Deep Learning Assisted mmWave Beam Prediction for Heterogeneous Networks: A Dual-Band Fusion Approach

Ke Ma, Shouliang Du, Haoming Zou, Wenqiang Tian,
Zhaocheng Wang, *Fellow, IEEE*, Sheng Chen, *Fellow, IEEE*

***Abstract*—In this paper, motivated by the inter-base station (BS) channel dependence due to the shared wireless environment, we propose to fuse sub-6 GHz channel information and mmWave low-overhead measurement to predict the optimal mmWave beam in heterogeneous networks (HetNets) and reduce the overhead of both mmWave BS selection and beam training. Moreover, deep learning is adopted to extract the complex dependence between sub-6 GHz and mmWave channels for achieving high prediction accuracy. Specifically, we propose to leverage a few user equipment (UE)-specific high-quality mmWave wide beams predicted by the sub-6 GHz channel state information (CSI) as the mmWave low-overhead measurement. In order to adapt to different confidences of the mmWave wide beam prediction for diverse UE, the sum-probability criterion is proposed to flexibly adjust the number of measured wide beams. Besides, to fully fuse the diversified features extracted from the sub-6 GHz CSI and mmWave wide beams, the attention mechanism is further exploited to adaptively weight the features for improving the prediction accuracy. Simulation results show that our proposed scheme achieves higher beamforming gain while imposing smaller mmWave measurement overhead over the conventional deep learning based schemes.**

***Index Terms*—Millimeter-wave communications, sub-6 GHz information, beam prediction, deep learning, heterogeneous networks**

I. Introduction

Since the millimeter-wave (mmWave) frequency band provides abundant bandwidth resources for supporting high-speed transmissions, mmWave communications have been regarded as one of the promising technologies in the next-generation wireless networks [1], [2]. However, mmWave signals suffer from more severe pathloss than its sub-6 GHz counterpart [3]. To address this issue, massive antenna arrays are integrated at both base station (BS) and user equipment (UE) for establishing directional transmissions, so that the beamforming gain could compensate for the serious pathloss [4]–[6]. Consequently, combining the sufficient bandwidth resources in the mmWave band and the efficient beamforming based on multiple-input multiple-output (MIMO) is considered as one of the crucial technologies to meet the massive high-speed access requirements for the fifth-generation wireless networks and beyond [7], [8].

To enhance the mmWave received power, beam training has been broadly applied to find the optimal transmit and receive beam pair with the highest beamforming gain. The most straightforward method is the exhaustive search [9], which sweeps all the candidate beam pairs and selects the optimal one with the maximum received power. However, this could impose excessive training overhead. To tackle this problem, the twin-level low-overhead beam search was further proposed [10], [11], where the first level searches for the optimal wide beam, and then the second level finds the best narrow beam within the angular range of the optimal wide beam. Furthermore, the work [12] proposed to separately estimate the strongest angles at the BS and UE sides by angular search for designing the beamforming matrices. Nevertheless, the angular search process in these schemes [10]–[12] still requires considerable overhead.

To further decrease the overhead of beam training, sub-6 GHz channel information has been exploited for assisting mmWave beam selection based on the spatial similarity between sub-6 GHz and mmWave channels in the co-located scenarios [13]–[16], where both sub-6 GHz and mmWave antennas are deployed at the same BS, as shown in Fig. 1(a). In [13], field experiments demonstrated that the power azimuth spectrums (PASs) of sub-6 GHz and mmWave channels are almost congruent in the co-located scenarios. Based on these characteristics, the works [13], [14] proposed to only sweep the mmWave beams angularly neighboring to a few dominant sub-6 GHz paths for reducing the training overhead. Furthermore, other studies [15], [16] formulated the beam selection as a sparse signal recovery problem weighted by the sub-6 GHz PAS, which can reach high achievable rates under the smaller size of sparse training codebooks.

Nevertheless, the works [13]–[16] generally rely on the conventional approaches of estimating spatial parameters, which makes the achievable performance sensitive to the estimation error due to the small number of sub-6 GHz antennas [17]. To tackle this problem, deep learning was adopted to adaptively extract the complex relationships between the sub-6 GHz channel and mmWave channel for predicting the optimal mmWave beam [18]–[22]. Specifically, the study in [18] proved the existence of the mapping that can predict the optimal mmWave beam from the sub-6 GHz channel state information (CSI), and exploited a fully-connected network (FCN) to implement the prediction. The work [18] further proposed to sweep a limited subset of candidate beams with the best predicted qualities and select the beam with the highest received power as the optimal one, which achieved higher beamforming gain at the cost of a modest extra overhead. The authors of [19] utilized a mmWave

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1801102, in part by the National Natural Science Foundation of China (Grant No. 61871253), and in part by OPPO Research Fund. *(Corresponding author: Zhaocheng Wang)*

K. Ma, S. Du, H. Zou and Z. Wang are with Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. Z. Wang is also with Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China (E-mails: ma-k19@mails.tsinghua.edu.cn, dusl19@mails.tsinghua.edu.cn, zouhm19@mails.tsinghua.edu.cn, zcwang@tsinghua.edu.cn).

W. Tian is with OPPO Mobile Telecommunications Corp., Beijing 100084, China (E-mail: tianwenqiang@oppo.com).

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

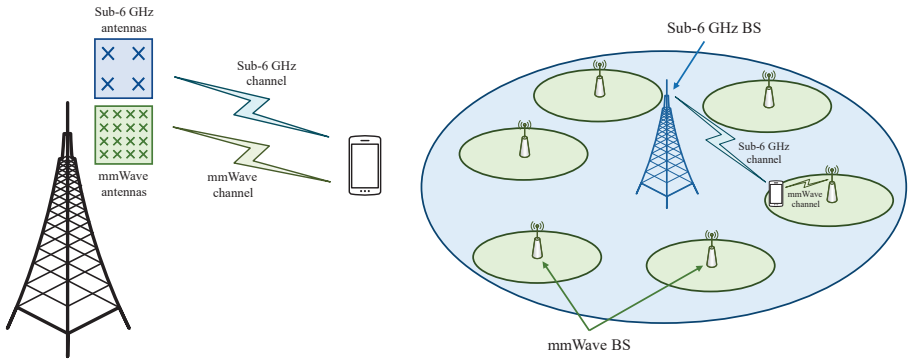


Fig. 1. Illustration of (a) co-located scenarios, and (b) heterogeneous networks, where blue and green circles denote sub-6 GHz and mmWave cells, respectively.

prototype to validate the effectiveness of the sub-6 GHz CSI assisted mmWave beam prediction using a FCN. By contrast, the works [20], [21] conceived a convolutional neural network (CNN) structure to process the variable-size sub-6 GHz CSI due to different user configurations. Furthermore, the study [22] proposed to fuse the CSI of sub-6 GHz antennas and a small number of mmWave antennas to jointly predict the optimal mmWave beam for enhancing the prediction accuracy.

However, the existing studies on the sub-6 GHz channel information assisted mmWave beam prediction [18]–[22] share a common limitation that they all rely on the spatial similarity in the co-located scenario. Since mmWave signals suffer from more serious pathloss than sub-6 GHz signals, the cellular radius of a mmWave BS is generally smaller than its sub-6 GHz counterpart. Consequently, mmWave BSs are usually deployed more densely than sub-6 GHz BSs, which forms the heterogeneous network (HetNet) [23], [24], as shown in Fig. 1(b). Particularly for the hot-spot areas with dense mmWave BSs, several accessible mmWave BSs may be located around the UE, which leads to the huge overhead of finding the optimal beam in a relatively large beam pool broadcasted from surrounding multiple mmWave BSs. However, in this more practical scenario, most UE are likely to access non-co-located sub-6 GHz and mmWave BSs, where the conventional schemes are difficult to achieve a satisfactory performance of the mmWave beam prediction due to the lack of the spatial similarity.

Fortunately, since the wireless channel is the result of the interaction between the transmitted signals and the wireless environment around BS and UE, the CSI matrices between a UE and multiple neighboring BSs have complex hidden relationships based on their shared propagation environment, which is termed as the inter-BS CSI dependence [25], [26]. Therefore, the channel of the source BS can be used to infer the channel features of the neighboring target BS. The work [26] proved the existence of this inter-BS CSI dependence theoretically, and then utilized a CNN to predict the optimal beam of the target mmWave BS according to the CSI of the neighboring mmWave BS. Nonetheless, such an approach of estimating mmWave CSI may impose excessive overhead owing to the huge number of mmWave antennas.

In contrast to the studies [25], [26], in this paper, we propose to exploit sub-6 GHz CSI to predict the optimal mmWave beam in the HetNet and reduce the huge overhead of both mmWave BS selection and beam training. Nevertheless, it is very hard to implement an accurate prediction by only using the sub-6 GHz CSI, since the small number of sub-6 GHz antennas cannot comprehensively reflect the complicated inter-BS CSI dependence. Therefore, we propose a dual-band fusion approach to integrate sub-6 GHz CSI with mmWave low-overhead measurement to facilitate reliable prediction, where deep learning is adopted to accurately model the complex inter-BS CSI dependence. Specifically, we propose to leverage the UE-specific high-quality mmWave wide beams predicted by the sub-6 GHz CSI as the mmWave low-overhead measurement for two reasons. Firstly, compared with exploiting the CSI of small-number mmWave antennas [22], the UE-specific high-quality beams can achieve higher received power for providing reliable mmWave channel information. Secondly, compared with additionally measuring the mmWave narrow beams [18], the wide beams enjoy stronger coverage capability, which is capable of effectively covering the large uncertain area from the sub-6 GHz CSI based prediction at low overhead.

Because the predictions are expressed as the probability that each candidate wide beam contains the optimal narrow beam, we design two criteria to select the wide beams to be measured. The first criterion is the fixed-number criterion, which selects the fixed number of the maximum (max)-probability wide beams. In order to adapt to different confidences of the mmWave wide beam prediction for diverse UE, the sum-probability criterion is further proposed to select the max-probability wide beam set whose probability sum is larger than a predefined threshold, so that the number of measured wide beams can be flexibly adjusted. Besides, to fully fuse the diverse features extracted from the sub-6 GHz CSI and mmWave wide beams, the attention mechanism is further leveraged to adaptively weight the features for improving the prediction accuracy. Finally, we extensively evaluate the performance of our proposed dual-band fusion scheme based on the open-source DeepMIMO dataset [27]. Simulation results demonstrate that the proposed scheme is capable of achieving higher beamforming gain while imposing smaller mmWave measurement overhead, compared to the conventional deep learning based schemes [18], [22].

The main contributions of this paper are recapped as fol-

lows.

- We propose a dual-band fusion approach to integrate sub-6 GHz CSI with mmWave low-overhead measurement to predict the optimal mmWave beam in the HetNet and reduce the huge overhead of both mmWave BS selection and beam training. Furthermore, we propose to adopt deep learning to extract the complex relationships between sub-6 GHz CSI and mmWave low-overhead measurement, where the attention mechanism is leveraged to adaptively weight the features for enhancing the prediction accuracy.
- We propose to adopt the UE-specific high-quality mmWave wide beams predicted by the sub-6 GHz CSI as the mmWave low-overhead measurement. We also design two criteria for the wide beam selection. The fixed-number criterion selects the fixed number of the max-probability wide beams, while the sum-probability criterion selects the max-probability wide beam set whose probability sum is larger than a predefined threshold, which can adapt to different confidences of the mmWave wide beam prediction for diverse UE.

The paper is organized as follows. Section II presents the system model. Our proposed dual-band fusion scheme is detailed in Section III. Section IV presents the simulation study. The conclusions are provided in Section V.

Notations: $\mathbb{N}^+$ denotes the positive integer set. $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ denote the $m\times n$ real space and complex space, respectively. Boldface capital and lower-case letters represent matrices and vectors, respectively, e.g., $\boldsymbol{A}$ and $\boldsymbol{a}$. Calligraphic capital letters denote sets, e.g., $\mathcal{A}$. $\Re(\cdot)$ and $\Im(\cdot)$ stand for the real and imaginary parts of a complex number, and $\mathrm{j}=\sqrt{-1}$. The conjugate and transpose operators are expressed by $(\cdot)^*$ and $(\cdot)^{\mathrm{T}}$, respectively, while $|\cdot|$ represents the magnitude operator. $\langle\cdot\rangle$ denotes the order statistics, e.g., for $\mathcal{A}=\{a_1,a_2,\cdots,a_n\}$, $\langle\mathcal{A}\rangle=\{a_{\sigma_1},a_{\sigma_2},\cdots,a_{\sigma_n}\}$ with $a_{\sigma_1}\le a_{\sigma_2}\le\cdots\le a_{\sigma_n}$. $\odot$ denotes element-wise product, and $\circ$ denotes the function composition operator, while $\lfloor\cdot\rfloor$ is the round-down operation. The variable with overline corresponds to the sub-6 GHz parameters, e.g., $\overline{\boldsymbol{h}}$, while the variable with hat $\widehat{\ }$ represents the estimated value, e.g., $\widehat{p}$.

II. System Model

Consider a HetNet consisting of one sub-6 GHz BS and J mmWave BSs, where each user accesses the sub-6 GHz BS and one of the J mmWave BSs, as shown in Fig. 1(b). For simplicity, we consider single UE equipped with one sub-6 GHz antenna and one mmWave antenna [18], but our scheme can be straightforwardly extended to the multiuser scenario with multiple UE antennas. Further assume the two-dimensional system model, namely, BSs are equipped with uniform linear arrays (ULAs), where only azimuth angles are considered.

A. Channel Model

For the mmWave link, we adopt the well-known Saleh-Valenzuela channel model [28], [29]. Specifically, the CSI vector $\boldsymbol{h}^{(j)}\in\mathbb{C}^{M^{(j)}\times 1}$ related to the j-th mmWave BS is expressed as

$$\boldsymbol{h}^{(j)}=\sum_{l=1}^{L^{(j)}}\sqrt{\frac{M^{(j)}}{\rho_l^{(j)}}}\alpha_l^{(j)}\boldsymbol{a}^{(j)^*}\big(\phi_l^{(j)}\big). \tag{1}$$

In this model, $M^{(j)}$ denotes the number of antennas in the j-th mmWave BS, $L^{(j)}$ represents the number of paths, and the l-th path has pathloss $\rho_l^{(j)}$, complex gain $\alpha_l^{(j)}$ and angle-of-departure (AoD) $\phi_l^{(j)}$. Furthermore, $\boldsymbol{a}^{(j)}(\phi)\in\mathbb{C}^{M^{(j)}\times 1}$ denotes the antenna response vector of the j-th BS at the AoD ϕ, which is expressed as

$$\boldsymbol{a}^{(j)}(\phi)=\sqrt{\frac{1}{M^{(j)}}}\Big[1\ e^{\mathrm{j}2\pi d^{(j)}\sin\phi/\lambda}\cdots e^{\mathrm{j}2\pi(M^{(j)}-1)d^{(j)}\sin\phi/\lambda}\Big]^{\mathrm{T}}, \tag{2}$$

where $d^{(j)}$ is the antenna spacing at the j-th mmWave BS, and λ is the wavelength. Without loss of generality, we set $d^{(j)}=\lambda/2$ and assume that all the mmWave BSs have the same number of antennas. Thus the BS index $^{(j)}$ of the corresponding variables are removed, i.e., M, $\boldsymbol{a}(\phi)$.

The sub-6 GHz CSI vector $\overline{\boldsymbol{h}}\in\mathbb{C}^{\overline{M}\times 1}$ is similar to its mmWave counterpart. To clearly distinguish two frequency bands, the variables of the sub-6 GHz link are marked with overline.

B. MmWave Beam Training Model

We assume that a single radio frequency (RF) chain and the same phase shifter based analog beamforming scheme are employed at each mmWave BS, where $\boldsymbol{f}\in\mathbb{C}^{M\times 1}$ denotes the transmit beam of BS. The beam is selected from the predefined codebook $\mathcal{F}$, which comprises N candidate beams. Further assume that the mmWave BS applies the discrete Fourier transform (DFT) codebook [30], [31]. Therefore, the candidate transmit beam $\boldsymbol{f}_n$, $n\in\{1,2,\cdots,N\}$, can be written as

$$\boldsymbol{f}_n=\sqrt{\frac{1}{M}}\Big[1\ e^{\mathrm{j}\pi\gamma_n}\cdots e^{\mathrm{j}\pi(M-1)\gamma_n}\Big]^{\mathrm{T}}, \tag{3}$$

where γ_n denotes the beam phase of the n-th candidate beam at BS. To cover the whole angular space, we further assume that the beam phases are uniformly sampled in $[-\pi,\ \pi]$ [30], i.e.,

$$\gamma_n=-1+\frac{2n-1}{N}. \tag{4}$$

Given the CSI vector $\boldsymbol{h}$ and the transmit beam $\boldsymbol{f}$, the received signal y can be expressed as

$$y=\sqrt{P}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{f}x+\mathtt{n}, \tag{5}$$

where x represents the transmitted signal with $|x|=1$ and P denotes the transmit power. Furthermore, $\mathtt{n}$ denotes the additional white Gaussian noise (AWGN) with the noise power σ^2, i.e., $\mathtt{n}\sim\mathcal{CN}\big(0,\sigma^2\big)$.

Let $\boldsymbol{f}_n^{(j)}$ be the n-th transmit beam of the j-th mmWave BS. In HetNets, beam training targets to find the optimal transmit beam $\boldsymbol{f}_{n^\star}^{(j^\star)}$ with the maximum beamforming gain

from the candidate beams of all the mmWave BSs, which can be expressed as

$$\{j^\star, n^\star\} = \arg \max_{\substack{j\in\{1,2,\cdots,J\},\\ n\in\{1,2,\cdots,N\}}} \left|\boldsymbol{h}^{(j)\mathrm{T}} \boldsymbol{f}_n^{(j)}\right|^2. \tag{6}$$

A direct method to solve this problem is the exhaustive beam search, which sweeps all the candidate transmit beams to find the beam with the maximum received power [9]. However, this scheme usually leads to excessive training overhead due to the huge-number of the candidate beams from all the J mmWave BSs.

To reduce the training overhead, the twin-level beam search based on the beams with different widths may be adopted [10], [11]. Since more antennas could generate narrower beamwidths, we use partial antennas to obtain the wide beams for the sake of conceptual simplicity. Concretely, consider switching on $M_{\mathrm{w}} \triangleq \lfloor M/s \rfloor$ BS antennas to generate $N_{\mathrm{w}} \triangleq \lfloor N/s \rfloor$ transmit wide beams, where $s \in \mathbb{N}^+$ defines the number of narrow beams within the angular range of each wide beam [32]. Therefore, the candidate transmit wide beam $\boldsymbol{f}_{\mathrm{w},n}$, $n \in \{1, 2, \cdots, N_{\mathrm{w}}\}$, can be written as

$$\boldsymbol{f}_{\mathrm{w},n} = \sqrt{\frac{1}{M_{\mathrm{w}}}} \left[1 \; e^{\mathrm{j}\pi\gamma_n^{\mathrm{w}}} \cdots e^{\mathrm{j}\pi(M_{\mathrm{w}}-1)\gamma_n^{\mathrm{w}}}\right]^{\mathrm{T}}, \tag{7}$$

where the phase of the n-th candidate transmit wide beam γ_n^{w} can be written as

$$\gamma_n^{\mathrm{w}} = -1 + \frac{2n-1}{N_{\mathrm{w}}}. \tag{8}$$

Accordingly, the beam search can be divided into two levels. The first-level targets to find the optimal transmit wide beam, which can be expressed as

$$\left\{j^\star, n_{\mathrm{w}}^\star\right\} = \arg \max_{\substack{j\in\{1,2,\cdots,J\},\\ n\in\{1,2,\cdots,N_{\mathrm{w}}\}}} \left|\boldsymbol{h}_{\mathrm{w}}^{(j)\mathrm{T}} \boldsymbol{f}_{\mathrm{w},n}^{(j)}\right|^2, \tag{9}$$

where $\boldsymbol{h}_{\mathrm{w}}^{(j)} \in \mathbb{C}^{M_{\mathrm{w}}\times 1}$ is the channel vector corresponding to the antennas for generating wide beams, and $\boldsymbol{f}_{\mathrm{w},n}^{(j)}$ is the n-th wide beam of the j-th BS. Recall that M antennas correspond to N candidate transmit narrow beams of (3) with the phases of (4), respectively. Then, the second-level search confirms the best transmit narrow beam within the angular range of the selected wide beam (9), which is expressed as

$$n^\star = \arg \max_{n\in\{(n_{\mathrm{w}}^\star-1)s+1,\cdots,n_{\mathrm{w}}^\star s\}} \left|\boldsymbol{h}^{(j^\star)\mathrm{T}} \boldsymbol{f}_n^{(j^\star)}\right|^2. \tag{10}$$

Although the twin-level beam search achieves a lower overhead than the exhaustive beam search, it still requires considerable measurements due to the large number of candidate wide beams.

III. Dual-Band Fusion for mmWave Beam Prediction in HetNets

A. Motivation and Problem Formulation

The wireless channel is the result of the interaction between the transmitted signals and the wireless environment around BS and UE, which manifests the information of the UE location and surrounding environment [33]–[35]. Since neighboring BSs share the common wireless environment, their CSI with the same UE have complex hidden relationships, which is termed as the inter-BS CSI dependence [25], [26]. This property allows us to use the sub-6 GHz CSI to predict the optimal mmWave beam from the large-number candidate beams of multiple neighboring mmWave BSs in the HetNet, so that the huge overhead of both BS selection and beam training can be effectively reduced.

To elaborate on the feasibility of the sub-6 GHz CSI for guiding the mmWave beam prediction, we investigate mapping the sub-6 GHz CSI vector $\overline{\boldsymbol{h}}$ to the indices of the optimal mmWave BS and beam $\{j^\star, n^\star\}$ via the UE location $\boldsymbol{b}$. Specifically, we define $\Phi_{\tilde{\boldsymbol{h}}} : \boldsymbol{b} \to \tilde{\boldsymbol{h}}, \tilde{\boldsymbol{h}} \in \{\overline{\boldsymbol{h}}, \boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, ..., \boldsymbol{h}^{(J)}\}$ as the position-to-channel mapping functions of the sub-6 GHz and mmWave CSI vectors. Then, based on the unique propagation environment shared by sub-6 GHz and mmWave BSs, we follow [18], [36], [37] and adopt the bijective assumption as below:

Assumption 1: The position-to-channel mapping functions, $\Phi_{\tilde{\boldsymbol{h}}} : \boldsymbol{b} \to \tilde{\boldsymbol{h}}$, are bijective.

Assumption 1 means that each position in the environment has its unique sub-6 GHz and mmWave channels. Based on this assumption, we have the following proposition on the existence of the mapping function from the sub-6 GHz CSI vector $\overline{\boldsymbol{h}}$ to the indices of the optimal mmWave BS and beam $\{j^\star, n^\star\}$:

Preposition 1: For any given communication environment, if Assumption 1 is satisfied, then there exists a mapping function Υ from the sub-6 GHz CSI vector to the indices of the optimal mmWave BS and beam $\Upsilon : \overline{\boldsymbol{h}} \to \{j^\star, n^\star\}$.

Proof: The proof follows from the existence of the mapping function from the sub-6 GHz CSI vector to the mmWave CSI vectors $\boldsymbol{h}^{(j)} = \Phi_{\boldsymbol{h}^{(j)}}(\boldsymbol{b}) = \Phi_{\boldsymbol{h}^{(j)}} \circ \Phi_{\overline{\boldsymbol{h}}}^{-1}(\overline{\boldsymbol{h}}), j = 1, 2, ..., J$. Then, the indices of the optimal mmWave BS and beam $\{j^\star, n^\star\}$ are found via beam search according to (6). □

It should be noted that the position-to-channel mapping is actually bijective with high probability in many practical wireless communication scenarios [36]–[38], which guarantees the feasibility to adopt the sub-6 GHz CSI for predicting the mmWave optimal BS and beam in the HetNet. However, to implement an accurate prediction by only using the sub-6 GHz CSI is very difficult for two reasons. First, the small number of sub-6 GHz antennas makes its CSI hard to comprehensively reflect the complex inter-BS relationships based on the shared high-dimensional environment features. Second, the significant frequency difference between sub-6 GHz and mmWave bands leads to sophisticated but weak inter-BS relationships. To tackle these two problems, we propose to fuse the sub-6 GHz CSI and mmWave low-overhead measurement to enhance the prediction accuracy.

An intuitive method of designing the mmWave low-overhead measurement is to acquire the CSI of a small number of mmWave antennas [22], but it cannot exploit high beamforming gain to compensate for the high pathloss of mmWave signals, so that the noise may severely degrade the reliability of the measured CSI. Another design is to sweep the subset of the predicted optimal mmWave narrow beams, and select

the beam with the maximum received power as the optimal one [18]. However, since it is difficult to extract the sufficient information of the UE location and propagation environment from the sub-6 GHz CSI, the narrow beam prediction usually leaves a large uncertain area, so that a large number of mmWave narrow beam measurements are required to cover this area for ensuring to find a satisfactory beam. By contrast, the wide beams, which simultaneously have wider coverage capability than the narrow beams and exploit the beamforming gain to obtain higher received power than the mmWave CSI, can be measured for predicting the optimal narrow beam [39], [40]. Therefore, we propose to leverage a few UE-specific high-quality mmWave wide beams predicted by the sub-6 GHz CSI as the mmWave low-overhead measurement, which can obtain more reliable mmWave channel information than exploiting mmWave CSI only [22], while simultaneously covering the large uncertain area with lower overhead than using narrow beams alone [18].

Specifically, we define the received signals of wide beam training of the j-th mmWave BS as $\boldsymbol{y}_{\mathrm{w}}^{(j)} = \left[y_{\mathrm{w},1}^{(j)}\ y_{\mathrm{w},2}^{(j)} \cdots y_{\mathrm{w},N_{\mathrm{w}}}^{(j)}\right]^{\mathrm{T}}$, where $y_{\mathrm{w},n}^{(j)}$ denotes the received signal of the corresponding n-th wide beam. Then, the concatenated received signal vector of wide beam training of all the J mmWave BSs is expressed as $\boldsymbol{y}_{\mathrm{w}} = \left[\boldsymbol{y}_{\mathrm{w}}^{(1)^{\mathrm{T}}}\ \boldsymbol{y}_{\mathrm{w}}^{(2)^{\mathrm{T}}} \cdots \boldsymbol{y}_{\mathrm{w}}^{(J)^{\mathrm{T}}}\right]^{\mathrm{T}} = \left[y_{\mathrm{w},1}^{(1)} \cdots y_{\mathrm{w},N_{\mathrm{w}}}^{(1)} \cdots y_{\mathrm{w},1}^{(J)} \cdots y_{\mathrm{w},N_{\mathrm{w}}}^{(J)}\right]^{\mathrm{T}}$. For convenience, we renumber the elements of $\boldsymbol{y}_{\mathrm{w}}$ and express it as $\boldsymbol{y}_{\mathrm{w}} = \left[y_{\mathrm{w},1}\ y_{\mathrm{w},2} \cdots y_{\mathrm{w},JN_{\mathrm{w}}}\right]^{\mathrm{T}} \in \mathbb{C}^{JN_{\mathrm{w}}\times 1}$. Further let $\mathcal{L}_{\mathrm{w}}$ denote the index set of the selected wide beams to be measured. Then the received signal vector of the measured wide beams $\boldsymbol{y}_{\mathrm{wp}}$ can be obtained according to

$$\boldsymbol{y}_{\mathrm{wp}}[n] = \begin{cases} y_{\mathrm{w},n}, & \text{if } n \in \mathcal{L}_{\mathrm{w}}, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Similarly, the indices of the candidate narrow beams of the J mmWave BSs are renumbered as $\{1, 2, \cdots, JN\}$.

Considering the finite number of candidate narrow beams, the prediction is formulated as a multi-class classification task, where each class represents one mmWave narrow beam. For convenience, we redefine $n^\star$ as the optimal narrow beam index from all the candidate narrow beams of the J mmWave BSs, and thus the prediction model can be expressed as

$$n^\star = f(\overline{\boldsymbol{h}}, \boldsymbol{y}_{\mathrm{wp}}), n^\star \in \{1, 2, \cdots, JN\}, \tag{12}$$

where $f(\cdot)$ denotes the classification function. This prediction relies on the complex dependence between the channels of the sub-6 GHz BS and its neighboring mmWave BSs, which is hard to be expressed by an explicit mathematical model. Consequently, we propose to utilize deep learning to adaptively model these complicated relationships for implementing the prediction [41]. Furthermore, it is noted that the channel relationships between two bands may be buried under serious noises, hence our proposed dual-band fusion approach is suitable for the high-SNR scenarios.

The mmWave narrow beam prediction scheme comprises two stages, training and predicting. In the training stage, the training dataset is constructed for optimizing the deep learning model, where each sample comprises the sub-6 GHz CSI together with the received signal vector of mmWave wide beams as the prediction input, and the index of the optimal mmWave narrow beam as the prediction label. These data can be acquired by conventional sub-6 GHz CSI estimation and mmWave beam training. During this stage, the loss function calculates the errors between the predicted results and the classification labels for optimizing the model parameters using the back propagation algorithm. Once the model has been sufficiently trained, it switches to the predicting stage. At this stage, the qualities of all the candidate mmWave wide beams are predicted based on the sub-6 GHz CSI, and only part of the predicted high-quality mmWave wide beams are measured. Then the sub-6 GHz CSI and the received signal vector of the measured mmWave wide beams are adopted to jointly predict the optimal mmWave narrow beam. Therefore, BS and UE only measure the sub-6 GHz CSI and a few mmWave wide beams, and thus the huge overhead of both mmWave BS selection and beam training is significantly reduced. In order to adapt to dynamic environmental fluctuations, online training can be employed during the predicting stage, where the deep learning model is continuously optimized by the training data collected from a small part of users that adopt conventional mmWave beam search.

The complete architecture of our proposed scheme of dual-band fusion based mmWave beam prediction for HetNets is depicted in Fig. 2, which consists of the four components: the sub-6 GHz feature extraction, mmWave wide beam selection, mmWave feature extraction, and feature fusion. We now detail them one by one.

B. Sub-6 GHz Feature Extraction

Intuitively, the most straightforward method for selecting the high-quality mmWave wide beams based on the sub-6 GHz CSI is to predict the probabilities of the optimal wide beam according to the wide beam label with the largest received power. However, since our final target is to predict the optimal narrow beam, the probability of each wide beam to contain the optimal narrow beam could be a more appropriate criterion to evaluate the qualities of wide beams, instead of the received power of wide beams.

Specifically, the optimal narrow beam index is first predicted based on the sub-6 GHz CSI, where the output is expressed as the probability that each candidate narrow beam is the optimal one. Concretely, the narrow beam prediction based on $\overline{\boldsymbol{h}}$ is formulated as a classification task similar to (12), which is expressed as

$$n^\star = f_{\text{sub-6}}(\overline{\boldsymbol{h}}), n^\star \in \{1, 2, \cdots, JN\}, \tag{13}$$

where $f_{\text{sub-6}}(\cdot)$ denotes the corresponding classification function. CNN is adopted to extract the features from the sub-6 GHz CSI $\overline{\boldsymbol{h}}$ owing to its outstanding performance in handling classification problems [42]. The proposed sub-6 GHz feature extraction model is illustrated in the top-left part of Fig. 2, which can be divided into three parts as below.

1) Preprocessing Module: Since the sub-6 GHz CSI $\overline{\boldsymbol{h}}$ is complex-valued, which cannot be processed by CNN directly,

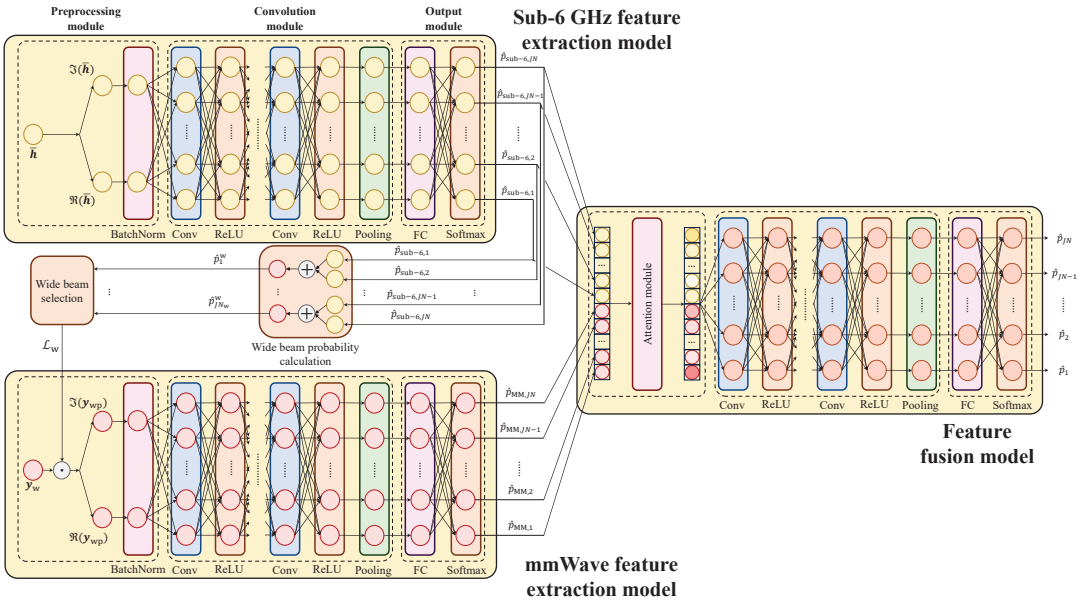


Fig. 2. Proposed prediction model consisting of sub-6 GHz feature extraction model, mmWave wide beam selection, mmWave feature extraction model and feature fusion. Each circle denotes one feature channel, and 'Conv' represents convolution layers.

the preprocessing module firstly divides $\overline{\boldsymbol{h}}$ into the two real-valued feature channels representing the real and imaginary parts $\{\Re(\overline{\boldsymbol{h}}), \Im(\overline{\boldsymbol{h}})\}$. Then, in order to handle the large dynamic ranges of $\overline{\boldsymbol{h}}$, batch normalization (BatchNorm) is utilized for transforming the two feature channels to the standard distribution $\{\Re(\overline{\boldsymbol{h}}^{\mathrm{N}}), \Im(\overline{\boldsymbol{h}}^{\mathrm{N}})\}$ with mean 0 and variance 1 [43].

2) Convolution module: We utilize convolution layers to extract the hidden features from $\overline{\boldsymbol{h}}^{\mathrm{N}}$, where the ReLU activation function is applied after each convolution layer to fit nonlinear relationships. In order to effectively reduce the feature space, the pooling layer is adopted after the ultimate convolution layer to downasmple the extracted features.

3) Output module: The fully-connected (FC) layer is exploited after the convolution module for transforming the extracted features to the candidate narrow beams. Then, a softmax activation layer is applied to normalize the outputs of the FC layer into probabilities. For convenience, we define $\widehat{p}_{\text{sub-6},n}$ as the probability predicted from the sub-6 GHz CSI that the n-th candidate narrow beam is the optimal one, and the corresponding probability vector of the narrow beams is denoted as $\widehat{\boldsymbol{p}}_{\text{sub-6}} = [\widehat{p}_{\text{sub-6},1}\ \widehat{p}_{\text{sub-6},2} \cdots \widehat{p}_{\text{sub-6},JN}]^{\mathrm{T}}$.

It is noted that the vast majority of the learnable parameters in our proposed model are from convolution layers and FC layers. Specifically, we define f_{i} and f_{o} as the numbers of input feature channels and output feature channels, while K denotes the kernel size of the convolution layer. Then, the parameter numbers of one convolution layer and one FC layer can be calculated as $f_{\mathrm{i}} f_{\mathrm{o}} K + f_{\mathrm{o}}$ and $f_{\mathrm{i}} f_{\mathrm{o}} + f_{\mathrm{o}}$, respectively.

C. MmWave Wide Beam Selection

In the mmWave wide beam selection as illustrated in the middle-left part of Fig. 2, the predicted probability of each wide beam is calculated as the sum of the narrow beam probabilities within its angular range, and the set of the max-probability wide beams are selected as the wide beams to be measured. More specifically, to obtain the predicted probability of the n_{w}-th wide beam $\widehat{p}^{\mathrm{w}}_{n_{\mathrm{w}}}$, the probabilities of the narrow beams within its angular range are added together, which can be expressed as

$$\widehat{p}^{\mathrm{w}}_{n_{\mathrm{w}}} = \sum_{n=1}^{s} \widehat{p}_{\text{sub-6},(n_{\mathrm{w}}-1)s+n},\ n_{\mathrm{w}} \in \{1, 2, \cdots, JN_{\mathrm{w}}\}. \tag{14}$$

Let K_{w} denote the number of the wide beams to be measured, which is independent to the number of mmWave BSs J. Therefore, the subset of the K_{w} max-probability wide beams are selected, i.e.,

$$\left\{\widehat{p}^{\mathrm{w}}_{\sigma_1}, \widehat{p}^{\mathrm{w}}_{\sigma_2}, \cdots, \widehat{p}^{\mathrm{w}}_{\sigma_{JN_{\mathrm{w}}}}\right\} = \left\langle \left\{\widehat{p}^{\mathrm{w}}_{1}, \widehat{p}^{\mathrm{w}}_{2}, \cdots, \widehat{p}^{\mathrm{w}}_{JN_{\mathrm{w}}}\right\} \right\rangle, \tag{15}$$

$$\mathcal{L}_{\mathrm{w}} = \{\sigma_{JN_{\mathrm{w}}-K_{\mathrm{w}}+1}, \sigma_{JN_{\mathrm{w}}-K_{\mathrm{w}}+2}, \cdots, \sigma_{JN_{\mathrm{w}}}\}. \tag{16}$$

How to select the measured wide beams for achieving higher prediction accuracy under the same average wide beam measurement overhead for various UE is one of the crucial issues for the narrow beam prediction. A straightforward criterion is the fixed-number criterion, i.e., to select the fixed number of the max-probability wide beams. However, the wide beam predictions of the UE at different locations usually have different confidences due to various surrounding environments. Therefore, the performance of selecting the fixed number of wide beams may suffer from high sensitivity to the UE location. By contrast, the predicted probabilities of wide beams provide a clear illustration of the UE-specific confidence. Specifically, when fewer wide beams possess larger sum probabilities, the prediction has higher confidence. Consequently, the sum-probability criterion is proposed to select the max-probability wide beam set whose probability sum is larger than a predefined threshold η_{w}, which can be formulated as the following optimization problem

$$\begin{aligned} \min_{K_{\mathrm{w}} \in \{1,2,\cdots,JN_{\mathrm{w}}\}} \quad & K_{\mathrm{w}}, \\ \text{s.t.} \quad & \sum_{n_{\mathrm{w}}=JN_{\mathrm{w}}-K_{\mathrm{w}}+1}^{JN_{\mathrm{w}}} \widehat{p}^{\mathrm{w}}_{\sigma_{n_{\mathrm{w}}}} > \eta_{\mathrm{w}}. \end{aligned} \tag{17}$$

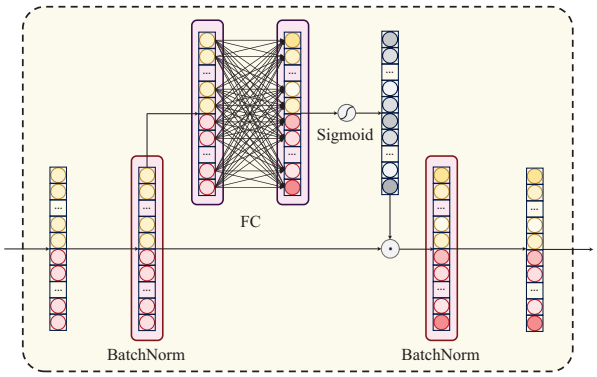


Fig. 3. Proposed attention module, where each circle denotes one feature.

Recall that K_{w} represents the number of the measured wide beams. (17) ensures that the number of the measured wide beams can be flexibly adjusted according to the prediction confidence.

It is important to point out that a synchronization process is required among different mmWave BSs to measure the UE-specific high-quality mmWave wide beams. The sub-6 GHz BS can schedule and notify both UE and mmWave BSs of the time slots for wide beam measurements.

D. MmWave Feature Extraction

Similar to the sub-6 GHz feature extraction model, CNN is adopted to extract the mmWave features from the received signal vector of the measured mmWave wide beams $\boldsymbol{y}_{\mathrm{wp}}$. In order to explicitly guide the parameter optimization of this CNN model, the mmWave feature extraction model is also formulated as the classification task for predicting the optimal mmWave narrow beam, where $\widehat{\boldsymbol{p}}_{\mathrm{MM}} = \left[\widehat{p}_{\mathrm{MM},1}\ \widehat{p}_{\mathrm{MM},2}\cdots\widehat{p}_{\mathrm{MM},JN}\right]^{\mathrm{T}}$ is used as the corresponding predicted probability vector of the narrow beams, as shown in the bottom-left part of Fig. 2.

E. Feature Fusion

Our feature fusion model is illustrated in the right part of Fig. 2. The two predicted probability vectors based on the sub-6 GHz CSI $\overline{\boldsymbol{h}}$ and the measured mmWave wide beams $\boldsymbol{y}_{\mathrm{wp}}$ can be regarded as the extracted features from the two frequency bands, which are further fused for enhancing the prediction accuracy. Specifically, we concatenate the two predicted probability vectors as the fused feature vector $\widehat{\boldsymbol{p}}_{\mathrm{f}} = \left[\widehat{\boldsymbol{p}}_{\text{sub-6}}^{\mathrm{T}}\ \widehat{\boldsymbol{p}}_{\mathrm{MM}}^{\mathrm{T}}\right]^{\mathrm{T}}$, which is fed into the feature fusion model to predict the optimal narrow beam.

However, the diversified features extracted from the sub-6 GHz CSI and mmWave wide beams usually have different importance [22]. Thus directly concatenating these features with the same weight may limit the prediction performance of the fusion model. To address this issue, the attention mechanism is introduced to adaptively weight the elements in the fused feature vector $\widehat{\boldsymbol{p}}_{\mathrm{f}}$, as depicted in Fig. 3. Concretely, the FC layer is adopted to learn the weights for the features $\boldsymbol{w} \in \mathbb{R}^{2JN\times 1}$ from the fused feature vector $\widehat{\boldsymbol{p}}_{\mathrm{f}}$, which can be expressed as

$$\boldsymbol{w} = \mathrm{sigmoid}\left(\boldsymbol{W}\widehat{\boldsymbol{p}}_{\mathrm{f}} + \boldsymbol{b}_{\mathrm{w}}\right), \tag{18}$$

where $\boldsymbol{W} \in \mathbb{R}^{2JN\times 2JN}$ and $\boldsymbol{b}_{\mathrm{w}} \in \mathbb{R}^{2JN\times 1}$ denote the weight matrix and bias vector of the FC layer, respectively. Besides, the sigmoid activation function normalizes each element in the output of the FC layer to be between 0 and 1, which evaluates the importance of each feature. Therefore, the weighted fused feature vector $\widehat{\boldsymbol{p}}_{\mathrm{f,w}}$ can be calculated as the element-wise product between $\widehat{\boldsymbol{p}}_{\mathrm{f}}$ and $\boldsymbol{w}$, i.e.,

$$\widehat{\boldsymbol{p}}_{\mathrm{f,w}} = \widehat{\boldsymbol{p}}_{\mathrm{f}} \odot \boldsymbol{w}. \tag{19}$$

Batch normalization is adopted both before the FC layer and after the weighting operation, in order to adjust the distribution of the fused feature vector for facilitating model learning.

To predict the optimal narrow beam according to the weighted fused feature vector $\widehat{\boldsymbol{p}}_{\mathrm{f,w}}$, CNN is adopted after the attention module to implement the classification, where the output is expressed as the narrow beam probability vector $\widehat{\boldsymbol{p}} = \left[\widehat{p}_1\ \widehat{p}_2\cdots\widehat{p}_{JN}\right]^{\mathrm{T}}$. Finally, the narrow beam having the highest predicted probability is chosen, i.e.,

$$\widehat{n}^{\star} = \arg\max_{n\in\{1,2,\cdots,JN\}} \widehat{p}_n. \tag{20}$$

Clearly, it is more difficult to accurately predict the optimal narrow beam from more mmWave BSs due to the larger number of candidate beams.

F. Model Training

The cross entropy loss is used to optimize the proposed model, which can be expressed as

$$\mathrm{loss} = -\sum_{n=1}^{JN} p_n \log \widehat{p}_n, \tag{21}$$

where $p_n = 1$ if the n-th candidate mmWave narrow beam is the actual optimal beam, and otherwise $p_n = 0$. However, considering the branched structure of our proposed network, only adopting the prediction loss of the feature fusion model may insufficiently facilitate the model convergence. In order to fully guide the parameter optimization of the whole model, the losses of the narrow beam predictions according to the sub-6 GHz CSI, the measured mmWave wide beams and the fused feature vector are added up to train the model, which can be written as

$$\mathrm{loss} = -\sum_{n=1}^{JN} p_n(\lambda_{\text{sub-6}} \log \widehat{p}_{\text{sub-6},n} + \lambda_{\mathrm{MM}} \log \widehat{p}_{\mathrm{MM},n} + \lambda_{\mathrm{f}} \log \widehat{p}_n), \tag{22}$$

where $\lambda_{\text{sub-6}}$, λ_{MM} and λ_{f} denote the corresponding weight coefficients. Based on the calculated loss, the model training is performed on the entire network in Fig. 2.

G. Additional Narrow Beam Measurement

Despite of achieving an excellent accuracy enhancement by fusing the sub-6 GHz CSI and mmWave wide beam measurement, it is still difficult to reach an almost perfect beam alignment due to the huge number of candidate beams. Therefore, we further exploit the additional narrow beam measurement to achieve a near perfect beam alignment at the

expense of a small extra overhead. Specifically, the narrow beam with higher probability is more likely to enjoy larger received power than other beam with smaller probability. Therefore, the additional narrow beam measurement based on the predicted probabilities can be performed, where the K_{n} narrow beams having the highest predicted probabilities are measured, whose indices $\mathcal{L}_{\mathrm{n}}$ are given by

$$\left\{\widehat{p}_{\sigma_1}, \widehat{p}_{\sigma_2}, \cdots, \widehat{p}_{\sigma_{JN}}\right\} = \langle\left\{\widehat{p}_1, \widehat{p}_2, \cdots, \widehat{p}_{JN}\right\}\rangle, \tag{23}$$

$$\mathcal{L}_{\mathrm{n}} = \left\{\sigma_{JN-K_{\mathrm{n}}+1}, \sigma_{JN-K_{\mathrm{n}}+2}, \cdots, \sigma_{JN}\right\}. \tag{24}$$

Let y_n represent the received signal of the n-th candidate mmWave narrow beam. The narrow beam having the highest received power is selected as the optimal beam, i.e.,

$$\widehat{n}^{\star} = \arg\max_{n \in \mathcal{L}_{\mathrm{n}}} |y_n|^2. \tag{25}$$

Similar to Subsection III-C, both the fixed-number criterion and the sum-probability criterion can be adopted to select the predicted optimal narrow beams to be measured.

H. Extensions to More Scenarios

Our proposed dual-band fusion approach can be extended to more practical scenarios, including the multi-user scenarios and highly-dynamic scenarios.

1) Multi-user scenarios: The dual-band fusion approach can be straightforwardly applied to the multi-user scenarios by separately predicting the optimal mmWave beam of each user. Based on the predicted optimal beams, hybrid beamforming is usually adopted at each BS to achieve a good tradeoff between the system capability and the hardware costs [44], [45], while flexible beam scheduling can be applied to avoid the interference among different mmWave BSs [46], [47].

2) Highly-dynamic scenarios: Although the mmWave channels usually vary rapidly, our proposed dual-band fusion approach could maintain a stable performance in the highly-dynamic scenarios. This is because the duration of measuring one mmWave beam is generally less than 100 μs [48]. As can be seen from the following simulation results, our proposed scheme can achieve an almost perfect beam alignment by only requiring less than 20 mmWave measurements within 2 ms, where the propagation environment and the UE location can be assumed to be unchanged. Therefore, the performance degradation of our proposed scheme in the highly-dynamic scenarios is negligible.

On the other hand, the stability of UE movements in the dynamic scenarios can be further utilized to track the UE trajectory and enhance the accuracy of beam prediction, where recurrent neural networks (RNNs), e.g., gated recurrent units [49] and long short-term memory networks [50], [51], have shown impressive performance in tracking the beam variations. Therefore, integrating the proposed dual-band fusion approach with RNN is a very interesting direction for future research in the highly-dynamic scenarios.

IV. Simulation Study

A. Simulation Setup

In order to accurately simulate the inter-BS CSI dependence, we adopt the DeepMIMO dataset [27] generated by the precise ray-tracing based on Wireless Insite [52]. Specifically, the outdoor scenario 'O1' shown in Fig. 4 is considered, where UE is randomly distributed in User Grid 1. The wireless HetNet consists of one sub-6 GHz BS and $J = 12$ mmWave BSs, where the sub-6 GHz BS is located at BS 5, and the mmWave BSs correspond to BS $1 \sim 12$. Unless otherwise stated, the parameters of the simulated HetNet are shown in Table I. The power σ^2 of the AWGN for mmWave signals is calculated as $(-174 + 10\log_{10} W + N_{\mathrm{F}})$ dBm, while the AWGN power $\overline{\sigma}^2$ for sub-6 GHz signals is calculated in the same way. The default numbers of mmWave candidate narrow beams and wide beams are set to $N = 64$ and $N_{\mathrm{w}} = 16$, respectively, and the ULA antennas of both sub-6 GHz and mmWave BSs are deployed in the y-axis direction.

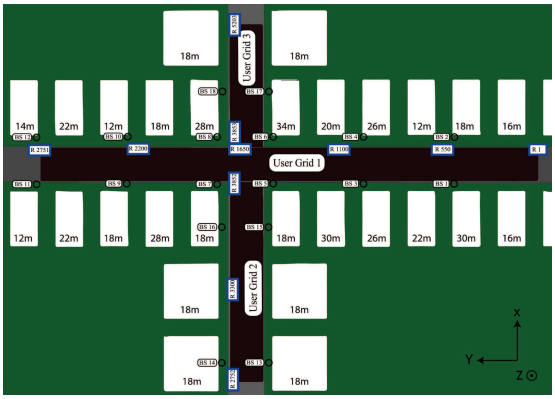


Fig. 4. Top view of 'O1' scenario [27].

TABLE I
DEFAULT PARAMETERS OF SIMULATED HETEROGENEOUS NETWORK.

Parameters	Sub-6 GHz	mmWave
Center frequency $\overline{f}_{\mathrm{c}}$, f_{c}	3.5 GHz	28 GHz
Bandwidth $\overline{W}$, W	10 MHz	100 MHz
BS antenna number $\overline{M}$, M	8	64
Transmit power $\overline{P}$, P	30 dBm	25 dBm
Noise factor $\overline{N}_{\mathrm{F}}$, N_{F}	3 dB	6 dB

The specific structures of our proposed deep learning models as well as the corresponding numbers of parameters are shown in Table II. Recall that f_{i} and f_{o} represent the numbers of input feature channels and output feature channels. L_{c} is the number of convolution layers, and $\{f_{\mathrm{o},1}, f_{\mathrm{o},2}, ..., f_{\mathrm{o},L_{\mathrm{c}}}\}$ denote the numbers of output feature channels in each convolution layer, while the corresponding CNN structure parameters (c_1, c_2, c_3) represent the kernel size K, sampling stride and zero-padding size, respectively. Furthermore, the dropout strategy is exploited in FC layers to abandon partial neurons randomly during model training for avoiding overfitting [53]. The training dataset comprising $100,000$ samples and the validation dataset comprising $5,000$ samples are constructed, respectively. The model is trained for 80 epochs, where Adam optimizer is adopted to optimize the model parameters [54].

We compare the performance of our proposed fixed-number criterion based (fixed-n based) and sum-probability criterion based (sum-p based) schemes with the following two baselines.

Baseline 1: The sub-6 GHz CSI assisted mmWave narrow beam prediction [18], which further measures the predicted

TABLE II
PROPOSED DEEP LEARNING MODELS.

Models	Layers	Structures	Numbers of parameters
Sub-6 GHz feature extraction model	Convolution	$L_{\mathrm{c}}=4, f_{\mathrm{i}}=2, f_{\mathrm{o}}=\{128,256,512,1024\}$, $(3,3,1)$, BatchNorm, ReLU	2.07×10^6
	Pooling	$f_{\mathrm{i}}=1024, f_{\mathrm{o}}=1024$, average-pooling	0
	FC	$f_{\mathrm{i}}=1024, f_{\mathrm{o}}=JN$, dropout $=0.3$, softmax	$1.02JN\times10^3$
mmWave feature extraction model	Convolution	$L_{\mathrm{c}}=4, f_{\mathrm{i}}=2, f_{\mathrm{o}}=\{128,256,512,1024\}$, $(3,3,1)$, BatchNorm, ReLU	2.07×10^6
	Pooling	$f_{\mathrm{i}}=1024, f_{\mathrm{o}}=1024$, average-pooling	0
	FC	$f_{\mathrm{i}}=1024, f_{\mathrm{o}}=JN$, dropout $=0.3$, softmax	$1.02JN\times10^3$
Attention module	FC	$f_{\mathrm{i}}=2JN, f_{\mathrm{o}}=2JN$, dropout $=0.3$, BatchNorm, sigmoid	$2JN(2JN+1)$
Feature fusion model	Convolution	$L_{\mathrm{c}}=4, f_{\mathrm{i}}=2, f_{\mathrm{o}}=\{128,256,512,1024\}$, $(3,3,1)$, BatchNorm, ReLU	2.07×10^6
	Pooling	$f_{\mathrm{i}}=1024, f_{\mathrm{o}}=1024$, average-pooling	0
	FC	$f_{\mathrm{i}}=1024, f_{\mathrm{o}}=JN$, dropout $=0.3$, softmax	$1.02JN\times10^3$

optimal K_{n} narrow beams for selecting the optimal beam with the maximum received power.

Baseline 2: The mmWave narrow beam prediction [22], which fuses the CSI of sub-6 GHz antennas and K_{ant} mmWave antennas pre-selected based on deep learning.

For a fair comparison, baseline 1 adopts the same structure of the proposed sub-6 GHz feature extraction model, while baseline 2 adopts a fully-connected network to select the mmWave antennas, and its sub-6 GHz feature extraction, mmWave feature extraction and feature fusion apply the same models of our proposed scheme.

Two metrics are utilized for performance evaluation as given below.

1) Normalized beamforming gain G_{N} which is defined as

$$G_{\mathrm{N}}=\frac{\left|\boldsymbol{h}^{(\widehat{j}^\star)}\boldsymbol{f}_{\widehat{n}^\star}^{(\widehat{j}^\star)}\right|^2}{\left|\boldsymbol{h}^{(j^\star)}\boldsymbol{f}_{n^\star}^{(j^\star)}\right|^2}. \tag{26}$$

Recall that $j^\star$ and $n^\star$ are the indices of the actual optimal mmWave BS and narrow beam, respectively, while $\widehat{j}^\star$ and $\widehat{n}^\star$ represent the corresponding predicted optimal indices.

2) mmWave measurement overhead O which is calculated as

$$O=K_{\mathrm{w}}+K_{\mathrm{n}}+K_{\mathrm{ant}}. \tag{27}$$

The measurement overhead O reflects the communication cost. It is noted that, without sub-6 GHz channel information, the conventional beam search requires $O=JN$ measurements according to (6).

Since the performance of our proposed deep model may depend on the model initialization in training, the average results of 3 training runs are actually used as the evaluation metrics.

B. Simulation Results

1) Complexity of the proposed model: Fig. 5 depicts the numbers of parameters for our proposed deep learning model under different numbers of mmWave BSs J and candidate narrow beams N. Obviously, the number of model parameters increases with J and N due to the larger size of the mmWave

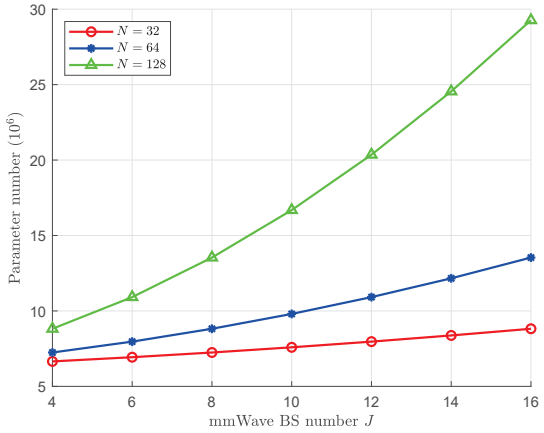
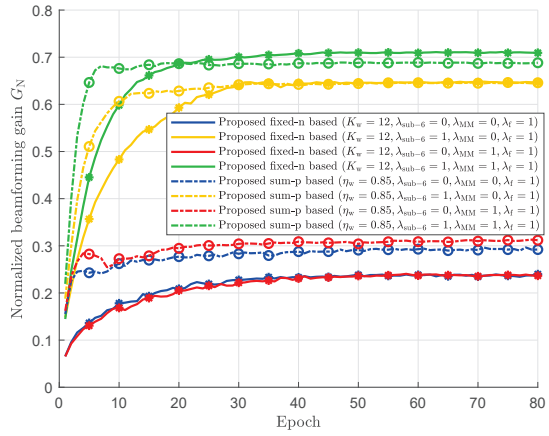


Fig. 5. Complexity of our proposed deep learning model in terms of parameter number under different numbers of mmWave BSs and candidate narrow beams.

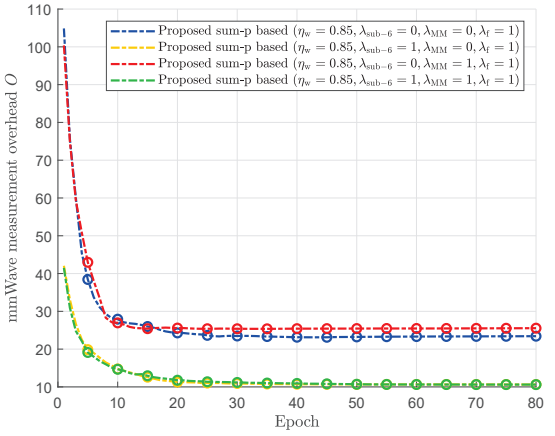
narrow beam output probabilities. Furthermore, the relatively small size of our proposed model with $\sim$ 10M parameters can be easily deployed at BS side.

2) Investigation of model training and structures: Firstly, Fig. 6 compares the convergence performance of our proposed schemes under different weight coefficients in the loss function (22), given $K_{\mathrm{w}}=12$ for the fixed-n based scheme and $\eta_{\mathrm{w}}=0.85$ for the sum-p based scheme. Specifically, Figs. 6(a) and 6(b) depict the normalized beamforming gain G_{N} and the mmWave measurement overhead O, respectively, and the fixed-n based scheme is omitted in Fig. 6(b) owing to its fixed overhead, i.e., $O=K_{\mathrm{w}}$.

It is clear that both the proposed fixed-n based and sum-p based schemes converge around 50 epochs. Interestingly, the number of the measured wide beams O of our proposed sum-p based scheme decreases with the training epoch in Fig. 6(b), because the deep learning model can predict the optimal mmWave beam more accurately via training, so that most predicted probabilities are more likely to be possessed by fewer wide beams. The results in Fig. 6(a) show that both our proposed schemes with $\lambda_{\text{sub-6}}=\lambda_{\mathrm{MM}}=\lambda_{\mathrm{f}}=1$ significantly outperform their counterparts with other loss function designs. Furthermore, Fig. 6(b) demonstrates that



(a)



(b)

Fig. 6. Convergence of our proposed schemes in terms of (a) normalized beamforming gain and (b) mmWave measurement overhead under different weight coefficients of the loss function.

the mmWave overhead O of the sum-p based scheme with $\lambda_{\text{sub-6}} = \lambda_{\text{MM}} = \lambda_{\text{f}} = 1$ is much less than the schemes with $\lambda_{\text{sub-6}} = 0$, and is almost the same as the scheme with $\lambda_{\text{sub-6}} = \lambda_{\text{f}} = 1, \lambda_{\text{MM}} = 0$. These results validate that adding up the loss function components in (22) could effectively facilitate the model optimization. Therefore, the weight coefficients for the loss function components in (22) are set to $\lambda_{\text{sub-6}} = \lambda_{\text{MM}} = \lambda_{\text{f}} = 1$ in the following simulations.

In Fig. 7, we further compare the two methods for calculating the wide beam probabilities in terms of the normalized beamforming gain performance G_{N} of our proposed fixed-n based and sum-p based schemes. Specifically, the first method predicts the probability of the optimal wide beam directly according to the wide beam label with the largest received power, while the second method, as proposed in (14), adds up the probabilities of the narrow beams within the angular range of each wide beam. The G_{N} performance is illustrated as the function of the mmWave measurement overhead O, which corresponds to the number of measured wide beams K_{w} for our proposed schemes, and each point in the curve of the sum-p based scheme depicts the corresponding overhead O and beamforming gain G_{N} at a given probability sum threshold from $\eta_{\text{w}} \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85\}$.

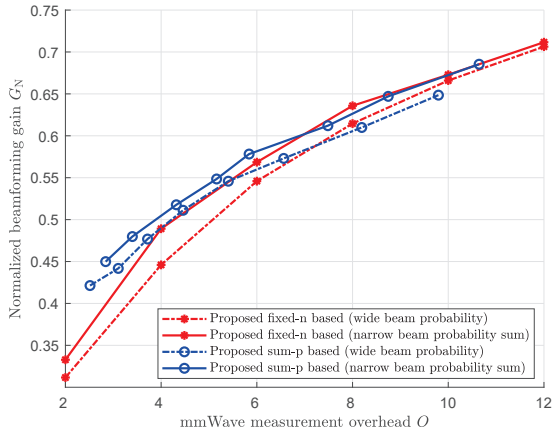


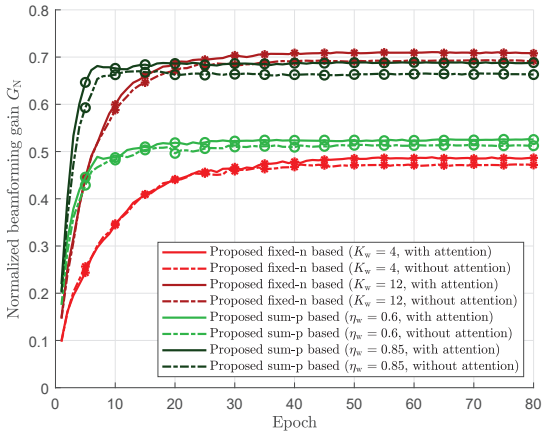
Fig. 7. Normalized beamforming gain as function of mmWave measurement overhead for our proposed schemes, where two methods for calculating wide beam probabilities are investigated.

It is clear that calculating the wide beams as the probability sum of narrow beams could achieve higher G_{N} performance under the same O for the both proposed schemes, which demonstrates that our proposed calculation method is a more suitable criterion to evaluate the qualities of wide beams for the narrow beam prediction.

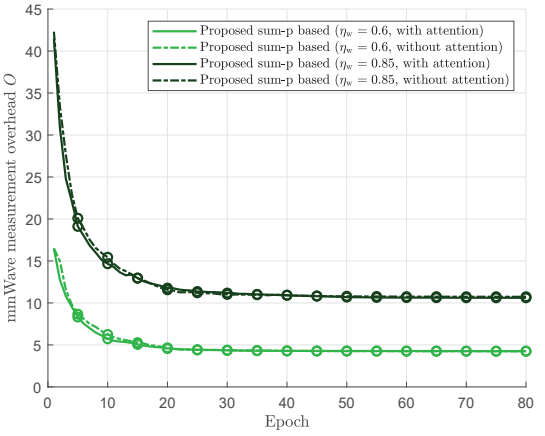
Next, Fig. 8 depicts the convergence performance of our proposed two schemes, in terms of the normalized beamforming gain G_{N} (Fig. 8(a)) and the mmWave measurement overhead O (Fig. 8(b)), where the impact of the attention mechanism is also investigated. For the fixed-n based scheme, we choose two values for the number of measured wide beams as $K_{\text{w}} = 4$ and 12, while for the sum-p based scheme, we set two values for the probability threshold as $\eta_{\text{w}} = 0.6$ and 0.85. Note that the fixed-n based scheme is omitted in Fig. 8(b) owing to its fixed overhead, i.e., $O = K_{\text{w}}$.

It can be seen from Fig. 8(a) that for the fixed-n based scheme with $K_{\text{w}} = 4$, G_{N} converges around 50 epochs, while the same scheme with a larger K_{w} of 12 achieves faster convergence of around 40 epochs. Clearly, the fixed-n based scheme with larger K_{w} achieves higher G_{N}. By contrast, for the sum-p based scheme with $\eta_{\text{w}} = 0.6$, G_{N} converges only around 30 epochs, and the same scheme with $\eta_{\text{w}} = 0.85$ achieves even faster convergence, only around 20 epochs. Obviously, the sum-p based scheme with higher η_{w} attends higher G_{N}. The results of Fig. 8(a) also show that the proposed schemes with the attention mechanism outperform the corresponding schemes without the attention mechanism, enhancing G_{N} by 1% to 2%. Furthermore, the results of Fig. 8(b) demonstrate that the sum-p based scheme with the attention mechanism imposes the same overhead O as its counterpart without the attention mechanism. Not surprisingly, the sum-p based scheme with lower η_{w} has smaller overhead O.

3) Performance comparison with the baselines in line-of-sight (LOS) scenarios: Based on the simulation setup in Subsection IV-A, we compare the performance of our proposed dual-band fusion scheme with the baselines in LOS scenarios. Firstly, Fig. 9 compares the normalized beamforming gain G_{N} performance as the functions of the mmWave measure-



(a)



(b)

Fig. 8. Convergence of our proposed schemes in terms of (a) normalized beamforming gain and (b) mmWave measurement overhead, where the impact of the attention mechanism is investigated.

ment overhead O for our fixed-n and sum-p based schemes as well as the both baselines. No additional narrow beam measurement is considered for our proposed schemes and baseline 2, i.e., O corresponds to the number of measured wide beams K_{w} for our proposed schemes, to the number of measured narrow beams K_{n} for baseline 1, and to the number of measured mmWave antennas K_{ant} for baseline 2, respectively. The comparison of various schemes is made under the same mmWave measurement overhead O. Each point in the curve of our sum-p based scheme illustrates the imposed mmWave overhead and achieved beamforming gain at a given probability sum threshold η_{w}, which is selected from $\eta_{\mathrm{w}} \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85\}$. As expected, the achievable G_{N} performance of all the schemes increase with O, because more mmWave measurements provide more mmWave channel information for facilitating accurate prediction. It can be seen that the G_{N} performance of our proposed schemes and baseline 1 is significantly better than that of baseline 2, since the UE-specific high-quality mmWave beams can achieve higher signal-to-noise ratios for acquiring more reliable mmWave channel information. Moreover, our proposed schemes outperform baseline 1 by around 7% in terms of achievable G_{N}, which validates that the wide beams

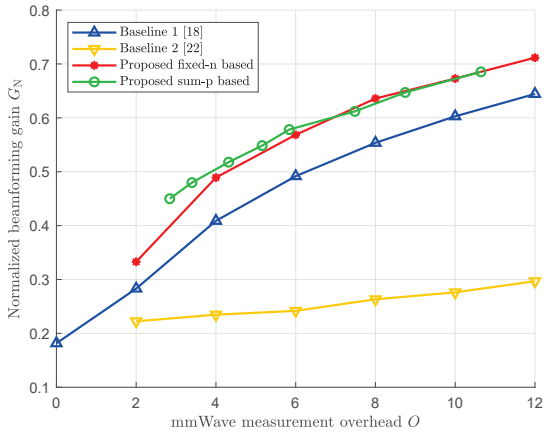


Fig. 9. Normalized beamforming gain as function of mmWave measurement overhead for various schemes, where the proposed schemes and baseline 2 do not take additional narrow beam measurement.

can effectively cover the large uncertain area under low overhead for enhancing the prediction performance, compared to the narrow beams. In addition, our sum-p based scheme achieves larger G_{N} than our fixed-n based scheme when O is small. In order to further illustrate the achievable performance under small to modest mmWave measurement overhead O, the cumulative distribution functions (CDFs) of the achievable G_{N} performance of various schemes are depicted in Fig. 10, where the mmWave measurement overhead is set to $O = 6$ for the both baselines and our fixed-n based scheme, while our sum-p based scheme has $O = 5.79$ with the probability threshold $\eta_{\mathrm{w}} = 0.7$. It can be seen that since our sum-p based scheme can adaptively adjust the number of measured wide beams according to the prediction confidence, the number of the UE suffering from severe performance degradation has been effectively reduced, compared to our fixed-n based scheme. However, as O increases, the achievable G_{N} of the fixed-n based scheme approaches that of the sum-p based scheme, which can be clearly seen from Fig. 9.

Taking the additional narrow beam measurement can obviously improve the performance. We further investigate the normalized beamforming gain performance G_{N} as the function of the mmWave measurement overhead O for various schemes

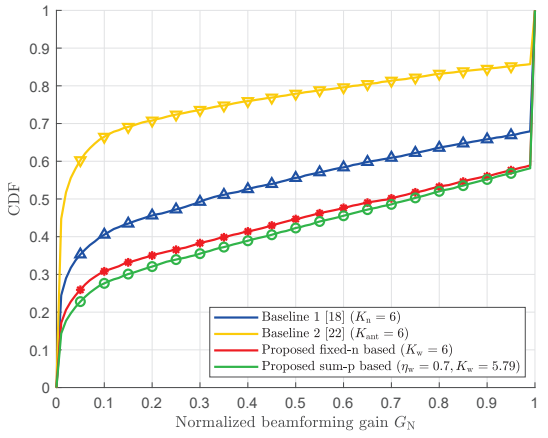


Fig. 10. Comparison of the CDFs of normalized beamforming gain for various schemes, where the proposed schemes and baseline 2 do not take additional narrow beam measurement.

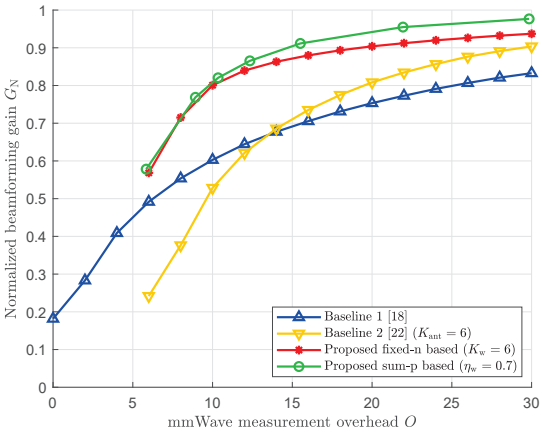
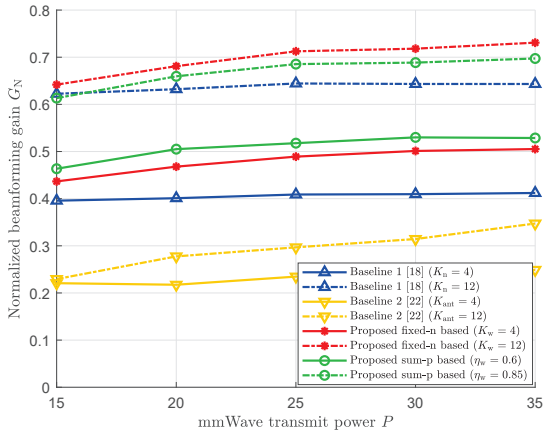


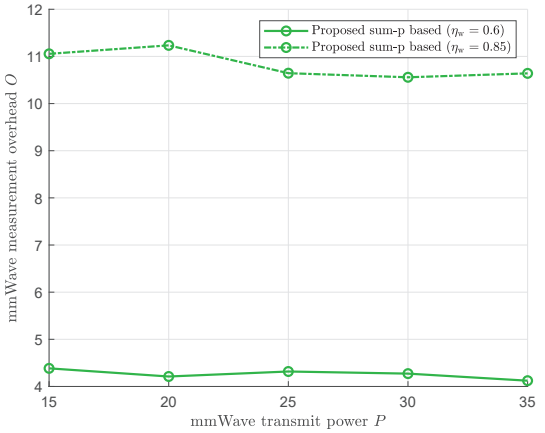
Fig. 11. Normalized beamforming gain as function of mmWave measurement overhead for various schemes, where the proposed schemes and baseline 2 take additional narrow beam measurement.

in Fig. 11, given $K_{\text{ant}} = K_{\text{w}} = 6$ for baseline 2 and our fixed-n based scheme, and $\eta_{\text{w}} = 0.7$ for our sum-p based scheme. The both baselines and our fixed-n based scheme use the fixed-number criterion to select the predicted optimal narrow beams, while our sum-p based scheme adopts the sum-probability criterion to select the narrow beams, given the probability threshold of the narrow beam prediction $\eta_{\text{n}} \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$. Therefore, the mmWave measurement overhead is $O = K_{\text{w}} + K_{\text{n}}$ for our two proposed schemes, and $O = K_{\text{ant}} + K_{\text{n}}$ for baseline 2, while the overhead is $O = K_{\text{n}}$ for baseline 1. Firstly, it can be seen that our proposed schemes achieve accurate beam alignment with $G_{\text{N}} > 80\%$ under $O = 10$, which significantly reduces the mmWave communication costs compared to the conventional beam search with $O = JN = 768$ measurements. Then, the comparison of various schemes is made under the same mmWave measurement overhead O, which can only start from $O \geq 6$. Both the proposed schemes significantly outperform the baselines. This is because, when the UE-specific additional narrow beam measurement has provided reliable mmWave channel information, the effective feature fusion of the sub-6 GHz and mmWave bands based on deep learning plays the dominant role for facilitating accurate beam prediction. Moreover, the G_{N} performance of our sum-p based scheme surpasses the fixed-n counterpart by around 4% when $O > 15$, and it obtains the almost perfect beam alignment of $G_{\text{N}} = 95.5\%$ with $O = 22$. Interestingly, baseline 1 outperforms baseline 2 when the total mmWave measurement overhead $O < 14$.

The impact of the mmWave transmit power P on the performance of various schemes is investigated in Fig. 12, where Fig. 12(a) illustrates the normalized beamforming gain G_{N} for various schemes, and Fig. 12(b) depicts the mmWave measurement overhead O for our proposed sum-p based scheme. It can be seen from Fig. 12(a) that the achievable G_{N} of baseline 2 and both the proposed schemes increases with P due to higher signal-to-noise ratios (SNRs) of the mmWave low-overhead measurements. By contrast, the G_{N} performance of baseline 1 is robust to P, since baseline 1 does not use the mmWave information as the prediction input. Furthermore,



(a)



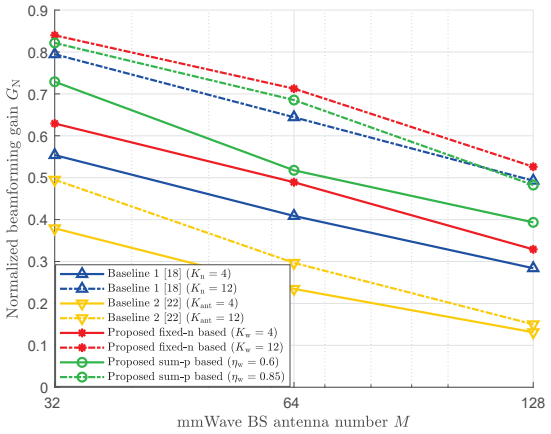
(b)

Fig. 12. Impact of mmWave transmit power on (a) normalized beamforming gains for various schemes, and (b) mmWave measurement overhead for the proposed sum-p based scheme.

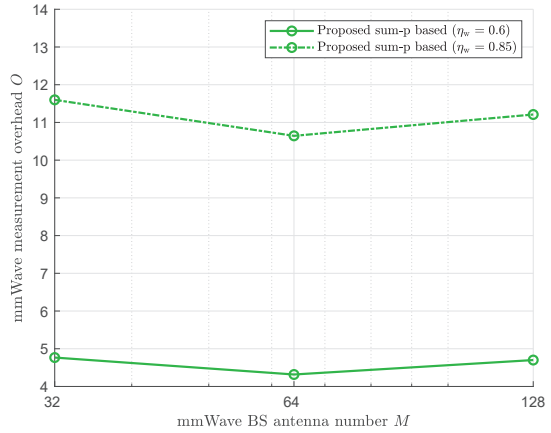
Fig. 12(b) shows that the mmWave transmit power P has little effect on the number of measured mmWave wide beams for our proposed sum-p based scheme, since the wide beam selection only relies on the sub-6 GHz CSI.

Fig. 13 depicts the impact of the number of mmWave BS antennas M on the performance of various schemes, where Fig. 13(a) illustrates the normalized beamforming gains G_{N} for various schemes, and Fig. 13(b) depicts the mmWave measurement overhead O for our sum-p based scheme. More specifically, Fig. 13(a) shows that the achievable G_{N} of all the schemes decrease with M. This is because more mmWave antennas could generate narrower beamwidth, making it more difficult to accurately predict the optimal narrow beam. Furthermore, the number of measured mmWave wide beams of our proposed sum-p based scheme maintains stable in Fig. 13(b), because the wide beam selection only adopts the sub-6 GHz CSI as the input.

Next, we investigate the impact of the sub-6 GHz antenna number $\overline{M}$ in Fig. 14. Specifically, Fig. 14(a) depicts the normalized beamforming gain G_{N} as the function of $\overline{M}$ for various schemes, while Fig. 14(b) illustrates the average mmWave measurement overhead O as the function of $\overline{M}$ for our sum-p based scheme. It can be seen from Fig. 14(a) that the

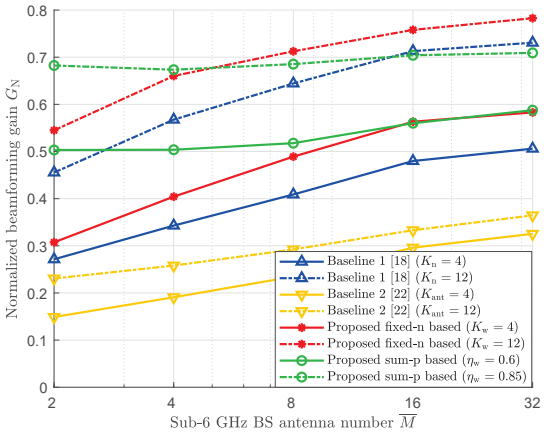


(a)

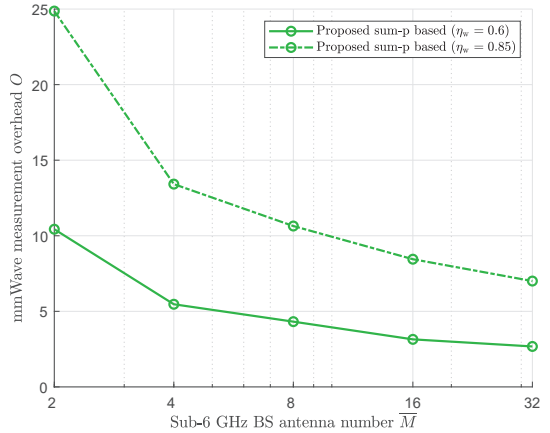


(b)

Fig. 13. Impact of mmWave BS antenna number on (a) normalized beamforming gains for various schemes, and (b) mmWave measurement overhead for the proposed sum-p based scheme.



(a)



(b)

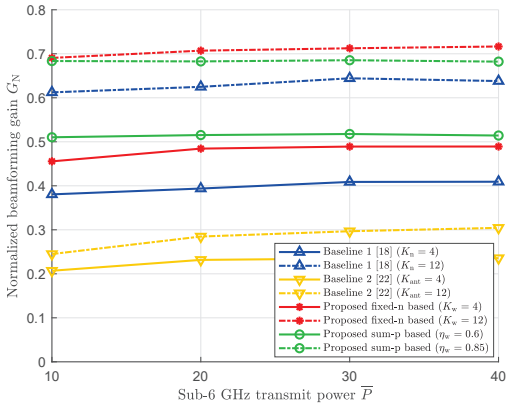
Fig. 14. Impact of sub-6 GHz BS antenna number on (a) normalized beamforming gains for various schemes, and (b) mmWave measurement overhead for the proposed sum-p based scheme.

achievable G_{N} increases with $\overline{M}$ for the both baselines and our fixed-n based scheme, given the same mmWave measurement overhead. This is because a larger number of sub-6 GHz antennas can provide more comprehensive information of the wireless environment and UE location for supporting accurate mmWave beam prediction. By contrast, the achievable G_{N} remains relatively unchanged for our sum-p based scheme as $\overline{M}$ increases. To understand why, we need to look Fig. 14(b), which shows that as $\overline{M}$ increases, our sum-p based scheme reduces the mmWave measurement overhead O accordingly to maintain the same G_{N}. This exactly demonstrates that our sum-p based scheme is capable of adaptively adjusting the number of selected wide beams to be measured based on different prediction confidences under various $\overline{M}$. Therefore, the probability sum threshold is a more robust criterion to different prediction confidences, which is capable of achieving the same excellent performance in diverse wireless environments.

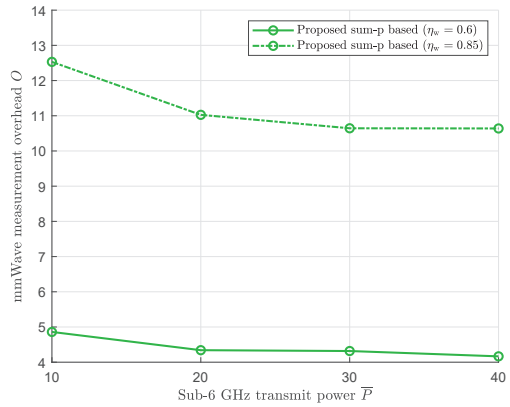
Finally, we investigate the impact of the sub-6 GHz transmit power $\overline{P}$ on the normalized beamforming gain G_{N} and the mmWave measurement overhead O in Fig. 15. Fig. 15(a) shows that the achievable G_{N} increases with $\overline{P}$ for the both baselines and our fixed-n based scheme due to higher sub-6 GHz SNRs. By contrast, it can be seen that the mmWave measurement overhead O of our proposed sum-p based scheme decreases as $\overline{P}$ increases in Fig. 15(b), while the corresponding G_{N} performance maintains stable in Fig. 15(a), which verifies that the sum-probability criterion can adaptively adjust the number of measured mmWave wide beams to achieve robust beamforming gain performance under different sub-6 GHz SNRs.

4) Performance comparison with the baselines in non-line-of-sight (NLOS) scenarios: To further validate the effectiveness of our proposed dual-band fusion approach, we investigate the beamforming gain performance of various schemes in the complex NLOS scenarios. Specifically, we assume that the UE is uniformly distributed in R1∼R1400 and R1900∼R2751 of User Grid 1 in Fig. 4, while one sub-6 GHz BS at BS 15 and 4 mmWave BSs at BS $15 \sim 18$ are considered in the wireless HetNet. The transmit power of sub-6 GHz and mmWave BSs is set to $\overline{P} = 40$ dBm and $P = 35$ dBm, respectively, and other simulation setups are consistent with Subsection IV-A.

Taking additional narrow beam measurement into consideration, the normalized beamforming gain performance G_{N} as the functions of the mmWave measurement overhead O are compared in Fig. 16, given $K_{\mathrm{ant}} = K_{\mathrm{w}} = 4$ for baseline 2 and our fixed-n based scheme, and $\eta_{\mathrm{w}} = 0.8$ for our



(a)



(b)

Fig. 15. Impact of sub-6 GHz transmit power on (a) normalized beamforming gain for various schemes, and (b) mmWave measurement overhead for the proposed sum-p based scheme.

sum-p based scheme. The fixed-number criterion is adopted to select the predicted optimal narrow beams for the both baselines and our fixed-n based scheme, while our sum-p based scheme uses the sum-probability criterion to select the narrow beams, where the probability threshold of the narrow beam selection is $\eta_n \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$. Clearly, our proposed schemes outperform the both baselines, and the G_N performance of our sum-p based scheme surpasses its fixed-n counterpart by around 5% under $6 < O < 8$. Moreover, both

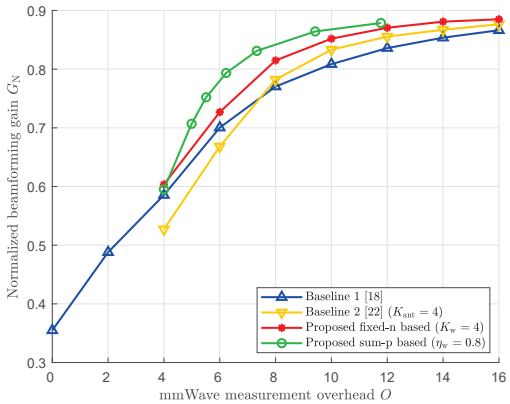


Fig. 16. Normalized beamforming gain as function of mmWave measurement overhead for various schemes in NLOS scenarios, where the proposed schemes and baseline 2 take additional narrow beam measurement.

the proposed schemes could achieve accurate beam alignment of $G_N > 80\%$ under $O = 8$, which demonstrates that our proposed dual-band fusion approach could efficiently find the optimal mmWave beam in the NLOS scenarios.

V. Conclusions

In this paper, by exploiting the inter-BS CSI dependence, we have proposed a dual-band fusion approach to integrate the sub-6 GHz CSI and mmWave low-overhead measurement to predict the optimal mmWave beam in the HetNet and reduce the overhead of both mmWave BS selection and beam training. Besides, deep learning is adopted to accurately model the complex inter-BS CSI dependence. Concretely, the UE-specific high-quality mmWave wide beams predicted by the sub-6 GHz CSI have been proposed as the mmWave low-overhead measurement. The sum-probability criterion, which selects the max-probability wide beams whose probability sum is larger than the predefined threshold, has been further developed to flexibly adjust the number of measured mmWave wide beams according to the prediction confidence. Furthermore, the attention mechanism has been utilized for adaptively weighting the features extracted from the sub-6 GHz CSI and mmWave wide beams for enhancing the prediction accuracy. Simulation results have demonstrated that our proposed scheme achieves higher beamforming gain with smaller mmWave measurement overhead, compared to the existing state-of-the-art sub-6 GHz CSI assisted mmWave beam prediction schemes.

References

[1] M. Xiao, *et al.*, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.

[2] S. K. Sharma, *et al.*, "Dynamic spectrum sharing in 5G wireless networks with full-duplex technology: Recent advances and research challenges," *IEEE Commun. Surv. Tuts.*, vol. 20, no. 1, pp. 674–707, Firstquarter 2018.

[3] T. S. Rappaport, *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[4] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[5] A. Alkhateeb, O. El 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.

[6] Z. Sha, Z. Wang, S. Chen, and L. Hanzo, "Graph theory based beam scheduling for inter-cell interference avoidance in mmWave cellular networks," *IEEE Trans. Veh. Tech.*, vol. 69, no. 4, pp. 3929–3942, Apr. 2020.

[7] T. E. Bogale and L. B. Le, "Massive MIMO and mmWave for 5G wireless HetNet: Potential benefits and challenges," *IEEE Veh. Tech. Mag.*, vol. 11, no. 1, pp. 64–75, Mar. 2016.

[8] L. Zhao, *et al.*, "Multi-cell hybrid millimeter wave systems: Pilot contamination and interference mitigation," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5740–5755, Nov. 2018.

[9] C. Jeong, J. Park, and H. Yu, "Random access in millimeter-wave beamforming cellular networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 180–185, Jan. 2015.

[10] J. Wang, *et al.*, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.

[11] J. Wang, *et al.*, "Beamforming codebook design and performance evaluation for 60 GHz wideband WPANs," in *Proc. VTC 2009 Fall* (Anchorage, AK, USA), Sep. 20–23, 2009, pp. 1–6.

[12] L. Zhao, D. W. K. Ng, and J. Yuan, "Multi-user precoding and channel estimation for hybrid millimeter wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1576–1590, Jul. 2017.

[13] 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.

[14] F. Maschietti, D. Gesbert, and P. de Kerret, "Coordinated beam selection in millimeter wave multi-user MIMO using out-of-band information," in *Proc. ICC 2019* (Shanghai, China), May 20–24, 2019, pp. 1–6.

[15] A. Ali, N. Gonzalez-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] Z. Li, C. Zhang, I. T. Lu, and X. Jia, "Hybrid precoding using out-of-band spatial information for multi-user multi-RF-chain millimeter wave systems," *IEEE Access*, vol. 8, pp. 50872–50883, Mar. 2020.

[17] P. Laxmikanth, *et al.*, "Enhancing the performance of AOA estimation in wireless communication using the MUSIC algorithm," in *Proc. SPACES 2015* (Guntur, India), Jan. 2–3, 2015, pp. 448–452.

[18] 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.

[19] 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, 2020.

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

[21] K. Ma, P. Zhao, and Z. Wang, "Deep learning assisted robust mmWave beam prediction using low-frequency information," to appear in *China Commun.*

[22] 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.

[23] A. S. Mubarak, H. Esmaiel, and E. M. Mohamed, "LTE/Wi-Fi/mmWave RAN-level interworking using 2C/U plane splitting for future 5G networks," *IEEE Access*, vol. 6, pp. 53473–53488, 2018.

[24] J. Zhao, *et al.*, "Multiband cooperation for 5G HetNets: A promising network paradigm," *IEEE Veh. Tech. Mag.*, vol. 14, no. 4, pp. 85–93, Dec. 2019.

[25] Z. Jiang, *et al.*, "Exploiting wireless channel state information structures beyond linear correlations: A deep learning approach," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 28–34, Mar. 2019.

[26] S. Chen, *et al.*, "Learning-based remote channel inference: Feasibility analysis and case study," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3554–3568, Jul. 2019.

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

[28] A. A. M. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 2, pp. 128–137, Feb. 1987.

[29] 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.

[30] S. Noh, M. D. Zoltowski, and D. J. Love, "Multi-resolution codebook and adaptive beamforming sequence design for millimeter wave beam alignment," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5689–5701, Sep. 2017.

[31] K. Chen, C. Qi, and G. Y. Li, "Two-step codeword design for millimeter wave massive MIMO systems with quantized phase shifters," *IEEE Trans. Signal Process.*, vol. 68, no. 1, pp. 170–180, Jan. 2020.

[32] Z. Xiao, T. He, P. Xia, and X. G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.

[33] A. Alkhateeb, *et al.*, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, p-p. 37328–37348, Jun. 2018.

[34] J. Vieira, *et al.*, "Deep convolutional neural networks for massive MIMO fingerprint-based positioning," in *Proc. PIMRC 2017* (Montreal, QC, Canada), Oct. 8-13, 2017, pp. 1–6.

[35] S. Zhang, *et al.*, "Deep learning based channel extrapolation for large-scale antenna systems: Opportunities, challenges and solutions," *IEEE Wireless Commun.*, vol. 28, no. 6, pp. 160–167, Dec. 2021.

[36] M. Alrabeiah and A. Alkhateeb, "Deep learning for TDD and FDD massive MIMO: Mapping channels in space and frequency," in *Proc. Asilomar 2019* (Pacific Grove, CA, USA), Nov. 3–6, 2019, pp. 1465–1470.

[37] B. Lin, *et al.*, "Deep learning-based antenna selection and CSI extrapolation in massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7669–7681, Nov. 2021.

[38] J. Vieira, *et al.*, "Deep convolutional neural networks for massive MIMO fingerprint-based positioning," in *Proc. PIMRC 2017* (Montreal, QC, Canada), Oct. 8–13, 2017, pp. 1–6.

[39] H. Echigo, Y. Cao, M. Bouazizi, and T. Ohtsuki, "A deep learning-based low overhead beam selection in mmWave communications," *IEEE Trans. Veh. Tech.*, vol. 70, no. 1, pp. 682–691, Jan. 2021.

[40] K. Ma, *et al.*, "Deep learning assisted calibrated beam training for millimeter-wave communication systems," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6706–6721, Oct. 2021.

[41] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS 2012* (Lake Tahoe, NV, USA), Dec. 3–6, 2012, pp. 1–9.

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

[44] 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, Jan. 2015.

[45] 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.

[46] W. Feng, *et al.*, "When mmWave communications meet network densification: A scalable interference coordination perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1459–1471, Jul. 2017.

[47] Z. Sha, S. Chen, and Z. Wang, "Near interference-free space-time user scheduling for mmWave cellular network," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6372–6386, Aug. 2022.

[48] X. Lin, *et al.*, "5G new radio: Unveiling the essentials of the next generation wireless access technology," *IEEE Commun. Standards Mag.*, vol. 3, no. 3, pp. 30–37, Sep. 2019.

[49] A. Mazin, M. Elkourdi, and R. D. Gitlin, "Accelerating beam sweeping in mmWave standalone 5G new radios using recurrent neural networks," in *Proc. VTC-Fall 2018* (Chicago, IL, USA), Aug. 27–30, 2018, pp. 1–4.

[50] S. H. Lim, S. Kim, B. Shim, and J. W. Choi, "Deep learning-based beam tracking for millimeter-wave communications under mobility," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7458–7469, Nov. 2021.

[51] K. Ma, D. He, H. Sun, and Z. Wang, "Deep learning assisted mmWave beam prediction with prior low-frequency information," in *Proc. ICC 2021* (Montreal, QC, Canada), Jun. 14–23, 2021, pp. 1–6.

[52] Remcom, Wireless InSite, *http://www.remcom.com/wireless-insite*.

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

[54] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.