# Joint RIS-Assisted Localization and Communication: A Trade-off Among Accuracy, Spectrum Efficiency, and Time Resource

Sanaz Kianoush, *Member, IEEE*, Alessandro Nordio, *Member, IEEE*, Laura Dossi, Roberto Nebuloni, Stefano Savazzi, *Member, IEEE*

***Abstract*—Integrated sensing and communication (ISAC) and reconfigurable intelligent surfaces (RISs) are viewed as promising technologies for the future 6G wireless networks. ISAC designs assisted by RISs are particularly attractive for tracking and localization problems in internet of everything (IoE) applications. In particular, RISs can be deployed to realize a smart radio environment (SRE) that tracks the user equipment (UE) in blind spaces, i.e., where the direct line-of-sight (LoS) wireless link is not available. This paper proposes a deep learning framework to integrate RIS-assisted mobile UE localization and communication in the 60 GHz band. The number of RIS and their electronic steering angles are investigated to support both localization and communication processes implemented on shared time resources. The UE localization is obtained through DL algorithms based on convolutional neural networks (CNN) and vision transformers (ViT) structures. The proposed algorithms are trained using a wide variety of physical parameters such as number of RIS steering angles, RIS area size, and number of antenna at the base station (BS). The system performance is measured in terms of achieved positioning root mean squared error (RMSE), algorithm complexity, and inference time. A Cramér-Rao bound for estimating the localization error based on RISs deployment, is also provided. Localization accuracy, frame efficiency and throughput tradeoffs are explored for different IoT setups.**

***Index Terms*—Integrated sensing and communication (ISAC), Reconfigurable intelligent surface (RIS), sub-THz communication, localization, deep learning, transformers.**

## I. Introduction

THE emergence of integrated sensing and communications (ISAC) as a cutting-edge technology is closely related to recent advancements in wireless communication, to growing demands for enhanced sensing capabilities, and to the need of real-time precise localization of the wireless network nodes. ISAC has garnered significant research attention and is recognized as a key technique for shaping the future of mobile communications, particularly in the context of the sixth generation (6G) of mobile communications [1]. A fundamental concept within ISAC involves leveraging communication signaling to infer user's position [2]. Within this framework, dynamically configured radio environments, also referred to as SRE, along with reconfigurable intelligent surfaces (RIS), hold considerable potential in enhancing both sensing and communication functionalities.

RISs, are typically two-dimensional arrays of small, passive, and reconfigurable elements [3] called meta atoms. These arrays are engineered using advanced metamaterials that exhibit unique properties, such as negative refractive index, which enable fine-grained control over the direction and strength of reflected waves. By phase-aligning the signals scattered by the meta atoms, a RIS can macroscopically behave as a steerable mirror that reflects the impinging signal towards a desired direction [4]. By manipulating the propagation of electromagnetic waves, RISs can improve the spectral efficiency of communication systems, extend the communication and sensing coverage [5], and improve the estimation of the signal direction of arrival (DOA), or the localization accuracy [6], [7]. With a proper deployment and configuration of RISs, connectivity between any two communicating devices can be granted even in the absence of a direct LoS path among them. It has been shown that RISs represent an efficient solution to solve wireless channel impairments, mitigate interference and realize energy-efficient beamforming [8].

RIS-assisted ISAC represents a new paradigm in 6G communications that has not been fully investigated. To fill this gap, in this work we explore the scenario depicted in Figure 1 where a base station (BS) communicates with a set of UEs and is assisted by RISs for both data communication and UEs localization.

A time-division allocation policy is proposed to integrate real-time localization and communication. On the one hand precise localization is crucial to enhance communication efficiency even though the localization process consumes time resources, which detracts from those allocated for communication. On the other hand low-accuracy localization requires less resources but also entails poor communication performance.

In this context, there is a trade-off with the aim of maximizing both the localization efficiency and the network throughput. Specifically, there is a loss in data throughput due to the necessity of transmitting pilot symbols dedicated to the localization task. We observe that extending the localization time results in a smaller localization error, leading to a higher Signal-to-Noise Ratio (SNR) and overall network throughput. However, a longer localization time decreases frame efficiency,

Authors are with Consiglio Nazionale delle Ricerche (CNR), IEIIT institute, Corso Duca degli Abruzzi 24, 10129 Torino, and P.zza Leonardo da Vinci 32, 20133, Milano, Italy. Corresponding author email:sanaz.kianoush@cnr.it

This work is partially supported by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Innovation Council and SMEs Executive Agency (EISMEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement No: 101099491, HOLDEN project. The work is also supported by the Italian National Recovery and Resilience Plan (NRRP) of NextGeneration EU, partnership on "Telecommunications of the Future" (PE0000001 program "RESTART").

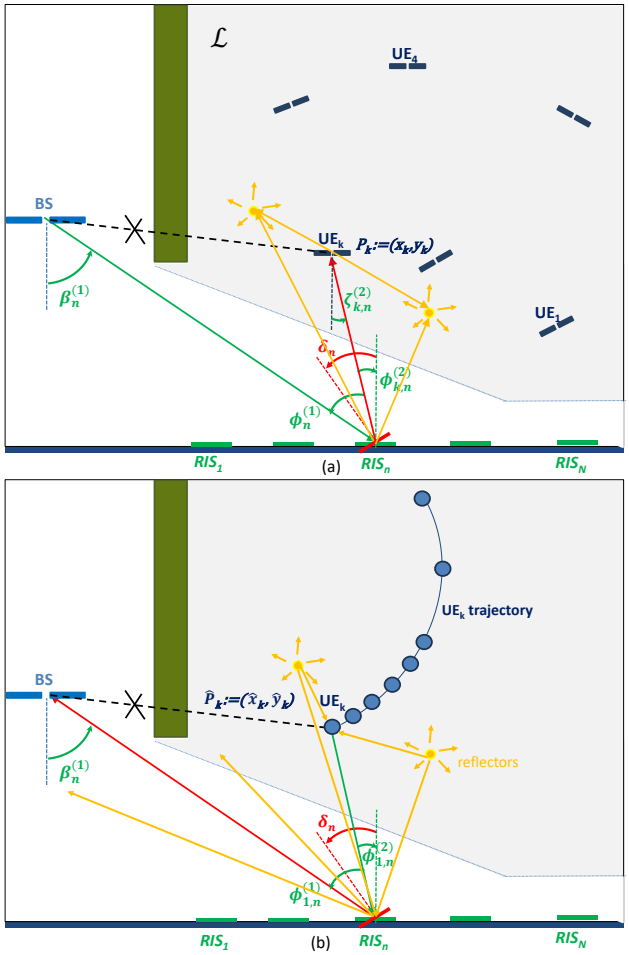


Fig. 1. Scheme of a RIS-aided communication and localization system where a BS exploits a set of RISs to interact with the UEs which are positioned in the shaded area. The LoS path connecting the BS to the UEs is blocked by an obstacle represented by the gree solid rectangle. Green and red solid lines refer to the BS-RIS and RIS-UE LoS paths, respectively, whereas the yellow lines represent the paths due to some reflecting objects. (a) RIS-assisted communication (downlink): $N$ RISs are deployed along a wall and reflect the BS signal towards the UEs; the optimum electronic rotation angle applied by RIS $n$ is denoted by $\delta_n$, (b) RIS-based localization (uplink): the BS collects the UE signal reflected by a set of dedicated RISs.

thereby reducing the network throughput. Therefore, it is expected that there is an optimal combination of localization error and frame efficiency that maximizes the network throughput.

### A. Related works

ISAC is an emerging technology that enables simultaneous sensing of targets and communication with users [9]. ISAC systems allow improving the efficiency of both sensing and communication services by sharing limited resources, and is expected to play a critical role in many future applications, such as the internet of everything, and smart cities [10].

Localization in ISAC systems is performed taking into account for different channel characteristics, including the angle of arrival, the time of arrival, the time difference of arrival, and the received signal strength (RSS) [11]. The implementation of ISAC systems faces several challenges. For example, when operating in the millimeter-wave (mmWave) band, localization signals suffer from high propagation attenuation and their performance severely degrades especially when the line-of-sight (LoS) between transmitting and receiving ends is obstructed [12]. Therefore, one of the main challenges is finding an optimal allocation of the available network resources to enhance the flexibility and capacity of ISAC systems [13].

The capability of RISs to provide virtual LoS links, modify the propagation environment and improve the localization accuracy is recently driving the development of RIS-aided ISAC systems [14].

Authors in [14] proposed a superimposed symbol scheme for double-RIS aided ISAC mmWave systems. Their aim was to enable simultaneous communication and localization for multiple UE. Their proposed approach involves estimating the initial target direction, followed by an iterative refinement through reduced-dimension matrix calculations. Simulation results demonstrate that their scheme improves the throughput and localization accuracy by 2 orders of magnitudes compared with a scenario without RIS.

In [15] we proposed a jointly optimized RIS-aided environment where the average network throughput is optimized considering the UE positioning error. Extending the work in [15], here we study the problem of integrating sensing and communication tasks in a multi-user time-variant scenario ensuring frame efficiency and network throughput.

In the following sections, we describe the main contributions of our paper regarding RIS-aided mobile UE localization and communication within the 60 GHz frequency band.

### B. Contributions

This paper extends our previous work [15] by adding the following key novel contributions:

- We propose a new solution for the localization of mobile UEs, exploiting multiple RISs deployed in an indoor environment. This solution makes use of deep learning (DL) algorithms. In particular, convolutional neural network (CNN) and vision transformer (ViT) are used for UE trajectory estimation. We measure the localization performance in terms of the achieved positioning root mean square error (RMSE) and we show its dependence on the system parameters.
- We derive an analytical expression for the Cramér-Rao Lower Bound (CRLB) on the localization RMSE. This bound is dependent on a) the number of RISs used for localization, b) the number of RIS rotations, and c) SRE geometry including the position of the BS and of the RISs. It serves as a tool to aid in pre-deployment design decisions.
- A time-division resource allocation strategy is proposed for the integrated UE localization and communication process in order to optimize localization accuracy and frame efficiency aiming to maximize the network throughput. We observed that, by increasing the localization accuracy, the network throughput is improved until an optimal operating point is reached, beyond which a further increase of the localization accuracy penalizes the network throughput.

- A Tradeoff analysis is considered for different computational capabilities of the deployed devices, namely low-power to high performance, and IoT relevant scenarios in terms of inference time and computational complexity.

The rest of paper is structured as follows. In Section II we address the details of the integrated RIS-aided localization and communication, introducing and exploring the challenges inherent in this context. Section III, delves into the channel model and characterization of RIS. In Sect. IV the localization process, the proposed DL algorithms and the localization scenario are described. Finally, in Sect. V, we conduct a thorough numerical analysis to assess the system's performance, thereby presenting conclusive insights into the effectiveness of the proposed approach.

### C. Mathematical notation

Throughout the paper we use the following mathematical notation: boldface uppercase and lowercase letters denote matrices and vectors, respectively. $\mathbf{I}_k$ is the $k \times k$ identity matrix and the conjugate transpose of matrix $\mathbf{A}$ is denoted by $\mathbf{A}^{\mathsf{H}}$. Moreover $\mathbf{A}^{+}$ and $\|\mathbf{A}\|_{\mathrm{F}}$ refer, respectively, to the Moore-Penrose pseudo-inverse and the Frobenius norm of $\mathbf{A}$.

## II. Integrated RIS-based sensing and communication system

In this section we describe our proposed integrated RIS-aided localization and communication system, designed to work in a time-variant indoor environment. First of all, we describe the system setup, the main system parameters and the resource allocation strategy. Finally, we discuss the system optimization, in terms of localization error and frame efficiency to maximize the network throughput.

### A. Localization and communication system

As depicted in Figure 1(a), we consider an indoor communication and localization scenario where a BS serves as access point (AP) and communicates with a set of UEs. The BS transmits $K$ data streams to as many mobile UEs that are moving within the shaded area on trajectories unknown to the BS. Differently from existing works on RIS-aided localization [16], which consider the presence of the direct LoS link between the BS and the UEs, we assume that such link is obstructed due to the presence of an obstacle, represented in the figure by the green solid rectangle. In our scenario, connectivity is granted by leveraging a set of $N$ RIS deployed on the bottom wall.

In our analysis a 2D model of the environment is assumed, where BS, UEs and RISs lie on the same plane. The design parameter of the $n$-th RIS is simply its electronic steering angle $\delta_n$ (or electronic rotation) as shown in Figure 1(a). A RIS electronically rotated by $\delta_n$ behaves as an anomalous mirror which appears to be rotated by an angle $\delta_n$ with respect to its mechanical orientation. Thus, at any time instant, the RIS-aided SRE can be configured by properly selecting the vector of the RIS rotation angles $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_N]^{\mathsf{T}}$ maximizing some performance parameters, such as the overall network throughput or the received signal strength.

When the UE position is exactly known, simple RIS-based strategies can be put in place to optimize the SRE. For instance, [4] proposes to electronically rotate RISs so that UEs and RISs are associated through a bijective map, that is, each RIS points towards a different UE receiver. On the other hand, if UE are moving along unknown trajectories, it is necessary to implement a UE localization strategy, sharing the available resources between the localization and the communication tasks. Localization must be periodically refreshed to update RIS rotation values with a rate high enough to track UE positional change. To manage the above dynamic scenario, we propose the time division strategy described in Section II-B.

### B. Frame structure and throughput optimization

A time division transmission scheme is assumed where each time frame hosts both sensing and communication tasks in separate slots. Figure 2 shows the structure of the frame, while Table I lists, among others, the frame parameters. The frame has duration $T_F$ and it is composed of a preamble time $T_P$, an actuation time $T_A$ and a communication time $T_C$, where:

- during the preamble the system estimates the positions of the $K$ users through uplink communication and the use of $N_s$ RIS dedicated to sensing;
- during the actuation time, the BS processes the estimated UEs positions and rotates $N_c$ RISs devoted to communication;
- during the communication time (downlink), the BS sends $K$ data streams to the UEs assisted by $N_c$ RISs.

The frame efficiency, $\eta_F$, is defined here as the time fraction dedicated to data communication, i.e.

$$\eta_F = \frac{T_C}{T_F} = 1 - \frac{T_P + T_A}{T_F}. \tag{1}$$

During the communication time the spectral efficiency achieved by the $k$-th UE is

$$\rho_k = \log_2\left(1 + \mathrm{SINR}_k(\eta_F)\right), \tag{2}$$

where the signal-to-interference plus noise ratio, $\mathrm{SINR}_k$ (13), is a function of the frame efficiency $\eta_F$, due to the interdependence between UE localization error, RISs configuration and achieved SINR.

For example, given the frame duration $T_F$, if we increase the communication time, $T_C$, we reduce the overhead required by UE position estimation, namely the preamble time $T_P$. This leads to a reduced localization accuracy and, thus, to suboptimal RIS configuration. If RISs are not correctly rotated they do not efficiently convey the signal energy towards the UEs, with detrimental effects on the $\mathrm{SINR}_k$. Conversely, by increasing $T_P$ we improve the UE localization accuracy and the achieved $\mathrm{SINR}_k$ but we also reduce the time dedicated to communication.

For a signal with bandwidth $B$, the average throughput, $T_k$, achieved by the $k$-th UE can be written as

$$\begin{aligned} T_k &= \eta_F \rho_k B & (3)\\ &= \eta_F B \log_2\left(1 + \mathrm{SINR}_k(\eta_F)\right). & (4) \end{aligned}$$

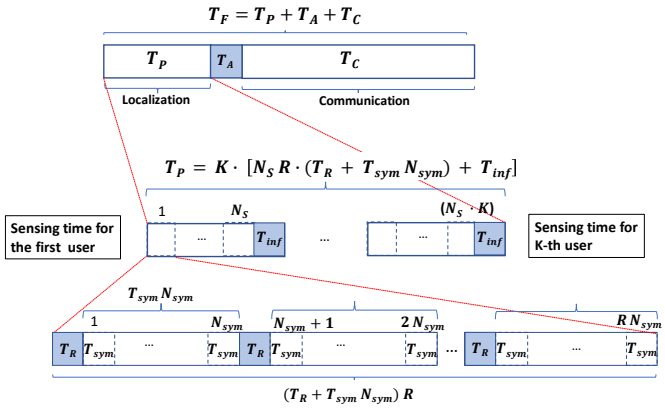


Fig. 2. Structure of a localization and communication time frame. The frame is constituted by a preamble time of duration $T_P$, devoted to the estimation of the UEs' positions, an actuation time of duration $T_A$, during which the estimated positions are elaborated by the BS and the RISs are optimally rotated, and a communication time $T_C$ during which the $K$ users simultaneously communicate in a spatially multiplexing mode. The parameters are defined in Table I.

TABLE I
MAIN SYSTEM PARAMETERS AND CORRESPONDING VALUES USED IN THE INTEGRATED LOCALIZATION AND COMMUNICATION TEST-BED

| Symbol | Description | Value |
|---|---|---|
| $f_0$ | Carrier frequency | 60 [GHz] |
| B | Bandwidth | 240 [MHz] |
| $\alpha$ | Roll-off | 0.25 |
| $T_{\text{sym}}$ | Symbol Time | 5.21 [ns] |
| $K$ | Number of users | 2 |
| $N_s$ | Number of RISs for sensing | 3 |
| $N_c$ | Number of RISs for communication | 2 |
| $T_R$ | RIS rotation time | 100 [$\mu$s] |
| $T_F$ | Frame time | 200 [ms] |
| $T_P$ | Preamble time | Eqn. (5) |
| $T_A$ | Actuation time | 1 [$\mu$s] |
| $T_C$ | Communication Time | $T_C = T_F - T_P - T_A$ |
| $R$ | Number of RIS rotations | 5 to 120 |
| $N_{\text{sym}}$ | Number of symbols per rotation | 250 |
| $T_{\text{inf}}$ | Inference time | Table V |
| $P_T$ | Transmit power | 20 [dBm] |
| $S_n$ | Power spectral density of noise | -174 [dBm/Hz] |

The maximum throughput is achieved for a certain optimal value of $\eta_F$ which trades-off communication performance and localization accuracy.

### C. Preamble and communication time slots

The preamble time $T_P$ is divided into $K$ uplink sensing slots, one for each UE. During the $k$-th sensing slot, the BS receives and processes the signal transmitted by the $k$-th UE and reflected by the $N_s$ RISs dedicated to sensing. UE localization is performed letting each of the above mentioned $N_s$ RISs assume, sequentially, $R$ electronic rotation angles, i.e., each RIS is reconfigured $R$ times during the $k$-th sensing slot. Each RIS configuration requires a time $T_R$ to be performed and it is held until the $k$-th UE has transmitted $N_{\text{sym}}$ symbols of duration $T_{\text{sym}}$ each. Thus, a RIS rotation is performed every $T_R + N_{\text{sym}} T_{\text{sym}}$ seconds.

Overall, the BS collects $N_s R N_{\text{sym}}$ samples of the signal transmitted by the $k$-th UE and elaborates them by using DL techniques, in order to infer the position of the $k$-th UE. The inference time is denoted by $T_{\text{inf}}$.

Therefore, the preamble time can be written as

$$T_P = K[N_s R(T_R + T_{\text{sym}} N_{\text{sym}}) + T_{\text{inf}}]\,. \tag{5}$$

The communication time $T_C$ is the frame portion devoted to the communication between UEs and BS. The UEs communicate simultaneously to the BS in space division access. $T_C$ is specified for downlink communication assuming Time Duplex Mode access.

We assume the wireless channel is static during the frame time. As $T_F$ is the interval between two consecutive estimates of the UE locations, it must be properly shaped according to the velocity of the mobile UEs as well as to the beamwidth of the radiation pattern generated by the RIS, which, in turn, is a function of the RIS area $A$. Specifically, each RIS generates a radiation pattern whose half power beamwidth (HPBW) is approximately equal to $\lambda/\sqrt{A}$, hence, larger surfaces generate narrower beams. We can determine the maximum interval between consecutive estimates of UE position as the time it takes to a UE to travel over an angular distance equal to half of the HPBW. Assuming that the UE is moving with velocity $v$ over an azimuthal path at distance $d$ from the RIS, the maximum frame time is given by

$$T_F = \frac{d\lambda}{2v\sqrt{A}}, \tag{6}$$

where $\lambda$ represents the signal wavelength. Considering the trajectory and speed of the $k$-th UE, both localization and communication tasks must occur within a precise time interval before the $k$-th UE moves to its next location. For instance, when the $k$-th UE has an average speed of $v = 0.5$ m/s, and follows the specific trajectory, for a $\lambda$ = 5 mm, distance $d = 4$ m and a RIS area $A$ = 100 cm$^2$ the localization process must be implemented within a time interval of $T_F$ = 200 ms allowing the sensing and communication procedure to conclude before the UE significantly changes its position.

## III. RIS-BASED SENSING AND COMMUNICATION CHANNEL MODEL

We now provide details on the wireless channel model we use for communication and localization. Figure 1(a), depicts the geometry of the downlink channel where the BS is simultaneously transmitting to the $K$ UEs. Instead, in Figure 1(b), we focus on localization, and consider the uplink scenario where the UEs are transmitting and the BS is receiving. In both cases, we assume that the UEs are not in LoS with the BS, but the signal is reflected by a set of dedicated RIS, as depicted in Figure 1.

### A. RIS model

The RIS model adopted in this work is the one reported in [17], and holds in the far-field regime. In such model, the RISs have a square shape and are composed of square-shape

meta-atoms [18] with a side-length $\Delta$ wavelength, arranged in a $L \times L$ grid of area

$$A = L^2 \Delta^2 \lambda^2. \tag{7}$$

We assume that the signal power collected by a RIS is proportional to $A \cos \phi^{(1)}$ for $\phi^{(1)} \in [-\pi/2, \pi/2]$, and zero otherwise, where $\phi^{(1)}$ is the angle of arrival (AoA) of the signal impinging on the RIS. Similarly, we assume that the RIS radiated power is proportional to $|\rho_{\mathrm{RIS}}|^2 A \cos \phi^{(2)}$, for $\phi^{(2)} \in [-\pi/2, \pi/2]$ and zero otherwise, where $\phi^{(2)}$ is the angle of departure (AoD) of the scattered field and $\rho_{\mathrm{RIS}}$ is the reflection efficiency of the RIS [4]. Both angles $\phi^{(1)}$ and $\phi^{(2)}$ are measured with respect to the normal to its physical surface. The above model holds in a 2D scenario where dependency on the elevation angle can be neglected.

Every RIS can be configured by properly setting the phase-shift that each meta-atom applies to the impinging signal. In this work we employ the model in [4], [19], [20] where the phase shift $\theta_{n,\ell,\ell'}$ applied by the meta-atom at position $\ell, \ell'$ in RIS $n$, obeys to the linear equation

$$\theta_{n,\ell,\ell'} = 2\pi \ell \Delta g_n + \psi_n \,. \tag{8}$$

for all $\ell' = 1 \ldots, L$, where $\psi_n$ and $g_n$ are parameters. The phase-shift setting in (8) allows the $n$-th RIS to macroscopically act as an anomalous mirror able to steer a signal, arriving from direction $\phi_n^{(1)}$, to an arbitrary direction, $\phi^{(2)}$, that can be decided by setting the parameter $g_n$ according to [4]

$$g_n = \sin(\phi_n^{(1)}) - \sin(\phi_n^{(2)}) \,. \tag{9}$$

The RIS then appears as electronically rotated by angle $\delta_n = (\phi_n^{(1)} + \phi_n^{(2)})/2$.

### *B. Communication phase (downlink)*

In the following, we assume the BS and the UE be equipped with ULAs composed of $M^{\mathrm{BS}}$ and $M^{\mathrm{UE}}$ isotropic antennas, respectively, spaced by $\Delta$ wavelengths. We also assume that the UE ULAs can only perform analog beamforming, since they are supposed to have limited hardware complexity.

The BS generates $K$ streams of information symbols to be transmitted to the $K$ UEs, denoted by the random complex vector $\mathbf{s}^{\mathrm{BS}} = [s_1^{\mathrm{BS}}, \ldots, s_K^{\mathrm{BS}}]^{\mathsf{T}}$ having zero-mean and covariance $\mathbb{E}[\mathbf{s}^{\mathrm{BS}} \mathbf{s}^{\mathrm{BS},\mathsf{H}}] = \mathbf{I}_K$. Then, the signal transmitted by the BS ULA can be described by the $M^{\mathrm{BS}} \times 1$ vector

$$\mathbf{t} = \mathbf{\Gamma} \mathbf{s}^{\mathrm{BS}} \,, \tag{10}$$

where $\mathbf{\Gamma}$ is a precoding matrix of size $M^{\mathrm{BS}} \times K$. The BS transmit power cannot exceed $\mathbb{E}[\|\mathbf{t}\|^2] = P^{\mathrm{BS}}$ and the signal received by the $k$-th UE can be described as

$$z_k^{\mathrm{UE}} = \mathbf{f}_k^{\mathsf{H}} \widetilde{\mathbf{H}}_k \mathbf{t} + \eta_k^{\mathrm{UE}} \,, \tag{11}$$

where $\mathbf{f}_k$ is the beamforming vector of the $k$-th UE, $\widetilde{\mathbf{H}}_k$ is the $M^{\mathrm{UE}} \times M^{\mathrm{BS}}$ channel matrix connecting the BS ULA to the $k$-th UE ULA, $\mathbf{t}$ is given by (10), and $\eta_k^{\mathrm{UE}}$ represents thermal noise, modeled as a complex random variable with distribution $\mathcal{N}(0, N_0 B)$ being $N_0$ the thermal noise power spectral density and $B$ the signal bandwidth.

As will be detailed in Section III-D, the matrix $\widetilde{\mathbf{H}}_k$ depends on the system geometry, including the position of the BS and of the RISs and, more importantly, the RIS setting and the (unknown) UE position $\mathbf{p}_k \in \mathcal{L}$, where $\mathcal{L}$ is the shaded area depicted in Figure 1. In general we can explicit this dependency by writing $\widetilde{\mathbf{H}}_k \equiv \widetilde{\mathbf{H}}_k(\mathbf{p}_k, \boldsymbol{\delta}, \boldsymbol{\psi})$ where $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_N]^{\mathsf{T}}$ is the vector of RIS rotation angles defined in Section III-A and $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_N]^{\mathsf{T}}$ are the coefficients appearing in (8). Note that $\boldsymbol{\delta}$ and $\boldsymbol{\psi}$ need to be properly chosen so as to reflect the BS signal towards the UE. This, in turn, requires a reliable estimation $\widehat{\mathbf{p}}_k$ of the true UE location $\mathbf{p}_k$.

In a practical communication system the matrix $\widetilde{\mathbf{H}}_k$ has to be estimated and employed so as to compensate for the channel effect. Estimates of the matrices $\widetilde{\mathbf{H}}_k$, $k = 1, \ldots, K$ are used at the BS as a component to build the precoding matrix $\mathbf{\Gamma}$. Among many possible choices for $\mathbf{\Gamma}$, we consider the zero-forcing solution proposed in [4] given by

$$\mathbf{\Gamma} = \sqrt{P^{\mathrm{BS}}} \frac{\widehat{\mathbf{H}}^{+}}{\|\widehat{\mathbf{H}}^{+}\|_{\mathrm{F}}} \,, \tag{12}$$

that holds under the condition $\min(M^{\mathrm{BS}}, N) \geq K$. The matrix $\widehat{\mathbf{H}}$ has size $K \times M^{\mathrm{BS}}$ and its $k$-th row is an estimate of the row vector $\mathbf{f}_k^{\mathsf{H}} \widetilde{\mathbf{H}}_k$ appearing in (11), $k = 1, \ldots, K$. Also, it should be noted that $\widehat{\mathbf{H}}$ depends on the RIS setting $\boldsymbol{\delta}$ and $\boldsymbol{\psi}$ which, in practice, need to be selected according to estimates $\hat{\mathbf{p}}_k$ of the UE locations.

With perfect channel estimation (i.e., when the $k$-th row of $\widehat{\mathbf{H}}$ equals $\mathbf{f}_k^{\mathsf{H}} \widetilde{\mathbf{H}}_k$) the precoder (12) nulls-out the multiuser interference at the UEs and, in general, allows a simple analytic representation of the SINR and of the achievable rate. Instead, with unperfect channel estimation the precoder in (12) leads to the achievable spectral efficiency in (2) where $\mathrm{SINR}_k$ is defined as

$$\mathrm{SINR}_k = \frac{|w_{k,k}|^2}{\sigma^2 + \sum_{j \neq k} |w_{k,j}|^2} \tag{13}$$

the vector $\mathbf{w}_k$ is given by $\mathbf{w}_k = \mathbf{f}_k^{\mathsf{H}} \widetilde{\mathbf{H}}_k \mathbf{\Gamma}$ and $w_{k,j}$ is its $j$-th component.

### *C. Localization phase (uplink)*

In the localization phase the $k$-th UE, located at position $\mathbf{p}_k$, employs a single antenna and emits a sequence of zero-mean random complex symbols, which are known at the receiver and denoted by the random variables, $s_k^{\mathrm{UE}}$ whose power is $\mathbb{E}[|s_k^{\mathrm{UE}}|^2] = P_k^{\mathrm{UE}}$. When the signal transmitted by the $k$-th UE is reflected on a RIS dedicated to localization, say the $n$-th, the signal received by the BS ULA is represented by the vector

$$\mathbf{z}_{k,n}^{\mathrm{BS}} = \widetilde{\mathbf{h}}_{k,n} s_k^{\mathrm{UE}} + \boldsymbol{\eta}_{k,n}^{\mathrm{BS}} \tag{14}$$

of size $M^{\mathrm{BS}}$, where $\widetilde{\mathbf{h}}_{k,n}$ is the channel connecting the UE antenna to the BS ULA through RIS $n$ and $\boldsymbol{\eta}_{k,n}^{\mathrm{BS}}$ is a complex random vector representing thermal noise, with distribution $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{M^{\mathrm{BS}}})$. Note that the vector $\widetilde{\mathbf{h}}_{k,n}$ depends on the $n$-th RIS setting and on the $k$-th UE position, so that we can write $\mathbf{z}_{k,n}^{\mathrm{BS}} \equiv \mathbf{z}_{k,n}^{\mathrm{BS}}(\mathbf{p}_k, \delta_n, \psi_n)$.

### D. Wireless channel model

We now specify the structure of the matrices $\widetilde{\mathbf{H}}_k$ in (11) and of the vectors $\widetilde{\mathbf{h}}_k$ in (14), $k = 1, \ldots, K$. In the scenario depicted in Fig. 1 the LoS link between the BS and the UEs is blocked by the presence of an obstacle and signal propagation is granted by the presence of the RISs. Therefore, the wireless channel connecting BS and the $k$-th UE, denoted by the matrix $\widetilde{\mathbf{H}}_k$ introduced in (11), is 2-hop and can be specified as

$$\widetilde{\mathbf{H}}_k = \sum_{n=1}^{N} \mathbf{H}_{k,n}^{(2)} \mathbf{\Theta}_n \mathbf{H}_n^{(1)}, \tag{15}$$

where:

- $\mathbf{H}_n^{(1)}$ is the $L^2 \times M^{\mathrm{BS}}$ channel matrix connecting the BS to the $n$-th RIS;
- $\mathbf{\Theta}_n$ is the diagonal matrix containing the phase shifts $\mathrm{e}^{\mathrm{j}\theta_{n,\ell,\ell'}}$ defined in (8);
- $\mathbf{H}_{k,n}^{(2)}$ is the $M^{\mathrm{UE}} \times L^2$ channel matrix connecting the $n$-th RIS to the $k$-th UE.

The matrices $\mathbf{H}_n^{(1)}$ and $\mathbf{H}_{k,n}^{(2)}$ are described by the superposition of a line-of-sight (LoS) path and $S$ non-LoS paths, each of them resulting from the reflection or scattering of the signal on an obstacle. We refer to [4, Section II.C] for details.

Similarly, in the localization phase (Sect. III-C) the channel vector $\widetilde{\mathbf{h}}_k$ introduced in (14) can be written as

$$\widetilde{\mathbf{h}}_{k,n} = {\mathbf{H}_n^{(1)}}^{\mathsf{T}} \mathbf{\Theta}_n \mathbf{h}_{k,n}^{(2)}, \tag{16}$$

where ${\mathbf{h}_{k,n}^{(2)}}^{\mathsf{T}}$ is the first row of the matrix $\mathbf{H}_{k,n}^{(2)}$. Note that in both (15) and (16) the dependency on $\delta_n$ and $\psi_n$ is hidden in the matrix $\mathbf{\Theta}$ whose elements are given by (8) and (9).

## IV. Localization system

In this section, we describe the algorithms and models used for UE localization. The proposed localization algorithms process the samples, $\mathbf{z}_{k,n}^{\mathrm{BS}}$, of the signals transmitted by the UEs, reflected by the RISs and received at BS, as specified in Section III-C, and obtain the estimated $k$-th UE position, $\widehat{\mathbf{p}}_k$, by classification. We discuss two DL models, namely convolutional neural network (CNN) and Vision Transformers (ViT) as well as training and validation processes. Transformer encoders have been viewed as viable alternatives to classical convolutional filtering in visual or image-based recognition/classification tasks [21]. CNN and ViT algorithms tailored for localization are trained off-line using labeled received signal at landmarks positions samples obtained from varying RISs, rotations and symbols.

The proposed algorithms have been applied to our scenario by conducting localization tests with different UE trajectories and indoor mobility patterns. The $k$-th user is assumed to move from a predefined position $\mathbf{p}_0$ and follows a trajectory $l$ (see Figure 4) with speed defined as $\mathbf{u} = [v, w]$ including linear and angular velocity.

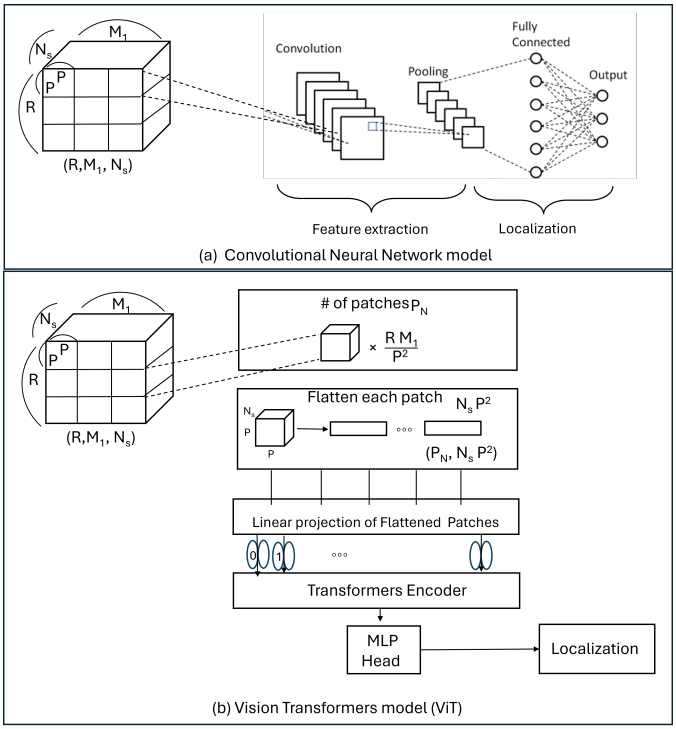


Fig. 3. Localization algorithms and input data structure; (a) CNN structure including input data structure, convolutional layers and output for the localization problem, (b) ViT structure replacing the convolutional layers with TF for the localization problem.

### A. Convolutional Neural Networks and Vision Transformers

A deep neural network (DNN) composed of $Q$ layers maps the signals transmitted by the $k$-th UE and observed at the BS into the estimated position

$$\widehat{\mathbf{p}}_k = \mathcal{H}(\mathbf{W}; \mathbf{Z}_k), \tag{17}$$

where $\mathbf{W} = \mathbf{W}^{(Q)}$ encapsulates the parameters of the DNN model, for all the $Q$ layers, and $\mathbf{Z}_k$

$$\begin{aligned} \mathbf{Z}_k \;=\; & [\mathbf{z}_{k,1}^{\mathrm{BS}}(\delta_{1,1}), \ldots, \mathbf{z}_{k,1}^{\mathrm{BS}}(\delta_{1,R}), \ldots \\ & \ldots, \mathbf{z}_{k,N_s}^{\mathrm{BS}}(\delta_{N_s,1}), \ldots, \mathbf{z}_{k,N_s}^{\mathrm{BS}}(\delta_{N_s,R})], \end{aligned} \tag{18}$$

is the matrix containing the signal samples, i.e. the input data to the DL models and algorithms.

The sequence of $R$ rotations of $n$-th RIS, $[\delta_{n,1}, ..., \delta_{n,r}, ..., \delta_{n,R}]$ is uniformly distributed between two limiting angles depending on the system geometry.

Figure 3 reports the DNN structures chosen to support the classification of UE position in the monitored area. Figure 3(a) depicts the CNN structure where the input signal samples are rearranged in a 3D matrix. The figure also shows the convolutional layers, pooling, and output fully connected layers. Figure 3(b) presents the ViT structure. Here transformers encoders, tailored for 3D input data [22], are employed instead of CNN layers. As shown in the figures convolutional layers and transformers encoders have different input data structures due to their different design principles. In particular, CNN operates on feature maps of the input data while ViT operates on a flattened sequence of the data patches. The parameters

TABLE II
CONVOLUTIONAL NEURAL NETWORK SETTING

| | | |
|---|---|---|
| Input size | $M^{\mathrm{BS}} \times R \times N_s$ | $8 \times \{50, 70, 100, 120\} \times 3$ |
| First layer | convolution+Relu+pooling | 8 filters with size ($3 \times 3$) |
| Second layer | convolution+Relu+pooling | 16 filters with size ($3 \times 3$) |
| Third layer | convolution+Relu+pooling | 64 filters with size ($3 \times 3$) |
| Output layer | dropout+fully connected | $\hat{\mathbf{p}}_k$ |

TABLE III
SOME PARAMETERS OF VIT

| | |
|---|---|
| Input size | $M^{\mathrm{BS}} \times R \times N_s$ |
| learning rate | 0.001 |
| Heads | 2 |
| Transformer layers | 3 |
| Patch size | $6 \times 6$ |
| Embedded dimension | 64 |

of the considered models $\mathcal{H}(\cdot)$ in (17), including the layers and the adopted optimizer, are detailed in Tables II and III for CNN and for ViT, respectively.

The DNN models are trained to classify $P$ UE marked positions (landmarks) to uniformly cover the monitored area. Landmarks are distributed on a 2D regular grid. DNN model learning is implemented via supervised methods. The localization accuracy is evaluated by assuming the UE moving on a random trajectory $l$. In order to evaluate the UE estimation performance, we replace the $Q$ layers in CNN with transformers (TF). For example the input data $\mathbf{Z}_k$ for the ViT algorithm is split into $P_N = \frac{RM^{\mathrm{BS}}}{P^2}$ patches. The sequence of linear embeddings (tokens) of the 2D patches ($\mathbf{X}_p \in \mathbb{R}^{P_N \times N_s P^2}$) is fed into repeated standard TF layers to model the global relations for classification (localization) [21].

### B. Online UE trajectory estimation

Figure 4, shows an example of a trajectory performed by a UE while moving inside the monitoring area. The trajectory parameters are shown in Table IV-B. The UE travels from the predefined position $\mathbf{p}_0$ to a specific destination by following a trajectory with linear/angular velocity defined by $\mathbf{u} = [v, w]$ in each $T_s$.

We consider the $k$-th UE following a trajectory $\mathbf{p}_k(t) = [x_k(t)\ y_k(t)\ \theta_k(t)]^T$ represented as a sequence of $p$ positions visited by the UE as travelling towards its destination. The set of positions that the UE assumes during its trajectory is computed iteratively as:

$$\begin{aligned} x_k(t) &= x_k(t-T_s) - \frac{v}{w}\sin\theta_k + \frac{v}{w}\sin(\theta_k + wT_s), \\ y_k(t) &= y_k(t-T_s) + \frac{v}{w}\cos\theta_k - \frac{v}{w}\cos(\theta_k + wT_s) \\ \theta_k(t) &= \theta_k(t-T_s) + wT_s. \end{aligned}$$

We consider the following constraints for the UE's trajectory $\{\mathbf{p}_k(t)\}$, also summarized in Table IV. First, the UE initial position is fixed to $\mathbf{p}_0 = (-1.2, 4, -\pi/5)$. Second, the maximum distance that the UE can travel in each time interval is $\mid \mathbf{p}_k(t+1) - \mathbf{p}_k(t) \mid \leq vT_s$, where $v$ represents the linear velocity of the UE. Finally, the BS should complete both localization and communication processes within the time frame/slot $T_s \leq T_F$.

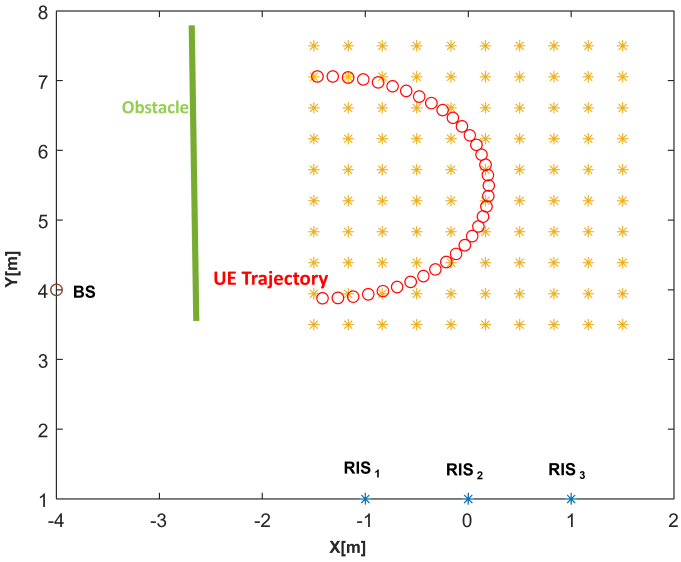


Fig. 4. RIS-aided UE localization scenario. $N_s$=3 RISs are exploited for the localization located at positions (-1 1), (0, 1), (1,1), and BS located at position (-4, 4), red circles shows the UE trajectory and yellow stars presents the training landmarks for the UE trajectory estimation.

TABLE IV
TRAJECTORY PARAMETERS

| | |
|---|---|
| Initial position | $\mathbf{p}_0 = (-1.2, 4, -\pi/5)$ |
| $T_s$ (time interval in seconds) | 0.3 |
| $v$ (linear velocity (m/s)) | 0.5 |
| $w$ (angular velocity ) | $\pi/10$ |
| $p$ (no. of visited positions) | 35 |
| $\theta$ (direction of UE on motion) | $[-\pi, \pi]$ |

## V. RESULTS

In this section we present the results obtained by simulating the RIS-aided ISAC scenario in Figure 1. We deployed a total of $N = 5$ RISs within the designated area. Among these, $N_s$=3 RISs are used for the localization task, while the remaining two serve for communication. In our simulation setup the UEs are equipped with a single antenna ($M^{\mathrm{UE}} = 1$) and are distributed within the shaded area $\mathcal{L}$ of Figure 1. The BS–RIS and RIS–UE links are assumed to be in LoS while no direct link between the BS and the UE is available. The RIS–UE propagation channel follows the model in (15). Moreover, $S = 3$ isotropic scatterers [23] are randomly positioned in $\mathcal{L}$. They are characterized by a reflection coefficient whose square magnitude is -20 dB. All the links also experience shadowing effects that we assume to be log-normally distributed with variance $\sigma_{\mathrm{sh}} = 2\,\mathrm{dB}$. Finally, the signal carrier frequency is set to $f_0 = 60\,\mathrm{GHz}$ (i.e. $\lambda = 5\,\mathrm{mm}$).

As for the UE localization, we consider the indoor test environment depicted in Figure 4, which represents a room of size $8\,\mathrm{m} \times 8\,\mathrm{m}$ with a BS located at coordinates $(-4, 4)$. The room is also equipped with $N_s$=3 RISs with area $A$=100 $\mathrm{cm}^2$, equally spaced along a wall coinciding with the $x$ axis. First,

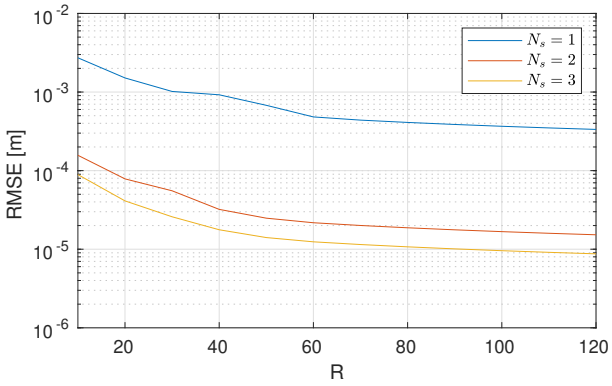


Fig. 5. CRLB for the scenario depicted in Figure 4, plotted versus the number of RIS rotations, as the number of RIS, $N_s$, varies.

we conducted a pre-deployment analysis based on the Cramer-Rao Lower Bound (CRLB) to verify the proper deployment of the RIS in the monitoring area, and to ensure a certain level of accuracy. Subsequently, we evaluated the localization performance in terms of RMSE, considering both RIS deployment and parameter settings such as electronic rotations of RIS, BS antenna numbers, and hyper-parameters used by CNN and ViT localization algorithms. The performance of CNN and ViT is analyzed in terms of positioning error, inference time and exchange parameters in training to identify the best performance based on available resource for sensing and communication in specific applications. Finally, we present the integrated localization and communication results in terms of network throughput and discuss the trade-off between localization accuracy and frame efficiency to maximize the throughput.

### A. RIS deployment and CRLB analysis

Figure 5 shows the CRLB bound to the localization accuracy plotted versus the number of RIS electronic rotations, $R$, as the number of RISs dedicated to localization, $N_s$, varies for $P^{\mathrm{UE}}$ =20 dBm, $N_{\mathrm{sym}} = 250$ samples per RIS rotation, and $M^{\mathrm{BS}} = 8$ antennas at the BS. The signal bandwidth is set to $B$ =1 GHz and the thermal noise power is $\sigma^2 = -84$ dBm.

The CRLB is computed as described in Appendix VI, and refer to the scenario depicted in Figure 4.

As expected, when $N_s$ increases the RMSE decreases since a larger number of signal samples are available. In particular, the RMSE decreases by about one order of magnitude when 2 RISs are employed instead of one. However, by adding another RIS to the system ($N_s = 3$) the RMSE does not significantly improve. We also observe that the RMSE significantly improves as the number of RIS rotations, $R$, increases. However for $R > 60$ it shows a floor. We therefore conclude that $N_s = 3$ RIS and $R = 60$ rotations are sufficent for providing excellent localization performance in the considered scenario. It's noteworthy that there exists a notable disparity between the RMSE obtained from the CRLB and that derived from our proposed algorithms for UE localization. This disparity arises from several factors. Firstly, the parameter setup for DL algorithms, such as number of convolutional layers, and filter number and size in CNN, or TF layers, and patch size in ViT, significantly impacts the accuracy of localization. Secondly, there is a discretization error floor since we treated the localization problem as a classification task and implemented the model training on a discrete grid. Therefore the achieved RMSE is strongly affected by the grid size. Lastly, the effects of multipath and shadowing are not accounted for in the CRLB calculation, leading to further errors in UE location estimation.

### B. Localization results

Figure 4 highlights the scenario considered for the localization process. LoS path is blocked by an obstacle (green), and the BS receives the signal transmitted by the UE and reflected from $N_s = 3$ RISs located at coordinates (-1 1), (0, 1), (1,1). The BS is equipped with a uniform linear array of $M^{\mathrm{BS}} = 8$ antennas while the RISs can implement up to $R$ electronic rotations between two consecutive localization updates. The UE moves on a trajectory which is generated as described in the Sect. IV-B: a trajectory example is also highlighted by red markers in the Figure 4 and the main parameters used to setup the UE movements inside the monitored area are defined in the Table IV. The training is done on a 2D regular grid comprising $P = 50$ landmarks, each spaced 30 cm apart from its consecutive counterpart.

The CNN and ViT algorithms main configuration parameters are defined in Table II and Table III, respectively. In what follows, the localization performance is verified for different numbers of RISs and their corresponding electronic rotations, namely $R = [5, 20, 40, 60, 80, 100, 120]$. To assess the UE positioning accuracy and communication performance, the CNN and ViT algorithms are also trained with various parameter sets, including samples $N_{\mathrm{sym}}$, and variable rotations $R$ as above. Note that the number of electronic rotations $R$ required to get an accurate scan of the UE region depends on the beamwidth of the RIS, which, in turn, is a function of its area $A$ in (7), here set to $A$ = 100 cm$^2$ [15]. However, since the surface area $A$ is inversely proportional to the HPBW of the RIS radiation pattern, an increase of $A$ requires a larger number of rotation angles $R$ to cover the monitored area and, hence, a higher computational effort for processing the input of the DL algorithms.

In Figure 6, the Root Mean Square Error (RMSE) (m) results using CNN and ViT algorithms are presented for various numbers of $R$. The localization error decreases with increasing values of $R$ when using CNN, reaching an RMSE value of approximately 0.65 m with $R = 120$ and $N_{sym} = 250$. The ViT algorithm outperforms CNN in almost all the explored cases in exchange for larger computational cost, as clarified in the following. The ViT model is particularly well performing as the observed RMSE reduces to 0.4 m.

Table V compares the inference time and the model size (model footprint), in terms of number of trainable parameters, using CNN and ViT algorithms and different numbers of $R$. In particular, the inference time measurements presented in Table V are obtained using both CNN and ViT models deployed on two devices with different computational capabilities and modelling low-power to high performance BS servers. We

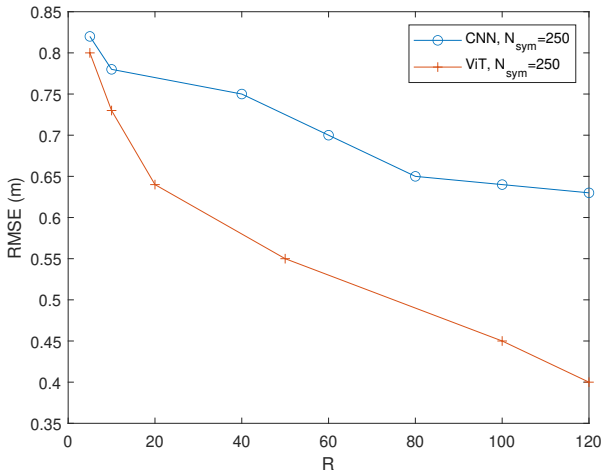


Fig. 6. Root Mean Square Error (RMSE) result of UE localization Using CNN and vision transformers (ViT) for the same scenario presented in figure 4. The RMSE result for $N_{sym}$=250 (i.e., sample) and number of electronic rotations R.

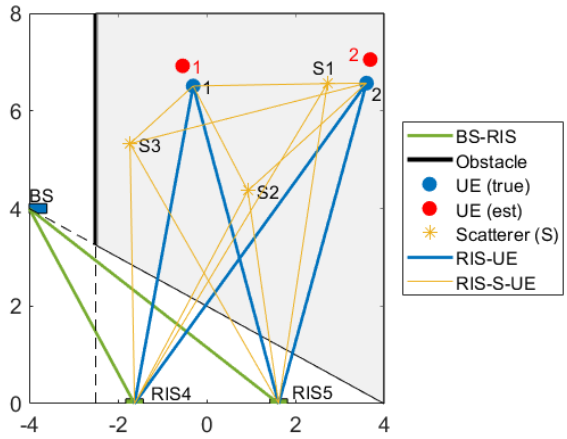


Fig. 7. Smart radio environment simulator example for communication performance assessment: 2 UEs are moving randomly in the monitored area and communicating with the BS.

considered a typical industrial IoT device (IoT) equipped with a dedicated low-power Tensor Processing Unit (TPU) and a higher performance computing system (HPC). The IoT device operates on 12-core ARM CPU and is equipped with a low-power NVIDIA Maxwell GPU (Jetson Nano IoT device model). The HPC utilizes a high performance 24 core processor (AMD Ryzen TR 3960X).

As presented in the result, ViT algorithm is more computational demanding in terms of inference time with respect to CNN using both IoT and HPC devices. Regarding the specific HPC hardware, the inference time decreases by approximately $80\%$ for CNN and $30\%$ for ViT, respectively.

Hence, there exists a trade-off among selecting the DL algorithms, their corresponding performance in terms of localization accuracy, and the computational capacity of available resources. In what follows, we discuss the impact of the proposed localization system on communication, throughput and frame efficiency performance. We also provide some general guidelines for efficient ISAC operations.

### C. Integrated localization and communication results

A RIS-aided communication network was simulated assuming $K = 2$ users served by as many RIS and the frame structure described in Sec. II-B. During the preamble time $(T_p)$, the UEs' locations were estimated by either CNN or ViT algorithm. Moreover, to quantify the impact of the localization inference time on frame efficiency, we analyzed the results obtained with IoT and HPC hardware (HW) setups, as described in Sec. IV, which model different BS designs. The SRE and frame parameter values used in the simulation are listed in Table I. The frame time value equal to 200 ms has been calculated from (6) assuming $d = 4$ m, $v = 0.5$ m/s and RIS area $A = 100$ cm$^2$, corresponding to a HPBW of about 0.05 rad. The resulting preamble time and its components, i.e. the required time for sensing the channel (red) and the inference time (green) are shown in Fig. 8 as a function of

TABLE V
COMPARISONS WITH CLASSIC CNN-BASED MODELS AND TRANSFORMER-BASED MODELS.

A: Inference time (ms), $N_{sym}$=250

| R | 5 | 20 | 40 | 80 | 100 | 120 |
|---|---|---|---|---|---|---|
| CNN (IoT/HPC) | 5/5 | 13/6 | 20/7 | 37/10 | 48/11 | 60/12 |
| ViT (IoT/HPC) | 23/20 | 66/20 | 72/49 | 144/51 | 161/135 | 205/145 |

B: Parameters millions (M), $N_{sym}$=250

| R | 5 | 20 | 40 | 80 | 100 | 120 |
|---|---|---|---|---|---|---|
| CNN | 0.02 | 0.03 | 0.1 | 0.2 | 0.28 | 0.34 |
| ViT | 2 | 2.8 | 2.8 | 5.6 | 15 | 15.8 |

the number of rotations $(R)$ for the four tested combinations of DL algorithms (CNN, ViT) and HW setups (IoT, HPC). The values of the inference time have been taken from Table V. The sensing time is proportional to $R$, as shown in (5) and it includes the RIS rotation time $T_R$ and the time to transmit $N_{sym}$ symbols from each UE to the BS through the RIS. With the chosen values of $T_R$ (100 $\mu$s), $N_{sym}$ (250) and $T_{sym}$ (5.21 ns), the $T_R$ is dominant over $T_{sym}N_{sym}$. Hence, in principle, $T_{sym}N_{sym}$ could be increased, without affecting the preamble time.

The localization inference time increases with $R$ following a rate dependent on the DL algorithm and on the HW setup employed. If ViT is used, the inference time dominates the preamble time $T_P$. Moreover, ViT can be operated only with a relatively small number of RIS rotations, especially when deployed on a IoT device with limited computing capabilities.

The performance of the communication system was assessed by generating J = 1000 snapshots featuring $K = 2$ single antenna ($M^{\mathrm{UE}} = 1$) UEs randomly deployed within the shaded area of Fig. 7, which illustrates an example snapshot. The RIS-UEs channel is also characterized by $S = 3$ scatterers, randomly deployed in the same area, which produce

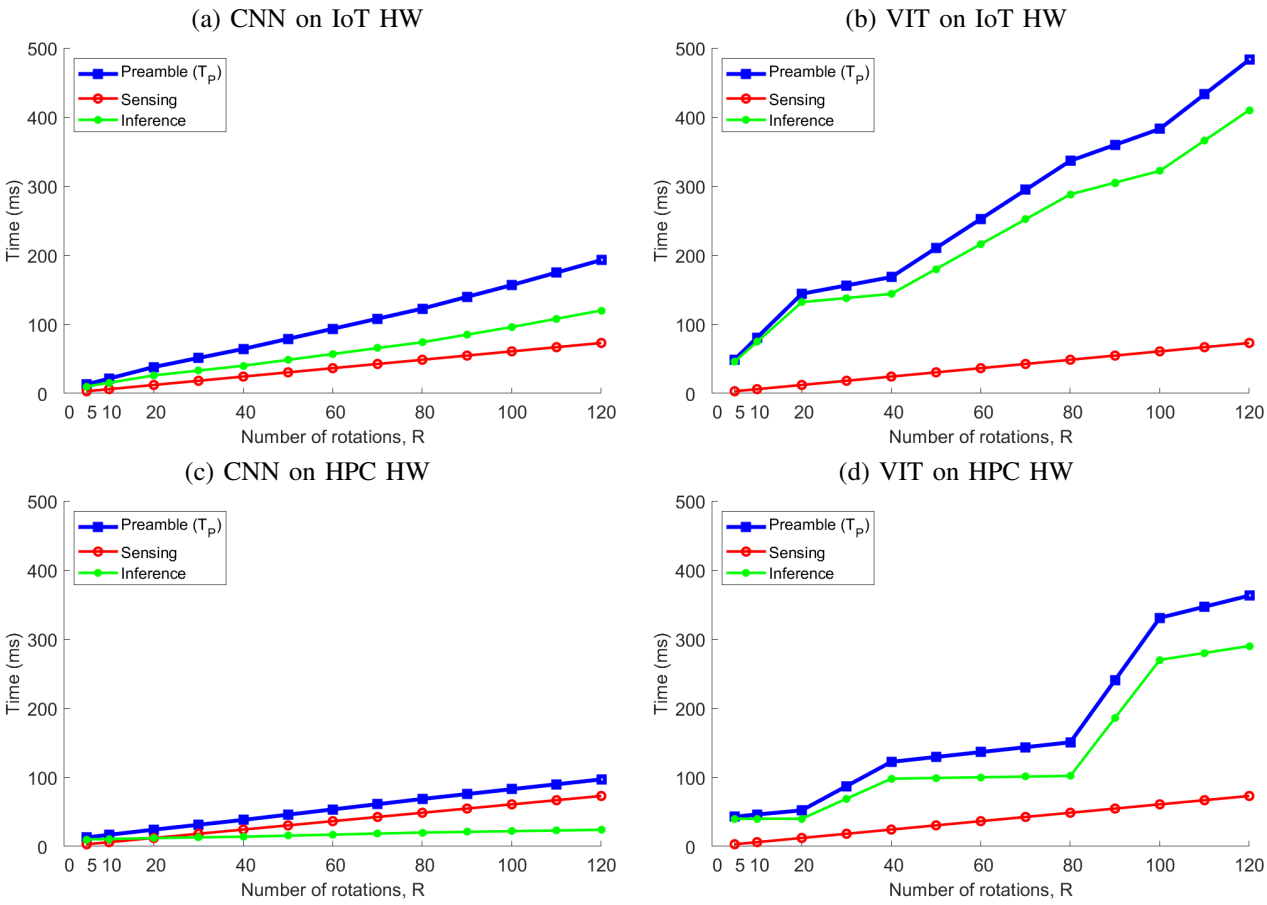


Fig. 8. Preamble time and its components as a function of the number of rotations when either CNN (left) or ViT (right) are implemented on two different types of HW to localize UEs.

interference signals received by the UEs. The above 1000 snapshots were generated for different values of the UE localization error sampled from the curves in Fig. 6 with $R$ ranging from 5 to 120 in steps of 10.

Fig. 9(a) shows the frame efficiency $\eta_F$ defined in (1) against the localization RMSE obtained by ViT and CNN algorithms under the two different HW setups. As expected from the previous analysis, high frame efficiencies $> 0.75$ can be obtained in exchange for increased localization errors ($> 0.6$ m), which, in turn, might deteriorate the system throughput.

The corresponding average system throughput defined in (4) is reported in Fig. 9(b). For $K$ UEs and J simulation runs, the average throughput is defined as:

$$\mathrm{T} = B \cdot \eta_F \cdot \frac{1}{\mathrm{J}} \sum_{j=1}^{\mathrm{J}} \sum_{k=1}^{K} \log_2 \left[1 + \mathrm{SINR}_k(j)\right] \qquad (19)$$

where $\mathrm{SINR}_k(j)$ defined in (13) for UE $k$ and snapshot $j$. The remaining quantities are defined in Table I.

The average throughput is shown in Fig. 9(b) as a function of the frame efficiency in (1) for the four combinations of DL algorithms and HW setups. ViT and CNN sample different intervals of the frame efficiency. For example, the CNN algorithm deployed on an IoT setup obtains an efficiency which ranges from $0.05$ to $0.85$ and throughput from $0.02$ Gbps up to $0.6$ Gbps. The same CNN model now deployed on HPC hardware provides a higher frame efficiency, between $0.5$ to $0.9$, and throughput from $0.4$ Gbps to $0.6$ Gbps. Considering both CNN and ViT algorithms, the throughput increases with the frame efficiency up to a maximum point above which the observed localization errors produce a degradation of the communication performance. The optimal operating point depends on the localization algorithm chosen and the computational capabilities of the device (IoT/HPC), which affect the inference time. For the considered scenario, the optimal point is observed at frame efficiencies of $0.75 - 0.87$ for ViT and $0.95 - 0.98$ for CNN. Using HPC HW does not turn into a significant gain on system throughput. In addition, the less resource demanding CNN-based localization algorithm awards very similar throughput performance as ViT, since it requires lower inference time for each localization update.

## VI. Conclusions

The paper proposed a novel approach for tackling the dual problem of UE localization and communication in a RIS-aided smart radio environment. The solution operates at 60 GHz frequency band while we considered a typical scenario where the Line-of-Sight (LoS) link between the UE and the multiple-antenna BS is blocked by obstacles. Multiple RISs are deployed to establish virtual links and support both localization and communication services. In the proposed indoor

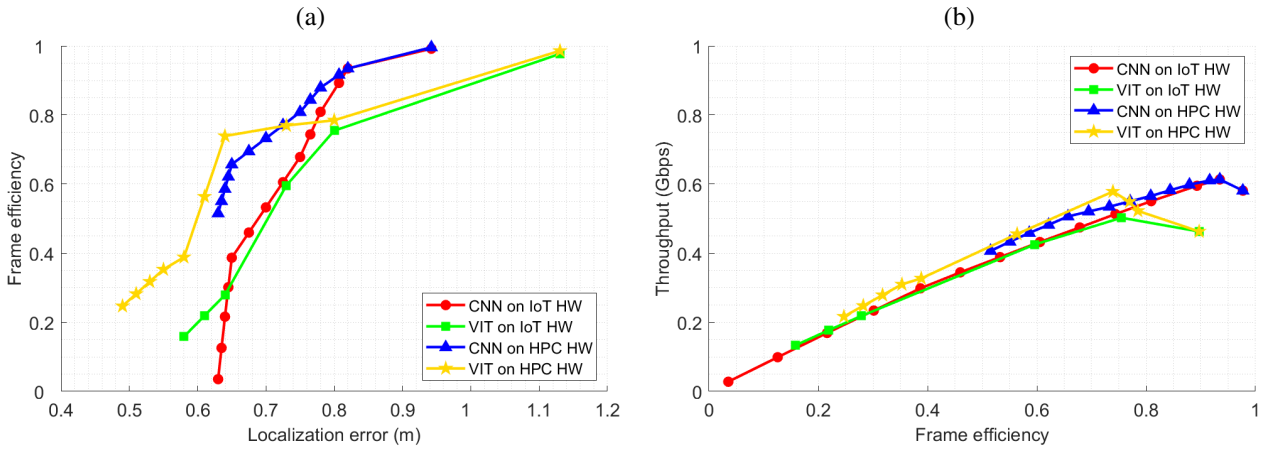


Fig. 9. Performance of the communication system: (a) frame efficiency $\eta_F$ against the localization error and (b) average system throughput (T) against frame efficiency for the four tested combinations of DL algorithm and HW type.

setup, each RIS has an area size of $A = 100\text{cm}^2$, the UEs have single antenna while the BS is equipped with a small array of $M^{\text{BS}} = 8$ antennas.

The UEs are localized by exploiting: (i) an optimized set of RIS which act as electronically steerable reflectors and configured by a set of rotation angles and (ii) 2 different algorithms based on Convolutional Neural Network (CNN) and Vision Transformer (ViT) deep learning models. We evaluate the performance of both algorithms in terms of RMSE for varying RIS parameters, inference time and complexity, considering also different computational capabilities of the deployed BS and devices, namely low-power (IoT) to high performance (HPC). Finally, the Cramer-Rao lower bound to positioning accuracy is evaluated and verified as a tool for pre-deployment assessment.

Targeting integrated localization and communication services as envisioned in Internet of Everything paradigms, we propose a time resource allocation strategy aiming to achieve the best compromise between UE positioning accuracy and spectral efficiency. While a longer localization time ensures a smaller positioning error, there is no apparent benefit in reducing such error below $0.5$ m as this subtracts resources useful for communication, decreasing the frame efficiency and the network throughput. An optimal tradeoff can be identified as a function of the RIS configuration, the computational capability of devices, the localization algorithm as well as the UE mobility pattern. Trading localization accuracy with frame efficiency is generally more beneficial for ViT-based localization algorithm rather than CNN.

Exploring the application of the proposed solution in real-world scenarios, such as smart cities or industrial IoT, would provide valuable insights into its practical feasibility and effectiveness. Lastly, delving deeper into the optimization of RIS deployment strategies, considering factors like RIS size, could contribute to the development of more practical and scalable implementations.

## APPENDIX: DERIVATION OF THE CRLB

In this section, we derive the Cramer-Rao Lower Bound (CRLB) on the estimate of the UE position, given the BS observations of the UE signals during the localization phase. The CRLB is employed in Sect.V for pre-deployment RIS performance assessment. Consider a set of $R_n$ setting (i.e. values for the pair $\delta_n, \psi_n$) for RIS $n$, and denote by $\delta_{n,r}, \psi_{n,r}$ the $r$-th setting, $r = 1, \ldots, R$. Then the observations $\mathbf{z}_{k,n}^{\text{BS}}(\delta_n, \psi_n)$ given by (14) obtained during the $r$-th setting can be denoted by $\mathbf{z}_{k,n,r}^{\text{BS}}$. For notation simplicity let us drop the superscript (BS) and the subscripts $k, n, r$ (they will be resumed later) and let us focus on a generic transmitting UE, reflecting RIS and RIS setting. The signal vector received at the BS can thus be rewritten as

$$\mathbf{z} = \mathbf{q} + \boldsymbol{\eta}, \tag{20}$$

where $\mathbf{q} \triangleq \widetilde{\mathbf{h}}_{k,n}(\delta_{n,r}, \psi_{n,r}) x_k^{\text{UE}}$. Note that $\mathbf{q}$ depends also on the user position, whose coordinates $(x, y)$ are the unknown parameters we would like to estimate. By recalling that the noise vector $\boldsymbol{\eta}$ has distribution $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{M^{\text{BS}}})$, the density of $\mathbf{z}$ given the user position is given by

$$f_{\mathbf{z}|x,y}(\mathbf{z}) = \frac{1}{(\pi\sigma^2)^{M^{\text{BS}}}} \exp\left( -\frac{(\mathbf{z}-\mathbf{q})^{\mathsf{H}}(\mathbf{z}-\mathbf{q})}{\sigma^2} \right) \tag{21}$$

The $2 \times 2$ Fisher information matrix (FIM) is defined as $\mathbf{F} = \mathbb{E}_{\mathbf{z}|x,y}[\mathbf{c}\mathbf{c}^{\mathsf{T}}]$ where

$$\mathbf{c} = \begin{bmatrix} \frac{\partial}{\partial x} \log f_{\mathbf{z}|x,y}(\mathbf{z}) \\ \frac{\partial}{\partial y} \log f_{\mathbf{z}|x,y}(\mathbf{z}) \end{bmatrix} \tag{22}$$

is the gradient of the log density, with respect to the user coordinates. We first observe that

$$\log f_{\mathbf{z}|x,y}(\mathbf{z}, x, y) = C - \frac{1}{\sigma^2}(\mathbf{z}-\mathbf{q})^{\mathsf{H}}(\mathbf{z}-\mathbf{q}), \tag{23}$$

where $C$ is a constant. After some algebra the FIM can be rewritten as

$$\mathbf{F} = \frac{2}{\sigma^2} \begin{bmatrix} \frac{\partial \mathbf{q}^{\mathsf{H}}}{\partial x}\frac{\partial \mathbf{q}}{\partial x} & \frac{\partial \mathbf{q}^{\mathsf{H}}}{\partial x}\frac{\partial \mathbf{q}}{\partial y} \\ \frac{\partial \mathbf{q}^{\mathsf{H}}}{\partial y}\frac{\partial \mathbf{q}}{\partial x} & \frac{\partial \mathbf{q}^{\mathsf{H}}}{\partial y}\frac{\partial \mathbf{q}}{\partial y} \end{bmatrix}. \tag{24}$$

Assume that for each reflecting RIS $n$ and for each rotation $r$ the BS observes $S$ independent signal samples. Then, overall the BS collects a set of $S\sum_{n=1}^{N'} R_n$ independent observations of the UE signal, where $N'$ is the number of RIS used in

the localization phase. Hence, by resuming the subscripts, the FIM for UE $k$ takes the form

$$\mathbf{F}_k = S \sum_{n=1}^{N'} \sum_{r=1}^{R_n} \mathbf{F}_{k,n,r}, \tag{25}$$

where $\mathbf{F}_{k,n,r}$ is the FIM computed using a single sample obtained through rotation $r$ of RIS $n$,

$$\mathbf{F}_{k,n,r} = \frac{2}{\sigma^2} \begin{bmatrix} \frac{\partial \mathbf{q}_{k,n,r}^{\mathsf{H}}}{\partial x_k} \frac{\partial \mathbf{q}_{k,n,r}}{\partial x_k} & \frac{\partial \mathbf{q}_{k,n,r}^{\mathsf{H}}}{\partial x_k} \frac{\partial \mathbf{q}_{k,n,r}}{\partial y_k} \\ \frac{\partial \mathbf{q}_{k,n,r}^{\mathsf{H}}}{\partial y_k} \frac{\partial \mathbf{q}_{k,n,r}}{\partial x_k} & \frac{\partial \mathbf{q}_{k,n,r}^{\mathsf{H}}}{\partial y_k} \frac{\partial \mathbf{q}_{k,n,r}}{\partial y_k} \end{bmatrix} \tag{26}$$

The elements on the diagonal of $\mathbf{F}_k^{-1}$ are the (per coordinate) CRLB on the variance of the estimation error achieved by any unbiased estimator. Therefore the CRLB on the variance of the position error is given by

$$\mathrm{CRB} = \mathsf{Tr}\left\{\mathbf{F}_k^{-1}\right\} \tag{27}$$

The analytic expression of the partial derivatives in $\mathbf{F}_{k,n,r}$ is quite cumbersome, although easy to obtain. Indeed, according to (16) the vector $\mathbf{q}_{k,n,r}$ can be written as

$$\mathbf{q}_{k,n,r} = \widetilde{\mathbf{h}}_{k,n,r} s_k^{\mathrm{UE}} = \mathbf{H}_n^{(1)^{\mathsf{T}}} \mathbf{\Theta}_{n,r} \mathbf{h}_{k,n}^{(2)} s_k^{\mathrm{UE}}, \tag{28}$$

where the dependence on the UE $k$ coordinates $(x_k, y_k)$ is in $\mathbf{h}_{k,n}^{(2)}$. Hence

$$\frac{\partial \mathbf{q}_{k,n,r}}{\partial x} = \mathbf{M}_{k,n,r} \frac{\partial \mathbf{h}_{k,n}^{(2)}}{\partial x_k}, \qquad \frac{\partial \mathbf{q}_{k,n,r}}{\partial y} = \mathbf{M}_{k,n,r} \frac{\partial \mathbf{h}_{k,n}^{(2)}}{\partial y_k} \tag{29}$$

and $\mathbf{M}_{k,n,r} = \mathbf{H}_n^{(1)^{\mathsf{T}}} \mathbf{\Theta}_{n,r} s_k^{\mathrm{UE}}$. If the link connecting UE $k$ with RIS $n$ has only the LoS path, the vector $\mathbf{h}_{k,n}^{(2)}$ is given by (see [4] for details) $\mathbf{h}_{k,n}^{(2)} = \frac{\sqrt{A_n \cos\phi_{k,n}}}{\sqrt{4\pi} d_{k,n}} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{\lambda} d_{k,n}} \mathbf{u}_{k,n}$ where $\mathbf{u}_{k,n}$ is the spatial signature of RIS $n$ as observed by UE $k$, whose $m$-th element is proportional to $\mathrm{e}^{\mathrm{j}2\pi m \Delta/\lambda \sin\phi_{k,n}}$, $d_{k,n}$ is the distance between UE $k$ and RIS $n$, and $\phi_{k,n}$ is the angle of UE $k$ as observed from RIS $n$. The dependence on the UE position $(x_k, y_k)$ is hidden in $d_{k,n}$ and in the angle $\phi_{k,n}$. Specifically $d_{k,n} = \sqrt{(x_n^{\mathrm{RIS}} - x_k)^2 + (y_n^{\mathrm{RIS}} - y_k)^2}$ and $\phi_{k,n} = \arctan \frac{x_n^{\mathrm{RIS}} - x_k}{y_n^{\mathrm{RIS}} - y_k}$ where $(x_n^{\mathrm{RIS}}, y_n^{\mathrm{RIS}})$ is the position of the $n$-th RIS.

## REFERENCES


[1] Fan Liu, Yuanhao Cui, Christos Masouros, Jie Xu, Tony Xiao Han, Yonina C. Eldar, and Stefano Buzzi. Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond. *IEEE Journal on Selected Areas in Communications*, 40(6):1728–1767, 2022.

[2] Carlos De Lima, Didier Belot, Rafael Berkvens, André Bourdoux, Davide Dardari, Maxime Guillaud, Minna Isomursu, Elena-Simona Lohan, Yang Miao, Andre Noll Barreto, Muhammad Reza Kahar Aziz, Jani Saloranta, Tachporn Sanguanpuak, Hadi Sarieddeen, Gonzalo Seco-Granados, Jaakko Suutala, Tommy Svensson, Mikko Valkama, Barend Van Liempd, and Henk Wymeersch. Convergent communication, sensing and localization in 6G systems: An overview of technologies, opportunities and challenges. *IEEE Access*, 9:26902–26925, 2021.

[3] Ertugrul Basar, Marco Di Renzo, Julien De Rosny, Merouane Debbah, Mohamed-Slim Alouini, and Rui Zhang. Wireless communications through reconfigurable intelligent surfaces. *IEEE Access*, 7:116753–116773, 2019.

[4] Alberto Tarable, Francesco Malandrino, Laura Dossi, Roberto Nebuloni, Giuseppe Virone, and Alessandro Nordio. Optimization of IRS-aided sub-THz communications under practical design constraints. *IEEE Transactions on Wireless Communications*, 21(12):10824–10838, 2022.

[5] Hongliang Zhang. Joint waveform and phase shift design for ris-assisted integrated sensing and communication based on mutual information. *IEEE Communications Letters*, 26(10):2317–2321, 2022.

[6] Baojia Luo, Miaomiao Dong, Hao Wu, Yue Li, Lu Yang, Xiang Chen, and Bo Bai. Reconfigurable intelligent surface assisted millimeter wave indoor localization systems. In *ICC 2022 - IEEE International Conference on Communications*, pages 4535–4540, 2022.

[7] Alexandros-Apostolos A. Boulogeorgos and Angeliki Alexiou. Performance analysis of reconfigurable intelligent surface-assisted wireless systems and comparison with relaying. *IEEE Access*, 8:94463–94483, 2020.

[8] Yuanbin Chen, Ying Wang, Jiayi Zhang, Ping Zhang, and Lajos Hanzo. Reconfigurable intelligent surface (ris)-aided vehicular networks: Their protocols, resource allocation, and performance. *IEEE Vehicular Technology Magazine*, 17(2):26–36, 2022.

[9] Yuanhao Cui, Fan Liu, Xiaojun Jing, and Junsheng Mu. Integrating sensing and communications for ubiquitous iot: Applications, trends, and challenges. *IEEE Network*, 35(5):158–167, 2021.

[10] Ertugrul Basar, Marco Di Renzo, Julien De Rosny, Merouane Debbah, Mohamed-Slim Alouini, and Rui Zhang. Wireless communications through reconfigurable intelligent surfaces. *IEEE Access*, 7:116753–116773, 2019.

[11] Ruhul Amin Khalil and Nasir Saeed. Convex hull optimization for robust localization in isac systems. *IEEE Sensors Letters*, 7(12):1–4, 2023.

[12] Tong Wei, Linlong Wu, Kumar Vijay Mishra, and M. R. Bhavani Shankar. Multiple IRS-assisted wideband dual-function radar-communication. In *2022 2nd IEEE International Symposium on Joint Communications & Sensing (JC&S)*, pages 1–5, 2022.

[13] Hangchuan He. BPNN localization method for a ISAC system under LOS/NLOS scenario. In *2023 4th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*, pages 444–448, 2023.

[14] Xu Gan, Chongwen Huang, Zhaohui Yang, Caijun Zhong, Xiaoming Chen, Jiguang He, Zhaoyang Zhang, Qinghua Guo, Chau Yuen, and Mérouane Debbah. Simultaneous communication and localization for double-RIS aided multi-UE ISAC systems. In *2023 IEEE 23rd International Conference on Communication Technology (ICCT)*, pages 422–427, 2023.

[15] Sanaz Kianoush, Laura Dossi, Roberto Nebuloni, Alessandro Nordio, and Stefano Savazzi. User location uncertainty in RIS-aided channel optimization. In *2023 26th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pages 20–26, 2023.

[16] Haobo Zhang, Hongliang Zhang, Boya Di, Kaigui Bian, Zhu Han, and Lingyang Song. Metalocalization: Reconfigurable intelligent surface aided multi-user wireless indoor localization. *IEEE Transactions on Wireless Communications*, 20(12):7743–7757, 2021.

[17] Özgecan Özdogan, Emil Björnson, and Erik G. Larsson. Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling. *IEEE Wireless Communications Letters*, 9(5):581–585, 2020.

[18] Christos Liaskos, Shuai Nie, Ageliki Tsioliaridou, Andreas Pitsillides, Sotiris Ioannidis, and Ian Akyildiz. A new wireless communication paradigm through software-controlled metasurfaces. *IEEE Communications Magazine*, 56(9):162–169, 2018.

[19] Jiguang He, Henk Wymeersch, Long Kong, Olli Silvén, and Markku Juntti. Large intelligent surface for positioning in millimeter wave MIMO systems. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–5, 2020.

[20] M. Dunna, C. Zhang, D. Sievenpiper, and D. Bharadia. ScatterMIMO: Enabling virtual MIMO with smart surfaces. *MobiCom '20: Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, (10):1–14, Oct. 2020.

[21] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2023.

[22] Krishna Teja Chitty-Venkata, Murali Emani, Venkatram Vishwanath, and Arun K. Somani. Neural architecture search for transformers: A survey. *IEEE Access*, 10:108374–108412, 2022.

[23] Amar Al-jzari, Jack Towers, and Sana Salous. Characterization of indoor environment in the 60 GHz band. In *2020 XXXIIIrd General Assembly and Scientific Symposium of the International Union of Radio Science*, pages 1–4, 2020.