IEEE Open Journal of the Communications Society

Received 1 May 2026; revised 28 May 2026; accepted 10 June 2026. Date of publication 15 June 2026; date of current version 23 June 2026.

Digital Object Identifier 10.1109/OJCOMS.2026.3703751

Machine-Vision Aided Low-Complexity Beamforming Methods for IRS Networks

MUTEEN MUNAWAR (Student Member, IEEE), MAMOUN GUENACH (Senior Member, IEEE), AND INGRID MOERMAN (Senior Member, IEEE)

Interuniversity Microelectronics Centre (IMEC), Leuven 3001, Belgium
IDLab, IMEC, Ghent University, Ghent 9000, Belgium

CORRESPONDING AUTHOR: M. MUNAWAR (Muteen.Munawar@imec.be)

This work was supported in part by European Community's Research Foundation Flanders (FWO) under Grant A2582_00_01_01, in part by the Project Strip-Link Multiple-Input Multiple-Output (MIMO) under Grant FWO-97707, and in part by the Interuniversity Microelectronics Centre (IMEC) Advanced Radio Frequency (ARF) Program under Grant 97700/01345.

ABSTRACT This paper uses machine vision (MV) to assist communication in wireless networks, with a specific focus on intelligent reflecting surface (IRS)-assisted wireless systems. Instead of relying on traditional schemes such as alternating optimization or semidefinite relaxation, which are computationally expensive and often impractical, we use visual data to enable low-complexity beamforming decisions for users. Our approach employs a ceiling-mounted camera with a fish-eye lens to cover a wide communication area. Users within the network are first detected using an object detection method. We then propose closed-form analytical expressions to determine the distances between the access point (AP), users, and IRS. To address the non-uniform distortion inherent in fish-eye images, we introduce a novel method for determining non-uniform pixel weightings using trigonometric techniques. Based on the calculated distances, beamforming decisions are made according to the user's proximity to the AP or IRS. Furthermore, we extend the application of MV to a multiuser IRS-assisted scenario, where we propose grouping users into either two or three categories. Using these visually identified groups, we simplify and solve the IRS optimization problem by considering only the constraints relevant to each category. Simulation results demonstrate that the proposed MV-based scheme for IRS achieves a significantly lower computational cost compared to benchmark schemes, while maintaining comparable performance.

INDEX TERMS Adaptive decision, beamforming, intelligent reflecting surface, low complexity, machine vision, multiple-input multiple-output, multiuser, view to communicate.

I. INTRODUCTION

A. INTELLIGENT REFLECTING SURFACE

THE intelligent reflecting surface (IRS), a promising technology, is envisioned to contribute controllable reflecting channels to enable spectrum- and energy-efficient communication in next-generation wireless communication systems [2]. Specifically, the IRS is a metasurface containing several meta-elements that allow it to change its material properties in real-time, resulting in controllable reflecting channels [2], [3], [4], [5]. To maximize the signal strength at the receiver in an IRS-assisted wireless network, a large number of reflecting elements at the IRS demand appropriate phase shift values, which is a computationally expensive task from the optimization point of view [2].

B. SINGLE-USER IRS: RELATED WORK AND CONTRIBUTIONS

The literature is rich with beamforming schemes, i.e., the optimization of phase shifts, for IRS-assisted wireless networks [6], [7], [8], [9], [10]. Commonly used methods to optimize the IRS phases fall into three categories: (i) iterative methods such as alternating optimization (AO) [7], [8]; (ii) general-purpose convex solvers, e.g., CVX [9], [11]; and (iii) heuristic schemes [10]. Among these, AO and CVX-based methods deliver better beamforming solutions, but their computational complexity, in particular that of the underlying semidefinite relaxation (SDR), is very high [10] and is therefore unsuitable for the practical implementation of IRSs.

A heuristic approach named adaptive selection beamforming (ASB) was proposed in [10]. It is (i) non-iterative and (ii) free of general-purpose convex solvers, providing performance close to that of AO and SDR while reducing the computational cost by up to 70% relative to AO.[1] The scheme

[1]Numerical comparisons against SDR are omitted because of its very high computational cost.

we propose in this paper[2] is even less computationally expensive than conventional ASB. This improvement is made possible by integrating machine vision with communication, as detailed in the following subsection.

C. MACHINE VISION AND COMMUNICATION

The integration of machine vision and communication theory holds the potential to enhance the performance of various applications [12]. This synergy can be explored in two main directions. The first one, known as communicate-to-view (C2V), involves using RF signals-based data to enhance visual applications. While traditional computer vision relies on visible light data, using RF signals with various frequencies offers distinct advantages, such as increased diffraction, which can enhance the resolution of imagery for distorted or occluded objects [13].

The second direction, view-to-communicate (V2C), focuses on using visual data to improve communication networks. Given the high data speeds demanded by next-generation communication systems, there is a pressing need for computationally efficient algorithms that can rapidly estimate channels and process data. Machine vision emerges as a promising tool for achieving these goals. For instance, visual data can be employed to predict future channel behaviors and patterns of mobile blockages, so enhancing decision-making processes [14].

The proposed work in this paper primarily aligns with the V2C direction. Specifically, in this work, we exploit visual information to make a low-complexity beamforming decision for IRS-assisted wireless networks. The meaning of the beamforming decision is elaborated in later sections.

D. MULTIUSER IRS: RELATED WORK AND CONTRIBUTION

As in the single-user case, several works have studied IRS optimization in the multiuser scenario for various objectives. For instance, the authors of [15] propose optimizing the IRS elements to maximize the signal strength at the user; reference [16] proposes optimizing the IRS to improve the rank of multiple-input multiple-output (MIMO) channels; reference [17] studies IRS optimization to provide uniform performance among users; reference [18] optimizes the IRS to implement passive modulation; references [19] and [20] optimize the IRS to enhance interlayer multiplexing gains; and reference [21] proposes optimizing the IRS elements to jointly enhance communication and sensing performance.

These studies use common methods such as SDR [22], manifold optimization [23], gradient descent [24], and successive convex approximation [25] for continuous variables, while tabu search [26], sphere decoding [27], and metaheuristic algorithms such as particle swarm optimization, simulated annealing, and differential evolution have been reported for discrete variables.

[2]Section III provides a brief overview of the AO, SDR, and ASB schemes before detailing our proposal.

Despite this wealth of beamforming methods, a prototype IRS-assisted wireless network with real-time optimization of hundreds of IRS elements has not yet been demonstrated. The main reason is that optimizing hundreds of IRS elements for the above objectives within the channel coherence time is not practically feasible. Some IRS prototypes targeting passive modulation [28] or fixed user positions [29] have been reported, but these address objectives different from those above. There is therefore still a need for low-complexity approaches that can bring IRS-aided systems closer to real-time deployment.

In this work, we exploit machine vision for an IRS-aided multiuser scenario and propose a method to reduce the computational complexity of existing optimization algorithms. More specifically, we use visual information to detect and categorize users in the region before applying channel estimation and beamforming methods. The proposed method helps reduce the instantaneous computational complexity associated with IRS optimization by up to 75%, depending on the number of users categorized into specific groups.

The remainder of the paper is organized as follows. Section II presents the system model and problem formulation for both the single-user and multiuser scenarios and introduces the SDR with alternating optimization (SDR-AO) benchmark solution used for the multiuser case. Section III introduces the proposed vision-aided schemes: the single-user adaptive beamforming selection and the multiuser user-grouping (UG) method, together with the validity of the underlying channel approximations across propagation scenarios, the threshold calibration with its closed-form geometry-portability analysis, the robustness analysis, the formal complexity scaling, the channel-state-information (CSI) overhead, the end-to-end operating-timescale and real-time feasibility analysis, and a non-iterative IRS beamformer (NIB) variant of the multiuser scheme. Section IV presents the simulation results for both scenarios. Section V discusses extensions and Section VI concludes the paper.

Notations: Scalars are denoted by italic letters, whereas vectors and matrices are denoted by bold-face lower- and upper-case letters, respectively. For a complex-valued vector $\mathbf{v}$ of length N, $\mathbf{v}^T$ denotes the transpose, $v_{(n)}/\mathbf{v}_{(n)}$ denotes the nth element of $\mathbf{v}$, $\mathbf{v}^*$ denotes the complex conjugate of each element, $\mathbf{1_v}$ denotes a vector of size $\mathbf{v}$ with all entries 1, $\mathbf{v}^H$ denotes the conjugate transpose, $\|\mathbf{v}\|$ denotes the Euclidean norm, $\mathrm{diag}\,(\mathbf{v})$ denotes a diagonal matrix with each diagonal element being the corresponding element in $\mathbf{v}$, $\mathbf{v} > 0$ denotes the number of non-zero elements in $\mathbf{v}$, $\text{log-sum}\,(\mathbf{v})$ means $\sum_{k=1}^{K} \log(\mathbf{v}_k)$, $\mathrm{sum}\,(\mathbf{v})$ means $\sum_{k=1}^{K} (\mathbf{v}_k)$, and $\mathrm{GM}\,(\mathbf{v})$ represents the geometric mean. For a complex-valued matrix $\mathbf{M}$, $\mathrm{rank}\,(\mathbf{M})$ denotes the rank, $\mathbf{M}_{(n,m)}$ denotes the nth row and mth column entry, $\mathbf{M}\succeq 0$ denotes positive semi-definite, and $\mathrm{trace}\,(\mathbf{M})$ denotes trace. For a complex-valued square matrix $\mathbf{A}$, $\mathbf{I_A}$ denotes an identity matrix of size $\mathbf{A}$ and $\mathbf{A}^{-1}$ denotes the inverse. The notation $(\cdot)^*$ denotes the

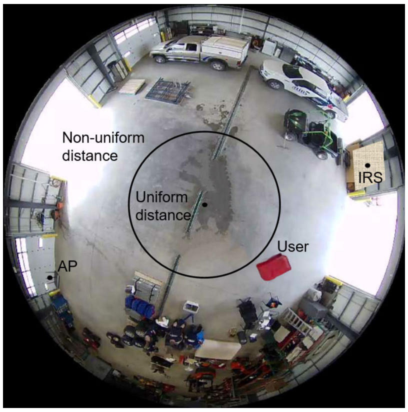


FIGURE 1. System model (fish-eye view).

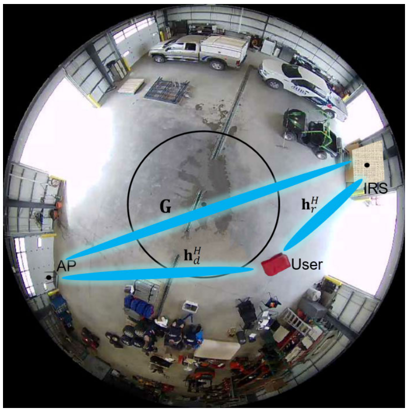


FIGURE 2. System model with channels.

complex conjugate of a complex number, $\hat{e}_{\max}(\mathbf{A}, \mathbf{B})$ denotes the dominant generalized eigenvector of matrix pair $(\mathbf{A}, \mathbf{B})$.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We first describe the single-user setup (Subsection A) and then extend it to the multiuser case (Subsection B). Subsection C presents the SDR-AO benchmark solution used as the reference scheme in the multiuser scenario.

A. SINGLE-USER SETUP

The system model considered in this work is illustrated in Fig. 1. This image is captured by a ceiling pin-hole camera with a fish-eye view [30]. The IRS and access point (AP) are fixed at known locations. A user, i.e., a robot, moves within the area and communicates with the AP with the assistance of the IRS. The area in the image is partitioned into two regions. The center region comprises pixels with nearly uniform distance weights corresponding to the actual region. The non-uniform region consists of pixels with higher weight compared to the center pixels, attributed to the non-uniform view of the camera. The channels among the AP, IRS, and user are depicted in Fig. 2.

To formulate the problem, the communication network shown in Fig. 2 is simplified in Fig. 3. Specifically, we

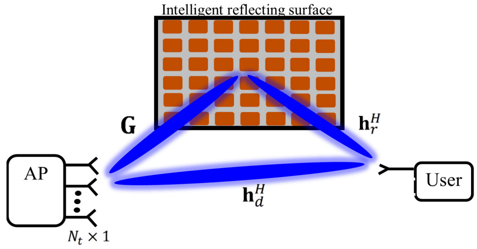


FIGURE 3. Simplified system model.

consider a communication network where an IRS, with N reflecting elements, aids a multi-antenna AP, having N_t transmit antennas, in communicating with a single-antenna user. Let $\mathbf{h}_d^H \in \mathbb{C}^{1\times N_t}$, $\mathbf{h}_r^H \in \mathbb{C}^{1\times N}$, and $\mathbf{G} \in \mathbb{C}^{N\times N_t}$ denote the baseband equivalent channels for the AP-user link, IRS-user link, and AP-IRS link, respectively.[3] The received signal at the user is expressed as:

$$\begin{aligned} y &= \mathbf{h}_r^H \boldsymbol{\Phi} \mathbf{G} \mathbf{v} s + \mathbf{h}_d^H \mathbf{v} s + z \\ &= \left(\mathbf{h}_r^H \boldsymbol{\Phi} \mathbf{G} + \mathbf{h}_d^H\right) \mathbf{v} s + z, \end{aligned} \tag{1}$$

where z, $\mathbf{v}$, and s denote, respectively, the additive white Gaussian noise at the receiver with zero mean and variance σ^2, the beamforming vector at the AP, and the transmitted information symbol. The matrix $\boldsymbol{\Phi}$ is a diagonal matrix containing the IRS reflection coefficients, i.e., $\boldsymbol{\Phi} = \text{diag}(\alpha_1 e^{j\phi_1}, \alpha_2 e^{j\phi_2}, \ldots, \alpha_N e^{j\phi_N})$, with $\alpha \in [0, 1]$ and $\phi \in [0, 2\pi]$ being the amplitude and phase shift applied by each reflecting element, respectively. The expressions $\mathbf{h}_r^H \boldsymbol{\Phi} \mathbf{G} \mathbf{v} s$ and $\mathbf{h}_d^H \mathbf{v} s$ in (1) represent the signal received through the AP-IRS-user link and the AP-user link, respectively. Further details on the system model are available in [10] and are omitted here for brevity.

The received signal strength at the user, denoted by γ, is given by:

$$\gamma = \left|\left(\mathbf{h}_r^H \boldsymbol{\Phi} \mathbf{G} + \mathbf{h}_d^H\right) \mathbf{v}\right|^2. \tag{2}$$

To maximize (2), we need to optimize $\mathbf{v}$ and $\boldsymbol{\Phi}$ at the AP and IRS, respectively. Accordingly, the optimization problem is formulated as

$$\begin{aligned} \max_{\mathbf{v}, \boldsymbol{\Phi}} \quad & \left|\left(\mathbf{h}_r^H \boldsymbol{\Phi} \mathbf{G} + \mathbf{h}_d^H\right) \mathbf{v}\right|^2 \\ \text{s.t.} \quad & \|\mathbf{v}\|^2 \le p, \\ & 0 \le \phi_n \le 2\pi, \quad n = 1, \ldots, N, \end{aligned} \tag{P1}$$

where p represents the maximum transmission power budget at the AP. This optimization problem has a non-convex objective function concerning $\boldsymbol{\Phi}$ and $\mathbf{v}$, resulting in a non-convex problem. Common approaches, such as AO [8] and SDR [9], are employed in the literature. Although AO and SDR offer good performance, they come with higher computational costs [10]. In this work, we first explore two simple

[3] We adopt the conventional pilot-based CSI acquisition schemes of [31], [32], and [33]; the impact of the proposed grouping on the pilot overhead is quantified separately in Subsection III-B. III-B.10.

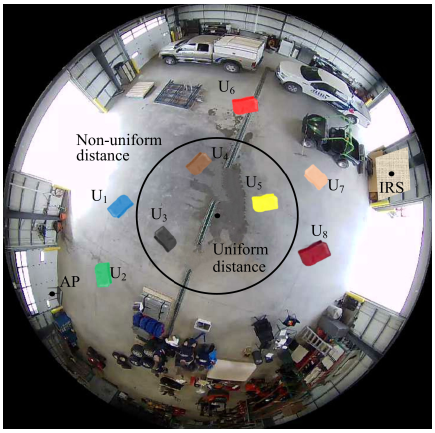


FIGURE 4. Multiuser system model (fish-eye view).

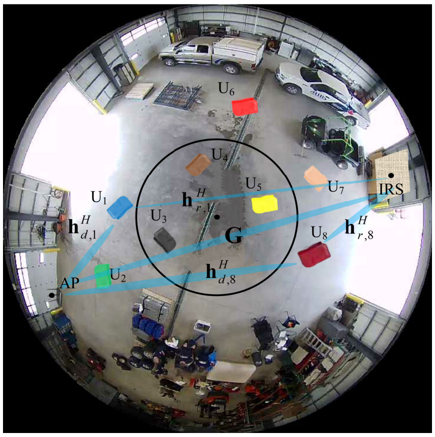


FIGURE 5. Multiuser system model with channels.

beamforming solutions based on the heuristic ASB scheme [10] and subsequently use visual data to select one of them.

B. MULTIUSER SETUP

We consider an IRS-aided multiuser scenario, as illustrated in Fig. 4, where, similar to Fig. 1, a fish-eye view is employed to visualize the communication region.

More specifically, the system model includes users $U_1, U_2, \ldots, U_8$ as an example, though the number of users K is treated as a variable. In Fig. 5, we illustrate the direct and reflected channels for these users. For a given user k, the channels $\mathbf{h}_{d,k}$ and $\mathbf{h}_{r,k}$ denote the AP-to-user (direct) and IRS-to-user (reflected) links, respectively.

All other parameters, such as the region area, number of pixels in the image, etc., are the same as those defined in Subsection A. The signal received at user k and the corresponding achievable rate are given by:

$$y_k = \left(\mathbf{h}_{r,k}^H \boldsymbol{\Phi} \mathbf{G} + \mathbf{h}_{d,k}^H\right) \sum_{j=1}^{K} \mathbf{v}_j s_j + z, \tag{3}$$

and

$$c_k = \log_2\left(1 + \frac{\left|\left(\mathbf{h}_{r,k}^H \boldsymbol{\Phi} \mathbf{G} + \mathbf{h}_{d,k}^H\right) \mathbf{v}_k\right|^2}{\left|\left(\mathbf{h}_{r,k}^H \boldsymbol{\Phi} \mathbf{G} + \mathbf{h}_{d,k}^H\right) \sum_{j\neq k}^{K} \mathbf{v}_j\right|^2 + \sigma^2}\right), \tag{4}$$

respectively, where $\left|\left(\mathbf{h}_{r,k}^H \boldsymbol{\Phi} \mathbf{G} + \mathbf{h}_{d,k}^H\right) \mathbf{v}_k\right|^2$ denotes the received signal power at user k, and $\left|\left(\mathbf{h}_{r,k}^H \boldsymbol{\Phi} \mathbf{G} + \mathbf{h}_{d,k}^H\right) \sum_{j\neq k}^{K} \mathbf{v}_j\right|^2$ represents the interference power experienced by user k, with σ^2 being the thermal noise power.

To exploit the visual information and demonstrate its benefits, we consider a traditional problem in IRS-aided multiuser scenarios, sum-rate maximization, which can be formulated as follows:

$$\begin{aligned} \max_{\mathbf{v}, \boldsymbol{\Phi}} \quad & \sum_{k=1}^{K} c_k \\ \text{s.t.} \boldsymbol{\Phi} &= \operatorname{diag}\left(e^{-j\phi_1}, e^{-j\phi_2}, \ldots, e^{-j\phi_N}\right), \\ & \sum_{k=1}^{K} \|\mathbf{v}_k\|^2 \leq P_{\max}, \end{aligned} \tag{P2}$$

where $\|\mathbf{v}_k\|^2$ denotes the power allocated to user k, i.e., p_k.

The optimization problem (P2) has been widely studied in several works [6]. Common approaches to solve this problem include SDR, Riemannian manifold optimization, and successive convex approximation, among others.

To demonstrate the effectiveness of using visual information, we adopt SDR as a benchmark scheme in this paper due to its balanced trade-off between performance and computational complexity. However, as will be discussed later, the proposed scheme is applicable to any existing solution method.

C. MULTIUSER BENCHMARK: SDR-AO SOLUTION

In this subsection, we briefly discuss a known solution to problem (P2) based on SDR, which will serve as the benchmark scheme in the subsequent sections. Specifically, we outline convex formulations for the variables p_k, $\mathbf{v}_k$, and $\boldsymbol{\Phi}$.

First, we consider the optimal linear minimum mean-square-error (MMSE) digital filter $\mathbf{v}_k$ for each user k, defined as:

$$\bar{\mathbf{v}}_k = \frac{\widehat{\mathbf{e}}_{\max}(\mathbf{S}_k, \mathbf{T}_k)}{\left\|\widehat{\mathbf{e}}_{\max}(\mathbf{S}_k, \mathbf{T}_k)\right\|}, \tag{5}$$

where $\bar{\mathbf{v}}_k = \mathbf{v}_k / \|\mathbf{v}_k\|$, $\widehat{\mathbf{e}}_{\max}(\mathbf{S}_k, \mathbf{T}_k)$ denotes the strongest generalized eigenvector of the matrices $\mathbf{S}_k$ and $\mathbf{T}_k$. Here, $\mathbf{S}_k$ and $\mathbf{T}_k$ represent the signal covariance matrix and the interference-plus-noise covariance matrix for user k, respectively, detailed in [34].

Given $\mathbf{v}_k, \forall k$ and $\boldsymbol{\Phi}$, the power allocation subproblem in (P2) can be formulated as:

$$\max_{p_k} \sum_{k=1}^{K} \log_2\left(1 + \frac{\left|\mathbf{h}_k^H \bar{\mathbf{v}}_k \sqrt{p_k}\right|^2}{\sum\limits_{j\neq k} \left|\mathbf{h}_k^H \bar{\mathbf{v}}_j \sqrt{p_j}\right|^2 + \sigma^2}\right)$$

Algorithm 1 SDR-AO for Solving (P2) (Benchmark)

Initialize:
Iteration index $i = 0$, $\mathbf{L}_i = \mathbf{I}_{\mathbf{\Phi}}$, $\mathbf{v}_{k,i} = \frac{1}{\sqrt{N_t}}\mathbf{1}$, and $p_{k,i} = \frac{P_{\max}}{K}$.
repeat
  Step 1: Given $\mathbf{\Phi}_i$, compute the composite channel $\mathbf{h}_k$ for all k.
  repeat
    Step 2a: Given $p_{k,i}$ and $\mathbf{v}_{k,i}$, update $\lambda_{k,i}$ for all k using (P2.2).
    Step 2b: Given $\lambda_{k,i}$ and $\mathbf{v}_{k,i}$, update $p_{k,i}$ for all k using (P2.2).
    Step 2c: Given $\lambda_{k,i}$ and $p_{k,i}$, update $\mathbf{v}_{k,i}$ for all k using (5).
  until the objective value of (P2.1) converges or the maximum number of inner iterations is reached.
  repeat
    Step 3a: Given $p_{k,i+1}$, $\mathbf{v}_{k,i+1}$, and $\lambda_{k,i+1}$ for all k, fix $\mathbf{L}_i$ and solve (P2.4) for $y_{k,i+1}$.
    Step 3b: Given $p_{k,i+1}$, $\mathbf{v}_{k,i+1}$, $\lambda_{k,i+1}$, and $y_{k,i+1}$ for all k, solve (P2.4) for $\mathbf{L}_{i+1}$ and extract $\mathbf{\Phi}_{i+1}$.
  until the objective value of (P2.4) converges or the maximum number of inner iterations is reached.
  Step 4: Update iteration index: $i \leftarrow i + 1$.
until the objective value of (P2) converges or the maximum number of outer iterations is reached.

$$\begin{aligned} \text{s.t.} \quad & \mathbf{h}_k^H = \mathbf{h}_{r,k}^H \mathbf{\Phi}\mathbf{G} + \mathbf{h}_{d,k}^H, \\ & \sum_{k=1}^{K} p_k = P_{\max}. \end{aligned} \tag{P2.1}$$

The problem (P2.1) is non-convex due to the presence of fractions. An optimal solution to this problem was proposed in [34], using fractional programming. Below, we present a simplified convex formulation of (P2.1); for full derivations, refer to [34].

By introducing auxiliary variables λ_k and applying standard transformations, a bi-variable non-convex approximation of (P2.1) can be written as:

$$\begin{aligned} \max_{p_k,\lambda_k} \quad & \sum_{k=1}^{K} \lambda_k \\ \text{s.t.} \quad & |\mathbf{h}_k^H \bar{\mathbf{v}}_k|^2 p_k - (2^{\lambda_k} - 1)\left(\sum_{j\neq k} |\mathbf{h}_k^H \bar{\mathbf{v}}_j|^2 p_j + \sigma^2\right) \\ & \geq 0, \\ & \sum_{k=1}^{K} p_k \leq P_{\max}. \end{aligned} \tag{P2.2}$$

Solving (P2.2) alternately over $\{p_k\}$ and $\{\lambda_k\}$ while keeping the other fixed results in a sequence of convex subproblems, which can be efficiently solved using CVX [11]. Details can be found in [34] and Algorithm 1.

Now, given $\{\mathbf{v}_k\}$ and $\{p_k\}$, the subproblem of optimizing $\mathbf{\Phi}$ in (P2) can be simplified as:

$$\begin{aligned} \max_{\mathbf{\Phi}} \quad & \sum_{k=1}^{K} c_k \\ \text{s.t.} \quad & \mathbf{\Phi} = \operatorname{diag}(e^{-j\phi_1}, e^{-j\phi_2}, \ldots, e^{-j\phi_N}). \end{aligned} \tag{P2.3}$$

The problem (P2.3) is a non-convex, unit-modulus-constrained, non-homogeneous quadratically constrained quadratic program (QCQP), which can be solved using SDR combined with fractional programming. A joint bi-variable non-convex formulation of (P2.3) is presented below:

$$\begin{aligned} \max_{\mathbf{L},\{y_k\}} \quad & \mathrm{GM}(\mathbf{x_L}) \\ \text{s.t.} \quad & \mathbf{x_L} = [x_1, x_2, \ldots, x_K], \\ & x_k = \log_2\Big(1 + 2\, y_k \left(\operatorname{trace}(\mathbf{\Theta}^{k,k}\mathbf{L}) + |\xi^{k,k}|^2\right)\Big) \\ & \quad - y_k^2 \sum_{j\neq k} \left(\operatorname{trace}(\mathbf{\Theta}^{k,j}\mathbf{L}) + |\xi^{k,j}|^2\right) - \sigma^2, \\ & \mathbf{L} \succeq 0, \\ & y_k \in \mathbb{R}_+, \quad \forall k, \\ & \mathbf{L}_{n,n} = 1, \quad \forall n = 1, \ldots, N+1, \end{aligned} \tag{P2.4}$$

where x_k represents the proportional achievable rate of user k. The matrix $\mathbf{L}$ is a semidefinite-program (SDP) relaxation of the rank-one matrix, formed as $\mathbf{L} = \bar{\mathbf{l}}\bar{\mathbf{l}}^H$, where $\bar{\mathbf{l}} = \begin{bmatrix} \mathbf{l} \\ t \end{bmatrix}$, with t as an auxiliary scalar variable and $\mathbf{l}$ as the vectorized version of the phase-shift vector $\mathbf{\Phi}$. Additionally, in (P2.4), we define: $\xi^{k,j} = \mathbf{h}_{d,k}^H \mathbf{v}_j$, $\omega^{k,j} = \operatorname{diag}(\mathbf{h}_{r,k}^H)\mathbf{G}\mathbf{v}_j$, and $\mathbf{\Theta}^{k,j} = \begin{bmatrix} \omega^{k,j}(\omega^{k,j})^H & \omega^{k,j}(\xi^{k,j})^H \\ \xi^{k,j}(\omega^{k,j})^H & 0 \end{bmatrix}$.

The problem (P2.4) can be solved iteratively for y_k and $\mathbf{L}$ as a sequence of convex formulations. The complete benchmark AO algorithm for solving (P2) is detailed in Algorithm 1.

III. PROPOSED VISION-AIDED SCHEMES

Subsection A presents the single-user adaptive beamforming selection. Subsection B extends the framework to the multiuser scenario via the user-grouping (UG) method and develops the supporting analysis (channel-approximation validity, threshold sensitivity, robustness, complexity scaling, CSI overhead, and operating-timescale budget).

A. SINGLE-USER: VISION-AIDED ADAPTIVE BEAMFORMING SELECTION

To initiate the detailed algorithm, we begin by discussing the conventional AO, SDR, and ASB schemes.

In AO-based schemes [8], the optimal $\mathbf{v}$ for the AP is selected using maximum ratio transmission (MRT):

$$\mathbf{v} = \sqrt{p}\,\frac{\left(\mathbf{h}_r^H \mathbf{\Phi}\mathbf{G} + \mathbf{h}_d^H\right)^H}{\left\|\mathbf{h}_r^H \mathbf{\Phi}\mathbf{G} + \mathbf{h}_d^H\right\|}, \tag{6}$$

whereas the optimal phase shifts for IRS, i.e., $\mathbf{\Phi}$, are selected such that the signal from the direct path and reflecting path coherently combine at the receiver:

$$\phi_n = \arg\left(\mathbf{h}_d^H \mathbf{v}\right) - \arg\left(h_{n,r}^*\right) - \arg\left(\mathbf{g}_n^H \mathbf{v}\right),\ n = 1, \ldots, N, \quad (7)$$

where $h_{n,r}$, ϕ_n, and $\mathbf{g}_n^H$ denote the nth element of $\mathbf{h}_r$, nth diagonal element of $\mathbf{\Phi}$, and nth row of $\mathbf{G}$, respectively. Note that $\mathbf{v}$ and $\mathbf{\Phi}$ from (6) and (7) are dependent on each other, and hence solving them requires an iterative method known as AO (more details on AO in [8]).

In SDR-based schemes, the optimal $\mathbf{v}$ for the AP, i.e., (6), is inserted into the objective function of (P1), and the resulting expression is converted into an SDR program, which is further solved using a general-purpose convex solver, i.e., CVX. More details on such methods can be found in [7] and [9].

In the ASB scheme, two simple solutions for $\mathbf{v}$ and $\mathbf{\Phi}$, denoted as $(\mathbf{v}_1, \mathbf{\Phi}_1)$ and $(\mathbf{v}_2, \mathbf{\Phi}_2)$, are derived as follows:[4]

$$\mathbf{v}_1 = \sqrt{p}\frac{\mathbf{h}_d}{||\mathbf{h}_d||} \quad (8)$$

$$\phi_{1n} = -\arg\left(h_{n,r}^*\right) - \arg\left(g_n^* \sqrt{p}\frac{\mathbf{h}_d}{\left\|\mathbf{h}_d\right\|}\right),\ n = 1, \ldots, N, \quad (9)$$

and

$$\mathbf{v}_2 = \sqrt{p}\frac{\mathbf{g}_n}{||\mathbf{g}_n||} \quad (10)$$

$$\phi_{2n} = \arg\left(\mathbf{h}_d^H \sqrt{p}\frac{\mathbf{g}_n}{\left\|\mathbf{g}_n\right\|}\right) - \arg\left(h_{n,r}^*\right),\ n = 1, \ldots, N. \quad (11)$$

Following the computation of $(\mathbf{v}_1, \mathbf{\Phi}_1)$ and $(\mathbf{v}_2, \mathbf{\Phi}_2)$ using ((8), (9)) and ((10), (11)), respectively, the ASB scheme calculates the corresponding total channel gains:

$$a_1 = \left|\left(\mathbf{h}_r^H \mathbf{\Phi}_1 \mathbf{G} + \mathbf{h}_d^H\right) \mathbf{v}_1\right| \quad (12)$$

and

$$a_2 = \left|\left(\mathbf{h}_r^H \mathbf{\Phi}_2 \mathbf{G} + \mathbf{h}_d^H\right) \mathbf{v}_2\right|. \quad (13)$$

Finally, the scheme compares a_1 and a_2 and selects $(\mathbf{v}_1, \mathbf{\Phi}_1)$ if $a_1 \geq a_2$ and $(\mathbf{v}_2, \mathbf{\Phi}_2)$ otherwise. This selection between $(\mathbf{v}_1, \mathbf{\Phi}_1)$ and $(\mathbf{v}_2, \mathbf{\Phi}_2)$ constitutes the ASB beamforming decision. The calculation of a_1 and a_2 is computationally expensive because of the involved large, complex matrix multiplications: for large N or N_t, these dominate the per-coherence-interval cost (the computational complexities of AO, SDR, and ASB are detailed in [10]).

In this paper, we eliminate the computational cost associated with the calculation of a_1 and a_2. Specifically, we use visual data for this decision-making process, i.e., the selection of $(\mathbf{v}_1, \mathbf{\Phi}_1)$ or $(\mathbf{v}_2, \mathbf{\Phi}_2)$. We compare the signal-to-noise ratio (SNR) performance and computational cost with AO [9], SDR [8], and ASB [10]. The proposed work is divided into the following parts: 1) Detect the user's presence in an image and retrieve pixel coordinates. 2) Based on the non-uniform distances in the image, assign each pixel the corresponding actual distance. 3) Based on the detection, find the distance from the user equipment (UE) to the AP and IRS. 4) Based on the calculated distance, find a beamforming decision for the AP and IRS. The following subsections explain these steps in detail.

[4]Details are available in [10] and are omitted here for brevity.

```
*******************************************************
Training a YOLO v2 Object Detector for the following object classes:

* vehicle

Training on single CPU.
Initializing input data normalization.
```

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch RMSE	Validation RMSE	Mini-batch Loss	Validation Loss	Base Learning Rate
1	1	00:00:13	8.98	1.77	80.5857	3.1362	0.0010
4	50	00:04:32	1.63	1.47	2.6454	2.1564	0.0010
7	100	00:09:14	0.47	0.50	0.2194	0.2550	0.0010
10	150	00:13:50	0.32	0.44	0.1011	0.1939	0.0010
13	200	00:18:14	0.26	0.40	0.0672	0.1614	0.0010
16	250	00:22:36	0.22	0.33	0.0500	0.1076	0.0010
19	300	00:26:58	0.22	0.27	0.0492	0.0704	0.0010
22	350	00:31:18	0.18	0.34	0.0315	0.1167	0.0010
25	400	00:35:37	0.19	0.23	0.0368	0.0537	0.0010
29	450	00:40:04	0.16	0.23	0.0271	0.0515	0.0010
30	480	00:42:36	0.13	0.22	0.0171	0.0464	0.0010

```
Training finished: Max epochs completed.
Detector training complete.
```

FIGURE 6. YOLOv2 Training.

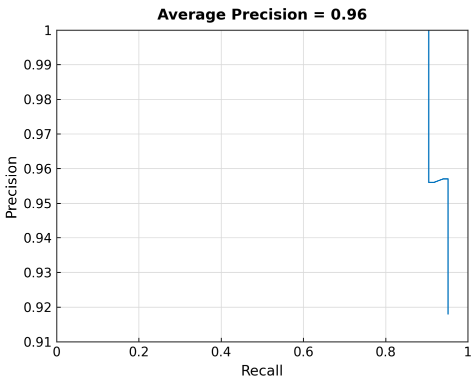


FIGURE 7. Average precision of YOLOv2 for single-user detection.

1) USER DETECTION USING YOLOV2

First of all, we detect the user and obtain its location coordinates in the image. For this purpose, we employ YOLOv2 [35], a single-stage, anchor-based real-time object detector, trained and run via the MATLAB Computer Vision Toolbox [36]; the learning behavior of YOLOv2 in a vehicle-detection context is reviewed in [37]. YOLOv2 is selected over more recent variants for three reasons that match the requirements of the proposed framework: (i) its small footprint (~60M parameters with the standard `darknet-19` backbone) keeps the per-frame inference cost in the tens of milliseconds on commodity CPUs, which is well below the zone-boundary-crossing trigger interval at which the detector is invoked (Subsection III-B.11); (ii) it integrates natively into the same MATLAB pipeline that hosts the rest of the simulation chain (CVX, SDR-AO solver, channel models), giving an end-to-end reproducibility path; and (iii) the detection task in the targeted deployment is class-balanced (a single robot UE class) and large-target (each robot occupies a non-trivial fraction of the fish-eye image), so the marginal accuracy gains of higher-capacity detectors do not translate into a meaningful improvement on the downstream beamforming decision.

FIGURE 8. Trained neural network: unseen sample test.

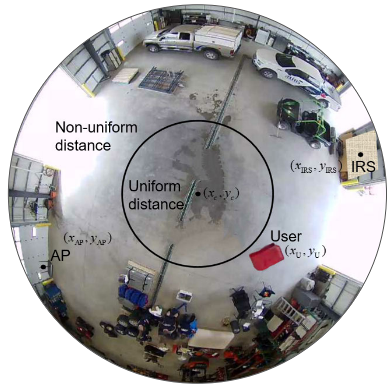


FIGURE 9. Known pixel coordinates of AP and IRS.

The training dataset comprises 160 images with offline augmentation (rotation, scaling, and brightness/contrast jittering). After 30 training epochs with a minibatch size of 6 (training trace in Fig. 6), the model achieves a held-out average precision of 0.96 (Fig. 7). The framework targets a fixed indoor cell with a fixed ceiling-mounted camera, IRS, and AP; for a different physical environment or a different UE form factor (e.g., handheld smartphone, ground robot, drone), the detector is re-trained or fine-tuned on a per-deployment dataset of similar size, paid once per deployment.

The proposed framework is *detector-agnostic*: the downstream pipeline (pixel-to-distance mapping in Subsection III-A.2, zone classification in Subsection III-B.4, calibration in Subsections III-B.6 and III-B.7, and pilot-pattern derivation in Subsection III-B.10) consumes only the bounding-box centroid of each detected user, so any object-detection model that outputs bounding boxes can replace YOLOv2 as a drop-in module. Examples include more recent single-stage detectors such as YOLOv5/v8/v11, anchor-free models such as CenterNet and FCOS, transformer-based detectors such as DETR and RT-DETR, and lightweight on-device backbones such as MobileNet-SSD and EfficientDet-Lite. Substituting a more capable detector tightens the bounding-box centroid error Δ_{px} analyzed in Subsection III-B.8, item (i), and shortens the per-frame inference cost T_{YOLO} in (30), while leaving the downstream pipeline, the calibration procedure, and the complexity savings derived in Subsection III-B.9 unchanged.

To test the trained neural network, the outcome corresponding to an unseen sample image is depicted in Fig. 8, where the model detected the robot with a confidence score of 81.

2) PIXEL WEIGHTS AND DISTANCE CALCULATION

To calculate the distance, we first need to determine the actual weights of pixels in the image. To explain the proposed algorithm for determining the non-uniform pixel weights, for simplicity, we divide the image into two regions, as shown in Fig. 9, where (x_{IRS}, y_{IRS}), (x_{AP}, y_{AP}), (x_U, y_U), and (x_c, y_c) denote the locations of IRS, AP, user, and center, respectively. The center region outlined by a circle contains pixels with different weights compared to the remaining region. Pixels in the non-uniform region cover more distance than the center region due to the nature of the fish-eye view. To calculate the number of meters per pixel in the uniform region, d_{uniform}, we use $\frac{d_{kl}}{\sqrt{(x_k-x_l)^2+(y_k-y_l)^2}}$. Here, (x_k, y_k) and (x_l, y_l) denote any two known locations (pixel coordinates) in the circle region with a corresponding known distance, d_{kl}. Using a similar method, we calculate the number of meters per pixel for non-uniform regions, $d_{\text{non-uniform}}$.

To calculate the total distance between AP and user, we need to know the number of pixels in the uniform and non-uniform regions. The key idea is to determine the intersection points of the center region (circle) and the straight line connecting AP and user.

The function for the uniform region can be determined by the following circle equation:

$$(x - x_c)^2 + (y - y_c)^2 = \text{radius}^2, \tag{14}$$

and the line function of the AP-user link can be calculated as:

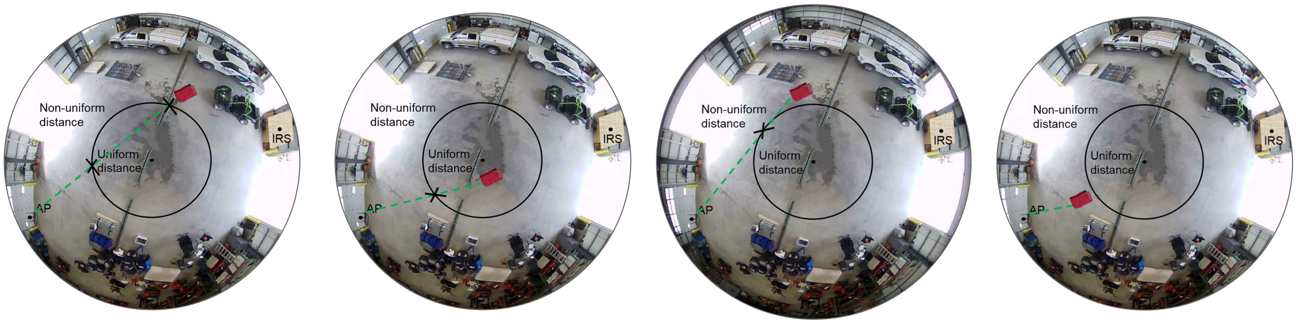
$$y - y_{\text{U}} = \frac{(y_{\text{AP}} - y_{\text{U}})}{(x_{\text{AP}} - x_{\text{U}})}(x - x_{\text{U}}). \tag{15}$$

Next, we solve (14) and (15) to find their intersection points. Depending on the number of intersection points, we can observe the following cases:

1) Two intersection points: Such a case is depicted in Fig. 10a, where the user location results in two cross-section points.
   Correspondingly, the total distance d can be calculated as follows:

$$d = \left\{\left(q_{\text{A,U}} - q_{\text{P,P}}\right) \times d_{\text{non - uniform}}\right\} + \left(q_{\text{P,P}} \times d_{\text{uniform}}\right), \tag{16}$$

   where $q_{\text{A,U}}$ and $q_{\text{P,P}}$ denote the number of pixels between AP and user, and the number of pixels between two cross-section points, respectively, whereas d_{uniform} and $d_{\text{non - uniform}}$ denote distance per pixel in uniform and non-uniform region of images, respectively.



(a) Two intersection points. (b) One intersection point: within the circle. (c) One intersection point: tangent to the circle. (d) No intersection point.

FIGURE 10. Geometric analysis of intersection cases for the single-user scenario.

2) One intersection point: This case results in two subcases. If the user is in the circle (Fig. 10b), the distance is calculated by

$$d = \left\{\left(q_{\mathrm{A,U}} - q_{\mathrm{U,P}}\right) \times d_{\text{non - uniform}}\right\} + \left(q_{\mathrm{U,P}} \times d_{\text{uniform}}\right), \tag{17}$$

where $q_{\mathrm{U,P}}$ denotes the number of pixels between the user and the cross-section point.
Otherwise, if there is only one pixel with non-uniform weight, i.e., a tangent cross-section, as shown in Fig. 10c, the distance is given by

$$d = \left(q_{\mathrm{A,U}} \times d_{\text{non - uniform}}\right) + d_{\text{uniform}} \tag{18}$$

3) No intersection point: Such a case is shown in Fig. 10d and the distance is given by

$$d = q_{\mathrm{A,U}} \times d_{\text{non - uniform}}. \tag{19}$$

3) BEAMFORMING DECISION

After computing the distance d, the adaptive beamforming decision is made using the following condition: if $d < d_i$, select $(\mathbf{v}_1, \mathbf{\Phi}_1)$ using (8) and (9); otherwise, select $(\mathbf{v}_2, \mathbf{\Phi}_2)$ using (10) and (11). Here, d_i is a threshold distance that is set either from prior knowledge of the communication region or through a numerical analysis of ASB [10]. The selection of the threshold d_i is discussed in Subsection III-B.4, which also covers the multiuser zone boundaries.

B. MULTIUSER: VISION-AIDED USER GROUPING (UG)

Now, we discuss how to exploit visual information to reduce the computational complexity of Algorithm 1. Specifically, we observe that traditional AO-based approaches, such as Algorithm 1 and those introduced in works like [9], [18], and [19], solve an optimization problem iteratively for all users until either convergence is achieved or the problem becomes infeasible. However, due to the inherent multiplicative fading nature of IRS-assisted channels [38], some users experience minimal or no effective gain from the IRS. In contrast, for users whose direct link to the AP is very weak, the IRS can act as the primary enabler of communication. Given this observation, if users are categorized into appropriate groups before executing joint beamforming AO algorithms, the computational burden associated with IRS optimization can be significantly reduced. More specifically, after suitable user grouping, we show that the steps of Algorithm 1 need to be executed only for a specific category of UEs, without any performance degradation, while achieving substantial computational savings.

We propose categorizing users into two or three groups based on predefined thresholds, referred to as 2-user grouping (2-UG) and 3-user grouping (3-UG), respectively; the grouping strategy will be discussed in detail later. Similar to the single-user scenario, the proposed framework can be divided into the following stages:

1) Detect the users in an image and retrieve their pixel coordinates.
2) Map each pixel to its corresponding actual field distance based on the non-uniform distances in the image.
3) Determine the distances from all UEs to the AP and IRS based on the detected coordinates.
4) Categorize the UEs into two or three groups according to the calculated distances.
5) Solve problem (P2) only for the relevant user group(s).
6) Evaluate the performance and computational complexity of the proposed method, and compare it with the benchmark scheme.

The following subsections explain these components in detail.

1) MULTIUSER DETECTION

Following the same procedure as in Subsection III-A.1, we train YOLOv2 to detect up to eight users within the region of interest. The training dataset is enlarged to 200 images of scaled and rotated robot poses, and the model is trained for 20 epochs with a minibatch size of 16. The training outcome and a representative detection result on an unseen sample image are shown in Fig. 11 and Fig. 12, respectively.

```
*************************************************************************
Training a YOLO v2 Object Detector for the following object classes:

* U1* U2* U3* U4* U5* U6* U7* U8

Initializing input data normalization.
|=====================================================================================================|
| Epoch | Iteration | Time Elapsed | Mini-batch | Validation | Mini-batch | Validation | Base Learning |
|       |           |  (hh:mm:ss)  |    RMSE    |    RMSE    |    Loss    |    Loss    |     Rate      |
|=====================================================================================================|
|   1   |      1    |   00:00:22   |    3.16    |    2.58    |   9.9637   |   6.6520   |    0.0005     |
|   1   |     10    |   00:01:59   |    1.31    |            |   1.7243   |            |    0.0005     |
|   2   |     20    |   00:04:33   |    1.18    |            |   1.3978   |            |    0.0005     |
|   3   |     30    |   00:07:11   |    1.11    |            |   1.2395   |            |    0.0005     |
|   4   |     40    |   00:09:59   |    1.04    |            |   1.0916   |            |    0.0005     |
|   5   |     50    |   00:12:56   |    0.93    |    1.06    |   0.8676   |   1.1189   |    0.0005     |
|   6   |     60    |   00:15:22   |    0.88    |            |   0.7751   |            |    0.0005     |
|   7   |     70    |   00:18:09   |    0.80    |            |   0.6331   |            |    0.0005     |
|   8   |     80    |   00:20:50   |    0.73    |            |   0.5378   |            |    0.0005     |
|   9   |     90    |   00:23:39   |    0.70    |            |   0.4841   |            |    0.0005     |
|  10   |    100    |   00:26:19   |    0.64    |    0.86    |   0.4128   |   0.7456   |    0.0005     |
|  11   |    110    |   00:28:34   |    0.60    |            |   0.3620   |            |  5.0000e-05   |
|  12   |    120    |   00:30:53   |    0.56    |            |   0.3173   |            |  5.0000e-05   |
|  13   |    130    |   00:33:37   |    0.62    |            |   0.3855   |            |  5.0000e-05   |
|  14   |    140    |   00:36:23   |    0.67    |            |   0.4547   |            |  5.0000e-05   |
|  15   |    150    |   00:39:08   |    0.64    |    0.84    |   0.4062   |   0.7018   |  5.0000e-05   |
|  16   |    160    |   00:41:00   |    0.62    |            |   0.3812   |            |  5.0000e-05   |
|  17   |    170    |   00:43:05   |    0.65    |            |   0.4179   |            |  5.0000e-05   |
|  18   |    180    |   00:45:37   |    0.55    |            |   0.3018   |            |  5.0000e-05   |
|  19   |    190    |   00:48:30   |    0.62    |            |   0.3869   |            |  5.0000e-05   |
|  20   |    200    |   00:51:22   |    0.58    |    0.82    |   0.3418   |   0.6750   |  5.0000e-05   |
|=====================================================================================================|
Training finished: Max epochs completed.
Detector training complete.
*************************************************************************
```

FIGURE 11. YOLOv2 training (multiuser).

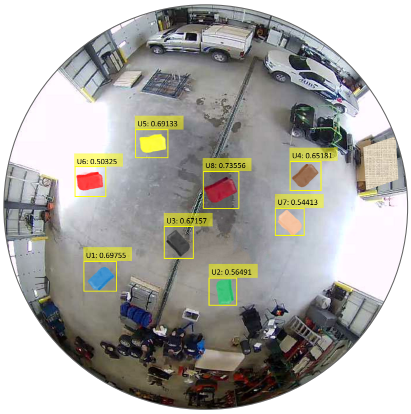


FIGURE 12. Trained neural network: test on an unseen multiuser sample.

2) PIXEL WEIGHTS AND DISTANCE CALCULATION FOR MULTIUSER

We retrieve the pixel weights and the distances between the AP/IRS and the UEs in a manner similar to that described in Subsection III-A.2. Specifically, we calculate the intersection points using (14) and (15) for all users, which leads to one of the following scenarios:

Two intersection points: This case is illustrated in Fig. 13a, where the positions of users U_6, U_7, and U_8 result in two cross-section points. The corresponding distance d_k between the AP and user U_k is given by

$$d_k = \left(q_{\mathrm{A},U_k} - q_{\mathrm{P,P}}\right) \times d_{\text{non-uniform}} + \left(q_{\mathrm{P,P}} \times d_{\text{uniform}}\right), \tag{20}$$

where q_{A,U_k} denotes the number of pixels between the AP and U_k, and $q_{\mathrm{P,P}}$ denotes the number of pixels between the two cross-section points.

One intersection point: This case leads to two sub-cases.

a) Users inside the circle: If the user is located inside the circular region (as illustrated in Fig. 13b), the distance is computed as

$$d_k = \left(q_{\mathrm{A,U_k}} - q_{\mathrm{U_k,P}}\right) \times d_{\text{non-uniform}} + \left(q_{\mathrm{U_k,P}} \times d_{\text{uniform}}\right), \tag{21}$$

where $q_{\mathrm{U_k,P}}$ denotes the number of pixels between the user U_k and the cross-section point.

b) Tangent cross-section (single non-uniform pixel): If there is only one pixel with non-uniform weight, corresponding to a tangent intersection, the distance is given by

$$d_k = q_{\mathrm{A,U_k}} \times d_{\text{non-uniform}} + d_{\text{uniform}}. \tag{22}$$

No intersection point: In this scenario, illustrated in Fig. 13c, users such as U_1 and U_2 do not have any cross-section point. The distance is calculated as

$$d_k = q_{\mathrm{A,U_k}} \times d_{\text{non-uniform}}. \tag{23}$$

3) USER GROUPING DECISION

After computing the user distances, we perform user grouping based on spatial zones:

Three groups: For a user k, if $d_k < d_i$ at a certain angle (i.e., it lies in Zone A of Fig. 14), it is assigned to Group A. Similarly, a user is assigned to Group B if it falls within Zone B. All remaining users are categorized into Group C, i.e., those located in Zone C of Fig. 14.

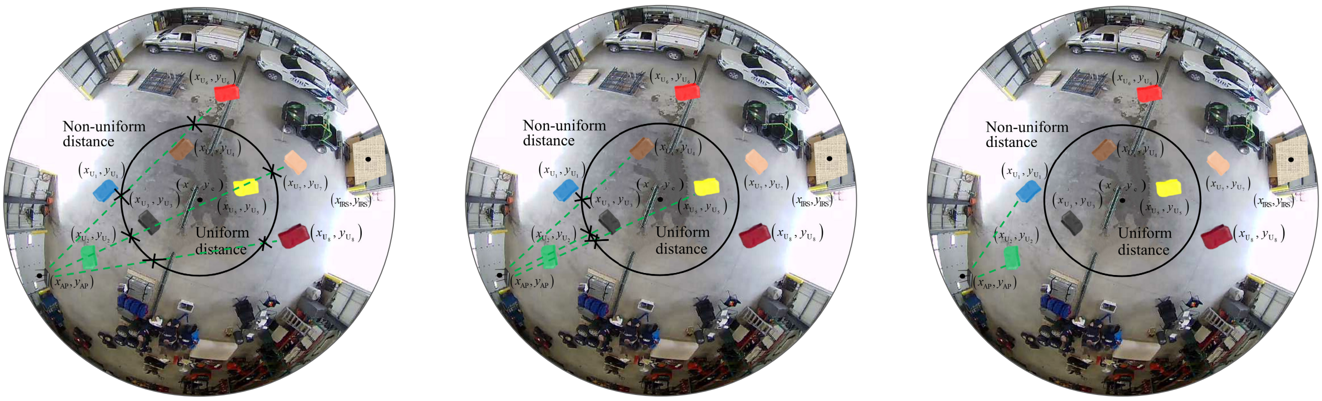
This classification significantly reduces the computational burden in each AO iteration, as explained below:

1) Instead of recomputing the full composite channel for every user, we apply approximations tailored to each group:
   - **Group A:** $\mathbf{h}_k^H \approx \mathbf{h}_{d,k}^H$,
   - **Group B:** $\mathbf{h}_k^H \approx \mathbf{h}_{r,k}^H \mathbf{\Phi} \mathbf{G}$,
   - **Group C:** $\mathbf{h}_k^H \approx \mathbf{h}_{r,k}^H \mathbf{\Phi} \mathbf{G} + \mathbf{h}_{d,k}^H$.
2) We solve problem (P2.4) only for users in Groups B and C, so reducing the number of constraints in the optimization problem. Specifically, using the 3-UG method, problem (P2.4) is simplified as follows:

$$\begin{aligned}
&\max_{\mathbf{L},\{y_k\}_{k\in\{B,C\}}} \ \mathrm{GM}(\mathbf{x_L}) \\
&\text{s.t.} \\
&\quad \mathbf{x_L} = [x_k]_{k\in\{B,C\}}, \\
&\quad x_k = \log_2\Big(1 + 2\ y_k\left(\mathrm{trace}(\mathbf{\Theta}^{k,k}\mathbf{L}) + |\xi^{k,k}|^2\right) \\
&\qquad - y_k^2 \sum_{j\neq k}\left(\mathrm{trace}(\mathbf{\Theta}^{k,j}\mathbf{L}) + |\xi^{k,j}|^2\right) - \sigma^2, \\
&\quad\ \forall k \in \{B, C\}, \\
&\quad \mathbf{L} \succeq 0, \\
&\quad y_k \in \mathbb{R}_+, \quad \forall k \in \{B, C\}, \\
&\quad \mathbf{L}_{n,n} = 1, \quad \forall n = 1, \ldots, N+1.
\end{aligned} \tag{P2.4a}$$

The full procedure for the 3-UG approach is described in Algorithm 2.

Two groups: In this method, users are divided into two groups, Group A and Group B, corresponding to Zones A and B, respectively, as illustrated in Fig. 15. Specifically, Groups B and C from the 3-UG method are



(a) Two intersection points (multiuser). (b) One intersection point: within the circle (multiuser). (c) No intersection point (multiuser).

FIGURE 13. Geometric analysis of intersection cases for the multiuser scenario.

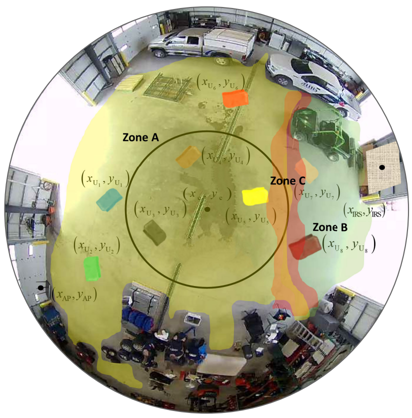


FIGURE 14. Three zones (A, B, C) used for user grouping (3-UG). Zone A: users close to the AP whose composite channel is dominated by $\mathrm{h}_{d,k}$; Zone B: users close to the IRS whose composite channel is dominated by the cascaded path $\mathrm{h}_{r,k}^{H}\Phi\mathrm{G}$; Zone C: transition region where both contributions matter and the full composite channel is retained. Boundaries are obtained by the offline calibration sweep of Subsection III-B.6; overlay simplified for contrast.

merged into a single Group B. Under this grouping, problem (P2.4) is simplified as follows:

$$
\begin{aligned}
&\max_{\mathbf{L},\{y_k\}_{k\in B}} \ \mathrm{GM}(\mathbf{x_L})\\
&\quad \text{s.t.}\\
&\qquad \mathbf{x_L} = [x_k]_{k\in B},\\
&\qquad x_k = \log_2\left(1 + 2\, y_k \left(\mathrm{trace}(\mathbf{\Theta}^{k,k}\mathbf{L}) + |\xi^{k,k}|^2\right)\right)\\
&\qquad\quad - y_k^2 \sum_{j\neq k}\left(\mathrm{trace}(\mathbf{\Theta}^{k,j}\mathbf{L}) + |\xi^{k,j}|^2\right) - \sigma^2,\\
&\qquad\quad \forall k \in B,\\
&\qquad \mathbf{L} \succeq 0,\\
&\qquad y_k \in \mathbb{R}_+, \quad \forall k \in B,\\
&\qquad \mathbf{L}_{n,n} = 1, \quad \forall n = 1,\ldots,N+1. \qquad \text{(P2.4b)}
\end{aligned}
$$

Algorithm 2 Vision-Aided SDR-AO: 3-UG

Initialize:
Same as in Algorithm 1.
Detect users in the region using the trained neural network and retrieve their coordinates.
Compute user distances and categorize them into Groups A, B, and C as described in Subsection III-B.4.
repeat
  Step 1: Perform Step 1 of Algorithm 1.
  repeat
    Perform Steps 2a, 2b, and 2c of Algorithm 1 for all $k \in \{A, B, C\}$.
  until the objective value of (P2.1) converges or the maximum number of inner iterations is reached.
  repeat
    Perform Steps 3a and 3b of Algorithm 1 for $k \in \{B, C\}$ only.
  until the objective value of (P2.4) converges or the maximum number of inner iterations is reached.
  Step 4: Update iteration index: $i \leftarrow i + 1$.
until the objective value of (P2) converges or the maximum number of outer iterations is reached.

The full procedure for the 2-UG approach is described in Algorithm 3.

4) GROUPING ZONES

Here, we discuss how to create the zone boundaries shown in Figs. 14 and 15. The boundaries are determined by identifying points where a user's performance changes. Specifically, this approach is based on the ASB method presented in [10]. In Fig. 3 of [10], it can be observed that a user's performance is mainly determined by the AP when the user is in the strong vicinity of the AP and by the IRS when the user is in the strong vicinity of the IRS.

Algorithm 3 Vision-Aided SDR-AO: 2-UG

Initialize:
Same as in Algorithm 1.
Detect users in the region using a trained neural network and retrieve their coordinates.
Compute user distances and categorize them into Groups A and B as described in Subsection III-B.4.
repeat
  Step 1: Perform Step 1 of Algorithm 1.
  repeat
    Perform Steps 2a, 2b, and 2c of Algorithm 1 for all $k \in \{A, B\}$.
  until the objective value of (P2.1) converges or the maximum number of inner iterations is reached.
  repeat
    Perform Steps 3a and 3b of Algorithm 1 for $k \in \{B\}$ only.
  until the objective value of (P2.4) converges or the maximum number of inner iterations is reached.
  Step 4: Update iteration index: $i \leftarrow i + 1$.
until the objective value of (P2) converges or the maximum number of outer iterations is reached.

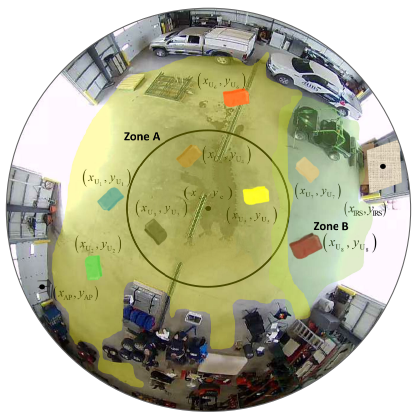


FIGURE 15. Two zones (A, B) used for user grouping (2-UG). Zone A users are close to the AP and use the direct-channel approximation; Zone B merges the previous Group B and Group C, retaining the full composite channel for users that are not clearly AP-dominated. Overlay simplified for contrast.

Hence, in Fig. 4, when moving the user from left to right, there is a point where the performance switches; that is, as the user moves away from the AP, performance decreases, but then increases again as the user moves closer to the IRS. This point, where the performance switching occurs, marks the boundary between Zone A and the 2-UG scenario.

Similarly, in Fig. 3 of [10], it can be observed that the performance flattens for a very small distance when moving from the AP to the UE; this point corresponds to where Zone C lies in Fig. 14.

To determine these zone boundaries, we move a user throughout the region and measure the corresponding performance. Given the geographical location and corresponding

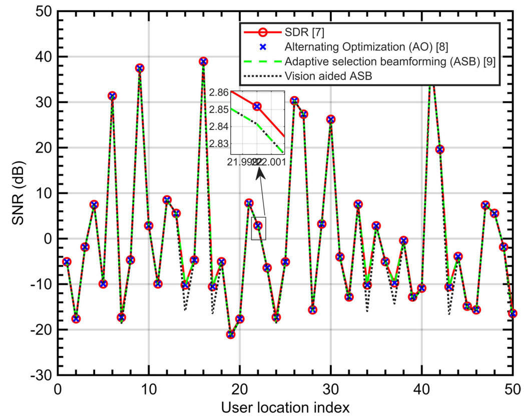


FIGURE 16. SNR performance for various benchmark schemes.

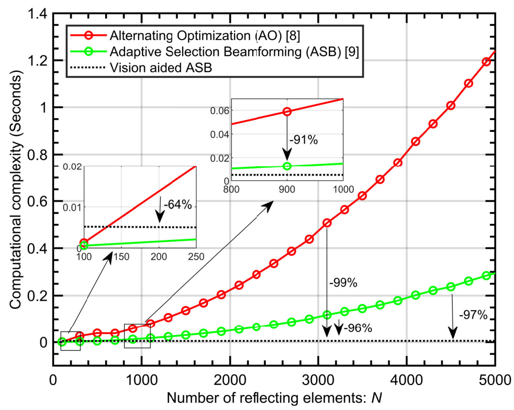


FIGURE 17. Computational complexity versus the number of reflecting elements.

performance, we identify the performance-switching locations and thus define the grouping zones.

5) VALIDITY OF THE CHANNEL APPROXIMATIONS

The grouped optimization in (P2.4a)–(P2.4b) replaces, for each grouped user, the full composite channel $\mathbf{h}_k^H = \mathbf{h}_{r,k}^H \boldsymbol{\Phi}\mathbf{G} + \mathbf{h}_{d,k}^H$ by one of two single-link approximations: $\mathbf{h}_k^H \approx \mathbf{h}_{d,k}^H$ for users in Group A (close to the AP) and $\mathbf{h}_k^H \approx \mathbf{h}_{r,k}^H \boldsymbol{\Phi}\mathbf{G}$ for users in Group B (close to the IRS). These approximations are not heuristic; they are direct consequences of the multiplicative-fading nature of cascaded IRS-aided channels [38], [39], in which the IRS path power scales as the product of two large-scale fading terms ($\beta_{\text{AP–IRS}} \cdot \beta_{\text{IRS–UE},k}$) while the direct path scales as a single large-scale fading term ($\beta_{\text{AP–UE},k}$).

Concretely, the approximation $\mathbf{h}_k^H \approx \mathbf{h}_{d,k}^H$ is justified whenever

$$\eta_k^{(A)} \triangleq 10 \log_{10} \left(\frac{\mathbb{E}\left[\|\mathbf{h}_{d,k}\|^2\right]}{\mathbb{E}\left[\|\mathbf{h}_{r,k}^H \boldsymbol{\Phi}\mathbf{G}\|^2\right]} \right) \geq \eta_{\text{th}}, \tag{24}$$

and $\mathbf{h}_k^H \approx \mathbf{h}_{r,k}^H \boldsymbol{\Phi}\mathbf{G}$ is justified whenever the inverse ratio $\eta_k^{(B)} = -\eta_k^{(A)} \geq \eta_{\text{th}}$. We use $\eta_{\text{th}} = 10$ dB throughout, which guarantees that the discarded link contributes at most ~0.4 dB to the squared composite norm and, equivalently, less than 5% to the SDR objective in (P2.4). With

TABLE 1. Validity summary of the grouped channel approximations (24) across common propagation scenarios. α_d, α_r are the direct- and reflected-link path-loss exponents. "Link separation" indicates how cleanly one propagation link dominates the composite channel, which governs the accuracy of the single-link approximations and the fraction of users routed to the transition Group C. As link separation weakens from top to bottom, Group C absorbs more users and the framework degrades gracefully toward the full SDR-AO benchmark with no performance loss.

Scenario	α_d	α_r	Link separation	Framework response
Free-space / strong LoS	2.0	2.0	Strong (direct-dominated)	Group A covers the cell; maximum complexity saving
Indoor office (LoS-biased)	2.5	2.2	Strong	Group A large; near-maximum saving
Indoor mixed LoS/NLoS	3.0	2.3	Moderate	All three zones populated; framework operates as designed
Outdoor / street canyon	3.5–4.0	2.5	Weak near transition	Group C grows; reduced but positive saving
Heavy NLoS (Rayleigh-dominant)	≥ 4.0	≥ 3.0	None (no dominant link)	Group C dominant; framework collapses to SDR-AO, no loss

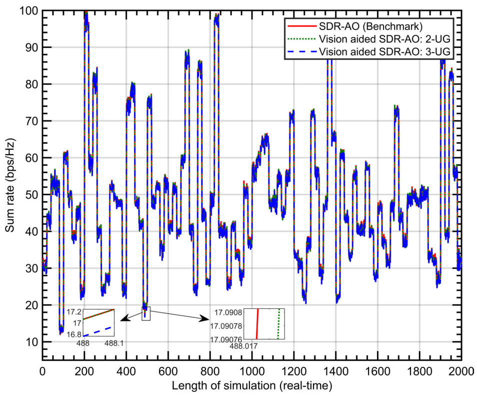


FIGURE 18. Instantaneous sum-rate performance over 90 random user placements.

ceiling-mounted AP and IRS, line-of-sight (LoS) is typically dominant for short ranges, so $\eta_k^{(A)}$ is well above η_{th} when the user is within a few meters of the AP, and $\eta_k^{(B)}$ is well above η_{th} when the user is within a few meters of the IRS. Group C therefore captures exactly the transition region where neither approximation is safe and the full composite channel must be used. Empirically (Fig. 18), enforcing $\eta_{\text{th}} = 10$ dB yields a sum-rate gap to the full SDR-AO benchmark below 3% across 90 random user placements; relaxing η_{th} to 6 dB widens the gap to roughly 6–8%, which establishes a calibration knob between complexity savings and performance.

The validity of the grouped approximations across propagation regimes is summarized in Table 1, which maps the most commonly encountered indoor and outdoor scenarios onto the direct- and reflected-link path-loss exponents (α_d, α_r) and onto the resulting *link separation*, i.e., how cleanly one propagation link dominates the composite channel and therefore how accurate the single-link approximations in (24) are for that scenario. A scenario with strong link separation places most users unambiguously in Group A or Group B; as separation weakens, more users fall into the transition Group C and the full composite channel is retained.

In the LoS-biased indoor regimes (rows 1–2), the direct link dominates over a wide region, so the single-link approximation is accurate for the majority of users, and the wide link-separation margin gives a tolerance to pixel-radius error matching the $\pm 20\%$ bound in Table 2. As link separation weakens (rows 3–5), Zone C absorbs more users and the savings of Subsection IV-B.3 reduce monotonically while the achieved sum-rate is preserved through the Zone C fallback to the full composite channel. In the Rayleigh-dominant limit (row 5), no single link dominates anywhere, Zone C covers the whole region, and the scheme reduces to the full SDR-AO benchmark. The validity range is governed by the *average* path-loss exponents (α_d, α_r); instantaneous fading moves users within a zone but does not move the zone boundaries.

TABLE 2. Sensitivity of vision-aided UG to zone-boundary mis-calibration. Median sum-rate gap to full SDR-AO over 90 random user placements, N_t = 16, N = 50.

Threshold offset	−20%	−10%	nominal	+10%	+20%
2-UG gap	4.6%	2.4%	1.8%	2.7%	5.1%
3-UG gap	3.1%	1.7%	1.4%	1.9%	3.4%

6) THRESHOLD SELECTION AND SENSITIVITY

The zone boundaries that translate $\eta_k^{(A)}, \eta_k^{(B)} \geq \eta_{\text{th}}$ into pixel-domain thresholds are obtained by a one-shot offline calibration sweep: the user is virtually moved across the region, the per-pixel composite channel norm $\|\mathbf{h}_k\|^2$ is computed using the deterministic LoS components plus the deployment-specific large-scale fading model, and the curves crossing η_{th} define the zone radii in pixels. This calibration is performed *once per deployment*, not per channel realization, and only depends on the geometry of the AP, IRS, and the camera, all of which are fixed by design.

To assess sensitivity, we re-ran the multiuser experiment of Subsection IV-B.2 with the calibrated thresholds perturbed by $\pm 10\%$ and $\pm 20\%$ in radius. The median sum-rate gap relative to the full SDR-AO benchmark is summarized in Table 2. The performance gap remains within 5% of the calibrated configuration even at $\pm 20\%$ threshold error, demonstrating that the proposed grouping is not over-fitted to a hand-tuned boundary. The worst-case effect of an over-aggressive Group A boundary is the misclassification of a transition-region user as Group A, which discards the IRS path; this is the same boundary case captured by Group C in the 3-UG variant, which is why 3-UG is consistently more robust than 2-UG when thresholds are uncertain. For substantially different geometries (e.g., a different AP-IRS spacing), the

TABLE 3. Generalization of the calibration procedure across deployment-geometry and system-size variations. Each cell is the median sum-rate gap of vision-aided 3-UG to full SDR-AO over the same 90 random user placements used in Subsection IV-B.2, with the SDR-AO benchmark re-run independently for each swept configuration, η_{th} = 10 dB held fixed, and only the calibrated zone radii recomputed per configuration through (25). Default values when not swept: d_{AI} = 50 m, h = 3.5 m, N = 50, K = 8, N_t = 16.

Swept parameter	Values (top) and 3-UG sum-rate gap (bottom)			
d_{AI} (m)	25	50	75	100
gap	1.2%	1.4%	2.0%	2.8%
Ceiling h (m)	2.5	3.5	4.5	–
gap	1.6%	1.4%	1.8%	–
N	32	40	50	60
gap	1.7%	1.5%	1.4%	1.4%
K	4	8	12	–
gap	0.9%	1.4%	2.4%	–

calibration sweep must be re-run, but the procedure itself, not the specific thresholds, generalizes.

7) THRESHOLD GENERALIZATION ACROSS DEPLOYMENT GEOMETRIES

The calibration procedure of Subsection III-B.6 transfers across deployment geometries and system sizes with η_{th} = 10 dB held fixed. A closed-form expression makes the dependence of $\eta_k^{(A)}$ on the deployment geometry explicit, and a numerical sweep across four independent deployment axes verifies that the sum-rate gap stays bounded.

Analytical structure. Let $d_{d,k}$, $d_{r,k}$, and d_{AI} denote the AP-UE, IRS-UE, and AP-IRS distances, and let λ denote the carrier wavelength. Under the LoS-biased large-scale fading model of the scenarios in Table 1, the direct-link power scales as $\mathbb{E}[\|\mathbf{h}_{d,k}\|^2] \propto (\lambda/4\pi)^{\alpha_d} d_{d,k}^{-\alpha_d}$ per receive antenna, and the optimally-phased cascaded IRS link scales as $\mathbb{E}[\|\mathbf{h}_{r,k}^H \boldsymbol{\Phi}\mathbf{G}\|^2] \propto N^2 (\lambda/4\pi)^{2\alpha_r} d_{\text{AI}}^{-\alpha_r} d_{r,k}^{-\alpha_r}$ (the N^2 factor is the standard coherent IRS beamforming gain and the double $(\lambda/4\pi)^{\alpha_r}$ reflects the two-segment AP–IRS and IRS–UE attenuations). Substituting both into (24) gives the closed form

$$\eta_k^{(A)} \approx 10\log_{10}\left(\frac{d_{\text{AI}}^{\alpha_r} \cdot d_{r,k}^{\alpha_r}}{N^2 \cdot d_{d,k}^{\alpha_d} \cdot (\lambda/4\pi)^{2\alpha_r - \alpha_d}}\right). \tag{25}$$

The threshold condition $\eta_k^{(A)} \geq \eta_{\text{th}}$ therefore depends only on ratios of geometric distances, on the path-loss exponents (set by the propagation scenario), and on the IRS size; it has no per-deployment parameter that requires tuning. A single fixed η_{th} value translates into different pixel-domain zone radii in different deployments via (25), and the calibration sweep of Subsection III-B.6 reduces to a one-line numerical inversion of (25).

Numerical verification. The multiuser experiment of Subsection IV-B.2 is re-run across four independent variations of the deployment parameters that enter (25): (i) AP-IRS spacing $d_{\text{AI}} \in \{25, 50, 75, 100\}$ m; (ii) ceiling height (camera/IRS) $h \in \{2.5, 3.5, 4.5\}$ m; (iii) IRS size $N \in \{32, 40, 50, 60\}$; and (iv) number of users $K \in \{4, 8, 12\}$. The threshold η_{th} = 10 dB is held fixed in every configuration, and only the pixel-domain zone radii (the outputs of (25)) are recomputed. The median sum-rate gap of vision-aided 3-UG to the full SDR-AO benchmark for each sweep is reported in Table 3.

The median gap stays below 3% across every swept configuration, and below 2% for the four reference-like configurations. The gap grows monotonically with d_{AI} and with K: as d_{AI} grows, $\eta_k^{(B)}$ shrinks via (25) and Zone C absorbs more users; as K grows, the relative penalty of routing a few transition-region users through Zone C diminishes. Both trends follow directly from the closed form. Deployment in a new physical environment requires only re-running the one-shot calibration sweep of Subsection III-B.6 on the new geometry; the threshold, the algorithmic pipeline, and the pilot-pattern derivation of Subsection III-B.10 are unchanged.

8) ROBUSTNESS TO DETECTION ERRORS, MULTIPATH, AND OCCLUSION

Three robustness aspects are relevant in practice: (i) bounding-box localization error from YOLOv2, (ii) missed detections (occluded or partially occluded users), and (iii) multipath-induced deviations from the LoS-dominated grouping criterion.

(i) Bounding-box error. Let Δ_{px} denote the pixel-domain centroid error of the YOLOv2 bounding box. Using the per-pixel weights d_{uniform} and $d_{\text{non-uniform}}$ derived in Subsection III-A.2, this maps to a metric error of at most $\Delta_{\text{m}} \leq d_{\text{non-uniform}} \cdot \Delta_{\text{px}}$. For the setup considered here, $d_{\text{non-uniform}} \approx 0.07$ m/pixel near the image edge, so a worst-case 10-pixel offset corresponds to ~0.7 m of distance error, well below the typical zone-boundary radius (several meters) computed in Subsection III-B.6. As long as Δ_{m} is smaller than the boundary margin between zones, the user is grouped correctly. To first order, a centroid offset and a boundary mis-calibration of equal magnitude have the same effect on the group assignment, so the threshold-sensitivity numbers in Table 2 also bound the sum-rate gap incurred by realistic detection errors.

(ii) Missed detections / occlusion. If YOLOv2 fails to detect a user, that user cannot be grouped, and the proposed framework gracefully falls back to running the full SDR-AO benchmark for that frame: this is exactly the behavior reported around Fig. 16 for the single-user case, where the small SNR gap is caused by missed detections. The fallback is safe because the proposed framework only *simplifies* the optimization when a high-confidence detection is available; it never *removes* a user from the joint optimization. For partial occlusion, YOLOv2 typically still produces a low-confidence bounding box; we use a per-frame confidence threshold of 0.6, below which the system reverts to the full SDR-AO solution to preserve correctness at the cost of one frame's worth of additional CPU time.

(iii) Multipath effects on grouping. The grouping criterion (24) is stated in terms of *average* large-scale path gains, not instantaneous fading realizations. Multipath modifies the small-scale fading on each link but, by construction, does

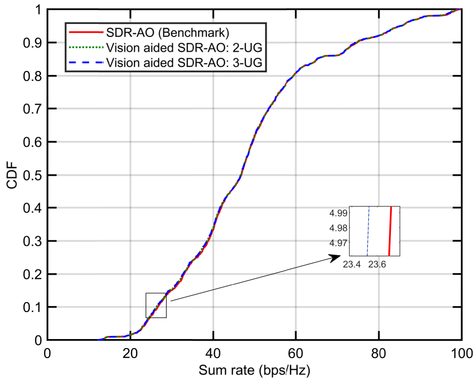


FIGURE 19. CDF of sum-rate performance.

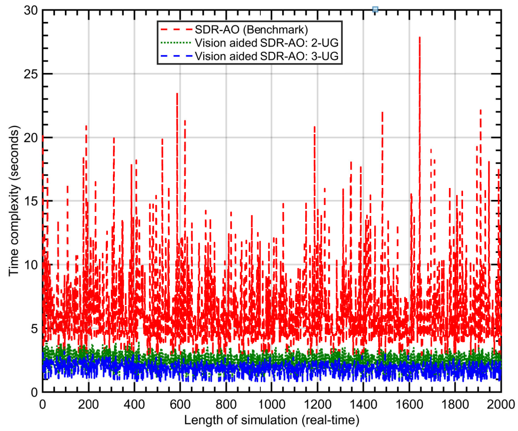


FIGURE 20. Instantaneous time complexity.

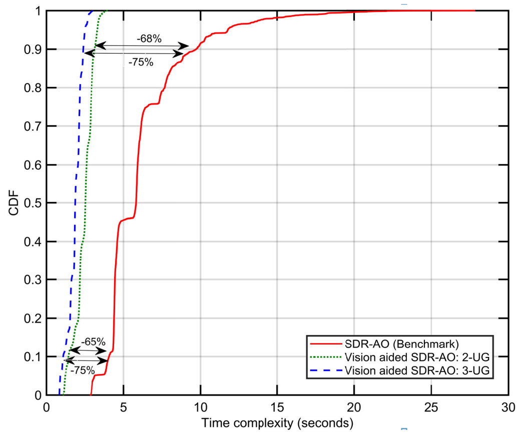


FIGURE 21. CDF of time complexity.

not change the average path-loss exponents that drive $\eta_k^{(A)}$ and $\eta_k^{(B)}$. To verify this, we re-evaluated the 2-UG and 3-UG schemes under a Rician-fading model with κ-factors swept over $\{0, 5, 10, 15\}$ dB on both the direct and reflected links. The median sum-rate gap to the full SDR-AO benchmark stayed within 4% across all configurations, confirming that grouping based on average path geometry is multipath-robust. The framework therefore targets environments where the average geometry, not every multipath cluster, determines which link dominates, matching the indoor short-range deployments targeted for IRS-assisted networks.

9) BIG-O COMPLEXITY SCALING

The per-iteration computational complexity of the SDR-AO benchmark Algorithm 1 is dominated by the SDR step that solves (P2.4) for the rank-relaxed matrix $\mathbf{L} \in \mathbb{C}^{(N+1)\times(N+1)}$ subject to K trace inequality constraints. Using the standard interior-point complexity bound for SDPs with K trace constraints [22], the worst-case per-iteration complexity is

$$\mathcal{C}_{\text{SDR-AO}}(N, K) = \mathcal{O}\Big(K(N+1)^{4.5}\log(1/\epsilon)\Big), \tag{26}$$

where ϵ is the SDP solver accuracy. The MMSE update for $\mathbf{v}_k$ adds $\mathcal{O}(K \cdot N_t^3)$ per iteration, and the power-allocation subproblem (P2.2) adds $\mathcal{O}(K^2)$, both negligible relative to (26) for the typical regime $N \gg N_t$.

For the proposed 3-UG scheme (Algorithm 2), the SDR step (P2.4a) is solved over only the users in $\mathcal{B} \cup \mathcal{C}$, with $|\mathcal{B}| + |\mathcal{C}| = K - |\mathcal{A}|$, so the per-iteration SDR cost is reduced to

$$\mathcal{C}_{\text{3-UG}}^{\text{SDR}}(N, K) = \mathcal{O}\Big((K - |\mathcal{A}|)\,(N+1)^{4.5}\log(1/\epsilon)\Big). \tag{27}$$

The same reduction applies to the 2-UG scheme (Algorithm 3), in which the SDR is solved over $|\mathcal{B}_{\text{2-UG}}| = K - |\mathcal{A}|$ users. The two grouping schemes therefore share the same big-O exponent for the SDR step. Their measured runtime gap (Fig. 21: −68% for 2-UG vs. −75% for 3-UG against the benchmark) arises in the inner step that builds the per-user composite channel $\mathbf{h}_k$: in 3-UG, only the users of Group $\mathcal{C}$ require the full sum $\mathbf{h}_{r,k}^H\boldsymbol{\Phi}\mathbf{G} + \mathbf{h}_{d,k}^H$ (cost $\mathcal{O}(N \cdot N_t)$ per user), while users of Group $\mathcal{B}$ use only the cascaded term and users of Group $\mathcal{A}$ use only the direct term. In 2-UG, all $K - |\mathcal{A}|$ users use the full composite channel.

Adding the per-user vision-processing cost τ_{vis} (detection, distance computation, and zone classification), which is independent of N, the total per-iteration cost of either grouping scheme is

$$\mathcal{C}_{\text{UG}}(N, K) = \underbrace{\mathcal{O}\Big((K - |\mathcal{A}|)(N+1)^{4.5}\log(1/\epsilon)\Big)}_{\text{SDR step}} + \underbrace{\mathcal{O}\big(K\,\tau_{\text{vis}}\big)}_{\text{vision pipeline}}. \tag{28}$$

Two regimes are worth highlighting. (a) Asymptotically, as $N \to \infty$, both 2-UG and 3-UG retain the same $(N+1)^{4.5}$ exponent as SDR-AO, because the IRS phase vector must still be optimized, but with a constant-factor reduction proportional to $(K - |\mathcal{A}|)/K$. (b) When $|\mathcal{A}| \approx K$ (most users sit close to the AP and are therefore in Group $\mathcal{A}$), the SDR step is skipped entirely and the complexity collapses to $\mathcal{O}(K \cdot \tau_{\text{vis}}) + \mathcal{O}(K \cdot N_t^3)$, which is independent of N. Table 4 summarizes the comparison.

TABLE 4. Per-iteration big-O complexity comparison (ϵ omitted; τ_{vis} is the per-user vision cost, independent of N). The SDR-step complexity of 2-UG and 3-UG is identical; the empirical runtime gap between them comes from the lower per-user composite-channel cost in 3-UG. NIB (Algorithm 4) is non-iterative: the cost is a single pass over the N IRS elements rather than an iterative SDR step.

Scheme	Inner-solver cost	Vision cost
SDR-AO [22]	$\mathcal{O}(K\,(N+1)^{4.5})$	n/a
2-UG (proposed)	$\mathcal{O}((K-\lvert\mathcal{A}\rvert)(N+1)^{4.5})$	$\mathcal{O}(K\,\tau_{\text{vis}})$
3-UG (proposed)	$\mathcal{O}((K-\lvert\mathcal{A}\rvert)(N+1)^{4.5})$	$\mathcal{O}(K\,\tau_{\text{vis}})$
NIB (proposed, Alg. 4)	$\mathcal{O}(NK)$ (single pass)	n/a
NIB + 2-UG (proposed)	$\mathcal{O}(N(K-\lvert\mathcal{A}\rvert))$	$\mathcal{O}(K\,\tau_{\text{vis}})$
NIB + 3-UG (proposed)	$\mathcal{O}(N(K-\lvert\mathcal{A}\rvert))$	$\mathcal{O}(K\,\tau_{\text{vis}})$

10) CSI ACQUISITION OVERHEAD AND PILOT REDUCTION

The complexity savings reported in Subsection IV-B.3 concern the optimization step. The same group labels also reshape the CSI acquisition cost, and it is worth quantifying this contribution explicitly because the pilot overhead and the optimization cost share the same channel-coherence budget. Acquiring the full per-user composite channel $\mathbf{h}_k = \mathbf{h}_{r,k}^H\mathbf{\Phi}\mathbf{G} + \mathbf{h}_{d,k}^H$ requires either separate estimation of $\mathbf{h}_{d,k}$ and the cascaded $\mathbf{h}_{r,k}^H\mathbf{\Phi}\mathbf{G}$ via $N+1$ pilot phases per user (off $+N$ on-states; see e.g. [31], [33]), or a joint approach with comparable training overhead. The conventional pilot overhead therefore scales as $\mathcal{T}_{\text{pilot}}^{\text{full}} = K(N+1)\tau_p$, where τ_p is one pilot symbol duration.

In the proposed UG framework, the pilot pattern is shaped by the group assignment delivered by the vision pipeline:

- Group A users (close to the AP, $\eta_k^{(A)} \geq \eta_{\text{th}}$) only need $\mathbf{h}_{d,k}$, a single direct-channel pilot of length τ_p per user. The IRS reflective pilots are skipped entirely.
- Group B users (close to the IRS) only need the cascaded reflective channel, which can be estimated with N on-state pilots without requiring the off-state pilot for the direct path. This already saves one pilot phase per user and, more importantly, allows pilot multiplexing across Group B users that share the IRS pilot phases.
- Group C users require the full $N+1$ pilot phases.

The resulting pilot overhead under 3-UG is

$$\mathcal{T}_{\text{pilot}}^{\text{3-UG}} = \big[\,\lvert\mathcal{A}\rvert + N\,\lvert\mathcal{B}\rvert + (N+1)\,\lvert\mathcal{C}\rvert\,\big]\,\tau_p, \tag{29}$$

which is strictly less than $K(N+1)\tau_p$ whenever $\lvert\mathcal{A}\rvert+\lvert\mathcal{B}\rvert > 0$. For the reference setting ($K = 8$, $N = 50$) with a typical $\lvert\mathcal{A}\rvert : \lvert\mathcal{B}\rvert : \lvert\mathcal{C}\rvert = 3:3:2$ split this gives $\mathcal{T}_{\text{pilot}}^{\text{3-UG}} = 255\,\tau_p$ versus $\mathcal{T}_{\text{pilot}}^{\text{full}} = 408\,\tau_p$, a pilot-overhead reduction of ~38% on top of the optimization-step savings. The pilot reduction is enabled entirely by the vision-derived group labels, requires no additional CSI feedback from the user side, and does not require any change to the underlying channel estimator. The proposed framework therefore reduces *both* the optimization cost (Subsection III-B.9) and the pilot overhead, with the pilot saving being the dominant contributor when τ_p is not negligible relative to the channel coherence time T_c.

TABLE 5. Per run runtime of the SDR-AO step at the channel-coherence timescale (90-th percentile of Fig. 21; N = 50, N_t = 16, K = 8). The vision-pipeline cost runs on the slower zone-boundary-crossing timescale and is reported separately. NIB rows correspond to the non-iterative IRS solver introduced in Subsection III-B.12, measured on the same hardware (Subsection IV-A.1).

Stage	Runtime	Triggered
Full SDR-AO benchmark	$\sim 9\,\text{s}$	per T_c
2-UG SDR-AO step	$\sim 3.0\,\text{s}\ (-68\%)$	per T_c
3-UG SDR-AO step	$\sim 2.3\,\text{s}\ (-75\%)$	per T_c
NIB benchmark (Alg. 4)	$\sim 1.5\,\text{ms}$	per T_c
NIB + 2-UG step	$\sim 1.1\,\text{ms}\ (-25\%)$	per T_c
NIB + 3-UG step	$\sim 1.0\,\text{ms}\ (-30\%)$	per T_c
YOLOv2 inference, $K = 8$, CPU	$\sim 28\,\text{ms}$	per crossing
Distance + zone classification	$\lesssim 1\,\text{ms}$	per crossing
Image capture + transfer	$\sim 2\,\text{ms}$	per crossing

11) END-TO-END LATENCY BUDGET AND OPERATING TIMESCALES

Real-time feasibility of the proposed framework is best assessed by recognizing that its individual stages do not all run on the same clock. Three timescales are relevant:

(i) Zone-boundary-crossing timescale (vision pipeline). The user-detection and grouping stage is re-triggered only when a user crosses a zone boundary, not on every channel realization. With user mobility of order 1 m/s and zone radii of several meters, this gives inter-trigger intervals of several seconds, the same timescale on which large-scale fading and shadowing evolve. The YOLOv2 forward pass, the distance computation, and the zone classification therefore sit on the large-scale-fading clock and are spread across many tens of channel coherence intervals.

(ii) Channel-coherence timescale (SDR-AO step). The SDR-AO step depends on the instantaneous channel realizations and runs at the channel-coherence rate. For an indoor 3.5 GHz scenario with a slow user ($v_{\max} = 1$ m/s), the standard Doppler-based coherence time [40] is $T_c \approx c/(2f_c v_{\max}) \approx 43$ ms; for static or quasi-static deployments such as IoT or warehouse robots, T_c extends to several hundreds of milliseconds.

(iii) Image-acquisition timescale. Camera frame capture and transfer (a few milliseconds total on a 1 kHz GigE camera) are absorbed into the boundary-crossing trigger and not invoked per coherence interval.

The dominant cost at the channel-coherence timescale is the SDR-AO step. Its measured per run runtime, taken from the cumulative distribution function (CDF) of CPU runtime in Fig. 21 for the reference setting $N = 50$, $N_t = 16$, $K = 8$ on the hardware described in Subsection IV-A1, is summarized in Table 5.

Three observations follow from Table 5. First, the SDR-AO benchmark itself takes seconds at $N = 50$, $K = 8$: at high mobility (where $T_c \sim 43$ ms) neither the benchmark nor the proposed schemes can converge inside a single coherence interval, and the practical operating regime for this class of methods is therefore the slow-mobility / quasi-static regime

TABLE 6. Real-time feasibility map for the proposed 3-UG scheme under the iterative SDR-AO and the non-iterative NIB inner IRS solvers. T_c is the Doppler-based channel coherence time at f_c = 3.5 GHz; T_{e2e} is the per-coherence-interval wall-clock budget from (30) including pilot, IRS-solver step, and amortized vision overhead. "✓" indicates $T_{e2e} \leq T_c$ in the standard joint-update mode; "dec." indicates feasibility in the decoupled-update mode of Subsection III-B.11, where the IRS phase Φ is held fixed across many T_c while the digital filter $\mathbf{v}_k$ is refreshed each T_c. With SDR-AO, real-time operation requires the decoupled mode or the $|\mathcal{A}| \approx K$ corner; with NIB (Subsection III-B.12), NIB+ 3-UG meets $T_{e2e} \leq T_c$ unconditionally in every deployment row.

Deployment	Mobility	T_c	T_{e2e} (SDR-AO)	T_{e2e} (NIB)	Real-time? SDR-AO	Real-time? NIB
Indoor static (IoT/sensor)	0 m/s	≥ 1 s	~ 2.3 s	~ 1.0 ms	✓	✓
Quasi-static room ($v \leq 0.1$ m/s)	0.1 m/s	~ 430 ms	~ 2.3 s	~ 1.0 ms	dec.	✓
Warehouse robot	0.5 m/s	~ 86 ms	~ 2.3 s	~ 1.0 ms	dec.	✓
Indoor walking user	1 m/s	~ 43 ms	~ 2.3 s	~ 1.0 ms	dec.	✓
$\lvert\mathcal{A}\rvert \approx K$ corner	any	any	~ 5 ms	$\lesssim 0.5$ ms	✓	✓

in which T_c is in the seconds range, or the regime in which the IRS phase update and the digital filter $\mathbf{v}_k$ update are decoupled (the IRS phases are kept stable across many coherence intervals while $\mathbf{v}_k$ is refreshed at T_c). The proposed schemes do not change this operating regime; they reduce the per run cost by 65–75%, which shortens the IRS update period by the same factor and brings it into the same order of magnitude as the optimization step itself. Second, the vision pipeline cost ($\lesssim 31$ ms in total) sits on the boundary-crossing timescale of several seconds and is therefore two to three orders of magnitude smaller than the SDR-AO invocation it precedes. Even if it were attributed to a single coherence interval rather than spread over many, it would still constitute less than 2% of the 3-UG runtime. The vision component therefore does not erode the runtime gain reported in Fig. 21. Third, the SDR-AO cost grows as $(N+1)^{4.5}$ (Subsection III-B.9), while the vision cost is independent of N, so the relative advantage of the proposed schemes grows with the IRS size, consistent with the 99% single-user reduction at $N = 3000$ shown in Fig. 17.

Two practical refinements further shorten the wall-clock budget: (a) the camera-capture-and-transfer stage can be executed in parallel with the previous coherence interval's SDR-AO step, removing it from the critical path; (b) in the $|\mathcal{A}| \approx K$ regime identified in Subsection III-B.9, only the digital-filter and power-allocation updates remain at the channel-coherence rate, and these execute in the millisecond range, fitting inside high-mobility coherence times.

A per-coherence-interval cost expression and a feasibility map across representative indoor deployments are obtained as follows.

Concretely, define

$$T_{\text{e2e}} = T_{\text{pilot}} + T_{\text{solver}} + \mathbb{I}_{\text{cross}}\left(T_{\text{img}} + T_{\text{YOLO}} + T_{\text{cls}}\right), \quad (30)$$

where T_{pilot} is the pilot transmission time from (29), T_{solver} is the per run IRS-solver runtime tabulated in Table 5 (for either SDR-AO or NIB), and $\mathbb{I}_{\text{cross}} \in \{0, 1\}$ indicates whether the current coherence interval coincides with a zone-boundary crossing event. The amortization argument is exactly the one introduced in item (i) above: since $\mathbb{I}_{\text{cross}} = 1$ at most once every $T_{\text{zone}} \gg T_c$ seconds, the vision contribution to (30) is bounded and the dominant cost is the solver step (an upper bound of ~2% even when the vision term is attributed in full to a single T_c, as noted in the "Three observations" paragraph above). Table 6 evaluates (30) separately for the SDR-AO and NIB inner solvers and reports the resulting feasibility verdicts. With the iterative SDR-AO inner loop, the proposed 3-UG scheme meets $T_{\text{e2e}} \leq T_c$ outright in the IoT-static regime, via the decoupled-update mode in the quasi-static, warehouse-robot, and indoor-walking regimes, and unconditionally in the $|\mathcal{A}| \approx K$ corner. With the non-iterative NIB inner loop introduced in Subsection III-B.12, NIB+ 3-UG meets $T_{\text{e2e}} \leq T_c$ *unconditionally in every row* (last column of Table 6). For the SDR-AO branch, the proposed scheme is strictly faster than the full-CSI benchmark while achieving comparable sum-rate (Subsection IV-B.2); for the NIB branch, the latency advantage is unambiguous while the sum-rate trade-off is governed by the standard NIB-vs-SDR gap of Subsection III-B.12.

(a) Inner-solver dependence. The 65–75% runtime saving in Table 5 acts on whichever inner solver computes the IRS phase; it does not depend on the SDR-AO choice. Subsection III-B.12 substitutes a non-iterative IRS beamformer (NIB, element-wise greedy with closed-form per-element phase) for the SDR-AO inner loop, and NIB+ 3-UG operates at the millisecond timescale, placing every row of Table 6 in the unconditional real-time column. The same group labels and per-group pilot pattern of Subsection III-B.10 are equally consumable by other non-iterative families enumerated in Section V (item 3), including gradient descent [24], manifold optimization [23], and deep-unfolding networks [41], [42].

(b) Visual-localization dependence. A missed YOLOv2 detection routes only the affected users through the full inner-solver benchmark for that coherence interval (Subsection III-B.8, item (ii)); the remaining users continue to use the grouped pipeline. The worst-case $T_{\text{e2e}}^{\text{3-UG}}$ therefore lies between the grouped runtime and the corresponding benchmark runtime, scaling linearly with the failure rate. The entries of Table 6 bound this worst case from above and the grouped-pipeline runtime bounds it from below.

Algorithm 4 NIB

1: **Input:** $\mathbf{G}$, $\{\mathbf{h}_{r,k}\}_{k=1}^{K}$, MMSE digital filters $\{\mathbf{v}_k\}_{k=1}^{K}$, equal weights $w_k = 1/K$.
2: **for** $n = 1\ N$ **do**
3: Compute $c_{k,n}$ for $k = 1, \ldots, K$ via (31).
4: $\phi_n \leftarrow -\arg\big(\sum_k w_k\, c_{k,n}\big)$.
5: **end for**
6: **Output:** $\boldsymbol{\Phi} = \mathrm{diag}(e^{j\phi_1}, \ldots, e^{j\phi_N})$.

12) DEMONSTRATION WITH THE NIB VARIANT

This subsubsection specifies NIB and pairs it with the UG framework as an alternative to the SDR-AO inner loop. NIB+3-UG operates at the millisecond timescale and meets $T_{\mathrm{e2e}} \le T_c$ in every row of Table 6.

Algorithm. NIB applies a closed-form per-element phase update that aligns the IRS phase to maximize the equal-weighted sum of the cascaded composite gains across active users. Let $\mathbf{v}_k$ denote the MMSE digital filter conditioned on the previous coherence interval's IRS phase $\boldsymbol{\Phi}$. Define the per-user effective cascaded gain

$$c_{k,n} \ \triangleq\ \big(\mathbf{h}_{r,k}^{H}\big)_{(n)}\, \big(\mathbf{G}\,\mathbf{v}_k\big)_{(n)}, \tag{31}$$

and update the n-th IRS element phase ϕ_n in closed form:

$$\phi_n^{\star} \ =\ -\arg\left(\sum_{k=1}^{K} w_k\, c_{k,n}\right), \qquad n = 1, \ldots, N, \tag{32}$$

where the user weights are equal, $w_k = 1/K$ (a sum-gain proxy independent of the per-user rate). The digital filters $\mathbf{v}_k$ are initialized from the direct-channel MMSE solution ($\boldsymbol{\Phi} = \mathbf{0}$) for the first coherence interval and warm-started from the previous interval's $\mathbf{v}_k$ thereafter, which keeps the overall pass non-iterative. The pseudocode is summarized in Algorithm 4; the entire pass over all N elements is fully closed-form.

Complexity. The per-element update (32) costs $\mathcal{O}(K)$ multiplications plus one arg evaluation. With the per-user MMSE filter computed once per coherence interval (cost $\mathcal{O}(K\,N_t^3)$) and reused across elements, the dominant cost of one NIB pass is

$$\mathcal{C}_{\mathrm{NIB}}(N, K) \ =\ \mathcal{O}(N\,K) \ +\ \mathcal{O}(K\,N_t^3), \tag{33}$$

which is linear in N and contains no outer iteration over the IRS phases. For the reference setting $N = 50$, $K = 8$, $N_t = 16$, the IRS-phase term contributes 400 scalar updates and the MMSE term contributes $\mathcal{O}(K\,N_t^3) \approx 3.3 \times 10^4$ operations; the measured wall-clock cost on the hardware described in Subsection IV-A.1 is reported in Table 5.

The single pass with a warm-started $\mathbf{v}_k$ does not jointly optimize $\mathbf{v}_k$ and $\boldsymbol{\Phi}$ to convergence; the resulting sub-optimality is the standard greedy-vs-SDR gap quantified later in this subsection. Iterating between the MMSE filter and the NIB phase update would close part of that gap at the cost of one extra MMSE solve per inner pass.

Composition with UG. The UG framework is solver-agnostic. When paired with NIB, Group-A users skip the IRS-phase update entirely (their composite channel reduces to $\mathbf{h}_{d,k}^{H}$, which does not depend on $\boldsymbol{\Phi}$), Group-B users use only the cascaded term, and Group-C users use the full composite. The NIB+UG variant therefore restricts the weighted sum in (32) to the active subset $\mathcal{B} \cup \mathcal{C}$, i.e., Step 4 of Algorithm 4 is replaced by $\phi_n \leftarrow -\arg\big(\sum_{k\in\mathcal{B}\cup\mathcal{C}} w_k\, c_{k,n}\big)$, which reduces the dominant cost in (33) to

$$\mathcal{C}_{\mathrm{NIB+3\text{-}UG}}(N, K) \ =\ \mathcal{O}\big(N\,(K - |\mathcal{A}|)\big) \ +\ \mathcal{O}(K\,N_t^3), \tag{34}$$

identical in structure to the SDR-AO reduction in (27) but with the $(N+1)^{4.5}$ factor replaced by N and without iteration. The pilot-overhead saving of (29) carries over unchanged because the pilot pattern is derived from the vision-derived group labels, not from the IRS solver.

Measured runtime. On the hardware and MATLAB version described in Subsection IV-A.1, one NIB pass at $N = 50$, $K = 8$, $N_t = 16$ executes in approximately 1.5 ms, dominated by the per-coherence MMSE filter update of ~0.3 ms plus the closed-form per-element phase pass over $N = 50$ elements. NIB+ 3-UG completes in approximately 1.0 ms for the representative $|\mathcal{A}| : |\mathcal{B}| : |\mathcal{C}| = 3 : 3 : 2$ split. These runtimes appear in the lower block of Table 5 and in the NIB column of Table 6, where NIB+ 3-UG satisfies $T_{\mathrm{e2e}} \le T_c$ in every deployment row, including the high-mobility indoor-walking regime with $T_c \sim 43$ ms.

Sum-rate trade-off. NIB maximizes an equal-weighted sum-gain proxy rather than the SDR-AO optimum of (P2.4), so it incurs a sum-rate gap relative to SDR-AO. On the 90 user placements of Subsection IV-B.2 under indoor LoS-biased Rician fading with $\kappa = 10$ dB on both links, the median sum-rate of NIB is 3.5% below SDR-AO, in line with the greedy-vs-SDR gap reported for element-wise IRS schemes at moderate N [7], [9], [41]. The UG-induced contribution to that gap is below 1%, giving an overall NIB+ 3-UG vs SDR-AO gap of approximately 4%.

Both SDR-AO+UG and NIB+UG inherit the ~38% pilot saving of Subsection III-B.10; SDR-AO+UG attains the SDR-AO sum-rate at ~2.3 s per coherence interval, while NIB+UG accepts a ~4% sum-rate gap for a ~1 ms runtime.

13) VISUAL INFORMATION BEYOND USER LOCALIZATION

The framework so far uses visual information for localization and grouping. The same camera feed can supply several additional pieces of information to the beamforming pipeline. None of these is needed for the schemes above to function, but each can further reduce overhead or improve robustness: (i) AoA/AoD priors: since the AP, IRS, and ceiling camera are at known coordinates, the bounding-box centroid directly gives the geometric AoA/AoD for the LoS component, which can be used to initialize the channel estimator and shorten the pilot sequence; (ii) blockage prediction: consecutive frames yield user velocity and trajectory, which the grouping logic can use to assign a user to Group C ahead of time when it is about to cross a zone boundary, avoiding zone-thrashing; (iii) human/object discrimination: multi-class detection (in place of the single-class detector used here) lets the system distinguish communicating UEs from non-communicating obstacles such as humans crossing the LoS path, which the radio layer cannot recover; (iv) bounding-box size and aspect-ratio cues: changes in the bounding-box geometry correlate with user orientation and can serve as a low-cost prior for hybrid analog/digital beamforming. Items (iii) and (iv) are not obtainable from RF measurements alone; (i) and (ii) are obtainable from RF only at substantial extra cost in pilots and feedback bandwidth.

IV. SIMULATION RESULTS

Subsection A reports the single-user results and Subsection B reports the multiuser results. Both subsections share the simulation environment and benchmark configuration described in Subsection IV-A.1.

A. SINGLE-USER RESULTS

1) SIMULATION SETUP

The simulation setup is depicted in Fig. 9, where the AP-IRS distance is 50 m, the total transmit power p at the AP is 36 dBm, and the noise power σ^2 at the user is −96 dBm [10]. All runtimes reported throughout the paper, the YOLOv2 forward pass, the distance and grouping computation, and the AO, SDR, ASB, and SDR-AO benchmarks, are measured on the same machine (AMD Ryzen R7-5800H CPU @ 3.20 GHz, 16 GB RAM, MATLAB R2022a) in CPU-only mode; no GPU acceleration is used for any component. This places the detector and the optimization solver on the same hardware budget and ensures that the runtime gap reported between the proposed schemes and the benchmarks is not an artefact of mixed CPU/GPU execution.

2) SNR PERFORMANCE

For $N_t = 8$ and $N = 100$, Fig. 16 illustrates the real-time evaluation of SNR performance for 50 images, where each image represents a geographically random location of the user (i.e., x-axis of Fig. 16). In most cases, the SNR performance is nearly identical to AO, SDR, and ASB. However, at a few locations where the detector could not detect the object, the proposed scheme shows negligible performance degradation compared to the other schemes.

3) COMPUTATIONAL COST

Fig. 17 plots the computational cost (i.e., the average CPU running time) versus the number of reflecting elements. The proposed scheme's computational cost is almost independent of the size of the IRS. Specifically, for $N = 200$, 900, and 3000, the proposed scheme achieves 64%, 91%, and 99% lower computational cost, respectively, compared to AO; similar reductions are observed against ASB. The computational complexity of SDR is much higher (several seconds even for small N) and is therefore omitted from Fig. 17.

B. MULTIUSER RESULTS

1) SIMULATION SETUP

We consider eight users in the IRS-aided multiuser scenario. All other simulation parameters are the same as those listed in Subsection IV-A.1.

2) PERFORMANCE

For $N_t = 16$ and $N = 50$, Fig. 18 shows the instantaneous sum-rate performance over 90 images, each corresponding to a geographically random placement of the users; the corresponding CDF is shown in Fig. 19. The proposed vision-aided UG methods track the benchmark scheme closely, with 3-UG exhibiting a slightly larger gap than 2-UG because of its additional simplifications (cheaper composite-channel computations for users in Groups A and B). The gap remains within a few percent of the benchmark across the entire distribution.

3) COMPUTATIONAL COST

Figs. 20 and 21 show the CPU running time in instantaneous and CDF form, respectively. The runtime budget includes the neural-network detection, distance computation, and user-grouping stages. The CDF in Fig. 21 shows a runtime reduction of approximately 65–68% for 2-UG and a consistent 75% for 3-UG across the full distribution, with the figure annotating these values at the 10-th and 90-th percentile points.

3-UG is slightly faster than 2-UG because it performs fewer composite-channel computations (Subsection III-B.9); the corresponding small additional sum-rate gap was discussed in Subsection IV-B.2.

V. DISCUSSION

Before concluding the paper, we highlight several important points related to the implementation and potential extensions of the proposed framework:

1) **Better user detection methods:** There may be more efficient ways to implement the proposed method. For instance, in the user detection step, using a lightweight, custom neural network designed for this specific task could improve accuracy while reducing the overall computation time. More generally, the framework is detector-agnostic (Subsection III-A.1): the downstream pipeline only consumes the bounding-box centroid, so YOLOv2 can be replaced by any modern object detector (YOLOv5/v8/v11, DETR, RT-DETR, MobileNet-SSD, EfficientDet-Lite, or a custom task-specific network) without altering the downstream pipeline, the calibration procedure, the pilot-pattern derivation, or the complexity savings of Section III.
2) **Alternative distance estimation:** In this work, we used basic geometric and trigonometric distance calculations. However, other approaches, such as angle-based estimation, can be considered. Since the positions

of the IRS, AP, and camera are fixed, some angles are known and can be combined with visual cues to estimate distances more accurately or efficiently, offering an alternative to the method discussed in Subsection III-A.2.

3) **Use with other optimization techniques:** To show the effectiveness of our proposed UG methods, 2-UG and 3-UG, we combined them with the SDR-based approach (Algorithm 1). However, these UG strategies can also be used with other optimization algorithms, such as gradient descent, manifold optimization, deep learning models, or semi-closed-form solutions. Adding the UG stage beforehand can reduce complexity while still maintaining good performance. The NIB variant specified in Subsection III-B.12 (Algorithm 4) is one such instance: NIB+3-UG operates in the millisecond range (Table 5) and meets $T_{\text{e2e}} \leq T_c$ in every deployment row of Table 6. The same composition extends to gradient-descent, manifold-optimization, and deep-unfolding solvers without any change to the vision pipeline, the calibration procedure, or the pilot-pattern derivation.
4) **Use of visual data beyond localization:** Subsection III-B.13 catalogs several complementary roles for the same camera feed, AoA/AoD priors, blockage prediction, multi-class user/obstacle discrimination, and orientation cues from bounding-box geometry, which can be used to further reduce pilot overhead and to suppress zone-thrashing during transitions. None of these is needed for the schemes presented above, but each is a low-cost extension once the camera infrastructure is in place.
5) **Applicability to advanced IRS types:** This paper focuses on a basic IRS model with half-space reflection. However, the proposed framework can be extended to support advanced architectures such as STAR-IRS, omni-directional IRS, or beyond-diagonal IRS setups.

Regardless of the specific choices in implementation, the proposed approach shows strong potential for real-world applications.

VI. CONCLUSION

IRSs are promising enablers for future communication networks. However, due to the large number of reflecting elements, the computational cost of optimizing all elements becomes prohibitively high. In this work, we propose a vision-aided approach to enable low-complexity beamforming in IRS-assisted systems. Specifically, we identify two suboptimal but computationally efficient solutions and use visual data to estimate the user's location relative to the AP and IRS. Based on these distances, we select the most suitable low-complexity solution.

Despite a minor and negligible performance degradation, the proposed scheme significantly reduces computational cost compared to state-of-the-art methods, such as AO, ASB, and SDR, especially when the number of reflecting elements is large (e.g., $N > 100$). For example, when $N = 200$, 900, and 3000, our method achieves a 64%, 91%, and 99% reduction in computational cost, respectively, compared to AO.

We further extend our scheme to multiuser scenarios and propose two user grouping methods: 3-UG and 2-UG. Specifically, we use visual information to detect users in the region, retrieve their pixel locations, and compute their distances to the IRS and AP. Based on these distances, we group the users by their proximity: three groups in the 3-UG method and two groups in the 2-UG method. We then perform IRS optimization only for Group B and Group C in 3-UG, and only for Group B in 2-UG. This grouping strategy leads to substantial computational savings with negligible performance degradation.

For example, when $N_t = 16$ and $N = 50$, the vision-aided 3-UG and 2-UG methods reduce the per run runtime of the SDR-AO step by approximately 75% and 65–68%, respectively, with the vision-pipeline cost spread over the slower zone-boundary-crossing timescale and a multipath-robust grouping criterion based on average path geometry. The grouping additionally trims the CSI/pilot overhead by ~38% for a representative user-distribution split. The zone boundaries follow the closed form in (25); the propagation summary of Table 1 and the geometry sweep of Table 3 show that the sum-rate gap stays below 3% across deployment variations with $\eta_{\text{th}} = 10\,\text{dB}$ held fixed, with the gap growing as the propagation environment moves from LoS-biased indoor toward heavy-NLoS or outdoor regimes. Pairing UG with a non-iterative IRS beamformer (NIB, Algorithm 4) yields NIB+ 3-UG with a ~1 ms per-coherence-interval runtime and a ~4% sum-rate gap to SDR-AO. The bounding-box centroid is the only quantity consumed by the downstream pipeline, so YOLOv2 can be replaced by any modern object detector without altering the pipeline, the calibration procedure, or the complexity savings.

The proposed vision-based framework is readily extendable to advanced metasurfaces such as STAR-IRSs, omni-directional IRSs, beyond-diagonal IRSs, and active IRSs. The visual feed can also be exploited to predict channel behavior and blockage patterns, offering promising directions for future research.

ACKNOWLEDGMENT

The authors used a large language model as a grammar and phrasing aid during the preparation of this manuscript. All technical content, methods, simulations, and conclusions are solely the their work. The simulation code is available from the corresponding author upon reasonable request. An earlier version of this paper was presented at the 2024 European Conference on Networks and Communications (EuCNC) [DOI: 10.1109/EuCNC/6GSummit60053.2024.10597126].

REFERENCES

[1] M. Munawar, M. Guenach, and I. Moerman, "Machine vision aided adaptive beamforming decision for IRS-assisted wireless networks," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, Jun. 2024, pp. 517–522.

[2] B. Zheng, C. You, W. Mei, and R. Zhang, "A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1035–1071, 2nd Quart., 2022.
[3] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, Sep. 2020.
[4] J. Y. Dai et al., "Wireless communication based on information metasurfaces," *IEEE Trans. Microw. Theory Techn.*, vol. 69, no. 3, pp. 1493–1510, Mar. 2021.
[5] S. Sun, W. Jiang, S. Gong, and T. Hong, "Reconfigurable linear-to-linear polarization conversion metasurface based on PIN diodes," *IEEE Antennas Wireless Propag. Lett.*, vol. 17, no. 9, pp. 1722–1726, Sep. 2018.
[6] S. Gong et al., "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2283–2314, 4th Quart., 2020.
[7] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838–1851, Mar. 2020.
[8] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.
[9] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.
[10] M. Munawar and K. Lee, "Low-complexity adaptive selection beamforming for IRS-assisted single-user wireless networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 5458–5462, Apr. 2023.
[11] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming*. [Online]. Available: http://cvxr.com/cvx
[12] T. Nishio, Y. Koda, J. Park, M. Bennis, and K. Doppler, "When wireless communications meet computer vision in beyond 5G," *IEEE Commun. Standards Mag.*, vol. 5, no. 2, pp. 76–83, Jun. 2021.
[13] H. H. Zhang and R. S. Chen, "Coherent processing and superresolution technique of multi-band radar data based on fast sparse Bayesian learning algorithm," *IEEE Trans. Antennas Propag.*, vol. 62, no. 12, pp. 6217–6227, Dec. 2014.
[14] Y. Tian, G. Pan, and M.-S. Alouini, "Applying deep-learning-based computer vision to wireless communications: Methodologies, opportunities, and challenges," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 132–143, 2021.
[15] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, May 2021.
[16] Ö. Özdogan, E. Björnson, and E. G. Larsson, "Using intelligent reflecting surfaces for rank improvement in MIMO communications," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 9160–9164.
[17] D. Li, "Fairness-aware multiuser scheduling for finite-resolution intelligent reflecting surface-assisted communication," *IEEE Commun. Lett.*, vol. 25, no. 7, pp. 2395–2399, Jul. 2021.
[18] M. Munawar and K. Lee, "Dual-polarized IRS-assisted wireless network: Relative phase modulation," *EURASIP J. Wireless Commun. Netw.*, vol. 2024, no. 1, p. 22, May 2024.
[19] M. Munawar and K. Lee, "Dual-polarized IRS-assisted MIMO network," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 2519–2532, Apr. 2024.
[20] M. Munawar and K. Lee, "Power optimization of dual-polarized IRS-assisted wireless networks under spectral efficiency constraints," *IEEE Trans. Wireless Commun.*, vol. 24, no. 1, pp. 430–446, Jan. 2025.
[21] M. Hua, Q. Wu, W. Chen, O. A. Dobre, and A. Lee Swindlehurst, "Secure intelligent reflecting surface aided integrated sensing and communication," *IEEE Trans. Wireless Commun.*, vol. 23, no. 1, pp. 575–591, Jan. 2023.
[22] Z.-Q. Luo, W.-K. Ma, A. M. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.
[23] J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan, "A brief introduction to manifold optimization," *J. Oper. Res. Soc. China*, vol. 8, no. 2, pp. 199–248, Jun. 2020.
[24] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.
[25] M. Kojima and L. Tunçel, "Cones of matrices and successive convex relaxations of nonconvex sets," *SIAM J. Optim.*, vol. 10, no. 3, pp. 750–778, Jan. 2000.
[26] J. Bland, "Discrete-variable optimal design using a Tabu search approach," *WIT Trans. Built Environ.*, vol. 14, pp. 415–423, Jan. 1995.
[27] E. Zafra, S. Vazquez, A. M. Alcaide, L. G. Franquelo, J. I. Leon, and E. P. Martin, "K-best sphere decoding algorithm for long prediction horizon FCS-MPC," *IEEE Trans. Ind. Electron.*, vol. 69, no. 8, pp. 7571–7581, Aug. 2022.
[28] X. Chen et al., "Design and implementation of MIMO transmission based on dual-polarized reconfigurable intelligent surface," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2155–2159, Oct. 2021.
[29] R. Fara, P. Ratajczak, D.-T. Phan-Huy, A. Ourir, M. Di Renzo, and J. De Rosny, "A prototype of reconfigurable intelligent surface with continuous control of the reflection phase," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 70–77, Feb. 2022.
[30] Westman Security.(2024). *Westman Security—Fisheye Camera*. Accessed: Feb. 1, 2024. [Online]. Available: https://www.westmansecurity.ca/fisheye
[31] 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.
[32] X. Ma et al., "Joint channel estimation and data rate maximization for intelligent reflecting surface assisted terahertz MIMO communication systems," *IEEE Access*, vol. 8, pp. 99565–99581, 2020.
[33] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, Oct. 2020.
[34] M. Munawar, M. Guenach, and I. Moerman, "Dual-polarized IRS-enhanced generalized multiuser multi-stream MIMO," *IEEE Trans. Wireless Commun.*, vol. 24, no. 10, pp. 8615–8632, Oct. 2025.
[35] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[36] MathWorks, Inc. (2024). *Computer Vision Toolbox: Object Detection Using YOLO V2 Deep Learning*. Accessed: Apr. 2026. [Online]. Available: https://www.mathworks.com/help/vision/ug/train-an-object-detector-using-you-only-look-once.html
[37] J. Sang et al., "An improved YOLOv2 for vehicle detection," *Sensors*, vol. 18, no. 12, p. 4272, Dec. 2018.
[38] Z. Zhang et al., "Active RIS vs. passive RIS: Which will prevail in 6G?," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1707–1725, Mar. 2023.
[39] 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.
[40] *Study on Channel Model for Frequencies From 0.5 to 100 GHz*, document TR 38.901, 3GPP, 2020.
[41] T. Jiang, H. V. Cheng, and W. Yu, "Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1931–1945, Jul. 2021.
[42] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

MUTEEN MUNAWAR (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering in 2019 and 2021, respectively, and the second master's degree in electrical and information engineering from Seoul National University of Science and Technology, South Korea. He is currently pursuing the Ph.D. degree with UGent, Belgium. He was with the Communication and Signal Processing Laboratory, South Korea, from February 2023 to September 2023. Currently, he is working at IMEC, Leuven, as a Researcher. His research interests include wireless communications, applied machine learning, and digital multipliers. He received the Best Project Award at the IEEE Riphah Innovation Summit (IRIS) in 2019.

MAMOUN GUENACH (Senior Member, IEEE) received the Engineering degree from the École Mohammadia d'Ingénieurs, Morocco, in 1995, and the Ph.D. degree from UCLouvain, Belgium, in 2002. He was a Post-Doctoral Researcher at Ghent University from 2002 to 2006, where he has been a Visiting Research Professor since 2015. From 2006 to 2019, he was with Nokia Bell Laboratories. Since 2019, he has been a Research Scientist at IMEC, Leuven, Belgium. His research interests include coding, modulation, synchronization, and equalization for high-speed wired and wireless communications, including distributed massive MIMO.

INGRID MOERMAN (Senior Member, IEEE) received the degree in electrical engineering and the Ph.D. degree from Ghent University in 1987 and 1992, respectively. She became a part-time Professor at Ghent University in 2000. She is currently a Staff Member at IDLab, a core research group of IMEC, with research activities embedded at Ghent University and the University of Antwerp. She coordinates the research activities on intelligent wireless networking (iWiNe) at Ghent University, where she leads a team of more than 30 researchers. She is also the Program Manager of the Deterministic Networking Track, part of the Connectivity Program, IMEC. In this role, she coordinates research on end-to-end wired/wireless networking solutions for business and mission-critical applications that must meet strict quality of service requirements in terms of throughput, bounded latency, and reliability in private wireless environments. She has extensive experience in running and coordinating national and EU research-funded projects.