Stacked Intelligent Metasurfaces for Holographic MIMO Aided Cell-Free Networks

Qingchao Li, *Graduate Student Member, IEEE*, Mohammed El-Hajjar, *Senior Member, IEEE*,
Chao Xu, *Senior Member, IEEE*, Jiancheng An, *Member, IEEE*, Chau Yuen, *Fellow, IEEE*,
and Lajos Hanzo, *Life Fellow, IEEE*

Abstract—Large-scale multiple-input and multiple-output (MIMO) systems are capable of achieving high date rate. However, given the high hardware cost and excessive power consumption of massive MIMO systems, as a remedy, intelligent metasurfaces have been designed for efficient holographic MIMO (HMIMO) systems. In this paper, we propose a HMIMO architecture based on stacked intelligent metasurfaces (SIM) for the uplink of cell-free systems, where the SIM is employed at the access points (APs) for improving the spectral- and energy-efficiency. Specifically, we conceive distributed beamforming for SIM-assisted cell-free networks, where both the SIM coefficients and the local receiver combiner vectors of each AP are optimized based on the local channel state information (CSI) for the local detection of each user equipment (UE) information. Afterward, the central processing unit (CPU) fuses the local detections gleaned from all APs to detect the aggregate multi-user signal. Specifically, to design the SIM coefficients and the combining vectors of the APs, a low-complexity layer-by-layer iterative optimization algorithm is proposed for maximizing the equivalent gain of the channel spanning from the UEs to the APs. At the CPU, the weight vector used for combining the local detections from all APs is designed based on the minimum mean square error (MMSE) criterion, where the hardware impairments (HWIs) are also taken into consideration based on their statistics. The simulation results show that the SIM-based HMIMO outperforms the conventional single-layer HMIMO in terms of the achievable rate. We demonstrate that both the HWI of the radio frequency (RF) chains at the APs and the UEs limit the achievable rate in the high signal-to-noise-ratio (SNR) region.

Index Terms—Holographic multiple-input and multiple-output (HMIMO), stacked intelligent metasurface (SIM), cell-free network, hardware impairment (HWI).

I. INTRODUCTION

IN the fifth generation (5G) wireless systems, large-scale multiple-input and multiple-output (MIMO) systems have been harnessed for providing significantly increased throughput by employing a large number of antennas at the base station (BS) [1], [2], [3], [4], [5]. However, they require

Qingchao Li, Mohammed El-Hajjar, Chao Xu and Lajos Hanzo are with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: qingchao.li@soton.ac.uk; meh@ecs.soton.ac.uk; cx1g08@ecs.soton.ac.uk; lh@ecs.soton.ac.uk).
Jiancheng An and Chau Yuen are with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jiancheng.an@ntu.edu.sg; chau.yuen@ntu.edu.sg).

numerous active radio frequency (RF) chains, which results in excessive hardware cost and energy consumption. Hence, some authors have focused their attention on the design of low-cost and energy-efficient solutions [6], [7], [8], [9], [10]. As an attractive design alternative, holographic MIMO (HMIMO) solutions rely on an intelligent software reconfigurable paradigm in support of improved hardware efficiency and energy efficiency. They achieve this ambitious objective by utilizing a spatially near-continuous aperture and holographic radios having reduced power consumption and fabrication cost [11], [12], [13], [14], [15], [16]. The recent progress in channel modeling and efficient channel estimation conceived for HMIMO systems is reported in [17].

Currently, the hardware architectures of HMIMO are mainly based on reconfigurable refractive surfaces (RRS) [18], reconfigurable holographic surfaces (RHS) [19], [20], [21], [22], [23], [24], [25], [26], and dynamic metasurface antennas (DMA) [27], [28], [29], [30].

1) Reconfigurable refractive surfaces: It is infeasible to realize HMIMO schemes relying on a large number of conventional RF chains and active antennas due to the excessive power consumption [16]. As a remedy, Zeng *et al.* [18] employed a RRS illuminated by a single RF chain at the BS for creating an energy-efficient HMIMO. A substantial beamforming gain can be achieved by adjusting the coefficient of each RRS element. Both their theoretical analysis and simulation results show that the RRS-aided HMIMO has higher energy efficiency than the conventional MIMO systems using phased arrays.

2) Reconfigurable holographic surfaces: In [19], [20], [21], Deng *et al.* proposed a RHS based architecture. Their digital beamformer relying on the state-of-the-art (SoA) zero-forcing (ZF) transmit precoding method and a holographic beamformer are jointly optimized at the BS and the RHS, respectively. The simulation results showed that the RHS-based hybrid beamformer achieves a higher sum-rate than the SoA massive MIMO based hybrid beamformer relying on phase shift arrays. To reduce the configuration complexity, Hu *et al.* [22] proposed a holographic beamforming scheme based on amplitude-controlled RHS elements having limited resolution. They showed that the holographic beamformer associated with 2-bit quantized amplitude resolution achieves a similar sum-rate as that associated with continuous amplitude values. Furthermore, in [23] the RHS beamformer is employed for ultra-dense low-Earth-orbit (LEO) satellite communications to compensate for the severe path-loss of satellite communications. The simulation results showed that the RHS provides a

more cost-effective solution for pursuing higher data rate than the phased array architecture. The impact of the number of radiation elements on the sum-rate of RHS-based LEO satellite communications is further investigated in [24]. Explicitly, the authors theoretically analyzed the minimum number of RHS elements required for the sum-rate of the RHS-aided system to exceed that of the phased array system. Furthermore, Wei *et al.* [25] proposed a low-complexity ZF beamformer based on utilizing Neumann series expansion to replace the matrix inversion operation in multi-user RHS-based MIMO communications. To reduce the channel state information (CSI) estimation overhead, Wu *et al.* [26] minimized the transmit power of the RHS-based holographic MIMO based on a two-time scale beamformer. Specifically, this holographic beamformer was designed based on the statistical CSI, and then the instantaneous CSI of the equivalent channel links was estimated and utilized for designing the digital transmit precoding matrix.

3) Dynamic metasurface antennas: The dynamic metasurface antenna consists of multiple microstrips and each microstrip is composed of a multitude of sub-wavelength and frequency-selective resonant metamaterial radiating elements, which can be employed to realize low-cost, power-efficient and size compact antenna arrays [27], [28], [29], [30]. In the DMA, the information is beamformed by linearly combining the radiation observation from all metamaterial elements in each microstrip. The mathematical model of DMA-based massive MIMO systems were firstly proposed by Shlezinger *et al.* in [27], where the fundamental limits of DMA-aided uplink communications was also investigated. To approach these limits, the alternating optimization algorithm was employed for designing practical DMAs for arbitrary multipath channels and frequency selectivity profiles. Furthermore, the achievable sum-rate of DMA-based downlink massive MIMO systems was characterized in [28], and an efficient alternating algorithm was proposed for dynamically configuring the DMA weights to maximize the achievable sum-rate. It was shown that the fundamental limits of DMA-based massive MIMO systems are comparable to those of conventional MIMO systems based on ideal antenna arrays. In [29], You *et al.* employed DMA to realize large-scale antenna arrays having reduced physical size, hardware cost, and power consumption. Specifically, the energy efficiency of the DMA-based massive MIMO system was maximized by the Dinkelbach transform, alternating optimization, and deterministic equivalent methods. The simulation results showed that higher energy efficiency can be achieved by the DMA-based massive MIMO architecture than by the conventional fully-digital and hybrid massive MIMO systems. Furthermore, Li *et al.* [30] proposed the power-efficient DMA, which operate at high frequencies and realize extremely large-scale MIMO (XL-MIMO) schemes. The DMA-based XL-MIMO architecture is composed of the conventional digital beamformer and the DMA-based holographic beamformer. Specifically, in the holographic beamformer, the DMAs can be configured based on three different modes, including continuous-amplitude configurations, binary-amplitude configurations and Lorentzian-constrained phase configurations [30]. An efficient

successive convex approximation based alternating direction method of multipliers (ADMM) aided algorithm was proposed for optimizing the digital beamformer and the DMA-based holographic beamformer in an alternating manner. The associated simulation results showed that the DMA-based array has lower hardware overhead and power consumption than the conventional hybrid massive MIMO beamformer.

The above holographic MIMO architectures are based on a single-layer metasurface. To further improve both the spatial-domain gain and the beamformer's degree-of-freedom, An *et al.* [31] proposed a HMIMO system based on stacked intelligent metasurfaces (SIM), which is composed of stacked multi-layer reconfigurable surfaces to carry out advanced signal processing directly in the native electromagnetic (EM) wave regime without digital beamformer. The gradient descent algorithm is employed for optimizing the phase shifts of the elements in all layers of the metasurfaces to maximize the sum-rate. The simulation results show that the SIM architecture outperforms its single-layer metasurface counterparts. The wave-based beamforming relying on the SIM is benefit of simplifying hardware architecture and improving computational efficiency. Furthermore, an SIM has the application of performing interference cancellation for multiple access and enabling integrated sensing and communications [32].

However, the above contributions assume idealized perfect RF hardware both at the BSs and the UEs, which is impractical. Furthermore, the above HMIMO architectures are tailored for cellular networks, where the cell-edge users (UEs) suffer from low data-rate both due to the increased BS-UE distance and owing to the inter-cell interference. As a design alternative, the cell-free concept mitigates the signal path loss and the inter-cell interference by deploying a set of distributed access points (APs) for cooperatively serving UEs without cell boundaries [33]. Specifically, the ZF method and the minimum mean square error (MMSE) method can be employed for designing the beamformers in the centralized algorithm. To reduce the required overhead of CSI-sharing between APs, a distributed algorithm can be employed based on the maximum ratio transmission (MRT) or the maximum ratio combining (MRC) criteria, albeit while at the cost of a performance degradation. Furthermore, in the distributed optimization algorithm of the cell-free system, the cooperation between APs is promising in terms of harnessing parallel computing resources and achieving almost the same data rate as the centralized algorithm [34]. To deal with these challenges, in this paper we propose an SIM assisted HMIMO architecture for cell-free networks, while directly considering the signal distortion resulting from the realistic hardware impairments (HWIs) of the RF chains both at the APs and the UEs. Against this background, our contributions are detailed as follows, while Table I explicitly contrasts them to the literature at a glance.

- We conceive an SIM-based HMIMO architecture for the uplink of a cell-free network, where the SIM is employed at the APs to attain high spectral- and energy-efficient information transfer. The distributed operation is employed for the SIM-based cell-free network. Specifically, at each AP the hybrid beamformer coefficients and the receiver combining (RC) vectors are jointly optimized to acquire a

TABLE I
CONTRASTING THE NOVELTY OF OUR PAPER TO THE EXISTING MIMO
TECHNIQUES IN THE LITERATURE [18], [19], [20], [21], [22], [23], [24],
[25], [26], [27], [28], [29], [30], [31], [32], [33], [34].

	Metasurface technique	Multi-user system	Cell-free network	Transceiver hardware impairments
Our paper	SIM	✔	✔	✔
[18]	RRS			
[19]–[24]	RHS	✔		
[25]	RHS	✔		
[26]	RHS	✔		
[27]	DMA	✔		
[28]	DMA	✔		
[29]	DMA	✔		
[30]	DMA	✔		
[31], [32]	SIM	✔		
[33]	Conventional antennas	✔	✔	
[34]	Conventional antennas	✔	✔	

Fig. 1. System model of the SIM-aided holographic MIMO in cell-free networks.

local estimate of the information arriving from the UEs. Afterwards, the central processing unit (CPU) uses the data detected by all APs to carry out the final detection of each UE's data.

- Since designing the hybrid beamformer coefficient of the SIM coefficients and the RC vectors at each AP is a non-convex problem, we propose a low-complexity layer-by-layer iterative optimization algorithm. Specifically, when the RC vectors are given, the coefficients of the intelligent metasurface are optimized on a layer-by-layer basis. By contrast, when the coefficients of all layers of the intelligent metasurface are given, the RC vectors are designed based on the MRC criterion. The SIM coefficients and the RC vectors of each AP are alternately optimized until reaching convergence.

- For recovering the information gleaned from the UEs, the weight vector of the CPU used for combining the local detections arriving from all APs is designed based on the MMSE criterion harnessed for maximizing the signal-to-noise-ratio (SNR) of the received signal. Furthermore, since having RF hardware impairments at APs and UEs is inevitable, we take them into account in the weight vector design by exploiting their statistics.

- Our numerical results show that the average achievable rate of our SIM-based HMIMO architecture in the cell-free network outperforms the conventional single-layer intelligent surface aided HMIMO. Furthermore, owing to the HWIs at APs and UEs, the achievable rate saturates at high SNRs without reaching its theoretical maximum.

The rest of this paper is organized as follows. In Section II, we present the system model, while the beamformer design is described in Section III. Our simulation results are presented in Section IV, while we conclude in Section V.

Notations: Vectors and matrices are denoted by boldface lower and upper case letters, respectively, $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\dagger}$ and $(\cdot)^{\mathrm{H}}$ represent the operation of transpose, conjugate and Hermitian transpose, respectively, $\odot$ represents the Hadamard product operation, $|a|$ and $\angle a$ denote the amplitude and angle of the complex scalar a, respectively, $\|\mathbf{a}\|$ denotes the norm of the
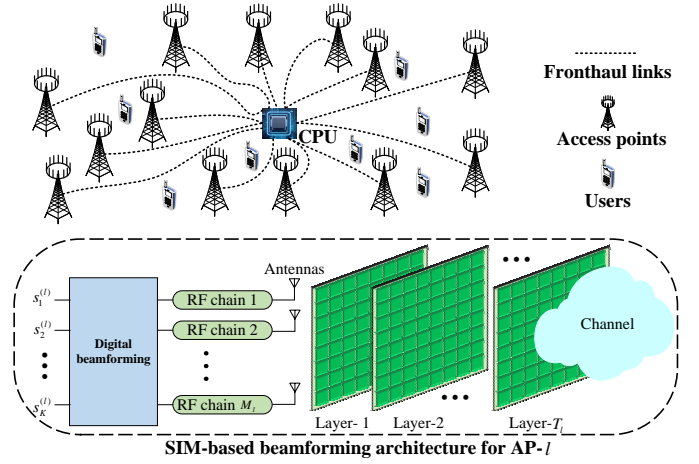
vector $\mathbf{a}$, $\mathbb{C}^{m \times n}$ denotes the space of $m \times n$ complex-valued matrices, $\mathbf{0}_N$ is the $N \times 1$ zero vector, $\mathbf{I}_N$ represents the $N \times N$ identity matrix, $\mathbf{Diag}\{\mathbf{a}\}$ denotes a diagonal matrix having elements of $\mathbf{a}$ in order, $[\mathbf{a}]_n$ is the nth element in the vector $\mathbf{a}$, $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a circularly symmetric complex Gaussian random vector with the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

II. SYSTEM MODEL

In this section, we describe our proposed SIM-aided HMIMO architecture, and then present the channel model of the cell-free network considered.

The system model of the SIM-aided HMIMO operating in narrowband cell-free networks is shown in Fig. 1[1]. In contrast to the conventional cellular network, where a central AP is deployed in each cell to support the surrounding UEs, the cell-free network deploys multiple distributed APs. We focus our attention on the uplink scenario, where we consider L distributed multi-antenna APs and K single-antenna UEs. AP-l ($l = 1, 2, \cdots, L$) has M antennas and a stacked intelligent metasurface containing T_l layers, with each layer composed of N reconfigurable passive elements. Specifically, the information is sent from the K UEs to L APs. In each AP, the received RF signals go through the passive beamformer of the multi-layer SIM, and then they are converted to baseband signals via the RF chains. To recover the desired information, we rely on distributed operation in order to reduce the detection complexity, where each AP locally detects the data of the UEs supported by it, and then the CPU fuses the data detected at all APs via the fronthaul links to carry out the final detection of the desired information.

A. SIM-Aided Holographic MIMO Architecture

Before describing the SIM-aided holographic MIMO architecture, we briefly review the conventional fully-digital

---

[1]In this paper, we investigate the SIM-based hybrid beamforming design of narrowband cell-free networks. In wideband networks, both spatial-wideband effects and frequency-selective effects should be considered [35]. The SIM-based hybrid beamforming design of wideband networks considering the spatial-wideband effect and the frequency-selective effect is set aside for our future work.

massive MIMO architecture and the hybrid digital-analog hybrid massive MIMO architecture. In the conventional fully-digital massive MIMO architecture, the number of the RF chains is the same as that of the transmit antennas. To reduce the power consumption of the RF chains, a hybrid digital-analog beamforming architecture is formed, where only a few RF chains is employed, and a phase shift array (PSA) is harnessed between the RF chains and AP antennas.

The SIM-aided HMIMO architecture of cell-free networks is shown in Fig. 1. Firstly, the signals received at each AP go through an SIM-based beamformer. The output signals of the SIM are received by the antennas, and then converted to baseband signals via multiple RF chains. The SIM is constituted by a closed vacuum container having several layers of stacked reconfigurable metasurfaces, each of which is composed of a large number of passive densely-spaced elements [31], [32]. By appropriately configuring the phase shifts of the elements in each layer of the metasurface with the aid of a software controllers, such as a field programmable gate array (FPGA) [31], [32], a substantial beamforming gain may be achieved. In practice, the SIM is enclosed in a supporting structure surrounded by wave-absorbing materials, to prevent interferences from undesired diffraction, scattering, and environmental noise [36], [37]. The AP antennas are $M_x \times M_y$ uniform rectangular planar arrays (URPA), while the intelligent surfaces in each layer are $N_x \times N_y$ URPAs, satisfying $M = M_x M_y$ and $N = N_x N_y$. We denote the size of each reconfigurable element as $\delta_x \times \delta_y$, and the distances between the adjacent antennas are d_x and d_y along the x and y axis, respectively.

Explicitly, we contrast our SIM-aided HMIMO to various MIMO technologies in Table II.

B. Channel Model

In this section, we describe the channel between the UEs and the AP antennas. As shown in Fig. 1, we have to consider the channel between the SIM and the antennas at the AP as well as the channel between the UEs and the SIM at the AP.

1) Channel links between the SIM and the antennas at APs: We assume that the antenna array and the SIM layers are parallel to the xoy plane of the Cartesian coordinate. We denote the Cartesian coordinates of the M antennas at the lth AP as $\mathbf{p}_1^{(l,0)} = (x_1^{(l,0)}, y_1^{(l,0)}, z^{(l,0)})$, $\mathbf{p}_2^{(l,0)} = (x_2^{(l,0)}, y_2^{(l,0)}, z^{(l,0)}), \cdots, \mathbf{p}_M^{(l,0)} = (x_M^{(l,0)}, y_M^{(l,0)}, z^{(l,0)})$, and that of the N elements in the SIM layer-t as $\mathbf{p}_1^{(l,t)} = (x_1^{(l,t)}, y_1^{(l,t)}, z^{(l,t)})$, $\mathbf{p}_2^{(l,t)} = (x_2^{(l,t)}, y_2^{(l,t)}, z^{(l,t)}), \cdots, \mathbf{p}_N^{(l,t)} = (x_N^{(l,t)}, y_N^{(l,t)}, z^{(l,t)})$ with $t = 1, 2, \cdots, T_l$. We employ the near-field model to describe the channel response between the antennas and the SIM layers.

As shown in Fig. 1, we denote the channel response between AP-l of the SIM layer-1 and the antennas as $\mathbf{A}^{(l,1)} \in \mathbb{C}^{M \times N}$, with the (m, n)th entry $a_{m,n}^{(l,1)}$ representing the response of the nth element in the SIM layer-1 to the mth antenna, while $a_{m,n}^{(l,1)}$ is given by

$$a_{m,n}^{(l,1)} = \sqrt{\Gamma_{m,n}^{(l,1)}} e^{-J \frac{2\pi}{\lambda} \left\| \mathbf{p}_n^{(l,1)} - \mathbf{p}_m^{(l,0)} \right\|}. \tag{1}$$

In (1), λ is the carrier wavelength, $\left\| \mathbf{p}_n^{(l,1)} - \mathbf{p}_m^{(l,0)} \right\|$ is the distance between the nth element in the SIM layer-1 and the mth antenna, while $\Gamma_{m,n}^{(l,1)}$ is the power radiated from the nth element in the SIM layer-1 to the mth antenna, which can be represented as [38]

$$\Gamma_{m,n}^{(l,1)} = \iint_{\mathcal{D}_n} \frac{\kappa \left(\frac{z^{(l,1)} - z^{(l,0)}}{\left\| \mathbf{p}_n^{(l,1)} - \mathbf{p}_m^{(l,0)} \right\|} \right)^{\frac{\kappa}{2}}}{4\pi \left\| \mathbf{p}_n^{(l,1)} - \mathbf{p}_m^{(l,0)} \right\|^2} \mathrm{d}x\mathrm{d}y, \tag{2}$$

where κ is the directional radiation gain of the reconfigurable elements and the integration region is formulated as:

$$\mathcal{D}_n = \left\{ (x, y) \in \mathbb{R}^2 : x_n - \frac{\delta_x}{2} \leq x \leq x_n + \frac{\delta_x}{2}, \right.$$
$$\left. y_n - \frac{\delta_y}{2} \leq y \leq y_n + \frac{\delta_y}{2} \right\}. \tag{3}$$

Next, we focus our attention on the channel model between the SIM layers. For AP-l we denote the response of the channel spanning from the SIM layer-t to layer-$(t-1)$ as $\mathbf{A}^{(l,t)} \in \mathbb{C}^{N \times N}$ $(t = 2, 3, \cdots, T_l)$ with the (n_2, n_1)th entry representing the response from the n_1th element in the SIM layer-t to the n_2th element in the SIM layer-$(t-1)$, given by

$$a_{n_2,n_1}^{(l,t)} = \sqrt{\Gamma_{n_2,n_1}^{(l,t)}} e^{-J \frac{2\pi}{\lambda} \left\| \mathbf{p}_{n_1}^{(l,t)} - \mathbf{p}_{n_2}^{(l,t-1)} \right\|}. \tag{4}$$

In (4), $\left\| \mathbf{p}_{n_1}^{(l,t)} - \mathbf{p}_{n_2}^{(l,t-1)} \right\|$ is the distance between the n_1th element in the SIM layer-t and the n_2th element in the SIM layer-$(t-1)$, while $\Gamma_{n_2,n_1}^{(l,t)}$ is the power radiated from the n_1th element in the SIM layer-t to the n_2th element in the SIM layer-$(t-1)$, which can be represented as [18]

$$\Gamma_{n_2,n_1}^{(l,t)} = \iint_{\mathcal{D}_n} \frac{\kappa \left(\frac{z^{(l,t)} - z^{(l,t-1)}}{\left\| \mathbf{p}_{n_1}^{(l,t)} - \mathbf{p}_{n_2}^{(l,t-1)} \right\|} \right)^{\frac{\kappa}{2}}}{4\pi \left\| \mathbf{p}_{n_1}^{(l,t)} - \mathbf{p}_{n_2}^{(l,t-1)} \right\|^2} \mathrm{d}x\mathrm{d}y. \tag{5}$$

Furthermore, the phase shifts of the reconfigurable elements in each layer can be adjusted to achieve a beamforming gain. In AP-l, we denote the phase shift of the nth element in layer-t as $\theta_n^{(l,t)}$. Upon considering the signal power attenuation resulting from the EM waves travelling through a SIM, we denote the radiation coefficient of the nth element in layer-t as $v_n^{(l,t)}$ satisfying $v_n^{(l,t)} \in [0, 1]$. Thus, by defining $\mathbf{\Upsilon}^{(l,t)} = \mathbf{Diag}\{v_1^{(l,t)}, v_2^{(l,t)}, \cdots, v_N^{(l,t)}\}$ and $\mathbf{\Theta}^{(l,t)} = \mathbf{Diag}\{e^{J\theta_1^{(l,t)}}, e^{J\theta_2^{(l,t)}}, \cdots, e^{J\theta_N^{(l,t)}}\}$, the response matrix of the SIM's layer-t can be represented as

$$\mathbf{\Xi}^{(l,t)} = \sqrt{\mathbf{\Upsilon}^{(l,t)}} \mathbf{\Theta}^{(l,t)}$$
$$= \mathbf{Diag}\left\{ \sqrt{v_1^{(l,t)}} e^{J\theta_1^{(l,t)}}, \sqrt{v_2^{(l,t)}} e^{J\theta_2^{(l,t)}}, \cdots, \right.$$
$$\left. \sqrt{v_N^{(l,t)}} e^{J\theta_N^{(l,t)}} \right\}. \tag{6}$$

Therefore, the equivalent channel response of the SIM-based

TABLE II
COMPARISON OF OUR SIM-AIDED HMIMO TO VARIOUS MIMO TECHNOLOGIES.

	Beamforming structure	Number of RF chains	Number of PSA (or metasurface) elements	Hardware cost	Energy efficiency
Fully digital MIMO	Fully digital beamformer	Large	No	High	Low
PSA-aided hybrid MIMO	Digital beamformer + Analog beamformer relying on PSA	Moderate	Large	Moderate	Moderate
SIM-aided HMIMO in [31], [32]	Wave-based beamformer relying on SIM	No	Very large	Low	High
Our SIM-aided HMIMO	Digital beamformer + Wave-based beamformer relying on SIM	Moderate	Large	Low	High

beamformer at the lth AP, denoted as $\mathbf{G}^{(l)} \in \mathbb{C}^{M \times N}$, is[2]

$$
\begin{aligned}
\mathbf{G}^{(l)} =& \mathbf{A}^{(l,1)} \mathbf{\Xi}^{(l,1)} \mathbf{A}^{(l,2)} \mathbf{\Xi}^{(l,2)} \cdots \mathbf{A}^{(l,T_l)} \mathbf{\Xi}^{(l,T_l)} \\
=& \mathbf{A}^{(l,1)} \sqrt{\mathbf{\Upsilon}^{(l,1)}} \mathbf{\Theta}^{(l,1)} \mathbf{A}^{(l,2)} \sqrt{\mathbf{\Upsilon}^{(l,2)}} \mathbf{\Theta}^{(l,2)} \cdots \mathbf{A}^{(l,T_l)} \\
& \sqrt{\mathbf{\Upsilon}^{(l,T_l)}} \mathbf{\Theta}^{(l,T_l)}.
\end{aligned}
\tag{7}
$$

2) Channel links between UEs and APs: We denote the large-scale fading and the small-scale fading between UE-k and the SIM at AP-l as $\varrho_k^{(l)}$ and $\mathbf{h}_k^{(l)} \in \mathbb{C}^{N \times 1}$, respectively. We assume the knowledge of $\mathbf{h}_1^{(l)}, \mathbf{h}_2^{(l)}, \cdots, \mathbf{h}_K^{(l)}$ can be attained at the AP-l. In practical metasurface-based holographic MIMO systems, this has to be acquired by CSI acquisition methods, such as the subspace-based channel estimator of [39] or the sparse channel estimator of [40]. We adopt the millimeter-wave (mmWave) channel model to characterize the propagation environment between the APs and each user[3]. Specifically, the mmWave uplink channel of UE-k is assumed to be the superposition of all propagation paths that are scattered in ζ_c clusters and each cluster contributes ζ_p paths, expressed as [41]

$$
\mathbf{h}_k^{(l)} = \sqrt{\frac{1}{\zeta_c \zeta_p}} \sum_{c=1}^{\zeta_c} \sum_{p=1}^{\zeta_p} \alpha_{c,p}^{(l,k)} \mathbf{f}(\psi_{c,p}^{(l,k)}, \varphi_{c,p}^{(l,k)}),
\tag{8}
$$

where $\alpha_{c,p}^{(l,k)}$ is the complex gain of the pth path in the cth cluster following $\alpha_{c,p}^{(l,k)} \sim \mathcal{CN}(0,1)$, and $\mathbf{f}(\psi_{c,p}^{(l,k)}, \varphi_{c,p}^{(l,k)})$ is

$$
\begin{aligned}
& \mathbf{f}\left(\psi_{c,p}^{(l,k)}, \varphi_{c,p}^{(l,k)}\right) = [1, \cdots, \\
& e^{j \frac{2\pi}{\lambda}\left(\delta_x n_x \sin \psi_{c,p}^{(l,k)} \cos \varphi_{c,p}^{(l,k)} + \delta_y n_y \sin \psi_{c,p}^{(l,k)} \sin \varphi_{c,p}^{(l,k)}\right)}, \cdots \\
& e^{j \frac{2\pi}{\lambda}\left(\delta_x (N_x-1) \sin \psi_{c,p}^{(l,k)} \cos \varphi_{c,p}^{(l,k)} + \delta_y (N_y-1) \sin \psi_{c,p}^{(l,k)} \sin \varphi_{c,p}^{(l,k)}\right)}]^{\mathrm{H}},
\end{aligned}
\tag{9}
$$

where $\psi_{c,p}^{(l,k)}$ and $\varphi_{c,p}^{(l,k)}$ are the elevation and azimuth angles of departure (AoD) from UE-k to the SIM in AP-l in the pth path of the cth cluster, respectively. Within the cth cluster, the random variables $\psi_{c,p}^{(l,k)}$ and $\varphi_{c,p}^{(l,k)}$ have uniformly distributed mean values of $\mu_{\psi_c^k}$ and $\mu_{\varphi_c^k}$, respectively, and have angular spreads, i.e. standard deviations, of $\sigma_{\psi_c^k}$ and $\sigma_{\varphi_c^k}$, respectively.

---

[2]In this paper, we assume that each SIM layer can be configured independently without mutual coupling. In reality, having mutual coupling among the metamaterial elements is inevitable due to their dense packaging. The holographic beamforming design considering the effect of mutual coupling is set aside for our future work.

[3]Note that here we consider a mmWave channel model as an example for describing the signal propagation between the APs and the UEs. But indeed, our proposed SIM-based beamforming architecture is also applicable to other channel models.

The signal received by the antennas at AP-l is given by

$$
\begin{aligned}
\mathbf{y}^{(l)} = \sum_{k=1}^{K} & \left(\sqrt{\rho_k \varepsilon_{u_k} \varepsilon_{\mathbf{v}^{(l)}}} \mathbf{q}_k^{(l)} s_k + \sqrt{\rho_k (1 - \varepsilon_{u_k}) \varepsilon_{\mathbf{v}^{(l)}}} \mathbf{q}_k^{(l)} u_k \right. \\
& \left. + \sqrt{\rho_k (1 - \varepsilon_{\mathbf{v}^{(l)}})} \mathbf{q}_k^{(l)} \odot \mathbf{v}_k^{(l)} \right) + \mathbf{w}^{(l)},
\end{aligned}
\tag{10}
$$

where $\mathbf{q}_k^{(l)} = \sqrt{\varrho_k^{(l)}} \mathbf{G}^{(l)} \mathbf{h}_k^{(l)}$ is the equivalent channel spanning from the UE-k to the antennas at AP-l, s_k is the desired information of UE-k, ρ_k denotes the transmit power of UE-k, and $\mathbf{w}^{(l)} \sim \mathcal{CN}(\mathbf{0}_M, \sigma_{w^{(l)}}^2 \mathbf{I}_M)$ is the additive noise at AP-l. Furthermore, $u_k \sim \mathcal{CN}(0,1)$ represents the contamination of the information symbol s_k due to HWIs at UE-k, resulting from the power amplifier non-linearities, amplitude/phase imbalance in the In-phase/Quadrature mixers, phase noise in the local oscillator, sampling jitter and finite-resolution quantization in the analog-to-digital converters. Furthermore, $\mathbf{v}_k^{(l)} \sim \mathcal{CN}(\mathbf{0}_M, \mathbf{I}_M)$ is the distortion of the information symbol s_k due to HWIs of the RF chains at AP-l. Finally, ε_{u_k} and $\varepsilon_{\mathbf{v}^{(l)}}$ represents the hardware quality factors of UE-k and AP-l satisfying $0 \leq \varepsilon_{u_k} \leq 1$ and $0 \leq \varepsilon_{\mathbf{v}^{(l)}} \leq 1$, respectively [42], [43]. Explicitly, a hardware quality factor of 1 indicates that the hardware is ideal, while 0 means that the hardware is completely inadequate.

III. BEAMFORMING DESIGN

As shown in the SIM-based cell-free network of Fig. 1, we rely on the distributed processing philosophy for reducing the required overhead of CSI sharing between APs. Specifically, for the information transmitted from the UE-k, each AP carries out a local detection based on its received signal, where the SIM coefficient matrices and the RC vectors of each AP are optimized based on the corresponding local CSI. Then, the CPU gathers the locally detected signals from all APs to generate a final estimate of the UE-k information. In the following, we first present the beamformer design in terms of the SIM coefficient matrices and combining vectors at each AP, followed by the optimization of the weight vectors at the CPU for generating the final detection of the UE information.

A. SIM Coefficient Matrices and RC Vectors Design at APs

To recover the information transmitted from UE-k, we denote the corresponding RC vector at AP-l as $\mathbf{b}_k^{(l)}$ satisfying $\|\mathbf{b}_k^{(l)}\|^2 = 1$. Thus, the locally recovered information of s_k at AP-l, denoted as $\hat{s}_k^{(l)}$, is given by

$$
\hat{s}_k^{(l)} = \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{y}^{(l)}
$$

$$= \sum_{k=1}^{K} \left(\sqrt{\rho_k \varepsilon_{u_k} \varepsilon_{\mathbf{v}^{(l)}}} \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{q}_k^{(l)} s_k \right.$$
$$+ \sqrt{\rho_k (1 - \varepsilon_{u_k}) \varepsilon_{\mathbf{v}^{(l)}}} \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{q}_k^{(l)} u_k$$
$$\left. + \sqrt{\rho_k (1 - \varepsilon_{\mathbf{v}^{(l)}})} \mathbf{b}_k^{(l)\mathrm{H}} \left(\mathbf{q}_k^{(l)} \odot \mathbf{v}_k^{(l)} \right) \right) + \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{w}^{(l)}, \tag{11}$$

and the corresponding signal-to-interference-plus-noise ratio (SINR), denoted as $\gamma_k^{(l)}$, can be formulated as

$$\gamma_k^{(l)} = \frac{\rho_k \varepsilon_{u_k} \varepsilon_{\mathbf{v}^{(l)}} \left| \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{q}_k^{(l)} \right|^2}{\varsigma_{k,k} + \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \left(\rho_{k'} \varepsilon_{u_{k'}} \varepsilon_{\mathbf{v}^{(l)}} \left| \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{q}_{k'}^{(l)} \right|^2 + \varsigma_{k,k'} \right) + \sigma_{w^{(l)}}^2}. \tag{12}$$

In (12) $\varsigma_{k,k'}$ represents the interference resulting from the distortion imposed on UE-k' signal for the recovery of s_k due to the HWIs, given by

$$\varsigma_{k,k'} = \rho_{k'} (1 - \varepsilon_{u_{k'}}) \varepsilon_{\mathbf{v}^{(l)}} \left| \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{q}_{k'}^{(l)} \right|^2 + \rho_{k'} (1 - \varepsilon_{\mathbf{v}^{(l)}}) \cdot$$
$$\mathbf{b}_k^{(l)\mathrm{H}} \left(\left(\mathbf{q}_{k'}^{(l)} \mathbf{q}_{k'}^{(l)\mathrm{H}} \right) \odot \mathbf{I}_M \right) \mathbf{b}_k^{(l)}. \tag{13}$$

For a specific UE-k, we aim for jointly optimizing the active beamformers $\mathbf{b}_1^{(l)}, \mathbf{b}_2^{(l)}, \cdots, \mathbf{b}_K^{(l)}$ and the passive SIM-based beamformers $\mathbf{\Theta}^{(l,1)}, \mathbf{\Theta}^{(l,2)}, \cdots, \mathbf{\Theta}^{(l,T_l)}$ to maximize the SINR of $\gamma_k^{(l)}$ with $l = 1, 2, \cdots, L$. The corresponding optimization problem can be formulated as

$$\text{(P1)} \quad \max_{\mathbf{b}_1^{(l)}, \mathbf{b}_2^{(l)}, \cdots, \mathbf{b}_K^{(l)}, \mathbf{\Theta}^{(l,1)}, \mathbf{\Theta}^{(l,2)}, \cdots, \mathbf{\Theta}^{(l,T_l)}} \gamma_k^{(l)}, \ l = 1, 2, \cdots, L \tag{14}$$

$$\text{s.t.} \quad \mathbf{\Theta}^{(l,t)} \mathbf{\Theta}^{(l,t)\mathrm{H}} = \mathbf{I}_N, \quad t = 1, 2, \cdots, T_l, \tag{15}$$

$$\left\| \mathbf{b}_{k'}^{(l)} \right\|^2 = 1, \quad k' = 1, 2, \cdots, K. \tag{16}$$

Since the L APs jointly support all K UEs, the beamforming of all APs focused on a specific UE-k to maximize the SINR $\gamma_k^{(l)}$ ($l = 1, 2, \cdots, L$) is carried out at the cost of disregarding the other $K - 1$ UEs. The main advantage of the cell-free architecture is that of reducing the path-loss from each AP to its nearest UE. Thus, at each AP we aim for maximizing the channel gain between the AP and its nearest user within the SIM-based holographic beamformer, while the digital beamformer is optimized for all K UEs to get their local information estimates. Specifically, we denote the AP set associated with the SIM-based beamforming focused on UE-k as $\mathcal{L}_k = \{l : D_k^{(l)} \leq D_k^{(l')}, l' = 1, 2, \cdots, L\}$ with $D_k^{(l)}$ being the distance from UE-k to AP-l. Therefore, problem (P1) can be formulated as

$$\text{(P2)} \quad \max_{\mathbf{b}_1^{(l)}, \mathbf{b}_2^{(l)}, \cdots, \mathbf{b}_K^{(l)}, \mathbf{\Theta}^{(l,1)}, \mathbf{\Theta}^{(l,2)}, \cdots, \mathbf{\Theta}^{(l,T_l)}} \gamma_k^{(l)}, \ l \in \mathcal{L}_k \tag{17}$$

$$\text{s.t.} \quad \mathbf{\Theta}^{(l,t)} \mathbf{\Theta}^{(l,t)\mathrm{H}} = \mathbf{I}_N, \quad t = 1, 2, \cdots, T_l, \tag{18}$$

$$\left\| \mathbf{b}_{k'}^{(l)} \right\|^2 = 1, \quad k' = 1, 2, \cdots, K. \tag{19}$$

Since (P2) is a non-convex problem, we can decouple it into a pair of sub-problems and optimize them iteratively as follows.

1) Design of RC vectors at APs: Once the SIM-based beamformer $\mathbf{\Theta}^{(l,1)}, \mathbf{\Theta}^{(l,2)}, \cdots, \mathbf{\Theta}^{(l,T_l)}$ is given, we can estimate the equivalent channels impinging from all K UEs to AP-l, i.e $\mathbf{q}_1^{(l)}, \mathbf{q}_2^{(l)}, \cdots, \mathbf{q}_K^{(l)}$. Then, the optimal active beamformer $\mathbf{b}_{k'}^{(l)}$ designed for the recovery of the information $s_{k'}$ at the AP-l can be optimized by using the MRC criterion as

$$\mathbf{b}_{k'}^{(l)} = \frac{\mathbf{q}_{k'}^{(l)}}{\left\| \mathbf{q}_{k'}^{(l)} \right\|}, \quad k' = 1, 2, \cdots, K. \tag{20}$$

2) Design of SIM coefficient matrices at APs: When the active beamformer $\mathbf{b}_k^{(l)}$ is given, the problem (P2) aims for optimizing the channel gain $\| \mathbf{G}^{(l)} \mathbf{h}_k^{(l)} \|^2$, and it is given by

$$\text{(P3)} \quad \max_{\mathbf{\Theta}_1^{(l)}, \mathbf{\Theta}_2^{(l)}, \cdots, \mathbf{\Theta}_{T_l}^{(l)}} \left\| \mathbf{G}^{(l)} \mathbf{h}_k^{(l)} \right\|^2, \ l \in \mathcal{L}_k \tag{21}$$

$$\text{s.t.} \quad \mathbf{\Theta}_t^{(l)} \mathbf{\Theta}_t^{(l)\mathrm{H}} = \mathbf{I}_N, \quad t = 1, 2, \cdots, T_l. \tag{22}$$

Since the sub-problem (P3) is still a non-convex one, we propose a layer-by-layer iterative optimization algorithm. Firstly, we optimize $\mathbf{\Theta}^{(l,1)}$ by fixing all the other $T_l - 1$ layers of the SIM. Therefore, the channel gain can be represented as $\overline{\mathbf{b}}_k^{(l,1)\mathrm{H}} \mathbf{\Theta}^{(l,1)} \overline{\mathbf{h}}_k^{(l,1)}$ in conjunction with

$$\overline{\mathbf{b}}_k^{(l,1)\mathrm{H}} = \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{A}^{(l,1)} \sqrt{\mathbf{\Upsilon}^{(l,1)}} \tag{23}$$

and

$$\overline{\mathbf{h}}_k^{(l,1)} = \mathbf{A}^{(l,2)} \sqrt{\mathbf{\Upsilon}^{(l,2)}} \mathbf{\Theta}^{(l,2)} \cdots \mathbf{A}^{(l,T_l)} \sqrt{\mathbf{\Upsilon}^{(l,T_l)}} \mathbf{\Theta}^{(l,T_l)} \mathbf{h}_k^{(l)}, \tag{24}$$

and the passive beamformer $\mathbf{\Theta}^{(l,1)}$ can be optimized as

$$\mathbf{\Theta}^{(l,1)} = \mathbf{Diag} \left\{ \mathrm{e}^{\jmath \left(\angle \overline{\mathbf{b}}_k^{(l,1)} - \angle \overline{\mathbf{h}}_k^{(l,1)} \right)} \right\}. \tag{25}$$

Afterwards, we optimize $\mathbf{\Theta}^{(l,2)}$ by fixing all the other $T_l - 1$ layers of the SIM with the channel gain represented as $\overline{\mathbf{b}}_k^{(l,2)\mathrm{H}} \mathbf{\Theta}^{(l,2)} \overline{\mathbf{h}}_k^{(l,2)}$ along with

$$\overline{\mathbf{b}}_k^{(l,2)\mathrm{H}} = \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{A}^{(l,1)} \sqrt{\mathbf{\Upsilon}^{(l,1)}} \mathbf{\Theta}^{(l,1)} \mathbf{A}^{(l,2)} \sqrt{\mathbf{\Upsilon}^{(l,2)}} \tag{26}$$

and

$$\overline{\mathbf{h}}_k^{(l,2)} = \mathbf{A}^{(l,3)} \sqrt{\mathbf{\Upsilon}^{(l,3)}} \mathbf{\Theta}^{(l,3)} \cdots \mathbf{A}^{(l,T_l)} \sqrt{\mathbf{\Upsilon}^{(l,T_l)}} \mathbf{\Theta}^{(l,T_l)} \mathbf{h}_k^{(l)}, \tag{27}$$

and the passive beamformer $\mathbf{\Theta}^{(l,2)}$ can be optimized as

$$\mathbf{\Theta}^{(l,2)} = \mathbf{Diag} \left\{ \mathrm{e}^{\jmath \left(\angle \overline{\mathbf{b}}_k^{(l,2)} - \angle \overline{\mathbf{h}}_k^{(l,2)} \right)} \right\}. \tag{28}$$

Then, $\mathbf{\Theta}^{(l,3)}, \mathbf{\Theta}^{(l,4)}, \cdots, \mathbf{\Theta}^{(l,T_l)}$ can be optimized in turn.

The details of the layer-by-layer iterative optimization of the hybrid beamformer is presented in Algorithm 1.

B. Weight Vector Design at the CPU

The CPU computes its estimate as a linear combination of the local estimates as [44]

$$\hat{s}_k = \sum_{l=1}^{L} \eta_k^{(l)\dagger} \hat{s}_k^{(l)}, \tag{29}$$

$$\hat{s}_k = \sum_{l=1}^{L} \eta_k^{(l)\dagger} \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{y}^{(l)}$$

$$= \underbrace{\sum_{l=1}^{L} \eta_k^{(l)\dagger} \sqrt{\rho_k \varepsilon_{u_k} \varepsilon_{\mathbf{v}^{(l)}}} \left\| \mathbf{q}_k^{(l)} \right\| s_k}_{\text{Desired signal for } s_k} + \underbrace{\sum_{l=1}^{L} \eta_k^{(l)\dagger} \sqrt{\rho_k (1 - \varepsilon_{u_k}) \varepsilon_{\mathbf{v}^{(l)}}} \left\| \mathbf{q}_k^{(l)} \right\| u_k}_{\text{Signal distortion resulting from the HWI at UE-}k} + \underbrace{\sum_{l=1}^{L} \eta_k^{(l)\dagger} \sqrt{\rho_k (1 - \varepsilon_{\mathbf{v}^{(l)}})} \frac{\mathbf{q}_k^{(l)\mathrm{H}} \left(\mathbf{q}_k^{(l)} \odot \mathbf{v}_k^{(l)} \right)}{\left\| \mathbf{q}_k^{(l)} \right\|}}_{\text{Signal distortion resulting from the HWI at APs}}$$

$$+ \underbrace{\sum_{\substack{k'=1 \\ k' \neq k}}^{K} \sum_{l=1}^{L} \eta_k^{(l)\dagger} \left(\sqrt{\rho_{k'} \varepsilon_{u_{k'}} \varepsilon_{\mathbf{v}^{(l)}}} \frac{\mathbf{q}_k^{(l)\mathrm{H}} \mathbf{q}_{k'}^{(l)} s_{k'}}{\left\| \mathbf{q}_k^{(l)} \right\|} + \sqrt{\rho_{k'} (1 - \varepsilon_{u_{k'}}) \varepsilon_{\mathbf{v}^{(l)}}} \frac{\mathbf{q}_k^{(l)\mathrm{H}} \mathbf{q}_{k'}^{(l)} u_{k'}}{\left\| \mathbf{q}_k^{(l)} \right\|} + \sqrt{\rho_{k'} (1 - \varepsilon_{\mathbf{v}^{(l)}})} \frac{\mathbf{q}_k^{(l)\mathrm{H}} \left(\mathbf{q}_{k'}^{(l)} \odot \mathbf{v}_{k'}^{(l)} \right)}{\left\| \mathbf{q}_k^{(l)} \right\|} \right)}_{\text{Inter-user interference}}$$

$$+ \underbrace{\sum_{l=1}^{L} \eta_k^{(l)\dagger} \frac{\mathbf{q}_k^{(l)\mathrm{H}} \mathbf{w}^{(l)}}{\left\| \mathbf{q}_k^{(l)} \right\|}}_{\text{Additive noise}} \tag{30}$$

Algorithm 1 Layer-by-layer iterative optimization algorithm for hybrid beamforming at AP-l ($l \in \mathcal{L}_k$).

Input: The power attenuation matrices of SIM elements $\mathbf{\Upsilon}^{(l,1)}, \mathbf{\Upsilon}^{(l,2)}, \cdots, \mathbf{\Upsilon}^{(l,T_l)}$, and the channel links $\mathbf{A}^{(l,1)}, \mathbf{A}^{(l,2)}, \cdots, \mathbf{A}^{(l,T_l)}$ and $\mathbf{h}_k^{(l)}$.

1: Set the random initial passive SIM-based beamforming matrices $\mathbf{\Theta}^{(l,1)}, \mathbf{\Theta}^{(l,2)}, \cdots, \mathbf{\Theta}^{(l,T_l)}$, satisfying $\mathbf{\Theta}^{(l,1)} \mathbf{\Theta}^{(l,1)\mathrm{H}} = \mathbf{\Theta}^{(l,2)} \mathbf{\Theta}^{(l,2)\mathrm{H}} = \mathbf{\Theta}^{(l,T_l)} \mathbf{\Theta}^{(l,T_l)\mathrm{H}} = \mathbf{I}_N$.

2: **repeat**

3:     The equivalent channel from the UE-k to AP-l antennas is $\mathbf{q}_k^{(l)} = \sqrt{\varrho_k^{(l)}} \mathbf{A}^{(l,1)} \sqrt{\mathbf{\Upsilon}^{(l,1)}} \mathbf{\Theta}^{(l,1)} \mathbf{A}^{(l,2)} \sqrt{\mathbf{\Upsilon}^{(l,2)}}$ $\mathbf{\Theta}^{(l,2)} \cdots \mathbf{A}^{(l,T_l)} \sqrt{\mathbf{\Upsilon}^{(l,T_l)}} \mathbf{\Theta}^{(l,T_l)} \mathbf{h}_k^{(l)}$.

4:     Design the active beamforming vector by the MRC method as $\mathbf{b}_{k'}^{(l)} = \frac{\mathbf{q}_{k'}^{(l)}}{\|\mathbf{q}_{k'}^{(l)}\|}$ for $k' = 1, 2, \cdots, K$.

5:     **for** $t = 1$ to T_l **do**

6:         $\overline{\mathbf{b}}_k^{(l,t)} = \mathbf{b}_k^{(l)\mathrm{H}} \mathbf{A}^{(l,1)} \sqrt{\mathbf{\Upsilon}^{(l,1)}} \mathbf{\Theta}^{(l,1)} \cdots \mathbf{A}^{(l,t)} \sqrt{\mathbf{\Upsilon}^{(l,t)}}$.

7:         $\overline{\mathbf{h}}_k^{(l,t)} = \mathbf{A}^{(l,t+1)} \sqrt{\mathbf{\Upsilon}^{(l,t+1)}} \mathbf{\Theta}^{(l,t+1)} \cdots \mathbf{A}^{(l,T_l)}$ $\sqrt{\mathbf{\Upsilon}^{(l,T_l)}} \mathbf{\Theta}^{(l,T_l)} \mathbf{h}_k^{(l)}$.

8:         The optimal SIM-based passive beamforming matrix for layer-t is $\mathbf{\Theta}^{(l,t)} = \mathbf{Diag}\{e^{\jmath(\angle \overline{\mathbf{b}}_k^{(l,t)} - \angle \overline{\mathbf{h}}_k^{(l,t)})}\}$.

9:     **end for**

10: **until** reaching the iteration times.

Output: The optimized active beamforming vectors $\mathbf{b}_k^{(l)}$ and the optimized passive SIM-based beamforming matrices $\mathbf{\Theta}^{(l,1)}, \mathbf{\Theta}^{(l,2)}, \cdots, \mathbf{\Theta}^{(l,T_l)}$.

---

where $\boldsymbol{\eta}_k = [\eta_k^{(l)}, \eta_k^{(2)}, \cdots, \eta_k^{(L)}]^{\mathrm{T}} \in \mathbb{C}^{L \times 1}$ is the weight vector that the CPU assigns to the local signal estimate of the signal arriving from UE-k satisfying $\|\boldsymbol{\eta}_k\|^2 = 1$. Therefore, according to (11), $\hat{s}_k$ can be written as in (30). The SINR of $\hat{s}_k$, denoted as γ_k, can be derived as

$$\gamma_k = \frac{\rho_k \left| \boldsymbol{\eta}_k^{\mathrm{H}} \mathbf{z}_{k,k} \right|^2}{\boldsymbol{\eta}_k^{\mathrm{H}} \mathbf{R}_k \boldsymbol{\eta}_k}, \tag{31}$$

where $\mathbf{R}_k$ is given by

$$\mathbf{R}_k = \rho_k \left(\mathbf{z}_{u_{k,k}} \mathbf{z}_{u_{k,k}}^{\mathrm{H}} + \left(\mathbf{z}_{\mathbf{v}_{k,k}} \mathbf{z}_{\mathbf{v}_{k,k}}^{\mathrm{H}} \right) \odot \mathbf{I}_L \right) + \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \rho_{k'} \left(\mathbf{z}_{k,k'} \right.$$
$$\left. \mathbf{z}_{k,k'}^{\mathrm{H}} + \mathbf{z}_{u_{k,k'}} \mathbf{z}_{u_{k,k'}}^{\mathrm{H}} + \left(\mathbf{z}_{\mathbf{v}_{k,k'}} \mathbf{z}_{\mathbf{v}_{k,k'}}^{\mathrm{H}} \right) \odot \mathbf{I}_L \right) + \mathbf{W} \tag{32}$$

in conjunction with

$$\mathbf{z}_{k,i} = \begin{bmatrix} \frac{\sqrt{\varepsilon_{u_k} \varepsilon_{\mathbf{v}^{(1)}}}}{\|\mathbf{q}_k^{(1)}\|} \mathbf{q}_k^{(1)\mathrm{H}} \mathbf{q}_i^{(1)} \\ \vdots \\ \frac{\sqrt{\varepsilon_{u_k} \varepsilon_{\mathbf{v}^{(L)}}}}{\|\mathbf{q}_k^{(L)}\|} \mathbf{q}_k^{(L)\mathrm{H}} \mathbf{q}_i^{(L)} \end{bmatrix}, \tag{33}$$

$$\mathbf{z}_{u_{k,i}} = \begin{bmatrix} \frac{\sqrt{(1-\varepsilon_{u_k}) \varepsilon_{\mathbf{v}^{(1)}}}}{\|\mathbf{q}_k^{(1)}\|} \mathbf{q}_k^{(1)\mathrm{H}} \mathbf{q}_i^{(1)} \\ \vdots \\ \frac{\sqrt{(1-\varepsilon_{u_k}) \varepsilon_{\mathbf{v}^{(L)}}}}{\|\mathbf{q}_k^{(L)}\|} \mathbf{q}_k^{(L)\mathrm{H}} \mathbf{q}_i^{(L)} \end{bmatrix}, \tag{34}$$

$$\mathbf{z}_{\mathbf{v}_{k,i}} = \begin{bmatrix} \frac{\sqrt{1-\varepsilon_{\mathbf{v}^{(1)}}}}{\|\mathbf{q}_k^{(L)}\|} \left\| \mathbf{q}_k^{(1)\dagger} \odot \mathbf{q}_i^{(1)} \right\| \\ \vdots \\ \frac{\sqrt{1-\varepsilon_{\mathbf{v}^{(L)}}}}{\|\mathbf{q}_k^{(L)}\|} \left\| \mathbf{q}_k^{(L)\dagger} \odot \mathbf{q}_i^{(L)} \right\| \end{bmatrix}, \tag{35}$$

and

$$\mathbf{W} = \mathbf{Diag}\{\sigma_{w^{(1)}}^2, \sigma_{w^{(2)}}^2, \cdots, \sigma_{w^{(L)}}^2\}. \tag{36}$$

Based on the generalized Rayleigh quotient, the maximum of γ_k in (31) can be attained as follows:

$$\gamma_k = \rho_k \mathbf{z}_{k,k}^{\mathrm{H}} \mathbf{R}_k^{-1} \mathbf{z}_{k,k}$$
$$= \rho_k \mathbf{z}_{k,k}^{\mathrm{H}} \left(\rho_k \left(\mathbf{z}_{u_{k,k}} \mathbf{z}_{u_{k,k}}^{\mathrm{H}} + \left(\mathbf{z}_{\mathbf{v}_{k,k}} \mathbf{z}_{\mathbf{v}_{k,k}}^{\mathrm{H}} \right) \odot \mathbf{I}_L \right) \right.$$

$$+ \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \rho_{k'} \left(\mathbf{z}_{k,k'} \mathbf{z}_{k,k'}^{\mathrm{H}} + \mathbf{z}_{u_{k,k'}} \mathbf{z}_{u_{k,k'}}^{\mathrm{H}} \right.$$

$$\left. + \left(\mathbf{z}_{\mathbf{v}_{k,k'}} \mathbf{z}_{\mathbf{v}_{k,k'}}^{\mathrm{H}} \right) \odot \mathbf{I}_L \right) + \mathbf{W} \right)^{-1} \mathbf{z}_{k,k}, \qquad (37)$$

by setting

$$\boldsymbol{\eta}_k = \mathbf{R}_k^{-1} \mathbf{z}_{k,k}. \qquad (38)$$

C. Computational Complexity

1) Computational complexity of the hybrid beamformer at each AP: The computational complexity of the proposed layer-by-layer iterative hybrid beamformer at AP-l ($l \in \mathcal{L}_k$) in Algorithm 1 depends both on the number of iterations in the alternating maximization, which is denoted as τ, and on the computational complexity required in each iteration. As for each iteration, the sub-problem of optimizing the combining vectors $\mathbf{b}_1^{(l)}, \mathbf{b}_2^{(l)}, \cdots, \mathbf{b}_K^{(l)}$ is solved in line 3 and 4 of Algorithm 1, while the sub-problem of optimizing the coefficients of all layers in the SIM $\boldsymbol{\Theta}^{(l,1)}, \boldsymbol{\Theta}^{(l,2)}, \cdots, \boldsymbol{\Theta}^{(l,T_l)}$ is evaluated in line 5 to 9 of Algorithm 1. The complexity of additions is neglected, since its operation is easily implemented in hardware. Hence, we quantify the computational complexity by counting the number of floating-point multiplication operations that are required. Specifically, the calculation of $\mathbf{q}_k^{(l)}$ in line 3 of Algorithm 1 requires

$$c_1' = (T_l - 1)\left(N^2 + 2N\right) + MN + 2N + M \qquad (39)$$

floating-point multiplication operations by exploiting the property that $\boldsymbol{\Upsilon}^{(l,1)}, \boldsymbol{\Upsilon}^{(l,2)}, \cdots, \boldsymbol{\Upsilon}^{(l,T_l)}$ and $\boldsymbol{\Theta}^{(l,1)}, \boldsymbol{\Theta}^{(l,2)}, \cdots, \boldsymbol{\Theta}^{(l,T_l)}$ are diagonal matrices. The calculation of $\mathbf{b}_1^{(l)}, \mathbf{b}_2^{(l)}, \cdots, \mathbf{b}_K^{(l)}$ in line 4 of Algorithm 1 requires

$$c_2' = 2KM \qquad (40)$$

floating-point multiplications. Furthermore, the calculation of $\overline{\mathbf{b}}_k^{(l,t)}$ and $\overline{\mathbf{h}}_k^{(l,t)}$ in line 6 and 7 of Algorithm 1 requires

$$c_{3,t1}' = (t - 1)\left(N^2 + 2N\right) + (M + 1)N \qquad (41)$$

and

$$c_{3,t2}' = (T_l - t)\left(N^2 + 2N\right) \qquad (42)$$

floating-point multiplications, respectively. The loop between line 8 of Algorithm 1 has no floating-point multiplications. Thus, the loop between line 5 to 9 of Algorithm 1 entails

$$c_3' = \sum_{t=1}^{T_L} \left(c_{3,t1}' + c_{3,t2}' \right)$$
$$= \left(T_l^2 - T_l\right)\left(N^2 + 2N\right) + T_l(M + 1)N \qquad (43)$$

floating-point multiplications. Therefore, according to (39), (40) and (43), the total number of floating-point multiplications in each iteration is

$$c' = c_1' + c_2' + c_3'$$
$$= \left(T_l^2 - 1\right)N^2 + \left(2T_l^2 + T_lM + T_L + M\right)N$$
$$+ (2K + 1)M. \qquad (44)$$

Hence for $N > M$ and $N > K$, the overall computational complexity of our proposed layer-by-layer iterative optimization of the hybrid beamformer at all APs is

$$\mathcal{O}\left(\tau \left(\sum_{l=1}^{L} T_l^2 \right) N^2 \right). \qquad (45)$$

This shows that the proposed layer-by-layer iterative optimization algorithm conceived for the SIM-based hybrid beamformer is of polynomial time-complexity with respect to the number of APs, the number of SIM layers at each AP, and the number of reconfigurable elements in each SIM layer.

2) Computational complexity of the CPU processing: To recover the information s_k, the computational complexity at the CPU includes the calculation of the matrix $\mathbf{R}_k$ defined in (32), of the linear combination vector $\boldsymbol{\eta}_k$ defined in (38) and of the information recovery $\hat{s}_k$ defined in (29). Specifically, the calculation of $\mathbf{R}_k$ requires

$$c_1'' = \left(L^2 + 2L + 3M\right)K \qquad (46)$$

floating-point multiplication operations according to (32), (33), (34) and (35). The calculation of the linear combining vector $\boldsymbol{\eta}_k$ requires

$$c_2'' = \frac{1}{3}\left(L^3 - L\right) + L^2 \qquad (47)$$

floating-point multiplication operations by employing the Cholesky decomposition of the Hermitian positive-definite matrix $\mathbf{R}_k^{-1}$. Furthermore, the calculation of the information recovery $\hat{s}_k$ requires

$$c_3'' = L \qquad (48)$$

floating-point multiplication operations according to (29). Therefore, according to (46), (47) and (48), the total number of floating-point multiplications required for recovering the information s_k is

$$c'' = c_1'' + c_2'' + c_3''$$
$$= \frac{1}{3}L^3 + (K + 1)L^2 + \left(2K + \frac{2}{3}\right)L + 3MK. \qquad (49)$$

Therefore, the overall computational complexity of recovering the information $s_1, s_2, \cdots, s_K$ at the CPU is

$$\mathcal{O}\left(KL^3\right) + \mathcal{O}\left(K^2L^2\right) + \mathcal{O}\left(K^2M\right). \qquad (50)$$

This shows that the CPU processing conceived for information recovery in the cell-free systems is of polynomial time-complexity with respect to the number of UEs, the number of APs and the number of RF chains at each AP.

D. Convergence Analysis

First, in line 3 and 4 of Algorithm 1 convinced for the digital beamforming design, the received SINR γ_k becomes non-decreasing after the digital beamformer $\mathbf{b}_1^{(l)}, \mathbf{b}_2^{(l)}, \cdots, \mathbf{b}_K^{(l)}$ has been optimized, given the holographic beamformer $\boldsymbol{\Theta}^{(l,1)}, \boldsymbol{\Theta}^{(l,2)}, \cdots, \boldsymbol{\Theta}^{(l,T_l)}$, i.e.,

$$\gamma_k(\ddot{\mathbf{b}}^{(l,i+1)}, \ddot{\boldsymbol{\Theta}}^{(l,i)}) \geq \gamma_k(\ddot{\mathbf{b}}^{(l,i)}, \ddot{\boldsymbol{\Theta}}^{(l,i)}), \qquad (51)$$

TABLE III
SIMULATION PARAMETERS.

Parameters	Values
Carrier frequency	$f_c = 30$ GHz
Number of APs	$L = 16$
Number of UEs	$K = 8$
Number of AP antennas	$M = 4 \times 4$
Number of SIM elements in each layer	$N = 16 \times 16$
Number of SIM layers	$T_l = 4 \ (l = 1, 2, \cdots L)$
Inter-layer distance	$z^{(l,0)} = z^{(l,1)} = \cdots = z^{(l,T_l-1)} = \lambda \ (l = 1, 2, \cdots L)$
Directional radiation gain	$\kappa = 10$dB
Distance between AP antennas	$d_x = d_y = \frac{\lambda}{2}$
Size of the SIM elements	$\delta_x = \delta_y = \frac{\lambda}{4}$
Noise power	$\sigma^2_{w^{(l)}} = -80$dBm $(l = 1, 2, \cdots L)$
Hardware quality factor	$\varepsilon_{u_k} = \varepsilon_{\mathbf{V}^{(l)}} = \varepsilon = 1 \ (k = 1, 2, \cdots, K,$ and $l = 1, 2, \cdots, L)$
Transmit power at UEs	$\rho = \rho_1 = \rho_2 = \cdots = \rho_K$
Path loss exponent	$\beta = 3.5$
Path loss at the reference distance of 1 meter	$C_0 = 10^{-3}$
Number of clusters	$\zeta_c = 4 \ (c = 1, 2, \cdots, \zeta_c)$
Number of paths in each cluster	$\zeta_p = 8 \ (p = 1, 2, \cdots, \zeta_p)$
Mean values of the elevation angles	$\mu_{\psi^k_c}$ are randomly distributed in $[0°, 180°] \ (c = 1, 2, \cdots, \zeta_c,$ and $k = 1, 2, \cdots, K)$
Mean values of the azimuth angles	$\mu_{\varphi^k_c}$ are randomly distributed in $[0°, 360°] \ (c = 1, 2, \cdots, \zeta_c,$ and $k = 1, 2, \cdots, K)$
Elevation angle spreads	$\sigma_{\psi^k_c} = 7.5° \ (c = 1, 2, \cdots, \zeta_c,$ and $k = 1, 2, \cdots, K)$
Azimuth angle spreads	$\sigma_{\varphi^k_c} = 7.5° \ (c = 1, 2, \cdots, \zeta_c,$ and $k = 1, 2, \cdots, K)$
SIM power radiation coefficients	$v_n^{(l,t)} = 1 \ (n = 1, 2, \cdots, N, t = 1, 2, \cdots, T_l,$ and $l = 1, 2, \cdots, L)$

Fig. 2. Comparison of the average achievable rate $\overline{R}$ versus the transmit power ρ with different number of APs.

Fig. 3. Comparison of the average achievable rate $\overline{R}$ versus the transmit power ρ with different number of SIM layers.

where $\gamma_k(\ddot{\mathbf{b}}^{(l,i_1)}, \ddot{\boldsymbol{\Theta}}^{(l,i_2)})$ represents the received SINR based on the digital beamformer $\mathbf{b}_1^{(l)}, \mathbf{b}_2^{(l)}, \cdots, \mathbf{b}_K^{(l)}$ in the i_1th iteration and on the holographic beamformer $\boldsymbol{\Theta}^{(l,1)}, \boldsymbol{\Theta}^{(l,2)}, \cdots, \boldsymbol{\Theta}^{(l,T_l)}$ in the i_2th iteration. Secondly, in line 5 to 9 of Algorithm 1, the received SINR γ_k becomes non-decreasing after the holographic beamformer $\boldsymbol{\Theta}^{(l,1)}, \boldsymbol{\Theta}^{(l,2)}, \cdots, \boldsymbol{\Theta}^{(l,T_l)}$ has been optimized, given the digital beamformer $\mathbf{b}_1^{(l)}, \mathbf{b}_2^{(l)}, \cdots, \mathbf{b}_K^{(l)}$, i.e.,

$$\gamma_k(\ddot{\mathbf{b}}^{(l,i)}, \ddot{\boldsymbol{\Theta}}^{(l,i+1)}) \geq \gamma_k(\ddot{\mathbf{b}}^{(l,i)}, \ddot{\boldsymbol{\Theta}}^{(l,i)}). \quad (52)$$

Therefore, (51) and (52) imply that in each iteration of the proposed layer-by-layer iterative optimization algorithm at each AP, the objective function value of the received SINR γ_k is non-decreasing. Additionally, the objective function value sequence obtained throughout the iteration process is monotonic and it is also bounded, hence the overall algorithm is guaranteed to converge.

IV. NUMERICAL AND SIMULATION RESULTS

In this section, the average achievable rate of the SIM-based holographic MIMO of cell-free networks is quantified. A total of L APs are uniformly distributed in the area $\mathcal{S}$ with the Cartesian coordinate of $\{(x, y) : -100\text{m} \leq x \leq 100\text{m}, -100\text{m} \leq y \leq 100\text{m}\}$. Furthermore, K UEs are uniform-randomly distributed in the area $\mathcal{S}$. We employ a distance-dependent path-loss model for the UE-AP channel,
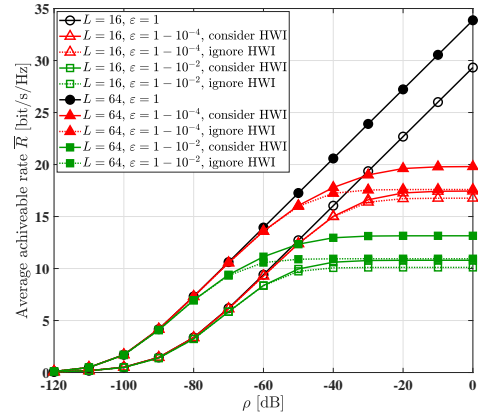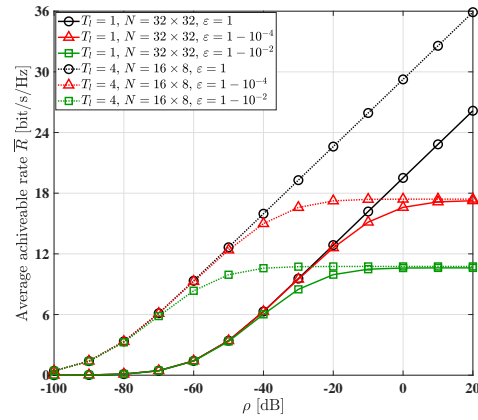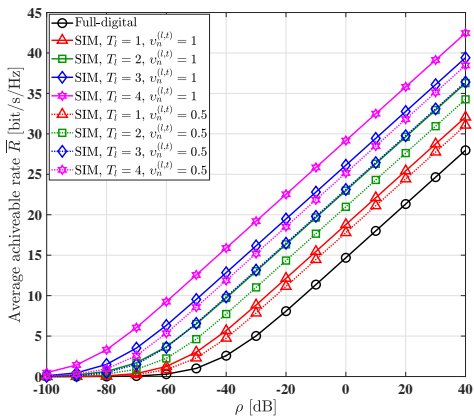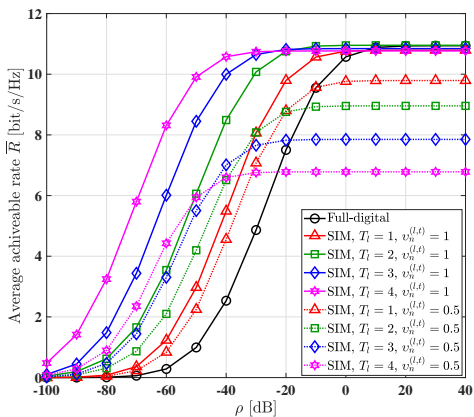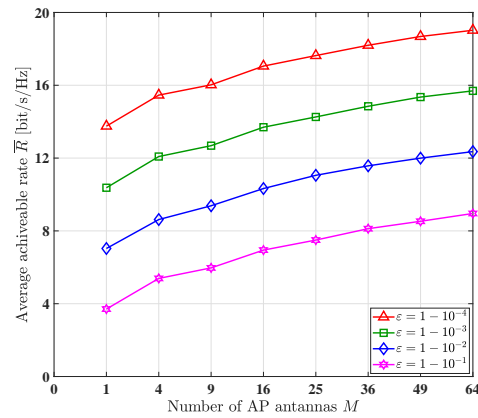
given by $\varrho_k^{(l)} = \min\{C_0, C_0(d_k^{(l)})^{-\beta}\}$, where $d_k^{(l)}$ is the length of the link spanning from AP-l to UE-k, β is the path loss exponent, and C_0 is the path loss at the reference distance of 1 meter [45]. Referring to [32], [46], the simulation parameters are those given in Table III, unless otherwise specified.

Fig. 2 compares the average achievable rate $\overline{R} = \frac{1}{K}\sum_{k=1}^K R_k = \frac{1}{K}\sum_{k=1}^K \log_2(1 + \gamma_k)$ versus the transmit power ρ for different number of APs in cell-free networks, with the number of AP $L = 16$ and $L = 64$, respectively. Observe from Fig. 2 that the average achievable rate can be improved by increasing the number of APs when the average distance between the APs and UEs can be reduced. When the hardware is non-ideal, i.e. $\varepsilon < 1$, the average achievable rate tends to a constant value upon increasing of the transmit power ρ. Furthermore, as seen in Fig. 2, an obvious performance degradation exists when the HWIs are ignored during information recovery, especially with the increase of the number of APs. Fig. 3 compares the average achievable rate $\overline{R}$ versus the transmit power ρ for different number of SIM layers in each AP, where the number N of elements in each SIM layer is set to $N = 16 \times 16$, $N = 8 \times 8$ and $N = 4 \times 4$ when the number of SIM layers T_l are 1, 2 and 4 respectively for ensuring that the same total number of SIM elements are employed. Observe in Fig. 3 that the average achievable rate

(a) Hardware quality factor $\varepsilon = 1$.

(b) Hardware quality factor $\varepsilon = 1 - 10^{-4}$.

(c) Hardware quality factor $\varepsilon = 1 - 10^{-2}$.

Fig. 4. Average achievable rate $\overline{R}$ versus the transmit power ρ in the conventional full-digital beamformer and the SIM-based hybrid beamformer.

Fig. 5. Average achievable rate $\overline{R}$ versus the number of AP antennas M.

rate than the conventional full-digital beamformer, even when the SIM has the signal radiation attenuation associated with the power radiation coefficients of $\upsilon_n^{(l,t)} = 0.5$. This is a benefit of the beamforming gain attained by the configuration of the SIM elements. Furthermore, the average achievable rate can be improved upon increasing the number of SIM layers. When the hardware quality is imperfect, i.e., $\varepsilon < 1$, it is illustrated in Fig. 4 (b) and Fig. 4 (c) that the SIM-based hybrid beamformer outperforms the full-digital beamformer in the low-SNR region, while the achievable rate in the high-SNR region is limited by the hardware quality. Moreover, in the high-SNR region, increasing the number of SIM layers even degrades the achievable rate when the SIM power radiation coefficients $\upsilon_n^{(l,t)} < 1$, since the signals are attenuated in each layer of the SIM and it cannot be compensated by the configuration of SIM elements due to the limitation of the hardware quality. It can be observed that although the SIM can increase the equivalent channel gain, but it cannot compensate for the deleterious effects of HWIs in the high-SNR region.

The performance comparison of the average achievable rate $\overline{R}$ versus the number of antennas M at each AP is presented in Fig. 5, for different hardware quality factors ε. It shows that the average achievable rate can be improved upon increasing the number of antennas at each AP, but only at the cost of higher energy consumption. However, the signal distortion resulting from the imperfect hardware quality can be compensated by employing more AP antennas.

Fig. 6 portrays the average achievable rate versus the inter-layer distance, where we have $z = z^{(l,0)} = z^{(l,1)} = \cdots = z^{(l,T_l-1)}$ for $l = 1, 2, \cdots, L$. The figure shows that there exists an optimal inter-layer distance in terms of the highest average achievable rate. Furthermore, it also shows that the optimal inter-layer distance increases upon employing more intelligent surface elements. Observe that when the number of intelligent surface elements is small, the beamforming gain attained by the SIM is predominantly limited by the path loss between layers, and thus a higher average achievable rate can be attained by reducing the inter-layer distance. By contrast, upon increasing the number of intelligent surface elements, the inter-layer path loss effect becomes negligible and increasing the inter-layer distance improves the signal radiation between
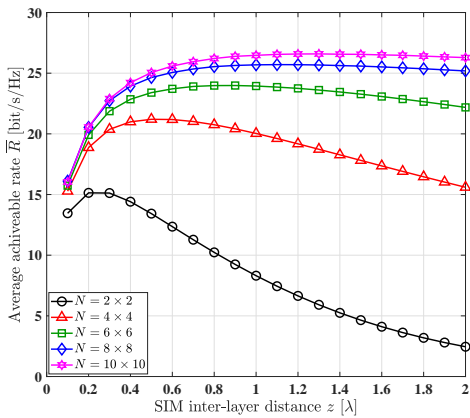
is improved upon increasing the number of SIM layers. It inspires our future research on the SIM design to determine how many SIM layers is optimal in terms of maximizing the performance gain given the total number of SIM elements.

To provide further insights, Fig. 4 characterizes the average achievable rate of both our SIM-based hybrid beamforming architecture and of the conventional cell-free full-digital beamforming architecture, for the hardware quality factors of $\varepsilon = 1$, $1 - 10^{-4}$ and $1 - 10^{-2}$, respectively. In terms of the perfect hardware quality factor, i.e., $\varepsilon = 1$, Fig. 4 (a) shows that the SIM-based hybrid beamformer has higher average achievable

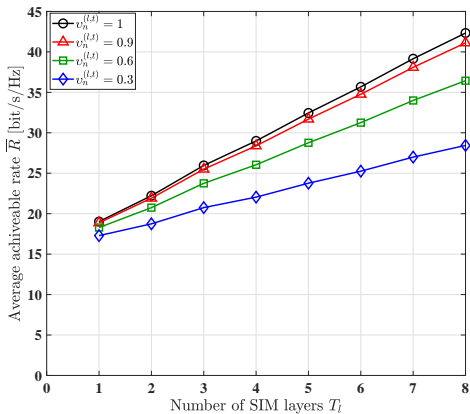Fig. 6. Average achievable rate $\overline{R}$ versus the SIM inter-layer distance z.

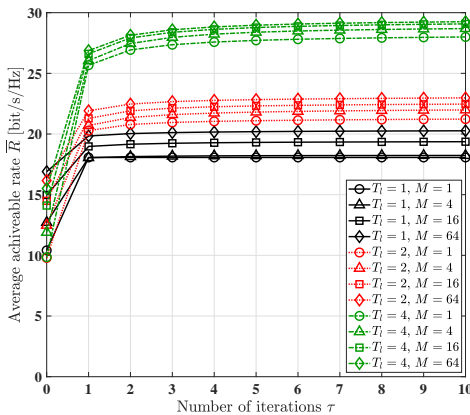Fig. 7. Average achievable rate $\overline{R}$ versus the number of SIM layers T_l.

Fig. 8. Average achievable rate $\overline{R}$ versus the number of iterations τ.

adjacent layers, increasing the achievable rate.

To investigate the impact of the signal attenuation caused by the signal travelling through each layer of the SIM on the achievable rate performance, Fig. 7 presents the average achievable rate $\overline{R}$ versus the number of SIM layers at each AP, characterize by different power radiation coefficients $v_n^{(l,t)}$. It shows that although the average achievable rate degrades with the reduction of the radiation coefficients $v_n^{(l,t)}$, this effect can be compensated by increasing the number of SIM layers.
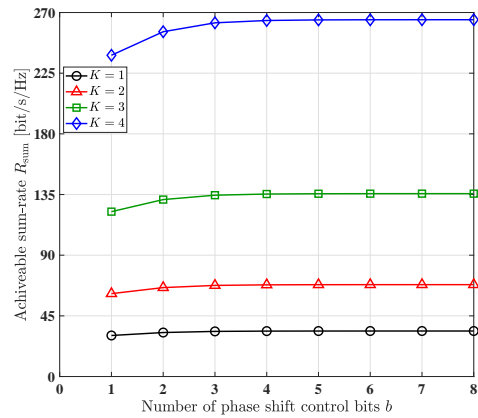
In Fig. 8 we portray the average achievable rate $\overline{R}$ versus

Fig. 9. Achievable sum-rate R_{sum} versus the number of phase shift control bit b.

the number of iterations τ attained by the proposed layer-by-layer iterative optimization algorithm used by the hybrid beamformer. Observe that although the convergence speed is reduced as the number of SIM layers increases, it achieves convergence within 10 iterations.

In the above simulations, we assume that the reconfigurable SIM elements have continuous phase shifts in the range of $[0, 2\pi)$. However, in practical hardware implementations, the phase shift of each reconfigurable element is limited to a finite number of discrete values. For simplicity, we assume that the discrete phase shift set is obtained by uniformly quantizing the interval $[0, 2\pi)$ into 2^b levels, with b respecting the number of phase shift control bits. Thus, we have $\theta_n^{(l,t)} \in \{2\pi \cdot \frac{0}{2^B}, 2\pi \cdot \frac{1}{2^B}, \cdots, 2\pi \cdot \frac{2^b-1}{2^b}\}$. Fig. 9 characterizes the achievable sum-rate, denoted as $R_{\mathrm{sum}} = \sum_{k=1}^{K} R_k = \sum_{k=1}^{K} \log_2(1 + \gamma_k)$ versus the number of phase shift control bits b, with different number of UEs k. It shows that 4-bit phase shift quantization can approach the rate of the infinite phase shift resolution.

V. CONCLUSIONS

We conceived an uplink SIM-based cell-free HMIMO architecture, where the distributed signal processing was employed. Each AP carries out a local detection of the UE information, where the hybrid beamformer and the RC vectors of each distributed AP are optimized based on our low-complexity layer-by-layer iterative optimization algorithm. The CPU recovers the final UE information by fusing the local detections gleaned from all APs, where the RC weight vector used for combining the local detections is designed based on the MMSE criterion, taking into account the HWIs of the RF chains at the APs and of the UEs. The simulation results showed that the achievable rate of the SIM-based cell-free HMIMO network improves upon increasing the number of SIM layers as well as the number of elements in each layer. Furthermore, due to the HWI of the RF chains at the APs and the UEs, the achievable rate saturates in the high-SNR region.

REFERENCES

[1] P. Wang, J. Fang, L. Dai, and H. Li, "Joint transceiver and large intelligent surface design for massive MIMO mmWave systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1052–1064, 2020.

[2] L. Dai, J. Tan, Z. Chen, and H. V. Poor, "Delay-phase precoding for wideband THz massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7271–7286, 2022.

[3] J. Du, S. Ye, L. Jin, X. Li, H. Q. Ngo, and O. A. Dobre, "Tensor-based joint channel estimation for multi-way massive MIMO hybrid relay systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 9, pp. 9571–9585, 2022.

[4] X. Li, M. Zhang, H. Chen, C. Han, L. Li, D.-T. Do, S. Mumtaz, and A. Nallanathan, "UAV-enabled multi-pair massive MIMO-NOMA relay systems with low-resolution ADCs/DACs," *IEEE Trans. Veh. Technol.*, vol. 73, no. 2, pp. 2171–2186, 2024.

[5] J. Gao, C. Zhong, G. Y. Li, J. B. Soriaga, and A. Behboodi, "Deep learning-based channel estimation for wideband hybrid mmWave massive MIMO," *IEEE Trans. Commun.*, vol. 71, no. 6, pp. 3679–3693, 2023.

[6] C.-J. Wang, C.-K. Wen, S. Jin, and S.-H. Tsai, "Finite-alphabet precoding for massive MU-MIMO with low-resolution DACs," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4706–4720, 2018.

[7] L. Yang, J. Yang, W. Xie, M. O. Hasna, T. Tsiftsis, and M. Di Renzo, "Secrecy performance analysis of RIS-aided wireless communication systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12 296–12 300, 2020.

[8] L. Yang, F. Meng, J. Zhang, M. O. Hasna, and M. Di Renzo, "On the performance of RIS-assisted dual-hop UAV communication systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10 385–10 390, 2020.

[9] L. V. Nguyen, A. L. Swindlehurst, and D. H. Nguyen, "Linear and deep neural network-based receivers for massive MIMO systems with one-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7333–7345, 2021.

[10] Q. Li, M. El-Hajjar, I. Hemadeh, A. Shojaeifard, A. A. Mourad, B. Clerckx, and L. Hanzo, "Reconfigurable intelligent surfaces relying on non-diagonal phase shift matrices," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6367–6383, 2022.

[11] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. Di Renzo, and M. Debbah, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, 2020.

[12] Q. Li, M. El-Hajjar, I. Hemadeh, D. Jagyasi, A. Shojaeifard, E. Basar, and L. Hanzo, "The reconfigurable intelligent surface-aided multi-node IoT downlink: Beamforming design and performance analysis," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 6400–6414, 2022.

[13] Q. Li, M. El-Hajjar, I. Hemadeh, A. Shojaeifard, A. A. Mourad, and L. Hanzo, "Reconfigurable intelligent surface aided amplitude-and phase-modulated downlink transmission," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 8146–8151, 2023.

[14] I. Yoo and D. R. Smith, "Sub-6-Ghz uplink massive MIMO system using holographic beamforming metasurfaces: A conceptual development," *IEEE Wireless Commun. Lett.*, vol. 12, no. 4, pp. 644–648, 2023.

[15] R. Deng, Y. Zhang, H. Zhang, B. Di, H. Zhang, and L. Song, "Reconfigurable holographic surface: A new paradigm to implement holographic radio," *IEEE Veh. Technol. Mag.*, vol. 18, no. 1, pp. 20–28, 2023.

[16] T. Gong, P. Gavriilidis, R. Ji, C. Huang, G. C. Alexandropoulos, L. Wei, Z. Zhang, M. Debbah, H. V. Poor, and C. Yuen, "Holographic MIMO communications: Theoretical foundations, enabling technologies, and future directions," *IEEE Commun. Surv. Tutor.*, vol. 26, no. 1, pp. 196–257, 2024.

[17] J. An, C. Yuen, C. Huang, M. Debbah, H. V. Poor, and L. Hanzo, "A tutorial on holographic MIMO communication–Part I: Channel modeling and channel estimation," *IEEE Commun. Lett.*, vol. 27, no. 7, pp. 1664–1668, 2023.

[18] S. Zeng, H. Zhang, B. Di, H. Qin, X. Su, and L. Song, "Reconfigurable refractive surfaces: An energy-efficient way to holographic MIMO," *IEEE Commun. Lett.*, vol. 26, no. 10, pp. 2490–2494, 2022.

[19] R. Deng, B. Di, H. Zhang, Y. Tan, and L. Song, "Reconfigurable holographic surface: Holographic beamforming for metasurface-aided wireless communications," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 6255–6259, 2021.

[20] R. Deng, B. Di, H. Zhang, and L. Song, "HDMA: Holographic-pattern division multiple access," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1317–1332, 2022.

[21] R. Deng, B. Di, H. Zhang, D. Niyato, Z. Han, H. V. Poor, and L. Song, "Reconfigurable holographic surfaces for future wireless communications," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 126–131, 2021.

[22] X. Hu, R. Deng, B. Di, H. Zhang, and L. Song, "Holographic beamforming for ultra massive MIMO with limited radiation amplitudes: How many quantized bits do we need?" *IEEE Commun. Lett.*, vol. 26, no. 6, pp. 1403–1407, 2022.

[23] R. Deng, B. Di, H. Zhang, H. V. Poor, and L. Song, "Holographic MIMO for LEO satellite communications aided by reconfigurable holographic surfaces," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 3071–3085, 2022.

[24] X. Hu, R. Deng, B. Di, H. Zhang, and L. Song, "Holographic beamforming for LEO satellites," *IEEE Commun. Lett.*, vol. 27, no. 10, pp. 2717–2721, 2023.

[25] L. Wei, C. Huang, G. C. Alexandropoulos, E. Wei, Z. Zhang, M. Debbah, and C. Yuen, "Multi-user holographic MIMO surfaces: Channel modeling and spectral efficiency analysis," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 5, pp. 1112–1124, 2022.

[26] H. Wu, Y. Chen, Y. Ming, and Z. Wang, "Two-timescale beamforming optimization for downlink multi-user holographic MIMO surfaces," *IEEE Trans. Veh. Technol.*, vol. 73, no. 3, pp. 4476–4481, 2023.

[27] N. Shlezinger, O. Dicker, Y. C. Eldar, I. Yoo, M. F. Imani, and D. R. Smith, "Dynamic metasurface antennas for uplink massive MIMO systems," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6829–6843, 2019.

[28] H. Wang, N. Shlezinger, S. Jin, Y. C. Eldar, I. Yoo, M. F. Imani, and D. R. Smith, "Dynamic metasurface antennas based downlink massive MIMO systems," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[29] L. You, J. Xu, G. C. Alexandropoulos, J. Wang, W. Wang, and X. Gao, "Energy efficiency maximization of massive MIMO communications with dynamic metasurface antennas," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 393–407, 2022.

[30] Y. Li, S. Gong, H. Liu, C. Xing, N. Zhao, and X. Wang, "Near-field beamforming optimization for holographic XL-MIMO multiuser systems," *IEEE Trans. Commun.*, vol. 72, no. 4, pp. 2309–2323, 2024.

[31] J. An, C. Xu, D. W. K. Ng, G. C. Alexandropoulos, C. Huang, C. Yuen, and L. Hanzo, "Stacked intelligent metasurfaces for efficient holographic MIMO communications in 6G," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2380–2396, 2023.

[32] J. An, C. Yuen, C. Xu, H. Li, D. W. K. Ng, M. Di Renzo, M. Debbah, and L. Hanzo, "Stacked intelligent metasurface-aided MIMO transceiver design," *IEEE Wireless Commun.*, 2024.

[33] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured massive access for scalable cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1086–1100, 2020.

[34] W. Xu, J. An, H. Li, L. Gan, and C. Yuen, "Algorithm unrolling-based distributed optimization for RIS-assisted cell-free networks," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 944–957, 2024.

[35] J. Xu, L. You, G. C. Alexandropoulos, X. Yi, W. Wang, and X. Gao, "Near-field wideband extremely large-scale MIMO transmissions with holographic metasurface-based antenna arrays," *IEEE Trans. Wireless Commun.*, 2024.

[36] C. Liu, Q. Ma, Z. J. Luo, Q. R. Hong, Q. Xiao, H. C. Zhang, L. Miao, W. M. Yu, Q. Cheng, L. Li *et al.*, "A programmable diffractive deep neural network based on a digital-coding metasurface array," *Nature Electronics*, vol. 5, no. 2, pp. 113–122, 2022.

[37] J. An, C. Yuen, Y. L. Guan, M. Di Renzo, M. Debbah, H. V. Poor, and L. Hanzo, "Two-dimensional direction-of-arrival estimation using stacked intelligent metasurfaces," *arXiv preprint arXiv:2402.08224*, 2024.

[38] H. Lu and Y. Zeng, "Communicating with extremely large-scale array/surface: Unified modeling and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4039–4053, 2021.

[39] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Channel modeling and channel estimation for holographic massive MIMO with planar arrays," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 997–1001, 2022.

[40] M. Cui and L. Dai, "Channel estimation for extremely large-scale MIMO: Far-field or near-field?" *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2663–2677, 2022.

[41] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 2, pp. 870–913, 2017.

[42] E. Björnson, J. Hoydis, L. Sanguinetti *et al.*, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[43] Q. Li, M. El-Hajjar, Y. Sun, I. Hemadeh, A. Shojaeifard, Y. Liu, and L. Hanzo, "Achievable rate analysis of the STAR-RIS aided NOMA uplink in the face of imperfect CSI and hardware impairments," *IEEE Trans. Commun.*, vol. 71, no. 10, pp. 6100–6114, 2023.

[44] Ö. T. Demir, E. Björnson, L. Sanguinetti *et al.*, "Foundations of user-centric cell-free massive MIMO," *Foundations and Trends® in Signal Processing*, vol. 14, no. 3-4, pp. 162–472, 2021.

[45] M. Haenggi, *Stochastic geometry for wireless networks*. Cambridge University Press, 2012.

[46] R. Deng, B. Di, H. Zhang, Y. Tan, and L. Song, "Reconfigurable holographic surface-enabled multi-user wireless communications: Amplitude-controlled holographic beamforming," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6003–6017, 2022.

Qingchao Li (Graduate Student Member, IEEE) received the B.Eng. degree from the University of Science and Technology Beijing, Beijing, China, in 2017, the M.S. degree from the University of Science and Technology of China, Heifei, China, in 2020, and the Ph.D. degree in wireless communications from the University of Southampton, Southampton, U.K., in 2024 under the University of Southampton-Interdigital Communications Inc. Fully Funded Scholarship. He is currently a Research Fellow with the School of Electronics and Computer Science, University of Southampton. His research interests include reconfigurable intelligent surfaces, holographic MIMO and non-terrestrial wireless networks. He was the recipient of 2022 Chinese Government Award for Outstanding Self-Financed Student Abroad and 2023 Doctoral College Research Award, University of Southampton.

Mohammed El-Hajjar (Senior Member, IEEE) received the Ph.D. degree in wireless communications from the University of Southampton, Southampton, U.K., in 2008. He is currently a Professor with the School of Electronics and Computer Science, University of Southampton. Following the Ph.D. degree, he joined Imagination Technologies as a Design Engineer, where he worked on designing and developing Imagination's multi-standard communications platform, which resulted in three patents. He has authored or coauthored a Wiley-IEEE book and in excess of 150 journal and conference papers. His research is funded by the U.K. government and many industrial collaborators. His research interests include the development of intelligent communications systems, energy-efficient transceiver design, millimeter wave communications, and radio over fiber network design. He was the recipient of several academic awards.

Chao Xu (Senior Member, IEEE) received the B.Eng. degree in telecommunications from the Beijing University of Posts and Telecommunications, Beijing, China, the B.Sc. (Eng.) degree (with First Class Hons.) in telecommunications from the Queen Mary, University of London, London, U.K., through a Sino-U.K. joint degree Program in 2008, and the M.Sc. degree (with distinction) in radio frequency communication systems and the Ph.D. degree in wireless communications from the University of Southampton, Southampton, U.K., in 2009 and 2015, respectively. He is currently a Senior Lecturer with Next Generation Wireless Research Group, University of Southampton. His research interests include index modulation, reconfigurable intelligent surfaces, noncoherent detection, and turbo detection. He was the recipient of the Best M.Sc. Student in Broadband and Mobile Communication Networks by IEEE Communications Society (United Kingdom and Republic of Ireland Chapter) in 2009. He was also the recipient of 2012 Chinese Government Award for Outstanding Self-Financed Student Abroad and 2017 Dean's Award, Faculty of Physical Sciences and Engineering, the University of Southampton, and Marie Sklodowska-Curie Actions (MSCA) Global Postdoctoral Fellowships with the highest evaluation score of 100/100 in 2023.

Jiancheng An (Member, IEEE) received the B.S. degree in Electronics and Information Engineering and the Ph.D. degree in Information and Communication Engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2016 and 2021, respectively. From 2019 to 2020, he was a Visiting Scholar with the Next-Generation Wireless Group, University of Southampton, U.K. From 2021 to 2023, he was with the Engineering Product Development (EPD) Pillar, Singapore University of Technology and Design. He is currently a Research Fellow with the School of Electrical and Electronics Engineering, Nanyang Technological University (NTU), Singapore. His research interests include stacked intelligent metasurfaces (SIM), wave-based computing, and near-field communications.

Chau Yuen (S'02-M'06-SM'12-F'21) received the B.Eng. and Ph.D. degrees from Nanyang Technological University, Singapore, in 2000 and 2004, respectively. He was a Post-Doctoral Fellow with Lucent Technologies Bell Labs, Murray Hill, in 2005. Since 2023, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, currently he is Provosts Chair in Wireless Communications. Dr. Yuen received IEEE Communications Society Leonard G. Abraham Prize (2024), IEEE Communications Society Best Tutorial Paper Award (2024), IEEE Communications Society Fred W. Ellersick Prize (2023), IEEE Marconi Prize Paper Award in Wireless Communications (2021), and IEEE APB Outstanding Paper Award (2023). He is a Distinguished Lecturer of IEEE Vehicular Technology Society, Top 2% Scientists by Stanford University, and also a Highly Cited Researcher by Clarivate Web of Science. He has 3 US patents and published over 400 research papers at international journals.

Lajos Hanzo (FIEEE'04) received Honorary Doctorates from the Technical University of Budapest (2009) and Edinburgh University (2015). He is a Foreign Member of the Hungarian Science-Academy, Fellow of the Royal Academy of Engineering (FREng), of the IET, of EURASIP and holds the IEEE Eric Sumner Technical Field Award. For further details please see http://www-mobile.ecs.soton.ac.uk, https://en.wikipedia.org/wiki/Lajos_Hanzo.