Going Beyond RF: How AI-enabled Multimodal Beamforming will Shape the NextG Standard

Debashri Roy*, Batool Salehi*, Stella Banou*, Subhramoy Mohanti*, Guillem Reus-Muns*, Mauro Belgiovine*, Prashant Ganesh†, Carlos Bocanegra*, Chris Dick‡, and Kaushik Chowdhury*

* Electrical and Computer Engineering
Northeastern University
Boston, MA 02115
droy, bsalehihikouei, sbanou, smohanti,
greusmuns, mbelgiovine, cbocanegra, krc}@ece.neu.edu

† Mechanical and Aerospace Engineering
University of Florida REEF
Shalimar, FL 32579
prashant.ganesh@ufl.edu

‡ Nvidia Inc.
cdick@nvidia.com

Abstract—Incorporating artificial intelligence and machine learning (AI/ML) methods within the 5G wireless standard promises autonomous network behavior and ultra-low-latency reconfiguration. However, the effort so far has purely focused on learning from radio frequency (RF) signals. Future standards and next-generation (nextG) networks beyond 5G will have two significant evolutions over the state-of-the-art 5G implementations: (i) massive number of antenna elements, scaling up to hundreds-to-thousands in number, and (ii) inclusion of AI/ML in the critical path of the network reconfiguration process that can access sensor feeds from a variety of RF and non-RF sources. While the former allows unprecedented flexibility in ‘beamforming’, where signals combine constructively at a target receiver, the latter enables the network with enhanced situation awareness not captured by a single and isolated data modality. This survey presents a thorough analysis of the different approaches used for beamforming today, focusing on mmWave bands, and then proceeds to make a compelling case for considering non-RF sensor data from multiple modalities, such as LiDAR, Radar, GPS for increasing beamforming directional accuracy and reducing processing time. This so-called idea of multimodal beamforming will require deep learning based fusion techniques, which will serve to augment the current RF-only and classical signal processing methods that do not scale well for massive antenna arrays. The survey describes relevant deep learning architectures for multimodal beamforming, identifies computational challenges and the role of edge computing in this process, dataset generation tools, and, finally, lists open challenges that the community should tackle to realize this transformative vision of the future of beamforming.

Index Terms—beamforming, beam selection, beam search, mmWave, multimodal, non-RF data, fusion, 5G, NextG.

I. INTRODUCTION

Today’s ultra-connected world is demanding high bandwidths, ultra-low latency, and autonomous network reconfiguration to accommodate new applications, heterogeneous devices and shared spectrum use. The number of users is also increasing at unprecedented levels, with predictions of the number of networked devices exceeding $3x$ the global population by 2023 [1]. To serve bandwidth-hungry application needs, the expected maximum 5G data rate is now revised to be $13x$ faster in 2023, a significant revision from earlier estimations made only a few years ago in 2018 [1]. Many exciting applications will leverage such high capacity wireless networks, such as relaying high-resolution three dimensional (3D) graphical content, VR/AR streams [2], vehicle-to-everything (V2X) links leading towards autonomous cars [3], [4], among others.

A key underlying technology that is essential for all of the above is transmit beamforming, where signals from multiple antenna elements combine constructively at the receiver. Consider a multi-antenna radio, with each of these antenna elements having a specific directional radiation pattern, referred to as a beam. The beams from transmitter and receiver antennas are steered to initiate communication via beamforming [5]. The communication link is then established through the periodic beam sweeping and beam selection [6]. Beamforming increases the signal strength at the receiver, which in turn raises the capacity limit, mitigates interference by avoiding undesirable signals at neighboring receivers, and combats the effect of pronounced path loss at high frequencies. Thus,
beamforming is considered as a critical component of all modern WiFi [7] standards and is steadily being integrated into 5G [8].

Our survey is motivated by this observation, and we strive to answer the following two questions: (i) are there fundamental limitations of traditional RF-only beamforming technology that will impact future standards evolution, and (ii) how can new data types (beyond RF) be harnessed in the future, and, given the possible information explosion by acquiring such multimodal sensor feeds, can they be analyzed through emerging machine learning methods to guide real-time beamforming decisions? To ensure a focused discussion, we emphasize use-cases that will shape the future standards beyond 5G (henceforth referred to as NextG), namely, beamforming scenarios that combine very large number of antenna elements and mobility. As an indicative example of a mmWave vehicular network that we cover in this survey, Fig. 1 shows moving vehicles beamforming towards a static base station by combining data from RF and non-RF modalities, and then using ML to identify a smaller set of beam-pairs for further optimization, instead of an exhaustive search. We begin our discussion by highlighting the need for beamforming with massive number of antennas and the use of AI/ML in beamforming communication systems.

**Need for Beamforming in NextG Standards:** The 5G New Radio (5G-NR) standard provisions for use of both sub-6 GHz as well as millimeter wave (mmWave) frequency bands from 24.25 GHz to 52.6 GHz [8]. The sub-6 GHz band is already congested, and this problem worsens when a large data transfer needs to occur at short contact times, typically seen in mobile environments with few antennas [6]. While mmWave-band transmission increases capacity using wider bandwidth (up to 2 GHz) [9], it also suffers from severe attenuation and penetration loss [6]. Phased-array antennas [10] address the attenuation problem by leveraging the highly directional gain of the antenna elements, thereby focusing radiated RF energy into beams. This capability is enhanced in higher frequencies given the dense packing of antenna elements, i.e., higher order phased arrays are possible with proportional increase in the number of beams. While theoretically hundreds of antenna elements can be packed in a 1cm × 1cm area for mmWave band operation, the bottleneck lies in the complexity of processing methods and the computational resource available to properly configure the beams. Even though it is economically feasible to create large phased arrays, scaling beyond 8-12 antennas while supporting real-time operation in small form factor wireless devices still remains an open challenge. Thus, there is need to revisit existing approaches to beamforming to potentially scale up to thousands of antenna elements, as is being envisaged in NextG standards [11].

**Beamforming with Massive Number of Antenna Elements:** While the possibility of having multiple antennas resulted in the multiple input and multiple output (MIMO) paradigm in WiFi networks starting with 802.11n, these were limited to 4x4 MIMO. Massive MIMO (mMIMO) scales this number up to the order of hundreds of antennas at least at the base station (BS) side, offering great flexibility in forming single directional beams [12] as well as multi-beams that can simultaneously target multiple users in what is known an multi-user MIMO (MU-MIMO) [13]. The implementation of mMIMO depends on acquiring accurate channel state information (CSI), which is then used to compensate for wireless channel distortions in a method called as precoding. The legacy CSI schemes inherited from 4G LTE consists of either: (a) codebook-based feedback for frequency division duplex (FDD) operation, or (b) reciprocity-based sounding for the time division duplex (TDD) operation at the base station [14].

The higher number of antenna elements in mMIMO increases the complexity of optimal codebook design, while the overhead of CSI measurement from downlink pilot and feedback also increases at the user devices. For these reasons, this survey focuses on ways to facilitate beamforming for mMIMO, going beyond CSI-based methods. Indeed, the many degrees of freedom in codebook design and then active selection of the actual code (i.e., selection of the beam as decided by that code) for a given situation make optimal deterministic and even most heuristic based solutions infeasible for deployment.

**Motivation for using AI-enabled Beamforming:** Artificial intelligence and machine learning (AI/ML) based algorithms have been effectively demonstrated to outperform classical approaches in wireless-centric tasks of modulation recognition [15], RF fingerprinting [16], rogue transmitter detection [17], etc. The use of AI-enabled algorithms to solve the above mentioned beamforming in nextG networks is still in a nascent stage. The general approach so far on using ML involves RF channel estimation followed by channel equalization by using different neural network-based architectures that accept a stream of in/quadrature phase (I/Q) samples collected by the receiver. We believe there is a vast untapped potential for AI-enabled techniques for extracting relevant information using different types of modalities, for e.g., images can recognize the location of the target BS and this can rapidly reduce the number of candidate beams to be explored. We refer to this emerging research trend in the domain of out-of-band beamforming as multimodal beamforming.

**Scope of this Survey:** The statistics presented in Fig. 2 comprise of the number of articles (including patents), from Google Scholar search results, that have referenced the terms beamforming in 5G and beamforming in mmWave. We believe this survey will serve the wireless research community work-
ing on beamforming in high frequency band (30-300 GHz), as in these frequencies, beamforming lies on the critical path to combat signal attenuation. We introduce and analyze the notion of multimodal beamforming for mmWave frequencies by recognizing the existing interest in the intersection of MIMO systems, wireless AI/ML, and the NextG bandwidth needs. Furthermore, we emphasize the vehicular scenario shown in Fig. [1] as it poses challenges caused by mobility that cannot be addressed in feasible time-scales through legacy methods for such large beamforming antenna arrays. As evidence of community interest on this general theme, we see a spike in citations (88 citations within 2 years) for the publicly available dataset called Raymobtime [18], which contains multimodal non-RF sensor data alongwith the corresponding RF ground truths for the purpose of mmWave beamforming in a V2X environment.

While we strive to produce a comprehensive survey on this subject matter, we skip the reviews on the basics of mmWave channel models, mMIMO, different beamforming system models and techniques, as there exist plethora of survey literature focusing on these fundamentals, and is out-of-scope considering our focus area. For example, the promise of mmWave communication in 5G is extensively reviewed in [19], the use of mmWave band for vehicular communication is surveyed in [20], applications of mMIMO are surveyed in [21] and [22], detailed analysis of general RF-only beamforming in indoor and outdoor mmWave communications can be found in [23]. RF-only beamforming can have digital and analog beamforming, as well as hybrid approaches that combine the two. Related models and system architectures that contrast these three approaches are described in [24] and [25]. A flow-graph summarizing the existing surveys related to the “beamforming in 5G/NextG” systems is shown in Fig. [3] and we explore each of these topics in their relevant sections later in this paper. We broadly categorize the trend of existing surveys on that topic in three groups: beamforming techniques for 5G, hybrid beamforming, and out-of-band beamforming; where the first two categories are related to the traditional beamforming process, while the last one is aligned towards out-of-the-box solutions. In this regard, the purpose of this survey is to identify the shortcomings in the traditional beamforming methods and identify the advantages of using non-RF modalities to facilitate the beamforming process, considering nextG communications. Ultimately, we make a case for expanding the research focus towards incorporating such non-RF sensor modalities in combination with AI/ML, as a feasible pathway for NextG networks.

- **Organization of this Survey:** The remainder of this article is organized as follows. High level differences between different traditional and non-RF based beamforming techniques for nextG networks are described in Sec. [II] followed by a comprehensive review of published surveys in related areas of beamforming in Sec. [III]. The drawbacks in the legacy signal processing approaches in the RF domain lead us towards the use of new tools, such as ML-based approaches, which we describe in Sec. [IV]. We then turn towards the use of out-of-band RF and non-RF data for beamforming is presented in Sec. [V] and Sec. [VI] with multiple data modalities available, we present few novel fusion techniques for fast beam selection in Sec. [VII]. We discuss different application areas of multimodal beamforming in Sec. [VIII] with additional emerging trends included in Sec. [IX]. The conclusions are drawn in the last section. Acronyms used in the survey are listed in Tab. [I].

## II. Beamforming Techniques

In this section, we first analyze the state-of-the-art in traditional beamforming techniques that may impair inclusion in future NextG standards operating in mmWave bands. We then explore the current research on non-RF based beamforming to motivate our intent of using these methodologies to address the shortcomings of traditional RF-only beamforming.

### A. Traditional Beamforming

Existing RF-based beamforming approaches (analog, digital, hybrid) have their unique advantages, and are applicable in specific scenarios. Indeed, the 5G-NR standard supports all three types of beamforming in the time domain [23].

A brief comparison study for these approaches is presented in Tab. [II]. Digital beamforming improves the spectral efficiency (SE) of a MIMO system by simultaneously transmitting data to multiple users. However, it needs a distinct RF chain per antenna, making it less cost-effective for higher order of antenna elements. This is one core reasons why there are few off-the-shelf mmWave radios [28] which support digital beamforming even with low order (1x4) of antenna elements. Unlike its digital counterpart, analog beamforming creates the beam using one element per set of antenna. Once the best beam, among all possible combinations of beam-pairs is identified, it is activated to mitigate the impact of high pathloss in mmWave band. This is why most of the off-the-shelf mmWave devices [29], [30], [31] support only analog beamforming. Also, analog beamforming is considered mandatory in 5G-NR [13] for mmWave operation.

Hybrid beamforming, on the other hand, is a combination of analog and digital beamforming. The idea of hybrid beamforming revolves around trading-off the hardware cost for the overhead of time involved in beam selection. Here, a subset of antennas is connected to a particular RF chain, as opposed to having individual RF chains for each antenna element in digital beamforming. Even though, hybrid beamforming promises faster communication with higher order antenna elements, this is still an area on ongoing research [32]. Additionally, for hybrid beamforming, the continuous beam management technique in a mobile environment involves periodic overhead [13]. Here, beam selection is done after the measurement of reference signals (RS) received in a specific direction by manipulating the beamforming weights applied across different antenna elements.

### B. Out-of-band Beamforming

As discussed before, analog beamforming technique involves time-overhead of beam selection due to exhaustive search among all possible transmitter-receiver (TX-RX) antenna elements. The decision is made based on a combination...
Beamforming in 5G/NextG Surveys

5G Networks
- mmWave Channel Models [9]
- mMIMO [21], [22]
- mmWave Communication [19], [20], [23]

Hybrid Beamforming
- Architecture and System Models [25]
- mMIMO [24]
- mmWave beamforming [26]
- Wireless Energy Harvesting [27]

Out-of-band Beamforming
- Out-of-band RF-based (Sec. V)
- Non-RF based (Sec. VI-A - VI-C)
- Multimodal (Sec. VI-D)
- AI-enabled Implementations (Sec. VI - VII)

Fig. 3: Existing surveys on different areas of beamforming for 5G and beyond. This survey mainly focuses on use of out-of-band beamforming in recent literature for the 5G and NextG networks. We describe different categories of out-of-band beamforming in the corresponding sections in the rest of the paper.

Fig. 4: The average beam initialization overhead of IEEE 802.11ad and 5G-NR standards with respect to different beam search spaces. The beamforming time significantly increases with increasing number of beam pairs.

of RF measurements, such as CSI, SNR etc., in the desired frequency band of communication. This overhead (Fig. 4) is exacerbated in the case of mobile users where the position of user equipments (UEs) are changing continuously, resulting in the exhaustive search being instantiated multiple times within a few seconds. Furthermore, the wireless channel varies 10x faster at 30 GHz as opposed to 3 GHz, even for the same UE mobility rate. This results in 10x more frequent beam sweeping and channel estimation [33]. Thus, we believe that out-of-band RF measurements and the use of environmental non-RF data appear offer an attractive alternative towards minimizing the overhead of exhaustive search. We refer to such approaches as out-of-band beamforming techniques. A visual representation of the existing traditional and out-of-band beamforming techniques are given in Fig. 5.

III. BEAMFORMING SURVEYS ON mMIMO FOR 5G AND BEYOND

We now review the state-of-the-art in research and analysis on mMIMO, beamforming, and mmWave communication in 5G and nextG wireless networks. The fundamentals of mMIMO and mmWave operation and the applications of mMIMO are comprehensively surveyed in earlier works [21][22]. The promise of mmWave communication in 5G is extensively reviewed in [19], and the use of mmWave band for vehicular communication is surveyed in [20].

From Fig. 4 we see that the research interest on beamforming in mmWave band and 5G standards are strongly coupled, as the advancements in the former are essential to meet operational requirements for the latter. Additionally, exploration of new spectrum, assigning more bandwidth, carrier aggregation, inter-cell interference mitigation techniques, integration of mMIMO antennas, etc., are all key features that have been extensively covered in [34]. Also, the authors state that providing accessibility, flexibility, and cloud based services through proper modulation and coding scheme (MCS), mmWave and device to device (D2D) communication is the key to realize functional nextG networks. Authors in [26] validate the notion that beamforming has a bigger role to play in mmWave bands, as compared to low frequency bands. Hence, there is great interest in beamforming optimization in mmWave bands for nextG standards.

For sake of completeness, we mention below surveys that describe beamforming advancements tailored for sub-1 GHz, sub-6 GHz as well as sub-30 GHz 5G bands. Authors in [35] focus on the frequency allocation, beamforming techniques and custom-designed integrated circuits for those specific bands. Kutty et al. [23] capture the evolution of different beamforming techniques in the context of mmWave communication. They describe different radio frequency system design and implementation for millimeter wave beamforming for indoor and outdoor communication scenarios. The authors describe the mmWave propagation characteristics in-terms of path loss and clustered multi-path structures, dominant LoS component, wideband communication and 3D spatio-temporal modeling. They also illustrate different phased array antenna architectures to support MIMO capability in mmWave beamforming. Finally, the authors concur that using hybrid beamforming in the mmWave band for MIMO to minimizing
In a survey on hybrid beamforming for mMIMO, Molisch et al. [24] analyze the trade-offs of using instantaneous or average (second-order) CSI in hybrid beamforming. Here, the authors evaluate current research on various types of hybrid multiple-antenna transceivers and consider how the channel sparsity in the mmWave band can be leveraged for optimizing channel estimation and beam training. However, to get broader aspects of hybrid beamforming, we review an extensive survey by Ahmed et al. in [25], which thoroughly track the progress in this domain till 2017. In this paper the authors present different architectures of hybrid beamforming and the techniques for optimization of phase shifters, DAC/ADC resolutions and antenna configurations. From the system model perspective, they examine eight variations of hybrid beamforming and identify many resource management aspects, particularly in beam management, MAC protocol variants, which can impact the performance of hybrid beamforming.

Researchers have also done surveys on beamforming in cognitive radio networks (CRN) [36] and wireless energy harvesting (EH) [27]. The basic principle of achieving high SINR through beamforming makes it a potential candidate for transmitters within EH networks. The efficacy of using CRN and EH for energy-constrained communication networks has been exhibited in [4, 37, 38, 39], along with focused research on the topics of military communications and submarines [27], sensor networks [40, 41], and medical implants [42, 43]. In [27], the authors also advocate for beamforming for CRN in nextG networks. Tab. III summarizes the published surveys related to beamforming in the mmWave frequency band. However, the goal of this survey paper is to describe beamforming techniques that exploit out-of-band RF and non-RF multimodal data for nextG networks. To motivate the case for out-of-band RF and multimodal data, we first identify the limitations of traditional beamforming methods using RF-only data.

IV. LIMITATIONS OF TRADITIONAL RF-ONLY BASED APPROACHES

The traditional RF-only beamforming approach utilizes one of these two options: (a) estimate the mmWave channel at the receiver, and send this information back to the transmitter for generating the precoding weights, (b) sweep through the antenna codebook elements of the transmitter and receiver. In this section, we discuss how these RF-only solutions for beamforming impose significant overhead for mmWave links. For the first option, we discuss the published literature related to continuous channel estimation and closed loop feedback to the transmitter for beamforming in mmWave mMIMO. For the second option, we review works on continuous beam sweeping and beam alignment needed for sustainable communication, and introduce the out-of-band beamforming as a solution to overcome the expensive channel estimation and beam sweeping tasks.

A. Channel Estimation

Modern MIMO wireless communication systems use spatial multiplexing to improve the data throughput in a rich scattering environment. In order to send multiple data streams through the channel, a set of precoding and combining weights are derived from the channel matrix to recover each data stream independently. These weights contain both the magnitude and phase of the channel and are normally applied in the digital domain. Depending on the application, uplink and downlink communication is performed either in Time Division Duplex (TDD) or Frequency Division Duplex (FDD). While FDD allows for full duplex wireless interfaces, TDD is more practical in dense cellular deployments.

Accurate channel estimation is challenging in mmWave mMIMO systems due to the sheer magnitude of number of antennas, low SNR channels, hardware constraints, etc. The basic idea behind channel estimation is to acquire the most current CSI at the receiver and forward it to the transmitter [44] within some application specific latency bound so as to allow both the channel sounding and data transfer phases to be completed within the channel coherence time. Such stringent thresholds on CSI feedback latency ensure that the transmitter can turn around its radio front-end and leverage channel reciprocity for the downlink transmission. CSI is also key for realizing spatial multiplexing, where independent paths are available in the channel between the transmitter and receiver. Typically, perfect channel estimation is assumed in literature, which is hard to achieve in practical real-life deployments [44]. While channel estimation is typically accomplished via classical methods, in recent years, ML based methods are being researched to overcome the limitations of these classical approaches. We next summarize the fundamental differences between these two approaches.

1) Traditional Channel Estimation: Least square (LS) estimation is one of the simplest and fastest channel estimators,
Although its performance is affected by high mean squared error (MSE) at low SNR levels. In such cases, an additional filtering based on minimum mean squared estimation (MMSE), typically a linear MMSE (LMMSE) [45], is adopted to improve the LS estimation by filtering out the noise. Even in its linear form, MMSE turns out to be an expensive and scales poorly [45]. This calls for further research on more efficient methods for mMIMO systems.

2) ML-based Channel Estimation: ML and, in particular, deep learning (DL) are increasingly considered for channel estimation in many areas of wireless communication [44]. An end-to-end orthogonal frequency-division multiplexing (OFDM) symbol decoding method using MLP is presented by [50] through the process of treating a single input single output (SISO) channel model as a black box.

Applying DL based approaches for CSI estimation in mMIMO is still at a developing stage. Given the high dimensionality in mMIMO, especially when involving OFDM techniques, the majority of existing solutions use complex and deep architectures to estimate large channel matrices. These solutions treat the multi-dimensional input signal as a single entity and often require additional prior or post-estimation steps. Although use of very deep architectures is a growing trend, their complexity usually limits use in edge devices that are typically constrained in power and processing capability. Dong et al. [47] use convolutional neural networks (CNN) to improve the quality of a coarse initial estimate of the channel matrix in a method called tentative estimation. To exploit adjacent sub-carrier frequency correlations, the coarse channel estimate matrices are concatenated in large input tensors and processed by a neural network consisting of 10 convolutional layers. He et al. [48] propose a 10-layer learned denoising-based approximate message passing (LDAMP) architecture, based on the unfolding of the iterative D-AMP algorithm. As the estimated channel is treated as a noisy 2D image, each layer relies on an additional denoising CNN, which is 20-layers deep and is used to update the channel estimated in the previous layer. Although CNNs are efficient in terms of number of parameters, the resulting complexity poses a challenge at the edge for the deep architectures, like those proposed in [47] and [48]. In the context of single-carrier systems [40] devises an uplink transmission for single antenna users and multi-antenna BS using a 6-layer MLP to first estimate the direction of arrival (DoA) and then determine the channel for each user, by expressing the channel estimate as a function of the DoA and solving an additional linear system of equations. Balevi et al. [49], describe an online

| Acronyms | Meanings |
|----------|----------|
| 2D       | Two Dimensional |
| 3D       | Three Dimensional |
| 5G       | 5th Generation |
| 5G NR    | 5th Generation New Radio |
| AI       | Artificial Intelligence |
| AoA      | Angle of Arrival |
| AP       | Access Point |
| AR       | Augmented Reality |
| BBS      | Blind Beam Steering |
| BS       | Base Station |
| CDF      | Cumulative Distribution Function |
| CNN      | Convolutional Neural Network |
| CRN      | Cognitive Radio Network |
| CS       | Compressive Sensing |
| CSI      | Channel State Information |
| DL       | Deep Learning |
| DoA      | Direction of Arrival |
| FDD      | Frequency Division Duplex |
| EH       | Energy Harvesting |
| EHF      | Extremely High Frequency |
| FANET    | Flying Ad-hoc NETworks |
| FML      | Fast Machine Learning |
| IA       | Initial Access |
| ID       | Identification |
| IoT      | Internet of Things |
| IR       | Infrared |
| LiDAR    | Light Detection and Ranging |
| LMMSE    | Linear Minimum Mean Squared Estimation |
| LSTM     | Long Short-Term Memory |
| MAB      | Multi Armed Bandit |
| MCS      | Modulation and Coding Scheme |
| MIMO     | Multiple Input Multiple Output |
| ML       | Machine Learning |
| MLP      | Multi-layer Perceptron |
| MMSE     | Minimum Mean Squared Estimation |
| mMIMO    | massive MIMO |
| MU-MIMO  | Multiple User MIMO |
| mmWave   | millimeter Wave |
| MR       | Mixed Reality |
| MSE      | Mean Squared Error |
| nExG     | Next Generation |
| NOMA     | Non-Orthogonal Multiple Access |
| NLOS     | Non Line of Sight |
| OFDM     | Orthogonal Frequency Division Multiplexing |
| RADAR    | Radio Detection And Ranging |
| RF       | Radio Frequency |
| RGB      | Red Green Blue |
| RMSE     | Root Mean Squared Error |
| RS       | Reference Signal |
| RSU      | Road Side Unit |
| RX       | Receiver |
| LoS      | Line of Sight |
| SE       | Spectral Efficiency |
| SHF      | Super High Frequency |
| SINR     | Signal-to-interference Noise Ratio |
| SISO     | Single Input Single Output |
| SLAM     | Simultaneous Localization And Mapping |
| SS       | Synchronization Signals |
| TDD      | Time Division Duplex |
| THz      | Terahertz |
| TX       | Transmitter |
| UAV      | Unmanned Aerial Vehicles |
| UE       | User Equipment |
| UGV      | Unmanned Ground Vehicles |
| V2I      | Vehicle to Infrastructure |
| V2X      | Vehicle to Everything |
| VR       | Virtual Reality |
| WID      | Wireless Infrastructure Drone |
| WiGig    | Wireless Gigabit |
| XR       | eXtended Reality |

TABLE I: The details of the frequent acronyms used in the article.

TABLE II: Comparison of different beamforming types.

| Metrics                  | Analog | Digital | Hybrid |
|--------------------------|--------|---------|--------|
| Degree of Freedom        | Limited| Highest | High   |
| Implementation           | Phase Shifter | AI/DSM | Everything |
| Architecture             | Simple | Complex | Complex |
| Baseband Chains          | Less   | Highest | High   |
| Complexity               | Less complex | Complex | Complex |
| Power Consumption        | Less   | Highest | High   |
| Cost                     | Less   | Highest | High   |
| Inter-user interference   | High   | Lowest  | Low    |
| MIMO Support             | No     | Yes     | Yes    |
| Flexibility              | Fixed delay | Flexible weight vector | Flexible weight vector |
TABLE III: Comparative analysis of existing surveys for beamforming in mmWave band.

| Paper | Year | Beamforming in mmWave | Inclusion of 5G NR Standard | Supports MIMO | Shortcomings of Traditional mmWave Beamforming | Aiding the Traditional mmWave Beamforming | non-RF Data for Beamforming |
|-------|------|------------------------|-----------------------------|--------------|-----------------------------------------------|-------------------------------------------|----------------------------|
| [23]  | 2019 | √                      | ×                           | ×            | ×                                             | ×                                         | ×                          |
| [23]  | 2016 | √                      | ×                           | ×            | ×                                             | ×                                         | ×                          |
| [23]  | 2017 | √                      | ×                           | ×            | ×                                             | ×                                         | ×                          |
| [23]  | 2018 | √                      | ×                           | ×            | ×                                             | ×                                         | ×                          |
| This  | 2021 | √                      | ×                           | ×            | ×                                             | ×                                         | ×                          |

TABLE IV: A coarse computational complexity comparison between existing methods and proposed channel estimator. Notation: \(N_T\) = number of transmitter antennas, \(N_R\) = number of receiver antennas, \(K\) = number of sub-carriers, \(L\) = number of hidden layers, \(I_i\) = number of input features of layer \(i\)-th, \(N_i\) = number of neurons (or kernels, in case of CNNs) in \(i\)-th layer, \(F_i\) = kernel size of \(i\)-th convolutional layer (assuming square kernels), \(W_i\) = width of input volume for \(i\)-th convolutional layer, \(H_i\) = height of input volume of \(i\)-th convolutional layer, \(E\) = number of epochs, \(L\) = complexity of LDAMP layer (linear system) in [48], \(T\) = complexity of tentative estimation (linear system, including matrix multiplications and inversions) in [47], \(G\) = complexity of additional linear system needed to compute complex channel coefficients from DoA estimation (requires matrix inversion) in [45]. † = method requires OFDM demodulation, § = method does not require OFDM demodulation.

training method based on deep image prior scheme, using a 6-layer architecture based on \(1 \times 1\) convolutions and upsampling, which performs denoising of the received signal before a traditional LS estimation. Although the number of parameters here is low, this method requires the network to be trained during every transmission for thousands of epochs, without any guarantee that this training step will be completed within the channel coherence time. For single-carrier solutions, \(K\) separate models, where \(K\) is the number of sub-carriers, should be trained for deployment in practical OFDM systems, resulting additional complexity over the LS estimation.

Recently, Belgiovine et al. [44], demonstrated an edge-oriented MLP with compact architecture that exploits similarities in each transmit-receive antenna pairs to estimate their channels at each sub-carrier independently. Due to the inherently parallel nature, DL models can complete channel estimation process with improved quality of estimation and reduced computational time. Tab. IV summarizes the time complexity of existing methods and compares how this DL approach results in a much simpler model that is suitable for edge architectures.

B. Feedback from Receiver

The next step after the CSI estimation is to send the feedback to transmitter from the receiver. If TDD is employed, then there are two phases involving the BS and UE: channel sounding, in which case the UE performs CSI estimation for the complete MIMO channel and sends it back to the BS, and data transfer, where the BS uses the received CSI estimation to compute the precoder and combiner’s weights for directional beams. On the other hand, FDD schemes allow for the upload of CSI to BS on a dedicated band. Transferring the entire CSI is impractical considering the available channel bandwidth, and therefore, its relative information increases linearly with the number of transmitter antennas in mMIMO systems. Methods based on compressive sensing (CS) [51], [52], [53], [54], [55], [56] focus on reducing feedback overhead by using spatial and temporal correlation of CSI. In particular, correlated CSI can be transformed into an uncorrelated sparse vector in some cases and CS can be used to obtain a sufficiently accurate estimate of such sparse vectors. However, channels are not usually sparse and may not always have an interpretable structure. In order to overcome this limitation and learn a better CSI compression function, DL is used in [57] to learn an encoder-decoder scheme to compress the CSI into a lower dimensional space and transmit a compressed information of it to the BS. Yet, CSI feedback ushers in complexity and overhead, which must otherwise be kept as low as possible. This motivates the need for further studies on advanced channel feature extraction mechanisms suitable for edge devices.

C. Beam Sweeping

The alternative of using CSI feedback for beamforming is to perform an exhaustive beam search based on transmitter and receiver codebook to establish the directional link in 5G and nextG networks. In general, the overall link establishment process comprises of 4 different steps [58]: (a) beam...
sweeping, which involves exploration of all the available beams through transmission/reception of reference signals; (b) beam measurement, which evaluates the quality of each beam through a predefined metric i.e. SNR; (c) beam selection, which is the process of selecting the best beam based on the beam measurement results; (d) beam reporting, which shares beam quality/decision information, usually from the UE to the BS. In this section, we briefly summarize the beam sweeping process defined by the 5G-NR standard during the initial access (IA), which assumes that no link has been previously established between the user equipment (UEs) and the base station (gNB). Notice that methods alternative to beam sweeping have attracted most of the research efforts due to the inefficiency and high overhead of legacy brute force algorithms. The 5G-NR standard [59] defines an exhaustive beam search process to find the best beam-pair configuration between the UE and the gNB. For larger antenna arrays, the time required to sweep through different sectors is not scalable for time sensitive applications.

- **Exhaustive Beam Search Time in 5G NR:** Consider a gNB-UE pair, with codebook sizes $C_{UE}$ and $C_{gNB}$ respectively. Then, the total number of beam directions to be scanned is $|C| = C_{UE} \times C_{gNB}$. During the initial access (IA), the gNB and the UE exchange a number of messages to find the best beam pair. During this process, the gNB sequentially transmits synchronization signals (SS) in each codebook element. Meanwhile, the UE also switches among sectors to receive in different codebook elements until all $|C|$ possible beam configurations are swept. The SS transmitted in a certain beam configuration are referred to as SS blocks, with multiple SS blocks from different beam configurations grouped into one SS burst. The SS burst duration ($T_{ssb}$) is fixed at 5 ms in the NR standard, and it is typically transmitted with a periodicity ($T_p$) of 20 ms [60], although different values are supported $T_p \in \{5, 10, 20, 40, 80, 160\}$ ms. In the mmWave band, a maximum of 32 SS blocks fit within a SS burst, and the number of explored beams per block is dependent on the beamforming technique. For example, hybrid and digital beamforming architectures allow transmitting or receiving multiple beams simultaneously. On the other hand, analog beamforming architectures only enable one beam per configuration, requiring as many SS blocks as beam configurations to perform the beam sweeping process. The total beam sweeping time ($T_{bs}$) can thus be expressed as:

$$T_{bs} = T_p \times \left\lfloor \frac{|C| - 1}{32} \right\rfloor + T_{ssb}. \quad (1)$$

As showcased in Fig. 4, the beam sweeping time increases linearly with increasing beam pairs, and this increment is larger in case of 5G-NR ($\approx 40$ ms for $|C| > 64$). Such delay has the potential for severely degrading the 5G NR performance, where time sensitive applications typically require latencies of $\leq 10$ ms [59].

In order to reduce the overhead of the complicated channel estimation and time-consuming beam-sweeping techniques, multiple out-of-band approaches have been explored in the recent literature, with the aim of achieving low overhead. These beamforming techniques can be broadly categorized into (a) RF-based and (b) non-RF based, with their different sub-categories illustrated in Fig. 5. In the next sections, we explore in detail each of these categories.

V. Out-of-band RF Based Beamforming

The main idea behind leveraging out-of-band RF frequencies during beamforming is to incorporate the cross channel correlation at mmWave bands with lower frequencies (2.4 GHz, radar bands, etc.). Such cross correlation is then utilized to reduce the beam search space by establishing a mapping between the channel measurements in the mmWave bands with lower frequencies (see Fig 6). Although the propagation characteristics in mmWave is different from lower frequencies, recent research reveals that the main direction of arrivals (DoAs) are comparable. Hence, the CSI at lower frequencies can be used to restrict the beam search space and avoid time-intensive exhaustive search, as proposed in the IEEE 802.11ad standard [61]. This is relevant as mmWave systems are very likely to be deployed in conjunction with lower frequency systems, where mmWave access points (APs) are envisioned to be paired with lower frequency APs that provide wide area control signalling and coordination. Moreover, multiband communication is one of the proposed solutions for providing high throughput communication systems with high reliability, thus reinforcing the interest in taking advantage of such systems in the near future [62]. Among the RF based out-of-band beamforming techniques, the use of radar signals and utilizing sub-6 GHz frequencies for mmWave beamforming have shown promising results.

A. Radar

For a vehicle to infrastructure (V2I) hybrid MIMO scenario, González-Prelcic et al. [63], derives the channel information from the infrastructure mounted radar that is used to obtain precoders/combiners at the vehicle and the infrastructure. The radar sensor operates at 76.5 GHz, which is close to the mmWave communication band at 65 GHz. Taking advantage of this close proximity of the operating frequencies, the computed covariance of the received signal at the radar is applied as an estimation of the covariance of the communication signal in the mmWave band. The authors then argue that the optimum...
Table V: Survey of literature on out-of-band RF based beamforming.

| Paper                     | Year | Dual Band Frequency | Approach                          | Validation | Multiple AP | Evaluation Metric             |
|---------------------------|------|---------------------|-----------------------------------|------------|-------------|-------------------------------|
| González-Prelcic et al.   | 2016 | 76.5 GHz, 65 GHz    | Compressive covariance estimation | Simulation | No          | Sum spectral efficiency       |
| Ali et al.                | 2017 | 76 GHz              | Compressive covariance estimation | Simulation | No          | Transmission rate             |
| Reus et al.               | 2019 | 60 GHz, 5.89 GHz    | DoA estimation                    | Simulation | No          | Beamforming time              |
| Nitsche et al.           | 2015 | 60 GHz              | LoS path detection               | Experiment | No          | Direct path estimation accuracy|
| Ali et al.                | 2017 | 60 GHz              | DoA estimation                    | Simulation | No          | Success percentage in detecting best beam |
| Hanhimi et al.            | 2018 | 30 GHz              | DoA estimation                    | Experiment | No          | mmWave throughput             |
| Mohamed et al.           | 2015 | 60 GHz              | Wi-Fi fingerprinting             | Experiment | Yes         | Average packet delay          |

B. Sub-6 GHz

Among the sub-6 GHz out-of-band beamforming techniques proposed in the state-of-the-art literature, Nitsche et al. [66] propose a blind beam steering (BBS) system which couples mmWave with legacy 2.4/5 GHz bands to estimate the direction for pairing nodes from passively overheard frames, as a replacement to the in-band trial-and-error beam initialization. Upon a beam training request, the proposed method first performs out-of-band direction inference to calculate angular profiles by broadcasting passively overheard detection band frames at the legacy sub-6 GHz band. In particular, an angular profile specifies received signal energy with respect to the azimuth incidence angle at the last known position. The LoS paths in all profiles remain nearly static, and appear as peaks at the same angle. However, the peaks resulting from reflections vary among profiles. Hence, the authors employ a profile history aggregation method over varying multipath conditions. By aggregating, the alternating reflection peaks are flattened and the remaining strongest peak is estimated to correspond to the direct path. Moreover, it gives an estimate of the uncertainty for the direct path estimate by measuring the deviation of the direct path angle over different profiles. Given the profile history for each device, a threshold for the peak-to-average ratio is defined to infer the LoS path and to reject the reflected paths. If the ratio for a direction estimate is below this threshold, BBS proceeds with the legacy IEEE 802.11ad beam training method. The experimental results depict that BBS successfully detects unobstructed direct path conditions with an accuracy of 96.5% and reduces the IEEE 802.11ad beam training overhead by 81%. Similarly, in [67] the authors propose using the sub-6 GHz digital beam scanning method for faster estimation of the optimal direction. The candidate mmWave beams are restricted only to those beams that overlap with the dominant paths at sub-6 GHz band. The angle of arrival (AoA) estimation on the 3 GHz channel is used in [68] to reduce the beam sweeping overhead for the mmWave in 30 GHz frequency. In particular, they experimentally show that in 94% of LoS conditions, the identified AoA in the 3 GHz band is within ±10° accuracy for the AoA of the mmWave signal. Hence, the authors propose using MUSIC for higher accuracy in AoA estimation. To estimate the AoA in the sub-6 GHz and running the exhaustive search only for angles in the corresponding direction of the mmWave band, while factoring in the error bound of ±10°.

One of the promising solutions for Gbps transmission in 5G is the use of wireless gigabit (WiGig) high frequency mmWave APs [71]. However, multiple WiGig APs are required to fully cover the target environment, due to their short ranges. In this regard, a comprehensive network architecture along with a dual-band MAC protocol is proposed in [69] for coordinated WiGig WLANs, which is based on tight coordination between the 5 GHz (WiFi) and the 60 GHz (WiGig) unlicensed frequency bands (see Fig. 9). In the proposed dual-band MAC protocol operation, the control frames to be shared among the APs are transmitted via the wide coverage sub-6 GHz WiFi band, while the high speed data frames are concurrently transmitted by the WiGig APs in the mmWave band. These control frames coordinate the beam training among the APs, so only one AP performs the beam training at a time, eliminating the probability of packet collisions due to simultaneous beamforming. Also, the link information consisting of the used beam identification (ID), modulation coding scheme (MCS) index and received power, is broadcasted in the sub-6 GHz WiFi frequencies, allowing other APs to effectively exclude those beam IDs that may interfere with the existing data link from their beamforming training beams. Moreover, since the location of a UE can be roughly estimated using WiFi channel information at WiFi frequencies through a process called fingerprinting, the authors propose this WiFi fingerprinting method to estimate the best and bad beam IDs of the WiGig links. Given a database of WiFi fingerprints and WiGig best beam IDs, an offline statistical learning is introduced where by comparing the current UE WiFi fingerprint with the pre-stored UE WiFi fingerprints, a best associated AP is selected for the UE, and a group of WiGig best sector IDs (beams) are estimated for the selected AP to effectively communicate with the UE at its current position. Among these estimated best beams, the beam IDs overlapping with the existing WiGig...
data links are recognized as bad beams and eliminated from the beamforming refinement process.

We conclude the discussion on out-of-band RF based beamforming techniques by providing a comprehensive overview of these processes in Tab. VI. Next, we explore the existing challenges in this area.

C. Challenges

While out-of-band RF assisted beamforming present promising improvements in beam initialization speed, there are some limitations associated, which we itemize as follows:

- The out-of-band RF channel measurements need to be acquired constantly in order to estimate the channel at the mmWave band. Hence, it requires an integrated protocol for multi-band coexistence that can be challenging in dense networks.
- An optimal mapping is required between mmWave and out-of-band channel measurements. The mmWave band has unique propagation characteristics that preserves sparsity. In particular, the number of reflections is limited in mmWave band, while in lower frequencies, multiple reflections are normally observed. As a result, translating the DoA for bands that are located far apart from each other can be challenging and are prone to errors.
- RF-based out-of-band beamforming requires simultaneous multi-band channel measurements that increases the complexity of mmWave transceivers. Although future mmWave devices will likely support lower frequencies as well, this feature is not widely deployed in commercial devices yet.
- The existing out-of-band methods do not yet support simultaneous beamforming at both the transmitter and receiver sides, which is required for effective directional transmissions.

D. Non-RF Modalities for RF Tasks

Considering these challenges in out-of-band RF based beamforming techniques, there is growing interest in studying different non-RF modalities for optimizing wireless links. These various non-RF data modalities, e.g., RGB/RGB-D (RGB-Depth) camera images, LiDAR etc., capture the situational information in the environment from different perspectives, which can be exploited to assist in a variety of wireless tasks, such as handover or channel quality prediction. While many sensing technologies have been proposed to enhance the reliability of wireless links, mostly in the mmWave bands, recent advances in computer vision offer an untapped potential for camera-aided communications. We discuss few examples of such applications and the proposed solutions, and present a comprehensive overview in Tab. VI.

Most of the existing research in using non-RF modalities focus on problems that arise from the unique propagation characteristics, as well as susceptibility to blockage in the mmWave band. Oguma et al. in [724] propose a proactive mmWave base station selection method that predicts human blockage based on the dynamics observed through RGB-D camera images. Other works have taken similar approaches by exploring the use of camera images with reinforcement learning for handover management using single [73] and multiple cameras systems [74]. Jointly considering vision and communication is discussed in [75], where Nishio et al. propose a network stack for a hybrid camera-communication system.

Recently there is also a rising interest in predicting the channel quality without RF measurements. Forecasting low SNR conditions or throughput reduction due to blockage or other channel metrics is the first step towards taking proactive measures before the link quality deteriorates, or worse, the connection is lost. In particular, Nishio et al. in [76] propose an RGB-D based received power prediction scheme for mmWave networks, based on multiple deep learning techniques to predict power losses up to hundreds of milliseconds ahead. Other works have analyzed specific challenges of this approach, such as the input data size [77] or the application of pre-trained models in new scenarios using transfer learning [78]. The fusion of in-band mmWave data with camera images is explored in [79] through a split-learning architecture, where the base station and the user run independent models and combine their predictions. Koda et al. show that a single pixel image can notably enhance the power prediction versus only-RF based approaches [80]. In [81], a method to estimate throughput solely based on RGB-D images is presented by Okamoto et al., with an RMS error of 114-178 Mbps in real time.

After motivating the utility of leveraging various non-RF sensor data for RF tasks, we next map these benefits to the use-case of beamforming in mmWave bands, when higher magnitude of antenna elements (i.e., mMIMO systems) are involved. Additionally, the challenges of using RF-based out-of-band beamforming, described in Sec. V-C, suggest the research community needs to explore the space of beyond RF-only solutions (be it traditional or RF-based out-of-band). We explore this direction in the next section.

VI. BEAMFORMING USING NON-RF SENSOR DATA

In mmWave beamforming, the location of the TX-RX pair and potential obstacles are the key factors that directly affect the optimal beam configuration. Out-of-band RF aided beamforming methods estimate the approximate location of TX-RX pair given the AoA in other RF bands, which helps to narrow down the beam search space. Interestingly, the situational state of the environment can also be acquired through data obtained from other sensor devices [83], without occupying limited sub-6GHz RF resources. This motivates the use of non-RF sensor data to speed up the beam initialization process in mmWave band [84]. Unlike the previously discussed out-of-band RF methods, the non-RF based beamforming does not require simultaneous multi-band channel measurements and optimal mapping between mmWave and CSI collected from another band. It is also capable of generating a mutually acceptable decision for both transmitter and receiver.

Typically, non-RF based beamforming utilizes inputs from a number of different sensors such as, GPS (Global Positioning
TABLE VI: Survey of literature leveraging the non-RF sensor data for solving different mmWave challenges.

| Paper | Problem          | Data Type | Approach                     | Evaluation Metric |
|-------|------------------|-----------|------------------------------|-------------------|
| Koda et al. [73] | Handover         | RGB images | Reinforcement learning       | Throughput        |
| Koda et al. [74] | Handover         | RGB images | Reinforcement learning       | Received power    |
| Oguma et al. [72] | Handover         | RGB-D images | Deterministic              | Throughput        |
| Nishio et al. [75] | mmWave camera architecture | Depth images | Network design           | Throughput        |
| Okamoto et al. [83] | Throughput estimation | RGB-D images | Adaptive regularization of weight vectors | Throughput |
| Oguma et al. [82] | Base station selection | RGB-D images | Transfer learning          | Throughput        |
| Makuma et al. [78] | Received power prediction | RGB-D images | CNN, Conv. LSTM, and Random forest | Received power |
| Nakashima et al. [71] | Input data size for received power prediction | Depth images | Convolutional LSTM         | Input data size   |
| Nishio et al. [76] | Received power prediction | RGB-D images | CNN, Conv. LSTM, and Random forest | Received power |
| Koda et al. [79][80] | Received power prediction | RGB-D images + RF | Split learning            | Throughput        |

TABLE VII: Sensing technologies for aiding the beamforming.

| Sensing Technology | GPS | Camera | LiDAR |
|--------------------|-----|--------|-------|
| Cost               | Low | Low    | High  |
| Horizontal Range   | Low | High   | High  |
| Resolution         | Low | High   | High  |
| Night Vision       | Yes | No     | Yes   |
| Affected by Weather| No  | Yes    | Yes   |

**Fig. 7:** Market revenue forecast for advanced driver assistance system (ADAS) sensors from the year 2020 to 2025.

System), camera, LiDAR (Light Detection and Ranging), which provide a 3-D representation of the surroundings, etc. This is further aided by the fact that with the wide proliferation of IoT, multiple sensors are embedded in the environment, thus making it feasible to obtain situational information from non-RF sources. As an example, consider the automotive sector with vehicles that have advanced driver-assistance systems (ADAS). Fig. [7] depicts the increase in the market revenue of the various sensors enabling ADAS, as reported by Yole Dévelopement [85]. It is expected that the global market for GPS, radar, cameras and LiDARs will reach $159.6 billion in 2025. With the easy availability of such multitude of sensors, we need to incorporate methods that leverage the heterogeneous sensor data to extract a rich understanding of the environment.

In LoS scenarios, even though the optimal beam configuration can be estimated using the location of transmitter and receiver, it is not trivial to employ such approaches when encountering irregular radiation patterns, for e.g., when devices have multiple side lobes. The problem becomes more challenging when estimating the strongest reflection from obstacles in NLoS conditions. Hence, a proactive method is required to learn the channel characteristics associated with the observed non-RF sensor modalities on a case-by-case basis. Both deterministic and AI-enabled methods are proposed in literature that consider either single sensor modalities or multiple modalities through deep learning. We next go through these state-of-the-art methods, covering different sensor acquisition techniques, available datasets, exploitation methods of single and multiple modalities, and identify future research trends.
obstacle? Each sensor type has its advantages and limitations. For example, GPS equipped objects can be utilized to track location, but these sensors cannot capture the presence of obstacles. LiDAR can collect the 3D state of the environment but fails to track the location of the target transceivers. Thus, GPS data can be merged with raw LiDAR data in the preprocessing step to mark the coordinates of the target receiver in the collected point clouds. Hence, data-level aggregation methods are one of the commonly used approaches to refine the raw data to be more informative. Second, the preprocessing steps are also beneficial for reducing the data complexity by either discarding the irrelevant information or reducing the dimensionality of the input data. As an example, using a low-pass filter on camera images can reduce the dimensionality of the image by averaging the adjacent pixels while preserving the integrity. ML-based solutions only accept the data arranged in a fixed size, while for some modalities such as LiDAR the number of point clouds is varying on a case-by-case basis, depending on the number of present objects. Hence, preprocessing can account for this issue by transforming the data to a constrained representation without degrading information content. Hence, it is important to design proper preprocessing steps before using the data for inference. It should be noted here that the preprocessing pipeline of each modality must be designed based on the unique properties of each sensor type, while maintaining the highest correlation with the ultimate task, which is mmWave beamforming in this case.

### B. Datasets

Next, we discuss the features of the available public datasets specific to beamforming using non-RF sensor modalities. These datasets enable the research community to explore different aspects of non-RF beamforming without incurring an individual effort of data collection.

1) ViWi: Alrabeiah et al. proposed a scalable synthetic framework called Vision-Wireless (ViWi) [86]. The scenario of interest is a V2I setting in 28 GHz mmWave band. The first release of this dataset contains four scenarios with different camera distributions (co-located and distributed) and views (blocked and direct). The channel characteristics and images are generated using the Remcom Wireless Insite ray-tracing [87] and Blender [88] software, respectively. For each scenario, a set of images and raw wireless data (signal departure/arrival angles, path gains, and channel impulse responses) are recorded. An extended version of this dataset is named ViWi vision-aided mmWave beam tracking (ViWi-BT) [89], which contains 13 pairs of consecutive beam indices and corresponding street view images. This dataset contains a training set with 281,100 samples, a validation set with 120,468 samples, and a test set with 10,000 samples.

2) Raymobtime: The Raymobtime multimodal dataset [18] captures a virtual V2X deployment with high fidelity in the urban canyon region of Rosslyn, Virginia for different traffic patterns. A static roadside BS is placed at a height of 4 meters, alongside moving buses, cars, and trucks. The traffic is generated using the Simulator for Urban MObility (SUMO) software [90], which allows flexibility in changing the vehicular movement patterns. The image and LiDAR sensor data are collected by Blender, and Blender Sensor Simulation (BlenSor) [91] 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. A python orchestrator invokes each software for each scene and collects synchronized samples of LiDAR point clouds, GPS coordinates and camera images mounted at the BS. The combined channel quality of different beam pairs are also generated using Wireless Insite ray-tracing [87] software. The number of codebook elements for BS and the receiver is 32 and 8, respectively, leading to 256 beam configurations overall.

3) Image-based: This dataset is obtained by Salehi et al. in [92] from a testbed composed of two Sibeam mmWave [29] antenna arrays mounted on sliders enabling horizontal movement. Using the mmWave transceivers from National Instruments, the mutual channel is measured for 13 beam directions at transmitter and receiver (169 beam configurations overall). Two GoPro cameras observe the movements in the environment and are synchronized with the mmWave channel measurements. In the designed scheme, an obstacle blocks the LoS path between the transmitter and receiver and the experiment is repeated for two types of obstacles, wood and cardboard, causing 30dB and 4dB attenuation while blocking the LOS path, respectively.

### C. Single non-RF Modalities

Next, we present detailed descriptions of different studies and algorithms that use a single non-RF sensor modality. These include either GPS coordinates, camera images or LiDAR point clouds to accelerate the beam selection, and by extension, the beamforming process. Multimodal fusion is described later in Sec. VI-D.

- **GPS Coordinates**: The knowledge of the location of target receiver has been used earlier to address the challenges of cell discovery [106]. The same idea can be used to speed up the beam initialization in mmWave band, which utilizes directional transmission. The authors in [93, 94] use the GPS based position of the receiver to estimate the optimum future beam directions. In particular, the proposed algorithms predict the future locations by tracking the mobility profile of the receiver and geometrical features of the environment. However, it should be noted that this approach only works when the LoS path is available. Alternatively, Wang et al. propose a framework for mmWave beam prediction by exploiting the situational awareness [95]. They use the location of all the vehicles in the same scene as features to extend the solution to NLoS scenarios. The simulation scenario consists of small cars and trucks, any of which can be the target receiver. The authors argue that the vehicle dynamics have the main effect on the optimum beam configuration, since the road side buildings and infrastructures are stationary, and pedestrians are small in size. Hence, a feature vector map \( v = [r, t_1, t_2, c_1, c_2] \) is generated where \( r \) depicts the location of RSU, \( t \) and \( c \) represent the truck and car vehicles. The subscripts 1 and 2 denote the lane index where the vehicle is located and each vector \((t_i, c_i), i \in \{1, 2\}\) includes the location of the corresponding vehicle type in ascending order for the lane \(i\). Since the ML algorithms
accept a fixed size input, the number of trucks/cars on each lane is constrained, and the vehicles which are far away are eliminated. This feature vector is then used to predict the received power for any beam in the codebook, by leveraging ML. Similarly, Va et al. [96] propose an algorithm where the location of all the vehicles on the road, including the target receiver, is used as input to an ML algorithm, to infer the best beam configuration. The proposed algorithm uses the power loss probability as a metric to estimate the misalignment probability that might occur when non-optimal beams are selected. In this case, a subset of the beam configurations are suggested by the authors to minimize this misalignment probability. In order to speed up the beam initialization, an online learning algorithm is proposed in [97], which exploits the coarse user location information in vehicular systems. In particular, the problem is modeled as a contextual multi armed bandit (MAB) problem and a lightweight context-aware online learning algorithm, namely fast machine learning (FML) is used to learn from and adapt to the environment. The proposed FML algorithm explores different beams over time while accounting for contextual information (i.e., vehicles’ direction of arrival) and adapts the future beams accordingly, in order to account for the system dynamics such as the appearance of blockages and changes in traffic patterns. In comparison, Aviles et al. in [98] first generate a database that captures the propagation characteristics at 28 GHz and the position of UE. Then, given the location of a UE, a hierarchical alignment scheme is proposed, which consults with this database and incorporates the position of the UE for faster beam alignment.

• Camera Images and Light Sensors: The cameras are one of the sensing modalities that capture the situational state of the environment with high resolution. With the recent progress in computer vision and deep learning, powerful algorithms are now available that can be used for processing the images in real time for beamforming. A baseline for ViWi-BT dataset is presented in [99] based on gated recurrent units (GRUs) without the images and only the sequence of beam indices. The authors argue that beam prediction accuracy is expected to improve significantly by leveraging both wireless and visual data. In [99], Tian et al. propose a framework to predict future beam indices from previously observed beam indices and images. The proposed approach consists of three steps as follows. The first step consists of feature extraction, where ResNet, ResNext and 3D ResNext modules, each proven to have powerful feature-representation abilities, are used to capture 2D and 3D spatio-temporal features from the images. In the second step, a long short-term memory (LSTM) [107] network is designed to incorporate the time-series data for prediction. Finally, a feature-fusion module aggregates features from ResNet and 3D ResNext to generate high-level features. The fusion module comprises of two LSTM networks and a simple cross-gating block that only support linear transformation. To validate their approach, the authors use WiWi-BT dataset where the first eight pairs of images are used to predict next five future beams. Similarly, in [100], Xu et al. propose a scheme where the images captured from different perspectives are used to construct a 3D scene that resembles the point cloud data collected by 3D sensors like LiDAR. Then, a CNN with 3D input is designed to predict the future beams to be selected. Results reveal that the proposed 3D scene based beam selection outperforms LiDAR in accuracy, without imposing the huge cost of LiDAR sensor. While the majority of current literature uses synthetic datasets, the authors in [99] deploy a testbed using National Instruments radio at 60 GHz [29] and camera generated images to predict the best beam configuration. Their proposed method consists of two main steps, namely detection and prediction. In the first step, the transmitter and receiver are detected in the image in the form of a bitmap. This step is important to detect the features which are relevant to the task and discard the irrelevant ones, such as static walls, etc. Finally, the bitmaps are fed to another CNN to predict the optimum beam configuration given the historical data from collected dataset. The LiSteer system proposed in [102] steers mmWave beams to mobile devices by re-purposing indicator light emitting diodes (LEDs) on wireless APs to passively track the direction to the AP using light intensity measurements with off-the-shelf light sensors. The proposed approach considers the pseudo-optical properties of mmWave signal, i.e., dominant LoS propagation, to approximate the APs’ AoA in the mmWave band. Hence, their approach requires the APs to be equipped with LEDs and to be situated close to the mmWave band antenna.

| Paper | Year | Frequency | Data Type | Approach | Validation | Feedback Required? | Evaluation Metric | Dataset Publicly Available? |
|-------|------|-----------|-----------|----------|------------|-------------------|------------------|-----------------------------|
| Kim et al. [92] | 2013 | 60GHz | GPS | Deterministic | Simulation | No | Achievable capacity | No |
| Va et al. [94] | 2016 | 60GHz | GPS | Deterministic | Simulation | No | Outage and average rate ratio | No |
| Wang et al. [95] | 2018 | 5GHz | GPS | Deep learning | Simulation | No | Alignment probability | No |
| Xu et al. [96] | 2017 | 60GHz | GPS | Deep learning | Simulation | Yes | Power loss probability | No |
| Sim et al. [97] | 2018 | 28GHz | GPS | MAB | Simulation | Yes | Cumulative RX data | No |
| Aviles et al. [98] | 2019 | 28GHz | GPS | Deterministic | Simulation | Yes | CDF of AoA estimation | No |
| Aljebraish et al. [99] | 2020 | 28GHz | Camera | Deep learning | Simulation | No | Top-1 accuracy | Yes |
| Tian et al. [100] | 2020 | 28GHz | Camera | Deep learning | Simulation | No | Top-1 accuracy | Yes |
| Xu et al. [101] | 2020 | 60GHz | Camera | Deep learning | Simulation | Yes | Top-K accuracy | Yes |
| Kadileh et al. [102] | 2020 | 60GHz | Camera | Deep learning | Simulation | No | Top-1 accuracy | Yes |
| Wonderlin et al. [103] | 2021 | 28GHz | LiDAR | Deterministic | Mixture | No | Link latency | No |
| Haider et al. [104] | 2018 | 60GHz | Light sensor | Deterministic | Experiment | No | CDF of AoA estimation | No |
| Klouta et al. [105] | 2019 | 60GHz | GPS and LiDAR | Deep learning | Simulation | Yes | Top-K accuracy | Yes |
| Dias et al. [106] | 2019 | 60GHz | GPS and LiDAR | Deep learning | Simulation | Yes | Top-K accuracy | Yes |
| Aljebraish et al. [107] | 2020 | 28GHz | Camera and sat-6 | Deep learning | Simulation | Yes | Top-K accuracy | Yes |

TABLE VIII: Survey of literature on non-RF data for out-of-band and multimodal beamforming. The highlighted three rows of the table depict the most recent effort of using multimodal non-RF sensor data to aid in beamforming, referred as multimodal beamforming in this article.
authors propose using an array of light sensors to combat the in-coherency of light-AoA estimation that also allows steering beams for both 2D and 3D beamforming codebooks. The experimental results demonstrate that LiSteer achieves direction estimates within 2.5° of ground truth on average with beam steering accuracy of more than 97% in tracking mode, without incurring any client beam training or feedback overhead.

• **LiDAR Point Clouds:** Woodford et al. [101] use LiDAR to build a 3D map of the surrounding physical environment and captures the characteristics of the physical materials. The proposed approach uses a customized ray-tracing algorithm that can identify real RF paths in a 3D mesh generated by LiDAR sensors, and reject false reflection paths caused by reconstruction noise. The output of this phase is a pre-computed look-up table to select the best beams for all mmWave links in the environment. It should be noted that the LiDAR sensors are not required during the ordinary operation of the system and are only used in advance to generate the lookup table. The proposed approach can recompute the complete lookup table for the environment within 15 minutes. The authors validate their approach using Azure Kinect LiDAR camera [108] and a commercial 802.11ad radio [109], yielding to 66% reduction in latency and 50% increase in throughput.

**D. Multimodal Beamforming**

Since, each of the above sensor modalities capture different aspects of the environment, using more than one sensor modality and intelligently fusing these multimodal data can result in more comprehensive understanding of the environment and can consequently enable the undertaking of robust decisions.

- **Benefits of Fusion:** The fusion of multimodal data over the single modalities has multiple advantages, as explained below:
  - **Enhanced Data Representation:** For the situational information to be effective during beamforming, it is crucial to differentiate between the transmitter, receiver and obstacles. However, some sensor modalities cannot provide such information by only relying on raw data. In this case, the data from different modalities can be fused together to improve the data representation. As an instance, it is not trivial to locate the receiver within a LiDAR point cloud. In this case, the GPS coordinates can be used to mark the target receiver.
  - **Compensate for the Missing Information:** Sometimes the captured data from each sensing modality reflect an aspect of the environment, yet none can provide a complete understanding by it’s own. For instance the dimensionality of objects is not reflected in GPS, and the accurate Cartesian coordinates of the target receiver cannot be acquired using LiDAR or image sensors.
  - **Improved Accuracy:** Using more than one modality enables a fine grained understanding of the environment which results in more accurate predictions. Hence, fusion reinforces the prediction accuracy by gathering the information from different sensors to make the final decision. In this case, the fusion algorithms can automatically adjust the weights of each modality towards the optimum performance.

- **Robustness to Errors:** Collecting data using sensor devices come with associated considerations, including the inherent error. Here, the accuracy of measurement is dependent on working with the nominal structure that the device is designed for. For instance, the accuracy of LiDAR sensor degrades in with sunlight reflections, while it does not affect the GPS data [110]. Hence, fusion increases prediction robustness in the case of inaccurate or unreliable data.

- **Availability:** In some applications, the sensor does not have to be co-located. Hence, secondary control channels are required to enable the connectivity between the different sensors and the computing unit. However, this control channel is also subject to saturation and loss. Using more than one modality with fusion helps the system to be robust to such scenarios and it guarantees that the prediction happens when at least one modality is available during the inference.

Below, we give some examples of state-of-the-art multimodal beamforming with different fusion approaches on multiple sensors.

- **GPS and LiDAR Fusion:** Consider a typical V2I setting, where a static BS wants to establish communication with a target vehicle-mounted receiver. The vehicle is assumed to be equipped with GPS and LiDAR sensors that enable the vehicle to acquire its location and detect blocking objects nearby. In this scenario, Klautau et al. propose a distributed architecture to reduce the mmWave beam selection overhead [102]. Here, the BS constantly broadcasts its position via a low-band control channel. The situational state of the environment is then collected using LiDAR, situated on the vehicle and is aggregated by BS location in the preprocessing pipeline, where a histogram is generated at the beginning to quantize the space. The LiDAR point clouds then lie in the corresponding bin of the histogram, and the location of BS and receiver is also marked with unique indicators. Using the proposed preprocessing step, the measured point clouds are mapped to a ridge represented by a fixed size. Note that the number of point clouds in the raw data varies depending on the number of objects present during the measurement. This refined data representation is then fed as input to a deep CNN to estimate a set of K most likely candidate beam pairs. The selected beam pairs are then sent to the BS, and beam training is performed to generate the suggested subset to obtain the optimum beam configuration. Similarly, [103] considers a V2I setting and compares the performance of the previously described distributed scheme with two centralized schemes: (i) using a single LiDAR located at the BS, and (ii) fusing LiDAR data from neighboring vehicles at the BS. The LiDAR data is then used for both LOS detection and beam selection for three competing scenarios. The experimental results in this work depicts that in LoS, distributed and centralized methods perform closely, while the LiDAR at BS results in lower top-K beam prediction accuracy, because of limited range of LiDAR. On the other hand, in NLoS scenarios, the distributed scheme outperforms the centralized method, and both are better than LiDAR at BS.
• Camera with Sub-6 GHz Fusion: The possibility of vision-aided wireless communications is evaluated in [105] where a camera at the BS observes the movements in the environment, and snapshots of the environment are paired with sub-6 GHz channels to help overcome the beam selection and blockage prediction overhead. The proposed method models the beam prediction from images as an image classification task. Hence, each user location in the scene is mapped to a class representing the associated beamforming codebook. However, the pure image input may be insufficient for blockage detection since the instances of ‘no user’ and ‘blocked user’ are visually the same. Hence, in order to identify blocked users, the images are fused by sub-6 GHz channels to account for the aforementioned challenge.

The concise overview of different state-of-the-art beamforming methods while using single or multimodal data is presented in Tab. VII.

E. Future Research Directions

While multimodal learning is an extremely interesting research field, there are some challenges that need to be addressed. First, in order to exploit more than one modality, the synchronized information of all modalities must be present during inference. This requires a precise network controller and back-channel to enable connectivity among different modules while accounting for privacy concerns. Second, the fusion scheme needs to be designed such that the different modalities result in a reinforced prediction. The fusion model can be as simple as a linear transformation, such as summation or multiplication. However, learning the relation between different modalities might require non-linear transformations such as deep learning on custom-made neural networks. In that regard, we explore few novel fusion techniques that use non-linear transformation in the following section.

VII. Detailed Description of Data Fusion Methods for Multimodal Beamforming

In the earlier sections, we make the case for using non-RF modalities for mmWave beamforming, and more specifically, the motivation for using multimodal data to capture the holistic information of the wireless environment. Successfully exploiting multiple modalities depends on the ability of how and when to fuse different modalities. The state-of-the-art deep learning based fusion techniques generally involve concatenation of layers [111], [112], [113], element wise operation [114], [113] and cross-modality analysis [115], [116] in an one-step manner. In other words, all the available modalities are generally fused once in an aggregated manner. One of the contributions of this survey is to explore the full potential of multimodal learning by analyzing diverse modality fusion techniques for beam selection algorithms.

Novel fusion methods based on deep learning flexibly assign weights on each modality based on their relevance from the situational state information. This results in a faster yet accurate beam selection solution. Our vision for multimodal beamforming is to generate novel feature-level and multi-level fusion frameworks for fast and accurate beamforming by minimizing the beam search space. By fusing the features/outputs from various unimodal models, we describe a platform to improve the individual prediction accuracy. The customized neural network architectures for each single modalities are referred as unimodal models in this paper. We propose fusion frameworks which work on the ultimate and penultimate layers of each unimodal model. The multi-level aspect of the proposed framework is as follows: a feature-based fusion takes place at the first level, followed by subsequent fusions at each level based on the prediction of the previous one.

A. Scenario of Interest

As before, consider a vehicular communication scenario need for autonomous cars, where seamless yet ultra-fast communication between moving vehicles is crucial. Multiple sensors are already included as standard installations for the majority of new vehicles, as well as in fixed roadside base stations [117]. For example, LiDAR sensors are an indispensable part of modern vehicles that are used for either automated driving or collision avoidance [118]. The GPS data are regularly collected and transmitted as part of basic safety messages frame in V2X applications [119], and surveillance cameras have been used for decades as part of smart-city initiatives [120]. The V2X network scenario using these three sensor data types and exploiting various fusion techniques for beamforming is illustrated in Fig. 8. Next, we formulate the problem of reduced space beam selection which leverages the non-RF multimodal data in the multimodal beamforming problem.

B. Formulation of Top-K Beam Selection for Multimodal Beamforming

We consider a codebook-based mMIMO transmitter and receiver, where each codebook element represents a particular beam direction. The transmitter and receiver codebooks are represented as: $C_{tx} = \{t_1, \ldots , t_M\}$, $C_{rx} = \{r_1, \ldots , r_N\}$, where $M$, $N$ are the number of transmitter and receiver codebook elements, respectively. Hence, the set of all possible beam pairs $\mathcal{B} = \{(t_m, r_n)|t_m \in C_{tx}, r_n \in C_{rx}\}$, with $|\mathcal{B}| = M \times N$. For a specific beam pair $(t_m, r_n)$, the normalized signal power is $p(t_m, r_n) = |w_{t_m}^* \mathbb{H} w_{r_n}|^2$, where $\mathbb{H} \in \mathbb{R}^{M \times N}$ is the channel matrix and $^*$ 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 set of normalized power for all beam configuration is defined by: $\mathcal{P} = \{p(t_m, r_n)|t_m \in C_{tx}, r_n \in C_{rx}\}$. The beam selection process for multimodal beamforming is defined to find top-$K$ best beam configuration, $\mathcal{B}_K$:

$$B_K = \arg \max_{A \subseteq \mathcal{B}, |A| = K} A \in \mathcal{P}. \quad (2)$$

The classical standards (both IEEE 802.11ad [7] and 5G-NR [121]) for beam selection corresponds to sweeping all beam pairs $(t_m, r_n) \in \mathcal{B}$ sequentially in order to find the best one
for beamforming. The selected set of $B_K$ restricts the search for the optimal pair to this set.

Recall that the multimodal data to be comprised of GPS, LiDAR and camera image sensor data. The dimensionality of the data matrices are defined as: $(d_L^3 \times d_I^3 \times d_C^3)$ and $(d_L^3 \times d_I^3)$ for 3D LiDAR and 2D image sensors, respectively. GPS coordinate has 2 elements: latitude and longitude. Next we define each unimodal data separately as: $X_L \in \mathbb{R}^{N \times d_L^3 \times d_I^3 \times d_C^3}$, $X_I \in \mathbb{R}^{N \times d_L^3 \times d_I^3}$, and $X_C \in \mathbb{R}^{N \times 2}$, for LiDAR, image, and coordinate, respectively, where $N$ is the number of training samples. The overall multimodal data is represented as: $X = [X_L; X_I; X_C]$. We configure the label matrix $Y \in \{0, 1\}^{N \times |B|}$ to represent the basis vector of $B$ beam pairs, where the optimum beam pair is set to 1, and rest are 0, following the Eq. (2). Next, we define different unimodal models and the necessary notations used for proposed fusion frameworks.

\section{C. Unimodal Models}

Each unimodal network works as a feature extractor for each modality. The details of each unimodal network are out of the scope of this survey. However, modeling the penultimate and ultimate layers of each candidate modality is necessary for formulating different fusion approaches. The latent embeddings from each unimodal feature extractor are captured at the penultimate layer, and the prediction from each unimodal network is captured at its ultimate layer.

- **Defining the Latent Embeddings**: We assume that the penultimate layers of the unimodal networks of LiDAR, image, and GPS coordinates have $d_L^4$, $d_I^4$, and $d_C^4$ neurons, respectively. Hence, sample space of the inputs of each sensor modality maps to a vector with a dimension equal to corresponding number of neurons.

We denote the feature extractor of each modality as $f_{L}^d$, $f_{I}^d$, and $f_{C}^d$ for LiDAR, image, and coordinate data, respectively, each parameterized by weight vectors $\theta^m$, for $m \in \{L, I, C\}$. We refer to the output of these feature extractors as the latent embedding of each modality. Formally,

$$z_L = f_{L}^d(X_L), \quad f_{L}^d : \mathbb{R}^{d_L^3 \times d_I^3} \mapsto \mathbb{R}^{d_L^4} \tag{3a}$$

$$z_I = f_{I}^d(X_I), \quad f_{I}^d : \mathbb{R}^{d_L^3 \times d_I^3} \mapsto \mathbb{R}^{d_I^4} \tag{3b}$$

$$z_C = f_{C}^d(X_C), \quad f_{C}^d : \mathbb{R}^{d_L^3 \times d_I^3 \times 2} \mapsto \mathbb{R}^{d_C^4} \tag{3c}$$

where $z_L$, $z_I$, and $z_C$ show the extracted latent embeddings for input data $X_L$, $X_I$, and $X_C$ respectively.

- **Defining the Ultimate Layers**: We formulate the output of the ultimate layers for each unimodal network with regard to the corresponding latent embeddings. The ultimate layers are just the transformation the penultimate embeddings with suitable activation functions. The output of ultimate layers for unimodal networks of LiDAR, image, and coordinate are denoted as $s_L$, $s_I$, and $s_C$, respectively,

$$s_L = \sigma(f_{L}^s(z_L)), \quad f_{L}^s : \mathbb{R}^{d_L^4} \mapsto \mathbb{R}^{d_L^5} \tag{4a}$$

$$s_I = \sigma(f_{I}^s(z_I)), \quad f_{I}^s : \mathbb{R}^{d_I^4} \mapsto \mathbb{R}^{d_I^5} \tag{4b}$$

$$s_C = \sigma(f_{C}^s(z_C)), \quad f_{C}^s : \mathbb{R}^{d_C^4} \mapsto \mathbb{R}^{d_C^5} \tag{4c}$$

where $\sigma$ is the softmax activation functions over the latent embeddings. Finally, the output of each ultimate layer maps to all possible beam pairs $B$ combining transmitter and receiver codebook elements.

\section{D. Novel Feature-level Fusions at Penultimate Layer}

1) **Aggregated Fusion**: The conceptual overview of the aggregated fusion network is shown in Fig. 9. The latent embeddings of the penultimate layers from each unimodal network are concatenated in an aggregated manner. The fusion...
network can be designed with multiple convolutional, pooling and fully connected layers afterwards. The overall concept is generalized as follows: first, given \( z_L \in \mathbb{R}^{d^L} \), \( z_I \in \mathbb{R}^{d^I} \) and \( z_C \in \mathbb{R}^{d^C} \), we concatenate them in aggregated manner and generate the combined latent feature matrix \( z \):

\[
z = [z_L; z_I; z_C] \in \mathbb{R}^{d^L+d^I+d^C}.
\]

Them, we denote the aggregated fusion network as \( f^{AGF} (\cdot) \). Finally, we use softmax activation function \( (\sigma) \) to predict the optimality of each beam as:

\[
s^{AGF} = \sigma (f^{AGF} (z)), \quad f^{AGF} : \mathbb{R}^{d^L+d^I+d^C} \mapsto \mathbb{R}^{|B|}
\]

where \( s^{AGF} \) is the ultimate layer output of the aggregated fusion network.

2) Incremental Fusion: For incremental fusion network, we first sort unimodal networks according to their performance. We concatenate the penultimate layer embeddings of the two best unimodal networks. In this step, we freeze the best unimodal model and only retrain the second best unimodal model and the fusion layers. This forces the second-best unimodal model and the fusion layer to learn different information as compared to the best unimodal model. Similarly, when incorporating the third-best unimodal model, we only retrain the third and fusion layers. A conceptual overview of the incremental fusion framework is illustrated in Fig. 10.

The deep fusion at the second level is defined using the ultimate layers of unimodal and penultimate fusion networks: \( s_L, s_I, s_C, \) and \( s^{DF} \in \mathbb{R}^{|B|} \), where \( s^{DF} \in \{s^{AGF}, s^{IRF}\} \). In this case, the representation of combined matrix \( s^{DF} \) is:

\[
s^{DF} = [s_L; s_I; s_C; s^{DF}] \in \mathbb{R}^{4 \times |B|}.
\]

The graphical representation of the deep fusion network at the second-level of fusion is shown in Fig. 11. The deep fusion network on the output of ultimate layer will intelligently assign higher weights to the outputs of the more relevant models of first-level. We use 4 fully connected layers as the deep fusion architecture of second-level. The details of this network architecture is presented in Fig. 11. The number of used filters in each level is flexible to be fine-tuned with the available type of dataset.

The deep fusion at the second level is defined using the ultimate layers of unimodal and penultimate fusion networks: \( s_L, s_{IRF}, s_{PNF} \in \mathbb{R}^{|B|} \), where \( s_{PNF} \in \{s^{AGF}, s^{IRF}\} \). In this case, the representation of combined matrix \( s^{DF} \) is:

\[
s^{DF} = [s_L; s_I; s_C; s_{PNF}] \in \mathbb{R}^{4 \times |B|}.
\]

The softmax activation function \( \sigma \) is used to predict the optimality of each vehicle as:

\[
s^{DF} = \sigma (f^{DF} (z)), \quad f^{DF} : \mathbb{R}^{4 \times |B|} \mapsto \mathbb{R}^{|B|}
\]

where \( s^{DF} \) is the ultimate layer output after the second-level of fusion.
Fig. 12: Performance of different unimodal and aggregated fusion network on multimodal Raymobtime dataset [13]. We report the top-$K$ ($K=1, 5, 10$) accuracy of the predicted labels and true labels. Fusion significantly improves the top-1 accuracy, whereas the difference gets lesser while evaluating top-10.

F. Preliminary Results

The preliminary results of predicting top-1, top-5, and top-10 best beams on the Raymobtime dataset [13] (details are in Sec. VI-B2) are presented in Fig. 12. The initial results show the effectiveness of aggregated penultimate fusion over individual modalities. LiDAR performs better than the other two sensors but fusion proves to be more effective instead of an LiDAR-only approach. The difference in performance of fusion and best performing modality decreases with prediction of more beam pairs. Hence, to strategically decrease the search space for faster beam selection, fusion is more effective than any standalone, single modality. Implementation of the proposed incremental and multi-level fusion techniques on the same dataset are fertile areas of further exploration by the research community.

G. Future Possibilities: from Theory to Practice

The pathway leading from theoretical conceptualization to real-world application is challenging given the four key steps involved: (i) numerical analysis of theory; (ii) simulation on a system model; (iii) emulation of the system model using realistic setup; and (iv) validating the theory and system model with real-world testbed experiments. A typical “path to reality” for wireless communication research is illustrated in Fig. 13. The state-of-the-art literature on beamforming with non-RF modalities are highly relevant. In this section, we suggest applications where the benefits are inclusive of diversified wireless environments by considering different types of: (i) road curvature; (ii) pedestrian traffic; (iii) vehicular traffic from 5 mph to 45 mph; (iv) vegetation; (v) weather condition. Such dataset can become the defacto standard for other researchers to comparatively benchmark and validate similar approaches for multimodal beamforming.

Next, we discuss possible application areas of using multimodal beamforming in AI-enabled technologies for NextG networks.

VIII. MULTIMODAL BEAMFORMING: APPLICATIONS

In this section, we suggest applications where the benefits of beamforming with non-RF modalities are highly relevant.

A. Visual IoT

Current IoT developments rarely include visual data, even though powerful camera sensors have been developed over the years. The state-of-the-art cameras use dynamic, neuromorphic sensors that mimic the eye processes of mammals and can record billions of pixels per second. The key components of visual IoT produce large amounts of video data [123]. This

![Graph showing performance of different unimodal and aggregated fusion network on multimodal Raymobtime dataset]
necessitates rapid communication between visual IoT sensor nodes (edge devices), gateways and visual cloud servers [124]. Communication between IoT sensors and the rest of the system becomes challenging when we consider the volume of visual data. Taking surveillance cameras (closed-circuit television) as an example, there are projections of 1 billion cameras being installed globally by 2025 - if we assume 12 hour recordings each day, we would produce 1250 petabytes of data [125]. If this volume of data only comes from surveillance cameras, we can imagine the speed of data streams needed to communicate visual IoT sensor data to gateways and servers.

Our topic of interest, multimodal beamforming with non-RF data, can offer a solution for the need of such ultra-fast communication between visual IoT sensors with high data rates. The ability to use non-RF data and leverage AI-based algorithms, to fuse the different data modalities, as part of the beamforming algorithms can revolutionize visual IoT and aid in implementation of smart cities. By enabling fast communication within networks of high-speed, high-resolution cameras with other modalities such as GPS or radar, multimodal beamforming can contribute to applications such as traffic management, emergency monitoring preparedness, air quality management and smart parking. Visual IoT sensors deployed for smart city applications include unmanned vehicles (UAVs and UGVs) traffic cameras, smart phone cameras and more. Non-RF multimodal beamforming enables those sensors to communicate in networks that provide city-wide coverage with low latency.

B. V2X Architecture

The Vehicle-to-Everything (V2X) market is estimated to be $689 million in 2020 and projected to reach $12,859 million by 2028 [126]. V2X will enable communication among vehicles as well as between vehicles and networks, infrastructure and pedestrians, aiming to improve traffic efficiency, road safety and individual vehicle energy efficiency [127]. V2X connectivity is also essential for the advancement of autonomous driving. Traffic efficiency improves by monitoring congested areas and providing alternative routes, while maintaining road safety by monitoring speed and identifying risky drivers. At the same time, V2X networks can improve energy efficiency by making vehicles more intelligent, choosing journeys with lower carbon emissions.

Different beamforming techniques have direct impact on the performance of 5G-V2X networks [128]. In [129], Lee et al. presents an object detection algorithm by fusing visual and LiDAR data to form 3D images of the vehicle surroundings. Combining these two concepts, we envision the application area of using multimodal beamforming for V2X architecture will extend from fast and reliable communication to object detection in urban scenarios. As mentioned in Sec. VI-A, availability of different types of sensor data forms the backbone of V2X communication. The multimodal beamforming using these sensors can be leveraged to provide low latency V2X communication as well as knowledge of selected beams at a specific position can be leveraged to detect objects or pedestrian using AI-enabled algorithms. An example use-case of pedestrian detection via beamforming is depicted in Fig. 14.
D. Extended Reality (XR)

The FCC plans to release up to 3 THz spectrum that will accelerate future standardization efforts beyond 5G [136]. By achieving extreme data rates and high device capacity in NextG networks, the requirements for augmented reality (AR), virtual Reality (VR), and mixed reality (MR) (comprehensively extended reality (XR)) applications can be met [137]. Terahertz (THz) frequencies may support the bandwidth for wireless cognition - the real-time wireless transfer of human brain computations. Such high-speed communication increases the quality of physical experience, an important factor for AR/VR/MR/XR applications.

At the THz frequencies, the limited transmission distance once again requires directional antennas and beamforming with a higher number of antenna elements. Multimodal beamforming with AI-enabled algorithms at the THz/sub-THz band can be used in XR systems to ensure fast beam alignment to meet the real-time, high-speed data communication requirements.

E. Multi-agent Robotics Applications

Autonomous agents are increasingly used in a variety of applications like mining [138], agriculture [139], military [140], aerospace [141] and medicine [142] to name a few. Many system entities need to collectively coordinate with each other to make decisions online and collaboratively in this paradigm. Some examples of multi-agent robots are simultaneous localization and mapping (SLAM) [143], [144], [145], warehouse robotics [146], [147], [148], surgical robotics [149], [150], autonomous driving [151], [152], [153], agricultural robotics [154], [155], [156] etc. In these applications, each of the agents in a multi-agent system may be equipped with sensors like LiDAR, RGB and IR cameras, GPS receiver etc., which enable them to function autonomously. Many applications rely on agents being able to communicate locally with other agents. For such applications and for real-time collaboration, high-speed communication is needed for sharing sensor information, decisions and actions. To support such large data-rate requirements, industries with highly automated process flows are pursuing high bandwidth communication links, including access to mmWave bands [157]. Multimodal beamforming using non-RF data can be an interesting approach to facilitate the faster communication between such autonomous entities using the integrated sensors within them.

IX. EMERGING RESEARCH FRONTIERS

In this section, we present selected research frontiers, where the AI-enabled multimodal beamforming techniques can make transformative difference.

A. Terahertz Communication

Early works that prove the feasibility of exploiting the THz frequency bands (0.3 THz to 10 THz) point towards an upcoming paradigm shift in the way wireless spectrum will be used. THz-band links bridge the gap between radio and optical frequency ranges, which may be game-changing for nextG wireless networks [158] by enabling transfer rates of 10Gb/s [159]. However, highly directional and fine-grained beams in the phased array antennas, which are essential to support the THz communication, come with their own challenges. Additionally, the beam search space increases with the increasing frequencies. Hence, there is urgent need to exploit out-of-the-box approaches, such as AI-enabled CSI estimation techniques, to decouple the number of antenna elements from the beamforming time overhead [44]. We believe the idea of multimodal beamforming can also be extended to THz communication to reduce the exploding search space of antenna codebook elements by leveraging the environmental multimodal data.

B. Virtual Presence

Since the start of the COVID-19 pandemic, we have quickly transitioned to using virtual communications platforms to aid in wellness and safety. However, platforms like Zoom, Teams can only do so much with respect to quality of user experience. Most of these platforms are still limited by the on-screen presence. This is where the recent development of AR/MR/VR/XR can make a difference by opening up the possibility of transforming on-screen presence to a virtual presence. The concept of holographic representation can emulate physical presence for meeting, gaming, or collaborating with others. Such virtual presence will supports mobility while giving a group presentation or multi-player gaming. AR/MR/VR/XR technologies will require multi-Gbps data-rates that may saturate a sub-mmWave band within seconds. Even the still-evolving 5G standard is not capable of supporting these data transfer rates. The standardization of ultra-fast beamforming in mmWave communication is integral for NextG standards [11]. The concept of using multimodal non-RF data in such applications is promising in this regard. The rich properties of XR or holographic images can be exploited for situational awareness to aid in the beamforming in high frequencies, where the codebook search space is generally too large to compute optimally in real time via exhaustive searching [160].

C. Hybrid Beamforming

mMIMO communications in hybrid transceivers is realized by a combination of high dimensional analog phase shifters and power amplifiers with lower-dimensional digital signal processing units [25]. For fully connected hybrid transceivers, the situational states through the non-RF modalities can be leveraged to select multiple phase shifters (multi-label prediction), which can be inferred to derive the best RF chains and aid in even-faster beamforming. Multimodal beamforming can be applied per RF chain to select best phase shifter, and this will enable the parallel inference of all the RF chains at the same time. Hence, the use of multimodal data has huge potential for improving the emerging hybrid beamforming technique it will allow seamless scaling to make it suitable for NextG networks.

D. Multiple User Massive MIMO (MU-MIMO)

Massive MIMO links for multiple concurrent users will soon become part of the 5G standard [21]. The open challenge of
making interference-free beam formation with multiple users with available RF chains is the main roadblock in MU-MIMO. Another challenge is to guarantee scalable, real-time signal processing in large MU-MIMO systems [161]. Similar to the other trends, the situational state information from different non-RF modalities (such as image, infrared, LiDAR) can be used to build trained models to address these challenges, while providing scalability with number of antennas and users.

X. CONCLUSIONS

This paper provides a comprehensive survey of using AI-enabled beamforming techniques for out-of-band and multi-modal non-RF data RF for mmWave-band operation in NextG networks. While the previous surveys on beamforming [24], [25], are focused more on analyzing and using mmWave channel characteristics, or channel state information for beamforming in massive MIMO leveraging the complicated hybrid beamforming process; our survey reviews recent trends in the literature that adopt an out-of-box approach for solving the same problem. We discuss the state-of-the-art in research trends, application areas, and open challenges of this exciting and emerging paradigm of multimodal sensor data-enabled beamforming. We also present novel AI-enabled fusion techniques which proves the effectiveness of exploiting multimodal data for non-RF data based beamforming. We identify open research challenges to motivate future researches and well as indicate the potential transformative impact of this area on different wireless applications.

ACKNOWLEDGMENT

This material is based upon the research sponsored by the Nvidia Inc.

REFERENCES

[1] Cisco, “Cisco Annual Internet Report (2018–2023) White Paper,” 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report-white-paper-c11-741490.html

[2] W. Haerick and M. Gupta, “5G and the Factories of the Future,” pp. 1–31, 2015. [Online]. Available: https://5g-ppo.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Factories-of-the-Future_Vertical-Sector.pdf

[3] Huawei, “5G: A technology vision,” https://www.huawei.com/mediaties/CORPORATE/PDF/Magazine/WinWin/HW_329327.pdf, 2014.

[4] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, “Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective,” IEEE Wireless Communications, vol. 21, no. 3, pp. 118–127, 2014.

[5] F. W. Vosk, A. Ghosh, and T. A. Thomas, “Mimo and beamforming solutions for 5G technology,” in 2014 IEEE MTT-S International Microwave Symposium (IMS2014), 2014, pp. 1–4.

[6] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, “A tutorial on beam management for 3gpp nr at mmwave frequencies,” IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 173–196, 2019.

[7] T. Nitsche, C. Cordeiro, A. B. Flores, E. W. Knightly, E. Perahia, and J. C. Widmer, “IEEE 802.11ad: Directional 60 GHz communication for multi-gigabit-per-second Wi-Fi,” IEEE Communications Magazine, vol. 52, no. 12, pp. 132–141, 12 2014.

[8] 3GPP, “Study on New Radio (NR) Access Technology - Physical Layer Aspects - Release 14,” TR 38.802, 2017.

[9] I. A. Hamadeh, K. Souyanyarana, M. El-Hajjar, and L. Hanzo, “Millimetre-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget,” IEEE Communications Surveys Tutorials, vol. 20, no. 2, pp. 870–913, 2018.

[10] R. Mailoux, Phased Array Antenna Handbook, Third Edition, 2017.

[11] J. Zhang, X. Yu, and K. B. Letaief, “Hybrid Beamforming for 5G and Beyond Millimeter-Wave Systems: A Holistic View,” IEEE Open Journal of the Communications Society, vol. 1, pp. 77–91, 2020.

[12] W. Zhang, J. Xiang, Y.-N. R. Li, Y. Wang, Y. Chen, P. Geng, and Z. Lu, “Field trial and future enhancements for tdd massive mimo networks,” in 2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2015, pp. 2339–2343.

[13] Y.-N. R. Li, B. Gao, X. Zhang, and K. Huang, “Beam management in millimeter-wave communications for 5g and beyond,” IEEE Access, vol. 8, pp. 13 282–13 293, 2020.

[14] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, “Lte-advanced: next-generation wireless broadband technology [invited paper],” IEEE Wireless Communications, vol. 17, no. 3, pp. 10–22, 2010.

[15] N. Soltni, K. Sankhe, S. Ioannidis, D. Jaisinghani, and K. Chowdhury, “Spectrum awareness at the edge: Modulation classification using smartphones,” in 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), 2019, pp. 1–10.

[16] K. Sankhe, M. Belgiovine, F. Zhou, L. Angioli, F. Restuccia, S. D’Oro, T. Melodia, S. Ioannidis, and K. Chowdhury, “No radio left behind: Radio fingerprinting through deep learning of physical-layer hardware impairments,” IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 1, pp. 165–178, 2020.

[17] D. Roy, T. Mukherjee, M. Chatterjee, E. Blasch, and E. Pasiliao, “Rfn: Adversarial learning for rf transmitter identification and classification,” IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 2, pp. 783–801, 2020.

[18] A. Klautau, P. Batista, N. Gonzalez-Prelcic, Y. Wang, and R. W. Heath, “‘5g mimo data for machine learning: Application to beam-selection using deep learning,” in 2018 Information Theory and Applications Workshop (ITA), 2018, pp. 1–9.

[19] Y. Niu, D. Li, L. Jin, and A. V. Vasilakos, “A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges,” Wireless Networks, vol. 21, pp. 2657–2676, 2015.

[20] R. Mailloux, Phased Array Antenna Handbook, Third Edition

[21] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, “An overview of massive mimo: Benefits and challenges,” IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 5, pp. 742–758, 2014.

[22] Y. Niu, D. Li, L. Jin, and A. V. Vasilakos, “A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges,” Wireless Networks, vol. 21, pp. 2657–2676, 2015.

[23] Facebook Inc., “Terragraph,” https://terragraph.com/, 2021.

[24] Facebook Inc., “Terragraph,” https://terragraph.com/, 2021.

[25] Facebook Inc., “Terragraph,” https://terragraph.com/, 2021.

[26] Facebook Inc., “Terragraph,” https://terragraph.com/, 2021.

[27] Facebook Inc., “Terragraph,” https://terragraph.com/, 2021.

[28] Facebook Inc., “Terragraph,” https://terragraph.com/, 2021.

[29] Facebook Inc., “Terragraph,” https://terragraph.com/, 2021.

[30] Facebook Inc., “Terragraph,” https://terragraph.com/, 2021.

[31] Facebook Inc., “Terragraph,” https://terragraph.com/, 2021.
S. Banou, M. Swaminathan, G. Reus Muns, D. Duong, F. Kulsoom, I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm,” Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, vol. 57, no. 11, pp. 1413–1457, 2004.

D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” Proceedings of the National Academy of Sciences, vol. 106, no. 45, pp. 18 914–18 919, 2009.

C. Li, W. Yin, and Y. Zhang, “User’s guide for tv3l: Tv minimization by augmented lagrangian and alternating direction algorithms,” CAAM report, vol. 20, no. 46-47, p. 4, 2009.

C. A. Metzler, A. Maleki, and R. G. Baraniuk, “From denoising to compressed sensing,” IEEE Transactions on Information Theory, vol. 62, no. 9, pp. 5117–5144, 2016.

C. K. Wen, W. T. Shih, and S. Jin, “Deep learning for massive mimo csf feedback,” IEEE Wireless Communications Letters, vol. 7, pp. 748–751, 2018.

M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, “A tutorial on beam management for 3gpp nr at mmwave frequencies,” IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 173–196, 2018.

3GPP, “Release 17,” https://www.3gpp.org/release-17, 2020.

C. N. Barati, S. Dutta, S. Rangan, and A. Sabharwal, “Energy and Latency of Beamforming Architectures for Initial Access in mmWave Wireless Networks,” Journal of the Indian Institute of Science, pp. 1–22, 2020.

Y. Yaman and P. Spasojevic, “Reducing the LOS Ray Beamforming Setup Time for IEEE 802.11 ad and IEEE 802.15.3c,” in MILCOM 2016-2016 IEEE Military Communications Conference. IEEE, 2016, pp. 448–453.

Y. Kishiyaama, A. Benjebbour, T. Nakamura, and H. Ishii, “Future steps of lite-a: evolution toward integration of local area and wide area systems,” IEEE wireless communications, vol. 20, no. 1, pp. 12–18, 2013.

N. González-Prelcic, R. Méndez-Rial, and R. W. Heath, “Radar Aided Beam Alignment in MmWave V2I Communications Supporting Antenna Diversity,” in 2016 Information Theory and Applications Workshop (ITA). IEEE, 2016, pp. 1–7.

A. Ali, N. González-Prelcic, and A. Ghosh, “Passive Radar at the Roadside Unit to Configure Millimeter Wave Vehicle-to-Infrastructure Links,” IEEE Transactions on Vehicular Technology, vol. 69, no. 12, pp. 14 903–14 917, 12 2020.

G. Reus, K. Vijay, C. Bocanegra, Y. C. Eldar, and K. R. Chowdhury, “Beam Alignment and Tracking for Autonomous Vehicular Communication using IEEE 802.11ad-based Radar,” in IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 4 2019.

T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, “Steering with eyes closed: Mm-Wave beam steering without in-band measurement,” in Proceedings - IEEE INFOCOM, vol. 26. IEEE, 4 2015, pp. 2416–2424.

A. Ali, N. González-Prelcic, and R. W. Heath, “Millimeter Wave Beam selection Using Out-of-band Spatial Information,” IEEE Transactions on Wireless Communications, vol. 17, no. 2, pp. 1038–1052, 2017.

M. Hashemi, C. E. Koskal, and N. B. Shroff, “Out-of-Band Millimeter Wave Beamforming and Communications to Achieve Low Latency and High Energy Efficiency in 5G Systems,” IEEE Transactions on Communications, vol. 66, no. 2, pp. 875–888, 2018.

E. M. Mohamed, H. Kusano, K. Sakaguchi, and S. Sampei, “WiFi assisted multi-WiGig AP coordination for future multi-Gbps WLANs,” in 2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). IEEE, 8 2015.

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

W. Alliance, “Wi-fi certified wigig - multi-gigabit, low latency connectivity,” https://www.wi-fi.org/discover-wi-fi/wi-fi-certified-wigig, 2019.

Y. Oguma, T. Nishio, K. Yamamoto, and M. Morikura, “Handover management for mmwave networks with proactive performance prediction using camera images and deep reinforcement learning,” IEEE Transactions on Cognitive Communications and Networking, vol. 6, no. 2, pp. 802–816, 2019.

“Cooperative sensing in deep RL-based image-to-decision proactive handover for mmWave networks,” in 2020 IEEE 17th Annual...
Consumer Communications & Networking Conference (CCNC). IEEE, 2020, pp. 1–6.

[75] T. Nishio and A. Ashok, “High-speed mobile networking through hybrid mmwave-camera communications,” in Proceedings of the 3rd Workshop on Visible Light Communication Systems, 2016, pp. 37–42.

[76] T. Nishio, H. Okamoto, K. Nakashima, Y. Koda, K. Yamamoto, M. Morikura, Y. Asai, and R. Miyatake, “Proactive received power prediction using machine learning and depth images for mmwave networks,” IEEE Journal on Selected Areas in Communications, vol. 37, no. 11, pp. 2413–2427, 2019.

[77] K. Nakashima, Y. Koda, K. Yamamoto, H. Okamoto, T. Nishio, M. Morikura, Y. Asai, and R. Miyatake, “Impact of input data size on received power prediction using depth images for mm wave communications,” in 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE, 2018, pp. 1–5.

[78] T. Mikuma, T. Nishio, M. Morikura, K. Yamamoto, Y. Asai, and R. Miyatake, “Transfer learning-based received power prediction using rgb-d camera in mmwave networks,” in 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring). IEEE, 2019, pp. 1–5.

[79] Y. Koda, J. Park, M. Bennis, K. Yamamoto, T. Nishio, M. Morikura, and K. Nakashima, “Communication-efficient multiport split learning for mmwave received power prediction,” IEEE Communications Letters, vol. 24, no. 6, pp. 1284–1288, 2020.

[80] Y. Koda, J. Park, M. Bennis, K. Yamamoto, T. Nishio, and M. Morikura, “The pixel image and rf signal based splitting for mmwave received power prediction,” in Proceedings of the 15th International Conference on Emerging EXPERIMENTS and Technologies, 2019, pp. 54–56.

[81] H. Okamoto, T. Nishio, M. Morikura, K. Yamamoto, D. Murayama, and K. Nakahira, “Machine-Learning-Based Throughput Estimation Using images for mmWave Communications,” IEEE Vehicular Technology Conference, vol. 2017-June, 2017.

[82] Y. Oguma, T. Nishio, K. Yamamoto, and M. Morikura, “Performance modeling of camera-assisted proactive base station selection for human blockage problem in mmwave communications,” in 2016 IEEE Wireless Communications and Networking Conference. IEEE, 2016, pp. 1–7.

[83] A. Ali, N. Gonzalez-Prelcic, R. W. Heath, and A. Ghosh, “Leveraging sensing at the infrastructure for mmwave communication,” IEEE Communications Magazine, vol. 58, no. 7, pp. 84–89, 2020.

[84] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, “Millimeter-wave vehicular communication to support massive automotive sensing,” IEEE Communications Magazine, vol. 54, no. 12, pp. 160–167, 2016.

[85] “Yole Développement.” [Online]. Available: http://www.yole.fr

[86] M. Alrabeiah, A. Hredzak, Z. Liu, and A. Alkhateeb, “Viwi: A deep learning dataset framework for vision-aided wireless communications,” in Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services, 2021, pp. 389–401.

[87] Y. Koda, J. Park, M. Bennis, K. Yamamoto, T. Nishio, and M. Morikura, “The pixel image and rf signal based splitting for mmwave received power prediction,” in Proceedings of the 15th International Conference on Emerging EXPERIMENTS and Technologies, 2019, pp. 54–56.

[88] H. Okamoto, T. Nishio, M. Morikura, K. Yamamoto, D. Murayama, and K. Nakahira, “Machine-Learning-Based Throughput Estimation Using images for mmWave Communications,” IEEE Vehicular Technology Conference, vol. 2017-June, 2017.

[89] Y. Oguma, T. Nishio, K. Yamamoto, and M. Morikura, “Performance modeling of camera-assisted proactive base station selection for human blockage problem in mmwave communications,” in 2016 IEEE Wireless Communications and Networking Conference. IEEE, 2016, pp. 1–7.

[90] A. Ali, N. Gonzalez-Prelcic, R. W. Heath, and A. Ghosh, “Leveraging sensing at the infrastructure for mmwave communication,” IEEE Communications Magazine, vol. 58, no. 7, pp. 84–89, 2020.
