Scalable Hybrid Beamforming for Multi-User MISO Systems: A Graph Neural Network Approach

Journal:	*IEEE Transactions on Wireless Communications*
Manuscript ID	Paper-TW-Nov-23-2057
Manuscript Type:	Original Transactions Paper
Date Submitted by the Author:	14-Nov-2023
Complete List of Authors:	Wan, Shaojun; ShanghaiTech University, School of Information Science and Technology Wang, Zixin; ShanghaiTech University, School of Information Science and Technology Zhou, Yong; ShanghaiTech University, School of Information Science and Technology
Keyword:	

Scalable Hybrid Beamforming for Multi-User MISO Systems: A Graph Neural Network Approach

Shaojun Wan, *Student Member, IEEE*, Zixin Wang, *Student Member, IEEE*, and Yong Zhou, *Senior Member, IEEE*

***Abstract*—Hybrid beamforming is a disruptive technology for enhancing the energy- and spectral-efficiency of wireless networks with large-scale antenna arrays, yet the current designs fall short of concurrently achieving low computational complexity and high communication scalability. In this paper, we propose a scalable and effective hybrid beamforming framework for multi-user systems, where the bipartite graph neural network (BGNN) is leveraged to exploit the graph topological structure for sum-rate maximization. To capture permutation properties of the sum-rate maximization problem, we model the wireless network as a bipartite graph, where two disjoint sets of vertices respectively model users and radio frequency (RF) chains, and the edges connecting adjacent vertices characterize interactions between users and RF chains. Based on the bipartite graph, we partition the hybrid beamforming optimization into the updates of feature vectors at user and RF chain vertices, which are realized by alternately activating four kinds of vertex operators that constitute the proposed BGNN. The inputs and outputs of each vertex operator are specifically designed to be independent of the user number and RF chain number in terms of dimension. Numerical results validate the superiority of the proposed BGNN framework from the perspectives of achievable sum rate, computation complexity, and scalability.**

***Index Terms*—Hybrid beamforming, bipartite graph neural network, multi-user communications.**

I. Introduction

Large-scale antenna arrays serve as an essential technology for enhancing the spectral- and energy-efficiency of millimeter wave (mmWave) [1], [2] and terahertz (THz) communication systems [3]. Deploying massive antennas on the base station (BS) can effectively reduce severe propagation loss and generate narrow beam patterns to provide users with high-quality and high-security communication services [4], [5]. However, deploying a massive number of antennas also poses a great challenge on the transceiver beamforming design.

In the fully digital beamforming architecture, one dedicated radio frequency (RF) chain is required by each antenna. Since the hardware implementation complexity and energy consumption caused by deploying massive RF chains can be prohibitively high, it is generally infeasible to adopt the fully digital beamforming architecture for systems with massive antennas. To address this problem, the hybrid beamforming architecture, where analog beamforming modules are deployed to connect large-scale antenna arrays with a few of RF chains, is proved to be an effective solution [6], [7]. By properly adjusting both the amplitude and phase of signal, hybrid beamforming is capable of achieving much higher energy-efficiency than fully digital beamforming at the cost of introducing minor loss in spectral-efficiency [8].

The appealing advantages of hybrid beamforming have attracted substantial attentions [9]–[21]. For the single-user multiple-input multiple-output (SU-MIMO) systems, the authors in [9] and [10] proposed a hybrid beamforming algorithm via minimizing the Frobenius norm of the error between a target fully digital beamforming matrix and hybrid beamforming matrices. To further improve the spectral-efficiency, the authors in [13]–[15] proposed to capture the largest antenna array gain by leveraging singular value decomposition (SVD). For multi-user multiple-input single-output (MU-MISO) systems, the authors in [16] and [17] used the equal gain transmission (EGT) and zero-forcing (ZF) to design the analog and digital beamforming matrices, respectively. To achieve a satisfactory performance, however, the proposed EGT-ZF algorithm requires the antenna number to be much larger than the user number. To alleviate this requirement, the authors in [18] proposed to alternately update the analog beamforming matrix through element-wise optimization that aims at maximizing the spectral-efficiency and update the digital beamforming matrix by using ZF. For more complex MU-MIMO communications, the existing works heavily rely on SVD to cancel inter-user interference and maximize the antenna array gain [19]–[21]. When the user number and antenna number are large, however, these algorithms incur extremely high computational complexity. Furthermore, some existing studies restrict the RF chain number to be equal to the user number [13], [16], [17]. The high computational complexity and the hard restriction on the RF chain number severely hamper these optimization-based methods from being applied to the beamforming design in dynamic wireless networks.

To enable effective hybrid beamforming design with low computational complexity, many deep learning (DL) based approaches have been proposed recently [22]–[28]. Specifically, the authors in [22] leveraged DL to achieve the approximation of SVD with varying levels of complexity, and used them to design hybrid beamforming matrices in SU-MIMO communications. To further improve the energy-efficiency, the authors in [23] developed a convolutional neural network (CNN) framework to jointly design hybrid beamforming and antenna selection. Aiming at reducing the signaling overhead due to the acquisition and feedback of perfect channel state information, the authors in [24] proposed to utilize DL technologies for the hybrid beamforming design based on the received signal strength indicator. In addition to the aforementioned studies that employed neural networks to generate both analog and digital beamforming matrices, some

Shaojun Wan, Zixin Wang, and Yong Zhou are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (E-mail: {wanshj, wangzx2, zhouyong}@shanghaitech.edu.cn).

other works leveraged neural networks to predict only the analog beamforming, while the closed-form solutions of digital beamforming were generated through SVD [26], ZF [27], and minimum mean square error (MMSE) algorithms [28]. These DL-based methods can achieve satisfactory performance with much lower computational complexity than the optimization-based methods. However, the parameter dimensions of these neural networks scale with the user number and antenna number. As a result, these neural networks are required to be re-trained whenever either the user number or antenna number changes. The lack of scalability to different systems severely limits the feasibility of these DL-based methods.

As a promising solution for scalable resource allocation, graph neural network (GNN) has attracted lots of attentions in the beamforming design [29]–[33]. The authors in [30] proposed a message passing GNN (MPGNN) for the scalable fully digital beamforming design in systems with multiple transmitter-receiver pairs. For intelligent reflecting surface (IRS) aided systems, the authors in [31] proposed to leverage GNN for the joint channel estimation and beamforming design. By modeling each user as one vertex, the proposed GNN is scalable with respect to the user number. For the fully digital beamforming design in MU-MISO systems, the authors in [32] proposed a bipartite GNN (BGNN) framework, which achieves similar performance as the locally-optimal weighted MMSE (WMMSE) algorithm. For the hybrid beamforming design in SU systems, the authors in [33] proposed a scalable BGNN framework that models the data streams and antenna elements as vertices. However, the investigations of GNN on the hybrid beamforming design for the MU communication systems are left undiscussed.

A. Contributions

The study on hybrid beamforming is of great importance for mmWave and THz communications to achieve both high spectral- and energy-efficiency. In this paper, we investigate the hybrid beamforming design for maximizing the sum rate in MU-MISO systems. The challenges of this problem mainly lie in two aspects. One is the non-convexity due to the hardware characteristics of the hybrid beamforming architecture. It incurs prohibitively high computational complexity to optimally solve this non-convex problem. The other is the requirement of scalability to the RF chain number and user number. The existing methods fall short of concurrently addressing these two challenges. On one hand, the existing optimization-based methods incur high computational complexity to achieve a satisfactory performance, and may have strict constraints on the number of RF chains and rely on an extremely large number of antennas. On the other hand, the existing DL neural networks are only applicable to the systems that have the same numbers of antennas and users as training samples, thereby are not adaptive to the dynamic communication scenarios. To concurrently deal with these two challenges, we propose a BGNN framework for efficient and scalable hybrid beamforming design. We summarize the main contributions of this paper as follows:

- We model the MU-MISO system as a bipartite graph by characterizing the users and RF chains as two disjoint sets of vertices, and channels as edges connecting adjacent vertices. Besides, the feature vectors of user and RF chain vertices are specifically designed to capture necessary information for digital and analog beamforming design, respectively. The modeled bipartite graph well exploits the graph topological structure of MU-MISO communication systems as well as the interactions between users and RF chains.
- Based on the modeled bipartite graph, we maximize the achievable sum rate by developing a BGNN that consists of four kinds of vertex operators, which are implemented by multi-layer perceptron (MLP). These operators are shared for all user and RF chain vertices, facilitating the proposed BGNN to realize *permutation equivalence* of the digital beamforming and *permutation invariance* of the analog beamforming. Besides, the input and output dimensions of these operators are specifically designed to be irrelevant with user number and RF chain number, which contributes to the scalability of the proposed BGNN framework.
- We evaluate the proposed framework through extensive simulations. Numerical results indicate that the proposed BGNN consumes much less computation time than the optimization-based methods during the inference stage. Meanwhile, the proposed BGNN achieves a satisfactory performance in terms of the sum rate under various system settings. Notably, the proposed BGNN outperforms all the benchmark algorithms when the signal-to-noise ratio (SNR) is low. Besides, the scalability of the proposed BGNN to different user number and RF chain number are also validated.

B. Organizations and Notations

We organize the rest of this paper as follows. Section II describes the system model and formulates a sum rate maximization problem. We present the proposed BGNN framework for achieving scalable hybrid beamforming design in Section III. Section IV describes the simulation results. Finally, we conclude the paper in Section V.

Throughout this paper, we adopt bold upper-case and lower-case letters for matrices and vectors, respectively. The notation $(\mathbf{a})_i$, $\Re_e(\mathbf{a})$, $\Im_m(\mathbf{a})$ represents the i-th entry, real component and imaginary component of vector $\mathbf{a}$, respectively. We denote the inverse, transpose and Hermitian transpose as $(\cdot)^{-1}$, $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$, respectively. Furthermore, $\mathbb{C}^{m\times n}$ and $\mathbb{R}^{m\times n}$ respectively denotes the m-by-n dimensional set of complex and real number, while $\mathbf{I}$ and $\mathbf{0}$ denote the identity and all-zero matrix, respectively. Notation $\mathcal{CN}(a,b)$ represents the circular symmetric complex Guassian distribution with mean a and variance b. Notation $\mathcal{U}\left[-\pi,\pi\right]$ and $\mathcal{L}_{\mathrm{a}}$ respectively denotes the uniformly distribution over the interval from $-\pi$ to π and Laplacian distribution. Notation $\mathbb{E}(\cdot)$, $\mathrm{Tr}(\cdot)$, $\|\cdot\|$, and $\circ$ denote the expectation, trace, Frobenius norm, and Hadamard product, respectively.

II. System Model and Probelm Formulation

In this section, we first present the MU-MISO system with hybrid beamforming and then formulate a sum rate maximization problem, taking into account the hardware characteristics of the hybrid beamforming architecture.

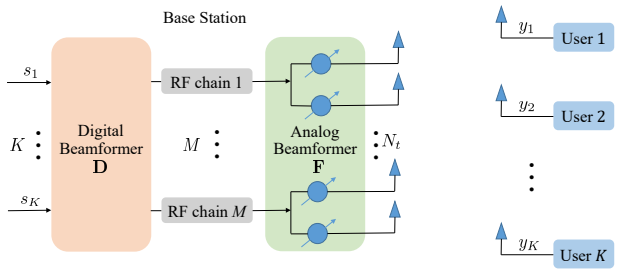

Fig. 1. MU-MISO system model with a partially-connected hybrid beamforming architecture at the BS.

A. System Model

Consider an MU-MISO mmWave communication system that consists of K users and a BS, as shown in Fig. 1. The BS is deployed with N_{t} antennas, while each user is equipped with a single antenna. To maintain low energy consumption and hardware implementation complexity, the BS adopts the partially-connected hybrid beamforming architecture, where N antennas are connected with each RF chain. By denoting the RF chain number as M, we have $N_t = NM$.

The channel matrix from the BS to all users is denoted as $\mathbf{H} = \left[\mathbf{h}_1^{\mathrm{T}}, \mathbf{h}_2^{\mathrm{T}}, \ldots, \mathbf{h}_K^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{K \times N_t}$, which is assumed to be perfectly known at the BS via various state-of-the-art channel estimation methods (e.g., compressed sensing [34], [35] and beam training [36], [37]). We denote the signal intended for the k-th user as s_k, which is assumed to be independent and identically distributed, i.e., $\mathbb{E}\left(\mathbf{s}^{\mathrm{H}}\mathbf{s}\right) = \mathbf{I}$ with $\mathbf{s} = [s_1, s_2, \ldots, s_K]^{\mathrm{T}} \in \mathbb{C}^{K \times 1}$. The hybrid beamforming at the BS consists of a digital beamforming $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k] \in \mathbb{C}^{M \times K}$ and an analog beamforming $\mathbf{F} \in \mathbb{C}^{N_t \times M}$. Through the hybrid beamforming and channel propagation, the signal received by the k-th user, denoted as y_k, is given by

$$y_k = \mathbf{h}_k \mathbf{F} \mathbf{d}_k s_k + \sum_{l \neq k} \mathbf{h}_k \mathbf{F} \mathbf{d}_l s_l + z_k, \tag{1}$$

where $z_k \in \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise at the k-th user.

To accurately capture the limited-scattering characteristic of mmWave channels, we consider the commonly adopted Saleh-Valenzuela channel model [9], [38]. For the MU-MISO system, the channel vector from the BS to the k-th user is given by

$$\mathbf{h}_k = \sqrt{\frac{N_{\mathrm{t}}}{N_{\mathrm{cl}} N_{\mathrm{ray}}}} \sum_{c=1}^{N_{\mathrm{cl}}} \sum_{r=1}^{N_{\mathrm{ray}}} \alpha_{cr}^k \mathbf{a}_t^{\mathrm{H}}(\phi_{cr}^k), \tag{2}$$

where N_{cl} and N_{ray} denote the numbers of clusters and paths belonging to each cluster, respectively, $\alpha_{cr}^k \in \mathcal{CN}(0, 1)$ denotes the complex channel coefficient of the r-th path in the c-th cluster between the k-th user and the BS, and ϕ_{cr}^k denotes the corresponding angle of departure (AOD). Refering to [39], the AODs are assumed to follow the Laplacian distribution, which is well applicable to various mmWave propagation scenarios. Accordingly, we have $\phi_{cr}^k \in \mathcal{L}_{\mathrm{a}}(\phi_c^k, \sigma_\phi)$, where mean cluster angle $\phi_c^k \in \mathcal{U}(-\pi, \pi]$ and angular spread σ_ϕ is set as 7.5 degrees [9]. Moreover, an uniform linear array (ULA) configuration, where the distance between any two adjacent antennas is equal to half-wavelength, is considered at the BS. Under such settings, we can express the transmit array response vector $\mathbf{a}_t(\phi)$ as

$$\mathbf{a}_t(\phi) = \frac{1}{\sqrt{N_{\mathrm{t}}}} [1, e^{j\pi \sin(\phi)}, \ldots, e^{j\pi(N_{\mathrm{t}}-1)\sin(\phi)}]^{\mathrm{H}}. \tag{3}$$

B. Problem Formulation

In this paper, we aim at maximizing the achievable sum rate under the hardware limitations of the partially-connected hybrid beamforming architecture and the constraint on total transmit power. According to (1), the achievable rate of the k-th user with normalized bandwidth, denoted as $r_k(\mathbf{H}, \mathbf{F}, \mathbf{D})$, is given by

$$r_k(\mathbf{H}, \mathbf{F}, \mathbf{D}) = \log\left(1 + \frac{|\mathbf{h}_k \mathbf{F} \mathbf{d}_k|^2}{\sum_{l \neq k} |\mathbf{h}_k \mathbf{F} \mathbf{d}_l|^2 + \sigma^2}\right). \tag{4}$$

For the partially-connected architecture as shown in Fig. 1, the matrix $\mathbf{F}$ characterizing the connections between N_{t} antennas and M RF chains can be expressed as

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{f}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{f}_M \end{bmatrix}, \tag{5}$$

where $\mathbf{f}_m \in \mathbb{C}^{N \times 1}$ denotes the analog beamforming vector dedicated to the m-th RF chain. Besides, we consider the scenario that matrix $\mathbf{F}$ is implemented by phase shifters, which are unable to adjust the amplitude of input signals. Thus, we have

$$|(\mathbf{f}_m)_i| = 1, \forall m \in \mathcal{M}, \forall i \in \mathcal{N}, \tag{6}$$

where $\mathcal{M} = \{1, 2, \ldots, M\}$ denotes the index set of RF chains and $\mathcal{N} = \{1, 2, \ldots, N\}$ denotes the index set of N antenna elements connected with each RF chain. The total transmit power of the BS is limited by P_{t}, i.e., $\|\mathbf{FD}\|_F^2 \leq P_{\mathrm{t}}$. Considering the block-diagonal property of matrix $\mathbf{F}$ as in (5) and the unit modulus constraint as in (6), we have $\mathbf{F}^{\mathrm{H}}\mathbf{F} = N\mathbf{I}$. Thus, we enforce the total transmit power constraint to the digital beamforming as

$$\|\mathbf{D}\|_F^2 \leq \frac{P_{\mathrm{t}}}{N}. \tag{7}$$

Based on the above discussions, we can formulate the problem that aims at maximizing the achievable sum rate as

$$\begin{aligned} \mathcal{P}1: \quad & \underset{\mathbf{F},\mathbf{D}}{\text{maximize}} \quad \textstyle\sum_{k=1}^{K} r_k(\mathbf{H}, \mathbf{F}, \mathbf{D}) \\ & \text{subject to} \quad \|\mathbf{D}\|_F^2 \leq \tfrac{P_{\mathrm{t}}}{N}, \\ & \qquad\qquad |(\mathbf{f}_m)_i| = 1, \forall m \in \mathcal{M}, \forall i \in \mathcal{N}. \end{aligned} \tag{8}$$

It is of great challenge to solve Problem $\mathcal{P}1$ due to the non-convex unit modulus constraint and the coupled optimization

variables. The existing solutions to problem $\mathcal{P}1$ can be categorized into two classes, i.e., optimization-based methods and DL-based methods. In terms of optimization-based methods, the authors in [16], [17] design matrices $\mathbf{F}$ and $\mathbf{D}$ by utilizing the EGT and ZF algorithms, respectively. These methods are of low computational complexity. However, for these methods to achieve satisfactory performance, the antenna number is required to be much larger than the user number, i.e., $N_{\mathrm{t}} \gg K$. Besides, these methods are only applicable to the systems that have an equivalent user number and RF chain number, i.e., $M = K$. In addition, the methods proposed in [14], [18] are developed based on the alternating optimization of matrices $\mathbf{F}$ and $\mathbf{D}$, and can achieve satisfactory performance without relying on the assumption of $N_{\mathrm{t}} \gg K = M$. However, these methods incur prohibitively high computational complexity when the antenna number N_{t} is large. To achieve fast beamforming design, many DL-based methods have been proposed [24], [27], [28]. Specifically, the authors in [24] leverage the DL neural network to optimize both matrices $\mathbf{F}$ and $\mathbf{D}$. Whereas in [27] and [28], only matrix $\mathbf{F}$ is predicted through DL neural networks, while $\mathbf{D}$ is generated with the ZF and MMSE algorithms, respectively. However, the parameter dimensions of these neural networks scale with the user number K and the antenna number N_{t}, which poses a great challenge for training these neural networks when K and N_{t} are large. In addition, these neural networks need to be retrained whenever either K or N_{t} changes, which hampers their applications in dynamic wireless networks.

In summary, although the aforementioned studies can be applied to solve problem $\mathcal{P}1$, they suffer from either of the following limitations, i.e., relying on the strong assumptions such as $N_{\mathrm{t}} \gg K$ and $K = M$, incurring prohibitively high computational complexity, and lacking the scalability to different values of K and M. In this paper, we aim to address all the aforementioned limitations by developing a GNN-based method with low computational complexity and high scalability with respect to different values of K and M.

III. Proposed BGNN Framework for Beamforming Design

In this section, we model the MU-MISO hybrid beamforming system as a bipartite graph, and then propose a BGNN framework for sum-rate maximization. Besides, the training policy and complexity analysis of the proposed BGNN are also presented.

A. Bipartite Graph Representation

To maximize the achievable sum rate of downlink data transmission, we model the MU-MISO system as a bipartite graph, which is explicitly shown in Fig. 2. There are two kinds of vertices in the bipartite graph, i.e., user vertices $\left\{v_k^{\mathrm{UE}}, \forall k \in \mathcal{K}\right\}$ and RF chain vertices $\left\{v_m^{\mathrm{RF}}, \forall m \in \mathcal{M}\right\}$. Vertex v_k^{UE} represents the k-th user and vertex v_m^{RF} represents the m-th RF chain. As there are no direct communications between users and no interference between RF chains, no direct edges connect different users vertices and different RF chain vertices. Every v_k^{UE} is connected with all v_m^{RF} through

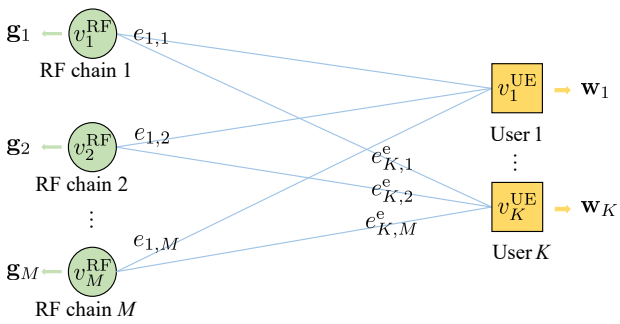

Fig. 2. Bipartite graph representation of the MU-MISO communication system.

edge $e_{k,m}$, which represents the channel between the k-th user and the m-th RF chain. It is worthy noting that all vertices and edges have their own properties, which are captured by their feature vectors [40].

We aim to leverage the vertex feature vectors to design the hybrid beamforming. For the proposed BGNN to be scalable to both user number K and RF chain number M, the dimensions of all vertex feature vectors should be independent of K and M. The feature vector of RF chain vertex v_m^{RF}, denoted as $\mathbf{g}_m$, should be carefully designed to capture sufficient information for the analog beamforming vector $\mathbf{f}_m \in \mathbb{C}^N$. As the existing software packages are incapable of supporting complex-valued operations, we express $\mathbf{g}_m \in \mathbb{R}^{2N}$ as

$$\mathbf{g}_m = \left[\left(\mathbf{f}_m^{\Re}\right)^{\mathrm{T}}, \left(\mathbf{f}_m^{\Im}\right)^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{9}$$

where $\mathbf{f}_m^{\Re} \in \mathbb{R}^N$ and $\mathbf{f}_m^{\Im} \in \mathbb{R}^N$ respectively capture the information of the real and imaginary components of $\mathbf{f}_m$, i.e., $\Re_e\left(\mathbf{f}_m\right)$ and $\Im_m\left(\mathbf{f}_m\right)$.

The feature vector of user vertex v_k^{UE}, denoted as $\mathbf{w}_k$, should capture sufficient information for the digital beamforming vector $\mathbf{d}_k \in \mathbb{C}^M$. One intuitive method is to set $\mathbf{w}_k$ as $\widetilde{\mathbf{w}}_k = \left[\left(\mathbf{d}_k^{\Re}\right)^{\mathrm{T}}, \left(\mathbf{d}_k^{\Im}\right)^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^{2M}$, where $\mathbf{d}_k^{\Re} \in \mathbb{R}^M$ and $\mathbf{d}_k^{\Im} \in \mathbb{R}^M$ respectively capture the information of $\Re_e\left(\mathbf{d}_k\right)$ and $\Im_m\left(\mathbf{d}_k\right)$. However, such a design is unable to achieve the scalability to different values of M. As the BS adopts the partially-connected hybrid beamforming architecture, the total transmit power constraint on matrix $\mathbf{FD}$ can be equivalently transformed into the power constraint on matrix $\mathbf{D}$ as in (7). According to [41], the optimal solution of $\mathbf{d}_k$ with fixed $\mathbf{H}$ and $\mathbf{F}$ is given by

$$\mathbf{d}_k = \sqrt{p_k}\frac{\left(\sigma^2\mathbf{I}_M + \sum_{l=1}^K q_l \hat{\mathbf{h}}_l^{\mathrm{H}}\hat{\mathbf{h}}_l\right)^{-1}\hat{\mathbf{h}}_k^{\mathrm{H}}}{\left\|\left(\sigma^2\mathbf{I}_M + \sum_{l=1}^K q_l \hat{\mathbf{h}}_l^{\mathrm{H}}\hat{\mathbf{h}}_l\right)^{-1}\hat{\mathbf{h}}_k^{\mathrm{H}}\right\|}, \tag{10}$$

where $\hat{\mathbf{h}}_k = \mathbf{h}_k\mathbf{F}$ denotes the effective channel between the k-th user and all the RF chains, and $p_k \geq 0$ and $q_k \geq 0$ satisfying $\sum_{k=1}^k p_k = \sum_{k=1}^K q_k = \frac{P_{\mathrm{t}}}{N}$ denote the primal downlink and virtual uplink power allocated to the k-th user, respectively. With the optimal solution as in (10), we can generate $\mathbf{d}_k$ with the normalized power p_k and q_k. Thus, we can design $\mathbf{w}_k$ as

$$\mathbf{w}_k = [\bar{p}_k, \bar{q}_k]^{\mathrm{T}}, \tag{11}$$

where elements $\bar{p}_k$ and $\bar{q}_k$ can be respectively used to generate p_k and q_k through power normalization. Compared to the intuitive solution $\widetilde{\mathbf{w}}_k \in \mathbb{R}^{2M}$, feature vector $\mathbf{w}_k \in \mathbb{R}^2$ not only achieves the scalability to different values of M, but also reduces the dimension of the solution space from $2M$ to 2, which simplifies the neural network architectures especially when M is large.

As shown in Fig. 2, edge $e_{k,m}$ characterizes the channel between the k-th user and the m-th RF chain, which can be expressed as

$$\mathbf{h}_{k,\mathcal{N}_m} = \left[(\mathbf{h}_k)_{N(m-1)+1}, (\mathbf{h}_k)_{N(m-1)+2}, \ldots, (\mathbf{h}_k)_{Nm}\right]. \tag{12}$$

Thus, we can set the feature vector of edge $e_{k,m}$ as

$$\mathbf{x}^{\mathrm{e}}_{k,m} = \left[\Re_e(\mathbf{h}_{k,\mathcal{N}_m}), \Im_m(\mathbf{h}_{k,\mathcal{N}_m})\right]. \tag{13}$$

We assume that the channels remain constant during each coherence block [31], [42]. With the fixed edge feature vectors, the update of vertex feature vectors is elaborated in the following subsection.

B. Proposed BGNN Architecture

The feature vectors of RF chain vertex v_m^{RF} and user vertex v_k^{UE}, i.e., $\mathbf{g}_m$ and $\mathbf{w}_k$, are designed to capture the information for generating $\mathbf{f}_m$ and $\mathbf{d}_k$, respectively. Based on the modeled bipartite graph in Fig. 2, we propose a BGNN for the hybrid beamforming design, as shown in Fig. 3.

The overall BGNN alternately updates the feature vectors $\mathbf{g}_m$ and $\mathbf{w}_k$ over T iterations by exploiting the interactions between the users and RF chains. We symbolize these interactions by messages, i.e., $\mathbf{c}_{m,k} \in \mathbb{R}^J$ and $\mathbf{b}_{k,m} \in \mathbb{R}^J$, that are exchanged between the corresponding user vertex v_k^{UE} and RF chain vertex v_m^{RF}. In the t-th iteration, as explicitly shown in Fig. 3(b), every RF chain vertex v_m^{RF} first aggregates messages $\mathbf{c}_{k,m}^{[t-1]}$ transmitted from all user vertices, and then updates its feature vector $\mathbf{g}_m^{[t]}$ by activating the RF chain feature decision operator $\mathcal{G}(\cdot)$, followed by generating messages $\mathbf{b}_{m,k}^{[t]}$ by activating the RF chain message generator $\mathcal{B}(\cdot)$. Similarly, as explicitly shown in Fig. 3(c), every user vertex v_k^{UE} first aggregates messages $\mathbf{b}_{m,k}^{[t]}$ transmitted from all RF chain vertices, and then activates the user vertex feature decision operator $\mathcal{W}(\cdot)$ to update its feature vector $\mathbf{w}_k^{[t]}$, followed by activating user message generator $\mathcal{C}(\cdot)$ to generate messages $\mathbf{c}_{k,m}^{[t]}$. Below, we describe the details of each step in each iteration.

1) RF Chain Vertex Feature Decision with $\mathcal{G}(\cdot)$: After receiving messages $\left\{\mathbf{c}_{k,m}^{[t-1]}, \forall k \in \mathcal{K}\right\}$ transmitted from neighboring user vertices, vertex v_m^{RF} endeavors to generate the optimal $\mathbf{g}_m^{[t]}$ by applying operator $\mathcal{G}(\cdot)$. Theoretically, feeding all the messages into operator $\mathcal{G}(\cdot)$ is able to achieve the best performance. However, the resultant dimension of corresponding input messages is KJ, which prohibits the proposed BGNN from realizing the scalability to different values of K. To address this issue, vertex v_m^{RF} first aggregates $\left\{\mathbf{c}_{k,m}^{[t-1]}, \forall k \in \mathcal{K}\right\}$ as

$$\mathbf{c}_m^{[t]} \triangleq \mathcal{P}_{\mathcal{S}}\left(\left\{\mathbf{c}_{k,m}^{[t]}, \forall k \in \mathcal{K}\right\}\right), \tag{14}$$

where $\mathcal{P}_{\mathcal{S}}(\cdot)$ denotes the sum pooling operator. Referring to [32], we in this paper implement the operator $\mathcal{P}_{\mathcal{S}}(\cdot)$ by summation, i.e., $\mathcal{P}_{\mathcal{S}}\left(\{\mathbf{a}_u, \forall u \in \mathcal{U}\}\right) = \sum_{u\in\mathcal{U}} \mathbf{a}_u$. Then, the aggregated message $\mathbf{c}_m^{[t]}$ is of dimension J, which is irrelevant to the value of K. Besides, the adopted $\mathcal{P}_{\mathcal{S}}(\cdot)$ is order-invariant, i.e., changes in the ordering of input vectors have no influence on the output. By further taking into account $\mathbf{g}_m^{[t-1]}$, the update of $\mathbf{g}_m^{[t]}$ is given by

$$\mathbf{g}_m^{[t]} = \mathcal{G}\left(\mathbf{g}_m^{[t-1]}, \mathbf{c}_m^{[t-1]}\right). \tag{15}$$

These processes are illustrated in Fig. 3(b), where $\mathbf{h}_{\mathcal{N}_m} = [\mathbf{h}_{1,\mathcal{N}_m}, \ldots, \mathbf{h}_{K,\mathcal{N}_m}]$ denotes the channel between all the users and the m-th RF chain. Obviously, operator $\mathcal{G}(\cdot)$ maps an input of dimension $J+2N$ to an output of dimension $2N$. Both the input and output dimensions are irrelevant to K and M.

2) RF Chain Message Generation with $\mathcal{B}(\cdot)$: One essential advantage of the proposed BGNN is the capability of exploiting the interactions between users and RF chains by exchanging messages between neighboring vertices. To help vertex v_k^{UE} generate optimal $\mathbf{w}_k^{[t]}$, vertex v_m^{RF} is designed to transmit an encoded message $\mathbf{b}_{m,k}^{[t]}$ to vertex v_k^{UE}. After the vertex feature decision, vertex v_m^{RF} has access to the latest feature vector $\mathbf{g}_m^{[t]}$ and the aggregated message $\mathbf{c}_m^{[t-1]}$. Since the message is transmitted through edge $e_{k,m}$ to vertex v_k^{UE}, operator $\mathcal{B}(\cdot)$ has access to the corresponding edge feature vector $\mathbf{x}^{\mathrm{e}}_{k,m}$. Thus, message $\mathbf{b}_{m,k}^{[t]}$ is generated as

$$\mathbf{b}_{m,k}^{[t]} = \mathcal{B}\left(\mathbf{g}_m^{[t]}, \mathbf{c}_m^{[t-1]}, \mathbf{x}^{\mathrm{e}}_{k,m}\right). \tag{16}$$

As shown in Fig. 3(b), vertex v_m^{RF} generates one tailored message for every user vertex through operator $\mathcal{B}(\cdot)$. The input and output dimensions of operator $\mathcal{B}(\cdot)$ are $J+4N$ and J, respectively.

3) User Vertex Feature Decision with $\mathcal{W}(\cdot)$: Vertex v_k^{UE} aims to generate optimal $\mathbf{w}_k^{[t]}$ by applying operator $\mathcal{W}(\cdot)$. As shown in Fig. 3(c), vertex v_k^{UE} first activates operator $\mathcal{P}_{\mathcal{S}}(\cdot)$ to aggregate messages $\left\{\mathbf{b}_{m,k}^{[t]} : \forall m \in \mathcal{M}\right\}$ as

$$\mathbf{b}_k^{[t]} \triangleq \mathcal{P}_{\mathcal{S}}\left(\left\{\mathbf{b}_{m,k}^{[t]} : \forall m \in \mathcal{M}\right\}\right). \tag{17}$$

The generated vector $\mathbf{b}_k^{[t]} \in \mathbb{R}^J$ in (17) is of the dimension invariant to the RF chain number M. Then, vertex v_k^{UE} updates its feature vector $\mathbf{w}_k^{[t]}$ as

$$\mathbf{w}_k^{[t]} = \mathcal{W}\left(\mathbf{w}_k^{[t-1]}, \mathbf{b}_k^{[t]}\right). \tag{18}$$

The input and output dimensions of operator $\mathcal{W}(\cdot)$ are $J+2$ and 2, respectively.

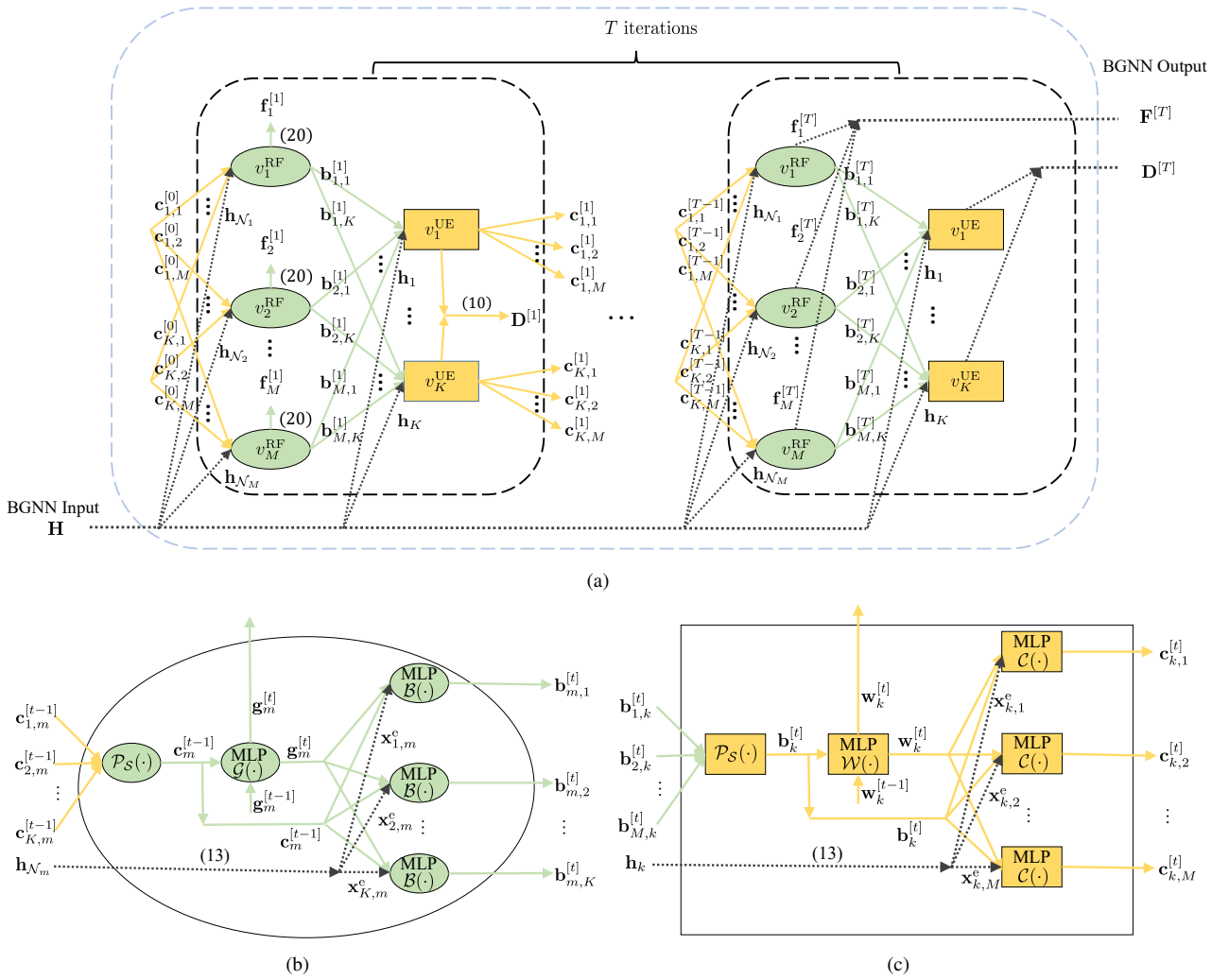

Fig. 3. Proposed BGNN architecture with T iterations. (a) Overall BGNN architecture. (b) RF chain vertex operators in v_m^{RF}. (c) User vertex operators in v_k^{UE}.

4) User Vertex Message Generation with $\mathcal{C}(\cdot)$: To help vertex v_m^{RF} generate an optimal $\mathbf{g}_m^{[t+1]}$ in the next iteration, vertex v_k^{UE} is required to generate an encoded message $\mathbf{c}_{k,m}^{[t]}$ by leveraging operator $\mathcal{C}(\cdot)$. As shown in Fig. 3(c), vertex v_k^{UE} generates the tailored message $\mathbf{c}_{k,m}^{[t]}$ for vertex v_m^{RF} as

$$\mathbf{c}_{k,m}^{[t]} = \mathcal{C}\left(\mathbf{w}_k^{[t]}, \mathbf{b}_k^{[t]}, \mathbf{x}_{k,m}^{\mathrm{e}}\right). \tag{19}$$

Consequently, the input and output dimensions of operator $\mathcal{C}(\cdot)$ are respectively $J+2N+2$ and J, which are irrelevant to K and M.

In the t-th iteration, $\left\{\mathbf{g}_m^{[t]}, \forall m \in \mathcal{M}\right\}$ and $\left\{\mathbf{w}_k^{[t]}, \forall k \in \mathcal{K}\right\}$ can be generated through the above four steps. Then, according to (9) and (6), the computation of $\mathbf{f}_m^{[t]}$ from the updated $\mathbf{g}_m^{[t]}$ can be expressed as

$$(\mathbf{f}_m^{[t]})_i = \frac{(\mathbf{g}_m^{[t]})_i}{\sqrt{(\mathbf{g}_m^{[t]})_i^2 + (\mathbf{g}_m^{[t]})_{i+N}^2}} + j\frac{(\mathbf{g}_m^{[t]})_{i+N}}{\sqrt{(\mathbf{g}_m^{[t]})_i^2 + (\mathbf{g}_m^{[t]})_{i+N}^2}}, \forall i \in \mathcal{N}. \tag{20}$$

Under the total transmit power constraint, the computations of $p_k^{[t]}$ and $q_k^{[t]}$ are given by

$$p_k^{[t]} = \frac{P_{\mathrm{t}}\left(\mathbf{w}_k^{[t]}\right)_1}{N\sum_{k=1}^{K}\left(\mathbf{w}_k^{[t]}\right)_1}, q_k^{[t]} = \frac{P_{\mathrm{t}}\left(\mathbf{w}_k^{[t]}\right)_2}{N\sum_{k=1}^{K}\left(\mathbf{w}_k^{[t]}\right)_2}, \forall k \in \mathcal{K}. \tag{21}$$

Finally, we can generate matrices $\mathbf{F}^{[t]}$ and $\mathbf{D}^{[t]}$ by substituting $\left\{\mathbf{f}_m^{[t]}, \forall m \in \mathcal{M}\right\}$ into (5) and substituting $\left\{p_k^{[t]}, \forall k \in \mathcal{K}\right\}$ and $\left\{q_k^{[t]}, \forall k \in \mathcal{K}\right\}$ into (10).

Compared with the conventional optimization and DL-based methods, the advantages of the proposed BGNN are summarized as follows:

• *Permutation Equivalence and Invariance:* Based on Fig. 3 and the aforementioned discussions, we observe three import characteristics of the proposed BGNN. Firstly, each user (RF chain) vertex shares the same operators with other user (RF

chain) vertices. Secondly, the update of $\mathbf{g}_m^{[t]}$ is based on encoded message $\mathbf{c}_m^{[t-1]}$ that does not change when the user indexes permute, as shown in (14) and (15). Thirdly, the generation of $\mathbf{w}_k^{[t]}$ is based on the encoded message $\mathbf{b}_k^{[t]}$ that is tailored for vertex v_k^{UE}, as shown in (17) and (18). Under such settings, the proposed BGNN successfully exploits the *permutation equivalence* and *permutation invariance* properties of problem $\mathcal{P}1$. Here, *permutation equivalence* means that the digital beamforming vectors $\{\mathbf{d}_k, \forall k \in \mathcal{K}\}$ would permute in the same way as the user channels permute. *Permutation invariance* means that the analog beamforming matrix $\mathbf{F}$ keeps the same regardless of how the user channels permute.

• *Scalability:* According to (14)-(19), both the input and output dimensions of the proposed vertex operators are irrelevant to K and M. Thus, the proposed BGNN is scalable to the variation of K and M by accordingly changing the number of user and RF chain vertices, respectively. However, the conventional neural networks are required to be retrained whenever either K or M changes, as their parameter dimensions scale with K and M [43].

• *Effectiveness:* The proposed BGNN successfully captures the interactions between the users and RF chains by exchanging messages between the neighboring vertices. Thus, the proposed BGNN can realize more effective beamforming design than the conventional neural networks, especially when K and M are large [30], [43].

• *Low Computational Complexity:* The proposed BGNN enjoys lower computational complexity of online beamforming design than the complex optimization-based methods [32], [33], which is verified via both the analytical and simulation results in Subsection IV-A. Besides, each user (RF chain) vertex has no straight connection with any other user (RF chain) vertices, as shown in Fig. 3. As a result, the message generation and vertex feature decision among user (RF chain) vertices can be parallelly performed, which helps reduce the running time.

C. Neural Network Training

The forward pass of training process is constructed by alternately activating the RF chain and user vertex operators through T iterations, as shown in Fig. 3. Generally, the proposed BGNN with a larger T can generate better beamforming matrices $\mathbf{F}^{[T]}$ and $\mathbf{D}^{[T]}$, thereby leading to a higher sum rate. However, a large T causes a great challenge for the backpropagation process to get valid gradients of BGNN parameter $\mathbf{\Theta}$. To address this problem, the loss function should be designed to capture a complete trajectory of $\left\{\mathbf{g}_m^{[t]}, \forall m \in \mathcal{M}\right\}$ and $\left\{\mathbf{w}_k^{[t]}, \forall k \in \mathcal{K}\right\}$ in each iteration. Therefore, the loss function is designed as

$$\mathbf{L}(\mathbf{\Theta}) = -\frac{1}{T}\mathbb{E}_H\left(\sum_{t=1}^{T}\sum_{k=1}^{K} r_k\left(\mathbf{H}, \mathbf{F}^{[t]}, \mathbf{D}^{[t]}\right)\right), \tag{22}$$

where $\mathbb{E}_H(\cdot)$ denotes expectation over different channel realizations.

Algorithm 1 Overall training policy.

Input: Given N, $\bar{\mathcal{K}}$, $\bar{\mathcal{M}}$, N_{e}, N_{b}, and N_{s}.

for $n_{\mathrm{e}} = 1, 2, \ldots, N_{\mathrm{e}}$ **do**
  for $n_{\mathrm{b}} = 1, 2, \ldots, N_{\mathrm{b}}$ **do**
    Decide user number $K \in \bar{\mathcal{K}}$ and RF chain number $M \in \bar{\mathcal{M}}$.
    Independently generate N_{s} channel matrices $\mathbf{H} \in \mathbb{C}^{K \times MN}$ according to (2).
    Perform the forward pass training of BGNN as in Fig. 3, and compute the loss function $\mathbf{L}(\mathbf{\Theta})$ according to (22).
    Update BGNN parameter $\mathbf{\Theta}$ according to (23).
  end for
  Update learning rate μ according to the cosine annealing scheme.
end for

We optimize the BGNN parameter $\mathbf{\Theta}$ to minimize the $\mathbf{L}(\mathbf{\Theta})$ as in (22) by using mini-batch stochastic gradient descent (SGD) methods with Adam optimizer [44]. As discussed in Section III-B, the well designed BGNN is scalable with respect to the user number K and RF chain number M, and thus can be trained over samples that have different values of K and M. The sets of possible values of K and M are denoted as $\bar{\mathcal{K}}$ and $\bar{\mathcal{M}}$, respectively. For each training mini-batch, we first decide the values of K and M, which are sequentially chosen from sets $\bar{\mathcal{K}}$ and $\bar{\mathcal{M}}$. Subsequently, according to the decided K and M, we generate the corresponding channel matrix $\mathbf{H} \in \mathbb{C}^{K \times MN}$ based on the Saleh-Valenzuela channel model in (2). Through the forward pass process of training BGNN as shown in Fig. 3, the BGNN parameter $\mathbf{\Theta}$ is updated as

$$\mathbf{\Theta} \leftarrow \mathbf{\Theta} + \mu \nabla_{\mathbf{\Theta}} \mathbf{L}(\mathbf{\Theta}), \tag{23}$$

where μ denotes the learning rate, and $\nabla_{\mathbf{\Theta}}$ denotes the gradient computation operator with respect to $\mathbf{\Theta}$. Besides, the learning rate μ keeps the same within each training epoch, and is updated according to the cosine annealing scheme. We denote the number of training epochs as N_{e}, and the number of mini-batches within each epoch as N_{b}. Besides, each mini-batch contains N_{s} training samples. The overall training policy is summarized in **Algorithm** 1.

D. Complexity Analysis

In this subsection, we analyze the computational complexity of the proposed BGNN. Generally, the computational complexity of the offline training is affordable [45], [46]. Thus, we focus on analyzing the computational complexity of the online calculations, which are realized by the linear matrix operations. As multiplications are much more complex than additions, we only count the required number of multiplications.

In this paper, all the vertex operators are implemented by MLP-based encoders and shared for all user and RF chain vertices. For the simplicity of expressions, we assume that all the adopted MLPs contain L hidden layers, which are of

the same dimension h. Then, each operator requires $\mathcal{O}(Lh^2)$ real multiplications. As shown in Fig. 3(b), each RF chain vertex contains one feature decision operator $\mathcal{G}(\cdot)$ and K message generators $\mathcal{B}(\cdot)$. As shown in Fig. 3(c), each user vertex contains one operator $\mathcal{W}(\cdot)$ and M operators $\mathcal{C}(\cdot)$. Thus, $\mathcal{O}\left((2MK+M+K)Lh^2\right) \approx \mathcal{O}(2Lh^2MK)$ real multiplications are required in each iteration. Since one complex multiplication contains four real multiplications, the proposed BGNN requires $\mathcal{O}(\frac{1}{2}TLh^2MK)$ complex multiplications. Besides, we generate the digital beamforming according to (10), which incurs the complexity of $\mathcal{O}(NM^2K)$. However, hidden layer dimension h is generally large such that $\frac{1}{2}TLh^2MK \gg NM^2K$. Thus, we can express the overall computational complexity of the proposed BGNN as $\mathcal{O}(\frac{1}{2}TLh^2MK)$.

IV. Simulation Results

In this section, we evaluate the performance of the proposed BGNN framework for the hybrid beamforming design by leveraging the deep learning library Pytorch. The simulations are implemented on an personal computer with AMD Ryzen 9 3900 CPU, 64 GB RAM, and a NVIDIA GeForce GTX 1660 SUPER GPU. In the simulations, we set the iteration number $T = 2$, and implement the proposed operators by the 3-layer MLP-based encoder, i.e., $L = 1$. The hidden layers are of dimension $h = 200$, and use the rectified linear unit (ReLU) as the activation function. For the output layer, operator $\mathcal{W}(\cdot)$ uses Sigmoid as its activation function, while other operators use tangent (Tanh) as activation functions. The settings of MLP-based operators are summarized as in Table I. For the value of J, simulation results indicate that setting $J = N$ can obtain satisfactory performance. For the mmWave channel as in (2), we set $N_{\text{cl}} = 5$ and $N_{\text{ray}} = 2$. Besides, noise z_k is assumed to have unit variance, i.e., $\sigma^2 = 1$.

In the training process, we set $N_{\text{e}} = 100$, $N_{\text{b}} = 300$, and $N_{\text{s}} = 32$. We initialize the learning rate as $\mu = 10^{-3}$, and initialize $\left\{\mathbf{c}_{k,m}^{[0]}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}\right\}$ according to the $\mathcal{CN}(0,1)$ distribution and $\mathbf{w}_k^{[0]}$ according to the equal power allocation as

$$\mathbf{w}_k^{[0]} = \left[\frac{P_{\text{t}}}{NK}, \frac{P_{\text{t}}}{NK}\right]^{\text{T}}, \forall k \in \mathcal{K}. \tag{24}$$

Besides, to follow the permutation invariance of matrix $\mathbf{F}$, we propose to initialize $\mathbf{g}_m^{[0]}$ by matching the phase of $\sum_{k=1}^{K} \mathbf{h}_{k,\mathcal{N}_m}^{\text{H}}$, which is denoted as $\boldsymbol{\psi}_m = \arg\left\{\sum_{k=1}^{K} \mathbf{h}_{k,\mathcal{N}_m}^{\text{H}}\right\}$. Specifically, $\mathbf{g}_m^{[0]}$ is given by

$$\mathbf{g}_m^{[0]} = \left[\Re_e\left(e^{j(\boldsymbol{\psi}_m)_1}\right), \ldots, \Re_e\left(e^{j(\boldsymbol{\psi}_m)_N}\right), \right. \tag{25}$$

$$\left. \Im_m\left(e^{j(\boldsymbol{\psi}_m)_1}\right), \ldots, \Im_m\left(e^{j(\boldsymbol{\psi}_m)_N}\right)\right]^{\text{T}}. \tag{26}$$

To verify the superiority of our developed BGNN, the following baseline methods are considered.

• **SDP-AltMin**: This algorithm proposed in [10] alternately updates the digital and analog beamforming matrices to approach to a target fully digital bemforming matrix. Specifically, the analog beamforming matrix is generated via a closed-form expression and the digital beamforming matrix is generated by solving a semidefinite programming (SDP) problem.

• **MM-AltMin**: This algorithm proposed in [12] approaches the target fully digital beamforming matrix by alternately updating analog beamforming using the major-minimization algorithm and digital beamforming using the MMSE solution.

• **Element-AltMax**: This algorithm proposed in [18] alternately updates the digital and analog beamforming matrices to directly maximize the achievable sum rate. Specifically, the digital beamforming matrix is updated according to ZF algorithm, and the analog beamforming matrix is updated through an element-wise optimization method. Notably, Element-AltMax is only applicable to systems where the number of RF chains is larger than that of users, i.e., $M > K$.

• **Conventional DNN**: To further investigate the effectiveness of our developed BGNN, we also consider a conventional DNN as a baseline. This conventional DNN takes the vectorized channel, i.e., $\mathbf{z} = [\Re_e(\mathbf{h}_1), \ldots, \Re_e(\mathbf{h}_K), \Im_m(\mathbf{h}_1), \ldots, \Im_m(\mathbf{h}_K)]^{\text{T}} \in \mathbb{R}^{2KNM}$ as input, and generates $\bar{\mathbf{z}} = \left[(\bar{\mathbf{z}}_1)^{\text{T}}, (\bar{\mathbf{z}}_2)^{\text{T}}\right]^{\text{T}}$ as output, where $\bar{\mathbf{z}}_1 = \left[(\Re_e(\mathbf{f}_1))^{\text{T}}, \ldots, (\Re_e(\mathbf{f}_M))^{\text{T}}, (\Im_m(\mathbf{f}_1))^{\text{T}}, \ldots, (\Im_m(\mathbf{f}_M))^{\text{T}}\right]^{\text{T}} \in \mathbb{R}^{NM}$ contains the information for analog beamforming, and $\bar{\mathbf{z}}_2 = [p_1, \ldots, p_K, q_1, \ldots, q_K]^{\text{T}} \in \mathbb{R}^{2K}$ contains both the primal downlink and virtual uplink power of users. Then, $\bar{\mathbf{z}}$ can be used to generate matrices $\mathbf{F}$ and $\mathbf{D}$ through (5) and (10), respectively. For a fair comparison, the conventional DNN is also developed based on the 3-layer MLP. We set the dimension of hidden layers in conventional DNN as $2Mh$, which incurs the complexity of $\mathcal{O}(h^2M^2)$ to ensure almost the same computational complexity as the proposed BGNN with $T = 2$ and $L = 1$.

All the three optimization-based algorithms require multiple iterations to reach convergence. In the simulations, we terminate the r-th iteration of optimization-based algorithms until $\gamma = \frac{|t_r - t_{(r-1)}|}{t_{(r-1)}} < 0.01$, where t_r denotes the Frobenius norm of the error between target fully digital beamforming and hybrid beamforming matrices in [10], [12] and the sum rate in [18] in the r-th iteration. Besides, it is worthy noting that the target fully digital beamforming matrix is generated according to the MMSE algorithm [47].

A. Complexity Comparison

We compare the computational complexity of the proposed BGNN with the benchmark algorithms from both the theoretical and numerical perspectives. As previously mentioned, the conventional DNN shares the similar computational complexity with the proposed BGNN. In this subsection, we focus on the complexity comparison between the proposed BGNN and optimization-based algorithms.

The dominant computational complexity of the SDP-AltMin algorithm lies in optimizing the digital beamforming matrix by solving the SDP problem, which has a minimum complexity of $\mathcal{O}(n^{3.5}\log_2(1/\epsilon))$ [48], where n and ϵ respectively denote the dimension of the optimization variable

TABLE I
SETUPS OF MLP-BASED OPERATORS.

Operator	Size	Hidden layer activation function	Output layer activation function
$\mathcal{G}$	$(J+2N)\times 200\times 200\times 2N$	ReLU	Tanh
$\mathcal{B}$	$(J+4N)\times 200\times 200\times J$	ReLU	Tanh
$\mathcal{W}$	$(J+2)\times 200\times 200\times 2$	ReLU	Sigmoid
$\mathcal{C}$	$(J+2N+2)\times 200\times 200\times J$	ReLU	Tanh

and solution accuracy. Accordingly, the computational complexity of SDP-AltMin is $\mathcal{O}(I_{\rm sdr}\log_2(1/\epsilon)(MK+1)^{3.5})\approx \mathcal{O}(I_{\rm sdr}\log_2(1/\epsilon)M^{3.5}K^{3.5})$, where $I_{\rm sdr}$ denotes the required number of iterations for alternately updating matrices $\mathbf{F}$ and $\mathbf{D}$. For both MM-AltMin and Element-AltMax, multiple iterations for updating matrix $\mathbf{F}$ with a given $\mathbf{D}$ are required. We denote the required number of iterations for updating matrix $\mathbf{F}$ with a given $\mathbf{D}$ as $I^{\rm in}$, and that for alternately updating matrices $\mathbf{F}$ and $\mathbf{D}$ as $I^{\rm out}$. For the MM-AltMin algorithm, the overall computational complexity can be expressed as $\mathcal{O}(I^{\rm out}_{\rm mm}I^{\rm in}_{\rm mm}N^3M^6)$, which is dominated by the eigenvalue decomposition when updating matrix $\mathbf{F}$. For the Element-AltMax algorithm, the computational complexity is given by $\mathcal{O}\left(2I^{\rm out}_{\rm ele}I^{\rm in}_{\rm ele}N^3M^3K\right)$, which is dominated by the element-wise optimization of matrix $\mathbf{F}$. We summarize the theoretical computational complexity of optimization-based algorithms and the proposed BGNN in Table II.

According to the aforementioned analysis, the proposed BGNN enjoys a much lower computational complexity than the conventional optimization-based algorithms, especially when K, M, and N are large. To directly show this advantage of the proposed BGNN, we present the running time of all these algorithms versus M in Fig. 4. As the available python packages fail to support the CVX programming on GPU, the SDP-AltMin algorithm is implemented on CPU, while other algorithms are implemented on GPU. It can be observed in Fig. 4 that, with the increase of M, the running time of all the benchmark algorithms increases rapidly, while that of the proposed BGNN increases slightly. The trends shown in Fig. 4 are consistent with the analytical results. Besides, the proposed method consumes far less running time than the benchmark algorithms. The running time ratio of the proposed BGNN to the benchmark algorithms is around 0.2% when $M=8$.

Both the analytical and simulation results verify the superiority of the proposed BGNN over the conventional optimization-based algorithms in terms of the computational complexity.

B. Effectiveness

To investigate the effectiveness of the proposed BGNN, we compare its performance with the benchmark algorithms under different values of $P_{\rm t}$ and N. In this subsection, the proposed BGNN is trained with data samples that have fixed user number K and RF chain number M.

Fig. 5 shows the impact of the total transmit power $P_{\rm t}$ on the achievable sum rates of the proposed BGNN and benchmark algorithms. From Fig. 5, we can observe that the proposed BGNN achieves a satisfactory performance under any $P_{\rm t}$. Benefiting from unsupervised learning that directly maximizes the sum rate, both the proposed BGNN and conventional DNN outperform the complex optimization-based SDP-AltMin and MM-AltMin methods that maximize the sum rate by approaching the target fully digital beamforming scheme. By successfully exploiting the graph topological structure and the interactions between users and RF chains, the proposed BGNN performs better than the conventional DNN with any $P_{\rm t}$. Moreover, the proposed BGNN obtains a significantly higher sum rate than the Element-AltMax algorithm when $P_{\rm t}\le -5$ dB. This is because the Element-AltMax algorithm designs the digital beamforming matrix by utilizing the ZF algorithm, which performs unsatisfactorily when SNR is low [49]. When $P_{\rm t}=5$, the proposed BGNN achieves around 94% of the sum rate attained by the Element-AltMax algorithm.

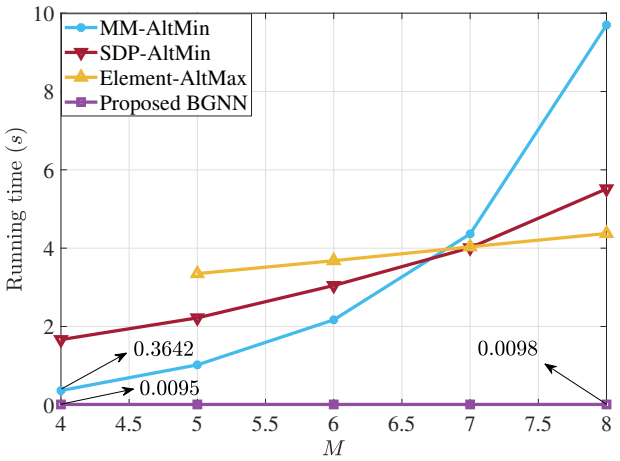

Fig. 4. Running time of the proposed BGNN and benchmark algorithms versus M when $K=4$, $N=8$, and $P_{\rm t}=10$ dB.

Fig. 6 illustrates the sum rate achieved by the proposed BGNN and benchmark algorithms versus the number of antennas connected with each RF chain, i.e., N. From Fig. 6, it can be seen that the sum rates of all methods increase with the increase of N. This is because a larger value of N with fixed M results in larger N_t, which provides a higher array gain to each user. Under any N, the proposed BGNN performs the best among these methods. The performance gain of the proposed BGNN over conventional DNN becomes larger, as the value of N increases. This is because the advantages of utilizing the interactions between users and RF chains tends to be more significant with a larger $N_{\rm t}$ [32].

The simulation results in Figs. 5 and 6 verify that the proposed BGNN outperforms the benchmark methods under various parameter settings. The advantage of the proposed

TABLE II
COMPUTATIONAL COMPLEXITY OF OPTIMIZATION-BASED BENCHMARK ALGORITHMS AND THE PROPOSED METHOD.

	Proposed BGNN	Element-AltMax	SDP-AltMin	MM-AltMin
Complexity	$\mathcal{O}(\frac{1}{2}TLh^2MK)$	$\mathcal{O}(2I_{\text{ele}}^{\text{out}}I_{\text{ele}}^{\text{in}}N^3M^3K)$	$\mathcal{O}(I_{\text{sdr}}\log_2(1/\epsilon)M^{3.5}K^{3.5})$	$\mathcal{O}(I_{\text{mm}}^{\text{out}}I_{\text{mm}}^{\text{in}}N^3M^6)$

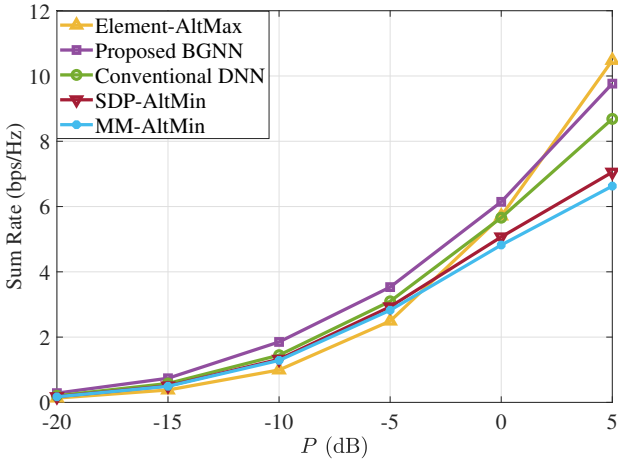

Fig. 5. Achievable sum rates of the proposed BGNN and benchmark algorithms versus P_{t} when $K = 4$, $M = 5$, and $N = 4$.

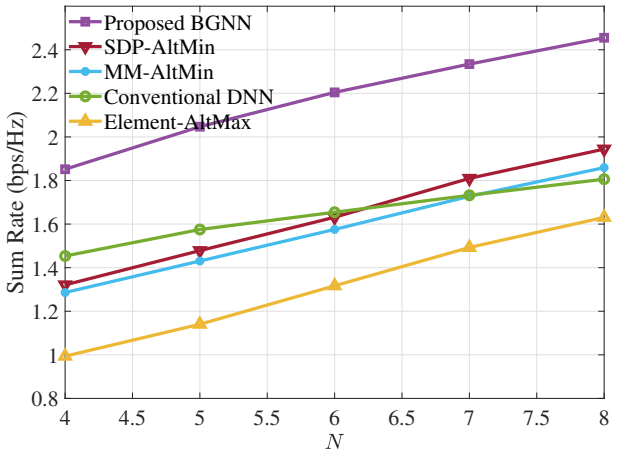

Fig. 6. Achievable sum rates of the proposed BGNN and benchmark algorithms versus N when $K = 4$, $M = 5$, and $P_{\text{t}} = -10$ dB.

BGNN is particularly significant in the low SNR regime when compared to the optimization-based algroithms, and with large values of N_{t} when compared to the conventional DNN.

C. Scalability

In this subsection, we investigate the scalability of the proposed BGNN with respect to the different RF chain number M and user number K. We train the proposed BGNN with data samples that have different values of M and K according to **Algorithm 1**. As a comparison, we also evaluate the performance of BGNN trained with fixed values of K and M. Since the conventional DNN is incapable of generalizing to scenarios with different values of K and M, we disregard its performance in the following simulations.

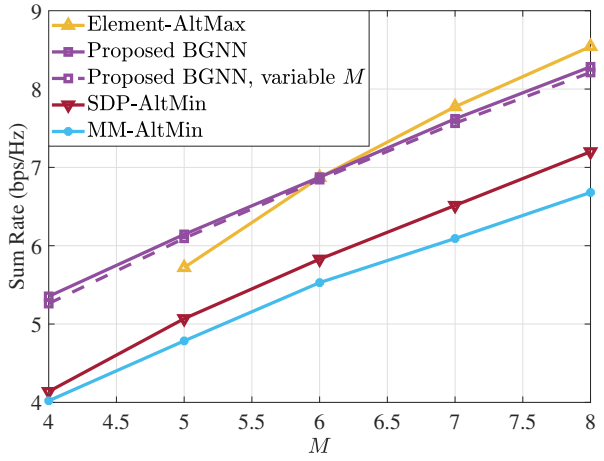

Fig. 7. Achievable sum rates of the proposed BGNN and benchmark algorithms versus M when $K = 4$, $N = 4$, and $P_{\text{t}} = 0$ dB.

Fig. 7 illustrates the sum rate under different values of M, where we train the BGNN with fixed $K = 4$ and variable $M \in \bar{\mathcal{M}} = \{4, 5, \ldots, 8\}$. We observe that the sum rates of all methods increase with the increase of M. This is because a larger value of M with fixed N results in a larger $N_{\text{t}} = NM$, which provides a larger antenna array gain. The sum rate of the BGNN trained with variable $M \in \bar{\mathcal{M}}$ is close to that of the BGNN trained with data samples that have fixed values of K and M. Besides, the BGNN trained with variable M achieves satisfactory performance under different values of M. These performance indicates that the proposed BGNN generalizes well to the systems with different values of M.

We now study the scalability of the proposed BGNN to different values of K. We train the BGNN with fixed $M = 8$ and variable $K \in \bar{\mathcal{K}} = \{3, 4, \ldots, 7\}$, and evaluate its performance versus K in Fig. 8. We observe that the sum rates of benchmark algorithms start to decrease with the increase of K when K is larger than 5. Since the proposed BGNN is trained through unsupervised learning that directly maximizes the sum rate and the digital beamforming is obtained based on the optimal solution as in (10) that has robust performance versus SNR, the proposed BGNN achieves a higher sum rate as K increases. Despite performing slightly worse than the BGNN trained with fixed values of K, the BGNN trained with variable $K \in \bar{\mathcal{K}}$ achieves a satisfactory sum rate compared to the benchmark algorithms, especially when $K \geq 5$. These performance demonstrates the ability of the proposed BGNN for generalizing to different values of K.

The numerical results in Figs. 7 and 8 well verify the scalability of the proposed BGNN to different values of K and M. Besides, these results further demonstrate the effectiveness of the proposed BGNN over various system setups.

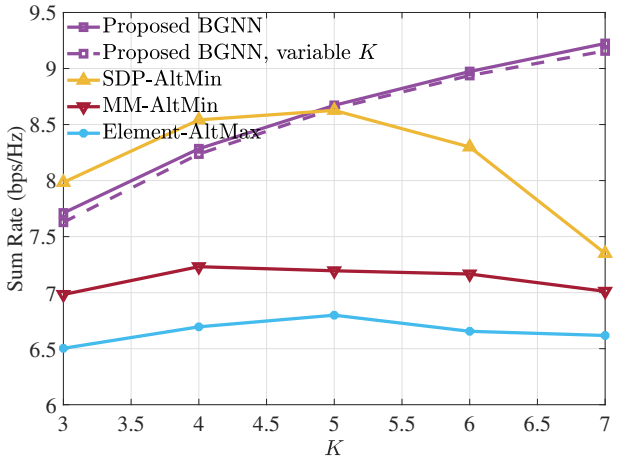

Fig. 8. Achievable sum rates of the proposed BGNN and benchmark algorithms versus K when $M = 8$, $N = 4$, and $P_{\mathrm{t}} = 0$ dB.

V. Conclusion

In this paper, we proposed an effective and scalable BGNN framework for low-complexity hybrid beamforming design to maximize the achievable sum rate in the MU-MISO systems. To capture the *permutation invariance* of analog beamforming and *permutation equivalence* of digital beamforming, we model the MU-MISO communication systems as a bipartite graph, where the users and RF chains are characterized as vertices and the corresponding channels as the edges. Benefiting from exploiting the graph topological structure and interactions between users and RF chains, the proposed BGNN can achieve a satisfactory performance with much lower computational complexity than the complex optimization-based algorithms. With the carefully designed vertex operators, the proposed BGNN enjoys good scalability to both the user number and RF chain number. Numerical results verified the advantages of the proposed BGNN framework on the computational complexity, effectiveness, and scalability.

References

[1] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, C. I, and A. A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, 2017.

[2] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 2, pp. 836–869, 2018.

[3] Z. Chen, X. Ma, B. Zhang, Y. Zhang, Z. Niu, N. Kuang, W. Chen, L. Li, and S. Li, "A survey on terahertz communications," *IEEE China Commun.*, vol. 16, no. 2, pp. 1–35, 2019.

[4] Y. Wu, R. Schober, D. W. K. Ng, C. Xiao, and G. Caire, "Secure massive MIMO transmission with an active eavesdropper," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3880–3900, 2016.

[5] P. Popovski, C. Stefanovic, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, 2019.

[6] S. Han, C. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, 2015.

[7] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, 2017.

[8] X. Zhao, T. Lin, Y. Zhu, and J. Zhang, "Partially-connected hybrid beamforming for spectral efficiency maximization via a weighted MMSE equivalence," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 12, pp. 8218–8232, 2021.

[9] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. H. Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wirel. Commun.*, vol. 13, no. 3, pp. 1499–1513, 2014.

[10] X. Yu, J. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 3, pp. 485–500, 2016.

[11] D. H. N. Nguyen, L. B. Le, T. Le-Ngoc, and R. W. H. Jr., "Hybrid MMSE precoding and combining designs for mmwave multiuser systems," *IEEE Access*, vol. 5, pp. 19 167–19 181, 2017.

[12] S. Huang, Y. Ye, and M. Xiao, "Hybrid beamforming for millimeter wave multi-user MIMO systems using learning machine," *IEEE Wirel. Commun. Lett.*, vol. 9, no. 11, pp. 1914–1918, 2020.

[13] Z. Zhang, X. Wu, and D. Liu, "Joint precoding and combining design for hybrid beamforming systems with subconnected structure," *IEEE Syst. J.*, vol. 14, no. 1, pp. 184–195, 2020.

[14] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for mmwave OFDM large-scale antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1432–1443, 2017.

[15] Y. Liu, Q. Feng, Q. Wu, Y. Zhang, M. Jin, and T. Qiu, "Energy-efficient hybrid precoding with low complexity for mmwave massive MIMO systems," *IEEE Access*, vol. 7, pp. 95 021–95 032, 2019.

[16] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wirel. Commun. Lett.*, vol. 3, no. 6, pp. 653–656, 2014.

[17] J. Du, W. Xu, H. Shen, X. Dong, and C. Zhao, "Hybrid precoding architecture for massive multiuser MIMO with dissipation: Sub-connected or fully connected structures?" *IEEE Trans. Wirel. Commun.*, vol. 17, no. 8, pp. 5465–5479, 2018.

[18] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 3, pp. 501–513, 2016.

[19] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201–211, 2016.

[20] X. Wu, D. Liu, and F. Yin, "Hybrid beamforming for multi-user massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3879–3891, 2018.

[21] J. Zhan and X. Dong, "Interference cancellation aided hybrid beamforming for mmwave multi-user massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2322–2336, 2021.

[22] T. Peken, S. Adiga, R. Tandon, and T. Bose, "Deep learning for SVD and hybrid beamforming," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 10, pp. 6621–6642, 2020.

[23] A. M. Elbir and K. V. Mishra, "Joint antenna selection and hybrid beamformer design using unquantized and quantized deep learning networks," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 3, pp. 1677–1688, 2020.

[24] H. Hojatian, J. Nadal, J. Frigon, and F. Leduc-Primeau, "Unsupervised deep learning for massive MIMO hybrid beamforming," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 11, pp. 7086–7099, 2021.

[25] Y. Shi, L. Lian, Y. Shi, Z. Wang, Y. Zhou, L. Fu, L. Bai, J. Zhang, and W. Zhang, "Machine learning for large-scale optimization in 6g wireless networks," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2023.

[26] A. M. Elbir, "CNN-based precoder and combiner design in mmwave MIMO systems," *IEEE Commun. Lett.*, vol. 23, no. 7, pp. 1240–1243, 2019.

[27] Q. Wang, X. Li, S. Jin, and Y. Chen, "Hybrid beamforming for mmwave MU-MISO systems exploiting multi-agent deep reinforcement learning," *IEEE Wirel. Commun. Lett.*, vol. 10, no. 5, pp. 1046–1050, 2021.

[28] Z. Bo, R. Liu, Y. Guo, M. Li, and Q. Liu, "Deep learning based low-resolution hybrid precoding design for mmwave MISO systems," in *IEEE Globecom Workshops*, Taiwan, Dec. 2020, pp. 1–6.

[29] Y. Zhou, Y. Shi, H. Zhou, J. Wang, L. Fu, and Y. Yang, "Towards scalable wireless federated learning: Challenges and solutions," *IEEE Internet of Things Mag.*, 2023, to appear.

[30] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "Graph neural networks for scalable radio resource management: Architecture design and theoretical analysis," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 101–115, 2021.

[31] T. Jiang, H. V. Cheng, and W. Yu, "Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1931–1945, 2021.

[32] J. Kim, H. Lee, S. Hong, and S. Park, "A bipartite graph neural network approach for scalable beamforming optimization," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 1, pp. 333–347, 2023.
[33] Y. Shen, J. Zhang, S. Song, and K. B. Letaief, "Graph neural networks for wireless communications: From theory to practice," *IEEE Trans. Wirel. Commun.*, pp. 1–1, 2022.
[34] W. U. Z. Bajwa, J. D. Haupt, A. M. Sayeed, and R. D. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, 2010.
[35] J. Lee, G. Gil, and Y. H. Lee, "Channel estimation via orthogonal matching pursuit for hybrid MIMO systems in millimeter wave communications," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2370–2386, 2016.
[36] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. H. Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 831–846, 2014.
[37] Z. Xiao, P. Xia, and X. Xia, "Codebook design for millimeter-wave channel estimation with hybrid precoding structure," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 1, pp. 141–153, 2017.
[38] Y. Zhou, V. W. S. Wong, and R. Schober, "Coverage and rate analysis of millimeter wave NOMA networks with beam misalignment," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 12, pp. 8211–8227, 2018.
[39] A. Forenza, D. J. Love, and R. W. H. Jr., "Simplified spatial correlation models for clustered MIMO channels with different array configurations," *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 1924–1934, 2007.
[40] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 1, pp. 4–24, 2021.
[41] E. Björnson, M. Bengtsson, and B. E. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, 2014.
[42] Z. Wang, J. Qiu, Y. Zhou, Y. Shi, L. Fu, W. Chen, and K. B. Letaief, "Federated learning via intelligent reflecting surface," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 2, pp. 808–822, 2022.
[43] Z. Wang, Y. Zhou, Y. Zou, Q. An, Y. Shi, and M. Bennis, "A graph neural network learning approach to optimize RIS-assisted federated learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6092 – 6106, Sep. 2023.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.
[45] B. Matthiesen, A. Zappone, K. Besser, E. A. Jorswieck, and M. Debbah, "A globally optimal energy-efficient power control framework and its efficient implementation in wireless interference networks," *IEEE Trans. Signal Process.*, vol. 68, pp. 3887–3902, 2020.
[46] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1866–1880, 2020.
[47] D. H. N. Nguyen, L. B. Le, and T. Le-Ngoc, "Hybrid MMSE precoding for mmwave multiuser MIMO systems," in *IEEE ICC*, Kuala Lumpur, Malaysia, May. 2016, pp. 1–6.
[48] Z. Luo, W. Ma, A. M. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, 2010.
[49] Y. Ma, Y. Shen, X. Yu, J. Zhang, S. Song, and K. B. Letaief, "Learn to communicate with neural calibration: Scalability and generalization," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 11, pp. 9947–9961, 2022.