2021 17th International Conference on Mobility, Sensing and Networking (MSN) | 978-1-6654-0668-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/MSN53354.2021.00087

Deep Learning on Visual and Location Data for V2I mmWave Beamforming

Guillem Reus-Muns, Batool Salehi, Debashri Roy, Tong Jian, Zifeng Wang, Jennifer Dy, Stratis Ioannidis,
Kaushik Chowdhury
Department of Electrical and Computer Engineering
Northeastern University, Boston, USA
E-mail: {greusmuns, bsalehihikouei, droy, jian, zifengwang, jdy, ioannidis, krc}@ece.neu.edu

***Abstract*—Accurate beam alignment in the millimeter-wave (mmWave) band introduces considerable overheads involving brute-force exploration of multiple beam-pair combinations and beam retraining due to mobility. This cost becomes often intractable under high mobility scenarios, where fast beamforming algorithms that can quickly adapt the beam configurations are still under development for 5G and beyond. Besides, blockage prediction is a key capability in order to establish mmWave reliable links. In this paper, we propose a data fusion approach that takes inputs from visual edge devices and localization sensors to (i) reduce the beam selection overhead by narrowing down the search to a small set containing the best possible beam-pairs and (ii) detect blockage conditions between transmitters and receivers. We evaluate our approach through joint simulation of multi-modal data from vision and localization sensors and RF data. Additionally, we show how deep learning based fusion of images and Global Positioning System (GPS) data can play a key role in configuring vehicle-to-infrastructure (V2I) mmWave links. We show a 90% top-10 beam selection accuracy and a 92.86% blockage prediction accuracy. Furthermore, the proposed approach achieves a 99.7% reduction on the beam selection time while keeping a 94.86% of the maximum achievable throughput.**

***Index Terms*—mmWave, beam selection, machine learning, deep learning, fusion, multi-modal data, 5G.**

I. Introduction

Connected and autonomous vehicles are quickly being integrated into multiple industrial and civil applications given the implicit advantages of energy consumption, comfort and safety. However, these new systems require to manage enormous amounts of data to enable real-time planning, video processing and sensing, among others. In addition, many of these applications are time-sensitive and require real-time data processing and highly reliable links. For instance, camera sensors may generate data up to 700MB/s and most automotive applications have latency requirements of <100ms.

A. Motivation and Challenges in use of V2I mmWave links

Millimeter-wave (mmWave) is considered as the main candidate to satisfy the needs for high data rates in V2I scenarios and has already been adopted by 5G and WLAN (802.11ad/ay) standardization groups. Furthermore, mmWave and TeraHertz systems, both requiring directional antenna arrays, are envisioned to play a major role in the future communication systems of 5G and beyond. However, transmission at these bands suffers from high attenuation associated with increasing carrier frequency and channel losses arising from natural phenomena such as atmospheric absorption. Highly directional antennas or arrays with multiple elements acting as beamformers are found to be the best solution to extend the range in these high frequencies. While these systems are proven to be efficient in static scenarios, no unifying solution for coordinated directional communications under mobility constraints has been widely accepted. Exhaustive search approaches where both transmitter and receiver explore all possible beam directions are time-consuming and not feasible under mobility scenarios where such process should be repeated constantly. Hence, techniques that avoid time-consuming beam selection procedures and predict blockage are required for mmWave V2I networks.

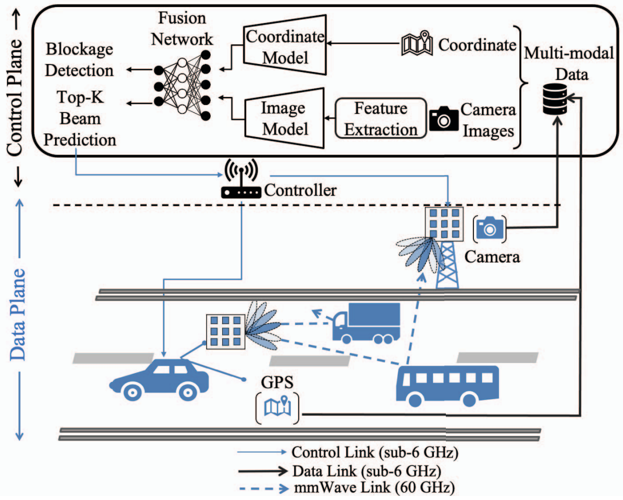

Fig. 1: Visual and location data are combined using data fusion techniques. Different neural networks are used to process each data modality, and are ultimately combined with the fusion neural network to (i) aid the beam selection and (ii) detect the blockage.

In addition to high attenuation, mmWave links are found to be susceptible to blockage due to their high penetration loss. Hence, Line-of-Sight (LoS) connections are strongly preferred over non-LoS (NLoS), specially for cellular systems where the distances between a base station and a user can be of a few hundreds of meters. Thus, blockage prediction is a key

978-1-6654-0668-0/21/$31.00 ©2021 IEEE
DOI 10.1109/MSN53354.2021.00087

capability required to establish reliable mmWave links and propose counteractive measures to NLoS conditions.

Interestingly, commercial devices are becoming more complex, integrating multiple sensors that provide data in a variety of sensing modalities, creating intelligent *Internet of Things* (IoT) systems. In particular, recent advances in image processing and computer vision open up a world of opportunities for new visual IoT (V-IoT) applications, including wireless communications.

B. Vision-Aided Beam Selection

Self-driving systems are often equipped with camera systems, providing autonomy and reliability through visual understanding of the surroundings. While such capabilities are widely exploited for assisted and autonomous driving, they have been under-explored in other domains that could also leverage from context-aware information, such as wireless communications [1]. For instance, camera images are able to enhance both beam selection and blockage detection for mmWave links by providing contextual information that other RF technologies cannot offer. However, visual data has its limitations as well (i.e. standalone images are insufficient for multi-user scenarios). Such limitations can be addressed through intelligent data fusion with other sensing modalities deployed into commercial vehicles (GPS modules, radars, LIDARs, etc). Thus, combining different out-of-band data aims to intelligently fuse each sensing modality by highlighting the advantages that each one provides, similar to Fig. 1, where images are combined with GPS data to enhance beam selection and blockage prediction. Nonetheless, data fusion faces challenges of its own and robust control channels are required if data collection and computation are not co-located.

C. Summary of the Contributions

In this paper, we propose a deep learning based fusion framework for beam selection which leverages the GPS location information along with visual data yielding to a low-overhead fast beamforming process under mobility scenarios. We list the novel contributions as follows:

- We design a custom image feature extractor for visual data that identifies information relevant to beam selection while filtering out background clutter.
- We propose a deep learning architecture that uses visual information top-K best beam-pairs in order to reduce the beam selection overhead. Additionally, we designed a data fusion neural network that intelligently combines the location and image data in order to improve the top-K beam selection accuracy.
- We analyze the impact of the LoS/NLoS conditions in terms of beam selection accuracy and propose a method to predict mmWave link blockage.
- We provide numerical results of the overhead reduction as well as the link quality for the proposed approach.

The rest of the paper is organized as follows. We summarize the related work on out-of-band beam selection in Sec. II. Sec. III and Sec. IV present the system model and contributing overheads in the proposed framework, respectively. The detailed description of the proposed solution is presented in Sec. V. We present an evaluation of our approach in Sec. VI. Finally, we discuss potential future research directions and conclude the paper in Sec. VII and Sec. VIII, respectively.

II. Related Work

We summarize the out-of-band techniques that leverage the data from different data modalities to achieve low-overhead beam alignment solutions in Fig. 2. We highlight that no prior work has explored data fusion of visual data with location information.

A. Cross Channel Correlation

Exploiting channel knowledge at lower/higher frequency bands such as sub-6 GHz and radar has shown promising results on aiding the beam selection.

1) Sub-6 GHz: To the best of our knowledge, exploring the sub-6 GHz channel properties as a single out-of-band technology for V2I beam selection was only investigated in [2]. However, multiple works have explored sub-6GHz in conjunction with other data modalities, as we summarize in Sec. II-B4.

2) Radar: In [3] the concept of radar-aided vehicular communication is introduced, where the radar is exploited as an additional source of information for V2I mmWave beamforming. The authors in [4] leverage the PHY layer 802.11ad frames to perform both radar operations and conventional communications using a standard-compliant Tx/Rx chain. The radar is employed to estimate the location of the vehicles and consequently assist the beam selection. In [5], a passive radar receiver is placed at the roadside unit to tap the transmissions from other automotive radar. The spatial covariance of the radar signals is explored to establish the communication link. Finally, [6] uses radars to estimate the azimuth power spectrum and compare it with the one obtained from a communication system.

B. Use of non-RF Sensor Data

Different kinds of sensors are being integrated into commercial and industrial systems in various ways. For example, smartphones are starting to be equipped with LIDARs on top of multi-camera systems and autonomous cars are assisted with multiple sensor and vision systems. In this subsection, we summarize how different data modalities can help reduce the overhead.

1) Localization: The authors in [7] propose a localization-based beam selection algorithm that explores geometrical patterns through the location of not only the vehicle involved in the communication but all the neighboring vehicles. In [8], the multipath channel fingerprints are characterized and stored in a database. Then, every newly obtained fingerprint is used to query such database to provide a set of potential beam directions for reliable and fast-changing beam alignment.

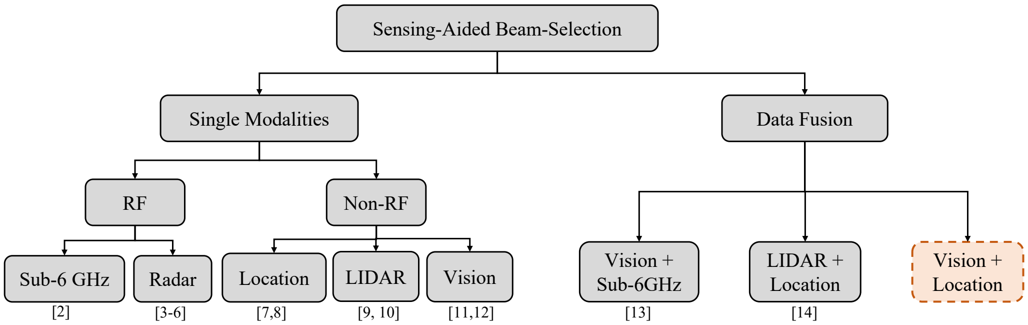

Fig. 2: Sensing-aided beam selection modalities. Multiple RF, non-RF and fusion of out-of-band sensing modalities has been explored. We highlight that no prior work combines vision and location information.

2) LIDAR: LIDARs are used in autonomous vehicles for obtaining accurate mappings of the environment and high-resolution positioning. The authors in [9] leverage LIDAR in different forms for LoS detection and beam selection overhead reduction. Similarly, Woodford *et al.* [10] generate a 3D map of the environment by multiple LIDAR measurements and feed it to ray-tracing software to predict the reflection patterns. However, practical considerations such as the costly pre-processing of high-dimensional point clouds raise the concern on whether the accurate LIDAR mapping is a feasible candidate for achieving low-overhead beamforming.

3) Vision: Image processing and computer vision have enormously improved their capabilities in recent years due to the latest advances in deep learning. Such algorithms are employed in multiple domains, providing multiple forms of automated understanding of visual data. Similarly, context-aware information obtained from images can also be exploited to assist wireless communications systems. In particular, visual data can help to address some of the challenges associated with establishing reliable mmWave links, such as beam selection or blockage prediction, as we propose in this paper. Vision-based beam prediction was first proposed in [11], where a simulation-based beam tracking solution is presented. The authors train a neural network to predict the next beam direction for different numbers of time-steps in the future. We also highlight the work in [12], while the authors do not face the challenges of V2I, mmWave radios and cameras are used to showcase the first implementation of a vision-based beam selection system.

4) Fusion: To overcome the potential weaknesses and combine the advantages of single sensing modalities, a variety of methods leveraging different sensory data are proposed in the literature. The authors in [13] explore the gains of employing images to overcome blockage and enhance the beam selection process, in combination with sub-6 GHz channel information. Cameras are mounted at the mmWave base station to provide visual information, which is fused with other out-of-band channels. In [14] the authors propose a distributed approach where the best beam-pair inference is computed at the LIDAR-equipped vehicle. The mmWave BS broadcasts its location to the vehicles, which is combined with LIDAR data to predict the best beam.

While multiple data fusion approaches have been explored, we believe that the combination of multiple sensing modalities has only scratched the surface of its possibilities. In this paper, we showcase the fusion of GPS and images to aid the V2I mmWave beam selection process, which remains unexplored up until now.

III. System Model

We consider a V2I cellular communication system mainly operating in the mmWave band with an available sub-6 GHz control channel. Thus, both transmitter and receiver are equipped with two RF chains each. We assume a discrete number of mmWave antenna array configurations that sectorizes the space into a set of possible directions. Predefined and fixed beam codebooks are assumed on both ends, expressed as $B_{tx} = \{1, ..., t, ..., N_{tx}\}$ and $B_{rx} = \{1, ..., r, ..., N_{rx}\}$, where *t* and *r* represent the transmitter and receiver beam indices, respectively. The total number of beam-pairs depends on the transmitter and receiver codebooks sizes, defined as N_{tx} and N_{rx}, respectively. Then, every beam-pair can be expressed as (*t*, *r*), where $t \in B_{tx}$ and $r \in B_{rx}$ represent the beam indices. Additionally, we define (t_b, r_b) as the beam-pair providing the best performance (highest received signal strength) out of all the $N_{tx} \times N_{rx}$ possible configurations.

We develop an ML-based method to reduce the beam-selection overhead for a V2I cellular system. We assume that the vehicle is equipped with a GPS module and broadcasts its absolute location and vehicle type periodically to the BS using the control channel in the sub-6 GHz band. Notice that this functionality will most likely be included in the most self-driving cars as part of collision avoidance mechanisms.

The vehicle location is then combined with the visual information obtained from the cameras located at the BS to assist the beam-selection. We run the inference model at the BS locally in order to avoid additional latency derived from communication with the cloud. This assumption is backed up by the recently proposed 5G functional splits between the Baseband Units (BBU) and the Remote Radio Heads (RRH) [15].

A. Dataset

We evaluate our approach using the publicly available mmWave Raymobtime multimodal datasets (s008 and s009) [16]. The simulation consists of an urban scenario with a single BS and multiple vehicle types (bus, car, or truck). Each scene is labeled with the best possible beam-pair out of the 256 available combinations (N_{tx} and N_{rx} are 32 and 8, respectively). We use the coordinate and image modalities to evaluate our proposed approach. The target receiver location is expressed as a 2-point coordinate. The images are captured from three co-located cameras with slightly different angles. Additionally, we notice that not all beam-pair combinations are equally represented in the dataset, for which we will explore corrective measures in future work.

IV. Beam Selection Overhead

For comparison purposes, we describe the beam selection time required to find the (t_b, r_b) following the 5G-NR standard compliant operation as well as our sensing-aided beam selection approach.

A. Beam Selection Overhead in 5G-NR

The beam selection process happens during the initial access, where the gNodeB and user exchange a number of messages to find the best beam-pair combination. In particular, the gNodeB sequentially transmits synchronization signals (SS) in each codebook element $t \in B_{tx}$. In the meanwhile, the user switches through all its codebook configurations $r \in B_{rx}$, until all possible configurations are explored.

The standard defines an SS block as the set of SS transmitted under the same beam configuration, with multiple SS blocks further grouped into SS burst. Hence, in order to explore all beam-pair combinations, a total of $|\mathcal{B}| = N_{tx} \times N_{rx}$ SS blocks need to be transmitted. 5G-NR defines the maximum SS burst duration (T_{ssb}) to 5ms, which is transmitted with a periodicity (T_p) of 20ms [17].

The mmWave band allows a maximum of 32 SS blocks within a SS burst, which enables exploring up to 32 different beams within one SS burst. Thus, given the limit on SS blocks per SS burst, the total time to explore all beam-pair combinations (T_{bs}^{nr}) can be formulated as:

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

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

B. Beam Selection Overhead of the Proposed Approach

Our proposed approach provides a reduction on the beam search by intelligently combining the image and location data. Thus, the beam search space is reduced from $|\mathcal{B}|$ to a subset of $K \ll |\mathcal{B}|$ likely beam candidates, which are the K candidates with the highest predicted probabilities. We recall that the NR standard assumes that up to 32 sectors can be swept within $5\,ms$. Thus, we define the time to explore one single beam as $T_b = 5\,ms/32 = 156\,ns$. Then, the required time for sweeping the selected top-K beam-pairs can be expressed as:

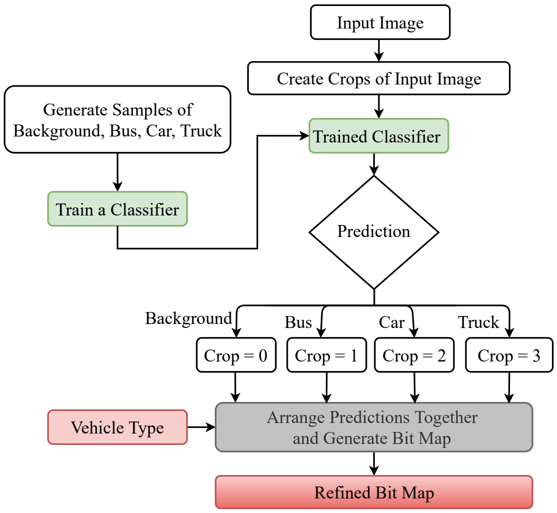

Fig. 3: Image feature extractor front end, the input images are fed to a model that segments the image into 4 possible labels: Background (0), Bus (1), Car (2), Truck (3).

$$T_{sweep}(K) = T_p \times \left\lfloor \frac{K-1}{32} \right\rfloor + T_b \times (1 + (K-1) \bmod 32). \qquad (2)$$

Notice that the image processing time and the vehicle positioning feedback are neglected, and only the beam sweeping time is considered in this analysis.

V. Data Fusion for Beam Selection

In this section, we describe our beam selection vision-based approach. Next, we propose a fusion network that intelligently combines GPS and visual data to further enhance the beam selection accuracy.

A. Visual Data Pre-processing

Visual images capture information from the overall scene. However, multiple vehicles are present in one snapshot and multiple regions of an image contain irrelevant background information. Thus, similar to [12], we explore design a feature extractor to provide simple contextual information, such as vehicle detection and background removal. To do so, we construct a dataset out of windowed images from the full dataset images. Each window is accurately selected so that it only contains a certain vehicle type (car, bus and truck) or background. We manually label a small set and extend the dataset by cropping each labeled image with $W \times W$ windows. Next, we train a neural network to classify these four different classes. We use a simple architecture with one convolutional layer with rectified linear units (ReLU) activation, a maxpool layer, dropout and two fully connected layers with ReLU and softmax activations, respectively. Next, we use the trained classifier to separate different vehicles from the background

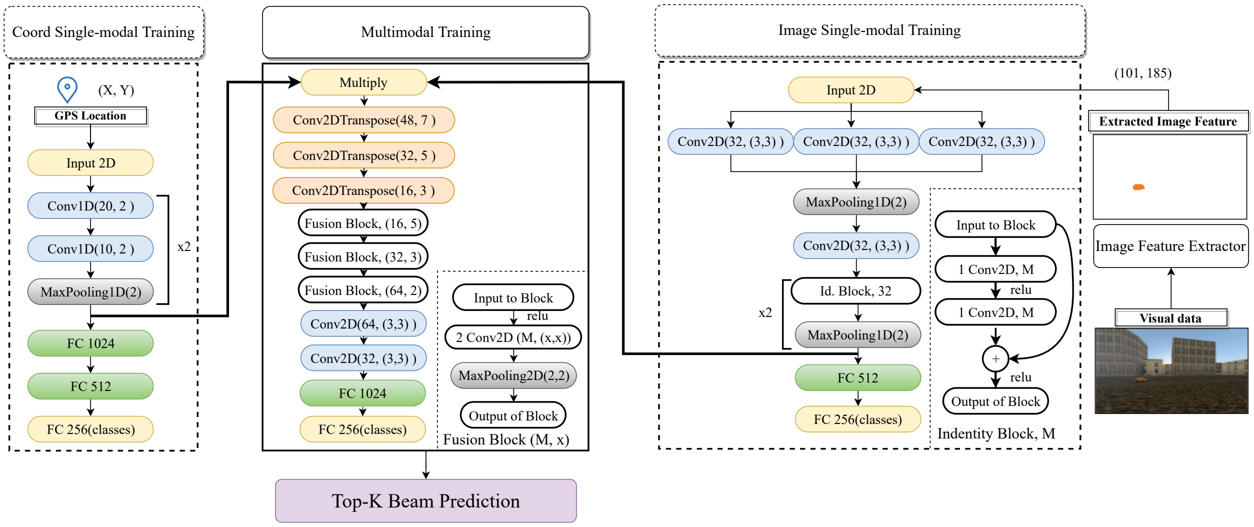

Fig. 4: Neural Network architectures for single-modal and multimodal implementations.

in the images collected by the camera on the BS. First, we quantize the input images into multiple $W \times W$ crops in steps of S, referred to as stride size. Then, using the above-described model, each crop is classified into Background (0), Bus (1), Car (2), Truck (3). Finally, we leverage the vehicle type knowledge of the receiver to set the different vehicle types as -1 (i.e., If a certain user is a car, each cropped area where a Bus (1) or a Truck(3) are detected, would be set to -1). At the output, we obtain a dimensionally reduced version of the original image, in form of a bitmap that separates the target vehicle type from the background and other vehicle classes.

In our evaluation, we only use the images from a single camera since the different angles are comparable and adding information from the remaining cameras does not result in increased performance. We set W and S to 40 and 5, respectively. We obtain an accuracy of 84% on background and vehicle type detection. The raw visual information is expressed in form of 960×540 RGB images. The output of our custom image feature extractor is a bitmap of size $(101, 185)$.

B. Beam Selection using Visual Data

In this subsection, we present our designed model architecture for predicting the best beam-pair based on the images extracted features. We show the model architecture in Fig. 4. In the first layer, we use an inception module. Specifically, we apply 3×3, 7×7, and 11×11 convolutional layers to the input features altogether. Different sized convolutions are employed to extract spatial features on different levels and all three feature maps are concatenated at the output of each layer. The next two layers are a max polling and a convolutional layer with $(3, 3)$ kernels. The next modules are inspired by ResNet against overfitting. Each module contains two convolutional layers with 32 kernels of size 3×3, and an identity shortcut connection that skips these two layers, followed by a 2×2 max pooling layer. The last two fully-connected layers act as a classification layer. We use dropout of 0.25 and ReLU activation function in each convolutional and fully-connected layers.

C. Proposed Fusion of Location Sensor with Visual Data

While location data provides a tremendous advantage for beam-selection, it does not provide any environmental information and becomes incomplete as a standalone sensor. In contrast, vision-aided beam-selection becomes challenging when multiple users are captured in the camera images and no other information is provided. Hence, in this paper, we pair a camera system with GPS sensors information into a fusion framework that intelligently combines each sensing modality.

Prior to the data fusion stage, the location information is processed by a dedicated neural network. We show the architecture used for location information in Fig. 4. Next, the extracted features from the image and GPS data are combined into the data fusion network. In order to maintain the spatial information, the proposed fusion network takes the features from the last convolutional layers of each single modality network. The different single-modal features are fused using an element-wise product operation. Next, transposed convolutional layers are used to expand the dimensionality of the original input. The rest of the architecture is presented in Fig. 4. In the following section, we evaluate the beam selection performance by using each single data modality independently as well as following the data fusion approach described above.

VI. Experiments and Results

In this section, we describe the training process and provide results for top-K beam selection accuracy, as well as throughput ratio and beam selection time in comparison to the 5G-NR standard approach. Finally, we provide insights on the importance of blockage detection and provide prediction results that justify the feasibility of the approach.

TABLE I: Performance of single-modal and multi-modal approaches.

Modality	Top-1				Top-2	Top-5	Top-10	Top-30	Top-50
	Acc	Precision	Recall	F_1 Score	Acc	Acc	Acc	Acc	Acc
Visual data	16.70%	0.0%	0.0%	0.0%	31.8%	58.2%	78.46%	91.88%	95.68%
Coordinate	54.72%	74.04%	27.39%	32.8%	71.4%	83%	87.71%	96.99%	98.91%
Fusion	**57.53%**	69.59%	42.94%	**45.82%**	**75.61%**	**87.96%**	**93.4 %**	**98.11%**	**99.07%**

A. Training Parameters and Evaluation Metrics

We use the *softmax* function as an activation in the classifier layer, and categorical cross-entropy as the loss function. We train the model with a batch size of 64, we use Adam optimizer with $\beta = (0.9, 0.999)$ and initialize the learning rate to 0.0001. To analyze the performance of the proposed approach, we use top-K accuracy, as well as F_1 score, as performance metrics. Notice that the F_1 score is relevant in this work given the imbalanced nature of the used dataset. Additionally, we also evaluate the overhead as beam selection time, and the link quality as throughput ratio.

B. Performance of Proposed Fusion Technique

The training is done similarly for both single and multi-modal models. The first row in Table I represents different test accuracies for top-K beam selection using only visual data. We observe how a standalone vision-based beam selection requires a considerably high *K* to achieve acceptable accuracy. One the other hand, the results (Table I) reveal that the fusion of coordinate modality with visual data significantly improves the accuracy for the top-K beam selection as well as the F_1 score. We observe that the coordinate gives better performance compared to visual data; however, the fusion of both modalities improves the performance over single-modal implementation. In particular, a 93% top-10 accuracy is achieved using the fusion model, which further reduces tje beam alignment time. The improvement in F_1 score is also very important for the considered imbalanced dataset. The source codes for our implementation are available in [18].

C. Throughput Ratio

While the proposed out-of-band method outperforms the state-of-the-art mmWave standard in time, we also need to compare them with respect to the received power strength. We define a metric *throughput ratio* to account for the the degradation in the performance of the system caused by not capturing the optimum beam direction in a single shot. Considering the set of all possible beam directions within the transmitter and receiver codebooks (t, r), the optimum beam direction is defined as (t_b, r_b) according to Sec. III. The 5G-NR standard detects the optimum direction by sending beacons over each sector as described in Sec. IV-A. On the other hand, the proposed method exploits the side information from optical cameras and GPS positioning to provide an optimality probability estimation of each beam configuration. We denote the top model prediction by (t_p, r_p) and define the throughput ratio as follows:

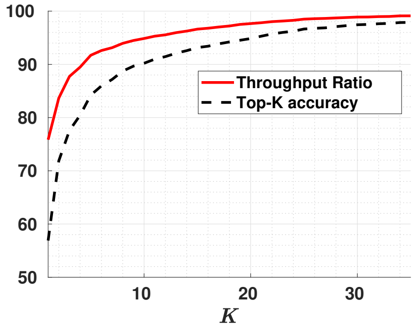

Fig. 5: Throughput ratio and top-K accuracy for different K values.

$$R_T = \frac{1}{N_t} \sum_{n=1}^{N_t} \frac{log_2[1 + y_{(t_b,r_b)}(n)]}{log_2[1 + y_{(t_p,r_p)}(n)]}, \tag{3}$$

where N_t is the total number of test samples and $y_{(t_b,r_b)}$ and $y_{(t_p,r_p)}$ denote the received power associated with the ground-truth and predicted beam-pairs, respectively. Fig. 5 shows top-K accuracy and throughput ratio with respect to increasing K. From this figure, we observe that top-K accuracy and throughput ratio starts with 57.5% and 75.8% for $K = 1$ and achieves $\sim 99\%$ throughput ratio when $K = 30$, which corresponds to a much smaller search space than the 5G-NR standard, i.e. 256 beam-pairs.

D. Beam Selection Overhead

As described in Sec. IV, the beam selection process introduces an overhead that is dependent on the number of available beam combinations (K). Here, we numerically analyze how the proposed reduction in the beam space translates into a faster initial access. In Fig. 6, it can be observed that the 5G-NR process requires a total 145ms to find the best beam-pair out of the 256 available options. In contrast, our approach achieves a throughput ratio of 94.8% and 97.8% with a total overhead of 1.6ms and 3.28ms, respectively. Thus, as mentioned previously, our beam selection approach provides a reduction in beam search space, which considerably reduces the mmWave initial access overhead.

E. The Importance of Line-Of-Sight (LoS): Blockage Prediction

As mentioned in the Sec. I, the presence of blockage can lead to massive drops in channel quality given the high attenuation in the mmWave band. Additionally, users might experience a considerable reduction in their quality of service

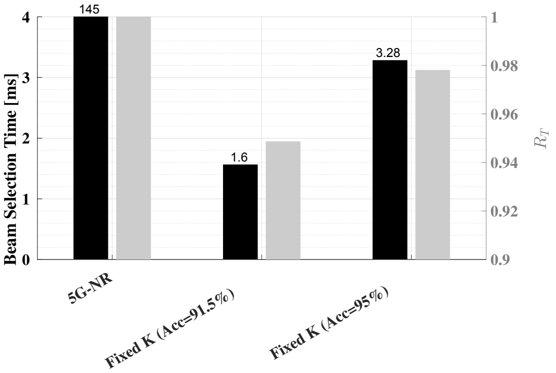

Fig. 6: Overhead beam selection analysis for 5G-NR and the proposed approach. K has been fixed to ensure a top-K accuracy of 91.5% and 95%. It can be observed how the brute force approach achieves a throughput ratio of 100%, as expected. However, the proposed approach provides an $R_T > 0.95$ in both cases.

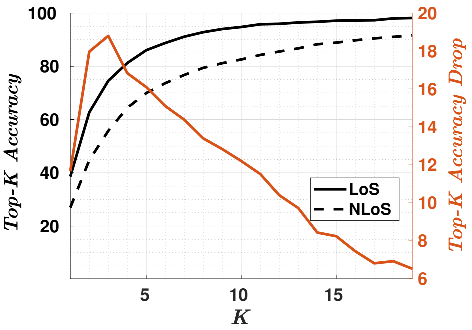

Fig. 7: LoS and NLoS *Top-K* accuracy comparison. NLoS face a drop in beam selection accuracy due to blockage. We observe how NLoS can be affected by >18% drop in beam selection accuracy (right y-axis). Thus, blockage prediction techniques are relevant to establish reliable mmWave links.

(QoS) to tens of Gbps. On top of the adversity nature of NLoS mmWave links under perfect beam alignment is assumed, finding the best beam-pair (t_b, r_b) in the first place is more challenging if a LoS link is not available. We show this in Fig. 7, where we compare the *Top-K* accuracy for NLoS and LoS links. As expected, predicting the complex reflections of NLoS links to find the best direction of transmission is more challenging and results in a drop in the beam selection accuracy. In particular, we observe a drop of $\approx 13\%$ in top-1 accuracy and a worse-case of 18.79% accuracy drop in the top-3 case. These results were generated using the same model and data fusion technique described earlier in this section. Thus, being able to detect blockage conditions is key to develop algorithms that can work well under those scenarios or predict the link reliability. Here, we train the same fusion network in Fig. 4 for blockage detection task. We just modify the final fully-connected layer to the new number of classes (*blockage* and *no-blockage*). We achieve a blockage prediction accuracy of 92.86%. While we trained a new neural network from scratch, we argue that multi-task learning could be exploited for joint beam selection and blockage detection predictions. We envision an architecture where multiple fusion networks combine features extracted from every single modality with different objectives, providing fast machine learning to solve a variety of problems in a joint manner.

VII. Open Research Challenges

There are several research opportunities for future work in data fusion V2I mmWave links, and we highlight some of them below. Data for different sensing modalities is likely to be collected in a distributed manner, whereas inference tends to happen at a centralized entity. Hence, reliable and low-latency control channels that enable data sharing among devices are needed, which might require accurate re-design of existing control channel technologies. Also, sharing certain sensing modalities might require bandwidths not available at all times. Thus, it is interesting to design modular fusion architectures that can operate under missing data or to include additional incoming data (multiple camera angles). Additionally, in this work, we explored two different tasks (blockage prediction and beam selection) independently. However, multi-task learning approaches where different fusion networks are trained for different tasks will increase efficiency of memory and computation resources, and inference time. Also, techniques that enable fast inference of deep learning models, such as pruning [19], are needed to enable real-time response times in high mobility scenarios. Finally, exploring data fusion of additional sensing modalities (i.e. LIDAR) will provide additional features to increase accuracy and reliability for multiple applications of wireless communications.

VIII. Conclusions

The widespread availability of camera sensors, in combination with the recent advances in computer vision is a potential candidate to reduce mmWave beam selection overhead. Additionally, data fusion of such visual information with location data, in combination with deep learning, opens up a new world of opportunities for problem-solving in the wireless communication [20]–[24] domain. In this paper, we propose a custom-designed data fusion network that successfully identifies the set of beams with highest link quality probabilities. In particular, the proposed approach provides a reduction of 97,7% beam selection overhead versus the brute force approach while maintaining a throughput ratio of 97,8%. Additionally, we show a 92.86% blockage detection while leveraging the same fusion network design.

Acknowledgement

This work is supported by the NSF grant CCF-1937500.

References

[1] N. González-Prelcic, A. Ali, V. Va, and R. W. Heath, "Millimeter-Wave Communication with Out-of-Band Information," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 140–146, 2017.

[2] 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.

[3] 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.

[4] G. R. Muns, K. V. Mishra, C. B. Guerra, Y. C. Eldar, and K. R. Chowdhury, “Beam Alignment and Tracking for Autonomous Vehicular Communication using IEEE 802.11 ad-based Radar,” in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2019, pp. 535–540.

[5] A. Anum, N. GonzalezPrelcic, and A. Ghosh, “Passive Radar at the Roadside Unit to Configure Millimeter Wave Vehicle-to-Infrastructure Links,” *IEEE Transactions on Vehicular Technology*, 2020.

[6] 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.

[7] Y. Wang, A. Klautau, M. Ribero, M. Narasimha, and R. W. Heath, “Mmwave Vehicular Beam Training with Situational Awareness by Machine Learning,” in *2018 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2018, pp. 1–6.

[8] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath, “Inverse Multipath Fingerprinting for Millimeter Wave V2I Beam Alignment,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4042–4058, 2017.

[9] A. Klautau, N. González-Prelcic, and R. W. Heath, “LIDAR Data for Deep Learning-Based mmWave Beam-Selection,” *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 909–912, 2019.

[10] T. Woodford, X. Zhang, E. Chai, K. Sundaresan, and A. Khojastepour, “SpaceBeam: LiDAR-driven One-Shot mmWave Beam Management,” in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 389–401.

[11] M. Alrabeiah, J. Booth, A. Hredzak, and A. Alkhateeb, “ViWi Vision-Aided mmWave Beam Tracking: Dataset, Task, and Baseline Solutions,” *arXiv preprint arXiv:2002.02445*, 2020.

[12] B. Salehi, M. Belgiovine, S. Garcia Sanchez, J. Dy, S. Ioannidis, and K. Chowdhury, “Machine Learning on Camera Images for Fast mmWave Beamforming,” in *IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2020.

[13] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, “Millimeter Wave Base Stations with Cameras: Vision-Aided Beam and Blockage Prediction,” in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, 2020, pp. 1–5.

[14] M. Dias, A. Klautau, N. González-Prelcic, and R. W. Heath, “Position and LIDAR-aided mmwave Beam Selection using Deep Learning,” in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[15] L. M. Larsen, A. Checko, and H. L. Christiansen, “A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 146–172, 2018.

[16] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang, and R. W. Heath, “5G MIMO Data for Machine Learning: Application to Beam-Selection Using Deep Learning,” in *2018 Information Theory and Applications Workshop (ITA)*, 2018, pp. 1–9.

[17] 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.

[18] G. Reus-Muns, B. Salehi, D. Roy, Z. Wang, and T. Jian, “VisualIoT,” https://github.com/debashriroy/Visual_Location_based_Beamforming, 2021.

[19] Z. Wang, T. Jian, K. Chowdhury, Y. Wang, J. Dy, and S. Ioannidis, “Learn-Prune-Share for Lifelong Learning,” in *IEEE International Conference on Data Mining (ICDM)*, 2020.

[20] D. Roy, T. Mukherjee, M. Chatterjee, and E. Pasiliao, “Detection of Rogue RF Transmitters using Generative Adversarial Nets,” in *IEEE Wireless Communications and Networking Conference*, 2019, pp. 1–7.

[21] D. Roy, T. Mukherjee, M. Chatterjee, E. Blasch, and E. Pasiliao, “RFAL: Adversarial Learning for RF Transmitter Identification and Classification,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 783–801, 2020.

[22] D. Roy, T. Mukherjee, M. Chatterjee, and E. Pasiliao, “Primary User Activity Prediction in DSA Networks using Recurrent Structures,” in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2019, pp. 1–10.

[23] M. Belgiovine, K. Sankhe, C. Bocanegra, D. Roy, and K. R. Chowdhury, “Deep Learning at the Edge for Channel Estimation in Beyond-5G Massive MIMO,” *IEEE Wireless Communications*, vol. 28, no. 2, pp. 19–25, 2021.

[24] D. Roy, T. Mukherjee, M. Chatterjee, and E. Pasiliao, “Defense against PUE Attacks in DSA Networks Using GAN Based Learning,” in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.