This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3324916

PARAMOUNT: Towards generalizable deeP leARning for mmwAve beaM selectiOn using sUb-6GHz chaNnel measuremenTs

Katarina Vuckovic, Mahdi Boloursaz Mashhadi, *Senior Member, IEEE*, Farzam Hejazi, Nazanin Rahnavard, *Senior Member, IEEE*, Ahmed Alkhateeb

***Abstract*—Deep neural networks (DNNs) in the wireless communication domain have been shown to be hardly generalizable to scenarios where the train and test datasets follow a different distribution. This lack of generalization poses a significant hurdle to the practical utilization of DNNs in wireless communication. In this paper, we propose a generalizable deep learning approach for millimeter wave (mmWave) beam selection using sub-6 GHz channel state information (CSI) measurements, referred to as PARAMOUNT. First, we provide a detailed discussion on physical aspects of the electromagnetic wave scattering in the mmWave and sub-6 GHz bands. Based on this discussion, we develop the augmented discrete angle delay profile (ADADP) which is a novel linear transformation for the sub-6 GHz CSI that extracts the angle-delay attributes and provides a semantic visual representation of the multi-path clusters. Next, we introduce a convolutional neural network (CNN) structure that can learn the signatures of the path clusters in the sub-6 GHz ADADP representation and transform it to mmWave band beam indices. We demonstrate by extensive simulations on several different datasets that PARAMOUNT can generalize beyond the training dataset which is mainly due to transfer learning principles that allow transferring information from previously learned tasks to the learning of new unseen tasks.**

I. Introduction

Fifth-generation (5G) and beyond mobile networks promise to usher in a new era of ultra high-speed communications that surpasses previous generations by several orders of magnitude in communication capacity [1]. One of the core technologies behind such a spectacular revolution is spatial user multiplexing enabled through massive multi-input-multi-output (MIMO). By providing the ability to focus energy on users' devices, massive MIMO empowers pushing the capacity of the network to such immense boundaries required for 5G and beyond communications [2]. Simultaneously, mobile mmWave communication is enabled through 5G networks and transforms directional communication from a promising aspect of next generation networks into a must-have feature [3]. MmWave communication experiences huge attenuation in the open air, requiring the transmitted energy to be directed in narrow rays to meet sufficient signal-to-noise-ratio (SNR) thresholds required at receivers [4].

As *directional* communication has gained importance in the next generation of communication systems, *beam selection* has obtained gravity as an enabler of directional communication. To clarify this necessity, consider that two devices that exploit directional antennas cannot communicate unless they ascertain in which direction they should send/receive signals to/from the other device. Moreover, this knowledge of direction of the other device should be maintained during the communication period; otherwise, the link will be disrupted [5]. To find the best direction for communication, both sides of the link should search for the beam pair that results in the highest SNR. This process is called *beam selection*. The straightforward method for beam selection is to conduct exhaustive search on all the possible beam pairs between the two sides of the link. The exhaustive search will result in a huge overhead, extensive latency [6], and reduced throughput for the communication systems. Speeding up and mitigating the burden of beam selection in mmWave communications has been a topic of extensive research recently [7]. Due to the high path loss in mmWave communications, there are usually a small number of propagation paths between the two sides of the link. Furthermore, due to the sparse nature of communication channel in the angular-delay domain [8], most early works on this topic propose employing techniques based on compressive sensing (CS) [9]–[11]. CS-based mmWave beam alignment techniques have been proposed to exploit sparsity of mmWave channels to reduce the beam search overhead [12], [13]. The authors in [12] proposed a modified orthogonal matching pursuit (OMP) algorithm called logit weighted - OMP (LW-OMP) to improve sparse recovery of mmWave channels using sub-6 GHz channel information. In [13], the authors used the spatial characteristics extracted from the sub-6 GHz band to construct the mmWave channel covariance. This mmWave covariance knowledge can be utilized to reduce the training overhead associated with designing the analog or hybrid analog/digital precoding matrices. However, CS-based methods typically depend on estimating certain spatial parameters, such as angular characteristics and path gains, at the sub-6 GHz band and then utilizing them at mmWave frequencies. This makes their

K. Vuckovic, F. Hejazi and N. Rahnavard are with the ECE Dep., University of Central Florida, USA. M. Boloursaz Mashhadi is with the 5GIC & 6GIC, Institute for Communication Systems (ICS), University of Surrey, UK. A. Alkhateeb is with the Dep. of Electrical Engineering, Arizona State University, USA.

This work is supported by the National Science Foundation under Grant No. CCF-1718195. M. Boloursaz Mashhadi is supported by a grant from the UK Department for Science, Innovation, and Technology under project TUDOR (Towards Ubiquitous 3D Open Resilient Network).

performance highly sensitive to estimation errors in the low-frequency bands. Moreover, CS-based methods still require a considerable amount of beam training overhead to acquire CS measurements at the mmWave band, which increases with the number of antennas. Many recent studies propose machine learning (ML) techniques [14]–[17]. Long et. al cast the beam selection problem as a multi-class classification problem and employ support vector machine (SVM) to achieve a statistical classification model that maximizes the sum rate [14]. Li et. al propose an end-to-end deep learning (DL) technique to design a structured CS matrix based on the underlying channel distribution, leveraging both sparsity and the particular spatial structure that appears in communication channels [16]. In recent years, the integrated sensing and communications (ISAC) models have been explored as a way to accommodate the dynamic changes in the environment. The idea of ISAC is the enhancement of situational awareness by fusing the information from various types of sensors including cameras, RAdio Detection And Ranging (RADAR), LIght Detection And Ranging (LIDAR), and user's Global Position Systems (GPS) information. Vision-aided approaches were proposed in [18], [19], where images taken from cameras mounted within the cellular coverage area are input to a neural network (NN) to predict the optimal beams. In [20], a DL algorithm uses RADAR data for beam prediction in mmWave and terahertz communication systems. Furthermore, NN-based approaches using input data from LIDAR sensors to identify the optimal beam directions have been proposed in [21]–[24]. Lastly, GPS aided beam selection has been used on its own in [25] and combined with cameras in [26].

In an experimental study, a comprehensive channel measurement campaign conducted in Europe in 2014-2016 in numerous indoor and outdoor scenarios showed that geometry of the main propagation paths at sub-6 GHz and mmWave bands are almost similar [27]. This experimental outcome has motivated several recent works to use sub-6 GHz channel information for mmWave beam selection. Additionally, digital beamforming can be challenging in the mmWave band due to hardware limitations. Nevertheless, it is still possible to implement and use it to estimate channel state information (CSI) at the sub-6 GHz band [28]. Ali et. al. introduce the possibility of using sub-6 GHz for mmWave beam selection in [29]. The authors introduce a non-parametric approach (based on interpolation/extrapolation) and a parametric approach (based on estimates of angle of arrival (AOA) and angle spreads) for the sake of mmWave channel estimation using the sub-6 GHz spatial correlation matrix. More recent studies consider using DL for learning the transformation between sub-6 GHz channel and mmWave beam selection. Alrabeiah et. al used a fully connected neural network (FCNN) for mmWave beam selection [15]. They show a 90% accuracy in beam selection using the proposed technique. In [30], the authors demonstrate the performance of FCNN with an experimental testbed in an indoor setup.

The current wireless communication technology has been developed using model-based techniques which have been proven to be successful in tackling real-world wireless communication challenges so far. On the other hand, data-driven approaches such as DL are inherently data dependent. Majority of DL techniques in wireless communication are tied to the training dataset, and there is no guarantee that they can generalize to perform well beyond that training dataset [31]. In contrast, the model-driven techniques can well generalize to many real-world environments. Data dependency constraint of DL techniques is a significant challenge in front of data-driven techniques to replace their model-driven counterparts. Previous studies have proposed deep transfer learning as a means of transferring general features learned from one environment to a new one [32], [33]. However, these models require fine-tuning using an adaptation dataset collected from the new environment. In contrast, our proposed model does not need to be fine-tuned to perform effectively in a new, unseen scenario.

A. Contribution

This paper mainly focuses on the generalization aspects of using DL for mmWave beam selection via sub-6 GHz CSI. We propose a generalizable deep learning approach for millimeter wave (mmWave) beam selection using sub-6 GHz channel state information (CSI) measurements, in short **PARAMOUNT**. We formulate the mmWave beam selection problem using sub-6 GHz CSI in a generalizable fashion such that we can train the proposed deep neural network (DNN) in a specific scenario and expect it to perform sufficiently well in an unseen scenario. Next, we introduce a new feature space for training deep networks. Then, we discuss why a CNN performs better in terms of generalization in comparison with the CSI + FCNN used in [15]. Finally, we present our proposed CNN structure and demonstrate its superior performance in unseen test scenarios through simulations. The main contributions of this paper can be summarized as follows:

- Based on insights and domain knowledge from the inherent properties of the wireless channels provided in Section V, we here introduce a new transfer (namely, ADADP) which extracts a general set of features/semantics from the input CSI that work sufficiently well for the beam selection task in various environments. The generalization capability of our proposed approach is mainly due to the transfer learning principle which allows transferring information from previously learned tasks to the learning of new unseen tasks without the need for fine-tuning the model.
- We propose a CNN architecture that can effectively extract the physical attributes of the propagation paths from the ADADP. We argue that the ADADP transformation is necessary for the CNN to learn physical characteristics of the communication channel and to extend this knowledge beyond the training dataset. We argue that by transforming the raw CSI data into a sparse ADADP, the CNN can focus on the propagation paths in the ADADP instead of shifting its focus on irrelevant details in the CSI and thus impeding generalizability. The combination of the CNN architecture and ADADP representation results in generalization.
- We show sufficient generalization capabilities of the proposed approach even in the case where the training and test datasets stem from separate scenarios by the CSI data

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3324916

collected at sub-6 GHz band to make beam predictions at mmWave band.

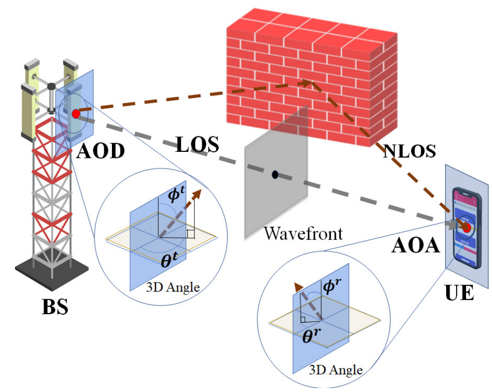
The rest of the paper is structured as follows. In Section II, we discuss physical aspects of electromagnetic wave propagation at sub-6 GHz and mmWave bands. Based on that, in Section III, we discuss massive MIMO CSI channel modeling and derive a mathematical representation of beam selection. In Section IV, we demonstrate learning of the transformation from sub-6 GHz CSI to mmWave Beam Space. In Section V, we discuss the requirement for a semantic representation of the MIMO wireless channel and we introduce a novel angular delay visual transformation of CSI, namely augmented discreet angular delay profile (ADADP) to achieve this. In Section VI, we discuss the structure of the proposed CNN and its generalization capability. We provide extensive simulations to demonstrate the generalization capability of PARAMOUNT in Section VII. Finally, we conclude the paper in Section VIII.

Notations: $\mathbb{C}^{m\times n}$ denotes the $m\times n$ complex space and $j=\sqrt{-1}$. Boldface capital and lower-case letters represent matrices and vectors, respectively (e.g. $\boldsymbol{A}$ and $\boldsymbol{a}$). Calligraphic letters denote sets (e.g. $\mathcal{A}$). Operators $(\cdot)^T$ and $(\cdot)^H$ represent the matrix transpose and Hermitian transpose, respectively. Furthermore, $\otimes$ is the Kronecker product, and $|\cdot|$ is the absolute value.

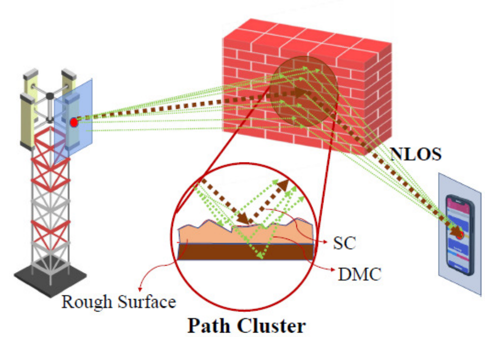
II. Physical Attributes of Wireless Communication Channels

Since there is a solid model-based background knowledge in wireless communications, there is a consensus within the community that DL models should be able to incorporate the available prior knowledge with their structures, otherwise they are doomed to be impractical [34].

A wireless communication channel normally consists of two sides, a base station (BS) (or an access point or a device in case of device-to-device communication) and a user equipment (UE) as shown in Fig. 1(a). Each side of the link is equipped with one or more antenna elements. We assume that the two sides of the link are located in the far-field zone of the antenna with respect to each other. Considering the far-field, the BS and the UE can be assumed to be points at the epicenters of the BS and the UE antennas. Typically, there is a line-of-sight (LOS) path (which may be blocked) and a number of non-LOS (NLOS) paths between the BS and the UE. Each propagation path is denoted by a line segment between the BS and UE in the 3D space. We assume that the propagation path between the BS and UE is the same regardless of which side transmits the signal. Each propagation path identifies a wave-front plane perpendicular to the path and can be characterized by a delay, two 3D angles (zenith and azimuth), and a power. The delay is the length of the path divided by the speed of light. The pair (ϕ^t,θ^t) indicate the angle of departure (AOD) of the path from the BS and the pair (ϕ^r,θ^r) denote the angle of arrival (AOA) of the path to the UE. The gain of the path shows how strong a signal can be transmitted over the path.The LOS path is the straight line between the BS and the UE whose geometry is frequency independent. The NLOS paths between the BS and the UE are created because of the interactions of the



(a)



(b)

Fig. 1: a) The wireless channel has LOS and NLOS propagation paths, each originating from the BS antenna center and terminating at the UE antenna center. These paths have a perpendicular wavefront and are characterized by AOD and AOA angle pairs, length, and complex power. b) The diffusion from the rough surface will result in a cluster of path between the BS and the UE. Each cluster consists of a strong SC and numerous DMCs.

electromagnetic wave and the surfaces of the scatterers in the environment. Since the surfaces within the environment may be rough, three phenomenon happen when an electromagnetic wave impinges on a surface [35]:

- A fraction of the wave reflects specularly from the surfaces (specular component (SC))
- A fraction of the energy is absorbed by or passes through the surfaces and
- A fraction of the energy diffuses from the surfaces.

The diffusion results in numerous reflection paths with slightly different angles and delays with respect to the angle and the delay of the SC path from the surfaces. This results in a dense distribution of weak propagation paths around the SC called dense multi-path components (DMC) (Fig. 1(b)). The SC plus DMCs from a surface create a cluster of paths. Each path cluster can be characterized by the angle and the delay of the SC, a delay spread, and an angle spread which specify DMCs. With respect to its definition, the geometrical characteristics of a SC is frequency agnostic, thus path clusters at different frequencies show similar SCs in the geometrical sense. On the other hand, the diffusion from a surface is directly proportional to the wavelength and hence, it is frequency dependent [27]. Due to the facts that (i) the SCs geometrical attributes are frequency agnostic, and (ii) DMCs have very similar angle and delay profile to those of their corresponding SC, the overall geometrical characteristics of path clusters at

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3324916

different frequency bands are similar.

The geometrical similarity between propagation paths at sub-6 GHz and mmWave bands has been experimentally observed and demonstrated via various empirical channel measurement campaigns. Further, the measurements campaigns show that the stronger cluster in sub-6 GHz tends to be the strongest in mmWave band as well [27], [30], [36]. Therefore, if we can train a model to learn the strongest path in sub-6 GHz, then we would also know the strongest path in the mmWave channel, which then can be mapped to the optimal mmWave beam. In this work, we propose a ADADP transformation of the CSI, which enables semantic representation of the LOS and NLOS path clusters. While the ADADP enables us to identify the strongest path cluster, the mapping from the path cluster to the optimal mmWave beam is best achieved through the training our proposed CNN model, as discussed in Section VI. Other than the scattering, blockage also results in path cluster creation with very weak DMCs. Path blockage at sub-6 GHz results in 15-25 dB loss and refraction [27], leading to DMC creations. At mmWave bands, blockage causes even higher loss, rendering the path completely blocked. This work focuses on propagation paths with significant energy between the BS and UE, and thus we will not consider blocked paths for comparing the communication channel at mmWave and sub-6 GHz bands.

III. Massive MIMO Channel Modeling And Problem Formulation

The antenna configuration at the BS and the UE has a direct impact on how meticulously the BS and the UE can estimate and then utilize the channel. Discussing the channel estimation mechanism for massive MIMO orthogonal frequency-division multiplexing (OFDM) systems, we will explain how massive MIMO provides information about the angular characteristics of the propagation paths and OFDM provides information about the delay of the propagation paths.

A. Channel Modeling at Sub-6 GHz

Let us consider a sub-6 GHz MIMO-OFDM wireless system in which both the BS and the UE are equipped with an antenna array with N_B and N_U elements, respectively; and they use OFDM signaling with N_c subcarriers and B_c subcarrier spacing. The received signal at the UE antenna array at the l^{th} subcarrier in the frequency domain can be written as

$$\boldsymbol{y}[l] = \boldsymbol{h}[l]\boldsymbol{s}[l] + \boldsymbol{n}[l]\,, \tag{1}$$

where $\boldsymbol{y}[l] \in \mathbb{C}^{N_U \times 1}$ denotes the received signal, $\boldsymbol{h}[l] \in \mathbb{C}^{N_U \times N_B}$ denotes the channel matrix, $\boldsymbol{s}[l] \in \mathbb{C}^{N_B \times 1}$ denotes the transmitted signal at the BS, and $\boldsymbol{n}[l] \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0},\, \sigma^2 \boldsymbol{I})$ denotes the receiver noise. We assume there are C distinguishable path clusters between the BS and the UE. Moreover, each cluster constitutes R_C distinguishable paths. Each path can be characterized by a delay $\tau_m^{(k)}, k \in \{1,\ldots,C\}, m \in \{1,\ldots,R_C\}$, an (azimuth, elevation) AOD from the BS's antenna characterized by $(\theta_m^{(t)(k)}, \phi_m^{(t)(k)})$, an (azimuth, elevation) AOA to the UE's antenna characterized by $(\theta_m^{(r)(k)}, \phi_m^{(r)(k)})$ and a complex gain $\alpha_m^{(k)}$ [37]. The path geometry does not change depending on weather the channel is an uplink or downlink. This means that the AOD and AOA at the UE are equivalent and the same is true for the BS. Therefore, AOD and AOA can be used interchangeably throughout this paper. Given a wide-band OFDM system, $\tau_m^{(k)} = n_m^{(k)} T_s$, where T_s and $n_m^{(k)}$ denote the sampling duration and the sampled delay belonging to the path m of the cluster k, respectively [38]. With Nyquist rate of sampling, $T_s = \frac{1}{B}$, where $B = N_c B_c$ is the total bandwidth of the system and B_c represents the subcarrier spacing. Therefore, channel frequency response (CFR) for each subcarrier l can be written as

$$\boldsymbol{h}[l] = \sum_{k=1}^{C} \sum_{m=1}^{R_C} \alpha_m^{(k)} (\boldsymbol{e_m}^{(r)} \otimes \boldsymbol{e_m}^{(t)}) e^{-j2\pi \frac{l\, n_m^{(k)}}{N_c}}\,, \tag{2}$$

such that

$$\boldsymbol{e_m}^{(r)} = \boldsymbol{e}^{(r)} \left(\theta_m^{(r)(k)}, \phi_m^{(r)(k)}\right) \tag{3}$$

$$\boldsymbol{e_m}^{(t)} = \boldsymbol{e}^{(t)} \left(\theta_m^{(t)(k)}, \phi_m^{(t)(k)}\right) \tag{4}$$

where $\boldsymbol{e_m}^{(r)} \in \mathbb{C}^{N_U \times 1}$ and $\boldsymbol{e_m}^{(t)} \in \mathbb{C}^{N_B \times 1}$ denote the array response vector of the BS and the UE, respectively [39]. The overall CFR matrix of the channel between the BS and the UE can be expressed as $\boldsymbol{H} = [\boldsymbol{h}[1]\ \boldsymbol{h}[2]\ \ldots\ \boldsymbol{h}[N_c]]$. In literature, this matrix is commonly referred to as CSI. As (2) shows, the AOA and AOD of each path is indirectly preserved in array response vectors ($\boldsymbol{e_m}^{(r)}$ and $\boldsymbol{e_m}^{(t)}$), and delay of the path is preserved in the subcarrier phase shift ($e^{-j2\pi \frac{l\, n_m^{(k)}}{N_c}}$). CSI is directly measurable via MIMO-OFDM systems equipped with fully digital beamforming which is essential for the wireless communication systems to gain understanding about the channel.

B. mmWave Channel Model and CSI Estimation

Let us consider a mmWave MIMO-OFDM wireless system in which the BS is equipped with an antenna array with N_B^{mm} elements and the UE is equipped with an antenna array with N_U^{mm} elements and use OFDM signaling with N_C^{mm} subcarriers and B_C^{mm} subcarrier spacing. One profound difference between mmWave band and sub-6 GHz band is that digital beamforming is not accessible at mmWave bands due to hardware limitations. This means that CSI is not directly measurable at mmWave band. MmWave MIMO-OFDM systems are typically equipped with analog beamforming. Thus, the received signal at the UE antenna array at the l^{th} subcarrier turns out to be [40]

$$y^{mm}[l] = \boldsymbol{q}\boldsymbol{h}^{mm}[l]\boldsymbol{w}s^{mm}[l] + n^{mm}[l]\,, \tag{5}$$

where $s^{mm}[l] \in \mathbb{C}$ denotes the signal, $\boldsymbol{h}^{mm}[l] \in \mathbb{C}^{N_U^{mm} \times N_B^{mm}}$ denotes the mmWave channel response, $\boldsymbol{q} \in \mathbb{C}^{1 \times N_U^{mm}}$ is the combining vector at the UE, $\boldsymbol{w} \in \mathbb{C}^{N_B^{mm} \times 1}$ is the precoder vector at the BS and $n^{mm}[l] \in \mathcal{N}(0, (\sigma^{mm})^2)$ is the noise. So instead of measuring a high-dimensional channel response matrix $\boldsymbol{h}^{mm}[l] \in \mathbb{C}^{N_U^{mm} \times N_B^{mm}}$, we can only measure a one dimensional $\boldsymbol{q}\boldsymbol{h}^{mm}[l]\boldsymbol{w}$ due to the lack of fully digital beamforming at mmWave.

CSI Estimation: Obtaining CSI in 6-GHz band is feasible and cost effective due to availability of fully digital transceiver chain and several techniques have been proposed to perform the task [41]–[43]. On the other hand, mmWave channel estimation is a challenging task due to high power consumption of base band mix signal components, small SNR before beamforming, and greater number of antennas [44]. Thus, mmWave massive MIMO channel estimation has remained under extensive research [45]–[47]. Since, only analog receiver chain with digital phase shifters has been feasible for mmWave implementation thus far, mmWave massive MIMO systems are only able to shape beams based on predefined beam codebooks. This puts a huge constraint on beamforming ability of mmWave MIMO systems in comparison to their sub-6 GHz counterparts equipped with fully digital transceivers [6].

Beam Selection Problem: Let the set of all precoders at the BS be denoted by $\mathcal{W}$ (precoder codebook) and the set of all combiners at the UE be denoted by $\mathcal{Q}$ (combiner codebook). The achievable rate for a mmWave channel $\boldsymbol{H}^{mm}$ and a given pair of precoder/combiner $\boldsymbol{q}, \boldsymbol{w}$ is derived as [15]

$$R(\boldsymbol{H}^{mm}, \boldsymbol{q}, \boldsymbol{w}) = \sum_{l=1}^{N_C^{mm}} \log_2 \left(1 + \mathrm{SNR}_l \, |\boldsymbol{q}\boldsymbol{h}^{mm}[l]\boldsymbol{w}|^2\right), \quad (6)$$

where $\mathrm{SNR}_l = \frac{\mathbb{E}(|s^{mm}[l]|^2)}{(\sigma^{mm})^2}$ denotes the per subcarrier SNR, $\mathbb{E}(.)$ shows the expectation operator, and $|.|$ represents the absolute value. For *beam selection*, we find the optimal precoder/combiner that maximize the rate as

$$(\boldsymbol{w}^*, \boldsymbol{q}^*) = \underset{\boldsymbol{w}\in\mathcal{W}, \boldsymbol{q}\in\mathcal{Q}}{\operatorname{argmax}} R(\boldsymbol{H}^{mm}, \boldsymbol{q}, \boldsymbol{w})\,. \quad (7)$$

Any given precoder/combinber pair (*beam pair*) points the antenna beam of BS/UE in a specific direction. Since the BS and the UE have no knowledge of neither CSI nor their relative direction, they should conduct a search on predefined sets of precoders and combiners to find the best beam pair with the highest rate. The beam selection of (7) is a non-convex optimization problem and hence should be solved by an exhaustive search on all possible precoder/combiner pairs. We define *beam space* as the set of all possible precoder/combiner pairs and beam space search as the exhaustive search on all possible beam pairs. The exhaustive search method checks for all q and w combinations the to find the representative combination that results in the largest gain at the receiver. Let $N_B^{mm} \times G_{tx}$ matrix $\mathcal{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_{G_{tx}}]$ denote the transmitter codebook consisting of G_{tx} precoding vectors and $N_u^{mm} \times G_{rx}$ matrix $\mathcal{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_{G_{rx}}]$ denote the receiver codebook, consisting of G_{rx} combining vectors. In the beam training phase, the BS uses a precoding vector $\boldsymbol{w}_m \in \mathcal{W}$ and the UE uses a combining vector $\boldsymbol{q}_n \in \mathcal{Q}$. The BS transmits the training OFDM blocks on G_{tx} precoding vectors. For each precoding vector, the UE uses G_{rx} distinct combining vectors. The number of total training blocks is $G_{rx} \times G_{tx}$. The UE determines the best precoder-combiner pair that generates the largest $|y^{nm}[l]|$ for $n = 1, ..., G_{rx}, m = 1, ..., G_{tx}$, and feeds back this information to the BS. The beam space search results in a huge overhead for the system which leads to higher latency and lower throughput. To reduce the overhead and latency of exhaustive search, various multi-resolution beam alignment techniques have been proposed [48], [49]. Multi-resolution beam alignment techniques involve transmitting and receiving signals at different resolutions to efficiently search for the best beam alignment. The alignment process is typically divided into two stages: coarse and fine alignment. In coarse alignment, the transmitter and receiver use relatively wide beamwidths to search for each other's signal. Once the coarse alignment is established, the fine alignment stage uses narrower beamwidths to refine the alignment and optimize the signal quality.

IV. Learning the Transformation from Sub-6 GHz CSI to mmWave Beam Space

The CSI model of (2) holds for both mmWave and sub-6 GHz bands. However, the parameters C, R_C, and $\|\alpha_m^{(k)}\|$ typically are smaller at the mmWave band. Unlike the mmWave band, the digital beamforming is easily accessible at sub-6 GHz band and therefore, CSI is available. Moreover, as we discussed in Section II, the geometry of propagation paths at mmWave and sub-6 GHz bands are very similar. Thus, given we have co-located sub-6 GHz and mmWave antennas at the BS and the UE, we can take advantage of sub-6 GHz CSI to mitigate the beam space search at mmWave band. In other words, since beam space search looks for the propagation path with the highest rate, we can extract the directions of propagation paths from sub-6 GHz CSI. Then, we can utilize this knowledge to reduce the beam space search only to those directions with an existing propagation path. As we discussed in Section III, the information about the geometry of propagation paths is hidden in sub-6 GHz CSI. If we can train a DNN to extract the geometrical information of propagation paths from the CSI and then transform this knowledge to find the optimal precoder/combiner pairs at mmWave, we can drastically mitigate the beam space search overhead.

In [15], the authors train a FCNN to learn the transformation between the sub-6 GHz CSI and the best beam pair for mmWave communication. To train the FCNN, the authors form a dataset of sub-6 GHz CSIs tagged by the best mmWave beam pair. Furthermore, they assume that the mapping from the position of the UE to the sub-6 GHz CSI is a bijective function. The importance of the bijectiveness is that it guarantees the existence of an inverse function that maps the sub-6 GHz channel to the corresponding location. Under this condition, the authors prove that there exists a mapping between the sub-6 GHz channel and optimal mmWave beam. Despite the existence of this mapping, the mapping is very hard to characterize analytically. Therefore, they propose utilizing DL to learn these non-trivial mapping functions. However, their approach is purely data driven and does not take into account the physical similarities between angular-delay distribution of propagation paths at different frequency bands. In Section VII, we empirically demonstrate that in fact the FCNN does not learn to extract the required information from CSI and therefore cannot generalize well beyond the training dataset. However, in our proposed method, we first perform feature engineering to convert the raw CSI into a more informative

ADADP feature. Then, we apply a DL model to map the path clusters in the sub-6 GHz ADADP to the optimal mmWave beam. While the mapping from the path clusters to optimal beam can be performed using classical (non-ML) methods, we demonstrate by simulations that the best performance is achieved by training our proposed CNN model.

V. Extracting Propagation Paths Information From Sub-6 GHz CSI

Herein, we investigate a transformation of the CSI that reveals the geometrical information of propagation paths in a semantic representation learnable by deep networks. For the ease of exposition, let us assume that the massive MIMO system can perfectly separate between two propagation paths in the angular domain which means (assuming infinite number of antennas at the UE and the BS [8])

$$\begin{aligned} &\operatorname{vec}\left(\boldsymbol{e_1}^{(r)} \otimes \boldsymbol{e_2}^{(t)})\right)^H .\operatorname{vec}\left(\boldsymbol{e_1}'^{(r)} \otimes \boldsymbol{e_2}'^{(t)}\right) \\ &\quad = \delta(\theta_1-\theta_1', \phi_1-\phi_1', \theta_2-\theta_2', \phi_1-\phi_1') \end{aligned} \tag{8}$$

where vec(.) denotes an operator that concatenates columns of a matrix ($M \times N$) into a vector ($MN \times 1$), (.) denotes inner product, and θ and ϕ values denote the azimuth and the elevation angles. Thus for subcarrier l and the path associated with AOA and AOD equal to $\left(\theta_m^{(r)(k)}, \phi_m^{(r)(k)}\right)$ and $\left(\theta_m^{(t)(k)}, \phi_m^{(t)(k)}\right)$, respectively. We define $r(l)$ as

$$r(l) = \operatorname{vec}\left(\boldsymbol{e_m}^{(r)} \otimes \boldsymbol{e_m}^{(t)}\right)^H .\boldsymbol{h}[l] = \alpha_m^{(k)} e^{-j2\pi\frac{l\, n_m^{(k)}}{N_c}}\delta(0)\,. \tag{9}$$

Let us concatenate $r(l)$ for all subcarriers to get

$$\begin{aligned} \boldsymbol{r} &= \alpha_m^{(k)}\delta(0)[1 \quad e^{-j2\pi\frac{n_m^{(k)}}{N_c}} \dots e^{-j2\pi\frac{(N_c-1)\, n_m^{(k)}}{N_c}}] \\ &= \alpha_m^{(k)}\delta(0)[1 \quad e^{-j2\pi\frac{\tau_m^{(k)}}{T_s N_c}} \dots e^{-j2\pi\frac{(N_c-1)\, \tau_m^{(k)}}{T_s N_c}}], \end{aligned} \tag{10}$$

where $\boldsymbol{r} \in \mathbb{C}^{1\times N_c}$. Therefore, $\alpha_m^{(k)}$ and $n_m^{(k)}(\mod N_C)$ can be estimated by applying Fourier transform on $\boldsymbol{r}$. For now assume that we have an ideal system with infinite bandwidth ($B \to \infty$, $T_s \to 0$), infinite number of subcarriers ($N_C \to \infty$) , and limited subcarrier spacing ($\frac{1}{B_c} = T_s N_C \to \dot{\tau}_c < \infty$) where $\dot{\tau}_c$ is larger than the maximum delay of the propagation paths. Defining $\boldsymbol{t}(\tau) = \frac{1}{\sqrt{N_c}}[1 \quad e^{j2\pi\frac{\tau}{\dot{\tau}_c}} \dots e^{j2\pi\frac{(N_c-1)\,\tau}{\dot{\tau}_c}}]^T \in \mathbb{C}^{N_c\times 1}; 0 \leqslant \tau \leqslant \dot{\tau}_c$, we have

$$\begin{aligned} \boldsymbol{rt}(\tau) &= \frac{\alpha_m^{(k)}\delta(0)}{\sqrt{N_c}}\sum_{i=0}^{N_c-1} e^{-i\, j2\pi\frac{(\tau-\tau_m^{(k)})}{\dot{\tau}_c}} \\ &= \frac{\alpha_m^{(k)}\delta(0)}{\sqrt{N_c}}\frac{\sin(\pi N_c\frac{(\tau-\tau_m^{(k)})}{\dot{\tau}_c})}{\sin(\pi\frac{(\tau-\tau_m^{(k)})}{\dot{\tau}_c})} \\ &\Rightarrow \lim_{N_c\to\infty}\boldsymbol{rt}(\tau) = \alpha_m^{(k)}\delta(0)\delta(\tau-\tau_m^{(k)}) \end{aligned} \tag{11}$$

which gives us the gain and delay of the path. To summarize the whole process of transforming CSI to the paths' geometrical information in a linear transformation process, let us redefine $\boldsymbol{H}$ as

$$\dot{\boldsymbol{H}} = [\operatorname{vec}(\boldsymbol{h}[1]) \quad \operatorname{vec}(\boldsymbol{h}[2])\dots\operatorname{vec}(\boldsymbol{h}[N])\dots]\,. \tag{12}$$

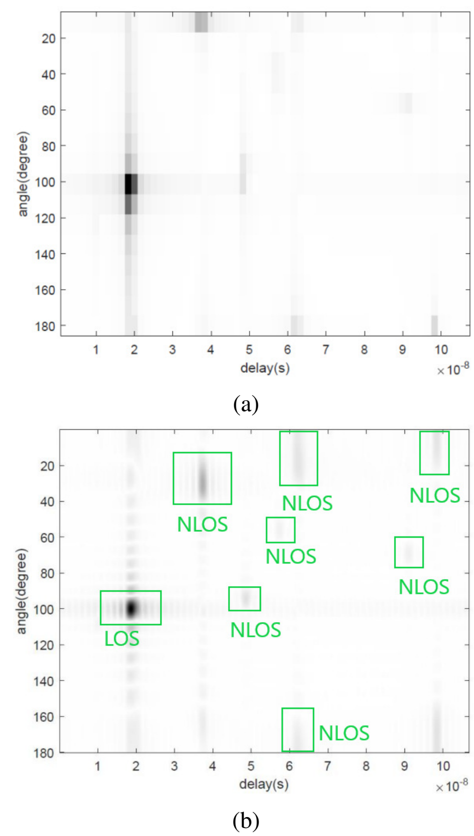


Fig. 2: (a) shows the original DADP introduced in [8] and (b) shows the ADADP introduced in this paper. The DADP is a blurred version of the ADADP. The position of path clusters in angular-delay domain are labeled as LOS and NLOS.

Next, we define the angular-delay transformation of CSI as

$$\boldsymbol{Q}\left(\theta^{(r)}, \phi^{(r)}, \theta^{(t)}, \phi^{(t)}, \tau\right) = \operatorname{vec}\left(\boldsymbol{e}^{(r)} \otimes \boldsymbol{e}^{(t)}\right)^H \dot{\boldsymbol{H}}\boldsymbol{t}(\tau)\,. \tag{13}$$

Assuming the condition of (8) and infinite bandwidth, we have

$$\begin{aligned} &\boldsymbol{Q}\left(\theta^{(r)}, \phi^{(r)}, \theta^{(t)}, \phi^{(t)}, \tau\right) \\ &= \sum_{k=1}^{C}\sum_{m=1}^{R_C}\alpha_m^{(k)}\delta\left(\Delta\theta^r, \Delta\phi^r, \Delta\theta^t, \Delta\phi^t, \Delta\tau\right). \end{aligned} \tag{14}$$

where $\Delta\theta^r = \theta^{(r)} - \theta_m^{(r)(k)}, \Delta\theta^t = \theta^{(t)} - \theta_m^{(t)(k)}, \Delta\phi^r = \phi^{(r)} - \phi_m^{(r)(k)}, \Delta\phi^t = \phi^{(t)} - \phi_m^{(t)(k)}$, and $\Delta\tau = \tau - \tau^{(t)(k)}$.

We call (14) the continuous angular-delay profile (CADP) of CSI. CADP is a linear transformation of CSI, ideally reveals all the geometrical attributes of all propagation paths with infinite resolution. The condition of (8) needs infinite number of antenna elements at both the BS and the UE. Moreover, obtaining (14) requires infinite bandwidth. Such conditions are not accessible in a practical setup. In a real-world system with finite angular resolution and limited bandwidth the transformation of (14) still holds yet the angular-delay resolution of the transformation is limited. In the next two subsections, we will discuss the angular-delay transformations for systems equipped by uniform linear array (ULA) antennas with finite number of antennas.

A. Discrete Angular-Delay Profile

Herein, we use the discrete angular-delay profile (DADP) as defined in [8] for a single-antenna UE when the BS is equipped with a ULA antenna with half-wavelength antenna spacing. A ULA with N elements has an array response

$$\boldsymbol{e}(\theta) = \frac{1}{\sqrt{N}}[1, e^{-j\pi cos(\theta)}, \ldots, e^{-(N-1)(j\pi cos(\theta))}]^T, \quad (15)$$

where $\boldsymbol{e}(\theta)$ has the characteristics of a discrete Fourier Transform (DFT) vector. The authors define the DFT matrix $\boldsymbol{V} \in \mathbb{C}^{N_B \times N_B}$ as

$$[\boldsymbol{V}]_{z,q} \triangleq \frac{1}{\sqrt{N_B}} e^{-j2\pi \frac{\left(z(q-\frac{N_B}{2})\right)}{N_B}},$$

and $\boldsymbol{F} \in \mathbb{C}^{N_c \times N_c}$ as

$$[\boldsymbol{F}]_{z,q} \triangleq \frac{1}{\sqrt{N_c}} e^{-j2\pi \frac{zq}{N_c}}.$$

Then DADP matrix $\boldsymbol{G} \in \mathbb{C}^{N_B \times N_C}$ is defined as [38]

$$\boldsymbol{G} = \boldsymbol{V}^H \boldsymbol{H} \boldsymbol{F}. \quad (16)$$

Then, authors in [8] prove that the $[\boldsymbol{G}]_{z,q}$ denote the power of path associated with the angle $\theta_q = \arccos\left(\frac{2q-N_B}{N_B}\right)$ and delay $\tau_z = zT_s$. In other words, DADP is a sampled version of CADP at angles equal to $\theta_q = \arccos\left(\frac{2q-N_B}{N_B}\right), q = 0, \ldots, N_B - 1$ and delays equal to $\tau_z = zT_s, z = 0, \ldots, N_C - 1$. Fig. 2 (a) shows a sample of DADP. As the figure shows, path clusters appear with $+$ like shape in DADP which conforms our understanding of path clusters that DMCs have slightly different angles and delays with respect to the SC. Authors in [38], discuss that due to the semantic nature of DADP and similar patterns of path clusters in DADP, the DADP can transform a localization problem to an image processing problem. Furthermore, authors in [37] demonstrate that the problem of localization in highly dynamic scenarios can be transformed to a video prediction problem using DADP time series.

B. Augmented Discrete Angular-Delay Profile

One significant issue with DADP is that, the size of DADP of (16) equals to the size of CSI matrix, which can be problematic since the number of antennas and the bandwidth of a sub-6 GHz system is much less than those of the co-located mmWave system. Since the wavelength of mmWave systems is one order of magnitude lower than that of sub-6 GHz and the available bandwidth at mmWave is orders of magnitude higher than bandwidth at sub-6 GHz, the size of CSI at sub-6 GHz is much less than the size of mmWave CSI. Thus the angular-delay information of propagation paths available via sub-6 GHz CSI tends to be less than the angular-delay information available via mmWave CSI. In Section VII, we will show this asymmetry between sub-6 GHz and mmWave CSI weakens the generalization capability of DNNs.

To tackle this issue we propose to use the augmented DADP (ADADP), which can be arbitrarily large in size. In our previous work [50], we briefly introduced the concept of ADADP and used it in the context of the fingerprinting localization problem. We define the Vandermonde matrix $\dot{\boldsymbol{V}} \in \mathbb{C}^{N_B \times N_{BB}}$ as

$$[\dot{\boldsymbol{V}}]_{z,q} \triangleq \frac{1}{\sqrt{N_B}} e^{-j\pi z \cos(\frac{q\pi}{N_{BB}})},$$

and $\dot{\boldsymbol{F}} \in \mathbb{C}^{N_c \times N_{cc}}$ as

$$[\dot{\boldsymbol{F}}]_{z,q} \triangleq \frac{1}{\sqrt{N_c}} e^{-j2\pi z \frac{q}{N_{cc}}}.$$

Then the ADADP matrix $\dot{\boldsymbol{G}} \in \mathbb{C}^{N_{BB} \times N_{CC}}$ can be defined as

$$\dot{\boldsymbol{G}} = \dot{\boldsymbol{V}}^H \boldsymbol{H} \dot{\boldsymbol{F}}, \quad (17)$$

where N_{BB} and N_{CC} can be arbitrarily large integers. ADADP is a sampled version of CADP at angles equal to $\theta_q = \frac{q\pi}{N_{BB}}, q = 0, \ldots, N_{BB} - 1$ and delays equal to $\tau_q = \frac{qN_CT_s}{N_{CC}}, q = 0, \ldots, N_{CC} - 1$. Since ADADP can be sampled arbitrarily more dense compared to DADP, it can potentially provide more collective information about the angular-delay aspects of the propagation paths and drastically improve the generalization ability of DNNs. Fig. 2 illustrates that both representations can reveal the path cluster parameters in the angular-delay domain. However, ADADP produces a clearer image compared to DADP.

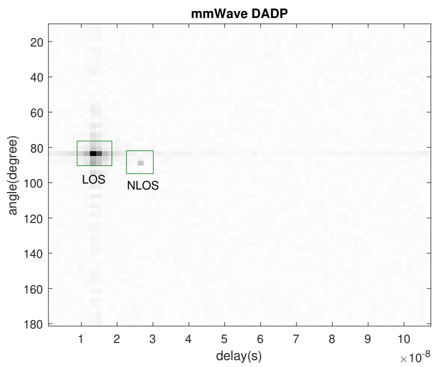
C. Sub-6 GHz and mmWave Discrete Angular-Delay Profile

The similarities between the mmWave DADP and sub-6 GHz ADADP of the co-located antennas are illustrated in Fig. 3. The CSIs used to generate the ADADPs are collected at the same time instance. While the sub-6 GHz ADADP may have more NLOS path clusters, it is clear that clusters in the mmWave DADP appear in the approximately same location as the clusters in sub-6 GHz ADADP, hence we can map the sub-6 GHz ADADP to the optimal mmWave beam using our proposed PARAMOUNT approach. Furthermore, that the strongest path cluster appears in approximately the same location in both ADADPs. Specifically, position of the strongest path cluster (LOS) in angular-delay domain shows up in the same location for sub-6 GHz and mmWave, which in this case is around 82°. There is also another NLOS path around 85° in both ADADPs.

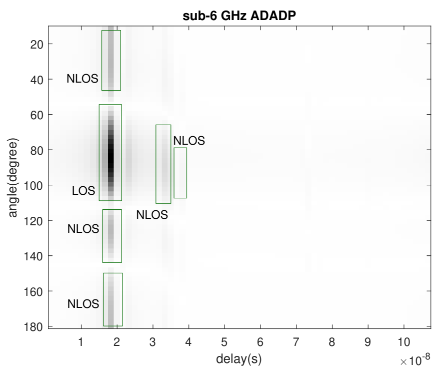
VI. DNN Structure and Generalization

Instead of inputting CSI and using a purely data-driven approach unlike many previous works [15], we propose to combine the data-driven model with the knowledge of the channel model. We attain this by converting an ambiguous representation of the channel model (raw CSI data), into an explicit representation of angle and delay distribution of the propagation paths (ADADP). In addition, the transformation of the dense CSI matrix into a sparse representation simplifies the CNN learning process by exploiting the sparsity of the ADADP. This technique has been shown to enhance the generalizability of models, as reported in [51]. Conversely, using dense matrix inputs like CSI requires a deeper NN that increases model complexity. Although it may be feasible to learn to correlate the geometric path propagation characteristics to the beam from raw CSI data, it would necessitate a more intricate model and significantly more data for training.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3324916



(a)



(b)

Fig. 3: a) LOS and NLOS path clusters in the mmWave DADP, b) LOS and NLOS path clusters in the sub-6 GHz ADADP. Data from the same user at the same time instance via the collocated mmWave and sub-6 GHz MIMO antennas.

Our proposal is to feed ADADP into a CNN, which utilizes multiple kernel filters at each layer and performs convolution on the previous layer's output. Because ADADP is a semantic and interpretable image, the CNN can learn the visual patterns of path clusters and their variations, separating the paths, and transforming them into mmWave beam space. The CNN can also learn the spatial correlations of the path clusters in the ADADP input.

In this way, by only inputting ADADP, the CNN can bring out all useful angular-delay information about path clusters from sub-6 GHz CSI. This way, even if we input a CSI outside the training dataset, the CNN will be able to separate path clusters, find the best path cluster at mmWave band for communication and output the best beam pair, hence generalize beyond the training dataset. The proposed structure for the CNN is illustrated in Fig. 4. Since the path clusters only appear with $+$-like shapes in ADADPs, we propose a CNN with only 3 convolutional layers. The proposed structure consists of an input layer, and convolutional layers followed by batch normalization [52], ReLU, and dropout. The next two convolutional layers have the same structure as the first layer and they are followed by a fully-connected layer and Softmax for classification [53]. We refer to our proposed algorithm with the proposed CNN and the ADADP input as *PARAMOUNT*. Specifically, we expect that the CNN learns the kernels matched to the path cluster patterns, thus separate them. The CNN can identify spatial correlations of the path clusters even if the patterns are shifted, scaled or distorted, while the FCNN is generally not invariant to translations or local distortions with respect to the inputs [54]. This enables the CNN to learn the visual pattern of the path clusters in the ADADP image even if they are shifted [55] and extend learned pattern features beyond the training dataset.

In [15], the authors introduce an FCNN which takes the sub-6 GHz CSI as the input and outputs the best beam with highest SNR at mmWave. Since FCNN can only take 1D inputs and CSI has a 2D structure, they first vectorize the CSI and then feed it to the FCNN. Since the FCNN is not regularized for the task at hand, we have no guarantee that the FCNN learns to extract angular-delay attributes of propagation paths, neither can generalize beyond the training dataset. Moreover, from the channel model described in Section II, we know that each path cluster has an angular-delay spread, while the FCNN structure does not have any internal mechanism to extract correlative dependencies between nearby angles and delays. Moreover, the number of trainable parameters, which includes both weights and biases, in the FCNN model far exceeds the number in the CNN model. Specifically, the FCNN model has 17,442,880 trainable parameters, while the CNN model has only trainable 340,368 parameters. Since FCNN is prone to overfitting data due to its large number of learnable parameters, it may *memorize* a dataset instead of learning the latent representation in the data, thus perform pretty well only when the input coming from the seen dataset, while it is unable to generalize beyond it.

VII. Simulation Results

In this section, we examine the generalization capability of the proposed CNN structure and compare it with the FCNN structure proposed in [15]. The common technique for training and testing a DNN is to take a given dataset measured within an environment, randomly separating the dataset to a test and a train dataset and then examine the performance of the trained DNN on the test dataset. In addition to that, to examine the generalization capability of the DNNs, we test the performance of the techniques in scenarios where the test dataset is coming from an unseen scenario.

A. Outdoor scenario

TABLE I: DeepMIMO Outdoor Dataset Parameters

Parameters	Sub-6 GHz Dataset	mmWave Dataset
Scenario name	O1_3p5	O1_28
Antenna Type	ULA	ULA
Number of BS Antennas	4	64
Antenna Spacing (wavelength)	0.5	0.5
Bandwidth(GHz)	0.08	0.5
Number of OFDM subcarriers	64	512
OFDM sampling factor	1	1
OFDM limit	64	64
Number of paths	15	5

Unseen Test Scenarios: For the sake of the testing, we use the traditional testing/training dataset separation which we call the seen test dataset. Moreover, we examine the performance of the trained DNN on two categories of unseen datasets:

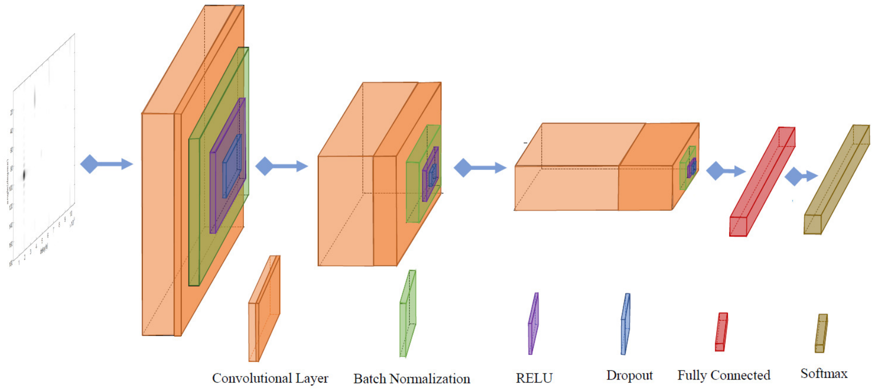


Fig. 4: The proposed CNN structure for extracting path clusters' information from ADADP.

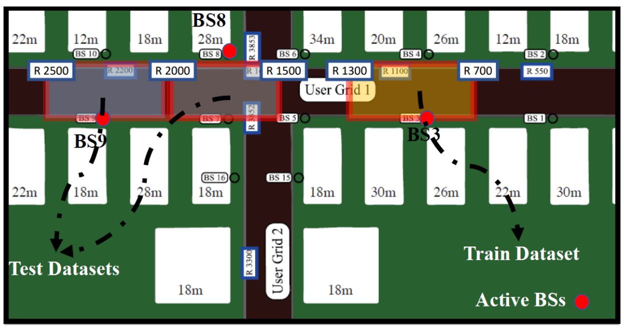


Fig. 5: Bird's-eye view of the DeepMIMO outdoor scenario. White rectangles illustrate buildings and black ribbons are streets. The training dataset is generated for the user grid around the active BS3. The LOS and NLOS blockage scenarios are generated for the same user grid. Two other test datasets are generated assuming BS8 and BS9 are active, respectively. The red bounding boxes show the user grid for each dataset.

- **Path Blockage:** in a real world setup, due to movement of the objects within an environment, it is very probable that some of the propagation paths get blocked. In path blockage scenarios, we assume that the BS3 is active and while propagation paths in the test dataset are subject to one of the following changes:
  - **LOS Blockage:** In this scenario, we assume that the strongest path cluster between the BS and the UE gets blocked for all the CSIs of the dataset.
  - **NLOS Blockage:** We assume that the second strongest path cluster between the BS and the UE gets blocked for all the CSIs of the dataset.
- **Unseen Base Stations:** As we explained in Section VII-A, for generating the train dataset we assume that BS3 is active. For testing we assume two additional unseen LOS scenarios:
  - **BS8:** We assume the BS8 is active and we generate the test detaset for rows R1500 to R2000. All other parameters from the training scenario remain the same.
  - **BS9:** We assume the BS9 is active and we generate the test dataset for rows R2000 to R2500. All other parameters from the training scenario remain the same.

Dataset Generation Using DeepMIMO [56]: For the sake of training and testing, we use the DeepMIMO framework for outdoor scenarios which simulates wireless channel within a dense urban scenario of an intersection of two streets using ray-tracing. Fig. 5 illustrates the bird's-eye view map of the whole scenario. There are 18 different BSs around the environment. For training, we assume that only BS3 is active. We generate a dataset of CSIs tagged by locations for the grid of users from row R700 to R1300 (collectively 108,600 datapoints). We assume that there are two co-located antennas at BS3, one operating at 3.5 GHz (sub-6 GHz band) and the other one at 28 GHz (mmWave band). The parameters for the sub-6 GHz and mmWave systems can be found in TABLE I. We assume that at both bands, the UE is equipped with one antenna. The DeepMIMO dataset provides us with the CSI at both bands. To find the best beam at mmWave band, we form a codebook at mmWave band consisting of N_U^{mm} precoding vectors. The i^{th} codeword is defined as $\boldsymbol{w}^{(i)} = \boldsymbol{e}(\frac{\pi i}{N_U^{mm}}), i = 0, \ldots, N_U^{mm} - 1$, where $\boldsymbol{e}$ is the ULA array response of (15). Using the mmWave CSI and the codebook, a DNN classifier finds the best beam class with highest rate. We tag each sub-6 GHz CSI with the mmWave beam class to form the dataset. To reproduce the scenario and benchmark the output, the training dataset and the mmWave dataset is exactly the same as the training dataset used in [15].

B. Performance Evaluation Metrics

For the purpose of comparison and benchmarking, we pick the *top-n* metric which is defined as the probability of the best beam being among the top-n beams suggested by the classifier. We also calculate the achievable rate of the mmWave communication system based on top-n beams. The top-n achievable rate is defined as the rate achieved using the best performing beam in the top-n classifier output.

C. Neural Network Training and Testing

For DNNs training, we split the training dataset to 80 percent training and 20 percent testing. For comparison, we train the FCNN network and the CNN network on the same training dataset. We train the CNN both on DADP (denoted by DADP + CNN) and ADADP (denoted by PARAMOUNT). To generate ADADPs, we set $N_{BB} = N_{CC} = 64$. The FCNN training hyperparameters are equal to the parameters introduced in [15]. For CNN training we use ADAM solver

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3324916

[57], and learning rate equals to 10^{-4}, other hyperparameters are the same as the ones for FCNN's. For testing, we randomly pick 40,000 datapoints from the mentioned unseen training datasets and average the results on them. All the simulations are conducted in MATLAB and running on a machine equipped with an RTX 3060 GPU and 64 GB of RAM. The codes for the simulations can be accessed in [58].

Noisy Channel Measurements: In all experiments, for testing, we assume noisy CSI measurements at sub-6 GHz. We assume the SNR changes from -5 dB to 20 dB. For calculating the achievable rate at mmWave, we assume the same SNR at mmWave as the sub-6 GHz band.

For training, in [15], the authors retrain the FCNN for each SNR level they use for experimentation at the testing dataset. This retraining approach is not practical. We do not have the luxury of network retraining for each SNR level in a real-world application. Instead of that, we add a noise with random SNR level between -5 dB and 20 dB to each training sample and train the DNNs (both the CNN and FCNN) once on the noisy dataset. In practice, we can make such a training dataset by measuring CSI at a high SNR level only once and then randomly adding noise to the measurements.

D. Beam Selection Generalization Performance Evaluation

1) Generalization Performance for Path Blockage Datasets: In the first simulation, we compare the performance of CSI + FCNN proposed in [15], DADP + CNN, and ADADP + PEAK SEARCH to the proposed PARAMOUNT (ADADP + CNN). Since the novelty of PARAMOUNT consists of two components, the CNN architecture and the ADADP formulation, we demonstrate the performance of models with only one of these novelties. The DADP + CNN model only considers the architecture novelty, while the ADADP + PEAK SEARCH focuses on the ADADP formulation. Furthermore, the DNNs (FCNN, DADP + CNN, and PARAMOUNT) are trained on dataset generated from the parameters in TABLE I. The DNNs are trained on LOS scenario datasets, and tested on both seen LOS scenario, and unseen LOS blockage and NLOS blockage scenarios.

TABLE II shows the performance of all the discussed methods in the 3 mentioned scenarios. As the results demonstrate, when the test scenario and the train scenario are the same (i.e. LOS Scenario), FCNN performs with higher top-1 and top-3 accuracy than the two other DNNs for SNRs greater than 0 dB. While for SNRs equals or less than 0 dB, PARAMOUNT shows the best performance among the DNNs. On the other hand, in the two unseen scenarios, LOS and NLOS blockage scenarios, the accuracy of FCNN drops sharply while PARAMOUNT shows a robust performance. In particular, PARAMOUNT overall performs 20% better than the CSI + FCNN in top-1 accuracy and 30-40% better in top-3 accuracy. The outstanding performance of CSI + FCNN in the seen scenario and its poor performance in the unseen scenarios, show that the FCNN has overfitted to the seen scenario and fails to generalize to the unseen scenarios. The FCNN appears to memorize the CSI to beam relationship without taking into account any knowledge of the wireless channel. Therefore, when the CSI input is altered by a LOS or NLOS blockage, the FCNN cannot match it to the correct beam. On the other hand, PARAMOUNT shows a very robust performance in all scenarios specifically when we consider the top-3 metric. The results prove that, as expected, the FCNN falls short of extracting propagation paths information from CSI while PARAMOUNT can extract the information from the ADADP.

Please note that the ADADP + PEAK SEARCH approach shows a slightly better performance in NLOS blockage scenario because this scenario is specifically in favor of the ADADP + PEAK SEARCH approach. This scenario refers to the case were the second strongest path cluster is blocked in mmWave. The second strongest path is the most problematic for the ADADP + PEAK SEARCH method as it can be mistakenly taken by ADADP + PEAK SEARCH as the one leading to the best mmWave beam. When the second strongest path is blocked, there is no such ambiguity for peak selection, hence leading to an exceptionally good performance for ADADP + PEAK SEARCH. The NLOS blockage is actually a very special case of the LOS scenario with a small probability of occurrence which is included to show robustness of PARAMOUNT in various scenarios. TABLE II shows that even in this exceptional scenario that is very much in favor of ADADP + PEAK SEARCH, our proposed PARAMOUNT approach outperforms it by 10% in top-1 accuracy. Since our main goal is to avoid any beam search at all, we care mostly about the top-1 performance. In other words, it is our significant improvement in the top-1 accuracy that enables mmWave connection establishment without any beam search at all. Using the values in TABLE II, an estimate of the top-1 accuracy of ADADP + PEAK SEARCH over PARAMOUNT for various SNR values, can be calculated $\frac{30+36+37+38+37+37}{19+39+61+72+76+76} = 62.7\%$ and $\frac{25+29+30+30+30+30}{18+38+49+51+51+53} = 66.9\%$ for the LOS and NLOS scenarios, respectively. Considering the overall accuracy, we get a significant gain by PARAMOUNT in comparison with a simple ADADP + PEAK SEARCH.

2) Ablation Study: The innovations in PARAMOUNT are two fold: 1) our proposed ADADP transform that not only provides good generalization performance, but also partially boosts the beam selection accuracy, and 2) our proposed efficient CNN-based classifier applied on top of the ADADP to further improve the beam selection accuracy. We use the results in TABLE II to assess the relative effectiveness of each individual innovation in PARAMOUNT. A performance comparison between DADP + CNN with PARAMOUNT (ADADP + CNN) allows us to asses the relative effectiveness of CNN. Comparing DADP + CNN with PARAMOUNT, we observe that PARAMOUNT consistently exceeds the performance in all instances. In top-1 for SNR greater than 0 dB, the PARAMOUNT outperforms DADP + CNN by 20% or more. This confirms that our proposed ADADP is partially boosting the beam selection accuracy, i.e. the CNN architecture alone does not achieve a high accuracy. Next, let us compare PARAMOUNT (ADADP + CNN) with ADADP + PEAK SEARCH to get insights on the performance of the CNN. As mentioned in previously, when averaged over various

TABLE II: Beam selection top-1 and 3 accuracy metrics for one seen and two unseen scenarios. The training LOS scenario is the LOS dataset around BS3 parameterized in Fig. 5. In the two unseen test scenarios, the LOS and the NLOS blockage scenarios, the user grid is the same as LOS scenario, however, we assume that the strongest path cluster and the second strongest path cluster gets blocked, respectively. To compare the generalization performance of the DNNs, we list the accuracy of DNNs for the scenario where the test dataset is seen (LOS scenario), as well as the two unseen scenarios.

		Top-1 Accuracy						Top-3 Accuracy					
	SNR(dB)	-5	0	5	10	15	20	-5	0	5	10	15	20
LOS Scenario	CSI + FCNN	11%	38%	**66%**	**81%**	**87%**	**89%**	30%	72%	**94%**	**99%**	**99%**	**99%**
	DADP + CNN	6%	17%	38%	52%	55%	55%	14%	44%	80%	90%	91%	91%
	ADADP + PEAK SEARCH	**30%**	36%	37%	38%	37%	37%	**71%**	**82%**	87%	89%	90%	90%
	PARAMOUNT (ADADP + CNN)	19%	**39%**	61%	72%	76%	76%	47%	79%	93%	96%	97%	97%
NLOS Blockage	CSI + FCNN	10%	24%	33%	34%	35%	35%	31%	59%	65%	64%	63%	63%
	DADP + CNN	5%	12%	22%	27%	29%	29%	14%	31%	52%	66%	70%	71%
	ADADP + PEAK SEARCH	**32%**	40%	43%	43%	43%	43%	**77%**	**89%**	**95%**	**98%**	**98%**	**99%**
	PARAMOUNT (ADADP + CNN)	17%	**41%**	**50%**	**50%**	**52%**	**54%**	44%	82%	92%	93%	93%	94%
LOS Blockage	CSI + FCNN	7%	16%	21%	22%	22%	23%	21%	41%	48%	49%	49%	48%
	DADP + CNN	5%	10%	15%	17%	19%	19%	12%	26%	42%	52%	56%	57%
	ADADP + PEAK SEARCH	**25%**	29%	30%	30%	30%	30%	**64%**	73%	78%	80%	81%	81%
	PARAMOUNT (ADADP + CNN)	18%	**38%**	**49%**	**51%**	**51%**	**53%**	43%	**74%**	**84%**	**86%**	**87%**	**87%**

SNR values, ADADP + PEAK SEARCH achieves only 62.7% and 66.9% top-1 accuracy of PARAMOUNT for LOS and NLOS, respectively. Thereby, it is obvious that proper use of CNN as proposed in PARAMOUNT significantly improves the top-1 beam selection accuracy which is considerably desirable as it fully omits the beam search overheads. Furthermore, ADADP + PEAK SEARCH method identifies the sub-6 GHz ADADP index corresponding to the strongest path's direction. This is achieved by upsampling the ADADP to 64 in the angular domain, where each sample corresponds to a beam index. As per Section II, the propagation paths at sub-6 GHz and mmWave bands share similar geometry. Therefore, it is possible to predict the mmWave band's beam using ADADP + PEAK SEARCH on the sub-6 GHz ADADP. The advantage of ADADP + PEAK SEARCH is its low complexity since it does not require any DNN training. For top-1 accuracy, PARAMOUNT outperforms ADADP + PEAK SEARCH anywhere between 7% and 23% in blockage scenarios and more than 20% on in LOS scenario when SNR is greater than 0 dB. For top-3 accuracy, PARAMOUNT outperforms ADADP + PEAK SEARCH by 6% or more in LOS blockage scenarios and more than 9% on in LOS scenario.

Although using DNN-based beam selection requires a training phase, it alleviates the need to send pilot signals on different beams to find the best (i.e. beam sweeping). With DNNs, we spend some computation resources (mostly for to training) to save the communication resources spent on beam search. Considering the technology trend which is quickly moving towards considerably cheaper computation and extremely expensive communication resources, NN-based beam selection seems to be definitely part of future wireless. It is already under study by 3GPP RAN1 for Release 18, 5G Advanced: AI/ML for NR Air Interface [59]. Additionally, collection of the dataset for training DNNs in a realistic scenario is not expensive either. Presently, these datasets are readily accessible at the BS within the current wireless networks. This is because the sub-6 GHz channel is estimated regularly in the current wireless protocols using periodic pilot transmissions to enable efficient communications. Similarly, the mmWave beam selection is also routinely performed using the conventional methods (beam sweeping, multi-resolution beam search, etc.) in the current wireless protocols. Therefore, by simply recording these sub-6 GHz channel estimates (i.e. input features to PARAMOUNT) and the corresponding optimal mmWave beams (i.e. output labels) during regular network operations, the training dataset is automatically generated given a specific period of time has passed since the BS was powered on. Consequently, the main idea in PARAMOUNT is to exploit the existing wealth of sub-6 GHz CSI and the corresponding optimal beam estimates to train a CNN that can later omit/reduce the mmWave beam search overheads.

One may also notice the poor performance of the ADADP + PEAK SEARCH in LOS scenario. This is due to the small number of sub-6 GHz antennas and large upsampling factor in the ADADP formulation. In Section II, we demonstrate that the strongest path cluster in the sub-6 GHz channel tends to be the strongest cluster in the mmWave channel. However, the number of sub-6 GHz antenna array elements (N_B) is significantly less than the number of array elements on the mmWave antenna. The resolution of the CSI matrix and consequently the DADP matrix is proportional to N_B as is evident from the formulation in (16). The sub-6 GHz DADP can be enlarged into a ADADP matrix using (17) to match the size of the mmWave DADP. The discrepancy between the peak in the sub-6 GHz ADADP and the mmWave DADP results from the upsampling error in the ADADP. The larger the upsampling factor, the lower the performance of ADADP + PEAK SEARCH. In case of the results in Table II, the sub-6 GHz antenna has only 4 elements while the mmWave antenna has 64. Therefore, the ADADP is enlarged 16 times which impacts the accuracy. We get a low top-1 accuracy of 37% with ADADP + PEAK SEARCH in the LOS case, and the accuracy further drops to 30% in the NLOS scenario. We would like to add that the role of the CNN in the proposed PARAMOUNT approach is enhancement of the ADADP resolution to counter the adverse effects of upsampling. One motivation of using CNNs in PARAMOUNT was their demonstrated success in image super-resolution [60]. In PARAMOUNT, the CNN extracts the local correlation patterns in ADADP and uses information from the entire path cluster to make the beam prediction, rather than focusing on a single peak. The subtle non-linear correlations between the path clusters and optimal beams is best captured using a CNN model. This ablation study demonstrates that using ADADP is a necessary but

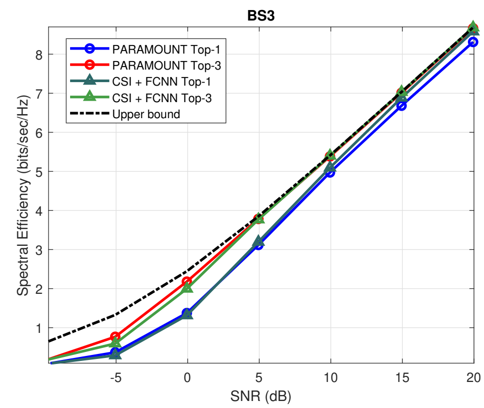
This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3324916

not sufficient condition for achieving the best beam selection performance. We show that by combining the ADADP with the CNN network in PARAMOUNT, we can achieve superior results, especially in the top-1 accuracy.

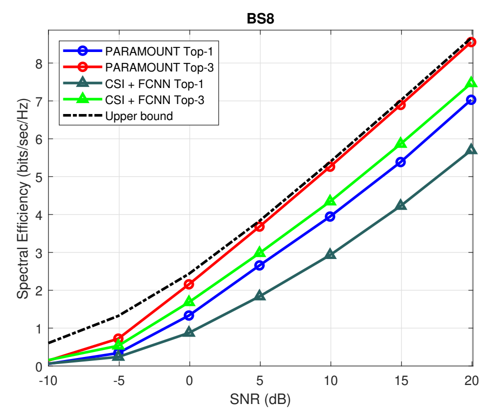
Additionally, the ADADP + PEAK SEARCH reaches its best performance for SNR greater than 0 dB and further increasing SNR does improve the performance as shown in Table II. As long as the signal power is higher than the noise level (positive SNR), the peak (maximum) in the ADADP will always be located on the same ADADP matrix element. Hence, further increasing SNR will not improve the ADADP + PEAK SEARCH performance.

3) Generalization Performance For Unseen LOS Datasets: In the next simulation, we test the trained DNNs in unseen LOS datasets of BS8 and BS9 as described in Section VII-A. Here the test and train datasets are both LOS, but the test datasets originate from totally different scenarios. Fig. 6 shows the achievable spectral efficiency (SE) using PARAMOUNT and CSI + FCNN in three scenarios. SE and beam alignment are closely related and interdependent. The purpose of beam alignment is to maximize the SE. By leveraging beamforming techniques and precise beam alignment, mmWave MIMO systems can achieve higher data rates and improved overall performance. Fig. 6(a) shows that both PARAMOUNT and CSI + FCNN top-3 rate approach the upper bound for SNR higher than 5 dB, while the CSI + FCNN top-1 rate shows a slightly higher performance in comparison with PARAMOUNT. The upper bound is obtained by an exhaustive search for optimal solution of (7). These outcomes are in agreement with the previous simulation results, in which we concluded that FCNN overfitted to the train dataset. On the other hand, when we look at the performance of the CSI + FCNN at the two unseen datasets (Fig. 6 (b) and (c)), we observe a significant performance drop in CSI + FCNN output. Particularly, in both cases, the CSI + FCNN top-3 rate can achieve the level of 85% of the upper bound, and the top-1 rate declined to 65% of the upper bound of the achievable rate, while in the seen scenario it comes very close to the achievable rate upper bound. On the other hand, PARAMOUNT shows a much more robust performance in both unseen scenarios and the top-3 rate can achieve more that 99% of the upper bound. The results show that in the case of the unseen LOS scenarios, CSI + FCNN fails to generalize properly to unknown inputs, while PARAMOUNT shows a strong generalization capability.

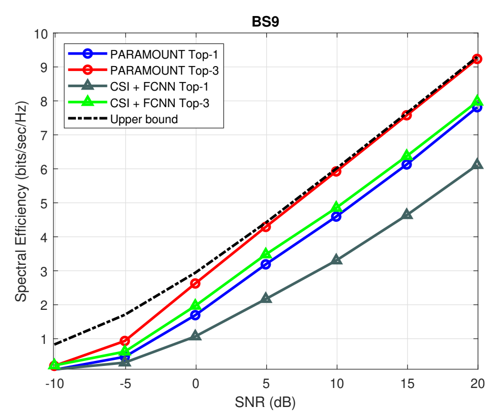
4) Effect of Training Dataset Size: In the next simulation, we explore the effect that the training dataset size has on the generalization performance of the DNNs. In this simulation, we assume that the dataset is the LOS blockage dataset (where both models have the poorest performance among all the test datasets). On the other hand, we vary the training dataset size as summarized in TABLE III. Referring to Fig. 7, the top-1 and top-3 accuracy of PARAMOUNT increase by 20% and 16%, respectively, when the training dataset size increases from 0.8 to 1.4 of the original dataset. The top-1 and top-3 accuracy of CSI + FCNN increase only 6% and 7%, respectively which shows that CSI + FCNN fails to take advantage of the increased size to improve its generalization capabilities. On the other hand, PARAMOUNT generalization performance



(a)



(b)



(c)

Fig. 6: The achievable rates using CSI + FCNN and PARAMOUNT are studied at three scenarios: (a) the test and train datasets come from the same environment around the active BS3. (b) the test scenario is different than the train scenario; for the test dataset we assume BS8 is active and the data is gathered around it (c) we assume BS9 is active and the test data is gathered around it.

significantly improves as the training dataset size increases.

5) Effect of the Number of Sub-6 GHz Antennas: In the next simulation, we study the effect of sub-6 GHz antenna array size on the generalization ability of the DNNs. So far, we have assumed that the number of antenna elements at the sub-6 GHz band is 4. In the following simulation, we change the number of antennas at sub-6 GHz to 8, 16, and 32. This way, the measured sub-6 GHz CSI provides us with more accurate information about the angles of path clusters as we

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3324916

TABLE III: To make different training datasets, we only change the user grid rows and keep all other parameters defined TABLE I the same.

Change Dataset Size	-20%	-10%	0%	10%	20%	30%	40%
User Grid Start Row	760	730	700	670	640	610	580
User Grid End Row	1240	1270	1300	1330	1360	1390	1420

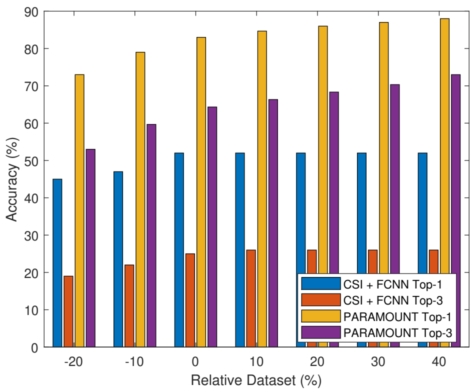


Fig. 7: shows the effect of changing the size of the training dataset on the classification accuracy of PARAMOUNT and CSI + FCNN. Here, the testing dataset is the LOS blocked dataset. Increasing the size of training dataset shows more impact on the generalization performance of PARAMOUNT in comparison with CSI + FCNN.

discussed in Section V. Based on the previous simulations, both DNNs have their poorest performance when the test dataset is the LOS blocked dataset. Therefore, we take this dataset for testing. In this simulation, we assume that the parameters described in Section VII-A remain the same except for the number of antennas at sub-6 GHz.

Fig. 8(a) shows the top-1 accuracy of PARAMOUNT and CSI + FCNN for different antenna sizes. As expected, ADADP shows a solid superior performance compared to CSI + FCNN. Furthermore, a significant performance improvement is observed when the number of antennas increases from 8 to 16 using the CNN technique. On the other hand, when the number of antennas increases from 16 to 32, we only observe performance improvements for SNRs less than 5 dB. This means that the antenna with 32 elements only provide the CNN with more SNR gain in comparison with 16 elements antenna, while the angular information provided by the latter is sufficient to separate the path clusters in the angular-delay domain. Thus, we can conclude that increasing the number of sub-6 GHz antennas from 16 to 32 is only required in low SNRs scenarios, while it does not provide us with any additional angular information about the propagation paths.

Looking at Fig. 8(b), we observe that with the 3 antenna configurations, PARAMOUNT shows almost similar performance for SNR larger than -5 dB. This observation has profound implications in showing the effectiveness of data augmentation for improving the performance of the DNN technique. As we discussed in Section V, ADADP can fundamentally improve the angular resolution of the antenna, while it provides us with more collective information of paths clusters angular-delay attributes. The results of Fig.8(b) show that the information provided by ADADP with only 8 antennas is enough to find the top 3 beam candidates and increasing the number of antennas will not significantly

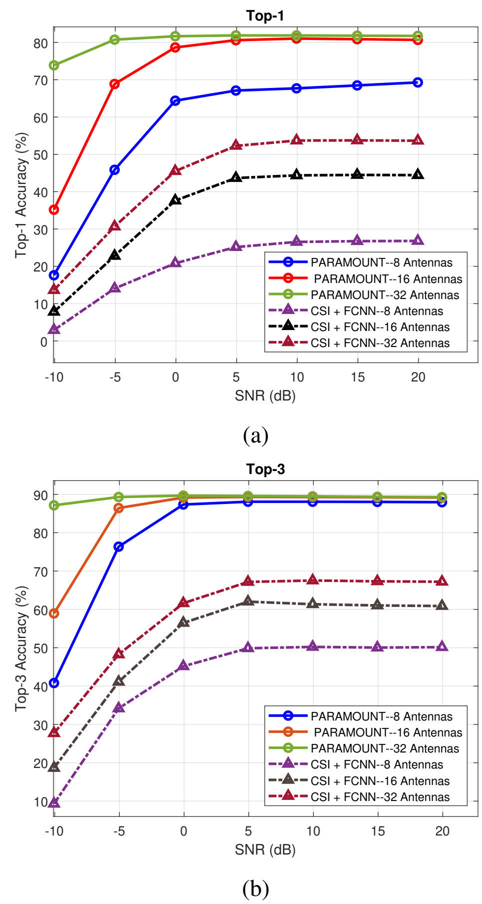
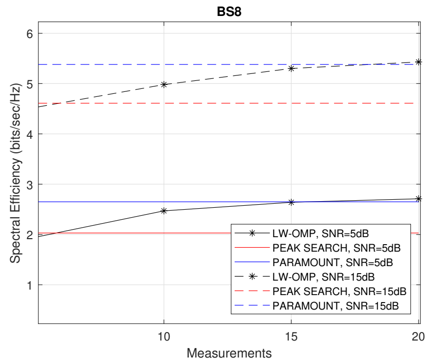


Fig. 8: In this experiment the testing dataset in the LOS blocked dataset. (a) shows the top-1 accuracy for different sub-6 GHz antenna size for PARAMOUNT and CSI + FCNN. (b) shows the top-3 accuracy.

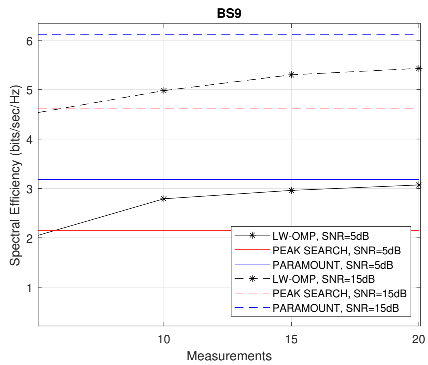
improve the performance of the CNN. According to the previous simulations, similarly we observe a pretty solid generalization performance when we consider the top-3 beams, which demonstrate the strong impact of data augmentation for the system performance.

E. Comparison with CS-based Beam Selection

In this simulation, we compare the performance of PARAMOUNT to the widely-adopted CS-based benchmark that uses a modified OMP algorithm called logit weighted - OMP (LW-OMP) [12] to predict the beam. We perform the simulations on the dataset from BS8 and BS9 for two different SNRs, namely 5 dB and 15 dB as shown in Fig. 9. To highlight the importance of our propose method, both PARAMOUNT and ADADP + PEAK SEARCH do not require any measurement; therefore, the results in Fig. 9 are for zero measurements. We spend some computation resources (mostly for training) to save the communication resources spent on beam search. From BS8 results in Fig. 9(a), we observe that LW-OMP requires more than 5 measurements to exceed the SE of ADADP + PEAK SEARCH and more than 15 measurements to exceed PARAMOUNT for both SNRs. In Fig. 9(b), we observe that LW-OMP requires more than 5 measurements to exceed the SE of ADADP + PEAK SEARCH on the BS9 dataset for both SNRs. Furthermore, we observe



(a)



(b)

Fig. 9: (a) Comparison between PARAMOUNT, LW-OMP, and ADADP + PEAK SEARCH for two different SNRs on BS8 dataset. (b) Comparison between PARAMOUNT, LW-OMP, and ADADP + PEAK SEARCH for two different SNRs on BS9 dataset.

that even with 20 measurements LW-OMP cannot exceed the SE of PARAMOUNT for neither SNRs.

F. Training Time

All training and experiments were performed using MATLAB. The simulations were carried out on a workstation equipped with 64 GB RAM, AMD Ryzen 9 5950X 16-cores and 32 logical processors CPU, and a NVIDIA GeForce RTX 3060 TI GPU. Fig. 10 shows the training duration of the CNN models for different input dimensions. The first four bars show the training time for the DADP + CNN with 4, 8, 16, and 32 sub-6 GHz antennas corresponding to DADP inputs of 4×64, 8×64, 16×64, and 32×64, respectively. The last bar shows the ADADP of size 64×64. The 4×64 DADP takes about 412 seconds to train while the 64x64 ADADP takes 1373 seconds which is approximately three times longer. However, since all training is performed offline, does not require frequent re-training, and is sufficiently generalizable, the training time is less of a concern.

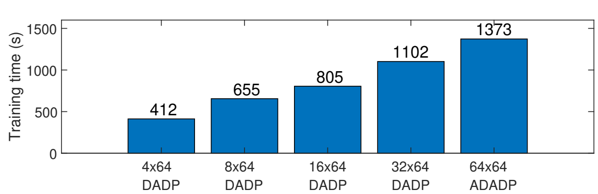


Fig. 10: The effect of enlarging the size of the input matrix with respect the training time.

G. Extension to Mobility Scenarios

The mobility scenario can be considered an extension of the dynamic scenarios simulated in Section VII-D as LOS and NLOS blockage scenarios. However, instead of predicting the current beam, the model tries to predict the future beams for a moving UE. The mobility scenarios are particularly important for vehicular communication networks (VCNs). Recurrent neural networks (RNNs) such as gate recurrent units (GRU) [61] and long-term short-term memory (LSTM) [62], [63] have shown promising results in beam prediction and beam tracking in VCNs. Specifically, this work can be extended to a mobility scenario by adding a RNN model that first predicts the sub-6 GHz ADADP at the next time stamp similar to the RNN model proposed in [37]. Then, the predicted sub-6 GHz ADADP can be used with the proposed PARAMOUNT framework to predict the mmWave beam. This implementation may be explored in future work. Additionally, an interesting extension to the mobility scenario is to transform the channel to a delay-Doppler-angle domain, following a similar approach to the DADP transformation.

VIII. Conclusion

In this paper, we first discuss the fundamental problem of generalization in using deep learning (DL) for wireless communication applications. We discuss that a solid history of well-developed model-based literature in wireless communication has provided us with exceptional performance in practical applications so far. On the other hand, the novel DL paradigms have been proved to be highly data-dependent and thus fall short of generalization beyond the training dataset. To make this matter worse, we discussed that very little effort has been devoted to study the generalization performance of DL techniques in wireless communication. We discussed that to be able to come up with generalization in DL, we need to incorporate the solid background knowledge in wireless communication with neural networks (NN). To this end, we took the problem of mmWave beam selection using sub-6 GHz CSI and discussed in details the physical aspects of electromagnetic wave scattering at mmWave and sub-6 GHz bands. From there, we have discussed that, to provide generalization, the NN needs to extract path clusters information from sub-6 GHz CSI and transform this information to mmWave band. To make sure that the NN can extract the required data, we introduce a novel augmented discrete angular-delay profile (ADADP) technique which provides us with a high resolution semantic image of the path clusters in angular delay domain. Due to the visual nature of the ADADP, we introduce a convolutional NN (CNN) for beam selection. To study the generalization capability, we train the CNN on a line-of-sight (LOS) dataset and test it on unseen datasets. Our simulation results show that the introduced data augmentation and the NN can secure generalization ability for the technique. We showed improved generalization capability for our proposed approach in comparison with previously proposed CSI + FCNN model.

References

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE communications magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[3] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE access*, vol. 1, pp. 335–349, 2013.

[4] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE access*, vol. 7, pp. 78 729–78 757, 2019.

[5] T. Nitsche, C. Cordeiro, A. B. Flores, E. W. Knightly, E. Perahia, and J. C. Widmer, "IEEE 802.11 ad: directional 60 GHz communication for multi-gigabit-per-second wi-fi," *IEEE Communications Magazine*, vol. 52, no. 12, pp. 132–141, 2014.

[6] M. Giordani, M. Mezzavilla, and M. Zorzi, "Initial access in 5G mmWave cellular networks," *IEEE communications Magazine*, vol. 54, no. 11, pp. 40–47, 2016.

[7] A. N. Uwaechia and N. M. Mahyuddin, "A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges," *IEEE Access*, vol. 8, pp. 62 367–62 414, 2020.

[8] X. Sun, X. Gao, G. Y. Li, and W. Han, "Single-site localization based on a new type of fingerprint for massive MIMO-OFDM systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6134–6145, 2018.

[9] A. Alkhateeb, G. Leus, and R. W. Heath, "Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?" in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2909–2913.

[10] C.-R. Tsai and A.-Y. Wu, "Structured random compressed channel sensing for millimeter-wave large-scale antenna systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 19, pp. 5096–5110, 2018.

[11] X. Song, S. Haghighatshoar, and G. Caire, "A scalable and statistically robust beam alignment technique for millimeter-wave systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4792–4805, 2018.

[12] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1038–1052, 2018.

[13] A. Ali, N. González Prelcic, and R. W. Heath, "Spatial covariance estimation for millimeter wave hybrid systems using out-of-band information," *IEEE Transactions on Wireless Communications*, vol. 18, no. 12, pp. 5471–5485, 2019.

[14] Y. Long, Z. Chen, J. Fang, and C. Tellambura, "Data-driven-based analog beam selection for hybrid beamforming under mm-wave channels," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 2, pp. 340–352, 2018.

[15] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5504–5518, 2020.

[16] X. Li and A. Alkhateeb, "Deep learning for direct hybrid precoding in millimeter wave massive MIMO systems," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 800–805.

[17] M. B. Mashhadi, Q. Yang, and D. Gündüz, "Distributed deep convolutional compression for massive MIMO CSI feedback," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2621–2633, 2021.

[18] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.

[19] W. Xu, F. Gao, S. Jin, and A. Alkhateeb, "3D scene-based beam selection for mmWave communications," *IEEE Wireless Communications Letters*, vol. 9, no. 11, pp. 1850–1854, 2020.

[20] U. Demirhan and A. Alkhateeb, "Radar aided 6G beam prediction: Deep learning algorithms and real-world demonstration," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 2655–2660.

[21] A. Klautau, N. González-Prelcic, and R. W. Heath, "LIDAR data for deep learning-based mmWave beam-selection," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 909–912, 2019.

[22] M. B. Mashhadi, M. Jankowski, T.-Y. Tung, S. Kobus, and D. Gündüz, "Federated mmWave beam selection utilizing LIDAR data," *IEEE Wireless Communications Letters*, vol. 10, no. 10, pp. 2269–2273, 2021.

[23] M. Zecchin, M. B. Mashhadi, M. Jankowski, D. Gündüz, M. Kountouris, and D. Gesbert, "LIDAR and position-aided mmWave beam selection with non-local CNNs and curriculum training," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 2979–2990, 2022.

[24] S. Jiang, G. Charan, and A. Alkhateeb, "LIDAR aided future beam prediction in real-world millimeter wave V2I communications," *IEEE Wireless Communications Letters*, vol. 12, no. 2, pp. 212–216, 2023.

[25] J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Position aided beam prediction in the real world: How useful GPS locations actually are?" *arXiv preprint arXiv:2205.09054*, 2022.

[26] G. Charan, T. Osman, A. Hredzak, N. Thawdar, and A. Alkhateeb, "Vision-position multi-modal beam prediction using real millimeter wave datasets," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 2727–2731.

[27] mmMAGIC, "Measurement campaigns and initial channel models for preferred suitable frequency ranges h2020-ict-671650-mmmagic/d2. 1 v1. 0," 2016.

[28] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501–513, 2016.

[29] A. Ali, N. González-Prelcic, and R. W. Heath, "Estimating millimeter wave channels using out-of-band measurements," in *2016 Information Theory and Applications Workshop (ITA)*, 2016, pp. 1–6.

[30] M. S. Sim, Y.-G. Lim, S. H. Park, L. Dai, and C.-B. Chae, "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. 51 634–51 646, 2020.

[31] H. He, S. Jin, C.-K. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 77–83, 2019.

[32] S. Rezaie, A. Amiri, E. de Carvalho, and C. N. Manchón, "Deep transfer learning for location-aware millimeter wave beam selection," *IEEE Communications Letters*, vol. 25, no. 9, pp. 2963–2967, 2021.

[33] C. Liu, Z. Wei, D. W. K. Ng, J. Yuan, and Y.-C. Liang, "Deep transfer learning for signal detection in ambient backscatter communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1624–1638, 2021.

[34] R. Rai and C. K. Sahu, "Driven by data or derived through physics? a review of hybrid physics guided machine learning techniques with cyber-physical system (CPS) focus," *IEEE Access*, vol. 8, pp. 71 050–71 073, 2020.

[35] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, F. Quitin, F. Tufvesson, and P. De Doncker, "The COST 2100 MIMO channel model," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92–99, 2012.

[36] M. Zarifneshat and L. Xiao, "Out-of-band multiple path discovery protocol for robust in-band millimeter wave links," in *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2020, pp. 383–391.

[37] F. Hejazi, K. Vuckovic, and N. Rahnavard, "DyLoc: Dynamic localization for massive MIMO using predictive recurrent neural networks," in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 2021, pp. 1–9.

[38] X. Sun, C. Wu, X. Gao, and G. Y. Li, "Fingerprint-based localization for massive MIMO-OFDM system with deep convolutional neural networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 10 846–10 857, 2019.

[39] A. Alkhateeb and R. W. Heath, "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 1801–1818, 2016.

[40] N. González-Prelcic, A. Ali, V. Va, and R. W. Heath, "Millimeter-wave communication with out-of-band information," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 140–146, 2017.

[41] M. B. Mashhadi and D. Gündüz, "Pruning the pilots: Deep learning-based pilot design and channel estimation for MIMO-OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6315–6328, 2021.

[42] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Transactions on Signal Processing*, vol. 68, pp. 764–779, 2020.

[43] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 15, pp. 4075–4089, 2017.

[44] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.

[45] P. Dong, H. Zhang, G. Y. Li, I. S. Gaspar, and N. NaderiAlizadeh, "Deep CNN-based channel estimation for mmWave massive MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 5, pp. 989–1000, 2019.

[46] S. Gao, P. Dong, Z. Pan, and G. Y. Li, "Deep learning based channel estimation for massive MIMO with mixed-resolution ADCs," *IEEE Communications Letters*, vol. 23, no. 11, pp. 1989–1993, 2019.

[47] Y. Jin, J. Zhang, B. Ai, and X. Zhang, "Channel estimation for mmWave massive MIMO with convolutional blind denoising network," *IEEE Communications Letters*, vol. 24, no. 1, pp. 95–98, 2019.

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

[49] J. Yang, Z. Wei, N. Li, L. Sang, and P. Li, "Enhanced multi-resolution hierarchical codebook design for adaptive compressed sensing based millimeter wave channel estimation," in *2016 IEEE/CIC International Conference on Communications in China (ICCC)*, 2016, pp. 1–5.

[50] K. Vuckovic, F. Hejazi, and N. Rahnavard, "MAP-CSI: Single-site map-assisted localization using massive MIMO CSI," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[51] T. Chen, Z. Zhang, pengjun wang, S. Balachandra, H. Ma, Z. Wang, and Z. Wang, "Sparsity winning twice: Better robust generalization from more efficient training," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=SYuJXrXq8tw

[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456.

[53] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[54] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[55] Y. Wu, X. Li, and J. Fang, "A deep learning approach for modulation recognition via exploiting temporal correlations," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2018, pp. 1–5.

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

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

[58] F. Hejazi, "https://github.com/farzamhejazik/gw2s," *Github*, 2021.

[59] "AI/ML for NR Air Interface — 3gpp.org," https://www.3gpp.org/technologies/ai-ml-nr, [Accessed 16-Jul-2023].

[60] J. J. Danker Khoo, K. H. Lim, and J. T. Sien Phang, "A review on deep learning super resolution techniques," in *2020 IEEE 8th Conference on Systems, Process and Control (ICSPC)*, 2020, pp. 134–139.

[61] A. Mazin, M. Elkourdi, and R. D. Gitlin, "Accelerating beam sweeping in mmWave standalone 5G new radios using recurrent neural networks," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, pp. 1–4.

[62] C. Liu, W. Yuan, S. Li, X. Liu, H. Li, D. W. K. Ng, and Y. Li, "Learning-based predictive beamforming for integrated sensing and communication in vehicular networks," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 8, pp. 2317–2334, 2022.

[63] X. Cheng, D. Duan, S. Gao, and L. Yang, "Integrated sensing and communications (ISAC) for vehicular communication networks (VCN)," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23 441–23 451, 2022.

Katarina Vuckovic received her B.S. in Aerospace Engineering (2017), B.S. in Electrical Engineering (2017), and M.S. in Electrical Engineering (2019) from Florida Institute of Technology. Presently, she is actively pursuing her Ph.D. in Electrical Engineering at the University of Central Florida, all while maintaining her steadfast position as an engineer at Collins Aerospace. Over the past seven years at Collins Aerospace, she has worked on a wide range of applications including wireless communication systems, automation, and data analytics. Currently, she holds a position as a research engineer at the Advanced Research and Technology (ART) department. In 2023, she received the Best Paper in Track 11 Award at the IEEE Aerospace Conference.

Mahdi Boloursaz Mashhadi (S'14-M'18, SM'23) is a Lecturer at the 5G/6G Innovation Centre (5G/6GIC) at the Institute for Communication Systems (ICS), University of Surrey (UoS), UK. Prior to joining ICS, he was a postdoctoral research associate at the Intelligent Systems and Networks (ISN) Research Group, Imperial College London, 2019-2021. He received B.S., M.S., and Ph.D. degrees in mobile telecommunications from the Sharif University of Technology (SUT), Tehran, Iran, in 2011, 2013, and 2018, respectively. He was a visiting research associate with the University of Central Florida, Orlando, USA, in 2018, and Queen's University, Ontario, Canada, in 2017. He has more than 40 peer-reviewed publications and patents in the areas of wireless communications, machine learning, and signal processing. He received the Best Paper Award from the IEEE EWDTS 2012 conference, and the Exemplary Reviewer Award from the IEEE ComSoc in 2021 and 2022. He has served as a panel judge for the International Telecommunication Union (ITU) to evaluate innovative submissions on applications of AI/ML in 5G and beyond wireless networks since 2021. He is an associate editor for the Springer Nature Wireless Personal Communications Journal.

Farzam Hejazi received a B.S., M.S., and Ph.D. degrees in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 2011, 2013, and 2018, respectively. He was working as a postdoctoral researcher at the University of Central Florida 2019-2021. He currently works as a senior machine learning engineer at Qualcomm. His current research interests lie in the areas of machine learning for wireless communication, Initial access in 6G/5G THz and mmWave systems, and radio frequency localization and sensing.

Nazanin Rahnavard (S'97-M'10, SM'19) received her Ph.D. in the School of Electrical and Computer Engineering at the Georgia Institute of Technology, Atlanta, in 2007. She is a Professor in the Department of Electrical and Computer Engineering and an Associate Faculty of Center for Research in Computer Vision (CRCV) at the University of Central Florida, Orlando, Florida. Dr. Rahnavard is the recipient of NSF CAREER award (2011), UCF Faculty Excellence in Mentoring Doctoral Students (2022), College of Engineering and Computer Science (CECS) Excellence in Research Award (2020), and UCF Women's History Month Honoree (2020). She has interest and expertise in a variety of research topics in deep learning, communications, networking, and signal processing areas. She serves on the editorial board of the Elsevier Journal on Computer Networks (COMNET) and on the Technical Program Committee of several prestigious international conferences.

Ahmed Alkhateeb received his B.S. degree (distinction with honor) and M.S. degree in Electrical Engineering from Cairo University, Egypt, in 2008 and 2012, and his Ph.D. degree in Electrical Engineering from The University of Texas at Austin, USA, in August 2016. Between Sept. 2016 and Dec. 2017, he was a Wireless Communications Researcher at the Connectivity Lab, Facebook, in Menlo Park, CA. He joined Arizona State University (ASU) in spring 2018, where he is currently an Assistant Professor in the School of Electrical, Computer and Energy Engineering. His research interests are in the broad areas of wireless communications, communication theory, signal processing, machine learning, and applied math. Dr. Alkhateeb is the recipient of the 2012 MCD Fellowship from The University of Texas at Austin, the 2016 IEEE Signal Processing Society Young Author Best Paper Award for his work on hybrid precoding and channel estimation in millimeter wave communication systems, and the 2021 NSF CAREER Award.