Deep Learning on Multimodal Sensor Data at the Wireless Edge for Vehicular Network

Batool Salehi, Guillem Reus-Muns, Debashri Roy, Zifeng Wang, *Graduate Student Member, IEEE*, Tong Jian, Jennifer Dy, *Member, IEEE*, Stratis Ioannidis, and Kaushik Chowdhury, *Senior Member, IEEE*

***Abstract*—Beam selection for millimeter-wave links in a vehicular scenario is a challenging problem, as an exhaustive search among all candidate beam pairs cannot be assuredly completed within short contact times. We solve this problem via a novel expediting beam selection by leveraging multimodal data collected from sensors like LiDAR, camera images, and GPS. We propose individual modality and distributed fusion-based deep learning (F-DL) architectures that can execute locally as well as at a mobile edge computing center (MEC), with a study on associated tradeoffs. We also formulate and solve an optimization problem that considers practical beam-searching, MEC processing and sensor-to-MEC data delivery latency overheads for determining the output dimensions of the above F-DL architectures. Results from extensive evaluations conducted on publicly available synthetic and home-grown real-world datasets reveal 95% and 96% improvement in beam selection speed over classical RF-only beam sweeping, respectively. F-DL also outperforms the state-of-the-art techniques by 20-22% in predicting top-10 best beam pairs.**

***Index Terms*—Beam selection, distributed inference, 5G, fusion, mmWave, multimodal data.**

I. Introduction

EMERGING vehicular systems are equipped with a variety of sensors that generate vast amounts of data and require *multi-Gbps* transmission rates [1]. These sensor inputs may be needed for safety-critical vehicle operation as well as for gaining situational awareness while in motion, which needs to be timely processed at a mobile edge computing (MEC) center to generate driving directives. Such a large data transfer volume at short contact times can quickly saturate the sub-6 GHz band. Thus, the millimeter-wave (mmWave) band is widely considered as the ideal candidate for vehicle-to-everything (V2X) communications [2], given the promise of 2 GHz wide channels and vast under-utilized spectrum resources in the 57–72 GHz band. However, transmission in the mmWave band has associated challenges related to severe attenuation and penetration loss.

Manuscript received July 10, 2021; revised November 29, 2021 and February 20, 2022; accepted April 17, 2022. Date of publication April 27, 2022; date of current version July 18, 2022. This work was supported in part by the National Science Foundation under Grants CCF-1937500 and CNS-2112471, in part by the Roux Institute, and in part by the the Harold Alfond Foundation. The review of this article was coordinated by Dr. Ming Li. *(Corresponding author: Kaushik Chowdhury.)*

The authors are with the Institute for the Wireless Internet of Things, Electrical and Computer Engineering Department, Northeastern University, Boston, MA 02115 USA (e-mail: bsalehihikouei@ece.neu.edu; greusmuns@ece.neu.edu; droy@ece.neu.edu; zifengwang@ece.neu.edu; jian@ece.neu.edu; jdy@ece.neu.edu; ioannidis@ece.neu.edu; krc@ece.neu.edu).

Digital Object Identifier 10.1109/TVT.2022.3170733

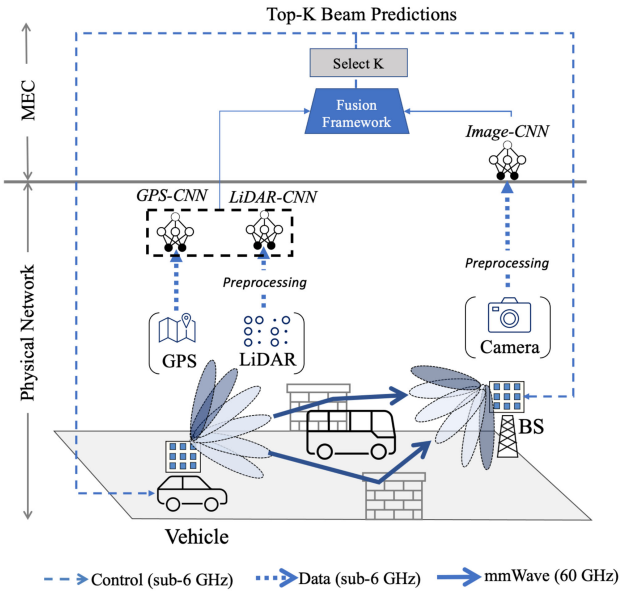


Fig. 1. Our fusion pipeline exploits GPS, camera and LiDAR sensor data to restrict the beam selection to top-K beam pairs.

Phased arrays with directional beamforming can compensate these issues by focusing RF energy at the receiver [3]. Hence, in the so called *beam selection* process, the nodes on either end of the link attempt to converge to the optimal *beam pairs*, where each beam pair is a tuple of transmitter and receiver beam indices, by mutually exploring the available space uniformly partitioned into discrete sectors [4]. However, exploring all possible beam directions in the existing IEEE 802.11ad [5] and 5G New Radio (5G-NR) [6] standards can consume up to tens of milliseconds and must be repeated constantly during vehicular mobility [7]. To address this problem, we propose to exploit the side out-of-band information to restrict the searching to a subset of most likely beam pair candidates. As shown in Table I, reducing the number of beam pairs from 60 to 30 significantly decreases the beam selection overhead by 50% and 80% for IEEE 802.11ad and 5G-NR standards, respectively.

A. Use of Sensors to Aid the Beam Selection

Due to the directional transmissions at mmWave band, the beam selection process can be interpreted as locating the paired user or detecting the strongest reflection in the case of line of

TABLE I
THE REDUCTION IN BEAM SELECTION TIME WHILE REDUCING THE BEAM SEARCH SPACE FROM 60 TO 30 BEAM PAIRS

Standard	Time (ms)	
	30 beam pairs	60 beam pairs
802.11ad	9.09	18.18
5G-NR	4.68	24.37

sight (LOS) and non-line of sight (NLOS) path, respectively. Hence, the location of the transmitter, receiver, and potential obstacles are the key factors in beam initialization. Interestingly, this information is also embedded in the situational state of the environment that can be acquired through monitoring sensor devices. Fig. 1 shows our scenario of interest with a moving vehicle and a road-side base station (BS) attempting to find the best beam pair with multiple reflectors and blocking objects. We assume the state of the environment is captured by a combination of GPS (Global Positioning System) and LiDAR (Light Detection and Ranging), which provides a 3-D representation of the surroundings, sensors in the moving vehicle, and a camera at the BS. We use a sub-6 GHz data channel for exchanging this sensor data between the vehicle and MEC. We then propose to use these non-RF sensor data to suggest a subset of "top-K" beam pairs and speed up the beam selection, consequently. The candidate set of selected beam pairs is communicated to both the BS and the vehicle over the sub-6 GHz control channel. After this, both the vehicle and the BS execute the standards-defined beam-searching algorithms, but only on the subset of top-K suggested beam pairs.

The Yole Dévelopment report anticipates that the global market for GPS, radar, cameras, and LiDARs will increase from \$67.14 in 2020 to \$159.6 in 2025 [8]. With the widespread of IoT devices, multiple sensors are now available as standard installations for the majority of electronic devices as well as fixed roadside infrastructures [9], [10]. LiDAR sensors are an indispensable part of modern vehicles that are used for either automated driving or collision avoidance [11]. The GPS data are regularly collected and transmitted as part of basic safety messages frame in V2X applications [12], and surveillance cameras have been in use for decades with the growth of smart cities [13].

B. Deep Learning on Multimodal Sensor Data

While using sensor data for out-of-band beam selection is an exciting new approach there some challenges that need to be addressed. First, since the physical environment influences signal propagation in ways that are hard to computationally model in real time, hand engineering features extracted from such sensor data that could be discriminative is infeasible, as there could be a vast multitude of reasons impacting the signal propagation. Second, a systematic approach is required to properly join the information from sensor modalities with different properties to predict the optimality of each beam pair. Note that while the beam pair can be inferred through basic geometry under ideal LOS conditions, such an approach fares poorly in scenarios with multiple reflections, such as in NLOS situations. Third, since the sensors are not all available at one site, both on the vehicle and BS, the secondary channels are required to maintain the connectivity between the vehicle and MEC. The communication constraints in these secondary channels need to be fully accounted for: the relaying cost of data exchange, especially massive LiDAR point cloud, might undermine the performance with respect to end-to-end latency. Finally, the beam search dimension K is a control parameter that needs to set prior to starting the beam-searching process. Hence, an algorithm is required to select the appropriate K to fully determine the system design.

Our approach directly addresses these challenges. First, we design a fusion-based deep learning (F-DL) framework operating on all these different modalities to predict a subset of top-K beam pairs that includes the globally optimal solution with high probability. Additionally, we adopt a distributed inference scheme to compress the raw data into high level extracted features at the vehicle to reduce the overhead on the wireless backchannel, accounting for end-to-end latency in the selection of the optimal beam. Finally, we take into account the prediction from our proposed F-DL framework along with mmWave channel efficiency to properly adjust the beam search space K, on a case-by-case basis.

C. Summary of Contributions

Our main contributions are as follows:

1) We design deep learning architectures that predict the set of top-K beam pairs using non-RF sensor data such as GPS, camera, and LiDAR, wherein the processing steps are split between the source sensor and the MEC. We validate the improvement achieved by fusing available modalities versus unimodal data on a simulation as well as a home-grown real-world datatset. Our results show that fusion improves the prediction accuracy by 3.32–43.9%. The proposed fusion network exhibits 20–22% improvement in top-10 accuracy with respect to the state-of-the-art techniques.
2) We formulate an optimization problem to appropriately select the set of K candidate beam pairs, which takes into account mmWave channel efficiency while trying to maximizing the alignment probability, i.e. the case where the optimum beam pair is included within the suggested subset. Thus, the control variable K is not arbitrarily chosen, but tightly coupled to scenario constraints.
3) We rigorously analyze the end-to-end latency of our proposed non-RF beam selection method and compare it with the state-of-the-art standard for mmWave communication, namely 5G-NR and demonstrate that the beam selection time decreases by 95–96% on average while maintaining 97.95% of the throughput, considering all the overhead of control/data signaling for both approaches.

II. RELATED WORK

Leveraging out-of-band data, both in RF and non-RF domains, can speed up the beam selection. RF-based out-of-band beam

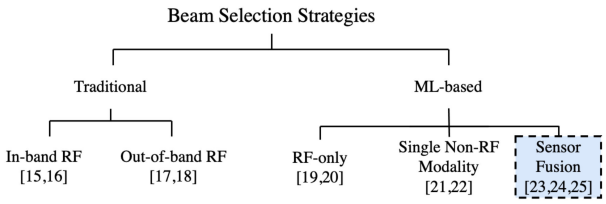


Fig. 2. Deterministic and ML-aided beam selection strategies.

selection is possible via simultaneous multi-band channel measurements, when there exists a mapping between mmWave and the channel state information (CSI) from the another band [14]. However, this method does not support simultaneous beamforming at both the transmitter and receiver ends. As opposed to the RF-only approach, non-RF out-of-band beam selection leverages data from different sensors and generates a mutual decision for both transmitter and receiver. Fig. 2 summarises the emphasis of this paper and different beam selection strategies.

A. Traditional

1) In-Band RF: Yang *et al.* [15] adopt a hierarchical search strategy where the mmWave channel is first tested with comparatively wider beams by using a reduced number of antenna elements. The beam width is then narrows until the best beam is obtained. Wang *et al.* [16] show that mmWave links preserve sparsity even across locations in mobile V2X scenarios. Hence, they utilize the angle of departure (AoD) to search for beams only within this range, thereby reducing beam selection overhead.

2) Out-of-Band RF: Steering with eyes closed [17] exploits omni-directional transmissions within the legacy 2.4/5GHz band to infer the LOS direction between the communicating devices to speed up the mmWave beam selection. GonzÃlez-Prelcic *et al.* [18] exploit the side information derived from RAdio Detection And Ranging (RADAR) data to adapt the beams in a vehicle to infrastructure network, where a compressive covariance estimation approach is used to establish a mapping between RADAR and mmWave bands.

B. Ml-Based

1) Rf-Only: He *et al.* [19] design a deep learning based channel estimation approach using iterative signal recovery, wherein the channel matrix is regarded as a noisy 2D natural image. Learnt denoising-based approximate message passing (LDAMP) neural networks are applied on the input for channel estimation. Hashemi *et al.* [20] model the mmWave beam selection as a MAB (Multi-armed bandit) and use the reinforcement learning to maximize the directivity gain (i.e., received energy) of the beam alignment policy.

2) ML Using Single non-RF Modality: Va *et al.* [21] consider a setting where the location of all vehicles on the road, including the target receiver, is used as input to a machine learning algorithm to infer the best beam configuration. Vision-aided mmWave beam tracking in [22] models a dynamic outdoor mmWave communication setting where the sequence of previous beams and visual images are used to predict future best beam pairs.

3) ML With Sensor Fusion: The proposed setting by Klautau *et al.* [23] and Dias *et al.* [24] comes closest to ours with GPS and LiDAR being used as the side information for LOS detection and also reducing the overhead in a vehicular setting. On the other hand, Muns *et al.* [25] use GPS and camera images to speed up the beam selection with a focus on designing preprocessing step for images and fusion scheme.

The state-of-the-art does not consider the deep learning based fusion for more than two non-RF modalities to fully exploit the latent features within the data. The GPS coordinates are only used in the preprocessing pipeline to identify the target receiver. There also has not been any effort to decouple the expert knowledge for dynamically reducing the beam search space depending on specific user constraints. Our proposed method exploits a customized deep learning fusion approach that is carefully designed to maximize the beam selection accuracy. Moreover, completed by an algorithm that automatically chooses a dynamic subset of beam pairs, our method can run end-to-end without any hand engineering.

TABLE II
NOTATION SUMMARY

Notation	Description
C_{Tx}	The codebook of transmitter with M beams
C_{Rx}	The codebook of receiver with N beams
$\mathcal{B}$	Set of all possible beam pairs
$y_{(t_m,r_n)}$	Normalized signal power for beam pair (t_m, r_n)
$\mathbb{H}$	Channel matrix
w_{t_m}	Beam weight of codebook element t_m at Tx
w_{r_n}	Beam weight of codebook element r_n at Rx
(t^*, r^*)	Optimum beam pair
$\mathcal{B}_K$	A subset of K beam pairs $\mathcal{B}_K \subseteq \mathcal{B}$
$X_{\mathtt{C}}, X_{\mathtt{L}}, X_{\mathtt{I}}$	Input samples from GPS, LiDAR and image
$N_t, N_t^{'}$	Number of train and test samples
$f^{\mathtt{C}}_{\theta^{\mathtt{C}}}, f^{\mathtt{L}}_{\theta^{\mathtt{L}}}, f^{\mathtt{I}}_{\theta^{\mathtt{I}}}$	Feature extractors of GPS, LiDAR and image
$\mathbf{z}_{\mathtt{C}}, \mathbf{z}_{\mathtt{L}}, \mathbf{z}_{\mathtt{I}}$	Latent embedding of GPS, LiDAR and image
$f^{\mathtt{F}}_{\theta^F}$	Fusion network
$\mathbf{z}$	Concatenated features of GPS, LiDAR and image
$\mathbf{s}$	Softmax score for all beam pairs
$c_K(\mathbf{s})$	Sum of the K largest scores for softmax score $\mathbf{s}$
$p(K;\mathbf{s})$	Probability of inclusion for softmax score $\mathbf{s}$
$\mathbf{s}_I$	Scores of a sample drawn from the training set
(t_I^*, r_I^*)	Optimum beam pair of sample $\mathbf{s}_I$ from training set
$\mu(K)$	Latency as a function of the number of beam pairs
$T_{bs}^{df}(K)$	End-to-end latency of F-DL method
T_{total}	Total time for which a certain beam pair is valid
α	Control parameter between probability of inclusion and latency
$Tp, Tssb$	Periodicity and duration of SS bursts
R_T	Throughput ratio

III. SYSTEM MODEL AND OVERVIEW

In this section, we first review classical beam selection and discuss it's limitations. We then propose to use non-RF data from multiple sensors to facilitate– and accelerate–beam selection. Table II summarizes our notation.

A. Beam Selection Problem Formulation

We consider an analog beamforming scheme with fixed size codebooks at transmitter and receiver radios as:

$$C_{Tx} = \{t_1, \ldots, t_M\}, C_{Rx} = \{r_1, \ldots, r_N\}, \tag{1}$$

where M, N are the number of transmitter and receiver codebook elements, respectively. Each element of the codebook represents a particular beam orientation that can be utilized by the radio. Thus, the set of all possible beam pairs $\mathcal{B}$ is:

$$\mathcal{B} = \{(t_m, r_n) | t_m \in C_{Tx}, r_n \in C_{Rx}\}, \tag{2}$$

with $|\mathcal{B}| = M \times N$. For a specific beam pair (t_m, r_n), the normalized signal power is obtained as:

$$y_{(t_m, r_n)} = |w_{t_m}^H \mathbb{H} w_{r_n}|^2, \tag{3}$$

where $\mathbb{H} \in \mathbb{R}^{M \times N}$ is the channel matrix and H is the conjugate transpose operator. The weights w_{t_m} and w_{r_n} indicate the corresponding beam weight vectors associated with the codebook element t_m and r_n, respectively ($|w_{t_m}| = M, |w_{r_n}| = N$). The goal of the beam selection process is to identify the best beam configuration, (t^*, r^*), that maximizes the normalized signal power, given by:

$$(t^*, r^*) = \underset{1 \le m \le M, 1 \le n \le N}{\arg\max} y_{(t_m, r_n)}. \tag{4}$$

In classical beam selection, such as the approach defined in the IEEE 802.11ad [26] and 5G-NR [27] standards, the transmitter and receiver sweep all beam pairs $(t_m, r_n) \in \mathcal{B}$ sequentially in order to select the best beam pair.

B. Subset Selection

While exhaustive searching through all candidate options ensures the beam alignment, the typical time to complete the entire procedure is in the order of $\sim$10 ms for IEEE 802.11ad [5] and $\sim$5 ms for 5G-NR [6] with only 30 beam pairs, respectively. To address this, we propose a beam selection framework that uses out-of-band multimodal data to identify a subset of candidate beams, which are subsequently swept to select the one that maximizes the normalized signal power [28]. More specifically, the key algorithmic component of our system amounts to proposing a means for identifying a subset $\mathcal{B}_K \subseteq \mathcal{B}$ of K beam pairs such that $(t^*, r^*) \in \mathcal{B}_K$ with high probability. Formally, assuming that we have a probability distribution for the optimal pair (t^*, r^*), we wish to find:

$$\mathcal{B}_K = \underset{A \subseteq \mathcal{B}, |A| = K}{\arg\max} \mathbf{P}((t^*, r^*) \in A). \tag{5}$$

Having obtained $\mathcal{B}_K$, we then restrict the search for the optimal pair to this set. Our solution uses a neural network to leverage out-of-band data to determine the probability distribution $\mathbf{P}$. Parameter K establishes a trade-off between throughput performance, obtained by the best beam in $\mathcal{B}_K$, and latency, as a larger K results in more processing time to search through the candidate options. Thus, our end-to-end design includes a means for appropriately determining K, where the boundary condition of $K = 1$ represents selecting the optimal beam pair. Overall, this auxiliary parameter K enables the users to adjust the system according to their specific constraints on establishing a low-latency or ultra-reliable communication. Moreover, it gives the flexibility to analyze the adjacent beam patterns with relatively closer performance or irregular radiation patterns under NLOS conditions.

C. System Overview

Overall our framework consists of three main components.

1) *Data Preprocessing:* For the collected data to be effective, it is crucial to mark the transmitter, target receiver, and blocking objects. Thus, we exploit the prepossessing step described in Sec. IV for image and LiDAR.
2) *Beam Prediction using Fusion-based Deep Learning:* Given the multimodal sensor data, we design a F-DL architecture that predicts the optimality of each beam pair. Our approach consists of custom-designed feature extractors for each sensor modality, followed by a fusion network that joins the information for the final prediction. Our proposed fusion approach is presented in Sec. V.
3) *Top-K Beam Pair Construction:* We select, the beam search space dimension, K by defining an optimization problem (see Sec. VI) that takes into account the mmWave channel efficiency and probability of including the globally optimum beam pair.

In summary, our proposed beam selection approach runs in four steps end-to-end. First, the sensors at the vehicle collect GPS and LiDAR data, and the camera at the BS captures an image. The collected raw data is then *preprocessed on site*. Second, having the feature extractors of GPS and LiDAR being deployed at the vehicle, the high level features are generated and *shared with the MEC over the sub-6 GHz data channel*. This approach avoids sharing unnecessary amounts of data and helps mitigating potential privacy concerns. The high-level features of the image are generated in parallel. Third, given the extracted features of all three modalities at MEC, our method suggest a set of top-K candidates for sweeping. The subset of K beam pair is *shared with the vehicle over the sub-6 GHz control channel*. Finally, the *beam sweeping runs at mmWave band (60 GHz)* in a reduced search space of selected top-K candidates to select the best beam pair and establish the link.

D. Sensor Modalities

The details of the three sensor modalities are given below:

1) *GPS:* This sensor generates readings in the decimal degrees (DD), where the separation between each line of latitude or longitude is expressed as a float number with 5 digit precision and pinpoints the location on the earth's surface. We do not assume any satellite link outages due to terrain or man-made structures.
2) *Image:* This sensor captures still RBG images of the environment. Although images allow comprehensive environmental assessment, they are impacted by low-light conditions and obstructions (such as a different vehicle in the LOS path)
3) *LiDAR:* This sensor generates a 3-D representation of the environment by emitting pulsed laser beams. The distance

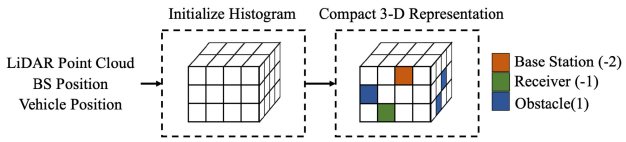


Fig. 3. The LiDAR preprocessing pipeline.

of each object from the origin (i.e., the sensor location) is calculated based on reflection times. The raw LiDAR point clouds are data intensive ($\sim$1.5 Mb for sparse settings), necessitating processing at the vehicle.

IV. Data Preprocessing

In this section, we describe our preprocessing pipeline for image and LiDAR.

A. Processing Images

The raw images collected at the BS provide a snapshot of the present objects in the scene. In this case, it is crucial to detect the region of the target receiver among other vehicles that correspond to the blocking objects. Hence, we design a preprocessing step as follows. First, we employ a multi-object detection approach that enables us to flexibly distinguish the spatial boundaries of different vehicle types in the same frame. Second, given the type of target vehicle, we separate the region of the target receiver and blocking vehicles. On the other hand, the background with static walls and buildings is invariant over different scenes and consequently does not affect the decision and can be further removed. In summary, our approach (i) detects multiple vehicle types present in the same scene, (ii) separates the receiver and obstacle regions, and (i) removes the static background. Since the focus of this paper is not directly on image processing, we include details of our custom designed approach in Appendix A. The output of this image preprocessing step is the *bit map* of the raw input camera image, and it serves as the input to our fusion pipeline.

B. Processing LiDAR Point Clouds

The raw LiDAR point cloud is a collection of (x, y, z) points that correspond to the location of detected objects in the environment. Directly exploiting the raw point cloud (with varying number of points depending on traffic density) not only comes with huge computational cost but also raises ML architecture design challenges as the input to a neural network must be preferably fixed in size. Hence, we use a preprocessing step as shown in Fig. 3 first proposed in [23] that considers a limited spatial zone for each axis. This space corresponds to coverage range of BS and is denoted as $(X_{\min}, X_{\max})$, $(Y_{\min}, Y_{\max})$, and $(Z_{\min}, Z_{\max})$. Then, we construct a 3-D histogram that corresponds to a quantized 3-D representation of the space. The histogram bin size along the three spatial dimensions (b_x, b_y, b_z) can be set based on desired resolution. The LiDAR point clouds lie in the corresponding bins of the histogram based on their location. Since the BS is fixed in our setting, it always occupies the same cell of the histogram with indicator (-2). The corresponding cell of the target receiver is also acquired with GPS data and indicated with (-1). The remaining elements are mapped to the corresponding histogram elements with (1), which implies the presence of obstacles. This leads to a compact 3-D representation of the environment that we use as input for our pipeline.

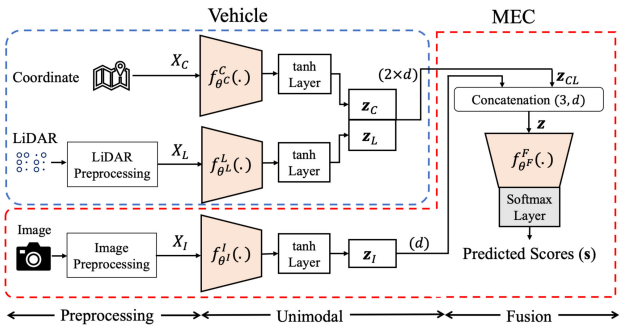


Fig. 4. Proposed fusion framework. In the training phase, the pipeline is trained offline, and during the distributed inference, the trained model is disseminated over the system.

V. Beam Prediction Using Fusion-Based Deep Learning

In the second step of our proposed framework, we design a multimodal data fusion pipeline to combine the available sensing modalities together and predict the optimality of each beam pair. First, we describe the methodology for training the fusion pipeline, followed by the proposed distributed inference approach as shown in Fig. 4.

A. Training Phase

We define the data matrices for GPS, LiDAR and images as: $X_C \in \mathbb{R}^{N_t \times 2}$, $X_L \in \mathbb{R}^{N_t \times d_0^L \times d_1^L \times d_2^L}$, $X_I \in \mathbb{R}^{N_t \times d_0^I \times d_1^I}$, respectively, where N_t is the number of training samples. Furthermore, $(d_0^L \times d_1^L \times d_2^L)$ and $(d_0^I \times d_1^I)$ give the dimensionality of preprocessed LiDAR and image data, while the GPS coordinate has 2 elements. We consider the label matrix $Y \in \{0, 1\}^{N_t \times |\mathcal{B}|}$ to represent the one-hot encoding of $\mathcal{B}$ beam pairs, where the optimum beam pair is set to 1, and rest are 0 as per Eq. (4). As mentioned in Sec. III-A, we have one optimal beam pair per sample, so we opted for one-hot encoding which enables having just one class per sample. Overall, we design a fusion framework to combine different data modalities that contains two main components: (i) base unimodal networks and (ii) the fusion network.

1) *Base Unimodal Neural Network:* We use the base unimodal neural network to (i) benchmark the performance of our fusion-based approach with respect to what can be achieved using only a single sensor type, and (ii) extract latent features from the penultimate (second last) layer of each that we use as input to our fusion network.

A deep neural network (DNN) can be considered as a combination of a non-linear feature extractor followed by a softmax classifier, i.e., the first layer until the penultimate layer of the DNN constitute the feature extractor [29]. The feature extractor

maps an input to a point in a multi-dimensional space called as the latent embedding space. The dimension of this high-level data representation is equal to the number of neurons in the penultimate layer. Then, in the final layer, the softmax activation function maps the high level representation of input data to a probability distribution over classes. As a result, the penultimate layer captures the unique properties of input data through a latent embedding space that is the key to making the final decision.

In this work, we propose to use the output of unimodal feature extractors as the high level data representation of each sensor modality. We assume that the penultimate layer of all three unimodal networks has d neurons. As a result, each sensor modality sample input maps to a vector with dimension d after passing through the feature extractors. We denote the feature extractor of each modality as $f^C_{\theta^C}$, $f^L_{\theta^L}$ and $f^I_{\theta^I}$ for coordinate, LiDAR, and image data, respectively, each parametrized by weight vectors θ^m, for $m \in \{C, L, I\}$. We refer to the output of these feature extractors as the latent embedding of each modality. Formally,

$$\mathbf{z}_C = f^C_{\theta^C}(X_C), \quad f^C_{\theta^C} : \mathbb{R}^2 \mapsto \mathbb{R}^d \tag{6a}$$

$$\mathbf{z}_L = f^L_{\theta^L}(X_L), \quad f^L_{\theta^L} : \mathbb{R}^{d^L_0 \times d^L_1 \times d^L_2} \mapsto \mathbb{R}^d \tag{6b}$$

$$\mathbf{z}_I = f^I_{\theta^I}(X_I), \quad f^I_{\theta^I} : \mathbb{R}^{d^I_0 \times d^I_1} \mapsto \mathbb{R}^d \tag{6c}$$

where $\mathbf{z}_C$, $\mathbf{z}_L$ and $\mathbf{z}_I$ show the extracted latent embeddings for input data X_C, X_L and X_I, respectively. We then apply a *tanh* activation on extracted latent features to regularize them in a range [-1, 1]. Note that the input to the base unimodal networks may contain negative values, which motivates the choice of *tanh* as the regularization function.

2) *Fusion Neural Network:* Each of the modalities capture different aspects of the environment. For instance, the GPS coordinates provide the precise location of the target receiver but it is blind to the shifts in the other objects in the environment and fails to provide any information about the dimensions of the vehicles. The LiDAR accuracy degrades in bright sunshine with many reflections [30]. Hence, fusing different modalities can compensate for the partial or inaccurate information and increase the robustness of the prediction.

Given the latent feature embedding of all modalities, we propose a fusion approach as follows: We explore that feature concatenation is an effective strategy for feature-level fusion in machine learning [31]. Hence, our fusion method is comprised of concatenation of latent feature embedding from each unimodal network to account for all sensor modalities, altogether. Thus, given $\mathbf{z}_C$, $\mathbf{z}_L$ and $\mathbf{z}_I \in \mathbb{R}^d$, we first concatenate them and generate the combined latent feature matrix $\mathbf{z}$ as:

$$\mathbf{z} = [\mathbf{z}_C; \mathbf{z}_L : \mathbf{z}_I] \in \mathbb{R}^{3\times d}. \tag{7}$$

Moreover, using multiple layers after concatenation of extracted features allows our fusion architecture to learn about the relevance of modalities, and therefore, it intelligently assigns higher weights to the features of the more relevant modalities. We pass the combined latent feature matrix $\mathbf{z}$ to another convolutional neural network (CNN) that we refer as *fusion network* to properly learn the relation of extracted latent embedding and the corresponding optimum beam pair. We denote the fusion network as $f^F_{\theta^F}(.)$. Finally, we use a softmax activation function to predict the optimality of each beam pair as:

$$\mathbf{s} = \sigma(f^F_{\theta^F}(\mathbf{z})), \quad f^F_{\theta^F} : \mathbb{R}^{3\times d} \mapsto \mathbb{R}^{|\mathcal{B}|} \tag{8}$$

where σ denotes the softmax activation function defined as $\sigma(\mathbf{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{|\mathcal{B}|} e^{x_j}}, i \in \mathcal{B}$, and $\mathbf{s} = [s_i]_{i\in\mathcal{B}} \in \mathbb{R}^{|\mathcal{B}|}$ indicates the predicted score of each beam pair. Note that $\mathbf{s}$ forms a probability distribution, with $s_i = \mathbf{P}((t^*, r^*) = i)$, $i \in \mathcal{B}$. We train this network offline using a cross-entropy penalty, over data in which the optimal (t^*, r^*) pair is one-hot encoded.

B. Distributed Inference Phase

Unlike the training phase that occurs offline, the inference needs to occur real-time. To that end, the MEC must receive instantaneous data from three sensor modalities, which is passed to the trained fusion pipeline for predicting the top-K beam pairs. Since the sensors are not co-located, to accelerate inference, we distribute the ML architecture taking account the limitations of the control channel delivering the sensor data to the MEC. Our distributed inference scheme is illustrated in Fig. 1. The trained base unimodal networks for GPS coordinates and LiDAR are deployed at the vehicle to locally generate the high level latent embeddings $\mathbf{z}_C$ and $\mathbf{z}_L$. The extracted features are then concatenated as $\mathbf{z}_{CL} = [\mathbf{z}_C; \mathbf{z}_L] \in \mathbb{R}^{2\times d}$ and sent over the sub-6 GHz data channel. Similarly, the base unimodal network of the image generates the features for this modality at the BS, which is then combined with $\mathbf{z}_{CL}$ at the MEC as $\mathbf{z} = [\mathbf{z}_{CL}; \mathbf{z}_I] \in \mathbb{R}^{3\times d}$. Note that this methodology results in the same combined latent feature matrix $\mathbf{z}$ as (7), we analyze the improvement in end-to-end latency with this distributed inference approach in Sec. VIII. Finally, given the latent feature embedding of all modalities available at the MEC, we use the fusion network, $f^F_{\theta^F}(\cdot)$, followed by a softmax activation to predict the score of each beam pair according to Eq. (8). Fig. 4 depicts the dissemination of the fusion pipeline over the system.

VI. Top-K Beam Pair Construction

The proposed fusion pipeline outputs a softmax score for each of the possible beam pairs given the different sensor modalities. Recall that our goal is to identify a subset of beam pairs $\mathcal{B}_K$ such that $(t^*, r^*) \in \mathcal{B}_K$ with high probability. We describe in this section how the neural network outputs are used for that purpose, as well as how we select parameter K.

A. K Selection Problem Formulation

Consider the softmax score vector $\mathbf{s} = [s_i]_{i\in\mathcal{B}} \in \mathbb{R}^{|\mathcal{B}|}$ outputed by the neural network via Eq. (8). Recalling that $\mathbf{s}$ provides a probability distribution for (t^*, r^*) over $\mathcal{B}$, the top-K beam configurations Eq. (5) becomes:

$$\mathcal{B}_K(\mathbf{s}) = \underset{A\subset\mathcal{B}, |A|=K}{\arg\max} \sum_{i\in A} s_i. \tag{9}$$

Hence, given scores $\mathbf{s}$ and parameter K, $\mathcal{B}_K$ can be easily constructed by sorting $\mathbf{s}$ and identifying the top-K elements.

B. Selecting K

Parameter K establishes a tradeoff between the probability that the optimal beam pair is in $\mathcal{B}_K$ and the time it takes to determine the best (but possibly sub-optimal) beam within $\mathcal{B}_K$. This suggests selecting K by optimizing an objective of the form:

$$\max_K \mathbf{P}\left((t^*, r^*) \in \mathcal{B}_K\right) + \mu(K)$$

where $\mu : \mathbb{N} \to \mathbb{R}_+$ is a penalty increasing with the latency incurred by the choice of K. We discuss how to set these terms, and additional constraints we introduce, in this section.

Modeling Probability of Inclusion: A simple way to model the probability of the event $(t^*, r^*) \in \mathcal{B}_K$ is via the softmax scores $\mathbf{s}$, as in Eq. (9). We observed however that this tends to overestimate the probability of this event in practice: even if softmax scores are good for selecting the set $\mathcal{B}_K$ quickly and efficiently, a more careful approach is warranted when selecting K.

To that end, we leverage the empirical distribution of scores in our training set. In particular, given a score vector $\mathbf{s} = [s_i]_{i\in\mathcal{B}} \in \mathbb{R}^{|\mathcal{B}|}$ and $K \in \mathbb{N}$ let

$$c_K(\mathbf{s}) = \max_{A\subset\mathcal{B}, |A|=K} \sum_{i\in A} s_i \tag{10}$$

be the sum of the K largest scores in $\mathbf{s}$. Let $I \in \{1, \ldots, N_t\}$ be a sample index selected uniformly at random from our training set. Let also $\mathbf{s}_I$ be the corresponding softmax output layer associated with I, and $(t_I^*, r_I^*) \in \mathcal{B}$ the optimal pair associated with this sample. Then, given a score vector $\mathbf{s}$ generated at runtime and the corresponding $\mathcal{B}_K$, we estimate the probability of the event $(t^*, r^*) \in \mathcal{B}_K$ via:

$$p(K) = \mathbf{P}\left((t_I^*, r_I^*) \in \mathcal{B}_K(\mathbf{s}_I)\right) \tag{11}$$

$$p(K; \mathbf{s}) = \mathbf{P}\left((t_I^*, r_I^*) \in \mathcal{B}_K(\mathbf{s}_I) \mid c_K(\mathbf{s}_I) \leq c_K(\mathbf{s})\right), \tag{12}$$

where the probability is w.r.t the random sample I in the dataset. Intuitively, this captures the empirical probability that (t^*, r^*) is in a random set $\mathcal{B}_K$ constructed in the training set, conditioned on the fact that our choice of K restricts these sets by bounding the quantity c_K to be at most $c_K(\mathbf{s})$. In some sense, this allows us to link softmax scores to the variability of confidence in the construction of $\mathcal{B}_K$, itself depending upon different LOS/NLOS conditions, vehicular traffic patterns, etc. The training set is used to statistically quantify this variability.

We note that Eq. (12) can be computed efficiently via Bayes rule, without the need to access the training set at runtime. In particular, for $c = c_K(\mathbf{s}) \in \mathbb{R}_+$, $p(K; \mathbf{s})$ is equal to:

$$\frac{\mathbf{P}\left(c_K(\mathbf{s}_I) \leq c \mid (t_I^*, r_I^*) \in \mathcal{B}_K(\mathbf{s}_I)\right) \mathbf{P}\left((t_I^*, r_I^*) \in \mathcal{B}_K(\mathbf{s}_I)\right)}{\mathbf{P}\left(c_K(\mathbf{s}_I) \leq c\right)}. \tag{13}$$

The constituent cumulative density functions can be computed directly from the dataset for each $K \leq |\mathcal{B}|$, and then used at runtime.

Algorithm 1: Top-K Beam Pair Selection.

Inputs: softmax score $\mathbf{s}$ generated by F-DL framework in Sec. V, T_{total};
Output: $\mathcal{B}_K$
1: Compute probability of inclusion (13)
2: Compute channel efficiency (14)
3: $K \leftarrow \max_K \; p(K; \mathbf{s}) + \alpha\mu(K)$;
4: Construct $\mathcal{B}_K$ according to Eq. (9)

Incorporating Latency: Since the transmitter and receiver sweep all suggested beam pairs in $\mathcal{B}_K$, we include a second term mmWave channel efficiency in the objective defined as:

$$\mu(K) = \frac{T_{total} - T_{bs}^{df}(K)}{T_{total}}, \tag{14}$$

with T_{total} and $T_{bs}^{df}(K)$ being the total time for which a certain beam pair is valid and the end-to-end latency imposed by our proposed fusion based beam selection approach, respectively. We precisely analyze the end-to-end latency of our proposed beam selection approach in Sec. VIII. Note that the T_{bs}^{df} is an increasing function of K. Hence, the mmWave channel efficiency is a decreasing function with respect to K.

Optimization: Combining the above terms, the final optimization problem we solve to determine K given a run-time score vector $\mathbf{s}$ is (see Algorithm 1):

$$\max_K \quad p(K; \mathbf{s}) + \alpha\mu(K), \tag{15a}$$

$$\text{s.t.} \quad T_{bs}^{df}(K) < T_{total}, \tag{15b}$$

$$\alpha > 0. \tag{15c}$$

In Eq. (15), the first term in objective enforces the algorithm to select higher values of K and ensure the alignment, when the optimum beam pair is included in the K suggested beams. On the contrary, the second item avoids selecting unnecessarily high K values. The control parameter α in (15) weights the importance between the two terms in the objective function.

VII. Dataset Description and DNN Architectures

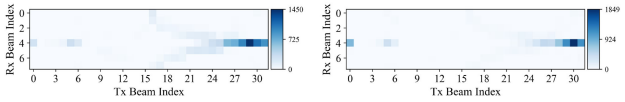
In this section, we introduce two datasets which we use to evaluate the F-DL framework. The Raymobtime dataset [32] is one of the widely used comprehensive multimodal dataset which has been basis of many state-of-the-art techniques. However, to give more perspective on applicability of the proposed F-DL architecture, we collect our own "real-world" multimodal data, which includes real sensors, urban environment, and RF ground-truth. Further, we detail the preprocessing and implementation steps used in the proposed framework.

A. Datasets

1) Simulation Dataset: The Raymobtime multimodal dataset captures virtually with high fidelity V2X deployment in the urban canyon region of Rosslyn, Virginia for different types of traffic. A static roadside BS is placed at a height of 4

TABLE III
STATISTICS OF S008 AND S009 DATASETS

dataset	# of Samples	LOS	NLOS	NLOS Percentage
S008	11194	6482	4712	42%
S009	9638	1473	8165	85%



(a) (b)

Fig. 5. Distribution of S008 and S009 datasets. (a) S008 (b) S009

meters, alongside moving buses, cars, and trucks. The traffic is generated using the Simulator for Urban MObility (SUMO) software [33], which allows flexibility in changing the vehicular movement patterns. The image and LiDAR sensor data are collected by Blender [34], a 3D computer graphics software toolkit, and Blender Sensor Simulation (BlenSor) [35] software, respectively. For a so called *scene*, the framework designates one active receiver out of three possible vehicle types i.e. car, bus and truck. For each scene, (i) the receiver vehicle collects the LiDAR point clouds and the GPS coordinates, (ii) a camera at the BS takes a picture, and (iii) the combined channel quality of different beam pairs are generated using Remcom' Wireless Insite ray-tracing software [36]. The BS and receiver vehicle have uniform linear arrays (ULAs) with element spacing of $\lambda/2$, where λ denotes the signal wavelength. The number of codebook elements for BS and the receiver is 32 and 8, respectively, leading to *256 beam pairs*. The gap between two consecutive scenes is 30 seconds which corresponds to sampling rate of 2 samples/minute. A python orchestrator is responsible for data flow across the system to ensure the different software operations are synchronized.

The simulation is repeated for the same scenario with two different traffic rates. We refer to these datasets as S008 and S009, which correspond to regular and rush-hour traffic, respectively. Since there are more vehicles in S009, the number of NLOS cases is higher. Table III denotes the number of LOS and NLOS cases for both datasets. We use the S008 dataset for training and validation and S009 as the testing set. Fig. 5 illustrates the distribution of the classes over S008 and S009. We observe that the dataset is highly imbalanced, i.e., there is a huge variation in the number of different classes, a property that is expected due to the sparsity of mmWave links.

2) Real-World NEU Dataset: This dataset contains multi-modal sensor observations collected in the greater metropolitan area of Boston. The experiment setting is an outdoor urban road with two-way traffic surrounded by high-rising buildings on both sides. An autonomous vehicle equipped with GPS (sampling rate 1 Hz) and Velodyne LiDAR (sampling rate 10 Hz) sensors establishes connection with a mmWave base station located at a road-side cart. The RF grand-truth is acquired using Talon AD7200 60 GHz mmWave routers with a codebook of 64 beam configurations [37]. Each dataset sample includes the synchronized recordings of GPS and LiDAR sensors along with

TABLE V
SUMMARY OF DIFFERENT CATEGORIES OF NEU DATASET

Category	Speed (mph)	Scenarios	Samples
LOS passing	10	–	1568
NLOS by pedstrain	15	standing walk right to left walk left to right	4791
NLOS by static car	15	in front	1506
NLOS by moving car	20	15mph same lane 15mph opposite lane	2988

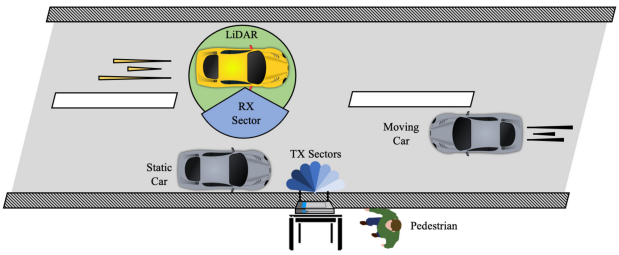


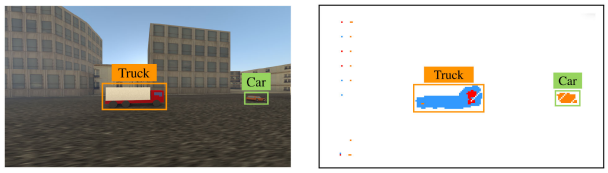
Fig. 6. NEU dataset collection environment includes for categories as: LOS passing, NLOS by pedestrian, NLOS by static car and NLOS by moving car.

TABLE IV
WEATHER FORECAST ON THREE DAYS OF DATA COLLECTION

Day	Temperature (F)	Humidity	Max Wind Speed (mph)	Atmospheric Pressure (Hg)	Precipitation (Inches)
1	53-75	48-74%	17	30.13	2.90
2	59-67	75-87%	13	30	3
3	56-68	54-84%	8	30.37	3.10

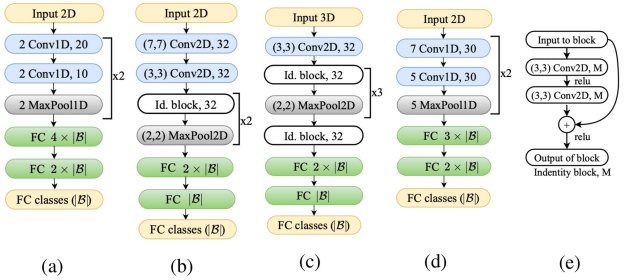
the grand-truth RF measurements. The data collection vehicle maintains speeds between 10-20 mph following the speed-limit of inner-city roads. The dataset setting spans a variety of four categories, including the LOS passing, blockage by pedestrian, static, and moving car with 10853 samples (116.7 GB) overall (see Table V). Fig. 6 denotes a diagram of the experiment setting top view. The dataset is collected during three days with different levels of humidity and weather conditions. The weather forecast information during data collection days is presented in Table V. In particular, the humidity and maximum wind speed change between 53–75% and 8–17 mph, respectively, resulting in a rich representation of weather in the dataset.

The NEU dataset is collected to expand the feasibility study of the F-DL architecture. However, to resemble the futuristic V2X architecture, the considered framework requires tower-mounted base stations equipped with a camera. As we did not have access to such infrastructure, we collect the NEU dataset with LiDAR and GPS sensors deployed in a car. This fact does not diminish the applicability of the collected dataset, as the processed fused features from LiDAR and GPS are transmitted from car to mmWave base station following the same architecture as mentioned in Fig. 1. Hence, we argue that the NEU dataset can be considered as a solid reference dataset for the beam selection task, considering the scarcity of real datasets for mmWave experiments. The real-world NEU dataset is released online in our public dataset repository [38].



(a) (b)

Fig. 7. An example of input and output of image preprocessing. (a) Raw image (b) Generated bit map



(a) (b) (c) (d) (e)

Fig. 8. Proposed architectures for unimodal and fusion networks. (a) GPS (b) Image (c) LiDAR (d) Fusion (e) Identity

B. Preprocessing

1) Image: To construct the dataset for the image preprocessing classifier, we manually identify and close in bounding boxes samples of bus, car, truck and background and quantize them by following the steps mentioned in Appendix A. We label these as background (0), bus (1), car (2), truck (3). The constructed dataset contains 22482 samples per class on average. We then train a classifier as follows. The input crops are first passed to a convolutional layer with 20 filters of kernel size (15, 15) followed by a max-pooling layer with the pool size of (3, 3) and stride size of (2, 2). The output is fed to two consecutive dense layers with 128 and 4 neurons (number of classes). Our trained classifier achieves 84% accuracy in separating the samples of each class. In the Raymobtime dataset, the camera generates (540, 960, 3) RGB images. We empirically choose the window size of 40 and stride size 3 for our task that results in the output bit map of size (101, 185). Fig. 7 shows a sample from the dataset and the generated bit map. Note that the multi-object detection algorithm can be easily extended to any type of vehicle by including the samples from new vehicles in the training set [39]. We evaluate the delay cost of image preprocessing in Sec. VIII-A.

2) LiDAR: The maximum distance for LiDAR is 100 meters in the Raymobtime dataset, and the zone of space is limited in each axis as, $(X_{\min}, X_{\max}) = (744, 767)$, $(Y_{\min}, Y_{\max}) = (429, 679)$, $(Z_{\min}, Z_{\max}) = (0, 10)$, where the static BS is located at $[746, 560, 4]$ within this Cartesian coordinate system. Moreover, the histogram bin size along the three spatial dimensions is set as $(1.15, 1.25, 1)$, respectively. Following the steps mentioned in Sec. IV-B, we generate a compact $(20, 200, 10)$ representation of the environment where the BS, target vehicle, and obstacles are marked with different indicators. For NEU dataset, we use the maximum LiDAR distance of 80 meters and map the LiDAR point clouds to a compact (20,20,20) representation in each axis.

C. Implementation Details

Our proposed fusion pipeline consists of three unimodal networks per modality followed by a fusion network as presented in Fig. 4. We first design each unimodal network tuned to each dataset which takes either raw (for coordinate) or preprocessed (LiDAR and image) data as input and generate the latent embeddings to be fed to the fusion network. For GPS unimodal network, we design a model that uses 1-D convolutional layers (see Fig. 8(a)). This enables capturing the correlation between the latitude and longitude, simultaneously. Our custom designed model for the preprocessed images (see Sec. IV-A) is inspired by ResNet [40] that uses identity connections to avoid the gradient vanishing problem commonly seen in deep architectures, by creating a direct path for the gradient during backpropagation. Each such *identity block* contains 2 convolutional layers and an identity shortcut that skips these 2 layers, followed by a max-pooling layer, as shown in Fig. 8(e). For LiDAR input, we also design a model structure similar to ResNet (see Fig. 8(c)). Note that while the input to image and LiDAR models are 2D and 3D, the majority of elements are zero due to filtering the irrelevant data during preprocessing. We also use max-pooling layers after convolutional layers for feature down-sampling and dropout of 0.25 after fully-connected layers to avoid overfitting.

The representation capacity of each network including latent embedding generators scales with the number of classes $|\mathcal{B}|$ in each dataset, 256 and 64 for Raymobtime and NEU, respectively. Though increasing the number of neurons generally improves the representation capacity of base unimodal architectures, we find having neurons equal to the number of classes to be sufficient for our task. We design a fusion network as depicted in Fig. 8(d) that takes as input the concatenated latent embedding of each modality. Ultimately, the last dense layer with the number of classes outputs the predicted score of each beam pair. For all models, we exploit categorical cross-entropy loss with batch size of 32 and training epochs of 100 and 400 for Raymobtime and NEU dataset with an earlier stopping point of patience 10. Moreover, we apply ℓ_1 and ℓ_2 kernel regularizers on dense layers with parameters 10^{-5} and 10^{-4}, respectively. We use Adam [41] as optimizer with $\beta = (0.9, 0.999)$ and initialize the learning rate to 0.0001.

VIII. End-to-End Latency Analysis With Distributed Inference

In this section, we explore the design details and performance trade-offs related to centralized/distributed inference. Moreover, we answer the following question: *What is the end-to-end latency of beam selection with our proposed method?*

A. Data Collection and Preprocessing

Current LiDAR sensors support pulse rate, i.e., the number of discrete laser "shots"per second that the LiDAR is firing, of

TABLE VI
THE REQUIRED TIME FOR SHARING THE DATA WITH MEC (T_{DATA}) FOR THREE DATA SHARING STRATEGIES FOR RAYMOBTIME AND NEU DATASETS

Method	Raymobtime					NEU				
		Min Reqd. time (ms)		Max Reqd. time (ms)			Min Reqd. time (ms)		Max Reqd. time (ms)	
	# Bytes	802.11p	LTE	802.11p	LTE	# Bytes	802.11p	LTE	802.11p	LTE
Preprocessed	326 KB	12.07	4.34	108.66	74.09	64 KB	2.37	0.85	21.33	14.55
High-level fused features	4 KB	0.148	0.053	**1.332**	0.90	1 KB	0.037	0.013	**0.33**	0.225

50,000 to 150,000 pulses per second, while 35 cm precision can be achieved with 8 pulses/m^2 [42]. The GPS sensor data does not require any preprocessing and the LiDAR preprocessing has a negligible latency that can be further reduced by exploiting parallel processing. For image sensor data, we measure the delay of our proposed object detection algorithm described in Appendix. A by passing a single sample 100 times and calculating the average required time for generating bit maps. Accordingly, our proposed image preprocessing pipeline generates the bit maps in 1.30 ms on average. As a result, our preprocessing pipeline runs in 1.30 ms on average ($T_{\text{process}} = 1.30\,ms$). Note that image preprocessing is applied on Raymobtime dataset only.

B. Sharing Features Between Vehicle and MEC

Data collected at vehicular locations can incur different relaying costs to the MEC, depending upon the sensor modality. For GPS coordinates, both latitude and longitude, can be expressed in 6 Bytes, while the raw LiDAR point cloud requires ∼1-1.5 MBytes for complete transfer. One possible approach is to relay the GPS measurements *as is* while subjecting the LiDAR data to additional preprocessing step as discussed in Sec. IV-B. This step maps the raw LiDAR point clouds to a ridge representation with size (20, 200, 10) that can be shown with ∼320 KBytes (78% less than raw LiDAR point clouds) for Raymobtime dataset. Using the aforementioned prepossessing reduces the data from 0.9 MByte to 64 KByte for NEU dataset as well. We can further improve the data transmission speed from vehicle to the MEC by sending the fused high level latent embeddings of LiDAR and GPS. Recall that we extract this information at an intermediate layer of the neural network (see Sec. V-A). With our proposed distributed inference design, the raw coordinates and LiDAR data is translated to an array with $2 \times |\mathcal{B}|$ elements that is expressed with only ∼4 KBytes and ∼1 KBytes for Raymobtime and NEU datasets, respectively (∼ 99% reduction in size than raw data), which is even more compressed and requires less bandwidth within the sub-6 GHz control channel.

Table VI illustrates the number of bytes and the minimum/maximum experienced delay while transmitting the compressed extracted features of coordinate and LiDAR over the sub-6 GHz data channel. The achievable throughput is assumed to be 3-27 Mbs and 4.4-75 Mbs for 802.11p [43] and single input single output (SISO) LTE [44], respectively.

Additionally, the fused features are difficult to interpret by third parties and provide a level of abstraction to the raw data. From Table VI, we observe that the data channel delay reduces drastically with the distributed inference. Without loss of generality, we use the maximum imposed delay of control signaling from vehicle to MEC being ($T_{\text{data}} = 1.332\,ms$) for Raymobtime and ($T_{\text{data}} = 0.33\,ms$) for NEU datasets to calculate the overall end-to-end latency.

C. Inference and Sharing Selected Beams With Vehicle

In order to evaluate the inference delay, we pass input data, i.e., the latent embedding of all modalities, through our pipeline and measure the prediction time by setting a timer and subtracting the timestamp before and after prediction. We note that the average inference time of our proposed fusion approach is 0.37 ms. On the other hand, sending the selected K beams from MEC to vehicle over the sub-6 GHz control channel requires at most 2 KB (256 elements) and 0.5 KB (64 elements) for Raymobtime and NEU datasets, respectively. That takes 0.66 ms and 0.16 ms as maximum required time, and results in a cumulative delay (T_{control}) of 1.03 ms and 0.53 ms for each dataset, respectively. Similar to the previous section, we consider the highest imposed delay related to using IEEE 802.11p standard as our reference.

D. Impact on Beam-Sweeping Latency: Case Study in 5G-NR

We first discuss the time requirement of exhaustive beam search in 5G-NR standard. Next, we calculate the required time for sweeping only the selected K beam pairs by following the same norms as 5G-NR standard.

1) Beam Selection Latency in 5G-NR: For evaluating a 5G-NR standard compliant beam selection process in the mmWave band, we consider a transmitter-receiver pair with the codebook sizes M and N, respectively. With analog beamforming, we have a total of $|\mathcal{B}| = M \times N$ combinations (see Sec. III). During the initial access, the gNodeB and user exchange a number of messages to find the best beam pair. In particular, the gNodeB sequentially transmits synchronization signals (SS) in each codebook element $t_m \in C_{Tx}$. Meanwhile, the receiver also tunes its array to receive in different codebook elements $r_n \in C_{Rx}$ until all possible beam configurations are swept. The SS transmitted in a certain beam configuration is referred as the SS block, with multiple SS blocks from different beam configurations grouped into one SS burst. The NR standard defines that the SS burst duration (T_{ssb}) is fixed to 5 ms, which is transmitted with a periodicity (T_p) of 20 ms [45]. In the mmWave band, a maximum of 32 SS blocks fit within a SS burst, which allows for 32 different beam pairs to be explored within one SS burst. Hence, in order to explore all beam pair combinations, a total of $|\mathcal{B}|$ SS blocks are required to be transmitted. Given the limit on SS blocks within a SS burst, the total time to explore all beam

pairs (T_{bs}^{nr}) can be expressed as:

$$T_{bs}^{nr}(|\mathcal{B}|) = T_p \times \left\lfloor \frac{|\mathcal{B}| - 1}{32} \right\rfloor + T_{ssb}, \tag{16}$$

where $T_p = 20\,ms$ and $T_{ssb} = 5\,ms$ correspond to periodicity and SS burst duration, respectively. Note that if a certain number of beam pairs are not explored within the first SS burst ($|\mathcal{B}| > 32$), there is an increasing delay given the separation T_p between SS bursts. On the other hand, exploring a number of pairs smaller than 32 will introduce the same overhead as if a total of 32 options were searched, given that T_{ssb} has a fixed duration of $5\,ms$. Similarly, this can be extended to any number $|\mathcal{B}|$ that is not a multiple of 32.

2) Improvement in Latency Through Proposed Approach: Our proposed approach reduces the beam search space from $|\mathcal{B}|$ to a subset of $K \ll |\mathcal{B}|$ most likely beam candidates, derived from Algorithm 1. We recall that the NR standard assumes that up to 32 can be swept within $5\,ms$. Thus, we define the time to explore one single beam as $T_b = 5\,ms/32 = 156\,ns$. Then, the required time for sweeping the selected top-K beam pairs can be expressed as:

$$T_{\text{sweep}}(K) = T_p \left\lfloor \frac{K - 1}{32} \right\rfloor + T_b(1 + (K - 1) \bmod 32). \tag{17}$$

E. End-to-End Latency Calculation

Considering the aforementioned four steps, the overall beam selection overhead following our proposed data fusion approach (T_{bs}^{df}) with distributed inference is expressed as:

$$T_{bs}^{df}(K) = T_{\text{process}} + T_{\text{data}} + T_{\text{control}} + T_{\text{sweep}}(K), \tag{18}$$

where the first three terms can be approximated by $3.662\,ms$ and $0.86\,ms$ for Raymobtime and NEU datasets, respectively. Note that the distributed inference play a pivotal role in reducing the overhead associated with sharing the situational state of the vehicle with the MEC (T_{data}). We validate the improvement in overall beam selection time using the proposed distributed inference (Eq. 18) approach rather than the traditional brute-force approach offered by the state-of-the-art 5G-NR (Eq. 16) standard in Sec. IX-E.

IX. Results and Discussions

In this section, we provide the results of our proposed method using the datasets described in Sec. VII-A. We use Keras 2.1.6 with Tensorflow backend (version 1.9.0) for implementation. To judge the efficiency of proposed beam selection approach on multi-class, highly-imbalanced, multimodal Raymobtime [32] and NEU datasets, we use four evaluation metrics that capture the performance from different aspects, including top-K accuracy, weighted F-1 score, KL divergence and throughput ratio. We provide the detailed definitions of these metrics in Appendix B. We first analyze the performance of proposed fusion deep learning method on Raymobtime dataset, and then further justify the performance on real-world NEU dataset in Sec. IX-F.

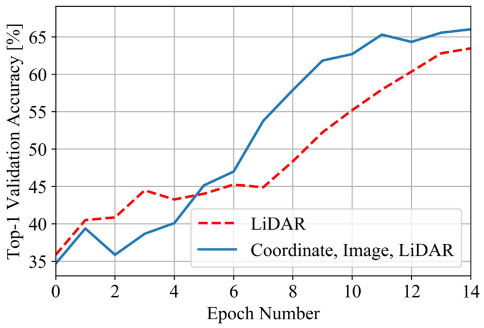


Fig. 9. Comparing top-1 validation accuracies of LiDAR-only and fusion with all three modalities on the Raymobtime dataset.

A. Performance of Base Unimodal Architectures

We assess the performance of beam selection by only relying on unimodal data. The experimental results of predicting top-K beam pairs are presented in Table VII, for each proposed unimodal architectures. In the table, we report the top-K (K=1, 2, 5, 10, 25, 50) accuracy along with weighted recall, precision and F1 score and the KL divergence of the predicted labels and true labels on Raymobtime dataset. We observe that the LiDAR outperforms coordinate and image in all metrics with 46.23% top-1 accuracy, which makes it the best single modality. Moreover, to justify the improvement achieved by using the image preprocessing step described in Appendix A, we compare the weighted recall on raw and preprocessed image data. Interestingly, we observed that by using the raw images, the model always predicts the class with the highest occurrence in the training set that results in the weighted recall of 0.01%. Intuitively, in the case of using raw images, the model cannot find a relation between the input image and the labels since from a raw image perspective any vehicle captured in the image can be the target receiver. On the other hand, using the image preprocessing step increases the weighted recall to 7% as presented in Table VII.

B. Performance of Fusion Framework

The results of fusion on different combinations of unimodal data are presented in Table VII for Raymobtime dataset. We observe that the fusion increases the beam prediction accuracy in all combinations. Moreover, the best result is achieved when all modalities are fused together with 9.99% improvement in top-1 accuracy in comparison with the best unimodal data i.e., LiDAR. The improvement with fusion can be also justified by the validation accuracy during training. Fig. 9 compares the top-1 validation accuracy of fusion of all three modalities with LiDAR-only (best single modality). We observe that although the top-1 validation accuracy of fusion is lower in early epochs, it outperforms the LiDAR after five epochs.

Since the dataset is highly imbalanced, we report results using metrics like weighted precision, recall, and F1 score to confirm the improvement. Furthermore, we use KL divergence metric to measure the overall performance of the fusion pipeline. The lower the divergence, the more is the similarity between true

TABLE VII
PERFORMANCE OF PROPOSED UNIMODAL AND FUSION WHEN TRAINED ON S008 AND TESTED ON S009 RAYMOBTIME DATASET

Modalities	Top-1 Accuracy	Top-2 Accuracy	Top-5 Accuracy	Top-10 Accuracy	Top-25 Accuracy	Top-50 Accuracy	Weighted Recall	Weighted Precision	Weighted F1 score	KL divergence
Coordinates	12.32%	31.51%	55.61%	77.93%	88.5%	95.14%	2%	12%	3%	3.02
Image	12.39%	26.84%	55.38%	71.65%	88.05%	95.01%	7%	12%	3%	2.9051
LiDAR	46.23%	64.67%	82.43%	89.95%	96.11%	98.13%	47%	46%	45%	0.1738
Coordinates, Image	25.76%	44.88%	74.18%	86.29%	94.78%	97.89%	21%	26%	22%	0.5432
Coordinates, LiDAR	55.42%	74.54%	85.51%	91.41%	96.75%	98.56%	55%	55%	54%	0.1357
Image, LiDAR	54.52%	73.08%	84.83%	91.23%	96.78%	98.50%	55%	55%	54%	0.1428
Coordinate, Image, LiDAR	**56.22%**	**74.08%**	**85.53%**	**91.11%**	**96.56%**	**98.60%**	**55%**	**56%**	**55%**	**0.1314**

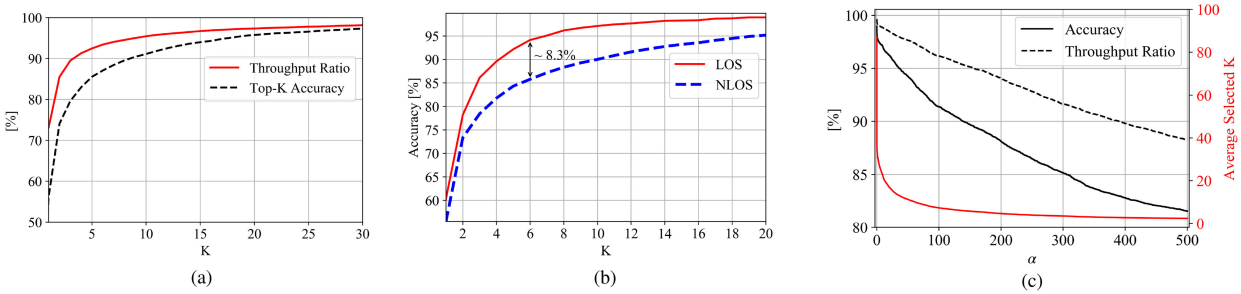


Fig. 10. (a) Comparison of throughput ratio and beam selection accuracy with varying K (b) LOS/NLOS accuracy for $K = 0, 1, \ldots, 20$ (c) Analysis of throughput ratio, accuracy and average selected K for different α values in (15).

and predicted labels. We also use KL divergence to show the relative entropy between train (S008) and test (S009) data labels (Shown in Fig. 5). We get KL divergence of 0.57 signifying high relative entropy between the train/test label distributions. From Table VII, we observe that the fusion with all unimodal data leads to the lowest KL scores. Hence, we deduce that fusion among all three modalities is the most successful scheme to capture the label distribution in the test set. Hence, we choose the proposed fusion-based approach comprising of all three modalities as beam selector for the rest of the performance evaluation.

C. Studying the Impact of K

To analyze the impact of different K values in the overall performance, we point out that failure in selecting the optimum beam pair within the suggested subset ($(t^*, r^*) \notin \mathcal{B}_k$) results in the drop in the received signal power. Hence, we choose the throughput ratio (see Appendix B) as our metric to assess the QoS of the system. Intuitively, the throughput ratio depicts the ratio of average throughput when sweeping only K beam pairs predicted by the model with reference to what could be achieved with exhaustive search. Fig. 10(a) compares the throughput ratio and normalized beam selection accuracy with K varying from 1 to 30 for Raymobtime dataset. As expected, both increase with K since it is more likely to include the optimum beam pair with higher K. We observe the gap between the accuracy and throughput ratio starts with 16.90% for K=1, and it decreases as K increases. We do not observe significant improvement in throughput ratio after $K = 10$; however, the accuracy keeps on improving until $K = 25$. Note that while increasing K improves the quality of service (QoS), it results in higher beam selection overhead. Hence, it is crucial to balance the tradeoff between the two as proposed in dynamic selection of top-K beam pairs algorithm in Sec. VI.

D. Impact of LOS and NLOS

The presence of obstacles leads to massive drops in channel quality given the high attenuation in the mmWave band. Additionally, users might experience a considerable reduction in their QoS to tens of Gbps. In LOS scenario, the corresponding best beam pair distinctively outperforms the others. However, the presence of blockage in LOS path causes unexpected beams to achieve the highest signal strength through multiple reflections. We show this in Fig. 10(b), which compares the accuracy of our proposed fusion where the sample of test data are separated based LOS/NLOS scenario in Raymobtime dataset. As expected, prediction in the case of complex reflections of NLOS links is more challenging, showing a maximum drop of 8.3% in beam selection accuracy against LOS scenarios.

E. Impact on Beam Selection Speed

As discussed in Sec. VIII-D1, the 5G-NR standard define a brute-force beam sweeping process that sequentially explores all possible directions. In addition, according to Eq. (16), only up to 32 directions can be explored within one SS burst, which creates additional waiting time within one beam selection process. In order to decrease such overhead, we propose a solution that selects a reduced set of K beam pairs and performs a brute-force search only on those ones. Also, given the different confidence levels of our prediction model due to potential scenario variations, we propose an algorithm that selects K flexibly to avoid unnecessary overhead.

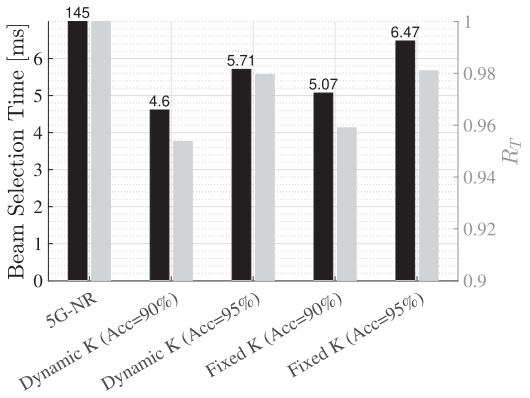


Fig. 11. Comparison of relative throughput and end-to-end beam selection time (*Eq. (18)*) of proposed approaches, Dynamic K (*Algorithm 1*) and Fixed K (*Eq. (9)*), with 5G-NR standard. The actual beam selection time of 145 ms for 5G-NR is scaled here, for better visibility and comparison purpose.

In the Raymobtime dataset, the road length is 200 meters and the BS is located in the middle. On the other hand, the 3-dB beam width of an uniform linear array antenna with N elements is approximately equal to $2/N$ radians [46] that results in span of 3.58° and 14.32° for each beam of transmitter and receiver codebooks, respectively. Hence, the overall BS coverage angle is equal to $\phi_{BS} = 114.56°$ and the contact time, i.e., the time that the vehicle remains in the span of one beam, is equal to $T_{total} = \frac{2h\tan(\frac{\phi_{BS}}{2})}{v_l}$ with h and v_l being the height of the BS and the velocity of the vehicle [47]. Consequently, the vehicle remains in the coverage region of each beam pair for $\sim 807\,ms$ while moving with the velocity of 32 km/h (average speed in urban roads). Therefore, the beam selection process needs to be repeated every $807\,ms$ (T_{total}). In Fig. 10(c), we analyze the impact of α in (15) on the throughput ratio (R_T), the accuracy and the average selected K. We observe how the triplet R_T, *accuracy* and *average selected* K decreases with α, the control parameter in Eq. (15). Intuitively, increasing α gives more weight to the second term in (15) that forces the algorithm to be faster and choose lower K which results in lower QoS and beam selection accuracy. Interestingly, we observe that for $\alpha = 0$ the maximum average selected K is equal to 87. In this scenario, the objective in (15) aims to maximize the alignment probability and increasing the K and yet it does not exceeds 87 out of 256. We conclude that our proposed fusion method achieves to ∼100% top-87 accuracy.

The control parameter in (15) enables us to slide between different accuracy and overhead conditions. Fig. 11 shows that the dynamic K selection approach achieves an average throughput ratio of 95.37% and 97.95% while targeting 90% and 95% accuracy, respectively. This implies that the capacity of the proposed F-DL approach is only 4.63% lower than the 5G-NR standard, while targeting the accuracy of 90% for instance. Moreover, the dynamic K selection approach offers the corresponding beam sweeping overhead of $0.94\,ms$ and $2.04\,ms$, Eq. (17) and the overall beam selection delay of $4.6\,ms$ and $5.71\,ms$. Note that the beam selection delay of our proposed dynamic K selection method in Fig. 11 corresponds to the end-to-end latency of the proposed F-DL method presented in Eq. (18). In contrast, the 5G-NR standard beam selection procedure requires $145\,ms$. Therefore, we notice 96% reduction in overall beam selection overhead while retaining 97.95% relative throughput associated with 95% accuracy. Furthermore, we compare the performance of proposed algorithm for constructing the subset $\mathcal{B}_K$, Algorithm 1, that is generated *dynamically* per case, with the fixed K one (Fig. 10(a)). Note, that fixed K selection is a posterior probability derived after observing all test samples; however, the dynamic K selection selects the K for each sample of test set, independently. From this figure, we observe that the proposed dynamic K selection approach outperforms the fixed K one, providing faster beam selection with close competing relative throughput while targeting the same accuracy. We use the same standard, i.e., 5G-NR for fair comparison (see Fig. 11). Note that our algorithm can be trivially extended to any other exhaustive beam search standards, such as IEEE 802.11ad by modifying Eq. (17), yet it does not negate the improvement achieved by restricting the beam selection to a lower dimension space.

TABLE VIII
PERFORMANCE OF PROPOSED UNIMODAL AND FUSION METHOD ON REAL-WORLD NEU DATASET

Modalities	Top-1 Accuracy	Top-2 Accuracy	Top-5 Accuracy	Weighted F1 score
Coordinates	39.94%	54.39%	81.05%	33.63%
LiDAR	74.86%	89.04%	97.57%	75.02%
Coordinates, LiDAR	78.18%	91.02%	98.02%	78.62%

F. Real-World Implementation

We validate the performance of the proposed fusion deep learning method on the home-grown NEU dataset. As mentioned in Sec. VII-A2, due to the infrastructural limitation, we use only LiDAR and GPS branch of the proposed F-DL (presented in Sec. V, Fig. 4) for this set of experiment. Table VIII compares the beam selection accuracy while using individual sensor inputs in contrast to the case where the information from GPS and LiDAR sensor are fused together. We observe that fusion improves the Top-1 prediction accuracy from 74.86% for the best modality, i.e., LiDAR to 78.18% for the fusion of GPS and LiDAR sensors. The weighted F1 score also increases by 3.6% denoting better handling of imbalances in ground-truth, which is common in mmWave beams.

G. Accuracy and End-to-End Latency Analysis

The Raymobtime and NEU datasets have 256 and 64 possible beam pairs each; hence, sweeping the entire codebook elements requires, $145\,ms$ and $25\,ms$, respectively, according to 5G-NR standard (see Sec. VIII-D1). On the other hand, the proposed beam selection method restricts the beam search space to a subset of K beam pairs. We study the trade-off between the accuracy and end-to-end beam selection time versus K in Fig. 12 for both datasets. Note that the 5G-NR standard defines $20\,ms$ waiting window between SS bursts, where each SS burst includes 32 SS

TABLE IX
COMPARISON OF PROPOSED BEST PERFORMING UNIMODAL AND F-DL ARCHITECTURES WITH TWO BENCHMARK DL BASED APPROACHES ON RAYMOBTIME DATASET [32] AND RESULTS ON THE REAL-WORLD NEU DATASET

Methods	Dataset	# Beams	Modalities	Inference	Top-1	Top-2	Top-5	Top-10
Dias *et al.* [24]	Raymobtime (S007)	264	LiDAR	Centralized	$20.5 \pm 1\%$	$25.5 \pm 1\%$	$54.5 \pm 1\%$	$68.5 \pm 1\%$
Klautau *et al.* [23]	Raymobtime (S008)	240	LiDAR	Centralized	$30.5 \pm 1\%$	$43.5 \pm 1\%$	$57.5 \pm 1\%$	$70 \pm 1\%$
Proposed LiDAR Network	Raymobtime (S008)	256	LiDAR	Centralized	46.23%	64.67%	82.43%	89.95%
Proposed F-DL	Raymobtime (S008)	256	GPS, Image, LiDAR	Distributed	**56.22%**	**74.08%**	**85.53%**	**91.11%**
	NEU	64	GPS, LiDAR	Distributed	**78.18%**	**91.02%**	**98.02%**	**99.37%**

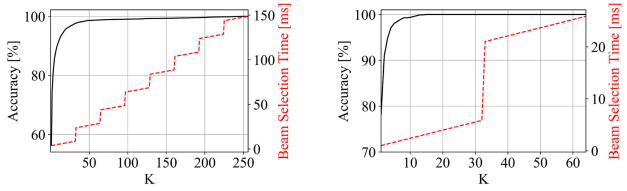


Fig. 12. Beam selection accuracy and end-to-end beam selection time versus K on the (a) Raymobtime and (b) NEU datasets.

blocks for sweeping 32 beam pairs (see Eq. (17)). This results in sudden increments in beam selection time at intervals of every 32 beams, observed in Fig. 12. We also notice that for the Raymobtime dataset the accuracy is $> 99\%$ for $K > 87$ while the end-to-end latency is still increasing. On the other hand, for the NEU dataset, the accuracy and end-to-end beam selection time starts with 78.18% and 3.818 ms for $K = 1$. The accuracy saturates at $K = 7$ and reaches $\sim 100\%$ for $K > 12$ while the beam selection time keeps on increasing and becomes 25.86 ms for $K = 64$. Specifically, Fig. 12 highlights the importance of the K selection method to choose the appropriate K and avoid unnecessary overhead imposed on the system.

H. Comparison With the State-of-The-Art

In Table IX, we compare the performance of our proposed models to the state-of-the-art DL based approaches by Klautau *et al.* [23] and Dias *et al.* [24], both evaluated on the Raymobtime dataset. To the best of our knowledge, these are the only methods that include equivalent scenarios to the ones considered in this paper. In particular, LiDAR sensor data collected on vehicles is used for beam prediction under both LOS and NLOS conditions. Other works that consider different evaluation metrics ([21], [48]), camera images under LOS-only scenarios ([22], [49]) or RF data [50] have been kept out of the comparison. As we show in Table IX, the proposed LiDAR model and the F-DL architecture outperform the state-of-the-art ([23], [24]) by 18.95-20.45% and 20.11-21.61% respectively in top-10 accuracy.

I. Discussion

We summarize below interesting observations from the experimental results:

- When LiDAR and GPS sensors are deployed over the vehicle and features are transmitted to the BS through sub-6 GHz data channel, the wireless control channel may impact the actual delivery at the MEC. On the other hand, cameras at the BS may have a reliable fiber connectivity to the MEC. Hence, in case of unreliable channel conditions or faulty sensors, our fusion framework is still able to make predictions based on any available sensor modality. This robustness to unreliable channel conditions is essential, even if there is no immediate gain from fusing a specific type of modality.
- Proposed beam selection technique with dynamically chosen K automatically selects the top-K best beam pairs, with performance closed to a fixed K when the latter is identified via expert knowledge. Thus our approach eliminates the need to include expert domain knowledge (know what K is needed to achieve certain amount of accuracy), by automating the beam selection process.
- We show that it is possible to reduce the beam-selection overhead in a practical and emerging 5G-NR standard by 95–96%, while maintaining 97.95% relative throughput.

X. CONCLUSION

Increasing softwarization and ability to automatically configure parameters [51] within 5G and beyond networks will necessitate the use of ML-based methods distributed at the MEC. In this paper, we propose an approach for ML-aided fast beam selection technique, where multimodal non-RF sensor data is exploited to reduce the search space for identifying best performing mmWave beam. Our proposed fusion method exploits the latent embeddings from each unimodal feature representation and the overall framework is evaluated in realistic emulated settings. We observe around 20-22% increase in performance for top-10 accuracy than the state-of-the-art using the proposed F-DL architecture. We also achieve 95–96% decrease in beam selection time compared to the exhaustive search defined by the 5G-NR standard in the high-mobility urban scenarios. We propose to extend this framework ahead to multiple-receiver scenarios [52], incorporate federated learning among the sensors [53], and handle different codebook sizes.

APPENDIX A
OBJECT DETECTION ALGORITHM

Our proposed image preprocessing step is a combination of a standard multi-object detection approach followed by a refinement step where each detected object is denoted by a unique indicator according to their role, i.e., target receiver or obstacle. It is constituted of a classifier that is capable to predict the presence of objects in the small bounding boxes. In the training phase, we separately label the examples from the

valid items in the environment. We then quantize the samples by filtering the images with a moving square-shaped window of size $W \times W$ pixels. Starting from the top left side of the image, and after generating the first crop, we move the window by X pixels. This process results in a dataset of cropped samples from each of possible items in the environment. Since the dimensions of items vary, we end up with different number of samples for each class. To achieve a balanced dataset, we augment the minority classes by applying different light conditions, until we reach the same number of samples per class. We split the final balanced dataset in (70%,15%,15%) proportion, and train the classifier.

Similarly, in testing phase, we quantize the image by sweeping it with a window of dimension $W \times W$ and step size X. Next, we feed each crop to the trained classifier and arrange the predictions in the same order as the crop generation. This process leads to a quantized representation of the image, where each element gives the prediction of the classifier for the object in the corresponding $W \times W$ window. We refer to this representation as the *bit map* of the raw input camera images. Given an input image with dimension $H \times L$, the shape of generated bit map will be $\lfloor \frac{H-W}{X} + 1 \rfloor \times \lfloor \frac{L-W}{X} + 1 \rfloor$.

We can refine our bit map further if the specific vehicle type is also transmitted directly by the receiver, as part of the basic safety message in IEEE 802.11p standard for instance. Therefore, given the generated bit map and the reported type of the target vehicle, we (i) keep the label of legitimate receiver vehicle type, (ii) map other vehicles to obstacles. This process designates the potential location of the target receiver as well as the location of obstacles with much more information than the raw images. Finally, to address the concern that the image preprocessing may introduce significant delay as it requires multiple forward passes, we convert the trained model to an equivalent fully convolutional network. We have previously explored such an approach in [54], which enables us to generate the entire bit map in a single forward pass.

APPENDIX B
EVALUATION METRICS

Top-K accuracy calculates the percentage of times that the model includes the correct prediction among the top-K probabilities. Given ground-truth beam pair (t^*, r^*) and the prediction score $S \in \mathbb{R}^{|\mathcal{B}|}$, top-$K$ accuracy is defined as:

$$Acc@K = \frac{1}{N'_t} \sum_{l=1}^{N'_t} 1_{((t^*,r^*) \in A' | \underset{A' \subset \{1,\ldots,|\mathcal{B}|\}, |A'|=K}{\arg\max} \sum_{j \in A'} s_j)}. \quad (19)$$

where N'_t denotes the number of test samples and ϕ is a Boolean predicate, with 1_ϕ to be 1 if ϕ is true, and 0 otherwise. For $K = 1$ we get the conventional top-1 accuracy that only the highest probability prediction is taken into account.

The F1 score measures a model's ability to perform with imbalanced class distribution and defined as the harmonic mean of precision and recall given as $F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$. Precision denotes how many of the predicted true labels are actually in the ground-truth, while recall denotes how many of the actual labels are predicted. To combine the per-class F1 scores into a multi-class version, we weight the F1-score, precision and recall of each class by the number of samples from that class. The KL divergence measures the divergence of the predicted probability distribution from the true one. Given the one-hot encoding $y \in \mathbb{R}^{|\mathcal{B}|}$ of the ground-truth labels and the prediction $\hat{y}$, KL divergence is defined as $KL(\hat{y}||y) = \sum_{i=1}^{|\mathcal{B}|} \{\hat{y}_i \log \frac{\hat{y}_i}{y_i}\}$. Finally, we evaluate the performance of our fusion based beam selector with respect to achieved throughput ratio that is defined as $R_T = \frac{1}{N'_t} \sum_{n=1}^{N'_t} \frac{\log_2[1+y_{\widehat{(t^*,r^*)}}(n)]}{\log_2[1+y_{(t^*,r^*)}(n)]}$, where (t^*, r^*) and $\widehat{(t^*, r^*)}$ show the best beam pair in $\mathcal{B}$ and $\mathcal{B}_k$ (as defined in Sec. III-A and III-B), respectively, and N'_t is the total number of test samples.

REFERENCES

[1] J. Choi, V. Va, N. González-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, Dec. 2016.

[2] I. Rasheed, F. Hu, Y. Hong, and B. Balasubramanian, "Intelligent vehicle network routing with adaptive 3D beam alignment for mmWave 5G-Based V2X communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 5, pp. 2706–2718, May 2021.

[3] W. Roh *et al.*, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.

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

[5] Y. Yaman and P. Spasojevic, "Reducing the LOS ray beamforming setup time for IEEE 802.11 ad and IEEE 802.15. 3c," in *Proc. IEEE Mil. Commun. Conf.*, 2016, pp. 448–453.

[6] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surv. Tut.*, vol. 21, no. 1, pp. 173–196, Jan.–Mar. 2019.

[7] L. Kong, M. K. Khan, F. Wu, G. Chen, and P. Zeng, "Millimeter-wave wireless communications for IoT-Cloud supported autonomous vehicles: Overview, design, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 62–68, Jan. 2017.

[8] "Yole Dévelopment," [Online]. Available: http://www.yole.fr

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

[10] D. Roy, Y. Li, T. Jian, P. Tian, K. R. Chowdhury, and S. Ioannidis, "Multi-modality sensing and data fusion for multi-vehicle detection," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2022.3145663.

[11] C. Premebida, G. Monteiro, U. Nunes, and P. Peixoto, "A lidar and vision-based approach for pedestrian and vehicle detection and tracking," in *Proc. IEEE Intell. Trans. Syst. Conf.*, 2007, pp. 1044–1049.

[12] A. Festag, "Standards for vehicular communication-from IEEE 802.11 p to 5G," *e & i Elektrotechnik und Informationstechnik*, vol. 132, no. 7, pp. 409–416, 2015.

[13] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2016, pp. 1–6.

[14] S. Sur, I. Pefkianakis, X. Zhang, and K.-H. Kim, "WiFi-assisted 60 GHz wireless networks," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, 2017, pp. 28–41.

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

[16] S. Wang, J. Huang, and X. Zhang, "Demystifying millimeter-wave V2X: Towards robust and efficient directional connectivity under high mobility," in *Proc. 26th Annu. Int. Conf. Mobile Comput. Netw.*, 2020, pp. 1–14.

[17] T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, "Steering with eyes closed: mm-Wave beam steering without in-band measurement," in *Proc. IEEE Conf. Comput. Commun.*, 2015, pp. 2416–2424.

[18] N. González-Prelcic, R. Méndez-Rial, and R. W. Heath, "Radar aided beam alignment in mmWave V2I communications supporting antenna diversity," in *Proc. Inf. Theory Appl. Workshop*, 2016, pp. 1–7.

[19] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 852–855, Oct. 2018.

[20] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2018, pp. 2393–2401.

[21] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath, "Inverse multipath fingerprinting for millimeter wave V2I beam alignment," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4042–4058, May 2017.

[22] M. Alrabeiah, J. Booth, A. Hredzak, and A. Alkhateeb, "ViWi vision-aided mmWave beam tracking: Dataset, task, and baseline solutions," 2020, *arXiv:2002.02445*.

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

[24] M. Dias, A. Klautau, N. González-Prelcic, and R. W. Heath, "Position and LIDAR-aided mmWave beam selection using deep learning," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun.*, 2019, pp. 1–5.

[25] G. Reus-Muns *et al.*, "Deep learning on visual and location data for V2I mmWave beamforming," in *Proc. 17th Int. Conf. Mobility, Sens. Netw.*, 2021, pp. 559–566.

[26] 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 Commun. Mag.*, vol. 52, no. 12, pp. 132–141, Dec. 2014.

[27] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "Standalone and non-standalone beam management for 3GPP NR at mmWaves," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 123–129, Apr. 2019.

[28] D. Roy *et al.*, "Going beyond RF: How AI-enabled multimodal beamforming will shape the nextg standard," 2022, *arXiv:2203.16706*.

[29] Z. Wang, B. Salehi, A. Gritsenko, K. Chowdhury, S. Ioannidis, and J. Dy, "Open-world class discovery with kernel networks," in *Proc. IEEE Int. Conf. Data Mining*, 2020, pp. 631–640.

[30] R. Heinzler, P. Schindler, J. Seekircher, W. Ritter, and W. Stork, "Weather influence and classification with automotive lidar sensors," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 1527–1534.

[31] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal fusion architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6959–6968.

[32] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang, and R. W. Heath, "5G MIMO data for machine learning: Application to beam-selection using deep learning," in *Proc. Inf. Theory Appl. Workshop*, 2018, pp. 1–9.

[33] P. A. Lopez *et al.*, "Microscopic traffic simulation using SUMO," in *Proc. Int. Conf. Intell. Transp. Syst.*, 2018, pp. 2575–2582.

[34] Blender. [Online]. Available: https://www.blender.org

[35] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "BlenSor: Blender Sensor Simulation Toolbox," in *Proc. Adv. Visual Comput.* Springer Berlin Heidelberg, 2011, pp. 199–208.

[36] Remcom, "Wireless insite," [Online]. Available: http://www.remcom.com/wireless-insite

[37] D. Steinmetzer, D. Wegemer, M. Schulz, J. Widmer, and M. Hollick, "Compressive millimeter-wave sector selection in off-the-shelf IEEE 802.11ad devices," in *Proc. Int. Conf. Emerg. Netw. EXperiments Technol.*, 2017, pp. 414–425.

[38] [Online]. Available: https://genesys-lab.org/multimodal-fusion-nextg-v2x-communications

[39] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] J. Carter *et al.*, "An introduction to LiDAR technology, data, and applications," *NOAA Coastal Serv. Center*, vol. 2, 2012.

[43] Q. Wang, S. Leng, H. Fu, and Y. Zhang, "An IEEE 802.11p-based multichannel MAC scheme with channel coordination for vehicular ad hoc networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 449–458, Jun. 2012.

[44] S. Sesia, I. Toufik, and M. Baker, *LTE-The UMTS Long Term Evolution: From Theory to Practice*. Hoboken, NJ, USA: John Wiley, 2011.

[45] C. N. Barati, S. Dutta, S. Rangan, and A. Sabharwal, "Energy and latency of beamforming architectures for initial access in mmWave wireless networks," *J. Indian Inst. Sci.*, vol. 100, no. 2, pp. 281–302, 2020.

[46] M. A. Richards, *Fundamentals of Radar Signal Processing*. New York, NY, USA: McGraw-Hill, 2014.

[47] G. R. Muns, K. V. Mishra, C. B. Guerra, Y. C. Eldar, and K. R. Chowdhury, "Beam alignment and tracking for autonomous vehicular communication using IEEE 802.11 ad-based radar," in *Proc. IEEE Conf. Comput. Commun. Workshops*, 2019, pp. 535–540.

[48] J. C. Aviles and A. Kouki, "Position-aided mm-Wave beam training under NLOS conditions," *IEEE Access*, vol. 4, pp. 8703–8714, 2016.

[49] Y. Tian, G. Pan, and M.-S. Alouini, "Applying deep-learning-based computer vision to wireless communications: Methodologies, opportunities, and challenges," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 132–143, Dec. 2021.

[50] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *Proc. IEEE 91st Veh. Technol. Conf.*, 2020, pp. 1–5.

[51] K. Li, U. Muncuk, M. Y. Naderi, and K. R. Chowdhury, "iSense: Intelligent object sensing and robot tracking through networked coupled magnetic resonant coils," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6637–6648, Apr. 2021.

[52] A. Vizziello, P. Savazzi, and K. R. Chowdhury, "A kalman based hybrid precoding for multi-user millimeter wave MIMO systems," *IEEE Access*, vol. 6, pp. 55 712–55 722, Sep. 2018.

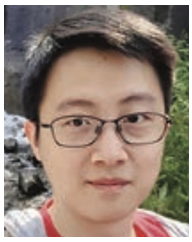
[53] B. Salehi, J. Gu, D. Roy, and K. Chowdhury, "FLASH: Federated learning for automated selection of high-band mmWave sectors," in *Proc. IEEE Conf. Comput. Commun.*, 2022.

[54] B. Salehi, M. Belgiovine, S. G. Sanchez, J. Dy, S. Ioannidis, and K. Chowdhury, "Machine learning on camera images for fast mmWave beamforming," in *Proc. IEEE 17th Int. Conf. Mobile Ad Hoc Sensor Syst.*, 2020, pp. 338–346.

Batool Salehi received the M.S. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2019. She is currently working toward the Ph.D. degree in computer engineering with Northeastern University, Boston, MA, USA, under the supervision of Prof. K. Chowdhury. Her research interests include mmWave beamforming, Internet of Things, and the application of machine learning in the domain of wireless communication.

Guillem Reus-Muns received the B.Sc. degree in telecommunications engineering from the Polytechnic University of Catalonia (UPC-BarcelonaTech), Barcelona, Spain, the M.Sc. degree in electrical and computer engineering from Northeastern University, Boston, MA, USA, where he is currently working toward the Ph.D. degree. He joined Northeastern University, in 2018. His research interests include cellular networks, machine learning for wireless communications, networked robotics, and spectrum access.

Debashri Roy received the M.S. and the Ph.D. degrees in computer science from the University of Central Florida, Orlando, FL, USA, in 2018 and 2020, respectively. She is currently an experiential AI Postdoctoral Fellow with Northeastern University, Boston, MA, USA. Her research interests include AI/ML enabled technologies in wireless communication, multimodal data fusion, network orchestration, and nextG networks.



Zifeng Wang (Graduate Student Member, IEEE) received the B.Sc. degree in electronic engineering from Tsinghua University, China, in 2014. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA. He works under the guidance of Prof. Jennifer Dy in machine learning. His research interests include lifelong learning, representation learning, and the application of machine learning in the domain of biostatistics, and wireless communication.

Tong Jian received the M.Sc. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, 2016. She is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA. She works under the guidance of Prof. Stratis Ioannidis in the field of machine learning. Her research focuses on the application of machine learning in the domain of wireless communication.

Jennifer Dy (Member, IEEE) is currently a Professor with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA, where she first joined the Faculty in 2002. Her research interests inlcude fundamental research in machine learning and their application to biomedical imaging, health, science and engineering, with research contributions in unsupervised learning, dimensionality reduction, feature selection, learning from uncertain experts, active learning, Bayesian models, and deep representations.

Stratis Ioannidis is currently an Associate Professor with the Electrical and Computer Engineering Department of Northeastern University, Boston, MA, USA, where he also holds a courtesy appointment with the Khoury College of Computer Sciences, Boston, MA, USA. Prior to joining Northeastern, he was a Research Scientist with the Technicolor Research Centers in Paris, France, and Palo Alto, CA, USA, and also with Yahoo Laboratories, Sunnyvale, CA. His research interests include machine learning, distributed systems, networking, optimization, and privacy.

Kaushik Chowdhury (Senior Member, IEEE) is currently a Professor with Northeastern University, Boston, MA, USA. He is currently the Co-Director of the Platforms for Advanced Wireless Research project office. His research interests include systems aspects of networked robotics, machine learning for agile spectrum sensing/access, wireless energy transfer, and large-scale experimental deployment of emerging wireless technologies.