A Scalable Deep Learning Framework for Dynamic CSI Feedback with Variable Antenna Port Numbers

Yu-Chien Lin, Ta-Sung Lee, and Zhi Ding

***Abstract*—Transmitter-side channel state information (CSI) is vital for large MIMO downlink systems to achieve high spectrum and energy efficiency. Existing deep learning architectures for downlink CSI feedback and recovery show promising improvement of UE feedback efficiency and eNB/gNB CSI recovery accuracy. One notable weakness of current deep learning architectures lies in their rigidity when customized and trained according to a preset number of antenna ports for a given compression ratio. To develop flexible learning models for different antenna port numbers and compression levels, this work proposes a novel scalable deep learning framework that accommodates different numbers of antenna ports and achieves dynamic feedback compression. It further reduces computation and memory complexity by allowing UEs to feedback segmented DL CSI. We showcase a multi-rate successive convolution encoder with under 500 parameters. Furthermore, based on the multi-rate architecture, we propose to optimize feedback efficiency by selecting segment-dependent compression levels. Test results demonstrate superior performance, good scalability, and high efficiency for both indoor and outdoor channels.**

***Index Terms*—CSI feedback, scalability, dynamic architecture, massive MIMO, deep learning**

I. Introduction

Massive multiple-input multiple-output (MIMO) technologies play an important role in improving spectrum and energy efficiency of 5G and future generation wireless networks. The power of massive MIMO hinges on the accurate downlink channel state information (CSI) at the base station or gNodeB (gNB). Without uplink/downlink (UL/DL) channel reciprocity assumed in time-division duplxing (TDD) systems, a frequency-division duplexing (FDD) base station typically relies on user equipment (UE) feedbacks for DL CSI acquisition. Yet, the increasing number of transmit antennas envisioned in millimeter wave bands or higher frequencies [1] requires a vast amount of uplink bandwidth and power for CSI feedback. To conserve bandwidth and UE battery, efficient compression of CSI feedback is vital to broad deployment of massive MIMO FDD communications systems.

The 3rd Generation Partnership Project (3GPP) recently released the features of Release 18 [2] which embraces artificial intelligence (AI) and machine learning (ML). In particular, AI and ML are expected for enhancement of CSI feedback (e.g., overhead reduction and improved estimation accuracy). From radio physics, cellular CSI exhibits a limited multipath delay spread (sparsity), making efficient CSI feedback possible. A deep autoencoder framework [3] was proposed with encoders at UEs and a matching decoder at serving station (e.g., gNB in 5G) for CSI compression and recovery, respectively. This and other related works [4]–[7] have demonstrated the potential for efficient CSI feedback and accurate CSI recovery enabled by deep learning technologies. Physical insights with respect to temporal variation of CSI, similar propagation conditions of nearby UEs, and UL/DL radio path reciprocity motivated more recent progresses that leverage various additional knowledge such as CSI time coherence [4], [8], CSI of nearby UEs [9], and UL/DL CSI reciprocity [10]–[14] to aid and improve DL CSI recovery.

Existing deep learning methods attempt to extract underlying mutual dependency among gNB antennas in massive MIMO configuration by simultaneously feeding CSI of all DL antennas into learning machine for joint compression. Such large input size makes it harder to develop a low-complexity and light-weight deep learning models. There have been attempts to directly reduce encoder's model complexity [15]–[18], achieving only limited success. In this work, we utilize the physical insight that *only nearby massive MIMO antennas exhibit non-negligible CSI correlation since gNB antenna array geometrically spans multiple wavelengths*. Accordingly, our light-weight autoencoder should harness the CSI correlation of neighboring antennas (or antenna ports) to compresses large-array CSI via a divide-and-conquer principle (DCP). The proposed DCP substantially decreases input size and, consequently, the deep learning model size. Moreover, since a DCP based learning model only needs CSI from adjacent antennas, the model is scalable to various sizes of massive MIMO antenna arrays.

Channel dissimilarity and inflexible model output size necessitate retraining of learning models for different channel scenarios and compression levels, respectively. To avoid training burdens for different channel scenarios, online training strategies have been suggested with aid of knowledge distillation [19] and transfer-learning [20], [21]. The authors in [22] applied transfer-learning to reduce training cost for multiple compression levels. A related work [23] designed a multi-rate CSI feedback framework and a matching classification model for selecting a *target compression ratio* according to the number of channel clusters. However, to the best of our knowledge, physical connection between compressibility and

Y.-C Lin and Z.Ding are with the Department of Electrical and Computer Engineering, University of California, Davis, CA, USA (e-mail: ycmlin@ucdavis.edu, zding@ucdavis.edu).

T.-S Lee is with the Institute of Communications Engineering, National Yang Ming Chiao Tung University, Taiwan (e-mail: tslee@nycu.edu.tw).

This work is based on materials supported by the National Science Foundation under Grants 2029027 and 2002937 (Lin, Ding) and by the National Science and Technology Council of Taiwan under grant 111-2218-E-A49-024, 111-2622-E-A49-011, 111-2634-F-A49-009 and 110-2221-E-A49-025-MY2. (Lee and Lin).

channel cluster number remains unconfirmed. In this work, we design a novel dynamic-rate CSI feedback framework. We propose a matching classifier to determine the optimal compression level for maximizing codeword efficiency.

Our primary goal is to reduce the deployment cost of deep learning models at the UE for CSI feedback. We systematically simplify deep learning architecture and make it reusable for different array geometries and compression ratios. We develop a scalable dynamic-rate CSI feedback framework along with a lightweight convolutional encoder for deployment at cost-sensitive UEs. Our contributions are summarized below.

- We develop a subarray based (SAB) CSI feedback framework as a new learning-based compression and recovery mechanism that systematically reduces learning model size by exploiting both strong and weak CSI correlations among adjacent and non-adjacent antennas, respectively.
- The SAB framework is scalable to accommodate large antenna sizes and applicable to most existing compression/recovery methods.
- This work provides a light-weighted encoder design for multi-rate CSI feedback framework which only requires hundreds of parameters for low cost UEs.
- Uplink feedback overhead can be further saved by a proposed feedback pruning mechanism with the aid of the sparsity in antenna port domain.
- To optimize codeword efficiency, we develop a dynamic CR for CSI feedback which adjusts codeword lengths to compress and recover segmented DL CSI according to their significance.

II. System Model

A. CSI Estimation via Pilots and Truncation

We consider a single-cell MIMO FDD link in which a gNB activates N_a antenna ports in communication with single-antenna UEs. Following 3GPP technical specifications, sparse pilot symbols (CSI-RS and DMRS) are distributed in frequency domain for downlink transmission. Assuming each subband contains N_f subcarriers with spacing of Δf and a downsampling rate DR_f, the frequency interval between consecutive pilots is $\text{DR}_f \cdot \Delta f$. We denote $\mathbf{h}_i \in \mathbb{C}^{M_f \times 1}$ as DL CSI of the i-th AP at M_f pilot positions. Let superscript $(\cdot)^H$ denote conjugate transpose. By collecting CSI of each AP, a pilot sampled DL CSI matrix $\widetilde{\mathbf{H}}$ relates to the full DL CSI matrix $\bar{\mathbf{H}}$ via

$$\widetilde{\mathbf{H}} = \bar{\mathbf{H}}\mathbf{Q}_{\text{DR}_f} = \left[\mathbf{h}_1\ \mathbf{h}_2\ \cdots\ \mathbf{h}_{N_a}\right]^H \in \mathbb{C}^{N_a \times M_f}, \quad (1)$$

where $\mathbf{Q}_{\text{DR}_f}$ is a downsampling matrix with pilot rate DR_f.

To reduce feedback overhead, we exploit the physical multipath delay sparsity of CSI by transforming full DL CSI into delay domain through discrete Fourier transform (DFT) or discrete cosine transform (DCT). We truncate the insignificant near-zero elements in trailing (large) delay indices as follows:

$$\mathbf{H} = \widetilde{\mathbf{H}} \cdot \mathbf{F} \cdot \underbrace{\begin{bmatrix} \mathbf{I}_{N_t \times N_t} \\ \mathbf{0} \end{bmatrix}}_{\mathbf{T}}, \quad (2)$$

where $\mathbf{F} \in \mathbb{C}^{M_f \times M_f}$ denotes a DFT/DCT matrix and $M_f \times N_t$ matrix $\mathbf{T}$ performs delay domain truncation. Note that the design of matrix $\mathbf{T}$ may varies according to transformation $\mathbf{F}$ and the CSI properties. Matrix $\mathbf{T}$ in Eq. (2) is an example for DCT transformation that drops the last $M_f - N_t$ columns of $\widetilde{\mathbf{H}} \cdot \mathbf{F}$ corresponding to large multipath delays.

B. Deep Learning Compression

Autoencoder has shown successes in several deep learning frameworks. An encoder at UE compresses its estimated DL CSI for uplink feedback and a decoder at gNB recovers the estimated CSI according to the feedback from UE. Assuming negligible CSI elements at large delays, many have exploited convolutional layers to compress and recover the truncated DL pilot CSI via

$$\text{Encoder:} \quad \mathbf{q} = f_{\text{en}}(\mathbf{H}), \quad (3)$$

$$\text{Decoder:} \quad \widehat{\mathbf{H}} = f_{\text{de}}(\mathbf{q}). \quad (4)$$

The decoder should replace the truncated DL CSI $\widehat{\mathbf{H}}$ via zero-padding to transform CSI back from delay domain to estimate DL CSI matrix $\tilde{\mathbf{H}}$ in the subcarrier domain as follows:

$$\widehat{\widetilde{\mathbf{H}}} = \left[\ \widehat{\mathbf{H}}\ \ \mathbf{0}_{N_a \times M_f - N_t}\ \right] \mathbf{F}^H. \quad (5)$$

The CSI recovery accuracy can be measured by the normalized mean square error (NMSE) of the full DL CSI:

$$\text{NMSE}(\widehat{\widetilde{\mathbf{H}}}, \widetilde{\mathbf{H}}) = \sum_{d=1}^{D} \left\| \widehat{\widetilde{\mathbf{H}}}_d - \widetilde{\mathbf{H}}_d \right\|_F^2 \Big/ \left\| \widetilde{\mathbf{H}}_d \right\|_F^2, \quad (6)$$

where subscript d denotes the d-th test. When training the autoencoder, the CSI error due to truncation is unavailable. Hence, autoencoder loss function can simply rely on the truncated DL CSI error

$$\text{NMSE}(\widehat{\mathbf{H}}, \mathbf{H}) = \sum_{d=1}^{D} \left\| \widehat{\mathbf{H}}_d - \mathbf{H}_d \right\|_F^2 \Big/ \|\mathbf{H}_d\|_F^2. \quad (7)$$

III. Multi-rate CSI Feedback Framework with Flexible Number of Antennas

There have been notable progresses in terms of recovery performance among the recent autoencoder-based CSI feedback frameworks [9], [12], [24], [25]. Since UEs often have limited resources [25], an important consideration is the computation complexity and storage needed by the CSI encoder at the UE. Unfortunately, naïve use of autoencoders from image compression for CSI compression requires direct input of full CSI matrix $\mathbf{H}$ as a 2D "image" to deep learning networks for feature extraction. The inevitably large input size necessitates large autoencoder learning models at both UE and gNB, thereby making it highly challenging to effectively reduce model complexity and storage need.

This raises an question: is it necessary to *simultaneously* feed full DL CSI matrix into the model for encoding CSI features across all ports? The answer may vary. In application

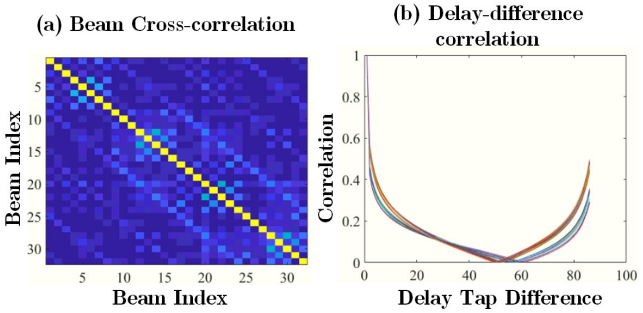

Fig. 1. (a) Cross-correlation between different beams (we consider a 8×4 orthogonal beam set), and (b) correlation versus various delay tap difference (we consider CSIs of 32 antennas denoted by curves with different colors). The low cross-correlation between beams and the high similarity of these curves in delay domain imply the possibility to compress and recovery CSI antenna-by-antenna.

when antenna configuration avoids spatial aliasing [1] (e.g., half-wavelength antenna spacings), CSI correlation across the multiple antenna ports tends to be weak and negligible. Thus, it may be unnecessary to import CSIs across many antennas of the same MIMO configuration to the UE encoder for compression and feedback.

We can gain some insights from the following test results. Figs. 1 (a) and (b) show the correlation between different antennas and the statistics at different delay taps for different antennas. It is apparent that correlation between antennas is weak and, in fact, CSI statistics at different delay taps even for different antennas appears similar. This recognition motives us to propose to apply a common and smaller deep-learning model to encode and decode the DL CSI across large number of antenna ports when distinct antennas serve as multiple activated ports.

A. Subarray Based (SAB) Framework

Previous works such as [3], [11], [12], [26] send full CSI matrix like an image as encoder input for compression. Such 2-D CSI structure in antenna and delay domains is akin to a natural 2-D image. However, from the preliminary results of Figs. 1 (a) and (b), the inter-antenna independence and similar statistics of delay profile of different antennas motivate a simpler subarray based (SAB) CSI encoding and decoding framework. In this section, we propose an SAB framework which divides a full DL CSI into non-overlapping several subarray pieces before their individual compression and gNB recovery.

We first define a new quantity, *subarray width*, as the spatial domain width of the new framework input. Let subarray width be K to capture K consecutive antenna ports among the N_a rows of the CSI matrix that exhibit correlation [25]. We concatenate real and imaginary parts of the full DL CSI matrix in an interleaving manner as an augmented real-value full DL CSI matrix $\mathbf{H}_{\text{aug}} = \left[\text{Real}(\mathbf{h}_1)\ \text{Imag}(\mathbf{h}_1)\ \text{Real}(\mathbf{h}_2) \ldots \text{Imag}(\mathbf{h}_{N_a}))\right]^T$ of size $2N_a \times N_t$ before partitioning the $2N_a$ rows to form $2N_a/K$ matrices of size $K \times N_t$ as follows:

$$\mathbf{H}_i = \mathbf{H}_{\text{aug}}(Ki{+}1 : Ki{+}K, :),\ i = 0,\ 1, ..., N_a/K - 1. \quad (8)$$

We train a common autoencoder for each of the K subarray CSIs. Each subarray matrix $\mathbf{H}_i$ enters the common encoder $\mathbf{q}_i = f_{\text{en}}(\mathbf{H}_i)$ at UE for compression and feedback. At the gNB, the decoder $\widehat{\mathbf{H}}_i = f_{\text{de}}(\mathbf{q}_i)$ recovers the subarray CSI before stacking them back into the full DL CSI matrix

$$\widehat{\mathbf{H}}_{\text{aug}} = \left[\widehat{\mathbf{H}}_1; \widehat{\mathbf{H}}_2; ...; \widehat{\mathbf{H}}_{N_a/K}\right] \quad (9)$$

By extracting rows at the odd and even indexes, we can obtain the estimate of the full DL CSI matrix $\widehat{\mathbf{H}}$.

B. Multi-rate CSI Feedback Framework

In practical applications, physical environment affects the MIMO CSI characteristics including its sparsity and entropy. Therefore, the degree to which an MIMO CSI can be compressed in a deep learning framework would vary with physical environment. Without knowing the actual CSI a priori, multiple encoder-decoder pairs may have to be deployed at UEs and gNB to achieve the required accuracy and feedback compression. Training multiple encoders would lead to higher memory use to store the models and possibly higher complexity to test the outcomes of different compression models (i.e., ratios).

To this problem, the authors of [15] proposed a multi-rate CSI framework as illustrated in Fig. 2. Its encoder of [15] can generate 4 different output arrays of 4 distinct compression ratios. The parameters of all layers in its encoder are common except for a final fully-connected (FC) layer. This framework of [15] reduces the total number of encoder parameters by enforcing convolutional layers for different compression ratios to remain the same so as to generate similar features. Only the final layer decides the encoder output for feedback at different compression ratios.

In this paper, we consider a similar architecture but proposing a new encoder design with fully convolutional layers and the proposed SAB framework. We name the new architecture "successive convolutional encoding network (SCEnet)" whose model complexity can be significantly tamed while preserving good recovery performance. To achieve a good tradeoff between performance and model complexity, we focus on complexity reduction at the encoder for low cost UEs. For the UE encoder, we introduce a fully-convolutional down-sizing block (FCDS) to lower the input size by half. The FCDS block consists of 1×7, 1×5 and 1×3 convolutional layers with 2 channels, respectively. Note that the stride lengths are all 1 except for the final horizontal stride in the last convolutional layer which is of length 2 to drop the input size by half. Fig. 3 shows an example of a CSI feedback framework using S FCDS blocks for dealing with 4 compression ratios (S= 4 throughout this paper). Specifically, the output of i-th block with size of $K \cdot N_t/2^i$ represents codewords with compression ratio = $2^i, i = 1, ..., S$.

[1] As a rule of thumb, CSI of antennas spaced more than one wavelength apart are nearly independent.

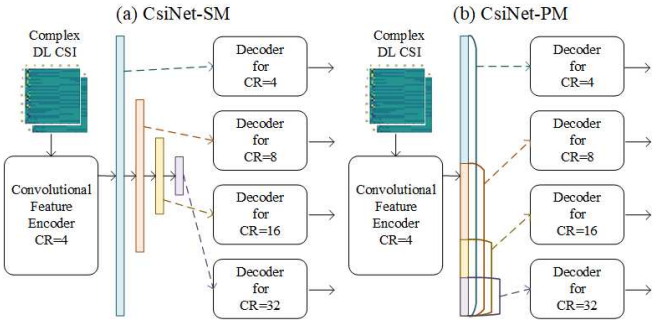

Fig. 2. Illustration of previous multi-rate CSI feedback frameworks, CsiNet-SM and CsiNet-PM. The encoders share model parameters at different compression ratios except for FC layers, which contribute the majority of model complexity.

Since gNBs are less resource constrained, individual CSI decoder is designed for each compression ratio. For the i-th decoder, the codeword is first fed to a $K \cdot N_t$ FC layer, a 1×3 convolutional layer and activation function after reshaping for initial estimation. An ensuing RefineBlock [15] provides refinement. RefineBlock uses a residual structure and consists of three 1×3 convolutional layers with 16, 8 and 1 channels, respectively. The RefineBlock is followed by a $K \cdot N_t$ FC layer for generating real/imaginary CSI estimates. To further improve recovery accuracy, we provide another SCNnet, called SCEnet+ by adding an additional FC layer at the end of each FCDS block which provides extra non-linearity at the same output size.

The parameters of the SCEnet are optimized according

$$\Omega_{en}, \Omega_{de} = \arg\min \sum_{i=1}^{D} \sum_{s=1}^{S=4} W_s \cdot \left\| \mathbf{H}_i - \widehat{\mathbf{H}}_{i,s} \right\|_{\mathrm{F}}^2, \quad (10)$$

$$\widehat{\mathbf{H}}_{i,1}, \widehat{\mathbf{H}}_{i,2}, \widehat{\mathbf{H}}_{i,3}, \widehat{\mathbf{H}}_{i,4} = f_{de}(f_{en}(\mathbf{H}_i)), \quad (11)$$

where subscript s denotes the outcome from the s-th compression ratio and Ω_{en}, Ω_{de} denote the trainable parameters of encoder f_{en} and decoder f_{de}. D is the training data size. In [15], hyper-parameters $\{W_1, W_2, W_3, W_4\}$ were chosen as $\{30/39, 6/39, 2/39, 1/39\}$.

IV. Multi-Rate CSI Feedback Framework with Flexible Number of Antenna Ports

The proposed SAB framework can effectively reduce the model size and computational complexity. However, the uplink feedback overhead is not lower with this framework. To reduce feedback information, we observe that CSI in beam domain (i.e., angular domain) appears to be sparse. For instance, outdoor propagation channels usually characterized with its low angular spread [**?**]. If we transform CSI matrices from antenna (i.e. spatial) domain to beam domain before compression and recovery with the proposed SAB framework, we may require fewer or even no codewords for those subarray CSIs with negligibly low energy. With this motivation, we propose a *DCP feedback pruning* mechanism to further reduce the uplink information for CSI feedback and the computational complexity of encoding/decoding at UE and gNB, respectively.

A. SAB framework in beam-delay (BD) domain

We reprsent the full CSI matrix in antenna-delay domain as

$$\mathbf{H}_{\mathrm{AP}} = \mathbf{M} \cdot \bar{\mathbf{H}} \cdot \mathbf{F} \cdot \underbrace{\begin{bmatrix} \mathbf{I}_{N_t \times N_t} \\ \mathbf{0} \end{bmatrix}}_{\mathbf{T}}, \quad (12)$$

where $\mathbf{M} \in \mathbb{C}^{N_a \times N_a}$ is an orthogonal transformation matrix transforming from antenna to antenna port (AP) domain. Without loss of generality, we can have a DL beam domain (BD) CSI matrix $\mathbf{H}_B$ by designing an orthogonal beam matrix $\mathbf{M} = \mathbf{B}$ which be found via the mechanism in [27]. Following the same preprocessing in the previous section, we first concatenate real and imaginary parts of CSIs as an augmented DL BD matrix $\mathbf{H}_{B,\mathrm{aug}}$ and divide the augmented DL BD matrix into $2N_a/K$ subarray CSI matrices of the same size $K \times N_t$ given below

$$\mathbf{H}_{B,i} = \mathbf{P}_i \mathbf{H}_{B,\mathrm{aug}}, \ \forall i = 1, 2, ..., N_a/K. \quad (13)$$

Thus, the parameters of the SCEnet are optimized according to criterion:

$$\Omega_{en}, \Omega_{de} = \arg\min \sum_{i}^{D} \sum_{s}^{S=4} W_s \cdot \left\| \mathbf{H}_i - \mathbf{B}^H \widehat{\mathbf{H}}_{B,i,s} \right\|_{\mathrm{F}}^2, \quad (14)$$

$$\widehat{\mathbf{H}}_{B,i,1}, \widehat{\mathbf{H}}_{B,i,2}, \widehat{\mathbf{H}}_{B,i,3}, \widehat{\mathbf{H}}_{B,i,4} = f_{de}(f_{en}(\mathbf{H}_{B,i})). \quad (15)$$

B. DCP Feedback Pruning

Due to small angular spread, outdoor CSIs in beam domain are usually sparse in angular domain. To take advantage of this physical property, we propose a DCP feedback pruning method to exploit the beam sparsity to further reduce the uplink feedback overhead and encoding/decoding computations by skipping feedback of those insignificant subarray CSI matrices of negligibly low Frobenius norm.

To evaluate whether a subarray CSI matrix is insignificant, we measure its relative energy ratio

$$R_{E,i} = \left\| \mathbf{H}_{B,i} \right\|_F^2 / \left\| \mathbf{H}_{B,\mathrm{aug}} \right\|_F^2. \quad (16)$$

Subarray CSI matrices with energy ratio below a predefined threshold T are regarded as insignificant and are ignored at the UE encoder. Importantly, UEs need to transmit extra information bits to indicate insignificant subarray to the gNB during feedback.

To minimize the information bits, as illustrated in Fig. 4, we suggest that UE could utilize a prefix bit indicating whether to send a *zero-skipping request* to base station. As depicted in Fig. 5, the additional bit is appended before the bit stream of each subarray CSI matrix as a prefix which is decoded first at gNB to avoid the subsequent CSI recovery for the insignificant subarray CSI matrix. For subarray CSI matrix with energy ratio $R_E \geq T$, UE encodes the CSI matrix and the codeword feedback on uplink to gNB with the indicator bit $= 1$. Otherwise, UE sends zero uplink feedback with indicator bit $= 0$. Alternatively, a $2N_a/K$ bitmap can lead or

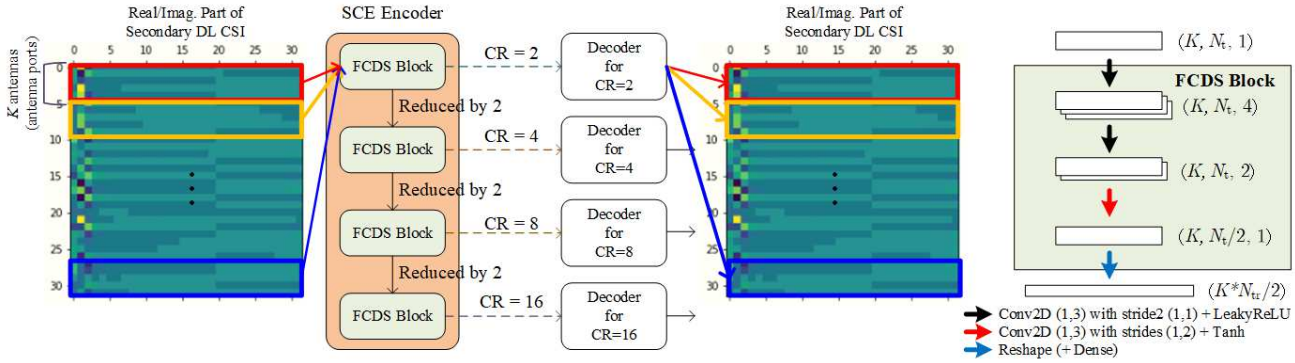

Fig. 3. SAB Framework and SCE Network Architecture. Input data are first split into real and imaginary and separated into subarray matrices. These matrices are fed to the SCE network and recovered in parallel. Note that, at encoder, after each FCDS block, the total size of input is reduced by half. The fully convoluted FCDS blocks share parameters. We also provide another alternative encoder (SCEnet+ encoder) where a FC layer is attached at the end of each FCDS block for enhancing performance.

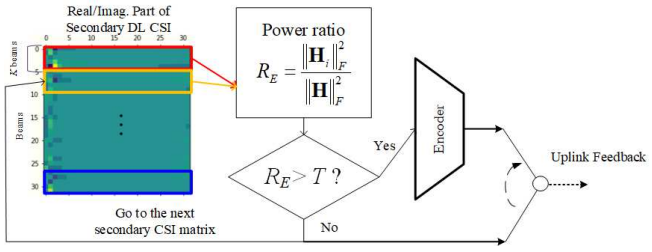

Fig. 4. Illustration of DCP Feedback Pruning. If the energy ratio of the i-th DL BD subarray CSI matrix is less than the predetermined threshold T, UE skips encoding and send only one bit to tell base station to fill zeros in the corresponding region of the DL BD subarray CSI matrix. Otherwise, UE operates SAB framework normally.

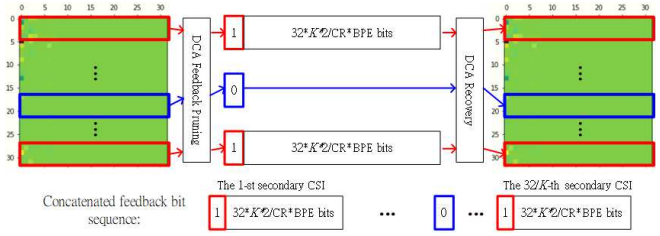

Fig. 5. DCP feedback pruning block diagram and ordered feedback bit sequence.

trail the CSI codeword feedback as indicators to the decoder. The gNB examines these indicator bits to decide whether to decode the corresponding subarray CSI codeword or to zeropad the corresponding subarray CSI before moving onto the next subarray CSI.

By doing so, a larger threshold T tends to skip more encoding/decoding process, use less uplink bandwidth for feedback, but possibly cause performance degradation due to the zero-skipping process. Thus, the selection of threshold T becomes a trade-off between the amount of uplink feedback overhead and recovery performance. Fortunately, due to the sparsity in angular domain, we can effectively reduce uplink feedback bandwidth and computations while not sacrificing too much recovery performance in general, especially for channels with low angular spreads.

C. Local Normalization

Recall that one assumption for SAB framework is the similar statistics of delay profile of different antennas. After transforming CSI from antenna to beam domain, although the relative delay profile is still similar for different beams, CSI energy concentrates in a few specific angles (directions) for in most propagation with low angular multipath spreads. As a result, CSI recovery may degrade because of training bias in which deep learning model endeavor to recover those stronger subarray CSI matrices better. This may lead to very poor recovery performance for subarray CSI matrices of modest energy. To tackle this problem, as depicted in Fig. 6(b), we let UE normalize each encoded subarray CSI matrix individually and encode the normalization factor as a feedback to gNB.

D. 2D Lightweight Encoder

In this section, we proposed a SAB framework in BD domain along with subarray row feedback and pruning to further reduce uplink feedback overhead by taking advantages of its beam domain sparsity. In fact, sparsity is also observed in the delay domain. A natural extension is develop a two-dimensional (2D) SAB framework as illustrated in Fig. 7 along with feedback pruning method to skip near-zero CSI matrix blocks for reducing uplink feedback bandwidth.

However, overly aggressive model reduction as such requires the CSI energy to be not only similarly distributed in the delay domain across antenna ports but also similarly distributed in spatial domain for each delay. Such property has not been experimentally verified. Therefore, although a 2D lightweight encoder admits a low complexity autoencoder structure, we must carefully weigh the complexity-accuracy tradeoff of such efforts.

V. SAB Framework with Dynamic CR

The proposed SAB feedback switches on/off the encoding of CSI subarrays for achieving a higher effective compression

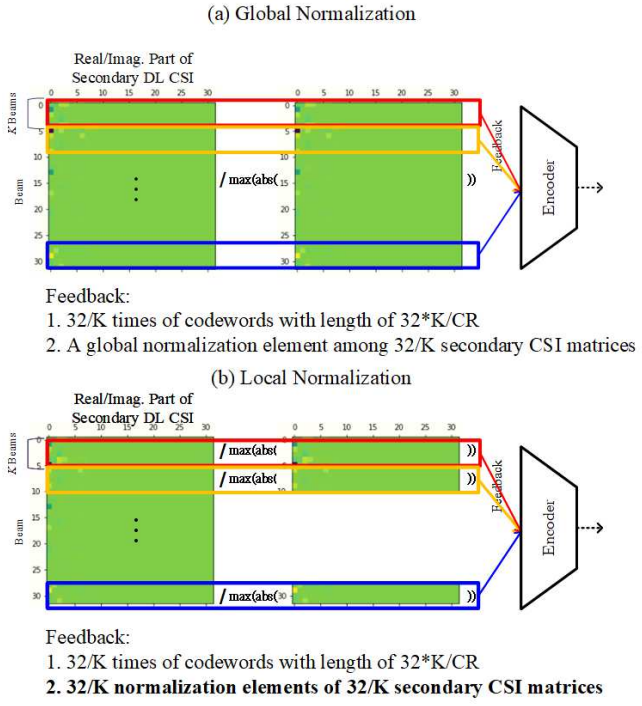

Fig. 6. Illustrations of (a) global normalization and (b) local normalization.

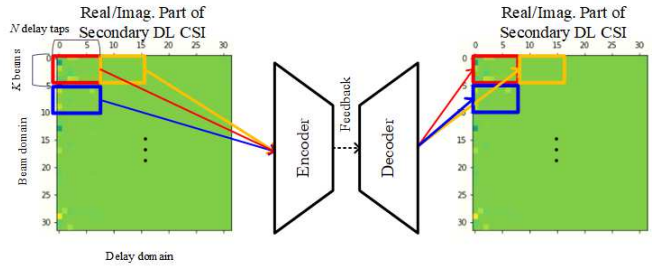

Fig. 7. Illustrations of 2D SAB framework (N and K are the numbers of delay taps and beams being considered in a single subarray CSI matrix, respectively).

ratio. We can utilize more feedback resource on high-energy subarray CSI matrices especially for channels with sparse distribution in beam or angular domain such as outdoor channels. Yet, instead of using a hard decision to determine whether to feedback or skip the encoding/recovery process, the multi-rate architecture motivates a softer decision approach. Here, we propose a dynamic CR CSI feedback framework which compresses subarray CSI matrices in a full DL CSI using dynamic CR by a energy-based CR selector according to their significance (i.e., normalized subarray CSI energy) to maximize the codeword efficiency (CE) of CSI feedback.

To define the efficiency of codeword, we measure the expected capacity provided by each codeword. With orthogonal multiple access, when using estimated full CSI $\widehat{\widetilde{\mathbf{H}}} = [\widehat{\widetilde{\mathbf{H}}}_1; \widehat{\widetilde{\mathbf{H}}}_2; ...; \widehat{\widetilde{\mathbf{H}}}_{2N_a/K}]$ as a maximum-ratio combining (MRC) precoder for DL transmission at gNB, the expected capacity for the i-th subarray CSI matrix in DL transmission can be reasonably set as

$$\mathcal{C}_i = \log_2(1 + \text{SNR}_i) \tag{17}$$

$$\text{SNR}_i = \frac{\left\|(\widehat{\widetilde{\mathbf{H}}}_i)^* \widetilde{\mathbf{H}}_i / \left\|\widehat{\widetilde{\mathbf{H}}}_i\right\|_{\text{F}}\right\|_{\text{F}}^2}{K \cdot N_f \cdot P_N} \tag{18}$$

where $\left\|(\widehat{\widetilde{\mathbf{H}}}_i)^* \widetilde{\mathbf{H}}_i / \left\|\widehat{\widetilde{\mathbf{H}}}_i\right\|_{\text{F}}\right\|_{\text{F}}^2 / (K \cdot N_f)$ and P_N denote the average signal and noise power, respectively, over N_f subcarriers and K antenna ports. $\widetilde{\mathbf{H}}_i$ denotes true subarray CSI matrix.

Let the sum length of uplink feedback codeword $\mathbf{q} = [\mathbf{q}_1; \mathbf{q}_2; ...; \mathbf{q}_{2N_a/K}]$ from UE to gNB be $L = \sum_{i=1}^{2N_a/K} L_i$. We can define the average CE as

$$\text{CE} = \sum_{i=1}^{2N_a/K} \frac{\mathcal{C}_i}{L_i} \frac{K}{2N_a} (\text{bits/s/Hz/codeword}). \tag{19}$$

This metric measures the contribution of each codeword to the eventual end-to-end CSI feedback performance.

Take the multi-rate CSI feedback framework, DCnet, as an example, it provides four distinct lengths of codewords (corresponding to four compression ratios) for different compressing/recovery quality. To achieve the best performance, we should compress CSI with the least compressive codewords and vice versa. There always exists a trade-off between uplink feedback cost and recovery performance. Yet, although there is no best choice of compression ratio, the most efficient one exists.

By dividing a full-size CSI matrix into several subarray CSI matrices, we discover that only a fraction of subarray CSI matrices dominate in terms of energy. That is, if we could recover those subarray CSI matrices well, we will have a high-quality CSI recovery even if other subarray CSI matrices are recovered with large errors. Hence, to improve feedback efficiency, we should utilize more resources (i.e., $CR = 2$) on subarray CSI matrices with larger significance (i.e., higher energy) and less resources ($CR = 16$) on those with less significance. We first evaluate the significance of the i-th subarray CSI matrix for each data sample according to its normalized CSI energy $R_{E,i}$ defined in (16).

We design a energy-based CR selector which selects CR according to the normalized energy of subarray CSI matrices. The CR determined by the CR selector for the i-th subarray CSI matrix is given by

$$\text{CR}_i = \begin{cases} 2 & a_0 \le R_{E,i} < a_1 \\ 4 & a_1 \le R_{E,i} < a_2 \\ 8 & a_2 \le R_{E,i} < a_3 \\ 16 & a_3 \le R_{E,i} \le a_4 \end{cases} \tag{20}$$

As illustrated in the Fig. 8, there are five anchor points $\mathbf{a} = [a_0 = 1, a_1, a_2, a_3, a_4 = 0]$ where $1 \ge a_1 \ge a_2 \ge a_3 \ge 0$ and a_1, a_2, a_3 are trainable. If we optimize the three anchor points by maximizing CE in Eq. 21, since the nominator does not grow proportionally as the denominator increases, we will have a trivial CR selector which always suggest to adopt the largest

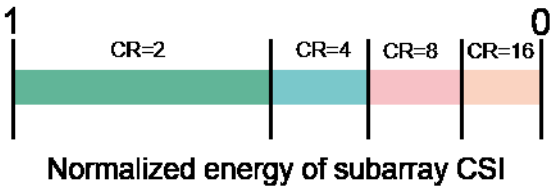

Fig. 8. Five anchor points of normalized energy of subarray CSI matrix for CR decision. Note that we only need to train the three anchor points a_1,a_2,a_3 for separating the operating regions of four CRs.

CR to achieve the highest codeword efficiency. Unfortunately, this induces a fairness problem since the CR selector tends to secure CE and ignore those CSI estimates with extremely poor performance. Those cases should be considered as recovery failure. Thus, using a standard step function $u(.)$, we define the mean outage capacity as

$$\text{CE} = E\left\{\sum_{i=1}^{2N_a/K} \frac{\mathcal{C}_i}{L_i} \cdot u(\text{NMSE}_i - T_{\text{out}})\right\} \tag{21}$$

We define an outage threshold T_{out} to reject cases when gNB totally fails to estimate DL CSI. In the training stage, as a rule of thumb, a typical value of T_{out} is set as -10 dB.

In this paper, we provide a heuristic training strategy for searching optimal points by following Alg. 1. Note that, since we consider four possible CRs, we need extra two-bit information for each subarray CSI matrix in the uplink feedback to gNB for correctly identifying the correct decoder of the corresponding CR as shown in Fig. 9.

Algorithm 1 Multi-point linear searching algorithm

Require: $\mathbf{a} = [1, 0, 0, 0, 0]$, N_{ter}, N, $\text{CE}_f = 0$, $\Omega = \{1, 2, 3\}$
Ensure: $\mathbf{a} = [1, a_1, a_2, a_3, 0], \text{CE}_f$
  for $i = 1 : 1 : N_{\text{ter}}$ **do**
    $j \leftarrow \text{mod}(i, \text{length}(\Omega)) + 1$
    $\mathbf{v}_f \leftarrow [a_j - \frac{a_{j-1}-a_j}{(N/2)+1}; ...; a_j - (N/2)\frac{a_{j-1}-a_j}{(N/2)+1}]$
    $\mathbf{v}_b \leftarrow [a_j + \frac{a_j-a_{j+1}}{(N/2)+1}; ...; a_j + (N/2)\frac{a_j-a_{j+1}}{(N/2)+1}]$
    $\mathbf{v} \leftarrow [\mathbf{v}_f; \mathbf{v}_b]$
    $\mathbf{a}_{\text{old}} \leftarrow \mathbf{a}$
    flag $\leftarrow$ False
    for $k = 1 : 1 : N$ **do**
      $a_{\Omega_j} \leftarrow \mathbf{v}[k]$
      Evaluate CE according to $\mathbf{a}$
      if $\text{CE} > \text{CE}_f$ **then**
        $\text{CE}_f \leftarrow \text{CE}$
        flag $\leftarrow$ True
      end if
    end for
    if $|a_2 - a_1| < 0.005$ **then**
      $\Omega = \{[1, 2], 3\}$
    else if $|a_3 - a_2| < 0.005$ **then**
      $\Omega = \{[1], [2, 3]\}$
    else if $|a_2 - a_1| < 0.005$ and $|a_3 - a_2| < 0.005$ **then**
      $\Omega = \{[1, 2, 3]\}$
    end if
    if flag $=$ False **then**
      $\mathbf{a} \leftarrow \mathbf{a}_{\text{old}}$
    end if
  end for

VI. Experimental Evaluations

A. Experiment Setup

In our experiments, we consider both indoor and outdoor cases. Using channel model software [28], we place a gNB of height equal to 20 m at the center of a circular cell with a radius of 30 m for indoor and 200 m for outdoor environment. The gNB equipped with a $8 \times 4 (N_H \times N_V)$ UPA for communicates with single-antenna UEs. UPA elements have half-wavelength uniform spacing.

For our proposed model and other competing models, we set the number of epochs to 1000. We use batch size of 200. For our model, we start with learning rate of 0.001 before switching to 5×10^{-4} after 300 epochs. Using the channel simulator, we generate several indoor and outdoor datasets, each containing 100,000 random channels. One seventh of these channels is test data for performance evaluation. Two and one thirds of the remaining are for training and validation. For both indoor and outdoor, we use the QuaDRiGa simulator [28] using the scenario features given in *3GPP TR 38.901 Indoor* and *3GPP TR 38.901 UMa* at 5.1-GHz and 5.3-GHz, and 300 and 330 MHz of UL and DL with LOS paths, respectively. To accurately assess recovery accuracy, we assume UEs are capable of exact CSI estimation. For each data channel, we consider $N_f = 1024$ subcarriers with $15K$-Hz spacing and place $M_f = 86$ pilots with downsampling ratio $\text{DR}_f = 12$ as illustrated in the Fig. 10. We set antenna type to *omni*. We use NMSE Eq. 6 as the performance metric.

B. SCEnet vs. SCEnet+

Figs. 11 (a) and (b) summarize NMSE performance for the two proposed models at different compression ratios in indoor and outdoor scenarios, respectively. We observe the benefits of the extra FC layer at encoder for low compression ratios. Considering the negligible error improvement in linear scale, SCEnet and SCEnet+ achieve similar performance. Yet, SCEnet+ has more flexible coding rate owing to the use of FC layers. For brevity, we use SCEnet+ as our benchmark in the rest of this section.

C. Performance, Complexity and Storage Comparison

For comparison, besides the proposed models SCEnet and SCEnet+, we also include two recent multi-rate CSI feedback alternatives which **take full DL CSI as model input** and are listed below:

- **CsiNet-SM** [15]: Fig. 2 (a) shows its general architecture. Note that we accommodate the model for desired compression ratios by adjusting the size of FC layers.
- **CsiNet-PM** [15]: Fig. 2 (b) shows the general architecture. Note that CsiNet-PM is a more compact model than

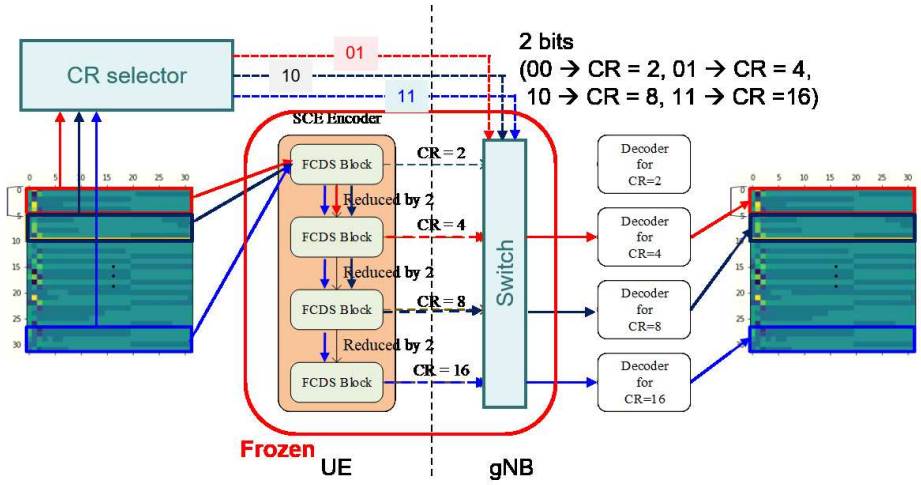

Fig. 9. UE sends extra two bits for each subarray CSI matrix to indicate the adopted CR. gNB selects the corresponding decoder according to the extra information.

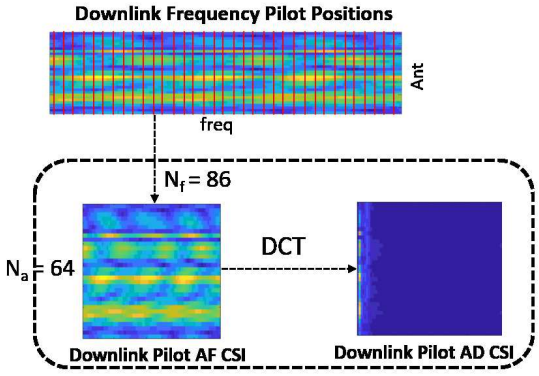

Fig. 10. Pilot placement illustration (Note that the red lines indicate the time-frequency resources to be placed pilot symbols. AF and AD stand for antenna-frequency and antenna-delay domains, respectively).

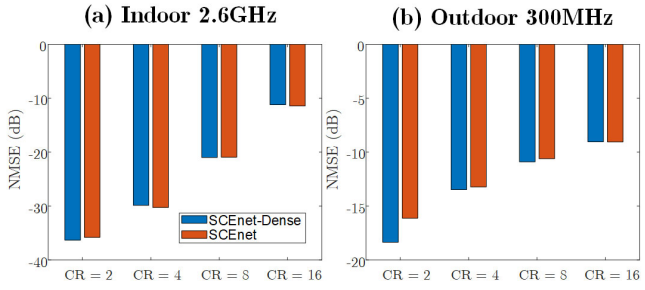

Fig. 11. NMSE performance at different compression ratios for SCEnet and SCEnet+ in indoor and outdoor scenarios.

CsiNet-SM but suffers slight performance degradation in general.

Note that the proposed models adopt a similar decoder as the alternatives in comparison with required accommodations such as reduced sizes of FC layers and one dimensional convolutional filter size (i.e., (1,3), (1,5) and (1,7)).

TABLE I
FLOATING-POINT OPERATION OF ALTERNATIVES IN COMPARISON. COMPRESSION RATIO IS 8 FOR CALCULATING DECDOER'S FLOP NUMBERS.

	SCEnet	SCEnet+	CsiNet-SM	CsiNet-PM
Encoder FLOPs	1.16M	1.4M	4.3M	2.2M
Decoder FLOPs (K=2)	10.7M		49.4M	
Decoder FLOPs (K=4)	12M			
Decoder FLOPs (K=8)	14.75M			

Most UEs have strict memory, computation and power constraints, thereby favoring light-weight and simpler encoders for deployment. Figs. 12 (a) and (b) model size of encoder and decoder, respectively, for SCEnet, SCEnet+, CsiNet-SM, and CsiNet-PM. Table I reveals computation complexity of encoder and decoder for alternatives in comparison. Table II shows the NMSE performance at different compression ratios and subarray width (K) for SCEnet+, CsiNet-SM and CsiNet-PM including both indoor and outdoor scenarios. We observe that SCEnet+ with $K = 64$ generally outperforms CsiNet-SM and CsiNet-PM and requires less FLOP number and storage at UE side. Leveraging the SAB framework of smaller subarray width K, we enjoy much lower complexity and storage with slight performance degradation. The selection of $K = 2$ yields an acceptable recovery performance and delivers several orders of encoder and decoder size reduction as well[2]. Moreover, SCEnet+ becomes scalable and can be a universal CSI feedback framework which can be applied to CSI feedback with various numbers of antenna ports (according to the 3GPP specification, 2, 4, 8, 16, 32 are possible antenna port number).

[2]The major model size reduction is attributed to smaller input size. However, smaller input size does not simplify the computation complexity by the same order. Although FLOP number grows proportionally with input size, the encoder is applied to multiple subarray CSI matrices.

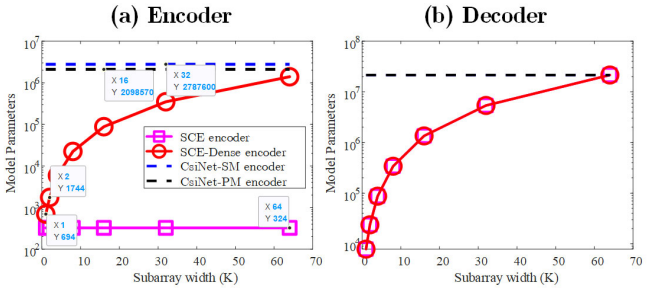

Fig. 12. (a) Encoder and (b) decoder model size comparison of SCEnet, SCEnet+, CsiNet-SM and CsiNet-PM.

TABLE II
NMSE PERFORMANCE OF THE CSINET-SM AND SCENET FOR DIFFERENT SELECTIONS OF SUBARRAY WIDTH (K).

CR	Scen.	SCEnet+				CsiNet -SM	CsiNet -PM
		K=2	K=4	K=8	K=64		
2	Ind.	-39.2	-38.8	-36.7	**-39.6**	-29.7	-29.8
	Out.	-17.8	-16.3	-16.1	**-19.8**	-18.9	-18.8
4	Ind.	-31.7	**-32.0**	-31.9	-31.5	-26.0	-25.9
	Out.	-13.6	-13.3	-12.6	-14.7	**-15.3**	-14.5
8	Ind.	-20.7	-21.8	-22.2	**-24.3**	-20.3	-19.1
	Out.	-11.5	-11.0	-10.6	**-12.7**	-12.3	-11.2
16	Ind.	-12.8	-12.3	-11.9	**-15.4**	-13.0	-12.0
	Out.	-10.3	-9.7	-9.5	**-11.5**	-10.2	-9.2

D. Testing Different Encoder/Decoder Pairs

To show the efficacy of SAB framework, Fig. 13 shows the NMSE performance at different compression ratios and three encoder/decoder pairs: 1) SAB encoder plus SAB decoder 2) SAB encoder plus pooling decoder 3) full-size encoder and decoder. We consider a subarray width of 2 for SAB encoder and decoder. Pooling decoder consists of 32 copies of SAB decoder and is followed by 2 residual blocks with 3×3 convolutional layers with 16, 8, 1 channels for pooling purpose. A full-size encoder and decoder are the SAB ones with $K = 64$. With respect to limited correlation between antennas, we can observe that the SAB encoder/decoder pair only causes slight performance degradation while requiring much less storage and computational burdens for UEs and base stations.

E. Testing Different Array Geometries

To show the scalibility of SCEnet+, Fig. 14 shows the NMSE performance at different compression ratios and array geometries (8-element ULA, 16, 32-element UPAs) in indoor and outdoor scenarios. The results show no obvious performance difference for arrays of different sizes. This demonstrates the scalibility of the proposed SAB framework.

F. BD SAB Framework in GN and LN approaches

The sparsity of CSI matrix in beam domain allows DCP feedback pruning for further uplink feedback reduction. On the other hand, it may cause power imbalance across antenna ports and performance degradation. Fortunately, this problem could be mitigated by LN.

Fig. 15 shows the NMSE performance by applying GN

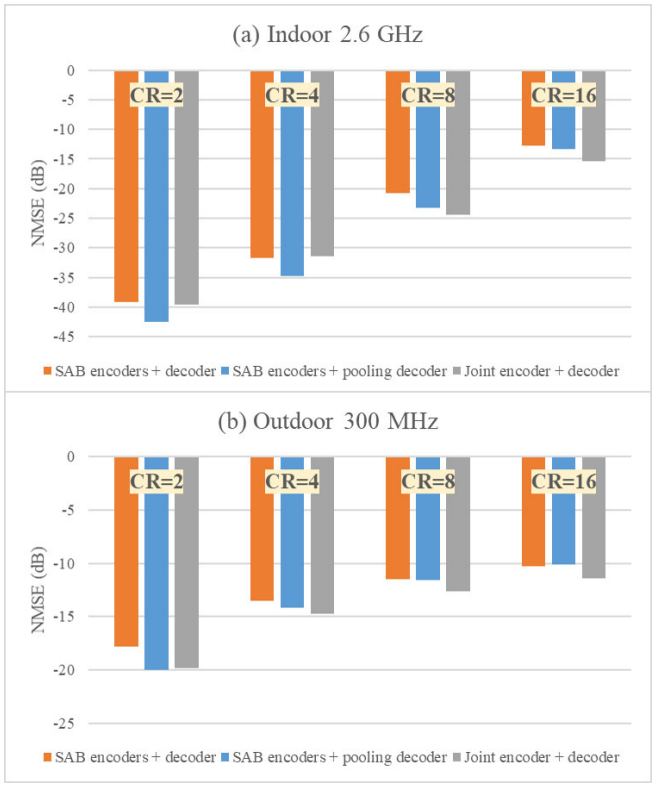

Fig. 13. NMSE performance versus compression ratios with different encoder/decoder pairs in (a) indoor and (b) outdoor scenarios.

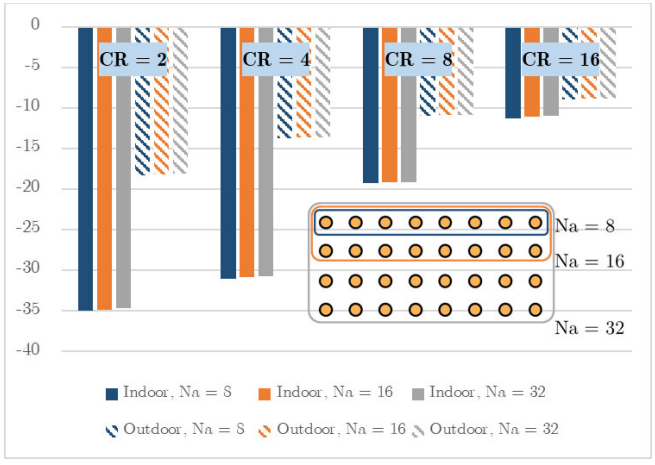

Fig. 14. NMSE performance of SCEnet+ for arrays with different array geometries. We consider 8-element ULA and 16- and 32-element UPAs.

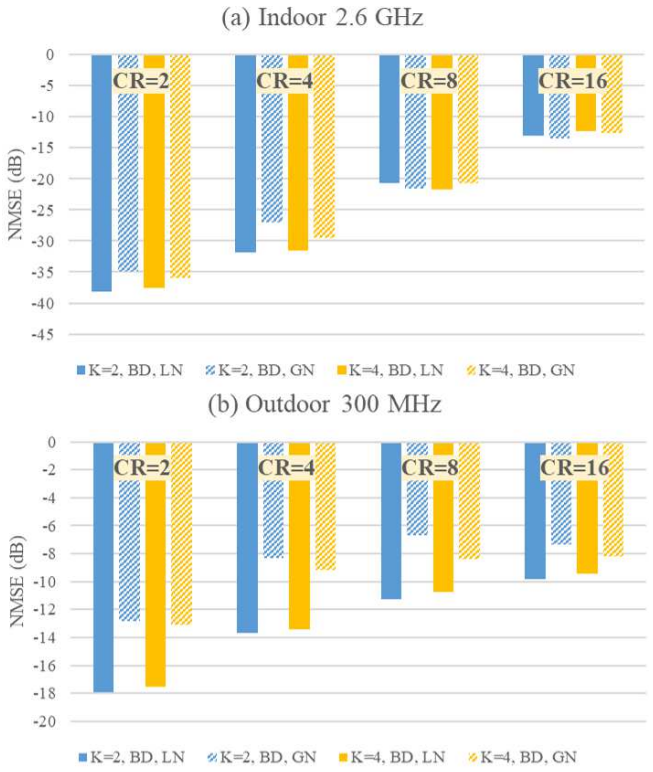

Fig. 15. NMSE performance versus compression ratios with or without local normalization (LN) in (a) indoor and (b) outdoor scenarios.

and LN to SCEnet+ when $K = 2$ and 4 in indoor and outdoor channels. We observe better performance by selecting a smaller subarray width K because of limited correlation between adjacent beams. Additionally, we also see that performance improvement, especially for outdoor scenario, is achieved by utilizing LN approach. Since outdoor channels characterize with its low angular spread, this causes severe power imbalance problem over different subarray BD CSI matrices when using GN approach. The experiment results show that LN can effective alleviate power imbalance problem. Note that LN is adopted in the following results.

G. DCP feedback pruning

In DCP feedback pruning, only subarray CSI matrices with energy ratio larger than T are encoded and fed back. The remaining are fed back to gNB with a bit "zero" as illustrated in Fig. 5. For a better understanding, we define a metric, called *pruning ratio*, to be the ratio of the number of encoded subarray CSIs to all. Note that a larger T can increase pruning ratio but cause performance degradation.

Figs. 16 and 17 show the NMSE performance under different pruning ratios in indoor and outdoor scenarios, respectively. The results suggest that the degradation of 20% pruning (pruning ratio = 0.2) is acceptable. Although low compression ratios appear to exhibit more severe performance loss in logarithm-scale, the actual discrepancy in MSE is quite small. From Fig. 17, we can observe that pruning exhibits more advantages in outdoor case. It is because its high sparsity

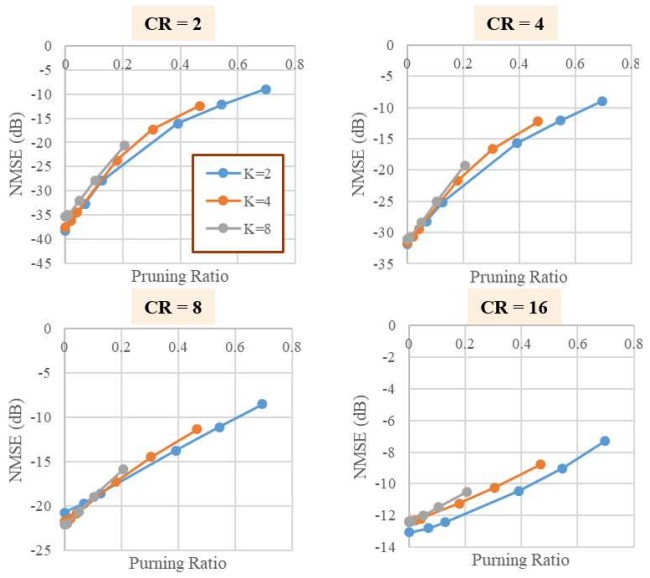

Fig. 16. NMSE performance versus pruning ratio for different selections of subarray width K in indoor scenario.

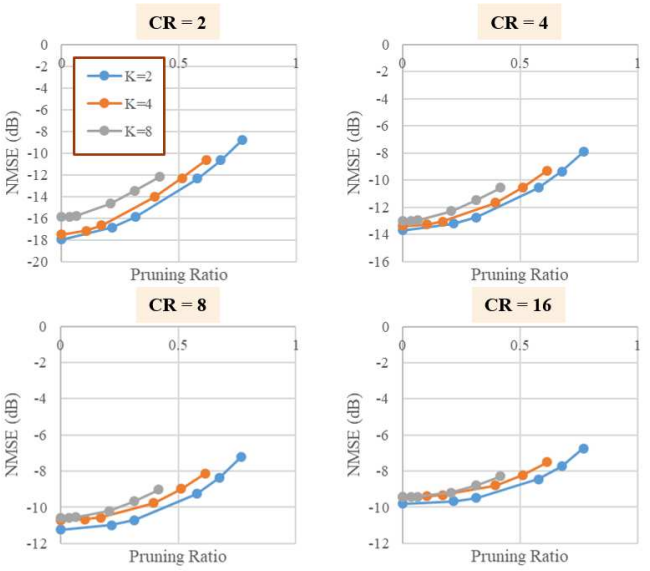

Fig. 17. NMSE performance versus pruning ratio for different selections of subarray width K in outdoor scenario

in beam domain gives rise to many near-zero subarray CSI matrices which can be skipped with little CSI distortion.

H. 2D SAB Framework

Fig. 18 shows the NMSE performance versus compression ratios for different settings of N under subarray width $K = 2$. We find that 2D SAB framework with a small N degrades less when increasing pruning ratio. However, due to the low sparsity for each subarray CSI matrix, the 2D SAB framework with a small N performs worse than that with a large N. Note that the 2D SAB framework with $N = 32$ is equivalent to the original SAB framework operating in BD domain. Performance degradation due to a small N can be attributed

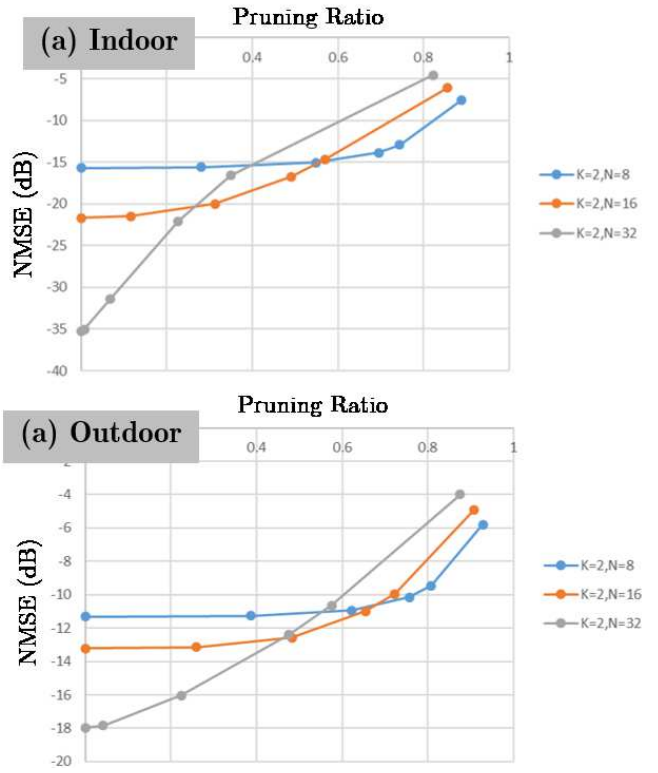

Fig. 18. NMSE performance versus pruning ratio for different N in (a) indoor and (b) outdoor scenarios.

to the aforementioned two factors: 1) incompatibility with the requirement of similar delay profile and 2) trade-off between sparsity and recovery performance. Yet, since the number of model parameters are nearly proportional to the input size squared, a smaller size of inputs in 2D SAB framework could further significantly reduce the model size of both encoder and decoder. However, the current model size using $K = 2$ is already under 1000 parameters, an extraordinarily small number for deep learning models. Further reduction of encoder model size appears to be less critical. However, since SAB framework can compress and recover in parallel, if the designer has strict computation time constraint, a 2D SAB framework may be a viable choice.

I. CSI feedback with dynamic CR

To show the benefits of the dynamic CR CSI feedback, we compare the recovery performance and codeword efficiency of the SAB CSI feedback frameworks with fixed and dynamic CRs. Since compression ratio cannot be perfectly controlled in dynamic CR CSI feedback, we define an effective CR below for fair comparison

$$\text{CR}_{\text{eff}} = \frac{1}{D}\sum_{d=1}^{D}\frac{2N_tN_a}{\sum_i^{2N_a/K}(L_{d,i}+L_{\text{CR}})}. \tag{22}$$

L_{CR} denotes the prefix codeword length to indicate adopted CR (i.e., 2 bits), which is equivalent to $2/B$ codeword elements. B denotes the quantization bits used for each codeword element. The beam-domain sparsity in outdoor channels reduces the cost of uplink feedback with minor performance loss via DCP feedback pruning. Furthermore, by properly assigning CRs to subarray CSIs, we can achieve performance improvement and codeword efficiency.

We consider four possible CRs ($= 2, 4, 16, \infty$), where $CR = \infty$ denotes the case of DCP feedback pruning. We define an outage CSI estimate when its NMSE is higher than a predetermined $T_{\text{out}} = -5$ dB, rending the CSI recovery unusable. We use an outage threshold $T_{\text{out}} = -10$ dB and $P_N = 0.01$ for training anchor points. Fig. 19 shows the average outage probability and codeword efficiency in outdoor scenario. The optimal anchor points are located at $\mathbf{a} = [1, 0.018, 0.018, 0, 0]$. This result reveals that two CRs (i.e., CR $= 2, 16$) is sufficient to maximize codeword efficiency. This further suggests that DCP feedback pruning is relatively inefficient owing to over-simplifying the low-energy subarray CSIs. Moreover, the SAB framework via dynamic CR feedback (effective CR is 5.9) can achieve comparable outage probability against a fixed low CR $= 2$ (requiring the most resources and achieving the best recovery).

TABLE III
THE RESULTING FIVE ANCHOR POINTS.

Noise Power	4 CRs	a_0	a_1	a_2	a_3	a_4
$P_N = 0.01$	[2,4,16,∞]	1	0.018	0.018	0	0
	[2,4,8,16]	1	0.018	0.018	0.018	0
$P_N = 0.0001$	[2,4,16,∞]	1	0.014	0.014	0	0
	[2,4,8,16]	1	0.014	0.014	0.014	0
$P_N = 1e-7$	[2,4,16,∞]	1	0.012	0.012	0	0
	[2,4,8,16]	1	0.012	0.012	0.012	0

J. Different Noise Powers and CR Selections

From the previous results, we know that CR $= \infty$ is unused in dynamic CR. Therefore, we attempt an additional combination of CRs $[2, 4, 8, 16]$. Table III shows the optimal points trained with different choices of $P_N = [0.01, 0.0001, 1e-7]$ and CR sets (i.e., $[2, 4, 8, 16]$ and $[2, 4, 16, \infty]$) to maximize codeword efficiency. The results show that the optimal anchor points are insensitive to P_N and continue to suggest that we only need two CRs (CR = 2 and 16) for maximizing codeword efficiency. We conclude that the most efficient strategy is to use the lowest CR to secure those subarray CSIs with high significance and keep the codeword stream as compact as possible for subarray CSIs with low energy. Also, we only need 1-bit information for acknowledging the adopted CR to gNB.

Fig. 20 shows the NMSE performance and outage probability via fixed CR and dynamic CR feedback. The anchor points shown in Table III are trained with different noise powers. The results show that dynamic CR manner not only improves the outage probability but also leads to better recovery performance than fixed CR for outdoor channels.

VII. CONCLUSIONS

This work proposes a lightweight deep-learning architecture for encoding and feeding back downlink CSI in massive MIMO wireless sytems. This new CSI feedback framework

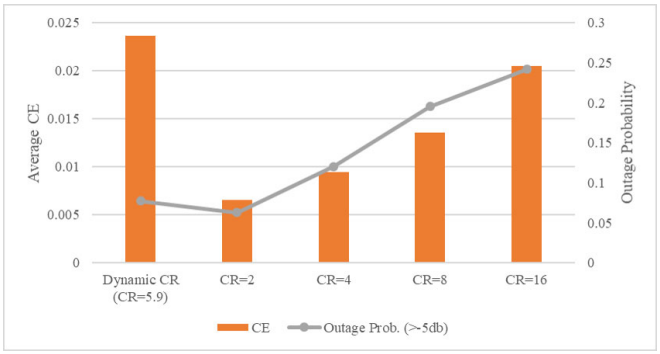

Fig. 19. Average CE and outage probability for dynamic CR and fixed CR CSI feedback framework in outdoor scenario.

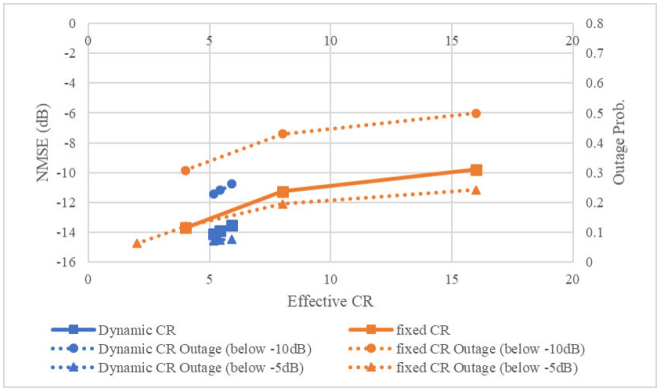

Fig. 20. NMSE performance and outage probability for dynamic CR and fixed CR CSI feedback framework in outdoor scenario.

flexibly accommodates different numbers of antenna ports in use and also requires lower computational and storage hardware at resource constrained UEs. By developing a subarray based (SAB) CSI feedback framework, a common encoder allows encoding of subarray CSI matrices separately. We further develop a dynamic encoding principle to flexibly compress subarray CSI matrices by applying dynamic compression ratios according to their significance. The new framework includes a channel-based CR selector at UE for determining CRs to achieve the maximum of codeword efficiency. Numerical results show the proposed framework generally outperforms the SOTAs, CsiNet-SM and CsiNet-PM. In summary, the proposed SAB framework heralds a simple and systematic CSI feedback manner with higher flexibility, and scalibility while requiring lower storage and computational complexity.

VIII. Acknowledgement

The authors would like to acknowledge Mason del Rosario for his useful discussions which helped the authors better understand of pilot placement and channel truncation.

References

[1] C.-H. Lin, S.-C. Lin, and E. Blasch, "TULVCAN: Terahertz Ultra-broadband Learning Vehicular Channel-aware Networking," in *IEEE INFOCOM workshop*, May 2021, pp. 1–6.

[2] X. Lin, "An overview of 5G Advanced evolution in 3GPP release 18," *arXiv preprint arXiv:2201.01358*, 2022.

[3] C. Wen, W. Shih, and S. Jin, "Deep Learning for Massive MIMO CSI Feedback," *IEEE Wirel. Commun. Lett.*, vol. 7, no. 5, pp. 748–751, 2018.

[4] T. Wang, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based csi feedback approach for time-varying massive mimo channels," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 416–419, 2019.

[5] Z. Lu, J. Wang, and J. Song, "Multi-resolution CSI Feedback with Deep Learning in Massive MIMO System," in *IEEE Intern. Conf. Communications (ICC)*, 2020, pp. 1–6.

[6] Q. Yang, M. B. Mashhadi, and D. Gündüz, "Deep Convolutional Compression For Massive MIMO CSI Feedback," in *IEEE Intern. Workshop Mach. Learning for Signal Process. (MLSP)*, 2019, pp. 1–6.

[7] S. Ji and M. Li, "CLNet: Complex Input Lightweight Neural Network Designed for Massive MIMO CSI Feedback," *IEEE Wirel. Commun. Lett.*, vol. 10, no. 10, pp. 2318–2322, 2021.

[8] Z. Liu, M. Rosario, and Z. Ding, "A Markovian Model-Driven Deep Learning Framework for Massive MIMO CSI Feedback," *IEEE Trans. Wirel. Commun.*, 2021, early access.

[9] J. Guo *et al.*, "DL-based CSI Feedback and Cooperative Recovery in Massive MIMO," *arXiv preprint arXiv:2003.03303*, 2020.

[10] Z. Liu, L. Zhang, and Z. Ding, "An Efficient Deep Learning Framework for Low Rate Massive MIMO CSI Reporting," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4761–4772, 2020.

[11] ——, "Exploiting Bi-Directional Channel Reciprocity in Deep Learning for Low Rate Massive MIMO CSI Feedback," *IEEE Wirel. Commun. Lett.*, vol. 8, no. 3, pp. 889–892, 2019.

[12] Y.-C. Lin, Z. Liu, T.-S. Lee, and Z. Ding, "Deep Learning Phase Compression for MIMO CSI Feedback by Exploiting FDD Channel Reciprocity," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2200–2204, 2021.

[13] Y. Ding and B. D. Rao, "Dictionary Learning-based Sparse Channel Representation and Estimation for FDD Massive MIMO Systems," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 8, pp. 5437–5451, 2018.

[14] X. Zhang, L. Zhong, and A. Sabharwal, "Directional Training for FDD Massive MIMO," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 8, pp. 5183–5197, 2018.

[15] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Convolutional Neural Network-Based Multiple-Rate Compressive Sensing for Massive MIMO CSI Feedback: Design, Simulation, and Analysis," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 4, pp. 2827–2840, 2020.

[16] H. Tang, J. Guo, M. Matthaiou, C.-K. Wen, and S. Jin, "Knowledge-distillation-aided Lightweight Neural Network for Massive MIMO CSI Feedback," in *IEEE Veh. Technol. Conf. (VTC2021-Fall)*, 2021, pp. 1–5.

[17] J. Guo, J. Wang, C.-K. Wen, S. Jin, and G. Y. Li, "Compression and Acceleration of Neural Networks for Communications," *IEEE Wirel. Commun.*, vol. 27, no. 4, pp. 110–117, 2020.

[18] X. Li, J. Guo, C.-K. Wen, S. Jin, and S. Han, "Multi-task Learning-based CSI Feedback Design in Multiple Scenarios ," *arXiv preprint arXiv:2204.12698*, 2022.

[19] H. Tang, J. Guo, M. Matthaiou, C.-K. Wen, and S. Jin, "Knowledge-distillation-aided Lightweight Neural Network for Massive MIMO CSI Feedback," in *IEEE Veh. Technol. Conf. (VTC2021-Fall)*, 2021, pp. 1–5.

[20] Y. Yang, F. Gao, Z. Zhong, B. Ai, and A. Alkhateeb, "Deep Transfer Learning-Based Downlink Channel Prediction for FDD Massive MIMO Systems," *IEEE Trans. Commun,*, vol. 68, no. 12, pp. 7485–7497, 2020.

[21] J. Zeng *et al.*, "Downlink CSI Feedback Algorithm With Deep Transfer Learning for FDD Massive MIMO Systems," *IEEE Trans. Cogn. Commun.*, vol. 7, no. 4, pp. 1253–1265, 2021.

[22] Y. Wang *et al.*, "Multi-Rate Compression for Downlink CSI Based on Transfer Learning in FDD Massive MIMO Systems," in *IEEE Veh. Technol. Conf. (VTC2021-Fall)*, 2021, pp. 1–5.

[23] S. Jo and J. So, "Adaptive Lightweight CNN-Based CSI Feedback for Massive MIMO Systems," *IEEE Wirel. Commun. Lett.*, vol. 10, no. 12, pp. 2776–2780, 2021.

[24] J. Guo, C.-K. Wen, and S. Jin, "CAnet: Uplink-Aided Downlink Channel Acquisition in FDD Massive MIMO Using Deep Learning," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 199–214, 2022.

[25] Y. Sun, W. Xu, L. Liang, N. Wang, G. Y. Li, and X. You, "A Lightweight Deep Network for Efficient CSI Feedback in Massive MIMO Systems," *IEEE Wirel. Commun. Lett.*, vol. 10, no. 8, pp. 1840–1844, 2021.

[26] J. Guo *et al.*, "Convolutional Neural Network-Based Multiple-Rate Compressive Sensing for Massive MIMO CSI Feedback: Design, Simulation,

and Analysis," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 4, pp. 2827–2840, 2020.
[27] R. L. Haupt, "Array Beamforming," *Timed Arrays: Wideband and Time Varying Antenna Arrays*, pp. 78–94, 2015.
[28] S. Jaeckel *et al.*, "QuaDRiGa: A 3-D Multi-Cell Channel Model with Time Evolution for Enabling Virtual Field Trials," *IEEE Trans. Antennas and Propag.*, vol. 62, no. 6, pp. 3242–3256, 2014.