Multi-Modality Sensing in mmWave Beamforming for Connected Vehicles Using Deep Learning

Muhammad Baqer Mollah, *Graduate Student Member, IEEE,* Honggang Wang, *Fellow, IEEE,* Mohammad Ataul Karim, *Life Fellow, IEEE,* and Hua Fang, *Senior Member, IEEE*

Abstract—Beamforming techniques are considered as essential parts to compensate severe path losses in millimeter-wave (mmWave) communications. In particular, these techniques adopt large antenna arrays and formulate narrow beams to obtain satisfactory received powers. However, performing accurate beam alignment over narrow beams for efficient link configuration by traditional standard defined beam selection approaches, which mainly rely on channel state information and beam sweeping through exhaustive searching, imposes computational and communications overheads. And, such resulting overheads limit their potential use in vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communications involving highly dynamic scenarios. In comparison, utilizing out-of-band contextual information, such as sensing data obtained from sensor devices, provides a better alternative to reduce overheads. This paper presents a deep learning-based solution for utilizing the multi-modality sensing data for predicting the optimal beams having sufficient mmWave received powers so that the best V2I and V2V line-of-sight links can be ensured proactively. The proposed solution has been tested on real-world measured mmWave sensing and communication data, and the results show that it can achieve up to 98.19% accuracies while predicting top-13 beams. Correspondingly, when compared to existing been sweeping approach, the beam sweeping searching space and time overheads are greatly shortened roughly by 79.67% and 91.89%, respectively which confirm a promising solution for beamforming in mmWave enabled communications.

Index Terms—Beamforming, Connected and Autonomous Vehicles, Deep Learning, Millimeter-Wave Communications, Multi-Modality Sensing, Out-of-Band Information.

I. INTRODUCTION

THE utilization of millimeter-Wave (mmWave) bands (e.g., 28 GHz and 57-71 GHz bands) has been shown to bring an abundance of spectrum resources, considered as a key enabler for communications technologies in connected and autonomous vehicles (CAVs) domain [2]–[4]. For example, in the context of cooperative perceptions such as in high-throughput and low latency demanding applications in CAVs,

the connected vehicles exchange a vast amount of 3D point cloud data from LiDAR sensors instead of lightweight textual data [5], [6]. Realizing the high throughput and low latency benefits, mmWave communications enabled by vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communications is expected to pave the way toward accessible and safe autonomous transportation systems.

On the other hand, an inherent property of mmWave band is having high path attenuation of its signals, whereby massive antenna arrays are typically employed, drastically impacts quality performance degradation [7], [8]. Beamforming techniques may be utilized to address this bottleneck, including formulating the narrow beams, thereby achieving the desired high throughput from mmWave communications. Likewise, the narrow beams are expected to be aligned precisely and required to re-direct in accordance with the environmental settings and any changes. Generally, in codebook-based beamforming [9], beam selection techniques are applied to find the optimal beams in a number of pre-defined beam codebooks.

In practice, with standard defined beam selection techniques, the vehicles usually select the best beam pairs through an exhaustive beam measurements process [10], which in turn introduces computing and latency overheads in applications involving highly dynamic moving vehicles. This overhead problem occurs mainly due to tight contact times (the time period the receiver received the correct packets), frequent beam realignments, and changes in channel state estimation to perform beam computing. Hence, it is crucial to avoid beam misalignment along with reduced and reasonable overheads in order to take the fully potential benefits from mmWave communications. For instance, deployment based the 3GPP 5G-NR (New Radio) and IEEE 802.11ad standards [11] might take around 25 and 19 milliseconds, respectively to select the best beams through searching over all beam directions for 60 beam pairs and need to run over repeatedly when the vehicles move forward.

Considering this challenge in vehicular settings, recent works have suggested effective approaches to configure the communications links by leveraging the out-of-band contextual information/side information. As such, the out-of-band contextual information has been applied in mmWave beamforming tasks by utilizing other lower bands (sub-6 GHz bands) [12]–[17], RADAR communications bands [18]–[21], or extracted useful sensing information from on-board vehicular non-RF sensing devices (i.e., position coordinates [22]–[24], visual data [25]–[28], and point clouds [29]–[32]) are the few notable example works. According to [33], sev-

Muhammad Baqer Mollah and Mohammad Ataul Karim are with the Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth, MA 02747 USA (Emails: mmollah@umassd.edu; mkarim@umassd.edu).

Honggang Wang is with the Department of Graduate Computer Science and Engineering, Katz School of Science and Health, Yeshiva University, NY 10016, USA (Email: honggang.wang@yu.edu).

Hua Fang is with the Department of Computer and Information Science, University of Massachusetts Dartmouth, MA 02747, USA (Email: hfang2@umassd.edu).

(Corresponding Author: Honggang Wang.)

eral approaches pursuing efficient beamforming have been introduced based on multi-modality sensing over the recent years. However, among these related research, most of the works are limited primarily to vehicle-to-infrastructure (V2I) communications, based on ray-tracing simulations, or considered single sensing modality, but utilizing real-world measured wireless data as well as considering the mobility of connected vehicles under V2V connectivity have not been investigated yet. Accordingly, we introduce a unified solution for mmWave beamforming for both V2I and V2V communications, unlike prior works, with the aid of a proposed multi-modal deep learning model. The main contributions of this work are listed in detail below:

- Drawing inspiration from the recent success of multi-modality sensing, we propose a 5G-NR standard compatible deep learning-based solution to predict a subset of beams, that is, top-M beams, thereby significantly reducing the searching space and latency of beam sweeping.
- Instead of single modality, we employ different sensing modalities with a common goal on providing more detailed information and robustness in terms of errors and missing information beyond the limitations of the individual sensing in practical situations. Specifically, the position information is leveraged to facilitate on identifying the targeted vehicles in urban scenarios.
- Unlike other approaches, the proposed solution fully leverages the sensing data obtained from sensors, commonly available on the modern vehicles and road-side units, subsequently lowering down the complicacy.
- The proposed solution is validated with experiments on real-world multimodal 60 GHz mmWave wireless data, and present the results in terms of two meaningful matrices, namely accuracies and received power ratios to demonstrate its efficacy and applicability in beamforming tasks in reality.
- The proposed solution is shown to integrate into existing 5G-NR standard and the benefits we can get after integration. In particular, the results show that the proposed solution can reduce the beam sweeping searching and time overheads by 79.67% and 91.89%, respectively for top-13 beams.

Paper Organization: The rest of this paper is organized as follows. In section II, we review the most relevant related works on efficient beamforming, in particular, that have been introduced in the literature. In section III, we present the considered system model and formulated problem. The detailed steps of the proposed multi-modal solution is introduced in section IV. Next, the multi-modal feature extraction architecture including its necessary components are described in section V. The experimental settings and results are presented in section VI. Finally, conclusions are summarized in section VII. The list of major notations and their descriptions used in this paper is summarized in Table I.

II. RELATED WORKS

In this section, we review the most relevant works on efficient beamforming based on the out-of-band information which have been introduced in the literatures as follows.

TABLE I: Summary of major notations and descriptions.

Notation	Description
v_1, v_2 & v_3	Moving vehicles
$\mathcal{A}$	gNodeB, base station, or road-side unit
N_R	The numbers of antenna array elements
$\mathbf{h}_n$	Channel vectors in each communication links
$\mathcal{Q}$	Beam codebooks
$\mathcal{I}$	Unique beam indices
$\mathcal{B}$	The set of recommended beams
X_{pos}	Position data matrix
X_{vis}	Visual data matrix
X_{lid}	3D point cloud data matrix
$\mathcal{D}_s$	Training dataset
$\psi(\cdot, w)$	Neural network
w	Trainable weights of $\psi(\cdot, w)$
ψ_{w*}	Parameterized by optimal weights w^*
$\hat{\mathcal{P}}_t$	Received power from predicted beams
$\mathcal{P}_t^{gt}$	Received power from ground-truth beams
$\mathcal{T}_{nr}$	Required time of beam selection in 5G-NR
T_{ssp}	Single burst in synchronizing signal blocks
t_{burst}	Synchronizing signal burst periodicity
$\mathcal{T}_{nr}^{cnn}$	Sweeping time for the predicted beams

A. Uni-Modal Sensing Data

Recent works have shown that positioning, visual, and 3D point clouds sensing information have emerged as potential enablers for mmWave beamforming. The solution in [34] has leveraged machine learning technique to investigate how position information from GPS sensors can help to reduce the beam training overheads. However, considering the localization errors of GPS data, the authors in [35] have proposed a beam alignment approach, where Bayesian approach has been applied on the radio maps at a geographical area's each spatial location. Meanwhile, the work in [36] has developed a data driven method by transformer networks with particular attention mechanisms to estimate the line of sight (LoS) time of arrival (ToA), and then, a model-driven method to locate the users in mmWave multiple-input multiple-output (MIMO) communications by approximating the maximum likelihood time difference of arrival (TDoA) with the estimated results. Likewise, the work in [24] emphasizes on both location and orientation information to perform 3D beam selection by utilizing deep learning technique.

Similar concepts have been also utilized in [37] and [38]. In [37], the authors utilize spatially indexed historical information so that statistically significant beam sectors can be chosen in specific locations. Their proposed work is particularly focused on IEEE 802.11ad standard with an aim of reducing the time of beam sweeping, thereby improving the available time for data communications. And, the work in [38] uses the information about spatial correlation of stronger channels in-between the receiver and sender at given locations so that the users are able to track the information beforehand to establish communication link quickly. Earlier, the authors in [39] had proposed an approach, where a multi-path fingerprint-based database is maintained which records fingerprints of necessary location information along with channel characteristics so that the knowledge of reliable beamforming can be known in advance.

In addition to the utilization of position information, the capabilities of scene understanding by visual data by RGB

or depth (RGB-D) sensing cameras have also been utilized in a number of studies. For instance, the authors in [40] and [41] have presented computer vision assisted mobility management techniques. In particular, Anh et al. [40] have been utilized visual sensing information to understand the moving user and potential blockages by extracting the distance and azimuth/elevation angle like geometric channel parameters, which later have been used to predict the downlink state (i.e., data rate) using a deep neural network. The simulation results show that their proposed work outperforms the beam sweeping based on the 5G-NR standard-defined technique by achieving 30% more gains in throughput while ensuring a seamless beamforming gain over the targeted area. Whereas, the authors in [41] have used computer vision techniques to understand the location and speed of the vehicles, at the same time, the potential obstacles, and later, this information has been utilized to determine the best handover decision. Here, the authors have incorporated a multivariate regression technique to play a vital role to predict the time until the moving vehicle arrives at the blockage area.

Meanwhile, the authors in [42] have designed a different computer vision assisted mmWave beamforming focusing on beam weight prediction, where the visual detection by a convolutional neural network and limited wireless feedback have been fused to enhance the spectral efficiency as well as a codeword rotation method implemented by Householder transformation to reduce the feedback overheads. Indeed, Osman et al. [43] have shown how visual sensing data collected from 360° cameras can help to contribute on beam predictions in V2V communication scenarios in their work. Besides, in [44], the authors have focused on user identification task together with a deep neural network and bounding box estimation so that the computer vision aided communications can be employed in realistic crowded scenarios.

The work in [45] has presented a multi-task learning module to predict the optimal road side units (RSU) and beam pairs by proposing deep learning based proactive RSU selection network (PRSN) and beam pair searching network (BPSN), respectively for dynamic V2I settings. Here, the authors have utilized multiple cameras installed on the RSUs for the purpose of getting multiple views, which essentially enriches the visual sensing information in terms of making more accurate and reliable decisions than single view techniques. However, utilizing raw visual sensing data directly introduces considerable storage and computational overheads. Instead, the authors in [46], [47] and [48] have utilized targeted object masks and bounding boxes like environment semantics information to address such overheads problems.

Although the aforementioned works have shown that the visual sensing data has been utilized as out-of-band information along with satisfactory scene understanding, having dark or glare visual images may lead to inaccurate beamforming decisions by the models/algorithms. In contrast, 3D structures captured by LiDAR point clouds can offer a better understanding of surrounding environment and scene understandings through high-resolution mapping and positioning. Accordingly, the authors in [32] and [31] have considered utilizing point clouds for detection of line-of-sight, and then,

predict the best beams by deep convolutional neural network and federated learning, respectively. While these two works have focused on predicting the current beams, similar work has been studied in [49] using by a recurrent neural network (RNN), and shown that with the aid of point clouds, it is possible to predict the future beams along with current ones.

Wu et al. [50] have shown in their work that leveraging convolutional neural network on point clouds and in-band mmWave signals can be a help in predicting the occurrence of line-of-sight link blockages in advance so that beam switching/hand-off decisions can be made to ensure reliable communications. Most importantly, the authors have validated their work on real-world datasets, showing that their approach can achieve around 95% and 80% accuracies while predicting the future blockages within 100 ms and 1, 000 ms, respectively. Besides, Ohta et al. in [51] have demonstrated with real experiments that it is possible to leverage the inherent properties of point clouds to understand the wireless propagation environment. They have designed a proposed 3D convolutional neural network with long short-term memory and gradient boosting decision tree for implementing on point clouds so that the link quality, more specifically, the mmWave received signal attenuations due to any blockages (e.g., pedestrians) can be predicted proactively. Further, the work in [52] has introduced a deep learning based technique, namely COPILOT, to identify the best road side unit having high bandwidth mmWave connectivity for cooperative perception purposes. The aim of this COPILOT is to support connected vehicles to recognise the transient blockages along the road with the help of point clouds, which eventually can help them to decide connecting the best road side unit by proactive handoffs. In particular, when compared to traditional reactive handoffs techniques, the experiment with real 60 GHz devices has shown that the COPILOT can offer upto 69.8% improvement on latency.

B. Multi-Modal Sensing Data

Beyond the single modality works, several solutions on leveraging muti-modality sensing to facilitate mmWave beamforming related tasks have been explored in recent years. For example, the works in [53] and [54] have focused on deep learning assisted fusion of positioning and visual sensing data. In particular, the work [53] utilized fine-tuned ResNet-50 model and data normalization technique on visual data and position data, respectively before feeding into a concatenation module, and then used a multi-layer perceptron network to get the predicted beam indices. However, the work [54] has incorporated transformer network and semantic localization techniques to contribute more on developing a robust beamforming approach while obtaining the visual data in varrying weather such as rainy, foggy, sunny, and snowy conditions. On the other hand, Zecchin et al. [55] have proposed a fusion between point clouds and GPS position data, where the authors have improved the beam prediction accuracies by utilizing non-local attention block particularly for non-line-of-sight scenarios, and also utilized a loss function motivated by knowledge distillation as well as a curriculum training strategy to further improve the training speed as well as

TABLE II: Summary of relevant state-of-the-art multi-modal sensing related works and our contributions.

Work	Frequency Band	Position	Visual	Point Clouds	Problem	Highlighting Method	V2I	V2V	Validation	Dataset and tool
[53]	60 GHz	✓	✓	✗	Beam selection	Computer vision	✓	✗	Real-world data	DeepSense6G
[54]	60 GHz	✓	✓	✗	Beam selection	Transformers network	✓	✗	Real-world data	DeepSense6G
[55]	28 GHz	✓	✗	✓	Beam selection	Deep neural network	✓	✗	Synthetic ray-tracing data	Raymobtime
[56]	60 GHz	✓	✗	✓	Beam selection	Deep neural network	✓	✗	Real-world data	DeepSense6G
[57]	28 GHz	✓	✓	✓	Beam selection	Deep neural network	✓	✗	Synthetic ray-tracing data	Raymobtime
[58]	28 GHz	✓	✓	✓	Multi-user beam selection	Deep neural network	✓	✗	Synthetic ray-tracing data	Raymobtime
[59]	60 GHz	✓	✓	✓	Sector selection	Federated learning	✓	✗	Real-world data	FLASH
[60]	60 GHz	✓	✓	✓	Beam selection	Digital twin	✓	✗	Real-world data	FLASH and RemCom [61]
Ours	**60 GHz**	✓	✓	✓	**Beam selection**	**Deep neural network**	✓	✓	**Real-world data**	**DeepSense6G**

final predicted results. Finally, this work has been validated with a synthetic ray-tracing dataset, namely Raymobtime[1]. In another work, Bian et al. [56] have presented a beam tracking method on fusing the measured point clouds and position data by a proposed deep neural network, however, the studies was limited to only beam tracking in vehicle-to-vehicle connectivity.

Meanwhile, multiple works in [57]–[60] have demonstrated how to fuse features from visual, point clouds, and position sensing data together. In particular, the work in [57] has addressed data redundancy issue by introducing a low redundancy data based optimal beam prediction approach. In this approach, the authors utilize three steps, namely, feature extraction and fusion, reduction of dimensionality and redundancy by principal component analysis, and prediction with multi-layer perceptron network. In comparison, the authors in [58] have focused on proposing a multi-user beamforming strategy for mmWave MIMO systems. For that, a multimodal deep neural network model including an alternative loss function, namely supervised soft-contrastive loss, has been proposed to estimate the channel beamspace representations from the sensing data. This estimated beamspaces have been utilized later to identify the radio frequency precoders for all users at the base station, which eventually can help to establish multi-user communication links.

In another work, Salehi et al. [59] have considered predicting the best sectors specifically in large codebooks, and to achieve this, a multimodal federated learning framework named FLASH-and-Prune has been proposed. In particular, their work starts with predicting the optimal sectors locally by multiple vehicles individually, then aggregates these local model parameters to make a globally optimized model, and finally, a model pruning strategy is applied to further reduce the overheads caused by model parameters exchanging in federated learning settings. In this work, the authors have also prepared a publicly available datasets with real-world measured wireless data, refereed as FLASH[2] and validated

their solution by it.

In addition, the work in [60] has designed a digital twin, which closely mimics the real wireless scenarios to speed up the beam selection procedures particularly in unseen environments. In their design, the digital twins are mounted on the vehicles, and collectively coined as multiverse of twins term. The goal of this multiverse of twins is to possibly obtain the necessary ray tracing components and complex interactions at different levels of fidelity from the real world and then utilize them to make decisions on finding the optimum twins based on latency constraints. The results have shown that the proposed Multiverse can offer roughly 85.22% for top-10 prediction accuracy for unseen scenarios.

To the best of our knowledge, the aforementioned multi-modality related solutions are the most relevant to ours, and accordingly, we show how this current work stands out from existing research in the table II.

III. SYSTEM MODEL AND PROBLEM STATEMENT

In this section, we first present the adopted communication system model, and then formally state the multi-modality sensing-aided beam prediction problem addressing in this work.

A. System Model

In this work, we consider a vehicular communications system model operating at mmWave frequency bands as presented in Fig. 1. The model consists of a moving vehicle v_2 providing communications services, e.g., cooperative perception, to another moving connected vehicle v_1 within its coverage area, and these two moving vehicles, v_1 and v_2, are connected with each other by vehicle-to-vehicle (V2V) communication links. We also consider a stationary base station (typically referred as gNodeB in 5G) unit $\mathcal{A}$ is communicating with the moving connected vehicle v_3 to provide communications services, for example, high-definition map updating, within its coverage area by vehicle-to-infrastructure (V2I) communication links. The vehicle v_1 and base station $\mathcal{A}$, acting as receivers R_x, are

[1]https://www.lasse.ufpa.br/raymobtime

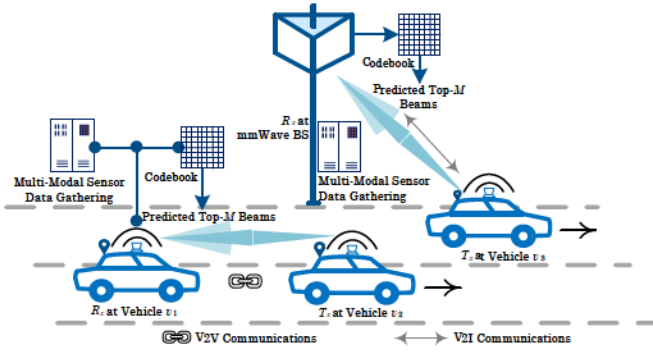[2]https://genesys-lab.org/multimodal-fusion-nextg-v2x-communications

Fig. 1: Illustration of our considered mmWave enabled V2I and V2V communications system model.

equipped with a number of sensing devices, i.e., visual and LiDAR sensors to obtain the multi-modal sensing information from the surroundings, for example, moving vehicles v_2 and v_3, whereas the vehicles v_2 and v_3, acting as transmitters T_x, have GPS sensing devices to gather the position information, such as latitude and longitude along with the timestamps, and share the information in turn with v_1 and $\mathcal{A}$ in real-time, respectively.

Additionally, both the vehicle v_1 and base station $\mathcal{A}$ have N_R numbers of antenna array elements, codebooks of $\mathcal{Q}$ predefined beams with, each elements of codebook referring to a specific beam orientation. The antenna array elements enable v_1 and $\mathcal{A}$ to perform beamforming so that v_1 and $\mathcal{A}$ can obtain adequate received powers. Each vehicle users v_2 and v_3 has, for simplicity, one antenna which is oriented always towards the receivers R_x. The beam codebooks can be expressed as $\mathcal{Q} = \{q_1, q_2, ..., q_{|\mathcal{Q}|} : q_i \in \mathbb{C}^{N_R \times 1}\}$, where $|\mathcal{Q}|$ is the cardinality of $\mathcal{Q}$. The beam vector can be expressed as $(\mathcal{Q} \times 1)$.

We further consider that orthogonal frequency-division multiplexing is used in the downlink communications with $\mathcal{N}$ subcarriers. For the given transmitted symbol $s_n \in \mathbb{C}$, the variance of the noise σ^2, and the received complex Additive White Gaussian noise $\omega_n \sim \mathcal{N}_\mathbb{C}(0, \sigma_n^2)$, the downlink transmitted received signal at n^{th} subcarrier at time instant t can be written as:

$$x_n[t] = \mathbf{h}_n^T[t]q_i[t]s_n + \omega_n[t], \quad (1)$$

where, $\mathbf{h}_n \in \mathbb{C}^{N_R \times 1}$ indicates the channel vectors in each V2V and V2I links. In particular, considering the 3D geometric channel model presented in [39], and denoting $p(\cdot)$ the shaping pulse, T_s the symbol period, α_l the complex path gain, τ_l the delay, $\mathbf{a}_R(\cdot)$ the complex steering vector at receiver end array, θ_l^A elevation angle of arrival, ϕ_l^A azimuth angle of arrival, of the path l, the $\mathbf{h}_n$ can be expressed as frequency-domain at m^{th} symbol vector as follows:

$$\mathbf{h}_n = \sqrt{N_R} \sum_{c=0}^{C-1} \sum_{l=0}^{L-1} \alpha_l e^{-j\frac{2\pi n}{\mathcal{N}}c} p(mT_s - \tau_l) \mathbf{a}_R(\theta_l^A, \phi_l^A), \quad (2)$$

where, L and C denote the number of total channel paths and cyclic prefix length, respectively.

B. Problem Statement

Given the system model, the primary task of this work is to make predictions for optimal beams q_i^* from the predefined codebooks $\mathcal{Q}$ such that the mmWave received powers are maximized. Hence, if the $\mathcal{A}$ and v_1 want to determine the optimal beams q_i^* from their beam codebooks $\mathcal{Q}$ to serve the v_3 and v_2, respectively, are determined which maximize the received powers. From the received signal $x_n[t]$ presented in (1), we can get the received power by summing over $\mathcal{N}$ subcarriers at the $\mathcal{A}$ and v_1 sides for the receiver codebook i. However, if we convert the beam codebooks $\mathcal{Q} = \{q_1, q_2, ..., q_{|\mathcal{Q}|}\}$ into a unique index as $\mathcal{I} \in \{1, 2, 3, ..., \mathcal{M} : \mathcal{M} \leq |\mathcal{Q}|\}$, predicting optimal beams from beam codebooks are same as predicting optimal beam index. Then, predicting the optimal beams problem can be formulated as:

$$\mathcal{I}^* = \arg\max_{i \in \{1,2,3,....,|\mathcal{Q}|\}} \sum_{n=0}^{\mathcal{N}-1} |\mathbf{h}_n^T[t]q_i[t]s_n|^2 \quad (3)$$

Specifically, our goal is to predict a set of recommended beams $\mathcal{B} = \{1, 2, 3, ..., M\}$ as top-M beams such that $\mathcal{I}^* \in \mathcal{B}$ as top-1 beam. Nevertheless, to solve the problem in (3), traditional beamforming approaches either leverage the knowledge of $\mathbf{h}_n$, which is rather difficult to acquire in mmWave enabled vehicular scenarios or carry out an exhaustive search over all codebooks thus introducing large training overheads. However, instead of relying on the knowledge of $\mathbf{h}_n$ and an exhaustive search, we aim to utilize deep learning models on the available multi-modal sensory information to solve the optimization problem.

IV. THE PROPOSED MULTI-MODAL SOLUTION

The proposed multi-modal solution is composed of the following three steps: (i) data acquisition, (ii) model training with deep learning, and (iii) deployment in real environment (inference step).

A. Data Acquisition

We first define the position, visual, and 3D point cloud data matrices at each time instant t as $X_{pos}[t] \in \mathbb{R}^{N_t \times 2}$ (two-dimensional latitude and longitude), $X_{vis}[t] \in \mathbb{R}^{N_t \times w_v \times h_v \times c}$ (width, height, and color channels of visual data), and $X_{lid}[t] \in \mathbb{R}^{N_t \times d_l \times h_l \times w_l}$ (depth, height, and width of point cloud data), respectively, where, N_t denotes the total number of samples for training.

In data acquisition step, the connected vehicle v_1 and basestation $\mathcal{A}$ collect the multi-modal sensing data along with corresponding true beams within their own coverage areas in different scenarios (e.g., day and night times), denoted by, $\mathcal{S} = \{s\}_{s=1}^S$, and each builds a training dataset $\mathcal{D}_s = \{(X_{pos}^{(s)}, X_{vis}^{(s)}, X_{lid}^{(s)}, Y^{(s)})\}_{s \in \mathcal{S}}$. Here, $Y^{(s)} = \{\mathcal{I}_j^*\}_{j=1}^{|\mathcal{D}_s|}$ represents the corresponding true beams, which can be determined either from the maximum received powers at codesbooks or by utilizing any existing beam selection approaches.

B. Model Training

This model training step leverages deep learning architectures to construct a multimodal neural network so that the mapping of input modalities from $\mathcal{D}_s$ to the predicted beam indices can be learned with good generalization abilities. The multimodal network has three unimodal models, which essentially act as feature extractors from the sensing modalities. For that, we define $\psi(\cdot, w)$ as a neural network, where w is its trainable weights which can be optimized to find the optimal weights w^* by solving the following problem:

$$w^* = \arg\min_{w} \frac{1}{N_t} \sum_{p=0}^{N_t-1} \mathcal{L}(\psi(X^{(p)}, w), Y^{(p)}), \qquad (4)$$

where, $X \in [X_{pos}, X_{vis}, X_{lid}]$ and $\mathcal{L}(\cdot)$ indicate the loss function to measure how far the neural network's predicted beams deviate from the true beams Y. Following that, a softmax function $\mathsf{Softmax}(\cdot)$ is applied on the output of the neural network $\psi(X^{(p)}, w)$ which helps to give the beam prediction result as $\hat{\mathcal{I}} = \mathsf{Softmax}(\psi(X^{(p)}, w))$ by converting the unnormalized output from the neural network into normalized values (a probability distribution). However, since our networks are particularly designed for predicting the top-M beams, the cross-entropy is utilized as a loss function, which can be calculated by $\mathcal{L}(\hat{\mathcal{I}}, Y) = -\frac{1}{N_t} \sum_{p=0}^{N_t-1} \sum_{q=0}^{|\mathcal{Q}|-1} Y_q^{(p)} \log \hat{\mathcal{I}}_q^{(p)}$, where, $|\mathcal{Q}|$ is the total number of beam indices, $Y^{(p)}$ is the true beam index for the pth sample, and $\hat{\mathcal{I}}^{(p)}$ is the predicted beam index probability by softmax for p^{th} sample. Further, we utilize a stochastic optimization technique namely adaptive moment estimation (Adam) [62] for faster convergence on minimizing the loss function while training the neural network models.

C. Inference: Deploying in Real Environment

While it may take certain amount of time for data acquisition and model training, after the data acquisition and training, the trained models must have the capability to put into the work in real environment for beams predictions with new data. This step is often referred as inference step. Consequently, the trained models is utilized in the inference step to make probabilistic predictions of top-M beams within the coverage areas as

$$\hat{\mathcal{I}} = \mathsf{Softmax}(\psi_{w^*}(X_{pos}^n, X_{vis}^n, X_{lid}^n)), \qquad (5)$$

where, ψ_{w^*} is parameterized by optimal weights w^*, and X_{pos}^n, X_{vis}^n, and X_{lid}^n denote the new input samples from GPS, Visual, and LiDAR sensors, respectively.

However, when it comes to real environment deployment, certain prediction errors may arise leading to performance degradation. To address this, the trained model may need to update with relatively smaller datasets than the original ones and fine-tune in accordance with the performance requirements.

V. FEATURE EXTRACTIONS ON MULTI-MODAL SENSING

In this section, the proposed feature extractions from multimodal sensory data inputs is described in detail highlighting the data processing steps, the architecture of feature extractions, and finally, beam selection procedures.

A. Data Preprocessing

According to Section IV, the training dataset is required to feed into multimodal neural networks as inputs, and the inputs are desired to be fixed and consistent in size. In this regard, feeding multimodal sensing data directly may require developing deep learning models with high complexity in terms of both architecture and computational cost. Thus, to make it fit as well as accelerate the convergence of the deep learning models, we process the sensing data into following preprocessing procedures.

Position Data Preprocessing: The raw position data captured from GPS sensors are basically in Decimal Degrees, specifically the latitude and longitude values are from $-90°$ to $+90°$ and $-180°$ to $+180°$, respectively. Hence, we pass through the raw position data into a preprocessing procedure, namely data normalization. For that, we calculate the normalized values of latitude x'_{lt} and longitude y'_{lg} at each rows of the matrix as: $(x'_{lt}, y'_{lg}) = ((x_{lt} - x_{lt\_min})/(x_{lt\_max} - x_{lt\_min}), (y_{lg} - y_{lg\_min})/(y_{lg\_max} - y_{lg\_min}))$, within Cartesian coordinate system, where x_{lt} and y_{lg} are the raw latitude and longitude values, while the max and min denote their maximum and minimum values, respectively.

Visual Data Preprocessing: Converting the obtained visual image data into the inputs for the expected model involves two procedures, i.e., resizing and normalization. In resizing, we fix the spatial dimensions of the visual images into specific $224 \times 224 \times 3$ shapes, where $224 \leq w_v$ and $224 \leq h_v$. Here, 224×224 indicates that each images has a resolution of 224 pixels in width and 224 pixels in height, for a total of $224 \times 224 = 49,536$ pixels. And, 3 represents the red, green, and blue color channels range from 0 to 255, which undergo a normalization by subtracting the mean values $(0.485, 0.456, 0.406)$ from them and then, dividing the results by the standard deviation values $(0.229, 0.224, 0.225)$ so that the normalized values become zero mean and unit variance.

Point Cloud Data Preprocessing: For each sample, as a result of raw measurements by LiDAR sensors output a large point cloud data $P = \bigcup_{i=0}^{n-1} \{(x_i, y_i, z_i)\}$, where n is the point numbers, and $\{(x_i, y_i, z_i)\}$ denotes the coordinate of the ith point representing the obstacle. Even though our expected model will be able to handle the raw point clouds directly, the point clouds samples may have varying number of points. Hence, we preprocess the P either by padding or downsampling while preserving the underlying essential features so that each sample will have consistency in terms of fixed number of points, and we set the fixed number of points as $15,000$. In padding, we pad the rest of points with zeros if the number of points in samples have lesser than the desired points whereas, in downsampling, we randomly eliminate the points otherwise.

Finally, the revised matrices with preprocessed data at time stamp t are then be denoted as $X'_{pos}[t]$, $X'_{vis}[t]$, and $X'_{lid}[t]$, respectively.

B. Multi-Modal Feature Extractions Architecture

The designed deep learning architecture of proposed feature extractions for mmWave beamforming task is shown in Fig. 2. In this architecture, we basically choose a type of deep learning techniques namely, convolutional neural networks. Due to having convolutional operations, such neural networks can achieve very good performance on visual data, however, we also utilize the same in other two modalities so that the resulting corresponding features can be fused in a straightforward way. In the following, we present the detailed feature extractions from each modalities.

1) Position Feature Extraction: The proposed feature extraction model for position data adopts two components: the convolutional blocks and a flatten layer. Specifically, the convolutional blocks, the key building blocks used in this proposed model, are made up of three convolutional blocks. These convolutional blocks are formed together with three sequence of 1D (one-dimensional) convolutional layers, however, each blocks are followed by a max-pooling layer.

Given the preprocessed position data X'_{pos} as an input to the proposed model, the convolutional block $\mathsf{ConvNetBlock}_1$ first compute the inputs to extract the features by learning the hidden meaningful representations. This block is repeated twice as $\mathsf{ConvNetBlock}_2$ and $\mathsf{ConvNetBlock}_3$ in the network to make enhanced feature extraction and improved performance. However, the max-pooling layers in all blocks contribute to downsize the feature maps while keeping the most significant features, leading to decreasing the risk of the model becoming overfitted. Beyond the convolutional blocks, the output from $\mathsf{ConvNetBlock}_3$ is passed though the flatten layer, which is employed to flatten the output received from the last convolutional block, while not effecting the batch size. The overall process is presented as follows.

$$
\begin{aligned}
c_1 &= \mathsf{ConvNetBlock}_1(g_t) \\
c_2 &= \mathsf{ConvNetBlock}_2(c_1) \\
c_3 &= \mathsf{ConvNetBlock}_3(c_2),
\end{aligned}
\tag{6}
$$

where, c_1, c_2, and c_3 are the hidden vectors, and g_t is considered as a variable representing the vector elements of $X'_{pos}[t]$.

2) Visual Feature Extraction: For the extraction of visual data, we employ fine-tuned EfficientNet-B0 model [63]. This model utilizes a compound scaling method, which applies an uniform scaling over depth, width, and resolution all together to the baseline model depending on the available resources. In particular, the baseline model of EfficientNet-B0 comprises two types of building blocks, namely, the 2D convolutional (Conv2D) and mobile inverted bottleneck convolution (MBConv) blocks. With these building blocks, the model works in the following ways.

At first, the Conv2D block performs an initial convolutional operation by applying 3×3 filter on the preprocessed visual data input X'_{vis} to capture the simple basic features, such as edges and textures. Besides, this operation also utilizes a stride of 2, which basically downsamples the feature maps with reducing the spatial dimensions, i.e., heights and widths, of the inputs by a factor of 2 so that the subsequent blocks will have ability to contribute on recognizing more complex and abstract features.

The resulting outputs are passed through a number of MBConv blocks. The MBConv blocks are designed by inverter residual block structure and utilizes depthwise separable convolution along with an additional squeeze and excitation layer. The depthwise separable convolutional process deals with both spatial dimensions (height and width) and depth dimensions (number of channels) by following depthwise convolution and pointwise convolution operations after one another.

Based on expanding the number of channels in expansion phase, the MBConv blocks have two variants including MBConv1 (no expansion) and MBConv6 (expansion by a factor of 6). From a structural point of view, each MBConv blocks start with expanding the channel inputs by a 1×1 convolution, which then utilizes 3×3 or 5×5 depthwise convolution, and finally a 1×1 pointwise convolution to adjust the dimensionality of the outputs so that the inputs and outputs can be added through the skip connection. The above processes can be represented as follows.

$$
\begin{aligned}
\mathsf{Conv}(x) &= \mathsf{Swish}(\mathsf{BNorm}(\mathsf{Conv1\times1}(x))) \\
\mathsf{DConv}(y) &= \mathsf{SE}(\mathsf{SiLU}(\mathsf{BNorm}(\mathsf{DConv}(\mathsf{Conv}(x))))) \\
\mathsf{PConv}(z) &= \mathsf{Swish}(\mathsf{BNorm}(\mathsf{PConv1\times1}(\mathsf{DConv}(y)))),
\end{aligned}
\tag{7}
$$

where, $\mathsf{Conv1\times1}(\cdot)$, $\mathsf{DConv}(\cdot)$, $\mathsf{PConv}(\cdot)$, $\mathsf{BNorm}(\cdot)$, and $\mathsf{SiLU}(\cdot)$ are the functions of 1×1 convolution, depthwise convolution, pointwise convolution, batch normalization, and activation, respectively.

Following the MBConv blocks, the outputs undergo a final convolutional block which involves a 1×1 convolution to aggregate the features learned from previous blocks and yield a large number of channels (1280 in our case), and then, a global average pooling to make an average of each feature map obtained from the final convolutional block, resulting in reducing the spatial dimensions to a single value. And, in this way, the detailed features from visual inputs can be captured. Besides, it should be noted that the batch normalization and SiLU activation are applied after each convolutional operations. Here, the Sigmoid Linear Unit (SiLU), also known as Swish, function is a type of smooth and non-linear activation function, and it can be defined as:

$$
\mathsf{SiLU}(x) = x \times \mathsf{Sigmoid}(x),
\tag{8}
$$

where, $\mathsf{Sigmoid}(x) = (1 + \exp(-x))^{-1}$ denotes as the sigmoid function.

3) Point Cloud Feature Extraction: Motivated by the capabilities of handling unordered 3D point clouds directly as well as having invariance properties under order and transformations, we utilize fine-tuned PointNet model [64]. In particular, the PointNet model consists of three parts: (a) input and feature transformations, (b) shared multi-layer perceptrons (1D convolution), and (c) max pooling. The model works in the following three ways:
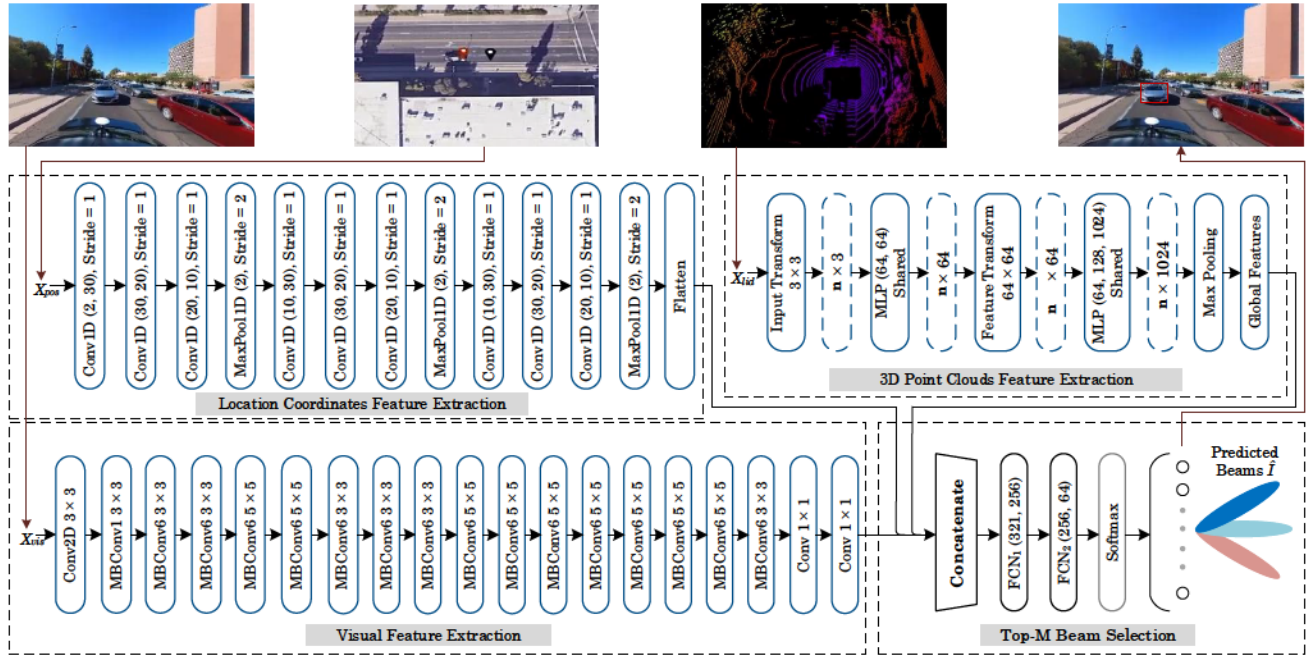
Fig. 2: The details diagram of the proposed deep learning model on multi-modality sensing for mmWave beamforming, which is mainly composed of three feature extraction and one top-M beam selection components.

The primary role of input transformation is to take point cloud inputs from X'_{lid}, and learns a 3×3 transformation matrix which is then multiplied to every input points to obtain invariance property to transformation, such as rotation and translation by aligning the input point clouds in a canonical coordinate space. In input points shape n×3, n represents the total number of points in each point cloud samples, whereas 3 refers the X, Y, and Z coordinates. After applying input transformation, the transformed data (3 dimensions) undergoes a series of multi-layer perceptrons $(64, 64)$ to learn the individual local features through processing each point independently, which returns an output with uplifting the dimensions to 64. While capturing such local features, the same weights are shared for each point, thereby achieving invariant to the order of the points.

Next, the feature transformation takes the individual point features obtained in first part as inputs (n×64). Similar to the input transformation, it repeats to learn a matrix again, which is then applied across all individual point features for further alignment and refining on learning the local features, allowing the outputs have invariance under different scales. Instead of 3×3, now the transformation matrix becomes 64×64. After feature transformation, let $P = \{p_1, p_2, ..., p_n\}$ as the outputs, where, p_i can be the ith point within the point clouds. These outputs are fed later into the an another shared multi-layer perceptrons $(64, 128, 1024)$ to emphasize on obtaining features for subsequent processing. Here, the 64, 128, 1024 refers the inputs mapping start with dimensions of 64, to 128, and at the end, to 1028. This can be denoted as $f_i = \text{MLP}(p_i)$, where MLP indicates multi-layer perceptrons, and f_i is the local feature of i_{th} point p_i. Particularly, the multi-layer perception is comprised of a number of fully connected layers along with a non-linear activation function

namely ReLU (Rectified Linear Unit). For all features, we can write as follows:

$$
\begin{aligned}
f_1 &= \text{ReLU}(W \times p_1 + \text{bias}) \\
f_2 &= \text{ReLU}(W \times p_2 + \text{bias}) \\
f_3 &= \text{ReLU}(W \times p_3 + \text{bias}) \\
&\cdot \\
&\cdot \\
&\cdot \\
f_n &= \text{ReLU}(W \times p_n + \text{bias}),
\end{aligned}
\tag{9}
$$

where, W and bias are the corresponding weights and biases, respectively.

Serving as the last operation in max pooling, it aggregates all individual point features, such as $f_1, f_2, f_3, ..., f_n$ obtained from last shared multi-layer perceptrons, which finally results in a constructed global feature representation, which can be denoted as

$$
\mathcal{G} = \text{MaxPool}(f_1, f_2, ..., f_n),
\tag{10}
$$

where, MaxPool is a max pooling function adhering a symmetric function operation, thereby making the model to work regardless of point orders.

C. Top-M Beams Selection

Assuming the outputs from each feature extractors as z_{pos}, z_{vis}, and z_{lid}, receptively, the beam selection module takes them as inputs. These output features z_{pos}, z_{vis}, and z_{lid} are then fused along the feature dimensions, and we utilize concatenation to fuse the features. The resulting concatenated features, after that, is forwarded to the two fully connected (dense) layers FCN_1 and FCN_2. The two fully connected layers together encompass a prediction by applying weights
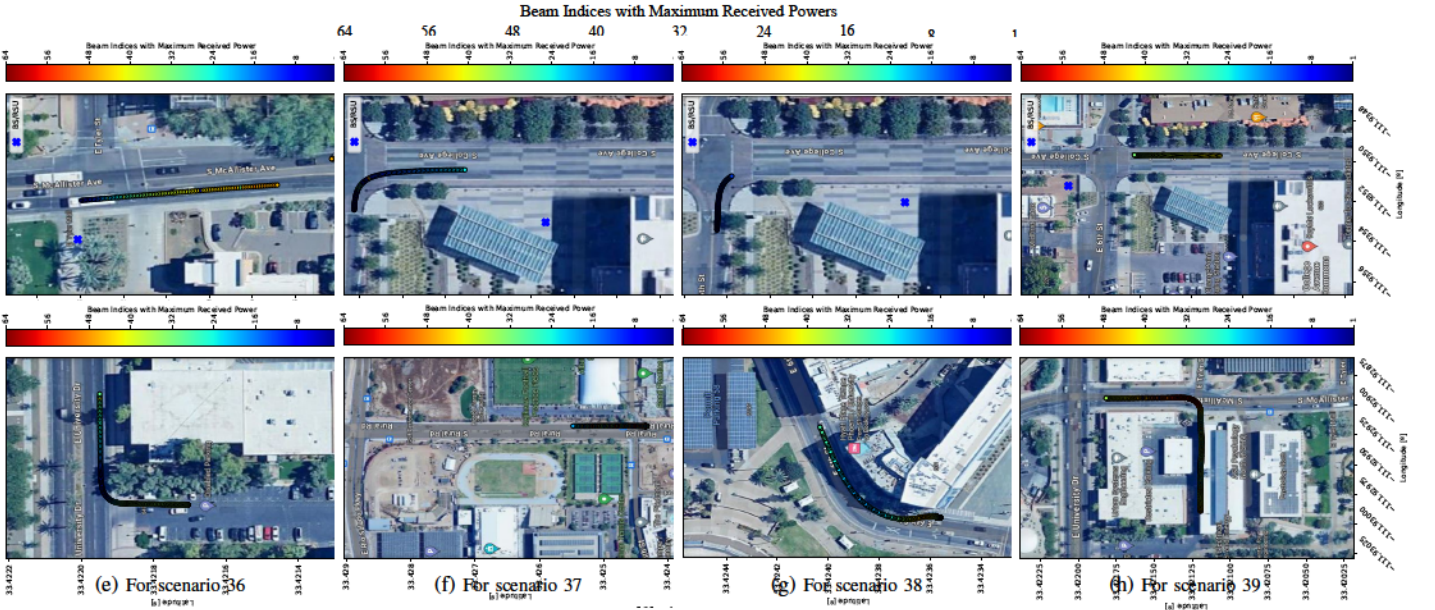
Fig. 3: Visual representation of receiver vehicle's GPS location data points (100 and 400 samples for V2I and V2V scenarios, respectively) along with corresponding best beam indices out of 64 beams on Google Map satellite view.

and biases. The weights and biases of these layers will be learned during the training process, enabling the multi-modal model to map input features to the desired output. We can describe as

$$
\begin{aligned}
\chi &= \mathsf{Concate}(\mathbf{z}_{pos}, \mathbf{z}_{vis}, \mathbf{z}_{lid}) \\
\mathcal{C}_1 &= \mathsf{FCN}_1(\chi) \\
\mathcal{C}_2 &= \mathsf{FCN}_2(\mathcal{C}_1),
\end{aligned}
\tag{11}
$$

where, $\mathsf{Concate}$ is the concatenation function, and χ, $\mathcal{C}_1$ and $\mathcal{C}_2$ denote the results from concatenation operation and the two connected layers, respectively.

In the end, the softmax layer is involved as last layer, where a softmax activation function ($\mathsf{Softmax}(\cdot)$) is applied on the output from the second fully connected layer to transform the values into probability scores within the range $[0, 1]$ with total value 1, which can be interpreted to eventually determine the top-M beams $\hat{\mathcal{I}} = \mathsf{Softmax}(\mathcal{C}_2)$.

VI. EXPERIMENTS AND EVALUATIONS

In this section, we present the datasets description, implementation settings, and the results obtained from a series of experiments. Subsequently, we also present the performance comparison with existing standard defined approach in terms of computational and communications overheads.

A. Datasets Description

In this work, we consider the DeepSense 6G dataset [65], [66], a collection of real-world mmWave sensing and communications measurements, to evaluate the effectiveness of proposed top-M beam prediction introduced in Section V. Specifically, consistent with the considered system model and formulated problem, we adopt scenario 31 ($7,012$ samples), scenario 32 ($3,235$ samples), scenario 33 ($3,981$ samples), scenario 34 ($4,439$ samples), scenario 36 ($24,800$ samples),

and 37 ($31,000$ samples), 38 ($36,000$ samples), and 39 ($20,400$ samples). In particular, the scenarios 31 to 34 are for V2I communications, whereas the scenarios 36 to 39 are for V2V communications. These scenarios are captured with two testbed setups including the stationary unit ($\mathcal{A}$) along with one moving vehicle (v_3), and two moving vehicles (namely, v_1 and v_2) deployed in diverse outdoor urban scenarios having long and shorter distances. And, the data has been collected at Tempe, Phoenix, and Scottsdale of Arizona during both day and night. Most importantly, this dataset has the mapping between beam codebooks and measured corresponding received powers for line-of-sight (LoS) conditions in both V2I and V2V scenarios. Other related datasets on multi-modal sensing and communications, such as Raymobtime [67] and FLASH [68] do not consider V2V scenarios, which limit their use in this work. However, it should be noted that since our proposed solution already utilizes multi-modal data, which could be further fine-tuned and revised to adjust beamforming prediction in non-line-of-sight (NLoS) scenarios, such as including alternative relay node selection task.

Here, the unit $\mathcal{A}$ and unit v_1 act as receivers R_x, whereas the other units v_3 and v_2 act as transmitters T_x, and both are operating at 60 GHz carrier frequencies. Additionally, the units $\mathcal{A}$ and v_1 employ mmWave phased arrays where, each of the phased arrays has 16 elements ($N_R = 16$) uniform linear arrays (ULA). The unit v_1, in particular, has four mmWave phased arrays which are facing right, left, back, and front directions on the vehicle. The number of pre-defined and over-sampled codebook beams for both phased array of units $\mathcal{A}$ and v_1 are 64 ($|\mathcal{Q}| = 64$), whereas the units v_3 and v_2 have one antenna element each with a total $64 \times 1 = 64$ beam pairs.

In particular, each of the units v_3 and v_2 utilizing its single antenna element continually performs omni-directional simultaneously, the units $\mathcal{A}$ and v_1 collect received powers via a full beam sweeping. The received powers at beams are

(a) The loss curve on the training data from all considered scenarios measuring the difference between predicted and true beams.

(b) The training accuracies obtained at each epochs representing the model performance improves over epochs.
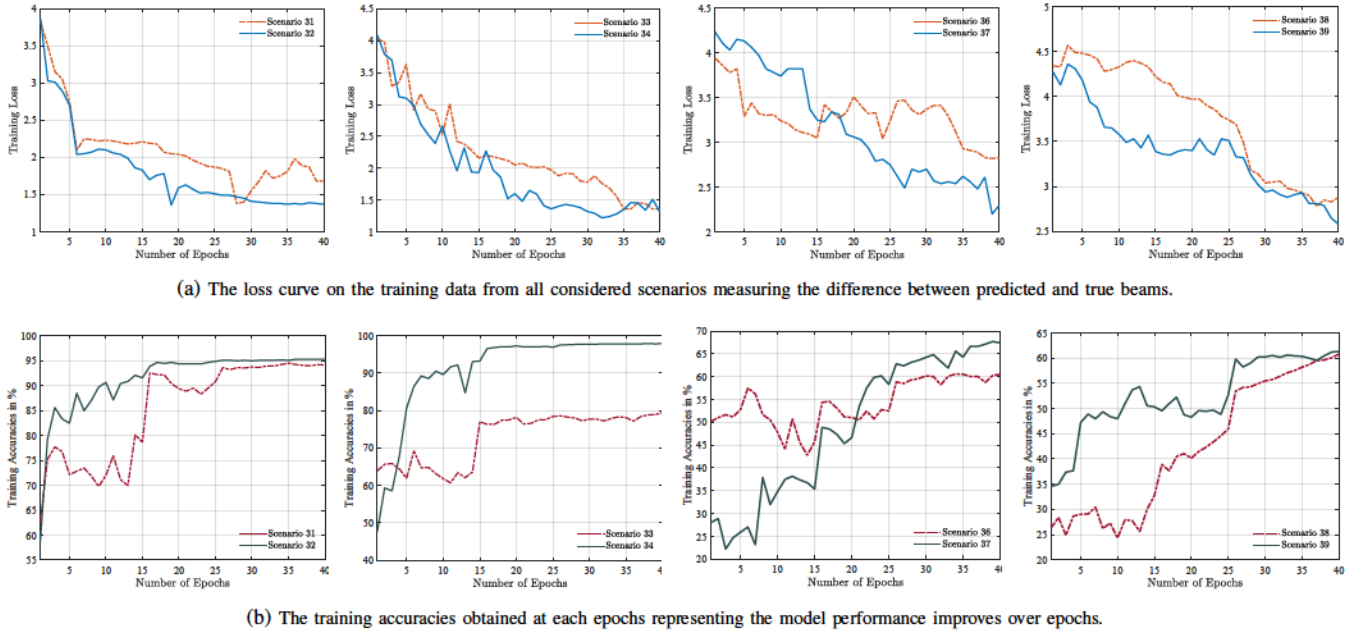
Fig. 4: The results of loss and accuracies while training the proposed model, indicating how the learning performances of the model on training data from all considered scenarios are improving over 40 number of epochs.

represented as beam indices ($\mathcal{I} \in \{1, 2, 3, ..., 64\}$). Each of the units $\mathcal{A}$, v_1, v_2, and v_2 is also equipped with GPS Real Time Kinematics sensors to obtain the vehicle position, i.e., latitude and longitude values in real-time. The time interval between two consecutive samples is 0.1 second corresponding to a rate of 10 samples/second. At each instant, the units $\mathcal{A}$ and v_1 record the synchronous data of location coordinates of units v_3 and v_2 as well as corresponding received powers and optimal beam indices (referred as ground truth indices). Figs. 3(a)-(h) illustrate the visual representation of 100 and 400 location data points for V2I and V2V scenarios, respectively of R_x corresponding to individual ground truth beams associated with maximum received powers for all scenarios. In fact, this indicates how the best beams change as the vehicles' change their respective position.

B. Implementations and Performance Evaluation

A series of experiments are carried out to assess the performance of the designed multi-modal model on the considered real-world datasets. The experiments are conducted on a computing device having Intel Core i7-10875H CPU with 32 GB RAM and NVIDIA GeForce RTX 2080 Super GPU with 8 GB memory. In particular, the deep learning model is developed by PyTorch and CUDA toolkit 11.7. We split the data resources of each scenario into 60%, 20%, and 20% respectively for training, validation, and testing. For the training step, the fixed learning rate and weight decay values of the Adam optimizer are set to 0.01 and 1×10^{-4}, respectively, while the batch size is 16. With these parameters, we then train the model on splitted training samples of each individual scenarios, and the model is trained up to a total of 40 epochs. Figs. 4(a) and (b) show the results of the training loss and accuracies, respectively, obtained at each epoch. The data preprocessing and implementation codes along with detailed instructions are

publicly available to facilitate the reproducibility of this work at GitHub[3].

For better analysis and illustration of the performance evaluation of the proposed solution, we consider following two matrices to measure the performances in terms of testing. Both matrices are averaged over total test samples. Besides, it should also be noted that all average results obtained from this experiments after running 5 times.

• *Top-M Accuracy:* The results of number of correct top-M predictions from the model during test in relation to the total number of predictions can be defined as follows.

$$Acc_{top-M} = \frac{1}{N_{test}} \sum_{t=0}^{N_{test}-1} \frac{|\, I_t^* \cap \hat{I}_t \,|}{|\, \hat{I}_t \,|}, \qquad (12)$$

where, N_{test}, I_t^*, and $\hat{I}_t$ denote the number of test samples, optimal beams at t^{th} time, and top-M beams at t^{th} time, respectively.

• *Received Power Ratio:* The ratio between received power in downlink associated with predicted beams and received power from ground-truth beams is a measure of received power that can be possibly received from the link. It can be defined as follows.

$$R_{\mathcal{P}_t} = \frac{1}{N_{test}} \sum_{t=0}^{N_{test}-1} \frac{\hat{\mathcal{P}}_t}{\mathcal{P}_t^{gt}}, \qquad (13)$$

where, $\hat{\mathcal{P}}_t$ and $\mathcal{P}_t^{gt}$ are the received power from predicted beams and ground-truth beams, respectively, at time instant t.

The results in terms of top-M beam selection accuracies for eight scenarios are depicted in Fig. 5(a) where the top beam

---

[3]https://github.com/mbaqer/V2X-mmWave-Beamforming

(a) Average achieved accuracies on the testing data considering all three sensing modalities

(b) Average achievable received power ratio on the testing data for all considered scenarios.
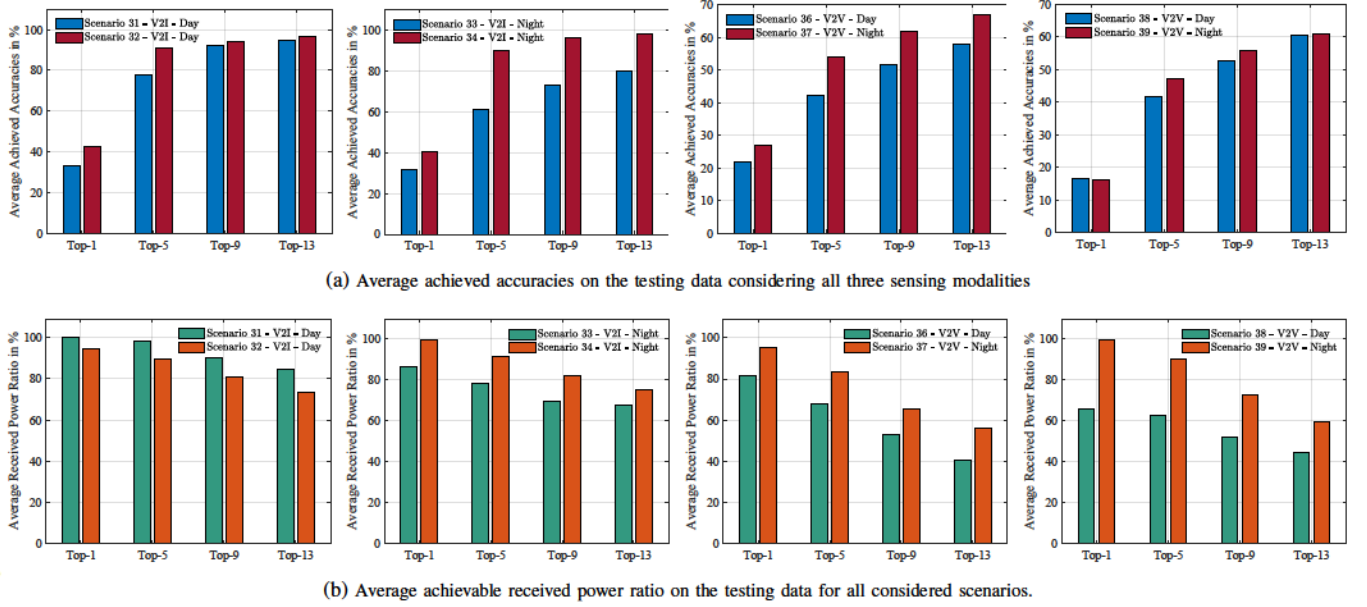
Fig. 5: The performances of average achieved top-M accuracies and received power ratios in percentages tested on the vehicle-to-infrastructure and vehicle-to-vehicle scenarios.

candidates are $M \in \{1, 5, 9, 13\}$. It can be noticed that the accuracies keep on increasing as M until the $M = 13$. For instance, while measuring the top-13 accuracies, it achieves up to 98.19% and 66.96% on V2I and V2V scenarios, respectively. For another, we also depict the results of received power ratio performances in Fig. 5(b), which is our emphasis of this proposed solution. From the results, we can particularly observe that our proposed solution performs very well while also accomplishing sufficient maximum received power from the mmWave communications links as desired. For example, according to the results, with top-1 predicted beams, 99.92% and 98.86% received power ratio can be possibly achieved on average in V2I and V2V scenarios, respectively.

However, even though the aforementioned performance results are obtained from measured data in urban settings during day and night times, addressing the adaptability of our work to suburban or rural scenarios is also necessary for real word applicability. Besides, the various weather conditions might impact the performance of sensing information, and hence play an important role. As such, we suggest that the model should be trained and updated continuously to get better generalization ability along with or without minor performance losses while considering the deployment in real world. Next subsection considers how the proposed solution might be integrated with 5G NR (New Radio) standard, and after integration, how it can contribute to reducing the beam sweeping latency and searching space overheads.

C. Comparison with Exhaustive Search in 5G-NR

To meet the 5G-NR standard specifications [69], [70], the beam selection procedure needs to incorporate a brute-force approach, where the transmitter and receiver nodes exhaustively explore the best beams with the strongest signal strength via sweeping of all beams predefined at the codebooks

of directions. During the initial access, the nodes exchange messages to establish a suitable directional link. In particular, the sweeping process starts with sending a set of synchronizing signal (SS) blocks originating from the transmitter node in a sequential manner. At the same time, the receiver node tunes its antenna arrays along the predefined beam directions until the best beams are located from the signal strengths. The set of SS blocks (each specific beam assigned under each SS block) are grouped into a single SS burst with a length of 5 ms, and repeated on every 20 ms as presented in Fig. 6(a). Considering up to 32 number of SS blocks fit into a single burst, we can express the total time requirements for beam selection process as follows:

$$\mathcal{T}_{nr}(|\mathcal{Q}|) = T_{ssp} \left\lfloor \frac{|\mathcal{Q}| - 1}{32} \right\rfloor + t_{burst}, \quad (14)$$

where, T_{ssp}, t_{burst}, and $|\mathcal{Q}|$ are the duration of a single burst, periodicity, and number of beams, respectively. We can also get a SSB duration, i.e., a single beam sweep duration as $t_{ssb} = t_{burst}/32 = 0.156$ ms. However, the proposed solution seeeks to reduce the number of beams to be explored from $|\mathcal{Q}|$ to identify predicted top-M beam candidates. Then, the time associated with sweeping the predicted beams can be given by:

$$\mathcal{T}_{nr}^{cnn}(M) = T_{ssp} \left\lfloor \frac{M - 1}{32} \right\rfloor + t_{ssb}(1 + (M - 1) \bmod 32). \quad (15)$$

Fig. 6(b) illustrates how 5G-NR defined exhaustive search beamforming works, whereas, Fig. 6(c) represents how our proposed approach integrates with such standard defined existing beam searching approach to find the best beam directions for communications between two units. In particular, the proposed approach enables the receiver base station/road-side
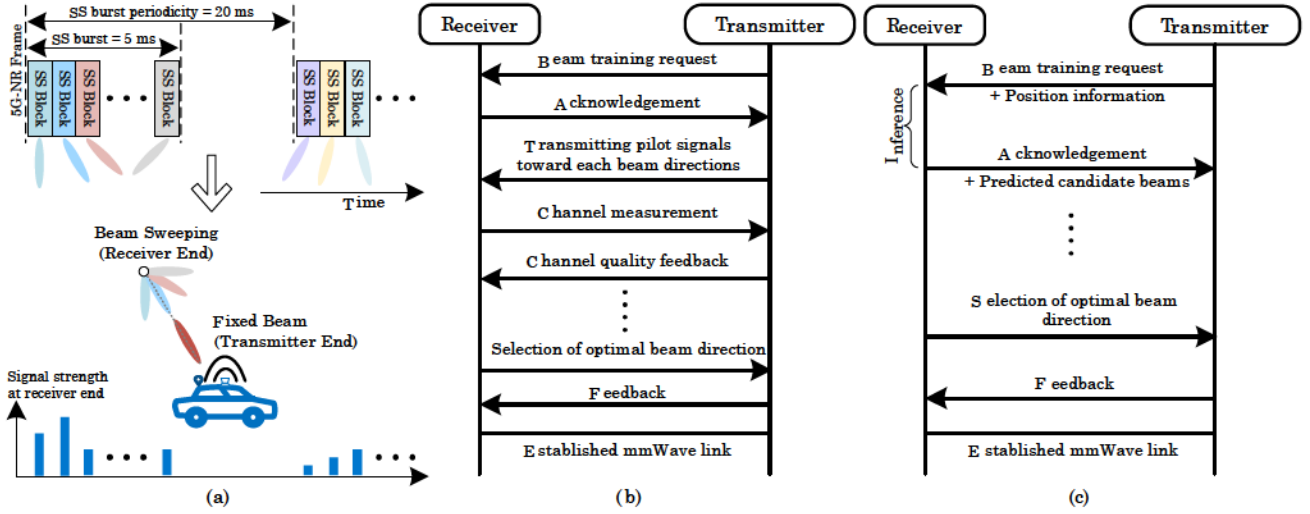
Fig. 6: (a) The frame structure of synchronizing signal (SS) burst within 5G-NR standard (each SS block corresponding to specific beam is represented as different colors) and (b)&(c) the timing diagrams of 5G-NR standardarized exhaustive beamforming and our approach with proposed multi-modal learning.

unit (V2I) or vehicle (V2V) to predict the candidate beams for the connected vehicles who wish to communicate with them. After that, 5G-NR standard based beam selection procedures can be employed, where both receiver and transmitter nodes systematically perform beam sweeping on the candidate beams over a range of pre-defined angular directions to find the most desired optimal directions. It should be noted that the beams are required to adjust according to the channel conditions and vehicular mobility over time. Hence, the procedure essentially involves measuring and giving feedback about the channel quality for each candidate beams. Upon analyzing the channel quality and feedback, the best beam direction can be determined eventually which in turn establish the mmWave communication link.

In Fig. 7, we compare the performances of our proposed approach against standard defined existing beam selection approach in terms of beam sweeping latency and beam searching space overheads. The beam sweeping times are calculated from the Eqs. (14) and (15). As shown, the proposed approach requires up to 2.03 ms to search top-13 candidate beams, a significant improvement over 25 ms needed otherwise for searching all possible 64 beams with existing exhaustive search. On the other hand, the beam search space overheads directly depend on the number of beams to be swept. For example, if we need to explore from top-13 predicted beams, the maximum searching overheads will be 13 number of beams. Since the proposed approach considered in this work provides top-M predicted beams to be explored, it could considerably shorten the beam searching overheads roughly to 20.31% when compared to exhaustive beam searching.

In a sum, we have employed deep learning on out-of-band sensor information for beam selection to avoid unnecessary beam searching overheads by providing a selected number of predicted candidate beams. Such reductions of beam searching space is particularly desirable for vehicular settings since it eventually enables to shorten the communications delays, and

possibly integrate also with 5G-NR standard. For instance, our evaluation results justify that the proposed solution greatly shortens beam sweeping searching and time overheads roughly by 79.67% and 91.89%, respectively for top-13 beams. However, despite the top-1 having minimal overheads, its low accuracy may mislead to identify the accurate best beam directions for communications. We can recall that we can get maximum accuracy from the model is 98.19% for top-13 predicted outcomes. As such, the simple intuition is employing top-13 beams would perform perfectly in predicting the correct beams, thereby ensuring to avoid possible beam misalignment with still very modest overheads. On the other hand, it is worth mentioning that while it takes longer time for training the proposed model, the average inference execution time (per sample) measured on NVIDIA GeForce RTX 2080 GPU is roughly between 0.5 ms and 1.0 ms, whereas the multi-modal data pre-processing times are negligible. This work is thus able to meet real-time requirements and be deployable in real world scenarios.
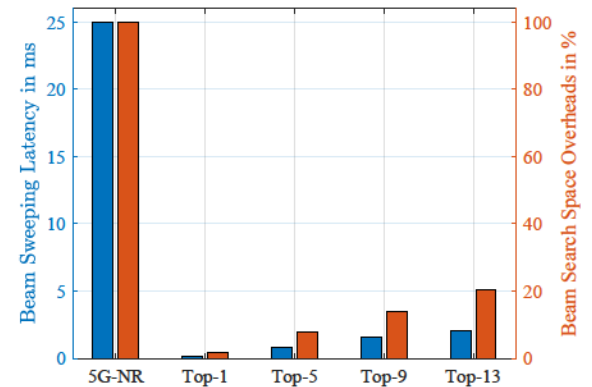

Fig. 7: Performance comparisons of standard defined existing beam selection approach and our proposed approach with respect to beam sweeping latency and beam searching space overheads.

VII. Conclusion

In this paper, we have introduced a 5G-NR compatible beam selection solution to improve on harnessing the downlink received power in 60 GHz mmWave enabled V2I and V2V communications. For that, we have presented a proposed multi-modal feature extractions architecture utilizing deep learning models to learn from the out-of-band multi-modality sensing information and predict the top-M beams (a subset of beams). Subsequently, these predicted beams in turn helps to reduce the beam searching space, thereby addressing the beam searching overheads limitation of mmWave communications. In the end, experiments on real-world measured wireless datasets have shown the proposed solution's effectiveness and applicability, indicating the benefits of exploring multi-modality sensing in mmWave beamforming for connected vehicles. Perhaps, exploring how to incorporate multi-modality sensing and communications in non-line-of-sight (NLoS) scenarios within mmWave vehicular communications could be considered as a future work.

Acknowledgments

References

[1] M. B. Mollah, H. Wang, and H. Fang, "Position aware 60 GHz mmWave beamforming for V2V communications utilizing deep learning," in *ICC 2024 - IEEE International Conference on Communications.* IEEE, 2024, pp. 4711–4716.

[2] M. B. Mollah, H. Wang, M. A. Karim, and H. Fang, "mmWave enabled connected autonomous vehicles: A use case with V2V cooperative perception," *IEEE Network*, vol. 38, no. 6, pp. 485–492, 2024.

[3] B. Xie, S. Chen, S. Zhou, Z. Niu, B. Galkin, and I. Dusparic, "Learning-assisted user scheduling and beamforming for mmWave vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 8, pp. 11 262–11 275, 2024.

[4] M. Feng and M. Krunz, "Online learning-based beamwidth optimization for initial access in millimeter wave cellular networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 1, pp. 231–242, 2025.

[5] C. Lin, D. Tian, X. Duan, J. Zhou, D. Zhao, and D. Cao, "V2VFormer: Vehicle-to-Vehicle cooperative perception with spatial-channel transformer," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 2, pp. 3384–3395, 2024.

[6] H. Ngo, H. Fang, and H. Wang, "Cooperative perception with V2V communication for autonomous vehicles," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 9, pp. 11 122–11 131, 2023.

[7] J. Osa, N. Björsell, I. Val, and M. Mendicute, "Measurement based stochastic channel model for 60 GHz mmWave industrial communications," *IEEE Open Journal of the Industrial Electronics Society*, vol. 4, pp. 603–617, 2023.

[8] S. Sun, T. S. Rappaport, M. Shafi, P. Tang, J. Zhang, and P. J. Smith, "Propagation models and performance evaluation for 5G millimeter-wave bands," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8422–8439, 2018.

[9] Q. Xue, J. Guo, B. Zhou, Y. Xu, Z. Li, and S. Ma, "AI/ML for management in 5G-Advanced: A standardization perspective," *IEEE Vehicular Technology Magazine*, vol. 19, no. 4, pp. 64–72, 2024.

[10] Q. Xue, C. Ji, S. Ma, J. Guo, Y. Xu, Q. Chen, and W. Zhang, "A survey of beam management for mmWave and THz communications towards 6G," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 3, pp. 1520–1559, 2024.

[11] B. Salehi, G. Reus-Muns, D. Roy, Z. Wang, T. Jian, J. Dy, S. Ioannidis, and K. Chowdhury, "Deep learning on multimodal sensor data at the wireless edge for vehicular network," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7639–7655, 2022.

[12] Y. Zhao, X. Zhang, X. Gao, K. Yang, Z. Xiong, and Z. Han, "LSTM-based predictive mmWave beam tracking via sub-6 GHz channels for V2I communications," *IEEE Transactions on Communications*, vol. 72, no. 10, pp. 6254–6270, 2024.

[13] Z. Huang, Z. Wang, and S. Chen, "Sub-6GHz assisted mmWave hybrid beamforming with heterogeneous graph neural network," *IEEE Transactions on Communications*, vol. 72, no. 11, pp. 6917–6928, 2024.

[14] C. He, Z. Zhang, Y. Wan, L. Zhao, H. Lu, and T. Shimizu, "An efficient sub-6 GHz assisted millimeter wave V2X communication scheduler," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 6, pp. 8915–8927, 2024.

[15] X. Xia, F. Wu, C. Yu, Z. Jiang, J. Xu, S.-Y. Tang, Z. Wang, Y. Yao, and W. Hong, "Millimeter-wave and sub-6 GHz aperture-shared antenna and array for mobile terminals accessing 5G/6G-enabled IoT scenarios," *IEEE Internet of Things Journal*, vol. 11, no. 10, pp. 18 808–18 823, 2024.

[16] L. Yashvanth and C. R. Murthy, "On the impact of an IRS on the out-of-band performance in sub-6 GHz & mmWave frequencies," *IEEE Transactions on Communications*, vol. 72, no. 12, pp. 7417–7434, 2024.

[17] K. Vuckovic, M. B. Mashhadi, F. Hejazi, N. Rahnavard, and A. Alkhateeb, "PARAMOUNT: Towards generalizable deep learning for mmWave beam selection using sub-6GHz channel measurements," *IEEE Transactions on Wireless Communications*, vol. 23, no. 5, pp. 5187–5202, 2024.

[18] Q. Zhou, Y. Gong, and A. Nallanathan, "Radar-aided beam selection in MIMO communication systems: A federated transfer learning approach," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 8, pp. 12 172–12 177, 2024.

[19] M. Al-Quraan, A. Zoha, A. Centeno, H. B. Salameh, S. Muhaidat, M. A. Imran, and L. Mohjazi, "Enhancing reliability in federated mmWave networks: A practical and scalable solution using radar-aided dynamic blockage recognition," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 10 146–10 160, 2024.

[20] A. Graff, Y. Chen, N. González-Prelcic, and T. Shimizu, "Deep learning-based link configuration for radar-aided multiuser mmWave vehicle-to-infrastructure communication," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 7454–7468, 2023.

[21] H. Luo, U. Demirhan, and A. Alkhateeb, "Millimeter wave V2V beam tracking using radar: Algorithms and real-world demonstration," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 740–744.

[22] J. Morais and A. Alkhateeb, "Localization in digital twin MIMO networks: A case for massive fingerprinting," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops).* IEEE, 2024, pp. 1–6.

[23] Z. Xing, R. Wang, X. Yuan, and J. Wu, "Location information assisted beamforming design for reconfigurable intelligent surface aided communication systems," *IEEE Transactions on Wireless Communications*, vol. 22, no. 11, pp. 7676–7695, 2023.

[24] S. Rezaie, E. De Carvalho, and C. N. Manchón, "A deep learning approach to location-and orientation-aided 3D beam selection for mmwave communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 11 110–11 124, 2022.

[25] S. Kim, Y. Ahn, D. Park, and B. Shim, "VOMTC: Vision objects for millimeter and terahertz communications," *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 1, pp. 243–257, 2025.

[26] Z. Hua, Y. Lu, G. Pan, K. Gao, D. B. da Costa, and S. Chen, "Computer vision-aided mmWave UAV communication systems," *IEEE Internet of Things Journal*, vol. 10, no. 14, pp. 12 548–12 561, 2023.

[27] W. Xu, F. Gao, X. Tao, J. Zhang, and A. Alkhateeb, "Computer vision aided mmWave beam alignment in V2X communications," *IEEE Transactions on Wireless Communications*, vol. 22, no. 4, pp. 2699–2714, 2023.

[28] N. Huang, C. Dou, Y. Wu, L. Qian, S. Zhou, and R. Lu, "Image analysis oriented integrated sensing and communication via intelligent reflecting surface," *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 1, pp. 274–287, 2025.

[29] M. Nerini and B. Clerckx, "Overhead-free blockage detection and precoding through physics-based graph neural networks: LIDAR data meets ray tracing," *IEEE Wireless Communications Letters*, vol. 12, no. 3, pp. 565–569, 2023.

[30] 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, 2022.

[31] 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.

[32] 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.

[33] D. Roy, B. Salehi, S. Banou, S. Mohanti, G. Reus-Muns, M. Belgiovine, P. Ganesh, C. Dick, and K. Chowdhury, "Going beyond RF: A survey on how AI-enabled multimodal beamforming will shape the nextG standard," *Computer Networks*, vol. 228, p. 109729, 2023.

[34] J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Position-aided beam prediction in the real world: How useful GPS locations actually are?" in *Proceedings of IEEE International Conference on Communications (ICC)*. IEEE, 2023, pp. 1824–1829.

[35] T. N. Ha, D. Romero, and R. López-Valcarce, "Radio maps for beam alignment in mmWave communications with location uncertainty," in *2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)*, 2024, pp. 1–7.

[36] F. Meng, S. Liu, S. Gao, Y. Yu, C. Zhang, Y. Huang, and Z. Lu, "TDoA positioning with data-driven LoS inference in mmWave MIMO communications," *Signal Processing*, vol. 220, p. 109447, 2024.

[37] M. Mattos, A. Rodrigues, R. Meireles, and A. Aguiar, "Geolocation-based sector selection for Vehicle-to-Infrastructure 802.11 ad communication," *Computer Communications*, vol. 193, pp. 224–233, 2022.

[38] S. Khosravi, H. S. Ghadikolaei, J. Zander, and M. Petrova, "Location-aided beamforming in mobile millimeter-wave networks," in *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*. IEEE, 2022, pp. 1–7.

[39] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath, "Inverse multipath fingerprinting for millimeter wave V2I beam alignment," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4042–4058, 2017.

[40] Y. Ahn, J. Kim, S. Kim, S. Kim, and B. Shim, "Sensing and computer vision-aided mobility management for 6G millimeter and terahertz communication systems," *IEEE Transactions on Communications*, vol. 72, no. 10, pp. 6044–6058, 2024.

[41] M. M. Al-Quraan, A. R. Khan, L. Mohjazi, A. Centeno, A. Zoha, and M. A. Imran, "Intelligent beam blockage prediction for seamless connectivity in vision-aided next-generation wireless networks," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 1937–1948, 2023.

[42] T. Xiang, J. Gu, Y. Wang, Y. Gao, and X. Zhang, "Computer vision aided beamforming fused with limited feedback," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 1, pp. 376–388, 2023.

[43] T. Osman, G. Charan, and A. Alkhateeb, "Vehicle cameras guide mmwave beams: Approach and real-world V2V demonstration," in *2023 57th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2023, pp. 225–232.

[44] G. Charan and A. Alkhateeb, "User identification: A key enabler for multi-user vision-aided communications," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 472–488, 2024.

[45] B. Lin, F. Gao, Y. Zhang, C. Pan, and G. Liu, "Multi-camera views based beam searching and BS selection with reduced training overhead," *IEEE Transactions on Communications*, vol. 72, no. 5, pp. 2793–2805, 2024.

[46] S. Imran, G. Charan, and A. Alkhateeb, "Environment semantic aided communication: A real world demonstration for beam prediction," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2023, pp. 48–53.

[47] ——, "Environment semantic communication: Enabling distributed sensing aided networks," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 7767–7786, 2024.

[48] F. Wen, W. Xu, F. Gao, C. Pan, and G. Liu, "Vision aided environment semantics extraction and its application in mmWave beam selection," *IEEE Communications Letters*, vol. 27, no. 7, pp. 1894–1898, 2023.

[49] 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.

[50] S. Wu, C. Chakrabarti, and A. Alkhateeb, "Proactively predicting dynamic 6G link blockages using LiDAR and in-band signatures," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 392–412, 2023.

[51] S. Ohta, T. Nishio, R. Kudo, K. Takahashi, and H. Nagata, "Point cloud-based proactive link quality prediction for millimeter-wave communications," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 1, pp. 258–276, 2023.

[52] S. Pradhan, D. Roy, B. Salehi, and K. Chowdhury, "COPILOT: Cooperative perception using lidar for handoffs between road side units," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*. IEEE, 2024, pp. 1–10.

[53] 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)*. IEEE, 2022, pp. 2727–2731.

[54] A. D. Raha, K. Kim, A. Adhikary, M. Gain, and C. S. Hong, "Advancing ultra-reliable 6G: Transformer and semantic localization empowered robust beamforming in millimeter-wave communications," *arXiv preprint arXiv:2406.02000*, 2024.

[55] 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.

[56] Y. Bian, J. Yang, S. Xia, and S. Jin, "3-D LiDAR and GPS aided beam tracking in millimeter wave vehicular communications," *IEEE Wireless Communications Letters*, vol. 13, no. 12, pp. 3290–3294, 2024.

[57] X. Niu, Y. Fu, M. Dong, N. Cheng, and C. Li, "Multi-source low redundancy data-aided beam prediction for V2I communication," in *2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall)*. IEEE, 2023, pp. 1–5.

[58] K. Patel and R. W. Heath Jr, "Harnessing multimodal sensing for multiuser beamforming in mmWave systems," *IEEE Transactions on Wireless Communications*, vol. 23, no. 12, pp. 18 725–18 739, 2024.

[59] B. Salehi, D. Roy, J. Gu, C. Dick, and K. Chowdhury, "FLASH-and-Prune: Federated learning for automated selection of high-band mmWave sectors using model pruning," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 11 655–11 669, 2024.

[60] B. Salehi, U. Demir, D. Roy, S. Pradhan, J. Dy, S. Ioannidis, and K. Chowdhury, "Multiverse at the edge: interacting real world and digital twins for wireless beamforming," *IEEE/ACM Transactions on Networking*, vol. 32, no. 4, pp. 3092–3110, 2024.

[61] "Wireless InSite 3D wireless propagation software)," Accessed: June 2024, Available: https://www.remcom.com/wireless-insite-propagation-software.

[62] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[63] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[64] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[65] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "Deepsense 6G: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 122–128, 2023.

[66] J. Morais, G. Charan, N. Srinivas, and A. Alkhateeb, "DeepSense-V2V: A Vehicle-to-Vehicle multi-modal sensing, localization, and communications dataset," *arXiv preprint arXiv:2406.17908*, 2024.

[67] "Raymobtime dataset," Accessed: June 2024, Available: https://www.lasse.ufpa.br/raymobtime.

[68] "FLASH dataset," Accessed: June 2024, Available: https://genesys-lab.org/multimodal-fusion-nextg-v2x-communications.

[69] B. Salehi, D. Roy, T. Jian, C. Dick, S. Ioannidis, and K. Chowdhury, "Omni-CNN: A modality-agnostic neural network for mmwave beam selection," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 6, pp. 8169–8183, 2024.

[70] R. Tuninato, D. G. Riviello, R. Garello, B. Melis, and R. Fantini, "A comprehensive study on the synchronization procedure in 5G NR with 3GPP-compliant link-level simulator," *EURASIP Journal on Wireless Communications and Networking*, vol. 2023, no. 1, p. 111, 2023.

Muhammad Baqer Mollah is a PhD candidate in the Department of Electrical and Computer Engineering at the University of Massachusetts Dartmouth. Before joining at University of Massachusetts Dartmouth, he has been working at Nanyang Technological University (NTU) as a full-time Research Associate. His research interests include advanced communications, security, and edge intelligence techniques for Internet of Things (IoT) and connected vehicles. He has a M.Sc. in Computer Science and B.Sc. in Electrical and Electronic Engineering from Jahangirnagar University, Dhaka and International Islamic University Chittagong, Bangladesh, respectively.

Honggang Wang is the founding department chair and Professor of the Department of Graduate Computer Science and Engineering at Katz School of Science and Health, Yeshiva University. He was a Professor in the Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth. He was the "Scholar of The Year" awardee in 2016, the highest research recognition at University of Massachusetts Dartmouth. He has served as the Editor-in-Chief of IEEE Internet of Things Journal (2020 - 2022). He has won several prestigious best paper awards throughout his career. He was the past Chair of the IEEE Multimedia Communications Technical Committee (2018-2020) and the IEEE eHealth Technical Committee Chair (2020-2021). His recently focused research topics involve AI and Internet of Things, mmWave communications, connected vehicles, smart health, and cyber security. He has a PhD in Computer Engineering from University of Nebraska - Lincoln.

Mohammad Ataul Karim is a Professor in the Department of Electrical and Computer Engineering at the University of Massachusetts Dartmouth. His areas of research encompass optical computing, pattern/target recognition, computer vision, and Internet of Things. He is the former Provost, Executive Vice Chancellor for Academic Affairs, and Chief Operating Officer at University of Massachusetts Dartmouth. He is author/editor of 19 books, over 365 research papers, 15 book chapters, and 3 US patents. He serves on the editorial boards of Optics and Laser Technology, International J. of Modern Physics B, and Modern Physics Letters B. He received his PhD in electrical engineering degree from the University of Alabama. Karim is a fellow also of Optica, Society of Photo-Instrumentation Engineers, Institute of Physics, Institution of Engineering and Technology, Bangladesh Academy of Sciences, and Asia-Pacific Artificial Intelligence Association.

Hua Fang is a Professor in the Department of Computer and Information Science at the University of Massachusetts Dartmouth. She has been leading Computational Statistics and Data Science Lab which aims at developing computational methods, tools and systems for broader health and behavior studies by integrating and advancing the theories and applications of computational statistics, which is a computational science at the interface of statistics and computer science. Her current research interests include real-time machine or statistical learning, and visual analytics of wearable biosensor data streams, broadly in E-/M-/Digital/Tele-/Virtual/Precision health, and Internet of Things.