulness in lowering the impact of sensing on the communication performance. Note that the specific overhead value depends on the number of communication packets that are to be transmitted in a processing window. Hence, Fig. 10 is intended to provide a comparison among the different methods rather than actual overhead estimates.

As a final note, we remark that, when communication traffic is extremely irregular, μD reconstruction may require the transmission of *additional* channel estimation fields for sensing, as highlighted in [12]. Therefore, it should be considered that, besides a larger overhead, existing methods also entail higher channel occupation, which may cause a communication rate reduction for *other* network nodes in the same radio cell.

VII. Concluding Remarks

In this paper, we tackled the problem of reconstructing μD spectrograms of human movement from *highly incomplete* channel estimates in a JCS system. To this end, we designed and evaluated STAR, an interpretable NN architecture that effectively combines deep unrolling of a single iteration of thresholding-based compressed sensing and an attention mechanism that exploits the temporal sequentiality of the μD. The key insight behind STAR design is to use attention to directly enhance the reconstructed spectrum, thus boosting the model's robustness to highly incomplete input measurements. Differently from existing standard or learning-based approaches, STAR provides accurate μD reconstructions even with 90%-incomplete channel estimates while retaining a low computational complexity.

STAR is thoroughly evaluated on the publicly available DISC dataset [14], containing standard-compliant 60 GHz IEEE 802.11ay CIR estimates from a physical environment where signal reflections are affected by moving people. STAR significantly outperforms existing algorithms from the literature in terms of RMSE and SSIM. Moreover, when using the reconstructed μD to perform human activity recognition, state-of-the-art approaches completely fail when only 10% of the channel measurements are available (0 to 0.3 F1-score), while STAR yields F1-scores from 0.5 to 0.8.

Future work includes exploring more advanced ways of integrating signal processing domain knowledge into the NN architecture. A possible research direction is to move beyond the deep unrolling paradigm to develop model-based and physics-informed neural networks, which naturally embed equations regulating signal propagation and human movement models into their structure.

References

[1] H. Wymeersch, D. Shrestha, C. M. De Lima, V. Yajnanarayana, B. Richerzhagen, M. F. Keskin, K. Schindhelm, A. Ramirez, A. Wolfgang, M. F. De Guzman, *et al.*, "Integration of communication and sensing in 6G: A joint industrial and academic perspective," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, (Helsinki, Finland), IEEE, September 2021.

[2] S. A. Shah and F. Fioranelli, "RF sensing technologies for assisted daily living in healthcare: A comprehensive review," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, pp. 26–44, Nov 2019.

[3] T. Guo, X. Li, M. Mei, Z. Yang, J. Shi, K.-K. Wong, and Z. Zhang, "Joint Communication and Sensing Design in Coal Mine Safety Monitoring: 3D Phase Beamforming for RIS-Assisted Wireless Networks," *IEEE Internet of Things Journal*, 2023.

[4] P. Kumari, J. Choi, N. González-Prelcic, and R. W. Heath, "IEEE 802.11 ad-based radar: An approach to joint vehicular communication-radar system," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3012–3027, 2017.

[5] S. D. Regani, C. Wu, B. Wang, M. Wu, and K. R. Liu, "mmWrite: Passive Handwriting Tracking Using a Single Millimeter Wave Radio," *IEEE Internet of Things Journal*, 2021.

[6] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, A. Bourdoux, W. De Neve, and T. Dhaene, "Indoor person identification using a low-power FMCW radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 7, pp. 3941–3952, 2018.

[7] T. Li, L. Fan, Y. Yuan, and D. Katabi, "Unsupervised Learning for Human Sensing Using Radio Signals," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3288–3297, January 2022.

[8] K. Wu, J. A. Zhang, X. Huang, R. W. Heath, and Y. J. Guo, "Green Joint Communications and Sensing Employing Analog Multi-Beam Antenna Arrays," *IEEE Communications Magazine*, 2023.

[9] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-Doppler effect in radar: phenomenon, model, and simulation study," *IEEE Transactions on Aerospace and electronic systems*, vol. 42, no. 1, pp. 2–21, 2006.

[10] A. Hanif, M. Muaz, A. Hasan, and M. Adeel, "Micro-Doppler based target recognition with radars: A review," *IEEE Sensors Journal*, vol. 22, pp. 2948–2961, Jan 2022.

[11] J. Pegoraro, J. O. Lacruz, F. Meneghello, E. Bashirov, M. Rossi, and J. Widmer, "RAPID: Retrofitting IEEE 802.11ay access points for indoor human detection and sensing," *IEEE Transactions on Mobile Computing*, 2023.

[12] J. Pegoraro, J. O. Lacruz, M. Rossi, and J. Widmer, "SPARCS: A Sparse Recovery Approach for Integrated Communication and Human Sensing in mmWave Systems," in *21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, (Milan, Italy), May 2022.

[13] Y. C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[14] J. Pegoraro, J. O. Lacruz, M. Rossi, and J. Widmer, "DISC: a dataset for integrated sensing and communication in mmWave systems," in *IEEE Dataport*, IEEE, 2022.

[15] J. Gérard, J. Tomasik, C. Morisseau, A. Rimmel, and G. Vieillard, "Micro-Doppler signal representation for drone classification by deep learning," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 1561–1565, IEEE, 2021.

[16] A.-K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward unobtrusive in-home gait analysis based on radar micro-Doppler signatures," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, pp. 2629–2640, 2019.

[17] V. C. Chen, "Analysis of radar micro-Doppler with time-frequency transform," in *Proceedings of the Tenth IEEE Workshop on Statistical Signal and Array Processing (Cat. No. 00TH8496)*, pp. 463–466, IEEE, 2000.

[18] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, pp. 51–56, 2019.

[19] G. Lai, X. Lou, and W. Ye, "Radar-Based Human Activity Recognition With 1-D Dense Attention Network," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[20] Z. Meng, S. Fu, J. Yan, H. Liang, A. Zhou, S. Zhu, H. Ma, J. Liu, and N. Yang, "Gait Recognition for Co-Existing Multiple People Using Millimeter Wave Sensing," in *AAAI Conference on Artificial Intelligence*, (New York, New York, USA), Feb 2020.

[21] 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*, vol. 59, no. 4, pp. 2994–3009, 2021.

[22] I. Djurović, V. Popović-Bugarin, and M. Simeunović, "The STFT-based estimator of micro-Doppler parameters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, pp. 1273–1283, Jun 2017.

[23] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-based dynamic hand gesture recognition using micro-Doppler signatures," in *Proc. IEEE Radar Conference (RadarConf17)*, (Seattle, WA, USA), IEEE, May 2017.

[24] G. Li and P. K. Varshney, "Micro-Doppler parameter estimation via parametric sparse representation and pruned orthogonal matching pursuit," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, pp. 4937–4948, Dec 2014.

[25] S. Stanković, I. Orović, T. Pejaković, and M. Orović, "Compressive sensing reconstruction of signals with sinusoidal phase modulation: application to radar micro-Doppler," in *2014 22nd Telecommunications Forum Telfor (TELFOR)*, pp. 565–568, IEEE, 2014.

[26] E. Sejdić, I. Orović, and S. Stanković, "Compressive sensing meets time–frequency: An overview of recent advances in time–frequency processing of sparse signals," *Digital signal processing*, vol. 77, pp. 22–35, Jun 2018.

[27] P. Kumari, N. J. Myers, and R. W. Heath, "Adaptive and fast combined waveform-beamforming design for mmWave automotive joint communication-radar," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, pp. 996–1012, Apr 2021.

[28] J. A. Zhang, A. Cantoni, X. Huang, Y. J. Guo, and R. W. Heath, "Framework for an innovative perceptive mobile network using joint communication and sensing," in *IEEE 85th Vehicular Technology Conference (VTC Spring)*, (Sydney, Australia), IEEE, Nov 2017.

[29] B. Li and A. P. Petropulu, "Joint transmit designs for coexistence of MIMO wireless communications and sparse sensing radars in clutter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, pp. 2846–2864, Dec 2017.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Processing Magazine*, vol. 38, pp. 18–44, Feb 2021.

[32] S. Wisdom, T. Powers, J. Pitton, and L. Atlas, "Building recurrent networks by unfolding iterative thresholding for sequential sparse recovery," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4346–4350, IEEE, 2017.

[33] Le, Hung Duy and Van Luong, Huynh and Deligiannis, Nikos, "Designing Recurrent Neural Networks by Unfolding an L1-L1 Minimization Algorithm," in *2019 IEEE International Conference on Image Processing (ICIP)*, (Taipei, Taiwan), September 2019.

[34] B. De Weerdt, Y. C. Eldar, and N. Deligiannis, "Designing Transformer Networks for Sparse Recovery of Sequential Data Using Deep Unfolding," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[35] G. Izacard, S. Mohan, and C. Fernandez-Granda, "Data-driven estimation of sinusoid frequencies," in *Advances in Neural Information Processing Systems*, (Vancouver, Canada), Dec 2019.

[36] Y. Jiang, H. Li, and M. Rangaswamy, "Deep learning denoising based line spectral estimation," *IEEE Signal Processing Letters*, vol. 26, pp. 1573–1577, Nov 2019.

[37] R. M. Dreifuerst and R. W. Heath, "SignalNet: A low resolution sinusoid decomposition and estimation network," *IEEE Transactions on Signal Processing*, vol. 70, pp. 4454–4467, Aug 2022.

[38] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[39] J. F. Mota, N. Deligiannis, A. C. Sankaranarayanan, V. Cevher, and M. R. Rodrigues, "Adaptive-rate reconstruction of time-varying signals with application in compressive foreground extraction," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3651–3666, 2016.

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

[41] S. Lohit, D. Liu, H. Mansour, and P. T. Boufounos, "Unrolled projected gradient descent for multi-spectral image fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Brighton, UK), IEEE, May 2019.

[42] S. A. H. Hosseini, B. Yaman, S. Moeller, M. Hong, and M. Akçakaya, "Dense recurrent neural networks for accelerated MRI: History-cognizant unrolling of optimization algorithms," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, pp. 1280–1291, Jun 2020.

[43] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *27th International Conference on Machine Learning (ICML)*, (Haifa, Israel), Jun 2010.

[44] B. Xin, Y. Wang, W. Gao, D. Wipf, and B. Wang, "Maximal sparsity with deep networks?," in *Advances in Neural Information Processing Systems*, (Barcelona, Spain), Dec 2016.

[45] J. A. Zhang, F. Liu, C. Masouros, R. W. Heath, Z. Feng, L. Zheng, and A. Petropulu, "An overview of signal processing techniques for joint communication and radar sensing," *IEEE Journal of Selected Topics in Signal Processing*, 2021.

[46] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems Magazine*, vol. 29, no. 6, pp. 82–100, 2009.

[47] D. Reid, "An algorithm for tracking multiple targets," *IEEE transactions on Automatic Control*, vol. 24, no. 6, pp. 843–854, 1979.

[48] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "mID: Tracking and Identifying People with Millimeter Wave Radar," in *15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, (Santorini Island, Greece), May 2019.

[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[50] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," in *ICLR: International Conference on Learning Representations*, pp. 1–15, 2015.

[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[52] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[53] IEEE 802.11 working group, "IEEE Draft Standard for Information Technology-Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks-Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications-Amendment: Enhanced Throughput for Operation in License-Exempt Bands Above 45 GHz," *IEEE P802.11ay/D3.0*, 2019.

[54] S. Blandino, T. Ropitault, A. Sahoo, and N. Golmie, "Tools, models and dataset for IEEE 802.11 ay CSI-based sensing," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 662–667, IEEE, 2022.