Pruning the Pilots: Deep Learning-Based Pilot Design and Channel Estimation for MIMO-OFDM Systems

Mahdi Boloursaz Mashhadi and Deniz Gündüz
Dept. of Electrical and Electronic Eng., Imperial College London, UK
Email: {m.boloursaz-mashhadi, d.gunduz}@imperial.ac.uk

***Abstract*—With the large number of antennas and subcarriers the overhead due to pilot transmission for channel estimation can be prohibitive in wideband massive multiple-input multiple-output (MIMO) systems. This can degrade the overall spectral efficiency significantly, and as a result, curtail the potential benefits of massive MIMO. In this paper, we propose a neural network (NN)-based joint pilot design and downlink channel estimation scheme for frequency division duplex (FDD) MIMO orthogonal frequency division multiplex (OFDM) systems. The proposed NN architecture uses fully connected layers for frequency-aware pilot design, and outperforms linear minimum mean square error (LMMSE) estimation by exploiting inherent correlations in MIMO channel matrices utilizing convolutional NN layers. Our proposed NN architecture uses a non-local attention module to learn longer range correlations in the channel matrix to further improve the channel estimation performance.We also propose an effective pilot reduction technique by gradually pruning less significant neurons from the dense NN layers during training. This constitutes a novel application of NN pruning to reduce the pilot transmission overhead. Our pruning-based pilot reduction technique reduces the overhead by allocating pilots across subcarriers non-uniformly and exploiting the inter-frequency and inter-antenna correlations in the channel matrix efficiently through convolutional layers and attention module.**

I. Introduction

Massive multiple-input multiple-output (MIMO) systems are considered as the main enabler of 5G and future wireless networks thanks to their ability to serve a large number of users simultaneously, achieving impressive levels of spectral efficiency. A base station (BS) with a massive number of antennas relies on accurate downlink channel state information (CSI) to achieve the promised performance gains. Therefore, massive MIMO systems are more amenable to time division duplex (TDD) operation, which, thanks to the reciprocity of the uplink and downlink channels, does not require downlink channel estimation at the users. FDD operation is more desirable due its improved coverage and reduced interference; however, channel reciprocity does not hold in FDD. In FDD MIMO, the BS broadcasts downlink pilot signals, the users estimate the channel from the received pilots and transmit the CSI feedback to the BS. The resulting overhead becomes significant due to the large number of antennas and users; and hence, efficient pilot design and channel estimation are crucial to reduce the overhead.

In massive MIMO systems where the pilot length is typically much smaller than the number of antennas, channel estimation becomes severely underdetermined. Hence, simple least squares (LS) or linear minimum mean square error (LMMSE) channel estimation and orthogonal FFT pilots perform poorly. To estimate the channel more efficiently and reduce the pilot overhead, many previous works take a model-based estimation approach assuming sparse [1]–[5] or low-rank [6], [7] models on the channel matrix and utilize compressive sensing (CS)-based reconstruction techniques to estimate the channel or design improved pilot sequences. CS-based approaches rely on sparse or low-rank properties of the channel, and do not take into account the inherent statistical correlations and structures beyond sparse or low-rank patterns. Moreover, CS-based reconstruction techniques employ computationally demanding iterative algorithms, which imposes an additional burden on the users.

More recently, deep learning (DL)-based approaches have been used for massive MIMO CSI acquisition and showed significant improvements in comparison with their counterparts based on sparsity and compressive sensing (refer to [8], [9], and references therein). In these works, neural network (NN) architectures are trained over large CSI datasets to learn complex distributions, structures, and correlations, and exploit them for data-driven pilot design [10], channel estimation [11]–[15], compression [16]–[21] and feedback [22], [23]. Many of these works focus on a single task and propose a NN architecture to achieve optimized performance for it. While designing a single NN architecture to simultaneously handle all or several of these tasks is desirable for an end-to-end optimized performance, the resulting NN may be more complex, and require a longer training process. In this paper, we consider joint pilot design and channel estimation for downlink FDD massive MIMO systems.

In [11], [24], the authors proposed a convolutional neural network (CNN)-based structure for massive MIMO channel estimation. Their proposed architecture outperforms non-ideal LMMSE-based channel estimation (where the required covariance matrices are estimated from a coarse initial estimate of the channel at the receiver) and approach ideal LMMSE (with perfect knowledge of the covariance matrices assumed at the receiver). In [10], the authors use dense layers (which represent the pilots) followed by subsequent convolutional layers for joint pilot design and channel estimation. However, they design the same pilots for all subcarriers, which not only neglects frequency specific features in the CSI, but also results in a large PAPR, which is practically undesirable.

In this paper, we propose a NN-based scheme consisting of convolutional and dense layers for downlink pilot design and channel estimation in FDD massive MIMO-OFDM systems. Our proposed NN-based scheme exploits frequency-specific features in a data-driven manner for more efficient pilot design and channel estimation. The specific contributions of this paper can be summarized as follows:

- We propose a NN architecture which learns MIMO channel statistics over a dataset during training and exploits it to jointly design pilots and estimate the channel without requiring covariance matrices or other prior statistical assumptions on the channel matrix. By joint optimization of the pilot signals and the channel estimator, our proposed NN structure improves the normalized mean square error (NMSE) in comparison with [11], [24], which uses simple FFT pilots with NN-based channel estimation.
- Our proposed NN architecture uses dense layers to design frequency-aware pilot signals followed by convolutional layers to learn the inherent correlations in the MIMO-OFDM channel, and to exploit them for efficient channel estimation. We also use a non-local attention module, which enables the NN to learn and exploit longer range correlations in the channel matrix.

This work was supported by the European Research Council (ERC) through project BEACON (grant no 677854).

- Our proposed frequency-aware NN-based pilot design scheme outperforms the simple frequency-independent pilot scheme used in [10] by a large margin. This is due to the fact that our NN-based structure learns frequency-specific features of the channel matrices over the training dataset and exploits them to optimize pilots over different frequencies in a data-driven manner. Frequency-aware pilot design also enables us to allocate pilots non-uniformly over different subcarriers by our proposed pilot reduction technique based on NN pruning.
- We propose a pilot reduction technique based on NN pruning, which effectively reduces the pilot overhead by allocating pilots across subcarriers non-uniformly; fewer pilots are transmitted on subcarriers that can be satisfactorily reconstructed by the subsequent convolutional layers utilizing inter-frequency correlations. According to the simulation results, this scheme effectively improves the estimation NMSE for the same amount of time-frequency resources allocated uniformly over subcarriers for pilot transmission.
- We show by extensive simulations that our proposed scheme significantly outperforms the ideal LMMSE channel estimation as well as the recent works in [10], [11], [24]. We provide an ablation study to investigate the performance of NN-based pilot design and channel estimation separately.

The rest of this paper is organized as follows. In Section II, we present the system model. In Section III we present the proposed NN architecture for joint pilot design and channel estimation. In Section IV we present our pilot reduction scheme by NN pruning. Section V provides the simulation results, and Section VI concludes the paper.

II. System Model

We consider an FDD massive MIMO system, where a BS with N antennas serves a single-antenna user utilizing orthogonal frequency division multiplexing (OFDM) over M subcarriers. We denote the downlink channel by $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M] \in \mathbb{C}^{N\times M}$, where $\mathbf{h}_m \in \mathbb{C}^N$ is the channel gain vector over subcarrier m, for $m = 1, \ldots, M$. We assume that the BS is equipped with a uniform linear array (ULA) with response vector:

$$\mathbf{a}(\phi) = \frac{1}{\sqrt{N}}[1, e^{-j\frac{2\pi d}{\lambda}\sin\phi}, \cdots, e^{-j\frac{2\pi d}{\lambda}(N-1)\sin\phi}]^T,$$

where ϕ is the angle of departure (AoD), and d and λ denote the distance between adjacent antennas and carrier wavelength, respectively. The channel gain is a summation of multipath components [25] given by

$$\mathbf{h}_m = \sqrt{\frac{N}{P}}\sum_{p=1}^{P}\alpha_p e^{-j2\pi\tau_p f_s \frac{m}{M}}\mathbf{a}(\phi), \tag{1}$$

where P is the number of multipath components, f_s is the sampling rate, τ_p is the delay, and α_p is the propagation gain of the pth path. According to Eq. (1), entries of the channel matrix $\mathbf{H}$ are correlated for nearby sub-carriers and antennas due to similar propagation paths, gains, and AoDs/AoAs. There also exist inherent characteristics in MIMO environments due to specific user distributions, scattering parameters, geometry, etc., that cause common structures among MIMO channel matrices.

Fig. 1 depicts the time-frequency resource grid structure, where we use a pilot block of size $L \times M$ (denoted by the black slots in Fig. 1) to estimate the channel over a time-frequency grid of size $T \times M$ and $L \leq T$. The coherence time of the channel is assumed much larger than T such that the channel can be assumed constant with time over the whole grid. The signal received on the (i, j)'th grid location is given by $y_{ij} = \mathbf{x}_{ij}^T\mathbf{h}_j + n_{ij}$ at time slot i and subcarrier j, where $\mathbf{x}_{ij} \in \mathbb{C}^N$ denotes the vector of downlink transmitted signal and $\mathbf{h}_m \in \mathbb{C}^N$ is the channel vector over the j'th subcarrier.

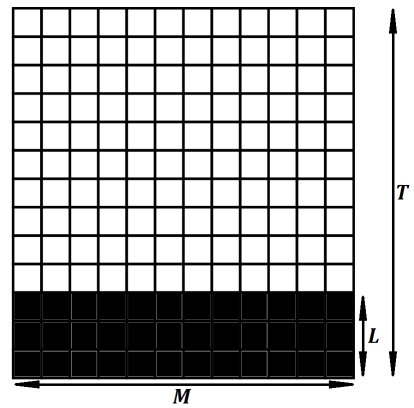

Fig. 1: The time-frequency resource grid structure.

Denoting downlink pilot signals transmitted by the BS over the mth subcarrier by $\mathbf{P}_m \in \mathbb{C}^{L\times N}$, where L is the pilot length, the received signal at the user is given by

$$\mathbf{y}_m = \mathbf{P}_m\mathbf{h}_m + \mathbf{n}_m, \tag{2}$$

where $\mathbf{n}_m \sim \mathcal{CN}(0, \sigma^2)$ is the complex additive Gaussian noise over the mth subcarrier, which is independent across subcarriers and time slots. A large pilot length L is infeasible not only because it increases the training overhead and computational complexity for channel estimation, but also because L should be much smaller than the channel coherence interval. Hence, in massive MIMO systems, where N is excessively large, the pilot length L is typically much smaller than the number of antennas N and Eq. (2) is severely underdetermined. In this work we propose a NN architecture to jointly design the optimum pilot signals $\mathbf{P}_1, \ldots, \mathbf{P}_m \in \mathbb{C}^{L\times N}$ and to estimate the downlink channel matrix $\mathbf{H}$ from the signals received over all subcarriers, i.e., $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M]^T \in \mathbb{C}^{M\times L}$.

Conventional channel estimation techniques are typically based on LMMSE estimation method, where the pilot signals are orthogonal discrete cosine transform (DFT) basis vectors or Zadoff-Chu sequences. The LMMSE channel estimate is given by $\hat{\mathbf{h}}_m^{MMSE} = \mathbb{E}[\mathbf{h}_m] + \mathbf{R}_{h_m y_m}\mathbf{R}_{y_m y_m}^{-1}(\mathbf{y}_m - \mathbb{E}[\mathbf{y}_m])$, where $\mathbb{E}[\mathbf{h}_m]$ and $\mathbb{E}[\mathbf{y}_m]$ are the corresponding expected values, and $\mathbf{R}_{h_m y_m}$ and $\mathbf{R}_{y_m y_m}$ are the covariance matrices over the m'th subcarrier. LMMSE is based on Gaussian channel assumption and requires the knowledge of the covariance matrices. In practical scenarios, the covariance matrices need to be empirically estimated over a dataset, which imposes additional computational load to the LMMSE technique. On the other hand, the simple choice of orthonormal DFT or Zadoff-Chu pilot signals regardless of the specific medium characteristics, e.g., scatterer positions, environment geometry, carrier frequencies, etc., is sub-optimal as will be discussed further in the next section. Our proposed approach utilizes a NN architecture for data-driven channel estimation and frequency-aware pilot design. The NN learns the required statistics over the dataset during training and leverages it for efficient channel estimation and pilot design; thereby reducing the pilot overhead.

III. NN-based Pilot Design and Channel Estimation

Fig. 2 depicts our proposed NN architecture for data-driven channel estimation and frequency-aware pilot design, where double channel inputs and outputs represent real and imaginary parts of the corresponding channel matrices. The network is composed of M fully connected branches, one for each subcarrier, followed by 3 convolutional layers, and is trained in an end-to-end fashion to minimize the mean square error (MSE). Each branch is composed of

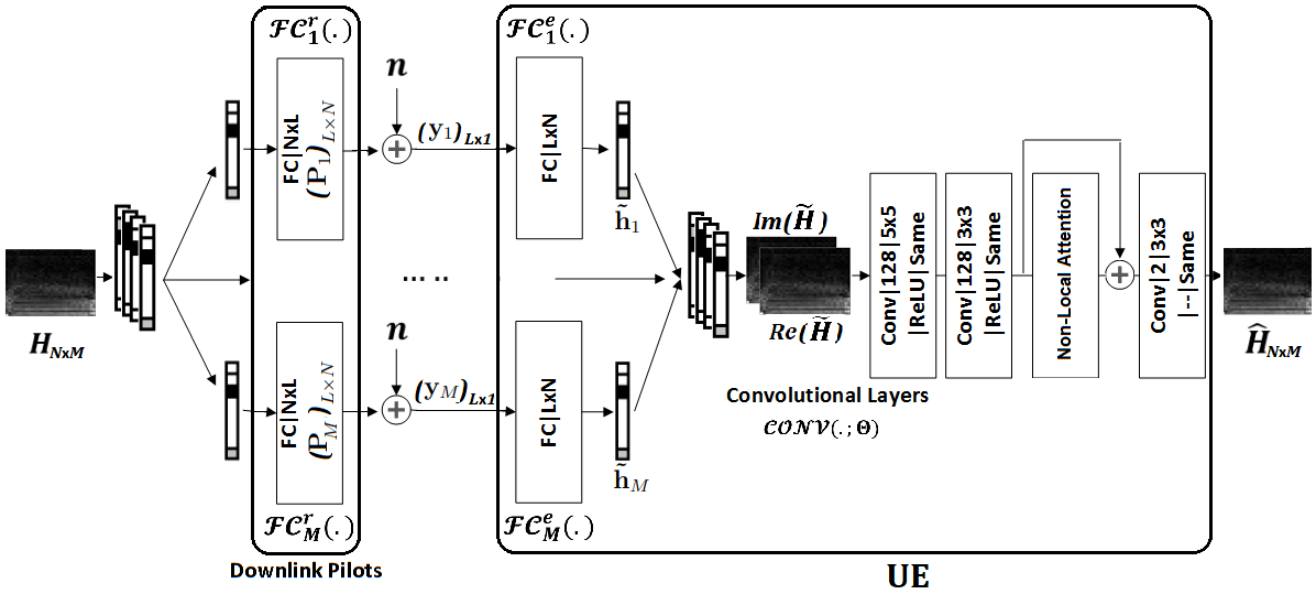

Fig. 2: Block diagram of the proposed scheme.

2 dense layers, a reduction and an expansion layer denoted by $\mathcal{FC}^r$ and $\mathcal{FC}^e$, respectively.

The input CSI matrix $\mathbf{H}$ is first divided into its subcarrier components $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M$, where $\mathbf{h}_m$ is input to the m'th fully connected branch. The reduction layer in each branch models downlink pilot transmission over that subcarrier, which is followed by downlink additive Gaussian noise according to $\mathbf{y}_m = \mathcal{FC}_m^r(\mathbf{h}_m; \mathbf{P}_m) + \mathbf{n}_m = \mathbf{P}_m\mathbf{h}_m + \mathbf{n}_m$. Hence, the weight parameters trained for $\mathcal{FC}_m^r$ correspond to the pilots to be transmitted over the m'th subcarrier, denoted by $\mathbf{P}_m^*$. Note that the dense layers $\mathcal{FC}_1^r, \ldots, \mathcal{FC}_M^r$ are used without bias or activations here, as they imitate the actual pilot transmission process, which is a simple matrix multiplication. A total power constraint on the pilot block is enforced by normalization, i.e. $\sum_{m=1}^{M} \|\mathbf{P}_m\|^2 = 1$. Here, we implement complex-valued fully connected layers, i.e., we use

$$\begin{pmatrix} \Re(\mathbf{y}_m) \\ \Im(\mathbf{y}_m) \end{pmatrix} = \begin{pmatrix} \Re(\mathbf{P}_m) & -\Im(\mathbf{P}_m) \\ \Im(\mathbf{P}_m) & \Re(\mathbf{P}_m) \end{pmatrix} \begin{pmatrix} \Re(\mathbf{h}_m) \\ \Im(\mathbf{h}_m) \end{pmatrix} + \begin{pmatrix} \Re(\mathbf{n}_m) \\ \Im(\mathbf{n}_m) \end{pmatrix}, \quad (3)$$

where $\Re(\mathbf{P}_m)$ and $\Im(\mathbf{P}_m)$ represent real and imaginary parts of the pilots, respectively, and are trainable NN parameters. Hence, each fully connected layer has $2LN$ real-valued trainable parameters.

The expansion layer in the m'th branch gives a coarse initial estimate of the channel at the m'th subcarrier according to $\tilde{\mathbf{h}}_m = \mathcal{FC}_m^e(\mathbf{y}_m; \mathbf{Q}_m, \mathbf{b}_m) = \mathbf{Q}_m\mathbf{y}_m + \mathbf{b}_m$, where the expansion layer is also implemented as a complex-valued fully connected layer, similarly to (3).This initial estimate imitates the familiar LMMSE estimate $\hat{\mathbf{h}}_m^{MMSE} = \mathbb{E}[\mathbf{h}_m] + \mathbf{R}_{h_m y_m}\mathbf{R}_{y_m y_m}^{-1}(\mathbf{y}_m - \mathbb{E}[\mathbf{y}_m])$, which is the best linear estimator that minimizes $\mathbb{E}\|\mathbf{h}_m - \hat{\mathbf{h}}_m\|_2^2$, and the NN learns the linear weights $\mathbf{Q}_m$ and $\mathbf{b}_m$ to minimize the empirical average MSE given by $1/K\sum_{k=1}^{K}\|\mathbf{h}_m^{(k)} - \hat{\mathbf{h}}_m^{(k)}\|_2^2$, where K is the size of the dataset.

We subsequently apply convolutional layers on these initial estimates to further improve the estimation accuracy by exploiting the inherent correlations within the channel matrix across subcarriers as well as antennas. This is motivated by our previous observations showing strong local correlations in MIMO channel matrices, which also explains the significant success of convolutional NNs in efficient MIMO CSI reduction and feedback in previous works [20], [23], [26]. To this end, we concatenate initial estimates from the dense branches to get $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_M]$, which is then input to the convolutional layers, as shown in Fig. 2. The convolutional layers implement real-valued convolution kernels and treat the complex output of the expansion layers as a two-dimensional input by concatenating its real and imaginary parts along the channel dimension.

In Fig. 2, "Conv|128| 5×5|ReLU|Same" represents a convolutional layer with 128 kernels of size 5×5 with rectified linear (ReLU) activation and "Same" padding technique. "$|--|$" means that the last layer has no activation. The final estimate is given by $\hat{\mathbf{H}} = \mathcal{CONV}(\tilde{\mathbf{H}}; \Theta)$, where $\mathcal{CONV}$ and Θ denote the convolutional layers and their corresponding set of parameters.

The convolutional layers learn and exploit local correlations to improve the estimation accuracy. As we will see in section V, local correlations account for a significant improvement in the reconstruction NMSE in comparison with simple LMMSE. However, we see that there still exists long range correlations in the channel matrix. In order to exploit long range correlations in an efficient manner, we have added a non-local attention module [27], [28]. We later show through simulations that the non-local attention module further improves the NMSE, specifically when fewer pilots are used and at lower downlink SNRs.

Eq. (4) provides the general input-output relation for non-local attention:

$$\mathbf{g}_i = \frac{1}{\pi(\mathbf{f})}\sum_{\forall j}\phi(\mathbf{f}_i, \mathbf{f}_j)\psi(\mathbf{f}_j), \quad (4)$$

in which i is the spacetime index of an output position whose response is to be computed, j is the index that enumerates all possible positions, $\mathbf{f}$ is the input, and $\mathbf{g}$ is the output of the same size as $\mathbf{f}$. A pairwise function ϕ computes a scalar representing the relationship between i and all j. The unary function ψ computes a representation of the input signal at position j. The response is normalized by a factor $\pi(\mathbf{f})$. We tried various popular choices for $\pi(\cdot), \phi(\cdot)$ and $\psi(\cdot)$ functions (refer to [28] for further details), and we found that the best performing ones in our problem are $\phi(\mathbf{f}_i, \mathbf{f}_j) = \exp([\mathbf{W}_{\phi_1}\mathbf{f}_i]^T[\mathbf{W}_{\phi_2}\mathbf{f}_j])$, $\psi(\mathbf{f}_j) = \mathbf{W}_\psi\mathbf{f}_j$, and $\pi(\mathbf{f}) = \sum_{\forall j}\phi(\mathbf{f}_i, \mathbf{f}_j)$, where $\mathbf{W}$s are trainable weight matrices. Fig. 3 provides the block diagram for our non-local attention module,

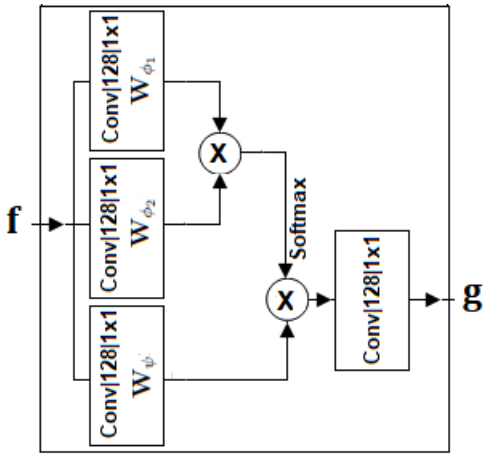

Fig. 3: Block diagram of the non-local attention module.

where 1×1 convolutions implement $\mathbf{W}$ weight multiplications, and "Softmax" activation implements the exponential function. Refer to [28] for more details on attention modules.

Finally, the network is trained to minimize the end-to-end empirical MSE cost given by

$$\mathcal{C}(\mathbf{H}; \{\mathbf{P}_i\}_{i=1}^{M}, \{\mathbf{Q}_i\}_{i=1}^{M}, \Theta) = \frac{1}{K}\sum_{k=1}^{K} MSE(\mathbf{H}^{(k)}, \hat{\mathbf{H}}^{(k)}) \tag{5}$$
$$= \frac{1}{K}\sum_{k=1}^{K} \|\mathbf{H}^{(k)} - \mathcal{CONV}(\tilde{\mathbf{H}}^{(k)}; \Theta)\|_2^2,$$

where $H^(k)$ denotes the k'th sample channel in the training dataset and K is the dataset size. Minimizing (5) end-to-end, we obtain all the network parameters including the pilot signals $\mathbf{P}_1, \ldots, \mathbf{P}_M$.

Remark 1: In comparison with the simple LMMSE estimator, our proposed NN-based approach is data-driven; in the sense that, it does not make any prior assumption on the channel statistics, but rather learns it from the dataset during training. Although LMMSE is optimum when channel and noise distributions are jointly Gaussian, the general MMSE estimate $\hat{\mathbf{H}} = \mathbb{E}[\mathbf{H}|\mathbf{Y}]$ has no closed-form expression for arbitrary channel distributions. Our proposed NN-based channel estimator is trained to minimize the empirical MSE, and hence, gives a tractable approximation of the general MMSE estimate regardless of the channel distribution. Moreover, our proposed NN architecture uses convolutional layers to exploit inter-frequency and inter-antenna correlations to improve the estimation performance. We later show through simulations that the proposed NN-based approach outperforms LMMSE channel estimation. The improvement is more significant for shorter pilot lengths and lower SNR values.

Remark 2: There has been a prior line of research to analytically design optimum pilot signals for downlink FDD MIMO systems [29]–[31]. When the channel distribution is assumed to be Gaussian, the resulting MSE of the LMMSE estimator has a closed-form expression. The authors in [30] use steepest descent optimization to design pilots that minimize the resulting MSE. This approach again relies on the Gaussian channel assumption. If this assumption is slightly generalized to a Gaussian mixture distribution, then closed-form expressions are available only for upper and lower bounds on the MSE [31]. The authors in [31] use steepest descent to design pilots that maximize the mutual information between the received noisy pilots and the channel, i.e., $I(\mathbf{Y}; \mathbf{H})$. Our NN approach, however, jointly designs channel estimator and pilot signals to minimize the end-to-end estimation error while avoiding computational difficulties.

Remark 3: The authors in [11], [24] utilize a CNN-based architecture for MIMO channel estimation, but they use simple FFT pilot signals. In this paper, we add the fully connected reduction layers to the CNN architecture to jointly optimize pilots and estimate the channel. We show later in the ablation study that the designed pilot signals significantly improve the performance. They also use the CNN structure to improve the NMSE of an initial LS estimate. We observed in simulations that replacing the initial LS estimate with the fully connected expansion layers lets the network to automatically learn the initial estimate and improves the end-to-end reconstruction MSE. Although the approach proposed in [24] and [11] outperforms non-ideal LMMSE (where the covariance matrices are estimated from a coarse initial LS estimate at the user), it still performs worse than the ideal LMMSE. We later show through simulations that our proposed NN-based approach outperforms ideal LMMSE (where the covariance matrices are estimated over the dataset). The improvement is more significant for shorter pilot lengths and lower SNR values.

Remark 4: The authors in [10] use a similar NN structure, but they utilize the same pilot signals over all the subcarriers. However, we observe different channel statistics over different subcarriers, which suggests that using the same pilots over all the subcarriers is sub-optimal. In this work, we use a frequency-aware pilot design approach, which utilizes different fully connected branches over different subcarriers to design pilots. We later show through simulations that our frequency-aware approach improves the MSE by a considerable margin. It also allows us to non-uniformly allocate pilots to different subcarriers by our proposed NN pruning technique, presented in the next section, which further reduces the MSE with the same pilot overhead.

IV. Pilot Allocation by NN pruning

In FDD operation mode of massive MIMO systems, downlink pilot transmission consumes a significant amount of radio resources, which in turn results in a significant loss of spectral efficiency. While the NN architecture proposed in the previous section significantly improves the end-to-end channel estimation performance by exploiting inherent correlations in the channel matrix using convolutional layers and designing pilot signals in a frequency-aware manner using the fully connected branches, our goal in this section is to further reduce the pilot overhead. To achieve this, we propose an efficient pilot allocation technique by pruning the least significant neurons from the fully connected layers in the proposed NN architecture. The reduction layer in the m'th fully connected branch consists of L neurons each of which corresponds to a single pilot transmission according to $\tilde{\mathbf{y}}_m = \mathcal{FC}_m^r(\mathbf{h}_m; \mathbf{P}_m) = \mathbf{P}_m\mathbf{h}_m$, and occupies one time-frequency resource over the downlink channel. We denote the pilot matrix by $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \ldots, \tilde{\mathbf{y}}_M]$, which represents a total of $L \times M$ time-frequency resources allocated to downlink pilot transmission during each coherence interval of the channel. Our idea is to gradually prune least sigificant neurons from the fully connected reduction layers $\mathcal{FC}_1^r, \ldots, \mathcal{FC}_M^r$ during training to reduce the pilot overhead by saving the corresponding time-frequency resources for data transmission, while causing the least possible degradation to the reconstruction MSE. This approach enables non-uniform pilot allocation and the transmission of fewer number of pilots over subcarriers that can be satisfactorily reconstructed by the subsequent convolutional layers utilizing inter-frequency correlations.

Consequently, we formulate the design and optimization of the pilots as a NN pruning problem, where the goal is to simplify a large NN by pruning some weights/neurons from it without significantly degrading the performance [32], [33]. We highlight here that pruning of NNs is typically employed to reduce the computation and memory complexity of NNs, where as in our case, pruning the fully-connected

TABLE I: Comparison of the NMSE (dB) values achieved by the proposed NN-based channel estimation scheme and the LMMSE technique for the indoor scenario ($N = 32, M = 256$).

		$\mathrm{SNR} = -5\mathrm{dB}$	$\mathrm{SNR} = 0\mathrm{dB}$	$\mathrm{SNR} = 5\mathrm{dB}$	$\mathrm{SNR} = 10\mathrm{dB}$	$\mathrm{SNR} = 15\mathrm{dB}$	$\mathrm{SNR} = 20\mathrm{dB}$
$L = 16$	NN+Attention	-14.32	-17.61	-20.58	-22.88	-23.65	-23.93
	NN	-13.40	-16.87	-19.95	-22.46	-23.40	-23.75
	FC Subnet	-8.08	-11.39	-15.00	-18.21	-21.45	-23.03
	LMMSE	-8.10	-11.47	-15.00	-18.27	-21.49	-23.06
$L = 12$	NN+Attention	-11.33	-13.61	-14.83	-15.35	-16.22	-16.79
	NN	-8.39	-11.53	-13.35	-14.41	-15.52	-16.40
	FC Subnet	-4.75	-8.00	-11.07	-13.10	-13.94	-14.80
	LMMSE	-4.82	-8.08	-11.17	-13.15	-13.96	-14.82
$L = 8$	NN+Attention	-13.23	-13.96	-14.41	-14.66	-14.82	-15.04
	NN	-7.99	-11.03	-12.10	-12.87	-13.55	-14.05
	FC Subnet	-3.75	-6.98	-8.95	-10.11	-11.09	-11.80
	LMMSE	-3.87	-7.05	-9.01	-10.15	-11.11	-11.82

reduction layers translates into reducing the pilot signals. Different works have heuristically proposed various saliency metrics to decide which connections/neurons can be pruned from a large NN without significantly degrading its performance. Saliency metrics heuristically approximate the relative importance of different weights or sets of weights in the NN. One of the most widespread used metrics is the l_1-norm of the weights [34], which leads to the magnitude-based pruning scheme [32]. This is motivated by the heuristic that the larger the weight value, the more likely the weight is to be contributing to the result. Other saliency metrics include average ratio of zero activations [35], the Fisher information [36], and 1st order Taylor expansions [37], or a composition of various metrics [38].

Inspired by [32], we take an l_1-regularized magnitude-based pruning approach by introducing two modifications to the training process. Note that the (i, j)'th pilot signal is given by

$$[\tilde{\mathbf{Y}}]_{ij} = [\tilde{\mathbf{y}}_j]_i = [\mathcal{FC}_j^r(\mathbf{h}_j; \mathbf{P}_j)]_i = [\mathbf{P}_j \mathbf{h}_j]_i = \sum_{n=1}^{N} [\mathbf{P}_j]_{in} [\mathbf{h}_j]_n, \quad (6)$$

where $[.]_{ij}$ represents the (i, j)'th element of a matrix. We define matrix $\mathbf{\Phi}$, where $[\mathbf{\Phi}]_{ij} = \sum_{n=1}^{N} \|[\mathbf{P}_j]_{in}\|^2$. Our idea is to prune neurons with the smallest sum squared weight connections (i.e., the smallest $[\Phi]_{ij}$) from the network. With this in mind, we make the following modifications to the training process:

Sparsity promoting regularization: We add a second term to our cost function to push $[\mathbf{\Phi}]_{ij}$ elements towards zero. In particular, we add the l_1-norm of the $\mathbf{\Phi}$ matrix to the cost function with a regularization parameter λ, which controls the trade-off between sparsity and the MSE according to

$$\begin{aligned} \mathcal{C}_{prun}(\mathbf{H}; \{\mathbf{P}_i\}_{i=1}^M, \{\mathbf{Q}_i\}_{i=1}^M, \Theta) &= \frac{1}{K} \sum_{k=1}^{K} MSE(\mathbf{H}^{(k)}, \hat{\mathbf{H}}^{(k)}) + \lambda \|\mathbf{\Phi}\|_1 \\ &= \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{H}^{(k)} - \hat{\mathbf{H}}^{(k)}\|_2^2 \\ &+ \lambda \sum_i \sum_j |\sum_{n=1}^{N} \|[\mathbf{P}_j]_{in}\|^2|, \end{aligned} \quad (7)$$

where λ is determined through numerical search until a desired sparsity level is achieved. Denoting our target sparsity level by S, which represents the percentage of the pruned neurons, a larger λ is required for larger S values.

Magnitude-based pruning: Define a pilot allocation mask $\mathbf{M} \in [0, 1]^{L \times M}$, in which the zeros represent the pruned pilots. Pruning is performed by element-wise multiplication of the pruning mask with the received pilot matrix during each optimization step, i.e., $\mathbf{Y} = \mathbf{M} \odot \tilde{\mathbf{Y}} + \mathbf{N}$. We initialize $\mathbf{M}$ to an all-one state and gradually update it during training according to a pruning schedule to achieve the final allocation mask $\mathbf{M}^*$. Our target sparsity S represents the ratio of zeros in the final mask $\mathbf{M}^*$ over its size, $L \times M$. As our regularization term gradually pushes more and more of the $[\mathbf{\Phi}]_{ij}$ values towards zero along the training steps, $[\mathbf{M}]_{ij}$ values corresponding to the smallest $[\mathbf{\Phi}]_{ij}$'s can be zeroed out. We use a linear pruning schedule where we apply 10 balanced updates on $\mathbf{M}$ during training, each of which prunes another $S/10$ of the pilots. It is important to train for a sufficient number of steps after each update of the pruning mask to allow the network to converge to the new optimum before the next pruning update.

Although the choice of the saliency metric has been mostly heuristic in the literature, we do have some intuition to support our magnitude-based approach for the specific channel estimation task. Utilizing Eq. (8), the pilot power received on the (i, j)'th resource grid is bounded by

$$\begin{aligned} \|[\tilde{\mathbf{Y}}]_{ij}\|^2 = \| \sum_{n=1}^{N} [\mathbf{P}_j]_{in} [\mathbf{h}_j]_n \|^2 &\leq (\sum_{n=1}^{N} \|[\mathbf{P}_j]_{in}\|^2)(\sum_{n=1}^{N} \|[\mathbf{h}_j]_n\|^2) \\ &= [\Phi]_{ij} (\sum_{n=1}^{N} \|[\mathbf{h}_j]_n\|^2). \end{aligned} \quad (8)$$

Denoting the received pilot SNR on the (i, j)'th resource grid by ρ_{ij}, we get $\rho_{ij} = \|[\tilde{\mathbf{Y}}]_{ij}\|^2 / \sigma_n^2 \leq [\Phi]_{ij} (\sum_{n=1}^{N} \|[\mathbf{h}_j]_n\|^2 / \sigma_n^2)$. Hence, the pilots located on grid locations with the smallest sum squared weight connections (i.e. the smallest $[\Phi]_{ij}$) are received at lower SNRs and hence contribute the least to the NMSE performance achieved by the NN.

V. Simulation Results

We have generated the datasets for training and testing using the COST 2100 channel model [39], which is a geometry-based stochastic channel model that reproduce the stochastic properties of MIMO channels over time, frequency and space. COST 2100 is a cluster-level model, and generates channel realizations that hold statistical consistency of the large scale channel parameters. We have considered an indoor picocellular scenario at 5.3 GHz and an outdoor rural scenario at 330 MHz band. The BS is equipped with a ULA of dipole antennas positioned at the center of a $20\mathrm{m} \times 20\mathrm{m}$ and $400\mathrm{m} \times 400\mathrm{m}$ square area for the indoor and outdoor scenarios, respectively. Note that we have presented the results for the outdoor scenario in Subsection V.B, but as the simulations revealed very similar results and conclusions for both the indoor and outdoor scenarios, we have included the results only for the indoor scenario

TABLE II: Comparison of the NMSE (dB) values achieved by the proposed NN-based channel estimation scheme and the LMMSE technique for the outdoor scenario ($N = 32, M = 256$).

		SNR $= -5$dB	SNR $= 0$dB	SNR $= 5$dB	SNR $= 10$dB	SNR $= 15$dB	SNR $= 20$dB
$L = 16$	NN+Attention	-16.84	-18.82	-20.37	-20.83	-21.34	-21.55
	NN	-15.31	-17.89	-19.68	-20.36	-20.84	-21.14
	LMMSE	-7.82	-11.92	-15.75	-18.91	-19.67	-20.47
$L = 12$	NN+Attention	-10.04	-11.05	-11.83	-12.64	-13.25	-13.82
	NN	-8.37	-9.99	-11.41	-12.25	-12.98	-13.61
	LMMSE	-3.76	-7.20	-10.33	-11.38	-12.15	-12.89
$L = 8$	NN+Attention	-7.37	-7.76	-8.06	-8.37	-8.57	-8.76
	NN	-4.91	-6.29	-7.34	-7.68	-7.92	-8.07
	LMMSE	-1.88	-3.51	-4.96	-6.39	-6.67	-6.89

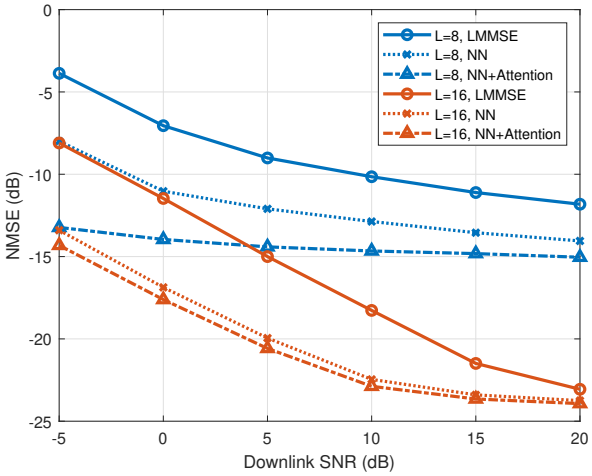

Fig. 4: NMSE (dB) vs. downlink SNR for the indoor scenario, $N = 32, M = 256$.

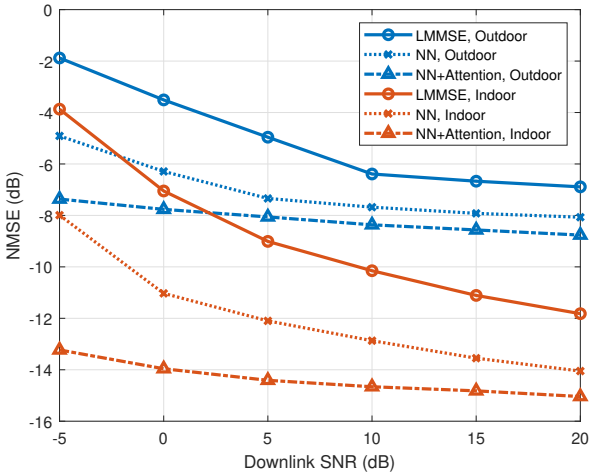

Fig. 5: NMSE (dB) comparisons for the indoor vs. outdoor scenarios, $L = 8, N = 32, M = 256$.

in the subsequent subsections to avoid tedious discussions of similar results.

The user is placed within the square area uniformly at random. All other parameters follow the default settings in [39]. Train and test datasets include 80000 and 20000 channel realizations, respectively, with $M = 256$ and $N = 32$. We train all NNs up to 110000 steps with a batch size of 100 utilizing the Adam optimizer [40]. We use the normalized MSE (NMSE) as the performance measure, defined by

$$\text{NMSE} \triangleq \frac{\mathbb{E}\{\|\mathbf{H} - \hat{\mathbf{H}}\|_2^2\}}{\mathbb{E}\{\|\mathbf{H}\|_2^2\}}. \tag{9}$$

A. NN vs. LMMSE performance

Table I and Fig. 4 compare the NMSE values achieved by the proposed NN-based channel estimation technique with those achieved by LMMSE for the indoor scenario at different pilot length and SNR values. We note that we have used the pilot signals optimized by the fully connected reduction layers for the LMMSE approach here. For LMMSE, we first estimate the channel covariance matrix over the training dataset utilizing the maximum likelihood (ML) covariance estimation for Gaussian vectors [41], and then use it to get the LMMSE channel estimate. Note that we have assumed that the user has perfect knowledge of these estimated covariance matrices. In practice, the user may need to estimate the covariance matrix from a coarse initial LS estimate of the channel [11], which degrades the LMMSE performance. Hence, the curves in Fig. 4 provide a lower bound on the performance of LMMSE as we have used covariance matrices estimated over the original training dataset. Table I also presents the contribution of each subnetwork of the proposed architecture towards the final NMSE value. In this table, we report the NMSE values for NN-based channel estimation with and without the attention module, denoted by "NN+Attention" and "NN", respectively. "FC subnet" corresponds to the NMSE obtained when only the fully connected layers are used for NMSE evaluations.

According to Fig. 4, the NN-based approach outperforms LMMSE for all SNR and pilot lengths. The proposed NN-based approach even without the attention module outperforms LMMSE, and the attention module further improves the performance specifically at low SNR regimes. For example, at a downlink SNR=-5dB, the "NN+Attention" scheme improves the reconstruction NMSE by 9.36dB , 6.51dB, and 6.22dB in comparison with LMMSE for $L = 8, 12$ and 16, respectively. For a target NMSE of ~ -15dB at a downlink SNR of 5dB, our proposed NN+Attention approach reduces the required pilot length from 16 to 8 in comparison with LMMSE, resulting in a 50% reduction in the pilot overhead.

Finally, comparing the "LMMSE" and "FC Subnet" results in Table I shows that the FC subnetwork achieves almost the same NMSE performance as the LMMSE estimate. We have also observed that the parameters learned by the FC layers, are very close to the LMMSE estimator, i.e., $\mathbf{Q}_m \sim \mathbf{R}_{h_m y_m}\mathbf{R}_{y_m y_m}^{-1}$ and $\mathbf{b}_m \sim \mathbb{E}[\mathbf{h}_m] - \mathbf{R}_{h_m y_m}\mathbf{R}_{y_m y_m}^{-1}\mathbb{E}[\mathbf{y}_m]$, which shows that the initial estimate at the output of the FC expansion layers (i.e., $\tilde{\mathbf{H}}$) imitates the LMMSE estimate.

TABLE III: Comparison of the NMSE (dB) achieved using same pilots (SP) over all subcarriers [10] and the proposed frequency-aware pilot design scheme (DP) for the indoor scenario at $N = 32, M = 256$.

		SNR $= -5$dB	SNR $= 0$dB	SNR $= 5$dB	SNR $= 10$dB	SNR $= 15$dB	SNR $= 20$dB
$L = 16$	DP+Attention	-14.32	-17.61	-20.58	-22.88	-23.65	-23.93
	DP	-13.40	-16.87	-19.95	-22.46	-23.40	-23.75
	SP	-11.84	-15.07	-18.08	-20.58	-22.34	-23.25
$L = 12$	DP+Attention	-11.33	-13.61	-14.83	-15.35	-16.22	-16.79
	DP	-8.39	-11.53	-13.35	-14.41	-15.52	-16.40
	SP	-7.92	-10.92	-11.74	-12.91	-13.30	-13.94
$L = 8$	DP+Attention	-13.23	-13.96	-14.41	-14.66	-14.82	-15.04
	DP	-7.99	-11.03	-12.10	-12.87	-13.55	-14.05
	SP	-7.74	-9.38	-10.69	-12.12	-12.99	-13.42

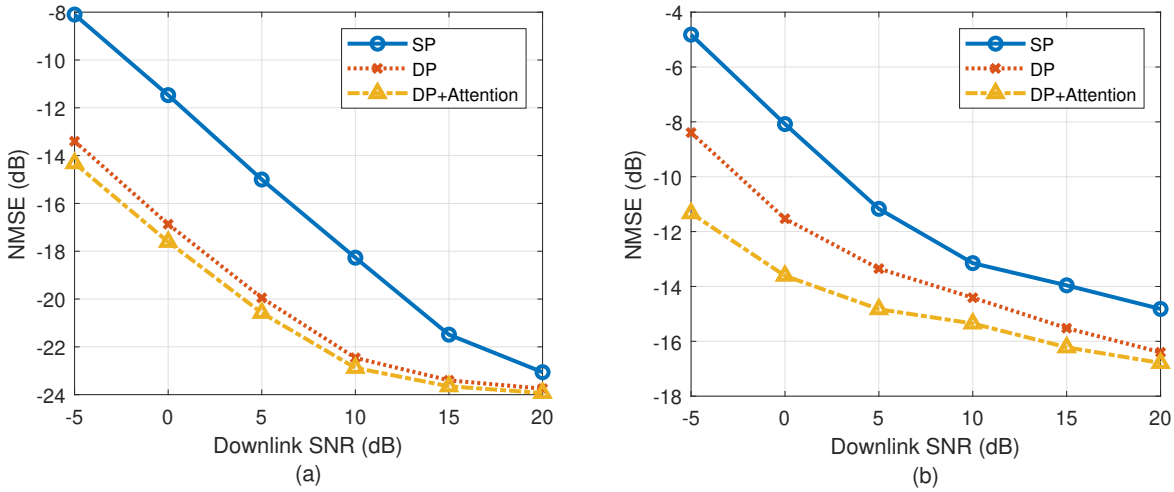

Fig. 6: Comparison between the NMSE (dB) values achieved using same pilots (SP) over all subcarriers [10] and the proposed frequency-aware pilot design scheme (DP) for $N = 32, M = 256$, (a) $L = 16$, (b) $L = 12$.

B. Impact of the scattering environment

Table II provides the reconstruction NMSEs for the outdoor scenario at different pilot length and SNR values. Note that we have used the pilot signals optimized by the fully connected reduction layers for the LMMSE approach here. Fig. 5 compares the NMSE values for the indoor and outdoor scenarios when $L = 8, N = 32, M = 256$. This figure shows a more significant performance improvement achieved by the NN-based approach for the outdoor scenario in comparison with LMMSE. According to Tables I and II, the "NN+Attention" approach improves the reconstruction NMSE by 5.49dB and 9.36dB in comparison with LMMSE for the outdoor and indoor scenarios, respectively at $L = 8, N = 32, M = 256$.

C. Frequency-aware pilot design

Unlike in [10], which optimizes the same pilots over all subcarriers, our proposed NN architecture utilizes M parallel fully connected branches to design different pilots over the subcarriers. This helps to improve the NMSE performance by allowing the NN to learn and exploit specific subcarrier statistics and structures in a frequency-aware manner. Table III provides a comparison between our proposed frequency-aware pilot design scheme with the approach proposed in [10], where the same pilots are used over different subcarriers. The acronyms "SP" and "DP" in this table represent "same pilots" and "different pilots" approaches, respectively.

According to Table III, the proposed frequency-aware approach improves the performance significantly even without the attention module, e.g., for a target NMSE of -12.9dB at a downlink SNR of 10dB, the "DP" approach reduces the required pilot length from 12 to 8 in comparison with "SP", which is a 33.3% reduction in the pilot overhead. The improvements are even more significant when the attention module is introduced to the proposed frequency-aware approach, e.g. for a downlink SNR of -5dB, the "DP+attention" approach reduces the required pilot length from 16 to 8 in comparison with "SP" and the reconstruction NMSE by 1.39dB. Fig. 6 compares the NMSE curves versus downlink SNR for "SP", "DP" and "DP+Attention" schemes for (a) $L = 16$ and (b) $L = 12$. This figure also shows a considerable NMSE improvement by the proposed "DP+Attention" scheme in comparison with the "SP" approach proposed in [10].

D. Impact of channel SNR

Fig. 7 studies the performance of the NN-based+Attention scheme when there is a mismatch between the training and test SNRs. This figure plots the NMSE versus test SNR curves for networks trained with each specific SNR value. We see that the NMSE degrades when the test SNR falls below the SNR value used for training. On the other hand, for a network trained with a lower SNR value, the performance saturates when test SNR improves above the training SNR. The grey curve represents the performance of a NN trained over a dataset with sample SNRs picked uniformly at random from the interval $[-5, 10]$dB. As observed in this figure, such a NN performs satisfactorily over the whole SNR range.

TABLE IV: Ablation results of NMSE (dB) for $L = 16$, $N = 32$, $M = 256$.

	SNR $= -5$dB	SNR $= 0$dB	SNR $= 5$dB	SNR $= 10$dB	SNR $= 15$dB	SNR $= 20$dB
Proposed (NN+NN+Attention)	-14.32	-17.61	-20.58	-22.88	-23.65	-23.93
NN+LMMSE	-8.10	-11.47	-15.00	-18.27	-21.49	-23.06
FFT+NN+Attention	-7.74	-10.87	-13.79	-15.02	-15.99	-16.48
FFT+LMMSE	-0.86	-0.92	-1.19	-1.77	-2.77	-4.17

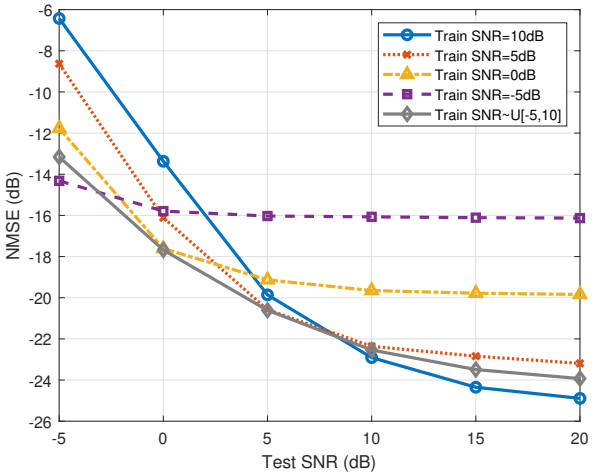

Fig. 7: NMSE (dB) vs. test SNR for networks trained at different SNR values with $L = 16, N = 32, M = 256$.

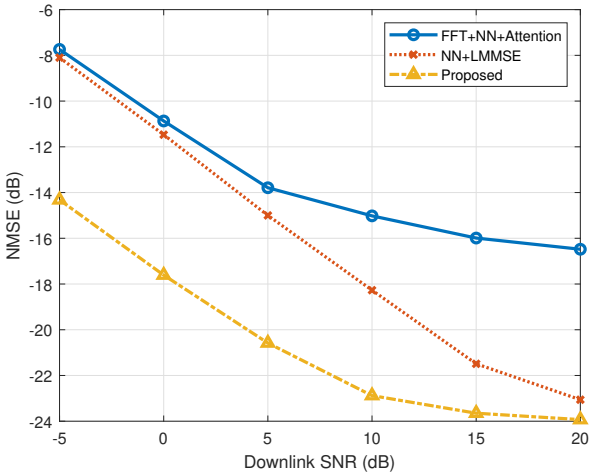

Fig. 8: Ablation results of NMSE (dB) for $L = 16$, $N = 32, M = 256$.

E. Ablation study

To study the improvements by each component of the proposed NN architecture, we provide ablation results in Table IV and Fig. 8. We compare 4 different schemes as below:

- Our proposed scheme, in which both the pilot design and channel estimation are handled by the NN including the attention module. This scheme is denoted by "NN+NN+Attention" in the table.
- A scheme, which uses LMMSE for channel estimation while the pilots are designed using a NN. In this case we take the received pilot vectors $\mathbf{y}_m = \mathbf{P}_m^* \mathbf{h}_m + \mathbf{n}_m$ from the fully connected branches and apply ideal LMMSE (utilizing covariances estimated over the dataset) on it, where $\mathbf{P}_m^*$ denotes the pilots optimized by the NN. This scheme is denoted by "NN+LMMSE" in the table.
- A scheme that employs the "NN+Attention" for channel estimation, but simple FFT pilots are used as proposed in [24] and [11]. Note that, in this scheme, we train the NN+Attention module replacing $\mathbf{P}_m$'s with a submatrix of DFT. This scheme is denoted by "FFT+NN+Attention" in the table.
- The conventional scheme using FFT pilots and LMMSE estimation. This scheme is denoted by "FFT+LMMSE" in the table.

According to Table IV and Fig. 8, the proposed fully NN-based scheme outperforms significantly all the three schemes with the NN-based pilot design leading to a remarkable improvement in the reconstruction NMSE. The conventional FFT+LMMSE scheme performs poorly at all channel SNRs as channel estimation is severely underdetermind for $N = 32$ and $L = 16$.

F. Comparison with extended LMMSE

We showed in the previous subsections that the proposed NN-based channel estimator outperforms the conventional LMMSE estimate by exploiting the inter-frequency correlations in the MIMO channel matrix. In particular, convolutional layers learn and exploit local correlations to improve the estimation accuracy. Local correlations account for a significant performance improvement in the reconstruction NMSE in comparison with simple LMMSE. However, we see that there still exist long range correlations in the channel matrix that are exploited by the proposed attention modules to further improve the reconstruction accuracy.

Note that one can similarly use an extended LMMSE approach which exploits all the inter-frequency and inter-antenna correlations in the channel matrix assuming a Gaussian channel model. This approach is computationally prohibitive as it requires an accurate estimate of a large covariance matrix of size $NM \times NM$ and becomes infeasible in wideband massive MIMO systems where both N and M are significantly larger. Hence, in practical scenarios, the subcarrier-wise LMMSE estimate, which exploits inter-antenna correlations in each subcarrier is used, and hence, this subcarrier-wise LMMSE estimate has been used for comparisons in previous subsections as well as previous works [11], [24].

However, to understand the full potential of the proposed datat-driven approach, in this subsection, we compare the performance of our proposed NN-based approach with the extended LMMSE bound. We are specifically interested in understanding the significance of long range correlations, and how well our proposed NN+Attention approach can learn and exploit these correlations. For the extended LMMSE we estimate the required covariance matrix over the training dataset utilizing the optimum estimator [41]. Hence, the number of real-valued trainable parameters for the extended LMMSE equals $2N^2M^2 + 2NM + 1$, where $2N^2M^2$ of the parameters account for the covariance matrix elements, $2NM$ account for the expected channel values and 1 is for the noise variance.

Table VI provides the number of trainable parameters for several channel estimators considered in this paper. Note that the number of parameters for the convolutional layers and the attention module is independent of the dimensions of the channel matrix. This is very

TABLE V: NMSE comparisons at $\mathrm{SNR} = -5\mathrm{dB}$ and $L = 16, N = 32, M = 256$.

	DP+Attention	DP	SP	LMMSE	Extended LMMSE
$\lvert\mathcal{D}\rvert = 10K$	-14.28	-13.41	-11.85	-8.10	-12.14
$\lvert\mathcal{D}\rvert = 20K$	-14.26	-13.38	-11.80	-8.10	-13.45
$\lvert\mathcal{D}\rvert = 40K$	-14.30	-13.43	-11.88	-8.08	-13.94
$\lvert\mathcal{D}\rvert = 80K$	-14.32	-13.40	-11.84	-8.10	-14.55
# Parameters	333K	267K	5.7K	541K	134M

TABLE VI: Parameter complexity of various channel estimators.

	Number of trainable parameters
DP+Attention	$2LNM + C_1 + C_2$
DP	$2LNM + C_1$
SP	$2LN + C_1$
LMMSE	$2N^2M + 2NM + 1$
Extended LMMSE	$2N^2M^2 + 2NM + 1$

TABLE VII: FLOPs for various channel estimators.

	FLOPs
DP+Attention	$\mathcal{O}(LNM) + \mathcal{O}(NM\rho^2 st) + \mathcal{O}(N^2M^2q)$
DP	$\mathcal{O}(LNM) + \mathcal{O}(NM\rho^2 st)$
SP	$\mathcal{O}(LNM) + \mathcal{O}(NM\rho^2 st)$
LMMSE	$\mathcal{O}(LNM)$
Extended LMMSE	$\mathcal{O}(N^2M^2)$

favorable as these modules can be used in wide-band massive MIMO systems without increasing the parameter complexity, while the model complexity increases with N^2 and M^2 for extended LMMSE and becomes infeasible in wideband massive MIMO scenarios. The number of trainable parameters for the convolutional layers (denoted by C_1) and the attention module (denoted by C_2) for our proposed architectures in Fig. 2 and 3 are given by:

$$C_1 = 128(5 \times 5 + 1) + 128(3 \times 3 + 1) + 2(3 \times 3 + 1) = 4628, \quad (10)$$
$$C_2 = 4 \times 128(128 \times 1 \times 1 + 1) = 66048.$$

Table V presents the reconstruction NMSE for various channel estimator when trained on datasets of various sizes. Note that we originally trained all NNs on a dataset of size $|\mathcal{D}| = 80K$, which includes 80K random realizations of the MIMO-OFDM channel matrices. In this table, we evaluate the performance of various channel estimators when less data is available for training. According to Table V, although the extended LMMSE slightly outperforms the NN-based approach when a large dataset is available to estimate the covariance matrix, its performance degrades significantly when less data is available for training. This is due to the fact that the extended LMMSE approach requires estimating a large covariance matrix with 134M trainable parameters and when less data is available for training, estimates of these parameters become less accurate leading to a significant performance loss. The performance of all the other approaches remains almost the same for smaller dataset sizes.

Table V shows that the proposed NN-based approach with an attention module outperforms extended LMMSE when less data is available for training. Even when trained with 80K samples, the proposed NN+Attention approach achieves a reconstruction NMSE very close to the extended LMMSE while reducing the parameter complexity by a factor of $\frac{134M}{333K} \sim 400$, learning to exploit the most significant long range correlations in the channel matrix.

G. Computational complexity

Our proposed NN-based channel estimation method is trained offline. The training complexity of various methods is compared in terms of the number of trainable parameters for them. In Table VI, we compared the number of real-valued trainable parameters for various methods in terms of the L, N and M values. Table VI includes a numerical comparison for $L = 16, N = 32, M = 256$ and shows that the proposed NN+Attention approach achieves a reconstruction NMSE very close to the extended LMMSE while reducing its parameter complexity by a factor of $\frac{134M}{333K} \sim 400$. Offline training of our proposed NN+attention channel estimation approach for $L = 16, N = 32$ and $M = 256$ up to 110000 steps takes 2 hours on a NVIDIA GEFORCE RTX 2080 Ti GPU.

After the network is trained, it is deployed for channel estimation. The computational complexity of various methods during deployment is compared in terms of the number of floating point operations (FLOPs) required for channel estimation in each method. The computational complexity for the forward path of our proposed NN-based scheme includes the number of FLOPs executed in the fully connected expansion layers, the convolutional layers and the attention module. The number of FLOPS executed for the fully connected expansion layers is $\mathcal{O}(LNM)$. The number of FLOPs required for each convolutional layer is $\mathcal{O}(NM\rho^2 st)$, where s and t are the number of input and output channels for the layer (e.g., $s = t = 128$ here) and ρ^2 is the kernel size (e.g. $\rho^2 = 5 \times 5, 3 \times 3$, etc.). Finally the number of FLOPs required for the matrix multiplications in the attention module is $\mathcal{O}(N^2M^2q)$, where q is the number of output channels in convolutions (we used $q = 128$ in simulations). Table VII compares the number of FLOPs required for various channel estimation techniques in terms of the channel dimensions N, M and L. According to this table, the "NN+attention" and extended LMMSE methods require the same order of FLOPs while without the attention module, the NN-based method significantly reduces the number of FLOPs required.

H. Pilot pruning

Table VIII provides a summary of the results obtained by our proposed pilot pruning scheme. In particular, we present the NMSE as a function of the sparsity level for different pilot lengths when SNR=10dB, $N = 32$ and $M = 256$. The numbers in parentheses show the λ values used to achieve the sparsity level in each simulation according to Eq. (7). We observe that the same λ value performs fairly well for a number of settings, e.g., we get good results with $\lambda = 10^{-6}$ for all $S \leq 50\%$ for $L = 8$ and 12. We observed that a well-performing λ value can always be found by trying only a few negative exponents of 10, i.e., $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$.

According to Table VIII, the NMSE value degrades slowly as we increase the sparsity; and hence, the proposed pruning scheme can be used to efficiently reduce the pilot overhead without significantly degrading the channel estimation accuracy. For example, when $L = 12$, our pilot pruning scheme can save up to 25% of the time-frequency resources while degrading the NMSE by only 1.07dB. Note that the number of time-frequency resources occupied by pilots is the same for $L = 16, S = 25\%$ and $L = 12, S = 0\%$, while the corresponding NMSE values are -19.43dB and -15.35dB for these two settings, respectively. Hence, our proposed pruning scheme can efficiently allocate pilots non-uniformly along subcarriers to improve

TABLE VIII: NMSE (dB) performance achieved for different sparsity levels by the pruning-based pilot reduction technique at SNR=10dB, $N = 32, M = 256$. Values of λ (in Eq. (7)) to achieve the specified sparsity level are given in parentheses.

	$S = 0\%$	$S = 5\%$	$S = 10\%$	$S = 15\%$	$S = 20\%$	$S = 25\%$
$L = 16$	-22.88	$-21.13\,(10^{-6})$	$-20.11\,(10^{-5})$	$-19.85\,(10^{-5})$	$-19.57\,(10^{-5})$	$-19.43\,(10^{-5})$
$L = 12\,(10^{-6})$	-15.35	-14.59	-14.64	-14.39	-14.36	-14.28
$L = 8\,(10^{-6})$	-14.66	-14.23	-13.78	-13.66	-13.55	-13.34

TABLE IX: Pilot allocation performance by NN pruning at SNR=10dB, $N = 32, M = 256$.

Pilot Scheme	A	B	C	D	E
NMSE (dB)	-12.18	-12.24	$-13.34\,(10^{-6})$	-11.96	$-14.41\,(10^{-6})$
SER	1.21×10^{-2}	1.19×10^{-2}	1.02×10^{-2}	1.25×10^{-2}	8.9×10^{-3}

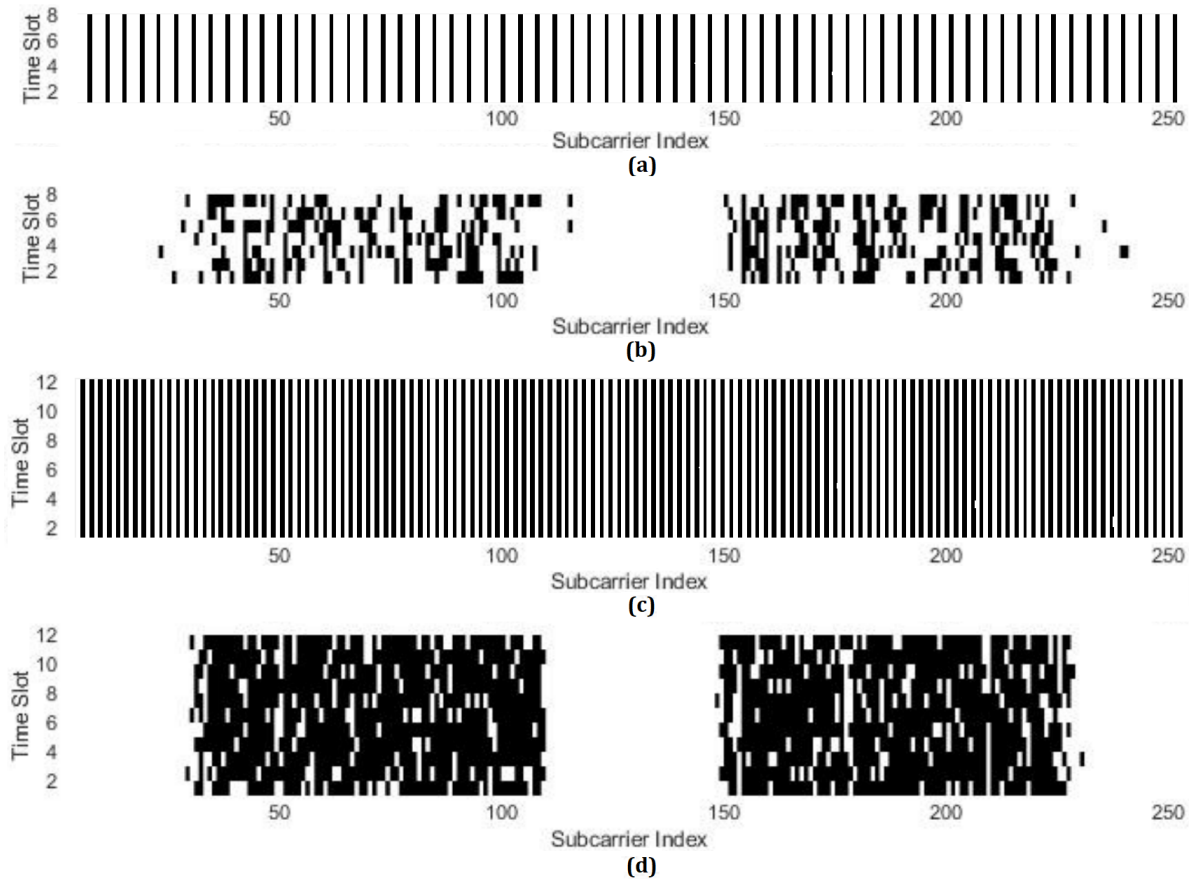

Fig. 9: Illustration of the pilot allocation masks compared in Table IX. White dots represent pilots and the black dots represent time-frequency resources saved for data transmission. (a) Periodic removal of pilots with $L = 8, S = 25\%$, (b) NN optimized pilot pruning with $L = 8, S = 25\%$, (c) Periodic removal of pilots with $L = 12, S = 50\%$, (d) NN optimized pilot pruning with $L = 12, S = 50\%$.

the performance while using the same number of time-frequency resources.

To elaborate further, we compare the following five settings in Table IX, all of which use $6 \times 256 = 1536$ time-frequency pilot resources to estimate the channel over a 70×256 resource grid:

-Pilot scheme A: A rectangular pilot block of size 6×256 is used for pilot transmission according to the frame structure depicted in Fig. 1.

-Pilot scheme B: In this scheme, one out of every 4 pilots is omitted periodically from a 8×256 rectangular block of pilots and saved for data transmission. Note that, we still use proposed NN scheme for pilot optimization and channel estimation, but without pruning. Fig. 9a depicts the pilot mask for scheme B.

-Pilot scheme C: In this scheme, we employ the proposed pruning-based pilot allocation scheme with $L = 8$ and $S = 25\%$. Fig. 9b depicts the pilot mask optimized by our proposed method for this scheme.

-Pilot scheme D: In this scheme one out of every 2 pilots is omitted periodically from a 12×256 rectangular block of pilots. Similarly to scheme B, we keep the pilot allocation mask M fixed during training. Fig. 9c depicts the pilot mask for scheme D.

-Pilot scheme E: In this scheme, we employ our proposed pruning-based pilot allocation scheme with $L = 12$, and $S = 50\%$. Fig. 9d depicts the pilot mask obtimized by our proposed method for this scheme.

Table IX shows considerable improvement by the proposed pruning-based pilot allocation scheme in comparison with periodic and rectangular allocations. For example, by starting from a pilot

block of 12×256 and pruning 50% of the pilots with our proposed scheme, we can improve the NMSE by 2.23dB in comparison with a 6×256 pilot block, which has the same pilot overhead. It should be noted that, $L = 12$ slightly increases the estimation delay at the receiver as the user cannot estimate the channel until all 12 resource blocks are received over the sub-carriers.

Table IX shows that when a fixed budget of time-frequency resources is available for pilot transmission, starting from a larger L value and pruning the network down to the available resource budget results in an improved NMSE performance. The NN achieves this by allocating less pilots to those sub-carriers which can be interpolated with sufficient accuracy by the NN exploiting the statistical CSI correlations, while allocating more pilots to other subcarriers. Fig. 9 depicts the allocation masks resulting from the proposed pilot pruning scheme for $L = 8, S = 25\%$ (Fig. 9b) and $L = 12, S = 50\%$ (Fig. 9d), where the white dots represent the time-frequency resources devoted to pilots, while the black dots represent those saved for data transmission. As observed in this figure, more resources can be saved for data over transmission subcarriers in the ranges (30-110) and (150-230) in the setting considered.

In Table IX, we also compare the symbol error rate (SER) values for 4-QAM modulation over a 70×256 time-frequency grid, where the channel is estimated using the five different pilot allocation schemes. We note that we use both the pruned and the data-dedicated time-frequency resources for 4-QAM transmission. For each allocation scheme, we estimate the channel with the trained NN, and use a zero-forcing equalizer over the resource grid, and the maximum likelihood (ML) detector to demodulate the received signal. We report the resulting SER over 10^8 random symbols. The downlink SNR is 10dB and $N = 32, M = 256$. Table IX shows that the proposed NN-based pilot pruning scheme improves the SER while using the same number of time-frequency resources for pilot transmission; that is, the channel resources saved from the pilot block by the proposed pruning approach can be used for reliable data transmission while reducing the SER.

VI. Conclusion

We have proposed a NN-based joint downlink pilot design and channel estimation scheme for FDD massive MIMO-OFDM systems. Our proposed network utilizes dense layers to design pilot signals in a frequency-aware structure followed by convolutional layers which utilize inherent correlations in the channel matrix to provide an accurate channel estimate in an efficient manner. We have also employed an attention module to exploit long-range correlations in the channel matrix, which cannot be inferred by the conventional convolutional layers.We also proposed an effective pilot reduction technique by gradually pruning less significant neurons from the dense layers to reduce the pilot overhead and to save time-frequency resources for data transmission. Our proposed NN-based pilot design and channel estimation scheme outperforms LMMSE estimation. Moreover, our pruning-based pilot reduction technique effectively reduces the pilot overhead by allocating pilots across subcarriers non-uniformly; allowing less pilot transmissions on subcarriers that can be satisfactorily reconstructed by the subsequent convolutional layers exploiting inter-frequency correlations.

References

[1] X. Rao and V. K. N. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3261–3271, 2014.

[2] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Transactions on Signal Processing*, vol. 63, no. 23, pp. 6169–6183, 2015.

[3] Z. Gao, L. Dai, W. Dai, B. Shim, and Z. Wang, "Structured compressive sensing-based spatio-temporal joint channel estimation for FDD massive MIMO," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 601–617, 2016.

[4] Y. Han, J. Lee, and D. J. Love, "Compressed sensing-aided downlink channel training for FDD massive MIMO systems," *IEEE Transactions on Communications*, vol. 65, no. 7, pp. 2852–2862, 2017.

[5] X. Kuai, L. Chen, X. Yuan, and A. Liu, "Structured turbo compressed sensing for downlink massive MIMO-ofdm channel estimation," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3813–3826, 2019.

[6] J. Fang, X. Li, H. Li, and F. Gao, "Low-rank covariance-assisted downlink training and channel estimation for FDD massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1935–1947, 2017.

[7] X. Li, J. Fang, H. Li, and P. Wang, "Millimeter wave channel estimation via exploiting joint sparse and low-rank structures," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1123–1133, Feb 2018.

[8] M. B. Mashhadi and D. Gündüz, "Deep learning for massive MIMO channel state acquisition and feedback," *arXiv:2002.06945 [cs.IT]*, Feb. 2020.

[9] Z. Liu, L. Zhang, and Z. Ding, "Overcoming the channel estimation barrier in massive MIMO communication systems," *arXiv:1912.10573 [cs.IT]*, Dec. 2019.

[10] X. Ma and Z. Gao, "Data-driven deep learning to design pilot and channel estimator for massive MIMO," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 5677–5682, 2020.

[11] P. Dong, H. Zhang, G. Y. Li, I. S. Gaspar, and N. NaderiAlizadeh, "Deep CNN-based channel estimation for mmwave massive MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 5, pp. 989–1000, Sep. 2019.

[12] Y. Zhang, M. Alrabeiah, and A. Alkhateeb, "Deep learning for massive MIMO with 1-bit ADCs: When more antennas need fewer pilots," *arXiv: 1910.06960 [cs.IT]*, 2019.

[13] E. Balevi and J. G. Andrews, "Two-stage learning for uplink channel estimation in one-bit massive MIMO," *arxiv: 1911.12461 [eess.SP]*, 2019.

[14] S. Gao, P. Dong, Z. Pan, and G. Y. Li, "Deep learning based channel estimation for massive MIMO with mixed-resolution ADCs," *IEEE Communications Letters*, vol. 23, no. 11, pp. 1989–1993, Nov 2019.

[15] H. He, C. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmwave massive mimo systems," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 852–855, 2018.

[16] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, 2018.

[17] T. Wang, C. Wen, S. Jin, and G. Y. Li, "Deep learning-based CSI feedback approach for time-varying massive MIMO channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 416–419, April 2019.

[18] Z. Liu, L. Zhang, and Z. Ding, "Exploiting bi-directional channel reciprocity in deep learning for low rate massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, pp. 1–1, 2019.

[19] J. Guo, C. Wen, S. Jin, and G. Y. Li, "Convolutional neural network based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2020.

[20] M. B. Mashhadi, Q. Yang, and D. Gündüz, "Distributed deep convolutional compression for massive MIMO CSI feedback," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2020.

[21] J. Guo, X. Yang, C.-K. Wen, S. Jin, and G. Y. Li, "DL-based CSI feedback and cooperative recovery in massive MIMO," *arXiv: 2003.03303 [cs.IT]*, 2020.

[22] Y. Jang, G. Kong, M. Jung, S. Choi, and I. Kim, "Deep autoencoder based CSI feedback with feedback errors and feedback delay in FDD massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 833–836, June 2019.

[23] M. B. Mashhadi, Q. Yang, and D. Gunduz, "CNN-based analog CSI feedback in FDD MIMO-OFDM systems," in *2020 IEEE 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[24] P. Dong, H. Zhang, G. Y. Li, N. NaderiAlizadeh, and I. S. Gaspar, "Deep cnn for wideband mmwave massive MIMO channel estimation using frequency correlation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 4529–4533.

[25] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive mimo with hybrid precoding over frequency-selective

fading channels," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1259–1262, 2016.

[26] Q. Yang, M. B. Mashhadi, and D. Gündüz, "Deep convolutional compression for massive MIMO CSI feedback," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019, pp. 1–6.

[27] A. Buades, B. Coll, and J. . Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 2005, pp. 60–65 vol. 2.

[28] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[29] J. Fang, X. Li, H. Li, and F. Gao, "Low-rank covariance-assisted downlink training and channel estimation for FDD massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1935–1947, 2017.

[30] S. Bazzi and W. Xu, "Downlink training sequence design for FDD multiuser massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4732–4744, 2017.

[31] Y. Gu and Y. D. Zhang, "Information-theoretic pilot design for downlink channel estimation in FDD massive MIMO systems," *IEEE Transactions on Signal Processing*, vol. 67, no. 9, pp. 2334–2346, 2019.

[32] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural networks," in *Advances in Neural Information Processing Systems (NIPS 2015)*, 2015, p. 1135–1143.

[33] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019.

[34] H. Mao, S. Han, J. Pool, W. Li, X. Liu, Y. Wang, and W. J. Dally, "Exploring the granularity of sparsity in convolutional neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1927–1934.

[35] H. Hu, R. Peng, Y. Tai, and C. Tang, "Network trimming: A data-driven neuron pruning approach towards efficient deep architectures," *CoRR*, vol. abs/1607.03250, 2016. [Online]. Available: http://arxiv.org/abs/1607.03250

[36] L. Theis, I. Korshunova, A. Tejani, and F. Huszár, "Faster gaze prediction with dense networks and fisher pruning," 2018, arXiv:1801.05787. [Online]. Available: https://arxiv.org/abs/1801.05787

[37] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[38] K. Persand, A. Anderson, and D. Gregg, "Composition of saliency metrics for channel pruning with a myopic oracle," *arXiv preprint arXiv:2004.03376v1*, 3 Apr 2020.

[39] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, F. Quitin, F. Tufvesson, and P. D. Doncker, "The COST 2100 MIMO channel model," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 92–99, December 2012.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[41] S. M. Kay, *Fundamentals of statistical signal processing*, ser. Prentice Hall signal processing series. Prentice Hall PTR, 1993.

Mahdi Boloursaz Mashhadi [S'14-M'20] received a dual major B.S. degree (Hons.) in electrical and industrial engineering and the M.S. and Ph.D. degrees in electrical engineering from Sharif University of Technology (SUT), Tehran, Iran, in 2011, 2013 and 2018, respectively. He worked as a research associate at the University of Central Florida (UCF), and Queens' University, Canada, and as a Lecturer at Sharif University of Technology (SUT), Tehran, Iran. He is currently a post-doctoral research associate with the Intelligent Systems and Networks (ISN) research group, Imperial College London. His research interests include the areas of wireless communications, machine learning, and sparse/statistical signal processing. His awards include the Best Paper Award at the EWDTS 2012 conference, and several national, regional, and IEEE grants. He serves as a reviewer for IEEE Transactions on Wireless Communications, IEEE Transactions on Communications, and IEEE Transactions on Signal Processing.

Deniz Gündüz [S'03-M'08-SM'13] received the B.S. degree in electrical and electronics engineering from METU, Turkey in 2002, and the M.S. and Ph.D. degrees in electrical engineering from NYU Tandon School of Engineering (formerly Polytechnic University) in 2004 and 2007, respectively. After his PhD, he served as a postdoctoral research associate at Princeton University, as a consulting assistant professor at Stanford University, and as a research associate at CTTC in Barcelona, Spain. In Sep. 2012, he joined the Electrical and Electronic Engineering Department of Imperial College London, UK, where he is currently a Professor of Information Processing, and serves as the deputy head of the Intelligent Systems and Networks Group. He is also a part-time faculty member at the University of Modena and Reggio Emilia, Italy, and has held visiting positions at University of Padova (2018-2020) and Princeton University (2009-2012).

His research interests lie in the areas of communications and information theory, machine learning, and privacy. Dr. Gündüz is a Distinguished Lecturer for the IEEE Information Theory Society (2020-22). He is an Area Editor for the IEEE Transactions on Communications and the IEEE Journal on Selected Areas in Communications (JSAC). He also serves as an Editor of the IEEE Transactions on Wireless Communications. He is an organiser and co-chair of the 2021 London Symposium on Information Theory. He is the recipient of the IEEE Communications Society - Communication Theory Technical Committee (CTTC) Early Achievement Award in 2017, a Starting Grant of the European Research Council (ERC) in 2016, and several best paper awards.