CNN-Based Hybrid Precoding Design with Geometric Mean Decomposition

Mahmoud A. Abugubba[1], Nagia M. Gaboua[1], Taissir Y. Elganimi[1] and Khaled M. Rabie[2]
[1]Department of Electrical and Electronic Engineering, University of Tripoli, Libya
[2]Department of Engineering, Manchester Metropolitan University, United Kingdom
Email: {ma.abugubba, n.gaboua, t.elganimi}@uot.edu.ly, k.rabie@mmu.ac.uk

***Abstract*—Communications over millimeter-wave (mmWave) frequencies is a key technology for the fifth generation (5G) cellular networks due to the large bandwidth available at mmWave bands. The short wavelength of mmWave bands enables large antenna arrays to be placed on the transceivers which forms massive multiple-input multiple-output (MIMO). Massive MIMO with conventional fully-digital (FD) beamforming is difficult to be implemented due to high power consumption and hardware cost. One of the most effective solutions to this problem is hybrid beamforming which can be used to balance the beamforming gain, hardware implementation cost, and the power consumption. However, due to the non-convex constraints imposed by phase shifters, finding the global optima for the hybrid beamforming system is very challenging with high computational complexity. To address this issue, deep learning (DL)-based hybrid precoding with geometric mean decomposition (GMD) algorithm for narrowband mmWave massive MIMO system is proposed in this paper, where it can directly estimate the hybrid analog and digital precoders (combiners) from a given optimal FD precoder (combiner). Simulation results demonstrated that the proposed hybrid precoding model can more accurately approximate the FD precoding performance.**

***Index Terms*—Massive MIMO, millimeter-wave, fully digital precoding, hybrid precoding, deep learning, CNN.**

I. Introduction

A key technology for the fifth generation (5G) of wireless communications and beyond is millimeter-wave (mmWave) communication, which offers greater data rates on the order of gigabit per second (Gbps), wider bandwidth, and higher spectral efficiency than conventional cellular communications [1]. However, the communication over the mmWave band is very challenging due to signal propagation and channel properties. To address these issues, massive multiple-input multiple-output (MIMO) with beamforming techniques is required, where a large antenna arrays are used to concentrate the radiated energy and steer it toward the receiver direction [2]. Higher diversity and multiplexing gains can be attained through massive MIMO with beamforming gain, which leads to higher spectral efficiency and higher radiated energy efficiency [3]. Massive MIMO is costly and very difficult to be combined with mmWave, though. One of the promising approaches to address these problems is hybrid beamforming, which utilizes significantly fewer power-hungry radio frequency (RF) chains to achieve the performance of fully digital (FD) beamforming, hence reducing the power consumption and the system implementation complexity [4].

The hybrid beamforming design problem is a non-convex optimization problem due to the constant modulus constraints that imposed by phase shifters. Most existing methods reduce the complexity by decoupling the optimization problem into two sub-problems, where the objective of each sub-problem is to approximate the FD precoding using matrix decomposition [5]. In [6], a phase-extraction (PE) and manifold optimization (MO)-based alternating minimization algorithm have been proposed. Although the MO algorithm can achieve near optimal performance by iteratively reducing the Euclidean distance between both the hybrid beamformer and the FD beamformer, its computational complexity prevents it from being used in implementations. In the PE algorithm, the computational complexity has been reduced with slight performance loss, but it still provides a better precoding algorithm than most existing algorithms. These algorithms are based on the singular value decomposition (SVD) which requires complicated bit allocation schemes to achieve the channel capacity. To avoid this issue, geometric mean decomposition (GMD) was proposed in [7] to decompose the channel matrix into several parallel subchannels with equal signal-to-noise-ratios (SNRs), and hence the simple identical bit allocation can be utilized for all subchannels. However, high computational complexity is still a great challenge in hybrid beamforming design.

Recently, embedding deep learning (DL) into wireless communications has had a great impact on solving complex problems and high computation issues. For example, the authors in [8] used deep recurrent neural network to optimize the resource allocation problem with low computational complexity for the non-orthogonal multiple access (NOMA) heterogeneous internet of things (IoT) network. In [9], two convolutional neural networks (CNNs) were proposed to address the problem of modulation classification, where it can accurately recognize various modulation types. In another work [10], a deep neural network (DNN) was employed to estimate the channel and direction-of-arrival for massive MIMO systems with better performance than the conventional methods.

DL-based hybrid precoding schemes for mmWave massive MIMO systems has also gained the attention of many researchers. For example, in [11], a DNN is trained to optimize the hybrid precoding using GMD. In another work [12], a CNN-based hybrid precoding is proposed with two CNNs, where each CNN is trained to estimate the hybrid precoders.

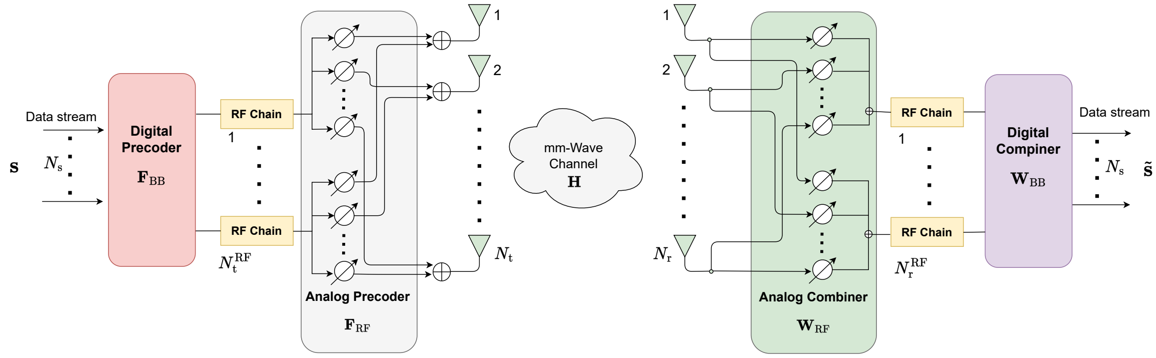

Fig. 1. Block diagram of mmWave massive MIMO system with hybrid precoding.

Later, a joint hybrid precoding framework based on DNN is proposed with end-to-end optimization [13]. In order to reduce the computation time, deep reinforcement learning has been applied to hybrid beamforming designs in [14], while an auto-encoder based on DNN is proposed in [15] for multi-user scenarios. In a recent work [16], the authors designed a hybrid precoding algorithm based on attention layer and CNN which is trained via unsupervised learning to maximize the spectral efficiency directly.

Motivated by the foregoing introduction and recent work, a CNN-based hybrid precoding with GMD-based algorithm is proposed in this paper. More specifically, the main contribution of this paper is to design hybrid precoding using DL approaches. The proposed hybrid preceding model simulate the working operation of phase shifters, and satisfy the power constraint on digital precoder. In addition, the model can directly estimate the hybrid analog and digital precoders and combiners from a given FD precoder and combiner, and also is trained with GMD preceding dataset to take advantages of GMD over SVD precoding. Simulation results verified the ability of the proposed hybrid preceding model to approximate the FD precoder performance.

The rest of this paper is organized as follows. In Section II, the system and channel models are presented for mmWave massive MIMO system with hybrid precoding. The system model of the proposed CNN-based hybrid precoding with GMD-based algorithm is presented in Section III. The simulation results are discussed in Section IV, and the conclusions and possible future work are given in Section V.

II. System Model

In this section, the system and channel models for mmWave massive MIMO system with hybrid precoding are described, and the problem formulation for maximizing the achievable rate is also discussed.

A. System Model

In Fig. 1, a narrowband mmWave massive MIMO system with hybrid precoding is considered. The N_t transmit antennas transmit N_s independent data streams collected by N_r receive antennas. The transmitter and the receiver are equipped with $N_\mathrm{t}^{\mathrm{RF}}$ and $N_\mathrm{r}^{\mathrm{RF}}$ RF chains, respectively, where $N_\mathrm{s} \leq N_\mathrm{t}^{\mathrm{RF}} \leq N_\mathrm{t}$ and $N_\mathrm{s} \leq N_\mathrm{r}^{\mathrm{RF}} \leq N_\mathrm{r}$ in order to enable data stream multiplexing between the transmitter and receiver [5]. In Fig. 1, the symbol vector $\mathbf{s} \in \mathbb{C}^{N_\mathrm{s}\times 1}$ that satisfy $\mathbb{E}[\mathbf{s}\mathbf{s}^*] = \frac{1}{N_\mathrm{s}}\mathbf{I}_{N_\mathrm{s}}$ is firstly precoded by low dimensional digital precoder $\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{N_\mathrm{t}^{\mathrm{RF}}\times N_\mathrm{s}}$, and then the high dimensional analog precoder $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_\mathrm{t}\times N_\mathrm{t}^{\mathrm{RF}}}$ is applied using phase shifters to shift the input phase and keep the amplitude as a constant. Therefore, all elements of $\mathbf{F}_{\mathrm{RF}}$ matrix are constrained to satisfy $(\mathbf{F}_{\mathrm{RF}}^{(i)}\mathbf{F}_{\mathrm{RF}}^{(i)H})_{i,j} = \frac{1}{N_\mathrm{t}}$ and can be represented by

$$\mathbf{F}_{\mathrm{RF}} = \frac{1}{\sqrt{N_\mathrm{t}}}\begin{bmatrix} e^{j\theta_{11}^{\mathbf{F}}} & e^{j\theta_{12}^{\mathbf{F}}} & \dots & e^{j\theta_{1N_{\mathrm{RF}}}^{\mathbf{F}}} \\ e^{j\theta_{21}^{\mathbf{F}}} & e^{j\theta_{22}^{\mathbf{F}}} & \dots & e^{j\theta_{2N_{\mathrm{RF}}}^{\mathbf{F}}} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j\theta_{N_\mathrm{t}1}^{\mathbf{F}}} & e^{j\theta_{N_\mathrm{t}2}^{\mathbf{F}}} & \dots & e^{j\theta_{N_\mathrm{t}N_{\mathrm{RF}}}^{\mathbf{F}}} \end{bmatrix}, \tag{1}$$

where $\theta_{ij}^{\mathbf{F}} \in [0, 2\pi]$ for phase shifters of the analog precoder. The transmitted signal at the transmit antenna array can be written as $\mathbf{x} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s}$, where $\mathbf{F}_{\mathrm{BB}}$ must be normalized to satisfy the total power constrain $\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F^2 = N_\mathrm{s}$. Therefore, the received signal $\tilde{\mathbf{y}} \in \mathbb{C}^{N_\mathrm{r}\times 1}$ can be written as

$$\tilde{\mathbf{y}} = \sqrt{\rho}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + \mathbf{n}, \tag{2}$$

where ρ denotes the average received power, $\mathbf{H} \in \mathbb{C}^{N_\mathrm{r}\times N_\mathrm{t}}$ is the channel matrix that satisfy $\mathbb{E}[\|\mathbf{H}\|_{\mathbf{F}}^2] = N_\mathrm{t}N_\mathrm{r}$, and $\mathbf{n}$ is the additive white noise vector that follows an independent and identical distribution, $\mathbf{n} \sim \mathcal{CN}(0, \sigma_n^2)$.

The receiver has the similar structure as the transmitter, where the received symbol vector $\tilde{\mathbf{s}}$ after the combining process can be represented by

$$\begin{aligned}\tilde{\mathbf{s}} &= \mathbf{W}_{\mathrm{BB}}^H\mathbf{W}_{\mathrm{RF}}^H\tilde{\mathbf{y}} \\ &= \sqrt{\rho}\mathbf{W}_{\mathrm{BB}}^H\mathbf{W}_{\mathrm{RF}}^H\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + \mathbf{W}_{\mathrm{BB}}^H\mathbf{W}_{\mathrm{RF}}^H\mathbf{n},\end{aligned} \tag{3}$$

where $\mathbf{W}_{\mathrm{RF}} \in \mathbb{C}^{N_\mathrm{r}\times N_\mathrm{r}^{\mathrm{RF}}}$ and $\mathbf{W}_{\mathrm{BB}} \in \mathbb{C}^{N_\mathrm{r}^{\mathrm{RF}}\times N_\mathrm{s}}$ are the analog and digital combiners, respectively. $\mathbf{W}_{\mathrm{RF}}$ has a similar property as the analog precoder $\mathbf{F}_{\mathrm{RF}}$ which means it satisfies the constraint $(\mathbf{W}_{\text{RF}}^{(i)}\mathbf{W}_{\text{RF}}^{(i)H})_{i,j} = \frac{1}{N_{\text{r}}}$, and the entire elements matrix can be represented by

$$\mathbf{W}_{\text{RF}} = \frac{1}{\sqrt{N_{\text{r}}}} \begin{bmatrix} e^{j\theta_{11}^{\mathbf{W}}} & e^{j\theta_{12}^{\mathbf{W}}} & \dots & e^{j\theta_{1N_{\text{RF}}}^{\mathbf{W}}} \\ e^{j\theta_{21}^{\mathbf{W}}} & e^{j\theta_{22}^{\mathbf{W}}} & \dots & e^{j\theta_{2N_{\text{RF}}}^{\mathbf{W}}} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j\theta_{N_{\text{r}}1}^{\mathbf{W}}} & e^{j\theta_{N_{\text{r}}2}^{\mathbf{W}}} & \dots & e^{j\theta_{N_{\text{r}}N_{\text{RF}}}^{\mathbf{W}}} \end{bmatrix}, \tag{4}$$

where $\theta_{ij}^{\mathbf{W}} \in [0, 2\pi]$ for phase shifters of analog combiner. The spectral efficiency (in bits/s/Hz) achieved by the mmWave hybrid precoding system when Gaussian symbols are conveyed over the wireless channel can be expressed as [17]

$$\begin{aligned} R = \log_2 \Bigg(\Big| \mathbf{I}_{N_{\text{s}}} + \frac{\rho}{N_{\text{s}}} \mathbf{C}_n^{-1} \mathbf{W}_{\text{BB}}^H \mathbf{W}_{\text{RF}}^H \mathbf{H} \mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}} \\ \times \mathbf{F}_{\text{BB}}^H \mathbf{F}_{\text{RF}}^H \mathbf{H}^H \mathbf{W}_{\text{RF}} \mathbf{W}_{\text{BB}} \Big| \Bigg), \end{aligned} \tag{5}$$

where $\mathbf{C}_n = \sigma_n^2 \mathbf{W}_{\text{BB}}^H \mathbf{W}_{\text{RF}}^H \mathbf{W}_{\text{RF}} \mathbf{W}_{\text{BB}} \in \mathbb{C}^{N_{\text{s}} \times N_{\text{s}}}$ is the noise covariance matrix after combining.

B. Channel Model

The mmWave propagation is characterized as limited scattering environment due to severe free-space pathloss. For this reason, the extended Saleh-Valenzuela model is used in this paper to cluster the channel into N_c scattering clusters, each of which contributes N_{ray} propagation paths. Therefore, the narrowband clustered channel matrix $\mathbf{H} \in \mathbb{C}^{N_{\text{r}} \times N_{\text{t}}}$ can be represented by [5]:

$$\mathbf{H} = \sqrt{\frac{N_{\text{t}} N_{\text{r}}}{N_c N_{ray}}} \sum_{i=1}^{N_c} \sum_{j=1}^{N_{ray}} \alpha_{ij} \mathbf{a}_{\text{r}}(\phi_{ij}^{\text{r}}, \theta_{ij}^{\text{r}}) \mathbf{a}_{\text{t}}(\phi_{ij}^{\text{t}}, \theta_{ij}^{\text{t}})^H, \tag{6}$$

where α_{ij} is the complex gain of the j-th ray in i-th cluster, $(\phi_{ij}^{\text{t}}, \theta_{ij}^{\text{t}})$ and $(\phi_{ij}^{\text{r}}, \theta_{ij}^{\text{r}})$ are the angle of departure and arrival in azimuth and elevation planes, respectively. $\mathbf{a}_{\text{t}}$ and $\mathbf{a}_{\text{r}}$ are the normalized array response vectors of the transmit and receive antenna arrays corresponding to j-th ray in i-th cluster, respectively.

The antenna array that enables the beamforming in azimuth and elevation planes is a uniform planar array (UPA) which has the following response vector

$$\begin{aligned} \mathbf{a}_{\text{UPA}}(\phi, \theta) = \frac{1}{\sqrt{N}} \Big[1, \dots, e^{jkd(m \sin(\phi)\sin(\theta) + n\cos(\theta))}, \\ \dots, e^{jkd((N_1 - 1)\sin(\phi)\sin(\theta) + (N_2 - 1)\cos(\theta))} \Big]^T, \end{aligned} \tag{7}$$

where $k = \frac{2\pi}{\lambda}$, d is the element spacing between array elements, the antenna array size is $N = N_1 N_2$, and $0 \leq m < N_1$ and $0 \leq n < N_2$ are the indices of an antenna element.

C. Problem Formulation

The hybrid precoder and combiner's design aims to maximize the spectral efficiency as follows

$$\begin{aligned} \max_{\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}, \mathbf{W}_{\text{RF}}, \mathbf{W}_{\text{BB}}} \quad & R, \\ \text{s.t.} \quad & \mathbf{F}_{\text{RF}} \in \mathcal{F}_{\text{RF}}, \\ & \mathbf{W}_{\text{RF}} \in \mathcal{W}_{\text{RF}}, \\ & \|\mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}, \end{aligned} \tag{8}$$

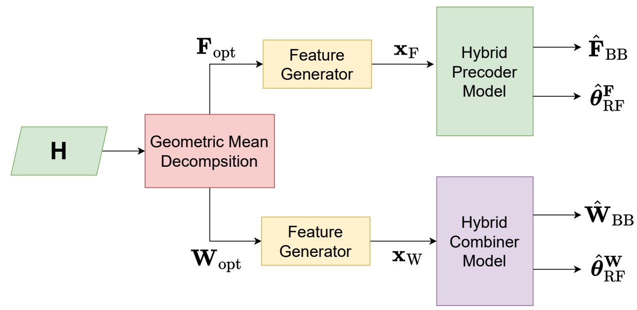

Fig. 2. The block diagram of the proposed hybrid precoding based on DL approach.

where $\mathcal{F}_{\text{RF}}$ and $\mathcal{W}_{\text{RF}}$ are the set of feasible analog precoders and combiners with constant amplitude, and $\|\mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}$ ensures the total transmitted power constraint. This optimization problem is non-convex and difficult to be solved, because it requires a joint optimization over four beamforming variable matrices ($\mathbf{F}_{\text{RF}}$, $\mathbf{F}_{\text{BB}}$, $\mathbf{W}_{\text{RF}}$ and $\mathbf{W}_{\text{BB}}$). Decoupling the joint design problem into hybrid precoding and combining problems can overcome the difficulty of the design and provide near-optimal hybrid precoding [5], where the objective of each problem is to minimize the Euclidean distance between the hybrid beamformer with an FD beamformer. The decoupled hybrid precoder design problem that can provide hybrid precoders ($\mathbf{F}_{\text{RF}}$, $\mathbf{F}_{\text{BB}}$) which approximately maximize R can be expressed as

$$\begin{aligned} \min_{\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}} \quad & \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}}\|_F, \\ \text{s.t.} \quad & \mathbf{F}_{\text{RF}} \in \mathcal{F}_{\text{RF}}, \\ & \|\mathbf{F}_{\text{RF}} \mathbf{F}_{\text{BB}}\|_F^2 = N_{\text{s}}, \end{aligned} \tag{9}$$

where $\mathbf{F}_{\text{opt}}$ is the optimal FD precoder.

D. Fully Digital Precoding Based on GMD

The channel matrix $\mathbf{H}$ can be decomposed using GMD into N_{s} sub-channels with equal SNR as [7]:

$$\mathbf{H} = \mathbf{Q}\mathbf{R}\mathbf{P}^H = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 & * \\ 0 & \mathbf{R}_2 \end{bmatrix} \begin{bmatrix} \mathbf{P}_1^H \\ \mathbf{P}_2^H \end{bmatrix}, \tag{10}$$

where $\mathbf{P}_1 \in \mathbb{C}^{N_{\text{t}} \times N_{\text{s}}}$ and $\mathbf{Q}_1 \in \mathbb{C}^{N_{\text{r}} \times N_{\text{s}}}$ are semi-unitary matrices (i.e., $\mathbf{Q}_1^* \mathbf{Q}_1 = \mathbf{P}_1^* \mathbf{P}_1 = \mathbf{I}_{N_s}$). However, $\mathbf{P}_1$ can be regarded as an optimal FD precoder such that $\mathbf{F}_{\text{opt}} = \mathbf{P}_1$, and $\mathbf{Q}_1$ as an optimal FD combiner such that $\mathbf{W}_{\text{opt}} = \mathbf{Q}_1$. In addition, $\mathbf{R}_1$ is an upper triangular matrix whose diagonal elements equal to the geometric mean of N_{s} positive singular values, and $*$ denotes an arbitrary matrix that can be neglected [18], [19].

III. Convolutional Neural Network-Based Hybrid Precoder and Combiner

This section presents the proposed hybrid precoding based on CNN approach as shown in Fig. 2. This figure shows the block diagram of the proposed hybrid precoding which composed of two models: hybrid precoder and combiner models, where all models have the same building blocks. As

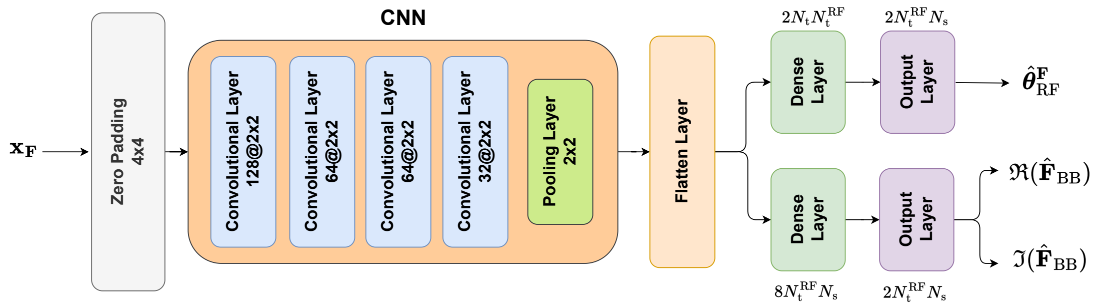

Fig. 3. The proposed hybrid precoder architecture.

shown in Fig. 2, the channel matrix $\mathbf{H}$ is firstly decomposed using GMD, and then $\mathbf{F}_{opt}$ and $\mathbf{W}_{opt}$ are fed to each hybrid precoder and hybrid combiner feature generator. The feature generator is a pre-process block that allows the model to extract more features of the input, which provides better training performance. The output of the feature generator regarded as the input raw data of the model.

The hybrid precoder (combiner) model has two output layers, where the first layer produces the phase angle of analog precoder (combiner) to simulate the working operation of phase shifters, while the second layer produces the normalized digital precoder (combiner). Each model is trained to minimize the decoupled joint optimization problem.

A. Feature Generator

Feature generator receives a complex-valued matrix such as the optimal FD precoder, $\mathbf{F}_{opt}$, and then it is converted to the real valued raw data by using the real, imaginary and the phase of the input matrix. For example, if the input matrix is $\mathbf{F}_{opt}$, the output vector $\mathbf{x_F} \in \mathbb{R}^{3N_tN_s\times 1}$ of the feature generator can be expressed as

$$\begin{aligned}\mathbf{x_F} = \Big[&\Re(\mathbf{F}_{opt})_{1,1}, \Re(\mathbf{F}_{opt})_{1,2}, \ldots, \Re(\mathbf{F}_{opt})_{N_t,N_s},\\ &\Im(\mathbf{F}_{opt})_{1,1}, \Im(\mathbf{F}_{opt})_{1,2}, \ldots, \Im(\mathbf{F}_{opt})_{N_t,N_s},\\ &\angle(\mathbf{F}_{opt})_{1,1}, \angle(\mathbf{F}_{opt})_{1,2}, \ldots, \angle(\mathbf{F}_{opt})_{N_t,N_s}\Big]^T.\end{aligned} \tag{11}$$

Similarly, the output vector is $\mathbf{x}_W \in \mathbb{R}^{3N_rN_s\times 1}$ when the input matrix is $\mathbf{W}_{opt}$.

B. Hybrid Precoder Model

The objective of the proposed CNN-based hybrid precoder model is to minimize the hybrid precoder design problem as stated in (9). Therefore, the proposed hybrid precoder model must be trained to minimize the Euclidean distance between the optimal FD precoder $\mathbf{F}_{opt}$ and hybrid precoders ($\mathbf{F}_{RF}$, $\mathbf{F}_{BB}$). Fig. 3 shows the hybrid precoder model architecture. As shown, the hybrid precoder model receives a feature vector $\mathbf{x}_F$ which contains a vectorized version of the real, imaginary and angular values of $\mathbf{F}_{opt}$. The first layer is zero-padding layer with 4×4 padding which adds zeros with size 4 around the input. Zero-padding layer is used to extract the features at the corner as well as the center. Four convolutional layers with different filter size and Leaky-ReLU activation functions are utilized to extract complex features of the input, the first layer has 128 filters with 2×2 filter size, the second and third layers have 64 filters with 2×2 size. The last convolutional layer has 32 filters with 2×2 size. In addition, pooling layer is applied to reduce the output size, and to speed the computation. Flatten layer is also required to make a connection between all pooling layer activation outputs and next dense layers neurons.

The output of the flatten layer is shared between two neural networks, the first neural network for analog precoder has a single dense layer with $2N_tN_t^{RF}$ neurons and Leaky-ReLU activation function, and output layer with $2N_t^{RF}N_s$ neurons which has been designed to meet the analog precoder constraint $(\mathbf{F}_{RF}^{(i)}\mathbf{F}_{RF}^{(i)H})_{i,j} = \frac{1}{N_t}$. Consequently, the estimated analog precoder with a constant amplitude $\hat{\mathbf{F}}_{RF}$ can be expressed as

$$\hat{\mathbf{F}}_{RF} = \frac{1}{\sqrt{N_t}} e^{j\hat{\boldsymbol{\theta}}^{\mathbf{F}}_{RF}}, \tag{12}$$

where $\hat{\boldsymbol{\theta}}^{\mathbf{F}}_{RF} \in \mathbb{R}^{N_t\times N_t^{RF}}$ is the phase angle of the estimated analog precoder, where each element must be between zero and 2π, i.e., $(\hat{\theta}^{\mathbf{F}}_{RF})_{i,j} \in [0, 2\pi]$. Therefore, the sigmoid function with 2π amplitude is used to limit the output in $[0,\ 2\pi]$ range. The second neural network for the digital precoder has single dense layer with $8N_t^{RF}N_s$ neurons and Leaky-ReLU activation function, and the output layer has $2N_t^{RF}N_s$ neurons that produces a vectorized real and imaginary components of the estimated digital precoder $\hat{\mathbf{F}}_{BB}$. The hybrid precoder must satisfy the total power constraint $\|\hat{\mathbf{F}}_{RF}\hat{\mathbf{F}}_{BB}\|_F^2 = N_s$. To meet this constraint, $\hat{\mathbf{F}}_{BB}$ should be normalized as

$$\hat{\mathbf{F}}_{BB} = \frac{\sqrt{N_s}}{\|\hat{\mathbf{F}}_{RF}\hat{\mathbf{F}}_{BB}\|_F}\hat{\mathbf{F}}_{BB}. \tag{13}$$

The model is trained with a defined loss function, that can be expressed as the hybrid precoder design problem as

$$L(\theta) = \|\mathbf{F}_{opt} - \hat{\mathbf{F}}_{RF}\hat{\mathbf{F}}_{BB}\|_F, \tag{14}$$

where θ denotes the parameters of hybrid precoder model, and the hybrid precoder matrix $\hat{\mathbf{F}}_{RF}\hat{\mathbf{F}}_{BB}$ should be a unitary matrix, where all columns are orthonormal vectors, such that

$$\hat{\mathbf{F}}_{BB}^H\hat{\mathbf{F}}_{RF}^H\hat{\mathbf{F}}_{RF}\hat{\mathbf{F}}_{BB} = \mathbf{I}_{N_s}. \tag{15}$$

Adding this constraint to loss function as penalty term can improve the model performance. Thus, the loss function can be rewritten as

$$L(\theta) = \|\mathbf{F}_{\text{opt}} - \hat{\mathbf{F}}_{\text{RF}}\hat{\mathbf{F}}_{\text{BB}}\|_F + \lambda_{\mathbf{F}}\|\hat{\mathbf{F}}_{\text{BB}}^H\hat{\mathbf{F}}_{\text{RF}}^H\hat{\mathbf{F}}_{\text{RF}}\hat{\mathbf{F}}_{\text{BB}} - \mathbf{I}_{N_s}\|_F, \tag{16}$$

where $\lambda_{\mathbf{F}}$ is a non-negative constant of the penalty term that used to satisfy $\hat{\mathbf{F}}_{\text{RF}}\hat{\mathbf{F}}_{\text{BB}}$ to be semi-unitary matrix with orthonormal vectors.

In order to minimize the lost function $L(\theta)$, Adam optimizer is used during the training process to update the model parameters in the direction of local minimum of $L(\theta)$. The layers parameters at the first iteration should be initialized at random points. In order to model a general problem as possible, each initialized parameters have different random distribution based on the layer activation function.

C. Hybrid Combiner Model

Hybrid combiner model is designed with the similar design objectives of hybrid precoder model, where it has the same building blocks: padding layer, four convolutional layers with single pooling layer, flatten layer, dense layers and output layers. The first output layer is designed to meet the analog combiner constraint $(\mathbf{W}_{\text{RF}}^{(i)}\mathbf{W}_{\text{RF}}^{(i)H})_{i,j} = \frac{1}{N_r}$. Consequently, the estimated analog combiner with a constant amplitude $\hat{\mathbf{W}}_{\text{RF}}$ can be expressed as

$$\hat{\mathbf{W}}_{\text{RF}} = \frac{1}{\sqrt{N_r}}e^{j\hat{\boldsymbol{\theta}}_{\text{RF}}^{\mathbf{W}}}, \tag{17}$$

where $\hat{\boldsymbol{\theta}}_{\text{RF}}^{\mathbf{W}} \in \mathbb{R}^{N_r \times N_r^{\text{RF}}}$ is the phase angle of the estimated analog combiner, where each element in $\hat{\theta}_{\text{RF}}^{\mathbf{W}}$ is limited in $[0,\, 2\pi]$ range using sigmoid function with 2π amplitude. The second output layer produces the estimated digital combiner $\hat{\mathbf{W}}_{\text{BB}}$.

The model is trained with a defined loss function, that can be expressed as the hybrid combiner design problem as

$$L(\theta) = \|\mathbf{W}_{\text{opt}} - \hat{\mathbf{W}}_{\text{RF}}\hat{\mathbf{W}}_{\text{BB}}\|_F + \lambda_{\mathbf{W}}\|\hat{\mathbf{W}}_{\text{BB}}^H\hat{\mathbf{W}}_{\text{RF}}^H\hat{\mathbf{W}}_{\text{RF}}\hat{\mathbf{W}}_{\text{BB}} - \mathbf{I}_{N_s}\|_F, \tag{18}$$

where $\lambda_{\mathbf{W}}$ is a non-negative constant of the penalty term that used to satisfy $\hat{\mathbf{W}}_{\text{RF}}\hat{\mathbf{W}}_{\text{BB}}$ to be semi-unitary matrix with orthonormal vectors. Adam optimizer is also used during the training process to update the model parameters in order to find a local minimum of $L(\theta)$.

IV. Simulation Results

In this section, simulation results are presented to illustrate the spectral efficiency of the proposed CNN-based hybrid precoding model compared to FD precoding based on GMD, PE and MO algorithms. Throughout the simulations, a massive MIMO with $N_t = 8 \times 8$ and $N_r = 4 \times 4$ UPA antennas with half-wave element spacing is considered and the channel is modeled as cluster environment with $N_c = 5$ clusters, each cluster with $N_{ray} = 10$ rays and 10° angle of spread.

The proposed model is constructed and processed using Keras Python package, and the training dataset is generated using the channel model expressed in (6) with 200,000 realizations. Adam optimizer is selected during the back propagation process with leaning rate of 0.00005. At each iteration, the model is trained with 64 mini batch size. Table I summarizes the training parameters.

TABLE I
A summary table of training parameters.

Parameter	Value
Optimizer	Adam (with β_1 = 0.9, β_2 = 0.999)
Learning rate (η)	0.00005
Leaky-ReLU constant (α)	0.2
$\lambda_{\mathbf{F}}$, $\lambda_{\mathbf{W}}$	0.15
Mini batch size	64
Size of dataset	200,000
Epochs	200

Fig. 4 shows the spectral efficiency performance comparison with N_t^{RF} = N_r^{RF} = 2 chains. It can be seen from Fig. 4 that the spectral efficiency of the proposed CNN-based hybrid precoding model, PE and MO algorithms are almost the same as the FD precoding in the case of N_s = 1, and are almost the same as PE and MO algorithms and within a small gap from the FD precoding in the case of N_s = 2 data streams. This reveals that the proposed hybrid precoder (combiner) model can more accurately approximate the optimal FD precoder (combiner).

To explore the performance of the proposed hybrid precoding model at larger input and output dimensions, Fig. 5 shows the spectral efficiency performance comparison with N_t^{RF} = N_r^{RF} = 4 chains. It can be noticed from Fig. 5 that the spectral efficiency performance of the proposed hybrid precoding model is almost the same as PE algorithm and within a small gap from MO algorithm and FD precoding in the case of N_s = 2, and within a small gap from the FD precoding, PE and MO algorithms in the case of N_s = 4 data streams.

In addition, the estimation time comparison of MO and PE algorithms with the proposed DL-based hybrid precoding model for N_t^{RF} = N_r^{RF} chains which equal to N_s = 2 and 4 multiplexed data streams and 1,000 realizations is shown in Fig. 6. It can be noticed from Fig. 6 that MO algorithm takes a long time to estimate the hybrid precoders and combiners compared with PE algorithm and the proposed hybrid precoding model, and the latter is almost 19 times faster than PE algorithm.

V. Conclusions

A novel CNN-based hybrid precoding model with GMD algorithm is proposed in this paper. The main purpose of implementing this proposed model in mmWave massive MIMO is that the conventional precoding methods are deployed numerically with high computational complexity. DL methods have been introduced to overcome this issue by generating the optimal hybrid precoders directly from a given optimal

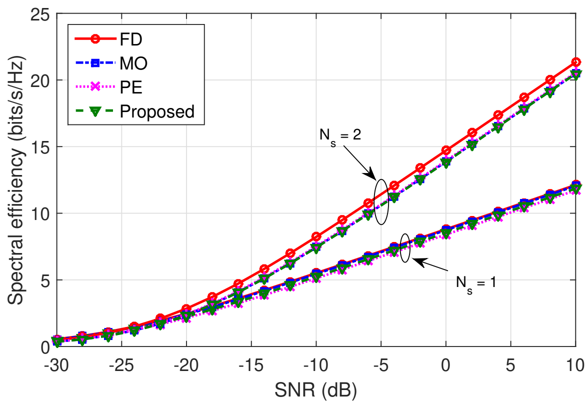

Fig. 4. Spectral efficiency performance versus SNR of different precoding algorithms with $N_t^{RF} = N_r^{RF} = 2$.

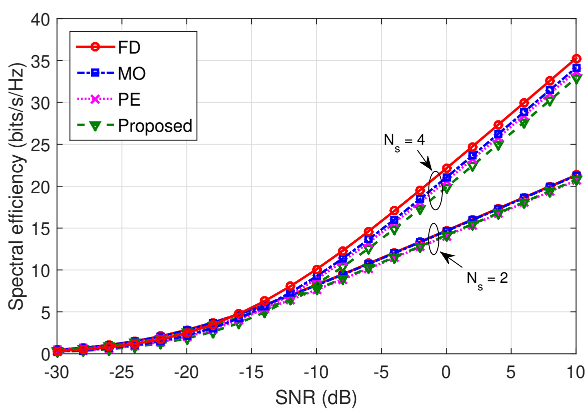

Fig. 5. Spectral efficiency performance versus SNR of different precoding algorithms with $N_t^{RF} = N_r^{RF} = 4$.

FD precoder. The precoder and combiner models trained with decoupled precoding and combining design problems, respectively, with addition term to satisfy the orthogonality in the estimated precoder. The spectral efficiency performance of the proposed DL-based hybrid precoding model is compared to FD precoder, PE and MO algorithms for different numbers of RF chains and multiplexed data streams. The results showed that the proposed hybrid precoding model can more accurately approximate the performance of FD precoding with lower time consumption than PE and MO algorithms. Our future studies will focus on integrating intelligent reflecting surfaces technology into the proposed system model.

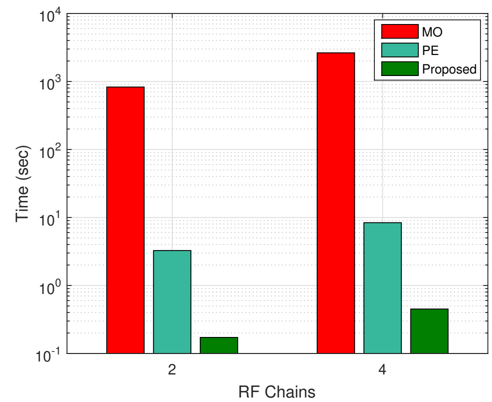

Fig. 6. Estimation time of different precoding algorithms with 1,000 realizations and $N_t^{RF} = N_r^{RF} = N_s = 2$ and 4.

References

[1] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," IEEE Commun. Mag., vol. 49, no. 6, pp. 101–107, 2011.

[2] G. R. MacCartney, M. K. Samimi, and T. S. Rappaport, "Exploiting directionality for millimeter-wave wireless system improvement," in 2015 IEEE Int. Conf. Commun. (ICC), London, UK, 2015, pp. 2416–2422.

[3] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, Fundamentals of Massive MIMO. Cambridge University Press, 2016.

[4] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," IEEE J. Sel. Topics Signal Process., vol. 10, no. 3, pp. 501–513, 2016.

[5] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R.W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," IEEE Trans. Wireless Commun., vol. 13, no. 3, pp. 1499–1513, 2014.

[6] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," IEEE J. Sel. Topics Signal Process., vol. 10, no. 3, pp. 485–500, 2016.

[7] Y. Jiang, W. W. Hager, and J. Li, "The geometric mean decomposition," Linear Algebra and its Applications, vol. 396, pp. 373-384, Feb. 2005.

[8] M. Liu, T. Song and G. Gui, "Deep cognitive perspective: Resource allocation for NOMA-based heterogeneous IoT with imperfect SIC," in IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2885-2894, April 2019.

[9] Y. Wang, M. Liu, J. Yang and G. Gui, "Data-driven deep learning for automatic modulation recognition in cognitive radios," in IEEE Trans. Veh. Technol., vol. 68, no. 4, pp. 4074-4077, April 2019.

[10] H. Huang, J. Yang, H. Huang, Y. Song and G. Gui, "Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system," in IEEE Trans. Veh. Technol., vol. 67, no. 9, pp. 8549-8560, Sept. 2018.

[11] H. Huang, Y. Song, J. Yang, G. Gui and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," in IEEE Trans. Veh. Technol., vol. 68, no. 3, pp. 3027-3032, March 2019.

[12] A. M. Elbir, "CNN-based precoder and combiner design in mmWave MIMO systems," in IEEE Commun. Lett., vol. 23, no. 7, pp. 1240-1243, July 2019.

[13] P. Dong, H. Zhang and G. Y. Li, "Framework on deep learning-based joint hybrid processing for mmWave massive MIMO systems," in IEEE Access, vol. 8, pp. 106023-106035, 2020.

[14] Q. Wang, K. Feng, X. Li and S. Jin, "PrecoderNet: Hybrid beamforming for millimeter wave systems with deep reinforcement learning," in IEEE Wireless Commun. Lett., vol. 9, no. 10, pp. 1677-1681, Oct. 2020.

[15] J. Chen et al., "Hybrid beamforming/combining for millimeter wave MIMO: A machine learning approach," in IEEE Trans. Veh. Technol., vol. 69, no. 10, pp. 11353-11368, Oct. 2020.

[16] H. Jiang, Y. Lu, X. Li, B. Wang, Y. Zhou and L. Dai, "Attention-based hybrid precoding for mmWave MIMO systems," 2021 IEEE Info. Theo. Workshop (ITW), Kanazawa, Japan, Oct. 2021, pp. 1-6.

[17] A. Goldsmith, S. A. Jafar, N. Jindal and S. Vishwanath, "Capacity limits of MIMO channels," in IEEE J. Sel. Areas Commun., vol. 21, no. 5, pp. 684-702, June 2003.

[18] T. Y. Elganimi and F. F. Alfitouri, "Norm-based user selection algorithm for single-RF space modulation techniques with geometric mean decomposition based precoding scheme," SN Appl. Sci., vol. 3, no, 1, pp. 1-13, 2021.

[19] T. Y. Elganimi and A. A. Aturki, "Joint user selection and GMD-based hybrid beamforming for generalized spatial modulation aided millimeter-wave massive MIMO systems," in the Proceedings of the 2020 IEEE 3rd Int. Conf. Info. Commun. Signal Process. (ICICSP), Shanghai, China, September 12–15, 2020, pp. 364–369.