MULTIMODAL TRANSFORMERS FOR WIRELESS COMMUNICATIONS: A CASE STUDY IN BEAM PREDICTION

Yu Tian[1], Qiyang Zhao[1], Zine el abidine Kherroubi[1], Fouzi Boukhalfa[1], Kebin Wu[1], Faouzi Bader[1]
[1]Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, United Arab Emirates

NOTE: Corresponding author: Qiyang Zhao, qiyang.zhao@tii.ae

Abstract – *Wireless communications at high-frequency bands with large antenna arrays face challenges in beam management, which can potentially be improved by multimodality sensing information from cameras, LiDAR, radar, and GPS. In this paper, we present a multimodal transformer deep learning framework for sensing-assisted beam prediction. We employ a convolutional neural network to extract the features from a sequence of images, point clouds, and radar raw data sampled over time. At each convolutional layer, we use transformer encoders to learn the hidden relations between feature tokens from different modalities and time instances over abstraction space and produce encoded vectors for the next-level feature extraction. We train the model on a combination of different modalities with supervised learning. We try to enhance the model over imbalanced data by utilizing focal loss and exponential moving average. We also evaluate data processing and augmentation techniques such as image enhancement, segmentation, background filtering, multimodal data flipping, radar signal transformation, and GPS angle calibration. Experimental results show that our solution trained on image and GPS data produces the best distance-based accuracy of predicted beams at 78.44%, with effective generalization to unseen day scenarios near 73% and night scenarios over 84%. This outperforms using other modalities and arbitrary data processing techniques, which demonstrates the effectiveness of transformers with feature fusion in performing radio beam prediction from images and GPS. Furthermore, our solution could be pretrained from large sequences of multimodality wireless data, on fine-tuning for multiple downstream radio network tasks.*

Keywords – Beam prediction, multimodal learning, transformer, wireless communications

1. INTRODUCTION

Wireless communications beyond 5G 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 have been leveraged to create ultra-narrow beams, so as to increase the received signal power and reduce interference on targeted users. Significant challenges in beam management arise in such systems and scenarios especially for high mobility vehicle users, to provide ultra-high reliable and low latency communications.

Multimodality sensory information has the potential to improve wireless communications in a challenging environment. Integrated sensing and communication has been actively studied for 6G [1]. In the vehicular network scenario, a roadside Base Station (BS) unit equipped with a camera, LiDAR, radar, and GPS can produce images, point clouds, radar signals, and location information of the road environment, objects, and vehicle users (UE). Such sensory data is potentially useful in assisting the BS to analyze the radio transmission scenario, so as to produce effective beam management.

1.1 Problem statement

In this paper, we present a transformer-based multimodal deep learning approach for sensing-assisted beam prediction, which is a solution to the DeepSense 6G problem statement in the ITU AI/ML for 5G challenge 2022 [2]. The challenge aims to develop machine learning-based models that can adapt to diverse environmental features and accurately predict optimal beam indices in entirely new locations using a multimodal training dataset. The objective is to enable effective generalization and sensing-aided beam prediction for improved wireless communication systems. The challenge provides large multimodal sensing datasets measured in a real environment. As shown in Fig. 1, each data sample contains five sequential instances of camera images, LiDAR point clouds, and radar signals, plus the first two instances of UE GPS position [3]. The ground truth in this context refers to the corresponding 64×1 power vectors obtained through beam training at the receiver using a 64-beam codebook, where omni-transmission is employed at the transmitter. The BS sensors capture LiDAR, radar, and visual data, while the positional data is collected from GPS receivers installed on the mobile vehicle. The dataset is measured in four scenarios (31, 32, 33, 34) shown in Fig. 5. Scenarios 31 and 32 are measured in the daytime while scenarios 33 and 34 are at night. Note that scenarios 32 and 33 are in the same location but at different times. A development dataset is provided with thousands of samples collected in scenarios 32, 33, and 34; and an adaptation dataset is provided with tens of samples collected in scenarios 31, 32, and 33. Both datasets have the ground-truth best beam of the UE associated with each sample.

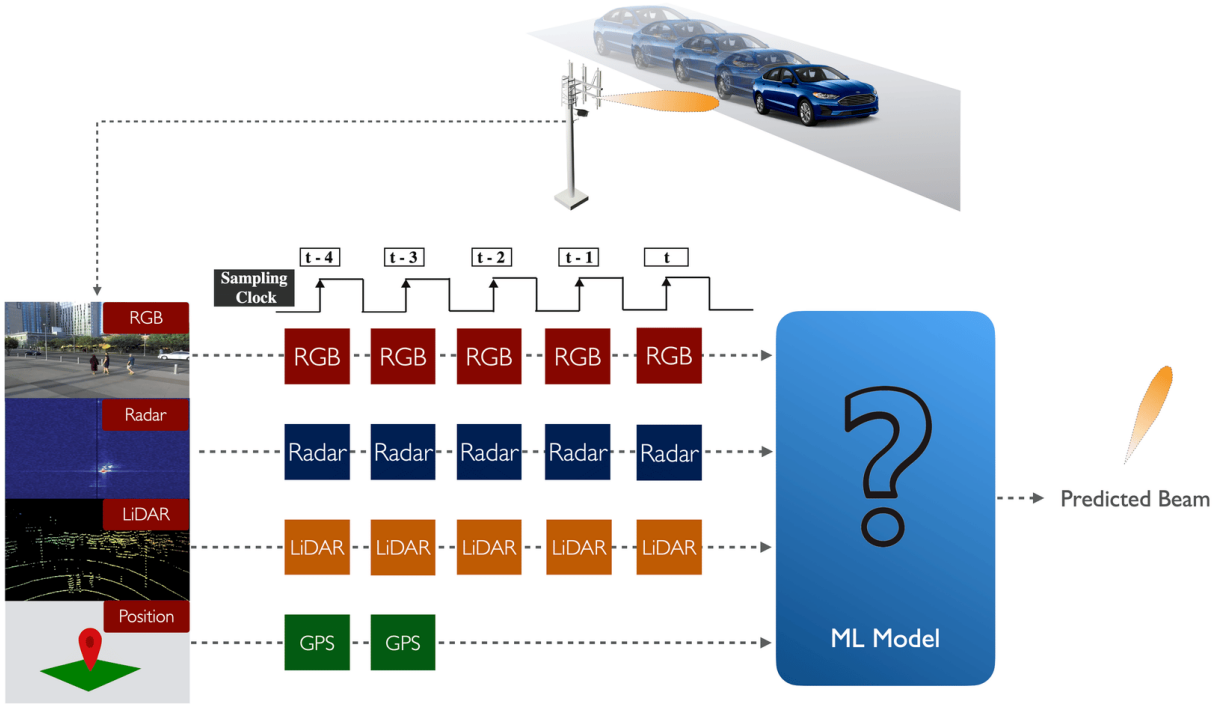


Fig. 1 – Schematic representation of the input data sequence utilized in this challenge task [2]

A test dataset with hundreds of samples is provided in all scenarios without labels. Specifically, most labeled data resides in scenarios 32, 33, and 34, whilst half of the unlabeled data resides in scenario 31. The sampling rate of the sequences in the test set is the same as that of the adaptation set but double that of the development set. The objective is to evaluate how the developed model can generalize to the unseen scenario, in which the sensing data is collected in a different location, Field of View (FoV), time (day, night), and sampling rate.

The evaluation metric is a "Distance-Based Accuracy Score (DBA Score)" with the top-3 predicted beams [2]. The DBA score is defined as:

$$\text{DBA score} = \frac{1}{3}(Y_1 + Y_2 + Y_3), \tag{1}$$

where $Y_K,\ K \in \{1, 2, 3\}$ is

$$Y_k = 1 - \frac{1}{N}\sum_{n=1}^{N} \min_{1 \le k \le K} \min\left(\frac{|\hat{y}_{n,k} - y_n|}{\Delta}, 1\right). \tag{2}$$

with y_n and $\hat{y}_{n,k}$ are respectively ground truth and the kth most-likely predicted beam indices of sample n in the dataset with a size of N. Δ is a normalization factor and set as 5.

1.2 Related work

There exists several solutions for multimodal sensor data fusion for multiple downstream tasks, 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, and 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 are no labels or bounding boxes indicating the UE. Secondly, we have also radar signals 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 that has not been well exploited with multimodal sensors. In particular, the relations between radio transmission scenarios and visionary data on abstraction space is not straightforward, making deep learning hard to generalize on unseen scenarios [5].

The use of visual data for wireless communications has been actively studied in recent years, including most work on beam prediction from the DeepSense group. This includes the use of images for beam tracking with Gated Recurrent Unit (GRU)-based deep learning [6]. Radar-aided beam prediction is studied in [7] using 2D Convolutional Neural Network (CNN). It also proposes FFT to transform radar signals to range angle and velocity maps for CNN. LiDAR-aided beam prediction is investigated in [8], which is also based on GRU plus an embedding block to convert 3D point clouds to 1D vectors. Position-aided beam prediction is studied in [9], which utilizes Multilayer Perceptron (MLP). A fusion of vision and position has been stud-

ied in [10], which concatenates the normalized position with extracted features from CNN to predict beams with MLP. Despite that, a number of solutions have been developed in this domain, most of which are not scalable to different modalities of sensory data. To achieve this we need to build a generalized ML model which can learn the abstracted features between multiple modalities, which is a key target of this paper.

1.3 Contributions

Our contribution can be summarized as follows. First, we develop a multimodal transformer framework for wireless communication application of beam prediction and prove that the model is flexible to adapt to different data modalities in the wireless domain. Second, we investigate several data processing and augmentation techniques in computer vision for wireless applications, alongside model training and validation methods for data imbalance and domain adaptation problems. Third, we validate with real measurement data that our framework is effective in producing beam prediction from multimodal sensory data, and generalize to unseen scenarios. Finally, we discuss that our framework could be extended to a generative model pretrained on sequences of multimodality data and fine-tuned for multiple tasks in radio air interfaces.[1]
The rest of this paper is structured as follows. Section 2 describes our developed methods for multimodal sensor data transformation and processing. Section 3 describes our proposed multimodal transformer framework for sensing-assisted beam prediction, with discussions on the training method and extension capabilities. Section 4 presents experiments of our solution on the multimodal beam prediction applications with discussions on the studied approaches. Finally, the work is concluded in Section 5 with some future research directions for the framework.

2. MULTIMODAL DATA TRANSFORMATION AND PROCESSING

2.1 Multimodal data transformation

We start by transforming LiDAR point clouds and radar signals into 2D vector space as well as calibrating GPS location data.
For LiDAR data, we convert the raw point clouds into an image-like representation through a Bird's-Eye View (BEV). Specifically, 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. Firstly, the point clouds within the Region Of Interest *(ROI)* are discretized into grid cells. Secondly, the height and intensity are encoded by their maximum values of the points in each grid cell. Finally, the density of the points is calculated [11]. The BEV representation for LiDAR point clouds has certain advantages. It can be used on CNN [12] to extract hidden features, which can be further processed with images. Moreover, it can preserve the basic structure of the point clouds and the depth information, while reducing the computational complexity in PointNet [11].

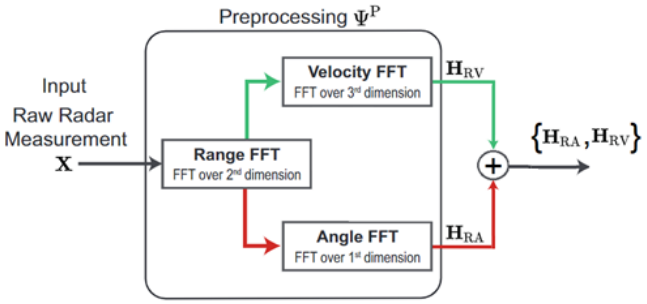


Fig. 2 – Combining radar Range-Angle $\mathbf{H}_{RA}$ and Range-Velocity $\mathbf{H}_{RV}$maps [7]

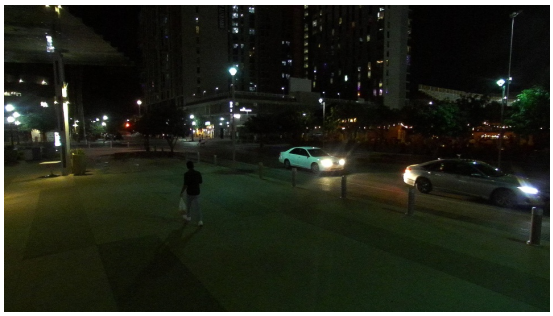
For radar data, we adopted the processing techniques used in [7]. The objective is to extract the range, the angles, and the velocity of the moving objects in the environment using *2D* Fourier transform, as described in [7]. Since the camera and LiDAR do not provide explicit velocity information, we concatenate the *Range-Angle Maps* with the *Range-Velocity Maps* of the radar to preserve the speed information of the moving cars, as illustrated in Fig. 2. Further, radar signals provide reliable speed measurement regardless of weather conditions and lightness level [13].

GPS data plays an important role in locating the UE's position. However, it is not always available or accurate in a practical system (caused by connection and delay issues). In this challenge, only data from the first two out of the five GPS instances are provided. We first transform the GPS coordinate of the UE and the BS from to the Cartesian, then calculate the relative position between the UE and the BS of the nt h GPS data, denoted as $(\Delta x_n, \Delta y_n)$. Afterward, we get the angle by $\arctan(\Delta y_n / \Delta x_n)$.

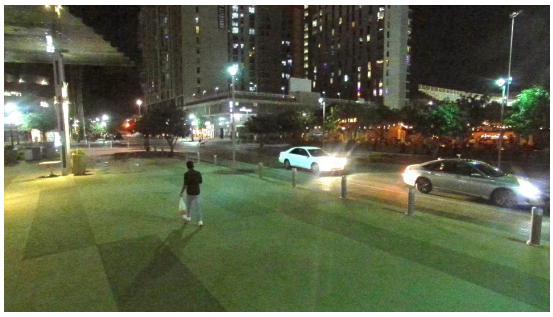
After exploring the dataset, we observed that the beam indices spread from 1 to 64 according to the UE's locations from left to right in the images. As the camera is located close to the BS, the beam indices are associated with the relative position (angle) between the UE and BS. However, the angles of the same beam index are different between scenarios, because roads are located in different positions with reference to the BS. Therefore, we calibrate the angle of the central pixel in the images of all scenarios.

We first manually select the data samples of these four scenarios where the UE is located in the middle of the images and their corresponding beam indices fall in the range of $[31, 34]$. We then calculate their angles according to their relative positions, as $[\theta_1 = -50.52^\circ, \theta_2 = 44.8^\circ, \theta_3 = 55.6^\circ, \theta_4 = -60^\circ]$. We rotate all the possible angles in each scenario with θ_i, $(i = 1, 2, 3, 4)$. Finally, we obtain the calibrated angles of the first two instances.

[1] https://github.com/ITU-AI-ML-in-5G-Challenge/DeepSense6G_TII.git



(a) The original image



(b) The enhanced image

Fig. 3 – Image enhancement in night scenario

2.2 Multimodal data processing

In this section, we introduce several data processing techniques on the multimodal data for training the multimodal transformers on the beam prediction task.

2.2.1 Camera data

Beam prediction from camera data is related to object detection and tracking tasks in computer vision. However, since there are no labels of the targeted UE in the image, we cannot distinguish it from other vehicles or pedestrians. Therefore, we tried to enhance the visual information of the vehicles in the images to allow the model to better recognize our targeted object.

Brightness enhancement: To overcome the darkness issue in the night scenarios 33 and 34, we utilize MIRNet [14] to enhance the brightness of these images. The vehicles become clearer as shown in Fig. 3b, compared to the raw image in Fig. 3a.

Segmentation: To highlight the vehicles in the camera data, we use the PIDNet [15] to segment the vehicles from images in the daytime scenarios 31 and 32 shown in Fig. 4. We also test this method on the brightened images in the night scenarios 33 and 34, but the performance is poor, which may be due to loss of background information.

Background masking: We also tried to mask the background with the black color and keep the street scene. The images in the same scenario have the same background because the camera is stable. Beam prediction can be partially seen as trajectory prediction over the horizontal axis. We can potentially make the neural network focus on the vehicle's trajectory by making it dominant in the images, as shown in Fig. 5.

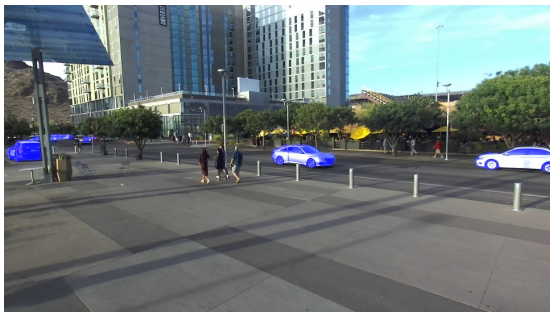
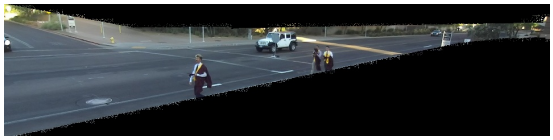
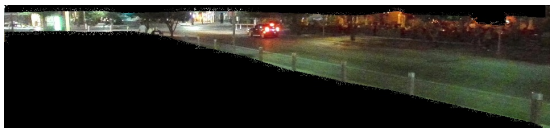


Fig. 4 – Image segmentation on vehicles (blue) in day scenario

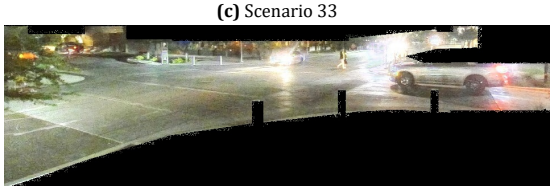


(a) Scenario 31

(b) Scenario 32



(c) Scenario 33



(d) Scenario 34

Fig. 5 – Image background masking

2.2.2 LiDAR data

The LiDAR produces on average more than *16000* 3D points in each time step. In order to reduce the size of the point-cloud data to speed up training the model, we preprocessed LiDAR data in the following ways:

Background filtering: We removed data points that correspond to static objects i.e. buildings. Similar to images these regions are not in the Line-Of-Sight *(LOS)* link between the BS and UE, which has less effect on the beam prediction. These points potentially add complexity and bias to the model. We subtract the background points from each point cloud frame using the moving average of all the frames in each scenario and then keep the desired region surrounding the moving vehicles.

FoV calibration: We crop the BEV projection of the LiDAR data to keep its FoV consistent with the view in the images. This could potentially assist the CNN focus on the region aligned with the images, to better allow transformers to learn the relations between them.

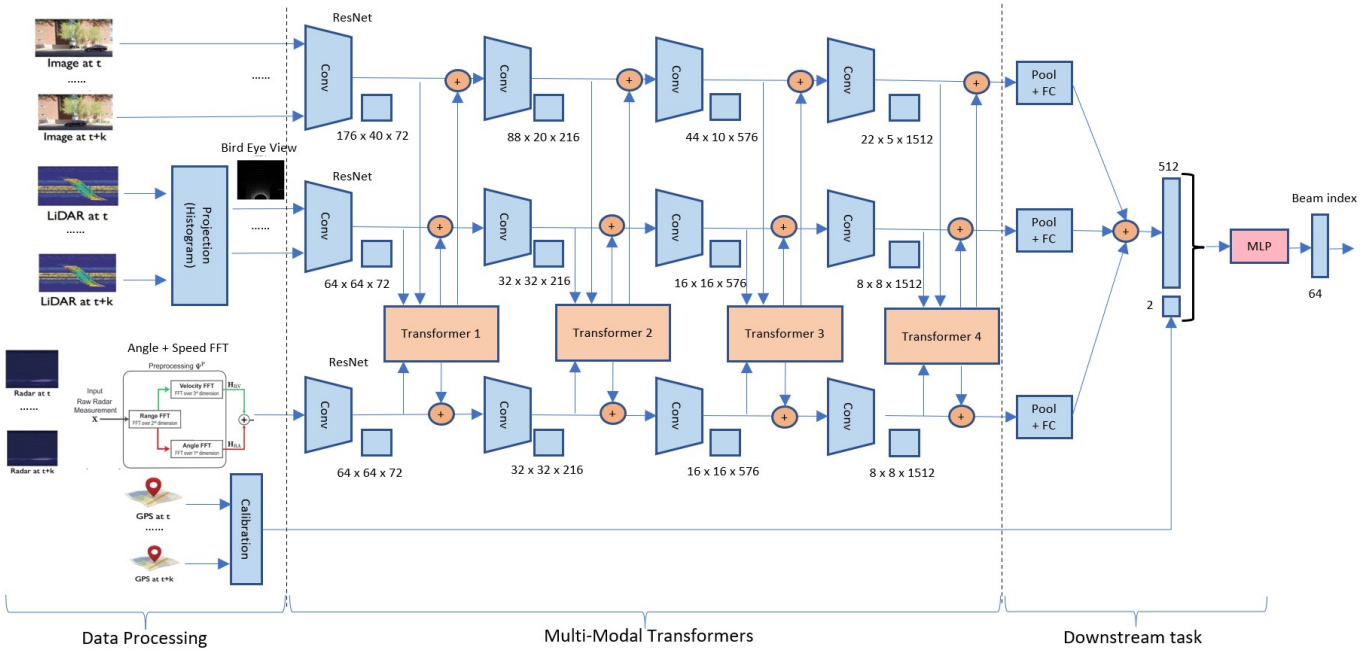


Fig. 6 – Multimodal transformers for sensing-assisted beam prediction

2.2.3 Data augmentation

Due to data imbalance between scenario *31* and others, we investigate data augmentation techniques to increase the dataset size for this scenario.

Image: Beam selection relies mainly on the transmitter/receiver locations and the geometry/characteristics of the surrounding environment [7]. In order to conserve this geometric information, we use only some *photometric* transformations that are *'safe'* for beam prediction application [16]. We augment each image by randomly changing the brightness, contrast, gamma correction, hue channel, color saturation, the sharpness, and performing *Gaussian* blurring on the image.

Point-cloud: 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).

Radar signal: 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.

Multimodal data flipping: According to the observations beam indices spread from 1 to 64 in the images, we horizontally flip the images, radar, and the point cloud data to achieve the augmentation purpose. Meanwhile, to keep the GPS data and beam indices consistent with the multimodal data, we reverse the calibrated GPS angles and get the new beam indices by subtracting the original indices from 65.

3. MULTIMODAL TRANSFORMERS FOR BEAM PREDICTION

In this section, we introduce our solution of a multimodal transformer framework for wireless communications and deep learning algorithms for sensing-assisted beam prediction.

3.1 Multimodal transformer architecture

With multimodality data transformed into 2D vector spaces, we leverage CNN to extract the higher-order features, and then learn the relations between them using transformers. Since the image, point-cloud, and radar signal raw data resides in very different representation spaces, it is difficult to create effective mathematical functions, i.e. through category theory, to transform them into a common abstraction space, in order to learn their structures. However, with multiple layers of learning extraction on CNN and relations on transformers, the deep learning model could potentially converge on an effective representation of multiple modalities. Such representation can be fine-tuned for different downstream tasks which is essentially a structure minimization process.

In this context, we build a multimodal transformer architecture as illustrated in Fig. 6. We first employ a deep residential network (ResNet) [17] to encode the image, point cloud, and radar signal on feature space. Specifically, the ResNet is used on each of the five instances of the RGB image, LiDAR BEV, and radar range angle-velocity map, after normalization and scaling to a 512×1 feature vector.

Each ResNet block of convolution, batch normalization, non-linear activation, and pooling produces an abstracted

feature vector as tokens. Note that for each modality we have five tokens sampled in different time steps. We use transformer encoder layers after each convolutional block to fuse the intermediate abstractions between the modalities of the image, point cloud, and radar map. The transformer uses linear projections for computing a set of queries, keys, and values. Scaled dot products are used between queries and keys to compute the attention weights and then aggregate 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 in attention layers with multiple heads to generate several queries, keys, and values. Since each convolutional block encodes 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 data at different modalities and time steps. In theory, the fusion of 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. In this manner, the transformer could estimate the position of the target UE in the scene at the 5^{th} instance.

The fused feature maps of different modalities are propagated to the next convolutional blocks and repeated several times with transformer blocks, and finally added together to be a 512×1 feature vector. Because the calibrated GPS locations (angles) have more apparent information than the other three pieces of data and only the first two instances are available, these two angles are concatenated with the 512×1 vector and passed through MLP layers to produce weights of 64 beam index using the softmax function.

3.2 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 neighboring five beams. This is to adapt the cross-entropy loss function to the DBA score, where higher weights are given if the beams are closer to the best beam.

We further apply a focal loss [18] 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 are much less than others. Moreover, some beams have much less probability to be served as the best beam than others. The adaptation dataset is with a 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 examples receiving a low loss.

We also employ several training methods to stabilize the convergence and make the model robust. We maintain the Exponential Moving Average (EMA) of the parameters during training, instead of utilizing the final trained values. This eliminates the fluctuation at the final steps and makes the model robust.

4. PERFORMANCE EVALUATIONS AND DISCUSSIONS

We performed experiments to train and evaluate our proposed multimodal transformer and data processing frameworks for beam prediction over the DeepSense challenge dataset [3]. The performance is measured in the DBA score defined in Eq. *(1)*, where the distances of the predicted beam to the ground-truth top three beams are averaged according to the mmWave received signal power from the vehicle UE to BS.

4.1 Beam prediction accuracy

We combine the development and adaptation datasets, then randomly split them into 90% for training and 10% for validation (to choose the best model weights and hyperparameters). The learning rate is set to start from 10^{-4}. We validate and compare the performance using different proposed data preprocessing, augmentation, ResNets (ResNet18 and ResNet34) and model training approaches, according to the accuracy scores evaluated on the test dataset provided by the organizer. The hyperparameter with the best performance on the validation dataset is submitted for evaluation. Since the training and test datasets have a large imbalance in scenario 31, it can show the generalization capability of the trained model performing in the unseen scenario.

The experimental results of different model training and data processing schemes are shown in Table 1. We compare the performances of the model on camera, radar, LiDAR, GPS, and multimodality data. We can first observe that the experiment using ResNet34 to encode images has a higher accuracy than that with ResNet18, and the experiments with camera data on all five instances of raw images already achieved an overall accuracy of 75%. This outperforms largely using only the last instance, indicating that the transformer can effectively utilize the relations between images sampled at different times to predict the beams, though car user is not indicated in the image. We can also see that its performance is better than, or similar to, most data preprocessing techniques, such as brightness enhancement, segmentation, and background masking, which further proves that the multimodal transformer model can generalize to different data domains without arbitrary processing. For example, its performance at unseen scenario 31 is close to the same day scenario 32 nearly 70%, without any data augmentation. Furthermore, we can see that it performs 10% better in night scenarios 33 and 33, mainly due to the mobility of

Table 1 – Distance-based accuracy of beam prediction on multimodal test dataset

Data Type[1]	Scheme[2]	Overall	Scenario 31	Scenario 32	Scenario 33	Scenario 34
Camera	Raw Image[18]	0.6535	0.5124	0.7457	0.7705	0.8137
	Raw Image[34]	**0.7548**	**0.6982**	**0.7160**	**0.8024**	**0.8494**
	5th instance[34]	0.6546	0.5171	0.7568	0.7548	0.8204
	Brightness Enhancement[34]	0.7327	0.6853	0.7469	0.7371	0.8305
	Background Masking[34]	0.7571	0.6896	0.7383	0.8157	0.8570
	Image Segmentation[34]	0.6979	0.5873	0.7556	0.7824	0.8372
	EMA Model Weights[34]	0.7146	0.6178	0.7642	0.7852	0.8402
	Cross Entropy Loss[34]	0.7395	0.7018	0.7420	0.7410	0.8234
Radar	Range - Angle & Velocity[34]	0.2807	0.1840	0.2827	0.4429	0.3282
	Range - Angle & Velocity [18]	0.3563	0.2936	0.3160	0.4800	0.3842
	Range - Angle[18]	0.3092	0.2462	0.1926	0.4686	0.3313
LiDAR	Raw Point-Cloud[34]	0.4362	0.3171	0.4037	0.6781	0.4636
	Raw Point-Cloud[18]	0.4422	0.3260	0.4272	0.6705	0.4707
	FoV Calibration[18]	0.4223	0.2964	0.4370	0.6781	0.4310
	Background Filtering[18]	0.2794	0.2598	0.2123	0.2986	0.3313
GPS	**Angle calibration**	**0.7425**	**0.6353**	**0.7704**	**0.8229**	**0.8906**
	Angle calibration + distance on 2nd instance	0.6266	0.4718	0.6704	0.8481	0.7262
Multimodal	Images[34] + Radar (Angle)[18]	0.6992	0.6304	0.6938	0.7533	0.8010
	Images[34] + Radar (Angle)[34]	0.7206	0.6378	0.7383	0.8033	0.8148
	Images[34] + Radar (Angle)[18] + Point-Cloud[18]	0.6356	0.5049	0.7333	0.7519	0.7705
	Images[34] + Radar (Angle)[34] + Point-Cloud[34]	0.7358	0.6649	0.7938	0.7919	0.8142
	Images[34] + GPS	**0.7767**	**0.7253**	**0.8000**	**0.8038**	**0.8560**
	Images[34] + GPS (Image Augmentation)	0.7127	0.5764	0.7654	0.8576	0.8483
	Images[34] + GPS (Flipping Augmentation)	**0.7844**	**0.7298**	**0.7852**	**0.8462**	**0.8433**
Best score on the leaderboard of the challenge[3]		0.7162	0.6536	0.7074	0.8576	0.712

[1] Data modalities with all 5 (GPS 2) instances unless specified.
[2] Data processing and model training schemes. Focal loss applied in all experiments unless specified.
[3] Leaderboard: `https://deepsense6g.net/ml-task-multimodal-beam-prediction`
[4] The superscript 18 indicates that ResNet18 is used for feature extraction, while 34 denotes the utilization of ResNet34.

car lights in the images being easier to identify, than multiple objectives appearing in the day scenarios.

In the performance of radar and LiDAR data, we can see that the model achieves the lower accuracy than images, and ResNet18 outperforms ResNet34 on encoding these two data. This is because the radar signals and point clouds received at the BS are reflected by all the moving vehicles and objects, making the model hard to detect the UE. Meanwhile, deeper residual layers lead to overfitting issues. Moreover, the signal has coverage constraints, causing issues in detecting UEs far away. Specifically, for the radar data, combining range-angle and range-velocity performs 5% better, which validates that velocity information can help the transformer to predict the UE mobility and select the beam. For the LiDAR data, filtering the background degrades the performance, because reference information of the UE in the environment could be cut out. This also explains that the model with LiDAR performs better than radar which only contains information about moving objects.

In the performance of GPS data, angle calibration on the first two instances achieve the best accuracy in scenario 34 at 89%, which outperforms the distance and angle calibration on the 2nd instance. This indicates that only two instances of GPS data can predict the beam very effectively, reaching an accuracy of 74% which is very close to using images. Our angle calibration scheme is very effective while the distance information is less useful.

The best performance is achieved on multimodal data using images and GPS. It can be observed that the transformer on these two modalities produces an overall accuracy of 77%, which is better than using them separately. This is much more significant in the unseen scenario 31, with 10% higher accuracy than using GPS only. This proves the advantage of multimodal fusion on the feature level of our transformer framework. The GPS information can assist the model to identify the UE in images, which improves accuracy in day scenarios. The fusion of images with radar and LiDAR data also largely outperforms using them separately by 30% to 45%. Moreover, when it comes to fusing images, employing ResNet34 for encoding radar and LiDAR data yields considerable performance enhancements as compared to using ResNet18. Furthermore, we also implement the data augmentation techniques in scenario 31. It demonstrates that flipping the images further enhances the performance, reaching the best overall accuracy of 78% than all other schemes. Finally, we compare our solution with the best score on the leaderboard of the challenge, which uses convolutional autoencoders to fuse the images and GPS data [19]. It can be seen that our multimodal transformer achieves 7% better accuracy in overall performance and scenarios 31 and 32, and 13% better in scenario 34. This further proves the effectiveness of this framework in solving the beam prediction problem.

4.2 Model complexity

We investigate the complexity of our proposed framework from aspects of Multiply–Accumulate Operations (MACs) and number of parameters (Params), and then compare ours with the best solution on the leaderboard of the challenge described in [19]. Note that the best solution of [19] utilized images, radar, and GPS data. Authors extract features from images by using CNN-based autoencoders. They get a threshold of radar heatmaps using a 2D Constant False Alarm Rate (CFAR), and then apply Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to obtain the object angle. They also calibrate GPS data in a similar way to ours. Finally, these three pieces of preprocessed data are concatenated and go through a dense model to predict the best beam. As we can't get the detailed parameters or the code of CFAR and DBSCAN in [19], we only calculate the MACs and Params of the feature extraction and dense model. For our model, we studied the main blocks in Fig. 6 with an input of five-instance data and three of the best schemes in Table 1. From Table 2, we observe that the MACs and Params of

Table 2 – MACs and Params

Source	Block or Method	MACs	Params
Main blocks in Fig. 6 (5 instances)	ResNet18	2,368,733,184	11,166,912
	ResNet34	4,784,652,288	21,267,648
	Transformer 1	127,221,760	400,000
	Transformer 2	506,101,760	1,586,432
	Transformer 3	2,018,836,480	6,318,592
	Transformer 4	8,064,204,800	25,220,096
This paper in Table 1	Overall best scheme (Images^{34}+GPS)	34,740,378,624	54,982,784
	Best scheme of camera data (5th instance^{34})	6,948,213,248	54, 982,272
	Best scheme of GPS data (Angle calibration)	41,472	41,920
In [19]	Feature extraction	191,949,184	39,998,304
	Dense model	303,616	304,704

ResNet34 are nearly double that of ResNet18. MACs and Params of 'Transformer 1', 'Transformer 2', 'Transformer 3', and 'Transformer 4' increase quadruply. The most complex block is the 4th transformer. Our overall best scheme with transformers is the one that is most computationally costly. However, when considering only the 5th image, the scheme experiences a substantial reduction of $\frac{4}{5}$ in terms of MACs and Params. On the other hand, the best scheme solely relying on GPS is the simplest among all the solutions. It is important to note that this scheme demonstrates significantly lower complexity while delivering superior performance compared to the best solution presented in [19]. Therefore, our low-complexity scheme is suitable for scenarios with limited computational resources, making it a viable option. On the other hand, the high complexity scheme is best suited for scenarios that demand high accuracy and possess abundant computational resources.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we present a multimodal transformer deep-learning solution for wireless communications and perform a case study in mmWave beam prediction for a target vehicle user. The transformer encoder is used to learn abstracted relations between features of images, point clouds, and radar signals at multiple time instances, extracted by convolutional layers. Multiple layers of transformers and ResNets are stacked to learn higher-level abstractions for downstream tasks. We employed data transformation techniques of point cloud projection, radar range-angle, and range-velocity FFT to convert the multimodal data on 2D vector space, as well as GPS calibration. Furthermore, we develop data processing

techniques to improve the task, including image brightness enhancement, segmentation, background masking, LiDAR field of view calibration, and background filtering. We also proposed data augmentation to reduce overfitting in training, including flipping the images. We trained the model by applying focal loss and exponential moving average techniques.

Experimental results show that our proposed multimodal transformer solution using image and GPS data achieves the best distance-based accuracy of predicted beams at 78.44%, with effective generalization to each of the scenarios at 73%, 78.5%, 84.6%, 84.3%, respectively. This outperforms significantly using LiDAR and radar, as well as each single modality. Specifically, the transformer effectively utilizes GPS information to detect the target UE in the images, whilst the images can assist GPS to generalize better in the unseen scenario. Furthermore, it also performs 7% better than the best state-of-the-art using autoencoders. We can conclude that our proposed multimodal transformer can effectively perform tasks between visual and radio domains, and generalize to different scenarios without customized data preprocessing and augmentation.

Further advanced deep learning models and techniques are worth studying for improving performance, especially feature extraction from data in other modalities. Domain generalization is an important issue in this task because the data in scenario 31 and the changed sampling rate in the test dataset have a different distribution than the training dataset. The Batchformer [20] algorithm is potentially efficient in making the model robust to imbalanced data, by exploring data sample relationships. Moreover, semi-supervised learning such as the FixMatch [21] 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.

The multimodal transformer framework can be utilized to build a foundation model to empower multiple downstream tasks in wireless communications. We can pretrain with self-supervised learning to build a generative model from sequences of images, LiDAR, radar, and radio signals collected at different times, frequencies, and locations, which learns high-level abstractions and relations among them. The transformer output can be stacked with classification or regression layers and fine-tuned for downstream tasks related to this data, such as channel prediction, beam management, and modulation. The pretrained model can also be used on devices with fewer sensors and less computing power by adapting the model branch and depth. It is a promising research direction to investigate such architecture for foundation models in the wireless communication domain.

ACKNOWLEDGMENT

We would like to thank the support from TII for our participation in the ITU AI/ML for 5G Challenge 2022.

REFERENCES

[1] Yu Tian, Gaofeng Pan, and Mohamed-Slim Alouini. "Applying Deep-Learning-Based Computer Vision to Wireless Communications: Methodologies, Opportunities, and Challenges". In: *IEEE Open Journal of the Communications Society* 2 (2021), pp. 132–143. DOI: 10.1109/OJCOMS.2020.3042630.

[2] G. Charan, U. Demirhan, J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb. "Multi-Modal Beam Prediction Challenge 2022: Towards Generalization". In: *arXiv preprint arXiv:2209.07519* (2022).

[3] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, and N. Srinivas. "DeepSense 6G: Large-Scale Real-World Multi-Modal Sensing and Communication Datasets". In: *to be available on arXiv* (2022). URL: https://www.DeepSense6G.net.

[4] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. "Multi-Modal Fusion Transformer for End-to-End Autonomous Driving". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[5] Takayuki Nishio, Yusuke Koda, Jihong Park, Mehdi Bennis, and Klaus Doppler. "When Wireless Communications Meet Computer Vision in Beyond 5G". In: *IEEE Communications Standards Magazine* 5.2 (2021), pp. 76–83. DOI: 10.1109/MCOMSTD.001.2000047.

[6] Shuaifeng Jiang and Ahmed Alkhateeb. *Computer Vision Aided Beam Tracking in A Real-World Millimeter Wave Deployment*. 2021. DOI: 10.48550/ARXIV.2111.14803.

[7] Umut Demirhan and Ahmed 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.

[8] Shuaifeng Jiang, Gouranga Charan, and Ahmed Alkhateeb. "LiDAR Aided Future Beam Prediction in Real-World Millimeter Wave V2I Communications". In: *IEEE Wireless Communications Letters* (2022), pp. 1–1.

[9] João Morais, Arash Behboodi, Hamed Pezeshki, and Ahmed Alkhateeb. *Position Aided Beam Prediction in the Real World: How Useful GPS Locations Actually Are?* 2022. URL: https://arxiv.org/abs/2205.09054.

[10] Gouranga Charan, Tawfik Osman, Andrew Hredzak, Ngwe Thawdar, and Ahmed Alkhateeb. "Vision-Position Multi-Modal Beam Prediction Using Real Millimeter Wave Datasets". In: *IEEE Wireless Communications and Networking Conference (WCNC)*. 2022, pp. 2727–2731.

[11] Shrayas Kapoor. "Point Cloud Data Augmentation for Safe 3D Object Detection using Geometric Techniques". MA thesis. Blekinge Institute of Technology, 371 79 Karlskrona, Sweden, 2021.

[12] Meng Liu and Jianwei 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. DOI: 10.1109/IROS51168.2021.9636810.

[13] Guohua Wei, Yuxiang Zhou, and Siliang 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. DOI: 10.1109/RADAR.2008.4720766.

[14] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. "Learning enriched features for real image restoration and enhancement". In: *European Conference on Computer Vision*. Springer. 2020, pp. 492–511.

[15] Jiacong Xu, Zixiang Xiong, and Shankar P Bhattacharyya. "PIDNet: A Real-time Semantic Segmentation Network Inspired from PID Controller". In: *arXiv preprint arXiv:2206.02066* (2022).

[16] Connor Shorten and Taghi M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". In: *Journal of Big Data* 6.1 (July 2019), p. 60. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal Loss for Dense Object Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 318–327. DOI: 10.1109/TPAMI.2018.2858826.

[19] Maximilian Arnold, Gouranga Charan, Umut Demirhan, Ahmed Alkhateeb, and Mohammed Alloulah. *Analysis of Multi-Modal Beam Prediction under Distribution Shift*. 2022. URL: https://github.com/ITU-AI-ML-in-5G-Challenge/BeamBench.

[20] Zhi Hou, Baosheng Yu, and Dacheng Tao. "BatchFormer: Learning to Explore Sample Relationships for Robust Representation Learning". In: *CVPR*. 2022.

[21] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. "FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence". In: *arXiv preprint arXiv:2001.07685* (2020).

AUTHORS

Yu Tian received a Ph.D. degree from King Abdullah University of Science and Technology (KAUST), Saudi Arabia in 2022. Since June 2022, he has been a researcher at the Technology Innovation Institute, Abu Dhabi, United Arab Emirates. His current research interests include deep learning and performance analysis of wireless communication systems.

Qiyang Zhao received Ph.D. degree in electronic engineering in 2013 from University of York, UK. He has been working with industry on research, development, and standardization of 5G and 6G wireless communication systems. Since February 2022, he has been with Technology Innovation Institute, UAE, where he is currently a lead researcher. His research interests cover various aspects of AI and telecommunications, with current emphasis on native AI networks, large language models, semantic communications, multi-agent systems.

Zine el abidine Kherroubi received a B.S. degree in automatic control systems engineering from École Nationale Polytechnique d'Alger, Algiers, Algeria in 2015 and an M.S. degree in mobility and electric vehicles from Art et Métier ParisTech, Lille, France in 2016, as a Fellow Student of the prestigious Renault Foundation Scholarship. He received a Ph.D. degree in computer sciences from Claude Bernard Lyon 1 University, France in 2020, in joint research between the LIRIS laboratory and Renault car manufacturer. He worked as an R&D engineer at Renault car manufacturer, France, between 2017 and 2020. He is currently with the Technology Innovation Institute, Abu Dhabi, UAE, as a researcher. His research interests include the application of AI/ML techniques for connected and autonomous vehicles technology, V2X network, collaborative driving, and vehicle-infrastructure cooperation.

Fouzi Boukhalfa received Ph.D. degree from the Sorbonne University, France in 2021, in joint research between INRIA PARIS and VEDECOM. After that, he joined Capgemini Group as a consultant on smart to grid (ENEDIS Nanterre). Since April 2022, he has been a researcher at the Digital Science Research Center in Technology Innovation Institute, Abu Dhabi, United Arab Emirates. His research interests include several aspects of V2X networks for connected and automated driving, with current emphasis on ISAC and AI solutions for 6G-V2X.

Kebin Wu received a B.S. degree in electronic and information engineering from the Harbin Institute of Technology in 2011 and a Ph.D. degree in information and communication engineering from Tsinghua University in 2018. She is now a senior researcher at the Technology Innovation Institute, with research interests including computer vision, multi-modality learning, and large language models.

Faouzi Bader received the Ph.D. degree (Hons.) in telecommunications from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2001. He joined the Centre Technologic de Telecomunicacions de Catalunya (CTTC), Barcelona, Spain, as a research associate in 2002, and from 2006 to 2013 as a senior research associate. From June to December 2013 he was an associate professor at CentraleSupélec, France. Since 2017 he has been as a honorary adjunct professor with the University of Technology Sydney, Australia, and from 2018 to 2019 as the Head of the Department of Signals and Communications, Institute of Electronics and Digital Technologies (IETR), Rennes, France. From 2020 to 2021 he took up the position of Director of Research at the Institut Supérieur D'Electronique de Paris (ISEP), France. Since December 2021, he has been the Director Telecom at the DSRC Centre of the Technology Innovation Institute (TII), Abu Dhabi, United Arab Emirates. His research interests include IMT-advanced systems such as 5G networks and systems, cognitive radio communication environment, and THz wireless communications (6G). He has been involved in several European projects from the 5th–7th EC research frameworks (eight EU funded projects and ten national projects), he has been the main coordinator of the BRAVE ANR French Project at CentraleSupélec, where the main goal was the achievement of efficient waveform for THz/terabits wireless communication devices. He has published over 45 journals, 136 papers in peer-reviewed international conferences, more than 13 book chapters, and four edited books. He served as a Technical Program Committee Member in major IEEE ComSoc and VTS conferences (ICC, PIMRC, VTC spring/fall, WCNC, ISWCS, GLOBECOM, and ICT).