Multi-Modal Sensing-assisted Beam Prediction using Real-World Dataset

Yerin Yeo, Junghyun Kim, Jihyung Kim, and Junhwan Lee

***Abstract*—This paper explores techniques for beam prediction using multi-modal sensing data. Specifically, our aim is to develop a deep learning model that predicts the optimal beam using information collected from camera, LiDAR, radar, and GPS sensors. For this purpose, we propose ResNet-SE, which integrates a squeeze-and-excitation network with ResNet, and PIformer, a transformer-based model we design using pooling layers and inception mixer modules. Experimental results demonstrate a 22% improvement in prediction accuracy and a 38% reduction in training time compared to the state-of-the-art model.**

***Index Terms*—Beam prediction, deep learning, multi-modal sensing, transformer, wireless communications.**

I. Introduction

Millimeter-wave (mmWave) communication systems utilize beamforming technology with large antenna arrays to achieve exceptional data rates. The use of narrow beams in beamforming minimizes interference between nodes and maximizes the received power of target beam. However, optimally managing narrow beams requires significant training overhead, especially in high-speed mobility environments, presenting considerable challenges in beam management. One effective approach to overcome this issue is to predict the optimal beam using multi-modal sensing data instead of in-band radio signals. The use of multi-modal sensing data in integrated sensing and communication holds potential for improving performance in challenging environments that support communication for high-speed moving devices, such as unmanned vehicles or autonomous cars. For instance, a base station (BS) equipped with camera, LiDAR, radar, and GPS sensors can collect images, point clouds, radar signals, and location information, utilizing this data for beam prediction. Using multi-modal data collected from various sensors simultaneously allows for a more robust response to environmental changes and obstacles, obtaining information that may not be available from a specific sensor through other sensors.

Manuscript received May 17, 2024; revised May 22, 2025; approved for publication by Mi, De, Guest Editor, June 21, 2025.

This study was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2021-0-00794, Development of 3D Spatial Mobile Communication Technology).

Yerin Yeo and Junghyun Kim are with the Department of Artificial Intelligence and Deep Learning Architecture Research Center, Sejong University, Seoul 05006, Republic of Korea, emails: 23110387@sju.ac.kr, j.kim@sejong.ac.kr.

Jihyung Kim and Junhwan Lee are with the Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea, emails: {savant21, junhwanlee}@etri.re.kr.

Junghyun Kim is the corressponding author.

Digital Object Identifier (DOI): 10.23919/JCN.2025.000049

The "Multi-modal beam prediction challenge 2022: Towards generalization" competition [1] was initiated to provide a platform for the design and evaluation of AI-based beam prediction models suitable for real-world wireless communication environments. This competition offers a comprehensive real-world multi-modal dataset, DeepSense 6G dataset [2], which includes communication and sensing data collected from various real-world locations and at different times of the day. In a significant research [3] utilizing this dataset, the authors proposed a neural network-based method that effectively utilizes user location information for beam prediction. Additionally, the authors in [4] suggested a multi-modal machine learning-based beam prediction technique that uses camera and GPS sensor data. A recent study [5] introduced a model that combines a convolutional neural network (CNN)-based neural network structure, specifically ResNet, with a transformer structure [6] to utilize data collected from camera, LiDAR, radar, and GPS sensors for beam prediction. The transformer blocks are used after each ResNet block to learn the correlations between different modalities and generate fused latent features for the next level of abstraction. This approach represents an advanced integration of diverse sensor data, aiming to enhance the accuracy and reliability of beam prediction in complex environments.

In this paper, we propose a new beam prediction model that builds upon and improves the state-of-the-art (SOTA) model [5] through two approaches. First, we enhanced the feature extraction performance by incorporating a squeeze-and-excitation (SE) network [7] into the ResNet structure. Second, we propose a transformer architecture called PIformer, which uses pooling layers [8] and Inception Mixer modules [9] instead of the self-attention mechanism used in traditional transformers [10] and vision transformers (ViT) [11]. With this structure, the proposed model can efficiently extract both low-frequency and high-frequency features from the data. In experimental results using the DeepSense 6G dataset, the proposed model demonstrated superior performance compared to the SOTA model in both Top-K beam prediction accuracy and distance-based accuracy (DBA) score [1]. Furthermore, it was observed that our model has lower complexity than the SOTA model, as evidenced by reductions in the number of parameters as well as training and testing times. This indicates that the model not only enhances prediction accuracy but also improves efficiency, making it a more practical option for real-world applications.

1229-2370/25/$10.00 © 2025 KICS

II. System Model and Problem Formulation

A. System Model

We consider a wireless communication system where a BS with N antennas and a user equipment (UE) with single antenna. The system adopts a pre-defined beamforming codebook $\mathcal{F} = \{\mathbf{f}_m\}_{m=1}^{M}$, where $\mathbf{f}_m \in \mathbb{C}^{N\times 1}$ and M is the number of beamforming vectors in the codebook. If x is the complex symbol transmitted by the BS using a beamforming vector $\mathbf{f} \in \mathcal{F}$, the received signal by UE in the downlink is given by

$$y = \mathbf{h}^H \mathbf{f} x + n, \tag{1}$$

where $\mathbf{h} \in \mathbb{C}^{N\times 1}$ is the complex channel vector, $(\cdot)^H$ denotes the Hermitian transpose operation, and $n \sim \mathcal{N}_c\left(0, \sigma^2\right)$ represents complex normally distributed noise.

B. Problem Formulation

The task of beam prediction can be defined as determining the optimal beamforming vector $\mathbf{f}^*$ out of candidate beams in the codebook $\mathcal{F}$, such that the received signal power is maximized. Mathematically, we can express the beam selection problem as

$$\mathbf{f}^* = \underset{\mathbf{f}\in\mathcal{F}}{\operatorname{argmax}} \|\mathbf{h}^H \mathbf{f}\|_2^2, \tag{2}$$

where $\mathbf{h}$ is the complex channel vector, $(\cdot)^H$ denotes the Hermitian transpose operation, and $\mathbf{f}$ is a beamforming vector.

In this paper, instead of relying on the explicit channel information, which is hard to acquire, we target designing a model to predict the optimal beam based solely on the sensor data. The model is developed to learn a prediction function f_Θ that utilizes the available sensor data to predict a candidate beam $\hat{\mathbf{f}} \in \mathcal{F}$ that maximizes the received power. Then, the well-designed prediction function can be formally written as

$$f_\Theta^* = \underset{f_\Theta}{\operatorname{argmax}}\, \mathbb{P}(\hat{\mathbf{f}} = \mathbf{f}^* | \mathcal{S}), \tag{3}$$

where $\mathcal{S}$ is the sensor data collected by the system and $\mathbf{f}^*$ is the optimal beamforming vector defined by (2).

III. Proposed Multi-Modal Beam Prediction Model

A. Data Preprocessing

We used the DeepSense 6G dataset, which includes multi-sensor data from diverse environments and wireless communication scenarios such as V2I, V2V, and drone communications, to evaluate the performance of sensing-assisted beam prediction under real-world conditions. In this paper, we utilized scenarios 32 and 33, which contain data collected using four sensors including RGB camera, LiDAR, radar and GPS in a V2I setting between a vehicle and a base station during daytime and nighttime, respectively. Approximately 7,000 data samples were used for training and evaluation, which is considered sufficient to ensure statistical reliability and stable model performance. Several preprocessing steps were applied to prepare the data for training the beam prediction model. We first downsampled a 64×1 power vector to 32×1. A 16-element antenna array operating in the 60 GHz frequency band received the transmitted signal. The BS utilized an oversampled codebook of 64 pre-defined beams. Therefore, even with downsampling, it does not affect the overall coverage area of the beams. Subsequently, from the downsampled power vector, we selected the beam index with the maximum received power to obtain the updated optimal beam for a specific sample.

1) Camera data: The dataset consists of image data collected during both daytime and nighttime. To ensure stable performance across various environmental conditions, this study utilizes four sensor modalities including camera, LiDAR, radar, and GPS. Since LiDAR, radar, and GPS data are unaffected by brightness variations, they help maintain overall model stability. However, image data may have lower visibility in dark environments, making it more challenging to identify vehicle positions, backgrounds, and objects compared to daytime images. Therefore, a preprocessing step was applied to adjust the brightness of nighttime images to further enhance model robustness and improve prediction accuracy. For illustration, Fig. 1 presents an original daytime image, an original nighttime image, and an enhanced nighttime image.

2) LiDAR data: The LiDAR sensor generates an average of over 16,000 3D points per scan, presenting challenges in training due to the extensive data size and computational costs. To address this, two preprocessing methods were applied to reduce the size of the point cloud data and increase model training speed. Firstly, to focus on the data points of moving vehicles, we removed points corresponding to backgrounds and buildings. This step eliminates data points related to fixed buildings and objects that could block the line-of-sight (LoS) between the BS and UE, thereby reducing complexity and bias during model training. The second method involved field-of-view (FoV) calibration. This process involves converting the LiDAR data into images and then projecting and trimming the BS's bird's-eye-view (BEV). This step effectively removes irrelevant data, enabling the CNN-based model to better focus on the useful information within the images. As a result, the transformer blocks can learn the relationships between images more effectively.

3) Radar data: The radar sensor provides information on the distance, angle, and speed of moving objects. Notably, it offers reliable speed information of vehicles that cannot be obtained from camera or LiDAR, independent of weather conditions or brightness levels. The preprocessing of the radar data is performed in two stages, as in [12]. In the first stage, 2D fast Fourier transform (FFT) is used to obtain chirp signals in the frequency domain. This process measures the frequency shift of signals transmitted and reflected back, which is then used to calculate the distance to the target. In the second stage, the velocity FFT is additionally applied to obtain speed information, and concurrently, the angle FFT is implemented to acquire angular information. Given the radar

(a) (b) (c)

Fig. 1. An example of preprocessing for camera data: (a) an original daytime image; (b) an original nighttime image; (c) an enhanced nighttime image.

TABLE I
NOTATION FOR RADAR DATA.

Notation	Description	Notation	Description
A	# of radar RX antennas	Ψ^{P}	Preprocessing function
S	# of samples per chirp	$\mathbf{H}_{\mathrm{RA}}$	Range-angle maps
C	# of chirps per frame	$\mathbf{H}_{\mathrm{RV}}$	Range-velocity maps

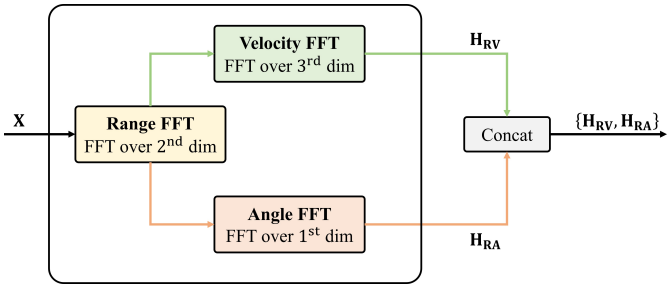


Fig. 2. The radar processing procedure.

data $\mathbf{X} \in \mathbb{C}^{A\times S\times C}$, the velocity FFT can be written as

$$\mathbf{H}_{\mathrm{RV}} = \Psi^{\mathrm{P}}_{\mathrm{RV}}\left(\mathbf{X}\right) = \sum_{a=1}^{A} \left|\mathcal{F}_{\mathrm{2D}}\left(\mathbf{X}_{\mathrm{a},:,:}\right)\right|. \tag{4}$$

The angle FFT can be written as

$$\mathbf{H}_{\mathrm{RA}} = \Psi^{\mathrm{P}}_{\mathrm{RA}}\left(\mathbf{X}\right) = \sum_{c=1}^{C} \left|\mathcal{F}_{\mathrm{2D}}\left(\mathbf{X}_{:,:,\mathrm{c}}\right)\right|. \tag{5}$$

The final preprocessing output of radar signals is generated in the concatenated form of range-angle maps and range-velocity maps as follows.

$$\mathbf{Y}_{\mathrm{Radar}} = \{\mathbf{H}_{\mathrm{RA}}, \mathbf{H}_{\mathrm{RV}}\}. \tag{6}$$

For ease of exposition, we summarize the notation for the radar data in Table I.

4) GPS data: The GPS data provides location information corresponding to the latitude and longitude of both the BS and UE. For selecting the optimal beam, relative positional information, rather than absolute locations, is crucial. Therefore, each angle is calculated using the relative positions between the BS and UE. To facilitate this, the magnitude of the relative positional information was first normalized. Additionally, since the correspondence between the angle information and beam indices varies for each scenario, a calibration was performed to adjust this relationship.

B. Proposed Model Architecture

The proposed beam prediction model extracts features from camera, LiDAR, and radar data by utilizing the ResNet-SE blocks, which are designed by adding SE network to the ResNet structure after preprocessing. Specifically, ResNet34 was used for camera data, while ResNet18 was used for LiDAR and radar data. By repeatedly passing through ResNet-SE blocks of varying sizes, more advanced features are extracted.

As data passes through each ResNet-SE block, it is processed by a newly designed block named PIformer that learns the relationships between the data. This process is performed sequentially four times, with the input size of the PIformer block gradually increasing to 64, 128, 256, and 512, respectively. Unlike existing self-attention-based transformers, PIformer utilizes pooling layers and Inception mixers to improve both performance and complexity.

After passing the final ResNet-SE block, the data is processed through a pooling layer and a reshape layer to transform the $5B \times 512 \times 8 \times 8$ pixel input into a $B \times 5 \times 512$ vector. Here, B is the batch size. The three output vectors and the $B \times 2 \times 512$ vector corresponding to the GPS data are concatenated in the channel axis direction and then summed. The model then predicts the optimal beam index using a multilayer perceptron (MLP) consisting of three layers with 256, 128, and 64 nodes respectively. The overall structure of the proposed model is shown in Fig. 3.

1) ResNet-SE block: Before applying the preprocessed camera, LiDAR, and radar data of size $B \times C \times 64 \times 64$ to the ResNet-SE block, they are transformed to $5B \times C \times 64 \times 64$ through a stack and reshape layer. Here, the batch size B is set to 8, and channel sizes C for the camera, LiDAR, and radar data are 3, 1, and 2, respectively. The resized data goes through convolutional layer, batch normalization, ReLU activation function, and max-pooling layer sequentially, just like the original ResNet structure, resulting in a size change to $5B \times 64 \times 64 \times 64$.

The camera data is applied to the ResNet34-SE structure, while LiDAR and radar data are applied to the ResNet18-SE structure to extract features. The ResNet18-SE comprises four blocks, each containing the same structure repeated twice, whereas the ResNet34-SE consists of blocks with the same structures repeated 3, 4, 6, and 3 times, respectively, as shown in Table II. Both ResNet18-SE and ResNet34-SE were used for feature extraction, excluding the final prediction.

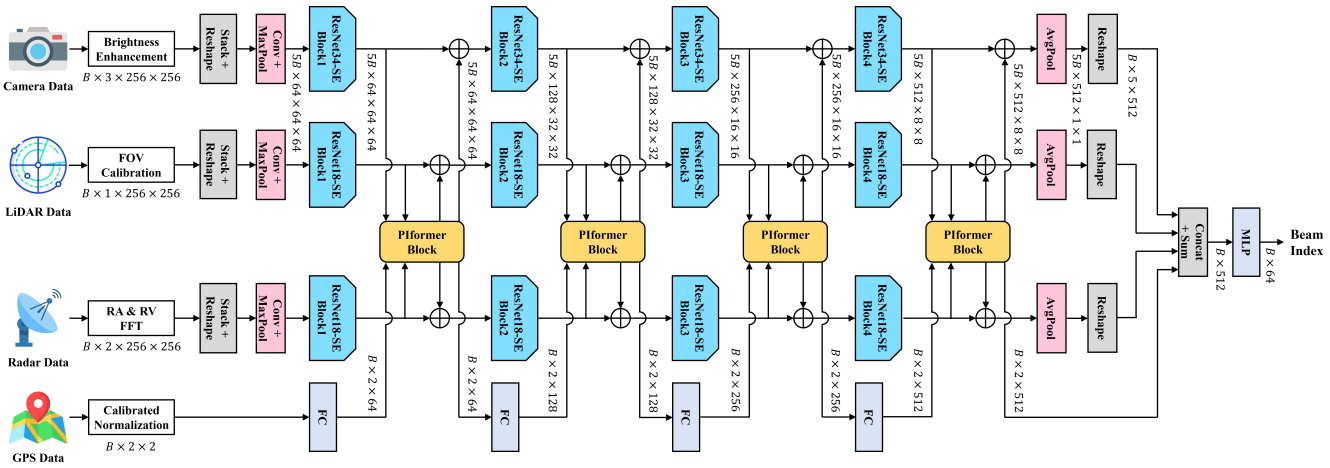


Fig. 3. The architecture of the proposed beam prediction model.

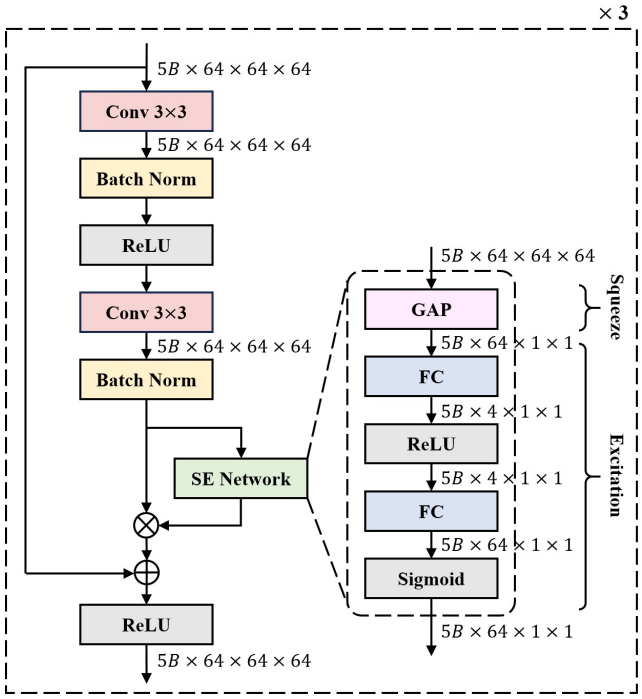


Fig. 4. The structure of ResNet34-SE block1 for camera data.

Here, the ResNet-SE block is an enhancement to the original ResNet structure, incorporating an SE network in the middle, as shown in Fig. 4. The SE network focuses on inter-channel interactions to improve image classification performance. It computes and reallocates channel weights through a squeeze stage, which summarizes the entire information of the feature map, and an excitation stage, which scales the significance of each summarized feature map.

2) PIformer block: In this paper, we propose PIformer, composed of 8 subblocks, to capture global features considering the interactions between each modality when extracting features from multi-modal data. Unlike traditional transformer models that use self-attention, the proposed PIformer utilizes the pooling layers [8] structure in the first and second subblocks, and the inception mixer [9] structure in the subsequent 6 subblocks. Replacing the self-attention layer in the traditional transformer architecture with pooling layers allows for more efficient feature extraction with significantly fewer parameters, as pooling does not require learnable weights. The inception mixer overcomes the limitations of traditional transformers in extracting high-frequency features by separating and capturing high- and low-frequency components within the image data. It consists of subblocks for high-frequency extraction, subblocks for low-frequency extraction, and fusion subblocks for integration. Additionally, an attention layer combining depth-wise convolution (DwConv) and multi-head self-attention (MSA) is used for effective feature extraction. This PIformer design enhances feature representation while maintaining high computational efficiency, reducing the number of parameters compared to conventional transformer-based architectures.

TABLE II
CONFIGURATION OF RESNET18-SE AND RESNET34-SE BLOCKS.

Type	ResNet18-SE block	ResNet34-SE block
Subblock 1	$\begin{bmatrix} 3\times 3, 64 \\ 3\times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times 3, 64 \\ 3\times 3, 64 \end{bmatrix} \times 3$
Subblock 2	$\begin{bmatrix} 3\times 3, 128 \\ 3\times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times 3, 128 \\ 3\times 3, 128 \end{bmatrix} \times 4$
Subblock 3	$\begin{bmatrix} 3\times 3, 256 \\ 3\times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times 3, 256 \\ 3\times 3, 256 \end{bmatrix} \times 6$
Subblock 4	$\begin{bmatrix} 3\times 3, 512 \\ 3\times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times 3, 512 \\ 3\times 3, 512 \end{bmatrix} \times 3$

For camera, LiDAR, and radar data, the feature maps of size $5B \times 64 \times 64 \times 64$ outputted from the ResNet-SE block are transformed to $B \times 5 \times 64 \times 8 \times 8$ through average-pooling and reshape layer. Three feature maps for camera, LiDAR, and radar data are combined into one feature map of size $B \times 15 \times 64 \times 8 \times 8$ through a concatenation layer, and reshaped to $B \times 960 \times 64$. The GPS data is then concatenated to create a feature map of size $B \times 962 \times 64$. These feature maps of the four types of data are added to positional embedding to provide

sequence information and pass through a dropout layer to prevent overfitting and enhance the generalization performance of the model, before being delivered to the PIformer.

The output of the PIformer is passed through a layer normalization, and then the data is split in the order it was concatenated during the embedding process, and reshaped to match the input size. To ensure that the input size remains the same, reshape and interpolation (Interp) layers were applied. The Interp layer, utilized for image resizing, performs interpolation based on surrounding pixel values to compute new pixel values. In this study, bilinear interpolation was employed, which uses the values of the surrounding four pixels to calculate the new pixel value. It retrieves the surrounding pixel values from the image at the given coordinates and uses them to compute the value of the new coordinates. Additionally, a scale factor, which is the ratio used to resize the data, is adjusted to 8, 4, 2, and 1 for each block, resulting in an image size increase of 8, 4, 2, and 1 times, respectively. Subsequently, the outputs of the PIformer are added to the feature map used as input and passed to the next ResNet-SE block. The entire structure of the PIformer block proposed in this paper is shown in Fig. 5.

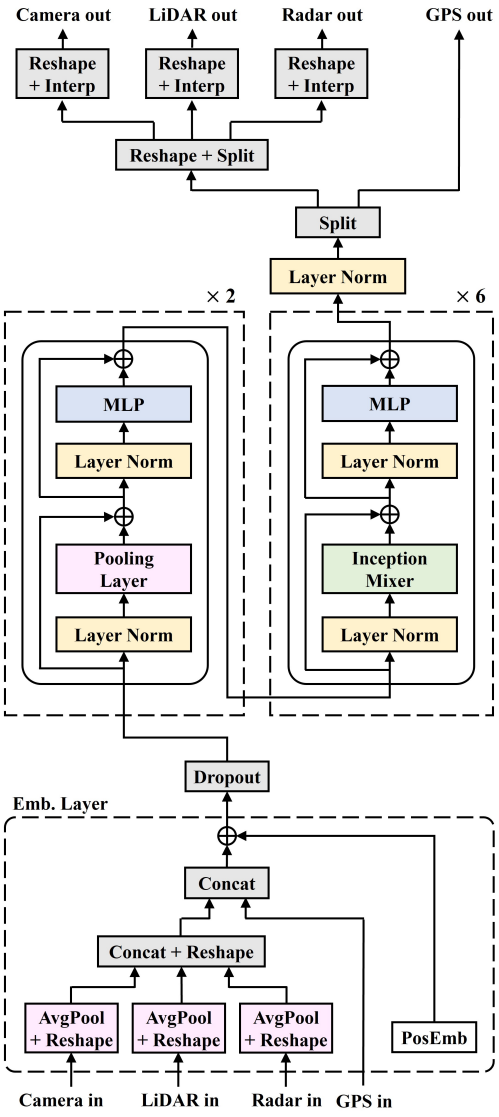


Fig. 5. The structure of PIformer block.

IV. Experiments and Results

A. Training and Optimization

We applied label smoothing to one-hot encoding to enhance the Top-K beam prediction performance. For a one-hot vector of 64 beams, the optimal beam is placed at the peak of a Gaussian distribution, and values corresponding to the Gaussian distribution are set for five beam positions on each side of this center, while the rest are set to zero. Furthermore, the dataset has an issue of data imbalance, as the proportion of beams in the labels is not uniform. To address this, we applied focal loss [13]. The focal loss is defined as the standard cross entropy criterion multiplied by a factor $(1-p_t)^{\gamma}$ as follows:

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t), \tag{7}$$

where p_t represents the probability of the ground-truth class, and γ is an adjustable focusing parameter. Setting γ greater than 0 reduces the relative loss for well-classified examples (where p_t is greater than 0.5), thus placing more emphasis on hard, misclassified examples. In this paper, we set γ to 2.

We trained the proposed model for 150 epochs with a batch size of 8 using the AdamW optimizer. When training the proposed model using the learning rate scheduler from [5], it was observed that the loss did not converge as shown in Fig. 7. To address this issue, a new scheduler was applied. In order to ensure stable convergence of the learning and to prevent overfitting, we utilized exponential moving average (EMA). Additionally, to derive the optimal convergence value, we utilized the whale shaped learning rate designed by extracting only one cycle from the cosine annealing warmup restart [14]. This method involves increasing the learning rate from 0 to 10^{-3} over 20 epochs, followed by a gradual decrease to 0 over the next 130 epochs. The learning rate schedulers used in [5] and in this paper are shown in Fig. 8.

B. Performance Metrics and Results

For performance evaluation, we chose the prediction accuracy for the Top-K beams and the DBA-score as performance metrics, demonstrating the superiority of our proposed model compared to the SOTA model. Additionally, to verify the contribution of each proposed approach to performance improvement, we conducted an ablation study.

We visualized the relationship between the beams predicted by the model and the actual ground truth beams using a confusion matrix and heat map, as illustrated in Fig. 9. The main diagonal represents correctly classified predictions, while the off-diagonal elements represent misclassified predictions. All four models exhibit a distribution predominantly along the main diagonal, indicating accurate predictions for the majority of the 64 beams. Additionally, the sparse distribution of values for beam indices above 50 indicates class imbalance. This observation suggests that improving class balance could further enhance the model's performance. Consideration of these factors could lead to performance improvements when

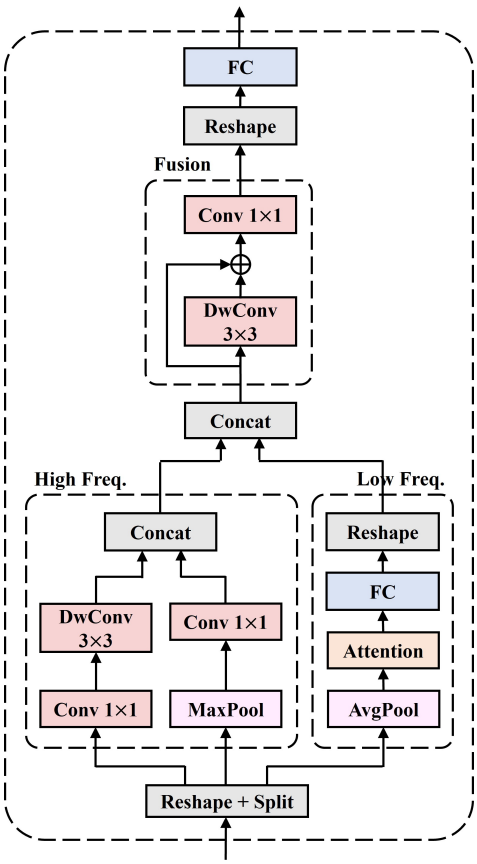


Fig. 6. The structure of inception mixer.

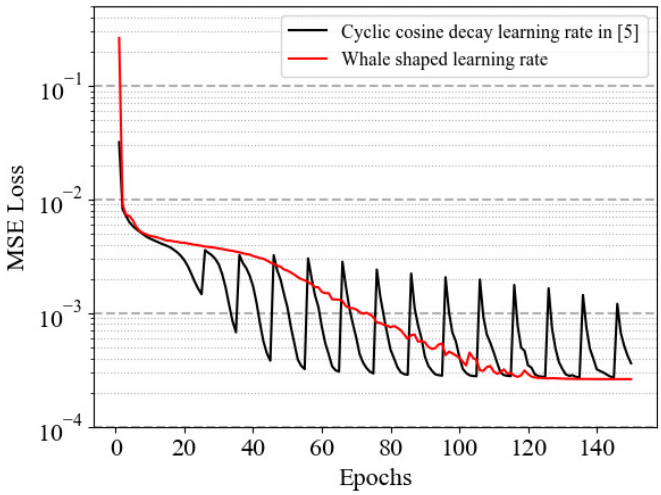


Fig. 7. The training MSE loss versus epochs.

refining the model.

Top-K accuracy refers to the proportion of instances where the actual optimal beam is included among the Top-K candidate beams predicted by the model. Top-K accuracy of the model can be written as

$$\text{Top-}K \text{ accuracy} = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K}\mathbf{1}_{\{\hat{m}_{l,k}^{*}=m_{l}^{*}\}}, \tag{8}$$

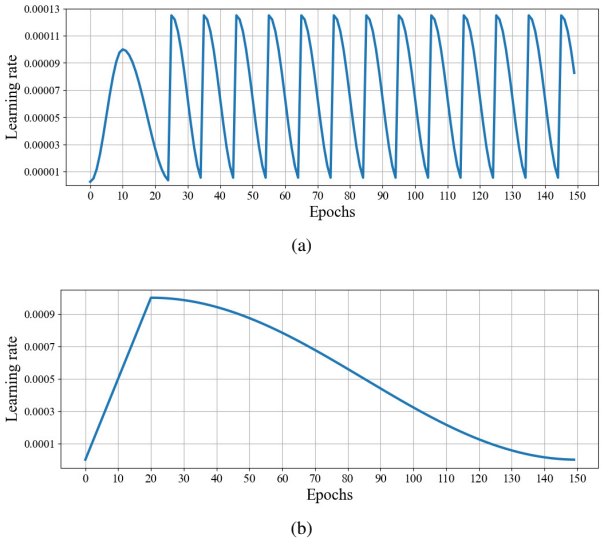


Fig. 8. Learning rate schedulers: (a) Cyclic cosine decay learning rate in [5]; (b) Whale shaped learning rate.

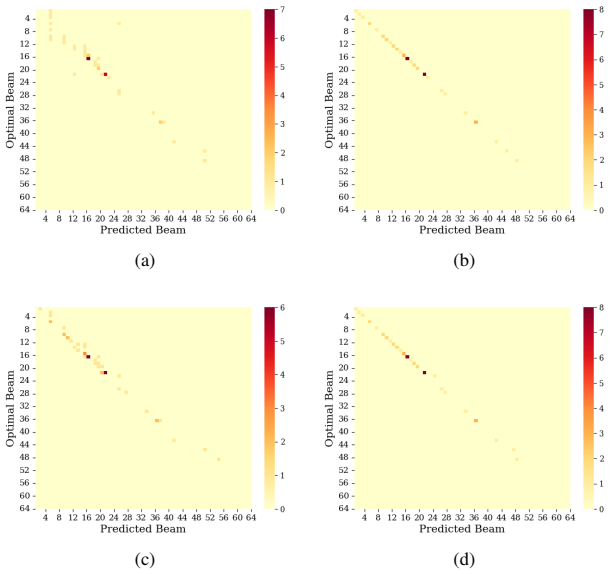


Fig. 9. The confusion matrices for predicted and optimal beams: (a) Reference [5]; (b) Proposed; (c) Only ResNet-SE; (d) Only PIformer.

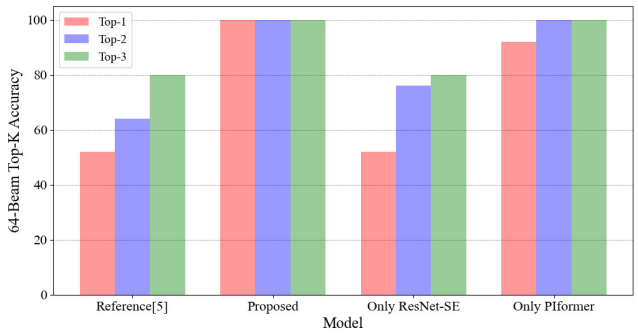
where L is the total number of samples in the test dataset, m^* is index of the optimal beam, $1_{\{\cdot\}}$is a function that returns one if satisfies the condition $\{\cdot\}$.

We conducted performance evaluations for scenarios 32 and 33, considering cases where $K = 1, 2, 3$. As shown in Table III, the accuracy of our proposed model surpassed that of the SOTA model in both scenarios. Additionally, through an ablation study, we were able to ascertain that the two improved structures in our proposed model, the ResNet-SE and PIformer blocks, each have a significant impact on the high beam prediction accuracy of the proposed model. Furthermore, it was observed that using these two structures together yields even greater effectiveness. In Fig. 10, we present the Top-K accuracy of the four models for scenarios 32 and 33.

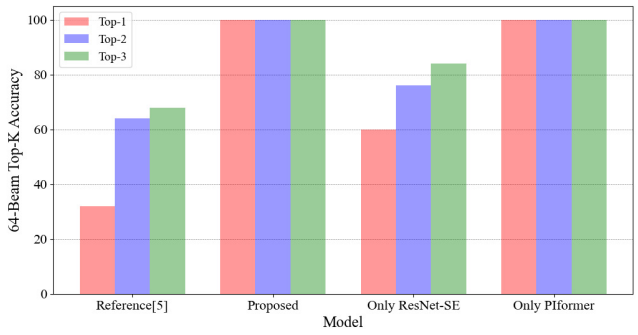
Another evaluation metric for Top-K beam prediction, the

TABLE III
COMPARISON OF TOP-K BEAM PREDICTION ACCURACY AMONG REFERENCE [5], PROPOSED, ONLY RESNET-SE, AND ONLY PIFORMER.

Model	Scenario 32			Scenario 33		
	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
Reference [5]	0.52	0.64	0.80	0.32	0.64	0.68
Proposed	1.00	1.00	1.00	1.00	1.00	1.00
Only ResNet-SE	0.56	0.76	0.80	0.60	0.76	0.84
Only PIformer	0.92	1.00	1.00	1.00	1.00	1.00



(a)



(b)

Fig. 10. Top-K Beam Accuracy: (a) Scenarios 32; (b) Scenarios 33.

DBA-score, is introduced to evaluate the distance between the predicted beams and the actual optimal beam. The DBA-score is a performance evaluation metric used in competitions to assess beam prediction accuracy. It is defined as follows:

$$\text{DBA-score} = \frac{1}{K}\sum_{k=1}^{K} Y_k. \tag{9}$$

Here, Y_k is defined as

$$Y_k = 1 - \frac{1}{L}\sum_{l=1}^{L} \min_{1\le k'\le k}\left[\min\left(\frac{\left|\hat{y}_{l,k'} - y_l\right|}{\Delta}, 1\right)\right]. \tag{10}$$

where y_l and $\hat{y}_{l,k'}$ are ground-truth beam index and the predicted beam index of sample l, respectively. Δ is used as a normalization factor, and we set $\Delta = 5$. As shown in Table IV, the DBA-score of our proposed model surpassed that of the SOTA model in both scenarios 32 and 33. The results of the ablation study also confirmed the effectiveness

TABLE IV
COMPARISON OF TOP-3 BEAM DBA-SCORES AMONG REFERENCE [5], PROPOSED, ONLY RESNET-SE, AND ONLY PIFORMER.

Model	Scenario 32	Scenario 33	Average
Reference [5]	0.83	0.81	0.82
Proposed	1.00	1.00	1.00
Only ResNet-SE	0.88	0.90	0.89
Only PIformer	0.99	1.00	0.99

TABLE V
COMPARISON OF MODEL COMPLEXITY AMONG REFERENCE [5], PROPOSED, ONLY RESNET-SE, AND ONLY PIFORMER.

Model	Parameters	Training Time	Test Time
Reference [5]	78,422,528	0.0381 sec	0.0363 sec
Proposed	74,472,152	0.0237 sec	0.0302 sec
Only ResNet-SE	78,753,792	0.0410 sec	0.0361 sec
Only PIformer	74,140,888	0.0230 sec	0.0359 sec

of the two improved structures in our proposed model, the ResNet-SE and PIformer blocks, in terms of the DBA-score.

As confirmed through the two performance metrics previously, our proposed model demonstrated superior performance to the SOTA model in both Top-3 prediction accuracy and DBA-score. To further assess the computational efficiency of our model, we compared its complexity with that of the SOTA model by analyzing the number of trainable parameters, training time, and testing time. Table V presents the results of this comparison. Our proposed model achieves a 5% reduction in the number of parameters while also decreasing training and testing times by 38% and 1.7%, respectively. Notably, we observed a significant decrease in complexity when replacing the transformer block in the existing model with our proposed PIformer structure.

For an additional comparison, we conducted additional experiments not only with [5] but also with [4], as summarized in Table VI. Given that [4] is a model that uses camera and GPS data, the proposed model was also tested using only camera and GPS data as input. The experimental results show that the proposed model achieved higher performance than [4] under the same input conditions. However, the performance was slightly lower than when using four types of sensor data, as shown in Table III. This indicates that utilizing four types of sensor data provides richer information, enabling the model to better learn complex beam patterns and improve prediction accuracy.

V. CONCLUSION

In this paper, we proposed a new deep learning model that predicts the optimal beam using multi-modal sensor data instead of in-band wireless signals. To improve both prediction accuracy and model complexity, we enhanced the traditional

TABLE VI
COMPARISON OF TOP-K BEAM PREDICTION ACCURACY USING IMAGE AND GPS DATA BETWEEN REFERENCE [4] AND PROPOSED.

Model	Scenario 32			Scenario 33		
	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
Reference [4]	0.48	0.68	0.76	0.32	0.60	0.72
Proposed	0.88	0.96	0.96	0.88	0.92	1.00

ResNet structure by adding an SE network and redesigned the transformer structure by combining pooling layers and Inception mixers. Experimental results show that our proposed model increased the DBA-score by about 22% compared to the SOTA model, and reduced the training time by 38% and testing time by 1.7%.

The proposed model demonstrated improved computational efficiency, enabling real-time inference and stable performance in large-scale environments. Additionally, the model can be flexibly adjusted depending on the combination of available sensor modalities, allowing simplified or customized versions to be constructed for different deployment scenarios. To enhance the model's generalization and robustness, future research will consider additional scenarios from the DeepSense 6G dataset, which reflects diverse real-world conditions and complex wireless environments.

REFERENCES

[1] G. Charan *et al.*, "Multi-Modal Beam Prediction Challenge 2022: Towards Generalization," 2022, *arXiv:2209.07519*.

[2] J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "DeepSense 6G: A Large-Scale Real-World Multi-Modal Sensing and Communication Dataset," *IEEE Comm. Mag.*, vol. 61, no. 9, pp. 122–128, 2023.

[3] J. Morais, A. Behboodi, H. Pezeshki, and A. Alkhateeb, "Position Aided Beam Prediction in the Real World: How Useful GPS Locations Actually Are?," 2022, *arXiv:2205.09054*.

[4] G. Charan, T. Osman, A. Hredzak, N. Thawdar, and A. Alkhateeb, "Vision-Position Multi-Modal Beam Prediction using Real Millimeter Wave Datasets," in *Proc. IEEE WCNC*, 2022.

[5] Y. Tian *et al.*, "Multimodal Transformers for Wireless Communications: A Case Study in Beam Prediction," 2023, *arXiv:2309.11811*.

[6] K. Chitta *et al.*, "TransFuser: Imitation With Transformer-Based Sensor Fusion for Autonomous Driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12878–12895, 2022.

[7] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE/CVF CVPR*, 2018.

[8] W. Yu *et al.*, "Metaformer is Actually What You Need for Vision," in *Proc. IEEE/CVF CVPR*, 2022.

[9] C. Si *et al.*, "Inception Transformer," in *Proc. NeurIPS*, 2022.

[10] A. Vaswani *et al.*, "Attention Is All You Need," in *Proc. NeurIPS*, 2017.

[11] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. ICLR*, 2021.

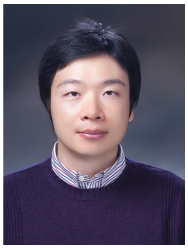
[12] U. Demirhan and A. Alkhateeb, "Radar Aided 6G Beam Prediction: Deep Learning Algorithms and Real-World Demonstration," in *Proc. IEEE WCNC*, 2022.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2018.

[14] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Restarts," 2016, *arXiv:1608.03983*.

Yerin Yeo received the B.S. degree in Bigdata Engineering from Soonchunhyang University, Asan, South Korea, in 2023, and the M.S. degree in Artificial Intelligence from Sejong University, Seoul, South Korea, in 2025. She is currently a Ph.D. student in the Department of Convergence Engineering for Artificial Intelligence at Sejong University. Her research interests include intelligent wireless communication systems, multi-modal learning, and semantic communications. She received the Best Paper Award at the KICS Winter Conference in 2023 and 2024.

Junghyun Kim received the B.S., M.S., and Ph.D. degrees in the Department of Electrical and Electronic Engineering from Yonsei University, South Korea, in 2006, 2008, and 2017, respectively. From 2010 to 2013, he was an Engineer with Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. From 2017 to 2019, he was a Senior Engineer with Samsung Research, Seoul, South Korea. He is currently an Assistant Professor with the Department of Artificial Intelligence and Data Science, Sejong University, Seoul, South Korea. Prior to joining Sejong University, he was a Faculty Member with Soonchunhyang University, Asan, South Korea, from 2019 to 2022. His research interests include intelligent wireless communication systems, post-quantum cryptography, drug discovery and development, and graph theory.



Jihyung Kim received the Ph.D. degree from the Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Republic of Korea, in 2007. Since 2007, he has been with the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea, where he is currently working as a Principle Researcher. His research interests include AI and RIS for 5G/6G mobile communication.

Junhwan Lee received a Ph.D. degree in Information & Computer Science from Keio University, Yokohama, Japan in 2009. He was a project research associate at Keio University from 2005 to 2006. He has been working at ETRI, Daejeon, Korea since 2000; 3GPP standardization more than 10 years. Currently, he is a Section Director of the Spatial Wireless Transmission Research Section of the Satellite Communication Research Division, where he covers aerial communications like UAS, UAM, and up to LEO satellite communications. He also serves as the Deputy Team Leader of Strategy & Global Collaboration of the 6G Forum. Through his global activities, he has participated in international collaborative projects such as KR-EU, KR-JP, KR-CN, KR-FI. With respect to research topics in digital and wireless communication systems, his key areas of interest particularly include physical layer design issues in LEO satellite communications. Additionally, he continues to make efforts to establish global joint projects with neighboring countries to promote and advance 5G/6G R&D activities.