Doctoral Dissertation

Quantization and Attention-based Hierarchical Deep Learning Models for Beam Training in mmWave Massive MIMO Systems

Haohui Jia
Program of Information Science and Engineering
Graduate School of Science and Technology
Nara Institute of Science and Technology

Supervisor: Professor Minoru Okada
Network Systems Lab. (Division of Information Science)

Submitted on September 15, 2023

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Engineering

Haohui Jia

Dissertation Committee:

Supervisor	Minoru Okada (Professor, Division of Information Science)
Co-supervisor	Shoji Kasahara (Professor, Division of Information Science)
	Takeshi Higashino (Associate Professor, Division of Information Science)
	Duong Quang Thang (Affiliate Associate Professor, Division of Information Science)
	Na Chen (Assistant Professor, Division of Information Science)

Quantization and Attention-based Hierarchical Deep Learning Models for Beam Training in mmWave Massive MIMO Systems[1]

Haohui Jia

Abstract

Millimeter wave (mmWave) massive multiple-input multiple-output (MIMO) serves as the critical technology in fifth/sixth generation (5G/6G) wireless communication systems due to its capacity to provide comprehensive spectrum and spatial resources for high transmission rate demands. Deep learning (DL)-based beam training schemes have been adopted to preserve spectral efficiency with fast optimal beam selection in mmWave massive MIMO systems. To achieve high prediction accuracy, these DL models rely on training with a tremendous amount of labeled environmental measurements, such as mmWave channel state information (CSI), under the supervised learning (SL) framework. However, the CSI is composed of multiple subsets of ray/clusters, in which a unified DL structure hardly expresses the varying spatial information of frequency domain and interrelations of frequency and spatial. Meanwhile, a complex environment also incurs critical performance degradation in the continuous output of beam training. Moreover, demanding a large volume of ground truth labels for beam training is inefficient and infeasible due to the high labeling cost and the requirement for expertise in practical mmWave massive MIMO systems.

This dissertation comprises a hierarchical paradigm based on the SL and ex-

[1] Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, September 15, 2023.

tends to a novel contrastive learning framework working on a tiny fraction of the labeled CSI dataset. We first shed light on developing a hierarchical DL structure containing a cascade encoder for frequency and spatial domain. Specifically, we organize two schemes for extracting the frequency information with the corrupt CSI. To get rid of noise and variation, we propose a non-deterministic encoder coding the channel intensities into a binary representation with the stochastic thresholds. In addition, we also perform an autoregressive encoder for the inherent time delay attribute of approached mmWave channel signals on the different antenna indexes. Benefiting from the semantic model, we perform a spatial attention encoder that contributes to the optimal beam decision by exploring the relation between latent beam directions and generated beam gains. By contrast, the non-deterministic scheme can save computational costs while sacrificing the accuracy of optimal beam prediction compared with the autoregressive encoder.

Leveraging the hierarchical paradigm, we propose a novel contrastive learning framework, named self-enhanced quantized phase-based transformer network (SE-QPTNet), for reliable beam training with only a small size of the labeled CSI dataset. We develop a quantized phase-based transformer network (QPTNet) to explore the essential features from frequency and spatial views and quantize the environmental components with a latent beam codebook to achieve robust representation. Next, we design the SE-QPTNet, including self-enhanced pre-training and supervised beam training. SE-QPTNet pre-trains by the contrastive information of the target user and others with the unlabeled CSI, and then it is utilized as the initialization to fine-tune with a reduced volume of labeled CSI.

Finally, experimental results show that this study outperforms existing DL-based schemes, obtaining higher capacity and highly reliable performance for mmWave massive MIMO systems. The proposed framework further enhances flexibility and breaks the limitation of the quantity of label information for practical beam training.

Keywords:

5G/6G, mmWave, massive MIMO, DL, non-deterministic, quantization, transformer, spatial attention, contrastive learning

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Millimeter wave (mmWave) massive multiple-input multiple-output (MIMO) is one of the most critical technologies in fifth/sixth-generation (5G/6G) wireless communication systems due to its capacity to provide wide range spectrum and spatial resources for high transmission rate demands [1, 5–7]. Benefiting from the short wavelength of mmWave signals, it permits a massive number of antenna elements to be integrated into limited equipment size at both base station (BS) and user equipment (UE) sides [8]. In addition, the massive MIMO array can compensate for the severe path loss of mmWave signals by highly directional beamforming, leading to stronger coverage, larger data rates, and improved reliability [9, 10].

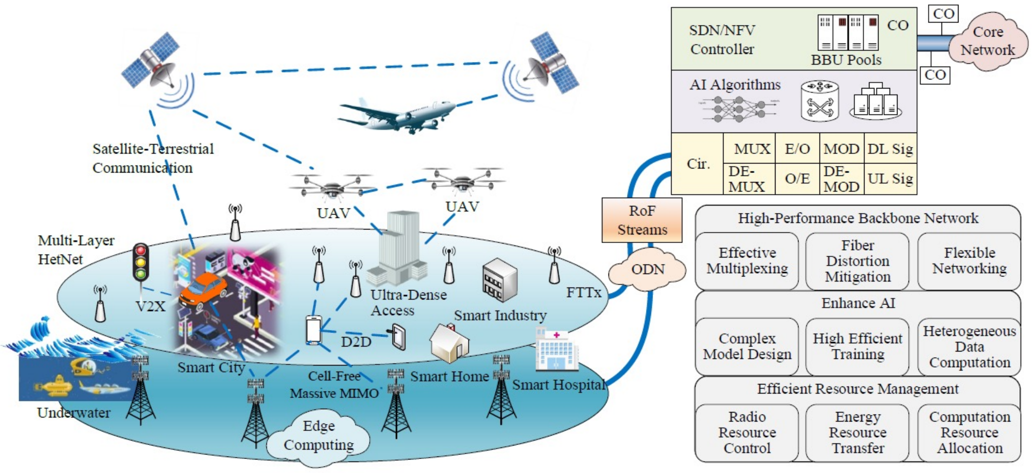


Figure 1: Overview of future intelligent wireless system [1].

Generally, mmWave massive MIMO arrays adaptively transmit signals via directional beams according to the wireless environment state [11–15]. To efficiently transmit beams with maximum gain, beam training is essential to identify the line-of-sight (LOS) or dominant channel path, yielding the optimal beam pair for the transceivers [11]. In practical beam training, candidate beams can be defined by a finite-size codebook covering the intended angle range and exhaustively retrieved to determine the optimal beam pair [12]. However, mmWave massive MIMO beam training is challenging due to the large codebook size, which results in high computational overhead. To reduce the training overhead, [13] proposes hierarchical multi-resolution codebook solutions, where a low-resolution sub-codebook detects the candidate transmitting direction, and the high-resolution codebook confirms the optimal beam pair. Another efficient beam training approach is interactive beam search. In [14] and [15], they detect the direction of the LOS/dominate path from the mmWave channel estimation and select optimal beam pairs.

The performance of beam training schemes, including alignment accuracy and overhead, is highly dependent on the codebook design. Literature has shown that an adaptive hierarchical codebook can decide the codewords based on previous beam training results with multiple mainlobes covering a spatial region for one or more user equipments (UEs) [16]. In [17], it is provided to efficiently generate the hierarchical codebook by jointly exploiting sub-arrays with the partial active antenna elements. An adaptive and sequential alignment scheme was proposed in [18], demonstrating the relation between fast search time and the probability of error in acquired beam directions through extrinsic Jensen-Shannon divergence. The study proposed in [19] developed a fast beam-sweeping algorithm based on compressive sensing (CS) to determine the minimum number of measurements required.

1.2 Related work

The conventional schemes can satisfy the user demands but can hardly inherit from the experience to further improve their ability. Deep learning (DL) has recently elevated the field of wireless systems research and beam training to new heights [20, 21]. These heuristic proposals depend on the well-labeled channel state information (CSI) corresponding to the aligned gain of a pre-defined codebook to construct a supervised learning (SL) framework. In this framework, the optimal beam can directly predict based on the large volume of labeled knowledge to reduce training overhead and effects of noise [2, 22–26]. A beam selection scheme based on deep neural networks (DNN) was proposed in [22] that recognized the desired beam from the elementary relation of position and received signals with low alignment overhead. Meanwhile, the potential of DL in predicting the optimal mmWave beam and correcting blockage status based on the sub-6 GHz channel information has been proposed by [23] to reduce beam training overhead and achieve reliable communication. [24] has trained a beamforming prediction network (BPNet) using supervised and unsupervised learning methods to optimize power allocation and predict virtual uplink beamforming (VUB) for improving computational efficiency. An online learning-based training strategy in [2] utilized a large volume DNN network to obtain the offline model parameters and fine-tune the DNN model according to the extra CSI measured in real time. An adaptive beam training scheme for calibration was proposed in [25] that estimated approximate CSI features by a convolutional neural network (CNN) and determined the optimal beam by self-criticism of the long short-term memory (LSTM). Moreover, a non-deterministic beam training was proposed in [26] that developed a binary coding scheme to represent the valid CSI and reduce the effect of noise. Although DL-based studies can obtain impressive achievements, severe multipath interference may impair prediction accuracy when the angles of paths in the local cluster are closely near the dominant channel path.

CSI is informative in deciding the optimal beam by exploring the dominant path during the training process. In [27], the authors proposed a hierarchical search by decomposing the multipath into several virtual components and using the hierarchical search to recover the dominant CSI for beam training. However, estimation performance highly relies on the training penalty. DL-based beamspace channel estimation was proposed in [28] to directly estimate the beamspace from the received signal, eliminating the need for a time-consuming beamforming process. A channel signature-based hybrid precoding design was proposed in [29] using DNN to estimate the channel and perform hybrid precoding with low computational complexity. Furthermore, a dual timescale variational framework for mmWave beam training and training [30] addressed beamforming direction in real-time by training a deep recurrent variational autoencoder, taking into account both the historical channel information and the current channel conditions. In [31] and [32], LSTM is shown to further improve the ability of beam training by the implicit channel signature. [31] inferred the optimal beam directions at a target BS in future time slots depending on the historical channel features for mobile UE. [3] indicated that spatial attention beam training can improve transmission reliability, and the associative LSTM encoder performs explicit channel features to improve training ability.

The existing solutions exploiting the DL model for beam training share the following limitations. Firstly, most existing solutions perform inefficient feature extraction for frequency and spatial domain CSI. Direct vectorization for frequency and spatial information leads to a coarse representation. It is thus hard for DNN-based beam training to extract the CSI feature information effectively, resulting in low learning efficiency and beam prediction performance. Meanwhile, CNN-based models show strong ability on 2-dimensional local feature extraction but suffer from a limited global view of beamspace awareness. In addition, sequence modeling can capture the relation of features from frequency and spatial domains but can hardly learn over an extensive range. Secondly, environmental measurements can affect the continuous output of the dense network. The continuous representation of CSI is sensitive to noise and channel variation, resulting in an incorrect beam prediction. Finally, the existing SL approaches require all CSI data to be labeled. However, labeling the large volume of CSI is unrealistic due to the high labeling cost and expertise requirement in practical mmWave massive MIMO systems. Although CSI is easily obtainable, handling the rapidly changing CSI when labeling the actual optimal beam is impractical. Therefore, the deficiency of labeled CSI can constrain the performance of existing DL-based beam training schemes.

1.3 Motivation

Typically, the CSI of massive MIMO contains the frequency dynamic response that exhibits spatial fluctuations in interconnected power grids. Because we concentrate on self-enhanced pre-training and SL with limited labeled data, it is critical to align the relation of preponderant paths corresponding with different subcarriers for well-explored beneficial features. Since the training signals result from the transmitted signals and the propagation environment, tracking the UE movement or capturing the local feature couples have achieved delightful results [25], [3]. However, these methods are incapable of fetching a global view of spatial and frequency features. Consequently, we absorb the benefits of existing works to develop a hierarchical DL architecture with two levels, where the first level tracks the spatial varying on different subcarriers and serves the second level to explore the global view for efficient beam training.

Moreover, the complexity and diversity of the wireless environment are challenging issues for DL-based beam training. Continuous feature extraction results in uncontrollable representation, sensitive to random disturbance terms. Existing work addresses phase quantization mainly in three ways, by designing with real-valued phase shifts and then applying quantization [33], by constructing an analog beamforming codebook [34], and by nonlinear mapping into binary phase quantization [4]. They lack flexibility and robustness to noise and channel variations. To address this problem, we develop a configurable codebook to quantize the continuous spatial feature satisfying the equal angele of departure (AoD) distribution in categorical beams. Specifically, discrete codewords can get rid of the effects of noise and channel variations and reflect the dominant factor of the CSI. Thus, we can obtain controllable quantized results from the codebook to improve the robustness of hierarchical DL architecture.

DL has been suggested as a promising approach to address the nonlinear relationship of CSI and optimal beam prediction. As indicated in [2,3], DL-integrated beam prediction is typically performed under an SL framework with perfect CSI annotation. However, it is challenging to acquire the exhaustive annotation of massive CSI for DL-based beam training in realistic massive MIMO mmWave systems, which leads to high labeling costs and expertise requirements. [35] proposes an unsupervised method that performs the CSI reconstruction, and accomplishes the online SL beam training with a large dataset of labeled CSI. It is inefficient for limited labeled CSI because of the uniqueness of the wireless environment, lacking extendability. We shed light on a novel contrastive learning framework with limited number of labeled CSI to mitigate the expertise requirements. In particular, we leverage the uniqueness of CSI of different UE locations to pre-train by identifying the contrastive information between the target UE and others.

1.4 Contributions

This dissertation comprehensively proposes a series of DL-integrated beam training schemes devoted to highly efficient, robust, and flexible beam training. In addition, we also consider reliable beam training with CSI measurements underlying limited labels for practical mmWave massive MIMO systems. We first develop a non-deterministic quantization to code the channel swelling as a binary sequence and also develop a wise scoring inference scheme to enhance the robustness of beam prediction against the effects of noise. Moreover, We further explore a hierarchical DL paradigm based on the frequency and spatial domain views to holistically perceive the co-varying of spatial changes in each sub-carrier index. Eventually, we extend it to a novel contrastive learning framework working on a tiny fraction of the labeled CSI dataset. Specifically, we organize non-deterministic and autoregressive encoders for extracting the frequency information with the corrupt CSI. The proposed non-deterministic encoder converts the channel intensities into a binary representation, and the autoregressive encoder handles the interrelation of frequency and spatial domain. Benefiting from the Transformer model, we apply a spatial attention encoder contributing to the optimal beam by scoring the relation between latent beam directions and generated beam gains. Leveraging the hierarchical DL paradigm, we further design a novel contrastive learning framework. We quantize the environmental components with a latent beam codebook to achieve robust representation. The proposed framework pre-trains by the contrastive information of the target user and others with the unlabeled CSI and then utilizes it as the initialization to fine-tune with negligible labeling cost. The main contributions of this study can be summarized as follows:

- We propose a novel binary representation of mmWave CSI to mitigate the noise and channel variations for massive MIMO wireless systems. To achieve a robust and efficient representation of CSI, we incorporate the ideas from non-deterministic quantization to convert the corrupt channel intensities into binary sequences. The non-deterministic method also allows the DNN model to compress the volume of parameters, making the proposed model easy to implement. Moreover, we develop a novel sparse feature-driven vision Transformer (SF-ViT) DL model. It inherits the advantage of binary quantization of CSI and extracts the spatial characteristics based on the semantic model from antenna indices.

- We develop a novel hierarchical DL paradigm devoted to improving the learning efficiency for beam training with the frequency and spatial mmWave CSI in massive MIMO systems. To overcome the noise and the mmWave channel fading attenuation, we propose a dual DL model that learns temporal delay features based on the autoregressive model from the frequency and extracts the spatial characteristics based on the semantic model from antenna indexes.

- We develop a contrastive learning framework SE-QPTNet benefiting from the hierarchical QPTNet and codebook-based phase quantization. This enhanced model further improves beam training accuracy with limited labeled CSI. To the best of our knowledge, this is the first study that introduces contrastive learning in beam training applications. SE-QPTNet performs two benefits based on contrastive environmental prediction. Firstly, it can pre-train without any label information by detecting the relationship between the global beam feature and positive/negative samples. Secondly, the similarity of a positive sample and beam signature can effectively capture the spatial dynamic changes under long inter-frequency spans. SE-QPTNet preserves the benefits of QPTNet and reduces the labeling cost.

Notations: $\boldsymbol{A}$ is a matrix; $\boldsymbol{a}$ is a vector; a is a scalar; $(\cdot)^T$ and $(\cdot)^H$ denote transpose and conjugate transpose, respectively, while $|\cdot|$ denotes the magnitude operator. $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary parts of a complex number, respectively. $\mathcal{CN}(0, \mathbf{\Sigma})$ represents the zero-mean complex Gaussian distribution with covariance matrix $\mathbf{\Sigma}$, respectively.

1.5 Dissertation layout

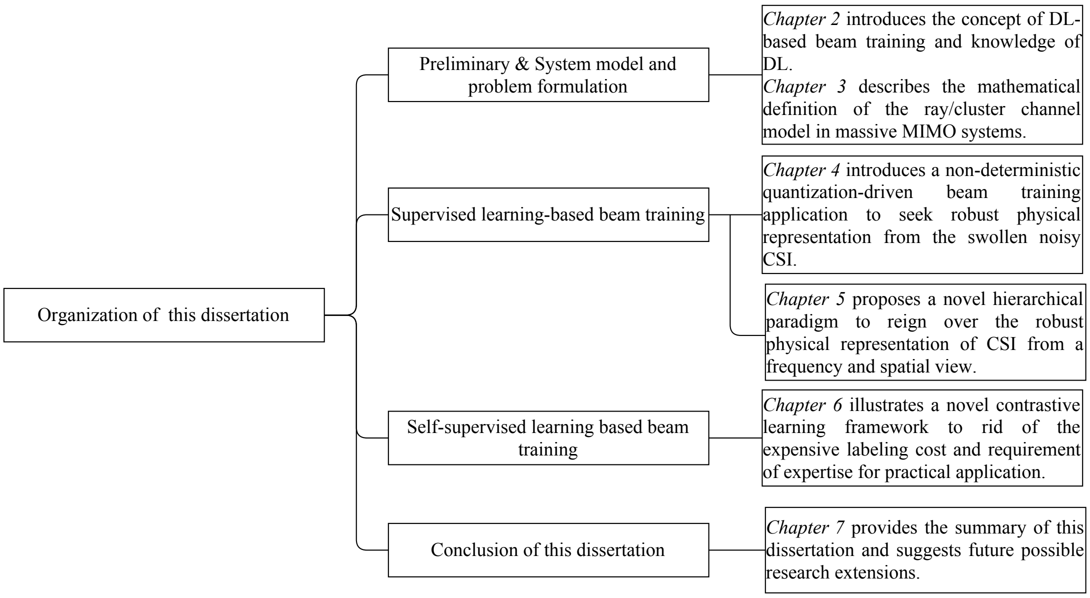

Organization of this dissertation
Preliminary & System model and problem formulation
Supervised learning-based beam training
Self-supervised learning based beam training
Conclusion of this dissertation
Chapter 2 introduces the concept of DL-based beam training and knowledge of DL. *Chapter 3* describes the mathematical definition of the ray/cluster channel model in massive MIMO systems.
Chapter 4 introduces a non-deterministic quantization-driven beam training application to seek robust physical representation from the swollen noisy CSI.
Chapter 5 proposes a novel hierarchical paradigm to reign over the robust physical representation of CSI from a frequency and spatial view.
Chapter 6 illustrates a novel contrastive learning framework to rid of the expensive labeling cost and requirement of expertise for practical application.
Chapter 7 provides the summary of this dissertation and suggests future possible research extensions.

Figure 2: Layout of this dissertation.

Fig. 2 shows the layout and the dissertation is organized as follows:

Chapter 2 describes the preliminary containing the principle of DL-integrated beam training and the concept of various DL models.

Chapter 3 introduces an overview of wireless communication systems and in particular the mmWave massive MIMO system model and the problem formulation.

Chapter 4 illustrates two non-deterministic quantization-driven proposals, which code the channel fluctuations into binary sequences to drain the effects of noisy intensities. The first proposal further explores a hardware-friendly model compression scheme. Besides, the second proposal draws inspiration from the attention mechanism, inducing the dense spatial representation for a robust beam training application.

Chapter 5 illustrates two hierarchical paradigm-based beam training proposals, aligning the frequency-varying with the autoregressive encoder and inducing the global beam signature with the spatial attention mechanism. Specifically, the first proposal tracks the frequency varying with the LSTM encoder from concatenated real and imaginary CSI. The second proposal conducts a memory-shared LSTM to capture the intensive frequency information from real and imaginary dimensions.

Chapter 6 illustrates a self-enhanced quantized phase-based Transformer network, which is composed of a latent beam codebook-driven quantization of environmental components and a contrastive learning framework for pre-training without annotation. The pre-training scheme contributes to ameliorating the demands of labeled CSI by identifying the inherent information of the target user and others with the unlabeled CSI for practical beam training applications with flexible label requirements.

Chapter 7 finally provides the conclusion of this dissertation and suggests future possible research extensions and directions based on the current works.

2 Preliminary

In this chapter, we introduce the basic concept of DL integrated beam training in mmWave massive MIMO systems and the primary DL modules applied in the following proposals. First, we describe the purpose of supervised beam training and the benefits of DL for tackling the non-linear relation of CSI and optimal beam response. Second, we shed light on a series of DL modules, including the representative autoregressive model (LSTM and GRU), Transformer, vector quantised-variational autoencoder (VQVAE), and the classic contrastive predictive coding (CPC) in the self-supervised learning field, respectively.

2.1 DL integrated beam training

The aim of DL-integrated beam training is to convert empirical searching into parametric design during the automatic training process. Fig. 3 shows the diagram of DL integrated beam strategy, including K uplink pilots $\hat{\boldsymbol{s}}^{pilot}$ for CSI acquisition and training to obtain the optimal beam prediction $\hat{\boldsymbol{f}}^{DL}$ downlink data transmission. In this phase, the system relies on training a DL model to determine the relation between feedback CSI and the expression of digital and analog beamformer for obtaining the optimal beam response. The DL model can train with labeled feedback CSI to learn the implicit relation between a defining signature for the user location/environment, and different optimal beam vectors for achieving a maximum data rate. Once the model is trained, the system operation moves to the second optimal beam prediction phase. According to the established implicit relation of CSI and optimal beams, the system can reach a fast beam response and high data transmission.

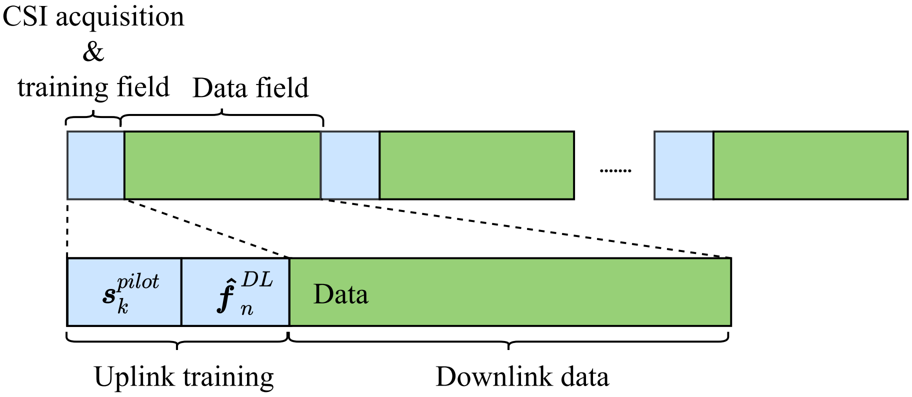


Figure 3: Diagram of DL integrated beam training strategy, including K uplink pilots for CSI acquisition and training to obtain the optimal beam prediction $\hat{\boldsymbol{f}}^{DL}$ downlink data transmission.

Fig. 4 shows the overview of DL-integrated beam training based on the SL framework and the operation for labeling the CSI with the actual optimal beam response based on the conventional beam sweeping. Because the candidate beams decided by the sweeping are finite, beam training can be regarded as a multi-class classification task, where the training process results predict the category corresponding to one candidate beam. Mathematically, the prediction can be represented by the probability results of the training function $\mathcal{F}(\cdot)$ as

$$n^* = \arg\max_{n \in \{1,2,\ldots,N_t\}} P(\boldsymbol{c}_t | \mathcal{F}(\boldsymbol{H}); \boldsymbol{W}) \tag{1}$$

where the optimal beam corresponding index n^* is the maximum probability from the output given the parameters $\boldsymbol{W}$ of training process.

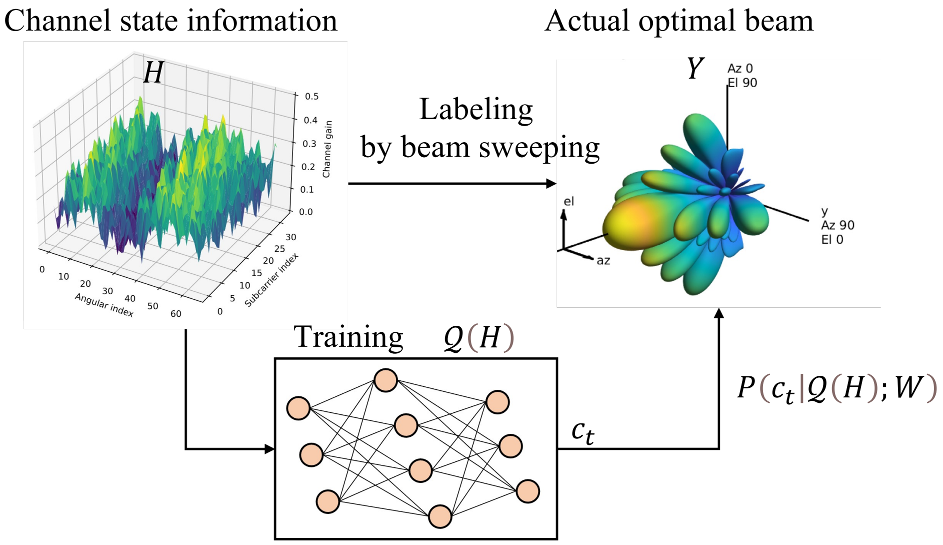


Figure 4: Overview of DL integrated beam training scheme, including the labeling and supervised training process.

The beam sweeping is the feasible method to obtain the ground-truth label information for supervised beam training. Conventionally, the analog beamformer $\boldsymbol{f}$ can be generated by a predefined codebook $\mathcal{F} \triangleq \{\boldsymbol{f}_n, n = 1, 2, \ldots, N_t\}$ that includes N_t codewords corresponding to different AoDs with the inherent transmitting spatial resolution. Identically, the analog combiner can be generated by $\mathcal{W} \triangleq \{\boldsymbol{\omega}_m, m = 1, 2, \ldots, N_r\}$ including N_r codewords with the inherent receiving spatial resolution with different AoAs. For each beam training test, the BS selects a codeword from $\mathcal{F}$ as the analog beamformer aligns with the analog combiner from $\mathcal{W}$ at the UE side. Generally, the discrete Fourier transform (DFT) codebook is a feasible option to apply for the candidate beamformer $\boldsymbol{f}_n$ and combiner $\boldsymbol{\omega}_m$, which can be described respectively as:

$$\boldsymbol{f}_n = \frac{1}{\sqrt{N_t}}[1, \mathrm{e}^{j\pi \sin \xi_{t,n}}, \cdots, \mathrm{e}^{j\pi(N_t-1)\sin \xi_{t,n}}]^T, \tag{2}$$

$$\boldsymbol{\omega}_m = \frac{1}{\sqrt{N_r}}[1, \mathrm{e}^{j\pi \sin \xi_{r,m}}, \cdots, \mathrm{e}^{j\pi(N_r-1)\sin \xi_{r,m}}]^T, \tag{3}$$

where the $\xi_{t,n}$ and $\xi_{r,m}$ are the beam directions of the nth possible beam at BS and mth possible received beams at the UE side. The DL integrated beam training include two stages:

Training stage: The input to the DL model includes the mmWave CSI, the ground-truth label information of the actual optimal beam, and the output would be the probability of optimal beam response. Once we collect the actual optimal pair of CSI and the ground-truth label, a DL-integrated beam training can be completed by making the loss between the actual optimal beam and the predicted beam as small as possible. The proposed DL models can be optimized by stochastic gradient descent with a given learning rate during backpropagation [36]. Practically, the training stage generally integrates into the BS because of the solid computational ability.

Predicting stage: A well-trained DL model can be utilized in real-time to predict the best beam direction given the current CSI. This would involve continuously feeding the current CSI into the DL model and adjusting the beam direction based on its output. Furthermore, the critical benefit of DL-integrated beam training is the comprehensive to deal with any mmWave CSI, which is faster than the conventional beam training schemes.

2.2 Auto-regressive model

In this part, we concentrate on the basic principle of RNN, which play an essential role in DL-integrated beam training. A central problem in sequence modeling for beam training is efficiently handling CSI and tracking beam varying that contains long-range frequency bands [37, 38]. It is feasible and efficient to capture the channel behavior by adopting the autoregressive model during the given state of the frequency band [39], such as LSTM and GRU. Consequently, we adopt the LSTM and GRU as the partial encoder to handle the CSI behavior from the frequency band and extract the stochastic knowledge of its AoA.

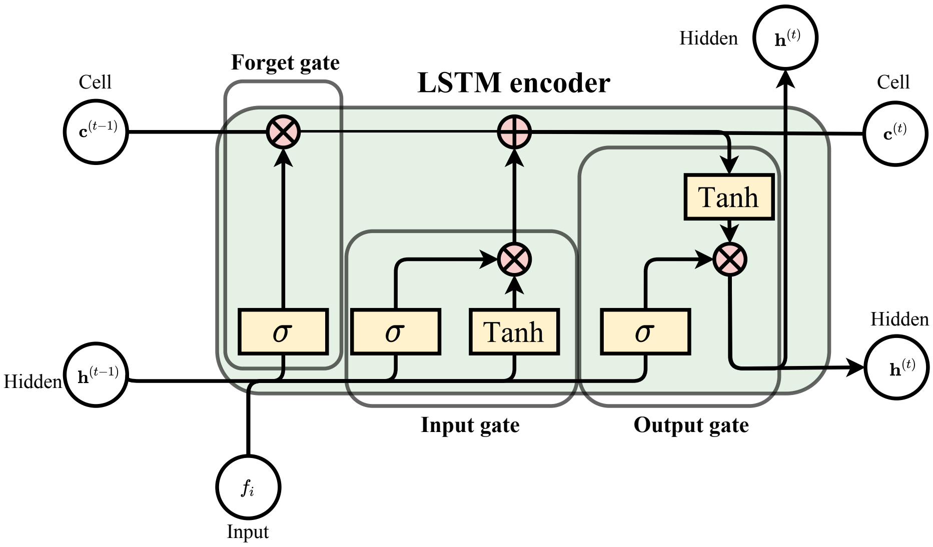


Figure 5: Overview of long short-term memory network.

Long short-term memory Fig. 5 shows the overview of LSTM, which is a popular type of RNN architecture used in DL. Unlike standard DNN model, LSTM has feedback connections. It can consider not only single data points but also entire data sequences, which is good at tackling the signal data. The LSTM consists of the memory cell and a series of gate operations. The memory cell is the central concept of the LSTM which is capable of maintaining its state over the entire signal period. The gate operations can control and modify the state of the memory cell to preserve the essential information. Specifically, gate operations are composed of a sigmoid neural net layer and a pointwise multiplication operation. Moreover, the forget gate keeps or forgets the content of the cell and the input gate decides what new information will be stored in the cell. Finally, the output gate decides the next hidden state.

An essential advantage of LSTMs is their ability to avoid the long-term dependency problem. This is a major challenge in training traditional RNNs, where early inputs in a sequence can have little influence on later outputs, making it hard for the model to learn how those early inputs should affect the prediction. LSTM networks are capable of learning long-term dependencies, making them useful for the various downstream tasks.

Gate recurrent unit The GRU module can be also treated as the instant feature extractor due to its competitive training efficiency and similarity to the LSTM network in temporal sequence learning. The GRU module efficiently handles long sequences of data using a gate mechanism to control the flow of information between the current and previous time steps, updating its hidden state via the following operation. This allows the GRU to learn features from prior inputs. In processing input data at each time step, the GRU takes in the current subcarrier input channel delay response as well as the shared hidden state and output from the previous subcarrier step. GRU updates its hidden state based on this information and outputs a new hidden state. The hidden state encodes the features that the GRU has learned from the input channel delay response thus far, and these features are used to make predictions about future subcarrier steps.

2.3 Transformer

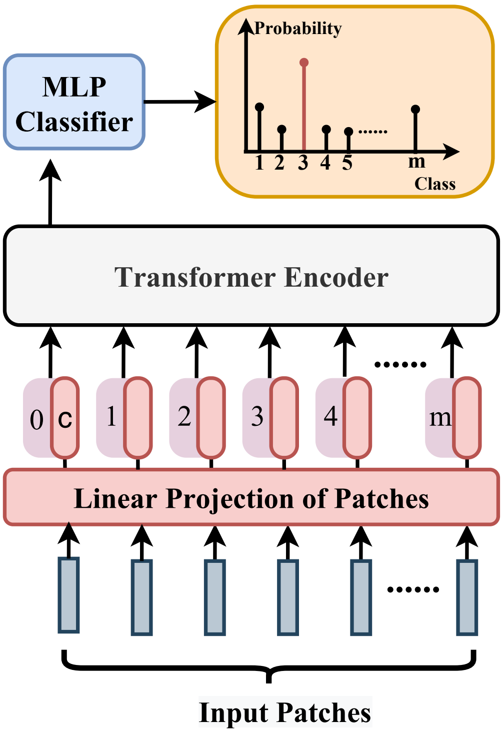
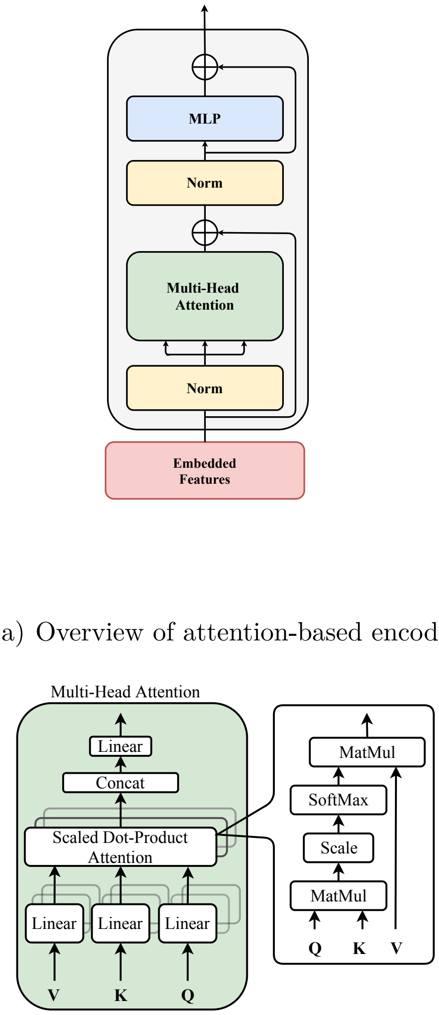


Figure 6: Overview of Transformer network.

Benefiting from the development of the semantic model, the Transformer model brings new insight for establishing the coherence of input data with attention strategy. The Transformer model is composed of stacked self-attention and point-wise, fully connected layers for both the encoder and decoder side, shown as Fig. 6. In the Transformer, the encoder is formed by N identical layers, and the encoder maps an input sequence of symbol representation with embedding layers $\boldsymbol{X} = (\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_d)$ to a memory matrix $\boldsymbol{Z}$ through the multi-head attention mechanism, and position-wise fully connected FFN [40]. Moreover, there is a residual connection between each sub-layer followed by layer normalization to enhance the details of data inputs. Similar to the encoder, the stacked decoder also utilizes the same processing, except for the addition of a cross-attention mechanism between the decoder self-attention sequences and memory $\boldsymbol{Z}$. Self-attention



(a) Overview of attention-based encoder

(b) Computing flows of multi-head attention.

Figure 7: Numerical experiment results among DNN, spatial attention, spatial attention with unshared weight LSTM encoder, and proposed model.

process (Fig. 7) could be described as a scaled dot-product function as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. In particular, we compute the dot-product of query matrix $\boldsymbol{Q}$ with all key matrix $\boldsymbol{K}$, and the dimensions must be identical, and apply a softmax activation to obtain the output matrix on the values matrix $\boldsymbol{V}$. In order to reduce the computational overhead, the attention function could be performed in parallel with the different attention heads.

2.4 Vector quantised-variational autoencoder

The nature of the transmitted beam is inherently discrete (aligning with the target UE with optimal direction). Nevertheless, most of the learning representations with continuous features are the basic procedure in the DL-based schemes. The discrete representations are potentially a more natural fit for many of the different dimensions [41, 42]. Furthermore, discrete representations are a natural fit for complex reasoning, planning, and predictive learning. In [43], authors introduce a new family of generative models successfully combining the variational autoencoder (VAE) framework with discrete latent representations through a novel parameterization of the posterior distribution of latent vectors given an observation. It relies on vector quantization (VQ), which is simple to train with a powerful decoder for avoiding the posterior collapse issue. Additionally, it is capable of offering the flexibility of discrete distributions.

Since VQ-VAE can make effective use of the latent space, it can successfully model important features that usually span the frequency band in mmWave channel space as opposed to focusing or spending capacity on noise and imperceptible details which are often local.

2.5 Self-supervised learning

Large-scale labeled data are generally required to train the DL model in order to obtain better performance in visual feature learning for different applications. To avoid the extensive cost of collecting and annotating large-scale datasets, as a subset of unsupervised learning methods, self-supervised learning methods are proposed to learn general image and video features from large-scale unlabeled data without using any human-annotated labels. To avoid time-consuming and expensive data annotations, self-supervised learning proposes to learn visual features from large-scale unlabeled data without using any human annotations. To learn visual features from unlabeled data, a popular solution is to propose the local-to-global pretext task for networks to solve, while the networks can be trained by learning objective functions of the pretext tasks, and the features are learned through this process [44].

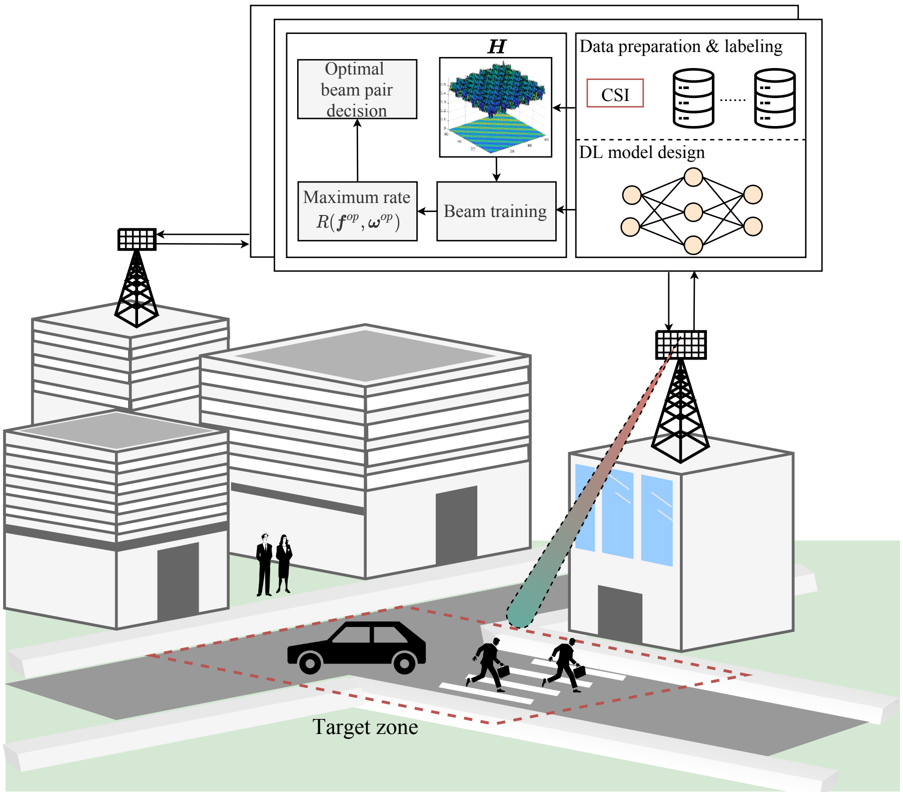
2.6 Concluding remark

In this chapter, we summarized the basic concept of DL integrated beam training in mmWave massive MIMO systems and the primary DL modules applied in the following proposals. Specifically, we introduced the principle of supervised beam training and the benefits of DL modules, including the representative LSTM and GRU for capturing the CSI behavior with the state of the frequency band, an efficient attention mechanism for seeking the desired beam direction based on the scores decision, a quantization based VAE to determine the approximate discrete expression with the learnable codebook, and the classic contrastive learning framework for self-supervised learning from the local-to-global view, respectively.

3 System model and problem formulation

In this chapter, we introduce the mathematical definition of mmWave massive MIMO wireless channel and formulate the main challenges for DL-integrated beam training. Specifically, we define the basic massive MIMO wireless system for the following proposals and the principle of beamforming and beam prediction in mmWave massive MIMO systems.

3.1 mmWave massive MIMO channel model



$\boldsymbol{H}$
Data preparation & labeling
Optimal beam pair decision
CSI
......
DL model design
Maximum rate $R(\boldsymbol{f}^{op}, \boldsymbol{\omega}^{op})$
Beam training
Target zone

Figure 8: Procedure of DL-based beam training schemes in mmWave massive MIMO wireless systems.

It is worth emphasizing that our proposals can be extended to various communication scenarios. First, the proposed contrastive learning framework can be developed for the multi-user hybrid beamforming design since the beam direction

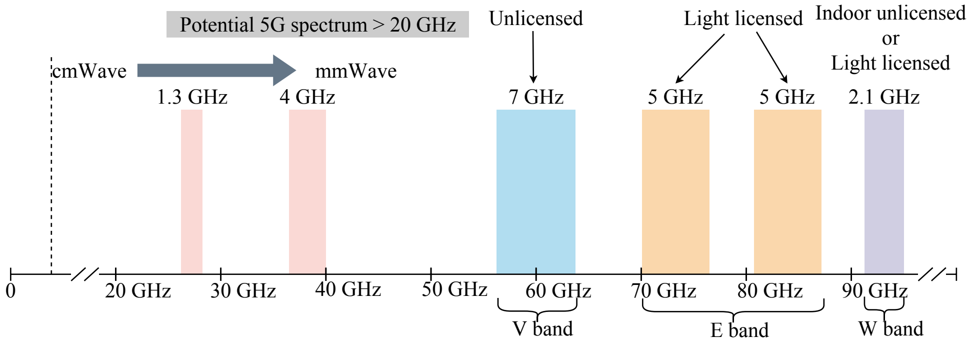


Figure 9: Overview of available mmWave bands.

is unique for each user separately to achieve fewer effects of multi-user interference, and the labeling cost is also intractable. Then, the proposed contrastive learning framework can be applied in the harsh wireless environment, e.g., the reconfigurable intelligent surface (RIS) scenario, where the beam direction of each user is aligned with the dominant AoD of RIS while others are treated as negative. The general procedure of the proposed beam training schemes is illustrated in Fig. 8. The proposed DL-integrated beam prediction is typically performed at the BS side to ensure fast prediction responses with high computational resources.

MmWave bands with significant amounts of band sources or moderately used sub-6 bandwidths are being considered as a key role in the current 5G systems. As shown in Fig. 9, the available bands in the range of 20-100 GHz make mmWave a bright prospect in the design of 5G networks. The authors in [45] explore the available mmWave frequency bands to design a 5G enhanced local area network. In [46], the authors propose a general framework to analyze the coverage and rate performance of the mmWave networks. However, mmWave cellular communication is heavily dependent on the propagation environment. MmWave signals are affected by several environmental factors and cannot penetrate through obstacles. Further, because of the high frequencies used in mmWave, the path loss with omnidirectional antennas increases with frequency. The authors in [47] analyze the performance of mmWave cellular systems using real-time propagation channel measurements. Blockage effects and angle spreads were also incorporated in [48] to analyze such systems. In a general communication system, LOS and NLOS measurements are composed of path loss and affect the communication

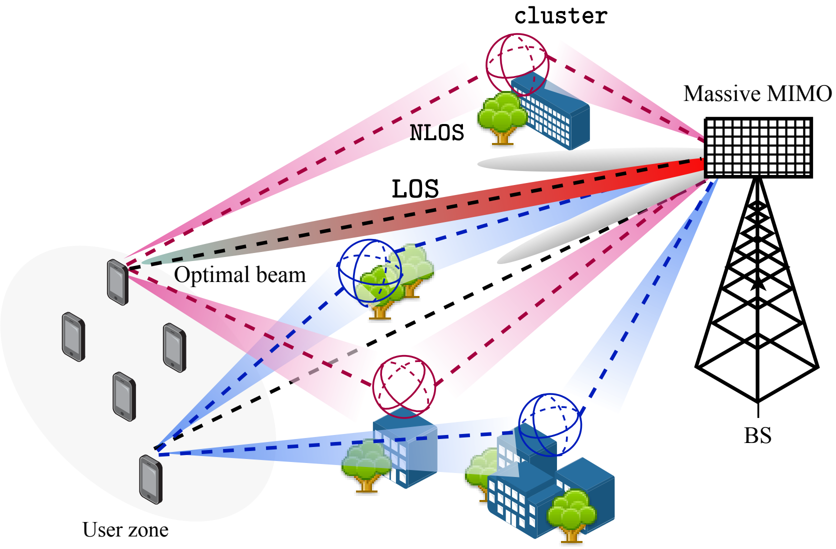


Figure 10: An illustration of an outdoor mmWave massive MIMO wireless system.

quantity. Hence, it is very necessary to explore the LOS and NLOS links in mmWave networks. It is a natural way to combat omnidirectional path loss by explicitly increasing the antenna aperture. The massive MIMO antenna arrays can overcome the frequency dependency on the path loss with the high array gain and allows mmWave systems to preserve a reasonable link margin.

Benefiting from the advantage of the massive MIMO technology, it can be considered to be an integral setup in the implementation of mmWave networks. The illustration of an outdoor mmWave massive MIMO wireless system is shown in Fig. 10. For analytical simplicity, we consider downlink transmission of a mmWave massive MIMO BS and a single antenna UE. For a 2-dimensional (2D) mmWave channel where only azimuth angles are considered at both BS and UE, the Saleh-Valenzuela channel model is typically adopted, which can be formulated as

$$\boldsymbol{H} = \sqrt{\frac{N_t N_r}{L}} \sum_{l=1}^{L} \beta_l \boldsymbol{a}_r(N_r, \theta_l) \boldsymbol{a}_t^H(N_t, \boldsymbol{\phi}_l), \tag{4}$$

where L, β_l, θ_l, and ϕ_l denote the number of channel paths, channel gain, angle-of-arrival (AoA), and angle-of-departure (AoD) of the lth channel path, respectively.

Since the first channel path, corresponding to the LOS path, is typically significant, recognizing the LOS path can be beneficial for improving the coverage of mmWave signals [3]. Although the number of resolvable channel paths is much smaller than the number of BS antennas, i.e., $L \ll N_t$, it is still challenging to efficiently distinguish the LOS path because of the limited scattering of mmWave channels and the non-line-of-sight (NLOS) [49] [50]. The AoA and AoD of the l th path can be defined as $\phi_l = 2d_t \sin \Phi_l/\lambda$ and $\theta_l = 2d_r \sin \Theta_l/\lambda$, where Θ_l and Φ_l are the set of LOS and NLOS paths, respectively; λ denotes the wavelength; $d_t = d_r = \lambda/2$ are the antenna spacing at the BS and UE. In particular, both Θ_l and Φ_l satisfy uniform distribution within $[-\frac{\pi}{2}, \frac{\pi}{2}]$. The transmit and receive array steering vectors can be expressed as

$$\boldsymbol{a}_t(\phi_l) = \frac{1}{\sqrt{N_t}}[1, \mathrm{e}^{j\pi\phi_l}, \cdots, \mathrm{e}^{j(N_t-1)\pi\phi_l}]^T, \tag{5}$$

$$\boldsymbol{a}_r(\theta_l) = \frac{1}{\sqrt{N_r}}[1, \mathrm{e}^{j\pi\theta_l}, \cdots, \mathrm{e}^{j(N_r-1)\pi\theta_l}]^T. \tag{6}$$

3.2 Beamforming and beam prediction

With the mmWave channel matrix $\boldsymbol{H}$ given in (4), the received signal can be described as

$$y = \sqrt{P}\boldsymbol{\omega}^H \boldsymbol{H} \boldsymbol{f} x + \boldsymbol{\omega}^H \boldsymbol{n}, \tag{7}$$

where P, $\boldsymbol{\omega} \in \mathbb{C}^{N_r \times 1}$, $\boldsymbol{f} \in \mathbb{C}^{N_t \times 1}$ denote the transmit power, combiner, and beamformer, respectively. x is the transmitted data with unit power, i.e., $|x| = 1$, while $\boldsymbol{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{N_r})$ denotes the additional white Gaussian noise (AWGN) vector with power σ^2. Typically, the beamformer and combiner do not increase or decrease the power gain, i.e., $\|\boldsymbol{\omega}\|_2 = \|\boldsymbol{f}\|_2 = 1$. The achievable rate can be described by

$$R = \log_2 \left(1 + \frac{P|\boldsymbol{\omega}^H \boldsymbol{H} \boldsymbol{f}|^2}{\sigma^2}\right). \tag{8}$$

To get the maximum achievable rate for the given $\boldsymbol{H}$, we conventionally perform the beam training to construct the optimal beam pair by optimizing $\boldsymbol{f}$ and $\boldsymbol{\omega}$ before data transmission (as shown in Fig. 8). This optimization issue can be implemented by the pre-defined codebooks $\mathcal{F}$ and $\mathcal{W}$ as the following equation

$$\{\boldsymbol{f}^{op}, \boldsymbol{\omega}^{op}\} = \arg\max_{\substack{\boldsymbol{f} \in \mathcal{F} \\ \boldsymbol{\omega} \in \mathcal{W}}} |\boldsymbol{\omega}^H \boldsymbol{H} \boldsymbol{f}|^2. \tag{9}$$

Practically, it is impossible to directly reveal the ideal pair of $\boldsymbol{f}^{op}$ and $\boldsymbol{\omega}^{op}$ since it is hard to tackle with the three independent matrices in (7).

A straightforward and ergodic solution of (9) is to enumerate all possible candidate codewords of $\boldsymbol{f}$ and $\boldsymbol{\omega}$ to determine the optimal solution with the largest achievable rate through the beam-sweeping. Conventionally, the beamformer $\boldsymbol{f}$ can be generated by a pre-defined codebook $\mathcal{F} \triangleq \{\boldsymbol{f}_n, n = 1, 2, \ldots, N_t\}$ that includes N_t codewords corresponding to different AoDs with the inherent transmitting spatial resolution. Identically, the combiner can be generated by $\mathcal{W} \triangleq \{\boldsymbol{\omega}_m, m = 1, 2, \ldots, N_r\}$ including N_r codewords with the inherent receiving spatial resolution with different AoAs. For each beam training test, the BS selects a codeword from $\mathcal{F}$ as the beamformer aligns with the combiner from $\mathcal{W}$ at the UE side. Generally, the discrete Fourier transform (DFT) codebook is a feasible option to decide the candidate beamformer $\boldsymbol{f}_n$ and combiner $\boldsymbol{\omega}_m$:

$$\boldsymbol{f}_n = \frac{1}{\sqrt{N_t}}[1, \mathrm{e}^{j\pi \sin \xi_{t,n}}, \cdots, \mathrm{e}^{j\pi (N_t-1) \sin \xi_{t,n}}]^T, \tag{10}$$

$$\boldsymbol{\omega}_m = \frac{1}{\sqrt{N_r}}[1, \mathrm{e}^{j\pi \sin \xi_{r,m}}, \cdots, \mathrm{e}^{j\pi (N_r-1) \sin \xi_{r,m}}]^T, \tag{11}$$

where the $\xi_{t,n}$ and $\xi_{r,m}$ are the beam directions of the n th possible beam at BS and m th possible received beams at the UE side. To span the whole angular domain in both BS and UE, $\xi_{t,n}$ and $\xi_{r,m}$ can be uniformly sampled in $(-\Xi_t, \Xi_t)$ and $(-\Xi_r, \Xi_r)$, i.e.,

$$\xi_{t,n} = (-1 + \frac{2n-1}{N_t})\Xi_t, \tag{12}$$

$$\xi_{r,m} = (-1 + \frac{2m-1}{N_r})\Xi_r. \tag{13}$$

To find the optimal solution of (9), the searching range is $K \triangleq N_t N_r$, while the candidate beam pair can be denoted as

$$\begin{gathered}\mathcal{B} = \{\boldsymbol{b}_k | k = 1, 2, \ldots, K\}, \boldsymbol{b}_k = \{\boldsymbol{f}_n, \boldsymbol{\omega}_m\}, \\ n = 1, 2, \ldots, N_t, m = 1, 2, \ldots, N_r.\end{gathered} \tag{14}$$

To evaluate the performance of beam training, the success rate is regarded as an important criterion in [17]. The index of solutions corresponding to the largest achievable rate can be treated as *successful*; otherwise, we consider the solutions are *fail* in the beam training. Hence, the success rate γ can be defined as the ratio of the number of successful trails N_{Suc} over the total number of trails N_{Tot} as the following equation

$$\gamma = \frac{N_{Suc}}{N_{Tot}}. \tag{15}$$

3.3 Problem formulation

This chapter proposes a novel contrastive learning-based SE-QPTNet for reliable beam training with low labeling cost and expertise requirements. Generally, the prediction of the optimal mmWave transmitting beam is operated at the BS side, and the same method can be easily extended to predict the optimal receiving beam. Considering severe multipath interference and inconsistent dominant beam prediction, we propose to adopt the phase quantization of periodically estimated CSI to reflect the relation of cluster AoDs/AoAs of UE and predict the optimal mmWave beam when mmWave beam training is required. Since the mmWave channels are considered to have identical LOS AoD/AoA and NLOS cluster AoDs/AoAs. For AoDs, we can rewrite the received signal (7) at the UE side as

$$\begin{aligned} y &= \sqrt{P}(\boldsymbol{H}_{LOS} + \boldsymbol{H}_{NLOS})\boldsymbol{f}_n x + \boldsymbol{n} \\ &= \sqrt{P}\boldsymbol{H}_{LOS}\boldsymbol{f}_n x + \underbrace{\sqrt{P}\boldsymbol{H}_{NLOS}\boldsymbol{f}_n x + \boldsymbol{n},}_{\boldsymbol{n}_{eq}} \end{aligned} \tag{16}$$

where the $\boldsymbol{n}_{eq}$ denotes the equivalent noise. By substituting into (4) and (16), it yields

$$y = \sqrt{\frac{N_t N_r P}{L}} \sum_{l=1}^{L} \beta_{LOS} \underbrace{\boldsymbol{a}_t^H(N_t, \boldsymbol{\phi}_{LOS})\boldsymbol{f}_n}_{correlation} x + \boldsymbol{n}_{eq}. \tag{17}$$

We can quantify the correlation between the mmWave beam in (10) and the array steering vector in (5) as

$$\begin{aligned} q_n(N_t, \phi_{LOS}) &= \boldsymbol{a}_t^H(N_t, \phi_{LOS})\boldsymbol{f}_n(\xi_{t,n}) \\ &= \frac{1}{N_t} \frac{\sin \frac{\pi N_t \psi_n}{2}}{\sin \frac{\pi \psi_n}{2}} \mathrm{e}^{j\pi \frac{N_t - 1}{2}\psi_n}, \end{aligned} \tag{18}$$

where $\psi_n = \sin \xi_{t,n} - \sin \phi_{LOS}$. The quantization $q_n(N_t, \phi_{LOS})$ illustrates the relation between angles of direction $\xi_{t,n}$ and ϕ_{LOS}, which is regarded as a quantization error [16]. However, the multipath interference and approximation of ϕ_{LOS} may lead to inaccurate predicted results.

3.4 Channel data generation with DeepMIMO platform

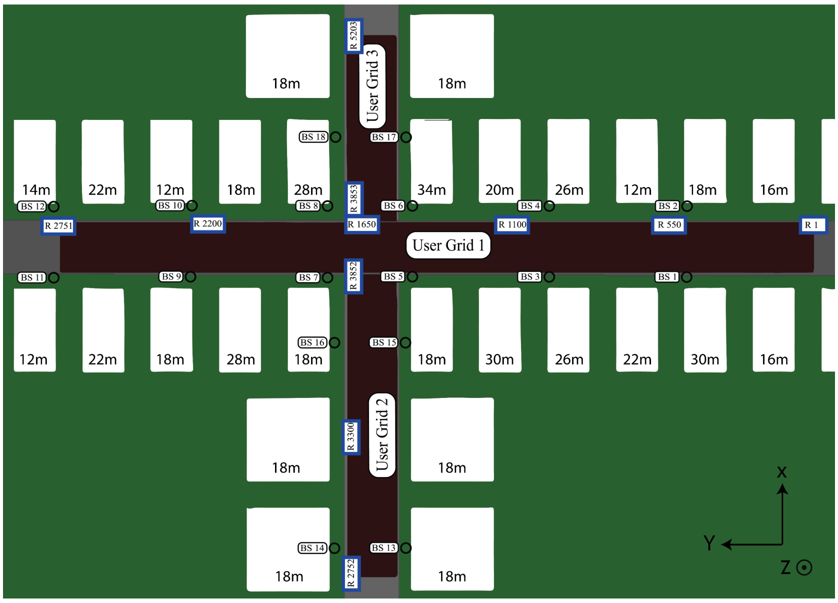


Figure 11: Overview of urban scenario in the DeepMIMO platform.

DeepMIMO is a platform for generating the mmWave massive MIMO channel data to explore various aspects, such as channel estimation, beamforming, precoding, and resource allocation, using DL algorithms to optimize and adapt the MIMO system parameters based on the characteristics of the wireless channel and user requirements. It supports various communication scenarios, including indoor, urban, and intelligent reflecting surface (IRS). We consider the urban scenario from the DeepMIMO platform [51] for data generation as shown in Fig.11. The scenario is constructed using the 3D ray-tracing software Wireless InSite [52], which captures the channel dependence on frequency and spatial domains. The mmWave signal is available at 28 GHz in an outdoor scenario, and we consider a single BS with ULA massive MIMO located at BS 3 of the scenario. We collect the mmWave channel data by the DeepMIMO generator and describe the details in Table 1.

Table 1: DeepMIMO dataset parameters

Parameters	Values
Scenario	Outdoor
mmWave (GHz)	28
Active BS	BS 3
Active UE range	Row 900 - 1300
Active UE position	100K
Number of BS antennas	64
Antenna spacing	0.5 λ
Bandwidth (GHz)	0.5
Number of OFDM subcarrier	512
OFDM sampling factor	1
OFDM limit	32
Number of LOS path	1
Number of NLOS paths	5

3.5 Concluding remark

In this chapter, we introduced the mathematical expression of mmWave massive MIMO systems, and formulated the main challenges. Specifically, we described the principle of the ray/cluster channel mode, which is widely considered in the mmWave massive MIMO systems. In addition, we illustrated the challenges of the DL-integrated beam training approaches, such as the effects of varying channel intensities and noise. Finally, we introduced the DeepMIMO platform and the details of mmWave channel data generation for the following experiments.

4 Non-deterministic quantization for beam training

In this chapter, we propose a novel binary representation of mmWave CSI to mitigate the noise and channel variations for massive MIMO wireless systems. We shed light on the low complexity and high reliability of beam training.

We first introduce a low complexity yet powerful DNN model that learns from binary quantified from the mmWave channel. The proposed method can achieve a highly reliable beam prediction and low hardware overhead, depending on the configurable DNN model in mmWave massive MIMO systems. To achieve a robust and efficient representation of CSI, we incorporate the ideas from non-deterministic quantization to convert the corrupt channel intensities into binary sequences. The non-deterministic method also allows the DNN model to compress the volume of parameters, making the proposed model easy to implement. Combining the non-deterministic scheme with the benefit of ensemble learning, we leverage the static layers to capture the multiple views of binary sequences and obtain robust beam prediction with the voting result by a wise scoring scheme.

We second propose a novel sparse feature-driven vision Transformer (SF-ViT) DL model. It inherits the advantage of binary quantization of CSI and extracts the spatial characteristics based on the semantic model from antenna indices. Specifically, we consider the peak values of the intensity as the essential components and further purify the sparse sequence by dynamic thresholds to remove the noise effects. Due to the spatial coherence of massive MIMO systems, we apply the self-attention mechanism to explore the spatial characteristics from the sparse features of antenna indices. Moreover, we also employ the wise voting scheme to improve the robustness of inference underlying the low signal-to-noise ratio (SNR) level.

4.1 Introduction

DL-integrated massive MIMO systems have brought vitality and a new vision to build intelligent communication systems for enabling high data rates and fast beam selection in the communication field [53], [54]. In practice, [55] proved that the DL model helps predict the optimal mmWave beam and correct blockage status based on the sub-6GHz CSI while reducing the beam training overhead and achieving reliable wireless systems. Moreover, the author of [56] trained the beamforming prediction network (BPNet) using the supervised and unsupervised learning method to optimize the power allocation and predict virtual uplink beamforming (VUB) for improving computational efficiency. On the other hand, [57] proposed to combine machine learning and situational awareness to predict mmWave beam power for improving the prediction performance and reducing the overhead at the cost of little performance.

Meanwhile, beam alignment refreshes every dozen milliseconds to accommodate the demands of numerous users and ensure a fast response [56]. Consequently, the DL module must predict the appropriate beam to maintain an achievable data rate within this period. Typically, the DL module is integrated into the system-on-chip (SoC) at the BS to meet the high computational requirements [58]. However, DL-integrated beam prediction schemes often utilize large-scale models to address accuracy issues caused by mmWave channel fading and noise. Integrating multiple accelerators into an SoC increases the runtime complexity. Hence, it is essential to investigate approaches for reducing the scale of DL model parameters and developing an efficient CSI representation to achieve a hardware-friendly SoC system with robust performance.

While the existing solutions exploit the DL model for optimal beam prediction, they share the following limitations. First, a regular scheme of complex mmWave channel data processing, such as the complex channel measurements, is typically divided into real and imaginary parts as input data. It makes learning efficiency sensitive to the time-variant and noisy mmWave signals. Second, the standard pre-defined methods compress the number of parameters with sparsification. The pruned connections can only be modified in advance, and incorrect pruning may cause severe accuracy loss under different channel conditions. Finally, the spatial-frequency mmWave channel measurements are vectorized as model inputs and ignore the spatial information among the antenna indices. So it needs an extensive enough DNN to obtain a high success rate in predicting the optimal mmWave beam.

To address those challenges, we develop a non-deterministic quantization encoder for describing the mmWave CSI in massive MIMO systems. It enables efficient DL model design processing and attempts to maintain reliable inference solutions with a wise scoring scheme inspired by ensemble learning. The binary sequence representation of the mmWave CSI can efficiently reduce the noise impacts and the DNN model size. Besides, we also consider exploring the inherent spatial characteristic of massive MIMO with the attention mechanism to efficiently acquire a global representation of the mmWave massive channel system from frequency and antenna indices to improve the probability of beam prediction. The main contributions of this chapter are summarized as follows:

- We develop a non-deterministic quantization for the mmWave channel to draw the relevant binary representation from unexpected noise and channel attenuation for training the DL model. This quantization relies on the peaks of mmWave channel measurements as coordinates. In addition, we infer the active components from coordinates by a non-deterministic strategy to eliminate interference. This non-deterministic method can efficiently reduce the impact of the corrupt mmWave channel on the training stage and the size of the DL model.

- We propose a DNN compression scheme to reduce the number of multiply-accumulate operations (MAC) with the binary sequences of CSI. The size of the DNN model can be effectively reduced by the binary sequence inputs, which extract through our non-deterministic quantization.

- We develop a SF-ViT model, which obtains the sparse feature (SF) based on a non-deterministic coding scheme. The proposed SF-ViT model can extract a global feature to represent the mmWave signal, considering the spatial and frequency variations to improve the reliability of beam prediction. The SF-ViT model can interactively learn the SF patterns varying in the frequency domain from the antenna indices according to the attention mechanism. The attention mechanism can analyze and capture the spatial variation of mmWave SF based on multiple factors, including frequency and a large number of antennas.

- We provide a wise scoring strategy to further improve the robustness. The performance of the DL-integrated beam training depends on the scale of a dataset and the data diversity. However, it is hard to improve the robustness by building an over-completed dataset with constantly changing CSI to improve the robustness. Besides, the training and inference stages are unbalanced since the main computation is completed in training and hardly considers the impacts of uncertainty in inference. The key idea of wise scoring is to generate the different individually and determine the final beam prediction through the scoring result to compensate for the lack of inference capability. The final beam prediction solution depends on the maximum scoring result.

4.2 Non-deterministic quantization for mmWave beam prediction

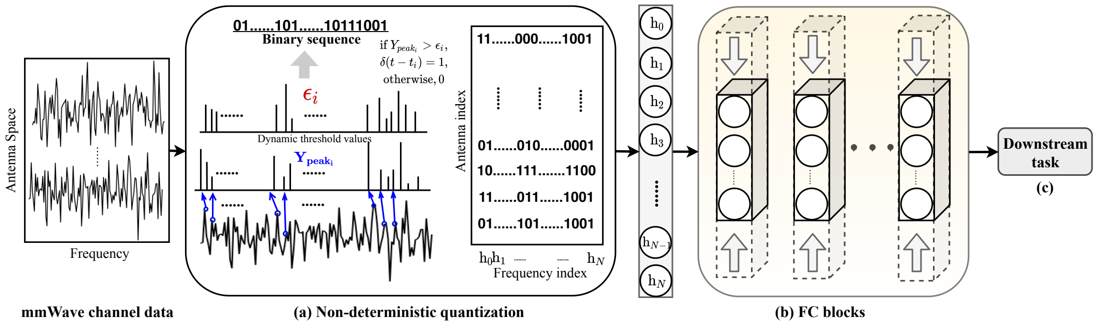


Figure 12: Overview of the proposed non-deterministic quantization for mmWave channel beam prediction.

The purpose of non-deterministic quantization is to code the mmWave channel signal for wisely representing the useful information and reduce the effects of noise as the binary sequence in [59]. As shown in Fig.12, the structure of non-deterministic quantization includes a peak value selection and a binary sequence generation. According to (4), we have the power constraint channel

$$\sum_{k=1}^{K} |\boldsymbol{h}_k|^2 = 1, \tag{19}$$

where K is the length of subcarriers for pilots in the frequency domain. Since the $\boldsymbol{h_k} \in (0, 1)$ and (19), we can obtain channel $\boldsymbol{H}$ has stochastic behavior and the value of (4) can regard as the probability of contribution for the DL training.

To generate the binary sequences for the mmWave channel, we first exploit the peak values of channel fluctuations and record the indices as p on each local antenna as (20).

$$\boldsymbol{Y}_{\text{PEAK}} = \boldsymbol{H}[p] \tag{20}$$

Secondly, the effective local values are remained by the dynamic threshold. If the values of the signal are more likely valid at the current frequency, we use 1 to represent the most likely event to keep, otherwise 0.

$$\epsilon_i \sim \mathcal{N}(0,\, \boldsymbol{\Sigma}_{sig}), \tag{21}$$

where $\boldsymbol{\epsilon}$ is the dynamic threshold,and the $\boldsymbol{\Sigma}_{sig}$ is the signal power of each antenna. Finally, the SF $\boldsymbol{s}$ can be expressed as

$$\boldsymbol{s} := \begin{cases} 1 & \text{if } \; Y_{\text{PEAK}_\text{i}} > \epsilon_i, \\ 0 & \text{else.} \end{cases} \tag{22}$$

Here the binary sequence input $\boldsymbol{h}_i = [\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_m]$, $\boldsymbol{s} \in \mathbb{R}^{N_f \times 1}$, where N_f is the length of channel measurements in frequency domain.

4.2.1 Neural network training

As shown in Fig. 12(b), the fully connected (FC) blocks are employed between the input layer and the output layer. Each FC block further includes a `dropout`$(\cdot)$ layer and the FC layer. For each FC layer, we apply the Relu activation function. In fact, the size of FC blocks directly determines the learning ability of the DNN. This consists of two fully connected networks with one rectified linear unit (`ReLU`$(\cdot)$) activation function:

$$F^1 = \texttt{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{X} + \boldsymbol{b}^{(1)}), \tag{23}$$

where $\boldsymbol{X}$ is the binary sequences as (22), which obtain from $\boldsymbol{W}^{(1)}$ is the matrix of weights and $\boldsymbol{b}^{(1))}$ is the matrix of bias in the first layer of FFN, and `ReLU`$(\cdot)$ activation function:

$$\texttt{ReLU}(x) = \max(0, x), \tag{24}$$

the equation of (23) would be express briefly as:

$$F^1 = f^{(1)}(\boldsymbol{X}; \theta^{(1)}), \tag{25}$$

where $\theta^{(1)}$ is the set of weight $\boldsymbol{W}^{(1)}$ and biases $b^{(1)}$ in the fully connected layer.Then, the output of FC blocks is represented as:

$$\boldsymbol{P} = (f^{(5)} \cdots (f^{(2)}(f^{(1)}(\boldsymbol{X}; \theta^{(1)}); \theta^{(2)}); \cdots \theta^{(5)}). \tag{26}$$

Moreover, we can apply automatic mixed precision (AMP) to further reduce the training time [60], since the model input is an integer binary sequence.

4.2.2 Wise scoring scheme

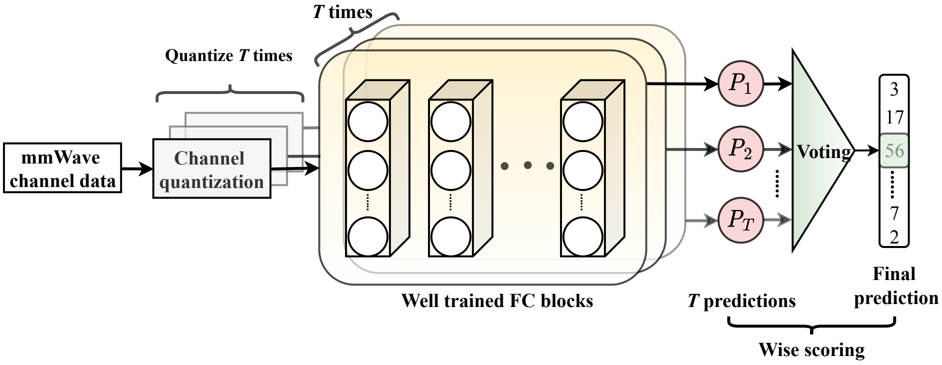


Figure 13: Illustration of the T votes wise scoring for reliable inferences.

Inspired by the benefit of ensemble learning [61], we develop a wise scoring with a voting scheme for the inference as shown in Fig 13. Specifically, the final prediction result depends on the maximum value of iterative voting as shown in Algorithm 1, where we quantize the mmWave channel signal individually with T votes to overcome the effects of fading and noise. We expect wise voting to enable the prediction to achieve a highly reliable beam prediction performance in the inference phase.

Algorithm 1 Inference based on wise scoring scheme

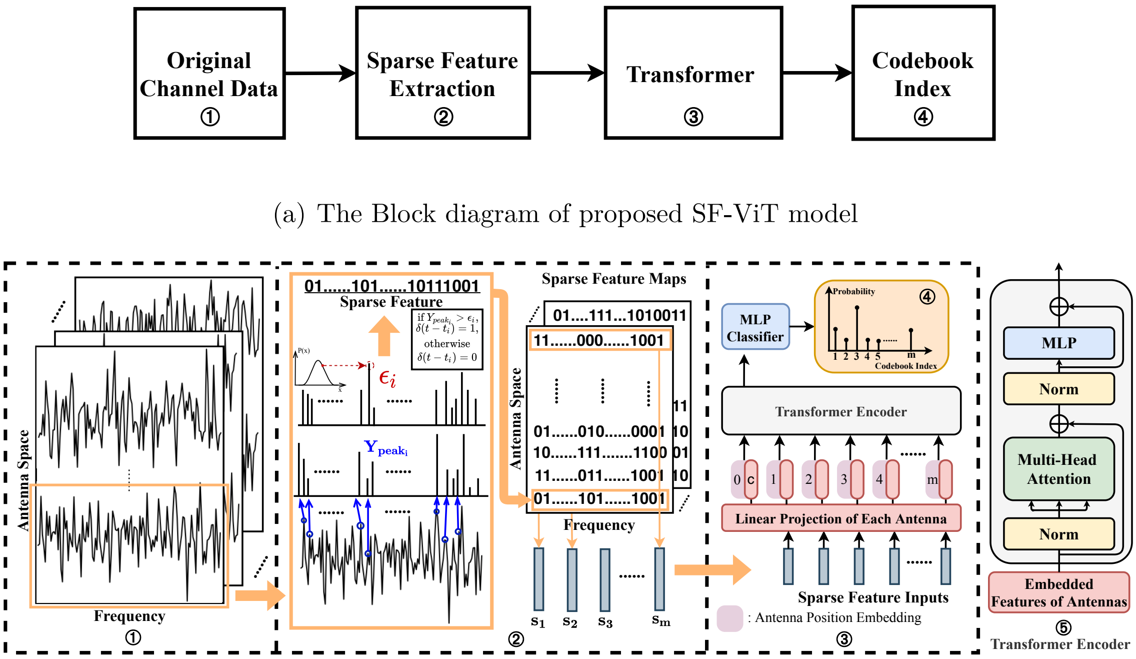
```
Input: $\boldsymbol{D}_{test}$, well trained DNN, number of votes: $T$, scoring $\mathcal{S} = \varnothing$.
repeat
    Obtain binary input $\boldsymbol{h}$ via (22).
    Obtain the probability vector $\boldsymbol{p}$ via (26).
    Obtain the index of the largest entry $i$ in $\boldsymbol{p}$.
    Obtain the prediction result $\mathcal{S} \cup \{i\}$.
until $T$ times
Output: maximum number of occurrences in $\mathcal{S}$
```

4.3 Non-deterministic sparse feature learning for reliable beam prediction

Benefiting from developing a non-deterministic quantization scheme and wise scoring, we propose a novel SF-ViT model for achieving reliable beam prediction. The proposed SF-ViT includes the operation of sparse feature (SF) extraction underlying non-deterministic quantization, spatial attention-based decision, and a wise scoring scheme for robust interference. The basic idea of the proposed SF-ViT model is shown in Fig. 14. This figure illustrates the procedure of the proposed non-deterministic SF extraction and SF-ViT model for predicting the optimal beam. Fig. 14 (a) describes the block diagram of the proposed SF-ViT model, and Fig. 14 (b) illustrates the specific operations that tackle the original channel data based on SF extraction and the SF-ViT model. Specifically, ① is the original spatial frequency channel data. ② is the SF extraction to obtain valid features using the peak value and stochastic coding from the corrupted CSI. ③ is the learning phase, in which the proposed SF-ViT model leverages the binary sequences extracted in ② on both frequency and antenna dimensions to train the SF-ViT. ④ is the probability of optimal beam. ⑤ is details of the Transformer encoder.

4.3.1 Proposed SF-ViT model

In this section, we first express the operation of SF extraction underlying non-deterministic quantization. Then we explain how to obtain the mutual feature



(a) The Block diagram of proposed SF-ViT model

(b) Overview of proposed SF-ViT contains the Transformer encoder to inquire the dense spatial diversity for reliable beam prediction

Figure 14: Procedure and overview of the proposed SF-ViT for reliable mmWave massive MIMO beam training.

from frequency and antenna space with the attention mechanism [40]. Finally, we show the main idea of the wise voting scheme, highlighting its advantages. The basic idea of the proposed SF-ViT model is shown in Fig. 14.

According to the (20),(21), and (22). We can acquire the binary representations of CSI as $\boldsymbol{s} = [\mathrm{s}_1, \mathrm{s}_2, \cdots, \mathrm{s}_m]$, $\boldsymbol{s} \in \mathbb{R}^{N_f \times 1}$, where N_f is the length of channel measurements. Firstly, we exploit the local linear projection of SF from antenna space through 1 fully connected layer. To preserve the relative distance among the antennas, the distinctive antenna position embedding parameters are employed with the linear projection results. Therefore, the linear projection of the SF can be defined as $\boldsymbol{X}_{\texttt{emb}}$ generated by the projection of $\boldsymbol{W}_{emb}^T \boldsymbol{s}$, with d_{model} dimension. The antenna position embedding (PE) is identical with dimension d_{model}, which is generated through the sine functions with different frequencies:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}), \tag{27}$$

where *pos* is the position and i is the dimension. By combing the (41) and (34), the inputs of Transformer encoder $\boldsymbol{X} = \boldsymbol{X_{emb}} + PE_{(pos,2i)}$.

The Transformer encoder enables the spatial-frequency feature to jointly learn the approximate codebook index representation by applying the stacked architectures and attention mechanism. Especially, the attention mechanism [40] enhances some parts of the frequency features while diminishing other parts by traversing all local antennas with the unique query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$, and value matrix $\boldsymbol{V}$. The Transformer encoder focuses more on mutually important mmWave channel components from frequency and local antenna, learning which part of the frequency information is more critical than others depending on variations of individual antennas. And the query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$, and value matrix $\boldsymbol{V}$ are generated by the wide fully connected layers with input signal $\boldsymbol{X}$ in the i th encoder:

$$\begin{aligned} \boldsymbol{Q}_i &= \boldsymbol{W}^{q_i}\boldsymbol{X} \\ \boldsymbol{K}_i &= \boldsymbol{W}^{k_i}\boldsymbol{X} \\ \boldsymbol{V}_i &= \boldsymbol{W}^{v_i}\boldsymbol{X}, \end{aligned} \tag{28}$$

where the $\boldsymbol{W}^{q_i}, \boldsymbol{W}^{k_i}, \boldsymbol{W}^{v_i}$ is the matrix of weights from the linear layer.

Attention operation can be described as a scaled dot-product function, which maps a query and a set of key-value pairs to an output. In particular, we compute the dot-product of query matrix $\boldsymbol{Q}$ with all key matrix $\boldsymbol{K}$, and the dimensions must be the same as d_k, and apply a $\texttt{softmax}(\cdot)$ activation to obtain the output matrix on the values matrix $\boldsymbol{V}$. Therefore, the matrix of output could be computed by (43):

$$\texttt{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \texttt{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{H}}}{\sqrt{d_k}})\boldsymbol{V}, \tag{29}$$

In order to reduce the computational overhead, the attention function could be performed in parallel with the different, learned linear projections of $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ to d_q, d_k, d_v dimensions with simple fully connected layers, respectively.

$$\begin{gathered} \texttt{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \texttt{Concat}(head_1, \cdots, head_h)\boldsymbol{W}^O \\ \text{where } head_i = \texttt{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V) \end{gathered}$$

where the projection matrices, $\boldsymbol{W}_i^Q \in \mathbb{R}^{d_{model} \times d_k}$,$\boldsymbol{W}_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $\boldsymbol{W}_i^V \in \mathbb{R}^{d_{model} \times d_v}$. In this chapter, we apply $h = 6$, which means there are 6 multi-heads in our model. For each of these we use $d_k = d_v = d_{model}/h = 64$. Due to the reduced dimension of each head, the total computational cost is similar to that of single-head attention with full dimensionality.

4.3.2 Wise voting scheme

We adopt a wise voting scheme to improve the robustness of inference suffering from the corrupt mmWave channel. Specifically, the final prediction result depends on the average of repeatedly voting, which we individually generate the SFs for overcoming the effects of the mmWave channel fading and noise. We expect that a wise voting scheme enables the beam prediction to be more believable with different statistical results, and achieve a highly reliable beam prediction performance.

4.4 Experimental results

The DNN architecture is described in Section 4.2. This neural network is trained using the binary sequences inputs, which are extracted by the non-deterministic quantization based on the mmWave channel. Specifically, the training process follows the training from scratch approach, where the weights are randomly initialized. Moreover, the dataset is split into 70% for training and 30% for testing. We use an initial learning rate of 10^{-3}, which is dropped by a factor of 0.3 with the 100 epochs. Besides, we employ the cross-entropy loss function for training the model as

$$loss = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}), \tag{30}$$

where M is the number of classes, y is the binary indicator (0 or 1) if class label c is the correct classification for observation o, and p is the predicted probability observation o is of class c. The other hyper-parameters are summarized in Table.2. All training and experiments are done in the Pytorch platform running on a machine with an RTX 3080 GPU.

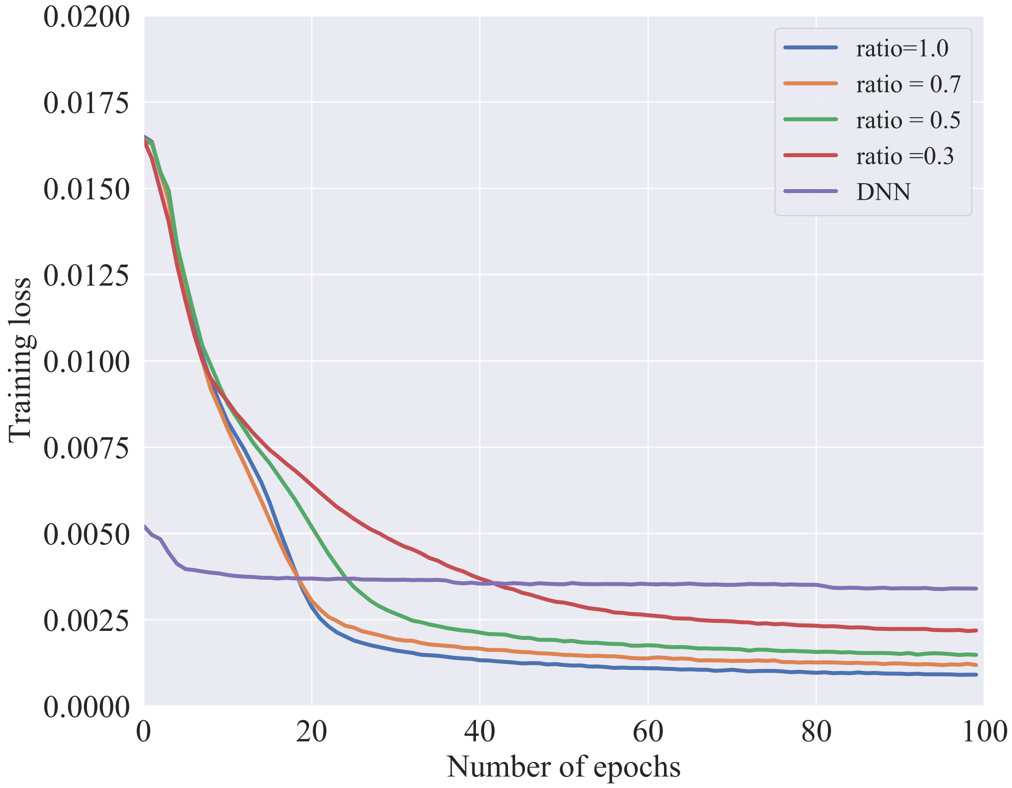


Figure 15: Training loss of the DNN and proposal with different compression ratio ρ.

Table 2: Neural network training hyperparameters

Parameters	Values
Initial dim. of DNN layers	512
Num. of DNN layers	4
Compression ratio	0.3, 0.5, 0.7, 1.0
Num. of votes	10, 30, 50, 70, 90
Optimizer	Adam
Learning rate	10^{-3}
Num. of epochs	100
Dropout percentage	30%
Dataset size	10^5
Dataset split	$70\%; 30\%$

4.4.1 Non-deterministic quantization for mmWave beam prediction

In this section, we provide numerical results to verify the effectiveness of the proposed beam prediction based on the DeepMIMO dataset. Firstly, we generate the binary sequence representation of mmWave based on the non-deterministic scheme for training a DNN and compress the parameters with a compression ratio ρ as

$$\rho = \frac{\textit{total parameters of proposal}}{\textit{total parameters of DNN}}. \tag{31}$$

The purpose of training with the proposed method is to confirm the function of binary sequence representation, and the architecture of the proposed method is totally identical to the conventional method. The DNN model in [53], is composed of 5 hidden layers, and each layer employs `ReLU`$(\cdot)$ and `dropout`$(\cdot)$. We also compare by applying different compression ratios to explore the ability of model compression of non-deterministic quantization. The performances are evaluated by training loss with the cross-entropy loss function, defined as (32). Finally, we investigate the accuracy performance between the DNN and wise scoring inference with different numbers of votes under different SNRs.

Figure 15 presents the training loss among the DNN, proposed method performance based on proposed non-deterministic quantization with different compression ratios ρ, where ρ equals 1.0, 0.7, 0.5, and 0.3. The vertical axis is the training loss values, and the horizontal axis is the training epochs. The training loss performance indicates that non-deterministic quantization can effectively explore the mmWave channel feature with binary sequences model input and the training loss value can achieve 0.0009, which is the smallest value compared with 0.003 using the original numerical channel value as model inputs, achieved by DNN. Specifically, it is shown our proposed non-deterministic quantization scheme is capable of exploring effective binary sequence representation from the mmWave channel. The noise and attenuation from the mmWave channel can also be purified by independent dynamic thresholds. Besides, we also observe model compression performance by applying different compression ratios ρ. The results show that even if we extremely compressed the DL model with 30% parameters, the training loss is still lower than the DNN.

Fig. 16 shows the test accuracy performance of the proposed method and DNN. When the compression ratio ρ is equal to 1.0, 0.7, 0.5, and 0.3, the accuracy of the proposed method can achieve 94%, 93%, 92%, 90%, and 81% compared to the DNN. The result shows that binary sequences representation of mmWave channel can approach a more stable performance and obtain a higher accuracy performance with the proposed method than DNN. This implies that the proposed method model has the potential to predict the approximate beam from the lower hardware cost while maintaining the achievable rate. This clearly illustrates the ability of the proposed Non-deterministic quantization to support the precise beam prediction based on the solutions. Furthermore, the result also shows that the proposed method using the binary representation of the mmWave channel is more stable than the DNN using the real channel amplitude values. That means binary inputs for the DL model can contribute to efficiently reducing the effects of noise and channel attenuation by relying on non-deterministic quantization. In addition, it can further improve simple and regular neural network performance.

Fig.17, Fig.18, Fig.19 shows the distribution of wise scoring performance under the different SNRs and attempts to find the optimal number of votes. We investigate the accuracy performance under different compression levels when the compression ratio ρ equals 0.7, 0.5, and 0.3. The prediction accuracy reveals that a wise scoring scheme can improve the prediction accuracy by increasing the

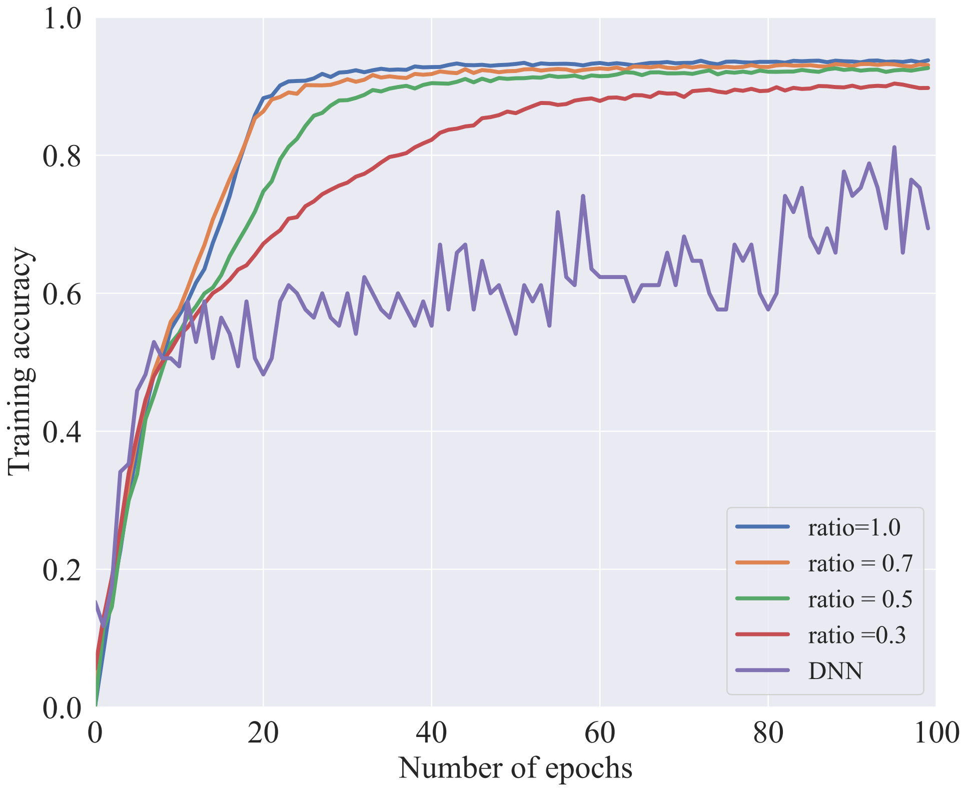


Figure 16: Test accuracy performance of proposed method and DNN.

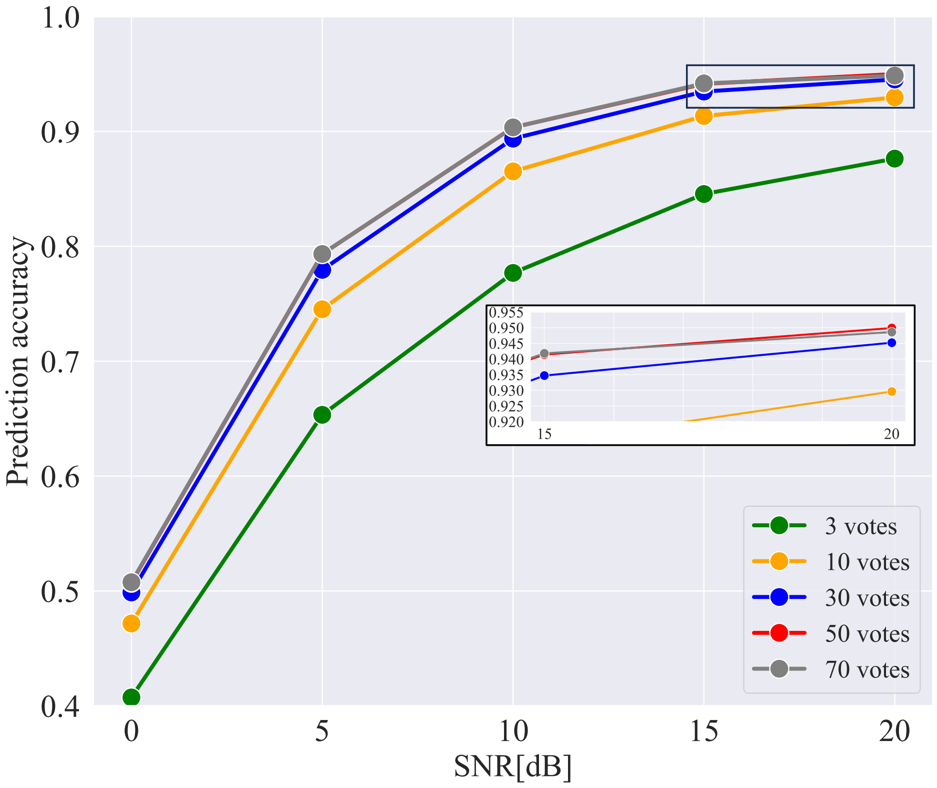


Figure 17: Prediction accuracy performance of the proposed method with 0.3 compression ratio ρ based on the wise scoring inference.

number of votes for the proposed method. According to the results, the accuracy can obtain robust inference results at 30 times, and the optimal accuracy can approach 50%, 59%, and 65% at 0 dB SNR when the compression ratio ρ is assigned different values. Moreover, the result shows that increasing the number of votes can not constantly improve prediction accuracy. Therefore, the wise scoring scheme can be equipped in SoC 30 times to approach a reliable solution with less inference time.

Fig. 20 shows the accuracy performance between DNN and the wise scoring inference with 30 votes using different compression ratios ρ under the SNR of 0-20 dB. The prediction accuracy reveals that the proposed model employs 50% and

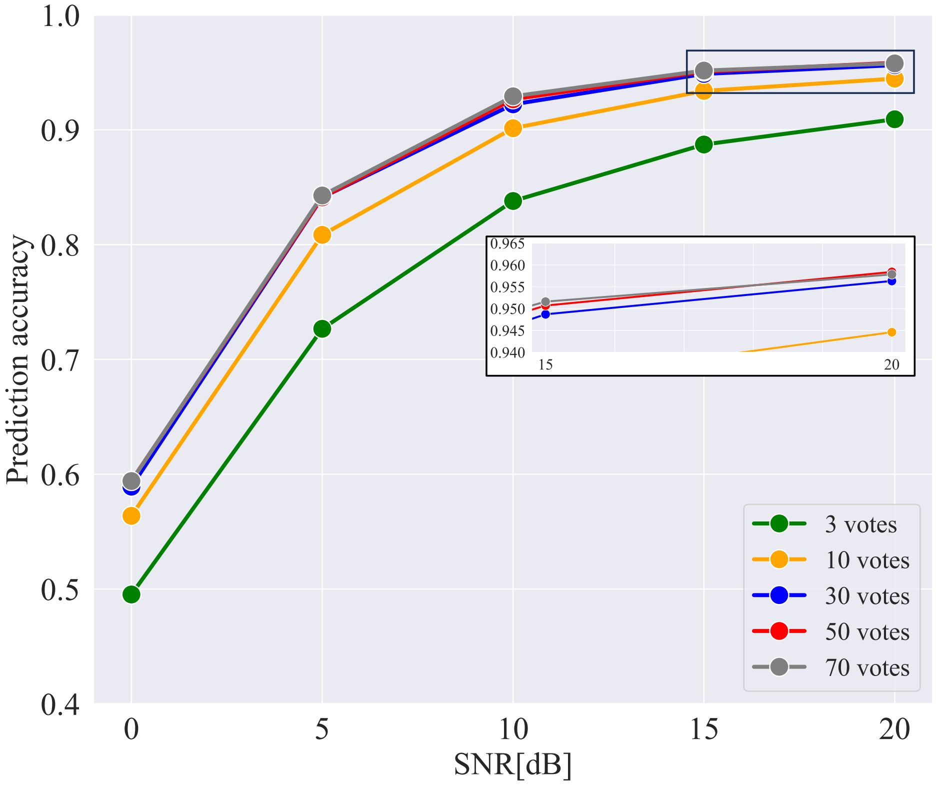


Figure 18: Test accuracy performance of the proposed method with 0.5 compression ratio ρ based on the wise scoring inference.

70% parameters, it can achieve almost the same inference accuracy 92% and 93% when SNR is larger than 10 dB. Besides, the result also shows that the accuracy of compression ratio 0.7 can approach 65%, which is better than 29% with a DNN scheme, 59% and 49% when ρ equals 0.5%, 0.3% at 0 dB SNR. Considering the trade-off between hardware cost and inference time, we can employ the 30 times-wise scoring scheme to achieve a reliable beam prediction solution with a 50% compression ratio.

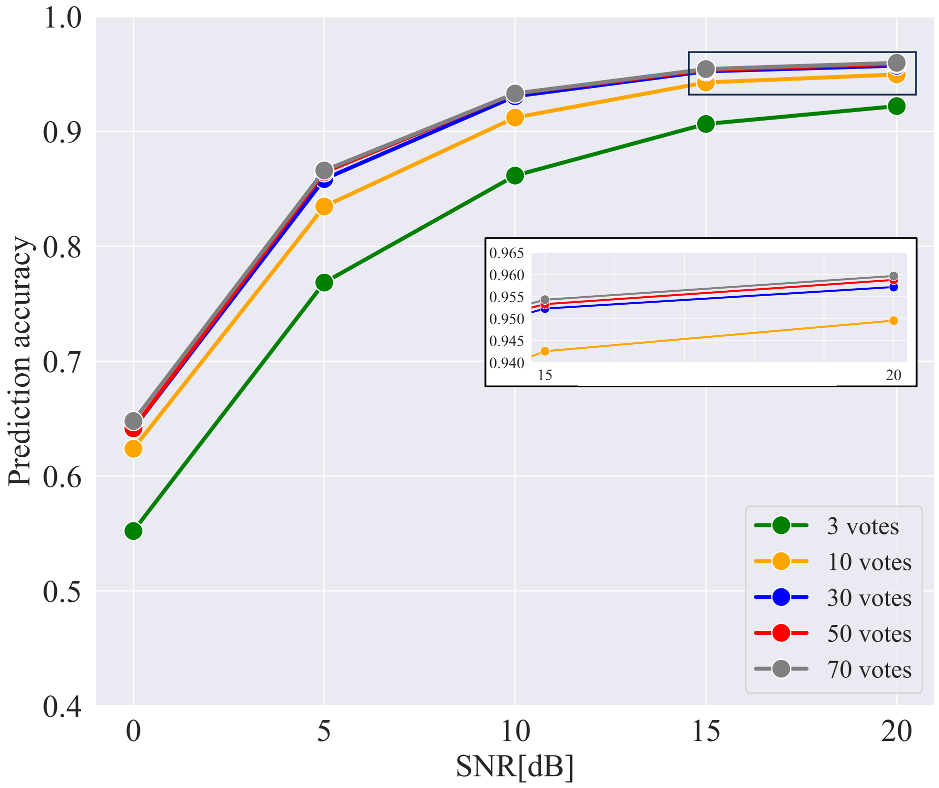


Figure 19: Test accuracy performance of the proposed method with 0.7 compression ratio ρ based on the wise scoring inference.

4.4.2 Non-deterministic sparse feature learning for reliable beam prediction

The proposed SF-ViT neural network architecture is described in Section 4.2 with embedded features of antennas and Transformer encoder. It is trained with the SF inputs, which are obtained by the non-deterministic feature extraction. Specifically, the training process follows the training from scratch approach, where the weights are randomly initialized. Moreover, the dataset is split into 70% for training and 30% for testing. We use an initial learning rate of 10^{-3} with the 100 epochs. Besides, we minimize the cross-entropy loss by adjusting the weights of

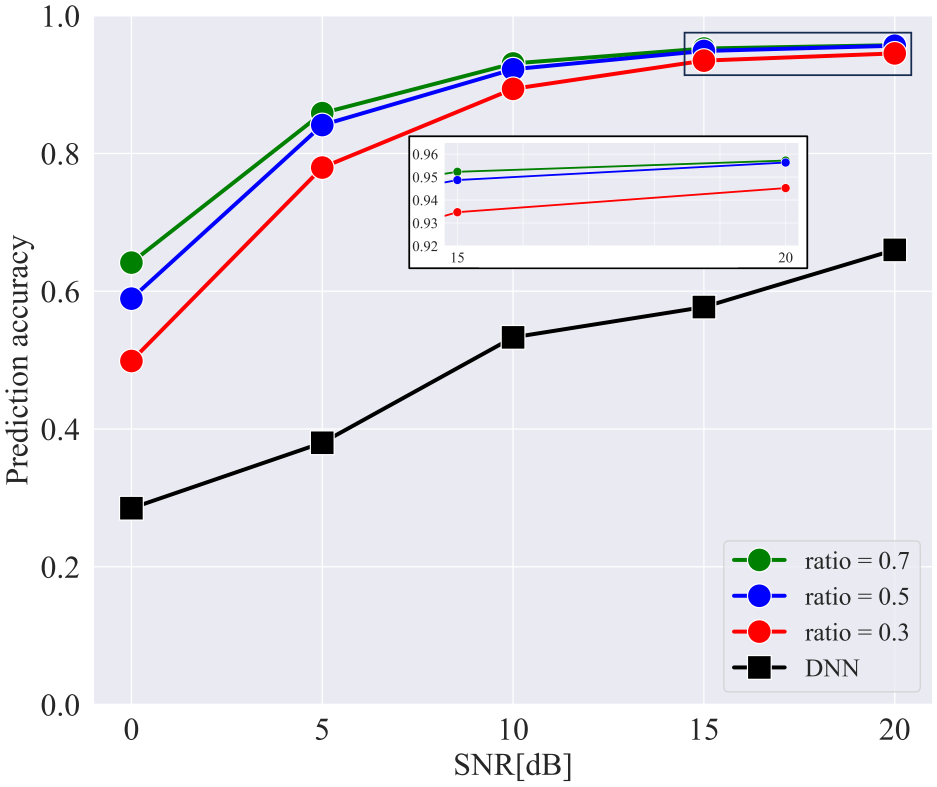


Figure 20: Test accuracy performance between DNN and the wise scoring inference with 30 votes using 0.7, 0.5, 0.3 compression ratio ρ.

SF-ViT as

$$loss = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}), \tag{32}$$

where M is the number of classes, y is the binary indicator (0 or 1) if class label c is the correct classification for observation o, and p is the predicted probability observation o is of class c. The other hyper-parameters are summarized in Table 3. All training and experiments are done in the Pytorch platform with RTX 3080 GPU.

Table 3: SF-ViT training hyperparameters

Parameters	Values
Dim. of embedded features	128
Multi-heads	6
Num. of encoder	8
Dim. of MLP	256
Optimizer	Adam
Learning rate	10^{-3}
Num. of epochs	100
Dropout percentage	20%
Dataset size	10×10^4
Dataset split	70%; 30%

In this section, we provide numerical results to validate the effectiveness of the proposed SF-ViT for beam prediction based on the DeepMIMO dataset. First, we generate SF inputs based on a non-deterministic scheme of training a normal DNN (which we define as SF-DNN) and our proposed SF-ViT. The purpose of training SF-DNN is to confirm the ability to reduce noise with the non-deterministic scheme, and the structure of SF-DNN is identical to that of the traditional method in [53]. The conventional DNN model consists of 5 hidden layers, and each layer employs `ReLU` and `drop out` operations. The performances are evaluated by training and the validation loss using the cross-entropy loss function, which is defined as (32). Finally, we investigate the accuracy performance of wise voting under different SNRs.

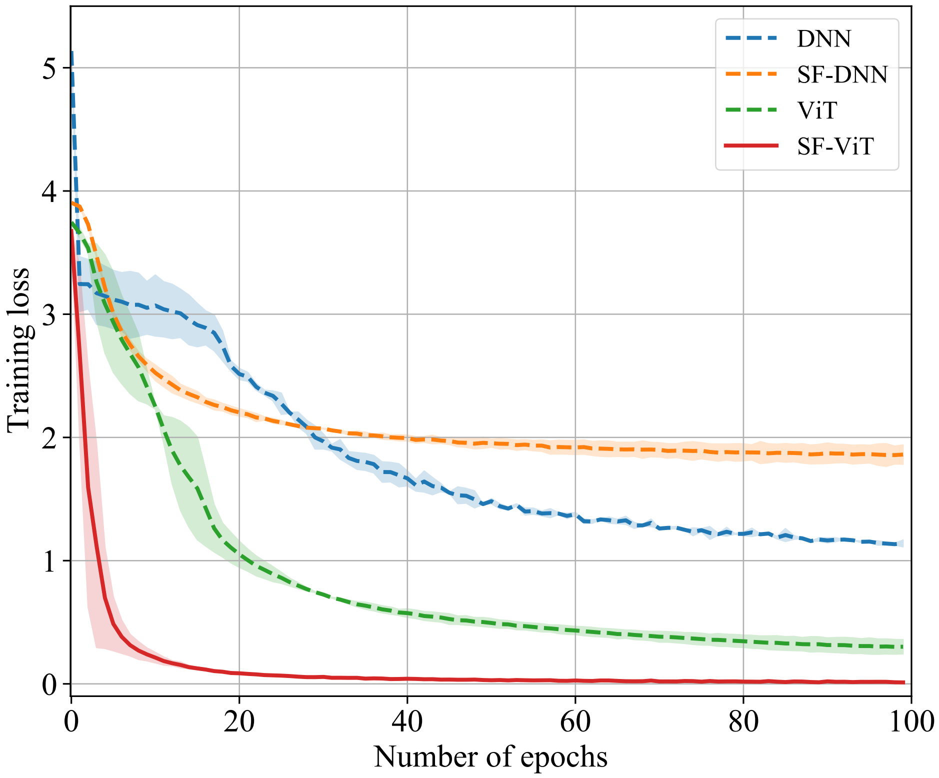


Figure 21: Training loss performance among the conventional DNN, SF-DNN, and SF-ViT.

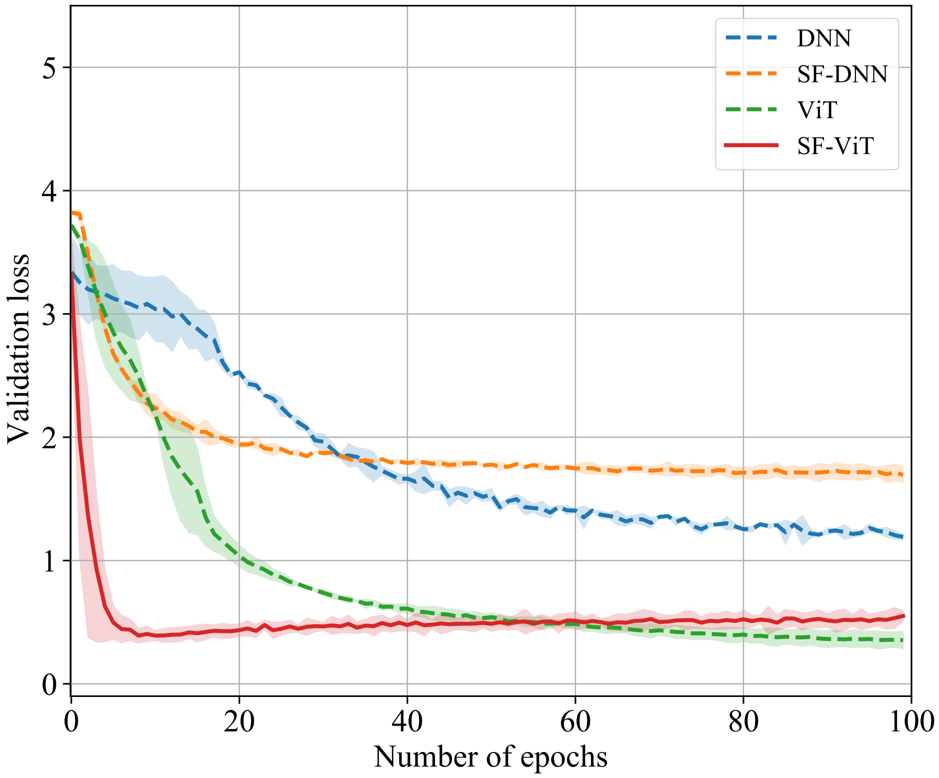


Figure 22: MSE performance of validation among the conventional DNN, SF-DNN, and SF-ViT.

Fig. 21 presents the training loss among the proposed SF-ViT, SF-DNN, ViT and conventional DNN. The Y-axis is the training loss values, and the X-axis is the training epochs. The training loss performance indicates that the SF-ViT model is the fastest convergence, and the loss value can achieve 0.2, which is the smallest value compared with 1.8, 2.1, and 0.8, achieved by DNN, SF-DNN, and ViT. Specifically, it is shown our proposed SF-ViT is capable of exploring more effective spatial features from individual antennae with the attention mechanism. The noise and attenuation from the mmWave channel can also be purified by dynamic thresholds. Besides, the fluctuation of the SF-DNN curve is lower than the conventional DNN.

Fig. 22 evaluates the generalization of DL models based on the validation data. We can observe that the loss curve of the proposed SF-ViT drops rapidly and converges to the minimum value after 20 epochs. It is shown, that the SF-ViT model can capture valid components from sparse features and learn efficiently. Moreover, the SF-DNN performance is more stable than the conventional scheme, which is trained by the original data points before 20 epochs. It proves that the sparse features, which depend on our non-deterministic feature extraction, can overcome the interference from the mmWave channel and promote the DL model performance more reliably and with better learning efficiency.

The test accuracy performances are shown in Fig. 23. This figure shows the average accuracy of the proposed SF-ViT and conventional proposals. The average accuracy of DNN, SF-DNN, ViT, and SF-ViT are 49%, 56%, 73%, and 87%. The result shows that our SF-ViT can approach the highest value among the conventional proposals. This means that the SF-ViT model is successfully predicting the codebook index from the attention mechanism with non-deterministic quantization. This clearly illustrates the ability of the proposed SF-ViT to support the precise beam prediction based on the solutions. Besides, the result also shows that SF-DNN and SF-ViT converge faster than the conventional DNN and ViT with non-deterministic quantization. That means SF inputs for the DL model can contribute to efficiently reducing the influences from noise and channel attenuation with non-deterministic quantization. Moreover, it also can improve simple and regular neural network performance.

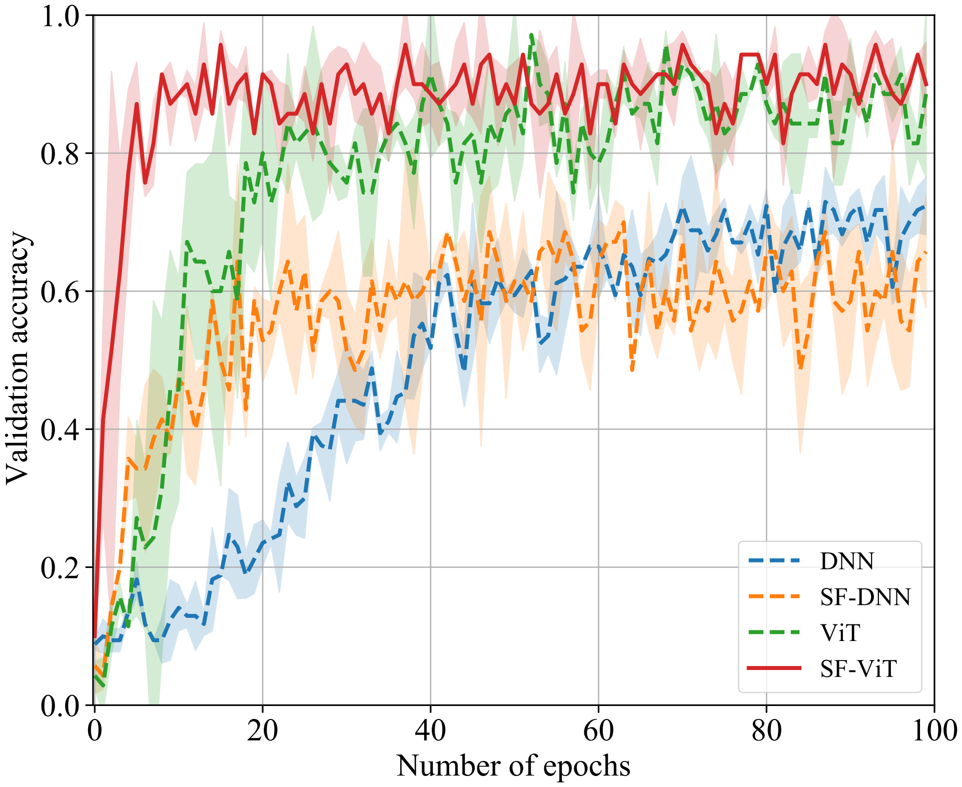


Figure 23: Validation accuracy performance among the conventional DNN, SF-DNN, and SF-ViT.

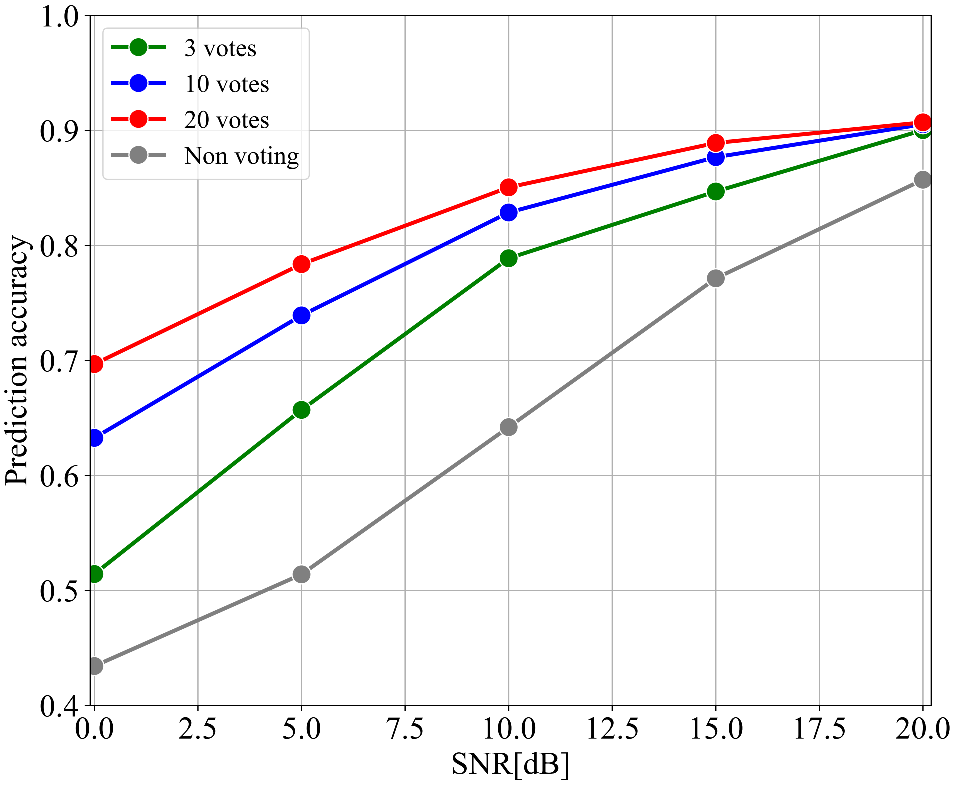


Figure 24: Test accuracy performance of SF-ViT based on the multi-round inference.

Figure. 24 shows the distribution of 3, 10, and 20 times-wise voting performance under the different SNRs. The prediction accuracy reveals that a wise voting scheme can improve the prediction accuracy from 43% to 69% under 0dB SNR with 20 votes to obtain robust inference results. Besides, the result shows that increasing the times of voting can also improve the prediction accuracy from 86% to 91% under the 20dB SNR. Therefore, the wise voting scheme can improve the reliability and robustness of inference without repetitive training.

4.5 Concluding remark

In this chapter, we developed non-deterministic encoder-based schemes for reliable beam prediction and DNN model compression with binary sequence inputs. The non-deterministic quantization enables highly reliable binary expression for mmWave CSI. The key idea of the developed strategy is to quantize the channel intensities with stochastic thresholds. Also, the binary sequence input can compress the model size to reduce the MAC elements and achieve a better performance than the conventional method. Moreover, the results demonstrated that the wise scoring scheme with 30 votes ensures high accuracy and reliable prediction results under the different SNR conditions with a compression ratio of 50%. Moreover, we also developed a sparse feature-driven DL model, which we define as SF-ViT. The proposed SF-ViT enables highly reliable and robust applications in large antenna array mmWave systems. The key idea of the developed strategy is to leverage the attention mechanism that learns from the spatial-frequency sparse features, which is obtained through a non-deterministic quantization. The results demonstrated that the proposed solution ensures high accuracy and reliable prediction results. Further, the non-deterministic quantization showed that the sparse feature can also be employed in the regular neural network, and improve the learning efficiency.

5 Hierarchical deep learning paradigm for beam training

In this chapter, we develop a novel hierarchical DL paradigm devoted to improving the learning efficiency for beam training with the frequency and spatial mmWave CSI in massive MIMO systems.

The DL-driven beam prediction is outperforming but sensitive to noise. In order to overcome the noise and the mmWave channel fading attenuation, we propose a dual DL model that learns temporal delay features based on the autoregressive model from the frequency and extracts the spatial characteristics based on the semantic model from antenna indexes. Specifically, we apply the LSTM encoder to explore the inherent temporal delay attribute of approached mmWave channel signals on the different antennas. Since the massive MIMO systems possess spatial coherence, we further use the self-attention mechanism to extract the spatial characteristics from the temporal delay features of each antenna. To verify the performance of the proposal, we compare the proposed model with conventional DL models underlying different SNR levels.

We then propose a memory-shared spatial attention neural network to further extend the capability of the hierarchical DL paradigm for mmWave massive MIMO beam prediction. The proposed method learns temporal features based on the autoregressive model from the frequency meanwhile extracting the spatial characteristics from the antenna indices. Specifically, we apply a weight-sharing LSTM to explore inherent delay features from real and imaginary mmWave channels among antenna indices. Due to the spatial coherence of massive MIMO, we further apply the attention mechanism to extract the spatial feature from the temporal features.

5.1 Introduction

Despite these tremendous successes, DL driven intelligent wireless system for mmWave massive MIMO beam prediction can achieve a high data rate with a well-trained DL model [53]. Particularly, [55] proved that DL can predict the optimal mmWave beam with low beam training overhead and obtain reliable communication. [56] proposes a beamforming prediction network (BPNet) without complex operations and iterations, which decomposes beamforming into power allocation and virtual uplink beamforming (VUB) for joint optimization, thereby improving computational efficiency. In addition to these, [57] also uses a beam learning method based on machine learning and state awareness, which greatly improves the prediction accuracy and reduces the overhead. [62] develops a machine learning-based method that utilizes location and visual data collected from wireless communication environments for fast beam prediction. The DL technology can also predict the optimal beam direction according to the low-frequency channel state information, which facilitates the initial beam training in millimeter-wave communication and reduces the overhead of beam training. [63] predicts the best narrow beam (high spatial resolution) using the less received signal strength (RSS) collected by the wide beam, which can achieve a higher data rate than the conventional hierarchical codebook design.

While the existing solutions exploit the DL model for optimal beam prediction, they share the following common limitations. Due to the noise effects and 2-dimensional performance (*frequency* and *spatial*) of the mmWave channel, the features may change dramatically once extracted from the vectorization of the mmWave channel. Firstly, millimeter waves are affected by noise in the actual environment, which often reduces the beam prediction accuracy. Secondly, in the context of the massive MIMO system, the spatial-frequency mmWave channel measurements are converted into a vector as model inputs, therefore the spatial information on the antenna space is neglected.

To break those issues, we develop a hierarchical paradigm for beam prediction, which enables the DL model to extract the feature separately from frequency and spatial view. Specifically, we incorporate an autoregressive encoder and a memory-shared LSTM encoder to extract the principle frequency features from the unexpected noise and channel fading. Moreover, the spatial coherence can be captured by the self-attention mechanism among the antenna indices.

To break these issues, we separately operate the mmWave channel from frequency and spatial domain from the autoregressive encoder and the distribution of spatial coherence from antenna indexes. The main contributions of this chapter are summarized as follows:

- We develop a two-dimensional feature extraction method to find the applicable inputs for training the DL model from frequency and spatial domain. This feature extraction can improve the learning efficiency by considering the frequency and spatial varying from the mmWave channel.

- We consider using an autoregressive model to extract principle temporal features of fluctuated mmWave channel from the unexpected noise and channel attenuation. Specifically, we apply a LSTM encoder to strain the useless components of the mmWave channel with the gate operations.

- The proposed DL model can learn from spatial and frequency space for promoting the success probability. It is able to interactively learn spatial coherence based on the frequency features from each antenna index depending on the attention mechanism. The attention mechanism is able to analyze and capture the appropriate representations for the mmWave channel.

5.2 Temporal sequence modeling with spatial attention for reliable beam prediction

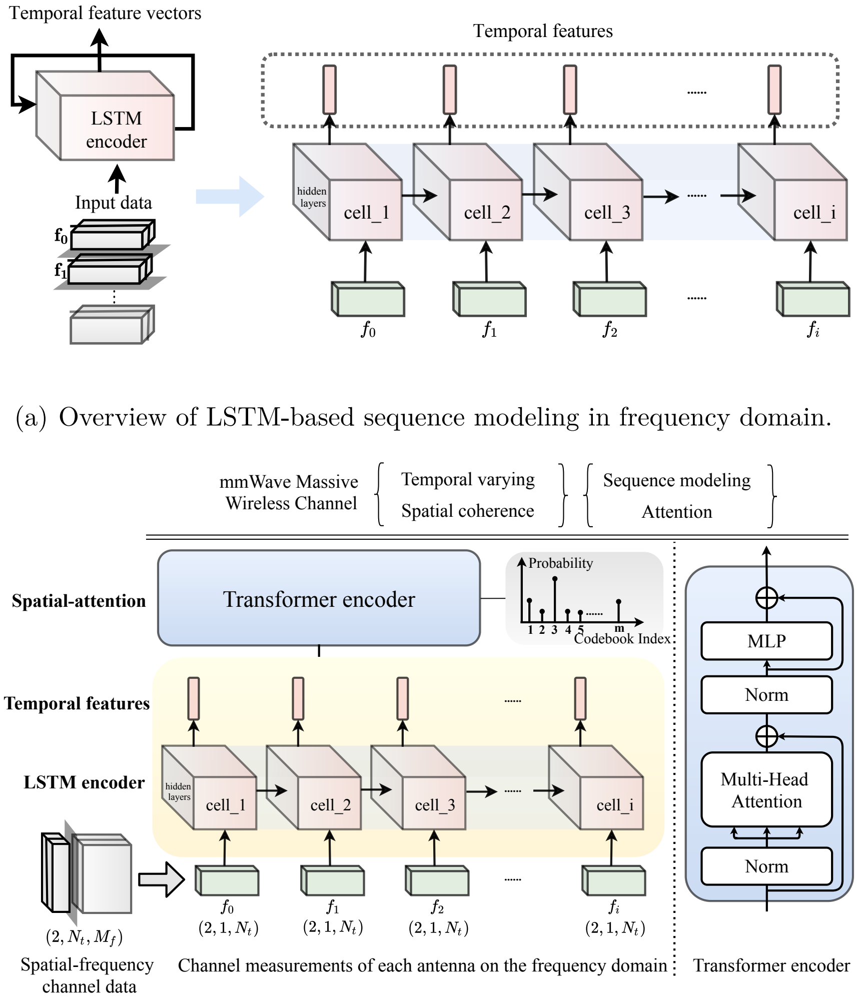
In this section, we first describe the extraction of the temporal principle feature underlying the LSTM encoder from the wildly fluctuating mmWave signals. Then we explain how to obtain the mutual feature from frequency and antenna space with the attention mechanism. The basic idea of the proposed DL model is shown in Fig. 25.

5.2.1 Long-short term memory encoder

Since the most benefit of LSTM module is to learn the relevant information by considering the past long-term dependencies [64]. We present the LSTM encoder to learn the prime temporal features from the real and imaginary parts of the complex mmWave channels.

5.2.2 Spatial attention

In this section, we explain how the proposed DL model learns the approximate spatial and frequency pattern using the attention module and the benefits of parallel computing [40]. Firstly, we exploit the local linear projection of the



(a) Overview of LSTM-based sequence modeling in frequency domain.

(b) Overview of proposal contains the LSTM encoder to extract the information from the concatenation of real and imaginary modeling in frequency domain and then inquire the spatial diversity.

Figure 25: Overview of proposed hierarchical DL paradigm

principle temporal feature from antenna indexes with a single fully connected layer. To preserve the relative distance among the antennas, the distinctive antenna position embedding parameters are employed with the linear projection results. Therefore, the linear projection of the SF can be defined as

$$\boldsymbol{X}_{\texttt{emb}} = \texttt{embedding}(\boldsymbol{s}), \tag{33}$$

where $embedding(\cdot)$ is a fully connected layer with d_{model} dimension. And the antenna position embedding is identical with dimension d_{model}, which is generated through the sine functions of different frequencies:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}), \tag{34}$$

where pos is the position and i is the dimension. By combing the (41) and (34), the inputs of Transformer encoder $\boldsymbol{X} = \boldsymbol{X}_{emb} + PE_{(pos,2i)}$.

The Transformer encoder enables the spatial-frequency feature to jointly learn the approximate codebook index representation by applying the stacked architectures and attention mechanism. Especially, the attention mechanism [40] enhances some parts of the frequency features while diminishing other parts by traversing all local antennas with the unique query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$, and value matrix $\boldsymbol{V}$. The Transformer encoder devotes more focus to mutually important mmWave channel components from frequency and local antenna, learning which part of the frequency information is more important than others depends on variations of individual antennas. The query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$, and value matrix $\boldsymbol{V}$ are generated by the wide fully connected layers with input signal $\boldsymbol{X}$ in i encoder:

$$\begin{aligned} \boldsymbol{Q}_i &= \boldsymbol{W}^{q_i}\boldsymbol{X} \\ \boldsymbol{K}_i &= \boldsymbol{W}^{k_i}\boldsymbol{X} \\ \boldsymbol{V}_i &= \boldsymbol{W}^{v_i}\boldsymbol{X}, \end{aligned} \tag{35}$$

where the $\boldsymbol{W}^{q_i}, \boldsymbol{W}^{k_i}, \boldsymbol{W}^{v_i}$ is the matrix of weights from the linear layer, the output of the multi-head attention of i encoder can be given as:

$$\boldsymbol{M}_i = \texttt{Concat}(\boldsymbol{Q}_i\boldsymbol{W}_i^{Q_i}, \boldsymbol{K}_i\boldsymbol{W}_i^{K_i}, \boldsymbol{V_i}\boldsymbol{W}_i^{V_i}), \tag{36}$$

where the $\boldsymbol{W}_i^{Q_i}, \boldsymbol{W}_i^{K_i}, \boldsymbol{W}_i^{V_i}$ is the weight matrix of projections. Then, the output of multihead would be applied to each position-wise FFN separately and independently.

Attention operation can be described as a scaled dot-product function, which maps a query and a set of key-value pairs to an output. In particular, we compute the dot-product of query matrix $\boldsymbol{Q}$ with all key matrix $\boldsymbol{K}$, and the dimensions must be the same as d_k, and apply a softmax activation to obtain the output matrix on the values matrix $\boldsymbol{V}$. Therefore, the matrix of output could be computed by (43).

To reduce the computational overhead, the attention function could be performed in parallel with the different linear projections of $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ to d_k, d_k, d_v dimensions with simple fully connected layers, which follows the (4.3.1).

In this chapter, we apply $h = 4$, which means there are 4 multi-heads in our model. For each of these we use $d_k = d_v = d_{model}/h$.

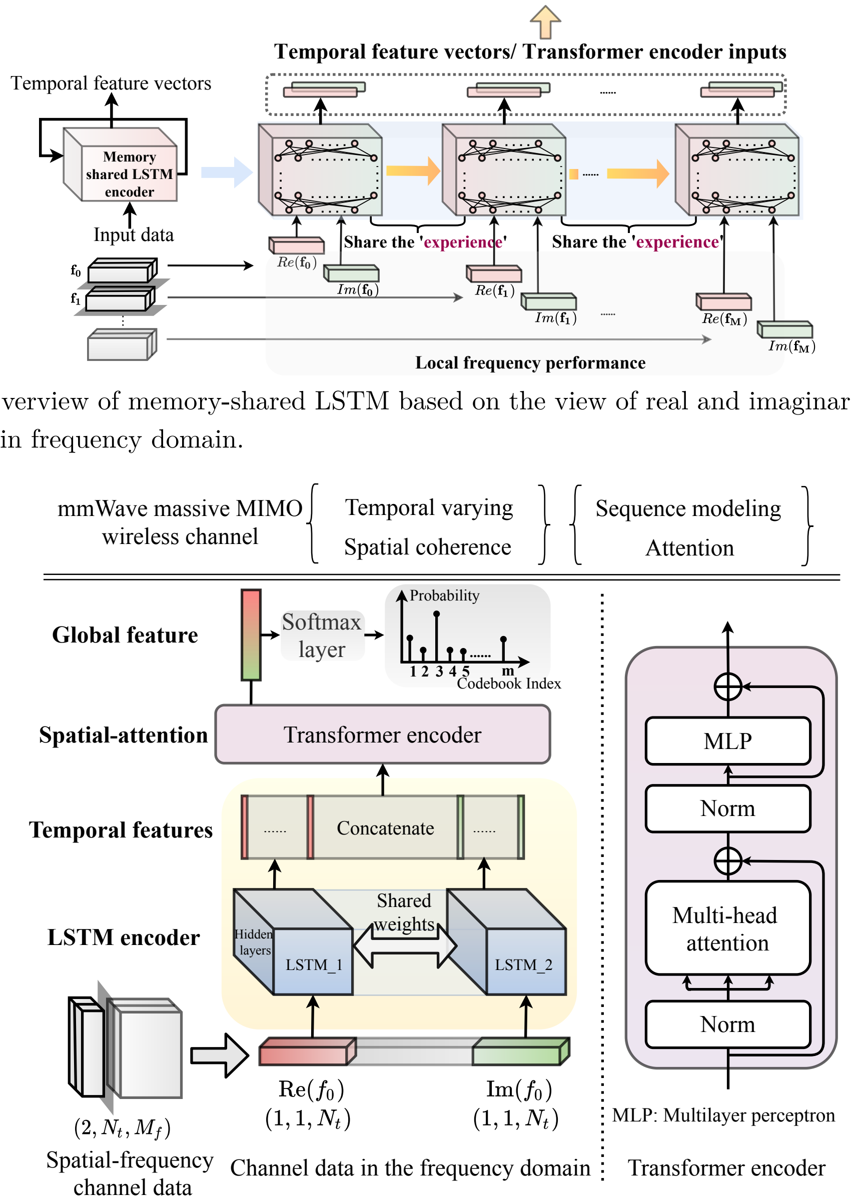
5.3 Memory shared spatial attention neural network for beam training

Despite these tremendous successes, deep learning (DL) driven intelligent wireless system for mmWave massive MIMO beam prediction can achieve a high data rate with a well-trained DL model [53]. Particularly, [55] proved that DL can predict the optimal mmWave beam with low beam training overhead and obtain reliable communication. However, due to the noise effects and 2-dimensional performance (*frequency* and *spatial*) of the mmWave channel, the features may change dramatically once extracted from the vectorization of the mmWave channel.

To break those issues, we develop a memory-shared spatial attention neural network for beam prediction as shown in Fig. 26, which enables the DL model to extract the feature separately from frequency and spatial domain. Specifically, we incorporate a shared weights LSTM encoder to extract the principle frequency features from the unexpected noise and channel fading. Moreover, the spatial coherence can be captured by the self-attention mechanism among the antenna indices.

Since the most benefit of the LSTM module is to learn the relevant information by considering the past long-term dependencies [64]. We present an LSTM encoder to learn the prime temporal features from the real and imaginary parts of the complex mmWave channel.

On the other hand, the attention operation can be described as a scaled dot-product function, which maps a query and a set of key-value pairs for the desired feature [40]. In particular, the dot-product function can be calculated by the query matrix $\boldsymbol{Q}$ with all key matrix $\boldsymbol{K}$, and apply a `softmax(·)` non-linear function to reveal the relationship of antenna indices with the values matrix $\boldsymbol{V}$.



(a) Overview of memory-shared LSTM based on the view of real and imaginary modeling in frequency domain.

(b) Overview of proposed memory-shared spatial attention network

Figure 26: Overview of memory-shared hierarchical DL paradigm

The operation can be computed by (43):

$$\texttt{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \texttt{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{H}}}{\sqrt{d_k}})\boldsymbol{V}. \tag{37}$$

To reduce the computational overhead, the attention function could be performed in parallel with the different, learned linear projections of $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ to d_k, d_k, d_v dimensions with simple fully connected layers, respectively.

$$\begin{aligned} &\texttt{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \\ &= \texttt{Concat}(head_1, \cdots, head_i, \cdots, head_N)\boldsymbol{W}^O \end{aligned} \tag{38}$$

where $head_i = \texttt{Attention}(\boldsymbol{QW}_i^Q, \boldsymbol{KW}_i^K, \boldsymbol{VW}_i^V)$, N is the number of heads, and the projection matrices $\boldsymbol{W}_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $\boldsymbol{W}_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $\boldsymbol{W}_i^V \in \mathbb{R}^{d_{model} \times d_v}$.

5.4 Experimental results

5.4.1 Temporal sequence modeling with spatial attention

The proposed DL model architecture is described in Section 3 with principal temporal features of antennas and Transformer encoder. This neural network is trained using the temporal feature inputs, which are extracted by the LSTM encoder. Specifically, we employ the cross-entropy loss function for training the model as

$$loss = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}), \tag{39}$$

where M is the number of classes, y is the binary indicator (0 or 1) if class label c is the correct classification for observation o, and p is the predicted probability observation o is of class c. The other hyper-parameters are summarized in Table.4. All training and experiments are done in the PyTorch platform with the RTX 3080 GPU.

In this section, we provide numerical results to verify the effectiveness of the proposed DL model for beam prediction based on the DeepMIMO dataset. Firstly, we drop the principle temporal features based on the LSTM encoder from the frequency domain, and the global representation of the mmWave channel can be extracted from different antenna indexes. The conventional DNN model in [53], contains 5 hidden layers, and each layer employs ReLU and drop-out operations. We also compare with the ViT model [65] as our baseline. The performances are evaluated by training loss by the cross-entropy loss function, defined as (39). Finally, we investigate the accuracy performance under different

Table 4: Proposed DL model training hyperparameters

Parameters	Beam Prediction
Dim. of embedded features	128
Num. of multi-heads	4
Num. of encoders	8
Dim. of MLP	256
Optimizer	Adam
Learning rate	$1e^{-3}$
Num. of epochs	100
Dropout percentage	20%
Dataset size	10×10^4
Dataset split	70%; 30%

SNRs. To obtain reliable results, we independently trained each model 2 times for applying the confidence interval, which is represented as the shadow regions in the result figures, to evaluate the model performance objectively.

Fig. 27 presents the training loss among the proposed DL model and regular methods. The vertical axis is the training loss values, and the horizontal axis is the training epochs. The training loss performance indicates that non-deterministic quantization can effectively explore the mmWave channel feature with binary sequences model input, and the training loss value can achieve 0.0009, which is the smallest value compared with 0.003 using the original numerical channel value as model inputs, achieved by normal DNN. Specifically, it is shown that the noise and attenuation from the mmWave channel can also be purified by the LSTM encoder.

Test accuracy performance of the proposed DL model and regular methods is shown in Fig. 28. The average accuracy of DNN, LSTM, ViT, and proposed are 48%, 55%, 72%, and 89%. The result shows that our proposal can approach the highest value among the conventional proposals. This clearly illustrates the ability of the proposed DL model to support the precise beam prediction based on the solutions. Besides, the result also shows that the proposal converges faster than the regular methods with the LSTM encoder. That means the temporal feature

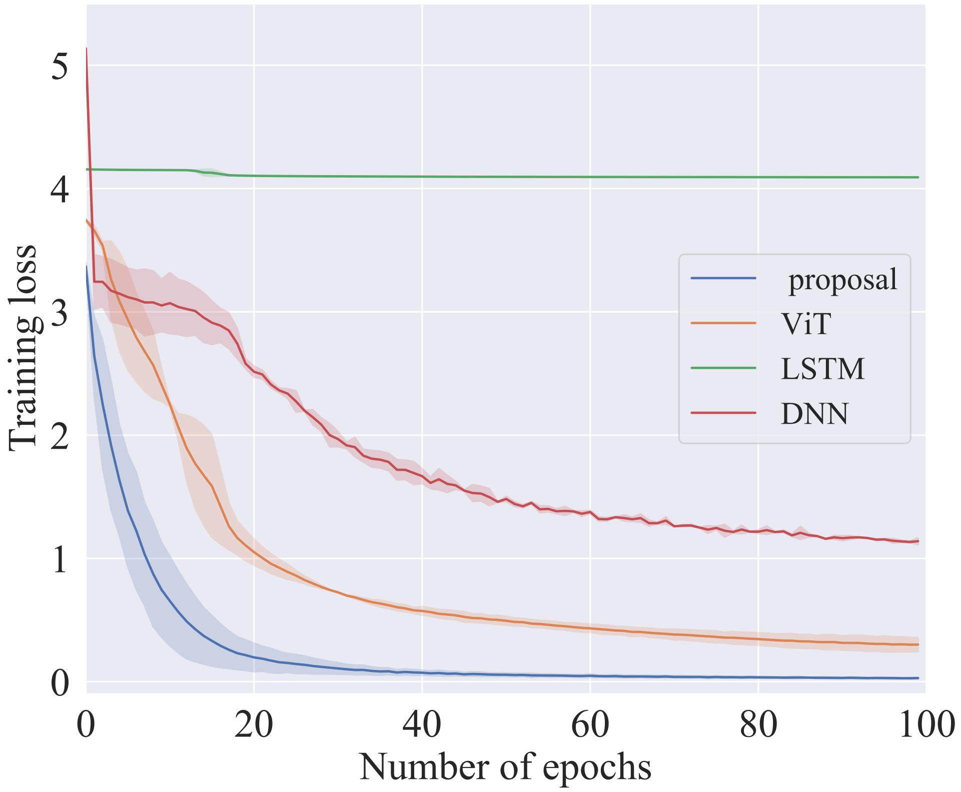


Figure 27: Training performances among the conventional DNN, LSTM, ViT, and proposed DL model.

inputs can efficiently reduce the influences from noise and channel attenuation gate operations.

Fig. 29 shows the prediction accuracy performance under the different SNRs. The prediction accuracy reveals that the proposed DL model can improve the prediction accuracy from 23% to 58% under 0dB SNR. Besides, the result also shows that the proposed DL model can be more stable, and the accuracy is 91% under the 20dB SNR. Therefore, the reliability and robustness of inference can be improved by simultaneously learning from frequency and spatial domain.

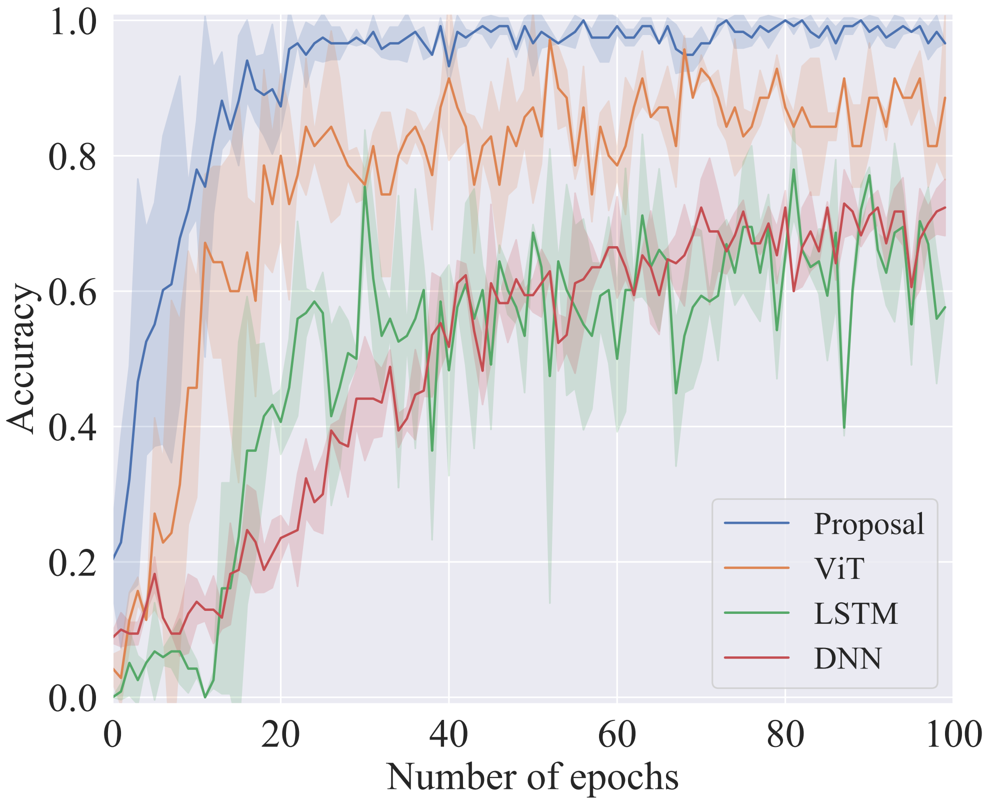


Figure 28: Test accuracy performances among the conventional DNN, LSTM, ViT, and proposed DL model.

5.4.2 Memory shared spatial attention neural network

The publicly available scenario of DeepMIMO [51] is considered as outdoor channel generation for the orthogonal frequency division multiplexing (OFDM) system. The detail of the parameters is described in Table.5.

Fig. 30 shows the training loss comparison among the DNN [53], spatial attention [40], spatial attention with unshared weight LSTM encoder, and proposed model. The vertical axis is the training loss values, and the horizontal axis is the training epochs. The training loss performance indicates that non-deterministic quantization can effectively explore the mmWave channel feature with binary se-

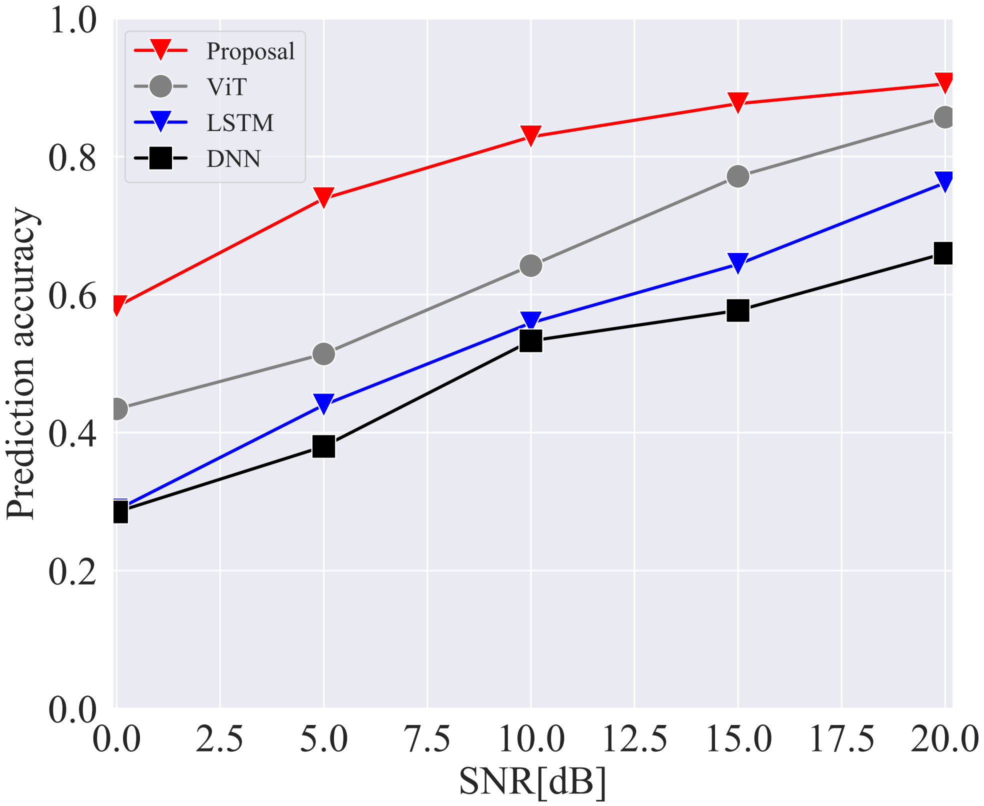


Figure 29: Prediction performances among the conventional DNN, LSTM, ViT, and proposed DL model under different SNR.

quences model input, and the training loss value can achieve 0.0009, which is the smallest value compared with 0.003 using the original numerical channel value as model inputs, achieved by normal DNN. Specifically, it is shown that the noise and attenuation from the mmWave channel can also be purified by the LSTM encoder. Fig. 30(a) presents the training loss and the loss value of our proposed memory-shared dual DL model that can rapidly achieve 0.021 after 100 epochs.

Test accuracy performance of the proposed DL model and regular methods is shown in Fig.31. The average accuracy of DNN, LSTM, ViT, and proposed are 48%, 55%, 72%, and 89%. The result shows that our proposal can approach the highest value among the conventional proposals. This clearly illustrates the

Table 5: DeepMIMO dataset parameters.

Parameters	Values
Carrier frequency (GHz)	28
Wavelength λ (mm)	10.7
Active BS	BS 3
Num. of active users	100K
Num. of BS antennas	64
Antenna spacing	0.5λ
Bandwidth (GHz)	0.5
Num. of OFDM subcarriers	512
OFDM limit	32
Num. of paths	5

ability of the proposed DL model to support the precise beam prediction based on the solutions. Besides, the result also shows that the proposal converges faster than the regular methods with the LSTM encoder. That means the temporal feature inputs can efficiently reduce the influences from noise and channel attenuation gate operations. Fig. 31 shows the mean accuracy of our proposal is the best and can approach 96%.

Fig.32 shows the prediction accuracy performance under the different SNRs. The prediction accuracy reveals that the proposed DL model can improve the prediction accuracy from 23% to 58% under 0dB SNR. Besides, the result also shows that the proposed DL model can be more stable, and the accuracy is 91% under the 20dB SNR. Therefore, the reliability and robustness of inference can be improved by simultaneously learning from frequency and spatial domain. Furthermore, the memory-shared proposal can obtain the most reliable prediction result from 0 to 20 dB in Fig. 32.

5.5 Concluding remark

In this chapter, we developed a temporal sequence modeling with a spatial attention scheme for reliable beam prediction with the LSTM encoder and atten-

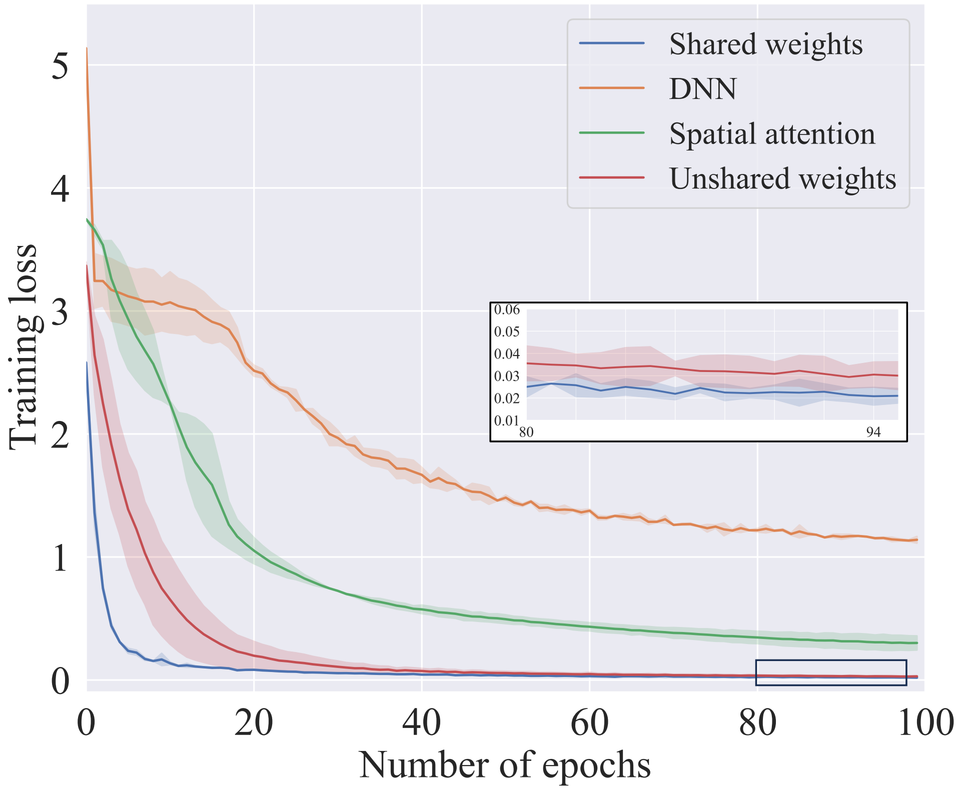


Figure 30: Training performances among the conventional DNN, unshared weight LSTM encoder, spatial attention, and proposed shared weights DL model.

tion mechanism. The LSTM encoder enables the extraction of the prime temporal features from the unexpected channel effects and noise. The key idea of the developed strategy is to leverage the attention mechanism that learns from spatial coherence based on temporal features from different antenna indexes. The results demonstrated that the proposed solution ensures high accuracy and reliable prediction performance.

Furthermore, we developed a shared weight spatial attention neural network to achieve a highly reliable and robust beam prediction for the mmWave massive MIMO system. The shared weight LSTM encoder extracted the joint frequency feature and reduced the impacts of unexpected noise and channel fading. Addi-

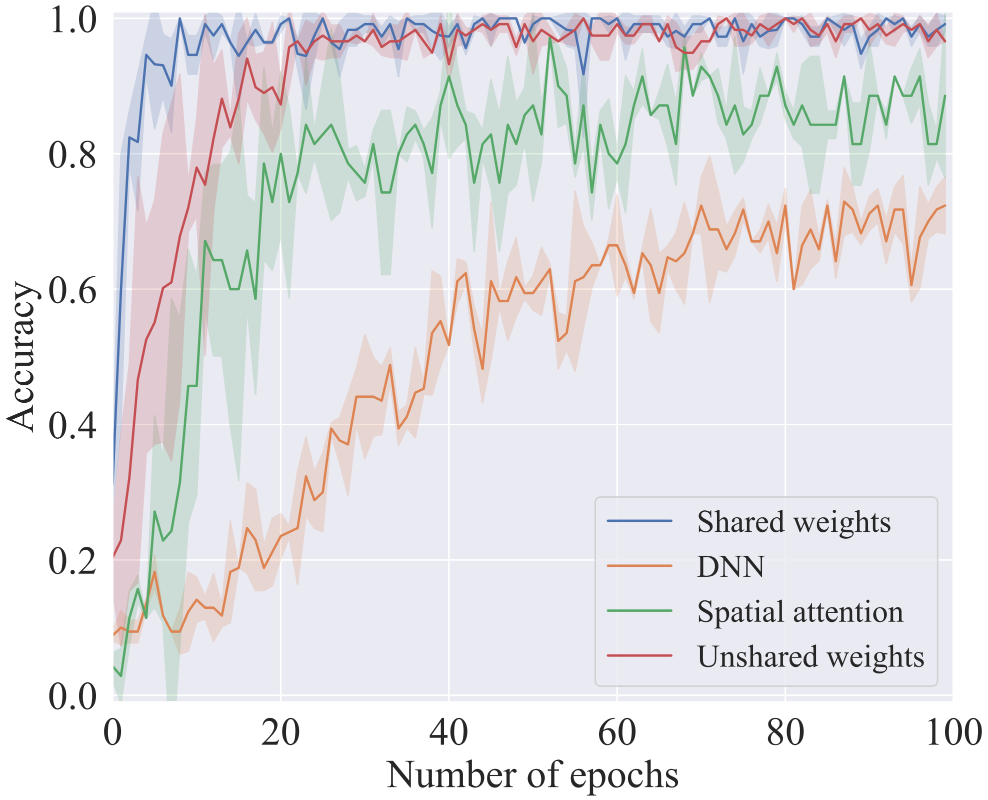


Figure 31: Beam prediction accuracy among the conventional DNN, LSTM, spatial attention, and proposed DL model.

tionally, the global feature was explored by the spatial attention from the joint frequency feature on each antenna index. Finally, the results demonstrated that the proposed solution ensures high accuracy and reliable prediction.

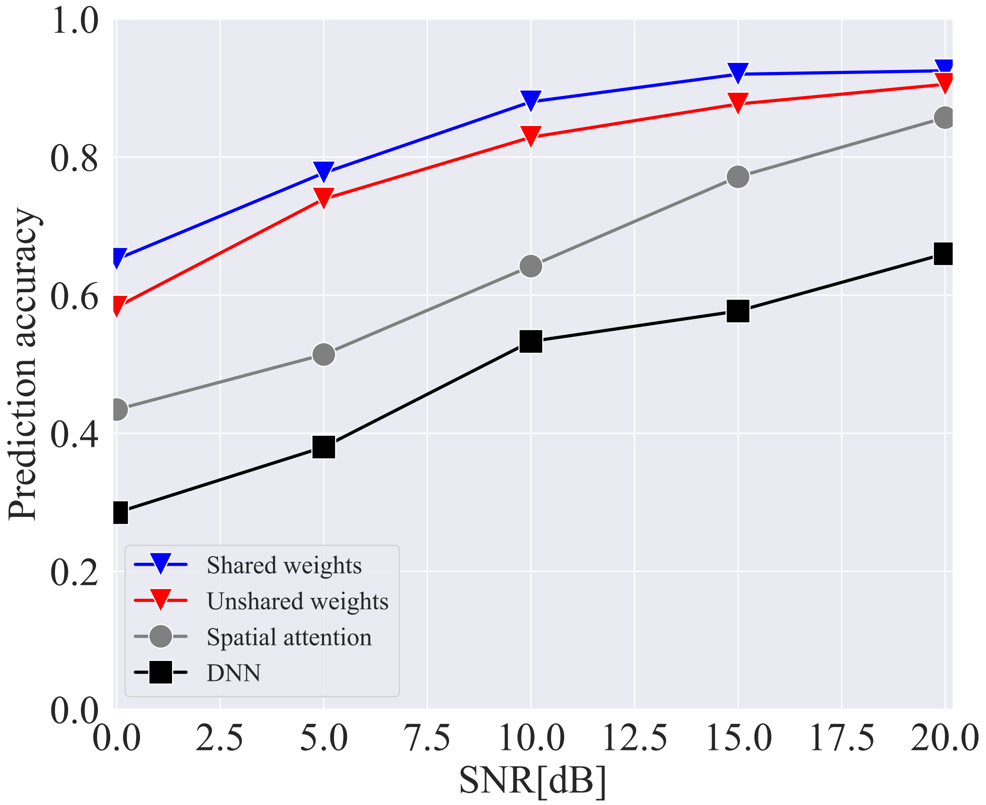


Figure 32: Beam prediction accuracy among the conventional DNN, unshared weight LSTM encoder, spatial attention, and proposed shared weights DL model with different SNRs.

6 Self-supervised learning for beam training

In this chapter, we extend the SL framework-based beam training to a self-supervised learning framework with low labeling cost and expertise requirements. We propose a novel contrastive learning framework based on our previous hierarchical DL paradigm for practical wireless systems.

As shown previously, DL-based beam training schemes have been exploited to improve spectral efficiency with fast optimal beam selection for mmWave massive MIMO systems. To achieve high prediction accuracy, these DL models rely on training with a tremendous amount of labeled environmental measurements, such as mmWave CSI. However, demanding a large volume of ground truth labels for beam training is inefficient and infeasible due to the high labeling cost and the requirement for expertise in practical mmWave massive MIMO systems. Meanwhile, a complex environment incurs critical performance degradation in the continuous output of beam training. This chapter proposes a novel contrastive learning framework, named self-enhanced quantized phase-based Transformer network (SE-QPTNet), for reliable beam training with only a tiny fraction of the labeled CSI dataset. We first develop a quantized phase-based Transformer network (QPTNet) with a hierarchical structure to explore the essential features from frequency and spatial views and quantize the environmental components with a latent beam codebook to achieve robust representation. Next, we design the SE-QPTNet, including self-enhanced pre-training and supervised beam training. SE-QPTNet pre-trains by the contrastive information of the target user and others with the unlabeled CSI, and then it is utilized as the initialization to fine-tune with a reduced volume of labeled CSI.

6.1 Introduction

This chapter proposes a novel contrastive learning framework, named self-enhanced quantized phase-based Transformer network (SE-QPTNet), in mmWave massive MIMO systems to enable reliable beam training with CSI measurements underlying limited labels. We first design a hierarchical DL architecture, named quantized phase-based Transformer network (QPTNet), which sequentially extracts frequency and spatial-domain features to enable effective representation.

In order to perform a reliable spatial representation, we also develop a latent beam codebook to align the similar phase features of the frequency domain for exploring the environmental components. Then we propose the contrastive learning framework SE-QPTNet extended from the QPTNet architecture. The SE-QPTNet framework is concerned with detecting the relationship between the global beam signature and the distinctive CSI corresponding to user locations using contrastive environmental prediction. By leveraging the contrastive information of the unlabeled CSI, SE-QPTNet is pre-trained as the initialization for fine-tuning with a limited amount of labels to provide more accurate performance. The main contributions of this chapter can be summarized as follows:

- We propose QPTNet, a hierarchical DL architecture with two levels, to enable effective representation of CSI in frequency and spatial domains. The first level performs gate recurrent unit (GRU) to model and extract the dependencies among clusters information in the frequency domain. The second level utilizes the spatial attention mechanism to extract a global beam signature based on the results of quantization.

- We design a codebook-based phase quantization to explore the complicated environmental components for reliable spatial representation. This quantization method can match and aggregate similar phase features with a codeword, improving the feature robustness from the effect of noise in the CSI. It converts the prior knowledge of continuous spatial features extracted from the frequency domain into the posterior knowledge of categorical beams.

- We develop a contrastive learning framework SE-QPTNet benefiting from the hierarchical QPTNet and codebook-based phase quantization. This enhanced model further improves beam training accuracy with limited labeled CSI. To the best of our knowledge, this is the first study that introduces contrastive learning in beam training applications. SE-QPTNet performs two benefits based on contrastive environmental prediction. Firstly, it can pre-train without any label information by detecting the relationship between the global beam feature and positive/negative samples. Secondly, the similarity of a positive sample and beam signature can effectively capture the spatial dynamic changes under long inter-frequency spans. SE-QPTNet preserves the benefits of QPTNet and reduces the labeling cost.

6.2 Spatial attention associated beam prediction

This section presents the efficient spatial attention associated beam prediction, depicted in Fig. 33. Efficient perception of the environment from the observed CSI can significantly improve beam prediction accuracy. Benefiting from the attention mechanism, we can infer the optimal beam response from implicit directions by its attention scoring [40]. Moreover, the attention mechanism is also good at simultaneously capturing features from entire implicit directions for a comprehensive beam signature.

Beam gain generation module: Since the initial spatial-frequency channel measurement $\boldsymbol{H}$ is complex-valued, we firstly convert normalized $\boldsymbol{H}$ into real-valued $\tilde{\boldsymbol{H}}$ and normalize with the maximum amplitude of its elements:

$$\tilde{\boldsymbol{H}} = [\Re(\frac{\boldsymbol{H}}{\|\max(\boldsymbol{H})\|})\Im(\frac{\boldsymbol{H}}{\|\max(\boldsymbol{H})\|})]. \tag{40}$$

Then real-valued $\tilde{\boldsymbol{H}}$ concatenates the real and imaginary components in the spatial domain to simultaneously generate the beam gain. Finally, the input modified channel can be denoted as $\tilde{\boldsymbol{H}} \in \mathbb{R}^{N_f \times 2N_t}$.

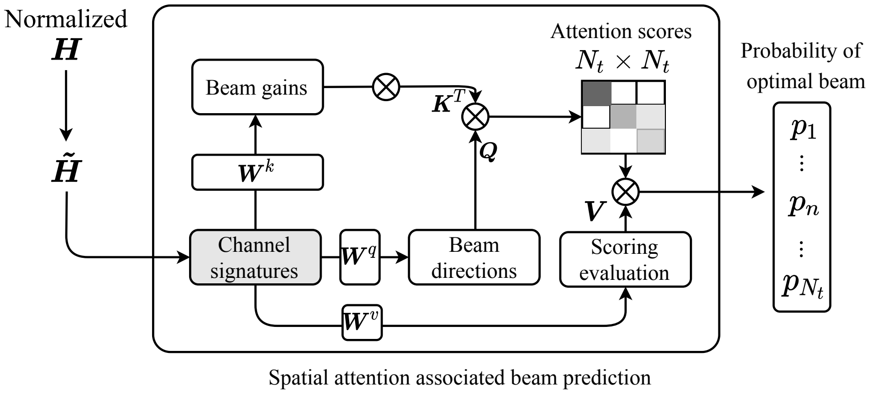


Figure 33: Overview of spatial attention associated beam prediction.The candidate beam responses can be performed by calculating the attention score with the beam gains and directions representation based on the channel signatures. And the predicted beam is decided by a scoring evaluation choosing the high pairing probability of beam gains and direction.

To generate the transmitting beam gain in each antenna direction, we exploit the linear projection of $\tilde{\boldsymbol{H}}$ to antenna space N_t through an embedding layer. To preserve the relative distance among the antennas, the distinctive beam gain embedding parameters employ with the linear projection results. The transmitting beam gains of $\tilde{\boldsymbol{H}}$ can be defined as

$$\boldsymbol{B}_g = \texttt{embedding}(\tilde{\boldsymbol{H}}), \tag{41}$$

where $\texttt{embedding}(\cdot)$ is a linear projection layer with $2N_t \times N_t$ transmit antenna dimension, and the size of $\boldsymbol{B}_g \in \mathbb{R}^{N_f \times N_t}$.

Beam direction tagging module: To symbolize the inherent direction for the beamforming gains, we consider a beam direction tagging $\boldsymbol{B}_t$ for each transmitting beam gain in (41), which is generated through the linear projection $\boldsymbol{B}_t \in \mathbb{R}^{N_f \times N_t}$ with each transmitter antenna index. By combing the (41), the inputs of Transformer encoder $\boldsymbol{X} = \boldsymbol{B}_g + \boldsymbol{B}_t$.

Stacked attention module: The Transformer encoder enables the spatial-frequency feature to deeply learn essential representation by applying the stacked attention module. Especially, the attention mechanism [40] may capture the relevant relation between the specific LOS/dominate path direction while diminishing the effects of NLOS/subordinate by traversing all transmission antennas with the beam direction query matrix $\boldsymbol{Q}$, beam gains key matrix $\boldsymbol{K}$, and corresponding scoring value matrix $\boldsymbol{V}$. The stacked attention module devotes more focus to mutually important alignment degrees from the coherence of candidate beam gains and the beam directions of the local antenna, learning which specific AoD information is more competitive than others, depending on the limited number of scatters of LOS and NLOS paths. Then the query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$, and value matrix $\boldsymbol{V}$ are generated by the wide fully connected (FC) layers with input signal $\boldsymbol{X}$ as

$$\begin{aligned}\boldsymbol{Q} &= \boldsymbol{W}^{q_i}\boldsymbol{X},\\ \boldsymbol{K} &= \boldsymbol{W}^{k_i}\boldsymbol{X},\\ \boldsymbol{V} &= \boldsymbol{W}^{v_i}\boldsymbol{X},\end{aligned} \tag{42}$$

where the $\boldsymbol{W}^{q_i}, \boldsymbol{W}^{k_i}, \boldsymbol{W}^{v_i}$ are the linear projection layers in i th attention module. Attention operation can be introduced as a scaled coherence function in Fig. 13, which maps a beam direction query and a bunch of candidate beam gains pairs as a dependency relation. Specifically, we compute the dot-product of beam direction query matrix $\boldsymbol{Q}$ with all key beam gains matrix $\boldsymbol{K}$ and apply a `softmax`$(\cdot)$ activation to obtain the pairing probabilities on the scoring value matrix $\boldsymbol{V}$, so that beam gains can be precisely aligned with LOS direction. Then the output can be computed by

$$\texttt{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \texttt{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{N_t}})\boldsymbol{V}. \tag{43}$$

The candidate pairs context of the beam gains and direction can be extracted by the forward stacked attention process and decided with the relevant score, which comes from the softmax probability result. Moreover, the superior collaboration of beam gains and direction can effectively improve the beam training performance.

Output module: Different from the dense FC layer based-prediction scheme, the global average pooling (GAP) in [66], is introduced to implement the average of each token from stacked attention operations to the candidate beams which can be written as

$$\boldsymbol{c}_t = \frac{1}{D}\sum_{d=1}^{D}\boldsymbol{c}_d, \tag{44}$$

where $\boldsymbol{c}_d \in \mathbb{R}^{N_t \times 1}, d = 1, 2, \cdots, D$ is the output from the stacked attention module, and the size of *beam signature vector* $\boldsymbol{c}_t \in \mathbb{R}^{N_t \times 1}$. One advantage of GAP over the FC layer is that it summarizes the aligned information between beam gains and transmitting direction with the weighted average results, thus it is more robust to spatial translations of the fusion local features of gains and direction. In addition, there are no extra parameters to optimize in the GAP layer, and overfitting can be avoided. The resulting vector $\boldsymbol{c}_t$ is directly fed into the $\texttt{softmax}(\cdot)$ layer and the optimal beam can be chosen with the maximum probability of the global beam signature $\boldsymbol{c}_t$ is

$$\begin{aligned} \boldsymbol{p} &= \texttt{softmax}(\boldsymbol{c}_t), \\ n^* &= \underset{n \in \{1,2,\ldots,N_t\}}{\arg\max} \; p_n, \end{aligned} \tag{45}$$

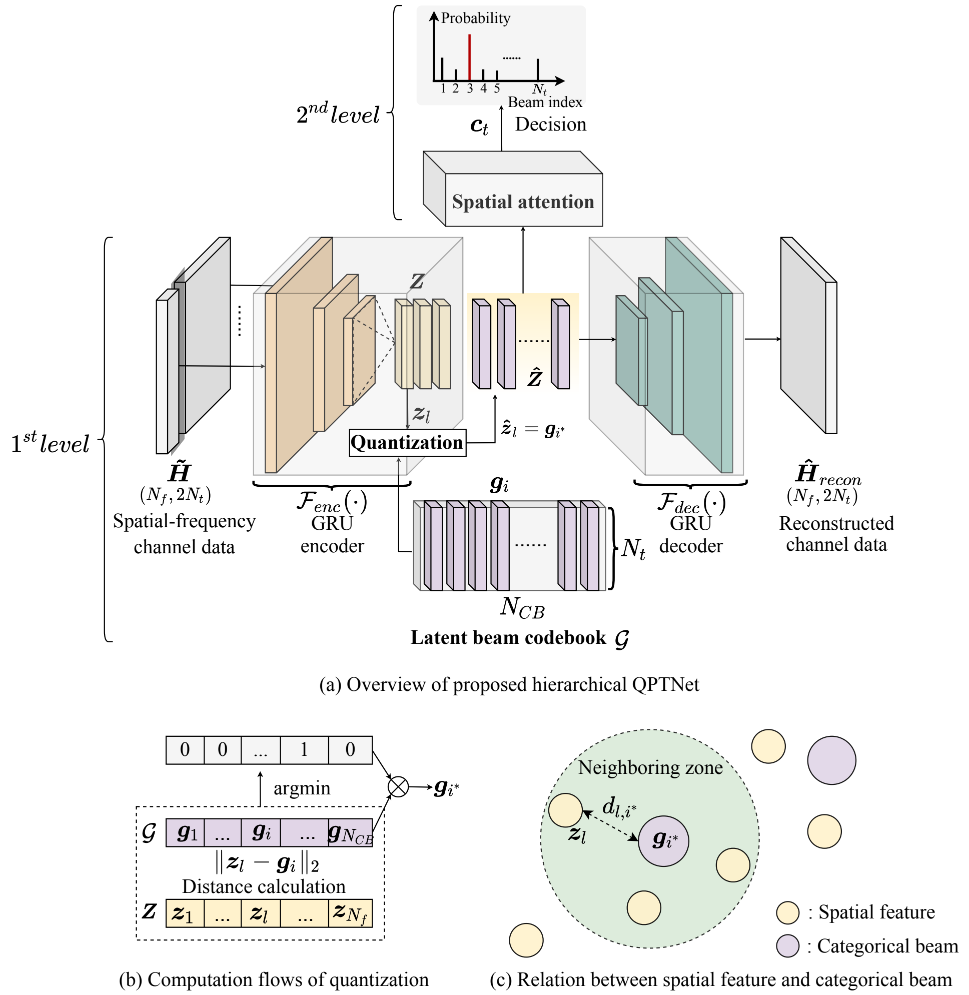
where $\boldsymbol{p}$ is the predicted probability, and p_n is the element of $\boldsymbol{p}$. To train our model, the cross entropy loss is applied as the evaluation metric for the classification problem. The cross entropy loss can be expressed as

$$\mathcal{L}_{cls} = -\sum_{n=1}^{N_t} y_c \log(p_n), \tag{46}$$

where y_c is the actual optimal mmWave beam described by one-hot encoding as the classification label. If label c is identical to the optimal beam, $y_c = 1$, otherwise, $y_c = 0$.

6.3 QPTNet associated beam training

This section presents the hierarchical QPTNet for efficiently processing the CSI. QPTNet has two hierarchical levels and a codebook-based phase quantization procedure. The first level is an autoregressive encoder based on GRU, and the second level is a spatial attention encoder based on the Transformer. The output of the GRU encoder can be quantized with a generated beam codebook. Finally, the spatial attention encoder extracts a global beam signature. The overview of QPTNet is illustrated in Fig. 34.



(a) Overview of proposed hierarchical QPTNet

(b) Computation flows of quantization

(c) Relation between spatial feature and categorical beam

Figure 34: Overview of proposed QPTNet. CSI is separately processed along the frequency subcarriers to capture the continuous spatial feature $\boldsymbol{z}_l$ via the GRU encoder. A latent beam codebook handles the perception of environmental variations by exploring the relation between the categorical beam $\boldsymbol{g}_{i^*}$ and the continuous spatial feature. Moreover, we reconstruct the channel data based on the selected latent beams via a GRU decoder to update the codebook and hold the consistency of channel and beamspace. The selected latent beams are fed into a spatial attention-associated beam prediction. $\boldsymbol{c}_t$ indicates the global beam signature.

6.3.1 Codebook-based phase quantization

To analyze the components of the propagation environment, we attempt to capture the implicated relation of cluster AoDs by the discrete phase-based quantization method. We define a latent beam codebook $\mathcal{G}$ including N_{CB} codewords $\boldsymbol{g}_i \in \mathbb{C}^{N_t \times 1}, i = 1, 2, \cdots, N_{CB}$ (i.e., N_{CB}-way categorical beams). The latent beamspace is initialized with randomly sampled angular $\bar{\phi}_i \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, and the normalized spatial frequency u_i is defined as

$$u_i = \frac{d \sin \bar{\phi}_i}{\lambda}, \tag{47}$$

where $u_i \in [-1/2, 1/2]$ for $\lambda/2$ element spacing. Intuitively, the beam vector can be generated by the DFT of $\boldsymbol{u}$ at points separated by $1/N_{CB}$. Note that $N_{CB} \geq N_t$ and there are N_{CB} categorical beam vectors $\boldsymbol{g}_i$. Thus, the latent beamspace can be described as

$$\boldsymbol{g}_i = \frac{1}{\sqrt{N_t}} [1, \mathrm{e}^{-j2\pi u_i}, \cdots, \mathrm{e}^{-j2\pi (N_{CB}-1) u_i}]^T. \tag{48}$$

Since we only consider the spatial power spectrum, the elements of (48) are adjustable with the network training.

As shown in Fig. 34(a) and (b), the GRU encoder yields a continuous spatial feature matrix $\boldsymbol{Z}$ with column vectors $\boldsymbol{z}_l \in \mathbb{C}^{N_t \times 1}, l = 1, 2, \cdots, N_f$. Next, we quantize the continuous feature $\boldsymbol{z}_l$ into categorical beam $\boldsymbol{g}_{i^*}$based on the Euclidean distance $d_{l,i}$. The categorical beam index is expressed as

$$i^* = \arg\min_i \underbrace{\|\boldsymbol{z}_l - \boldsymbol{g}_i\|_2}_{d_{l,i}}. \tag{49}$$

The minimal d_{l,i^*} ensures that the continuous spatial features corresponding to the categorical beams are within a neighboring zone of latent beamspace. Meanwhile, distance calculation can efficiently reflect the relation of cluster AoDs to represent the components of the environment. The quantization approach can also perform the clustering by mapping to the nearest codeword. The posterior categorical beam distribution $p(\hat{\boldsymbol{z}}_l = \boldsymbol{g}_i | \tilde{\boldsymbol{H}}, \bar{\boldsymbol{\phi}}_i)$ is formulated as

$$p(\hat{\boldsymbol{z}}_l = \boldsymbol{g}_i | \tilde{\boldsymbol{H}}, \bar{\boldsymbol{\phi}}_i) = \begin{cases} 1, & \text{if } \ i = i^*, \\ 0, & \text{else.} \end{cases} \tag{50}$$

During the forward pass, we define $\hat{\boldsymbol{Z}} = [\hat{\boldsymbol{z}}_1, \hat{\boldsymbol{z}}_2, \cdots, \hat{\boldsymbol{z}}_{N_f}]$ to present discrete angular beam responses. $\hat{\boldsymbol{z}}_l = \boldsymbol{g}_{i^*}$ are then selected corresponding to the approximate AoDs $\bar{\boldsymbol{\phi}}_i$. Different from the setup in Section 6.2, we acquire the optimal beam prediction based on the attention-scoring results of $\hat{\boldsymbol{Z}}$ to enhance the ability of environmental representation.

6.3.2 Hierarchical learning procedure

The proposed QPTNet contains two hierarchical levels for extracting the features from frequency and spatial domains. For the first level, we apply a GRU encoder to extract the relations between two domains and the spatial feature $\boldsymbol{Z}$. We employ the GRU module as the instant feature extractor due to its high training efficiency and similarity to the LSTM network in temporal sequence learning. GRU encoder proposes to synchronize the CSI estimation based on the gate mechanism to control the spatial feature in different subcarriers. In processing $\tilde{\boldsymbol{H}}$ at each subcarrier index, the GRU encoder takes in the channel delay response at the current subcarrier as well as the shared hidden state and output from the previous subcarrier step $(\boldsymbol{z}_{l-1}, \boldsymbol{\Theta})$ and updates its hidden state $\boldsymbol{\Theta}$ at the current subcarrier index, resulting into a new $\boldsymbol{z}_l$. The process can be described by

$$\boldsymbol{z}_l = \mathbb{P}(\tilde{\boldsymbol{H}}_{f_l} \mid \tilde{\boldsymbol{H}}_{f_{l-1}}, \cdots, \tilde{\boldsymbol{H}}_{f_2}, \tilde{\boldsymbol{H}}_{f_1}, \boldsymbol{\Theta}), \tag{51}$$

where $\mathbb{P}(\cdot)$ represents the conditional probability distribution over the subcarriers. Once the continuous spatial feature $\boldsymbol{z}_l$ is extracted, we quantize it by a latent beam codebook to perceive the environmental component. According to (49) and (50), we perform a nearest neighbor search to decide the posterior latent beam vector related to the discrete angular index $\bar{\boldsymbol{\phi}}_{i^*}$.

The input to the decoder corresponds to the selected latent beam matrix $\hat{\boldsymbol{Z}}$. We reconstruct the channel matrix based on $\hat{\boldsymbol{Z}}$ to guarantee the consistency between the latent beamspace and channel space. The reconstruction loss is formulated as

$$\mathcal{L}_{recon} = \mathbb{E}\{\frac{1}{N}\sum_{n=1}^{N} \|\mathcal{F}_{enc}(\tilde{\boldsymbol{H}}) - \mathcal{F}_{dec}(\hat{\boldsymbol{Z}})\|_2^2 / \|\mathcal{F}_{dec}(\hat{\boldsymbol{Z}})\|_2^2\}. \tag{52}$$

$\mathcal{F}_{enc}/\mathcal{F}_{dec}$ are GRU encoder/decoder mapping and the reconstructed channel $\hat{\boldsymbol{H}} = \mathcal{F}_{dec}(\hat{\boldsymbol{Z}})$, respectively. The forward computation pipeline can be regarded as the regular autoencoder with a specific non-linearity that corresponds to the 1-of-N_{CB} latent beam vectors. The optimization of QPTNet follows the back-propagating in [43]. To make sure the encoder commits to a latent beamspace and its output does not grow, we add a commitment loss. Thus, the total training loss becomes

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{recon} + \|sg[\mathcal{F}_{enc}(\tilde{\boldsymbol{H}})] - \boldsymbol{G}\|_2^2 + \beta\|\mathcal{F}_{enc}(\tilde{\boldsymbol{H}}) - sg[\boldsymbol{G}]\|_2^2, \quad (53)$$

where $sg[\cdot]$ means the stop gradient operator defined as an identity at forward computation time and has zero partial derivatives, $\boldsymbol{G} = [\boldsymbol{g}_1, \boldsymbol{g}_1, \cdots, \boldsymbol{g}_{N_{CB}}]$, and β is a hyperparameter. The first term is the beam prediction loss in (46) and the optimization does not change the latent beam codebook because of the non-linearity of quantization. The second term is the CSI reconstruction loss to update the latent beam codebook. The third term utilizes the l_2 norm loss to move the latent beam codebook vectors close to the encoder output $\mathcal{F}_{enc}(\tilde{\boldsymbol{H}})$, and the fourth term ensures that the encoder outputs toward the codeword.

6.4 SE-QPTNet associated beam training

This section proposes SE-QPTNet to provide an unsupervised pre-training strategy for improving the performance of QPTNet given the small training dataset. In SE-QPTNet, we construct positive and negative samples based on the anchor operation. Then the pre-training process can be completed by detecting the relationship among a positive sample of target UE, negative samples of others, and global beam signature with the contrastive environmental prediction. Therefore, SE-QPTNet can constantly perceive the environment component under a contrastive learning framework to further improve the performance of beam training. Fig. 35 describes the procedure of self-enhanced pre-training in SE-QPTNet.

To pre-train the proposed SE-QPTNet under the contrastive learning framework, we present an anchor operation to acquire the positive and negative samples. Specifically, we obtain forward latent beam features by splitting the outputs of quantization with the anchor t and explore the global beam signature $\boldsymbol{c}_t$ based on the attention-scoring evaluation. Meanwhile, the reconstructed channel data streams recovered by the GRU decoder divide as contrastive samples with length k, in which the positive sample is determined by the CSI of the target UE location while the others determine negative samples. SE-QPNet enhances the

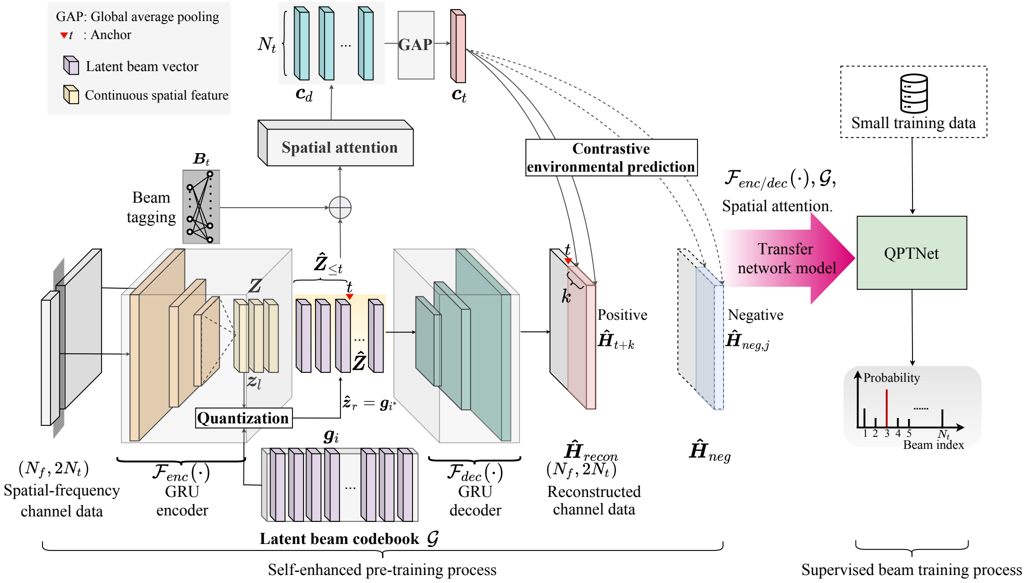

GAP: Global average pooling
: Anchor
Latent beam vector
Continuous spatial feature
N_t
$\boldsymbol{c}_d$
GAP
$\boldsymbol{c}_t$
Spatial attention
Contrastive environmental prediction
Small training data
$\boldsymbol{B}_t$
Beam tagging
$\mathcal{F}_{enc/dec}(\cdot), \mathcal{G}$,
Spatial attention.
Transfer network model
QPTNet
$\hat{\boldsymbol{Z}}_{\leq t}$
$\boldsymbol{Z}$
t
k
Positive
$\hat{\boldsymbol{H}}_{t+k}$
Negative
$\hat{\boldsymbol{H}}_{neg,j}$
$\boldsymbol{z}_l$
$\hat{\boldsymbol{Z}}$
Quantization
$\hat{\boldsymbol{z}}_r = \boldsymbol{g}_{i^*}$
$\boldsymbol{g}_i$
Probability
Beam index
$\hat{\boldsymbol{H}}_{recon}$
$\hat{\boldsymbol{H}}_{neg}$
$(N_f, 2N_t)$
Spatial-frequency channel data
$\mathcal{F}_{enc}(\cdot)$
GRU encoder
$\mathcal{F}_{dec}(\cdot)$
GRU decoder
$(N_f, 2N_t)$
Reconstructed channel data
Latent beam codebook $\mathcal{G}$
Self-enhanced pre-training process
Supervised beam training process

Figure 35: Overview of proposed SE-QPTNet. The proposed SE-QPTNet compensates for the effects of propagation delay by separately dealing with the sub-antenna arrays. Practically, the intermediate beam codes $\hat{\boldsymbol{z}}$ are divided into $\hat{\boldsymbol{z}}_{\leq t}$ as *small scale* parts and $\hat{\boldsymbol{z}}_{>t}$ as *large-scale* parts by an anchor t. We feed $\hat{\boldsymbol{z}}_{\leq t}$ to the spatial attention-associated beam prediction for identifying the beam signature vector $\boldsymbol{c}_t$. Contrastive environmental prediction measures the similarity between the $\boldsymbol{c}_t$ and the reconstructed $\hat{\boldsymbol{H}}_{recon}$, and makes the negative samples $\hat{\boldsymbol{H}}_{neg}$ dissimilar to its beam signature.

similarity between the target UE and the prediction of $\boldsymbol{c}_t$ while making the other UE responses dissimilar by the contrastive environmental prediction. Benefiting from this distinction, we can facilitate latent angular learning within the latent beams corresponding to the dominant path, discard the noises, and enhance the representation of the environmental component.

Considering the mmWave massive MIMO system, the amplitudes and phases of the received signals suffer from spatial differences due to a significant fraction of propagation delay [67]. To achieve robustness, the proposed SE-QPTNet compensates for the effects of propagation delay by separately dealing with the selected latent beams. We denote the global view $\hat{\boldsymbol{Z}}$ output from the quantization for notational convenience. We select an anchor t and define $\hat{\boldsymbol{Z}}_{\leq t}$ as the forward component. Taking inspiration from recent advancements [44], we introduce a token $\hat{\boldsymbol{z}}_0$ at the beginning of $\hat{\boldsymbol{Z}}_{\leq t}$ and feed them to the spatial attention associated beam prediction to identify the beam signature vector $\boldsymbol{c}_t$. We then predict the k reconstructed channels by a linear mapping $\bar{\boldsymbol{H}}_{t+k} = \boldsymbol{W}_{t+k}\boldsymbol{c}_t$, where $\boldsymbol{W}_{t+k}$ denotes the weight of the linear mapping $\mathcal{T}$. If the *large-scale* dense channel signatures can be successfully predicted, it indicates that the beam signature vector $\boldsymbol{c}_t$ contains a global and robust view of the propagation environment. By combining the positive and negative samples, the loss function of contrastive environmental prediction is

$$\mathcal{L}_N = -\sum_k \frac{\exp\left(\hat{\boldsymbol{H}}_{t+k}^T \bar{\boldsymbol{H}}_{t+k}\right)}{\exp\left(\hat{\boldsymbol{H}}_{t+k}^T \bar{\boldsymbol{H}}_{t+k}\right) + \sum_l \exp\left(\hat{\boldsymbol{H}}_{t+k}^T \bar{\boldsymbol{H}}_{neg,l}\right)}, \tag{54}$$

where $\bar{\boldsymbol{H}}_{neg,l}$ denotes negative samples taken from other channel data, and the positive sample is $\hat{\boldsymbol{H}}_{t+k}$. Considering the latent beam codebook updating, the total pre-training loss is expressed as

$$\mathcal{L}_{pre} = \mathcal{L}_N + \|sg[\mathcal{F}_{enc}(\tilde{\boldsymbol{H}})] - \boldsymbol{G}\|_2^2 + \beta\|\mathcal{F}_{enc}(\tilde{\boldsymbol{H}})] - sg[\boldsymbol{G}]\|_2^2. \tag{55}$$

The pre-training process is illustrated in Algorithm 2.

Algorithm 2 Pre-training process for SE-QPTNet

```
Input: Pre-training dataset $\mathcal{D}$, batch size $N$, anchor $t$, prediction step $k$, latent beam codebook $\mathcal{G}$, structure of $\mathcal{F}_{enc}$, $\mathcal{F}_{dec}$, spatial attention $\mathcal{S}$, and linear operation $\mathcal{T}$.
for sampled minibatch $\left\{\tilde{\boldsymbol{H}}_{\boldsymbol{b}}\right\}_{b=1}^{N}$ from $\mathcal{D}$ do
    for all $b \in \{1, \cdots, N\}$ do
        Obtain spatial feature $\{\boldsymbol{z}_l\}_{l=1}^{N_f} = \mathcal{F}_{enc}(\tilde{\boldsymbol{H}}_{\boldsymbol{b}})$.
        for all $l \in \{1, \cdots, N_f\}$ do
            Obtain the latent beam index $i^*$ based on (49)
            $\hat{\boldsymbol{z}}_l \leftarrow \boldsymbol{g}_{i^*}$
        end for
        Obtain $\hat{\boldsymbol{Z}} \leftarrow \{\hat{\boldsymbol{z}}_l\}_{l=1}^{N_f}$
        Select the forward component $\hat{\boldsymbol{Z}}_{\leq t}$
        Global beam signature $\boldsymbol{c}_t = \mathcal{S}(\hat{\boldsymbol{Z}}_{\leq t})$
        Reconstruct $\hat{\boldsymbol{H}} = \mathcal{F}_{dec}(\hat{\boldsymbol{Z}})$
        $\hat{\boldsymbol{H}}_{t+k} = [\hat{\boldsymbol{h}}_t, \cdots, \hat{\boldsymbol{h}}_{t+k}]$
        $\bar{\boldsymbol{H}}_k = \mathcal{T}(\boldsymbol{c}_t)$
    end for
    Calculate the loss $\mathcal{L}_{pre}$ based on (55)
    Update $\mathcal{F}_{enc}$, $\mathcal{F}_{dec}$, $\mathcal{S}$, $\mathcal{T}$, and $\mathcal{G}$ to minimize $\mathcal{L}_{pre}$
end for
return the network model
```

Once the pre-training process is completed, we transfer the parameters of the GRU encoder/decoder and spatial attention to the QPTNet for supervised beam training. According to the (46) and (55), the total beam training loss of the proposed SE-QPTNet can be described as

$$\mathcal{L}_{tot} = \mathcal{L}_{cls} + \mathcal{L}_{pre}. \tag{56}$$

Benefiting from the pre-trained model, the network can achieve high performance with a small training dataset. It allows mitigation of preprocessing tasks for labeling the actual optimal beam knowledge, reduces training overhead, and reaches stronger adaptability to environmental variation.

6.5 Experimental results

6.5.1 Simulation set up

We evaluate the training and predicting performances of the proposed QPTNet and SE-QPTNet on three tasks: training performance, success rate, and achievable rate. The main parameters of the proposed framework are represented in Table 6.

Table 6: Training hyper parameters

Parameters	Values
Dim. of embedded $\boldsymbol{B}_g$	128
Dim. of $\boldsymbol{B}_t$	128
Multi-heads	6
Num. of attention modules	2
Dim. of MLP	256
Length of beam codeword	64
Option of latent beam codebook N_{CB}	$64, 128, 256$
Batch size N	128
Num. of positive sample	1
Num. of negative samples	127
Anchor t	16
Upper bound of k	8
Optimizer	Adam
Learning rate	10^{-3}
Num. of epochs	100
Dropout percentage	20%
Dataset size (100%)	10×10^4
Dataset split	$70\%; 30\%$

6.5.2 Training performance

We first compare the training performance of QPTNet and SE-QPTNet with different sizes of latent beam codebook options, and then we evaluate the proposed

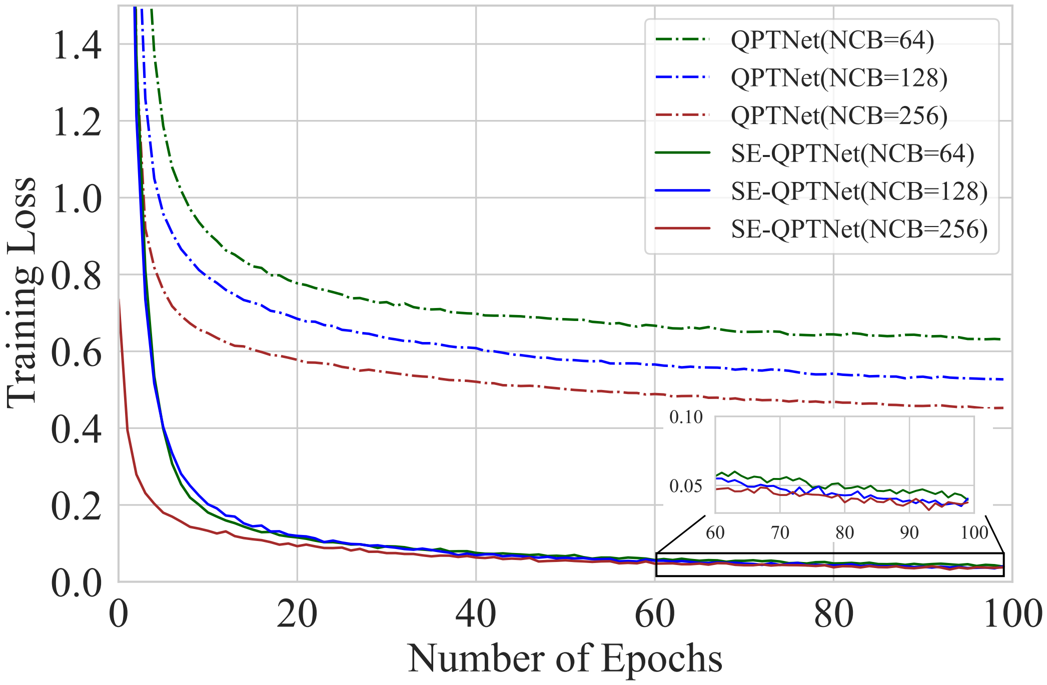


Figure 36: Training performance of QPTNet and SE-QPTNet.

schemes against existing related works in [20, 32].

Figure 36 presents the training performance of different latent beam codebook options between QPTNet and SE-QPTNet. The results indicate that the latent beam codebook with 256 spatial resolution results in the smallest training loss values for both QPTNet and SE-QPTNet. Additionally, SE-QPTNet outperforms QPTNet with arbitrary latent beam codebook options, achieving a much lower training loss curve. These results highlight the importance of higher spatial-resolution options in precisely quantifying the continuous channel features into categorical beams to achieve better beam training performance and faster convergence. It also shows that the contrastive learning framework based SE-QPTNet can improve the learning efficiency by consistently interacting with the propagation environment and accounting for the mmWave channel fluctuation in the re-identification of QPTNet. Therefore, the higher spatial resolution option and the contrastive environmental prediction can enhance training performance and learning efficiency. Consequently, we set the latent beam codebook option

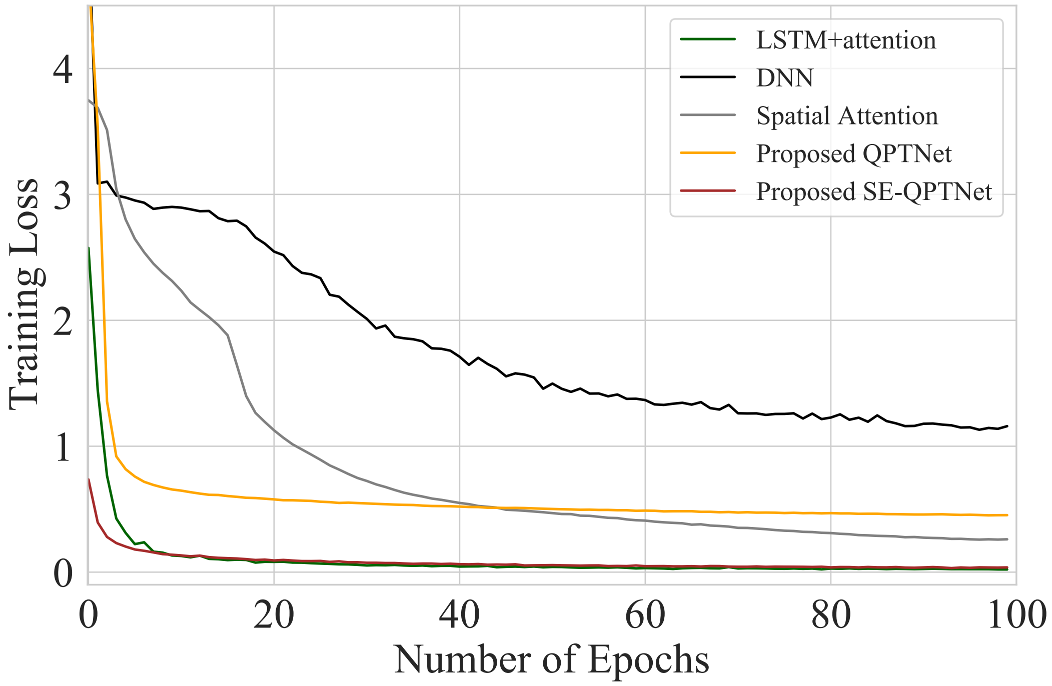


Figure 37: Training performance of QPTNet, SE-QPTNet, spatial attention associated beam prediction, and DNN [2] and LSTM attention [3].

$N_{CB} = 256$ for the following results.

Figure 37 illustrates the training comparisons among the proposed SE-QPTNet, QPTNet, spatial attention associated beam prediction, and DNN [2] and LSTM attention [3]. The training loss performance indicates that the SE-QPTNet achieves the fastest convergence with the loss value of 0.12, outperforming LSTM attention (0.15), DNN (1.8), spatial attention (0.31), and QPTNet (0.4). It is observed that the learning efficiency of SE-QPTNet is the best among the comparisons, benefiting from its competitive initial loss value (0.78) and smooth convergence. Looking closely at QPTNet and spatial attention, the latent beams can quantify the components of the propagation environment and significantly improve the accuracy of beam prediction, leading to faster convergence. However, the latent beam codebook updated by the mmWave channel reconstruction restricts the learning efficiency of QPTNet, because of the effects of noise and channel variations.

6.5.3 Success rate performance

To evaluate the robustness and low training overhead, we evaluate the performance of beam prediction in terms of the success rate under the different SNRs ranging from 0 to 20 dB and with various sizes of training datasets.

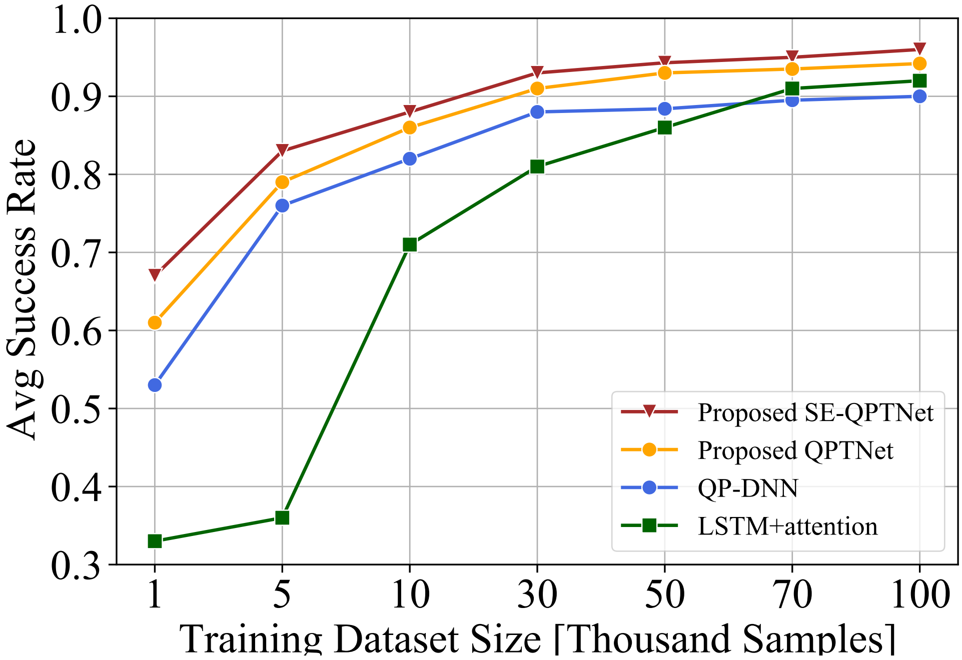


Figure 38: Average success rate performance of SE-QPTNet, QPTNet, QP-DNN, and LSTM attention [3] with different prior label information.

Figure 38 shows the average success rate of proposed SE-QPTNet, QPTNet, and related works with different training data sizes. It is visible that the proposed SE-QPTNet can achieve a success rate of around 0.69 only with 1% training data and rise to 0.91 given 30% training data. To illustrate the advantage of the proposed phase quantization scheme, we apply the DNN instead of spatial-attention associated beam prediction, named QP-DNN. It demonstrates that the unsupervised pre-training strategy can perceive informative environmental components to distinguish the implicit representations of AoDs so that SE-QPTNet can promote learning ability with a small training dataset. Moreover, the proposed contrastive learning framework can enhance the global beam signature with multipath interference to improve reliability. Compared with the LSTM attention, the proposed SE-QPTNet obtains almost the same performance with only 5% data, which improves learning efficiency by six times. The result shows the potential to utilize less training time and mitigation of labeling the actual optimal beam knowledge to perform a flexible beam training process.

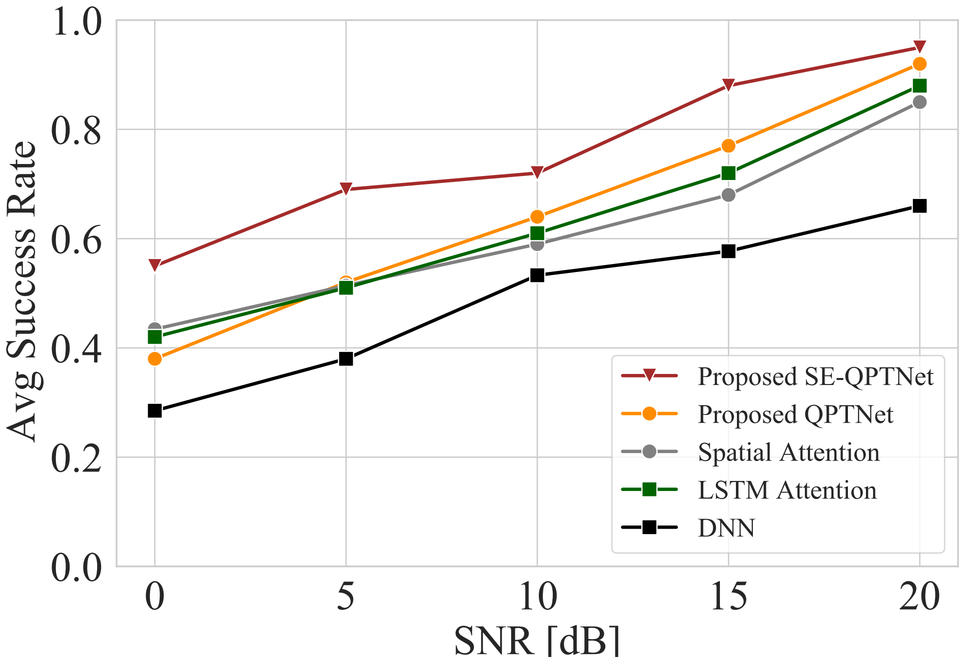


Figure 39: Average success rate of QPTNet, SE-QPTNet, spatial attention associated beam prediction, and DNN [2] and LSTM attention [3] with different SNRs from 0 to 20 dB.

To further investigate the robustness and lower training overhead, we evaluate the average success rate using 50% of the training dataset with the proposed SE-QPTNet. The results are presented in Fig. 39. It shows the success rate performance of the proposed SE-QPTNet (50% training data), QPTNet, spatial attention associated beam prediction, DNN, and LSTM attention. Our proposed SE-QPTNet achieves a success rate of 0.58, demonstrating its robustness at 0 dB SNR compared to LSTM attention, DNN, and spatial attention, as well as

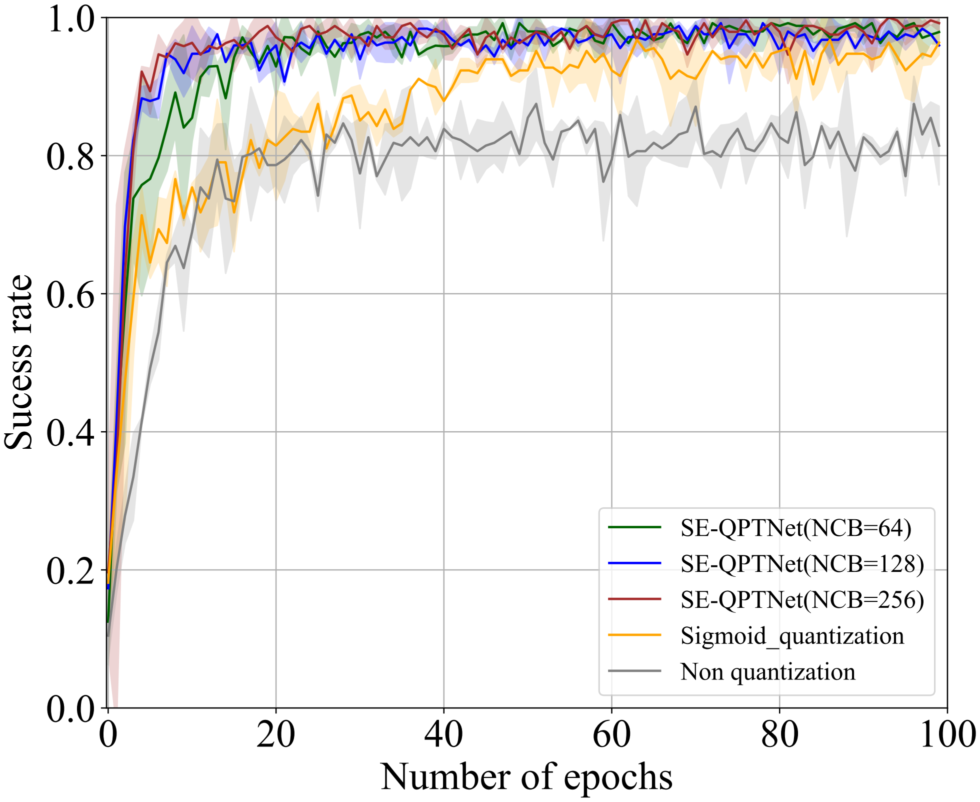


Figure 40: Success rate of SE-QPTNet with different codebook sizes, sigmoid-based quantization [4], and SE-QPTNet without quantization.

the proposed QPTNet, with achievable rates of 0.42, 0.43, 0.26, and 0.39, respectively. The proposed SE-QPTNet exhibited robustness against the effect of noise with low SNR levels in the comparison due to its contrastive learning framework. Closely looking at QPTNet, LSTM attention, and spatial attention; the performance of QPTNet is sensitive at the low SNR level because the latent beam codebook updates with the reconstruction of the mmWave channel. It makes QPTNet incorrectly quantize the spatial feature resulting in fuzzy environmental components.

Figure 40 illustrates the effectiveness of the proposed codebook-based phase quantization. We compare the success rate performance of different latent beam codebook options and sigmoid-based quantization [4]. The baseline is the proposed SE-QPTNet without quantization. It is visible that the proposed SE-QPTNet can obtain a higher success rate performance and faster convergence than the sigmoid-based method and the baseline method. The sigmoid-based quantization in [4] can map each frequency information into the range of 0 and 1 to establish a phase relationship. The continuous and monotonically increasing properties make the results of quantization hardly determine the discrete phase option. Unlike the sigmoid-based quantization, the proposed codebook-based method leads to a rich diversity of beam features corresponding to the discrete phases. Moreover, the proposed quantization method can boost robust performance (smaller error band) by increasing the size of the codebook.

Table 7: Complexity of the proposed SE-QPTNet, QPTNet, and related works.

Scheme	Total parameters	Success rate	Achievable rate
SE-QPTNet	34.4K	95.1%	7.37
QPTNet	34.4K	92.5%	7.24
CNN LSTM [25]	13K	80.4%	6.51
LSTM attention [3]	41.9K	87.6%	7.11
Spatial Attention [40]	48K	83.5%	6.80
DNN [2]	186K	50.7%	5.38

In Table 7, we compare the complexity of SE-QPTNet, QPTNet, and related works. Both the success rate and the achievable rate of SE-QPTNet can obtain the best performance against the related works. Although the achievable rate of QPTNet and SE-QPTNet are almost identical, the success rate of SE-QPTNet is higher. We can observe an improvement of 2.6%, owing to the contrastive environmental prediction that enables it to update the latent beam codebook efficiently. The number of parameters of SE-QPTNet is lower than those of LSTM attention [3], spatial attention [40], as well as the DNN [2], which can save 18%, 28%, 82% parameter requirements. The results demonstrate that the proposed contrastive learning framework can improve DL performance with low overhead. Additionally, contrastive environmental prediction is beneficial for improving learning efficiency by distinguishing the target UE from others, maintaining a low parameter overhead, and better global beam signature representation. Consequently, the proposed contrastive learning framework contributes to promoting learning efficiency and getting rid of high complexity.

Table 8: Performances of different attention module options.

Num of attention	Total parameters	Success rate	Achievable rate
1	31.0K	76.9%	6.43
2	34.4K	95.1%	7.37
4	41.0K	95.5%	7.36
6	47.7K	95.5%	7.36
8	54.3K	95.9%	7.38

Table 8 investigates the effects of spatial attention modules for beam prediction. We compare the complexity, success rate, and achievable rate of different spatial attention options. According to the results, the success rate can be improved by 20% when the number of attention modules is larger than 1. Increasing the number of attention modules does not significantly improve the accuracy of beam prediction, but increases the overhead. Therefore, we adopt 2 spatial attention modules as the optimal option for SE-QPTNet.

6.5.4 Achievable rate performance

The proposed SE-QPTNet can obtain extra performance by pre-training without label information, and achieve a high transmission rate with a reduced amount of labeled CSI. To evaluate the training overhead and robustness in practical systems, we plot the achievable rate for different training dataset sizes and their performance under different SNR levels.

In Table 9, we compare the achievable rate of QPTNet and SE-QPTNet against different latent beam codebook options. The achievable rates of SE-QPTNet and QPTNet are close but still higher than that of QPTNet without noise effects. Although the performance of QPTNet and SE-QPTNet are almost identical, SE-QPTNet is more robust. We can observe an improvement of 0.41 in 10 dB SNR, owing to the contrastive environmental prediction that enables it to update the latent beam codebook efficiently. The results demonstrate that a larger spatial resolution can comprehensively quantify environmental variations to improve beam prediction accuracy. Additionally, contrastive environmental prediction is beneficial for improving robustness by the inherent representation of the global beam signature among different UEs.

Table 9: Achievable rate of the proposed QPTNet and SE-QPTNet with different SNRs and N_{CB} options.

Scheme	SNR [dB]	N_{CB}		
		64	128	256
QPTNet	No noise	7.18	7.20	7.24
	10	3.25	5.22	6.20
SE-QPTNet	No noise	7.32	7.34	7.37
	10	5.13	5.23	6.61

To illustrate the low training overhead of proposed solutions, Fig. 41 shows the achievable rate of the proposed SE-QPTNet, QPTNet, and DNN approaches, defined in (8) for different dataset sizes. Note that the dot-dash line in Fig. 41 represents the upper bound on the performance of the given system and channel models. The horizontal axis represents the training data size, and the achievable rate results achieve with $1\%, 5\%, 10\%, 30\%, 50\%, 70\%, 100\%$ training data. The results of SE-QPTNet and QPTNet are close to the upper bound with only 10% training data, indicating that the proposed solutions allow the BS to apply the desired beam with an efficient training scheme. It is visible that SE-QPTNet can successfully predict the optimal beam based on the contrastive learning framework with only 1% training data at the BS. It clearly illustrates the ability of the self-enhancement to efficiently represent the desired angular sector with positive and negative samples, achieving negligible training overhead. Moreover, the proposed solutions are flexible in customizing the beam solution with small

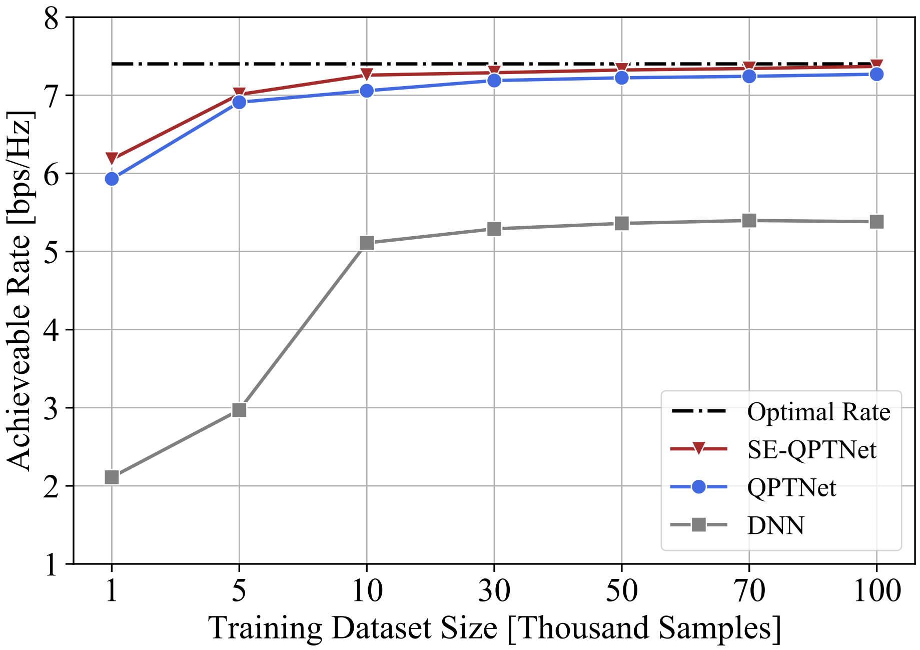


Figure 41: Achievable rate performance of QPTNet, SE-QPTNet, and DNN [2] with different training samples.

training data. Therefore, approaching this bound illustrates the optimality of the proposed solutions.

To verify the robustness of the proposed SE-QPTNet and QPTNet, we investigate the achievable rate of different SNR levels for generating the imperfect CSI, as shown in Fig. 42. The dot-dash line represents the upper bound performance, the optimal rate without noise for the given system and channel models. The results show that SE-QPTNet and QPTNet are consistently better than the simple DNN beam training approach. Benefiting from the contrastive environmental prediction to update the latent beam codebook, SE-QPTNet can achieve a higher rate than the QPTNet at low SNR levels. Both proposed beam training schemes are near the optimal rate when SNR is higher than 15 dB. It illustrates that the hierarchical strategy and the quantization can deal with the frequency and spatial channel efficiently, and the ability of self-enhancement can provide a better perception of the complicated propagation environment by the contrastive

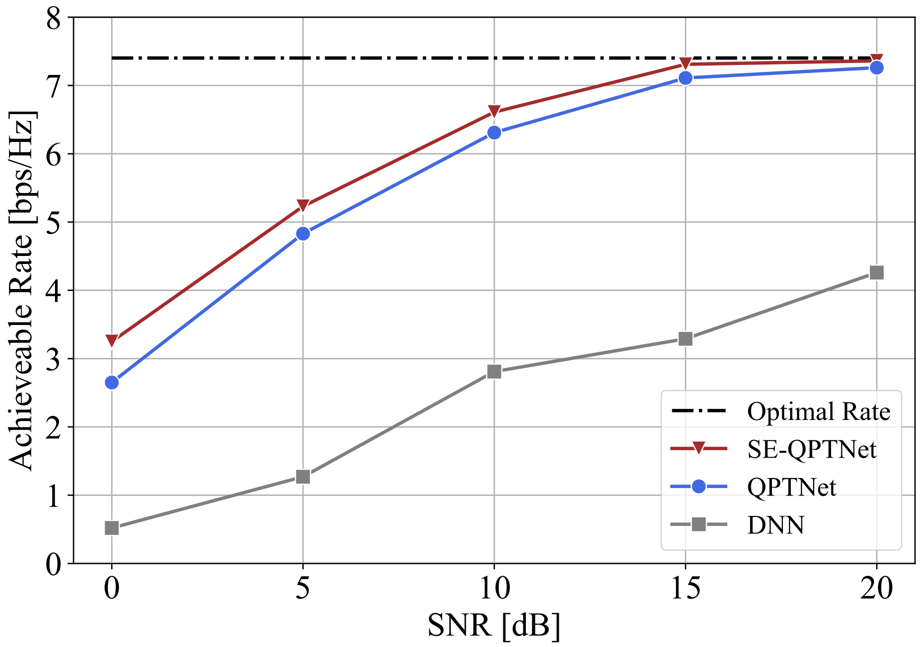


Figure 42: Achievable rate performance of QPTNet, SE-QPTNet, and DNN [2] (imperfect CSI) with different SNRs from 0 to 20 dB.

learning framework. Moreover, the contrastive environmental prediction can improve the robustness relying on the distinct global beam signature by considering the similarity among global beam signature and positive/negative samples.

To explore the performance of the proposed SE-QPTNet and QPTNet with different options of labeled CSI, we provide the achievable rate of different SNR levels, as shown in Fig. 43. The dashed line represents the upper bound performance, which is generated by the actual beam vectors (ground truth). The results reveal that SE-QPTNet and QPTNet can preserve a similar achievable rate with a small labeled CSI option and are consistently better than the simple DNN beam training approach. Benefiting from the proposed contrastive learning framework to pre-train, SE-QPTNet permits the practical networks to train with friendly requirements of labeling. Moreover, both proposed beam training schemes are better than the baseline DNN approach. It illustrates that the hierarchical paradigm can provide a better global beam signature with the views of

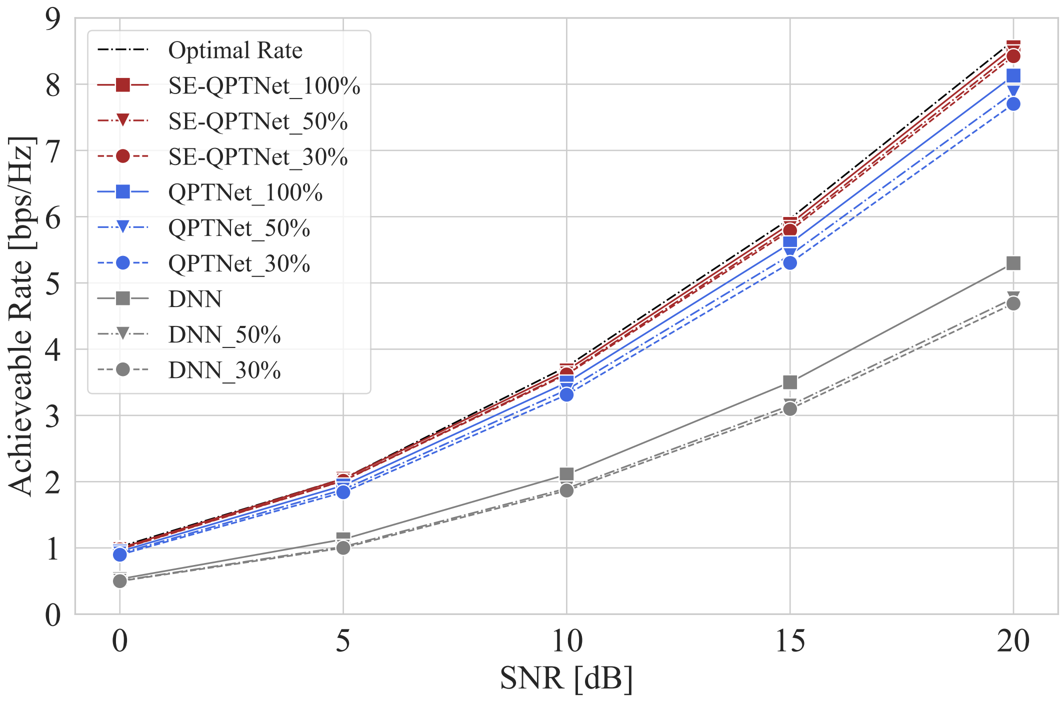


Figure 43: Achievable rate performance of QPTNet, SE-QPTNet, and DNN [2] (different labeled CSI for fine-tuning) with different SNRs from 0 to 20 dB.

frequency and spatial efficiently, and the quantization can enhance the robustness by the discrete spatial representation.

6.6 Concluding remark

This chapter proposed a novel framework SE-QPTNet for beam training based on spatial attention and quantization, utilizing the contrastive environmental prediction to detect the relationship between the beam signature of the target UE and others. The proposed hierarchical scheme QPTNet quantized the continuous spatial features into controllable latent beam responses within a wide range of discrete phases achieving higher robustness. The proposed contrastive learning framework SE-QPTNet enhanced the capability of identifying the beam signature by contrastive environmental prediction with a small fraction of the labeled CSI dataset. The proposed SE-QPTNet permitted DL-integrated beam training to adapt flexibly to wireless environments with low labeling costs. The simulation results showed that the proposed SE-QPTNet framework can achieve a data rate of 7.01 bps/Hz with only 5% labeled data, which represents 95.1% of the performance utilizing the full-size dataset. The proposed SE-QPTNet framework outperformed existing DL-based schemes, obtaining higher capacity with lower expertise requirements and highly reliable performance for mmWave massive MIMO systems.

7 Conclusion

Future wireless communication networks will predominantly be formed with mobile devices such as smartphones, tablets, wearables, the Internet of Things (IoT), etc. This has resulted in exponential growth in the amount of wireless data being created. However, spectrum resources in the current microwave regime are almost expended and it is evident that the wave of data requirement will not be met with the current state-of-the-art technologies. Hence, the DL-based intelligent design of future wireless networks become extremely important. This doctoral dissertation has presented a comprehensive investigation into the effects of DL-based beam training on mmWave massive MIMO systems. Through inherent characteristics of complicated propagation environment analysis and empirical studies related to DL-based proposals, several key findings have emerged.

In Chapter 4, we provided a non-deterministic quantization-based DL proposal and a wise voting scheme that enables highly reliable and hardware-friendly SoC applications for the massive MIMO mmWave systems. The key idea of the developed strategy was to quantize the noise and fading affected mmWave channel as binary sequences through a non-deterministic quantization instead of modeling from the noisy intensities of the wireless channel directly. Moreover, in Chapter 5 we proposed a hierarchical DL paradigm to capture the environmental information from the frequency and spatial views to conduct the optimal beam prediction. Benefiting from the dual views of CSI, the proposed methods can achieve a highly reliable and robust beam prediction compared with the existing unitary modeling schemes for the mmWave massive MIMO system. Finally, in Chapter 6 we provided a novel self-supervised learning scheme to get rid of the expensive labeling cost and requirement of expertise for the DL-integrated beam training. The proposed contrastive learning framework enhanced the capability of identifying the beam signature by contrastive environmental prediction and performed the beam prediction with a small fraction of the labeled CSI dataset. The proposed contrastive learning framework permitted DL-integrated beam training to adapt flexibly to wireless environments with low labeling costs.

While this dissertation provides valuable insights into the current state of DL integrated beam training in the face of 5G/beyond 5G systems, further research is encouraged to address certain knowledge gaps. When the wise-voting scheme is considered for achieving a robust beam application, the constraint on the multiple results of inference is time-consuming and more feasible for the static UE. Furthermore, the massive number of antennas provides holistic spatial sensing but it will increase the burden for the current DL-based beam training schemes. The current proposals contribute to beam training for general massive MIMO but they lack dealing with the extremely complicated massive MIMO systems (more than 1024). It is mandatory to process the redundancy of frequency and spatial CSI in advance to relieve the overhead of DL design. Moreover, the current proposals confine the static or low mobile UE which fall into the same cellular. The proposed schemes may moderate the beam prediction of the high mobile UE or multi-cellular systems under the distortion of Doppler effects.

Extending the possibility of the current SSL framework to the 6G wireless systems, we devote to exploring future works for the integrated sensing and communication (ISAC) networks. In the ISAC, the network can be regarded as a huge sensor to sense physical information, provide propagation details, and detect user positions [68]. The DL plays a key role in joint sensing and communication, sensing-aided communication, and communication-aided sensing systems by fusing the physical information that came from the different sensors [69, 70]. Future studies concentrate on establishing an identical relation of sensing the target UE from different sensors and enhancing its capacity to represent and distinguish with the other physical information.

8 Publication List

Journal

[1] **H. Jia**, N. Chen, H. Gao and M. Okada, "Spatial Attention and Quantization based Contrastive Learning Framework for mmWave Massive MIMO Beam Training," (accepted) EURASIP Journal on Wireless Communications and Networking.

[2] N. Chen, C. Liu, **H. Jia** and M. Okada, "Intelligent Reflecting Surface Aided Network under Interference toward 6G Applications," in IEEE Network, vol. 36, no. 4, pp. 18-27, July/August 2022, doi: 10.1109/MNET.001.2100675.

International conference

[1] **H. Jia**, C. Liu, Z. Wang, N. Chen and M. Okada, "Non-deterministic Sparse Feature Learning for Reliable Beam Prediction in mmWave Massive MIMO Systems," 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Kyoto, Japan, 2022, pp. 837-842, doi: 10.1109/PIMRC54779.2022.9977796

[2] **H. Jia**, N. Chen, H. Gao and M. Okada, "Feature Quantization Convolutional Neural Networks for CSI Feedback in Massive MIMO Systems," 2022 IEEE 24th Int Conf on High Performance Computing & Communications, Hainan, China, 2022, pp. 725-732, doi: 10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00 122.

[3] **H. Jia**, N. Chen, R. Zhang and M. Okada, "Non-deterministic Quantization for mmWave Beam Prediction," 2022 IEEE 35th International System-on-Chip Conference (SOCC), Belfast, United Kingdom, 2022, pp. 1-6, doi: 10.1109/SOCC 56010.2022.9908091.

[4] **H. Jia**, N. Chen and M. Okada, "Memory Shared Spatial Attention Neural Network for Reliable Beam Prediction," 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 2022, pp. 107-108, doi: 10.1109/GCCE56475.2022.10014249.

[5] T. Urakami, **H. Jia**, N. Chen and M. Okada, "Individual Memory Driven Transformer Deep Learning Model for Multi-Cell Massive MIMO Beam Prediction," 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 2022, pp. 1731-1736, doi: 10.23919/APSIPAASC55919.2022.9980016.

[6] Z. Chen, L. Zhu, **H. Jia**, and T. Matsubara, "A two-view EEG representation for brain cognition by composite temporal-spatial contrastive learning," 2023 SIAM International Conference on Data Mining (SDM). SIAM, 2023, pp. 334–342.

Domestic conference

[1] **H. Jia**, N. Chen and M. Okada, "Convolutional Radio Modulation Recognition Networks with Attention Models in Wireless Systems," ITE Technical Report, vol.45, no.24, pp.1-4, Vol.45, Sept. 2021.

[2] **H. Jia**, Z. Wang, N. Chen and M. Okada, "Temporal Sequence Modeling and Spatial Attention for High Reliable mmWave Massive MIMO Beam Prediction," IEICE Technical Report, vol. 122, pp. 57-62, Jul. 2022.

Acknowledgements

I would like to express my deepest gratitude to the exceptional individuals who supported and guided me throughout this long journey. Without their kind support, it is impossible to come to the long end.

Firstly, I wish to express my profound gratitude to my supervisor, Prof. Minoru Okada for affording me a hard-won opportunity and the continuous support throughout my PhD research. His dedication to the pursuit of knowledge and commitment to my academic growth have been instrumental in shaping the direction of my research.

I would also like to extend my heartfelt thanks to Professor Shoji Kasahara, for his insightful feedback, constructive criticism, and valuable suggestions. His expertise in the field and willingness to share the knowledge have enriched this research immensely.

I am also thankful to Associate Professor Takeshi Higashino, and Associate Professor Duong Quang Thang for their kind guidance and support. This dissertation would not be complete without their advice, comments as well as countless discussions.

I am deeply grateful to Assistant Professor Na Chen for her unwavering assistance throughout every stage of the research project, starting from the first year of my doctoral program. Her wisdom helps me to shape the direction of my research, enhance my writing skills, and clear the obstacles of my road.

I would like to thank all laboratory members, especially research group members, who supported me in my experiments as well as in daily life. That made my life in NAIST more memorable. I also would like to extend sincere thanks to my group-mates (Taisei Urakami and Zaoshi Wang). I would not have been able to have a great research life without their supports.

My thanks goes to my dear friend Dr. Zheng Chen, as a token of respect for his influence on my academic pursuits. His fresh idea (also his inquires) always inspires me to think from another aspect.

Finally, I am deeply indebted to my family for their unwavering support and understanding during this journey. Their love, encouragement, and belief in me sustained my motivation and determination to step this journey through to completion.

References

[1] N. Chen and M. Okada, "Toward 6G internet of things and the convergence with RoF system," *IEEE Internet of Things J.*, vol. 8, no. 11, pp. 8719–8733, 2021.

[2] C. Qi, Y. Wang, and G. Y. Li, "Deep learning for beam training in millimeter wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2020.

[3] L. Dai, X. Gao, S. Han, I. Chih-Lin, and X. Wang, "Beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array," in *Proc. IEEE ICCC*, pp. 1–6, 2016.

[4] H. Hojatian, J. Nadal, J.-F. Frigon, and F. Leduc-Primeau, "Flexible unsupervised learning for massive MIMO subarray hybrid beamforming," in *Proc.IEEE GLOBECOM*, pp. 3833–3838, IEEE, 2022.

[5] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks with a focus on propagation models," *IEEE Trans. Antennas Propag*, vol. 65, no. 12, pp. 6213–6230, 2017.

[6] M. Alsabah, M. A. Naser, B. M. Mahmmod, S. H. Abdulhussain, M. R. Eissa, A. Al-Baidhani, N. K. Noordin, S. M. Sait, K. A. Al-Utaibi, and F. Hashim, "6G wireless communications networks: A comprehensive survey," *IEEE Access*, vol. 9, pp. 148191–148243, 2021.

[7] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!," *IEEE Access*, vol. 1, pp. 335–349, 2013.

[8] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, 2013.

[9] J. Mo, B. L. Ng, S. Chang, P. Huang, M. N. Kulkarni, A. Alammouri, J. C. Zhang, J. Lee, and W.-J. Choi, “Beam codebook design for 5G mmWave terminals,” *IEEE Access*, vol. 7, pp. 98387–98404, 2019.

[10] F. A. Pereira de Figueiredo, “An overview of massive MIMO for 5G and 6G,” *IEEE Lat. Am. Trans.*, vol. 20, no. 6, pp. 931–940, 2022.

[11] J. Li, Y. Niu, H. Wu, B. Ai, S. Chen, Z. Feng, Z. Zhong, and N. Wang, “Mobility support for millimeter wave communications: Opportunities and challenges,” *IEEE Commun. Surv. Tutor.*, 2022.

[12] Y. Li, J. G. Andrews, and F. h. Baccelli, “Design and analysis of initial access in millimeter wave cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6409–6425, 2017.

[13] J. Kim and A. F. Molisch, “Fast millimeter-wave beam training with receive beamforming,” *J.Commn.Net*, vol. 16, no. 5, pp. 512–522, 2014.

[14] M. E. Eltayeb, A. Alkhateeb, R. W. Heath, and T. Y. Al-Naffouri, “Opportunistic beam training with hybrid analog/digital codebooks for mmWave systems,” in *Proc.IEEE GlobalSIP*, pp. 315–319, 2015.

[15] J. Wang, Z. Lan, C.Pyo, T. Baykas, C.Sum, M. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada, and S. Kato, “Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems,” *IEEE J. Sel. Areas in Commun.*, vol. 27, no. 8, pp. 1390–1399, 2009.

[16] C. Qi, K. Chen, O. A. Dobre, and G. Y. Li, “Hierarchical codebook-based multiuser beam training for millimeter wave massive mimo,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8142–8152, 2020.

[17] Z. Xiao, T. He, P. Xia, and X.-G. Xia, “Hierarchical codebook design for beamforming training in millimeter-wave communication,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, 2016.

[18] S.-E. Chiu, N. Ronquillo, and T. Javidi, “Active learning and csi acquisition for mmwave initial alignment,” *IEEE J. Sel. Areas in Commun.*, vol. 37, no. 11, pp. 2474–2489, 2019.

[19] I. Aykin and M. Krunz, "Efficient beam sweeping algorithms and initial access protocols for millimeter-wave networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2504–2514, 2020.

[20] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37328–37348, 2018.

[21] H. Echigo, Y. Cao, M. Bouazizi, and T. Ohtsuki, "A deep learning-based low overhead beam selection in mmwave communications," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 682–691, 2021.

[22] S. Rezaie, C. N. Manchón, and E. de Carvalho, "Location- and orientation-aided millimeter wave beam selection using deep learning," in *Proc.IEEE ICC*, pp. 1–6, 2020.

[23] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, 2020.

[24] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, "Fast beamforming design via deep learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1065–1069, 2019.

[25] K. Ma, D. He, H. Sun, Z. Wang, and S. Chen, "Deep learning assisted calibrated beam training for millimeter-wave communication systems," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6706–6721, 2021.

[26] H.Jia, Liu, Z.Wang, N.Chen, and M.Okada, "Non-deterministic sparse feature learning for reliable beam prediction in mmWave massive MIMO systems," in *Proc. IEEE PIMRC*, pp. 837–842, 2022.

[27] Z. Xiao, H. Dong, L. Bai, P. Xia, and X.-G. Xia, "Enhanced channel estimation and codebook design for millimeter-wave communication," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9393–9405, 2018.

[28] J. Palacios, D. De Donno, and J. Widmer, “Tracking mm-wave channel dynamics: Fast beam training strategies under mobility,” in *Proc. IEEE INFOCOM*, pp. 1–9, 2017.

[29] W. Ma, C. Qi, Z. Zhang, and J. Cheng, “Sparse channel estimation and hybrid precoding using deep learning for millimeter wave massive MIMO,” *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2838–2849, 2020.

[30] M. Hussain and N. Michelusi, “Learning and adaptation for millimeter-wave beam tracking and training: A dual timescale variational framework,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 37–53, 2022.

[31] S. Chen, Z. Jiang, S. Zhou, and Z. Niu, “Time-sequence channel inference for beam alignment in vehicular networks,” in *Proc. IEEE GlobalSIP*, pp. 1199–1203, 2018.

[32] H. Jia, N. Chen, and M. Okada, “Memory shared spatial attention neural network for reliable beam prediction,” in *Proc. IEEE GCCE*, pp. 107–108, 2022.

[33] H. Hojatian, V. N. Ha, J. Nadal, J.-F. Frigon, and F. Leduc-Primeau, “RSSI-based hybrid beamforming design with deep learning,” in *Proc.IEEE ICC*, pp. 1–6, 2020.

[34] K. Chen, J. Yang, Q. Li, and X. Ge, “Sub-array hybrid precoding for massive MIMO systems: A CNN-based approach,” *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 191–195, 2021.

[35] H. Hojatian, J. Nadal, J.-F. Frigon, and F. Leduc-Primeau, “Unsupervised deep learning for massive MIMO hybrid beamforming,” *IEEE Trans. Wirel. Commun.*, vol. 20, no. 11, pp. 7086–7099, 2021.

[36] S. ichi Amari, “Backpropagation and stochastic gradient descent method,” *Neurocomputing*, vol. 5, no. 4, pp. 185–196, 1993.

[37] S. H. Lim, S. Kim, B. Shim, and J. W. Choi, “Deep learning-based beam tracking for millimeter-wave communications under mobility,” *IEEE Trans. on Commun.*, vol. 69, no. 11, pp. 7458–7469, 2021.

[38] H. Akaike, "Autoregressive model fitting for control," *Annals of the Institute of Statistical Mathematics*, vol. 23, pp. 163–180, 1971.

[39] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeuralIPS*, pp. 5998–6008, 2017.

[41] A. Mnih and D. Rezende, "Variational inference for monte carlo objectives," in *International Conference on Machine Learning*, pp. 2188–2196, PMLR, 2016.

[42] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics* (D. van Dyk and M. Welling, eds.), vol. 5 of *Proceedings of Machine Learning Research*, (Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA), pp. 448–455, PMLR, 16–18 Apr 2009.

[43] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," in *Proc. NeuralIPS*, vol. 30, 2017.

[44] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[45] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. R. MacCartney, S. Sun, and S. Nie, "Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1152–1163, 2014.

[46] A. Thornburg, T. Bai, and R. W. Heath, "Performance analysis of outdoor mmwave ad hoc networks," *IEEE Trans. on Signal Process.*, vol. 64, no. 15, pp. 4065–4079, 2016.

[47] M. R. Akdeniz, Y. Liu, S. Rangan, and E. Erkip, "Millimeter wave picocellular system evaluation for urban deployments," in *Proc. IEEE Globecom (Wkshps.)*, pp. 105–110, 2013.

[48] S. Rajagopal, S. Abu-Surra, and M. Malmirchegini, "Channel feasibility for outdoor non-line-of-sight mmwave mobile communication," in *Proc. IEEE VTC (Fall)*, pp. 1–6, 2012.

[49] C. Gustafson, K. Haneda, S. Wyne, and F. Tufvesson, "On mmWave multipath clustering and channel modeling," *IEEE Trans. Antennas Propag.*, vol. 62, no. 3, pp. 1445–1455, 2014.

[50] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—with a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, 2017.

[51] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," *CoRR*, vol. abs/1902.06435, 2019.

[52] Remcom, "Wireless insite, http://www.remcom.com/wireless-insite," May, 2023.

[53] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37328–37348, 2018.

[54] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surv. and Tutor.*, vol. 21, no. 3, pp. 2224–2287, 2019.

[55] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmwave beam and blockage prediction using sub-6 ghz channels," *IEEE Trans. on Commun*, vol. 68, no. 9, pp. 5504–5518, 2020.

[56] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, "Fast beamforming design via deep learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1065–1069, 2019.

[57] Y. Wang, M. Narasimha, and R. W. Heath, "Mmwave beam prediction with situational awareness: A machine learning approach," in *Proc. IEEE SPAWC*, pp. 1–5, IEEE, 2018.

[58] T. Erpek, T. J. O'Shea, Y. E. Sagduyu, Y. Shi, and T. C. Clancy, "Deep learning for wireless communications," in *Development and Analysis of Deep Learning Architectures*, pp. 223–266, Springer, 2020.

[59] L. Zhu, K. Odani, Z. Yang, G. Shi, Y. Kan, Z. Chen, and R. Zhang, "Adaptive spike-like representation of eeg signals for sleep stages scoring," in *arXiv*, 2022.

[60] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, *et al.*, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.

[61] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, pp. 241–258, 2020.

[62] 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*, pp. 2727–2731, 2022.

[63] T. S. Cousik, V. K. Shah, J. H. Reed, T. Erpek, and Y. E. Sagduyu, "Fast initial access with deep learning for beam prediction in 5g mmwave networks," in *Proc. IEEE MILCOM*, pp. 664–669, IEEE, 2021.

[64] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, p. 1735–1780, nov 1997.

[65] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[66] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[67] B. Wang, F. Gao, S. Jin, H. Lin, G. Y. Li, S. Sun, and T. S. Rappaport, "Spatial-wideband effect in massive MIMO with application in mmWave systems," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 134–141, 2018.

[68] Z. Wei, F. Liu, C. Masouros, N. Su, and A. P. Petropulu, “Toward multi-functional 6G wireless networks: Integrating sensing, communication, and security,” *IEEE Commun. Mag.*, vol. 60, no. 4, pp. 65–71, 2022.

[69] U. Demirhan and A. Alkhateeb, “Integrated sensing and communication for 6G: Ten key machine learning roles,” *IEEE Commun. Mag.*, 2023.

[70] U. Demirhan and A. Alkhateeb, “Radar aided 6G beam prediction: Deep learning algorithms and real-world demonstration,” in *Proc. IEEE WCNC*, pp. 2655–2660, 2022.