Analysis of Multi-Modal Beam Prediction under Distribution Shift

Maximilian Arnold*†, Gouranga Charan‡, Umut Demirhan‡, Ahmed Alkhateeb‡, Mohammed Alloulah†
†Bell Labs
‡Wireless Intelligence Lab, Arizona State University

***Abstract*—Directional communications in millimetre-wave and sub-terahertz bands require dynamic beam management in order to track mobile users. To achieve fast and scalable beam management, directional communication systems in these bands may leverage additional sensing modalities for *learning* high-fidelity spatial state information beyond what can be gleaned from radio alone. However, building generalisable multi-modal beam management remains an open challenge, especially under mobility pattern and environmental distribution shifts.**

This paper presents BeamBench, a multi-modal beam prediction benchmarking study as formulated and facilitated by [1]. BeamBench builds over 100 different learning configurations for beam management, spanning representation learning and sensing modalities, as well as a classical baseline. Importantly, BeamBench characterises beam management performance on unseen data not encountered during training, which is crucial for real-world systems. We find that multi-modal learning outperforms classical estimation. Further BeamBench configurations exhibiting qualitative and quantitative differences, particularly w.r.t. robustness. We conclude with practical tips and guidelines for robust and generalisable multi-modal beam prediction.

I. Introduction

Millimetre-wave (mmWave) and sub-terahertz (THz) bands offer much needed spectrum to meet increased user demands for wireless communication. However, these bands require directional communications in order to compensate for their high propagation losses [2]. As such, suboptimal beam management becomes the limiting factor for harnessing mmWave and sub-THz bands [3]. Typically, the optimal beam between two communicating nodes is found through exhaustive search over the beamforming codebook—a process known as beam sweep. Currently standardised beam sweeps necessitate large beam training overhead. This excessive overhead is problematic, especially under high-mobility scenarios and/or latency-critical applications. As a result, state-of-the-art research proposes to use spatial information, obtained from other sensing modalities orthogonal to radio, to find the best beam setup without the radio search overhead [4]–[7]. Recently, Charan et al. have articulated this proposition, and provided the communication community with a dataset, a baseline, and a metric for advancing multi-modal beam prediction research [1].

Gouranga Charan, Umut Demirhan, and Ahmed Alkhateeb, who were co-organizers of the DeepSense ML competition in [1], have contributed to this work after the completion of the competition.

In this paper, we tackle the multi-modal beam prediction challenge as stipulated in [1]. We further introduce BeamBench, a multi-modal beam prediction benchmark to study the performance nuances of a range of configurations spanning (i) representation learning methods and (ii) different sensing modalities. Concretely, BeamBench investigates the following research questions:

- How do configurations of sensing modalities and representation learning methods impact beam prediction performance?
- How do these *learnt* methods compare to a classical line of sight (LoS) baseline?
- How important are time series data to beam prediction?
- How prone are these *learnt* methods to data distribution shifts arising from changes in mobility patterns and environmental factors?

II. Related Work

The promise of vast spectrum in the mmWave and sub-THz bands is contingent on overcoming the challenge of robust and scalable beam management. A number of beam management methods has been reported in prior art. Direct methods use classical feedback optimisation [3], possibly with sparsity-based optimisation [2]. Learning-based methods use data-driven optimisation [8]. Beam management is essentially a prediction problem. As such, multiple sensory data can be leveraged for learning: (1) vision [4]–[7], (2) radar [9], [10], (3) lidar [11], [12], (4) GPS [13], [14], or (5) a combination thereof within a multi-modal formulation for enhanced robustness and accuracy [4], [12]. However, there is little prior art on the generalisability of these learning-based methods, which has motivated the challenge set forth in [1] and whose dataset and metric we use for BeamBench.

III. A primer on beam prediction

We begin by summarising the beam prediction problem statement. We then derive an analytic baseline, and review the beam prediction accuracy metric we use later in BeamBench.

A. Problem statement

We consider the experimental setup of DeepSense 6G [15]. An mmWave basestation is equipped with a suite of sensors (i.e., camera, lidar, and radar) and an M-element ULA

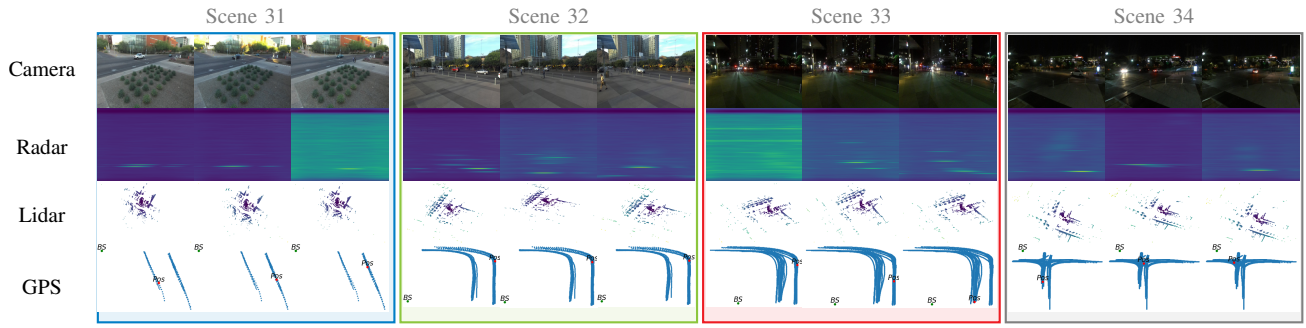

Fig. 1: Dataset examples for the four scenarios of DeepSense 6G beam prediction challenge [1].

antenna array. The user has a single-element antenna and transmits its current real-time location estimated from GPS. The basestation uses OFDM signalling with K subcarriers in order to serve a mobile user. The basestation has a beamforming codebook $\boldsymbol{\mathcal{F}} = \{\mathbf{f}_q\}_{q=1}^{Q}$, where $\mathbf{f}_q \in \mathbb{C}^{M\times 1}$ and Q is the total number of beamforming vectors. Let $\mathbf{h}_k[t] \in \mathbb{C}^{M\times 1}$ denote the channel between the basestation and the user at the k-th subcarrier and time t. The received signal at the basestation is

$$y_k[t] = \mathbf{h}_k^T[t]\mathbf{f}_{q[t]}x + v_k[t], \tag{1}$$

where $\mathbf{f}_{q[t]} \in \boldsymbol{\mathcal{F}}$ is the beamforming vector applied at time t and $v_k[t]$ is the receiver noise with a complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}(0,\sigma^2)$. Beam prediction finds the index $q^\star[t] \in \{1,\ldots,Q\}$ of the optimal beamforming vector $\mathbf{f}_{q^\star[t]}$, such that

$$q^\star[t] = \underset{q\in\{1,\ldots,Q\}}{\text{argmax}} \frac{1}{K}\sum_{k=1}^{K} |\mathbf{h}_k^T[t]\mathbf{f}_q|^2, \tag{2}$$

Typically in mmWave, accurate channel state information (a) is hard to estimate and (b) consumes large amount of communication resources. Alternatively, simple angle estimates α's of N strongest beams are substituted for channel state information. These angles are estimated by using an omnidirectional receiver, spatial sweeps of a directional transmitter, and maximising received power. However, such exhaustive sweeps have high-latency that would place limits ability to support highly-mobile users. Some mitigating techniques rely on additional sensory information to (a) detect a target in another modality, e.g., a visual image and (b) estimate the target's angle as a proxy to its beamforming vector. We note that these techniques either require LoS propagation conditions, or learn a scenario-specific optimisation that does not generalise to other spatially-different scenarios.

B. Analytic baseline

Fig. 2 illustrates the correspondence between a user's angle (left) and beam index (right), across space and relative to a basestation. Specifically, the colour map encodes the angle (in degrees) and beam index for all samples of the user

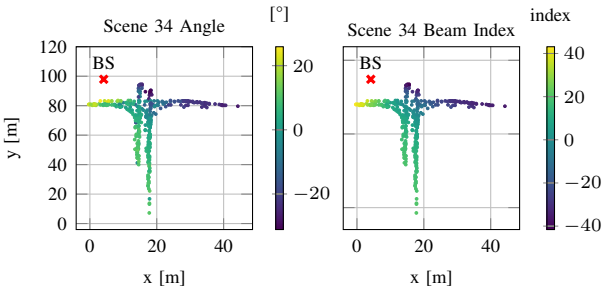

Fig. 2: Correspondence between calibrated GPS angles (left) and beam indices (right) for Scenario 34.

spatial trajectory. Note that we have centred the beam indices for a better colour correspondence with angles (i.e., beam index 31 corresponds in reality to 0 degree). Fig. 2 shows that under dominant LoS conditions, the angle translates to the beam index by scaling and shift only. We designate this analytic angle-beam mapping as our classical baseline. That said, the optimal beam search formulation of Eq. (2) would allow for reaching users via reflections under LoS occlusion and/or blockage. In other words, a truly general beam search would always require a certain amount of *radio* samples for optimisation.

Calibration. Fig. 2 demonstrates the analytic mapping between a user's angle and beam index. This analytic mapping is valid *iff* the angle can be centred w.r.t. beam index. DeepSense 6G does not include the rotation angle of the basestation. We thus rely on a calibration procedure, assuming that the camera and mmWave antenna array have overlapped focal centres. The calibration procedure is as follows:

1) identify the car with mounted GPS
2) find the GPS measurement at the centre pixel of the corresponding visual image
3) use the GPS position to calculate the angle of this centre position relative to the basestation
4) average the angle

This procedure centres mobile user measurements relative to the boresight of the basestation. For the four scenarios we

treat later in the paper, the centred user angles are: -0.72, -0.76, 0.59, and -0.51 radians, respectively for Scene 31, 32, 33, and 34.

Least squares fitting. The one-to-one correspondence between angle and beam index allows us to fit a least squares model to capture this mapping

$$\begin{aligned}\text{beam}_{\text{id}} = &-6.97\times10^{-5}\alpha^3 - 1.15\\ &\times10^{-3}\alpha^2 + 0.6885\alpha + 0.175,\end{aligned}$$

where we use a 3rd order polynomial. That is, simple regression is able to convert angles to beam indices according to the above model.

C. Performance metric

Charan et al. propose the distance-based accuracy (DBA) score for evaluating the performance of beam prediction [1]

$$\text{DBA} = \frac{1}{L}\sum_{\ell=1}^{L} B_\ell \;;$$

$$B_\ell := 1 - \frac{1}{N}\sum_{n=1}^{N} \min_{1\le\ell'\le\ell}\left[\min\left(\frac{\left|\hat{b}_{n,\ell'} - b_n\right|}{\Delta}, 1\right)\right] \quad (3)$$

where DBA averages L top predictions B_ℓ, b_n is the groundtruth beam index at sample n, and $\hat{b}_{n,\ell'}$ is its ℓ'th top predictor. See [1] for a justification of how DBA tracks power faithfully. In our evaluation, we use $L = 3$, i.e., top-3 accuracy score.

IV. Dataset

Sensory data. As depicted in Fig. 1, this challenge adopts Scenes 31-34 of the DeepSense 6G dataset to enable the study of the multi-modal beam prediction problem [15]. The sensor entries of Scenes 31-34 are recorded for each time step t in the trajectory of the mobile transmitter. A temporal snapshot of the following sensor entries is assumed to be available for prediction per time step t:

- 2× GPS position coordinates $\in \mathbb{R}^2$ converted from the transmitter longitude & latitude.
- 5× RGB images $\in \mathbb{R}^{W\times H\times 3}$, with width W and height H.
- 5× lidar point clouds $\in \mathbb{R}^{N_{\text{lidar}}\times 3}$, with N_{lidar} valid points per time step.
- 5× radar image $\in \mathbb{R}^{640\times 480}$, with 640 azimuth and 480 range bins.

Using only 2 GPS readings per time step t is meant to emulate sporadic user logs at the basestation.

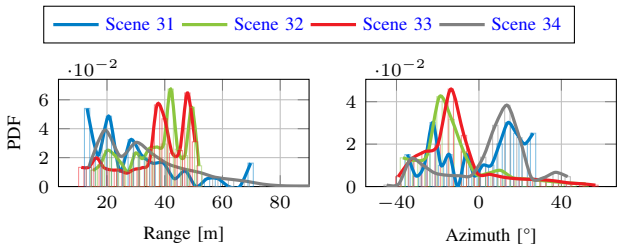

Fig. 3: User range and azimuth distributions across Scenes.

Distribution shift. Different user mobility patterns and surrounding environment across Scenes give rise to distribution shift in sensory data at the basestation. For example, Fig. 3 shows the range and azimuth histograms of user locations across Scenes 31-34. Note that user trajectories are converted from Cartesian to Polar coordinates as per the calibration procedure outlined in Sec. III-B. Fig. 3 shows that Scenes 32 and 33 have roughly similar user trajectories, while Scenario 31 is distinctly dissimilar. These distribution shifts allow us to gauge the generalisability of model configurations in BeamBench.

Multi-modal matching. We employ a matching procedure that allows us to associate multi-modal data to a common target label. Multi-modal matching acts as a qualitative pre-filtering of unreliable data in order to improve the training efficacy.

1) Image: Matching begins at the vision modality. Using Yolo v5 [16], [17], we estimate object bounding boxes present in an image. For each found object, we calculate an estimate of the object's angle α_{cam} w.r.t. its bounding box centre point $c_{\text{bb}} = (x_{\text{bb}}, y_{\text{bb}})$

$$\alpha_{\text{cam}} = \frac{|c_{\text{bb}} - (W/2, H/2)^T|}{\sqrt{W^2 + H^2}} \quad (4)$$

We then match α_{cam} to angles derived from GPS positions α_{gps}. A cross-modal object match is declared if $|\alpha_{\text{cam}} - \alpha_{\text{gps}}| < 10°$ in a given sensing scene. We devise this simple method for cross-modal matching because DeepSense 6G dataset does not include calibration information for the camera.

2) Radar: We threshold radar heatmaps using 2D constant false alarm rate (CFAR). We then apply density-based spatial clustering of applications with noise (DBSCAN) to obtain cluster centres. We calculate object angles α_{rad} and similarly match against α_{gps}.

3) Lidar: We calculate object angles α_{lid} similar to radar, and match against α_{gps}.

Scenario	Sensor	# Datapoints	Matchable points
Scene 34	Radar	4191	3226
Scene 34	Camera	4191	4153
Scene 34	Lidar	4191	3817
Scene 33	Radar	3862	2830
Scene 33	Camera	3862	3830
Scene 33	Lidar	3862	3820
Scene 32	Radar	3140	2169
Scene 32	Camera	3140	3132
Scene 32	Lidar	3140	3082
Scene 31	Radar	50	36
Scene 31	Camera	50	50
Scene 31	Lidar	50	49

TABLE I: Cross-modal matching of sensing modalities to GPS on DeepSense 6G data. Matching uses an angle threshold of 10°.

Tab. I lists the results of matching angular target estimates per sensing modality to GPS measurements, and across Scenes. Camera has the highest matchable data ratio. Lidar and radar come as second and third, respectively. This can be understood noting that camera maintains dense spatial detail throughout the field of view (i.e., range & angle). In comparison, lidar's and radar's coverage density of the field of view drops drastically after about 100 and 40 metres, respectively.

V. BeamBench

We build BeamBench, a suite of 10 beam prediction configurations, spanning sensing modalities and state-of-the-art learning architectures. The following reviews these configurations.

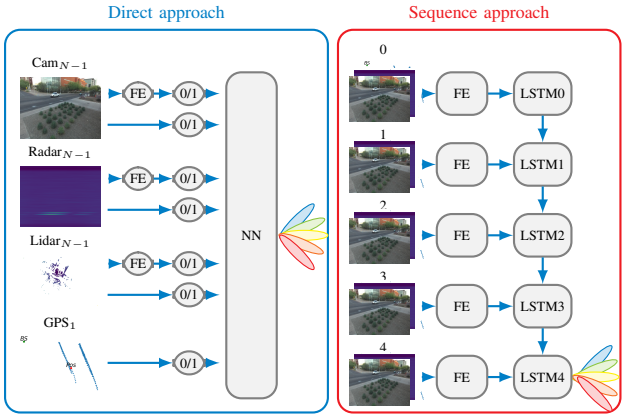

Fig. 4: BeamBench compiles a variety of neural configurations, where a direct approach (from the last sequence element) and an long-term short-term memory (LSTM) approach is configurable.

Fig. 4 illustrates two modelling approaches for beam prediction. The direct approach builds a model that ingests a snapshot of various combination of sensing modalities. Fig. 4 denotes model combinations as zero-one switches. A sensing modality can either be fed to the model as raw data or after feature extraction (FE). Alternatively, the sequence modelling approach captures the temporal dependencies of consecutive measurements for beam prediction. Sequential modelling uses a long short-term memory (LSTM) neural network. Sequential always encodes raw data with FE first in order to compress their high dimensionality, producing 1D encodings. One or more 1D encodings of multiple modalities can be stacked before feeding these to an LSTM.

Supervised learning. Training a beam predictor network uses a cross-entropy (CE) loss. CE effectively maps a high-dimensional input into a categorical selection of indices from the codebook $\mathcal{F} = \{\mathbf{f}_q\}_{q=1}^{Q}$ (cf., Sec. III-A). Concretely, this mapping can be denoted as $\mathbb{R}^{N_{\text{input}}} \to \mathbb{R}^{Q}$, where $\mathbb{R}^{Q}$ is sparse due to one-hot encoding and $\mathbb{R}^{N_{\text{input}}}$ varies with model configurations.

Unsupervised learning. Feature extraction uses unsupervised pretraining for any input modality x. BeamBench supports two widely used flavours.

First, an autoencoder (AE) learns an encoder function g and a decoder g^{-1}, such that the reconstruction error is minimised

$$\mathcal{L}_{\text{AE}} = \mathop{\mathbb{E}}_{x} \|x - g^{-1}(g(x))\|_2^2 \tag{5}$$

The encoder function g is typically designed to produce features $z = g(x)$ that are lower dimensional than x.

Second, contrastive learning (CL) produces features that preserve the semantic similarity of input data. Based on [18], we train two Siamese neural nets to obtain features that are dissimilar if the radio heatmaps are sufficiently so. Typically for vision, elaborate data augmentation schemes allow for preserving the semantic meaning of images in a straightforward manner. For heatmaps, we incorporate domain knowledge from radio to rank the similarity of heatmaps according to a cross-correlation factor that we utilise during CL. Let u and v_i be two raw heatmaps encoded by two nets f_θ and g_θ such that $q = f_\theta(u)$ and $k_i = g_\theta(v_i)$, where θ denoting weight parametrisation. With each u, use $K+1$ samples of v of which one sample v^+ is a true semantic match to v and K samples $\{v_i^-\}_{i=0}^{K-1}$ are false matches. The one-sided contrastive loss is [19]

$$\mathcal{L}_c^{v\to u} = -\mathop{\mathbb{E}}_{u,v} \log \left[\frac{\exp\left(q \cdot k^+/\tau\right)}{\exp\left(q \cdot k^+/\tau\right) + \sum_i \exp\left(q \cdot k_i^-/\tau\right)} \right] \tag{6}$$

where $\cdot$ is the dot product. The encodings $k^{+/-} = g_\theta(v^{+/-})$ correspond to true and false heatmaps. Further, vector $\mathbf{k}^- = \{k_i^-\}_{i=0}^{K-1}$ holds K false encodings, and τ is a temperature hyper-parameter. The bidirectional CL becomes $\mathcal{L}_c = (\mathcal{L}_c^{v\to u} + \mathcal{L}_c^{u\to v})/2$.

Once features are pretrained, supervised learning can be used on top in order to build the downstream beam predictor. We optimise all our network configurations with the neural network intelligence (NNI) AutoML tool [20]. We summarise our beam prediction configurations in Tab. II.

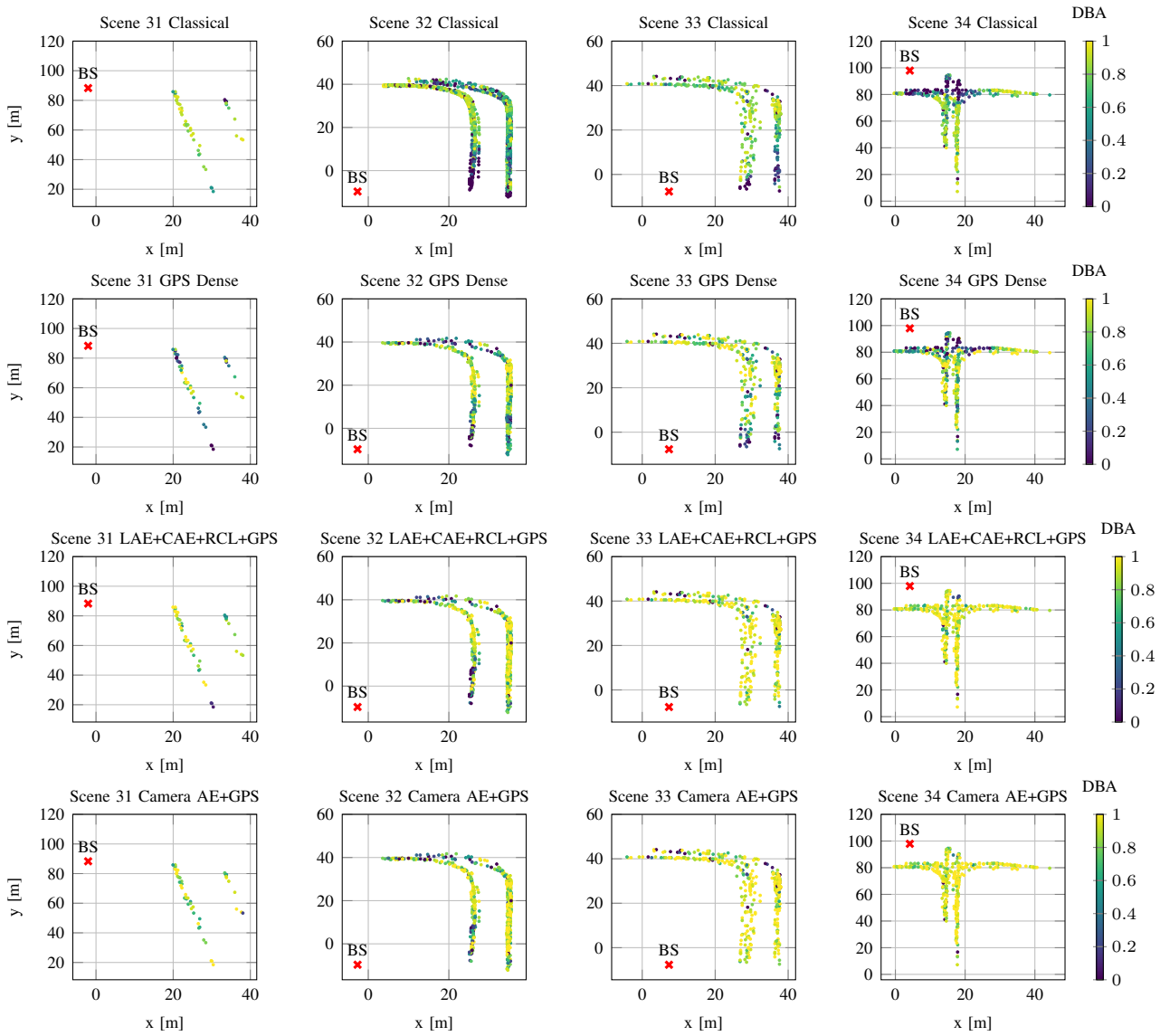

Fig. 5: DBA score illustrated spatially for four model configurations.

VI. Experiments

A. Setup

We evaluate in excess of 100 different combinations of sensory data, but only show the results for the top 10 configurations. Note that we use weight pruning to regularise learning throughout, as well as early stopping. The challenge supplied an adaptation set to fine-tune models in order to quantify how various models perform on new unseen data. Fine-tuning model configurations on adaptation sets uses reduced learning rates.

B. Takeaways

Performance under distribution shift. Fig. 6 depicts the cumulative distribution functions (CDFs) of DBA scores for various neural configurations, evaluated across Scenes. Without using the adaptation dataset, all models perform reasonably well on Scenes 32-34 seen during training, roughly irrespective of sensing modalities and the learning approach. With fine-tuning on the adaptation dataset, performance gains are limited on seen Scenes. Further, camera-based beam prediction achieves excellent performance with or without adaptation. However, the picture is more nuanced for Scene 31 not seen during training, where we see clear differentiation in performance across configurations. Both the car type, colour, and user mobility pattern are different in Scene 31 compared to Scenes 32-34. Configurations Camera AE+GPS and LAE+CAE+RCL+GPS have the best performance without adaptation, which suggests that pretraining enhances generalisability.

Spatial analysis. Fig. 5 analyses the DBA score spatially

TABLE II: Beam predictor nets and their valid architecture, input-output configurations, and training objective implemented in BeamBench.

Detail	Supervised	Unsupervised
Mapping	$\mathbb{C}^M \rightarrow \mathbb{R}^{64}$	$\mathbb{C}^M \rightarrow \mathbb{R}^{M'}$
Type	CNN	AE / CL
Input	GPS/Image/Radar/Lidar/Features	Image/Radar/Lidar
Output	Beam probability	M' Features
Configuration	**Description**	
Classical	Least squares fit	
Sensor Dense	Sensor fully-connected to beamvector	
Sensor CNN	Sensor CNN to beamvector	
Sensor AE	Sensor autoencoder, e.g. Radar AE RAE	
Sensor CL	Sensor contrastive learning, e.g. Radar CL, RCL	
Sensor AE + ...	Sensor fusion to beamvector	
Adapt ...	Fine-tuned on adaptation set	

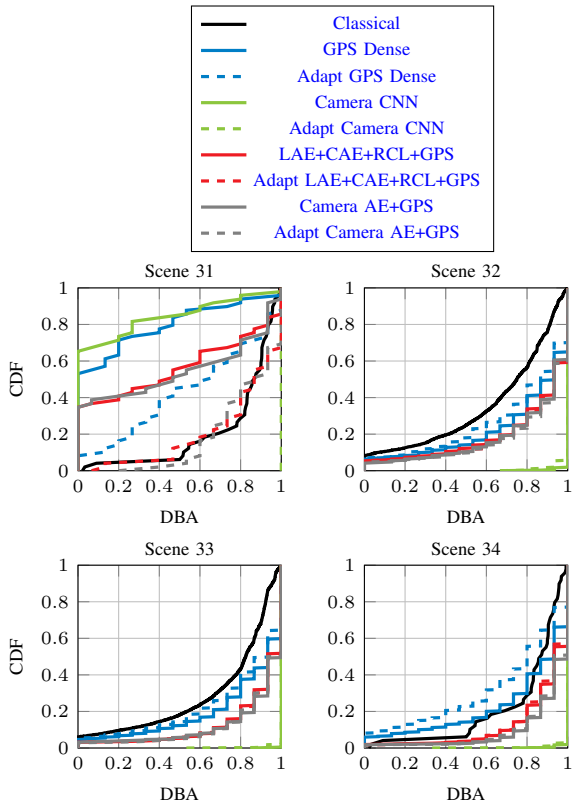

Fig. 6: DBA CDF's of Scenes 31-34 using various neural configurations (cf., Tab. II).

for the different Scenes and the different configurations. For Classical, the angles at the edges of user trajectories or those closet to the basestation have low DBA scores. This is to be expected as the coverage of the communication is limited by the antenna aperture. Its corresponding neural variant (GPS Dense) exhibits broadly a similar behaviour, albeit with mild enhancements. Similar to above analysis, configurations Camera AE+GPS and LAE+CAE+RCL+GPS and CAE+GPS offer, on the other hand, much improved spatial consistency of DBA scores. This suggests that neural multi-modal beam predictors have clear benefits w.r.t. spatial generalisability.

Top performers. Tab. III lists the DBA scores of top 10 configurations in BeamBench, all on the completely unseen test dataset. We have found multiple other configurations that work well on the training dataset, but do not generalise on test dataset. Tab. III shows that Camera AE+GPS is a solid beam predictor configuration. Further, Camera AE with and without LSTM have similar performance, which suggests that temporal modelling offers little benefit. A possible explanation is that training sequences and test sequences are different, which is impeding the generalisability of temporal modelling. Pretrained configurations (i.e., with latent space) outperform the supervised learning on the matched labels. We hypothesise that multi-modal matching is noisy, and that learning intermediate features per modality followed by "late fusion" is a better strategy on DeepSense 6G. Generally, multiple configurations achieve enhanced performance over that of the baseline. A good complexity-performance trade-off appears to be Camera AE+GPS.

C. Practical tips

Based on our analyses, we would recommend the following.

1 – Representation learning is important. Multiple analyses point to qualitative and quantitative differentiation w.r.t. BeamBench configurations. Robust beam prediction seems to be linked to feature extraction via label-agnostic pretraining, and not supervised learning.

2 – Pretraining per modality is powerful. Due to user coverage inconsistencies across modalities, cross-modal matching has proved necessary (cf., Tab. I). In the relatively "small data regime" of DeepSense 6G, competitive configurations rely on pretraining per modality, followed by a "late multi-modal fusion" of pretrained features for the downstream beam prediction task.

3 – Use camera AE and GPS. Uniformly across analyses, we found that using camera AE with GPS the results in a strong multi-modal configuration for beam prediction.

4 – Use regularisation. Reguralisation via weight pruning proved key to surpassing the performance of Classical baseline on DeepSense 6G beam prediction dataset.

VII. Limitations

We took first steps towards investigating the space of learning-based multi-modal beam predictors. Naturally, there remains many more avenues for further research. Examples include accounting for the elevation angle of the basestation, image augmentation for vision model variants, and active learning strategies. Time series modelling should offer performance enhancements, and future work could investigate more elaborate techniques to integrate temporal information during learning irrespective of dynamic sampling variabilities.

VIII. Conclusion

In this work, we build a comprehensive benchmark of learning-based beam predictors. Specifically, we (i) devise

Camera	Radar	Lidar	GPS	Fusion	Scene 31	Scene 32	Scene 33	Scene 34	Overall
			Classical*	No	0.5574	0.6505	0.7065	0.7618	0.6417
			Dense†	No	0.5980	0.6691	0.7971	0.6758	0.6651
AE	Dense	AE	Direct	Yes	0.6022	0.6494	0.8424	0.7394	0.6889
AE	AE	AE	Direct	Yes	0.5883	0.6741	0.8390	0.7003	0.6749
AE	AE		Direct	Yes	0.6044	0.6531	0.8557	0.7229	0.6911
AE	**CL**	**AE**	**Direct**	**Yes**	**0.6458**	**0.7002**	**0.8538**	**0.7003**	**0.7087**
AE			**Direct**	**Yes**	**0.6731**	**0.6173**	**0.8171**	**0.7313**	**0.7127**
AE	**Dense**		**Direct**	**Yes**	**0.6536**	**0.7074**	**0.8576**	**0.7120**	**0.7162**
AE	**AE**	**Dense**	**Direct**	**Yes**	**0.6469**	**0.6741**	**0.8571**	**0.6947**	**0.7063**
CNN				No	0.3720	0.6815	0.7490	0.7298	0.5582
AE			Direct	LSTM	0.6701	0.6010	0.7908	0.7313	0.6994

*Classical is the least square solution using calibrated GPS

†Dense uses calibrated GPS in a downstream task

TABLE III: DBA score for different configurations, evaluated on the test dataset

an analytic baseline, (ii) calibrate and match the DeepSense 6G multi-modal data, and (iii) train over 100 different configurations. Our characterisation shows that a vision-based autoencoder aided by GPS is a strong beam predictor configuration on DeepSense 6G. Other configurations based on pretrained multi-modal features are also strong beam predictors, and likely to offer unique advantages under inclement weather conditions (not covered in DeepSense 6G) where vision would struggle. We summarise our findings in a set of concrete takeaways, a number of practical tips, and open research directions.

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] H. Hassanieh, O. Abari, M. Rodriguez, M. Abdelghany, D. Katabi, and P. Indyk, "Fast millimeter wave beam alignment," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, 2018, pp. 432–445.

[3] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 2393–2401.

[4] G. Charan, T. Osman, A. Hredzak, N. Thawdar, and A. Alkhateeb, "Vision-Position Multi-Modal Beam Prediction Using Real Millimeter Wave Datasets." in *WCNC*, 2022, pp. 2727–2731.

[5] S. Jiang and A. Alkhateeb, "Computer Vision Aided Beam Tracking in A Real-World Millimeter Wave Deployment," *arXiv preprint arXiv:2111.14803*, 2021.

[6] M. A. L. Sarker, I. Orikumhi, J. Kang, H.-K. Jwa, J.-H. Na, and S. Kim, "Vision-Aided Beam Allocation for Indoor mmWave Communications," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2021, pp. 1403–1408.

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

[8] Y. Heng and J. G. Andrews, "Grid-Free MIMO Beam Alignment through Site-Specific Deep Learning," *arXiv preprint arXiv:2209.08198*, 2022.

[9] N. Gonzalez-Prelcic, R. Méndez-Rial, and R. W. Heath, "Radar aided beam alignment in mmWave V2I communications supporting antenna diversity," in *2016 Information Theory and Applications Workshop (ITA)*. IEEE, 2016, pp. 1–7.

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

[11] S. Jiang, G. Charan, and A. Alkhateeb, "LiDAR Aided Future Beam Prediction in Real-World Millimeter Wave V2I Communications," *arXiv preprint arXiv:2203.05548*, 2022.

[12] M. Zecchin, M. B. Mashhadi, M. Jankowski, D. Gündüz, M. Kountouris, and D. Gesbert, "LIDAR and Position-Aided mmWave Beam Selection with Non-local CNNs and Curriculum Training," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 3, pp. 2979–2990, 2022.

[13] A. Hu and J. He, "Position-Aided Beam Learning for Initial Access in mmWave MIMO Cellular Networks," *IEEE Systems Journal*, 2020.

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

[15] 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

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[17] G. J. et. al., "ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support," Oct. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5563715

[18] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[20] Microsoft, "Neural Network Intelligence," 1 2021. [Online]. Available: https://github.com/microsoft/nni