Sub-6GHz Assisted mmWave Hybrid Beamforming with Heterogeneous Graph Neural Network

Zhaohui Huang, Zhaocheng Wang, *Fellow, IEEE*, and Sheng Chen, *Life Fellow, IEEE*

***Abstract*—In next-generation communications, sub-6GHz and millimeter-wave (mmWave) links typically coexist, with the sub-6GHz link always active and the mmWave link active when high-rate transmission is required. Due to the spatial similarities between sub-6GHz and mmWave channels, sub-6GHz channel information can be utilized to support hybrid beamforming in mmWave communications to reduce overhead costs. We consider a multi-cell heterogeneous communication network where both sub-6GHz and mmWave communications co-exist. Multiple mmWave base stations (BSs) in the heterogeneous network simultaneously transmit signals to multiple users in their own mmWave cells while interfering with each other. The challenging problem is to design hybrid beamformers in the mmWave band that can maximize the system spectral efficiency. To address this highly complex programming using sub-6GHz information, a novel heterogeneous graph neural network (HGNN) architecture is proposed to learn the intrinsic relationship between sub-6GHz and mmWave and design the hybrid beamformers for mmWave BSs. The proposed HGNN consists of two different node types, namely, BS nodes and user equipment (UE) nodes, and two different edge types, namely, desired link edge and interfering link edge. In addition, the attention mechanism and the residual structure are utilized in the HGNN architecture to improve the performance. Simulation results show that the proposed HGNN can successfully achieve better performances with sub-6GHz information than traditional learning methods. The results also demonstrate that the attention mechanism and residual structure improve the performances of the HGNN compared to its unmodified counterparts.**

***Index Terms*—Hybrid beamforming, millimeter wave communications, out-of-band information, graph neural network (GNN), machine learning**

I. Introduction

Because of its huge accessible bandwidth, millimeter-wave (mmWave) transmission has been identified as a major approach for next-generation wireless communications [1]. To overcome the high path loss of mmWave signals, mmWave communication systems typically employ huge antenna arrays and directional beamforming/precoding [2]. However, huge antennas result in high power consumption of radio frequency (RF) components in fully digital baseband precoding modules and high overhead costs caused by channel state information (CSI) estimation for beamforming/precoding.

This work was supported in part by the National Natural Science Foundation of China under Grant U22B2057, in part by Guangdong Optical Wireless Communication Engineering and Technology Center, in part by Shenzhen VLC System Key Laboratory, and in part by Shenzhen Solving Challenging Technical Problems (JSGG20191129143216465) *(Corresponding author: Zhaocheng Wang)*.

Z. Huang, and Z. Wang are with Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. Z. Wang is also with the Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (E-mails: hzh21@mails.tsinghua.edu.cn, zcwang@tsinghua.edu.cn).

Sheng Chen is with School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (E-mail: sqc@ecs.soton.ac.uk).

In order to reduce the power consumption of RF components while achieving satisfactory performance, the hybrid beamforming technique is utilized in the mmWave MIMO systems [3], [4]. Typically, either fully-connected or partially-connected architectures are adopted, depending on whether each RF chain is connected to all antennas (the former) or to a disjoint subset of antennas (the latter) [5], [6]. There are several techniques to obtain the hybrid beamforming configuration [7]–[11]. The traditional methods, such as [7], [9], [11], usually rely on optimization theory to derive optimal or near optimal iterative algorithms for solutions, which impose high computational complexity and require the whole antenna array's CSI or the optimal fully digital beamforming configuration. To reduce the computational complexity, the work [10] proposed a deep unfolding based architecture, which replaces some original components with their learnable counterparts, and achieves faster convergence than traditional methods. However, these techniques all require the entire antenna array's CSI, leading to high overhead cost.

To reduce the overhead of establishing a mmWave link, various approaches exploiting channel sparsity were proposed [12]–[14]. These methods focus on obtaining a compressed sensing channel estimation that trades the measurement overhead for performance. Another technique to reduce the training overhead is to leverage the out-of-band information extracted from low-frequency channels [15], [16]. Experiments in [15], [17] demonstrated the feasibility of using the low-frequency information due to the similarities of spatial characteristics of sub-6 GHz and mmWave channels. The study [15] also showed that the power azimuth spectrums (PASs) are almost consistent in low-frequency and mmWave channels. In addition, the work [18] showed the close relationship of the power delay profiles (PDPs) between the sub-6GHz channel and the mmWave channel. With the aforementioned similar characteristics of low-frequency and mmWave channels, the authors of [19] demonstrated the construction of the overhead-free multi-Gbps mmWave link with out-of-band inference, while the authors of [20] showed the selection of the mmWave beam based on the estimated line-of-sight (LOS) direction of sub-6GHz channel. In addition, machine learning methods were also utilized to take advantage of the out-of-band information [16], [21], [22]. For example, the work [21] developed a dedicated deep learning model based on 3-dimensional (3D) convolutional neural network (CNN) to provide the mmWave beam selection with the aid of low-frequency information. The authors of [16] proposed a dual-input neural network to predict the optimal beam using both the sub-6GHz channel and a few pilots in the mmWave band transmitted from a few active antennas. The work [22] proved the existence of the mapping functions that can predict the optimal mmWave beam

and blockage status directly from the sub-6GHz channel under certain conditions, and developed a deep learning model that can predict the optimal mmWave beam and blockage with high success probability. However, the application of [16], [21], [22] is limited to the single-user scenario.

It is challenging to efficiently utilize the sub-6GHz CSI in mmWave hybrid beamforming, because to our knowledge the correlation between sub-6GHz and mmWave has not be derived in closed-form formulas, and hybrid beamforming itself is usually a non-convex problem due to mutual interference among links, coupling, and constraints of optimization variables.

Graph neural network (GNN) [23] provides a structural learning framework for graph-based problems. Through convolution operations between adjacent nodes, GNN can extract local and global information on the graph data, achieving classification, regression or prediction tasks. From a theoretical perspective, there are three advantages of GNNs compared to CNNs and multilayer perceptions (MLPs) in many communication problems.

1) Scalability and parallelization: The wireless communication network can be naturally modeled as a graph, where nodes correspond to base stations (BSs) and user equipment (UEs) with edges representing the channels between them. The graph convolution structure of GNN ensures its adaptability to input graph data with varying numbers of nodes (i.e., UEs and/or BSs), whereas MLP and CNN are constrained by strict dimensional requirements for their inputs. In addition, GNNs allow for parallel computation on graph data, contributing to a more efficient runtime performance.
2) Permutation invariance (PI) and permutation equivalence (PE): A function or model has PI property if its output remains unchanged when the order of its input elements is altered. Conversely, a function or model has PE property if the order of its outputs is correspondingly altered when its input elements are permuted. The PI and PE properties are proved to be universal in many wireless communication problems, such as power allocation, beamforming, and interference mitigation. The GNN can achieve PI or PE properties, distinguishing itself from CNN and MLP, which lack these properties. Consequently, the GNN emerges as a more suitable solution for numerous wireless communication problems.
3) Requiring less training samples: It is theoretically proved that GNNs require fewer training samples than MLPs to achieve the same performance, and the training-samples demand gap grows with the number of nodes in the graph [24]. A reduced number of training samples means lower overhead and less computational complexity, and therefore the GNNs are advantageously in wireless communication problems.

Empirically, compared with other deep learning neural networks (DNNs), such as CNN and MLP, GNN provides high performance in various wireless communication problems, such as power allocation, beamforming, reconfigurable intelligent surface (RIS) configurations [25]–[28]. For example, the work [25] proposed a wireless channel graph convolutional network (WCGCN) to solve the power allocation and beamforming problem in device-to-device (D2D) communication systems. The study [26] proposed a parameter sharing structure as a heterogeneous GNN (HGNN) for learning power control in cellular systems. Moreover, the authors of [28] unfolded a power allocation enabled iterative weighted minimum mean squared error (WMMSE) algorithm with a distributed GNN architecture, which reduces the computational complexity and has robustness and generalizability in different densities and sizes. The research [28] also showed the relationship between the GNN and deep unfolding techniques.

For heterogeneous wireless networks, the GNNs with some specific modifications, such as HGNNs and bipartite GNNs (BGNNs), offer advantages over traditional methods or other deep learning methods. For example, the work [29] proposed an unsupervised HGNN to solve the power control/beamforming in heterogeneous D2D networks, showing the applicability of dedicated GNNs in heterogeneous networks. In addition, the authors of [27] utilized the GNN to appropriately schedule users and design RIS configurations to achieve high overall throughput while considering fairness among the users, demonstrating the potential of GNNs in asymmetric communication problems. Furthermore, the study [30] proposed a BGNN to solve the scalable multi-antenna beamforming optimization problems, exhibiting the GNN's flexibility with respect to the system size, i.e., the number of antennas or users.

Inspired by these related works on GNNs for heterogeneous wireless networks, in this paper, we propose a HGNN, which consists of two different node types, namely, BS nodes and UE nodes, and two different edge types, namely, desired link edge and interfering link edge. Our HGNN achieves the fully-connected or partially-connected hybrid beamforming configurations in mmWave with the aid of both sub-6GHz CSI and partial mmWave CSI. We utilize this HGNN to solve the challenging hybrid beamforming problem in heterogeneous mmWave and sub-6GHz networks. The main contributions are summarized as follows.

1) The hybrid beamforming problem is formulated to maximize the spectral efficiency of the mmWave system in the heterogeneous cellular network (HCN) with power constraints. We model the HCN as a heterogeneous graph, where BSs and UEs are modeled as nodes, while desired links and interfering links represent the edges between the corresponding nodes. Based on this heterogeneous graph, the HGNN is proposed to solve the fully-connected or partially-connected hybrid beamforming problem. To reduce the overhead, we leverage the sub-6GHz CSI and the partial mmWave CSI in solving this optimization design.
2) To improve the performance of the proposed HGNN, the attention mechanism and the residual structure are utilized in the GNN structure. The attention mechanism is added in the aggregation procedure to adaptively learn the importance of different messages. The residual structure is added in the combination procedure to eliminate the degradation phenomenon. The introduction

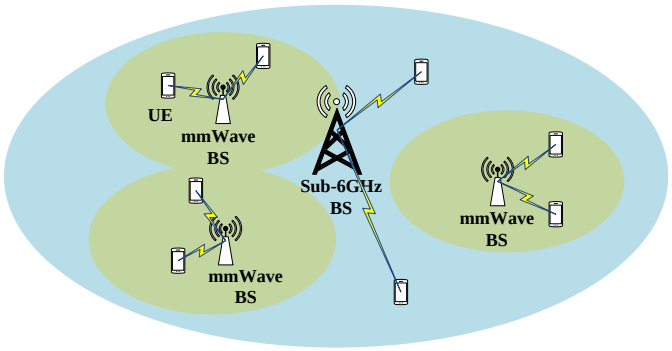

Fig. 1. The heterogeneous mmWave and sub-6GHz network.

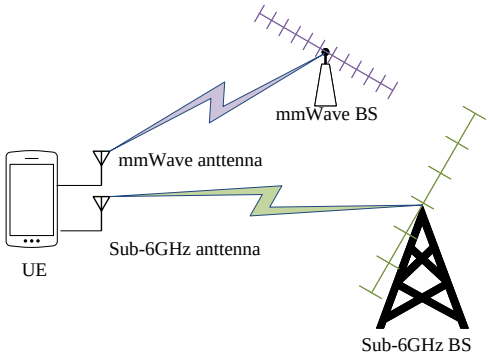

Fig. 2. UE communicates over both sub-6GHz and mmWave bands with corresponding BSs.

of attention mechanism and residual structure does not affect the PI and PE properties as well as scalability of GNNs.

3) Numerical results verify that the proposed HGNN outperforms other machine learning methods in various scenarios. Moreover, the utilization of attention mechanism and residual structure is shown to enhance the achievable performance. Besides, the strong scalability and low running complexity of the proposed HGNN are demonstrated.

The rest of this paper is organized as follows. Section II describes the heterogeneous mmWave and sub-6GHz network and formulates the system spectral efficiency maximization problem. The proposed HGNN is detailed in Section III. In Section IV, numerical results are presented to verify the effectiveness of the proposed HGNN. Finally, Section V concludes the paper.

Notations: Scalars, vectors and matrices are represented by normal face lowercase letters, boldface lowercase letters and boldface uppercase letters, respectively, e.g., a, $\mathbf{a}$ and $\mathbf{A}$. $\mathbb{C}^{m\times n}$ and $\mathbb{R}^{m\times n}$ denote the m by n dimensional complex space and real space, respectively. $\mathbf{I}_k$ is the $k\times k$ identity matrix, and $\mathrm{j}=\sqrt{-1}$ is the imaginary axis. The transpose and Hermitian transpose are denoted by $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$, respectively. The complex normal distribution is represented by $\mathcal{CN}(\mu,\sigma^2)$ with mean μ and variance σ^2. $\mathbb{E}(\cdot)$ and $\|\cdot\|_F$ are the expectation and Frobenius norm, respectively, while $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$. The mth-row and nth-column entry of $\mathbf{A}$ is denoted by $\mathbf{A}_{[m,n]}$. The concatenation of two vectors $\mathbf{a}$ and $\mathbf{b}$ is given by $[\mathbf{a}\,\|\,\mathbf{b}]=\left[\mathbf{a}^{\mathrm{T}},\mathbf{b}^{\mathrm{T}}\right]^{\mathrm{T}}$.

II. System Model and Problem Formulation

Consider a heterogeneous sub-6GHz and mmWave communication system, with K BSs equipped with N_m mmWave antennas and one BS equipped with N_s sub-6GHz antennas, as illustrated in Fig. 1. The coverage of sub-6GHz signals is the entire network while the coverage of mmWave signals is only in the corresponding mmWave cell. The sub-6GHz antenna array is fully digital, with each antenna connected to an independent RF chain, and the mmWave antenna array is a hybrid architecture. All the users are equipped with both a sub-6GHz antenna and a mmWave antenna, as shown in Fig. 2. Both sub-6GHz and mmWave communications are considered under single-carrier system for simplicity. Nevertheless, our proposed method can be easily extended to an OFDM system. It is assumed that all the UEs are constantly connected to the central sub-6GHz BS in sub-6GHz, and the mmWave connection is only active when a high transmission rate is required. For simplicity, we omit the sub-6GHz links in Fig. 1 for the UEs with both mmWave and sub-6GHz links. MmWave BS k, $k=1,2,\cdots,K$, can serve up to I_k users in cell k with mmWave channel. The total number of UEs is hence given by $I_{sum}=\sum_{k=1}^{K}I_k$. The ith user in mmWave cell k is denoted as i_k. Our goal is to find the mmWave hybrid beamforming configurations by utilizing sub-6GHz CSI and some estimation of partial mmWave CSI. The sub-6GHz beamforming can be handled by the WMMSE [7] or other traditional methods.

A. Hybrid beamforming in the transmitter at mmWave band

We assume that mmWave BS k with $N_{RF,k}$ RF chains communicates with each UE via only one stream. Thus, the number of data streams of BS k is I_k. Moreover, the number of RF chains of BS k is usually larger than the number of users that can be served simultaneously by BS k, i.e., $I_k\le N_{RF,k}$. For simplicity, we also assume that BS k will utilize the I_k RF chains to serve the corresponding I_k UEs.

Let $\mathbf{s}_k=[s_k[1],s_k[2],\cdots,s_k[I_k]]^{\mathrm{T}}\in\mathbb{C}^{I_k\times 1}$ denote the transmitted symbols of BS k, where $\mathbb{E}\left(\mathbf{s}_k\mathbf{s}_k^{\mathrm{H}}\right)=\mathbf{I}_{I_k}$ and $s_k[i]$ is the transmitted data for UE i_k. Assuming that the hybrid precoder of BS k is $\mathbf{F}_k\in\mathbb{C}^{N_m\times I_k}$, the precoded signal of BS k is given by

$$\mathbf{x}_k=\mathbf{F}_k\mathbf{s}_k. \tag{1}$$

The hybrid precoder $\mathbf{F}_k=\mathbf{F}_{RF,k}\mathbf{F}_{BB,k}$ is composed of the analog precoder $\mathbf{F}_{RF,k}\in\mathbb{C}^{N_m\times I_k}$ and the baseband precoder $\mathbf{F}_{BB,k}\in\mathbb{C}^{I_k\times I_k}$. We denote $\mathbf{F}_k=[\mathbf{f}_k[1],\mathbf{f}_k[2],\cdots,\mathbf{f}_k[I_k]]$ and $\mathbf{F}_{RF,k}=[\mathbf{f}_{RF,k}[1],\mathbf{f}_{RF,k}[2],\cdots,\mathbf{f}_{RF,k}[I_k]]$.

In our system model, we design GNNs for both fully-connected hybrid beamforming structure and partially-connected structure, which are shown in Fig. 3.

1) Fully-connected structure: In the fully-connected hybrid beamforming structure [31], each RF chain is connected to all antennas via phase shifters, which implies that each entry of $\mathbf{F}_{RF,k}$ has a constant modulus. All the entries are normalized to satisfy $\left|\mathbf{F}_{RF,k}{}_{[m,n]}\right|^2=\frac{1}{N_m}$, i.e., $\mathbf{F}_{RF,k}{}_{[m,n]}=$

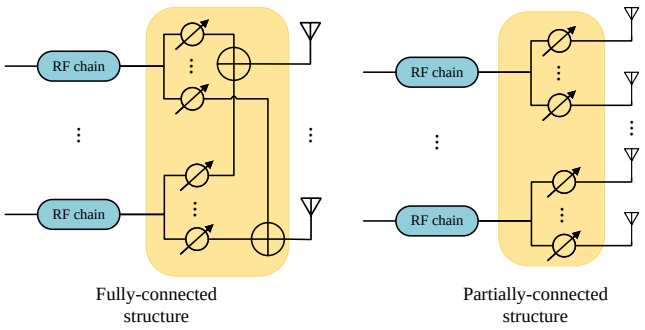

Fig. 3. The fully-connected and partially-connected hybrid beamforming structures in the transmitter.

$\frac{1}{\sqrt{N_m}}e^{\mathrm{j}\varphi_{k,m,n}}$, where $\varphi_{k,m,n} \in \mathbb{C}$ represents the phase of the m-th phase shifter in the n-th RF chain of the k-th BS. $\mathbf{F}_{BB,k}$ is forced to satisfy $\left\|\mathbf{F}_{RF,k}\mathbf{F}_{BB,k}\right\|_F^2 \le P_k$, where P_k is the maximum transmitting power of BS k.

2) Partially-connected structure: The partially-connected structure, also known as array of subarray structure, uses significantly fewer phase shifters for energy efficiency [32]–[34]. This structure connects each RF chain output signal to $N_m/N_{RF,k}$ antennas, which reduces the RF hardware complexity. The analog precoder $\mathbf{F}_{RF,k}$ can be written as a block diagonal matrix as

$$\mathbf{F}_{RF,k} = \begin{pmatrix} \underline{\mathbf{f}}_{1,k} & \underline{\mathbf{0}} & \cdots & \underline{\mathbf{0}} \\ \underline{\mathbf{0}} & \underline{\mathbf{f}}_{2,k} & \cdots & \underline{\mathbf{0}} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\mathbf{0}} & \underline{\mathbf{0}} & \cdots & \underline{\mathbf{f}}_{N_{RF,k},k} \end{pmatrix}, \tag{2}$$

where the row vector $\underline{\mathbf{f}}_{i,k} = \frac{1}{\sqrt{N_m}}\left[e^{\mathrm{j}\varphi_{k,(i-1)\frac{N_m}{N_{RF,k}}+1}}, \cdots, e^{\mathrm{j}\varphi_{k,i\frac{N_m}{N_{RF,k}}}}\right] \in \mathbb{C}^{1\times\frac{N_m}{N_{RF,k}}}$ represents the phases of the i-th RF chain at the k-th BS, and $\underline{\mathbf{0}} \in \mathbb{C}^{1\times\frac{N_m}{N_{RF,k}}}$ is the zero row vector. Similar to the fully-connected case, $\mathbf{F}_{BB,k}$ is forced to satisfy the power constraint of $\left\|\mathbf{F}_{RF,k}\mathbf{F}_{BB,k}\right\|_F^2 \le P_k$.

The received signal y_{i_k} at the receiver i_k can be formulated as

$$\begin{aligned} y_{i_k} &= \sum_{m=1}^{K} \mathbf{h}_{i_k,m}^{\mathrm{H}}\mathbf{x}_m + n_{i_k} \\ &= \mathbf{h}_{i_k,k}^{\mathrm{H}}\mathbf{f}_k[i_k]\mathbf{s}_k[i_k] + \sum_{l=1,l\neq i_k}^{I_k} \mathbf{h}_{i_k,k}^{\mathrm{H}}\mathbf{f}_k[l]\mathbf{s}_k[l] + \\ &\quad \sum_{m=1,m\neq k}^{K}\sum_{l=1}^{I_m} \mathbf{h}_{i_k,m}^{\mathrm{H}}\mathbf{f}_m[l]\mathbf{s}_m[l] + n_{i_k}, \end{aligned} \tag{3}$$

where $\mathbf{h}_{i_k,k} \in \mathbb{C}^{N_m\times 1}$ denotes the channel from BS k to UE i_k, n_{i_k} denotes the additive white Gaussian noise (AWGN) with the distribution $\mathcal{CN}(0,\sigma_{i_k}^2)$. On the right hand side of the last equation of (3), the four terms are the desired signal, intracellular interference, intercellular interference and noise, respectively.

B. MmWave channel model

The mmWave channel from BS k to UE i_k can be formulated as [16]

$$\mathbf{h}_{i_k,k} = \sqrt{\frac{N_m}{N_c}}\sum_{l=1}^{N_c}\beta_l e^{\mathrm{j}(\theta_l+2\pi\tau_l B)}\mathbf{a}(\phi_l), \tag{4}$$

where N_c is the number of paths, B is the bandwidth, β_l, θ_l and τ_l are the attenuation coefficient, phase and propagation delay of the l-th path, respectively, while ϕ_l is the angle of departure (AoD) of the l-th path, and $\mathbf{a}(\phi_l)$, $1 \le l \le N_c$, are the steering vectors at the departure side. In (4), all the index marks of BS k and UE i_k are omitted for clarity. For simplicity, we assume that each BS is equipped with a uniform linear array (ULA) for mmWave communications. Therefore, the array response vector can be formulated as

$$\mathbf{a}(\phi) = \frac{1}{\sqrt{N_m}}\left[1, e^{\mathrm{j}\frac{2\pi}{\lambda}d\sin(\phi)}, \cdots, e^{\mathrm{j}\frac{2\pi}{\lambda}(N-1)d\sin(\phi)}\right]^{\mathrm{T}}, \tag{5}$$

where λ is the signal wavelength and d is the antenna spacing, which is set to $d = \lambda/2$.

C. Problem formulation

Our target is to design the hybrid precoders to maximize the spectral efficiency of the whole system. Let R_{i_k} be the capacity of UE i_k, which is formulated as (6) shown at the top of the next page. Then, the spectral efficiency maximization problem can be formulated as (7) shown at the top of the next page, where the constraint (7a) is the transmitting power constraint for each BS and the constraint (7b1) or (7b2) is the constant modulus constraint for the phase shifters.

Due to the power constraint (7a) and the constant modulus constraint (7b1) or (7b2), the problem (7) is nonconvex, which is challenging to solve. Moreover, we want to solve this nonconvex problem not relying on the full mmWave CSI but relying on the sub-6GHz CSI and some estimation of the partial mmWave CSI, which makes the problem even more challenging.

D. Partial mmWave CSI

For simplicity, we assume that the perfect sub-6GHz CSI is available. Denote the sub-6GHz channel from BS 0 to UE i_k as $\widetilde{\mathbf{h}}_{i_k} \in \mathbb{C}^{N_s\times 1}$. Some works [35], [36] showed the similarities between sub-6GHz channel and mmWave channel and the feasibility of predicting the optimal mmWave channel via sub-6GHz CSI in the single user scenario. But the sub-6GHz CSI alone, which does not contain direct interference information or desired link information between BSs without sub-6GHz antennas and their serving UEs, is completely insufficient to optimize the problem (7). However, it is expensive to obtain the full mmWave CSI, since the full CSI estimation at mmWave band requires changing the phase shifters $\mathbf{F}_{RF,k}$ or baseband precoders $\mathbf{F}_{BB,k}$ at the transmitter N_m times, which imposes huge overhead cost.

Therefore, we follow [16] to use the partial channel information on $\bar{N}_m \le N_m$ antennas at mmWave band, namely, partial mmWave CSI, to assist the hybrid precoding. In the

$$R_{i_k} = \log_2\left(1 + \frac{\|\mathbf{h}_{i_k,k}^{\mathrm{H}}\mathbf{f}_k[i_k]\|^2}{\sum_{l=1,l\neq i_k}^{I_k}\|\mathbf{h}_{i_k,k}^{\mathrm{H}}\mathbf{f}_k[l]\|^2 + \sum_{m=1,m\neq k}^{K}\sum_{l=1}^{I_m}\|\mathbf{h}_{i_k,m}^{\mathrm{H}}\mathbf{f}_m[l]\|^2 + \sigma_{i_k}^2}\right) \tag{6}$$

$$\max_{\forall \mathbf{F}_{RF,k},\mathbf{F}_{BB,k}} \sum_{k=1}^{K}\sum_{i=1}^{I_k} R_{i_k} \tag{7}$$

$$\text{s.t.} \quad \left\|\mathbf{F}_{RF,k}\mathbf{F}_{BB,k}\right\|_F^2 \leq P_k, \forall k \tag{7a}$$

$$\text{for fully-connected structure: } \left\|\mathbf{F}_{RF,k[m,n]}\right\|^2 = \frac{1}{N_m}, \forall k,m,n \tag{7b1}$$

$$\text{for partially-connected structure: } \mathbf{F}_{RF,k} \text{ defined in (2)} \tag{7b2}$$

training process, $\bar{N}_m$ mmWave antennas are active while the others are inactive for each BS. In addition, all the RF chains are only connected to the active antennas through the corresponding phase shifters. Denote the partial channel from BS k to UE i_k as $\bar{\mathbf{h}}_{i_k,k}$. The partial mmWave CSI offers a rough description of the full mmWave CSI and can provide the candidate strong directions for the whole mmWave channel [15], which motivates us to adopt the partial mmWave CSI for hybrid precoding. The channel estimation of $\bar{\mathbf{h}}_{i_k,k}$ can be achieved via the least squares (LS) or linear minimum mean squared error (LMMSE) methods with much fewer pilots than what needed for the estimation of the full mmWave channel $\mathbf{h}_{i_k,k}$. We adopt the partial mmWave CSI and sub-6GHz CSI to solve the optimization problem (7).

III. Proposed HGNN

A. Heterogeneous graph representation of heterogeneous network

According to [37], an information graph can be defined as a directed graph $G = (\mathcal{V}, \mathcal{E})$ with a node type mapping function $\tau : \mathcal{V} \to \mathcal{A}$ and a link type mapping function $\phi : \mathcal{E} \to \mathcal{R}$, where each node $\mathbf{v} \in \mathcal{V}$ belongs to a particular node type $\tau(\mathbf{v}) \in \mathcal{A}$, each link $\mathbf{e} \in \mathcal{E}$ belongs to a particular relation $\phi(\mathbf{e}) \in \mathcal{R}$, and if two links belong to the same relation type, the two links share the same starting node type as well as the same ending node type. If the types of nodes $|\mathcal{A}| > 1$ or the types of relations $|\mathcal{R}| > 1$, the network is called a heterogeneous network; otherwise it is a homogeneous network. The neighborhood of a node $\mathbf{v}_m$ is defined as $\mathcal{N}(\mathbf{v}_m) = \{\mathbf{v}_n | \mathbf{e}_{m,n} \in \mathcal{E}\}$. In addition, we define the subset of the neighborhood for node $\mathbf{v}_m$ with node type $\tau(\mathbf{v}_n) = t$ by $\mathcal{N}_t(\mathbf{v}_m) = \{\mathbf{v}_n | \mathbf{e}_{m,n} \in \mathcal{E}, \tau(\mathbf{v}_n) = t\}$, and the subset of the neighborhood $\mathcal{N}_{t,w}(\mathbf{v}_m) = \{\mathbf{v}_n | \mathbf{e}_{m,n} \in \mathcal{E}, \phi(\mathbf{e}_{m,n}) = w, \tau(\mathbf{v}_n) = t\}$ is defined as the subset of $\mathcal{N}_t(\mathbf{v}_m) = \{\mathbf{v}_n | \mathbf{e}_{m,n} \in \mathcal{E}, \tau(\mathbf{v}_n) = t\}$ where all links have the same relation $\phi(\mathbf{e}_{m,n}) = w$.

As illustrated in Fig. 4, it is natural to model the heterogeneous mmWave and sub-6GHz cellular network described in Section II as a heterogeneous graph. Specifically, there are two different node types, namely, BS node and UE node, and two different edge types, namely, desired link and interfering link. The desired links represent the communication links at mmWave band from BS k to its serving UEs i_k, while the interfering links represent the interfering links at mmWave band from BS k to its non-serving UEs $i_{k'}, k' \neq k$. The links at sub-6GHz band are omitted in this graph. Since each UE has a sub-6GHz link to the BS with sub-6GHz antennas, the sub-6GHz link features can be included in the UE node features.

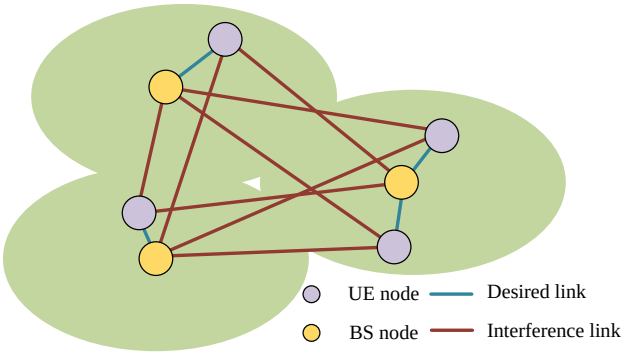

Fig. 4. The heterogeneous graph representation of the communication system described in Section II.

In this heterogeneous graph model for the system described in Section II, the node feature for BS k, which is denoted as $v_k = P_k \in \mathbb{R}^{1\times 1}$, consists of the maximum power of BS k, while the node feature vector of UE i_k, which is denoted as $\mathbf{v}_{i_k} = \left[\sigma_{i_k}^2, \widetilde{\mathbf{h}}_{i_k}^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{(1+N_c)\times 1}$, consists of the noise power and the sub-6GHz CSI from the central (sub-6GHz) BS to UE i_k. The edge feature of the link from BS k to UE i_k consists the partial mmWave CSI, i.e., $\mathbf{e}_{i_k,k} = \bar{\mathbf{h}}_{i_k,k} \in \mathbb{C}^{\bar{N}_m\times 1}$. The node types are $\mathcal{A} = \{\text{b(BS)}, \text{u(UE)}\}$, and the edge types are $\mathcal{E} = \{\text{d(desired link)}, \text{i(interfering link)}\}$.

B. Architecture of the proposed HGNN

In this subsection, we detail the proposed HGNN for solving the optimization problem (7). The architecture of the HGNN, depicted in Fig. 5, consists of three main components: 1) attention based aggregation, 2) res-based combination, and 3) output normalization.

1) Attention based aggregation: The aggregation procedure is utilized to aggregate the information of neighboring nodes and edges. In the traditional HGNN, the aggregated outputs of the $t \in \mathcal{A}$ type of nodes and $w \in \mathcal{E}$ type of edges for node $\mathbf{v}_i$ in layer l are formulated as [26]

$$\mathbf{a}_{i,t,w}^{(l)} = AGG_{j\in\mathcal{N}_{t,w}(\mathbf{v}_i)}\left(p_{t,w}\left(\mathbf{v}_j^{(l-1)}, \mathbf{e}_{i,j}\right)\right), \tag{8}$$

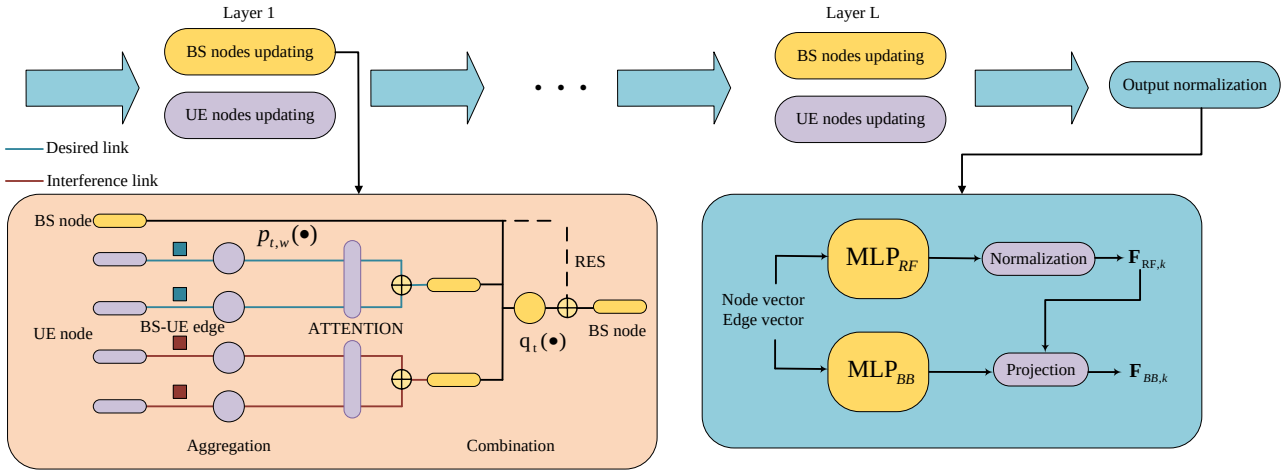

Fig. 5. The architecture of the proposed HGNN.

$$\alpha_{i,j,t,w} = \frac{\exp\left(\mathrm{ReLU}\left(\mathbf{att}_{t,w}^{\mathrm{T}}\left[\mathbf{v}_i^{(l-1)} \,\|\, \mathbf{v}_j^{(l-1)} \,\|\, \mathbf{e}_{i,j}\right]\right)\right)}{\sum_{k\in\mathcal{N}_{t,w}(\mathbf{v}_i)} \exp\left(\mathrm{ReLU}\left(\mathbf{att}_{t,w}^{\mathrm{T}}\left[\mathbf{v}_i^{(l-1)} \,\|\, \mathbf{v}_k^{(l-1)} \,\|\, \mathbf{e}_{i,k}\right]\right)\right)} \tag{10}$$

$$\mathbf{F}_{RF,k[i,j]} = \frac{MLP_{RF}\left(\mathbf{v}_k^{(L)}\right)_{[i,j]}}{\sqrt{N_m}\left|MLP_{RF}\left(\mathbf{v}_k^{(L)}\right)_{[i,j]}\right|}, 1 \le i \le N_m, 1 \le j \le N_{RF,k} \tag{13}$$

$$\mathbf{F}_{RF,k[i,j]} = \begin{cases} \dfrac{MLP_{RF}\left(\mathbf{v}_k^{(L)}\right)_{[i,j]}}{\sqrt{N_m}\left|MLP_{RF}\left(\mathbf{v}_k^{(L)}\right)_{[i,j]}\right|}, & (j-1)\frac{N_m}{N_{RF,k}} + 1 \le i \le j\frac{N_m}{N_{RF,k}}, 1 \le j \le N_{RF,k} \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where $\mathbf{v}_j^{(l-1)}$ denotes the node features in layer $l-1$, $AGG_{j\in\mathcal{N}_{t,w}(\mathbf{v}_i)}(\cdot)$ denotes the aggregated function and $p_{t,w}(\cdot)$ is an MLP, which is identical for all node-edge pairs with node type t and edge type w. $AGG_{j\in\mathcal{N}_{t,w}(\mathbf{v}_i)}(\cdot)$ is a function that satisfies the commutative law and is usually adopted as summation or maximization.

In the traditional aggregation procedure (8), the information of all the neighboring nodes and edges are aggregated with the same importance. However, in the heterogeneous graph, the importance of different messages may be different. Therefore, we further incorporate the attention mechanism into the aggregated function $AGG_{j\in\mathcal{N}_{t,w}\left(\mathbf{v}_i^{(l-1)}\right)}(\cdot)$. The attention function maps the inputs, a query and a set of key-value pairs, onto an output, which is computed as a weighted sum of the input values, where all the query, keys and values are vectors [38]. In the aggregated function $AGG_{j\in\mathcal{N}_{t,w}(\mathbf{v}_i)}(\cdot)$, query is set as $\mathbf{v}_i^{(l-1)}$, keys and values are set as $p_{t,w}(\mathbf{v}_j^{(l-1)}, \mathbf{e}_{i,j})$. By adopting attention, (8) can be rewritten as

$$\mathbf{a}_{i,t,w}^{(l)} = \sum_{j\in\mathcal{N}_{t,w}(\mathbf{v}_i)} \alpha_{i,j,t,w} \cdot p_{t,w}\left(\mathbf{v}_j^{(l-1)}, \mathbf{e}_{i,j}\right), \tag{9}$$

where the weights $\alpha_{i,j,t,w}$, which satisfy $0 \le \alpha_{i,j,t,w} \le 1$ and $\sum_{j\in\mathcal{N}_{t,w}(\mathbf{v}_i)} \alpha_{i,j,t,w} = 1$, represents the importance of the neighborhood $\mathbf{v}_j$ to the original node $\mathbf{v}_i$, and they are calculated as (10) shown at the top of the page, where $\mathrm{ReLU}(x) = \max\{x, 0\}$ is the activation function and $\mathbf{att}_{t,w}$ are the trainable parameters in the attention function.

2) Res-based combination: The combination procedure is utilized to combine the node information with the aggregated information from its neighborhood. In the traditional HGNN, the combination outputs in layer l are formulated as [26]

$$\mathbf{v}_i^{(l)} = COMB_{\tau(\mathbf{v}_i)}\left(\mathbf{v}_i^{(l-1)}, \left\{\mathbf{a}_{i,t,w}^{(l)}, t \in \mathcal{A}, w \in \mathcal{E}\right\}\right), \tag{11}$$

where $COMB_{\tau(\mathbf{v}_i)}(\cdot)$ represents the combination function for node type $\tau(\mathbf{v}_i)$ with trainable parameters.

The traditional combination procedure (11) is a simple concatenation of the node features and the aggregated features. However, when the layers of the HGNN are deep, this structure of combination may suffer from the severe degradation problem [39] and the over-smoothing problem where the outputs of all nodes may have the same features [40]. Therefore, we adopt the residual structure [39] to solve the degradation problem. The output of the res-based structure is a summation of the input and output of the original network. Therefore, with the residual structure, (11) is rewritten as

$$\mathbf{v}_i^{(l)} = \mathbf{v}_i^{(l-1)} + q_{\tau(\mathbf{v}_i)}\left(\mathbf{v}_i^{(l-1)}, \left\{\mathbf{a}_{i,t,w}^{(l)}, t \in \mathcal{A}, w \in \mathcal{E}\right\}\right), \tag{12}$$

where $q_{\tau(\mathbf{v}_i)}(\cdot)$ with $\tau(\mathbf{v}_i) = t$ is an MLP with trainable parameters.

3) Output normalization: After L layers of aggregation and combination, the node features of the last layer are fed into an output normalization layer to obtain the solutions of the problem (7), which consists of two MLPs, $MLP_{RF}(\cdot)$ and $MLP_{BB}(\cdot)$, for obtaining the analog beamforming matrix $\mathbf{F}_{RF,k}$ and the baseband precoding matrix $\mathbf{F}_{BB,k}$, respectively.

3.1) Analog beamforming solution: For the fully-connected structure, the MLP $MLP_{RF}(\cdot)$ with a normalization module is utilized to obtain $\mathbf{F}_{RF,k}$ with input $\mathbf{v}_k^{(L)}$ according to (13) shown at the top of previous page. For the partially-connected structure, the normalization module is only forced on the block diagonal elements and the other elements are set to zero, which can be formulated as (14) shown at the top of previous page.

3.2) Baseband precoding solution: After obtaining $\mathbf{F}_{RF,k}$, the baseband precoding matrix is generated as follows. First, the MLP $MLP_{BB}(\cdot)$ is utilized to produce:

$$\mathbf{f}'_{BB,k}[i_k] = MLP_{BB}\left([\mathbf{v}_k^{(L)} \parallel \mathbf{v}_{i_k}^{(L)}]\right). \tag{15}$$

Denote $\mathbf{F}'_{BB,k} = \left[\mathbf{f}'_{BB,k}[1], \mathbf{f}'_{BB,k}[2], \cdots, \mathbf{f}'_{BB,k}[I_k]\right]$. If the power constraint is satisfied, i.e., $\left\|\mathbf{F}_{RF,k}\mathbf{F}'_{BB,k}\right\|_F^2 \le P_k$, $\mathbf{F}_{BB,k} = \mathbf{F}'_{BB,k}$. Otherwise, the power constraint is imposed on $\mathbf{F}'_{BB,k}$ to produce a feasible baseband precoding solution as:

$$\mathbf{F}_{BB,k} = \frac{\mathbf{F}'_{BB,k}}{\sqrt{P_k}\left\|\mathbf{F}_{RF,k}\mathbf{F}'_{BB,k}\right\|_F}. \tag{16}$$

For all the above MLPs ($p_{t,w}(\cdot), q_t(\cdot), MLP_{RF}, MLP_{BB}, t \in \mathcal{A}, w \in \mathcal{E}$), batch normalization (BatchNorm) is utilized for the inputs to scale them into a similar range and the dropout strategy is adopted to avoid overfitting [41], [42]. The whole architecture of our proposed HGNN is shown in Fig. 5, where the node update process is divided into BS nodes updating and UE nodes updating, both of which consist of aggregation and combination for BS nodes and UE nodes. Note that as each UE is served by only one BS, the attention mechanism in the aggregation for UEs through desired links is degraded into a direct link.

C. Proposed HGNN based system operation

The proposed HGNN based system adopts a centralized deployment, where a central processing unit (CPU) is attached to the sub-6GHz BS and responsible for the hybrid beamforming in the heterogeneous sub-6GHz and mmWave communication system. The CPU is equipped with the proposed HGNN and the parameters of the HGNN are trained offline and fine tuned online.

In the training phase, each mmWave BS estimates the mmWave channel to each UE and transmits the mmWave CSI to the CPU in the sub-6GHz BS. Meanwhile, the sub-6GHz BS estimates the sub-6GHz channel to each UE. The HGNN in the CPU is trained with the both the mmWave and sub-6GHz CSI. The goal of the HGNN is to maximize the system's spectral efficiency, i.e., the problem (7), an unsupervised learning strategy is adopted in the training phase with the loss function given by

$$Loss = -\sum_{k=1}^{K}\sum_{i=1}^{I_k} R_{i_k}. \tag{17}$$

Adam algorithm [43], a gradient descent method, is used to train the parameters of the HGNN. The training procedure is summarized in Algorithm 1.

In the evaluation phase, the parameters of the HGNN and the selection of active mmWave antennas are fixed. The mmWave BSs estimate the partial mmWave CSI and the sub-6GHz BS estimates the sub-6GHz CSI. Both partial mmWave CSI and sub-6GHz CSI are transmitted to the CPU, which utilizes the well-trained HGNN to obtain the hybrid beamforming solution. The hybrid beamforming solution is transmitted to the mmWave BSs, which subsequently perform the hybrid beamforming and transmit the signals to the UEs.

D. Complexity analysis and overhead cost

In the aggregation procedure, we have the MLPs $p_{t,w}(\cdot)$ with $t \in \{\mathrm{b}, \mathrm{u}\}$ and $w \in \{\mathrm{d}, \mathrm{i}\}$, while in the combination procedure, we have the MLPs $q_t(\cdot)$ with $t \in \{\mathrm{b}, \mathrm{u}\}$. Let $C_{p_{\mathrm{b,d}}}$, $C_{p_{\mathrm{u,d}}}$, $C_{p_{\mathrm{b,i}}}$, $C_{p_{\mathrm{u,i}}}$, $C_{q_{\mathrm{b}}}$ and $C_{q_{\mathrm{u}}}$ denote the numbers of floating point operations (FLOPs) of the MLP functions

Algorithm 1 Training of the proposed HGNN

Input: Training dataset $\mathcal{D}$, number of iterations MAX_{iter}, number of GNN layers L, and initialized trainable parameters in $MLP_{BB}(\cdot)$, $MLP_{RF}(\cdot)$, $\mathbf{att}_{t,w}$, $p_{t,w}(\cdot)$, $q_t(\cdot)$, for $t \in \mathcal{A}$, $w \in \mathcal{E}$.

Output: Well-trained parameters in $MLP_{BB}(\cdot)$, $MLP_{RF}(\cdot)$, $\mathbf{att}_{t,w}$, $p_{t,w}(\cdot)$, $q_t(\cdot)$, for $t \in \mathcal{A}$, $w \in \mathcal{E}$.

for $n_{iter} = 1 : MAX_{iter}$ **do**
  Draw a random subset of $\mathcal{D}$.
  Build heterogeneous graph with $v_k^{(0)} = P_k$, $\mathbf{v}_{i_k}^{(0)} = \left[\sigma_{i_k}^2, \widetilde{\mathbf{h}}_{i_k}^{\mathrm{T}}\right]^{\mathrm{T}}$, $\mathbf{e}_{i_k,k} = \bar{\mathbf{h}}_{i_k,k}$.
  for $l = 1 : L$ **do**
    Update BS node features $\mathbf{v}_k^{(l)}$ with $\mathbf{v}_k^{(l-1)}$, $\mathbf{v}_{i_k}^{(l-1)}$ and $\mathbf{e}_{i_k,k}$ through aggregation and combination.
    Update UE node features $\mathbf{v}_{i_k}^{(l)}$ with $\mathbf{v}_k^{(l-1)}$, $\mathbf{v}_{i_k}^{(l-1)}$ and $\mathbf{e}_{i_k,k}$ through aggregation and combination.
  end for
  Compute analog precoder $\mathbf{F}_{RF,k}$ with MLP_{RF}.
  Compute baseband precoder $\mathbf{F}'_{BB,k}$ with MLP_{BB}.
  for $k = 1 : K$ **do**
    if $\|\mathbf{F}_{RF,k}\mathbf{F}'_{BB,k}\|_F^2 \le P_k$ **then** $\mathbf{F}_{BB,k} = \mathbf{F}'_{BB,k}$;
    else
      Project $\mathbf{F}'_{BB,k}$ onto feasible space $\mathbf{F}_{BB,k} = \frac{\mathbf{F}'_{BB,k}}{\sqrt{P_k}\left\|\mathbf{F}_{RF,k}\mathbf{F}'_{BB,k}\right\|_F}$.
    end if
  end for
  Compute loss function $Loss = -\sum_{k=1}^{K}\sum_{i=1}^{I_k} R_{i_k}$.
  Use Adam or gradient descent method to update trainable parameters.
end for

$p_{\mathrm{b,d}}(\cdot)$, $p_{\mathrm{u,d}}(\cdot)$, $p_{\mathrm{b,i}}(\cdot)$, $p_{\mathrm{u,i}}(\cdot)$, $q_{\mathrm{b}}(\cdot)$ and $q_{\mathrm{u}}(\cdot)$, respectively. The FLOPs of an MLP are determined by the number of its layers and the number of neurons in each layer. Furthermore, let $C_{att_{\mathrm{b,d}}}$, $C_{att_{\mathrm{u,d}}}$, $C_{att_{\mathrm{b,i}}}$ and $C_{att_{\mathrm{u,i}}}$ denote the numbers of FLOPs in the corresponding attention mechanisms. In addition, Let C_{BB} and C_{RF} be the numbers of FLOPs in $MLP_{BB}(\cdot)$ and $MLP_{RF}(\cdot)$, respectively. Then the total number of FLOPs for the HGNN is expressed as

$$\begin{aligned} C_{\mathrm{total}} =& L\Big(K\big(I_k\big(C_{p_{\mathrm{b,d}}}+C_{att_{\mathrm{b,d}}}\big) \\ &+\big(I_{sum}-I_k\big)\big(C_{p_{\mathrm{b,i}}}+C_{att_{\mathrm{b,i}}}\big)+C_{q_{\mathrm{b}}}\big) \\ &+I_{sum}\big(\big(C_{p_{\mathrm{u,d}}}+C_{att_{\mathrm{u,d}}}\big) \\ &+(K-1)\big(C_{p_{\mathrm{u,i}}}+C_{att_{\mathrm{u,i}}}\big)+C_{q_{\mathrm{u}}}\big)\Big) \\ &+KC_{RF}+I_{sum}C_{BB}. \end{aligned} \tag{18}$$

In practice, the number of layers L for the HGNN is set to 2 to 4, and the total complexity is mainly dominated by the number of FLOPs in the aggregation part of the HGNN.

The overhead pilots consist of the pilots for sub-6GHz channel estimation and partial mmWave channel estimation. As the sub-6GHz CSI is needed for communication at sub-6GHz band, the extra pilot cost only includes the pilots for partial mmWave CSI estimation, which can be calculated as $K \cdot I_{sum} \cdot Pilots(\bar{N}_m)$, where $Pilots(\bar{N}_m)$ represents the number of pilots for $\bar{N}_m$ mmWave antennas.

The information required to be transmitted among different BSs consists of the partial mmWave CSI, the cost of which can be calculated as $K \cdot I_{sum} \cdot \bar{N}_m$, and the hybrid beamforming matrices, the cost of which can be calculated as $\sum_{k=1}^{K}(N_m \cdot I_k + I_k \cdot I_k)$ for fully-connected beamforming and $\sum_{k=1}^{K}(N_m + I_k \cdot I_k)$ for partially-connected beamforming. It is worth noting that information transmission between BSs is achieved through (wired) backhaul communication, resulting in relatively low communication costs. However, the actual implementation of such backhaul communication is challenging due to hardware costs and synchronization issues, especially in OFDM systems, where the CPU needs to send many precoding matrices to the BSs in a frequent way. Reducing the information exchange between BSs for the proposed HGNN is a challenging objective for the future study.

IV. Simulation Results and Discussions

In this section, we evaluate the performance of our proposed HGNN.

A. Simulation setup

As shown in Fig. 6, we use the 'O1' scenario in the DeepMIMO dataset [44] to generate the sub-6GHz and mmWave channels. The DeepMIMO dataset is built by precise ray-tracing data from Remcom Wireless InSite and relays on the environment geometry/materials, which implies the reliability for machine learning. The sub-6GHz channel and the mmWave channel generated by DeepMIMO share the common environment, which determines the relevance of the sub-6GHz CSI and the mmWave CSI.

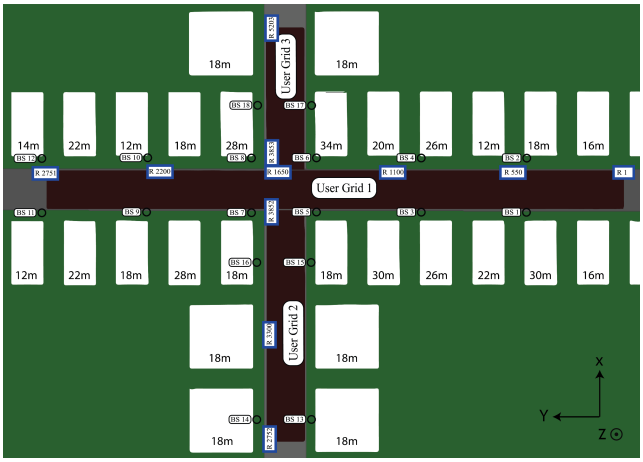

Fig. 6. The 'O1' scenario in [44].

TABLE I
Parameters to generate CSI in DeepMIMO dataset

Parameters	mmWave	sub-6GHz
carrier frequency	28GHz	3.5GHz
activate BSs	BS5 BS6 BS7 BS8	BS5
BS antennas	32	8
Bandwidth	100 MHz	10 MHz
antenna spacing	0.5 wavelength	0.5 wavelength
number of rays	25	25

In the 'O1' scenario, we divide the district from row 1400 to row 2000 in user grid 1 into 4 cells. Each cell is served by one mmWave BS, and all the cells are served by one sub-6GHz BS. The number of active antennas for estimation of partial mmWave CSI is set to $\bar{N}_m = 4$. $K = 2, 3, 4$ BSs are set active and each BS serves $I_k = 2, 4$ UEs. The ULA antennas of both sub-6GHz and mmWave BSs are deployed in the y-axis direction. Other system parameters used in data generation are listed in Table I. We construct a dataset with $11,000$ independent samples, of which $10,000$ samples are used as training dataset and $1,000$ samples are used as test dataset. In each sample, the locations of UEs are randomly and uniformly generated in each cell. The sub-6GHz channels and mmWave channels are generated according to the locations of UEs by the DeepMIMO dataset.

B. Parameters of the proposed HGNN

For both fully-connected hybrid beamforming and partially-connected beamforming, the same network parameters are utilized except for the normalization layer for analog precoders $\mathbf{F}_{RF,k}$. The inputs of both heterogeneous graph networks are a structural heterogeneous graph data consisting of a graph G, a node type mapping function τ and a link type mapping function ϕ, as defined in Subsection III-A. In the following simulations, we use HGNN-FULLY to refer to the HGNN used for fully-connected hybrid precoding, and HGNN-PARTIALLY to refer to the HGNN used for partially-connected hybrid precoding. For the both HGNNs, we set $L = 4$ and the detailed parameters are listed in Table II

TABLE II
PARAMETERS OF THE PROPOSED HGNN AND TRAINING ALGORITHM

Layers L	4
Batch size	10
Learning rate	0.001, $\times 0.9$ every 5 epoch
Sample number of training set	10000
Sample number of test set	1000
Epoch	50
Dropout	0.3
Dimension of node features	600
$p_{t,w}(\cdot), t \in \mathcal{A}, w \in \mathcal{E}$	(608,800,600)
$q_t(\cdot), t \in \mathcal{A}$	(1200,800,600)
MLP_{RF}	(600,600,64)
MLP_{BB}	(1208,50,4)

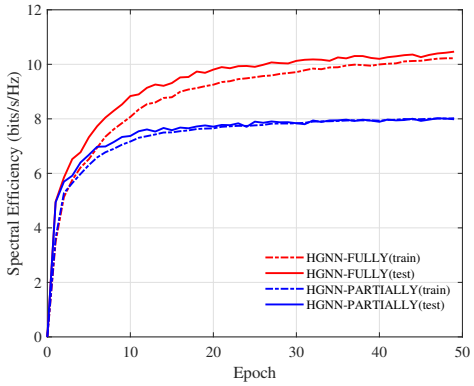

Fig. 7. Training and test learning curves for HGNN-FULLY and HGNN-PARTIALLY, in terms of spectral efficiency.

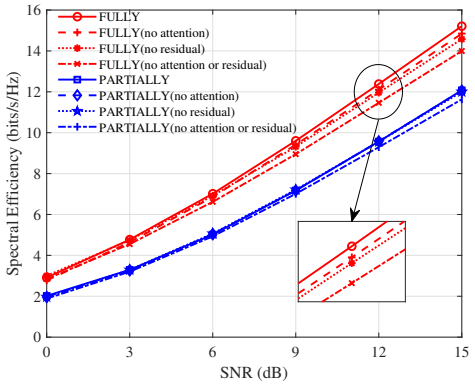

Fig. 8. Impact of attention mechanism and residual structure on the achievable performance of HGNN.

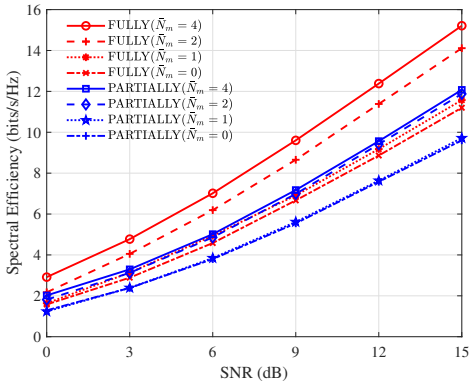

Fig. 9. The spectral efficiency comparison of HGNN with various numbers of active antennas $\bar{N}_m$, through which partially mmWave CSI is estimated.

C. Evaluation of the proposed HGNN

1) Convergence: We first evaluate the convergence of our HGNNs, and Fig. 7 depicts the spectral efficiency learning curves of HGNN-FULLY and HGNN-PARTIALLY. It can be seen that the training of HGNN-FULLY converges at about 40-th epoch while HGNN-PARTIALLY converges at about 20-th epoch. As expected, HGNN-PARTIALLY converges faster than HGNN-FULLY due to fewer outputs of the former. It is interesting to see that the test spectral efficiency on the test set is slightly higher than the training spectral efficiency for both HGNN-FULLY and HGNN-PARTIALLY, which is owing to *Dropout*[1]. In addition, the fully-connected structure achieves higher spectral efficiency than the partially-connected structure as expected.

2) Effectiveness of attention mechanism and residual structure: Fig. 8 investigates the impact of attention mechanism and residual structure on the achievable performance of the HGNN. The aggregation procedure of the HGNN without attention is obtained by replacing (9) with $\mathbf{a}_{i,t,w}^{(l)} = \sum_{j \in \mathcal{N}_{t,w}(\mathbf{v}_i)} p_{t,w}(\mathbf{v}_j^{(l-1)}, \mathbf{e}_{i,j})$, while the combination procedure of the HGNN without residual structure is obtained by replacing (12) with $\mathbf{v}_i^{(l)} = q_{\tau(\mathbf{v}_i)}\left(\mathbf{v}_i^{(l-1)}, \{\mathbf{a}_{i,t,w}^{(l)}, t \in \mathcal{A}, w \in \mathcal{E}\}\right)$. It can be seen that the performance gain of the attention mechanism and the residual structure is more significant in the fully-connected structure than in the partially-connected structure. This is because the optimization problem (7) for the fully-connected structure is more complicated, and there are more degrees of freedom in design to be exploited by the attention mechanism and the residual structure in the fully-connected structure. The results of Fig. 8 hence verify the effectiveness of attention mechanism and residual structure.

3) The effect of amount of partial mmWave CSI: The amount of partial mmWave SCI is measured by $\bar{N}_m$, the number of active antennas in the partial mmWave CSI estimation procedure. More active antennas means more information about the mmWave channel. Fig 9 shows the performance of the HGNN with different numbers of active antennas $\bar{N}_m$. It can be seen that the HGNN works even with only sub-6GHz CSI and no partial mmWave CSI. As expected, with more partial mmWave CSI, the HGNN achieves higher spectral efficiency, at the expense of higher training overhead. It can also be observed that the performance gain from increasing the partial mmWave CSI is more significant for the fully-connected structure than for the partially-connected structure.

[1]Dropout works by randomly dropping out (i.e., setting to zero) some of the neurons in a neural network during training. This is a common technique to assist a deep neural network avoiding overfitting and help improving the generalization, i.e., improving the performance on the data unseen in training. Although the random dropout of neurons may miss some information over connection of neurons, which would cause the training performance 'degradation', this kind of regularization can often avoid overfitting into the features only related to the training data, e.g., noise, and therefore enhances the model generalization capability. This reflects in Fig. 7 where the test spectral efficiency is slightly higher than the training spectral efficiency.

TABLE III
COMPARISON OF TEST SPECTRAL EFFICIENCY (BITS/S/HZ) FOR DIFFERENT METHODS WITH DIFFERENT NUMBERS OF UES AND MMWAVE BSS.

	HGNN-FULLY	HGNN-PARTIALLY	MLP-FULLY	MLP-PARTIALLY	MO-AltMin	SDR-AltMin
2BSs 2UEs	10.2	8.2	4.1	3.8	11.2	8.8
2BSs 8UEs	16.5	10.6	4.3	4.3	18.7	9.0
3BSs 6UEs	16.0	12.0	5.1	4.1	16.8	12.9
3BSs 12UEs	25.1	15.7	4.2	4.3	27.6	12.8
4BSs 8UEs	21.0	15.7	5.3	5.2	21.7	16.6
4BSs 16UEs	32.0	20.2	4.0	3.7	35.0	15.8

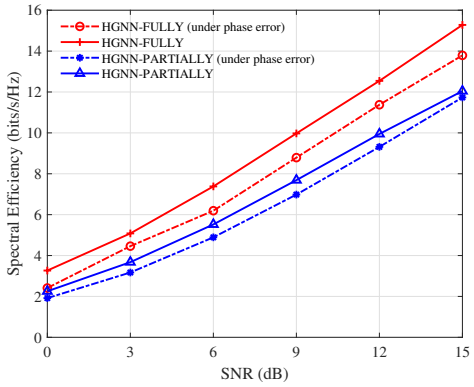

Fig. 10. The spectral efficiency of HGNN with/without the random phase error of CSI.

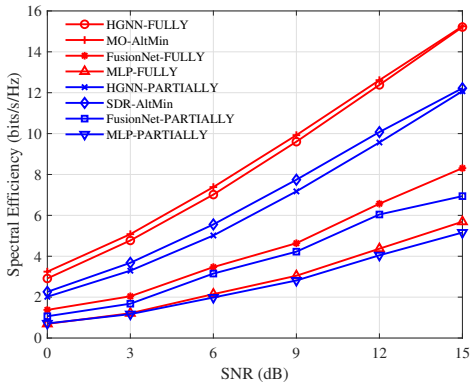

Fig. 11. The spectral efficiency of HGNN, MLP, FusionNet and alternating methods as the functions of SNR.

4) The robustness of the proposed HGNN under phase error: To evaluate the robustness of our proposed method, we further test the performance of our HGNN under the CSI with random phase error. The random phase error is generated by $\mathbf{e}_{i,j} = \mathbf{e}_{i,j} \cdot e^{j\theta}$, where θ is a random variable, following the normal distribution $\mathcal{N}(0, (5^\circ)^2)$. Fig. 10 shows the spectral efficiency of our HGNN with/without the random phase error of CSI. It can be seen that the performance of the HGNN degrade slightly due to the random phase error, which implies that our HGNN is robust to the random phase error of CSI.

D. Performance comparison

1) Comparison benchmarks: We compare our method with the MLP method, FusionNet method [16] and the alternating methods, i.e., MO-AltMin for fully-connected structure and SDR-AltMin for partially-connected structure [34].

In the MLP method, a vanilla MLP is utilized to generate both analog beamforming matrices $\mathbf{F}_{RF,k}$ and digital matrices $\mathbf{F}_{BB,k}$, which are subsequently normalized to satisfy the constant modulus constraint and power constraint. Similarly, MLP-FULLY refers to the MLP for fully-connected hybrid precoding while MLP-PARTIALLY refers to the MLP for partially-connected hybrid precoding. The number of hidden layers for the both MLPs is 3, and the numbers of neurons in the three hidden layers are $200, 300, 500$, respectively. Since MLPs cannot directly deal with heterogeneous graph data, all the node features, edge features, node types, and edge types are concatenated into a vector as the input of MLPs. For fairness, both MLPs are trained in an unsupervised manner with the same loss function (17) as our HGNNs.

In the FusionNet method, we adopt the same network structures as FusionNet in [16], where the sub-6GHz information and mmWave information are fused in an attention layer of the network. The original FusionNet is designed for beam selection in the heterogeneous network where sub-6GHz and mmWave coexist, and we modify its outputs as hybrid precoding matrix. Similarly, the FusionNet-FULLY refers to the FusionNet for fully-connected hybrid precoding while FusionNet-PARTIALLY refers to the FusionNet for partially-connected hybrid precoding. The number of hidden layers for the both FusionNets is 4, and the dimensionality of hidden features for both sub-6GHz and mmWave before fusion is 300. FusionNet-FULLY and FusionNet-PARTIALLY are trained in an unsupervised manner with the same loss function (17).

In [34], the alternating methods were introduced to approximate the optimal fully digital precoding matrix in the metric of Frobenius norm by alternately iterating between the baseband precoding matrix and the analog precoding matrix based on the traditional optimization theory. Specifically, the MO-AltMin is an optimization algorithm based on the manifold optimization theory for fully-connected hybrid beamforming, while the SDR-AltMin is based on the semi-definite relaxation (SDR) algorithm for partially-connected hybrid beamforming. The optimal fully digital precoding matrices for the MO-AltMin and SDR-AltMin are generated by the WMMSE method [7] in the simulation. Note that unlike the MLP method and our HGNN, the entire mmWave CSI is required for the both MO-AltMin and SDR-AltMin methods.

2) Comparison for different SNRs: We first compare the performance of the HGNN, FusionNet, MLP and traditional alternating methods at different SNR levels in Fig. 11. It can be seen that the performance of HGNN-FULLY and HGNN-PARTIALLY are very close to the performance of MO-AltMin

TABLE IV
COMPARISON OF RUN TIME (S) FOR DIFFERENT SCHEMES WITH DIFFERENT NUMBERS OF UES AND MMWAVE BSS.

	HGNN-FULLY	HGNN-PARTIALLY	MLP-FULLY	MLP-PARTIALLY	MO-AltMin	SDR-AltMin
2BSs 4UEs	58.36	2.08	0.22	0.26	96.87	420.22
2BSs 8UEs	62.06	2.35	0.23	0.25	253.29	939.62
3BSs 6UEs	65.17	1.95	0.13	0.26	136.76	562.46
3BSs 12UEs	65.08	2.18	0.24	0.26	356.03	1452.97
4BSs 8UEs	58.38	2.52	0.24	0.25	178.71	775.45
4BSs 16UEs	61.65	2.36	0.24	0.34	629.99	1900.43

TABLE V
OVERHEAD COMPARISON FOR DIFFERENT SCHEMES.

	HGNN	MLP/FusionNet	MO/SDR-AltMin
mmWave pilot overhead	$K \cdot I_{sum} \cdot Pilots(\bar{N}_m)$	$K \cdot I_{sum} \cdot Pilots(\bar{N}_m)$	$K \cdot I_{sum} \cdot Pilots(N)$
Backhaul overhead (fully)	$K \cdot I_{sum} \cdot \bar{N}_m + \sum_{k=1}^{K}(N_m \cdot I_k + I_k^2)$	0	$K \cdot I_{sum} \cdot N_m + \sum_{k=1}^{K}(N_m \cdot I_k + I_k^2)$
Backhaul overhead (partially)	$K \cdot I_{sum} \cdot \bar{N}_m + \sum_{k=1}^{K}(N_m + I_k^2)$	0	$K \cdot I_{sum} \cdot N_m + \sum_{k=1}^{K}(N_m + I_k^2)$

and SDR-AltMin, respectively, over the whole SNR range tested. This is very significant considering that only sub-6GHz and partially mmWave CSI are fed to the HGNNs, while the whole mmWave CSI is fed to the alternating methods. This result also implies that the HGNN can learn the relationships between the sub-6GHz channel and the mmWave channel well and is capable of solving the challenging precoding optimization problem (7) effectively. Also observe that the performances of the HGNNs are significantly better than the FusionNets and MLPs, due to the capability of HGNN to deal with heterogeneous graph data and interference between UEs.

3) Comparison for scalability of different BSs and UEs: We further test the scalability of the proposed HGNN, MLP and alternating methods by adjusting the number of mmWave BSs and the number of UEs per mmWave BS served, and the results are presented in Table III. It can be seen that the performance of HGNN-FULLY remains close to the performance of MO-AltMin method for all the configurations. It can also be seen that the performance of HGNN-PARTIALLY are even better than the performance of SDR-AltMin method in some configurations, especially when there are more UEs per mmWave BS served. Furthermore, the spectral efficiency gap between HGNN and MLP methods increases as the number of BSs and UEs increase. This is because the MLP methods are unable to effectively deal with the interference between UEs and have low scalability for the network size.

4) Comparison of run time: We run MO-AltMin and SDR-AltMin on 12*th Gen Intel (R) Core(TM) i7-12700KF* with 20 processors while the MLPs and the proposed HGNNs on *GeForce RTX 3080 Ti* to demonstrate the computational advantages of the proposed HGNN over the traditional alternating methods. Note that both MO-ALtMin and SDR-AltMin are unable to exploit the parallel computation of GPU due to their sequential computational flows. The total run times of various methods on the test dataset are compared in Table IV. It is obvious that the proposed HGNNs consume much less run time than the alternating methods. In addition, the run times of the HGNNs are almost constant with different numbers of BSs and UEs, while the run times of MO-AltMin and SDR-AltMin increase as the numbers of BSs and UEs increase. The MLP methods impose the lowest run time but their performance are the worst.

5) Comparison of overhead: The total communication overhead consists of the additional mmWave pilot overhead for channel estimation and the backhaul overhead for information exchange between BSs. The comparison of overhead for different schemes is shown in Table V. It can be seen that HGNN has the same pilot overhead as MLP and FusionNet methods, while the MO/SDR-AltMin methods require the whole CSI, resulting in higher pilot overhead. The backhaul overhead of HGNN is lower than that of MO/SDR-AltMin methods, as HGNN only requires the partial CSI to be exchanged between BSs, while MLP and FusionNet methods do not require the information exchange between BSs, and hence have no backhaul overhead.

V. CONCLUSIONS

In this paper, we have focused on the multi-user multi-cell hybrid beamforming problem at mmWave band under the co-existence of sub-6GHz and mmWave communications. In order to reduce the overhead cost, the sub-6GHz CSI and only partial mmWave CSI have been utilized. We have modeled the system as a heterogeneous graph and have proposed a HGNN with the attention mechanism and residual structure to solve this challenging problem. The numerical results have demonstrated that our proposed HGNN outperforms other machine learning methods in various scenarios. Moreover, the utilization of attention mechanism and residual structure has been shown to enhance the performance of the proposed HGNN. For the future work, it would be interesting to investigate the reduction of the communication overhead between BSs in more complex scenarios including the non-orthogonal multiple access and RIS aided systems.

REFERENCES

[1] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeterwave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100-1114, Feb. 2015.

[2] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101-107, Jun. 2011.

[3] 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, Jul. 2017.

[4] J. Li, L. Xiao, X. Xu, and S. Zhou, "Robust and low complexity hybrid beamforming for uplink multiuser mmWave MIMO systems," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1140-1143, Jun. 2016.

[5] O. E. Ayach, *et al.*, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499-1513, Mar. 2014. .

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

[7] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-Utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331-4340, Sep. 2011.

[8] H. Sun, *et al.*, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438-5453, Oct. 2018.

[9] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831-846, Oct. 2014.

[10] K. Y. Chen, H. Y. Chang, R. Y. Chang, and W. H. Chung, "Hybrid beamforming in mmWave MIMO-OFDM systems via deep unfolding," in *Proc. VTC2022-Spring* (Helsinki, Finland), Jun. 19-22, 2022, pp. 1-7.

[11] L. Zhu, *et al.*, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065-5079, Nov. 2019.

[12] A. Alkhateeb, G. Leus, and R. W. Heath, "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?," in *Proc. ICASSP 2015* (South Brisbane, QLD, Australia), Apr. 19-24, 2015, pp. 2909-2913.

[13] J. Choi, "Beam selection in mm-wave multiuser MIMO systems using compressive sensing," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2936-2947, Aug. 2015.

[14] P. Dong, H. Zhang, and G. Y. Li, "Machine learning prediction based CSI acquisition for FDD massive MIMO downlink," in *Proc. GLOBECOM 2018* (Abu Dhabi, United Arab Emirates), Dec. 9-13, 2018, pp. 1-6.

[15] A. Ali, N. Gonzlez-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless. Commun.*, vol. 17, no. 2, pp. 1038-1052, Feb. 2018.

[16] F. Gao, *et al.*, "FusionNet: Enhanced beam prediction for mmWave communications using sub-6 GHz channel and a few pilots," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8488-8500, Dec. 2021.

[17] M. Peter, *et al.*, "Measurement campaigns and initial channel models for preferred suitable frequency ranges," Deliverable D2.1, document H2020-ICT-671650-mmMAGIC/D2.1, 2016.

[18] T. Jiang, *et al.*, "The comparative study of S-V model between 3.5 and 28 GHz in indoor and outdoor scenarios," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 2351-2364, Mar. 2020.

[19] T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, "Steering with eyes closed: Mm-Wave beam steering without in-band measurement," in *Proc. INFOCOM 2015* (Hong Kong, China), Apr. 26-May 1, 2015, pp. 2416-2424.

[20] M. Hashemi, C. E. Koksal, and N. B. Shroff, "Out-of-band millimeter wave beamforming and communications to achieve low latency and high energy efficiency in 5G systems," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 875-888, Feb. 2018.

[21] K. Ma, P. Zhao, and Z. Wang, "Deep learning assisted beam prediction using out-of-band information," in *Proc. VTC2020-Spring* (Antwerp, Belgium), May 25-28, 2020, pp. 1-5.

[22] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504-5518, Sep. 2020.

[23] F. Scarselli, *et al.*, "The graph neural network model," *IEEE Trans. Neural Networks*, vol. 20, no. 1, pp. 61-80, Jan. 2009.

[24] Y. Shen, J. Zhang, S. H. Song, and K. B. Letaief, "Graph neural networks for wireless communications: From theory to practice," *IEEE Trans. Wireless Commun.*, vol. 22, no. 5, pp. 3554-3569, May 2023.

[25] 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, Jan. 2021.

[26] J. Guo and C. Yang, "Learning power control for cellular systems with heterogeneous graph neural network," in *Proc. WCNC 2021* (Nanjing, China), Mar 29-Apr. 1, 2021, pp. 1-6.

[27] Z. Zhang, T. Jiang, and W. Yu, "User scheduling using graph neural networks for reconfigurable intelligent surface assisted multiuser downlink communications," in *Proc. ICASSP 2022* (Singapore, Singapore), May 23-27, 2022, pp. 8892-8896.

[28] A. Chowdhury, *et al.*, "Unfolding WMMSE using graph neural networks for efficient power allocation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 6004-6017, Sep. 2021.

[29] X. Zhang, *et al.*, "Scalable power control/beamforming in heterogeneous wireless networks with graph neural networks," in *Proc. GLOBECOM 2021* (Madrid, Spain), Dec. 7-11, 2021, pp. 1-6.

[30] J. Kim, H. Lee, S.-E. Hong, and S.-H. Park, "A bipartite graph neural network approach for scalable beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 333-347, Jan. 2023.

[31] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481-6494, Nov. 2015.

[32] L. Dai, *et al.*, "Near-optimal hybrid analog and digital precoding for downlink mmWave massive MIMO systems," in *Proc. ICC 2015* (London, UK), Jun. 8-12, 2015, pp. 1334-1339.

[33] S. Han, C.-L. 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, Jan. 2015.

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

[35] M. S. Sim, *et al.*, "Deep learning-based mmWave beam selection for 5G NR/6G with sub-6 GHz channel information: Algorithms and prototype validation," *IEEE Access*, vol. 8, pp. 51634-51646, Mar. 2020.

[36] Z. Ren, S. Wu, and A. Zhao, "Coexist design of sub-6GHz and millimeter-wave antennas for 5G mobile terminals," in *Proc. ISAP 2018* (Busan, South Korea), Oct. 23-26, 2018, pp. 1-2.

[37] Y. Sun and J. Han. "Mining heterogeneous information networks: A structural analysis approach," *ACM SIGKDD Explorations Newsletter*, vol.14, no. 2, pp. 20-28, Apr. 2013.

[38] A. Vaswani, *et al.*, "Attention is all you need," in *Proc. NIPS 2017* (Long Beach, CA, USA), Dec. 4-9, 2017, pp. 1-11.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR 2016* (Las Vegas, NV, USA), Jun. 27-30, 2016, pp. 770-778.

[40] K. Oono and T. Suzuki, "Graph neural networks exponentially lose expressive power for node classification," in *Proc. ICLR 2020*, Apr. 26-May 1, 2020, pp. 1-37.

[41] S. Ioffe and S. Christian, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML 2015* (Lille France), Jul. 6-11, 2015, pp. 448-456

[42] N. Srivastava, *et al.*, "DropOut: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929-1958, 2014.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR 2015* (San Diego, CA, USA), May 7-9, 2015, pp. 1-15.

[44] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," in *Proc. ITA 2019* (San Diego, CA, USA), Feb. 10-15, 2019, pp. 1-8.