Transformer-Based Multi-Modal Deep Learning for Sensing-Assisted Beam Prediction for Beyond 5G

Qiyang Zhao, Yu Tian, Zine el abidine Kherroubi, Fouzi Boukhalfa*
Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, United Arab Emirates
Email: {qiyang.zhao, yu.tian, zine.kherroubi, fouzi.boukhalfa}@tii.ae

***Abstract*—Wireless communications at high frequency bands with large antenna arrays face challenges in beam management, which can potentially be improved by multi-modality sensing information from camera, LiDAR, radar and GPS. In this article, we propose a transformer-based multi-modal deep learning framework for sensing assisted beam prediction. We first preprocess sensory data by enhancing and segmenting images, filtering point-clouds, transforming radar signal and encoding user's GPS location. We then employ ResNet CNN model to extract the features from image, point-cloud and radar raw data. The GPT transformer model is used after each convolutional block to learn the correlation of different modalities and produce fused latent features to the next level abstraction. We train the model on augmented sequential multi-modal sensory data with softened top beam index for a given car user, by utilizing focal loss, cosine decay and exponential moving average to make the model generalized and robust on imbalanced data. Experimental results on sequential multi-modal sensor data shows that our solution produces distance-based accuracy score on measured scenario up to 0.82, and unsceen scenario up to 0.53, yielding overall score at 0.67, with different sampling rate than the training dataset. It performs significantly better than using any single modality or smaller model. Furthermore, our solution is flexible to extend on other applications of sensing and communications.**

***Index Terms*—Multimodal Deep Learning, Transformer, Sensing assisted Communications, Beam Prediction**

I. Introduction

Communication beyond 5G and 6G is exploiting high frequency bands such as mmWave and THz, in order to boost the system capacity by utilizing large available bandwidth. Massive antenna arrays has been leveraged to create ultra-narrow beams, so as to increase the received signal power and reduce interference for targeted users. Significant challenges in beam management arises in such system and scenario to find the best beams for users in high mobility under large propagation lost and fast changing channels, with the goal to provide ultra-high reliable and low latency services, such as autonomous driving, mixed-reality, digital twin.

Multi-modality sensory information has the potential to improve wireless communications in challenging environments, as part of the integrated sensing and communication (ISAC) topic being actively studied for 6G. In the vehicular network scenario, road side base station unit (BS) equipped with camera, LiDAR, radar and GPS can produce sequential image, point-cloud, signal and location information of the road environment, objects, and vehicle users (UE). Such sensory data is potentially useful in assisting the RSU to analyse the radio transmission scenario, so as to produce effective beam management.

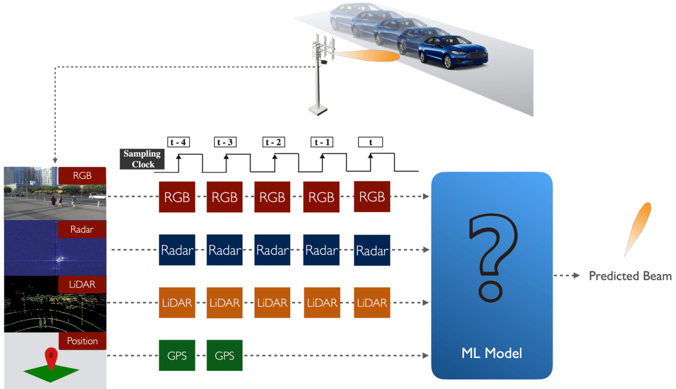


Fig. 1: Schematic representation of the input data sequence utilized in this challenge tasks [3]

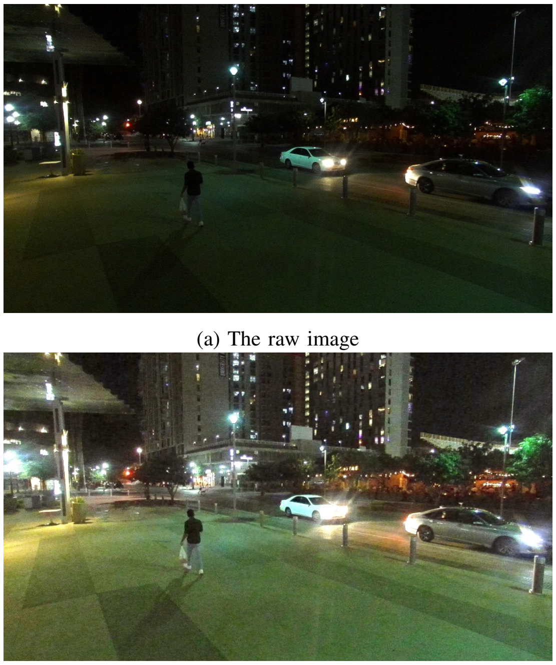
In this article, we present a transformer based multi-modal deep learning approach for sensing assisted beam prediction, which is a solution to the DeepSense6G problem statement in the ITU AI/ML for 5G challenge 2022 [1]. The challenge provides large multi-modal sensing and communicaiton datasets of camera image, LiDAR point-cloud, radar signal sampled at the BS, plus the GPS coordinates of BS and UE. As shown in Fig. 1, each data sample contains a sequence of 5 image, point-cloud, radar signal instances, plus the first 2 instances of UE position [2]. The dataset is measured in 4 scenarios (31, 32, 33, 34). A development dataset is provided with thousands of samples collected in scenario 32, 33, 34; and an adaptation dataset is provided with tens of samples collected in scenario 31, 32, 33. Both datasets have the groundtruth best beam of the UE associated with each sample. A test dataset with hunderds of samples is provided in all scenarios without labels. Specifically, most labeled data resides in scenario 32, 32, 34, and half of the test data resides in scenario 31. Moreover, the sampling rate of the sequences in the test dataset is same as the adaptation dataset, however, double of the development dataset. The objective is to evaluate how the developed model can be generalized to unseen scenario, in which the sensing data is collected in different location, field of view (FoV), time (morning, afternoon, evening), and sampling rate.

Our solution is developed upon a transformer-based multi-modal deep learning framework customized for sensing as-

*The authors contributed equally to this work.

sisted beam prediction. There exists several solutions for multi-modal sensor data fusion, such as the TransFuser framework proposed for autonomous driving [4]. However, the work is developed for computer vision applications such as semantic segmentation, object detection, recognition, localization. The data is collected from sensors equipped on the moving vehicles. In comparison, our task has several unique challenges where the TransFuser model is difficult to solve. Firstly, our sensors equipped on the BS produce much wider FoV than those on vehicles. There are many static and moving objects in the scene, but there is no labels or bounding box indicating the UE. Secondly, we have also radar signal and GPS location, and how to utilize these modalities to assist our task is unclear. Thirdly, beam prediction is a unique application in wireless communications which has not been well exploited with multi-modal sensors. In particular, the correlated abstractions between radio transmission scenarios and visionary data is not straingtforward, making deep learning hard to generalize on unseen scenarios [5].

The rest of this article is structured as follows. Section II describes our proposed methods on preprocessing and augmentation of multi-modality sensor data. Senction III describes our developed approach for transformer based multi-modal sensing assisted beam prediction. Section IV presents experiments of our solution on the multi-modal beam prediction challenge dataset. Finally, the work is concluded in section V by discussing some ongoing research aspects and extension to wider applications.


(a) The raw image

(b) The enhanced image

Fig. 2: ‘Image_281.jpg’ in scenario 33

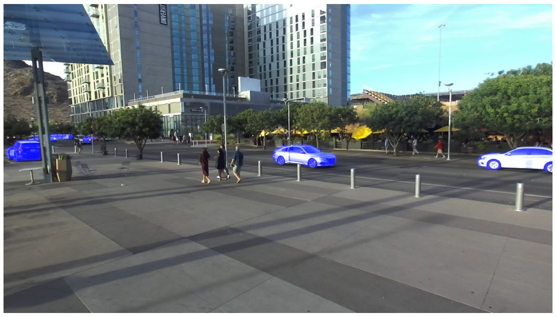


Fig. 3: ‘Image_116.jpg’ and its vehicle segmentation (blue) in scenario 32

II. Multi-Modal Sensing Data Analysis

A. Data Preprocessing

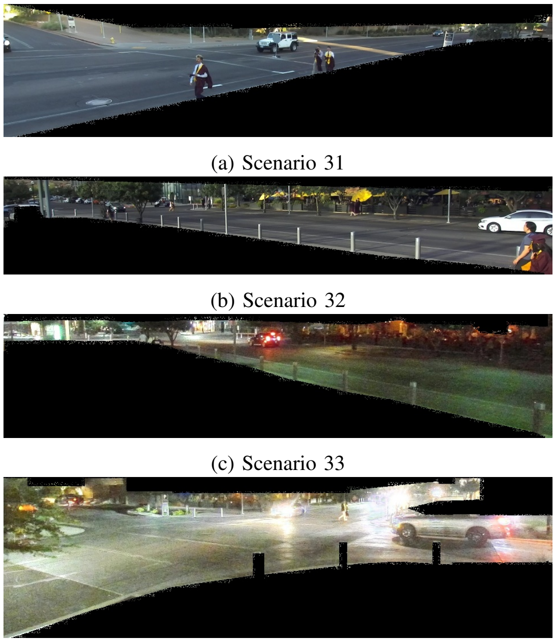
We developed several data preprocessing units on the multi-modal sensor data for training the sensing assisted beam prediction model.

1) Camera Data: The beam prediction task in this challenge is related to the object detection and tracing tasks in computer vision where Camera data plays a vital role. Meanwhile, from the LIDAR and radar data, we can’t distinguish the UE and other vehicles. Therefore, we tried many ways to preprocess the camera images data to enhance the visual features of the vehicles for recognizing and distinguishing, including enhancing the lightness, semantic segmentation, and masking the background.

- *Enhancing the lightness*: To overcome the darkness issue in the night scenarios 33 and 34, we utilize MIRNet [6] to enhance the lightness of these images. The appearances of vehicles become clear after enhancing shown in Fig. 2b compared with the raw image in Fig. 2a.
- *Semantic segmentation*: To highlight the vehicles in the camera data, we use the PIDNet [7] to segment out the vehicles from the images in the daytime scenarios 31 and 32 shown in Fig. 3. We also tried to segment the image in the nighttime scenarios 33 and 34. However the performance is very poor even with the enhanced lightness. Therefore, we abandon the vehicle segmentation in these two nighttime scenarios. After getting the vehicles’ pixels, we can enhance their lightness and darken the background to make vehicles dominant.
- *Masking the background*: We also tried to mask the useless background with the black color and just keep the streets in the images. Because the camera is stable, all the images in the same scenario share the same background. We just need to guide the Neural Network to focus on the objects on the street. Meanwhile, observing that the beam indices are distributed in the horizontal dimension, we crop the useless masked parts with all black cross the whole horizontal dimension. Finally we get the masked images shown in Fig. 4.

2) LiDAR Data: Each LiDAR’s point cloud frame has, on average, more than *16000* 3D points. In order to reduce the size of the input data and speed up the training of the model, we preprocessed LiDAR data by three ways:

- *Filtering background objects:* we removed data points that



(a) Scenario 31

(b) Scenario 32

(c) Scenario 33

(d) Scenario 34

Fig. 4: Masked camera images

corresponds to static buildings and infrastructure, since these regions are always out of the line-of-sight *(LOS)* between the BS and the end user *(UE)*, and do not affect the beam prediction. These points cloud add complexity and bias when training our model. To implement this, we used a kind of moving average points cloud for each scenario across the entire dataset. We substract the background point cloud from each point cloud frame to filter out the static building and objects, and extract just the desired region for the beam prediction.

- *Birds-Eye-View(BEV) Conversion:* This technique converts raw point cloud data into an image-like representation. The height, intensity, and density of the 3D point cloud are mapped to the Red, Green and Blue channels of a color image to generate the *BEV* image. First, the point-clouds within the region of interest *(ROI)* are discretized into a grid cells. Then, the height is encoded by considering the max height of the points in each grid cell. Also, the intensity is encoded by considering the maximum intensity in the grid cell. Finally, the density of the points is calculated for each grid cell [8]. The representation of a bird's eye view *(BEV)* LiDAR point-cloud has certain advantages. Firstly, this representation can be conveniently combined with the technologies of image-based deep convolutional neural networks *(CNN)* [9], which will be used in our model for beam prediction. Moreover, this image-like representation has the benefit of preserving the basic structure of the point clouds and the depth information, while reducing the computational effort needed to process the data [8].
- *Custom LIDAR FoV:* We crop the BEV projection of the LIDAR data to keep its FoV consistent with the view in the images. Because the LIDAR data contains so many useless contents, this processing will help the CNN focus on these useful stuffs and enhance the fusion of the LIDAR and camera data.

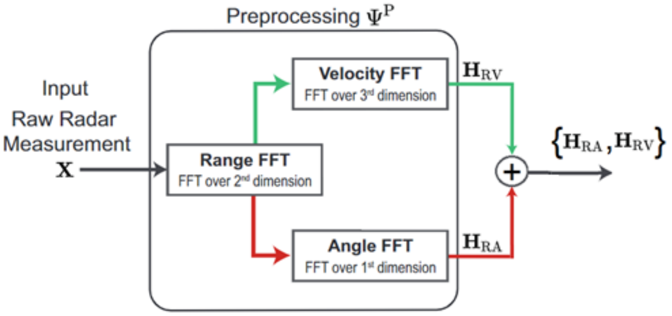

Preprocessing Ψ^{P}
Input
Raw Radar
Measurement
$\mathbf{X}$
Velocity FFT
FFT over 3rd dimension
Range FFT
FFT over 2nd dimension
Angle FFT
FFT over 1st dimension
$\mathbf{H}_{\mathrm{RV}}$
$\mathbf{H}_{\mathrm{RA}}$
$\{\mathbf{H}_{\mathrm{RA}}, \mathbf{H}_{\mathrm{RV}}\}$

Fig. 5: Combining Range-Angle $\mathbf{H}_{RA}$ and Range-Velocity $\mathbf{H}_{RV}$ maps [10].

3) Radar Data: We adopted the preprocessing techniques used in [10]. The objective is to extract the range, the angles, and the velocity of the moving objects in the environment using *2D* Fourier transform, as described in [10]. Since the Camera and LiDAR do not provide explicitly velocity measurements, we propose to concatenate the *Range-Angle Maps* with the *Range-Velocity Maps* of the Radar to preserve the information about the speed of the moving cars as illustrated in Figure 5. Indeed, Radar provides reliable speed measurement regardless of weather conditions and lightness level [11].

4) GPS Data: GPS data plays an important role in locating the UE's position. However, it is not always available or accurate caused by the connection and delay issues. In this challenge, only data of first two out of the five instances are provided to mimic the a real-world scenario. As to the processing, we first translate the GPS locations of the UE and the BS from the GPS coordinate to the Cartesian coordinate. Then we calculate the relative position, denoted as $(\Delta x_n, \Delta y_n)$, between the UE and the BS of the nth GPS data. There are two next processing ways.

- *Min-max normalization:* According to [12], min-max normalization can achieve better performance than other approaches, such as using the relative coordinates to the BS, with divide-by-max normalizations that would not distort the scale of the data, and with the usage of polar coordinates, i.e., distance and angle. Thus we adopt this normalization as the first way. The detailed expressions of the new relative position, denoted as $(\Delta \hat{x}_n, \Delta \hat{y}_n)$, after min-max normalization are show as Eqs. (1) and (2).

$$\Delta \hat{x}_n = \frac{\Delta x_n - \min\left(\{\Delta x_n\}\right)}{\max\left(\{\Delta x_n\}\right) - \min\left(\{\Delta x_n\}\right)}, \qquad (1)$$

and

$$\Delta \hat{y}_n = \frac{\Delta y_n - \min\left(\{\Delta y_n\}\right)}{\max\left(\{\Delta y_n\}\right) - \min\left(\{\Delta y_n\}\right)}. \qquad (2)$$

where $\{\Delta x_n\}$ and $\{\Delta y_n\}$ are sets of the all the Δx_n and Δy_n in the training dataset.
- *Calibrated normalization:* Because beam indices depend on the relative position between the UE and BS, the

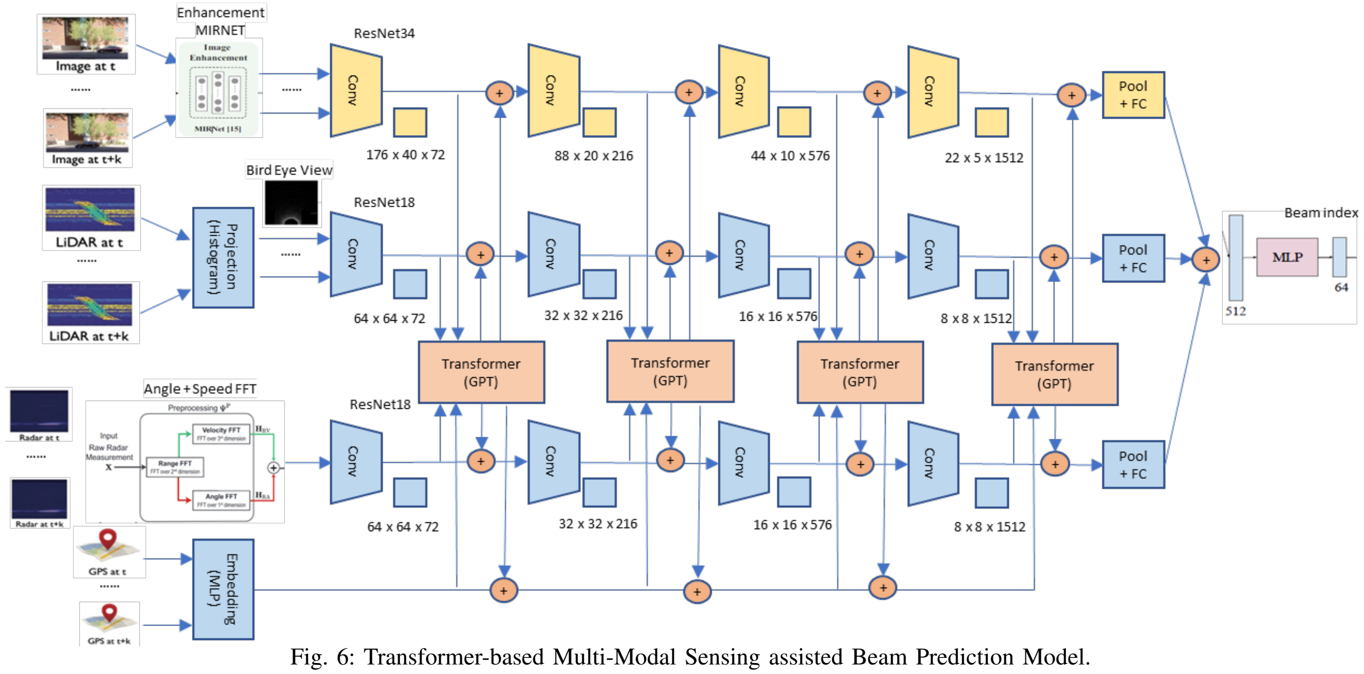


Fig. 6: Transformer-based Multi-Modal Sensing assisted Beam Prediction Model.

angles between the UE-BS connection and the x axis in the Cartesian coordinate play an important role in locating the UE in the image. However, the angles w.r.t the same beam index are different in different scenarios. Therefore, we calibrate the angle of the central pixel in the images of all scenarios as 0°. To achieve this goal, we first manually select the images where the UE is located in the middle of the images from the four scenarios and calculate their angles θ_1, θ_2, θ_3, and θ_4 according to their relative positions. Then we subtract all the angles in the $3i$th scenario with θ_i $(i = 1, 2, 3, 4)$. Afterwards, we transform the new calibrated angles to the new $(\Delta x'_n, \Delta y'_n)$ with $(\Delta x'_n)^2 + (\Delta y'_n)^2 = 1$. Finally, we obtain the calibrated and normalized GPS data.

B. Data Augmentation

Due to data imbalance between scenario *31* and remaining scenarios, we use some data augmentation techniques to increase the dataset size for this scenario.

1) Camera Data: Beam selection relies mainly on the transmitter/receiver locations and the geometry/characteristics of the surrounding environment [10]. In order to conserve these geometric information, we use only some *photometric* transformations that are *'safe'* for the beam prediction application [13]. We augment each image by randomly changing the brightness, the contrast, gamma correction, hue channel, color saturation, the sharpness, and performing *Gaussian* blurring on the image.

2) LiDAR Data: Similar to the Camera data, we perform two *'safe'* data augmentation techniques for each point cloud frame without deteriorating the geometric information of the environment: Randomly down-sampling the point cloud by a factor of *10 %*, and adding small and random 3D position deviation for each point. These transformations conserve the position and general shape of the objects in the environment (cars, buildings, pedestrians ... etc).

3) Radar Data: In order to augment the Radar Data, we add a small and random noise to each normalized *FFT* coefficient. The added noise is limited to *10%* of each *FFT* component amplitude in order to conserve the shape of the spectrum. Hence, this transformation is *'safe'* in the spectral domain.

III. Transformer-based Multi-Modal Sensing assisted Beam Prediction

In this section, we introduce our solution of a transformer based deep learning model for multi-modality sensing and algorithms for beam prediction.

A. Transformer-based Multi-Modal Sensing Model

With the preprocessing unit that transforms different modalities of sensory data into 2D vector space, we can leverage Convolutional Neural Network (CNN) as encoders to extract their high order features and learn correlations between them. Inspired by the TransFuser [4] framework built on autonomous driving, we develop a Generative Pre-trained Transformer (GPT) based deep learning model that fuse a sequence of camera image, LiDAR point-cloud, radar signal and GPS locations measured at RSU, to predict the top mmWave beams for communication with the UE.

The model architecture is illustrated in Fig. 6. We first employ deep residential network (ResNet) [14] to encode image, point-cloud and radar signal on abstraction space. Specifically, the ResNet34 is used on each of the 5 instances of the RGB image, after normalized and resized to a 512 feature vector. The ResNet18 is used to encode the LiDAR BEV, and radar range angle-velocity map. An multi-layer preception (MLP) layer is used to encode the first 2 instances of UE location.

Each ResNet block of convolution, batch normalization, activation and pooling produces an abstracted feature map. We use GPT based transformer modules after each convolution block to fuse the intermediate abstractions between the modalities of image, point-cloud, radar map. The transformer uses linear projections for computing a set of queries, keys and value. Scaled dot products are used between queries and keys to compute the attention weights and then aggregates the values for each query. Finally a non-linear transformation is used to calculate the output features. It applies the attention mechanism multiple times throughout the structure, resulting attention layers with multiple heads to generate several queries, keys and values. Since each convolutional block encode different aspects of the scene at different layers, thus several transformer blocks are used to fuse these features at multiple scales throughout the encoder.

The transformer learns the correlation between modalities. In particular, the fusion image and point-cloud can better represent the scene, especially in some dark and night scenarios. Furthermore, the radar velocity and angle map can position the mobility objects in the scene. Since we have only the first 2 instances of user location, it is passed through the attention layer with maps with each of the 5 instances of other modalities. In this manner, the transformer could estimate the position of the user in the scene at the 5^{th} instance.

The fused feature maps of different modalities are propagated to the next convolutional blocks are repeated several times with transformer blocks, then finally concatenated and passed through MLP layers to produce weights of 64 beam index using the softmax function.

B. Training and Optimization for Beam Prediction

We develop a number of training and optimization mechanisms to customize the model to the beam prediction task. Firstly, we transform the one-hot beam indexes to Gaussian distribution, by positioning the peak at the best beam and cutoff to 0 at its neighbouring 5 beams. This is to adapt the cross-entropy loss function to the distance based accuracy (DBA) score as defined in [1], where higher weights are given if the beams are closer to the best beam.

We further apply a focal loss [15] method to improve training on a sparse set of hard examples. Data imbalance is a significant challenge in this task. The data samples from scenario 31 is much less than others. Moreover, some beams have much less probability to be served as best beam than others. The adaptation dataset is with different sampling rate than the development dataset. To differentiate between easy and hard examples, a modulating factor $(1 - p_t)^\gamma$ is added to the cross entropy loss, with tunable focusing parameter $\gamma \geq 0$. Intuitively, it reduces the loss contribution from easy examples and extends the range of example receiving low loss.

We also employ several training methods to stabilize the convergence and make the model robust. We use cosine decay scheduler to start training with high learning rate (LR), rapidly decreased to minimum value, then do warm restart by resetting the LR to a smaller value [16]. This stabilize the weights as with more epoches. Furthermore, we maintain exponential moving average (EMA) of the parameters during training, instead of utilizing the final trained values. This eliminates the flucturation at final steps and make the model robust.

IV. Performance Evaluation and Discussion

We performed experiments to train and evaluate our proposed models and approaches over the challenge dataset [2]. We first combine the development and adaptation dataset, then randomly split into 90% for training and 10% for validation. The learning rate is set to start from 10^{-4}. We validate and compare the performance using different proposed data analysis and model training approaches with the scores evaluated on the test dataset by the organizer.

The DBA scores on the test dataset of different developed schemes are shown in Table I. The experiments are performed by incrementally adding different enhanced schemes proposed in this work, on top of the base schemes, to evaluate their effectiveness in this task. The results are presented as the DBA scores of predicted beams in each scenario and the overall average, as defined by Eqs. (5) and (6) in [1].

In the baseline scheme **A**, we use only the 5^{th} timestamp data, adding image enhancement and using range-angle map of radar signal after FFT. We can observe that the scheme performs very well in scenarios 32, 33 and 34. However, it produces very low score at 0.12 in scenario 31 which is lack of training data, making the overall score low. In scheme **B**, we fine tune the model on scenario 31 data, which we can see the score is significantly improved to 0.47. This makes the overall score improved from 0.46 to 0.59, though there is slight reduction in other scenarios.

In the second set of experiments, we apply the sensor data from all the 5 timmestapms (including 2 timestamps of GPS). In scheme **C**, we add the range-velocity map in addition to angle from radar signals. We also applied focal loss, cosine decay LR and soft beam index over scheme **A**. The results show that we have obtained slightly higher score than scheme **B**, but without further fine tuning on scenario 31. This indicates that the generalization of our model is improved, especially we can see higher scores in other scenarios.

We further conducted several improvements to the solution. We utilized GPS calibrated angle normalization, and data augmentation in all modalities as introduced previously. It can be observed in scheme **D** that the scores in scenario 31 and 32 are slightly improved. This is mainly contributed by more accurate mapping the relative coordinates of UE directly to the positions in the images. Furthermore, we apply EMA over the training process to obtain the best model parameters. It shows in scheme **E** that the scores in scenario 33 and 34 are significantly improved, making the over score improved from 0.60 to 0.63. This is due to the EMA has the capability to make the model more robust, especially in the night scenarios which is difficult to recognize the UE. Finally, we obtained the best score of this challenge from scheme **F**, which we align the FoV of LiDAR with the camera. It significantly improve the scores in almost all scenarios, with the overall score reaching 0.6671.

TABLE I: DBA score on test dataset of developed schemes

Test	Base	Enhance	Overall	31	32	33	34
A		Timestamp 5 Image enhance Radar angle	0.4618	0.1147	0.6864	0.7848	0.8188
B	A	Fine tune 31	0.5891	0.4718	0.6222	0.6933	0.7328
C	A	Timestamp 1 to 5 Radar velocity Focal loss Cosine decay LR Soft beam index	0.5989	0.4509	0.6852	0.7538	0.7369
D	C	GPS angle norm Data augment	0.5997	0.4713	0.7000	0.7424	0.6997
E	D	EMA	0.6325	0.4760	0.7123	0.7819	0.7985
F	D	LiDAR FoV	**0.6671**	**0.5331**	**0.7173**	**0.7910**	**0.8209**

TABLE II: DBA score on test dataset of experimental preprocessing

Test	Base	Enhance	Overall	31	32	33	34
G	F	LiDAR filter	0.6398	0.4856	0.7000	0.7914	0.8061
H	G	EMA	0.6458	0.5347	0.6951	0.7505	0.7679
I*	F	Image segment Image mask	0.6298	0.4709	0.7284	0.7810	0.7684
J*	I	EMA	0.6433	0.4947	0.7506	0.7890	0.7837

* No image enhancement in scenario 33 and 34.

We can conclude from here that positioning the vision views on the most useful context is very benefital to the transformer model for beam prediction.

In Table II, we present experimental results by exploiting further preprocessing on LiDAR and camera data from the best scheme **F**. We first apply background filtering to the point-clouds by averaging the points on the data samples. We can observe that the scores in scenario 31 and 34 are reduced. We can also see similar results by applying enhancements to images. In scheme **I** we perform segmentation on the moving objects, add mask to the background scenarios, and remove enhancement in the night scenarios. We can see that the scores decrease further than scheme **G**. This indicates that the background scene provide essential information of the relative moving position of the UE, plus potential obstacles and reflectors over the radio transmission environment. Furthermore, we apply EMA in training on both preprocessing schemes, which we can see that scores are increased especially in scenario 31. This validates the robustness of EMA in training small data.

V. Conclusions and Future Works

In this article, we presents a solution to sensing assisted beam prediction for beyond 5G communication system, based on transformer empowered multi-modal deep leaning. We propose a GPT based architecture to fuse the feature maps learnt from ResNet on mulit-modality sensory data from camera, LiDAR, radar signal and GPS locations. We developed effective preprocessing techniques on image enhancement, point-cloud projection, filed-of-view alignment, radar singal preprocessing, and GPS angle normalization. We also proposed data augmentation methods on each modality. We train the deep learning model by applying focal loss, cosine decay learning rate, soft beam index and exponential moving average. The results show that our proposed deep learning model, data preprocessing and training schemes significantly improves sensor aid beam prediction from accuracy of 0.46 to 0.67, with effective generalization in unseen scenario 31 improved from 0.12 to 0.54. We also exploited that the background visionary information is beneficial in improving beam prediction.

We are further developing advanced deep learning techniques to improve the current results. Domain generalization is an important issue in this task, because the data in scenario 31 and the changed sampling rate in test dataset have different distribution than the training dataset. The Batchformer [17] algorithm is potentially efficient in making the model robust to imbalance data, by exploring data sample relationships. Moreover, semi-supervised learning such as the FixMatch [18] algorithm can improve the model on unlabeled data by training on pseudo labels from evaluation confidences. These methods are useful in practice with no additional computing complexity.

Furthermore, our transformer-based multi-modal deep learning framework is very useful for different applications of integrated communication and sensing. For example, one can modify the last fusion layer to other tasks in communications, such as beam switching, power control, resource allocation, link adaption, and so on. On the other hand, the input modalities can be changed according to the availability of different sensors, for related task such as sensing aid positioning, trajectory prediction, environment reconstruction. The framework of using attention mechanism from GPT is effective to learn abstraction of different modality data, which can further produce semantics for targeted tasks. We will extend this work to much wider applications and explore more challenging problems as proposed.

Acknowledgment

We would like to thank Dr Kebin Wu for her guidance and fruitful discussions in this work, and the support from TII on our participation in the ITU AI/ML for 5G challenge 2022.

References

[1] G. Charan, U. Demirhan, J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Multi-modal beam prediction challenge 2022: Towards generalization," *arXiv preprint arXiv:2209.07519*, 2022.

[2] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, and N. Srinivas, "DeepSense 6G: Large-scale real-world multi-modal sensing and communication datasets," *to be available on arXiv*, 2022. [Online]. Available: https://www.DeepSense6G.net

[3] "Multi Modal Beam Prediction Challenge 2022: Towards Generalization," accessed: 2022-11-27. [Online]. Available: https://deepsense6g.net/multi-modal-beam-prediction-challenge/

[4] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] T. Nishio, Y. Koda, J. Park, M. Bennis, and K. Doppler, "When wireless communications meet computer vision in beyond 5g," *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 76–83, 2021.

[6] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *European Conference on Computer Vision*. Springer, 2020, pp. 492–511.

[7] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired from pid controller," *arXiv preprint arXiv:2206.02066*, 2022.

[8] S. Kapoor, "Point cloud data augmentation for safe 3d object detection using geometric techniques," Blekinge Institute of Technology, 371 79 Karlskrona, Sweden, 2021.

[9] M. Liu and J. Niu, "Bev-net: A bird's eye view object detection network for lidar point cloud," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 5973–5980.

[10] U. Demirhan and A. Alkhateeb, "Radar aided 6g beam prediction: Deep learning algorithms and real-world demonstration," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 2655–2660.

[11] G. Wei, Y. Zhou, and S. Wu, "Detection and localization of high speed moving targets using a short-range uwb impulse radar," in *2008 IEEE Radar Conference*, 2008, pp. 1–4.

[12] J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Position aided beam prediction in the real world: How useful gps locations actually are?" 2022. [Online]. Available: https://arxiv.org/abs/2205.09054

[13] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, Jul 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[16] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[17] Z. Hou, B. Yu, and D. Tao, "Batchformer: Learning to explore sample relationships for robust representation learning," in *CVPR*, 2022.

[18] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.