Vision-Aided Positioning and Beam Focusing for 6G Terahertz Communications

Seungnyun Kim, *Member, IEEE*, Jihoon Moon, *Member, IEEE*, Jiao Wu, *Member, IEEE*, Byonghyo Shim, *Senior Member, IEEE*, and Moe Z. Win, *Fellow, IEEE*

***Abstract*— To meet the ever-increasing data rate demand expected in 6G networks, terahertz (THz) ultra-massive (UM) multiple-input multiple-output (MIMO) systems have gained much attention recently. One notable aspect of these systems is that the deployment of an extremely large-scale antenna array and high transmission frequency result in an expansion of the near-field region where the electromagnetic (EM) radiation is modeled as a spherical wave. In the near-field region, the channel becomes a function of a position of a user equipment (UE) rather than the direction, giving rise to a beam focusing operation that focuses the signal power onto the specific position. However, the traditional approaches relying on the sweeping of discretized beam codewords cannot support this ultra-sharp beam focusing operation in THz UM-MIMO systems. This paper proposes a novel beam focusing technique based on sensing and computer vision (CV) technologies. The essence of the proposed scheme is to estimate the UE's position from the vision information using the CV technique and then generates the beam heading towards the estimated position. By replacing the discretized and time-consuming beam sweeping operation with a highly precise CV-based positioning, the positioning accuracy as well as the beam focusing gain can be improved significantly. Numerical results show that the proposed scheme achieves significant positioning accuracy and data rate gains over the conventional codebook-based beam focusing schemes.**

***Index Terms*— 6G, terahertz, near-field, positioning, beam focusing, computer vision.**

I. Introduction

TERAHERTZ (THz) communication has received much attention as a key enabling technology to support a wide range of data-demanding applications for 6G [1]. By exploiting the abundant frequency spectrum resource in the terahertz (THz) bands ($0.1-10$ THz), THz communications can support truly immersive services such as digital twin, metaverse realized by extended reality (XR) devices, and high-fidelity holograms. As the operating frequency increases, the beamforming technique realized by array of multiple antennas becomes more essential to compensate for the severe attenuation of signal power caused by propagation, reflection, diffuse scattering, and atmospheric absorption losses [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. Note that the goal of the beamforming is to control the phase (and/or amplitude) of the signal transmitted at each antenna such that the difference in phase delays of signals are compensated. Since the beamforming gain is maximized only when the beamforming vector is properly aligned with the array steering vector (i.e., a set of phase delays of antenna elements) of signal propagation paths, the base station (BS) needs to acquire information on the signal radiation pattern.

In conventional microwave or millimeter wave (mmWave) systems, the radiation pattern of a signal is a function of elevation angle θ and azimuth angle φ. This is because the array aperture (usually on the order of centimeters) is much smaller than the communication distance (e.g., the mmWave microcell coverage is 500 m) so the electromagnetic (EM) radiation can be readily approximated as the plane wave [14]. In this so-called far-field region, the array steering vector is expressed as a function of azimuth and elevation angles so that one should consider the *far-field beam steering*, an approach to focus the signal power onto the specific direction towards a user equipment (UE). In the THz systems equipped with hundreds of antennas, however, the plane wave approximation might not be effective due to the increased array aperture (e.g., 1.5 m in 1024-antenna systems operating at 0.1 THz band) and the reduced communication distance (e.g., a few tens of meters) [15]. This implies that in ultra-massive (UM)-multiple-input multiple-output (MIMO) THz systems, the EM radiation is performed through spherical waves. In this so-called near-field region, the array steering vector depends on the angles (θ, φ) as well as the distance r. Thus, a new type of beamforming operation called *near-field beam focusing* that focuses the signal power towards the specific position of the UE is needed (see Fig. 1) [16], [17].

For the focused beam generation, a codebook-based approach has been widely used [18], [19], [20], [21], [22], [23], [24], [25], [26], [27]. In this approach, the BS transmits the sequence of pre-defined beam codewords carrying pilot signals, such as the synchronization signal block (SSB) and the channel state information-reference signal (CSI-RS). In the UE, an index of the beam codeword corresponding to the largest reference signal received power (RSRP) is sent to the BS. For example, in 5G New Radio (NR), a two-dimensional (2D) discrete Fourier transform (DFT)

Manuscript received 14 November 2023; revised 20 March 2024; accepted 18 April 2024. Date of publication 24 June 2024; date of current version 21 August 2024. The fundamental research described in this paper was supported, in part, by the National Research Foundation of Korea under Grant RS-2023-00252789 and Grant RS-2023-00208985, by the National Science Foundation under Grant CNS-2148251, and by the federal agency and industry partners in the RINGS program. *(Corresponding author: Moe Z. Win.)*

Seungnyun Kim is with the Wireless Information and Network Sciences Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: snkim94@mit.edu).

Jihoon Moon and Byonghyo Shim are with the Institute of New Media and Communications, Seoul National University, Seoul 08826, Republic of Korea (e-mail: jhmoon@islab.snu.ac.kr; bshim@snu.ac.kr).

Jiao Wu is with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia (e-mail: jiao.wu@kaust.edu.sa).

Moe Z. Win is with the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: moewin@mit.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2024.3413949.

Digital Object Identifier 10.1109/JSAC.2024.3413949

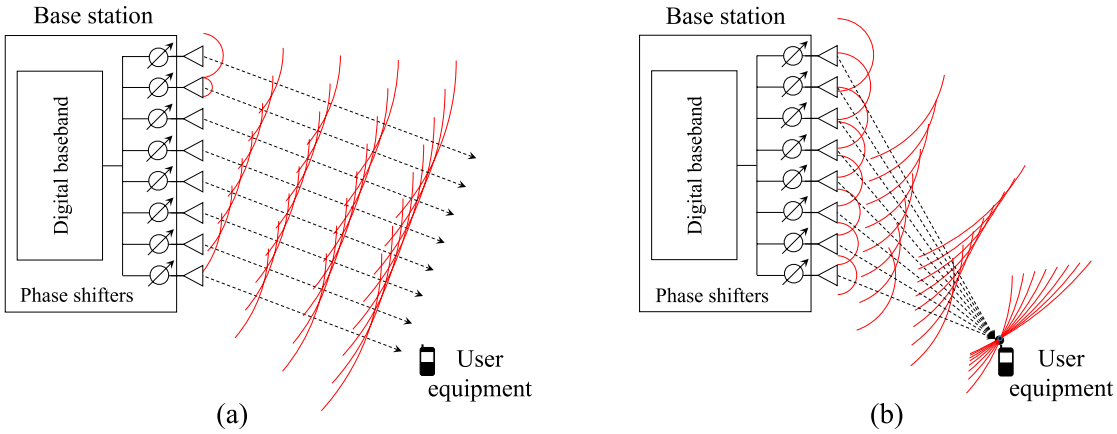

Fig. 1. Comparisons of (a) the far-field beam steering and (b) the near-field beam focusing.

matrix-based beam codebook is used [18], [19]. Recently, advanced codebook-based beam focusing schemes have been proposed [20], [21], [22], [23], [24], [25], [26], [27]. In [20], [21], and [22], hierarchical beam codebook designs for the mmWave MIMO systems have been proposed. In [23] and [24], beam training schemes exploiting the beam squint effect of wideband THz near-field systems have been proposed. In [25] and [26], beam codebook designs for reconfigurable intelligent surfaces (RIS)-assisted systems have been proposed. Also, in [27], a deep learning (DL)-based beam training scheme for UM-MIMO systems has been proposed.

The major shortcoming of the codebook-based beam focusing techniques is the mismatch between the pre-defined beam codeword and the desired beam focusing vector directed towards the UE's position (i.e., beam discretization error) which causes a significant degradation of the beam focusing gain [28]. The beam discretization error as well as its impact on the beam focusing gain will become even more pronounced in the THz near-field systems since, in this case, the beam search space extends to three-dimensional (3D) space represented by the distance as well as the azimuth and elevation angles [15]. Frequent transmission of beams to mitigate the beam discretization error will increase the resource overhead, latency, and power consumption.

Two fundamental questions for the design of the near-field beam focusing technique are as in the following:

- how to accurately identify the positions of UEs while minimizing the beam training overhead; and
- how to design the near-field beam focusing vectors maximizing the sum-rate using the estimated positions?

The answers to these questions will not only lead to the improved signal strength and enhanced interference mitigation but also ensure the precise targeting and spectral efficiency maximization in THz UM-MIMO systems, thereby realizing the data-intensive applications envisioned for 6G.

The aim of this paper is to propose a novel near-field beam focusing framework based on the sensing and computer vision (CV) technologies for THz UM-MIMO systems. Our approach is justified by two crucial observations: 1) in line with the recent trend of communication area being close to the human visual area (a few tens of meters) due to the use of high-frequency bands, sensing technologies that observe the surrounding environments through various sensing modalities (e.g., red-green-blue (RGB) camera, depth camera, and radar) have gained considerable attention [29], [30], [31], [32]; and 2) CV technique analyzing the sensing information has made a significant advancement in performing the object classification, detection, and tracking from raw images with aid of DL [33], [34]. Motivated by these, in the proposed technique called *vision-aided beam focusing* (VBF), we estimate the position of a UE from the vision information using the CV technique and then generate the beam heading towards the estimated position. Since the beam focusing vector is generated directly from the position extracted from the image, VBF is free from the beam discretization error, thereby achieving the maximization of beam focusing gain. Also, by minimizing the complicated handshaking operations (i.e., pilot transmission and channel feedback) between the BS and UE, the beam training overhead is reduced substantially. As a main DL engine, we use Transformer, a DL model specialized for extracting the temporally and spatially correlated features [35]. Using the attention mechanism quantifying the correlations between the input and outputs values, Transformer assigns relatively large weights to the input values (e.g., pixel values) which are more relevant to the output values (e.g., target objects), thereby facilitating the feature extraction of the UE.

The key contributions of this paper can be summarized as in the following:

- we develop a near-field beam focusing framework for THz UM-MIMO systems that utilizes the CV technique for UE positioning;
- we present a position-aware hybrid analog-digital beam focusing technique maximizing the system throughput of THz UM-MIMO systems; and
- we demonstrate through extensive simulations that VBF significantly improves positioning accuracy and system throughput.

The rest of this paper is organized as in the following. Section II presents the THz near-field systems and then reviews the conventional codebook-based schemes. Section III explains the vision-aided UE positioning. Section IV presents a position-aware near-field beam focusing technique. Section V discusses the practical implementation issues of VBF.

Section VI presents the simulation results. Section VII concludes the paper.

Notations: Random variables are displayed in sans serif, upright fonts; their realizations in serif, italic fonts. Vectors and matrices are denoted by bold lowercase and uppercase letters, respectively. For example, a random variable and its realization are denoted by x and x for scalars, $\mathbf{x}$ and $\boldsymbol{x}$ for vectors, and $\mathbf{X}$ and $\boldsymbol{X}$ for matrices. Sets and random sets are denoted by upright sans serif and calligraphic font, respectively. For example, a random set and its realization are denoted by X and $\mathcal{X}$, respectively. The m-by-n matrix of zeros is denoted by $\mathbf{0}_{m\times n}$; when $n=1$, the m-dimensional vector of zeros is simply denoted by $\mathbf{0}_m$. The m-by-m identity matrix is denoted by $\boldsymbol{I}_m$. The operators $\mathrm{tr}(\boldsymbol{x})$, $\|\boldsymbol{x}\|_2$, and $\|\boldsymbol{X}\|_{\mathrm{F}}$ denote the trace, the Euclidean norm, and the Frobenius norm, respectively. The operations $\otimes$ and $\odot$ denote the Kronecker product and element-wise product, respectively. The ith row and jth column of $\boldsymbol{X}$ is denoted by $[\boldsymbol{X}]_{i,j}$. The transpose, conjugate, and conjugate transpose of $\boldsymbol{X}$ are denoted by $(\cdot)^{\mathrm{T}}$, $(\cdot)^*$, and $(\cdot)^{\mathrm{H}}$, respectively. The real part of a complex number is denoted by $\Re\{\cdot\}$. The notation $\mathrm{diag}(\cdot)$ represents a diagonal matrix with the arguments being its diagonal elements.

II. THz Near-Field UM-MIMO Systems

In this section, we present the THz UM-MIMO system model and then discuss the THz near-field line-of-sight (LOS) channel model. We also provide a brief overview of the conventional codebook-based beam focusing schemes.

A. THz Near-Field UM-MIMO System Model

We consider a multi-user THz UM-multiple-input single-output (MISO) system where a BS equipped with a uniform planar array (UPA) of $N = N_{\mathrm{h}} \times N_{\mathrm{v}}$ antennas serves K single-antenna UEs. The set of UEs is denoted as $\mathcal{K} = \{1, 2, \ldots, K\}$. In our work, we consider the 3D coordinate systems where the $(0,0)$th antenna is located at the origin and the antenna array is located at the XZ-plane. Note that the (m,n)th antenna denotes the antenna element located at mth row and nth column of the antenna plane. The RGB-depth (RGB-d) camera is attached at the BS to identify the wireless environments. To reduce the hardware complexity, we consider a hybrid analog-digital architecture with $N_{\mathrm{RF}} = K$ radio-frequency (RF) chains,[1] each of which is connected with N phase shifters. Specifically, the hybrid beam focusing vector for the kth UE is expressed as $\boldsymbol{f}_k = \boldsymbol{F}_{\mathrm{RF}}\boldsymbol{f}_{\mathrm{BB},k} \in \mathbb{C}^N$ where $\boldsymbol{F}_{\mathrm{RF}} \in \mathbb{C}^{N\times K}$ is the analog RF beam focusing matrix and $\boldsymbol{f}_{\mathrm{BB},k} \in \mathbb{C}^K$ is the digital beam focusing vector for the kth UE. The set of feasible RF beam focusing matrices is denoted as $\mathcal{F}_{\mathrm{RF}} \triangleq \big\{\boldsymbol{F}_{\mathrm{RF}} \in \mathbb{C}^{N\times K} \mid [\boldsymbol{F}_{\mathrm{RF}}]_{n,k} = e^{jw_{n,k}}, w_{n,k} \in \Theta\big\}$ where Θ is the set of feasible phase shifts.[2] By combining the beam focusing vectors of K UEs, we obtain the hybrid beam focusing matrix $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_K] = \boldsymbol{F}_{\mathrm{RF}}[\boldsymbol{f}_{\mathrm{BB},1}, \boldsymbol{f}_{\mathrm{BB},2}, \ldots, \boldsymbol{f}_{\mathrm{BB},K}] = \boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}} \in \mathbb{C}^{N\times N_{\mathrm{RF}}}$ where $\boldsymbol{F}_{\mathrm{BB}} = [\boldsymbol{f}_{\mathrm{BB},1}, \boldsymbol{f}_{\mathrm{BB},2}, \ldots, \boldsymbol{f}_{\mathrm{BB},K}] \in \mathbb{C}^{K\times K}$ is the digital beam focusing matrix. Note that $\boldsymbol{F}$ is bounded by the BS transmission power P_{tx} as $\|\boldsymbol{F}\|_{\mathrm{F}}^2 = \|\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\|_{\mathrm{F}}^2 \leqslant P_{\mathrm{tx}}$.

The received signal $\mathsf{y}_k \in \mathbb{C}$ of the kth UE is given by

$$\mathsf{y}_k = \boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{f}_{\mathrm{BB},k}\mathsf{s}_k + \sum_{j\neq k}\boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{f}_{\mathrm{BB},j}\mathsf{s}_j + \mathsf{n}_k \tag{1}$$

where $\boldsymbol{h}_k \in \mathbb{C}^N$ is the downlink channel vector from the BS to the kth UE, s_k is the data symbol intended for the kth UE such that $\mathbb{E}\{|s_k|\} = 1$, and $\mathsf{n}_k \sim \mathcal{CN}(0, \sigma_{\mathrm{n}}^2)$ is the Gaussian noise. Then, the achievable rate of the kth UE is

$$R_k = \log_2\big(1 + \gamma_k(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})\big) \tag{2}$$

where $\gamma_k(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})$ is the downlink signal-to-interference-plus-noise ratio (SINR) of the kth UE defined as

$$\gamma_k(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}) \triangleq \frac{\left|\boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{f}_{\mathrm{BB},k}\right|^2}{\sum_{j\neq k}\left|\boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{f}_{\mathrm{BB},j}\right|^2 + \sigma_{\mathrm{n}}^2}. \tag{3}$$

B. THz Near-Field LOS Channel Model

In THz systems, due to the high directivity and path loss of THz band signal, the scattering and refraction of signal are negligible so the LOS path becomes the dominant means of propagation [12]. Indeed, the number of propagation paths in THz band is less than 4 [12], [37]. Moreover, the power gap between the LOS and non-line-of-sight (NLOS) path signals is significant due to the huge reflection and diffuse scattering losses.[3] For example, in the 0.4 THz band, the Rician K-factor, a ratio of the powers of the LoS component to the diffuse component, is around 20 dB [12].

Another key aspect of THz UM-MIMO channel is the near-field characteristics. In general, the EM radiation field can be divided into two regions: 1) far-field region where the EM radiation can be approximated as the plane waves and 2) near-field region where the EM radiation is modeled as the spherical waves. To distinguish these regions, the Fraunhofer distance $Z \triangleq \frac{N^2c}{2f}$ is widely used where f is the signal frequency and c is the speed of light [15]. While Z is typically a few meters in the traditional systems, it can reach up to a hundred meters in the THz UM-MIMO systems due to the large number of antennas. In the near-field region, due to the spherical wavefront, the phase delay between two antenna elements is affected by the spherical coordinates (r, θ, φ), meaning that the near-field array steering vector is a function of (r, θ, φ) [38].

Based on these observations, we use the THz near-field LOS channel model where the downlink channel vector $\boldsymbol{h}_k \in \mathbb{C}^N$

[1] When the number of RF chains N_{RF} is smaller than the number of UEs K (i.e., $N_{\mathrm{RF}} < K$), the BS cannot support all UEs simultaneously since the number of data streams is limited by the number of RF chains. In this case, to accommodate all UEs, one can use a user grouping strategy that segments the UEs into several groups and then serves these groups sequentially [36]. When N_{RF} is greater than K (i.e., $N_{\mathrm{RF}} > K$), the BS can allocate multiple data streams to each UE. Note that, even in this case, the proposed scheme can be used to identify the optimal hybrid beam focusing matrix with minor modifications.

[2] For example, Θ is $[0, 2\pi)$ in case of analog phase shifter with infinite phase shift levels and Θ is $\big\{\frac{2\pi b}{2^B} \mid b = 0, 1, \ldots, 2^B - 1\big\}$ in case of B-bit digital phase shifter with 2^B phase shift levels.

[3] In the THz band, the wavelength (e.g., 100 μm in the 3 THz band) is smaller than the surface roughness of objects (e.g., the roughness of concrete wall is $300 - 1000\,\mu$m) so the diffuse scattering is the dominant means of reflection. Since the reflected signal is scattered over an area, the power of NLOS path signal is much smaller than that of the LOS path signal.

from the BS to the kth UE is expressed as [13], [39]

$$\boldsymbol{h}_k = \sqrt{\alpha(f, r_k)}\, e^{-j\frac{2\pi f}{c} r_k} \boldsymbol{a}(r_k, \theta_k, \varphi_k) \tag{4}$$

where r_k is the distance between the $(0,0)$th BS antenna and the kth UE, θ_k and φ_k are the elevation and azimuth angles of departure (AODs), respectively, and $\alpha(f, r_k) = G_{\text{free}}(f, r_k) G_{\text{abs}}(f, r_k)$ is the path gain consisting of the free-space path loss $G_{\text{free}}(f, r_k) = (\frac{c}{4\pi r_k f})^2$ and the molecular absorption $G_{\text{abs}}(f, r_k) = e^{-k(f) r_k}$ with $k(f)$ being the absorption coefficient [40]. Also, $\boldsymbol{a}(r_k, \theta_k, \varphi_k) \in \mathbb{C}^N$ is the near-field UPA array steering vector whose (m,n)th element is

$$[\boldsymbol{a}(r_k, \theta_k, \varphi_k)]_{m,n} = e^{-j\frac{2\pi f}{c}\left(r_k^{(m,n)} - r_k\right)} \tag{5}$$

for $m = 1, 2, \ldots, N_{\text{h}}$ and $n = 1, 2, \ldots, N_{\text{v}}$ where $r_k^{(m,n)}$ is the distance from the (m,n)th BS antenna to the kth UE.

Lemma 1: The distance $r_k^{(m,n)}$ between the (m,n)th BS antenna and the kth UE can be approximated as a function of the UE position $(r_k, \theta_k, \varphi_k)$ as

$$\begin{aligned} r_k^{(m,n)} \approx\ & r_k - d\big((m-1)\sin\theta_k \cos\varphi_k + (n-1)\cos\theta_k\big) \\ & + \frac{d^2}{2r_k}\big((m-1)^2 + (n-1)^2 \\ & - \big((m-1)\sin\theta_k\cos\varphi_k + (n-1)\cos\theta_k\big)^2\big). \end{aligned} \tag{6}$$

Proof: See Appendix A. □

Note that $\boldsymbol{h}_k$ can be readily expressed as a function of $(r_k, \theta_k, \varphi_k)$. Thus, to generate the optimal beam focusing matrix $\boldsymbol{F}^{\text{opt}}$ maximizing the sum-rate, an accurate UE positioning is imperative [41], [42]. For the UE positioning, techniques relying on the time-based measurements (e.g., time-of-arrival (TOA)) and angle-based measurements (e.g., angle of arrival (AOA)) have been widely used [43], [44], [45], [46], [47], [48], [49], [50], [51]. For example, in 5G NR, a new reference signal known as positioning reference signal (PRS) is introduced to measure TOA and AOA [52], [53]. However, due to the limited wireless resources (e.g., bandwidth), these techniques achieve only meter-level positioning accuracy.

C. Conventional Codebook-Based Beam Focusing

To determine the beam focusing vector, codebook-based approaches have been proposed [20], [21], [22], [23], [24], [25], [26], [27]. The codebook-based approach consists of two major steps: 1) beam sweeping where a BS transmits a sequence of beam codewords chosen from the B-bit codebook $\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_{2^B}\}$ and 2) beam selection where each kth UE feeds back the index $\hat{\mathsf{i}}_k$ of the best beam codeword $\boldsymbol{c}_{\hat{\mathsf{i}}_k}$ maximizing the RSRP to the BS:

$$\hat{\mathsf{i}}_k = \arg \max_{i=1,2,\ldots,2^B} \left|\boldsymbol{h}_k^{\text{H}} \boldsymbol{c}_i + \mathsf{n}_{k,i}\right|^2 \tag{7}$$

where $\mathsf{n}_{k,i}$ is the additive noise. Since the beam focusing vector is chosen from the beam codebook, the performance of the codebook-based schemes depends heavily on the beam codebook design. In conventional far-field systems, the channel is a function of the azimuth angle φ and elevation angle θ so the beam codebook is designed to sample the 2D angular space. In the near-field systems, however, the channel is a function of the spherical coordinates (r, θ, φ), and thus the beam search space expands to the 3D space. For instance, the near-field beam codebook scheme in [25] first generates uniformly sampled Cartesian coordinates $\Xi_{\text{car}} = \left\{(\bar{x}_i, \bar{y}_i, \bar{z}_i) \mid i = 1, 2, \ldots, 2^B\right\}$ and then converts them to the spherical coordinates $\Xi_{\text{sph}} = \left\{(\bar{r}_i, \bar{\theta}_i, \bar{\varphi}_i) \mid i = 1, 2, \ldots, 2^B\right\}$. Then, the near-field beam codebook $\mathcal{C}_{\text{near}}$ is obtained from the near-field array steering vectors corresponding to the spherical coordinates in Ξ_{sph} as $\mathcal{C}_{\text{near}} = \left\{\boldsymbol{a}(\bar{r}_i, \bar{\theta}_i, \bar{\varphi}_i) \mid (\bar{r}_i, \bar{\theta}_i, \bar{\varphi}_i) \in \Xi_{\text{sph}}\right\}$.

One major issue of the codebook-based beam focusing scheme is the mismatch between the pre-defined beam direction and the real UE direction. For example, when we use 6-bit DFT-based beam codebook in the near-field 256-antenna systems, the beam focusing gain degradation in the worst-case is around 30%. Another serious problem is the latency caused by the complicated handshaking process between the BS and the UE. The beam sweeping latency is expected to increase even further in the 6G near-field THz systems since the search space is expended to the 3D space.

III. Vision-Aided UE Positioning

The main goal of VBF is to extract the geometric information (e.g., position and class) of a UE from the RGB-d image and then utilize it for the THz focused beam generation. To do so, we use the object detection, a CV technique specialized for detecting instances of semantic objects in a certain class (e.g., humans, cars, or UEs). By replacing the discretized and time-consuming beam sweeping operation with the precise CV-based positioning, the positioning accuracy can be improved significantly, resulting in an enhancement of the beam focusing gain. Also, since the UE location is acquired directly using the object detector without pilot transmission and channel feedback operations, the beam training overhead such as resource overhead, power consumption, and latency can be reduced substantially. The proposed vision-aided UE positioning consists of three major steps (see Fig. 2):

- **Vision information acquisition**: The BS acquires the rough estimate of the UE's location from the SSB beam index feedback and then the RGB-d camera captures the image of the area covered by the SSB beam index.
- **Transformer-based object detection**: From the captured RGB image, the Transformer-based object detector extracts the 2D pixel-wise coordinates $(\hat{x}_{\text{px}}, \hat{y}_{\text{px}})$ of the UE from the captured image.
- **Coordinate transformation from image to real world**: the BS converts the 2D pixel-wise coordinates $(\hat{x}_{\text{px}}, \hat{y}_{\text{px}})$ in the image to the elevation/azimuth angles (θ, φ) in the real world and then acquires the distance r using the LiDAR sensor of RGB-d camera.

A. Vision Information Acquisition

To extract the UE's position from the image, the sensing direction should be determined properly so that the captured image contains the UE. To do so, the BS can leverage the SSB beam index obtained at the initial access process. During the initial access, the BS transmits the set of SSB beam codewords, each of which covers a relatively wide physical area [54], [55]. Then the UE measures the RSRP of the SSB

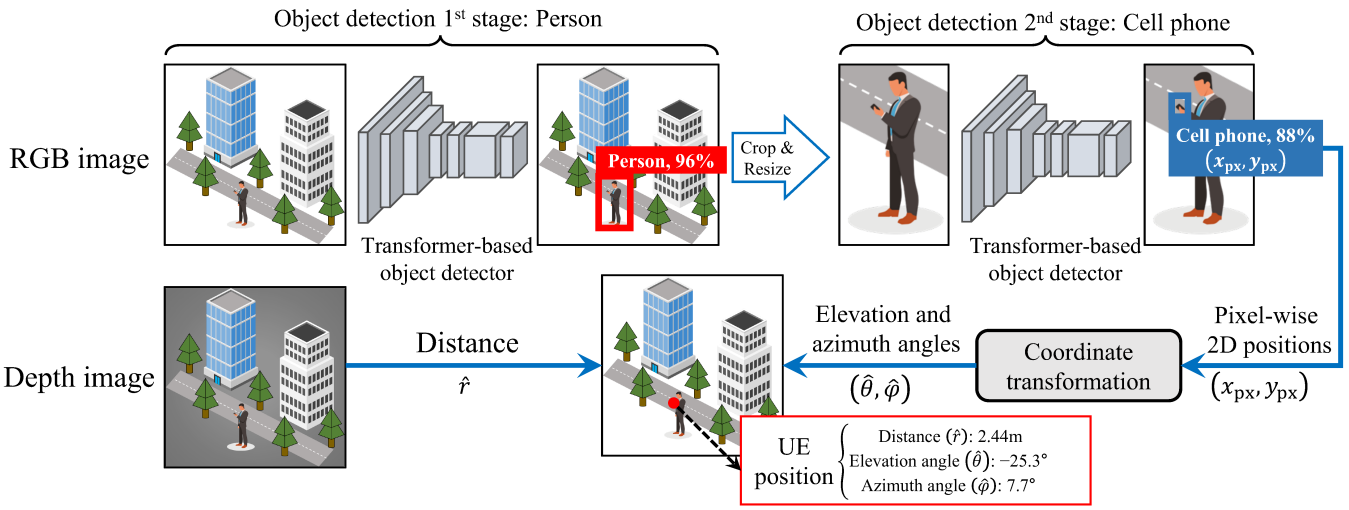

Fig. 2. Illustration of vision-aided UE positioning.

beams and then feeds back the index of the SSB beam corresponding to the largest RSRP. Therefore, the BS can acquire a rough estimate of the UE's position using the SSB beam index. Based on this observation, in VBF, the RGB-d camera takes a shot for the area covered by the SSB beam index.[4]

B. Transformer-Based Object Detection

Once the image is acquired, the BS identifies the position of the UE from the image using the object detection technique. In the object detection, the deep neural network (DNN) learns the end-to-end mapping between the input $W \times H$-pixel RGB image $\mathcal{D}_{\text{RGB}} \in \mathbb{R}^{W \times H \times 3}$ and the geometric information, i.e., object class $\hat{c}_{\text{class}}$ and 2D pixel-wise coordinates $(\hat{x}_{\text{px}}, \hat{y}_{\text{px}})$ of the centroid of the bounding box. The object detection problem to find out the mapping function g is formulated as

$$\left(\hat{x}_{\text{px}},\, \hat{y}_{\text{px}},\, \hat{c}_{\text{class}}\right) = g\left(\mathcal{D}_{\text{RGB}}; \boldsymbol{\eta}\right) \tag{8}$$

where $\boldsymbol{\eta}$ denote the DNN parameters.

For the object detection, convolutional neural network (CNN) architectures have been popularly used due to its simplicity and ability to extract spatial features from the visual information [56]. In the CNN architecture, the features are extracted by performing the convolution operation of a weight matrix (called kernel) and a part of the input image. While the CNN architecture is effective in extracting local features to some extent, it is not that efficient in extracting global features due to the locality of the filter kernel. Recently, there has been considerable interest in object detection techniques utilizing a DL architecture known as the *Transformer* [35]. The key ingredient of Transformer is the attention block that quantifies the correlations between the pixel values of the input image and then assigns varying degrees of importance (i.e., attention weight) to each pixel value based on the calculated

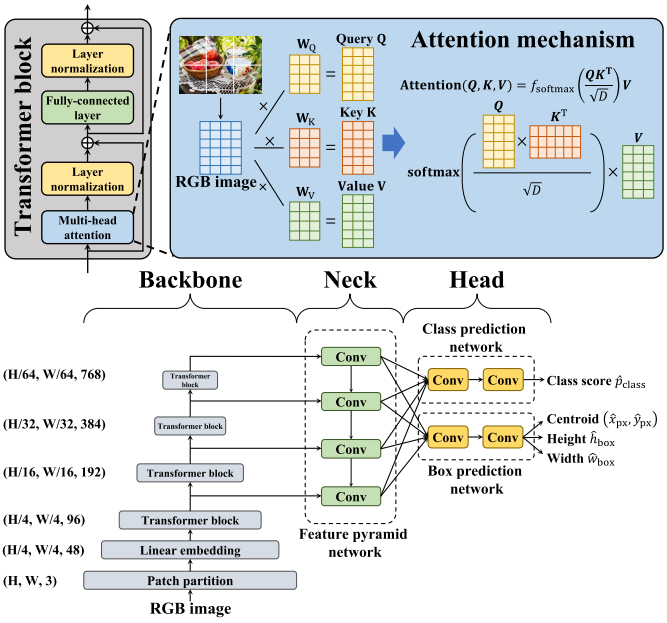

Fig. 3. Structure of Tranformer-based object detector.

correlation. Using the attention mechanism, Transformer can extract both the correlations of the adjacent pixels and those of the spaced-apart pixels, thereby generating the local and global features in the image. For these reasons, the object detection techniques using Transformer are achieving the state-of-the-art (SOTA) performances these days (e.g., DETR [57] and Swin Transformer [58]).

1) Basics of Transformer: In Transformer, the input is the RGB image consisting of pixel values and the output is the spatially-correlated features of the input image (see Fig. 3). First, the input image passes through multiple attention blocks connected in parallel (i.e., multi-head attention). After the multi-head attention layer, the pixel values of the input images are scaled and shifted to have zero mean and unit variance (this process is called the layer normalization) [35]. Using the layer normalization process, one can enforce the input distribution to have fixed means and variances and therefore, increase the stability of network training.

As mentioned, the role of the attention block is to quantify the correlation between the pixel values (i.e., *attention score*)

[4]To acquire the sensing information of multiple UEs simultaneously, one can use a circular array of multiple cameras, each of which covering a specific angular sector. In this system, when the UE reports its SSB beam index to the BS, the camera covering the area designated by the reported SSB beam index captures the sensing information. Using this circular camera array, the BS can simultaneously acquire sensing information of multiple UEs without rotating the camera. For example, if the BS is equipped with a circular array of 8 cameras and the number of SSBs is 64, then each camera would cover an angular sector of $\frac{360}{8} = 45°$, encompassing coverage of $\frac{64}{8} = 8$ consecutive SSB beams.

and then generate the weighted pixel values by multiplying the original pixel values by the obtained attention score. To do so, the attention block constructs three different embedding matrices from the input image matrix $\boldsymbol{X}$, i.e., the query $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_{\mathrm{Q}}$, the key $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_{\mathrm{K}}$, and the value $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_{\mathrm{V}}$ where $\boldsymbol{W}_{\mathrm{Q}}$, $\boldsymbol{W}_{\mathrm{K}}$, $\boldsymbol{W}_{\mathrm{V}}$ are the weight matrices (see Fig. 3). Since the query $\boldsymbol{Q}$ and the key $\boldsymbol{K}$ contain the features of original input image, by performing the inner product of $\boldsymbol{Q}$ and $\boldsymbol{K}$, we can obtain the attention score $\boldsymbol{M}$:

$$\boldsymbol{M} = f_{\mathrm{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{D}}\right) \tag{9}$$

where D is the number of hidden layer units and $f_{\mathrm{softmax}}(\boldsymbol{X})$ is the row-wise softmax function defined as $[f_{\mathrm{softmax}}(\boldsymbol{X})]_{i,j} = \frac{e^{[\boldsymbol{X}]_{i,j}}}{\sum_j e^{[\boldsymbol{X}]_{i,j}}}$. Finally, by multiplying the attention score $\boldsymbol{M}$ by the value $\boldsymbol{V}$, we obtain the weighted input as

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \triangleq \boldsymbol{M}\boldsymbol{V}\,. \tag{10}$$

The obtained weighted input passes through the fully-connected layer and the layer normalization layer again, generating the features of the input image.

2) Transformer-Based Object Detection: The Transformer-based object detector consists of three main components [58] (see Fig. 3): 1) *backbone* extracting the features from the image, 2) *neck* aggregating the extracted multi-scale features, and 3) *head* performing the object detection and classification.

- **Backbone**: The backbone consisting of multiple Transformer blocks hierarchically extracts the spatially-correlated features (e.g., color, shape) in different scales from the input RGB image $\mathcal{D}_{\mathrm{RGB}}$. For example, the local features (e.g., edge and curve) are extracted at the bottom of backbone and the global features (e.g., face and wall) are extracted at the top of backbone.
- **Neck**: Once the features are extracted at the backbone, the neck aggregates the extracted features in different scales. This process is necessary since the local features contain geometric information for the bounding box identification while the global features contain semantic information for the classification. Thus, to achieve high performance in both positioning and classification, a feature pyramid network (FPN) structure where the global features extracted at the top of backbone are delivered to the bottom is employed.
- **Head**: Using the aggregated features as input, the head performs the bounding box detection and the class identification. Specifically, the box prediction network generates the centroid pixel $(\hat{x}_{\mathrm{px}}, \hat{y}_{\mathrm{px}})$, the height $\hat{h}_{\mathrm{box}}$, and the width $\hat{w}_{\mathrm{box}}$ of the bounding box. Also, the class prediction network generates the class score $\hat{p}_{\mathrm{class}}(c)$ (probability of object belonging to the specific class c) and then choose the class with the highest class score:

$$\hat{c}_{\mathrm{class}} = \arg\max_{c}\ \hat{p}_{\mathrm{class}}(c)\,. \tag{11}$$

It is worth mentioning that the performance of the Transformer-based object detector depends heavily on the size of the target object. When the target object (in our case, a cell phone) is small, only a few pixels would represent the target object so that it is not easy to find out the pixels representing the target object in the captured image. To address this issue, we use a two-stage object detection process that first detects a large object (e.g., person holding a UE) from the whole image and then identifies a small object (e.g., UE) from the detected bounding box containing a person (see Fig. 2).

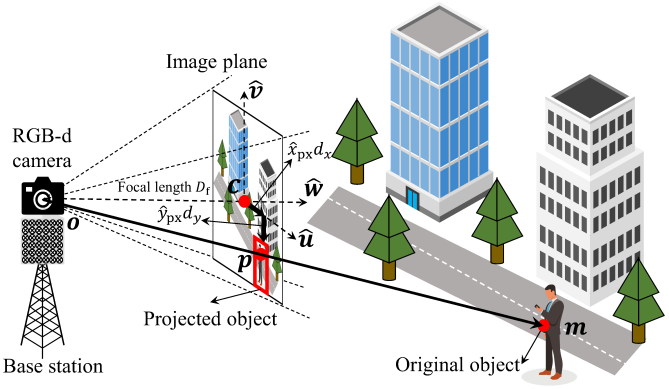

Fig. 4. Illustration of the coordinate transformation.

3) Loss Function Design and Network Training: To assess the quality of the object detection model, we use the weighted sum $L \triangleq \lambda_{\mathrm{p}} L_{\mathrm{p}} + \lambda_{\mathrm{c}} L_{\mathrm{c}}$ of two training losses for bounding box prediction and class prediction where λ_{p} and λ_{c} are the weights. First, the mean absolute error (MAE) L_{p} evaluating the positioning error of the bounding box is given by[5]

$$\begin{aligned} L_{\mathrm{p}} \triangleq\ & |x_{\mathrm{px}} - \hat{x}_{\mathrm{px}}| + |y_{\mathrm{px}} - \hat{y}_{\mathrm{px}}| \\ & + |w_{\mathrm{box}} - \hat{w}_{\mathrm{box}}| + |h_{\mathrm{box}} - \hat{h}_{\mathrm{box}}|. \end{aligned} \tag{12}$$

Second, the negative log-likelihood loss L_{c} measuring the class score of the ground-truth class c is given by

$$L_{\mathrm{c}} \triangleq -\log \hat{p}_{\mathrm{class}}(c)\,. \tag{13}$$

Then the network parameter $\boldsymbol{\eta}$ is trained in a direction to minimize L using the gradient descent method.

C. Coordinate Transformation From Image to Real World

Transformer-based object detector provides only 2D pixel-wise coordinates $(\hat{x}_{\mathrm{px}}, \hat{y}_{\mathrm{px}})$ of the UE on the image plane but for the beam focusing operation, we need the spherical coordinates $(\hat{r}, \hat{\theta}, \hat{\varphi})$ in the real world space. For the acquisition of r, we use the RGB-d camera which measures the distance to the point in each pixel using the LiDAR sensor. Subsequently, we perform a coordinate transformation to convert $(\hat{x}_{\mathrm{px}}, \hat{y}_{\mathrm{px}})$ in the image plane to $(\hat{\theta}, \hat{\varphi})$ in the real world space. To do so, we exploit the fact that the object in an image is a projection of the real-world object onto the image plane (see Fig. 4).[6] Thus, by finding out the position vector of the projected object, we can extract the elevation and azimuth angles $(\hat{\theta}, \hat{\varphi})$.

[5]When detecting multiple objects, to match the ground-truth objects with the detected objects, the Transformer-based object detector uses the bipartite matching that finds out the optimal permutation σ^* of object indices minimizing the matching loss as $\sigma^* = \arg\min_{\sigma} \sum_i (-p_{\sigma(i)}(c_i) + \|\boldsymbol{b}_i - \boldsymbol{b}_{\sigma(i)}\|_1)$ where $\boldsymbol{b} = [x_{\mathrm{px}}, y_{\mathrm{px}}, w_{\mathrm{box}}, h_{\mathrm{box}}]$.

[6]In practice, due to the lens distortion (e.g., barrel distortion or pincushion distortion), the position vector of the UE and that of the centroid of the detected bounding box may differ, particularly at the boundary of image. However, note that most of modern cameras come equipped with built-in lens correction features, which automatically compensate for such distortions during image processing. In case the built-in lens correction is unavailable, one can manually correct lens distortion by utilizing lens correction software (e.g., DxO Optics), which is based on the mathematical models (e.g., radial distortion model and Brown-Conrady model) describing how pixels are shifted away from the center of the image.

Let us consider the Cartesian coordinate system where the RGB-d camera is located at the origin $\boldsymbol{o}$. Also, let $\boldsymbol{m} \in \mathbb{R}^3$ and $\boldsymbol{p} \in \mathcal{I}_{\text{img}}$ be the positions of the UE in the real world and the projected UE in the image plane $\mathcal{I}_{\text{img}} \subseteq \mathbb{R}^3$. Then $\boldsymbol{m}$ and $\boldsymbol{p}$ have the same orientation, i.e., $\boldsymbol{m} \parallel \boldsymbol{p}$, meaning that we can obtain $(\hat{\theta}, \hat{\varphi})$ by finding out $\boldsymbol{p}$. To do so, we use the triangular law of vector addition with the centroid $\boldsymbol{c}$ of $\mathcal{I}_{\text{img}}$:

$$\boldsymbol{p} = \boldsymbol{c} + (\boldsymbol{p} - \boldsymbol{c}) \tag{14}$$

First, $\boldsymbol{c}$ is obtained from the focal length D_{f} as $\|\boldsymbol{c}\|_2 = D_{\text{f}}$ and elevation/azimuth angles $(\theta_{\text{c}}, \varphi_{\text{c}})$ of RGB camera direction:[7]

$$\boldsymbol{c} = \left(D_{\text{f}} \sin\theta_{\text{c}} \cos\varphi_{\text{c}},\ D_{\text{f}} \sin\theta_{\text{c}} \sin\varphi_{\text{c}}, D_{\text{f}} \cos\theta_{\text{c}}\right). \tag{15}$$

Then we can re-express $\mathcal{I}_{\text{img}}$ as $\mathcal{I}_{\text{img}} = \{\boldsymbol{i} \in \mathbb{R}^3 \mid \boldsymbol{c} \perp (\boldsymbol{i} - \boldsymbol{c})\}$.

Second, to calculate $\boldsymbol{p} - \boldsymbol{c}$, we exploit the fact that $\mathcal{I}_{\text{img}}$ is a parallel translation of 2D subspace of the Euclidean space. To be specific, by defining the orthogonal bases $\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}} \in \mathbb{R}^3$ of the 2D subspace, $\mathcal{I}_{\text{img}}$ can be re-expressed as

$$\begin{aligned} \mathcal{I}_{\text{img}} &= \boldsymbol{c} + \{\boldsymbol{j} \in \mathbb{R}^3 \mid \boldsymbol{c} \perp \boldsymbol{j}\} && (16)\\ &= \boldsymbol{c} + \{x d_x \hat{\boldsymbol{u}} + y d_y \hat{\boldsymbol{v}} \mid x, y \in \mathbb{Z}\} && (17) \end{aligned}$$

where (x, y) are the 2D pixel-wise coordinates and $d_x \times d_y$ is the size of each pixel. Since $\boldsymbol{p} \in \mathcal{I}_{\text{img}}$ is the projected UE in the image, $\boldsymbol{p} - \boldsymbol{c}$ can be expressed as a function of $(\hat{x}_{\text{px}}, \hat{y}_{\text{px}})$ as

$$\boldsymbol{p} - \boldsymbol{c} = \hat{x}_{\text{px}} d_x \hat{\boldsymbol{u}} + \hat{y}_{\text{px}} d_y \hat{\boldsymbol{v}}. \tag{18}$$

Now, what remains is to calculate $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$. Let $\hat{\boldsymbol{w}}$ be the unit normal vector of $\mathcal{I}_{\text{img}}$ given by

$$\hat{\boldsymbol{w}} = \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2} = \left(\sin\theta_{\text{c}} \cos\varphi_{\text{c}},\ \sin\theta_{\text{c}} \sin\varphi_{\text{c}},\ \cos\theta_{\text{c}}\right). \tag{19}$$

Then $\{\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{w}}\}$ forms the left-handed Cartesian coordinate system. Since the image is not rotated, $\hat{\boldsymbol{u}}$ is parallel to the XY-plane, which means that $\hat{\boldsymbol{u}}$ is perpendicular to both $\hat{\boldsymbol{w}}$ and the z-axis unit vector $\hat{\boldsymbol{z}} = (0, 0, 1)$. Thus, $\hat{\boldsymbol{u}}$ can be obtained from the cross product of $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{z}}$ as

$$\hat{\boldsymbol{u}} = \frac{\hat{\boldsymbol{w}} \times \hat{\boldsymbol{z}}}{\|\hat{\boldsymbol{w}} \times \hat{\boldsymbol{z}}\|_2} = \left(\sin\varphi_{\text{c}},\ -\cos\varphi_{\text{c}},\ 0\right). \tag{20}$$

Similarly, $\hat{\boldsymbol{v}}$ is obtained from the cross product of $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{w}}$ as

$$\hat{\boldsymbol{v}} = \frac{\hat{\boldsymbol{u}} \times \hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{u}} \times \hat{\boldsymbol{w}}\|_2} = \left(-\cos\theta_{\text{c}} \cos\varphi_{\text{c}},\ -\cos\theta_{\text{c}} \sin\varphi_{\text{c}},\ \sin\theta_{\text{c}}\right). \tag{21}$$

By plugging (20) and (21) into (18), we get

$$\begin{aligned} \boldsymbol{p} - \boldsymbol{c} = \big(&\hat{x}_{\text{px}} d_x \sin\varphi_{\text{c}} - \hat{y}_{\text{px}} d_y \cos\theta_{\text{c}} \cos\varphi_{\text{c}},\\ &-\hat{x}_{\text{px}} d_x \cos\varphi_{\text{c}} - \hat{y}_{\text{px}} d_y \cos\theta_{\text{c}} \sin\varphi_{\text{c}},\\ &\hat{y}_{\text{px}} d_y \sin\theta_{\text{c}}\big). \end{aligned} \tag{22}$$

Finally, by plugging (15) and (22) into (14), we obtain the desired position vector $\boldsymbol{p}$ of the projected UE as

$$\begin{aligned} \boldsymbol{p} = \big(&\hat{x}_{\text{px}} d_x \sin\varphi_{\text{c}} - \hat{y}_{\text{px}} d_y \cos\theta_{\text{c}} \cos\varphi_{\text{c}} + D_{\text{f}} \sin\theta_{\text{c}} \cos\varphi_{\text{c}},\\ &-\hat{x}_{\text{px}} d_x \cos\varphi_{\text{c}} - \hat{y}_{\text{px}} d_y \cos\theta_{\text{c}} \sin\varphi_{\text{c}} + D_{\text{f}} \sin\theta_{\text{c}} \sin\varphi_{\text{c}},\\ &\hat{y}_{\text{px}} d_y \sin\theta_{\text{c}} + D_{\text{f}} \cos\theta_{\text{c}}\big). \end{aligned} \tag{23}$$

[7]Recall that in the vision information acquisition stage, the BS takes a shot for the area covered by the SSB beam index. Thus, the camera direction $(\theta_{\text{c}}, \varphi_{\text{c}})$ can be obtained from the SSB beam index.

Since $\boldsymbol{m} \parallel \boldsymbol{p}$, the elevation and azimuth angles $(\hat{\theta}, \hat{\varphi})$ of the position vector $\boldsymbol{m}$ of the UE in the real world can be obtained by converting $\boldsymbol{p}$ to the spherical coordinates.

Proposition 1: The Cartesian coordinates $\boldsymbol{p} = [p_x, p_y, p_z]$ of the projected UE in the image are given by

$$\boldsymbol{p} = \begin{bmatrix} \sin\varphi_{\text{c}} & -\cos\theta_{\text{c}} \cos\varphi_{\text{c}} & \sin\theta_{\text{c}} \cos\varphi_{\text{c}} \\ -\cos\varphi_{\text{c}} & -\cos\theta_{\text{c}} \sin\varphi_{\text{c}} & \sin\theta_{\text{c}} \sin\varphi_{\text{c}} \\ 0 & \sin\theta_{\text{c}} & \cos\theta_{\text{c}} \end{bmatrix} \begin{bmatrix} \hat{x}_{\text{px}} d_x \\ \hat{y}_{\text{px}} d_y \\ D_{\text{f}} \end{bmatrix} \tag{24}$$

where $(\hat{x}_{\text{px}}, \hat{y}_{\text{px}})$ is the 2D pixel-wise coordinates, $(\theta_{\text{c}}, \varphi_{\text{c}})$ is the elevation and azimuth camera directions, D_{f} is the focal length, and $d_x \times d_y$ is the pixel size. Also, the elevation and azimuth angles $(\hat{\theta}, \hat{\varphi})$ of the UE in the real world are

$$\hat{\theta} = \arccos\left(\frac{p_z}{(p_x^2 + p_y^2 + p_z^2)^{\frac{1}{2}}}\right) \tag{25}$$

$$\hat{\varphi} = \text{sgn}(p_y) \arccos\left(\frac{p_x}{(p_x^2 + p_y^2)^{\frac{1}{2}}}\right). \tag{26}$$

where $\text{sgn}(\cdot)$ is the sign function. □

Proof: (24) is directly from (23). Also, (25) and (26) are from the conversion equations between two coordinate systems. ⊠

Finally, by combining the elevation and azimuth angles of UEs $\left\{\left(\hat{\theta}_k, \hat{\varphi}_k\right)\right\}_{k=1}^K$ in (25) and (26) with the distance $\{\hat{r}_k\}_{k=1}^K$ acquired from the depth camera, we obtain the positions $\left\{\left(\hat{r}_k, \hat{\theta}_k, \hat{\varphi}_k\right)\right\}_{k=1}^K$ of the UEs. Then the BS reconstructs the near-field LOS channel vectors $\{\hat{\boldsymbol{h}}_k\}_{k=1}^K$ as

$$\hat{\boldsymbol{h}}_k = \sqrt{\alpha(f, \hat{r}_k)}\, e^{-j\frac{2\pi f}{c}\hat{r}_k} \boldsymbol{a}(\hat{r}_k, \hat{\theta}_k, \hat{\varphi}_k). \tag{27}$$

IV. Position-Aware Near-Field Beam Focusing

In this section, we explain the generation of hybrid beam focusing matrix $\boldsymbol{F}$ using the reconstructed channel information $\{\hat{\boldsymbol{h}}_k\}_{k=1}^K$. Specifically, the optimization problem to determine $\boldsymbol{F}^{\text{opt}} = \boldsymbol{F}_{\text{RF}}^{\text{opt}} \boldsymbol{F}_{\text{BB}}^{\text{opt}}$ maximizing the sum-rate is formulated as

$$\begin{aligned} \mathscr{P}_0: \quad &\underset{\boldsymbol{F}_{\text{RF}}, \boldsymbol{F}_{\text{BB}}}{\text{maximize}} && \sum_{k=1}^K \log_2\left(1 + \gamma_k(\boldsymbol{F}_{\text{RF}}, \boldsymbol{F}_{\text{BB}})\right) && (28a)\\ &\text{subject to} && \boldsymbol{F}_{\text{RF}} \in \mathcal{F}_{\text{RF}} && (28b)\\ & && \|\boldsymbol{F}_{\text{RF}} \boldsymbol{F}_{\text{BB}}\|_{\text{F}}^2 \leqslant P_{\text{tx}} && (28c) \end{aligned}$$

where γ_k is the SINR of the kth UE in (3). Since γ_k is a non-convex quadratic fractional function of $\boldsymbol{F}_{\text{RF}}$ and $\boldsymbol{F}_{\text{BB}}$, it is challenging to determine the optimal solution of $\mathscr{P}_0$. Also, the non-convexity of (28b) makes solving $\mathscr{P}_0$ more challenging.

To obtain a tractable solution to $\mathscr{P}_0$, we convert the intricate sum-rate maximization problem to a series of unconstrained subproblems. Since the closed-form solutions of the subproblems are available, one can derive the suboptimal solution with significantly reduced computational complexity. Specifically, the proposed beam focusing algorithm consists of three major steps: 1) Lagrangian dual transform to convert the sum-of-logarithms-of-ratios problem to the sum-of-ratios problem [59], 2) fractional programming (FP) to decompose the sum-of-ratios problem to a series of constrained quadratic

programs (QPs) [60], and 3) alternating direction method of multipliers (ADMM) to convert the constrained QPs to the unconstrained problems with penalty terms [61].

A. Lagrangian Dual Transform

We first present the Lagrangian dual transform to isolate the quadratic fractional function from its logarithm form [59].

Proposition 2: $\mathscr{P}_0$ can be equivalently converted to $\mathscr{P}_1$ by defining auxiliary variables $\boldsymbol{\omega} \triangleq [\omega_1, \omega_2, \ldots, \omega_K]^{\mathrm{T}} \in \mathbb{R}^K$ as

$$\mathscr{P}_1: \quad \underset{\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{\omega}}{\text{maximize}} \quad f_1(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{\omega}) \tag{30a}$$

$$\text{subject to} \quad \boldsymbol{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{RF}} \tag{30b}$$

$$\|\boldsymbol{F}_{\mathrm{RF}} \boldsymbol{F}_{\mathrm{BB}}\|_{\mathrm{F}}^2 \leqslant P_{\mathrm{tx}} \tag{30c}$$

with

$$f_1(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{\omega}) \triangleq \sum_{k=1}^{K} \log_2(1+\omega_k) - \sum_{k=1}^{K} \omega_k + \sum_{k=1}^{K} \frac{(1+\omega_k)\,\gamma_k(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})}{1+\gamma_k(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})}. \tag{31}$$

Also, the optimal solution $\boldsymbol{\omega}^{\mathrm{opt}} \triangleq [\omega_1^{\mathrm{opt}}, \omega_2^{\mathrm{opt}}, \ldots, \omega_K^{\mathrm{opt}}]^{\mathrm{T}} \in \mathbb{R}^K$ for a given $(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})$ is given by

$$\omega_k^{\mathrm{opt}} = \frac{\left|\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{F}_{\mathrm{RF}} \boldsymbol{f}_{\mathrm{BB},k}\right|^2}{\sum_{j \neq k} \left|\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{F}_{\mathrm{RF}} \boldsymbol{f}_{\mathrm{BB},j}\right|^2 + \sigma_{\mathrm{n}}^2} \qquad \forall k \in \mathcal{K}. \tag{32}$$

□

Proof: See Appendix B. ⊠

Using Proposition 2, we can obtain the suboptimal solution of $\mathscr{P}_1$ in an alternating fashion: 1) fix $(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})$ and update $\boldsymbol{\omega}$ as (32) and 2) fix $\boldsymbol{\omega}$ and solve the reduced problem $\mathscr{P}_{1,\mathrm{a}}$:

$$\mathscr{P}_{1,\mathrm{a}}: \quad \underset{\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}}{\text{maximize}} \quad \sum_{k=1}^{K} \frac{(1+\omega_k)\,\gamma_k(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})}{1+\gamma_k(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})} \tag{33a}$$

$$\text{subject to} \quad \boldsymbol{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{RF}} \tag{33b}$$

$$\|\boldsymbol{F}_{\mathrm{RF}} \boldsymbol{F}_{\mathrm{BB}}\|_{\mathrm{F}}^2 \leqslant P_{\mathrm{tx}}. \tag{33c}$$

B. Fractional Programming

Now we convert the quadratic fractional problem $\mathscr{P}_{1,\mathrm{a}}$ to a sequence of QPs by leveraging the FP technique [60].

Proposition 3: $\mathscr{P}_{1,\mathrm{a}}$ can be equivalently converted to $\mathscr{P}_2$ by defining the auxiliary variables $\boldsymbol{\nu} \triangleq [\nu_1, \nu_2, \ldots, \nu_K]^{\mathrm{T}} \in \mathbb{C}^K$ as

$$\mathscr{P}_2: \quad \underset{\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{\nu}}{\text{maximize}} \quad f_2(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{\nu}) \tag{34a}$$

$$\text{subject to} \quad \boldsymbol{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{RF}} \tag{34b}$$

$$\|\boldsymbol{F}_{\mathrm{RF}} \boldsymbol{F}_{\mathrm{BB}}\|_{\mathrm{F}}^2 \leqslant P_{\mathrm{tx}} \tag{34c}$$

with

$$f_2(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{\nu}) \triangleq 2\Re\{\mathrm{tr}(\boldsymbol{W}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}})\} - \left\|\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\right\|_{\mathrm{F}}^2 - \sigma_{\mathrm{n}}^2\|\boldsymbol{\nu}\|_2^2 \tag{35}$$

where $\boldsymbol{H} \triangleq [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_K]$ is the multi-user channel matrix, $\boldsymbol{V} \triangleq \mathrm{diag}(\boldsymbol{\nu})$, and $\boldsymbol{W} \triangleq \left(\boldsymbol{I}_K + \mathrm{diag}(\boldsymbol{\omega})\right)^{\frac{1}{2}}$. Also, the optimal solution $\boldsymbol{\nu}^{\mathrm{opt}} \triangleq [\nu_1^{\mathrm{opt}}, \nu_2^{\mathrm{opt}}, \ldots, \nu_K^{\mathrm{opt}}]^{\mathrm{T}} \in \mathbb{C}^K$ for a given $(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})$ is

$$\nu_k^{\mathrm{opt}} = \frac{\sqrt{1+\omega_k}\,\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{F}_{\mathrm{RF}} \boldsymbol{f}_{\mathrm{BB},k}}{\sum_{j=1}^{K} \left|\boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{F}_{\mathrm{RF}} \boldsymbol{f}_{\mathrm{BB},j}\right|^2 + \sigma_{\mathrm{n}}^2} \qquad \forall k \in \mathcal{K}. \tag{36}$$

□

Proof: The proof is similar to that of Proposition 2. ⊠

We can derive the suboptimal solution of $\mathscr{P}_2$ through the following alternating steps: 1) fix $(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}})$ and update $\boldsymbol{\nu}$ as (36) and 2) fix $\boldsymbol{\nu}$ and solve the reduced problem $\mathscr{P}_{2,\mathrm{a}}$:

$$\mathscr{P}_{2,\mathrm{a}}: \quad \underset{\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}}{\text{maximize}} \quad f_3(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}) \tag{37a}$$

$$\text{subject to} \quad \boldsymbol{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{RF}} \tag{37b}$$

$$\|\boldsymbol{F}_{\mathrm{RF}} \boldsymbol{F}_{\mathrm{BB}}\|_{\mathrm{F}}^2 \leqslant P_{\mathrm{tx}} \tag{37c}$$

where

$$f_3(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}) \triangleq 2\Re\{\mathrm{tr}(\boldsymbol{W}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}})\} - \left\|\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{BB}}\right\|_{\mathrm{F}}^2. \tag{38}$$

C. Alternating Direction Method of Multipliers

One can see from (38) that the objective function f_3 of $\mathscr{P}_{2,\mathrm{a}}$ is a concave quadratic function of $\boldsymbol{F}_{\mathrm{RF}}$ and $\boldsymbol{F}_{\mathrm{BB}}$. However, $\mathscr{P}_{2,\mathrm{a}}$ is still a nonconvex problem due to (37b). To solve the problem, we use the ADMM technique that converts a complicated constrained problem to an unconstrained problem by adding a quadratic penalty term to the objective function [61].

We first reformulate $\mathscr{P}_{2,\mathrm{a}}$ by introducing the auxiliary variable $\boldsymbol{Q} \in \mathbb{C}^{N \times K}$ to replace $\boldsymbol{F}_{\mathrm{RF}}$ and the indicator function $\mathbb{1}_{\mathcal{F}_{\mathrm{RF}}}(\cdot)$ (i.e., $\mathbb{1}_{\mathcal{F}_{\mathrm{RF}}}(\boldsymbol{F}_{\mathrm{RF}}) = \infty$ if $\boldsymbol{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{RF}}$ and $\mathbb{1}_{\mathcal{F}_{\mathrm{RF}}}(\boldsymbol{F}_{\mathrm{RF}}) = 0$ otherwise) to enforce $\boldsymbol{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{RF}}$:

$$\mathscr{P}_3: \quad \underset{\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{Q}}{\text{maximize}} \quad f_3(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}) + \mathbb{1}_{\mathcal{F}_{\mathrm{RF}}}(\boldsymbol{F}_{\mathrm{RF}}) \tag{39a}$$

$$\text{subject to} \quad \boldsymbol{F}_{\mathrm{RF}} = \boldsymbol{Q} \tag{39b}$$

$$\|\boldsymbol{Q}\boldsymbol{F}_{\mathrm{BB}}\|_{\mathrm{F}}^2 \leqslant P_{\mathrm{tx}}. \tag{39c}$$

By adding a quadratic penalty term for (39b) and a linear penalty term for (39c) to the objective function of $\mathscr{P}_3$, we obtain the augmented Lagrangian L, which is expressed in (29), as shown at the bottom of the page, where $\boldsymbol{\Lambda} \in \mathbb{C}^{N \times K}$ and $\mu \geqslant 0$ are the Lagrangian multipliers for (39b) and (39c), respectively, and $\rho > 0$ is the scaling factor. Using the augmented Lagrangian, the dual problem $\mathscr{P}_4$ is formulated as

$$\mathscr{P}_4: \quad \underset{\boldsymbol{\Lambda}, \mu \geqslant 0}{\text{minimize}} \quad \underset{\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{Q}}{\text{maximize}} \quad L(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{Q}, \boldsymbol{\Lambda}, \mu). \tag{40}$$

$$L(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{Q}, \boldsymbol{\Lambda}, \mu) \triangleq 2\Re\{\mathrm{tr}(\boldsymbol{W}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{Q}\boldsymbol{F}_{\mathrm{BB}})\} - \left\|\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{Q}\boldsymbol{F}_{\mathrm{BB}}\right\|_{\mathrm{F}}^2 + \mathbb{1}_{\mathcal{F}_{\mathrm{RF}}}(\boldsymbol{F}_{\mathrm{RF}}) - \rho\|\boldsymbol{F}_{\mathrm{RF}} - \boldsymbol{Q} + \boldsymbol{\Lambda}\|_{\mathrm{F}}^2 - \mu\left(\|\boldsymbol{Q}\boldsymbol{F}_{\mathrm{BB}}\|_{\mathrm{F}}^2 - P_{\mathrm{tx}}\right) \tag{29}$$

Since $\mathscr{P}_4$ is an unconstrained problem, it is much easier to handle than the primal problem $\mathscr{P}_3$. Also, based on the weak duality, the optimal value of $\mathscr{P}_4$ corresponds to the upper bound of the optimal value of $\mathscr{P}_3$ [61]. As L is a joint function of $\boldsymbol{F}_{\mathrm{RF}}$, $\boldsymbol{F}_{\mathrm{BB}}$, $\boldsymbol{Q}$, $\boldsymbol{\Lambda}$, and μ, in solving $\mathscr{P}_4$, we use an alternating approach that optimizes one variable at a time while fixing the other variables.

1) RF Beam Focusing Matrix Update: When $\boldsymbol{F}_{\mathrm{BB}}$, $\boldsymbol{Q}$, $\boldsymbol{\Lambda}$, and μ are fixed, the update equation of $\boldsymbol{F}_{\mathrm{RF}}$ is given by

$$\begin{aligned}\boldsymbol{F}_{\mathrm{RF}}^{(t+1)} &= \arg\max_{\boldsymbol{F}_{\mathrm{RF}}} \; L\big(\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}^{(t)}, \boldsymbol{Q}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mu^{(t)}\big) \\ &= \arg\max_{\boldsymbol{F}_{\mathrm{RF}}} \; \Big(\mathbb{1}_{\mathcal{F}_{\mathrm{RF}}}(\boldsymbol{F}_{\mathrm{RF}}) \qquad (41)\\ &\qquad\qquad - \rho\big\|\boldsymbol{F}_{\mathrm{RF}} - \boldsymbol{Q}^{(t)} + \boldsymbol{\Lambda}^{(t)}\big\|_{\mathrm{F}}^2\Big) \qquad (42)\\ &= \mathrm{proj}_{\mathcal{F}_{\mathrm{RF}}}\big(\boldsymbol{Q}^{(t)} - \boldsymbol{\Lambda}^{(t)}\big) \qquad (43)\end{aligned}$$

where $\mathrm{proj}_{\mathcal{F}_{\mathrm{RF}}}(\cdot)$ denotes the projection onto $\mathcal{F}_{\mathrm{RF}}$. Using the set of feasible phase shifts Θ (see footnote 2), we obtain

$$\big[\boldsymbol{F}_{\mathrm{RF}}^{(t+1)}\big]_{n,k} = e^{j \arg\min_{w\in\Theta} \left|e^{jw} - \left[\boldsymbol{Q}^{(t)} - \boldsymbol{\Lambda}^{(t)}\right]_{n,k}\right|}. \qquad (44)$$

2) Baseband Beam Focusing Matrix Update: When $\boldsymbol{F}_{\mathrm{RF}}$, $\boldsymbol{Q}$, $\boldsymbol{\Lambda}$, and μ are fixed, the update equation of $\boldsymbol{F}_{\mathrm{BB}}$ is given by

$$\begin{aligned}\boldsymbol{F}_{\mathrm{BB}}^{(t+1)} &= \arg\max_{\boldsymbol{F}_{\mathrm{BB}}} \; L\Big(\boldsymbol{F}_{\mathrm{RF}}^{(t+1)}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{Q}^{(t)}, \boldsymbol{\Lambda}^{(t)}, \mu^{(t)}\Big) \qquad (46)\\ &= \arg\max_{\boldsymbol{F}_{\mathrm{BB}}} \; L_1(\boldsymbol{F}_{\mathrm{BB}}) \qquad (47)\end{aligned}$$

where $L_1(\boldsymbol{F}_{\mathrm{BB}})$ is the reduced Lagrangian defined as

$$\begin{aligned}L_1(\boldsymbol{F}_{\mathrm{BB}}) \triangleq\; & 2\Re\big\{\mathrm{tr}\big(\boldsymbol{W}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{Q}^{(t)}\boldsymbol{F}_{\mathrm{BB}}\big)\big\} \\ & - \big\|\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{Q}^{(t)}\boldsymbol{F}_{\mathrm{BB}}\big\|_{\mathrm{F}}^2 - \mu^{(t)}\big\|\boldsymbol{Q}^{(t)}\boldsymbol{F}_{\mathrm{BB}}\big\|_{\mathrm{F}}^2. \qquad (48)\end{aligned}$$

Since $L_1(\boldsymbol{F}_{\mathrm{BB}})$ is a concave quadratic function of $\boldsymbol{F}_{\mathrm{BB}}$, one can easily obtain $\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}$ by solving $\frac{\partial L_1}{\partial \boldsymbol{F}_{\mathrm{BB}}} = \boldsymbol{0}_{K\times K}$ as

$$\begin{aligned}\boldsymbol{F}_{\mathrm{BB}}^{(t+1)} = & \Big(\big(\boldsymbol{Q}^{(t)}\big)^{\mathrm{H}}\big(\boldsymbol{H}\boldsymbol{V}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}} + \mu^{(t)}\boldsymbol{I}_N\big)\boldsymbol{Q}^{(t)}\Big)^{-1} \\ & \times \big(\boldsymbol{Q}^{(t)}\big)^{\mathrm{H}}\boldsymbol{H}\boldsymbol{V}\boldsymbol{W}^{\mathrm{H}}. \qquad (49)\end{aligned}$$

3) Auxiliary Matrix Update: Similar to the update of $\boldsymbol{F}_{\mathrm{BB}}$, L becomes a concave quadratic function of $\boldsymbol{Q}$ when $\boldsymbol{F}_{\mathrm{RF}}$, $\boldsymbol{F}_{\mathrm{BB}}$, $\boldsymbol{\Lambda}$, and μ are fixed as

$$\begin{aligned}\boldsymbol{Q}^{(t+1)} &= \arg\max_{\boldsymbol{Q}} \; L\Big(\boldsymbol{F}_{\mathrm{RF}}^{(t+1)}, \boldsymbol{F}_{\mathrm{BB}}^{(t+1)}, \boldsymbol{Q}, \boldsymbol{\Lambda}^{(t)}, \mu^{(t)}\Big) \qquad (50)\\ &= \arg\max_{\boldsymbol{Q}} \; L_2(\boldsymbol{Q}) \qquad (51)\end{aligned}$$

where $L_2(\boldsymbol{Q})$ is the reduced Lagrangian defined as

$$\begin{aligned}L_2(\boldsymbol{Q}) \triangleq\; & 2\Re\big\{\mathrm{tr}\big(\boldsymbol{W}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{Q}\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big)\big\} \\ & - \big\|\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{Q}\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big\|_{\mathrm{F}}^2 - \rho\big\|\boldsymbol{F}_{\mathrm{RF}}^{(t+1)} - \boldsymbol{Q} + \boldsymbol{\Lambda}^{(t)}\big\|_{\mathrm{F}}^2 \\ & - \mu^{(t)}\big\|\boldsymbol{Q}\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big\|_{\mathrm{F}}^2. \qquad (52)\end{aligned}$$

Unfortunately, one cannot directly obtain $\boldsymbol{Q}^{(t+1)}$ from (52) since $\boldsymbol{H}$ and $\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}$ are multiplied at both sides of $\boldsymbol{Q}$. To address this issue, we first vectorize $\boldsymbol{Q}$ to $\boldsymbol{q} \triangleq \mathrm{vec}(\boldsymbol{Q}) \in \mathbb{C}^{NK}$ and then re-express L_2 as a function of $\boldsymbol{q}$ as

$$\begin{aligned}L_2(\boldsymbol{q}) \triangleq\; & 2\Re\Big\{\mathrm{vec}\Big(\boldsymbol{H}\boldsymbol{V}\boldsymbol{W}^{\mathrm{H}}\big(\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big)^{\mathrm{H}}\Big)^{\mathrm{H}}\boldsymbol{q}\Big\} \\ & - \Big\|\Big(\big(\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big)^{\mathrm{T}} \otimes \big(\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\big)\Big)\boldsymbol{q}\Big\|_2^2 \\ & - \rho\big\|\mathrm{vec}\big(\boldsymbol{F}_{\mathrm{RF}}^{(t+1)} + \boldsymbol{\Lambda}^{(t)}\big) - \boldsymbol{q}\big\|_2^2 \\ & - \mu^{(t)}\Big\|\Big(\big(\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big)^{\mathrm{T}} \otimes \boldsymbol{I}_N\Big)\boldsymbol{q}\Big\|_2^2. \qquad (53)\end{aligned}$$

We then obtain $\boldsymbol{q}^{(t+1)}$ by solving $\frac{\partial L_2}{\partial \boldsymbol{q}} = \boldsymbol{0}_{NK}$, which is expressed in (45), as shown at the bottom of the next page. Finally, by de-vectorizing $\boldsymbol{q}^{(t+1)}$, we obtain $\boldsymbol{Q}^{(t+1)} = \mathrm{devec}(\boldsymbol{q}^{(t+1)})$.

4) Lagrangian Multiplier Update: Once the primal variables $\boldsymbol{F}_{\mathrm{RF}}$, $\boldsymbol{F}_{\mathrm{BB}}$, and $\boldsymbol{Q}$ are updated, the Lagrangian multipliers are updated using the dual descent method as [61]

$$\boldsymbol{\Lambda}^{(t+1)} = \boldsymbol{\Lambda}^{(t)} + \big(\boldsymbol{F}_{\mathrm{RF}}^{(t+1)} - \boldsymbol{Q}^{(t+1)}\big) \qquad (54)$$

$$\mu^{(t+1)} = \mu^{(t)} + \max\Big(\big\|\boldsymbol{Q}^{(t+1)}\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big\|_{\mathrm{F}}^2 - P_{\mathrm{tx}}, 0\Big). \qquad (55)$$

The update procedures (44), (49), (54), (55) and (45) are repeated until $\boldsymbol{F}_{\mathrm{RF}}$ and $\boldsymbol{F}_{\mathrm{BB}}$ converge. The proposed position-aware hybrid beam focusing algorithm is summarized in Algorithm 1.

Algorithm 1 Position-Aware Hybrid Beam Focusing Algorithm

Input: Position estimates $\{(\hat{r}_k, \hat{\theta}_k, \hat{\varphi}_k)\}_{k=1}^K$, BS transmission power P_{tx}, set of feasible RF precoders $\mathcal{F}_{\mathrm{RF}}$
Initialize:
$\boldsymbol{F}_{\mathrm{RF}} \leftarrow \big[\boldsymbol{a}(\hat{r}_1, \hat{\theta}_1, \hat{\varphi}_1), \boldsymbol{a}(\hat{r}_2, \hat{\theta}_2, \hat{\varphi}_2), \ldots, \boldsymbol{a}(\hat{r}_K, \hat{\theta}_K, \hat{\varphi}_K)\big]$,
$\boldsymbol{F}_{\mathrm{BB}} \leftarrow \mathrm{diag}\big(\sqrt{\alpha(f,\hat{r}_1)}e^{-j\frac{2\pi f \hat{r}_1}{c}}, \sqrt{\alpha(f,\hat{r}_2)}e^{-j\frac{2\pi f \hat{r}_2}{c}}, \ldots, \sqrt{\alpha(f,\hat{r}_K)}e^{-j\frac{2\pi f \hat{r}_K}{c}}\big)$, $\boldsymbol{Q} \leftarrow \boldsymbol{F}_{\mathrm{RF}}$, $\boldsymbol{\Lambda} \leftarrow \boldsymbol{0}_{N\times K}$, $\mu \leftarrow 0$
Reconstruct $\boldsymbol{h}_k$ using (27) $\quad \forall k \in \mathcal{K}$
while $\boldsymbol{F}_{\mathrm{RF}}$ or $\boldsymbol{F}_{\mathrm{BB}}$ do not converge **do**
  Update $\boldsymbol{\omega}$ using (32)
  while $\boldsymbol{F}_{\mathrm{RF}}$ or $\boldsymbol{F}_{\mathrm{BB}}$ do not converge **do**
    Update $\boldsymbol{\nu}$ using (36)
    while $\boldsymbol{F}_{\mathrm{RF}}$ or $\boldsymbol{F}_{\mathrm{BB}}$ do not converge **do**
      Update $\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB}}, \boldsymbol{Q}, \boldsymbol{\Lambda}, \mu$ using (44), (49), (45), (54), and (55)
    end while
  end while
end while
Output: $\boldsymbol{F}_{\mathrm{RF}}$, $\boldsymbol{F}_{\mathrm{BB}}$

D. Asymptotically Optimal Beam Focusing Matrix Design

Since $\boldsymbol{F}_{\mathrm{RF}}$ and $\boldsymbol{F}_{\mathrm{BB}}$ are updated using closed-form expressions, the computational complexity of the proposed near-field hybrid beam focusing algorithm is notably lower compared to conventional approaches relying on intricate optimization techniques (e.g., semidefinite programming (SDP)). In fact, while the most computationally intensive operation in the proposed scheme is matrix inversion, which has a complexity

of $\mathcal{O}(N^3)$, the complexity of SDP is $\mathcal{O}(N^6)$. However, even the matrix inversion operation might be burdensome when the number of antennas N grows exceptionally large in THz UM-MIMO systems. Thankfully, the increase in N leads to the mutual orthogonality among the near-field channel vectors, which can greatly simplify the beam focusing matrix design to achieve the capacity. In the following proposition, we provide the favorable propagation property in the near-field region.

Proposition 4: [Near-field favorable propagation property] When the number of antennas N goes to infinity, distinct near-field array steering vectors are asymptotically orthogonal:

$$\frac{1}{N}\lim_{N\to\infty}\left|\boldsymbol{a}^{\mathrm{H}}(r_i,\theta_i,\varphi_i)\,\boldsymbol{a}(r_j,\theta_j,\varphi_j)\right|=\delta_{i,j}\,. \tag{56}$$

□

Proof: See Appendix C. ⊠

By exploiting the favorable propagation property, one can easily see that the asymptotic optimal beam focusing matrix takes the form of the weighted channel matrix.

Proposition 5: In the ideal THz near-field UM-MISO systems employing continuous phase shifters, the asymptotic optimal solution of $\mathscr{P}_0$ is given by

$$\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{opt}}=\left[\boldsymbol{a}\big(\hat{r}_1,\hat{\theta}_1,\hat{\varphi}_1\big),\boldsymbol{a}\big(\hat{r}_2,\hat{\theta}_2,\hat{\varphi}_2\big),\ldots,\boldsymbol{a}\big(\hat{r}_K,\hat{\theta}_K,\hat{\varphi}_K\big)\right] \tag{57}$$

$$\boldsymbol{F}_{\mathrm{BB}}^{\mathrm{opt}}=\sqrt{\frac{P_{\mathrm{tx}}}{N}}\mathrm{diag}\big(\sqrt{p_1},\sqrt{p_2},\ldots,\sqrt{p_K}\big) \tag{58}$$

where $p_k\geqslant 0$ is the power weight for the kth UE given by

$$p_k=\max\left\{0,\,\frac{1}{\nu}-\frac{\sigma_{\mathrm{n}}^2}{NP_{\mathrm{tx}}\alpha(f,\hat{r}_k)}\right\} \tag{59}$$

and ν is obtained by solving the following equation.

$$\sum_{k=1}^{K}\max\left\{0,\,\frac{1}{\nu}-\frac{\sigma_{\mathrm{n}}^2}{NP_{\mathrm{tx}}\alpha(f,\hat{r}_k)}\right\}=1\,. \tag{60}$$

□

Proof: See Appendix D. ⊠

V. Practical Implementation Issues of VBF

In this section, we briefly discuss practical issues for the successful realization of VBF. These include seamless coverage provision, multi-user identification, and resource usage.

- **Seamless coverage provision**: One of the major challenges of VBF is to ensure seamless service quality even in demanding scenarios with obstacles or low light conditions (e.g., nighttime and rainy weather).[8] To do so, one can exploit a multi-modal sensing, which utilizes multiple sensing modalities simultaneously. For instance, in the NLOS scenario where the UE is visually blocked by obstacles, multiple RGB cameras with different orientations or sensors employing relatively long-wavelength light (e.g., ultrasonic sensor and radar) can be used to detect the UEs hidden behind the obstacles. Also, in the low light environment or adverse weather conditions (e.g., rain, fog, and snow), integration of the RGB camera and non-camera sensors (e.g., radar and LiDAR) can substantially enhance the positioning accuracy of VBF. Since radar and LiDAR generate their own signals (i.e., radio waves and laser pulses) to interact with the environment, they are less susceptible to visibility issues.
- **Multi-user identification**: Since the BS extracts the positions of the UEs without any feedback operation, distinguishing the UE requiring the service is challenging. To identify the target UE, the BS can perform a small range beam sweeping onto the positions identified by the object detector. Specifically, using the extracted positions $\{(\hat{r}_k,\hat{\theta}_k,\hat{\varphi}_k)\}_{k=1}^{K}$, the BS first constructs the beam codebook $\mathcal{C}=\{\boldsymbol{a}(\hat{r}_1,\hat{\theta}_1,\hat{\varphi}_1),(\hat{r}_2,\hat{\theta}_2,\hat{\varphi}_2),\ldots,\boldsymbol{a}(\hat{r}_K,\hat{\theta}_K,\hat{\varphi}_K)\}$ and then sequentially transmits the beam codewords. After that, the UE feeds back the index of best beam codeword through the scheduled physical uplink shared channel (PUSCH). Since the PUSCH scheduling information is distinct for each UE, the BS can discern the target UE from the PUSCH and match it with the position acquired from the beam codeword index feedback.
- **Resource usage**: One natural concern when implementing the DL technique for wireless systems is the resource usage, i.e., latency and power consumption. In VBF, the position information is derived from the image using the CV technique so the traditional beam sweeping latency is replaced by the DNN inference latency. For example, when utilizing the latest artificial intelligence (AI)-focused system-on-chip (SoC) *Qualcomm Snapdragon 888*, the inference time is around $10-15$ ms whereas the beam sweeping latency of 5G NR is 20 ms. Also, considering that the SSB beam transmission power is 20 W while the power consumption of Qualcomm Snapdragon 888 and IntelRealSense L515 RGB-d camera is 5 W and 4 W, respectively, VBF exhibits promising potential for energy savings. We anticipate that this inference power consumption and latency can be further reduced with the implementation of dedicated vision processors designed with a few nano-scale CMOS technology.

[8]Adverse weather conditions (e.g., rain, snow, and fog) can affect the localization performance due to factors like reduced visibility, motion blur, and color distortion. For instance, rain can obscure image clarity and introduce additional noise through droplets on the camera lens. Similarly, fog can decrease visibility by dispersing light and diminishing the contrast, which might blur object outlines and reduce their visual distinctiveness. Also, snow can both impede camera views and change the visual presentation of objects.

VI. Numerical Results

A. Simulation Setup

In our simulations, we consider THz UM-MISO systems where the BS equipped with $N=16\times 16$ UPA antennas

$$\boldsymbol{q}^{(t+1)}=\left(\left(\big(\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big)^{*}\big(\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big)^{\mathrm{T}}\right)\otimes\left(\boldsymbol{H}\boldsymbol{V}\boldsymbol{V}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}+\mu^{(t)}\boldsymbol{I}_N\right)+\rho\boldsymbol{I}_{NK}\right)^{-1}\mathrm{vec}\Big(\boldsymbol{H}\boldsymbol{V}\boldsymbol{W}^{\mathrm{H}}\big(\boldsymbol{F}_{\mathrm{BB}}^{(t+1)}\big)^{\mathrm{H}}+\rho\big(\boldsymbol{F}_{\mathrm{RF}}^{(t+1)}+\boldsymbol{\Lambda}^{(t)}\big)\Big) \tag{45}$$

TABLE I
UE POSITIONING PERFORMANCE OF VBF

	Person		Cell phone		Positioning error		
	precision (%)	**recall** (%)	**precision** (%)	**recall** (%)	**distance** (cm)	**elevation angle** (deg.)	**azimuth angle** (deg.)
VBF (DETR)	97.61	99.80	97.15	97.56	13.2	0.44	0.79
VBF (EfficientDet)	95.50	99.59	93.44	89.76	30.8	0.74	1.37
VBF (ResNet)	91.96	99.80	91.54	93.03	39.0	0.92	1.84
Spherical codebook	-	-	-	-	967	0.88	1.45
Cartesian codebook	-	-	-	-	1,059	0.92	1.53
5G NR beamforming	-	-	-	-	-	6.23	11.08

serves $K = 10$ single-antenna UEs. The UEs are located randomly around the BS within the cell radius of $r = 80\,\text{m}$. For the analog-digital hybrid architecture, we set the number of RF chains to $N_{\text{RF}} = K$ and employ $B = 4$-bit discrete phase shifters to generate the RF beam focusing matrix. In practice, due to the high cost and power consumption of analog phase shifters, low-cost discrete phase shifters are typically adopted. We use the sub-THz near-field LOS channel model with carrier frequency $f_c = 0.1\,\text{THz}$ [13], [39]. Throughout the simulations, we set the signal-to-noise ratio (SNR) to 30 dB. As a performance metric, we use the sum-rate $R_{\text{tot}} = \sum_{k=1}^{K} R_k$. In each point of the plots, we test at least $N_{\text{iter}} = 10,000$ randomly generated near-field THz systems. For the vision-aided UE positioning, we use DETR [57], the state-of-the-art Transformer-based object detector pre-trained on the MS-COCO 2017 dataset consisting of 80 classes of objects and 200,000 training images [62]. To evaluate the performance of VBF, we use VOBEM1, a sensing dataset tailored for wireless communications using Intel RealSense L515 RGB-d camera consisting of 135 pairs of RGB and depth images acquired from 21 distinct wireless environments (see https://github.com/islab-github/VOBEM1).

To compare the UE positioning performance, we employ two benchmark schemes: 1) EfficientDet-based object detector [28] and 2) ResNet-based object detector [63]. Note that these two object detectors are based on the CNN architecture.

Also, to compare the sum-rate performance, we employ five benchmark schemes: 1) fully-digital weighted minimum mean squared error (WMMSE) beam focusing scheme,[9] 2) hybrid beam focusing scheme using the manifold optimization (MO) technique [39], 3) spherical codebook-based scheme [21], 4) Cartesian codebook-based scheme [25], and 5) 2D DFT codebook-based scheme in 5G NR [19]. Note that in the beam focusing vector generation, we utilize the THz near-field channel reconstructed based on the position estimates. To make a fair comparison, for the non-codebook-based schemes, we use the same position estimates obtained from the DETR-based UE positioning.

[9]WMMSE precoding is a type of linear precoding that aims to minimize the sum-mean squared error (MSE) between the transmitted signal and the received signal. Based on the uplink-downlink duality that the solution to the uplink sum-MSE minimization problem is equivalent to the solution to the downlink sum-rate maximization problem, WMMSE precoding achieves the optimal performance among the linear precoding techniques [64].

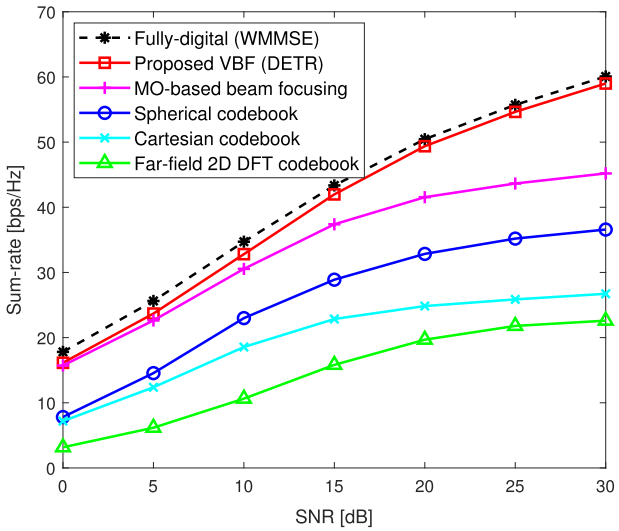

Fig. 5. Sum-rate as a function of SNR.

B. Simulation Results

Table I presents the UE positioning performances of the proposed VBF. The precision is the percentage of correctly detected objects among total detected objects and the recall is the percentage of detected objects among all target objects. We observe that the precision and recall performances of DETR-based object detector (more than 97%) are much higher than those of EfficientDet-based object detector. We also observe that while the conventional codebook-based schemes achieve meter-level positioning accuracy, the proposed VBF achieves a centimeter-level positioning accuracy by using the DETR-based object detector. This is because in the codebook-based schemes, due to a finite number of beam codewords, a mismatch between the pre-defined beam direction and the real UE direction is unavoidable but no such behavior occurs in VBF since the beam is generated from the extracted positions.

Fig. 5 shows the sum-rate as a function of the transmit SNR. We observe that the proposed VBF scheme outperforms the conventional beam focusing schemes by a large margin. For example, when SNR = 30 dB, VBF achieves more than 61.7% sum-rate enhancements over the conventional codebook-based schemes. Even when compared to the MO-based scheme, the sum-rate gain of VBF is more than 30.5%. This is because the MO-based scheme might not perform well in the practical discrete phase shifter scenario since the set of discrete phase

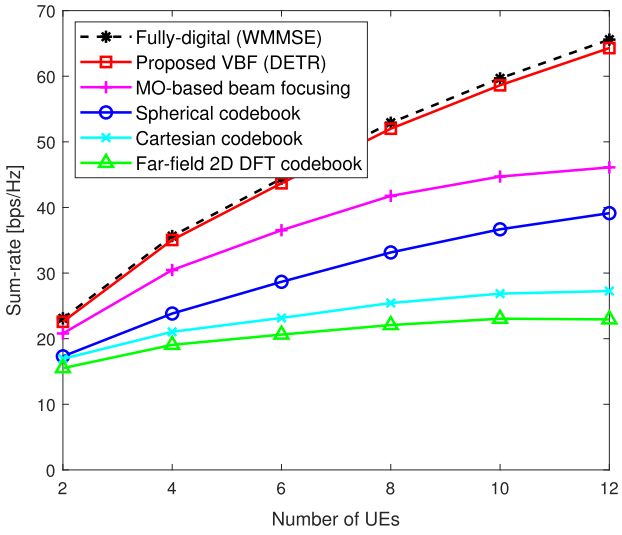

Fig. 6. Sum-rate as a function of the number of UEs.

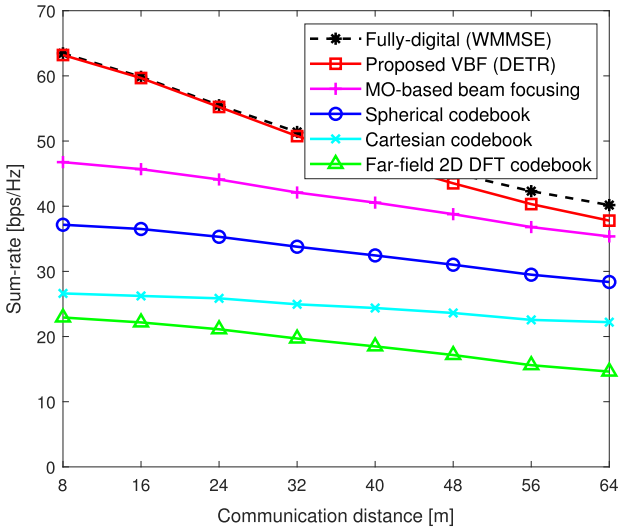

Fig. 7. Sum-rate as a function of communication distance.

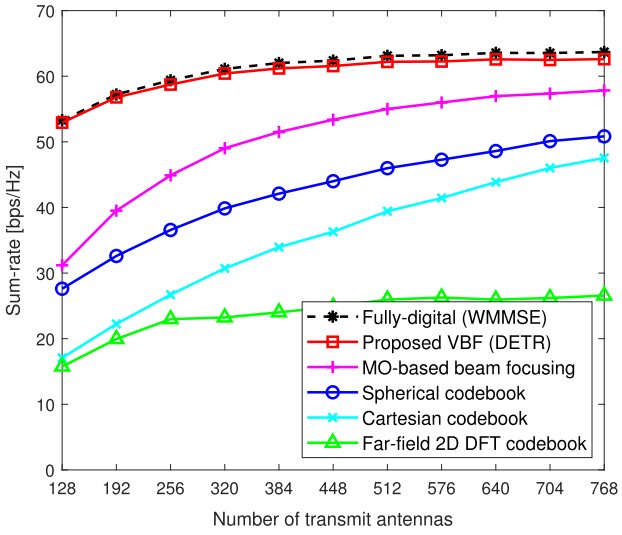

Fig. 8. Sum-rate as a function of the number of antennas.

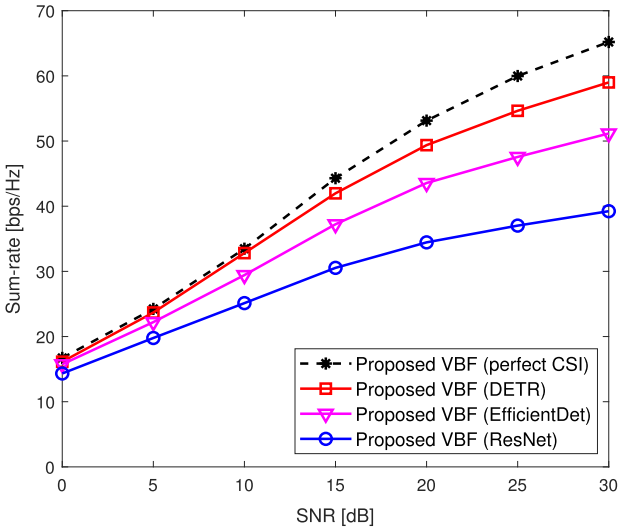

Fig. 9. Sum-rate as a function of SNR.

shifts does not form a smooth Riemannian manifold. Whereas, VBF can effectively handle the discrete phase shift constraints by using the ADMM technique. In fact, the performance of VBF is similar to that of the fully-digital beam focusing scheme.

Fig. 6 shows the sum-rate of VBF as a function of the number of UEs K when SNR $= 30$ dB. From the simulation results, we see that the data rate gain of VBF increases with the number of UEs. Furthermore, we observe that as the number of UEs increases, the sum-rate performances of the conventional beam focusing schemes gradually converge whereas that of the proposed VBF scheme increases sharply. This is because when the number of UEs is large, inter-user interference (IUI) cannot be properly suppressed in the conventional beam focusing schemes due to the finite phase shift levels. In contrast, such is not the case for VBF since we simultaneously achieve the maximization of sum-rate and the satisfaction of the feasible RF beam focusing matrix constraint by maximizing the augmented Lagrangian.

Fig. 7 presents the sum-rate as a function of the communication distance r. We observe that the performance gain of VBF over the codebook-based beam focusing schemes increases as the communication distance decreases. For example, when $r = 64$ m, the performance gain of VBF over the spherical codebook-based scheme is around 33.2% but it increases to 70.1% when $r = 8$ m. Recall that the quadratic term of the exponents of near-field array steering vector is proportional to the inverse of r (see Lemma 1). This means that when the distance is short, the near-field channel vector will be highly affected by the value of r. Thus, the codebook-based schemes suffer a severe performance degradation caused by the relatively high positioning error.

Fig. 8 shows the sum-rate as a function of the number of antennas N. We observe that the proposed VBF achieves a significant data rate gain over the conventional beam focusing schemes. For example, when $N = 512$, the sum-rate gains of VBF over the conventional spherical and Cartesian codebook-based beam focusing schemes are around 35.2% and 57.7%, respectively. Interestingly, we observe that the performances of VBF and benchmark schemes gradually converge as the number of antennas increases. As shown in Proposition 4 and 5, when the number of antennas goes to infinity, the distinct near-field channel vectors become mutually orthogonal so the optimal hybrid beam focusing matrix asymptotically

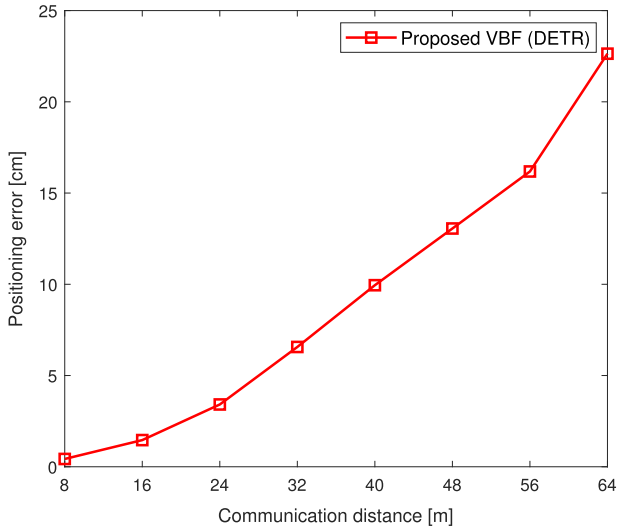

Fig. 10. Positioning error as a function of communication distance.

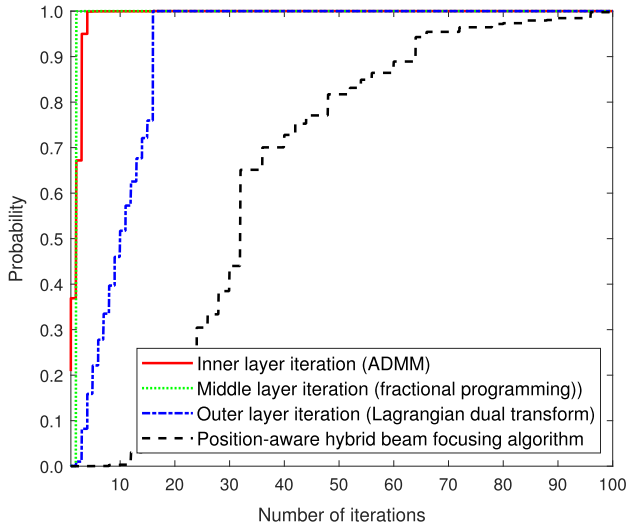

Fig. 11. Cumulative distribution of the number of iterations required to converge.

takes a form of the weighted channel matrix. Since the channel matrix is used as an initial value in the iterative beam focusing optimization techniques, the performances of these techniques ultimately converge with the growing number of antennas.

To investigate the impact of positioning error on the beam focusing performance, we compare the sum-rate performance of the proposed scheme with the ideal beam focusing scheme using the perfect channel state information (CSI) and the schemes using the EfficientDet-based object detector and the ResNet-based object detector in Fig. 9. We observe that the performance gap between the ideal scheme and the proposed scheme utilizing the DETR-based object detector is minimal. In contrast, the performance gap between the ideal scheme and those employing the EfficientDet-based object detector and ResNet-based object detector is relatively large. This arises from the difference of positioning accuracies between the DETR-based object detector and those based on EfficientDet or ResNet.

Fig. 10 shows the positioning error as a function of the communication distance r. We observe that although the positioning error increases with the communication distance, VBF still achieves a centimeter-level positioning accuracy even when the UE is far away ($r = 64\,\mathrm{m}$) from the BS. Considering that the effective communication range of THz systems is up to several tens of meter due to the severe path loss of THz band signals, this implies that VBF is a promising solution for pinpointing the UE's location and generate a proper focusing beam in the THz near-field systems.

Fig. 11 shows the cumulative distributions of the number of iterations needed for the convergence of inner layer iteration (ADMM), middle layer iteration (fractional programming), outer layer iteration (Lagrangian dual transform), and the total position-aware hybrid beam focusing algorithm. We observe that all three layer iterations converge within 20 iterations.

VII. Conclusion

This paper proposed a vision-aided beam focusing technique, called VBF, for 6G THz near-field communications. In contrast to conventional approaches relying exclusively on the sweeping of pre-defined beam codewords, we exploit the sensing and CV technologies in determining the beam direction. By extracting the geometric information (distance, azimuth angle, and elevation angle) of a target device from the visual sensing data via CV technique, VBF accurately identifies the UE position, using which the beam focusing vectors maximizing the sum-rate are generated. Since the position information is derived from the captured image, complicated handshaking operations for the pilot transmission and channel feedback can be minimized, thereby reducing the beam training overhead considerably. From the numerical evaluations on realistic 6G environments, we demonstrated that VBF is effective in improving the positioning accuracy and the sum-rate. The proposed VBF will ensure accurate UE targeting and spectral efficiency maximization in THz UM-MIMO systems, thereby enabling the data-heavy applications anticipated for 6G.

Appendix A
Proof of Lemma 1

Proof: Using the fact that the coordinates of the (m,n)th BS antenna and the kth UE are $((m-1)d, 0, (n-1)d)$ and $(r_k \sin\theta_k \cos\varphi_k, r_k \sin\theta_k \sin\varphi_k, r_k \cos\theta_k)$, respectively, $r_k^{(m,n)}$ can be calculated as

$$\begin{aligned}\left(r_k^{(m,n)}\right)^2 &= (r_k \sin\theta_k \cos\varphi_k - (m-1)d)^2 \\ &\quad + (r_k \sin\theta_k \sin\varphi_k)^2 \\ &\quad + (r_k \cos\theta_k - (n-1)d)^2 \qquad (61)\\ &= r_k^2\left(1 - 2\frac{A_k(m,n)}{r_k} + \frac{B(m,n)}{r_k^2}\right) \qquad (62)\end{aligned}$$

where $A_k(m,n) \triangleq d((m-1)\sin\theta_k \cos\varphi_k + (n-1)\cos\theta_k)$ and $B(m,n) \triangleq d^2((m-1)^2 + (n-1)^2)$. Using the second-order Taylor expansion $\sqrt{1+x} \approx 1 + \frac{x}{2} - \frac{x^2}{8}$, we obtain

$$r_k^{(m,n)} \approx r_k - A_k(m,n) + \frac{1}{2r_k}\left(B(m,n) - \left(A_k(m,n)\right)^2\right) - \frac{\left(A_k(m,n) - \frac{B(m,n)}{2r_k}\right)^2 - \left(A_k(m,n)\right)^2}{2r_k}. \qquad (63)$$

In (63), considering that the antenna spacing $d = \frac{\lambda}{2}$ is in the order of millimeter or even micrometer whereas the horizontal and vertical antenna indices m and n are less than hundred, $A_k(m,n)$ is generally much smaller than $\frac{B(m,n)}{2r_k}$. Thus, the cubic and quartic terms of (63) can be readily neglected, leading to the result in Lemma 1. ☒

Appendix B
Proof of Proposition 2

Proof: Since f_1 is a concave differentiable function of $\boldsymbol{\omega}$, $\boldsymbol{\omega}^{\rm opt}$ in (32) can be obtained by solving $\frac{\partial f_1}{\partial \boldsymbol{\omega}} = \mathbf{0}_K$. Also, one can verify that by plugging $\boldsymbol{\omega} = \boldsymbol{\omega}^{\rm opt}$, we get $f_1(\boldsymbol{F}_{\rm RF}, \boldsymbol{F}_{\rm BB}, \boldsymbol{\omega}^{\rm opt}) = \sum_{k=1}^{K} \log_2\big(1+\gamma_k(\boldsymbol{F}_{\rm RF}, \boldsymbol{F}_{\rm BB})\big)$. Thus, $\mathscr{P}_0$ and $\mathscr{P}_1$ are equivalent in the sense that $(\boldsymbol{F}_{\rm RF}, \boldsymbol{F}_{\rm BB})$ is the optimal solution of $\mathscr{P}_0$ if and only if it is the optimal solution of $\mathscr{P}_1$, and the optimal values of $\mathscr{P}_0$ and $\mathscr{P}_1$ are the same. ☒

Appendix C
Proof of Proposition 4

Two lemmas to be used in the proof are presented first.

Lemma 2: For every $H = 1,2,\ldots,N_{\rm h}$ and $V = 1,2,\ldots,N_{\rm v}$ and some $C>0$, we have

$$\left|\frac{1}{N}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{jx_{m,n}}\right|^2 \leqslant \frac{C}{NHV}\left|\sum_{\substack{h=1-H\\h\neq 0}}^{H-1}\sum_{\substack{v=1-V\\v\neq 0}}^{V-1}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{j\partial_{h,v}x_{m,n}}\right| + \mathcal{O}\Big(\frac{HV}{N}\Big) + \mathcal{O}\Big(\frac{H^2V^2}{N^2}\Big) + \mathcal{O}\Big(\frac{1}{HV}\Big). \tag{64}$$

□

Proof: For every $h=1,2,\ldots,H$ and $v=1,2,\ldots,V$, we get

$$\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{jx_{m,n}} = \sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{jx_{m+h,n+v}} + \sum_{i=1}^{h}\sum_{j=1}^{v}\left(e^{jx_{i,j}} - e^{jx_{i+N_{\rm h},j+N_{\rm v}}}\right) \tag{65}$$

$$= \sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{jx_{m+h,n+v}} + \mathcal{O}(HV). \tag{66}$$

By adding (66) for all $h = 1,2,\ldots,H$ and $v = 1,2,\ldots,V$, we get

$$\left|\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{jx_{m,n}}\right|^2 = \left|\frac{1}{HV}\sum_{h=1}^{H}\sum_{v=1}^{V}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{jx_{m+h,n+v}} + \mathcal{O}(HV)\right|^2 \tag{67}$$

$$\leqslant \frac{2}{H^2V^2}\left|\sum_{h=1}^{H}\sum_{v=1}^{V}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{jx_{m+h,n+v}}\right|^2 + \mathcal{O}\big(H^2V^2\big) \tag{68}$$

$$\overset{(a)}{\leqslant} \frac{2N}{H^2V^2}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}}\left|\sum_{h=1}^{H}\sum_{v=1}^{V} e^{jx_{m+h,n+v}}\right|^2 + \mathcal{O}\big(H^2V^2\big) \tag{69}$$

where (a) is from the Cauchy-Schwarz inequality. Note that

$$\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}}\left|\sum_{h=1}^{H}\sum_{v=1}^{V} e^{jx_{m+h,n+v}}\right|^2 = \sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}}\sum_{h,h'=1}^{H}\sum_{v,v'=1}^{V} e^{j\left(x_{m+h,n+v}-x_{m+h',n+v'}\right)} \tag{70}$$

$$= \sum_{h,h'=1}^{H}\sum_{v,v'=1}^{V}\sum_{m=1+h'}^{N_{\rm h}+h'}\sum_{n=1+v'}^{N_{\rm v}+v'} e^{j\left(x_{m+h-h',n+v-v'}-x_{m,n}\right)} \tag{71}$$

$$= CHV\sum_{\substack{h=1-H\\h\neq 0}}^{H-1}\sum_{\substack{v=1-V\\v\neq 0}}^{V-1}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{j\left(x_{m+h,n+v}-x_{m,n}\right)} + \mathcal{O}(NHV) + \mathcal{O}\big((HV)^3\big) \tag{72}$$

where $C>0$ is some scalar. Finally, by plugging (72) into (69), we obtain the desired result in (64). ☒

Lemma 3: If $\partial_{h,v}x : (m,n) \mapsto x_{m+h,n+v} - x_{m,n}$ satisfies

$$\lim_{N\to\infty}\left|\frac{1}{N}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{j\partial_{h,v}x_{m,n}}\right| = 0 \tag{73}$$

for every $h \in \{1-N_{\rm h}, 2-N_{\rm h}, \ldots, N_{\rm h}-1\}\setminus\{0\}$ and $v \in \{1-N_{\rm v}, 2-N_{\rm v}, \ldots, N_{\rm v}-1\}\setminus\{0\}$, then

$$\lim_{N\to\infty}\left|\frac{1}{N}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{jx_{m,n}}\right| = 0\,. \tag{74}$$

□

Proof: Due to (73), the first, second, and third terms of the right-hand side of (64) goes to zero as $N\to\infty$. Also, since H and V are arbitrary, the last term of the right-hand side of (64) goes to zero as $H,V\to\infty$, thereby obtaining the desired result in (74). ☒

We then prove the statement in Proposition 4.

Proof: From the definition of $\boldsymbol{a}(r,\theta,\phi)$ and Lemma 1, one can see that the correlation $f_N \triangleq \left|\frac{1}{N}\boldsymbol{a}^{\rm H}(r_i,\theta_i,\varphi_i)\boldsymbol{a}(r_j,\theta_j,\varphi_j)\right|^2$ is the exponential sum of quadratic functions given by

$$f_N \triangleq \left|\frac{1}{N}\boldsymbol{a}^{\rm H}(r_i,\theta_i,\varphi_i)\,\boldsymbol{a}(r_j,\theta_j,\varphi_j)\right|^2 \tag{75}$$

$$= \left|\frac{1}{N}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{j\frac{2\pi f}{c}\left(\Delta r_i^{(m,n)}-\Delta r_j^{(m,n)}\right)}\right| \tag{76}$$

$$= \left|\frac{1}{N}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{j\left(am^2+bmn+cn^2+dm+en\right)}\right| \tag{77}$$

$$= \left|\frac{1}{N}\sum_{m=1}^{N_{\rm h}}\sum_{n=1}^{N_{\rm v}} e^{jx_{m,n}}\right| \tag{78}$$

where $x_{m,n}$ is the quadratic exponent function of m and n defined as

$$x_{m,n} \triangleq am^2 + bmn + cn^2 + dm + en \tag{79}$$

where a, b, c, d, and e are the coefficients determined by $(r_i, \theta_i, \varphi_i)$ and $(r_j, \theta_j, \varphi_j)$ (see Lemma 1).

When $(a, b, c) = \mathbf{0}_3$, $x_{m,n} = dm + en$ becomes a linear function of m and n so one can see that $\lim_{N\to\infty} f_N = 0$ as

$$\lim_{N\to\infty} f_N = \lim_{N\to\infty} \left| \frac{1}{N} \sum_{m=1}^{N_\mathrm{h}} \sum_{n=1}^{N_\mathrm{v}} e^{j\left(dm+en\right)} \right| \tag{80}$$

$$= \lim_{N\to\infty} \left| \frac{1}{N} \frac{(1-e^{jN_\mathrm{h}d})(1-e^{jN_\mathrm{v}e})}{\left(1-e^{jd}\right)\left(1-e^{je}\right)} \right| \tag{81}$$

$$= 0\,. \tag{82}$$

When $(a, b, c) \neq \mathbf{0}_3$, $x_{m,n} = am^2 + bmn + cn^2 + dm + en$ becomes a quadratic function of m and n, meaning that $\partial_{h,v} x_{m,n}$ is a linear function of m and n given by

$$\partial_{h,v} x_{m,n} = x_{m+h,n+v} - x_{m,n} \tag{83}$$

$$= (2ah + bv)m + (2cv + bh)n + (dh + ev). \tag{84}$$

From (82), one can see that (73) holds true for every $h \in \mathbb{Z} \setminus \{0\}$ and $v \in \mathbb{Z} \setminus \{0\}$. Thus, by exploiting Lemma 3, (74) holds true for $x_{m,n} = am^2 + bmn + cn^2 + dm + en$, which leads to $\lim_{N\to\infty} f_N = 0$. ☒

It is worth mentioning that even in the cases where the near-field array steering vector is modeled by a polynomial of dimension higher than 2, one can show that $\lim_{N\to\infty} \sum_{m=1}^{N_\mathrm{h}} \sum_{n=1}^{N_\mathrm{v}} e^{jx_{m,n}} = 0$ remains valid by sequentially employing Lemma 3.

Appendix D
Proof of Proposition 5

Proof: Using the mutual orthogonality between the near-field array steering vectors, the achievable rate R_k of the kth UE in (2) can be re-expressed as

$$R_k = \log_2 \left(1 + \frac{\left| \boldsymbol{h}_k^\mathrm{H} \boldsymbol{F}_\mathrm{RF} \boldsymbol{f}_{\mathrm{BB},k} \right|^2}{\sigma_\mathrm{n}^2} \right) \tag{85}$$

$$= \log_2 \left(1 + \frac{\alpha(f, r_k) \left| \boldsymbol{a}^\mathrm{H}(r_k, \theta_k, \varphi_k) \boldsymbol{F}_\mathrm{RF} \boldsymbol{f}_{\mathrm{BB},k} \right|^2}{\sigma_\mathrm{n}^2} \right). \tag{86}$$

Thus, one can easily see that the optimal hybrid beam focusing vector $\boldsymbol{F}_\mathrm{RF}^\mathrm{opt} \boldsymbol{f}_{\mathrm{BB},k}^\mathrm{opt}$ maximizing R_k is given by

$$\boldsymbol{F}_\mathrm{RF}^\mathrm{opt} \boldsymbol{f}_{\mathrm{BB},k}^\mathrm{opt} = \sqrt{\frac{p_k P_\mathrm{tx}}{N}} \boldsymbol{a}(r_k, \theta_k, \varphi_k) \tag{87}$$

where $p_k \geqslant 0$ is the power weight such that $\sum_{k=1}^K p_k \leqslant 1$. From (87), one can also see that $\boldsymbol{F}_\mathrm{RF}^\mathrm{opt}$ and $\boldsymbol{F}_\mathrm{BB}^\mathrm{opt}$ are those in (57) and (58).

By plugging (87) to the sum-rate maximization problem (28), we obtain the power allocation problem $\mathscr{P}_\mathrm{power}$:

$$\mathscr{P}_\mathrm{power}: \quad \underset{\boldsymbol{p}}{\text{maximize}} \quad \sum_{k=1}^K \log_2 \left(1 + \frac{N P_\mathrm{tx} \alpha(f, r_k)}{\sigma_\mathrm{n}^2} p_k \right) \tag{88a}$$

$$\text{subject to} \quad \mathbf{1}^\mathrm{T} \boldsymbol{p} = 1 \tag{88b}$$

$$p_k \geqslant 0, \qquad \forall k \in \mathcal{K}\,. \tag{88c}$$

where $\boldsymbol{p} \triangleq [p_1, p_2, \ldots, p_K]^\mathrm{T} \in \mathbb{R}^K$. Note that in (88b), we recast the original inequality constraint $\mathbf{1}^\mathrm{T}\boldsymbol{p} \leqslant 1$ in (28c) into the equality constraint. This is due to the fact that the data rate increases with the power weight so the optimal power weight vector $\boldsymbol{p}^*$ should satisfy the equality condition. It is worth mentioning that $\mathscr{P}_\mathrm{power}$ is a concave water-filling optimization problem that satisfies the Slater's condition [65]. Thus, one can solve $\mathscr{P}_\mathrm{power}$ by finding out the primal solution $\boldsymbol{p}^\mathrm{opt}$ and dual solutions $\boldsymbol{\lambda}^\mathrm{opt}$ and ν^opt satisfying the Karush–Kuhn–Tucker (KKT) conditions:

$$p_k^\mathrm{opt} \geqslant 0 \qquad \forall k \in \mathcal{K} \tag{89a}$$

$$\mathbf{1}^\mathrm{T} \boldsymbol{p}^\mathrm{opt} = 1 \tag{89b}$$

$$\lambda_k^\mathrm{opt} \geqslant 0 \qquad \forall k \in \mathcal{K} \tag{89c}$$

$$\lambda_k^\mathrm{opt} p_k^\mathrm{opt} = 0 \qquad \forall k \in \mathcal{K} \tag{89d}$$

$$\nu^\mathrm{opt} = \frac{N P_\mathrm{tx} \alpha(f, r_k)}{N P_\mathrm{tx} \alpha(f, r_k)\, p_k^\mathrm{opt} + \sigma_\mathrm{n}^2} + \lambda_k^\mathrm{opt} \qquad \forall k \in \mathcal{K}. \tag{89e}$$

The optimal power weight vector $\boldsymbol{p}^\mathrm{opt}$ satisfying (89a)-(89e) is that with elements given by

$$p_k^\mathrm{opt} = \max \left\{ 0, \frac{1}{\nu^\mathrm{opt}} - \frac{\sigma_\mathrm{n}^2}{N P_\mathrm{tx} \alpha(f, r_k)} \right\} \qquad \forall k \in \mathcal{K} \tag{90}$$

where ν^opt is obtained by solving the following equation.

$$\sum_{k=1}^K \max \left\{ 0, \frac{1}{\nu^\mathrm{opt}} - \frac{\sigma_\mathrm{n}^2}{N P_\mathrm{tx} \alpha(f, r_k)} \right\} = 1\,. \tag{91}$$

☒

References

[1] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electron.*, vol. 3, no. 1, pp. 20–29, 2020.

[2] J. H. Winters, "Optimum combining in digital mobile radio with cochannel interference," *IEEE J. Sel. Areas Commun.*, vol. JSAC-2, no. 4, pp. 528–539, Jul. 1984.

[3] A. Conti, W. M. Gifford, M. Z. Win, and M. Chiani, "Optimized simple bounds for diversity systems," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2674–2685, Sep. 2009.

[4] J. Winters, "On the capacity of radio communication systems with diversity in a Rayleigh fading environment," *IEEE J. Sel. Areas Commun.*, vol. JSAC-5, no. 5, pp. 871–878, Jun. 1987.

[5] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 41–59, Aut. 1996.

[6] J. Winters, "Optimum combining for indoor radio systems with multiple users," *IEEE Trans. Commun.*, vol. COM-35, no. 11, pp. 1222–1230, Nov. 1987.

[7] J. H. Winters, J. Salz, and R. D. Gitlin, "The impact of antenna diversity on the capacity of wireless communication systems," *IEEE Trans. Commun.*, vol. 42, no. 234, pp. 1740–1751, Feb. 1994.

[8] M. Z. Win, N. C. Beaulieu, L. A. Shepp, B. F. Logan, and J. H. Winters, "On the SNR penalty of MPSK with hybrid selection/maximal ratio combining over i.i.d. Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 51, no. 6, pp. 1012–1023, Jun. 2003.

[9] P. F. Driessen and G. J. Foschini, "On the capacity formula for multiple input-multiple output wireless channels: A geometric interpretation," *IEEE Trans. Commun.*, vol. 47, no. 2, pp. 173–176, Feb. 1999.

[10] G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 11, pp. 1841–1852, Nov. 1999.

[11] M. Chiani, M. Z. Win, and H. Shin, "MIMO networks: The effects of interference," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 336–349, Jan. 2010.

[12] C. Han et al., "Terahertz wireless channels: A holistic survey on measurement, modeling, and analysis," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 3, pp. 1670–1707, 3rd Quart., 2022.

[13] B. Ning et al., "Beamforming technologies for ultra-massive MIMO in terahertz communications," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 614–658, 2023.

[14] A. Boriskin and R. Sauleau, *Aperture Antennas for Millimeter and Sub-Millimeter Wave Applications*. Berlin, Germany: Springer, 2018.

[15] M. Cui, Z. Wu, Y. Lu, X. Wei, and L. Dai, "Near-field MIMO communications for 6G: Fundamentals, challenges, potentials, and future directions," *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 40–46, Jan. 2023.

[16] A. A. D'Amico, A. de Jesus Torres, L. Sanguinetti, and M. Z. Win, "Cramér-Rao bounds for holographic positioning," *IEEE Trans. Signal Process.*, vol. 70, pp. 5518–5532, 2022.

[17] H. Zhang, N. Shlezinger, F. Guidi, D. Dardari, and Y. C. Eldar, "6G wireless communications: From far-field beam steering to near-field beam focusing," *IEEE Commun. Mag.*, vol. 61, no. 4, pp. 72–77, Apr. 2023.

[18] *Technical Specification Group Radio Access Network; NR; Physical Layer Procedures for Data (Release 18)*, document TS 38.214, V18.0.0, 3GPP, Sep. 2023.

[19] Y. R. Li, B. Gao, X. Zhang, and K. Huang, "Beam management in millimeter-wave communications for 5G and beyond," *IEEE Access*, vol. 8, pp. 13282–13293, 2020.

[20] G. C. Alexandropoulos, V. Jamali, R. Schober, and H. V. Poor, "Near-field hierarchical beam management for ris-enabled millimeter wave multi-antenna systems," in *Proc. IEEE 12th Sens. Array Multichannel Signal Process. Workshop (SAM)*, Jul. 2022, pp. 460–464.

[21] Z. Wu et al., "Multiple access for near-field communications: SDMA or LDMA?" *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1918–1935, Jun. 2023.

[22] J. Chen, F. Gao, M. Jian, and W. Yuan, "Hierarchical codebook design for near-field mmWave MIMO communications systems," *IEEE Wireless Commun. Lett.*, vol. 12, no. 11, pp. 1926–1930, Nov. 2023.

[23] M. Cui, L. Dai, Z. Wang, S. Zhou, and N. Ge, "Near-field rainbow: Wideband beam training for XL-MIMO," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3899–3912, Jun. 2022.

[24] Y. Zhang, X. Wu, and C. You, "Fast near-field beam training for extremely large-scale array," *IEEE Wireless Commun. Lett.*, vol. 11, no. 12, pp. 2625–2629, Dec. 2022.

[25] X. Wei, L. Dai, Y. Zhao, G. Yu, and X. Duan, "Codebook design and beam training for extremely large-scale RIS: Far-field or near-field?" *China Commun.*, vol. 19, no. 6, pp. 193–204, Jun. 2022.

[26] F. Wang et al., "Ring-type codebook design for reconfigurable intelligent surface near-field beamforming," in *Proc. IEEE 33rd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2022, pp. 391–396.

[27] X. Zhang, H. Zhang, C. Li, Y. Huang, and L. Yang, "Environment-specific beam training for extremely large-scale MIMO systems via contrastive learning," *IEEE Commun. Lett.*, vol. 27, no. 10, pp. 2638–2642, Oct. 2023.

[28] Y. Ahn et al., "Towards intelligent millimeter and terahertz communication for 6G: Computer vision-aided beamforming," *IEEE Wireless Commun.*, vol. 30, no. 5, pp. 179–186, Oct. 2022.

[29] M. Chiani, A. Giorgetti, and E. Paolini, "Sensor radar for object tracking," *Proc. IEEE*, vol. 106, no. 6, pp. 1022–1041, Jun. 2018.

[30] H. Sarieddeen, N. Saeed, T. Y. Al-Naffouri, and M.-S. Alouini, "Next generation terahertz communications: A rendezvous of sensing, imaging, and localization," *IEEE Commun. Mag.*, vol. 58, no. 5, pp. 69–75, May 2020.

[31] R. Zhang, B. Shim, W. Yuan, M. D. Renzo, X. Dang, and W. Wu, "Integrated sensing and communication waveform design with sparse vector coding: Low sidelobes and ultra reliability," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4489–4494, Apr. 2022.

[32] R. Zhang et al., "Integrated sensing and communication with massive MIMO: A unified tensor approach for channel and target parameter estimation," *IEEE Trans. Wireless Commun*, early access, 2024.

[33] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Found. Trends Comput. Graph. Vis.*, vol. 12, nos. 1–3, pp. 1–308, 2020.

[34] 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.

[35] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5998–6008.

[36] A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[37] G. Torsoli, M. Z. Win, and A. Conti, "Blockage intelligence in complex environments for beyond 5G localization," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1688–1701, Jun. 2023.

[38] Z. Wang, Z. Liu, Y. Shen, A. Conti, and M. Z. Win, "Location awareness in beyond 5G networks via reconfigurable intelligent surfaces," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 7, pp. 2011–2025, Jul. 2022.

[39] H. Zhang et al., "Beam focusing for near-field multiuser MIMO communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7476–7490, Sep. 2022.

[40] I. E. Gordon et al., "The HITRAN2020 molecular spectroscopic database," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 277, Jan. 2022, Art. no. 107949.

[41] G. Kwon, A. Conti, H. Park, and M. Z. Win, "Joint communication and localization in millimeter wave networks," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1439–1454, Nov. 2021.

[42] G. Kwon, Z. Liu, A. Conti, H. Park, and M. Z. Win, "Integrated localization and communication for efficient millimeter wave networks," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 12, pp. 3925–3941, Dec. 2023.

[43] A. Conti et al., "Location awareness in beyond 5G networks," *IEEE Commun. Mag.*, vol. 59, no. 11, pp. 22–27, Nov. 2021.

[44] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal, "Locating the nodes: Cooperative localization in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 54–69, Jul. 2005.

[45] M. Z. Win et al., "Network localization and navigation via cooperation," *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 56–62, May 2011.

[46] A. H. Sayed, A. Tarighat, and N. Khajehnouri, "Network-based wireless location: Challenges faced in developing techniques for accurate wireless location information," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 24–40, Jul. 2005.

[47] M. Z. Win, Y. Shen, and W. Dai, "A theoretical foundation of network localization and navigation," *Proc. IEEE*, vol. 106, no. 7, pp. 1136–1165, Jul. 2018.

[48] M. Z. Win, W. Dai, Y. Shen, G. Chrisikos, and H. V. Poor, "Network operation strategies for efficient localization and navigation," *Proc. IEEE*, vol. 106, no. 7, pp. 1224–1254, Jul. 2018.

[49] U. A. Khan, S. Kar, and J. M. F. Moura, "Linear theory for self-localization: Convexity, barycentric coordinates, and Cayley–Menger determinants," *IEEE Access*, vol. 3, pp. 1326–1339, 2015.

[50] A. Conti, S. Mazuelas, S. Bartoletti, W. C. Lindsey, and M. Z. Win, "Soft information for localization-of-things," *Proc. IEEE*, vol. 107, no. 11, pp. 2240–2264, Sep. 2019.

[51] S. Safavi, U. A. Khan, S. Kar, and J. M. Moura, "Distributed localization: A linear theory," *Proc. IEEE*, vol. 106, no. 7, pp. 1204–1223, Jul. 2018.

[52] A. Conti, G. Torsoli, C. A. Gómez-Vega, A. Vaccari, G. Mazzini, and M. Z. Win, "3GPP-compliant datasets for xG location-aware networks," *IEEE Open J. Veh. Technol.*, vol. 5, pp. 473–484, 2024.

[53] A. Conti, G. Torsoli, C. A. Gómez, A. Vaccari, and M. Z. Win, "xG-Loc: 3GPP-compliant datasets for xG location-aware networks," *IEEE Dataport*, Dec. 2023, doi: 10.21227/rper-vc03.

[54] *Technical Specification NR; Requirements for Support of Radio Resource Management (Release 15)*, document TS 38.133, V15.3.0, 3GPP, Oct. 2018.

[55] *Technical Specification Group Radio Access Network; Study on New Radio (NR) Access Technology (Release 17)*, document TR 38.912, V17.0.0, 3GPP, Mar. 2022.

[56] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.
[57] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[58] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[59] K. Shen and W. Yu, "Fractional programming for communication systems—Part II: Uplink scheduling via matching," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2631–2644, May 2018.
[60] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.
[61] S. Boyd et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.
[62] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. ECCV*, vol. 14, 2014, pp. 740–755.
[63] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *Proc. IEEE 91st Veh. Technol. Conf.*, May 2020, pp. 1–5.
[64] S. Shi, M. Schubert, and H. Boche, "Rate optimization for multiuser MIMO systems with linear processing," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 4020–4030, Aug. 2008.
[65] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Seungnyun Kim (Member, IEEE) received the B.S. (with honor) and Ph.D. degrees in electrical and computer engineering from the Seoul National University (SNU), Seoul, South Korea, in 2016 and 2023, respectively.

He is currently a Postdoctoral Fellow with the Wireless Information and Network Sciences Laboratory, Massachusetts Institute of Technology, Cambridge, USA. His main areas of research are in information theory, optimization methods, and machine learning with applications to real-world problems, including wireless communications, network localization and navigation, and non-terrestrial networks.

Dr. Kim was a recipient of the Sejong Science Fellowship from the Korean Government in 2023, the Best Ph.D. Dissertation Award from SNU in 2023, the Qualcomm Innovation Fellowship Finalist in 2021, and the Samsung Humantech Paper Award Gold Prize in 2019.

Jihoon Moon (Member, IEEE) received the B.S. degree in electrical and computer engineering from the Seoul National University (SNU), Seoul, South Korea, in 2019.

He is currently pursuing the Ph.D. degree in electrical and computer engineering at SNU. His research interests include sensing and deep learning techniques for 6G.

Jiao Wu (Member, IEEE) received the B.S. degree in communication engineering from North China Electric Power University (NCEPU), China, in 2015, the M.S. degree in electronics and communication engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2018, and the Ph.D. degree in electrical and computer engineering from Seoul National University (SNU), Seoul, South Korea, in 2023.

She is currently a Post-Doctoral Researcher with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. Her research interests include signal processing and optimization techniques for the 5G and 6G wireless communications.

Byonghyo Shim (Senior Member, IEEE) received the B.S. and M.S. degrees in Control and Instrumentation Engineering from Seoul National University, South Korea, in 1995 and 1997, respectively, and the M.S. degree in mathematics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Champaign, IL, USA, in 2004 and 2005, respectively.

From 1997 to 2000, he was an Officer (First Lieutenant) and an Academic full-time Instructor in the Department of Electronics Engineering, Korean Air Force Academy. From 2005 to 2007, he was a Staff Engineer with Qualcomm Inc., San Diego, CA, USA. From 2007 to 2014, he was an Associate Professor with the School of Information and Communication, Korea University, Seoul. Since 2014, he has been with Seoul National University (SNU), where he is currently a Professor of the Department of Electrical and Computer Engineering and Director of Institute of New Media and Communications. His research interests include wireless communications, statistical signal processing, and deep learning.

Dr. Shim was a recipient of the M. E. Van Valkenburg Research Award from the ECE Department, University of Illinois, in 2005, the Hadong Young Engineer Award from IEIE in 2010, the Irwin Jacobs Award from Qualcomm and KICS in 2016, the Shinyang Research Award from the Engineering College of SNU in 2017, the Okawa Foundation Research Award in 2020, the IEEE Comsoc AP Outstanding Paper Award in 2021, and the JCN Best Paper Award in 2024. He was an Elected Member of the Signal Processing for Communications and Networking (SPCOM) Technical Committee of the IEEE Signal Processing Society. He has been served as an Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (TWC), IEEE TRANSACTIONS ON COMMUNICATIONS (TCOM), IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (TVT), IEEE TRANSACTIONS ON SIGNAL PROCESSING (TSP), IEEE WIRELESS COMMUNICATIONS LETTERS (WCL), and *Journal of Communications and Networks* (JCN) and a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC).

Moe Z. Win (Fellow, IEEE) is the Robert R. Taylor Professor at the Massachusetts Institute of Technology (MIT) and the founding director of the Wireless Information and Network Sciences Laboratory. Prior to joining MIT, he was with AT&T Research Laboratories and with the NASA Jet Propulsion Laboratory.

His research encompasses fundamental theories, algorithm design, and network experimentation for a broad range of real-world problems. His current research topics include ultra-wideband systems, network localization and navigation, network interference exploitation, and quantum information science. He has served the IEEE Communications Society as an elected Member-at-Large on the Board of Governors, as elected Chair of the Radio Communications Committee, and as an IEEE Distinguished Lecturer. Over the last two decades, he held various editorial positions for IEEE journals and organized numerous international conferences. He has served on the SIAM Diversity Advisory Committee.

Dr. Win is an elected Fellow of the AAAS, the EURASIP, the IEEE, and the IET. He was honored with two IEEE Technical Field Awards: the IEEE Kiyo Tomiyasu Award (2011) and the IEEE Eric E. Sumner Award (2006, jointly with R. A. Scholtz). His publications, co-authored with students and colleagues, have received several awards. Other recognitions include the MIT Frank E. Perkins Award (2024), the MIT Everett Moore Baker Award (2022), the IEEE Vehicular Technology Society James Evans Avant Garde Award (2022), the IEEE Communications Society Edwin H. Armstrong Achievement Award (2016), the Cristoforo Colombo International Prize for Communications (2013), the Copernicus Fellowship (2011) and the *Laurea Honoris Causa* (2008) from the Università degli Studi di Ferrara, and the U.S. Presidential Early Career Award for Scientists and Engineers (2004).