Attention-Based Deep Learning for Hybrid Beamforming in OFDM Systems with Phase Noise

Faramarz Jabbarvaziri[1] and Lutz Lampe[1]

[1]Affiliation not available

January 05, 2025

Attention-Based Deep Learning for Hybrid Beamforming in OFDM Systems with Phase Noise

Faramarz Jabbarvaziri and Lutz Lampe, *Senior Member, IEEE*

***Abstract*—We introduce a deep learning-based hybrid beamforming (HBF) strategy for millimeter-wave transmission systems, specifically addressing the challenges posed by phase noise of local oscillators. Our approach utilizes a deep neural network to optimize precoding and combining matrices based on channel state information. We incorporate the symbol index through an adaptive attention mechanism and employ a self-supervised learning approach with a phase-noise-aware loss function to mitigate the effects of phase noise. While primarily focused on phase noise, our method also accommodates other practical constraints, such as limited-resolution phase shifter and imperfect channel estimation. Simulation results demonstrate that our design outperforms traditional and deep-learning based HBF methods in terms of data rate both in scenarios impacted only by phase noise and compounded distortion scenarios including low-resolution phase shifters and channel estimation errors.**

I. Introduction

Today's millimeter-wave (mmWave) multiple-input and multiple-output (MIMO) transceivers employ hybrid beamforming (HBF) architectures to strike a balance between performance, device cost, and power consumption [1]–[8]. An HBF system features an analog and digital precoder at the transmitter and an analog and digital combiner at the receiver to create optimized antenna patterns using the channel state information (CSI). HBF optimization algorithms are divided into codebook-based [1], [2], [9]–[12] and codebook-free schemes [3]–[7], [13], [14]. Codebook-based HBF, prevalent in 4G and 5G technologies, offers low computational complexity by using predefined matrices for precoding and combining. However, its performance may fall short due to limited adaptability to channel conditions. Conversely, codebook-free schemes allow for optimization of precoder and combiner pairs based on objective functions such as channel capacity, achievable rate or signal-to-interference-plus-noise ratio (SINR) in a continuous space. This paper concentrates on codebook-free HBF optimization methods, and the term HBF hereafter specifically refers to codebook-free HBF systems.

The design of HBF for mmWave systems has been investigated in numerous works, addressing problems such as reducing the computational complexity associated with beamforming optimization [15]–[19], mitigating the impact of channel estimation error [13], [20]–[22], and studying the effects of hardware-related impairments such as power amplifier nonlinearity [23], [24], low-resolution phase shifters [6], [25], and oscillator phase noise [26], [27].

F. Jabbarvaziri and L. Lampe are with the Department of Electrical and Computer Engineering, University of British Columbia, BC V6T 1Z4, Canada (e-mail: jabbarva@ece.ubc.ca, lampe@ece.ubc.ca).

Phase noise originating from local oscillators is a critical hardware impairment in mmWave systems often modeled as Wiener random process. Phase noise induces signal constellation rotation, known as common phase error (CPE), and inter-carrier interference (ICI) in orthogonal frequency-division multiplexing (OFDM) transmission. Additionally, it impacts the spatial selectivity of beamforming. The body of research on CPE and ICI mitigation techniques is substantial [26]–[30]. Nevertheless, addressing phase-noise effects on beamforming optimality remains a challenge. This issue is important due to its direct impact on beamforming array factor mismatch and gain loss, leading to a reduction in achievable rate [31]–[35]. Current model-driven [3]–[5] and data-driven [8], [36], [37] HBF optimization techniques primarily assume ideal local oscillators without phase noise. This assumption makes such HBF systems susceptible to performance degradations in real-world scenarios where phase noise exists. The detrimental impact of phase noise is compounded by the fact that antenna patterns of both the transmitter and the receiver of HBF systems are jointly optimized based on channel measurements taken prior to data transmission. Phase noise variations during data transmission alter the effective channel matrix and render the beamforming matrices mismatched. In MIMO systems, phase-noise-aware combining techniques are conventionally employed to mitigate the effects of phase noise [38]–[40]. However, we note a gap in the literature concerning the adverse effects of phase noise in two-sided beamforming systems, which necessitate jointly optimized precoders and combiners. This unresolved issue can significantly impact the achievable rate of MIMO systems.

Recently, deep learning-based methods have emerged as an effective tool for mmWave HBF optimization, offering efficient parallelized inference [41]–[43] and adaptability to specific environments via fine-tuning [44]–[47]. In this paper, we present a deep learning-based HBF optimization solution that adaptively pre-distorts the antenna radiation pattern to mitigate beamforming mismatches caused by phase noise. Building upon state-of-the-art methods [43], [48]–[51], we employ a convolutional neural network with residual connections (ResNet) at the core of our design. We demonstrate that progressively reducing the directivity of antenna pattern lobes significantly enhances robustness against beamforming mismatches. To achieve this, we propose an adaptive attention mechanism called AdaSE-ResNet—a time-adaptive version of the squeeze-and-excitation residual network (SE-ResNet) from [52]—which dynamically adjusts antenna pattern smoothing for each OFDM symbol. This is implemented through self-supervised learning using a phase-noise-aware

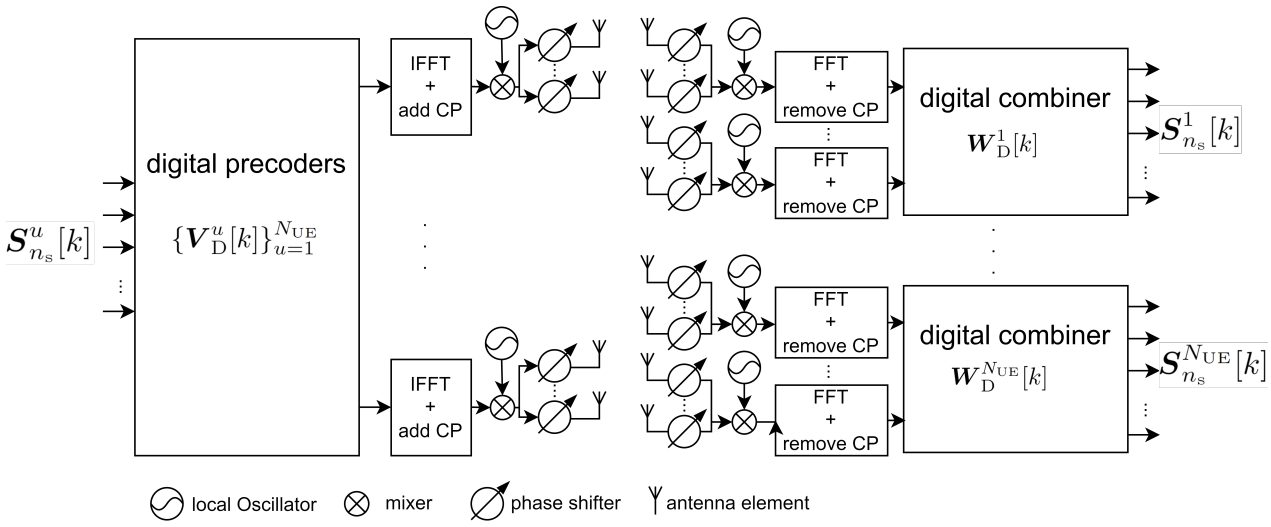

Figure 1. Block diagram of the multiuser hybrid beamforming system with one base station transmitting to N_{UE} users.

loss function that accounts for the time-dependency of the effective channel resulting from phase noise fluctuations and phase-noise-induced ICI. Furthermore, the proposed method addresses practical challenges such as low-resolution analog phase shifters and channel estimation errors, ensuring robust performance under practical constraints.

A critical factor underpinning the success of our methodology is the extensive use of transfer learning during training. We employ an approach in which initial training occurs on a series of quantized datasets, each increasing in size, followed by further training on the actual dataset. This technique significantly accelerates the training process, which otherwise appeared to be intractable, allowing us to efficiently train the proposed model.

For clarity and emphasis, we summarize our main contributions:

- We introduce AdaSE-ResNet, which recalibrates the attentional weights of the squeeze-and-excitation module for each OFDM symbol index.
- We propose a deep learning-based HBF optimization method that adaptively pre-distorts the antenna radiation pattern using the proposed AdaSE-ResNet to mitigate beamforming mismatches caused by phase noise. By progressively reducing the directivity of the antenna pattern lobes, our method enhances robustness to beamforming mismatches and mitigates signal-level loss.
- The data-driven nature of our method allows for its application across diverse phase noise models, including scenarios based on lab-measured phase noise samples. This adaptability allows for fine-tuning to specific deployment environments, enhancing its applicability in real-world scenarios.

The remainder of this paper is organized as follows. Section II presents the system model considered in this paper. In Section III, we derive the proposed HBF optimization using deep learning. Numerical results are presented and discussed in Section IV. Conclusions are provided in Section V.

II. System Model

We consider the downlink of a multi-user (MU) MIMO-OFDM HBF system. For concreteness, we assume operation in a frequency division duplex (FDD) mode, although the methodologies described are also equally applicable to time division duplex (TDD) communication. The difference lies in the CSI measurement procedure, which is essential but independent of the beamforming optimization process, provided the CSI is accessible at the base station (BS). In FDD systems, there is an inherent delay as CSI is measured in the downlink by user equipment (UE) and then transmitted to the base station. In contrast, TDD systems benefit from channel reciprocity, allowing the base station to directly measure CSI in the uplink, thereby eliminating any delay. This distinction in CSI acquisition impacts the timeliness but not the applicability of our beamforming approach across duplex modes.

We assume that the base station has $N_{\mathrm{B}}^{\mathrm{a}}$ antennas and $N_{\mathrm{B}}^{\mathrm{RF}}$ RF chains, and each user equipment has $N_{\mathrm{U}}^{\mathrm{a}}$ antennas and $N_{\mathrm{U}}^{\mathrm{RF}}$ RF chains, operating on K subcarriers. As depicted in Fig. 1, the system applies digital precoding and combining within the discrete Fourier transform (DFT) domain and applies analog precoding and combining to the up-converted signal through a network of adjustable analog phase shifters and signal adders.

Local oscillators used for up/down-conversion in the RF chain can operate in two distinct configurations: the independent local oscillator (ILO) structure, where each oscillator feeds only a single RF chain, and the common local oscillator (CLO) structure, where one oscillator feeds multiple RF chains [22]. Although the CLO architecture is appealing due to its singular phase-noise source, the ILO architecture is used in large antenna array systems where the spacing between antennas is large and distributing the signal from a single oscillator to all RF chains in such settings can be prohibitively expensive and technically challenging [39]. Thus, we consider the more general ILO configuration in this work. Having outlined the overall system model, we now discuss specific

aspects in more detail.

A. Phase Noise

Let $\phi[t]$ denote a phase noise sample at sample time t. A typical model for the temporal evolution of phase noise is the first-order recursion $\phi[t] = \phi[t-1] + \Delta\phi[t]$, where $\Delta\phi[t]$ is the phase noise innovation, which is typically assumed as an i.i.d. Gaussian process. This model is known as Wiener phase noise [53]. Other common models are often defined by their power spectral density, typically modeled using a Lorentzian spectrum that exhibits low-pass signal characteristics [54], [55]. For the derivation of our HBF design, we make no assumption about the statistics of the phase noise process other than that the phase noise innovation is small to permit an approximation which is explained later.

B. Channel Model

We adopt the geometric mmWave channel model proposed in [3] with N_{C} scattering clusters and N_{L} scatterers in each cluster. The $N_{\mathrm{U}}^{\mathrm{a}} \times N_{\mathrm{B}}^{\mathrm{a}}$ channel matrix experienced by the k^{th} OFDM subcarrier for the transmission from the BS to a UE u can be written as

$$\boldsymbol{H}^{u}[k] = \sum_{c=1}^{N_{\mathrm{C}}} \sum_{l=1}^{N_{\mathrm{L}}} \alpha_{c,l} \boldsymbol{a}_{\mathrm{r}}(\beta_{c,l}^{\mathrm{r}})(\boldsymbol{a}_{\mathrm{t}}(\beta_{c,l}^{\mathrm{t}}))^{\mathrm{H}} \mathrm{e}^{-\mathrm{j}2\pi\eta_c \frac{k}{K}} \tag{1}$$

where $\boldsymbol{a}_{\mathrm{r}}$ and $\boldsymbol{a}_{\mathrm{t}}$ are the antenna array responses of the UE receiver and the BS transmitter, respectively. For concreteness, we consider a uniform linear array configuration with antenna spacing d and wavelength λ at both the transmitter and receiver sides, and thus

$$\boldsymbol{a}_{\mathrm{r}}(\beta) = \frac{1}{\sqrt{N_{\mathrm{U}}^{\mathrm{a}}}}[1, \mathrm{e}^{\mathrm{j}2\pi d \sin(\beta)/\lambda}, \ldots, \mathrm{e}^{\mathrm{j}2\pi(N_{\mathrm{U}}^{\mathrm{a}}-1)d\sin(\beta)/\lambda}]^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{a}_{\mathrm{t}}(\beta) = \frac{1}{\sqrt{N_{\mathrm{B}}^{\mathrm{a}}}}[1, \mathrm{e}^{\mathrm{j}2\pi d \sin(\beta)/\lambda}, \ldots, \mathrm{e}^{\mathrm{j}2\pi(N_{\mathrm{B}}^{\mathrm{a}}-1)d\sin(\beta)/\lambda}]^{\mathrm{T}}. \tag{3}$$

Furthermore, $\alpha_{c,l} \sim \mathcal{CN}(0, \frac{N_{\mathrm{B}}^{\mathrm{a}} N_{\mathrm{U}}^{\mathrm{a}}}{N_{\mathrm{C}} N_{\mathrm{L}}})$ is the path gain, $\beta_{c,l}^{\mathrm{r}}$ and $\beta_{c,l}^{\mathrm{t}}$ are the angle of arrival (AoA) and angle of departure (AoD) of the l^{th} reflecting element of the c^{th} cluster, respectively, and η_c denotes the time lag of the c^{th} cluster.

In communication systems equipped with codebook-free beamforming, channel estimation is typically conducted in two stages: 1) before the start of data transmission frame for beamforming optimization and 2) during data transmission frame for data detection. Before the start of the data frame, each UE utilizes the channel state information reference signal (CSI-RS) to estimate channel conditions, and then sends this information back to the base station to facilitate beamforming optimization [12]. Within the data frame, demodulation reference signals (DMRS) are used for estimating and equalizing the channel for data detection [12]. Let N_{CSI} denote the periodicity of the CSI-RS, indicating the number of OFDM symbols within each data frame. We assume that the air interface channel, excluding the effects of phase noise, remains constant during transmission of N_{CSI} OFDM symbols. Moreover, in realistic FDD systems, the delay introduced by measuring and transmitting the downlink CSI to the BS via the uplink, compared to N_{CSI}, is negligible [12]. In TDD systems, thanks to channel reciprocity, downlink CSI is measured directly at the BS, thus eliminating any delay. Therefore, we assume that the CSI used for beamforming optimization in both duplex modes is not outdated.

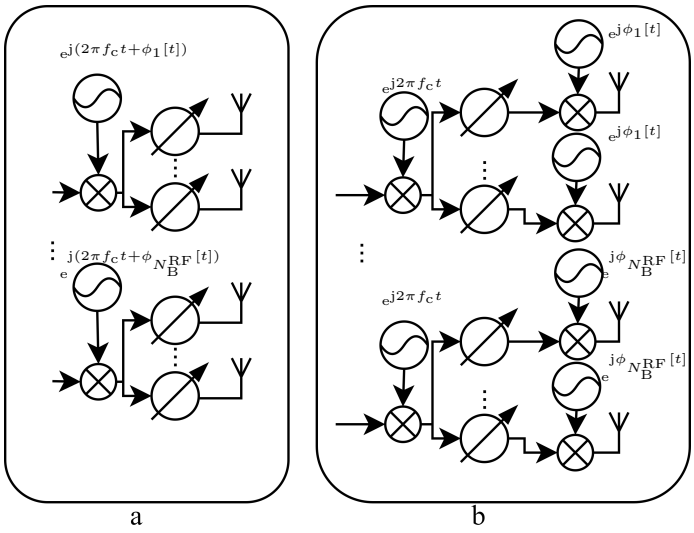

Figure 2. (a) Illustration of a partially connected analog precoder with $N_{\mathrm{B}}^{\mathrm{RF}}$ RF chains and LOs with carrier frequency f_{c} and phase noise $\phi_i[t]$. (b) Equivalent representation of the structure noting that the effect of phase noise can be considered after the precoding.

C. MIMO-OFDM HBF

As in most related literature [3]–[5], [7], we consider a partially connected analog beamforming structure as shown in Fig. 2, for its lower hardware complexity and power consumption compared to a fully-connected architecture. In MIMO-OFDM systems, the digital precoder at the transmitter and combiner at the receiver can be optimized separately for each subcarrier. However, this is not the case for the analog counterparts, as they interface with the aggregate OFDM signal in the time domain, meaning they must be designed to function effectively across all subcarriers [3], [8].

Input-Output Relation: Let $\boldsymbol{S}_{n_{\mathrm{s}}}^{u}[k] \in \mathcal{A}^{N}$ denote the signal sent from the BS to UE u over subcarrier k of the $n_{\mathrm{s}}^{\mathrm{th}}$ OFDM symbol, where $\mathcal{A}$ is the quadrature amplitude modulation (QAM) signal constellation of size $M = |\mathcal{A}|$, and N is the number of data streams. Furthermore, let $\boldsymbol{V}_{\mathrm{D}}^{u}[k] \in \mathbb{C}^{N_{\mathrm{B}}^{\mathrm{RF}} \times N}$ and $\boldsymbol{V}_{\mathrm{RF}} \in \mathcal{G}^{N_{\mathrm{B}}^{\mathrm{a}} \times N_{\mathrm{B}}^{\mathrm{RF}}}$ be the digital and analog precoders and $\boldsymbol{W}_{\mathrm{D}}^{u}[k] \in \mathbb{C}^{N_{\mathrm{U}}^{\mathrm{RF}} \times N}$ and $\boldsymbol{W}_{\mathrm{RF}}^{u} \in \mathcal{G}^{N_{\mathrm{U}}^{\mathrm{a}} \times N_{\mathrm{U}}^{\mathrm{RF}}}$ be the digital and analog combiners of the u^{th} user, respectively, where $\mathcal{G} = \{1, \mathrm{e}^{\mathrm{j}2\pi/2^{N_{\mathrm{b}}}}, \ldots, \mathrm{e}^{\mathrm{j}2\pi(2^{N_{\mathrm{b}}}-1)/2^{N_{\mathrm{b}}}}\}$ for N_{b}-bit phase shifters. Moreover, $\boldsymbol{I}_D$ represents the $D \times D$ identity matrix, $\mathbb{E}(\cdot)$ signifies the expected value operator, the function $\mathrm{BlkDiag}(\cdot)$ is used to create a block-diagonal matrix from its inputs, and $(\cdot)_K$ denotes the modulo-K operation. Then, we

can express the received vector $\boldsymbol{Y}^u_{n_\mathrm{s}}[k] \in \mathcal{C}^N$ for UE u as

$$\begin{aligned}\boldsymbol{Y}^u_{n_\mathrm{s}}[k] =&(\boldsymbol{W}^u_{\mathrm{RF}}\boldsymbol{W}^u_{\mathrm{D}}[k])^{\mathrm{H}}\sum_{q=1}^{K'}\Lambda^{n_\mathrm{s}}_u[(k-q)_K]\boldsymbol{H}^u[q]\times\\ &\sum_{m=1}^{K'}\boldsymbol{\Lambda}^{n_\mathrm{s}}_{\mathrm{B}}[(q-m)_K]\boldsymbol{V}_{\mathrm{RF}}\sum_{i=1}^{N_{\mathrm{UE}}}\boldsymbol{V}^i_{\mathrm{D}}[m]\boldsymbol{S}^i_{n_\mathrm{s}}[m]\\ &+(\boldsymbol{W}^u_{\mathrm{RF}}\boldsymbol{W}^u_{\mathrm{D}}[k])^{\mathrm{H}}\sum_{q=1}^{K'}\Lambda^{n_\mathrm{s}}_u[(k-q)_K]\boldsymbol{Z}_{n_\mathrm{s}}[q],\end{aligned} \tag{4}$$

where

$$\begin{aligned}\boldsymbol{\Lambda}^{n_\mathrm{s}}_{\mathrm{B}}[k] &\triangleq \mathrm{BlkDiag}(\theta^{n_\mathrm{s}}_1[k]\boldsymbol{I}_{\frac{N^{\mathrm{a}}_{\mathrm{B}}}{N^{\mathrm{RF}}_{\mathrm{B}}}},\ldots,\theta^{n_\mathrm{s}}_{N^{\mathrm{RF}}_{\mathrm{B}}}[k]\boldsymbol{I}_{\frac{N^{\mathrm{a}}_{\mathrm{B}}}{N^{\mathrm{RF}}_{\mathrm{B}}}}),\\ \boldsymbol{\Lambda}^{n_\mathrm{s}}_{\mathrm{U}}[k] &\triangleq \mathrm{BlkDiag}(\phi^{n_\mathrm{s}}_{u,1}[k]\boldsymbol{I}_{\frac{N^{\mathrm{a}}_{\mathrm{U}}}{N^{\mathrm{RF}}_{\mathrm{U}}}},\ldots,\phi^{n_\mathrm{s}}_{u,N^{\mathrm{RF}}_{\mathrm{U}}}[k]\boldsymbol{I}_{\frac{N^{\mathrm{a}}_{\mathrm{U}}}{N^{\mathrm{RF}}_{\mathrm{U}}}}).\end{aligned} \tag{5}$$

In (5), $\theta^{n_\mathrm{s}}_n[k]$ represents the phase noise at the k^{th} subcarrier of the $n^{\mathrm{th}}_{\mathrm{s}}$ OFDM symbol, originating from the local oscillator of the n^{th} RF chain of the transmitter. Similarly, $\phi^{n_\mathrm{s}}_{u,n}[k]$ represents the phase noise at the k^{th} subcarrier of the $n^{\mathrm{th}}_{\mathrm{s}}$ OFDM symbol, stemming from the local oscillator of the n^{th} RF chain of the receiver. In (4), $\boldsymbol{Z}_{n_\mathrm{s}}[k] \in \mathbb{C}^{N^{\mathrm{a}}_{\mathrm{U}}}$ denotes additive white Gaussian noise (AWGN) at the user side. We further assume that $\mathbb{E}\{\boldsymbol{S}^u_{n_\mathrm{s}}[k]\} = \mathbb{E}\{\boldsymbol{Z}_{n_\mathrm{s}}[k]\} = 0$, $\mathbb{E}\{\boldsymbol{S}^u_{n_\mathrm{s}}[k](\boldsymbol{S}^u_{n_\mathrm{s}}[k])^{\mathrm{H}}\} = \boldsymbol{I}_N$, and $\mathbb{E}\{\boldsymbol{Z}_{n_\mathrm{s}}[k](\boldsymbol{Z}_{n_\mathrm{s}}[k])^{\mathrm{H}}\} = \sigma^2\boldsymbol{I}_N$, where σ^2 is known to the BS. Furthermore, we applied the following approximations regarding phase noise in (4). Firstly, in line with established practices in the literature [26], [53], [56], we substitute the actual phase noise observed during OFDM signal transmission with its cyclically extended counterpart. Secondly, acknowledging the predominantly lowpass characteristic of the phase noise process [53], [57], we truncate the summations in (4) to include only $K' \ll K$ terms.

We rewrite (4) more compactly as

$$\boldsymbol{Y}^u_{n_\mathrm{s}}[k] = \boldsymbol{G}^u_{n_\mathrm{s}}[k]\boldsymbol{S}^u_{n_\mathrm{s}}[k] + \boldsymbol{Q}^u_{n_\mathrm{s}}[k], \tag{6}$$

where

$$\boldsymbol{G}^u_{n_\mathrm{s}}[k] =(\boldsymbol{W}^u_{\mathrm{RF}}\boldsymbol{W}^u_{\mathrm{D}}[k])^{\mathrm{H}}\tilde{\boldsymbol{H}}^u_{n_\mathrm{s}}[k]\boldsymbol{V}_{\mathrm{RF}}\boldsymbol{V}^u_{\mathrm{D}}[k], \tag{7}$$

represents the transmission path gain of the user's data signal,

$$\tilde{\boldsymbol{H}}^u_{n_\mathrm{s}}[k] = \sum_q \Lambda^{n_\mathrm{s}}_u[(k-q)_K]\boldsymbol{H}^u[q]\boldsymbol{\Lambda}^{n_\mathrm{s}}_{\mathrm{B}}[(q-k)_K]. \tag{8}$$

is the effective channel including the effect of phase noise, and $\boldsymbol{Q}^u_{n_\mathrm{s}}[k]$ is the total interference plus noise consisting of inter-user interference, ICI due to phase noise, and AWGN.

Demodulation: Considering the model in (6) and (7), channel estimation via DMRS will produce an estimate of $\tilde{\boldsymbol{H}}^u_{n_\mathrm{s}}[k]$. For the purpose of the problem formulation in Section III, we assume that a perfect estimate is available for demodulation. Imperfect channel estimation will be considered for numerical results in Section IV.

During data transmission, user u processes $\boldsymbol{Y}^u_{n_\mathrm{s}}[k]$ to compute log-likelihood ratios (LLRs) for the bits represented by $\boldsymbol{S}^u_{n_\mathrm{s}}$. For this, we adopt a pragmatic approach as in [50] and separate the data streams using a linear minimum-mean squared error (LMMSE) equalizer. Following [58], during equalization, we assume that the interference plus noise $\boldsymbol{Q}^u_{n_\mathrm{s}}[k]$ has a known covariance matrix $\boldsymbol{C}$. Then, the LMMSE equalized signal for the u^{th} user is obtained as [59]

$$\begin{aligned}&\boldsymbol{R}^u_{n_\mathrm{s}}[k] =\\ &\mathrm{diag}\Big((\boldsymbol{G}^u_{n_\mathrm{s}}[k])^{\mathrm{H}}(\boldsymbol{G}^u_{n_\mathrm{s}}[k](\boldsymbol{G}^u_{n_\mathrm{s}}[k])^{\mathrm{H}}+\boldsymbol{C})^{-1}\boldsymbol{G}^u_{n_\mathrm{s}}[k]\Big)^{-1}\\ &\times(\boldsymbol{G}^u_{n_\mathrm{s}}[k])^{\mathrm{H}}(\boldsymbol{G}^u_{n_\mathrm{s}}[k](\boldsymbol{G}^u_{n_\mathrm{s}}[k])^{\mathrm{H}}+\boldsymbol{C})^{-1}\boldsymbol{Y}^u_{n_\mathrm{s}}[k],\end{aligned} \tag{9}$$

where $\mathrm{diag}(\cdot)$ returns a square matrix in which the diagonal elements of the input matrix are placed on the main diagonal, and all off-diagonal elements are zero. The LLR for the m^{th} bit associated with the n^{th} data stream in $\boldsymbol{S}^u_{n_\mathrm{s}}[k]$ is calculated as

$$L^{u,n,m}_{n_\mathrm{s}}[k] = \log\left(\frac{\sum_{x\in\mathcal{A}_{m,1}}\mathrm{e}^{-|r^{u,n}_{n_\mathrm{s}}[k]-x|^2/\sigma^2_{u,n_\mathrm{s}}[k]}}{\sum_{x\in\mathcal{A}_{m,0}}\mathrm{e}^{-|r^{u,n}_{n_\mathrm{s}}[k]-x|^2/\sigma^2_{u,n_\mathrm{s}}[k]}}\right), \tag{10}$$

where $\mathcal{A}_{m,b}$ represents the subset of constellation points with the m^{th} bit label equal to b and $r^{u,n}_{n_\mathrm{s}}[k]$ denotes the n^{th} element of $\boldsymbol{R}^u_{n_\mathrm{s}}[k]$. Furthermore, $\sigma^2_{u,n_\mathrm{s}}[k]$ represents the post-equalization noise power of the LMMSE equalizer for the $n^{\mathrm{th}}_{\mathrm{s}}$ symbol on the k^{th} subcarrier at the u^{th} user. This is given by the diagonal elements of the covariance matrix of the residual interference and noise [59, Lemma B.19]

$$\begin{aligned}&\boldsymbol{C}^{\mathrm{r}}_{u,n_\mathrm{s}}[k] =\\ &\Big((\boldsymbol{G}^u_{n_\mathrm{s}}[k])^{\mathrm{H}}(\boldsymbol{G}^u_{n_\mathrm{s}}[k](\boldsymbol{G}^u_{n_\mathrm{s}}[k])^{\mathrm{H}}+\boldsymbol{C})^{-1}\boldsymbol{G}^u_{n_\mathrm{s}}[k]\Big)^{-1}-\boldsymbol{I}_N.\end{aligned} \tag{11}$$

Achievable Information Rate (AIR): The LLRs obtained from the demodulation in (10) can be used to define a posterior distribution for the associated bit as

$$P(b^{u,n,m}_{n_\mathrm{s}}[k]|r^{u,n}_{n_\mathrm{s}}[k]) = \frac{1}{1+\mathrm{e}^{(-1)^{b^{u,n,m}_{n_\mathrm{s}}[k]}L^{u,n,m}_{n_\mathrm{s}}[k]}}. \tag{12}$$

Following [60], we can compute the empirical binary cross entropy (BCE)

$$\begin{aligned}&E[k] = -\frac{1}{N_{\mathrm{UE}}N_{\mathrm{CSI}}N}\times\\ &\sum_{u=1}^{N_{\mathrm{UE}}}\sum_{m=1}^{\log_2 M}\sum_{n_\mathrm{s}=1}^{N_{\mathrm{CSI}}}\sum_{n=1}^{N}\log_2\left(P(b^{u,n,m}_{n_\mathrm{s}}[k]|r^{u,n}_{n_\mathrm{s}}[k])\right)\end{aligned} \tag{13}$$

for subcarrier k and average it over several frames with independent channel realizations to obtain $\bar{E}[k]$. The empirical approximation of the AIR is

$$R[k] = \log_2(M) - \bar{E}[k]. \tag{14}$$

III. Proposed HBF Optimization

In this section, we introduce our proposed deep-learning-based method for hybrid beamforming optimization. The digital precoding and combining matrices are optimized at the BS independently for each subcarrier. This methodology is consistent with state-of-the-art practices [3]–[5], [8], which leverage the orthogonality of signals across subcarriers. In our implementation, this strategy allows the deep neural network (DNN) to manage a significantly reduced input and output size, necessitating fewer parameters and thus enabling more efficient training.

In practical systems, phase noise presents two significant challenges for HBF optimization: 1) the loss of signal orthogonality across subcarriers due to ICI, and 2) the progressive misalignment of beamforming throughout the frame, caused by phase noise-induced variations. Therefore, HBF designs that are unaware of phase noise are necessarily suboptimal.

To address the first challenge, we propose a HBF optimization method that processes a neighborhood of subcarriers simultaneously to mitigate the effects of ICI. This approach is effective because the interference is generally limited to a small group of subcarriers around the subcarrier of interest k, as a result of the low-pass nature of phase noise innovations [53], [57]. Therefore, we consider the K' adjacent subcarriers $\mathcal{K}(k) = \{k - \frac{K'}{2}, \ldots, k, \ldots, k + \frac{K'}{2}\}$ to generate effective beamforming matrices for subcarrier k, where K' is an even number.

Before addressing the second challenge, we remark that beamforming cannot benefit from any post-signal-detection information. Hence, phase noise estimation and tracking implemented at the data detection stage can only be used for equalization and improving the ICI and cannot reduce the phase-noise-induced beamforming mismatch. We propose a variant of the attention mechanism [61] which we refer to as adaptive attention to enhance robustness against unknown phase noise variations. The primary function of this module is to progressively smooth the antenna patterns as the OFDM symbol index n_{s} within each data frame increases, thereby reducing sensitivity to phase noise. To ensure the smoothing level is aligned with the phase-noise-induced mismatch at each n_{s}, we integrate n_{s} as an auxiliary input to the proposed DNN alongside the CSI matrix. Additionally, we select a loss function that specifically emphasizes the dependency on n_{s}. The components of the proposed DNN will be detailed further in the subsequent sections.

A. Deep Neural Network Architecture

In this section, we discuss the structure of our proposed DNN and explain the specific design choices that help the algorithm tackle phase-noise-induced beamforming mismatch and ICI.

To address the design requirements for subcarrier-specific and aggregate OFDM signal processing in HBF systems for digital and analog parts, respectively, we choose a neural network architecture comprising two specialized branches. The first branch is dedicated to processing subcarrier-specific channel information, i.e., $\tilde{\boldsymbol{H}}_0^u[k]$, yielding digital precoders and combiners. The second branch processes matrix product $(\tilde{\boldsymbol{H}}_0^u[k])^{\mathrm{H}}\tilde{\boldsymbol{H}}_0^u[k]$, aggregated over all subcarriers to produce analog precoder and combiners. This methodological choice to utilize aggregated channel squared matrices to design the analog beamforming allows the neural network to incorporate all paths in the analog beamforming, regardless of the phase differences among these paths across subcarriers. This approach ensures that the analog beamforming process is comprehensive, capturing essential channel components to optimize performance.

Denoting by F_η the input-output function of a DNN characterized by the parameter set η, we can express the operation of the proposed DNN as

$$\left\{\begin{bmatrix}\boldsymbol{V}_{\mathrm{D}}^1[k]\\ \boldsymbol{V}_{\mathrm{D}}^2[k]\\ \vdots\\ \boldsymbol{V}_{\mathrm{D}}^{N_{\mathrm{UE}}}[k]\end{bmatrix}, \boldsymbol{V}_{\mathrm{RF}}, \begin{bmatrix}\boldsymbol{W}_{\mathrm{D}}^1[k]\\ \boldsymbol{W}_{\mathrm{D}}^2[k]\\ \vdots\\ \boldsymbol{W}_{\mathrm{D}}^{N_{\mathrm{UE}}}[k]\end{bmatrix}, \begin{bmatrix}\boldsymbol{W}_{\mathrm{RF}}^1\\ \boldsymbol{W}_{\mathrm{RF}}^2\\ \vdots\\ \boldsymbol{W}_{\mathrm{RF}}^{N_{\mathrm{UE}}}\end{bmatrix}\right\} = F_\eta(\tilde{\mathcal{H}}_0[k], \tilde{\boldsymbol{P}}_0, n_{\mathrm{s}}), \tag{15}$$

where,

$$\tilde{\mathcal{H}}_0[k] = \left\{\begin{bmatrix}\tilde{\boldsymbol{H}}_0^1[\kappa]\\ \tilde{\boldsymbol{H}}_0^2[\kappa]\\ \vdots\\ \tilde{\boldsymbol{H}}_0^{N_{\mathrm{UE}}}[\kappa]\end{bmatrix} \mid \kappa \in \mathcal{K}(k)\right\}, \tag{16}$$

is the phase-noise-affected CSI at $n_{\mathrm{s}} = 0$ for the set of adjacent subcarriers $\mathcal{K}(k)$ around k, and

$$\tilde{\boldsymbol{P}}_0 = \sum_k \begin{bmatrix}(\tilde{\boldsymbol{H}}_0^0[k])^{\mathrm{H}}\tilde{\boldsymbol{H}}_0^0[k]\\ (\tilde{\boldsymbol{H}}_0^1[k])^{\mathrm{H}}\tilde{\boldsymbol{H}}_0^1[k]\\ \vdots\\ (\tilde{\boldsymbol{H}}_0^{N_{\mathrm{UE}}}[k])^{\mathrm{H}}\tilde{\boldsymbol{H}}_0^{N_{\mathrm{UE}}}[k]\end{bmatrix}, \tag{17}$$

is the phase-noise-affected channel matrix products stacked in the user domain at $n_{\mathrm{s}} = 0$.

To implement (15), we introduce a DNN consisting of Adaptive SE-ResNet (AdaSE-ResNet) modules. These modules are organized into a tree-like structure as illustrated in Fig. 3. In this figure, each dashed orange box represents a distinct sub-network that outputs the precoding and combining matrices for a user. Furthermore, drawing on insights from [62], we propose parameter binding for these sub-networks to prevent the overall neural network's parameter count from increasing with the number of users. As long as the users share the same antenna array structure and each user-specific sub-network possesses adequate learning capacity, the aforementioned parameter binding process does not adversely impact performance. It is important to note that each sub-network has a user-specific channel input, thereby generating a user-specific output.

Table III in the appendix provides detailed information about the modules shown in Fig. 3, which are used for the numerical results in Section IV. The architecture choices summarized in this table were carefully determined through extensive testing and iterative refinements to achieve strong performance while maintaining a compact DNN structure. In the following, we discuss the specifics for the AdaSE-ResNet blocks as well as the chosen activitation functions.

1) AdaSE-ResNet Block: To enhance robustness against phase-noise-induced beamforming mismatch, we integrate the OFDM symbol index n_{s} along with the CSI as an additional input. This index serves as auxiliary data that can be linked to the phase-noise effects on beamforming. For the joint processing of random CSI and deterministic symbol-index inputs, we draw inspiration from the methodology described in [63], where the authors employ a soft attention mechanism to

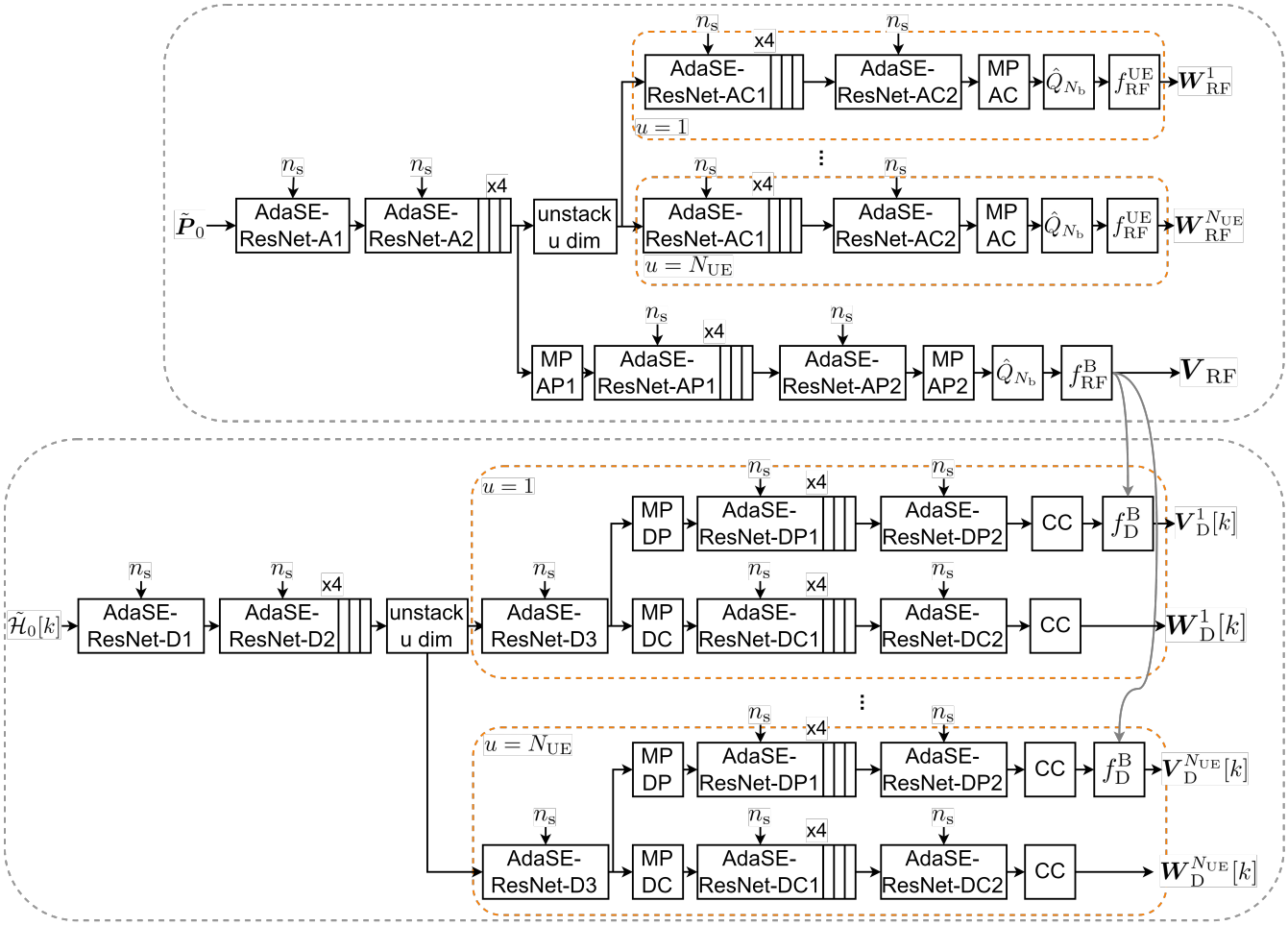

Figure 3. Architecture of the proposed DNN. Table III in the appendix provides details and parameter choices for all the depicted modules. Each orange box encloses a sub-network that produces the user-specific matrices $\boldsymbol{V}_{\mathrm{D}}^{u}[k]$, $\boldsymbol{W}_{\mathrm{D}}^{u}[k]$, and $\boldsymbol{W}_{\mathrm{RF}}^{u}$. The notation "x4" next to AdaSE-ResNet indicates a series of four consecutive AdaSE-ResNet blocks with identical configurations.

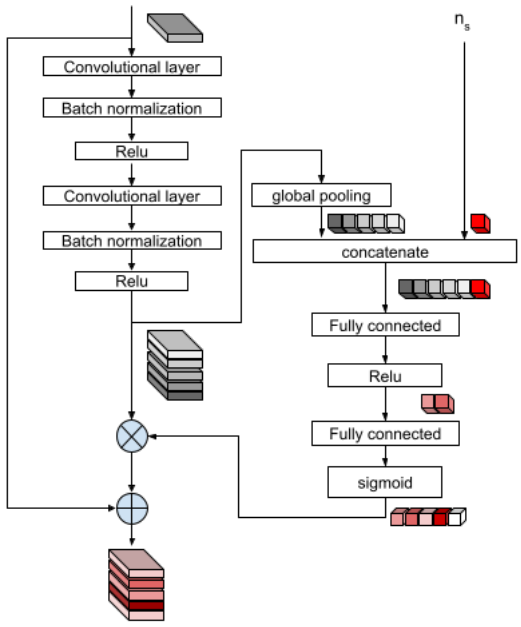

Figure 4. Structure of the proposed AdaSE-ResNet used in the neural network architecture shown in Figure 3.

use the signal-to-noise ratio (SNR) as auxiliary information for adjusting the compression ratio in image data source coding. Expanding on this concept and the principles of squeeze-and-excitation residual network, we introduce the adaptive version of SE-ResNet, named AdaSE-ResNet, shown in Fig. 4.

The AdaSE-ResNet architecture features two complementary paths to optimize HBF under the effects of phase noise. The main path processes the CSI instance to generate beamforming matrices, while the auxiliary path employs an attention mechanism to dynamically adapt to the symbol index n_{s} throughout the transmission frame. This dual-path design allows for simultaneous processing of CSI and time-adaptive beamforming calibration.

The attention mechanism functions as follows: the right-hand-side branch of Fig. 4 generates an output through the sigmoid function, referred to as the excitation vector (depicted with rearranged colors). This excitation vector represents symbol-index-specific attentional weights that tailor the ResNet's output to each symbol index n_{s}. By calibrating feature maps for each n_{s} and utilizing a frame-wise loss function, as detailed in Section III-B, these attentional weights aim to maximize the total AIR across the entire frame. Notably, these weights are not pre-defined, but they are learned

during backpropagation along with the other parameters of the DNN. This adaptive process equips the module with sufficient flexibility to effectively anchor antenna patterns to the symbol index.

By integrating both the CSI context $\tilde{\mathcal{H}}_0[k]$ and the symbol index n_s, the AdaSE-ResNet learns a unique, time-adaptive feature-map calibration strategy for each symbol index. This capability allows the network to dynamically predistort the antenna radiation patterns, thereby ensuring optimal beamforming performance and robustness against mismatches caused by phase noise throughout the transmission frame.

2) Activation Functions: To optimize the performance of the proposed DNN under the constraints of limited-resolution phase shifters for $\boldsymbol{V}_\mathrm{RF}$ and $\boldsymbol{W}_\mathrm{RF}^u$, we employ the tanh-approximated quantizer activation function [64]

$$\begin{aligned}\hat{Q}_{N_\mathrm{b}}(x) =& \frac{\pi}{2^{N_\mathrm{b}-1}}\left[\left(\left\lfloor \frac{x}{\frac{2\pi}{2^{N_\mathrm{b}}}}\right\rfloor - 0.5\right)\right.\\ &\left. + 0.5\tanh\left(\alpha\left(\frac{x_i}{\frac{2\pi}{2^{N_\mathrm{b}}}} - \left\lfloor \frac{x}{\frac{2\pi}{2^{N_\mathrm{b}}}}\right\rfloor\right) + 1\right)\right]\end{aligned} \tag{18}$$

during training. In (18), α is the steepness factor that modulates the proximity of the function to a step function, thereby controlling the accuracy of the approximation. This method ensures that the network effectively learns to manage resolution limitations while maintaining differentiability. In the inference phase, the phase shifter angles are quantized to the nearest elements in the set $\mathcal{G}$.

Furthermore, in order to meet the unit amplitude constraint of the elements of the analog precoding and combining matrices, in the last layer of the $\boldsymbol{V}_\mathrm{RF}$ and $\boldsymbol{W}_\mathrm{RF}^u$ branches, we apply the activation functions

$$f_\mathrm{RF}^\mathrm{B}(x) = \mathrm{BlkDiag}\left(\mathrm{e}^{\mathrm{j}x_1}\mathbf{1}_{\frac{N_\mathrm{B}^\mathrm{a}}{N_\mathrm{B}^\mathrm{RF}}}, \mathrm{e}^{\mathrm{j}x_2}\mathbf{1}_{\frac{N_\mathrm{B}^\mathrm{a}}{N_\mathrm{B}^\mathrm{RF}}}, \ldots, \mathrm{e}^{\mathrm{j}x_{N_\mathrm{B}^\mathrm{RF}}}\mathbf{1}_{\frac{N_\mathrm{B}^\mathrm{a}}{N_\mathrm{B}^\mathrm{RF}}}\right), \tag{19}$$

and

$$f_\mathrm{RF}^\mathrm{UE}(x) = \mathrm{BlkDiag}\left(\mathrm{e}^{\mathrm{j}x_1}\mathbf{1}_{\frac{N_\mathrm{U}^\mathrm{a}}{N_\mathrm{U}^\mathrm{RF}}}, \mathrm{e}^{\mathrm{j}x_2}\mathbf{1}_{\frac{N_\mathrm{U}^\mathrm{a}}{N_\mathrm{U}^\mathrm{RF}}}, \ldots, \mathrm{e}^{\mathrm{j}x_{N_\mathrm{U}^\mathrm{RF}}}\mathbf{1}_{\frac{N_\mathrm{U}^\mathrm{a}}{N_\mathrm{U}^\mathrm{RF}}}\right), \tag{20}$$

where $\mathbf{1}_D$ is the $D \times 1$ all-one vector. In (19) and (20), the all-one vectors scale the DNN outputs to match the dimensional requirements and connectivity constraints of the analog beamforming hardware in a partially connected architecture, as detailed in [3, Eq. (4)]. Additionally, to normalize the per-subcarrier transmit power, we apply

$$f_\mathrm{D}^\mathrm{B}(\boldsymbol{V}_\mathrm{D}[k]) = \boldsymbol{V}_\mathrm{D}[k]\sqrt{\frac{P}{\mathrm{Tr}(\boldsymbol{V}_\mathrm{RF}\boldsymbol{V}_\mathrm{D}[k]\boldsymbol{V}_\mathrm{D}^\mathrm{H}[k]\boldsymbol{V}_\mathrm{RF}^\mathrm{H})}}, \tag{21}$$

where P is the allowed transmit power per subcarrier.

B. Training Procedure

In line with [8], [60], we adopt the empirical BCE from (13) as our loss function. This choice is motivated by its dependence on the ICI caused by phase noise, which guides the gradient signals during backpropagation to mitigate ICI, as well as the computational simplicity of the BCE when using LLRs. Hence, the training process aims to solve the following optimization problem with respect to the DNN parameters η:

$$\begin{aligned}&\min_{\eta} E[k]\\ &\text{s.t.}\quad \text{equation (15)}.\end{aligned} \tag{22}$$

This training approach is independent of the specific subcarrier selected and can be conducted solely for $k = 0$. Once trained, the proposed DNN can be applied across all subcarriers to generate the corresponding HBF during the inference phase.

The distribution of the network input is determined through the distributions of the random variables contributing to the effective channel, i.e., AoAs, AoDs, path gains, users-BS distances, and phase noises. Our experiments indicate that training the DNN directly on the distribution corresponding to the full spectrum of these variables is prohibitively time-intensive. To overcome this limitation, we employ a variant of transfer learning that utilizes a sequence of quantized proxy datasets. These datasets are constructed through the Cartesian product of increasingly finer quantized AoAs, AoDs, channel gains, and user distances. In our method, we begin training with the coarsest dataset and progressively transfer the learned weights to training on finer datasets. This incremental process enables gradual adaptation to the unquantized dataset. Although this approach reduces overall training time, the use of simplified proxy datasets may introduce artifacts that adversely affect final performance, making its success dependent on carefully selecting the quantization levels.

C. Scalability

For practical applicability, the scalability with respect to the number of subcarriers, antennas, and users is critical.

1) Scalability with the number of subcarriers: As mentioned in the previous section, during the training phase, the proposed model learns from a single subcarrier and generalizes to the remaining subcarriers due to assumed identical distribution across subcarriers. Consequently, the number of parameters of F_η and the computational cost of training are independent of the number of subcarriers.

During the inference phase, F_η is applied separately to the CSIs of overlapping sets $\mathcal{K}(k)$ of fixed size K' to generate the beamforming matrices for all K subcarriers. As a result, the computational load during the inference phase scales linearly with K.

2) Scalability with the number of antennas: To ensure a sufficiently large receptive field for effectively processing the CSI matrix, the depth of the overall neural network must scale with the number of antennas. As discussed in [65], the receptive field of a convolutional neural network at each input dimension is linearly proportional to its depth. The proposed adaptive attention mechanism only recalibrates the feature maps, thus it does not alter the theoretical receptive field of the underlying convolutional layers. Consequently, both the depth of the neural network and the computational costs during the inference phase scale linearly with the number of transmit or receive antennas. Additionally, the computational cost of the training phase increases with the number of antennas at least

linearly. Precisely quantifying this relationship is challenging, as the number of training epochs is generally determined through empirical methods. We employ early stopping in our training methodology, a widely recognized pragmatic approach, which results in a distinctly linear relationship between computational cost of the training phase and the number of antennas.

3) Scalability with the number of users: As mentioned in Section III-A1, we use parameter binding for the user-specific sub-networks. Therefore, the number of parameters of the proposed DNN does not increase with the number of users. However, the computational cost of the inference phases scales linearly with the number of users, as each user-specific sub-network is processed individually. Similar to the case of scaling with the number of antennas, the computational cost of the training phase increases at least linearly with the number of users. This increase is strictly linear when employing early-stopping during the training process.

IV. Numerical Results

In this section, we present and discuss simulation results to assess the performance of the proposed data-driven beamformer design. We also aim to provide some insight into the type of HBF solutions the trained DNN generates in the presence of phase noise.

A. Simulation Setting

For our simulations, we use the following specifications.

1) System and Channel: We consider an MU-MIMO scenario where the transmitter is equipped with 16 antennas and 4 RF chains. The OFDM, propagation channel, and phase noise parameters are specified in Table I. The BS serves 4 users, each equipped with 4 antennas and 1 RF chain. The users are located within a disc around the BS with an inner radius of 20 m and an outer radius of 400 m. In each simulation iteration, users are redistributed uniformly at random within the coverage area. The path loss for a user at a distance d from the base station is modeled as $128.1+37.6\log_{10}(d/1\text{ km})$. The power spectral densities of the noise and transmitted signal are set to be -174 dBm/Hz and -55 dBm/Hz, respectively. Furthermore, following the approach in [3], we generate the AoAs $\beta^r_{c,l}$ and AoDs $\beta^t_{c,l}$ in (1) using a Laplacian distribution, with the cluster mean value uniformly between 0 and 2π and an angular spread of 10 degrees within each cluster. The phase noises at LOs are generated as independent Wiener processes with a phase noise level L in dBc/Hz at $f_0 = 100$ kHz (see Table I).

2) DNN: The modules used in the DNN architecture are described in detail in Table III in the appendix. We apply 10 AdaSE-ResNet layers for the analog beamforming branch and 11 for the digital beamforming branch, resulting in 20 and 22 convolutional residual layers, respectively, each with 64 feature maps. This setup ensures deep feature extraction, which is essential for beamforming optimization. Additionally, the use of 3x3 kernels strikes a balance between computational efficiency, fine-grained spatial feature extraction, and the number of trainable parameters. The proposed neural network has approximately 4.3 million parameters. For training, we employ the adaptive moment estimation (ADAM) optimization algorithm [68], complemented by a reduce-on-plateau learning rate scheduler [69]. We conduct a hyperparameter search to determine the optimal learning rate tailored to our specific problem. Consequently, the initial learning rate is set to 10^{-4}, the minimum possible rate is set to 10^{-7}, with a decay rate of 0.5, and patience of one epoch. For the training, we employ $32,000$ samples and an equivalent number of phase noise vectors where each phase noise vector is of size $N_{\text{CSI}} \times K$ for each RF chain. For both the initial and subsequent training we use a small batch size of 8 to introduce gradient noise for improved generalization. To monitor generalization performance and prevent overfitting, a validation set of size 1024 samples is utilized. Moreover, we implement early stopping with a specified patience to avert over-training and enhance the computational efficiency of the training process. Training continues until the maximum number of 20 epochs is reached.

Table I
Simulation Parameters

Property	Variable	Value
Signal constellation	$\mathcal{A}$	16QAM
Subcarrier spacing	Δf	15 kHz
FFT size	K	1024
Sampling time	T_{s}	65.104 ns
Symbols per slot	N_{slot}	14
CSI-RS symbol periodicity	N_{CSI}	$20N_{\text{slot}}$ [66]
Frequency offset from the carrier	f_0	100 kHz
Phase noise level	L	$\{-95, -100, -105\}$ dBc/Hz
Phase noise innovation variance	-	$4\pi^2 f_0^2 10^{L/10} T_{\text{s}}$ [67]
Size of influential subcarriers set	K'	4
Number of scattering clusters	N_{C}	5
Number of scatterers	N_{L}	10
Antenna spacing	-	$\lambda/2$
Phase shift in c^{th} cluster	η_c	$c-1$ [3]
Steepness factor of $\tanh$ in (18)	α	10

B. Performance Evaluation

To highlight and assess the performance of our proposed method, we compare it against three state-of-the-art HBF techniques. The first two methods, higher-order singular value decomposition (HOSVD) [4] and constrained tensor decomposition-based hybrid beamforming (CTDH) [5], use an optimization-based approach. HOSVD employs a low-complexity disjoint optimization where the analog precoder and combiners are designed using the channel's SVD, ignoring multi-user interference effects. The digital precoders and combiners are then optimized, with fixed analog components, to maximize the SINR. CTDH involves a two-stage optimization process. Initially, a low-rank constrained Tucker2 decomposition is used to optimize the analog precoder and combiners. This is followed by deriving the digital precoder and combiner from the CSI's SVD in the subsequent stage. The third method, called two-timescale (TTS) end-to-end learning [8], is a deep learning-based approach developed for a single-user scenario that jointly optimizes beamforming, pilots, and CSI compression. At each OFDM symbol, the TTS method recalculates the digital precoding at the BS and the digital combining at the UE using DNNs. This approach assumes (i) the availability of an

auxiliary uplink channel to communicate the low-dimensional "equivalent CSI" matrix $\boldsymbol{W}_{\mathrm{RF}}^{\mathrm{H}}\tilde{\boldsymbol{H}}_{n_\mathrm{s}}^{u}[k]\boldsymbol{V}_{\mathrm{RF}}$ at every OFDM symbol, and (ii) that the UE has sufficient processing power to run a DNN during the inference phase. For the purpose of comparison with the proposed method, we make the idealized assumption that perfect equivalent CSI is available for both precoder and combiner recalculation at every OFDM symbol. Therefore, the results represent an upper bound on TTS's performance. Since the TTS method from [8] is applied to single-user MIMO (SU-MIMO) scenarios, we do not include it in the set of comparisons for *MU*-MIMO transmission discussed in Section IV-B2 below.

1) Performance Illustration: First, we illustrate the impact of phase-noise-induced beamforming mismatch on the received constellations. Fig. 5 shows scatter plots of the equalizer outputs at the UE as a function of the OFDM symbol index n_s. We recall that the beamforming design for the proposed method, HOSVD, CTDH, and the analog part of the TTS method rely on the CSI at index $n_\mathrm{s} = 0$, while the digital part of the TTS method is based on an updated equivalent channel. The phase noise level is set to $L = -100$ dBc/Hz, and the same level is used for training and testing of the learned methods. We observe that, as n_s increases, the point clouds for the proposed method and the TTS method remain more compact around the true locations for transmitted signals compared to those for HOSVD and CTDH. This is further confirmed by the mean-squared error (MSE) values reported for each scatter plot. We attribute the improved robustness of the proposed learned solution to the beam adaptation as a function of n_s, facilitated by the AdaSE-ResNet block. Additionally, the better performance of the TTS method can be attributed to its partial refreshing mechanism, achieved through the recalculation of digital beamformers using a low-dimensional equivalent channel.

To further highlight the antenna radiation pattern adaptation of the proposed method, Fig. 6 shows the beamforming pattern at the BS as a function of n_s for the case that the BS serves one UE and for four different channel realizations. It demonstrates how the AdaSE-ResNet dynamically adjusts the pattern to alleviate beamforming mismatch. This adjustment process is governed by the n_s-dependent attention mechanism within the AdaSE-ResNet block. The progression observed in the figure suggests a trend towards a smoother beamforming pattern over time. This adaptation helps to reduce the system's sensitivity to phase-noise-induced variations in the channel, which causes fluctuations in the angular distribution of paths throughout the frame. Accordingly, we expect that the degree of smoothing increases with the phase noise level.

2) Performance Results: We next report MU-MIMO simulation results for the AIR from (13) averaged for the four UEs, comparing the proposed method and the optimization-based benchmark methods HOSVD and CTDH. As in the previous figures, we plot AIR as a function of n_s, to delineate the performance degradation due to beamforming mismatch.

Figure 7 shows the average AIRs for the three methods and the three different phase-noise levels $L = -95\mathrm{dBc/Hz}$, $L = -100\mathrm{dBc/Hz}$, and $L = -105\mathrm{dBc/Hz}$. We assume a perfect CSI for beamformer optimization and equalization, and

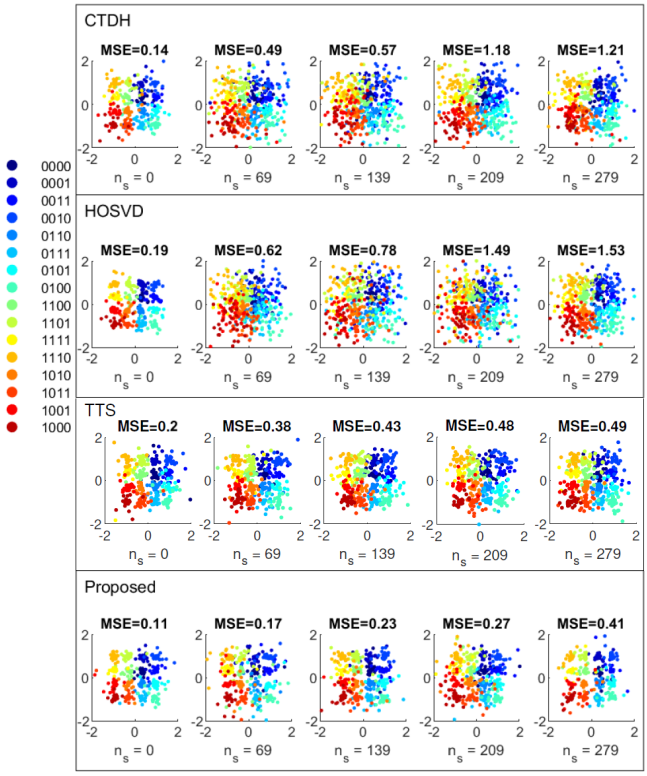

Figure 5. Scatter plot of the equalizer outputs at the UEs as a function of OFDM symbol index n_s. Comparison of HBF with the proposed DNN and HBF with CTDH, HOSVD, and TTS, respectively.

infinite-resolution phase shifters. Furthermore, the proposed DNN is specifically trained for each phase noise scenario to achieve optimal performance. We observe that the DNN solution consistently outperforms the benchmarks throughout the frame, with the performance gap widening over time. The results across various phase-noise levels indicate a favorable balance between achieving near-optimal performance in the absence of phase noise (almost realized at $n_\mathrm{s} = 0$) and maintaining robustness against increasing phase-noise distortions (at higher n_s values).

As mentioned, in Fig. 7 the DNN solution was trained for each level of the phase noise strength. In Table II, we illustrate the impact of potential discrepancies in phase noise levels between training and inference, as it might occur in realistic scenarios, on the performance of our proposed method. The values shown in this table represent the relative change in average frame AIR between the cases of DNN trained at L_{train} and evaluated at $L_{\mathrm{inference}}$ and the ideal case of $L_{\mathrm{train}} = L_{\mathrm{inference}}$. We observe that the DNN solution demonstrates considerable robustness to discrepancies in L. For example, a 10 dB discrepancy in phase-noise level results in a 15% decline in AIR, which we deem as relatively mild.

We now constrain phase shifters to have finite resolution. Fig. 8 shows the average AIRs when using phase shifters with 2-bit and 4-bit resolution. The curves for infinite-resolution phase shifters are included as a reference. Both the HOSVD and CTDH methods quantize the phase shifter angles post-convergence of their respective beamforming optimization

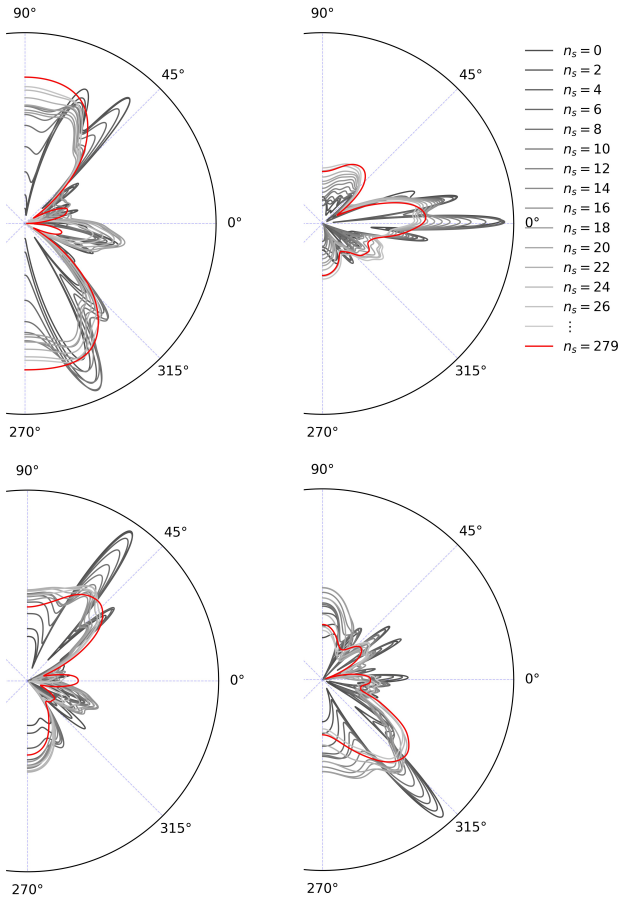

Figure 6. Progression of beamforming pattern at the BS over time (n_{s}) for the proposed DNN. In this example, the BS serves one UE. The four plots correspond to four different realizations of the propagation channel and user location at a phase noise level of $L = -100$ dBc/Hz.

Table II
RELATIVE PERFORMANCE LOSS BECAUSE OF MISMATCH BETWEEN TRAINING WITH L_{train} AND TESTING AT $L_{\mathrm{inference}}$.

$L_{\mathrm{inference}}$ \ L_{train}	-105 dBc/Hz	-100 dBc/Hz	-95 dBc/Hz
-105 dBc/Hz	0%	-2%	-5%
-100 dBc/Hz	-13%	0%	-5%
-95 dBc/Hz	-15%	-12%	0%

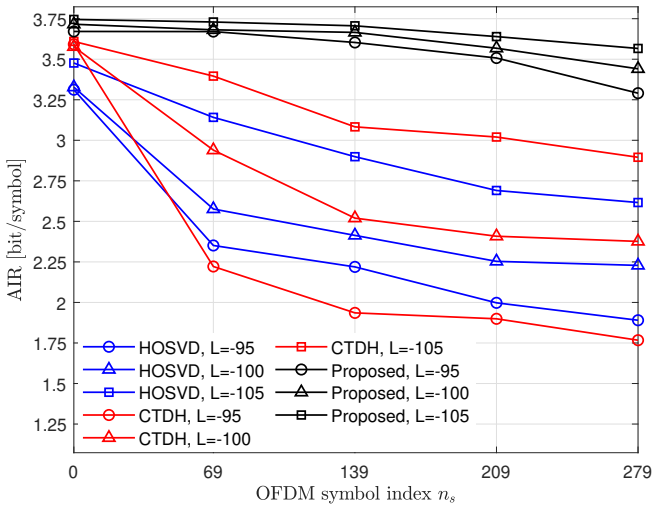

Figure 7. Average user AIR as a function of OFDM symbol index n_{s} for three different phase-noise levels L in MU-MIMO. Optimization and learning based HBF systems with infinite-resolution analog phase shifters and perfect CSI.

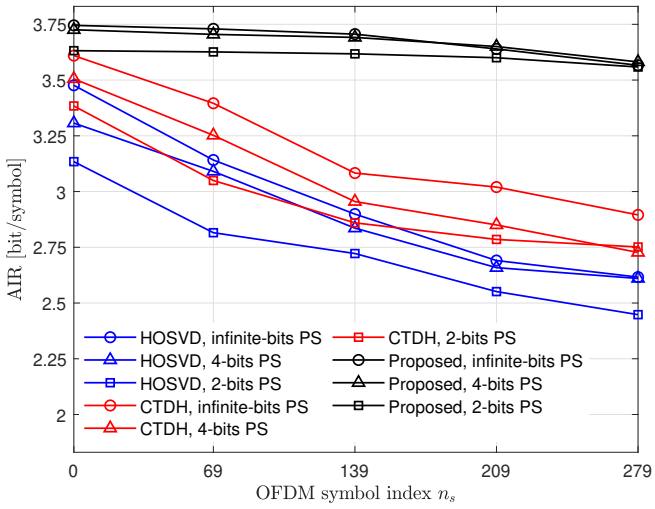

Figure 8. Average user AIR as a function of OFDM symbol index n_{s} for phase noise with $L = -105$ dBc/Hz in MU-MIMO. Optimization and learning based HBF systems with 2-bit and 4-bit analog phase shifters. Perfect CSI is assumed.

algorithms, and then adjust the transmit power by scaling the digital precoder considering the quantized analog precoder. We observe a degradation in the proposed DNN's performance at smaller OFDM symbol indices and a consistent degradation throughout the frame across benchmark methods when finite resolution phase shifters are introduced. Recall from Fig. 6 that the beamforming patterns generated by the DNN are smoother at higher indices, which do not require high-resolution phase shifters to produce. Consequently, the degradation is less pronounced towards larger n_{s} values. Additionally, incorporating the phase-shifter resolution constraint during the DNN's training phase through the tanh-approximated quantizer in (18) as an activation function proves effective. This approach results in only a moderate performance degradation when using practical phase shifters with low resolution.

Figure 9 presents the AIR results under imperfect CSI and 4-bit phase shifters. Imperfect CSI is modeled as an additive Gaussian channel estimation error with a normalized MSE (NMSE) of -10 dB, and the proposed DNN is trained on channel realizations with this NMSE. In addition to the proposed method and benchmarks HOSVD and CTDH, the figure includes results from an ablation study where the adaptive attention component (AdaSE) was removed, and only the remaining neural network, consisting of ResNets, was trained. The results show a significant performance decline of approximately 1 bit/symbol across the proposed and

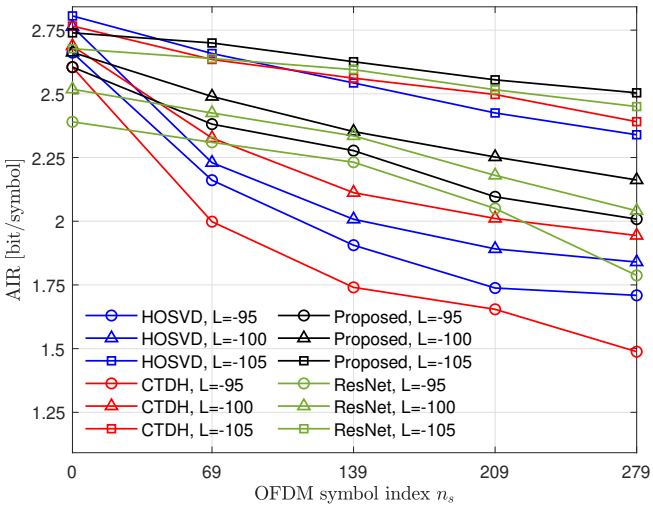

Figure 9. Average user AIR as a function of OFDM symbol index n_{s} for three different phase-noise levels L in MU-MIMO. Optimization and learning based HBF systems with 4-bit analog phase shifters. Imperfect CSI with an NMSE of -10 dB is assumed. "ResNet" represents the DNN obtained by excluding the AdaSE component from the proposed DNN, retaining only the ResNet structure to explicitly assess the effectiveness of the AdaSE component.

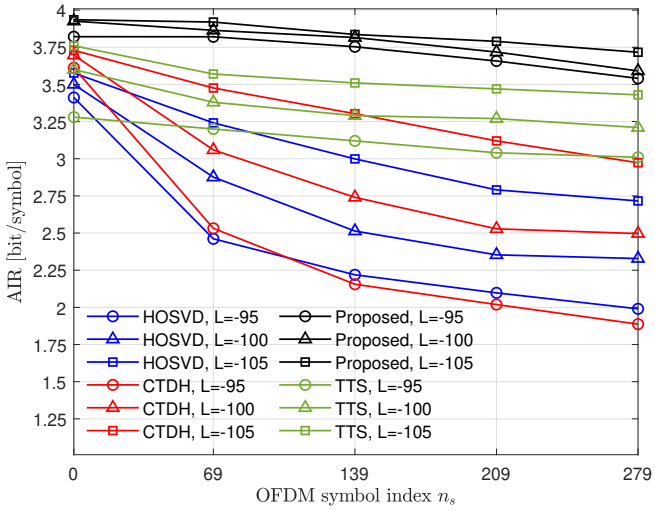

Figure 10. Average AIR as a function of OFDM symbol index n_{s} for three different phase-noise levels L in a SU-MIMO scenario. Optimization and learning based HBF systems with infinite-resolution analog phase shifters and perfect CSI.

benchmark methods due to imperfect CSI. Near $n_{\mathrm{s}} = 0$, distortion from imperfect CSI dominates because phase-noise-induced beamforming mismatch has not yet accumulated, leading to similar performance across all methods. However, as n_{s} increases, the AdaSE-ResNet solution consistently outperforms the benchmarks, particularly at higher phase noise levels, where it maintains a substantial performance advantage. Additionally, the ResNet-only DNN, obtained by removing the AdaSE component, experiences a decline in AIR, especially at the beginning and end of the frame. Without n_{s} as an anchor to tailor beam patterns for each symbol index, the DNN learns a non-adaptive, universal beam pattern that balances the needs of highly directive lobes for early symbols and smooth lobes for late symbols, resulting in a mid-level smoothed pattern most effective for the middle symbols.

To compare our proposed method with the TTS method, as well as HOSVD and CTDH, we evaluate a single-user scenario under three phase-noise levels: $L = -95, \mathrm{dBc/Hz}$, $L = -100, \mathrm{dBc/Hz}$, and $L = -105, \mathrm{dBc/Hz}$, assuming perfect CSI and infinite-resolution phase shifters. Fig. 10 depicts the average AIR in this scenario. We observe that the TTS method exhibits superior robustness to beamforming mismatches compared to the other benchmarks. This is achieved through the recalculation of digital beamformers using the equivalent channel $\boldsymbol{W}_{\mathrm{RF}}^{\mathrm{H}} \tilde{\boldsymbol{H}}_{n_{\mathrm{s}}}^{u}[k] \boldsymbol{V}_{\mathrm{RF}}$. However, this advantage comes at the cost of requiring (i) uplink transmission of the low-dimensional CSI for every OFDM symbol and (ii) additional processing power at the UE to execute part of the DNN for generating the digital combiner during inference. The latter can be limiting in legacy user devices or cost-sensitive systems compared to HOSVD, CTDH, and our DNN solution. Moreover, like the other benchmarks, TTS does not model the phase-noise-induced ICI in its problem formulation and loss function. Consequently, it suffers from a significant performance gap relative to our proposed method, particularly under high phase noise levels.

C. Implementation Remarks

While the proposed HBF method improves over state-of-the-art approaches, we acknowledge necessary practical considerations for real-world deployment. Implementing machine-learning (ML)-based HBF requires specialized hardware, such as graphics processing units (GPUs) or ML accelerators, to execute inference efficiently. Although modern base station deployments increasingly incorporate such hardware, legacy systems and cost-sensitive deployments may face significant challenges due to limited computational resources. To mitigate these constraints, computational cost optimizations such as model quantization, pruning, and distillation can be applied [70], [71]. These techniques reduce computational and memory overhead, enhancing the feasibility of the proposed method across a broader range of deployment scenarios.

Another critical factor is the dynamic nature of wireless environments, particularly in high-mobility scenarios where channel conditions vary rapidly. Addressing this requires periodic evaluation of the model's performance and adaptation to real-world conditions to prevent degradation. The data-driven nature of the proposed method enables on-line retraining or fine-tuning using updated samples, ensuring adaptability and robustness under varying operating conditions [44].

Online fine-tuning for ML-based HBF has primarily utilized methods such as model-agnostic meta-learning [45], [46] and direct transfer learning with online retraining [44], [47]. These approaches are explicitly designed to achieve efficient fine-tuning with a minimal number of samples, thereby minimizing the computational cost of online training. Despite significant advancements, they remain an active area of research, attracting substantial interest, particularly among practitioners.

V. Conclusion

This paper introduced and evaluated a new data-driven approach for optimizing HBF in multi-user MIMO communication systems at mmWave frequencies. The proposed solution distinguishes itself by considering the practical implementation challenge of phase noise impairments originating from distributed local oscillators, and also accounts for limitations in analog phase shifter resolution. We developed a self-supervised learning algorithm equipped with a phase-noise-aware loss function and an attention mechanism that facilitates time-adaptive beamforming calibration. This approach pre-distorts the antenna radiation pattern specifically for each OFDM symbol, thereby mitigating beamforming mismatches induced by phase noise. Accordingly, we observed a trend in beamforming radiation patterns towards smoother configurations across symbol indices, which effectively reduces sensitivity to these mismatches. Simulation results demonstrated the resilience of our proposed model against individual and compounded distortions, highlighting its benefits for practical use cases.

Appendix

Table III provides the detailed information about the components of the DNN used in our proposed method.

References

[1] X. Shi, J. Wang, Z. Sun, and J. Song, "Spatial-chirp codebook-based hierarchical beam training for extremely large-scale massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 23, no. 4, pp. 2824–2838, 2024.

[2] Y. Han, S. Jin, J. Zhang, J. Zhang, and K.-K. Wong, "DFT-based hybrid beamforming multiuser systems: Rate analysis and beam selection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 3, pp. 514–528, 2018.

[3] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1432–1443, 2017.

[4] D. Zhang, Y. Wang, X. Li, and W. Xiang, "Hybrid beamforming for downlink multiuser millimetre wave MIMO-OFDM systems," *IET Communications*, vol. 13, no. 11, pp. 1557–1564, 2019.

[5] G. M. Zilli and W.-P. Zhu, "Constrained tensor decomposition-based hybrid beamforming for mmWave massive MIMO-OFDM communication systems," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 5775–5788, 2021.

[6] H. Li, M. Li, Q. Liu, and A. L. Swindlehurst, "Dynamic hybrid beamforming with low-resolution PSs for wideband mmWave MIMO-OFDM systems," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 9, pp. 2168–2181, 2020.

[7] H. Gao, D. Liu, and Z. Zhang, "Energy efficiency maximization for partially-connected hybrid beamforming architecture with low-resolution DACs," *IEEE Transactions on Communications*, vol. 72, no. 9, pp. 5765–5780, 2024.

[8] Q. Hu, Y. Cai, K. Kang, G. Yu, J. Hoydis, and Y. C. Eldar, "Two-timescale end-to-end learning for channel acquisition and hybrid precoding," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 163–181, 2022.

[9] K. Xu, F.-C. Zheng, H. Xu, X. Zhu, and K.-C. Leung, "Codebook-based hybrid beamforming using combined phase shifters of high and low resolutions," *IEEE Wireless Communications Letters*, vol. 10, no. 12, pp. 2683–2687, 2021.

[10] S.-S. Wong, C.-F. Teng, and A.-Y. Wu, "Two-step codebook-assisted alternating minimization (CA-AltMin) for low-complexity hybrid beamforming design," *IEEE Communications Letters*, vol. 25, no. 6, pp. 1989–1993, 2021.

[11] S.-G. Yoon and S. J. Lee, "Improved hierarchical codebook-based channel estimation for mmwave massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 11, no. 10, pp. 2095–2099, 2022.

[12] 3GPP, "TS 38 214 V18.3.0: Technical Specification," Tech. Rep., 2024. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3216

[13] H. Iimori, G. T. F. De Abreu, O. Taghizadeh, R.-A. Stoica, T. Hara, and K. Ishibashi, "A stochastic gradient descent approach for hybrid mmWave beamforming with blockage and CSI-Error robustness," *IEEE Access*, vol. 9, pp. 74 471–74 487, 2021.

[14] L.-H. Shen, Y.-C. Lo, K.-T. Feng, S.-H. Wu, and L.-L. Yang, "MARS: Message passing for antenna and RF chain selection for hybrid beamforming in MIMO communication systems," *IEEE Transactions on Communications*, vol. 72, no. 11, pp. 7198–7214, 2024.

[15] Y. Chen, Y. Huang, C. Li, Y. T. Hou, and W. Lou, "Turbo-HB: A sub-millisecond hybrid beamforming design for 5G mmWave systems," *IEEE Transactions on Mobile Computing*, vol. 22, no. 7, pp. 4332–4346, 2023.

[16] Z. Hong, T. Li, F. Li, and R. Ju, "Channel estimation-free deep direct beamforming with low complexity in mmWave massive MIMO," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 7790–7802, 2021.

[17] M. S. Ibrahim, A. Konar, and N. D. Sidiropoulos, "Fast algorithms for joint multicast beamforming and antenna selection in massive MIMO," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1897–1909, 2020.

[18] Q. Deng, X. Liang, X. Wang, M. Huang, C. Dong, and Y. Zhang, "Fast converging iterative precoding for massive MIMO systems: An accelerated weighted neumann series-steepest descent approach," *IEEE Access*, vol. 8, pp. 50 244–50 255, 2020.

[19] J.-C. Chen, "Low-cost and power-efficient massive MIMO precoding: Architecture and algorithm designs," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 7, pp. 7429–7442, 2020.

[20] K. P. Rajput, P. Maity, S. Srivastava, V. Sharma, N. K. D. Venkategowda, A. K. Jagannatham, and L. Hanzo, "Robust linear hybrid beamforming designs relying on imperfect CSI in mmWave MIMO IoT networks," *IEEE Internet of Things Journal*, vol. 10, no. 10, pp. 8893–8906, 2023.

[21] M. Jafri, A. Anand, S. Srivastava, A. K. Jagannatham, and L. Hanzo, "Robust distributed hybrid beamforming in coordinated multi-user multi-cell mmWave MIMO systems relying on imperfect CSI," *IEEE Transactions on Communications*, vol. 70, no. 12, pp. 8123–8137, 2022.

[22] R. Zhang, B. Shim, and H. Zhao, "Downlink compressive channel estimation with phase noise in massive MIMO systems," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5534–5548, 2020.

[23] R. Mushini and J. Dooley, "Strategic initialization of genetic algorithm used in digital pre-distortion of mmWave power amplifiers for hybrid beamforming," in *IEEE Topical Conference on RF/Microwave Power Amplifiers for Radio and Wireless Applications (PAWR)*, 2022, pp. 50–53.

[24] N. N. Moghadam, G. Fodor, M. Bengtsson, and D. J. Love, "On the energy efficiency of MIMO hybrid beamforming for millimeter-wave systems with nonlinear power amplifiers," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7208–7221, 2018.

[25] J. Li, Z. Wang, Y. Zhang, P. Zhu, D. Wang, and X. You, "Robust hybrid beamforming for outage-constrained multigroup multicast mmWave transmission with phase shifter impairments," *IEEE Systems Journal*, vol. 17, no. 1, pp. 869–880, 2023.

[26] X. Cheng, K. Xu, and S. Li, "Compensation of phase noise in uplink massive MIMO OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1764–1778, 2019.

[27] V. V. Ratnam, "Performance of analog beamforming systems with optimized phase noise compensation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5334–5348, 2020.

[28] X. Chen, C. Fang, Y. Zou, A. Wolfgang, and T. Svensson, "Beamforming MIMO-OFDM systems in the presence of phase noises at millimeter-wave frequencies," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2017, pp. 1–6.

[29] M. E. Rasekh, M. Abdelghany, U. Madhow, and M. Rodwell, "Phase noise in modular millimeter wave massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6522–6535, 2021.

[30] P. Aggarwal, A. Pradhan, and V. A. Bohara, "A downlink multiuser MIMO–OFDM system with nonideal oscillators and amplifiers: Characterization and performance analysis," *IEEE Systems Journal*, vol. 15, no. 1, pp. 715–726, 2021.

[31] B. Chatelier and M. Crussière, "On the impact of phase noise on beamforming performance for mmWave massive MIMO systems," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 1563–1568.

[32] M. U. Aminu, J. Lehtomäki, and M. Juntti, "Beamforming and transceiver optimization with phase noise for mmWave and THz

Table III
MODULES AND THEIR PARAMETERS USED IN THE PROPOSED NEURAL NETWORK SHOWN IN FIGURES 3 AND 4.

Module	Abbreviation	Description	Input size, channels
AdaSE-ResNet of common analog beamforming branch type 1	AdaSE-RN-A1	Conv2D layer 1: [kernel=(3x3), channel=2], Conv2D layer 2: [kernel=(3x3), channel=64], Fully connected:[layer1=64, layer2=2]	$[N_{\mathrm{UE}}N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 2$
AdaSE-ResNet of common analog beamforming branch type 2	AdaSE-RN-A2	Conv2D layer 1: [kernel=(3x3), channel=64], Conv2D layer 2: [kernel=(3x3), channel=64], Fully connected:[layer1=64, layer2=2]	$[N_{\mathrm{UE}}N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 64$
AdaSE-ResNet of analog combining branch type 1	AdaSE-RN-AC1	Conv2D layer 1: [kernel=(3x3), channel=64], Conv2D layer 2: [kernel=(3x3), channel=64], Fully connected:[layer1=64, layer2=2]	$[N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 64$
AdaSE-ResNet of analog combining branch type 2	AdaSE-RN-AC2	Conv2D layer 1: [kernel=(3x3), channel=64], Conv2D layer 2: [kernel=(3x3), channel=1], Fully connected:[layer1=64, layer2=2]	$[N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 1$
AdaSE-ResNet of analog precoding branch type 1	AdaSE-RN-AP1	Conv2D layer 1: [kernel=(3x3), channel=64], Conv2D layer 2: [kernel=(3x3), channel=64], Fully connected:[layer1=64, layer2=2]	$[N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 64$
AdaSE-ResNet of analog precoding branch type 2	AdaSE-RN-AP2	Conv2D layer 1: [kernel=(3x3), channel=64], Conv2D layer 2: [kernel=(3x3), channel=1], Fully connected:[layer1=64, layer2=2]	$[N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 1$
AdaSE-ResNet of common digital beamforming branch type 1	AdaSE-RN-D1	Conv3D layer 1: [kernel=(3x3x3), channel=2], Conv3D layer 2: [kernel=(3x3x3), channel=64], Fully connected:[layer1=64, layer2=2]	$[K', N_{\mathrm{UE}}N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 2$
AdaSE-ResNet of common digital beamforming branch type 2	AdaSE-RN-D2	Conv3D layer 1: [kernel=(3x3x3), channel=64], Conv3D layer 2: [kernel=(3x3x3), channel=64], Fully connected:[layer1=64, layer2=2]	$[K', N_{\mathrm{UE}}N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 64$
AdaSE-ResNet of common digital beamforming branch type 3	AdaSE-RN-D3	Conv3D layer 1: [kernel=(3x3x3), channel=64], Conv3D layer 2: [kernel=(3x3x3), channel=64], Fully connected:[layer1=64, layer2=2]	$[K', N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 64$
AdaSE-ResNet of common digital combining branch type 1	AdaSE-RN-DC1	Conv3D layer 1: [kernel=(3x3x3), channel=64], Conv3D layer 2: [kernel=(3x3x3), channel=64], Fully connected:[layer1=64, layer2=2]	$[K', N_{\mathrm{U}}^{\mathrm{RF}}, N], 64$
AdaSE-ResNet of common digital combining branch type 2	AdaSE-RN-DC2	Conv3D layer 1: [kernel=(3x3x3), channel=64], Conv3D layer 2: [kernel=(3x3x3), channel=2], Fully connected:[layer1=64, layer2=2]	$[K', N_{\mathrm{U}}^{\mathrm{RF}}, N], 2$
AdaSE-ResNet of common digital precoding branch type 1	AdaSE-RN-DP1	Conv3D layer 1: [kernel=(3x3x3), channel=64], Conv3D layer 2: [kernel=(3x3x3), channel=64], Fully connected:[layer1=64, layer2=2]	$[K', N_{\mathrm{B}}^{\mathrm{RF}}, N], 64$
AdaSE-ResNet of common digital precoding branch type 2	AdaSE-RN-DP2	Conv3D layer 1: [kernel=(3x3x3), channel=64], Conv3D layer 2: [kernel=(3x3x3), channel=2], Fully connected:[layer1=64, layer2=2]	$[K', N_{\mathrm{B}}^{\mathrm{RF}}, N], 2$
Max pooling of analog combining branch	MP-AC	pool size = $[1, N_{\mathrm{B}}^{\mathrm{a}}]$, padding = yes	$[N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 1$
Max pooling of analog precoding branch type 1	MP-AP1	pool size = $[N_{\mathrm{UE}}, 1]$, padding = yes	$[N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 64$
Max pooling of analog precoding branch type 2	MP-AP2	pool size = $[N_{\mathrm{U}}^{\mathrm{a}}, 1]$, padding = yes	$[N_{\mathrm{B}}^{\mathrm{a}}], 1$
Max pooling of digital combining branch	MP-DC	pool size = $[1, \frac{N_{\mathrm{U}}^{\mathrm{a}}}{N_{\mathrm{U}}^{\mathrm{RF}}}, \frac{N_{\mathrm{B}}^{\mathrm{a}}}{N}]$, padding = yes	$[K', N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 64$
Max pooling of digital precoding branch	MP-DP	pool size = $[1, \frac{N_{\mathrm{U}}^{\mathrm{a}}}{N_{\mathrm{B}}^{\mathrm{RF}}}, \frac{N_{\mathrm{B}}^{\mathrm{a}}}{N}]$, padding = yes	$[K', N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{a}}], 64$
tanh-approximated quantizer	$\hat{Q}_{N_{\mathrm{b}}}$	equation (18)	any, 1
Cast to complex	CC	Creates a complex-valued tensor using the two channels of the input as real and imaginary parts, respectively.	any, 2
Enforce constraints of analog precoding branch	$f_{\mathrm{RF}}^{\mathrm{B}}$	equation (19)	$[N_{\mathrm{B}}^{\mathrm{a}}, N_{\mathrm{B}}^{\mathrm{RF}}], 1$
Enforce constraints of analog combining branch	$f_{\mathrm{RF}}^{\mathrm{UE}}$	equation (20)	$[N_{\mathrm{U}}^{\mathrm{a}}, N_{\mathrm{U}}^{\mathrm{RF}}], 1$
Normalize transmit power	$f_{\mathrm{D}}^{\mathrm{B}}$	equation (21)	any, 1

bands," in *International Symposium on Wireless Communication Systems (ISWCS)*, 2019, pp. 692–696.

[33] T. Höhne and V. Ranki, "Phase noise in beamforming," *IEEE Transactions on Wireless Communications*, vol. 9, no. 12, pp. 3682–3689, 2010.

[34] R. Krishnan, M. R. Khanzadi, N. Krishnan, Y. Wu, A. Graell i Amat, T. Eriksson, and R. Schober, "Linear massive MIMO precoders in the presence of phase noise- A large-scale analysis," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 3057–3071, 2016.

[35] R. Nissel, "Correctly modeling TX and RX chain in (distributed) massive MIMO-new fundamental insights on coherency," *IEEE Communications Letters*, vol. 26, no. 10, pp. 2465–2469, 2022.

[36] F. Liu, Z. Duan, L. Zhang, B. Shi, Y. Liu, and R. Du, "DPC-CNN algorithm for multiuser hybrid precoding with dynamic structure," *IEEE Transactions on Green Communications and Networking*, vol. 8, no. 4, pp. 1361–1370, 2024.

[37] Y. Zhang, J. Yang, Q. Liu, Y. Liu, and T. Zhang, "Unsupervised learning-based coordinated hybrid precoding for mmWave massive MIMO-enabled HetNets," *IEEE Transactions on Wireless Communications*, vol. 23, no. 7, pp. 7200–7213, 2024.

[38] A. Pitarokoilis, S. K. Mohammed, and E. G. Larsson, "Uplink performance of time-reversal MRC in massive MIMO systems subject to phase noise," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 711–723, 2015.

[39] M. R. Khanzadi, G. Durisi, and T. Eriksson, "Capacity of SIMO and MISO phase-noise channels with common/separate oscillators," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3218–3231, 2015.

[40] Y. Wang and J. Lee, "A simple phase noise suppression scheme for massive MIMO uplink systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 4769–4780, 2017.

[41] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[42] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, pp. 1–74, 2021.

[43] J. M. J. Huttunen, D. Korpi, and M. Honkala, "DeepTx: Deep learning beamforming with channel prediction," *IEEE Transactions on Wireless Communications*, vol. 22, no. 3, pp. 1855–1867, 2023.

[44] Y. Yuan, G. Zheng, K.-K. Wong, B. Ottersten, and Z.-Q. Luo, "Transfer learning and meta learning-based fast downlink beamforming adaptation," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1742–1755, 2021.

[45] Q. Hou, M. Lee, G. Yu, and Y. Cai, "Meta-gating framework for fast and continuous resource optimization in dynamic wireless environments,"

IEEE Transactions on Communications, vol. 71, no. 9, pp. 5259–5273, 2023.

[46] T. Raviv, S. Park, O. Simeone, Y. C. Eldar, and N. Shlezinger, "Online meta-learning for hybrid model-based deep receivers," *IEEE Transactions on Wireless Communications*, vol. 22, no. 10, pp. 6415–6431, 2023.

[47] A. H. Karkan, H. Hojatian, J.-F. Frigon, and F. Leduc-Primeau, "SAGE-HB: Swift adaptation and generalization in massive MIMO hybrid beamforming," in *IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*, 2024, pp. 323–328.

[48] K. M. Attiah, F. Sohrabi, and W. Yu, "Deep learning for channel sensing and hybrid precoding in TDD massive MIMO OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 839–10 853, 2022.

[49] M. Zhang, J. Gao, and C. Zhong, "A deep learning-based framework for low complexity multiuser MIMO precoding design," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 11 193–11 206, 2022.

[50] M. Goutay, F. A. Aoudia, J. Hoydis, and J.-M. Gorce, "Machine learning for MU-MIMO receive processing in OFDM systems," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2318–2332, 2021.

[51] Z. Liu, Y. Yang, F. Gao, T. Zhou, and H. Ma, "Deep unsupervised learning for joint antenna selection and hybrid beamforming," *IEEE Transactions on Communications*, vol. 70, no. 3, pp. 1697–1710, 2022.

[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[53] D. Petrovic, W. Rave, and G. Fettweis, "Effects of phase noise on OFDM systems with and without PLL: Characterization and compensation," *IEEE Transactions on Communications*, vol. 55, no. 8, pp. 1607–1616, 2007.

[54] A. Mehrotra, "Noise analysis of phase-locked loops," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 49, no. 9, pp. 1309–1316, 2002.

[55] T. A. Thomas, M. Cudak, and T. Kovarik, "Blind phase noise mitigation for a 72 GHz millimeter wave system," in *IEEE International Conference on Communications (ICC)*, 2015, pp. 1352–1357.

[56] A. Pitarokoilis, E. Björnson, and E. G. Larsson, "Performance of the massive MIMO uplink with OFDM and phase noise," *IEEE Communications Letters*, vol. 20, no. 8, pp. 1595–1598, 2016.

[57] V. Syrjälä, T. Levanen, T. Ihalainen, and M. Valkama, "Pilot allocation and computationally efficient non-iterative estimation of phase noise in OFDM," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 640–643, 2019.

[58] X. Yu, X. Gao, A.-A. Lu, J. Zhang, H. Wu, and G. Y. Li, "Robust precoding for HF skywave massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 22, no. 10, pp. 6691–6705, 2023.

[59] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[60] F. Ait Aoudia and J. Hoydis, "End-to-end learning for OFDM: From neural receivers to pilotless communication," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 1049–1063, 2022.

[61] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[62] F. Sohrabi, K. M. Attiah, and W. Yu, "Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4044–4057, 2021.

[63] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2315–2328, 2022.

[64] Z. Bo, R. Liu, Y. Guo, M. Li, and Q. Liu, "Deep learning based low-resolution hybrid precoding design for mmWave MISO systems," in *IEEE Globecom Workshops (GC Wkshps)*, 2020, pp. 1–6.

[65] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[66] 3GPP, "TS 38 211 V18.3.0: Technical Specification," Tech. Rep., 2024. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3213

[67] P. Neshaastegaran and A. H. Banihashemi, "Log-likelihood ratio calculation for pilot symbol assisted coded modulation schemes with residual phase noise," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3782–3790, 2019.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[69] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[70] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3579–3605, 2021.

[71] Z. Li, H. Li, and L. Meng, "Model compression for deep neural networks: A survey," *Computers*, vol. 12, no. 3, p. 60, 2023.