Computer Vision Aided Beam Tracking in A Real-World Millimeter Wave Deployment

Shuaifeng Jiang and Ahmed Alkhateeb
Arizona State University - Emails: {s.jiang, alkhateeb}@asu.edu

***Abstract*—Millimeter-wave (mmWave) and terahertz (THz) communications require beamforming to acquire adequate receive signal-to-noise ratio (SNR). To find the optimal beam, current beam management solutions perform beam training over a large number of beams in pre-defined codebooks. The beam training overhead increases the access latency and can become infeasible for high-mobility applications. To reduce or even eliminate this beam training overhead, we propose to utilize the visual data, captured for example by cameras at the base stations, to guide the beam tracking/refining process. We propose a machine learning (ML) framework, based on an encoder-decoder architecture, that can predict the future beams using the previously obtained visual sensing information. Our proposed approach is evaluated on a large-scale real-world dataset, where it achieves an accuracy of 64.47% (and a normalized receive power of 97.66%) in predicting the future beam. This is achieved while requiring less than 1% of the beam training overhead of a corresponding baseline solution that uses a sequence of previous beams to predict the future one. This high performance and low overhead obtained on the real-world dataset demonstrate the potential of the proposed vision-aided beam tracking approach in real-world applications.**

***Index Terms*—beam tracking, vision, sensing, machine learning, DeepSense 6G, real-world data**

I. Introduction

The millimeter-wave(mmWave) and terahertz (THz) have been considered as key enabler for the high data rate communication in future wireless networks [1]. The high carrier frequencies provide an order of magnitude more bandwidth compared with existing wireless communication systems. The move to higher frequencies, however, brings new challenges such as the higher path-loss. To overcome that and ensure sufficient receive power, mmWave/sub-THz communication systems need to deploy large antenna arrays at the transmitters/receivers and use narrow beams. Nevertheless, obtaining the optimal narrow beams often requires large beam training overhead, which occupies wireless resources and decreases spectral efficiency. This becomes more significant for high-mobility applications such as autonomous vehicles and vehicle-to-everything (V2X) communications [2], which are considered key applications for future wireless communication systems [3]. All that motivates the need to develop novel approaches that can find the optimal beams with low or negligible beam training overhead.

An important observation is that the use of narrow beams at mmWave/sub-THz networks and the reliance on line-of-sight (LoS) links give a special importance to the knowledge of the physical location of the transmitters/receivers and the geometry of the environment around the communication systems. This motivates the use of position/environment sensing devices (such as wireless and position sensors, cameras, etc.) at the communication terminals to guide the different link establishment/resource allocation tasks.

Prior works have studied improving mmWave/THz beam selection, blockage detection, and beam tracking based on sensing information of different modalities [4]–[9]. In [4], the authors propose that the sub-6 GHz channel contains useful information of mmWave channel, therefore, this out-of-band information can be used to establish mmWave link. In [5] the position information of the UE is utilized to guide beam training. [6] proposes to deploy passive radar receivers at the BSs to help establish communication links and reduce beam training overhead.

The vision/camera sensing modality has been also increasingly studied [7]–[9]. [7] employs cameras at the mmWave base stations and leverages the visual data to guide beam selection and detect potential blockages for the current time instance. However, this beam selection and blockage detection may not be adequate for the sensing-aided mmWave communication systems as the latency of capturing and processing the sensory data will unlikely enable current beam selection. To that end, beam tracking, which aims to proactively predict the future beams, is particularly important for the sensing-aided mmWave communication systems. [8], [9] investigated leveraging the camera signals to predict future optimal beams, i.e, vision-aided beam tracking. It is worth mentioning that, in [8], [9], the simulation and evaluation are conducted on a synthetic dataset [13]. However, these result on synthetic data are hard to scale to real-world scenarios considering the complexity and dynamics of wireless communication channels and the impairment of the communication hardware devices.

In this paper, we propose to utilize visual sensing information to enable fast and low-overhead mmWave/THz beam tracking. The main contribution can be summarized as follows.

- We propose a universal problem formulation for the auxiliary data-aided beam tracking, which could be used for different auxiliary data modalities, such as leveraging a sequence of beams, or a sequence of RGB images.
- We propose a machine learning (ML) framework for sensing information aided mmWave/THz beam tracking exploiting an encoder-decoder architecture.
- We evaluate the proposed vision-aided beam tracking on the real-world DeepSense 6G dataset [10], which

978-1-6654-5975- 4/22

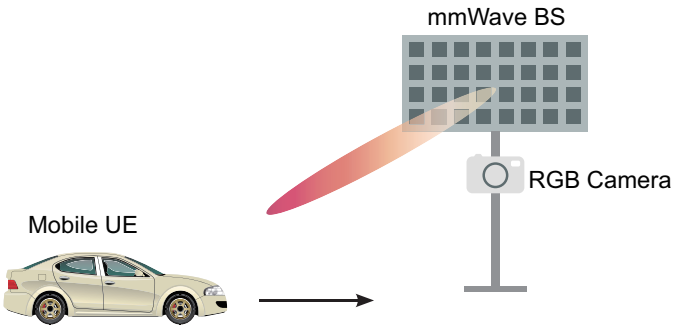

Fig. 1. This figure illustrates the considered system model: the BS senses the environment and the moving UE with an RGB camera. The obtained sensing information is then utilized for the BS beam management.

comprises co-existing visual and wireless beam data. To the best of our knowledge, this is the first time vision-aided beam tracking is investigated on real-world data.

Evaluation results demonstrate the capability of the proposed vision-aided beam tracking approach in achieving high accuracy and receive power. This highlights the potential gains of incorporating visual sensors (cameras) in real-world mmWave/THz communication systems.

II. System and Problem Formulation

In Section II, we first introduce the considered system model for mmWave communications. Then, we formulate beam tracking into an optimization problem. After that, we clearly define the vision-aided beam tracking machine learning task. Lastly, we also present a baseline beam tracking ML task using the previous optimal beam sequence.

A. System Model

Fig. 1 shows the considered system model for mmWave communications, where the base station (BS) is serving a mobile user equipment (UE). The BS is equipped with an antenna array of N elements and an RGB camera (visual data sensor). Using the antenna array, the BS performs beamforming to achieve adequate receive power. We assume that the BS has a pre-defined beamforming codebook $\boldsymbol{\mathcal{F}} = \{\mathbf{f}_1, \dots, \mathbf{f}_{|\boldsymbol{\mathcal{F}}|}\}$ containing $|\boldsymbol{\mathcal{F}}|$ beams $\mathbf{f}_m \in \mathbb{C}^{N\times 1}$. For the sake of simplicity, the UE is assumed to have a single antenna. At time step t, the BS transmits a complex symbol $s[t] \in \mathbb{C}$. We assume the downlink signal $s[t]$ satisfies the power constraint $\mathbb{E}\left[s^H[t]s[t]\right] = P$ with P denoting the transmit power and $(\cdot)^H$ denoting the Hermitian transpose. Then, the corresponding downlink receive signal $y[t]$ can be written as

$$y[t] = \mathbf{h}^H[t]\mathbf{f}[t]s[t] + n[t], \tag{1}$$

where $\mathbf{h}[t] \in \mathbb{C}^{N\times 1}$ denotes the channel between the BS and the UE at time step t. $\mathbf{f}[t] \in \boldsymbol{\mathcal{F}}$ is the transmit beamforming vector used at the BS at time step t. $n[t]$ is the receive noise which satisfies $\mathbb{E}[n[t]n^H[t]] = \sigma_n^2$, and σ_n^2 denotes the receive noise power.

B. Problem Formulation

This paper focuses on the beam tracking problem at the base station, which is defined as follows: Given the available sensing information up to time $t-1$, the BS attempts to determine the optimal beams of $\xi \in \mathbb{Z}^+$ future time steps, that is, optimal beams of $t, \dots, (t+\xi-1)$. First, we define the **optimal beam** at time step t as the one which gives the **highest beamforming gain**. The optimal beam at time step t is then represented by

$$\mathbf{f}^\star[t] = \underset{\mathbf{f}[t]\in\boldsymbol{\mathcal{F}}}{\arg\max} \left|\mathbf{h}^H[t]\mathbf{f}[t]\right|^2. \tag{2}$$

With this pre-defined codebook constraint, the optimal beam $\mathbf{f}^\star[t]$ can be uniquely represented by its beam index in the codebook. The optimal beam index at time step t satisfies

$$p^\star[t] = \underset{p[t]\in[1,2,\dots,|\boldsymbol{\mathcal{F}}|]}{\arg\max} \left|\mathbf{h}^H[t]\mathbf{f}_{p[t]}\right|^2, \tag{3}$$

where $|\boldsymbol{\mathcal{F}}|$ denotes the cardinality of $\boldsymbol{\mathcal{F}}$. Note that, under the codebook constraint, obtaining the optimal beam is equivalent to obtaining the optimal beam index. With the definition of the optimal beam index in (3), we formulate the beam tracking problem as follows:

$$\begin{aligned} \max_{\hat{p}[t]} \quad & \mathbb{P}\left\{\hat{p}[t] = p^\star[t] \mid \mathcal{O}_{t-\xi}\right\} \\ \text{s.t.} \quad & \hat{p}[t] \in [1, 2, \dots, |\boldsymbol{\mathcal{F}}|], \end{aligned} \tag{4}$$

where $\mathbb{P}\{\cdot \mid \cdot\}$ denotes the conditional probability. $\hat{p}[t]$ is the predicted optimal beam index for time step t. $\mathcal{O}_{t-\xi}$ is any auxiliary obtained **before** time step $(t-\xi+1)$ that contains the partial information of the optimal beam at time step t.

C. Vision-aided Beam Tracking

In this paper, We propose exploiting the visual sensing information to achieve accurate beam tracking. The BS uses its RGB camera to capture the visual sensing information of the mobile UE and the surrounding environment. Let $\mathbf{X}[t] \in \mathbb{R}^{H\times W\times 3}$ denote the visual sensing information (RGB image) captured at time step t. H and W are the height and width of the captured image, and the last dimension represents the 3 RGB channels. To utilize this visual sensing information for beam tracking, our objective then becomes obtaining a function which can predict the optimal future beams starting from time step t based on the visual sensing information obtained up to time step $t-1$. Let $\mathcal{X}_{t,i} = \{\mathbf{X}[t-i+1], \dots, \mathbf{X}[t]\}$ denote a sequence of visual sensing information with i representing the number of time steps in the observation window. Then, from (4), the vision-aided beam tracking optimization problem can be written as

$$\begin{aligned} \max_{\hat{p}[t]} \quad & \mathbb{P}\left\{\hat{p}[t] = p^\star[t] \mid \mathcal{X}_{t-\xi,i}\right\} \\ \text{s.t.} \quad & \hat{p}[t] \in [1, 2, \dots, |\boldsymbol{\mathcal{F}}|] \end{aligned} \tag{5}$$

Since the precise joint probability distribution of $p^\star[t]$ and $\mathcal{X}_{t-\xi,i}$ is difficult to model, we propose to leverage the powerful learning capabilities of ML models to solve (5) in an data-driven approach. Let $f(;\theta)$ denote an ML model with

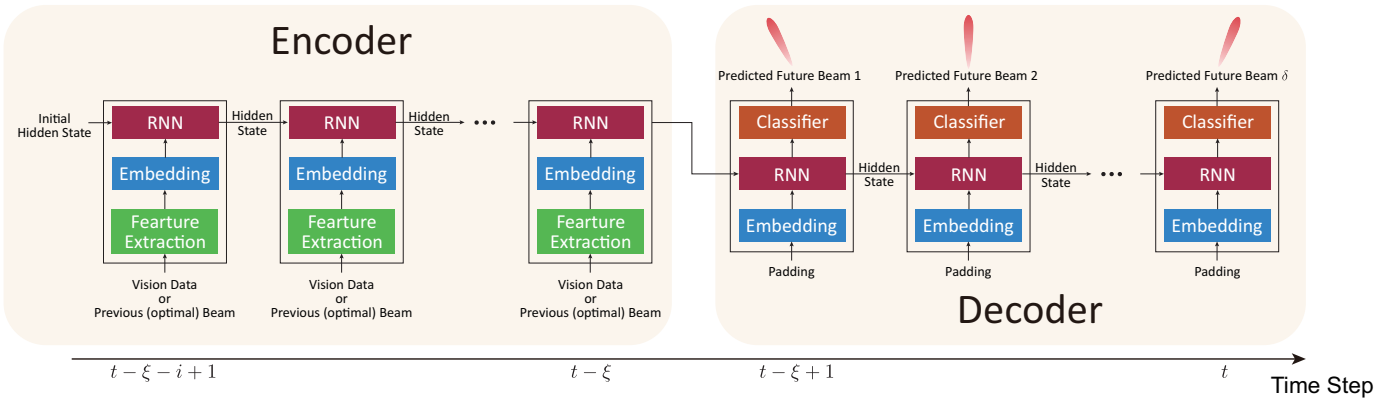

Fig. 2. This figure shows the block diagram of the proposed ML framework for sensing aided beam tracking. The ML framework adopts an encoder-decoder architecture, and incorporates the feature extraction block, the embedding block, the RNN block, and the classifier block.

θ representing its trainable parameters. To solve the vision-aided beam tracking in (5), the ML model aims at predicting the optimal beam index $p^\star[t]$ using the side information $\mathcal{X}_{t-\xi,i}$. Therefore the optimal ML model for vision-aided beam tracking can be mathematically represented by

$$f_v^\star(;\theta_v^\star) = \underset{f_v(;\theta_v)}{\arg\max}\ \mathbb{P}\left\{f_v\left(\mathcal{X}_{t-\xi,i};\theta_v\right) = p^\star[t]\right\}, \quad (6)$$

where $\theta_v^\star$ is the associated optimal parameters of $f^\star$.

D. Baseline Beam Tracking

The sequence of beams resulting from exhaustive search beam training in the previous time steps may also carry information of the mobile UE's future optimal beam. Therefore, it can be exploited as the side information $\mathcal{O}_{t-\xi}$ in (4) to predict the future beam (in the beam tracking problem). In this paper, we employ this approach as a baseline for the beam tracking task. If this approach is implemented using machine learning, then we define this ML task as:

$$f_b^\star(;\theta_b^\star) = \underset{f_b(;\theta_b)}{\arg\max}\ \mathbb{P}\left\{f_b\left(F_{t-\xi,i}^\star;\theta_b\right) = p^\star[t]\right\}, \quad (7)$$

where $F_{t,i}^\star = \{\mathbf{f}^\star[t-i+1],\ldots,\mathbf{f}^\star[t]\}$ denotes the optimal beam sequence from time step $(t-i+1)$ to time step t. $f_b^\star$ and $\theta_b^\star$ are the optimal ML model and trainable parameters associated to this baseline beam tracking task which uses the previous optimal beam sequence as input.

In Section III, we will explain in detail the proposed ML models for the vision-aided beam tracking and the baseline beam tracking tasks.

III. Proposed Solution

Fig. 2 shows the block diagram of the proposed ML framework for the sensing-aided beam tracking task. The ML framework adopts an encoder-decoder architecture featuring four components: the feature extraction block, the embedding block, the recurrent neural network (RNN) block, and the classifier block. The encoder processes the previously obtained information, and passes the information to the decoder. The decoder predicts the future beams based on the information it receives from the encoder.

Feature Extraction Block: The first component of the proposed ML framework is the feature extraction block which directly processes the raw input data. The raw input data often contain extra information which does not contribute to the beam tracking tasks. This irrelevant information can be detrimental since the ML model can overfit on them while neglecting the useful information. Moreover, dropping this unnecessary information also results in a smaller feature space, thus, results in a more stable training process and lower computational overhead. Therefore, the feature extraction block is designed to filter out the unwanted information and extract the features that are informative for the downstream task.

To process the raw RGB data for the vision-aided beam tracking task, we detect the potential transmission target (UE) in the RGB image. We take advantage of the advanced computer vision ML models and employ the YOLOv4 [12] object detector. YOLOv4 is a state-of-the-art convolutional neural network (CNN) based ML model designed to detect thousands of classes of objects from real-world RGB images. The YOLOv4 object detector can achieve real-time prediction and high accuracy which is suitable for the vision-aided beam tracking task. Given an input RGB image, the YOLOv4 object detector predicts a class index $c \in \mathbb{Z}$, a confidence score $s_c \in [0,1]$, and a bounding box vector $\mathbf{b} \in \mathbb{R}^{4\times 1}$. Note that the bounding box vector $\mathbf{b} = [x_c, y_c, w, h]^T$ consists of the x-center, y-center, width, and height of the detected object in the RGB image. For more detailed information related to the YOLOv4 model, we refer the readers to [12]. As discussed in Section I, the optimal beam is highly dependent on the direction/position of the transmission target. Therefore, we exploit the bounding boxes as the extracted feature for the successive blocks in the framework.

The feature extraction block is skipped in the baseline approach since the inputs are the optimal beam indices.

Embedding Block: The embedding layer transforms the input feature into a different vector space. The embedded vector ideally captures some of the semantics of the input vector such that semantically similar input vectors form clusters in the embedding space. For the vision feature (the bounding

boxes) embedding, we employ a fully connected layer which linearly transforms the bounding box vector $\mathbf{b} \in \mathbb{R}^{4\times 1}$ into $\tilde{\mathbf{b}} \in \mathbb{R}^{E_v \times 1}$. To embed the previous beam indices, we apply the same approach as the natural language processing (NLP) ML models embed the word token. For the $|\boldsymbol{\mathcal{F}}|$ beam indices in the codebook $\boldsymbol{\mathcal{F}}$, we employ a look-up table of $|\boldsymbol{\mathcal{F}}|$ trainable embedding vectors $\left\{\tilde{\mathbf{p}}_1, \ldots, \tilde{\mathbf{p}}_{|\boldsymbol{\mathcal{F}}|}\right\}$. In the simulations, we set E_v and E_b to 64.

RNN Block: The third component of the proposed ML model is the RNN block. The RNNs have been extensively studied for processing sequential signals and data such as the NLP and speech recognition tasks. Empirical results have shown that RNNs can effectively capture and process sequential features. Due to the sequential nature of the beam tracking task, we employ the RNN architecture to process the sequential input $\mathcal{O}_{t-i}$ and predict the future beams. We adopt single-layer gated recurrent units (GRUs) with ξ units to process the sequence of visual information or the previous optimal beam index. The hidden state size of the GRU is set to 64.

Classifier Block: A fully connected layer is used as the classifier block. This block predicts the future optimal beam index from the high-level features obtained by the RNN block. The softmax activation function is applied to this fully connected layer to output a confidence score vector $\hat{\mathbf{s}} = [s_1, \ldots, \mathbf{s}_{|\boldsymbol{\mathcal{F}}|}]^T$ of the beam indices in $\boldsymbol{\mathcal{F}}$. The beam index with the highest score is predicted as the optimal future beam

$$\hat{p} = \underset{p\in[1,|\boldsymbol{\mathcal{F}}|]}{\arg\max}\ \mathbf{s}_p. \tag{8}$$

Learning Phase: The encoder of the proposed ML framework processes the previously obtained information. The input sequence to the encoder is $X_{t-\xi,i}$ for the vision-aided beam tracking task or $F_{t-\xi,i}$ for the baseline beam tracking task. Based on the information received from the encoder, the decoder predicts the future optimal beams. A padding sequence of ξ zero vectors Z_ξ is input to the decoder as a placeholder.

Since the ML framework is designed to solve a classification problem, we employ the cross-entropy loss. The loss function can be written as

$$J = \sum_{j=t-\xi+1}^{t} \sum_{m=1}^{|\boldsymbol{\mathcal{F}}|} p^\star_m[j] \log_2\left(\hat{s}_m[j]\right), \tag{9}$$

where $p^\star_m[j]$ is the m-th element of the one-hot coded vector of $\mathbf{p}^\star$ at time step j. $\hat{s}_m[j]$ is the m-th element of the output vector $\hat{\mathbf{s}}[j]$ at time step j.

IV. Experimental Setup

Our proposed vision-aided beam tracking approach and the ML framework are designed to manage real-world mmWave beam tracking. Therefore, we need a high-quality real-world dataset consisting of co-existing RGB images and beam data to evaluate our proposed approach. In this paper, we adopt the DeepSense 6G dataset [10] in our simulation and performance evaluation. The DeepSense 6G is a **multi-modal** dataset comprising **real-world** measurements. The DeepSense 6G dataset incorporates co-existing including wireless beam data, visual sensing data, among other modalities.

Fig. 3. System setup of the DeepSense 6G Scenario 8. The mmWave BS antenna arrays operating at 60GHz receives the signal transmitted by the moving UE. A Camera is placed under the BS to obtain sensing information.

A. DeepSense 6G Scenario 8

We adopt Scenario 8 of the DeepSense 6G dataset for our simulation. The system setup of Scenario 8 is shown by Fig. 3. In Scenario 8, a fixed BS receives signals from a moving UE. The BS is equipped with a uniform linear array (ULA) of 16 elements and an RGB camera installed below the ULA. The BS adopted a beamforming codebook consisting of 64 predefined beams. The codebook is horizontal-only and 4-time oversampled. The UE is a moving vehicle equipped with a 60 GHz quasi-omni transmitter that is always oriented towards the BS. During the data collection process, the UE passes by the BS multiple times. At each time step, the BS measures the receive power of all beams in the codebook by beam sweeping, and captures the UE with the RGB camera. The RGB data and the beam receive power data are synchronized by downsampling. After synchronization, the time interval between each time step is 128 milliseconds. Note that, in our experiment, we only consider the vehicle as UE. Extension to other types of UE and multi-UE scenarios is an interesting future research direction.

B. Develop Dataset Generation

We evaluated our proposed vision-aided beam tracking approach on the official development dataset split of DeepSense 6G Scenario 8. The training dataset and the validation dataset consist of 70% and 20% of the raw dataset. The DeepSense 6G Scenario 8 dataset consists of multiple data sequences. In a data sequence, the vehicle completes its path passing by the BS for one time. Each data sequence consists of co-existing RGB images and beam receive powers of multiple time steps. We follow the official development dataset split of DeepSense 6G Scenario 8. The training dataset and the validation dataset consist of 70% and 20% of the raw dataset data sequences. Note that the data in two datasets come from different vehicle passes to assure there is no data leakage. For each data sequence, we break it into data samples using a sliding window size of 13. One data sample consists of 13 time steps and can be written as $\{(\mathbf{X}[1], p^\star[1]), \ldots, (\mathbf{X}[13], p^\star[13])\}$.

In the training process, we use an observation window size $i = 8$, and we train the models to predict the five future beams ($\xi \in [1, 5]$). Therefore, the input to the encoder is $\{\mathbf{X}[1], \ldots, \mathbf{X}[8]\}$ for the vision-aided beam tracking model, and $\{p^\star[1], \ldots, p^\star[8]\}$ for the baseline beam tracking model.

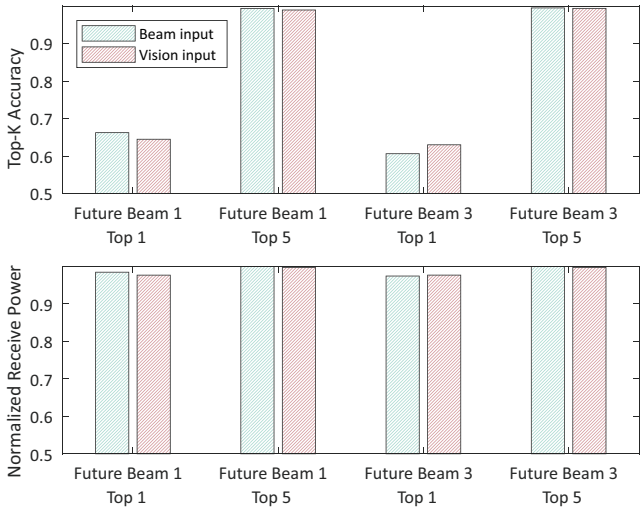

Fig. 4. This figure compares the accuracy and normalized receive power of the top-k predictions of future beam 1 and 3. For both the vision-aided and baseline beam tracking, the accuracy increases as k increases. However, the normalized receive power saturates to the optimum even when $k = 1$.

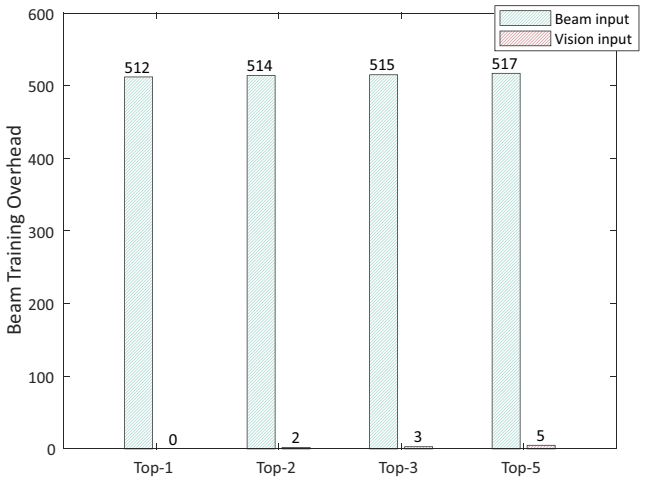

Fig. 5. This figure compares the training overhead required in nine time steps of the vision-aided and the baseline approaches. The vision-aided approach only consumes 1% beam training overhead compared to the baseline.

In both two beam tracking approaches, the decoder is expected to output $\{\hat{\mathbf{s}}[9], \ldots, \hat{\mathbf{s}}[13]\}$.

V. Evaluation Results

In this section, we evaluate the proposed vision-aided beam tracking approach and compare its performance with the baseline beam tracking's performance. The metrics adopted in the evaluation are the following.

- Top-k accuracy: the percentage of the time steps of all validation samples where the beam corresponding to the top-k confidence scores include the optimal beam.
- Normalized receive power: the ratio between the highest receive power achieved by the top-k predicted beams and the receive power of the optimal beam. This metric is averaged over all time steps and all validation samples.
- Beam training overhead: the number of beam power measurements including the beam power measurements required in the observation window and the beam power measurements that will be conducted to sweep top-k beams in the future time step.

A. Do the ML Models Learn to Predict Future Beams?

In Fig. 4, we present the top-k accuracy and normalized receive power of the future beam 1 and future beam 3. The top-k accuracy improves significantly as k increases for both the vision-aided and baseline beam tracking approaches. It can also be observed that the accuracy decreases when predicting beams in the further future. The top-5 accuracy of the vision-aided beam tracking for future beam 1 is 99.37%. This accuracy implies that the proposed vision beam tracking approach can find the optimal future beam 1 with 99.37% probability by testing 5 beams in the beam sweeping process. However, in terms of receive power, testing more beams (increasing k) may not be worth the price of the training overhead. For the vision-aided beam tracking, $\mathbf{97.66\%}$ **receive power can already be obtained for future beam 1 even without any beam training** ($k = 1$), leaving little room for testing 5 beams for improvement. Despite the relatively low top-1 accuracy obtained by both beam tracking approaches, the near-optimal receive power highlights that the **ML models effectively learn to predict future beams**, and most of the mistakes occur at the sub-optimal beam with near-optimal receive power. Overall, **the vision-aided beam tracking approach can achieve comparable performance to the baseline approach in terms of the two metrics.** Note that the baseline beam tracking approach is a strong baseline since it inputs the optimal beams of the 8 previous time steps. This highlights the capability of the proposed vision-aided approach in accurate beam tracking. It is worth mentioning that the adopted DeepSense 6G scenario 8 mainly consists of LoS data. However, the baseline beam tracking approach could be more sensitive to NLoS scenarios since the previous optimal beam may not contain enough information on the blockages, reflectors, and scatterers. The baseline beam tracking is also expected to further degrade when the surrounding environment becomes more dynamic. On the contrary, The visual data obtained by the camera captures rich information on the surrounding object and the dynamics. Moreover, the baseline beam tracking approach requires information on the optimal previous beams, which may not always be applicable in practice. The baseline model can instead rely on the beams it previously predicted. This may cause the baseline approach to deviate more from the optimal beam as the beam tracking goes on without calibration. The vision-aided approach, however, keeps capturing the latest information on the environment. Therefore, it is not likely to suffer from this deviation.

B. Vision-aided vs. Baseline: Beam Training Overhead

In Section V-A, we analyzed the accuracy and received power performance of the two beam tracking approaches. However, in our evaluation, the baseline approach requires

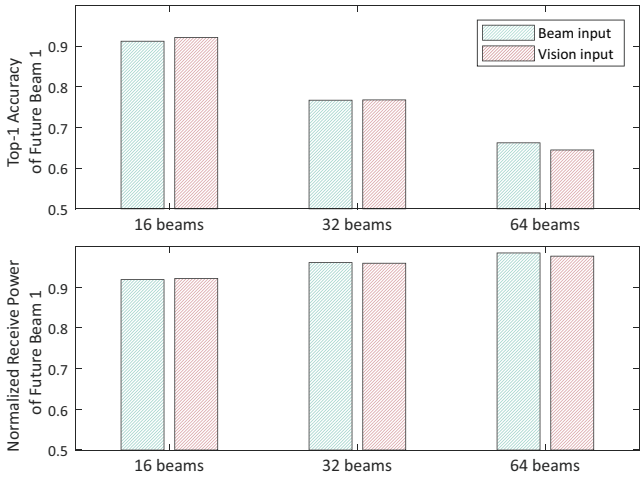

Fig. 6. This figure shows the accuracy and receive power performance of the vision-aided and baseline beam tracking approaches with different codebook sizes of 16, 32, and 64.

knowing the previous eight optimal beams to predict the optimal beam of the future time step and match the performance of the vision-aided approach. Therefore, Fig. 5 studies the beam training overhead required by the vision-aided and the baseline approaches in these nine time steps when top-k beams are predicted. It is assumed that both approaches will conduct beam sweeping over the predicted top-k beams at the future time step when $k \geq 2$. **For the 5 cases shown in Fig. 5, the beam training overhead required by the vision-aided approach is less than 1% of the baseline approach.** Furthermore, when top-1 beam is predicted the vision-aided approach completely eliminates the beam training overhead.

C. What is the effect of the Beamforming Codebook Size

In this section, we study the effect of the codebook size on the performance of the vision-aided and baseline beam tracking approaches. Fig. 6 shows the top-1 accuracy and normalized receive power of the future beam 1. It can be seen that, as codebook size increases, the top-k accuracies of both beam tracking approaches decrease as can be expected. However, the normalized receive power increases when a larger beam codebook size is adopted. Using the vision-aided beam tracking approach with 16 pre-defined beams, the normalized receive power of the top-1 prediction is 92.17% for future beam 1. Exploiting the codebook with 64 beams, a normalized receive power of 97.66% can be achieved, which is a 6% relative improvement. This demonstrates that **reasonable receive power improvement can be achieved by using oversampling codebook** at a price of slightly more computational complexity.

VI. Conclusion

This paper proposes a machine learning based vision-aided beam tracking framework. Exploiting this framework, we also develop an efficient baseline beam tracking approach that utilizes the previous optimal beams. The proposed approaches are evaluated using a large-scale real-world dataset comprising co-existing visual and wireless mmWave data. Evaluation results demonstrate that the proposed vision-aided beam tracking approach can learn to accurately predict future beams and achieve comparable performance to the baseline solution. It achieves a top-1 accuracy of 64.47% and a top-5 accuracy of 98.95% in predicting the future beam. The robustness of the proposed vision-aided beam tracking is illustrated by the 97.66% normalized receive power of the top-1 prediction. Moreover, the proposed vision-aided beam tracking only requires 1% of the beam tracking overhead of the baseline approach. These results highlight the potential of leveraging visual sensors in improving real-world mmWave communications.

VII. Acknowledgment

This work is supported in part by the National Science Foundation under Grant No. 2048021.

References

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

[2] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina, "V2X access technologies: Regulation, research, and remaining challenges," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 1858–1877, 2018.

[3] J. He, K. Yang, and H.-H. Chen, "6G cellular networks and connected autonomous vehicles," *IEEE Network*, vol. 35, no. 4, pp. 255–261, 2021.

[4] A. Ali, N. González-Prelcic and R. W. Heath, "Millimeter Wave Beam-Selection Using Out-of-Band Spatial Information," in *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1038-1052, Feb. 2018, doi: 10.1109/TWC.2017.2773532.

[5] T.-H. Chou, N. Michelusi, D. J. Love, and J. V. Krogmeier, "Fast position-aided MIMO beam training via noisy tensor completion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 3, pp. 774–788, 2021.

[6] A. Ali, N. González-Prelcic and A. Ghosh, "Passive Radar at the Roadside Unit to Configure Millimeter Wave Vehicle-to-Infrastructure Links," in *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14903-14917, Dec. 2020, doi: 10.1109/TVT.2020.3027636.

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

[8] Y. Tian and C. Wang, "Vision-Aided Beam Tracking: Explore the Proper Use of Camera Images with Deep Learning," *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pp. 01-05, 2021.

[9] H. Zhifeng, and C. Han. "Image and index fused sequence-to-sequence algorithm for vision-aided millimeter-wave beam tracking," *Proceedings of the 5th ACM Workshop on Millimeter-Wave and Terahertz Networks and Sensing Systems*, 2021.

[10] A. Alkhateeb, G. Charan, M. Alrabeiah, T. Osman, A. Hredzak, N. Srinivas, and M. Seth, "DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset," *available on arXiv*, 2021. [Online]. Available: https://www.DeepSense6G.net

[11] 3GPP TR 38.802 version 14.1.0 Release 14, "Study on New Radio Access Technology—Physical Layer Aspects," Tech. Rep., Jun 2017. [Online]. Available: http://ftp.3gpp.org

[12] B. Alexey, W. Chien-Yao, and L. Hong-Yuan Mark, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[13] M. Alrabeiah, A. Hredzak, Z. Liu and A. Alkhateeb, "ViWi: A Deep Learning Dataset Framework for Vision-Aided Wireless Communications," *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pp. 1-5, 2020, doi: 10.1109/VTC2020-Spring48590.2020.9128579.