Sensing Aided Reconfigurable Intelligent Surfaces for 3GPP 5G Transparent Operation

Shuaifeng Jiang, Ahmed Hindy, and Ahmed Alkhateeb

***Abstract*— Can reconfigurable intelligent surfaces (RISs) operate in a standalone mode that is completely transparent to the 3GPP 5G initial access process? Realizing that may greatly simplify the deployment and operation of these surfaces and reduce the infrastructure control overhead. This paper investigates the feasibility of building standalone/transparent RIS systems and shows that one key challenge lies in determining the user equipment (UE)-side RIS beam reflection direction. To address this challenge, we propose to equip the RISs with multi-modal sensing capabilities (*e.g.*, using wireless and visual sensors) that enable them to develop some perception of the surrounding environment and the mobile users. Based on that, we develop a machine learning framework that leverages the wireless and visual sensors at the RIS to select the high-performance beams between the base station (BS) and UEs and enable standalone/transparent RIS operation for 5G high-frequency systems. Using a high-fidelity synthetic dataset with co-existing wireless and visual data, we extensively evaluate the performance of the proposed framework. Experimental results demonstrate that the proposed approach can accurately predict the BS and UE-side candidate beams, and that the standalone RIS beam selection solution is capable of realizing near-optimal achievable rates with significantly reduced beam training overhead.**

***Index Terms*— Reconfigurable intelligent surface, sensing, computer vision, standalone operation, beam selection.**

I. Introduction

RECONFIGURABLE intelligent surfaces (RISs) and holographic multiple-input multiple-output surfaces (HMIMOS) have the potential to extend the coverage and reliability of millimeter wave (mmWave) and terahertz (THz) communication networks in 5G and beyond [2], [3], [4], [5], [6], [7], [8]. In particular, the RISs employ large numbers of reflecting elements that can reflect and focus the incident signals toward the wireless receiver with proper control. This enables the network to bypass blockages and maintain reliable link connections. Figuring out the right configuration of these reflecting elements, however, is a challenging task that requires large channel estimation, beam training, and feedback overhead [9], [10], [11]. Prior work has mainly assumed that the RIS is controlled by the base station (BS), simplifying its operation. However, this approach requires high channel estimation and control signaling overhead, which needs to be captured in future releases of the 5G standard to ensure interoperability across different user equipments (UEs) and network vendors. In this paper, we address the following question: **Can we build *standalone RISs* of which the operation is transparent to the 3GPP 5G protocol?** In particular, can the RIS assist the BS-UE link without any coordination or feedback from neither of them? We present the key challenges in realizing this standalone/transparent RIS operation vision and show how these challenges can be relaxed by employing sensing at the RIS surfaces.

A. Prior Work

Prior work has extensively studied various aspects of the RIS system operation, including its channel estimation and beamforming [9], [10], [11], [12], [13] and the application of machine learning (ML) to enhance RIS-aided communication systems [9], [14]. In [12], a channel estimation procedure for the RIS-aided communication systems was proposed relying on activating the RIS elements one by one, and estimating the end-to-end BS-RIS-UE channel at the BS. To reduce the estimation overhead, [13] proposed to divide the reflecting elements into groups and estimate the effective channel for each group. In [10] and [11], the author investigated the channel estimation for the RIS-aided uplink multi-user systems. These channel estimation procedures in [10], [11], [12], and [13], however, can only be used if the RIS is controlled by the BS and do not support standalone RIS operation because: (i) these approaches implicitly assume that the RIS reflection configuration is synchronized with the pilot transmissions, which requires the RIS coordination with the UE and infrastructure and (ii) after channel estimation, the BS and/or UE needs to feedback the estimated channel or the beamforming configuration to the RIS, which also needs dedicated signaling.

Towards standalone RIS operation, [9] developed what is known as the semi-passive RIS architecture which uses sparse active antenna elements to estimate the channels of the incident signals. In [14], the authors optimized the equipment of these sparse active antenna elements at the RIS, which can

Manuscript received 4 December 2022; revised 20 April 2023 and 18 June 2023; accepted 30 July 2023. Date of publication 15 August 2023; date of current version 20 November 2023. This work was supported by the National Science Foundation (NSF) under Grant No. 2048021. An earlier version of this paper was presented in part at the 2023 IEEE International Conference on Communications [DOI: 10.48550/arXiv.2211.07563]. The associate editor coordinating the review of this article and approving it for publication was L. Yang. *(Corresponding author: Ahmed Alkhateeb.)*

Shuaifeng Jiang and Ahmed Alkhateeb are with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: s.jiang@asu.edu; alkhateeb@asu.edu).

Ahmed Hindy is with Motorola Mobility LLC (a Lenovo Company), Chicago, IL 60654 USA (e-mail: ahmedhindy@motorola.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2023.3305478.

Digital Object Identifier 10.1109/TCOMM.2023.3305478

further improve the channel estimation and beamforming performance. [15] used a super-resolution neural network (NN) to predict the full RIS channel from the partial channel at the sparse active antenna elements. In [16], the authors proposed an efficient RIS beamforming approach that does not require explicit channel estimation. However, it still needs the UE to feed back the receive power measurements of the probing beams. In [17], the authors proposed a novel self-configuring RIS beamforming approach to establish communication links between BS and UEs by measuring the power of the pilot signals transmitted from both the BS and the UEs. In the 3GPP 5G initial access process [18], however, the UEs do not transmit any pilot signals before they receive the BS synchronization signals. Therefore, the above approaches alone are not sufficient to enable fully transparent RIS operation with respect to the 3GPP 5G.

In another context, integrating multi-modal sensing at the infrastructure and mobile UEs to aid wireless communications has been recently attracting interest for different use cases [19], [20], [21], [22], [23], [24], [25], [26], [27], and [28]. For example, in [19], [20], and [22], the authors showed that position and orientation data could enable the mobile UEs to predict their optimal beam directions and reduce the beam search overhead. Further, in [23], [24], and [28], the visual data collected by cameras installed at the BS was utilized to narrow down the beam search space, improve the support for high-mobility users, and enhance the system reliability. In particular, [28] built real-world proof-of-concept prototypes that demonstrated the feasibility of using visual information to aid the mmWave beam selection process and significantly reduce the beam training overhead. Similarly, in [25], [26], and [27], LiDAR and radar information are leveraged to aid mmWave beam prediction and tracking. Reference [29] investigated joint communication and environment sensing for multi-user RIS systems, where the communication signals are utilized for environment object detection.

B. Contribution

We propose a sensing-aided RIS operation that is transparent to the 3GPP 5G initial access process. In particular, the proposed RIS performs efficient beam selection without dedicated and additional signaling from the BS and UE according to 3GPP 5G protocols. To the best of our knowledge, **this work is the first to introduce transparent RIS operations and considers the compatibility with 3GPP 5G protocols.** Our contribution to enabling the transparent RIS beam selection can be summarized as follows.

- We introduce a sensing-based framework for the RIS to enable 3GPP 5G transparent beam selection operations. In particular, the RIS utilizes sparse wireless channel receivers to obtain information about the BS-side channel and leverages visual sensors (cameras) to obtain information about the UE-side channel.
- We develop a decoupled BS-side and UE-side RIS beam design formulation that has low complexity yet leads to near-optimal performance in realistic propagation scenarios. This also provides the flexibility to efficiently process the sensing information in different (wireless and visual) modalities for the BS-side and UE-side beam selections, respectively.
- We develop a ML framework that predicts the UE-side candidate beam sets based on the visual information. The proposed learning framework first detects the candidate UEs using an object detector. Then, an efficient NN architecture is proposed to predict the corresponding UE-side candidate beam set.
- We build a high-fidelity synthetic dataset that incorporates co-existing wireless and visual data, which enables the research of the proposed sensing-aided RIS system. Using this high-fidelity dataset, we extensively evaluate the performance of the proposed sensing-aided RIS beam selection algorithms.

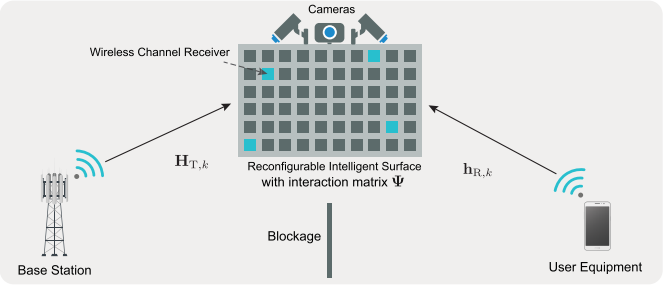


Fig. 1. The considered system model consists of a BS, a UE, and a RIS. The link between the BS and UE is blocked, and their communication is aided by the RIS. The RIS is equipped with cameras and sparsely-distributed receivers.

Simulation results demonstrate the efficiency of the proposed algorithms in reducing the beam training overhead, achieving near-optimal data rates, and enabling standalone RIS operation that is compliant with the 3GPP 5G initial access process, highlighting a promising framework for future RIS-aided wireless communication systems.

The rest of the paper is organized as follows. Section II explains the system and channel models. Section III briefly introduces the 5G NR initial access procedure. Section IV presents the key challenge of transparent 3GPP RIS. Section V demonstrates the idea of obtaining environment awareness with sensing-based perception, and presents the ML framework and the standalone RIS beam selection. Section VI explains the deep learning (DL) models in the ML framework. Simulation setup and results are presented in Section VII and Section VIII. In Section IX, we discuss the enabling features to extend the standalone RIS operation to more complex scenarios. Section X concludes the paper.

II. System and Channel Models

In this section, we describe the system and channel models for the RIS-aided wireless communication scenario considered in this paper.

A. System Model

We consider a high-frequency (*e.g.* mmWave and sub-THz) wireless communication system where a RIS aids the communication between a BS and a UE. For ease of exposition,

we assume that a blockage exists between the BS and the UE. Consequently, the BS can only communicate with the UE through the RIS, and the direct links between the BS and UE have negligible gain compared to the RIS links. This could be a reasonable assumption in high-frequency systems due to the high pathloss and penetration loss [9]. Note that, if a LoS direct link between the BS and the UE exists, then this direct link will be much more dominant than the RIS link. In this case, it may be preferable to use the direct link between the BS and the UE. We assume that the BS has an antenna array of N elements, and the UE has a single antenna. The RIS has M reconfigurable reflecting elements. Further, the RIS is equipped with RGB cameras and sparse wireless channel receivers to obtain sensing information about the surrounding environment. There are three differently-oriented cameras deployed at the center of the RIS surface. These cameras provide a central view and two side views of the surrounding environment. Four reflecting elements at the corners of the RIS are active (connected to baseband), and they act as the wireless receivers [9]. It is worth noting that the considered system model makes the following assumptions. (i) The RIS lies within the coverage area of only one BS. (ii) The area covered by the RIS cameras is within the coverage area of only one BS, which is the BS serving the RIS. (iii) There is only one UE that can be active and served by the RIS. (iv) The UE is supported by the service provider corresponding to the BS.

For the uplink and downlink communication, we adopt orthogonal frequency-division multiplexing (OFDM) with K subcarriers. Let $\mathbf{H}_{\mathrm{T},k} \in \mathbb{C}^{M\times N}$ and $\mathbf{h}_{\mathrm{R},k} \in \mathbb{C}^{M\times 1}$ denote the channel matrix from the BS to the RIS and the channel vector from the UE to the RIS at the k-th subcarrier, respectively. If the BS transmits a signal $s_k \in \mathbb{C}$ on the k-th subcarrier, then we can write the downlink received signal as

$$y_k = \mathbf{h}_{\mathrm{R},k}^T \boldsymbol{\Psi} \mathbf{H}_{\mathrm{T},k} \mathbf{f} s_k + n_k, \tag{1}$$

where $\mathbf{f} \in \mathbb{C}^{N\times 1}$ denotes the beamforming vector of the BS. The transmitted signal s_k satisfies the power constraint $\mathbb{E}\left[s_k^H s_k\right] = \frac{p_t}{K}$ with p_t representing the total transmit power. $n_k \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_n^2)$ is the complex receive noise at the UE. We use $\boldsymbol{\Psi} \in \mathbb{C}^{M\times M}$ to denote the RIS interaction matrix, which can be written as $\boldsymbol{\Psi} = \mathrm{diag}(\boldsymbol{\psi})$. The $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_M]^T$ is the diagonal vector of $\boldsymbol{\Psi}$ with $\psi_m \in \mathbb{C}$ denoting the phase shift of the m-th reflecting element. ψ_m satisfies $|\psi_m|^2 = 1$ to capture the constant-modulus constraint. We call $\boldsymbol{\psi}$ the reflecting beamforming vector of the RIS. The same $\boldsymbol{\psi}$ is applied to all subcarriers due to the time-domain implementation.

B. Channel Model

We adopt a wideband geometric channel model for the channels $\mathbf{H}_{\mathrm{T},k}$ and $\mathbf{h}_{\mathrm{R},k}$. With this model, if $\mathbf{h}_{\mathrm{R},k}$ consists of L clusters, and each cluster $\ell \in [1, L]$ contributes with one ray of time delay $\tau_\ell \in \mathbb{R}$, then the delay-d channel vector between the UE and the RIS can be written as

$$\mathbf{h}_{\mathrm{R},d} = \sqrt{\frac{M}{\rho}} \sum_{\ell=1}^{L} \alpha_\ell p(dT_s - \tau_\ell) \mathbf{a}(\phi_\ell^R, \theta_\ell^R), \tag{2}$$

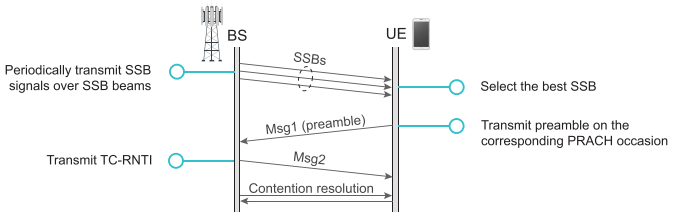


Fig. 2. This figure summarizes the signaling and message exchange during the typical 3GPP 5G initial access process.

where ρ denotes the pathloss and $p(\tau)$ denotes the pulse shaping function which represents a T_s-spaced signaling evaluated at τ seconds, $\mathbf{a}(\phi_\ell^R, \theta_\ell^R)$ is the array response vector of the RIS. ϕ_ℓ^R and θ_ℓ^R are the azimuth and elevation angles of arrival (AoA) associated with the ℓ-th cluster. $\alpha_\ell \in \mathbb{C}$ is a complex coefficient of the ℓ-th cluster. Given the delay-d channel in (2), the frequency domain channel vector at subcarrier k can be written as

$$\mathbf{h}_{\mathrm{R},k} = \sum_{d=0}^{D-1} \mathbf{h}_{\mathrm{R},d} e^{-j\frac{2\pi k}{K} d}, \tag{3}$$

where D represents the maximum delay of the channel. The channel $\mathbf{H}_{\mathrm{T},k}$ is similarly defined.

III. 3GPP 5G Initial Access: A Brief Background

We assume that the BS and the UE try to initiate a link using the 3GPP 5G protocol [18]. Here, we briefly revisit the 3GPP 5G initial access process, which is illustrated in Fig. 2.

A. SSB Signals

The BS transmits periodic synchronization signal blocks (SSBs) using a predefined set of beams (codebook) [18]. When a UE wants to access the 5G wireless network, it listens to these SSB signals and blindly decodes them with its set of initial access beams. Based on this beam training process, the UE selects the pair of the SSB beam and receive beam that results in the maximum reference signal receive power (RSRP).

B. Message 1 UE Preamble

After successfully decoding the SSBs, the UE initiates a random access process by transmitting an uplink preamble sequence using the selected receive beam (which maximizes the RSRP of the SSBs) at a physical random access channel (PRACH) occasion. Based on the transmitted preamble sequence and the PRACH occasion where the preamble sequence is transmitted, the BS knows which SSB beam (direction) was selected by the UE [18]. The preamble sequence is also called Message 1 (Msg1).

C. Message 2 BS Random Access Response (RAR)

The BS always listens to the PRACH. Upon detecting a UE preamble, the BS transmits RAR using the same SSB beam. The RAR is also known as Message 2 (Msg2). It contains a temporary cell radio network temporary identifier (TC-RNTI),

which is calculated based on the PRACH occasion of the UE preamble. By receiving the TC-RNTI, the UE knows if its preamble was decoded by the BS (more details in [18]).

IV. Transparent 3GPP 5G RIS: The Key Challenge

Section III briefly introduced the 3GPP 5G initial access. Now, if the direct link between the BS and UE is blocked, the RIS needs to configure its beamforming to aid this communication. But how can this be done if both the BS and the UE do not know about the existence of the RIS? We provide an initial investigation to address this question and highlight its key challenges.

To start, we adopt the achievable rate as the communication performance metric of interest. Given the system model in Section II and the downlink received signal in (1), the downlink achievable rate for the adopted RIS-based communication system can be written as

$$R = \frac{1}{K}\sum_{k=1}^{K}\log_2\left(1+\text{SNR}\left|\left(\mathbf{h}_{\text{R},k}\odot\mathbf{H}_{\text{T},k}\mathbf{f}\right)^T\boldsymbol{\psi}\right|^2\right), \quad (4)$$

where $\text{SNR}=\frac{p_t}{K\sigma_n^2}$ denotes the signal-to-noise ratio. The $\odot$ denotes the element-wise multiplication of two vectors. Therefore, for a given BS beamforming vector $\mathbf{f}$, the optimal reflecting beam for this pair of BS and UE is the one that maximizes the achievable rate as shown by

$$\boldsymbol{\psi}^\star = \arg\max_{\boldsymbol{\psi}\in\mathcal{O}}\frac{1}{K}\sum_{k=1}^{K}\log_2\left(1+\text{SNR}\left|\left(\mathbf{h}_{\text{R},k}\odot\mathbf{H}_{\text{T},k}\mathbf{f}\right)^T\boldsymbol{\psi}\right|^2\right), \quad (5)$$

where $\mathcal{O}$ is the set of all the reflecting beamformers $\boldsymbol{\psi}$ that satisfy the constant modulus phase-only constraint, *i.e.*, $|\psi_m|^2 = 1$. Note that the RIS reflecting beam $\boldsymbol{\psi}$ can be decomposed into the BS-side and the UE-side beam as $\boldsymbol{\psi} = \mathbf{p}\odot\mathbf{q}$, where $\mathbf{p},\mathbf{q}\in\mathcal{O}$ denote the BS-side and UE-side beamforming vectors. The optimization problem in (5) can then be equivalently written as

$$(\mathbf{p}^\star,\mathbf{q}^\star) = \arg\max_{\mathbf{p},\mathbf{q}\in\mathcal{O}}\frac{1}{K}\sum_{k=1}^{K}\log_2\left(1+\text{SNR}\left|\left(\mathbf{h}_{\text{R},k}\odot\mathbf{H}_{\text{T},k}\mathbf{f}\right)^T(\mathbf{p}\odot\mathbf{q})\right|^2\right). \quad (6)$$

Now, to account for the practical constraint of quantized phase shifters [30], we limit the search space of $\mathbf{p}$ and $\mathbf{q}$ to pre-designed finite-size codebooks $\mathcal{P}$ and $\mathcal{Q}$. It is important to note here that (the new optimization problem) can act as an upper bound of the designed codebooks. Given the two codebooks, the optimization problem in (5) can be re-written as

$$(\mathbf{p}^\star,\mathbf{q}^\star) = \arg\max_{\substack{\mathbf{p}\in\mathcal{P},\\ \mathbf{q}\in\mathcal{Q}}}\frac{1}{K}\sum_{k=1}^{K}\log_2\left(1+\text{SNR}\left|\left(\mathbf{h}_{\text{T},k}\odot\mathbf{p}\right)^T\left(\mathbf{h}_{\text{R},k}\odot\mathbf{q}\right)\right|^2\right), \quad (7)$$

where $\mathbf{h}_{\text{T},k} = \mathbf{H}_{\text{T},k}\mathbf{f}$ is the effective channel vector between the BS and the RIS (accounting for the BS beamforming). It can be seen from (7) that the optimal BS-side beam $\mathbf{p}^\star$ and the optimal UE-side beam $\mathbf{q}^\star$ depend on both the BS-side channel $\mathbf{h}_{\text{T},k}$ and the UE-side channel $\mathbf{h}_{\text{R},k}$. To further simplify the RIS operation, we decouple the optimization problem in (7) to the sub-optimal approximation as follows:

$$\begin{cases}\mathbf{p}^\star = \arg\max_{\mathbf{p}\in\mathcal{P}}\frac{1}{K}\sum_{k=1}^{K}\left|\left(\mathbf{h}_{\text{T},k}\odot\mathbf{p}\right)^H\mathbf{a}^*\right|^2 & \text{(8a)}\\ \mathbf{q}^\star = \arg\max_{\mathbf{q}\in\mathcal{Q}}\frac{1}{K}\sum_{k=1}^{K}\left|\left(\mathbf{h}_{\text{R},k}\odot\mathbf{q}\right)^H\mathbf{a}\right|^2, & \text{(8b)}\end{cases}$$

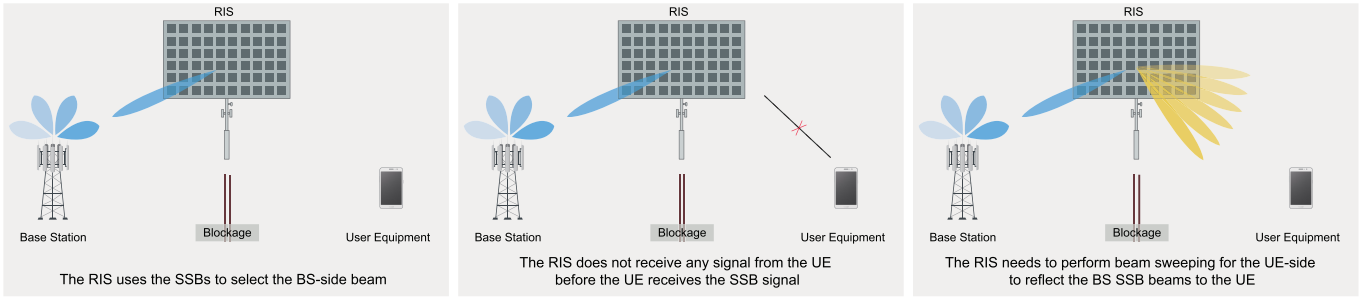
where $\mathbf{a}\in\mathbb{C}^{M\times 1}$ is an arbitrary reference vector. It is worth noting that the $\mathbf{p}^\star$ and $\mathbf{q}^\star$ obtained in (8) is the *optimal* solution to (7) for a very important case, *i.e.*, when the BS-side channel and the UE-side channel only contain the LoS path, and $\mathcal{P}=\mathcal{Q}=\mathcal{O}$ (proof in Appendix).

A. BS-Side RIS Beam Selection

The optimal BS-side beam in (8a) depends on the effective BS-RIS channel $\mathbf{h}_{\text{T},k}$. Here, we propose that the RIS can predict the BS-side beam by exploiting the SSBs as pilot signals. Since the SSBs are transmitted with or without the RIS in 3GPP 5G [18], dedicated signaling for RIS is not required. Note that the effective BS-RIS channel $\mathbf{h}_{\text{T},k}$ may vary with different BS beams $\mathbf{f}$. To enhance clarity, we first explain the BS-side beam selection with a constant $\mathbf{f}$. Then, we extend it to align with the 3GPP 5G, which allows for variation in the BS SSB beam $\mathbf{f}$.

With a constant BS beam $\mathbf{f}$, the RIS first blindly decodes the SSBs with its sparse wireless channel receivers. Next, the RIS employs the SSBs as predefined pilot signals and estimates the channel between the BS and the sparse wireless channel receivers, which is denoted by $\overline{\mathbf{h}}_{\text{T},k}$. The $\overline{\mathbf{h}}_{\text{T},k}$ can be considered a sub-sampled version of the effective BS-RIS channel $\mathbf{h}_{\text{T},k}$ because the wireless channel receivers are a subset of the RIS elements. In [9], the author showed that it is possible to estimate $\mathbf{h}_{\text{T},k}$ from the sub-sampled channel $\overline{\mathbf{h}}_{\text{T},k}$ when the channel experiences sparse scattering, which is typically the case in the considered high-frequency system. The literature has extensively studied estimating the $\mathbf{h}_{\text{T},k}$ from the sub-sampled channel $\overline{\mathbf{h}}_{\text{T},k}$ using compressive sensing or DL approaches [9], [14], [15]. With the existing solutions, we can obtain the estimate of $\mathbf{h}_{\text{T},k}$, which can then be used to find the optimal BS-side beam $\mathbf{p}\in\mathcal{P}$ via an offline exhaustive search using (8a).

As explained in Section III, each SSB can be transmitted with a different BS beam in the predefined BS beam codebook. In this case, the RIS can estimate all the BS-RIS effective channels (accounting for different BS beams) using the corresponding SSBs. Then, the optimal BS-side beam can be obtained via an offline exhaustive search over all BS-RIS effective channels using (8a). Once the RIS beam is determined, the BS can select the optimal BS beam that maximizes the UE's RSRP based on the beam sweeping and the UE feedback in Msg1 (explained in Section III).



(a) BS-side beam selection (b) Key challenge (c) UE-side beam sweeping

Fig. 3. This figure illustrates the key challenge in applying the 3GPP 5G initial access process for a RIS-aided communication system; the UE sends the preamble sequence only after it receives and decodes the SSBs. Hence, the RIS cannot have any information about the UE channel and it needs to perform beam sweeping over a very large codebook.

Since the BS-side RIS beam selection can be solved with existing approaches, we do not focus on this problem. Instead, we assume that the optimal BS-side beam is employed.

B. The Key Challenge

Selecting the UE-side beam from the codebook $\mathcal{Q}$ is the key challenge. According to the 3GPP 5G initial access, the UE does not transmit any signal/preamble until it detects the BS SSB signals. When there is a blockage between the BS and the UE, the UE may not be able to detect the BS SSBs without a communication link established by the RIS. **Therefore, the RIS is required to configure its interaction matrix to guarantee sufficient receive signal power at the UE *before* the UE can detect any signal from the BS, and before it (the RIS) receives any signals from the UE.** In other words, the RIS needs to configure the interaction matrix *without* receiving any signals from the UE. A trivial solution to design the UE-side beamforming vector at the RIS is the exhaustive search (beam sweeping) over all the beams in the codebook $\mathcal{Q}$ until the UE responds with an index corresponding to one of the beams. However, the number of beams that the RIS supports generally grows proportionally to the number of reflecting elements to fully exploit them. **Due to the large number of RIS reflecting elements, the beam sweeping requires tremendous training overhead.** Further, the beam sweeping becomes more impractical for mobile UEs as the beam sweeping needs to be frequently repeated in short periods of time.

Then, our objective is to solve the optimization problem in (8b), *i.e.*, finding the UE-side beamforming vector $\mathbf{q}^\star \in \mathcal{Q}$, with the smallest number of trials. In Section V, we propose to leverage sensing to achieve this objective.

V. Sensing for Standalone RIS Operation

The convergence of communication, sensing, and localization is considered one of the key features in 6G and beyond [31], [32]. The sensing and localization capabilities may provide rich information and awareness about the surrounding environment to the communication systems. In this section, we propose to utilize sensing at the RIS to build up environment awareness and leverage this awareness in enabling efficient transparent operation for these surfaces.

A. Key Idea: Observe With Sensing

This paper mainly focuses on high-frequency RIS-aided 5G communications. The high-frequency systems often rely on the beamforming gain of highly directional beams to achieve sufficient receive SNR. Moreover, in high-frequency systems, RIS is primarily used to aid communications through the reflected LoS path, where the channels between the RIS and the BS/UE are likely to be dominated by LoS paths. Therefore, the RIS beams highly rely on the geometric topology, such as the position and direction of the communication devices. This motivates employing sensors at RIS to obtain sensing information about the communication environment. While the sensing information may not fully reveal the cluster information for wideband channel modeling, it captures the geometric topology information which can be used to aid the RIS beamforming. Although there are various modalities of sensors that can capture information about the communication environment, we are particularly interested in adopting visual sensors at the RIS for the following reasons. (i) In high-frequency RIS systems, the communication devices are likely to be LoS with the RIS. Cameras can provide fine-grained spatial and visual information about LoS objects, which can aid in identifying and locating them. (ii) Compared to other types of sensors, such as LiDAR and radar, cameras typically have the advantage of low hardware cost. (iii) Advanced computer vision and image processing algorithms, *e.g.*, object detection, can be adopted to aid communications with little modifications.

The visual sensing information can help the RIS-integrated wireless communication system in several ways. (i) The RIS can leverage the sensing information for identifying the promising beamforming directions and avoid extensive (blind) beam training. (ii) The sensing information can potentially help the RIS manage its beams to avoid causing interference to adjacent users associated with neighboring BSs. (iii) By periodically monitoring the locations/directions of the candidate UEs, the RIS can track the UEs' beams and model their mobility patterns. Next, we focus on the first potential gain, which is reducing the beam training overhead.

B. RIS Beam Set Prediction

As discussed in Section IV, the key challenge of realizing standalone RIS operation lies in the high beam training overhead associated with UE-side RIS beam sweeping. To tackle

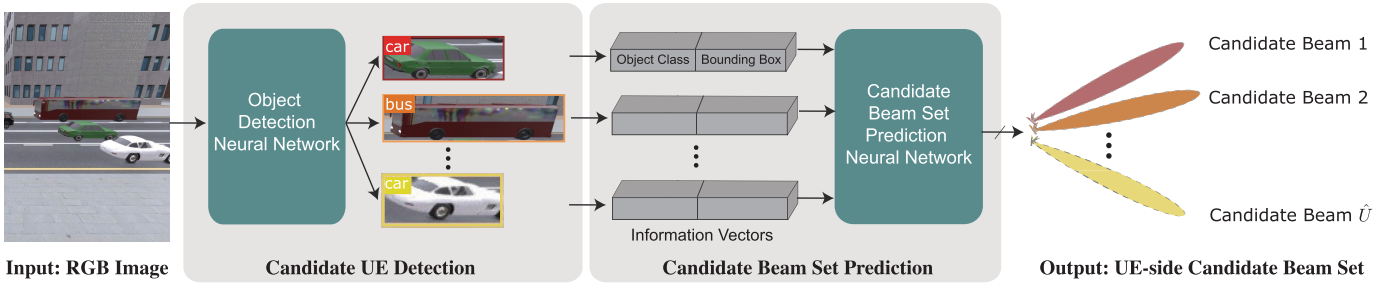


Fig. 4. Overview of the proposed ML framework for the RIS UE-side candidate beam set prediction. Given the visual information of the surrounding environment of the RIS, the framework first detects the candidate UEs with an object detector. Then the information about the candidate UEs (bounding boxes and object classes) are used to predict the candidate beams.

this problem, we propose an ML framework that (i) utilizes visual sensors at the RIS to obtain information about the candidate UEs in the scene, and (ii) exploits this information to reduce the UE-side beam training overhead of the RIS. Fig. 4 shows the overview of the proposed ML framework. With the equipped cameras, the RIS first obtains visual information, namely, RGB images, of the surrounding environment. Then, the RIS leverages computer vision object detection models to identify the candidate UEs in the scene. The information of all the detected candidate UEs is then used to predict the set of the most promising beams for the candidate UEs. Note that the number of these promising/candidate beams is proportional to the number of candidate UEs in the scene, which is typically much smaller than the size of the RIS beam codebook. This highlights the potential of significantly reducing the beam training overhead by leveraging the available visual information.

The objective of the ML framework is then to accurately predict a set of UE-side candidate beams that correspond to the candidate UEs in a given scene. More specifically, for each UE, the objective is to accurately predict the UE-side RIS beam $\mathbf{q}^\star$ that satisfies (8b). The optimal UE-side candidate beam set of an image can then be defined as

$$\mathbf{Q}^\star = \{\mathbf{q}_1^\star, \ldots, \mathbf{q}_U^\star\}, \tag{9}$$

where U is the total number of ground-truth candidate UEs in the image, and $\mathbf{q}_u^\star$ $(u = 1, \ldots, U)$ is the optimal UE-side beamforming vector for the u-th ground-truth UE. The optimal framework $f^\star(\cdot)$ is then defined as the one which can perfectly predict the optimal UE-side candidate beam set for any given image. Let $\mathbf{X} \in \mathbb{R}^{w \times h \times 3}$ denote the input RGB image to the ML framework with w and h denoting the width and height of $\mathbf{X}$. The optimal framework can be expressed as

$$f^\star(\mathbf{X}) = \mathbf{Q}_{\mathbf{X}}^\star, \tag{10}$$

where $\mathbf{Q}_{\mathbf{X}}^\star$ is the optimal UE-side candidate beam set of image $\mathbf{X}$. Deriving the exact expression of $f^\star(\cdot)$ is very difficult since it depends on the channel model, the visual model, and the environment around the UE and RIS. This motivated using DL models to learn the complex function $f^\star(\cdot)$ in a data-driven manner. The adopted DL models will be explained in Section VI.

Algorithm 1 Proposed Sensing-Aided Transparent RIS Operation

Step 0: RIS predicts the BS-side beam (explained in Section IV-A)
The RIS synchronizes with the BS by receiving and blind decoding SSBs
RIS estimates the channels at the sparse wireless receivers using SSBs as pilots
RIS predicts BS-side beam $\hat{\mathbf{p}}$ using the channel estimates at the sparse wireless receivers [9], [14], [15]
Step 1: RIS predicts the UE-side candidate beam set
RIS detects candidate UEs using the DL-based object detection (explained in Section VI-A)
RIS predicts the UE-side candidate beam set $\hat{\mathbf{Q}}$ using the DL model (explained in Section VI-B)
# Step 2: RIS beam sweeping over the UE-side candidate beam set
for each UE-side beam $\hat{\mathbf{q}}$ in the predicted UE-side candidate beam set $\hat{\mathbf{Q}}$ **do**
  RIS configures its elements using the RIS beam $(\hat{\mathbf{p}} \odot \hat{\mathbf{q}})$
  if RIS detects a successful initial access **then**
    Jump to Line 13
  end if
end for
Jump to Line 2
# Step 3: Maintaining the link
while Stopping criterion is not met **do**
  RIS performs beam tracking starting from the current RIS beam. (This is not implemented in this paper)
end while
Jump to Line 2

C. Vision-Aided Transparent RIS for 3GPP 5G

Here, we describe the proposed transparent 3GPP 5G operation of the vision-aided RIS system. The proposed RIS operation mainly comprises four steps, which are summarized in Algorithm 1.

Step 0. RIS Predicts the BS-Side Beam: Using the sparse wireless channel receivers, the RIS blindly decodes the SSBs periodically transmitted by the BS. These SSBs are first used to synchronize with the BS. Then, the RIS exploits the SSBs as pilot signals to predict the BS-side beamforming vector $\hat{\mathbf{p}}$ as explained in Section IV-A. Note

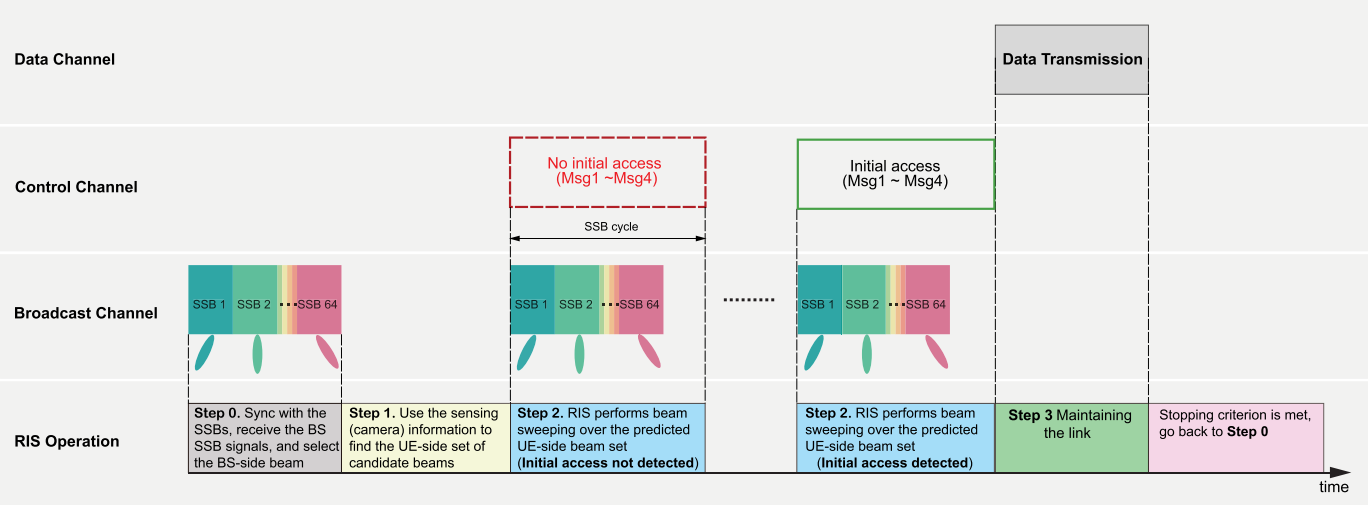


Fig. 5. This figure shows the proposed standalone RIS system operation and its compliance with the 3GPP 5G protocol. Exploiting the visual information, the ML framework reduces the UE-side beam training overhead for the RIS (step 1).

that we assume the RIS knows beforehand the initial access configuration, including the SSB periodicity, bandwidth, and carrier frequency. This can be achieved if the RIS has the same capability as a UE device. Note that we assume that the RIS lies within the coverage area of only one BS, otherwise, it may not be able to correctly align its BS-side beam with the intended BS.

Step 1. RIS Predicts UE-Side Candidate Beam Set With Camera Information: The RIS obtains the visual information (images) of the surrounding environment with its visual sensors. Given this visual information, the RIS detects the candidate number of UEs $\hat{U}$ and predicts the corresponding UE-side candidate beam set $\hat{\mathbf{Q}}$ with the ML framework shown in Fig. 4. Note that we assume that the area covered by the images is within the coverage area of only one BS, which is the BS that the RIS is aligned with in Step 0.

Step 2. RIS Performs Beam Sweeping Over the Predicted UE-Side Beam Set: The RIS applies the BS-side beam vector obtained in Step 0, and performs beam training within the UE-side candidate beam set $\hat{\mathbf{Q}}$ obtained in Step 1. For each UE-side candidate beam $\hat{\mathbf{q}}_u \in \hat{\mathbf{Q}}$, the RIS holds the beam for a time window longer than one SSB cycle (*e.g.* 20 milliseconds). Meanwhile, using its sparse wireless channel receivers, the RIS tries to detect a successful initial access (Msg1 $\sim$ Msg4) by detecting the signal power on the wireless bands where the initial access messages are transmitted.[1] If the RIS cannot detect a successful initial access, it concludes that no UE can use the currently tested UE-side beam. The RIS then switches to the next UE-side candidate beam in the set $\hat{\mathbf{Q}}$, and repeats the same process. As shown by Fig. 5, when testing the UE-side beams, the RIS ensures its beam switching is synchronized with the BS's SSB beam sweeping. In such a manner, the UE can have a chance to receive one or more complete SSBs. Note that the proposed transparent RIS operation focuses on the single-user MIMO scenario. The RIS may detect multiple candidate UEs in Step 1, however, it serves only one of the candidate UEs. We briefly discuss the multi-user scenario in Section IX.

Step 3. Maintaining the Link: If the RIS detects a successful initial access, it indicates that one UE is communicating with the BS via the current RIS beam. The RIS should perform beam tracking (starting from this beam) to serve this UE as it moves. To achieve beam tracking, the sensing information can be exploited to infer the optimal UE-side beam for the current time instance. Furthermore, enabled by the sensing capability, the RIS can predict the dynamics of the environment and the UE to achieve proactive beam tracking (predict the future optimal UE-side beam). During the beam tracking, the RIS keeps monitoring certain stopping criteria. After a stopping criterion is met, the RIS stops beam tracking and goes back to Step 0. For the stopping criteria, we expect that the RIS beam tracking stops when the communication between the UE and the BS is terminated. More specifically, the RIS beam tracking stopping criteria include the following: (i) There is a blockage between the RIS and the BS/UE. (ii) The BS has switched to serve another UE. (iii) The communication session between the BS and the UE is terminated. Note that the RIS beam tracking and stopping criteria detection functionalities are not currently handled in this work. We highlight them as future research directions in Section IX-B.

Fig. 5 shows the timing of each step of the proposed vision-aided standalone RIS system operation with respect to the 3GPP 5G initial accessed process described in Section III. This demonstrates the compatibility of the proposed beam selection with the 3GPP 5G protocol.

D. Deployment Considerations

Here, we discuss several essential considerations for the implementation of the proposed RIS system, including (i) power consumption and implementation cost, (ii) sensing and processing delay, and (iii) compatibility with different 5G deployment methods.

[1]By monitoring the signal power on the wireless band, the RIS does not need to decode the messages exchanged between the BS and the UE in the initial access process.

1) Power Consumption and Implementation Cost: One motivation of the RIS is to reduce power consumption and implementation cost [33]. We anticipate that the proposed sensing-aided RIS will have reasonable power consumption and implementation cost due to the following reasons. (i) The proposed RIS adopts a power-efficient array architecture where only a few baseband processing chains are needed for the sparse channel receivers. Moreover, the sparse wireless channel receivers do not need the transmitter functionality. These reduce power consumption and implementation cost. (ii) The proposed RISs do not need dedicated and additional signaling with the BS or user, which helps improve wireless resource efficiency and overall energy efficiency. (iii) Compared to other modalities of sensors such as radar and LiDARs, the proposed RIS adopts cameras that often have lower implementation costs. (iv) Although the adopted DL models consume some power, they can be more efficiently implemented using techniques model quantization and pruning techniques [34] and efficient DL hardware [35]. Moreover, the object detector may utilize more advanced and lightweight models [36].

2) Sensing and Processing Delay: The proposed RIS uses sensors to capture information about the communication environment and select the RIS beam. Obtaining and processing the sensing information, however, may lead to time delay. The impact of this delay depends on the dynamics in the system. For example, the time delay may pose challenges to serving highly mobile users. In this case, it might be interesting to leverage the sensing information to predict future RIS beams [24], which can help compensate for the time delay. Note that this delay is not modeled in this paper.

3) Compatibility With Different 5G Deployment Methods: There are several popular deployment methods for 5G. For instance, the non-standalone deployment utilizes the existing 4G LTE infrastructure including the evolved packet core (EPC) and the eNodeB. The pure 5G deployment, however, adopts the 5G gNodeB and next-generation core network (NGC). Since the proposed standalone RIS operation does not rely on explicit signaling with the infrastructure, it is compatible with both EPC and NGC. Since we mainly consider high-frequency systems, the proposed RIS operation is compatible with the gNodeBs that operate on the 5G frequency range 2 (FR2) high-frequency bands. However, the proposed RIS operation may not be compatible with the eNodeBs that only support the frequency range 1 (FR1) sub-6 GHz bands.

VI. Deep Learning Modeling

The proposed sensing-aided transparent RIS approach depends on the capability of the RIS to leverage its sensors (cameras in this paper) to determine the set of candidate beams. We employ the powerful learning capabilities of computer vision and DL to achieve this task, which we divide into two sub-tasks, namely candidate UE detection in the field of view, and candidate beam set prediction. Next, we describe the adopted DL models for these two sub-tasks.

A. Candidate UE Detection

Convolutional neural networks (CNNs) have been extensively investigated for visual object detection and have demonstrated promising performance. Therefore, we adopt a CNN-based object detection model. Since the wireless environment is typically changing quickly, the object detector in the proposed framework for UE-side candidate beam set prediction needs to satisfy an essential requirement: the capability to produce fast object detections of high quality. To that end, the YOLOv3 object detector [37] is selected due to its fast prediction speed, high accuracy, and ease of implementation, which make it well-suited for real-world deployments. Given an image, the YOLOv3 model outputs a class index $c \in \mathbb{Z}$, and a bounding box vector $\mathbf{b} \in \mathbb{R}^{4\times 1}$ for each detected candidate UEs. The $\mathbf{b}$ consists of the x-center, the y-center, the width, and the height of the bounding box. We refer readers to [37] for more details. Note that the object detection algorithm may detect some non-user objects. These objects will not respond to the SSBs but they may cause adding additional beams in the predicted UE-side candidate beam set.

B. Candidate Beam Set Prediction

Based on the class and bounding box information of the candidate UEs, we now design an NN that can predict the UE-side candidate beam set. Next, we will describe the key components of the proposed NN for the UE-side candidate beam set prediction, namely the input/output representation, the NN architecture, and the loss function and learning model.

1) Input/Output Representation and Normalization: Given one image, the YOLOv3 model detects $\hat{U}$ candidate UEs. For each candidate UE, the YOLOv3 model outputs a class index $c \in \mathbb{Z}$, and a bounding box $\mathbf{b} \in \mathbb{R}^{4\times 1}$. To make the training process of the NN faster and more stable, we convert the class c to a one-hot vector $\bar{\mathbf{c}}$, and we normalize the bounding box $\mathbf{b}$ by the size of the image, w and h. The normalized bounding box is denoted by $\bar{\mathbf{b}}$. Then the one-hot representation of the class $\bar{\mathbf{c}}$ and the normalized bounding box $\bar{\mathbf{b}}$ are concatenated to the candidate UE information vector $\mathbf{v} = [\bar{\mathbf{c}}^T, \bar{\mathbf{b}}^T]^T$. Finally, the input matrix $\mathbf{V}$ to the NN architecture is written as $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_{\hat{U}}, \mathbf{0}, \ldots, \mathbf{0}]$. Note that we pad $(U_{max} - \hat{U})$ zero-vectors since the number of detected UEs varies from image to image. U_{max} denotes the maximum number of candidate UEs that exist in any image. To construct the desired output of the NN, we first obtain the optimal UE-side beam set $\mathbf{Q}^\star$ corresponds to the image as shown by (9) with an exhaustive search over the codebook $\mathcal{Q}$. Then we convert $\mathbf{Q}^\star$ into a multi-hot vector $\mathbf{t}^\star = \left[t_1^\star, \ldots, t_{|\mathcal{Q}|}^\star\right]$. With $\mathcal{Q}_j$ denoting the j-th beam in $\mathcal{Q}$, the j-th element of $\mathbf{t}^\star$, $t_j^\star$ satisfies

$$t_j^\star = \begin{cases} 1 & \mathcal{Q}_j \in \mathbf{Q}^\star, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

2) NN Architecture: As shown by Fig. 6, we propose an NN architecture that effectively predicts the UE-side candidate beams from the class and bounding box information of all detected candidate UEs in an image. The proposed NN architecture first applies the same stack of fully connected NN layers on each candidate UE's information vector (each column of $\mathbf{V}$) to extract high-level features. These fully-connected layers adopt the ReLU activation function given by $f_{\text{ReLU}}(x) = \max(x, 0)$. After this feature extraction,

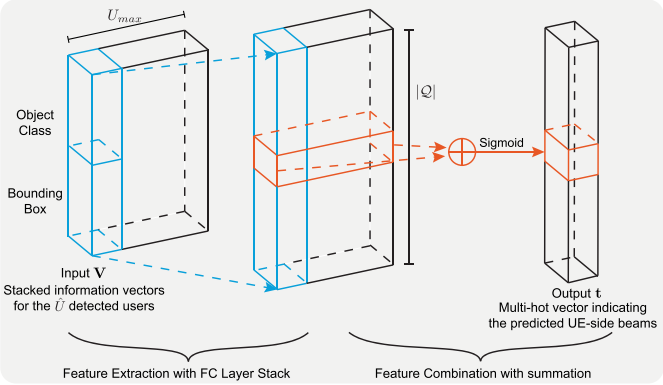


Fig. 6. The proposed NN architecture for predicting candidate beam set from detected UE information. The NN architecture first applies the same stack of FC layers on all the UE information vectors to project them into $|\mathcal{Q}|$-dimensional vectors. The $|\mathcal{Q}|$-dimensional vectors are then added up and activated by the sigmoid function.

each input candidate UE information vector is transformed to a $|\mathcal{Q}|$-dimensional vector. Then these $|\mathcal{Q}|$-dimensional vectors are combined by a summation operation. Since the desired output of the NN is designed as a multi-hot vector, the sigmoid activation is applied to the combined vector to restrict its value into the range $(0, 1)$. After the sigmoid activation, we obtain the output vector, $\mathbf{t} \in \mathbb{R}^{|\mathcal{Q}|\times 1}$.

The proposed NN architecture has two advantages for predicting the UE-side candidate beam set from the candidate UE information vectors. (i) It reuses the same stack of fully connected layers to extract features from different candidate UE information vectors. This aligns with the intuition that all candidate UEs are equivalent for the NN architecture, thus, they should be processed in the exact same way. Reusing this same stack of fully connected layers also reduces the complexity of the proposed NN architecture, which stabilizes the training process and reduces the computational complexity of the inference process. (ii) The output of the proposed NN architecture does not rely on the order of the input candidate UE information vectors. This is achieved by reusing the same stack of fully connected layers and the summation operation that combines features from all candidate UE information vectors. Thus, the proposed NN architecture can be more robust by not overfitting to the order of the input candidate UE information vectors. We validate our intuitions on the NN structure by numerical results presented in Section VIII-B.

3) Loss Function and Learning Model: The NN is designed to predict the UE-side candidate beam set based on the candidate UEs detected by the YOLOv3 model. This can be modeled as a multi-class classification problem. Therefore, we adopt a classification learning model. We train the NN by supervised learning and employ the cross-entropy loss function expressed by

$$L_{CE}(\omega_{NN}) = -\sum_{i=1}^{N_{tr}} \sum_{j=1}^{|\mathcal{Q}|} t_{i,j}^{\star} \log_2(t_{i,j}), \tag{12}$$

where $t_{i,j}^{\star}$ is the j-th element of the *desired* output vector $\mathbf{t}_i^{\star}$ of the i-th training sample, and $t_{i,j}$ is the j-th element of the NN's output vector of the i-th training sample. N_{tr} is the number of training samples. ω_{NN} denotes the trainable parameters of the NN.

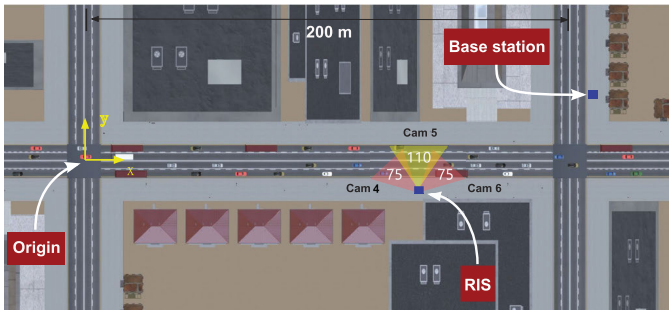


Fig. 7. A top-view of the adopted simulation scenario of the ViWi dataset. This scenario models a busy downtown area with a variety of objects including cars, buses, trucks, trees, and buildings.

VII. Dataset and Performance Metrics

In this section, we first explain in detail the considered simulation setup. Second, we elaborate on the generation process of the dataset, which is later used to train and evaluate the NN. Then, we introduce the metrics used to evaluate the UE-side beam set prediction performance.

A. Simulation Setup

We propose to utilize sensing to enable beamforming for standalone RIS. Hence, realistic wireless and visual modelings are essential for our simulation. Therefore, we generate the training and test data with the ViWi dataset [38]. The ViWi dataset provides co-existing wireless and visual data based on accurate ray tracing. It comprises sequences of RGB frames, wireless channels, and user link statuses. They are generated from a large synthetic outdoor environment depicting a downtown street with multiple moving objects. To simulate our RIS-aided system, we construct a new scenario based on the ViWi scenario 1. A top view of our scenario is presented in Fig. 7. The BS is located on the vertical street at the upper right. The UEs are the moving vehicles on the main street. The RIS is installed at the side of the main road to aid communication between the BS and the UEs. In the simulation, since we focus on the RIS operations, we assume the BS and the UEs to be single-antenna for simplicity. Note that, however, the proposed RIS operation is compatible with the multi-antenna BSs in the 3GPP 5G. The RIS is equipped with a uniform planar array (UPA) with 32 columns and 8 rows of reflecting elements, *i.e.*, the total number of reflecting elements of the RIS is 256. Three cameras ("Camera 4", "Camera 5" and "Camera 6") are deployed at the RIS as shown in Fig. 7. The central "Camera 5" has a 110° field of view while the side cameras' field of views are 75°. In the following simulations, we focus on the UEs in the views of "Camera 4" and "Camera 5". Note that, we adopt this camera setting as an example to demonstrate the proposed RIS operations. However, the optimal setting for the camera may need further investigation.

B. Data Generation

We first generate 10000 scenes with the ViWi dataset. The BS and RIS are fixed in positions for all scenes. The

vehicles are placed at different positions in each scene, thus, different scenes produce different wireless environments and UE channels. Then, for "Camera 4", each scene consists of an image $\mathbf{X}$, and the channels of all UEs in the view of the camera. Note that we only keep the data where the UE does not have direct links with the BS, *i.e.*, all the paths between the BS and the UE go through the RIS. We obtain the optimal UE-side candidate beam set $\mathbf{Q}^\star$ for each image according to (8b) and (9) by exhaustively searching over the codebook $\mathcal{Q}$. The optimal UE-side candidate beam set for each image is then converted to the multi-hot representation $\mathbf{t}^\star$. Then $\mathbf{X}$ and $\mathbf{t}^\star$ form a data point $(\mathbf{X}, \mathbf{t}^\star)$. We apply the same process to the data from "Camera 5". After processing all scenes, we have a dataset for "Camera 4" with 9384 data points, and a dataset for "Camera 5" with 5955 data points.

To generate the datasets for the beam set prediction NNs, a fine-tuned YOLOv3 model is first applied to all images in the two datasets to obtain the candidate UE information vectors. Then, each data point $(\mathbf{X}, \mathbf{t}^\star)$ is converted to $(\mathbf{V}, \mathbf{t}^\star)$. The two datasets for "Camera 4" and "Camera 5" are split into training and test datasets using an 80%-20% data split. To evaluate the generalizability of our proposed approach, we ensure that the test datasets consist of data that are unseen in the training datasets. Note that two NNs are separately trained and evaluated on the two datasets for "Camera 4" and "Camera 5" since they have different camera angles.

C. Performance Metrics

Here, we present the metrics followed to evaluate the quality of the NNs' UE-side beam set prediction. In the NNs' output vector $\mathbf{t}$, each element represents a promising score of the corresponding beam in $\mathcal{Q}$. To evaluate the prediction performance, we first apply the following unit step function on each element of the output vector, $f_{\text{step}}(x) = u(x-\delta)$, where we use $\delta = 0.5$ as the threshold. By applying the threshold to $\mathbf{t}$ as (11), we obtain $\hat{\mathbf{t}}$. Let $\hat{t}_j$ denote the j-th element of $\hat{\mathbf{t}}$, the predicted UE-side candidate beam set $\hat{\mathbf{Q}}$ can then be written as

$$\hat{\mathbf{Q}} = \left\{ \mathcal{Q}_j | \hat{t}_j = 1 \right\}. \tag{13}$$

The metrics adopted to evaluate the performance of the UE-side candidate beam set prediction are the accuracy and the recall. They are defined as follows:

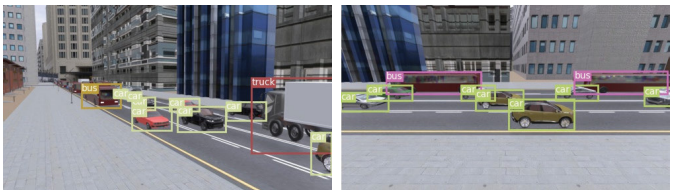
$$\text{Acc} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \frac{\left| \mathbf{Q}_i^\star \cap \hat{\mathbf{Q}}_i \right|}{\left| \hat{\mathbf{Q}}_i \right|} \tag{14}$$

$$\text{Recall} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \frac{\left| \mathbf{Q}_i^\star \cap \hat{\mathbf{Q}}_i \right|}{|\mathbf{Q}_i^\star|}, \tag{15}$$

where N_{test} is the number of data samples in the test dataset. $\mathbf{Q}_i^\star$ and $\hat{\mathbf{Q}}_i$ denote the optimal and the predicted UE-side candidate beam set for the i-th data sample, respectively.

VIII. Simulation Results

Here, we evaluate the performance of the proposed standalone RIS beam selection. First, we present the performance of the candidate UEs detection. Then, we present the accuracy and recall performance of the proposed NN structure in predicting candidate beam sets. Next, we study the amount of data required to train the proposed NN. After that, we show the efficacy of the proposed RIS beam selection in terms of the achievable rate and beam training overhead.



(a) An image from "Camera 4" (b) An image from "Camera 5"

Fig. 8. This figure shows two example images taken by the RIS cameras and illustrates the UE object class and bounding box information annotated by the fine-tuned YOLOv3 model.

A. Can YOLOv3 Detect Candidate UEs?

Candidate UE detection is the first step of the proposed ML framework for predicting UE-side candidate beam sets. Therefore, the quality of the candidate UE detection is essential for the downstream task and the performance of the framework. Hence, we first demonstrate the performance of the YOLOv3 object detector. In Fig. 8, we apply the fine-tuned YOLOv3 model on two images from "Camera 4" and "Camera 5", and let the YOLOv3 model annotate the class and bounding box information on the detected candidate UEs. This figure shows that the YOLOv3 model can accurately detect the candidate UEs in the two cameras and produce high-quality information (class and bounding box) on these UEs.

B. Does the Proposed NN Structure Learn Better?

In Section VI-B, we mentioned two key features of the proposed NN structure: (i) Reusing the fully connected layer stack on all the UE information vectors, and (ii) combining the information of the detected UEs by the summation operation. These two features are expected to improve the performance of the candidate beam set prediction. To analyze the effectiveness of the proposed NN structure and verify the intuitions used in its design, we study the training process of the NN. Fig. 9 presents the learning curves of the proposed NN structure compared with two variants trained on the "Camera 5" dataset. The first variant adopts vanilla fully connected NN. The second variant reuses the same fully connected layer stack on the information vectors from all candidate UEs, but it concatenates the resulting high-level feature vectors instead of applying the summation operation. From the training and the test loss in Fig. 9, we see that the vanilla fully connected NN overfits to the training dataset and its loss diverges on the test set. For the second variant, the test loss can converge along with training iterations. This indicates that **reusing the fully connected layer stack stabilizes the training process**. The proposed NN structure achieves the lowest loss on the test dataset, and the gap between the training and the test losses is the smallest. This implies that **combining the information from different UEs with the summation operation improves**

TABLE I
ACCURACY AND RECALL PERFORMANCE OF THE ML FRAMEWORK WITH THREE DIFFERENT NN STRUCTURES TRAINED ON 80% OF THE DATA

Model	Accuracy		Recall	
	Cam4	Cam5	Cam4	Cam5
Fully connected	71.2%	32.6%	70.1%	25.5%
Reusing fully connected layer stack	92.3%	92.1%	87.6%	71.3%
Reusing fully connected layer stack + sum operation	**94.2%**	**92.9%**	**92.1%**	**86.4%**

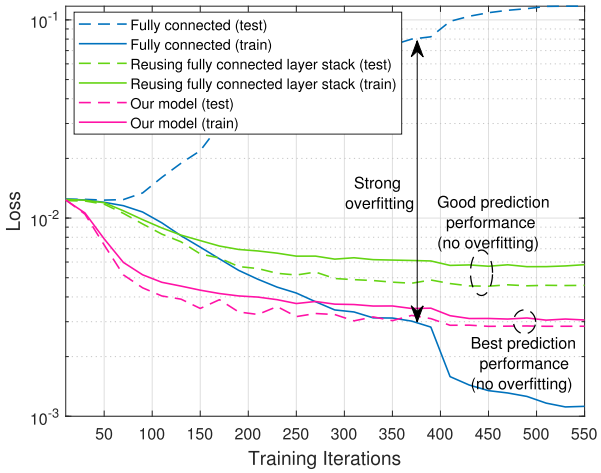


Fig. 9. This figure compares the learning curves (train and test) of the proposed neural network structure and two other baseline models when trained on the "Camera 5" dataset.

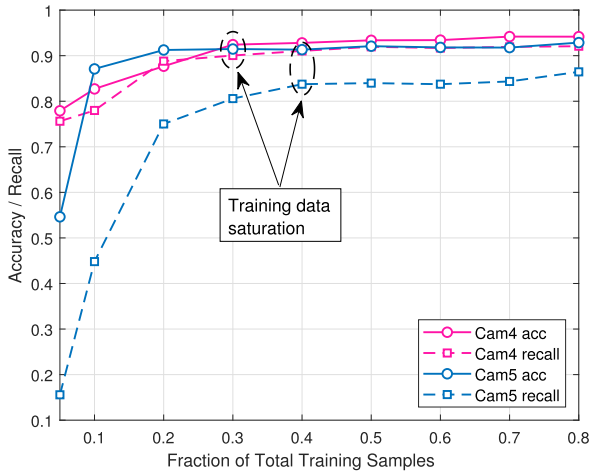


Fig. 10. The accuracy and recall performance of the proposed neural network structure when trained on different training dataset sizes (as fractions of the full training set). The figure shows that only 30%-40% of the dataset (which correspond to 2500-3000 data points) is enough to achieve around 90% accuracy and recall.

the robustness of the model. These results highlight the effectiveness of the proposed NN structure. Next, we will further evaluate the proposed NN structure in terms of the accuracy and recall performance.

C. Can the Proposed NN Structure Predict the Candidate Beams More Accurately?

After the YOLOv3 detects the candidate UEs, the proposed NN structure predicts the UE-side candidate beam set. The quality of the predicted candidate beam set directly determines the final transmission rate and the required beam training overhead. Therefore, in Table I, we present the accuracy and recall performance of the proposed NN structure compared with its two variants. Our proposed NN structure achieves 94.2% and 92.9% on the "Camera 4" dataset for accuracy and recall, respectively. On the "Camera 5" dataset, the accuracy and recall performances of the proposed NN structure are 92.1% and 86.4%. These results highlight that the proposed NN structure can accurately predict the UE-side candidate beam set.

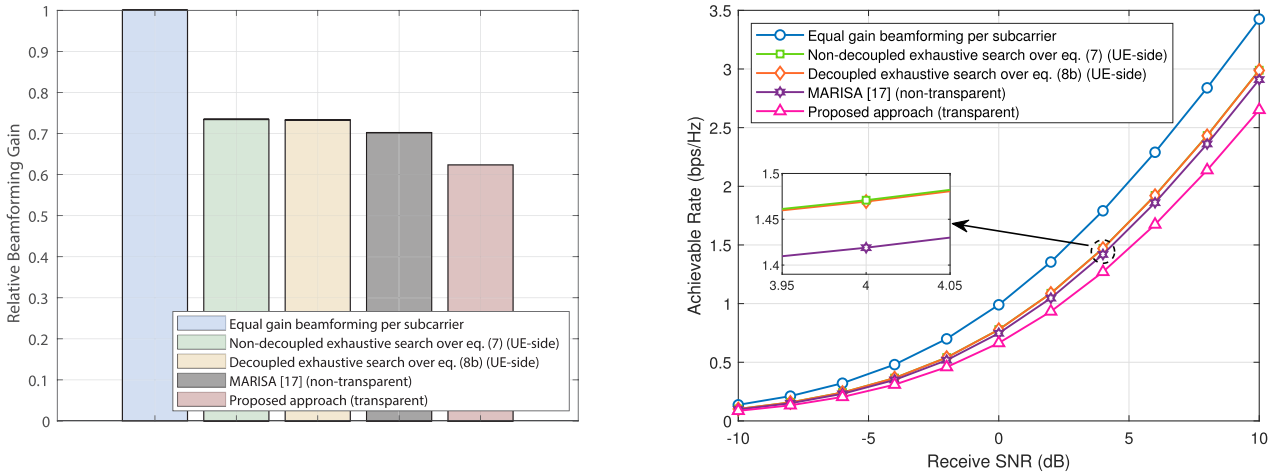
Comparing the proposed NN structure with the two variants, it can be seen that reusing the fully connected stack offers significant improvements on both datasets. For the dataset of "Camera 5", the accuracy increases by **59.5%**. Moreover, by reusing the fully connected stack and combining information of candidate UEs with the summation operation, our NN structure results in the highest performance on the test datasets. The recall performance for the "Camera 5" dataset is improved by **15.1%**. This again emphasizes that the two key features of the proposed NN structure help stabilize the training and achieve better performance in the beam set prediction.

D. How Much Data Is Needed to Train the Beam Prediction NN?

The size of the training data set is crucial for ML models when deployed in the real world. To that end, we draw insights on the dataset size required to train the proposed ML framework. Fig. 10 plots the test accuracy and recall obtained on datasets "Camera 4" and "Camera 5" versus the fraction of data used to train the proposed UE-side candidate beam prediction NNs. As can be seen from this figure, more training data helps improve the accuracy and recall performance. The accuracy and recall start to saturate after 30% and 40% of data are used in the training process for "Camera 4" and "Camera 5", respectively. This corresponds to 2815 data points for "Camera 4" and 2382 data points for "Camera 5". Training the proposed NN structure only requires a relatively small dataset since the proposed NN predicts the candidate beam set from the UEs detected by the YOLOv3 model instead of the raw RGB images. Besides, only 500 samples are used to fine-tune the YOLOv3 model for each dataset. These results show that **the proposed ML framework is data-efficient in the training process**.

E. How Good Are the Beamforming Gain and Achievable Rate?

Fig. 11 shows the relative beamforming gain and achievable rate performance of the proposed transparent RIS beam



(a) Relative Beamforming Gain (b) Achievable Rate

Fig. 11. This figure compares the beamforming gain and achievable rates of the proposed approach and four benchmark methods. The proposed approach does not rely on dedicated and additional signaling with the BS/UE, which is transparent to the 3GPP 5G initial access. The benchmark methods require channel information, large training overhead, and/or dedicated signaling.

selection compared with four baseline beamforming methods. (i) The first baseline method is the equal gain beamforming per subcarrier. It uses the conjugate of the channel vectors as the beamforming vectors, which serves as an upper bound of the achievable rate. (ii) The second baseline method is the non-decoupled UE-side exhaustive search with a codebook size of 256. Given the optimal BS-side beam, this method tries all UE-side beams in the UE-side codebook to obtain the highest achievable rate in (7). (iii) The third baseline method adopts the optimal BS-side beam, and performs the UE-side exhaustive search over the decoupled UE-side beam selection problem in (8b) with a codebook size of 256. The above three baseline beamforming methods are not practical. The equal gain combine applies different beamforming on different subcarriers, which violates the implementation constraint of the RIS. The two exhaustive search methods require either the channel information on both the BS-side and UE-side or the large beam training (sweeping) overhead. (iv) The fourth baseline method is a self-configuring RIS beamforming approach called MARISA [17]. This method configures the RIS beam by measuring the power of the pilot signals transmitted by the BS and the UEs. Since it relies on the pilot signal from both the BS and the UEs, this approach is not transparent to 3GPP 5G initial access. Moreover, this approach adopts beam sweeping over 256 probing beams, which leads to large beam training overhead. In our approach, we assume that the RIS uses the optimal BS-side beam (explained in Section IV-A), and the best UE-side beam within the predicted UE-side candidate beam set.[2] Our approach is transparent to 3GPP 5G initial access, which does not rely on dedicated and additional signaling from the BS and the UE.

Fig. 11(a) presents the beamforming gain relative to the equal gain beamforming. The proposed transparent approach achieves 85% beamforming gain compared to the exhaustive search methods that require 256 beam training iterations. Compared to [17], our proposed transparent approach achieves 89% beamforming gain which corresponds to only -0.4 dB loss in SNR. Fig. 11(b) shows the achievable rate performance that increases with the SNR. Notably, the performance gap between the non-decoupled (7) and decoupled (8b) exhaustive search methods is minimal. This suggests that the performance degradation due to decoupling the RIS beam selection problem in (7) is small in the considered scenario. Furthermore, **our proposed transparent approach obtains high performance in the SNR range from -10 to 10 dB compared to the exhaustive search methods (upper bounds). For example, at 0 dB receive SNR, it achieves 86.1% of the exhaustive search data rate**. Moreover, Compared to [17], our approach achieves only slightly lower achievable rate performance. Again, [17] is a non-transparent RIS beamforming approach that requires the pilot signals from both the BS and UE. In contrast, the proposed approach is transparent to 3GPP 5G initial access, which does not rely on pilot signals and other dedicated and additional signalings from the BS and UE.

F. How Much Beam Training Overhead Is Required?

One goal of this paper is to reduce the beam training overhead of the standalone RIS. In Fig. 12, we study the effect of the size of the UE-side candidate beam set on the achievable rate at 0 dB receive SNR. In the previous simulations, we apply the step function in Section VII-C to the output vector of the NN, and construct the candidate beam set as shown by (13). Here, the candidate beam set consists of the k beams corresponding to the top-k highest value in $\mathbf{t}$, the output vector of the NN. As shown in Fig. 12, when the candidate beam set size increases, the achievable rate performance of our proposed standalone RIS beam selection approaches the performance of the exhaustive search. **When the RIS sweeps over only 12 beams in the UE-side candidate beam set, the proposed approach can achieve 96.4% of the data rate achieved by the RIS exhaustive**

[2]In practice, the UE may respond to the first decodable SSBs, and then the RIS selects the UE-side beam corresponding to those SSBs. However, the decodable SSBs vary from one UE to another depending on the receive sensitivity of the UE manufacturer. In the simulation, we assume that the RIS selects the best UE-side beam within the UE-side candidate beam set.

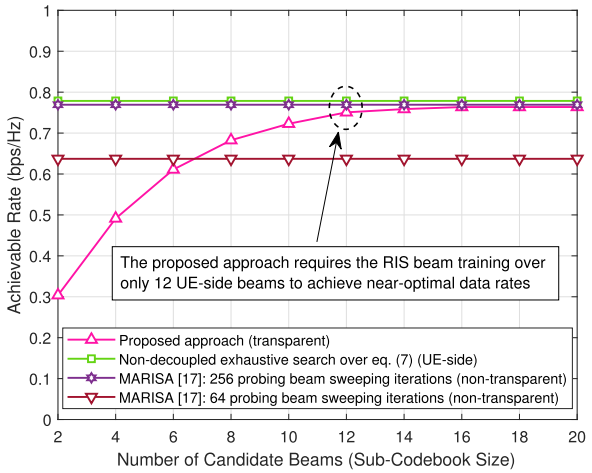


Fig. 12. This figure shows the achievable rate of the proposed standalone RIS approach with different sizes of the UE-side candidate beam set. The beam training overhead of the proposed approach is compared with two baseline approaches.

search over 256 UE-side beams. Note that this exhaustive search provides the optimal achievable rate given the BS-side beam. Therefore, approaching this upper bound indicates that the proposed standalone RIS beam selection can efficiently reduce the beam training overhead with little negative effect on the achievable rates of the RIS-aided system. [17] can achieve high achievable rate performance when sweeping over 256 probing beams. However, when 64 probing beams are employed, the performance of [17] degrades noticeably. [17] typically requires a large number of probing beams (*e.g.* 256) to sense the communication environment with high spatial resolution and obtain high-performance RIS communication beams. In contrast, the proposed approach senses the communication environment with wireless and visual sensors. With this sensing information, the RIS can achieve high achievable rates when sweeping over only 12 UE-side candidate beams. This again highlights the potential of leveraging sensing to guide RIS beam selection and reduce beam training overhead.

IX. Discussion and Future Work

In this section, we provide some insights on how to extend the proposed 3GPP 5G transparent RIS operation to more dynamic and complex scenarios, and discuss the enabling features.

A. Beam Tracking

Efficient beam tracking is essential for the RIS to support mobile UEs in dynamic environments. The 3GPP 5G beam refinement process includes the following: (i) The BS performs beam sweeping within a subset of transmit beams while the UE is maintaining a receive beam, and (ii) the UE reports the beam measurement results and the selected beam(s) to the BS. This beam refinement process, however, cannot be directly employed by the standalone RIS operation. The beam refinement requires signaling between the RIS and the BS/UE because the RIS lacks knowledge of the content of some messages exchanged between the BS and UE, *e.g.*, the Msg 1 and Msg 2. As discussed in Section V-C, however, the RIS's sensing capability can be utilized to achieve transparent RIS beam tracking. With the sensing capability, the RIS can obtain rich information about the UE position, its mobility pattern, and the environment layout, from which the RIS can infer its future UE-side beam from the current/previous beam sequence.

B. Beam Tracking Stopping Criteria

The RIS should stop the beam tracking when it detects that the communication between the BS and the UE ends. As discussed in Section V-C, the RIS beam tracking stopping criteria include: (i) There is a blockage between the RIS and the BS/UE, (ii) the BS has switched to other UEs, and (iii) the communication session is terminated. Note, however, that since the standalone RIS does not communicate with the BS or the UE, it can not directly know that the session has ended and it also can not access the configuration of the wireless resource allocation. Therefore, the RIS needs to infer and keep track of the wireless resource associated with the UE. Having a spectrum sensing capability at the RIS may be one step towards this objective. It remains, however, an interesting and open research problem. Apart from the spectrum sensing approach, other sensing modalities at the RIS can also be leveraged to detect the beam tracking criteria. For instance, the sensors (such as camera, radar, LiDAR, etc.) can detect and proactively predict the potential blockages between the BS/UE and the RIS and assist the stopping criteria detection for the standalone RIS beam tracking.

C. Multi-User Scenario

The proposed transparent RIS operations focus on the single-user scenario. However, it could be interesting to extend that to support the multi-user scenarios. Since different UEs can be allocated to different time and frequency resources in an uncontrolled way from the transparent RIS's perspective, the RIS may use one reflection beam to serve multiple UEs. This multi-user beam design problem may be addressed in various ways, such as (i) dividing the RIS into multiple sub-arrays with each sub-array configured to support the communication for one UE, and (ii) employing multi-lobe and composite beams [39], [40]. Using the multi-user beam design, the proposed transparent RIS operations may be extended to support the initial access for multiple UEs. Take the sub-array beam design as an example, the RIS can first serve a single user "UE 1" using its first sub-array. After that, to establish connections for other UEs, the RIS predicts the candidate UE-side beam set from the visual information. Then, the RIS sweeps over the candidate UE-side beam set using its second sub-array while maintaining the beamforming of the first sub-array to serve UE 1. By monitoring the UE preamble on the PRACH, the RIS can start to serve a new UE with its second sub-array. While the proposed transparent RIS operation may be extended to the multi-user initial access process, maintaining the links for multiple users is challenging. For the multi-user beam tracking and stopping criteria detection, the RIS may need to identify the users being served from the sensory data and associate them with the previously and currently

used UE-side beams. This user identification problem has been studied in [41] for a visual sensing-aided mmWave BS. However, it remains an open problem for the transparent RIS which lacks dedicated signaling with the BS and UEs.

X. Conclusion

In this paper, we investigated the feasibility of enabling 3GPP 5G transparent RIS operation using sensing-based perception. Utilizing the sensing capability at the RIS, we proposed a standalone RIS beam selection operation that does not need any dedicated signaling with the BS and the UE, and is compatible with the 3GPP 5G initial access process. To guide the RIS beam selection in the proposed standalone RIS operation, we developed an ML framework and an NN architecture that leverage the visual data captured by cameras installed at the RIS to predict the candidate set of beams. To evaluate the developed solution, we conducted extensive simulations based on a high-fidelity synthetic dataset gathering co-existing wireless and visual data. When benchmarked against other NN architectures, the proposed one shows clear advantages in both learning stability and prediction accuracy. In particular, the simulation results demonstrated that the proposed ML framework can accurately predict the UE-side candidate beam set. Further, these results showed that the proposed standalone RIS system can achieve near-optimal spectral efficiency with significantly reduced beam training overhead. This highlights the potential of leveraging sensing-based perception to develop 3GPP 5G transparent RIS operations.

Appendix

When the BS-side channel and UE-side channel only contain the LoS path, all the subcarriers have the same channels, *i.e.*, $\mathbf{h}_{\mathrm{T},k} = \mathbf{h}_{\mathrm{T}}$, $\forall k = 1,\ldots,K$, and $\mathbf{h}_{\mathrm{R},k} = \mathbf{h}_{\mathrm{R}}$, $\forall k = 1,\ldots,K$. Without loss of generality, let us assume

$$\begin{aligned}\mathbf{h}_{\mathrm{T}} &= \left[a_1 e^{j\phi_1},\ldots,a_M e^{j\phi_M}\right]^T\\ \mathbf{h}_{\mathrm{R}} &= \left[b_1 e^{j\omega_1},\ldots,b_M e^{j\omega_M}\right]^T\\ \mathbf{a} &= \left[c_1 e^{j\beta_1},\ldots,c_M e^{j\beta_M}\right]^T,\end{aligned} \tag{16}$$

where $a_m, b_m, c_m \geq 0$, $\phi_m, \omega_m, \beta_m \in [-\pi,\pi)$, $\forall m = 1,\ldots,M$. Since $\boldsymbol{\psi}^\star, \mathbf{p}^\star, \mathbf{q}^\star \in \mathcal{O}$, they can be written as

$$\begin{aligned}\boldsymbol{\psi}^\star &= \left[e^{j\lambda_1},\ldots,e^{j\lambda_M}\right]^T\\ \mathbf{p}^\star &= \left[e^{j\lambda_{p,1}},\ldots,e^{j\lambda_{p,M}}\right]^T\\ \mathbf{q}^\star &= \left[e^{j\lambda_{q,1}},\ldots,e^{j\lambda_{q,M}}\right]^T,\end{aligned} \tag{17}$$

where $\lambda_m, \lambda_{p,m}, \lambda_{q,m} \in [-\pi,\pi)$, $\forall m = 1,\ldots,M$.

With $\mathbf{h}_{\mathrm{T},k} = \mathbf{h}_{\mathrm{T}}$ and $\mathcal{P} = \mathcal{O}$, (8a) can be re-written as

$$\begin{aligned}\mathbf{p}^\star &= \arg\max_{\mathbf{p}\in\mathcal{O}} \frac{1}{K}\sum_{k=1}^{K}\left|\left(\mathbf{h}_{\mathrm{T}}\odot\mathbf{p}\right)^H\mathbf{a}^*\right|^2\\ &= \arg\max_{\mathbf{p}\in\mathcal{O}}\left|\left(\mathbf{h}_{\mathrm{T}}\odot\mathbf{p}\right)^H\mathbf{a}^*\right|^2\\ &= \arg\max_{\phi_m,\beta_m,\lambda_{p,m}}\left|\sum_{m=1}^{M} a_m c_m e^{-j(\phi_m+\beta_m+\lambda_{p,m})}\right|^2\end{aligned} \tag{18}$$

$$\Leftrightarrow\quad \phi_{m_1}+\beta_{m_1}+\lambda_{p,m_1} = \phi_{m_2}+\beta_{m_2}+\lambda_{p,m_2},\ \forall m_1, m_2. \tag{19}$$

Similarly, (8b) can be re-written as

$$\begin{aligned}\mathbf{q}^\star &= \arg\max_{\mathbf{q}\in\mathcal{O}}\left|\left(\mathbf{h}_{\mathrm{R}}\odot\mathbf{q}\right)^H\mathbf{a}\right|^2\\ &= \arg\max_{\phi_m,\beta_m,\lambda_{q,m}}\left|\sum_{m=1}^{M} b_m c_m e^{-j(\phi_m-\beta_m+\lambda_{q,m})}\right|^2\end{aligned} \tag{20}$$

$$\Leftrightarrow\quad \omega_{m_1}-\beta_{m_1}+\lambda_{q,m_1} = \omega_{m_2}-\beta_{m_2}+\lambda_{q,m_2},\ \forall m_1, m_2. \tag{21}$$

Adding up (19) and (21), we can derive

$$\begin{aligned}&\phi_{m_1}+\omega_{m_1}+\lambda_{p,m_1}+\lambda_{q,m_1}\\ &\qquad = \phi_{m_2}+\omega_{m_2}+\lambda_{p,m_2}+\lambda_{q,m_2},\ \forall m_1, m_2.\end{aligned} \tag{22}$$

Similarly, (5) can be re-written as

$$\begin{aligned}\boldsymbol{\psi}^\star &= \arg\max_{\boldsymbol{\psi}\in\mathcal{O}}\left|\left(\mathbf{h}_{\mathrm{R}}\odot\mathbf{h}_{\mathrm{T}}\right)^T\boldsymbol{\psi}\right|^2\\ &\Leftrightarrow\ \phi_{m_1}+\omega_{m_1}+\lambda_{m_1} = \phi_{m_2}+\omega_{m_2}+\lambda_{m_2},\ \forall m_1, m_2.\end{aligned} \tag{23}$$

Since $\mathbf{p}^\star\odot\mathbf{q}^\star = \left[e^{j(\lambda_{p,1}+\lambda_{q,1})},\ldots,e^{j(\lambda_{p,M}+\lambda_{q,M})}\right]^T$, it can be seen from (22) that $\mathbf{p}^\star\odot\mathbf{q}^\star$ satisfies (23), therefore, $\mathbf{p}^\star\odot\mathbf{q}^\star$ is an optimal solution of (5). Thus, they are an optimal solution of (7).

References

[1] S. Jiang, A. Hindy, and A. Alkhateeb, "Camera aided reconfigurable intelligent surfaces: Computer vision based fast beam selection," presented at the IEEE Int. Conf. Commun., May 30, 2023.

[2] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019.

[3] C. Pan et al., "Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 14–20, Jun. 2021.

[4] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Reconfigurable intelligent surfaces: Three myths and two critical questions," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 90–96, Dec. 2020.

[5] G. C. Trichopoulos et al., "Design and evaluation of reconfigurable intelligent surfaces in real-world environment," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 462–474, 2022.

[6] Y. Liu et al., "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1546–1577, 3rd Quart., 2021.

[7] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 3, pp. 990–1002, Sep. 2020.

[8] C. Huang et al., "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[9] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," *IEEE Access*, vol. 9, pp. 44304–44321, 2021.

[10] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang, and M. Debbah, "Channel estimation for RIS-empowered multi-user MISO wireless communications," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 4144–4157, Jun. 2021.

[11] L. Wei et al., "Joint channel estimation and signal recovery for RIS-empowered multiuser communications," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4640–4655, Jul. 2022.

[12] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 4659–4663.

[13] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, "Intelligent reflecting surface meets OFDM: Protocol design and rate maximization," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4522–4535, Jul. 2020.

[14] S. Zhang, S. Zhang, F. Gao, J. Ma, and O. A. Dobre, "Deep learning optimized sparse antenna activation for reconfigurable intelligent surface assisted communication," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6691–6705, Oct. 2021.

[15] Y. Jin, J. Zhang, X. Zhang, H. Xiao, B. Ai, and D. W. K. Ng, "Channel estimation for semi-passive reconfigurable intelligent surfaces with enhanced deep residual networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11083–11088, Oct. 2021.

[16] S. Ren, K. Shen, Y. Zhang, X. Li, X. Chen, and Z.-Q. Luo, "Configuring intelligent reflecting surface with performance guarantees: Blind beamforming," *IEEE Trans. Wireless Commun.*, vol. 22, no. 5, pp. 3355–3370, May 2023.

[17] A. Albanese, F. Devoti, V. Sciancalepore, M. Di Renzo, and X. Costa-Pérez, "MARISA: A self-configuring metasurfaces absorption and reflection solution towards 6G," in *Proc. IEEE Conf. Comput. Commun.*, May 2022, pp. 250–259.

[18] *5G; NR; Physical Layer Procedures for Control Version 15.3.0 Release 15*, 3GPP, document TR 38.213, 2018.

[19] Y. Wang, M. Narasimha, and R. W. Heath Jr., "Towards robustness: Machine learning for mmWave V2X with situational awareness," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 1577–1581.

[20] M. Arvinte, M. Tavares, and D. Samardzija, "Beam management in 5G NR using geolocation side information," in *Proc. 53rd Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2019, pp. 1–6.

[21] J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Position aided beam prediction in the real world: How useful GPS locations actually are?" 2022, *arXiv:2205.09054*.

[22] S. Rezaie, J. Morais, E. de Carvalho, A. Alkhateeb, and C. N. Manchón, "Location- and orientation-aware millimeter wave beam selection for multi-panel antenna devices," 2022, *arXiv:2203.11714*.

[23] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *Proc. IEEE 91st Veh. Technol. Conf.*, May 2020, pp. 1–5.

[24] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided 6G wireless communications: Blockage prediction and proactive handoff," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10193–10208, Oct. 2021.

[25] U. Demirhan and A. Alkhateeb, "Radar aided 6G beam prediction: Deep learning algorithms and real-world demonstration," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 2655–2660.

[26] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 212–216, Feb. 2023.

[27] U. Demirhan and A. Alkhateeb, "Integrated sensing and communication for 6G: Ten key machine learning roles," *IEEE Commun. Mag.*, vol. 61, no. 5, pp. 113–119, May 2023.

[28] S. Jiang and A. Alkhateeb, "Computer vision aided beam tracking in a real-world millimeter wave deployment," in *Proc. IEEE Globecom Workshops*, Dec. 2022, pp. 142–147.

[29] X. Tong, Z. Zhang, J. Wang, C. Huang, and M. Debbah, "Joint multi-user communication and sensing exploiting both signal and environment sparsity," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1409–1422, Nov. 2021.

[30] R. W. Heath Jr., N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[31] T. S. Rappaport et al., "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.

[32] C. De Lima et al., "Convergent communication, sensing and localization in 6G systems: An overview of technologies, opportunities and challenges," *IEEE Access*, vol. 9, pp. 26902–26925, 2021.

[33] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[34] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, Oct. 2021.

[35] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.

[36] L. Jiao et al., "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.

[37] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[38] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, "ViWi: A deep learning dataset framework for vision-aided wireless communications," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, May 2020, pp. 1–5.

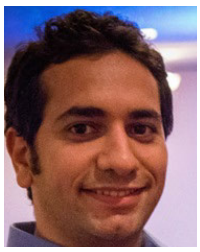
[39] N. Torkzaban and M. A. Amir Khojastepour, "Shaping mmWave wireless channel via multi-beam design using reconfigurable intelligent surfaces," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2021, pp. 1–6.

[40] N. Torkzaban, M. A. Amir Khojastepour, and J. S. Baras, "Codebook design for composite beamforming in next-generation mmWave systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 1545–1550.

[41] G. Charan and A. Alkhateeb, "User identification: A key enabler for multi-user vision-aided communications," 2022, *arXiv:2210.15652*.

Shuaifeng Jiang received the B.S. degree in information technology from Southeast University, China, in 2018, and the M.Eng. degree in information and communication engineering from the Tokyo Institute of Technology, Japan, in 2020. He is currently pursuing the Ph.D. degree in electrical engineering with Arizona State University. His research interests include wireless communications, wireless sensing, and machine learning.

Ahmed Hindy received the B.Sc. degree from Alexandria University, Egypt, in 2010, the M.Sc. degree from Nile University, Egypt, in 2012, and the Ph.D. degree in electrical engineering from The University of Texas at Dallas, USA, in 2017. He is currently an Advisory Researcher with Lenovo, Motorola Mobility LLC, Chicago, IL, USA. He is also a 3GPP delegate for 5G standardization. His research interests are in the general area of wireless communication. He was a recipient of the 2023 Lenovo 5G Individual Excellence Award, the 2021 Lenovo Research Rising Star Award, the 2015 Ericsson Graduate Fellowship, and the 2012 Erik Jonsson Graduate Fellowship.



Ahmed Alkhateeb received the B.S. (Hons.) and M.S. degrees in electrical engineering from Cairo University, Egypt, in 2008 and 2012, respectively, and the Ph.D. degree in electrical engineering from The University of Texas at Austin, USA, in August 2016. Between September 2016 and December 2017, he was a Wireless Communications Researcher with the Connectivity Laboratory, Facebook, Menlo Park, CA, USA. He joined Arizona State University (ASU) in Spring 2018, where he is currently an Assistant Professor with the School of Electrical, Computer and Energy Engineering. He has held research and development internships with FutureWei Technologies, Chicago, IL, USA, and Samsung Research America (SRA), Dallas, TX, USA. His research interests are in the broad areas of wireless communications, communication theory, signal processing, machine learning, and applied mathematics. He was a recipient of the 2012 MCD Fellowship from The University of Texas at Austin, the 2016 IEEE Signal Processing Society Young Author Best Paper Award for his work on hybrid precoding and channel estimation in millimeter wave communication systems, and the 2021 NSF Career Award.