Deep Learning-Based Robust Channel Estimation for MIMO IoT Systems

Jae-Mo Kang, *Member, IEEE*

***Abstract*—When the second-order statistics of channel and noise, such as their covariance matrices, are not exactly known, the acquisition of accurate channel state information (CSI) for a wireless propagation environment becomes quite challenging. In this paper, we tackle the problem of robust channel estimation for multiple-input multiple-output (MIMO)-aided Internet-of-Things (IoT) systems in the presence of uncertainties in the channel and noise covariance matrices. Our goal is to minimize the mean square error (MSE) of the channel estimation under the channel and noise covariance uncertainties by jointly optimizing the channel estimator and pilot signal, which is however highly nonconvex and mathematically intractable. To effectively and intelligently cope with this issue, we exploit a deep learning (DL) technique and propose a novel network architecture with two modules, namely, the pilot optimizer and channel predictor, both of which are designed by neural networks with their own local connections and weight sharings. Moreover, a novel and effective training strategy for the proposed DL model is devised in a self-supervised manner, in which samples obtained by properly compensated channel and noise covariance matrices are utilized to overcome any adverse impacts of the underlying uncertainties on the channel estimation. Through extensive numerical results simulated in realistic propagation environments, we substantiate the superior performance and effectiveness of the proposed scheme.**

***Index Terms*—Channel estimation, covariance uncertainty, deep learning, MIMO, IoT, robust training design.**

I. Introduction

Internet-of-Things (IoT) is now becoming an essential part of ubiquitous connections between various devices and services, supporting seamless interactions in our daily lives [1]–[3]. Particularly, the extensive connectivity of IoT together with the substantial data collected by diverse devices will be practically very useful in many deployment scenarios such as smart cities, smart homes, smart factories, and smart transportation [2], [3]. However, the rapid proliferation of IoT and its applications in various fields have led to ever-increasing demands for reliable wireless communications. Multiple-input multiple-output (MIMO) exploiting the spatial diversity of a wireless channel through the use of multiple antennas at both the transmitter and receiver [4], [5], has recently emerged as a key enabler technology to address such challenges [1], [2]. Accordingly, MIMO-aided IoT systems have the great potential and synergies to substantially enhance the reliability, coverage, and energy-efficiency of communication links.

However, the merits the MIMO IoT systems can offer are fully realizable only when accurate channel state information (CSI) is available. In practice, the CSI needs to be acquired by sending a priori known pilot (or training) signal, for which the accuracy of the resulting CSI estimate critically relies on how well the channel training process is designed according to the statistical knowledge of a given wireless propagation environment such as second-order statistics or covariance matrices of channel and noise [6]–[13].

In practice, the channel and noise covariance matrices need to be estimated as well based on real samples for channel estimate and noise measurement, respectively. As a result, the estimates of these covariance matrices are inevitably erroneous (especially, in some extreme propagation environments such as in unmanned aerial vehicle (UAV) or satellite communication scenarios with high mobility) due to imperfections in the channel estimation and noise measurement processes [14]–[22]. Moreover, the estimated covariance matrices may even be further distorted for the certain purposes such as quatization, compression, data embedding, feedback transmission, etc., inducing additional discrepancies with the actual values [17]–[22]. It is obvious that only with the knowledge of the estimated channel and noise covariance matrices, the traditional signal processing approach such as the linear minimum mean square error (LMMSE) channel estimation method may fail to accurately estimate the CSI due to the mismatches between the actual and estimated covariance matrices [19]. Even though the CSI can be estimated without any knowledge of the channel and noise statistics such as via the least squares (LS) channel estimation method, the resulting performance of such an approach might not be satisfactory, especially at low signal-to-noise ratio (SNR), due to the noise amplification [8]. To properly cope with these issues while guaranteeing the robustness, therefore, the uncertainties involved in the channel and noise covariance matrices have to be taken into account during the channel estimation.

In the literature, in the presence of the covariance uncertainties, the robust pilot signal design techniques were investigated for multiple-input single-output (MISO) systems [16], MIMO systems [17]–[20], UAV-assisted communication systems [21], and MIMO relaying systems [22]. These techniques, however, have several major limitations. First of all, in [16]–[22], an idealistic assumption was made that the LMMSE channel estimation could be performed with the exact knowledge of the

This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1A4A1033830), and in part by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-2020-0-01808) supervised by the IITP(Institute of Information & Communications Technology Planning Evaluation).

J.-M. Kang is with the Department of Artificial Intelligence, Kyungpook National University, Daegu 41566, South Korea (e-mail: jmkang@knu.ac.kr).

channel and noise covariance matrices. Thus, the channel estimation methodology considered therein is not actually robust. Furthermore, in the majority of the works [16]–[21], only the impact of the channel covariance uncertainty was considered in the pilot design, while that of the noise covariance uncertainty was neglected. In addition, the design approaches of [17]–[22] all presumed a specific channel model (namely, the Kronecker channel model), and thus, there is lack of generality in their applicability, especially for the scenarios where the presumed channel model is violated. Even if the presumed channel model is validated, the pilot signal designs need to be carried out through complicated iterative algorithms with high computational complexities (except for some special cases investigated in [16] and [20]), which hinders their applicability to practical IoT systems with stringent real-time operations.

Recently, deep learning (DL) techniques have emerged as promising solutions to improve the performance of wireless communication systems by effectively and intelligently overcoming various technical challenges [23]–[31]. Particularly, in [32]–[39], DL-based channel estimation techniques have been developed for various MIMO systems by considering different design approaches. Specifically, in [32], a beamspace channel estimation technique has been devised for millimeter-wave massive MIMO systems based on a learned denoising-based approximate message passing (LDAMP) neural network, which incorporated a denoising convolutional neural network (CNN) into an iterative sparse signal recovery algorithm. In [33], two algorithms for direction-of-arrival (DOA) estimation and channel estimation have been developed for massive MIMO systems based on (deep) feedforward neural networks (FNNs) by leveraging the spatial structure. In [34], a deep learning compressed sensing (DLCS) channel estimation scheme has been proposed for multi-user millimeter-wave massive MIMO systems and a deep learning quantized phase (DLQP) hybrid precoder design method has been developed subsequent to the channel estimation. The authors in [35] proposed a sparse complex-valued neural network (SCNet) for the downlink CSI prediction in frequency division duplex (FDD) massive MIMO systems via the uplink-to-downlink mapping function. Moreover, joint channel estimation and pilot signal design schemes with different DL architectures have been suggested for massive MIMO systems via data-aided iterative channel estimation [36], multi-user MIMO systems via successive interference cancellation [37], MIMO systems via received SNR feedback [38], and MIMO-OFDM systems via neural network pruning [39]. Unfortunately, however, the aforementioned works [32]–[39] did not consider the impacts of the channel and noise covariance uncertainties in the channel estimation (as well as pilot design) process, and thus, their performance will be deteriorated seriously in the real-world scenarios with imperfect covariance information (i.e., only with the knowledge of inexact covariance matrices). Accordingly, it is highly necessary to develop an innovative channel estimation scheme with high accuracy of the channel estimate even under the channel and noise covariance uncertainties.

To the best of our knowledge, all the aforementioned critical issues have not been addressed yet in the literature, which motivated our work. In this paper, we study the problem of robust channel estimation for MIMO-aided IoT systems in the presence of uncertainties in both the channel and noise covariance matrices, based on deep learning.[1] Note that our work is the first to present a DL framework for the robust channel estimation and pilot signal design, to the best of our knowledge, which is not resorting to the assumptions invoked in [16]–[22], and hence, has a wider applicability. To accomplish our design goal, we aim to minimize mean square error (MSE) of the channel estimation under the channel and noise covariance uncertainties by jointly optimizing the channel estimator and pilot signal, which is however generally hard to tackle due to nonconvexity and intractability. To break through this technical challenge, we develop an effective and intelligent DL technique. The main contributions of this paper are as follows:

- We propose a novel and effective DL model for the robust MIMO channel estimation with two modules, namely, the pilot optimizer and channel predictor, which has never been reported in the literature to the best of our knowledge.[2] The pilot optimizer is constructed by a locally-connected and weight-shared FNN with a specifically designed layer, called the *pilot layer*, such that the shared weights between the locally-connected nodes correspond to the pilot signal, enabling the optimization of the pilot signal through training. Moreover, we construct the channel predictor by adopting a CNN structure such that useful features for the robust MIMO channel estimation are efficiently learnable.
- In addition, we devise a novel and effective training strategy for the proposed DL model in a self-supervised manner, in which the channel and noise covariance matrices are appropriately compensated to overcome any adverse impacts of the underlying uncertainties on the channel estimation, and then, the samples drawn from the compensated covariance matrices are used to jointly train the two modules of the proposed DL model based on two different gradient descent approaches such that the MSE loss function of the prediction is minimized.
- We present extensive simulation results, through which the superior performance and better effectiveness of the proposed DL model is demonstrated compared to baseline schemes and some useful engineering insights into the

[1]The robust channel estimation in this paper means the robust estimation of CSI of the MIMO system with imperfect knowledge of the channel and noise covariance matrices under uncertainties in the channel and noise covariance matrices, the goal of which is to acquire the CSI estimate of the MIMO system that is robust to the channel and noise covariance uncertainties. The CSI estimate acquired by the robust channel estimation can be used for the subsequent tasks such as the robust beamforming design. Our proposed scheme can also be used for this purpose.

[2]The key novelty of our work lies in constructing the network structure of the pilot optimizer based on our own innovative construction inspired by the MIMO system model for transmission and reception of the pilot signal. In turn, the whole network structure of the proposed DL model combining the pilot optimizer and the channel predictor is entirely new and specialized in dealing with the robust channel estimation. In addition, a new finding from our work is that the constructed pilot optimizer (i.e., a very special neural network) still works well with the channel predictor, further supporting and demonstrating the universal superiority and effectiveness of constructing the prediction part with the CNN structure.

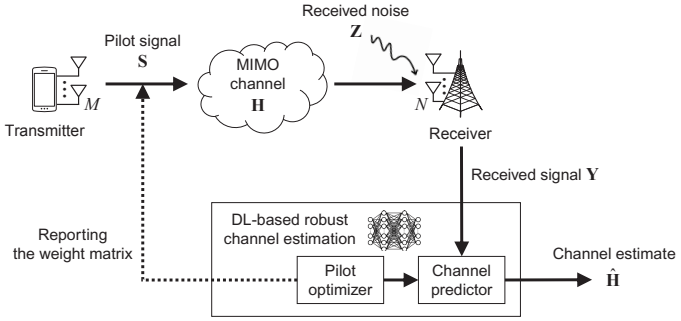

Fig. 1. A MIMO IoT system with the proposed DL model for the robust channel estimation.

practical design are drawn. We also analyze the computational complexities of the proposed and baseline schemes.

This paper is organized as follows. In Section II, the system model is described and the robust channel estimation problem under consideration is formulated. In Section III, the proposed DL architecture and training strategy are elaborated. Section IV presents the simulation results along with thorough discussions and complexity analysis. Section V concludes this paper.

Notations: $\mathbb{R}^{a\times b}$ and $\mathbb{C}^{a\times b}$ stand for the sets of $a\times b$ real- and complex-valued matrices, respectively. Also, $\mathbf{A}^{\mathsf{T}}$, $\mathbf{A}^{\mathsf{H}}$, $\mathbf{A}^{\frac{1}{2}}$, $\mathbf{A}^{-1}$, and $\mathbf{A}^{\dagger}$ denote the transpose, conjugate (or Hermitian) transpose, (Hermitian) square root, inverse, and pseudo-inverse of a matrix $\mathbf{A}$, respectively. $\mathbf{a}=\mathsf{vec}(\mathbf{A})$ is the vectorization of a matrix $\mathbf{A}$, which stacks all column vectors of $\mathbf{A}$ into a long column vector $\mathbf{a}$, and its inverse operator is denoted by $\mathbf{A}=\mathsf{vec}^{-1}(\mathbf{a})$. Also, $\mathbf{A}\otimes\mathbf{B}$ is the Kronecker product between matrices $\mathbf{A}$ and $\mathbf{B}$. The maximum eigenvalue of a matrix $\mathbf{A}$ is denoted by $\lambda_{\max}(\mathbf{A})$, and the (a,b)th entry of $\mathbf{A}$ by $\left[\mathbf{A}\right]_{a,b}$. The expectation of a random variable is denoted by $\mathbb{E}\left[\cdot\right]$. The real and imaginary parts of a complex-valued argument a is denoted by $\mathsf{Re}\{a\}$ and $\mathsf{Im}\{a\}$, respectively. The cardinality of a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. An $a\times a$ identity matrix is denoted by $\mathbf{I}_a$, and a zero matrix with an appropriate size by $\mathbf{0}$. In addition, $\mathbf{A}\succeq\mathbf{0}$ means that a Hermitian matrix $\mathbf{A}$ is positive semi-definite. The probability distribution of a circularly symmetric complex Gaussian (CSCG) random vector with mean $\mathbf{a}$ and covariance matrix $\mathbf{B}$ is denoted by $\mathcal{CN}(\mathbf{a},\mathbf{B})$. Also, $\mathcal{W}(a,\mathbf{B})$ denotes a Wishart distribution with a degrees of freedom and a scale matrix $\mathbf{B}$.

II. System Model and Problem Formulation

A. System Model

As shown in Fig. 1, we consider a MIMO IoT system composed of a transmitter (e.g., a mobile or IoT device) and a receiver (e.g., a base station or gateway),[3] which are equipped with M and N antennas, respectively. For the purpose of CSI acquisition, the transmitter sends a priori known pilot signal of length L, denoted by $\mathbf{S}\in\mathbb{C}^{L\times M}$, to the receiver, of which transmission power is constrained such that $\mathrm{Tr}(\mathbf{S}^{\mathsf{H}}\mathbf{S})\le P$, where P denotes the maximum power budget. The received pilot signal at the receiver, denoted by $\mathbf{Y}\in\mathbb{C}^{N\times L}$, is then given by

$$\mathbf{Y}=\mathbf{H}\mathbf{S}^{\mathsf{T}}+\mathbf{Z} \tag{1}$$

where $\mathbf{H}\in\mathbb{C}^{N\times M}$ and $\mathbf{Z}\in\mathbb{C}^{N\times L}$ are the matrices of MIMO channel coefficients and received additive noises (possibly accounting for interference from other links), respectively. Using $\mathrm{vec}(\mathbf{AXB})=(\mathbf{B}^{\mathsf{T}}\otimes\mathbf{A})\mathsf{vec}(\mathbf{X})$ [40], the received signal in (1) can be written in a vector form as

$$\mathbf{y}=\big(\mathbf{S}\otimes\mathbf{I}_N\big)\mathbf{h}+\mathbf{z} \tag{2}$$

where $\mathbf{y}=\mathsf{vec}(\mathbf{Y})$, $\mathbf{h}=\mathsf{vec}(\mathbf{H})$, and $\mathbf{z}=\mathsf{vec}(\mathbf{Z})$.

Let $\mathbf{C_h}=\mathbb{E}[\mathbf{h}\mathbf{h}^{\mathsf{H}}]\succeq\mathbf{0}$ and $\mathbf{C_z}=\mathbb{E}[\mathbf{z}\mathbf{z}^{\mathsf{H}}]\succeq\mathbf{0}$ denote the (actual) channel and noise covariance matrices, respectively. In practice, the values of $\mathbf{C_h}$ and $\mathbf{C_z}$ are not exactly known as they have to be estimated or acquired from the real (yet erroneous) samples for the channel estimate and noise measurement, respectively.[4] On top of such an incomplete acquisition, further errors may also arise in the subsequent processes such as quantization, compression, data embedding, feedback transmission, etc. Consequently, in practice, the estimated channel and noise covariance matrices are generally inaccurate and unavoidably subject to some errors. Considering such imperfection, in this paper, the mismatches between the actual values of $\mathbf{C_h}$ and $\mathbf{C_z}$, and their estimated values (denoted by $\hat{\mathbf{C}}_\mathbf{h}\succeq\mathbf{0}$ and $\hat{\mathbf{C}}_\mathbf{z}\succeq\mathbf{0}$, respectively) are modeled as follows [16]–[22]:

$$\mathbf{C_h}=\hat{\mathbf{C}}_\mathbf{h}+\mathbf{E_h}, \tag{3}$$

$$\mathbf{C_z}=\hat{\mathbf{C}}_\mathbf{z}+\mathbf{E_z} \tag{4}$$

where $\mathbf{E_h}\in\mathcal{E}_\mathbf{h}$ and $\mathbf{E_z}\in\mathcal{E}_\mathbf{z}$ denote the corresponding error matrices, both of which are (generally indefinite) Hermitian matrices (i.e., $\mathbf{E_h}=\mathbf{E_h^H}$ and $\mathbf{E_z}=\mathbf{E_z^H}$) such that $\hat{\mathbf{C}}_\mathbf{h}+\mathbf{E_h}\succeq\mathbf{0}$ and $\hat{\mathbf{C}}_\mathbf{z}+\mathbf{E_z}\succeq\mathbf{0}$, respectively. Furthermore, $\mathcal{E}_\mathbf{h}$ and $\mathcal{E}_\mathbf{z}$ denote unitarily-invariant sets of the channel and noise covariance uncertainties, respectively, such that if $\mathbf{E_h}\in\mathcal{E}_\mathbf{h}$ and $\mathbf{E_z}\in\mathcal{E}_\mathbf{z}$, then $\mathbf{U}\mathbf{E_h}\mathbf{U}^{\mathsf{H}}\in\mathcal{E}_\mathbf{h}$ and $\mathbf{V}\mathbf{E_z}\mathbf{V}^{\mathsf{H}}\in\mathcal{E}_\mathbf{z}$ for arbitrary unitary matrices $\mathbf{U}$ and $\mathbf{V}$. Examples of such unitarily-invariant sets are norm-bounded sets such as $\mathcal{E}_\mathbf{a}=\left\{\mathbf{E_a}:\mathsf{Tr}(\mathbf{E_a^H}\mathbf{E_a})\le\epsilon_\mathbf{a}\right\}$ (i.e., Frobenius norm-bounded set),

[3]This single user scenario is very fundamental, and even useful and insightful (and thus, still meaningful) for in-depth inspection of the robust channel estimation with deep learning. Nevertheless, our work is not confined to the single user scenario, but can be readily extended to the multi-user or massive access scenario. Specifically, the proposed DL model developed for the single user scenario can be readily extended to the multi-user or massive access scenario in uplink only with a very minor modification on the parameter update for the pilot optimizer to deal with individual transmission power constraints on pilot signals of multiple transmitters. Also, the proposed scheme can be directly applied to the multi-user or massive access scenario in downlink by treating the whole of multiple receivers as a large-size single receiver. Due to the scalability issue, however, the maximum number of accessible users should be limited or judiciously determined in practice according to the model capacity of the constructed DL network as well as the system requirements on the computational cost/capability/burden and the inference latency/delay.

[4]If the actual covariance matrices $\mathbf{C_h}$ and $\mathbf{C_z}$ were exactly known, the system CSI could be acquired via the linear channel estimation technique such as the LMMSE channel estimation, which would be given by (27) with $\hat{\mathbf{C}}_\mathbf{h}$ and $\hat{\mathbf{C}}_\mathbf{z}$ replaced by $\mathbf{C_h}$ and $\mathbf{C_z}$, respectively.

$\mathcal{E}_{\mathbf{a}} = \left\{\mathbf{E}_{\mathbf{a}} : \sqrt{\lambda_{\max}(\mathbf{E}_{\mathbf{a}}^{\mathsf{H}}\mathbf{E}_{\mathbf{a}})} \leq \epsilon_{\mathbf{a}}\right\}$ (i.e., spectral norm-bounded set), and $\mathcal{E}_{\mathbf{a}} = \left\{\mathbf{E}_{\mathbf{a}} : \mathsf{Tr}\big((\mathbf{E}_{\mathbf{a}}^{\mathsf{H}}\mathbf{E}_{\mathbf{a}})^{\frac{1}{2}}\big) \leq \epsilon_{\mathbf{a}}\right\}$ (i.e., nuclear norm-bounded set) for $\mathbf{a} \in \{\mathbf{h}, \mathbf{z}\}$, where $\epsilon_{\mathbf{a}} \geq 0$ denotes an upper bound [20], [21].

Remark 1: In this paper, we consider a more realistic and general, yet much more challenging, scenario than the previous works [16]–[21] considered, in the following aspects:

- In [16]–[22], the actual values of the channel and noise covariance matrices were assumed to be exactly known at the receiver, which is idealistic and impractical. While, in this paper, we make a more practical assumption that those actual values are *not* knowable even at the receiver side.
- Most of the previous works [16]–[21] assumed that there existed the uncertainty only in the channel covariance matrix, whereas there was no uncertainty in the noise covariance matrix. More generally, in this paper, we assume that there exist the uncertainties in *both* the channel and noise covariance matrices.
- In [17]–[22], the channel and noise covariance matrices were assumed to be Kronecker-separable; that is, each of them can be factorized into the Kronecker product of two smaller matrices such that $\mathbf{C}_{\mathbf{h}} = \mathbf{A}_{\mathbf{h}} \otimes \mathbf{B}_{\mathbf{h}}$ and $\mathbf{C}_{\mathbf{z}} = \mathbf{A}_{\mathbf{z}} \otimes \mathbf{B}_{\mathbf{z}}$ for some $\mathbf{A}_{\mathbf{h}} \succeq \mathbf{0}$, $\mathbf{B}_{\mathbf{h}} \succeq \mathbf{0}$, $\mathbf{A}_{\mathbf{z}} \succeq \mathbf{0}$, and $\mathbf{B}_{\mathbf{z}} \succeq \mathbf{0}$. However, this assumption is generally inaccurate in practice [14] and may even be violated in certain scenarios where a strong coupling between transmitter and receiver exists due to proximity and/or in certain types of propagation environments where the transmitter and receiver shares a part of scatterers [14]. In this paper, for generality and universality of the practical applicability, it is assumed that the channel and noise covariance matrices are *Kronecker-inseparable*, i.e., $\mathbf{C}_{\mathbf{h}} \neq \mathbf{A}_{\mathbf{h}} \otimes \mathbf{B}_{\mathbf{h}}$ and $\mathbf{C}_{\mathbf{z}} \neq \mathbf{A}_{\mathbf{z}} \otimes \mathbf{B}_{\mathbf{z}}$, respectively.

B. Problem Formulation

The goal of the robust channel estimation in this paper is to estimate the MIMO channel vector $\mathbf{h}$ as accurately as possible in the presence of the uncertainties in the channel and noise covariance matrices, which is a rather challenging task. Obviously, only with the knowledge of the estimated covariance matrices $\hat{\mathbf{C}}_{\mathbf{h}}$ and $\hat{\mathbf{C}}_{\mathbf{z}}$, the traditional signal processing approach such as the LMMSE channel estimation method may fail to achieve this goal due to the mismatches between the actual and estimated covariance matrices.

To mathematically formalize an optimization problem for the robust channel estimation under consideration, let $\hat{\mathbf{h}} = f_\theta(\mathbf{y}; \mathbf{S})$ denote a MIMO channel estimate, which is specified by a (possibly nonlinear) function of the received signal $\mathbf{y}$ for a given pilot signal $\mathbf{S}$, parameterized by a set θ of parameters. In this paper, we aim to find the channel estimator f_θ as well as to design the pilot signal $\mathbf{S}$ with the transmission power constraint such that the MSE of the channel estimation is minimized under the channel and noise covariance uncertainties as follows:

$$\text{(P1)}: \quad \underset{f_\theta, \mathbf{S}}{\text{minimize}} \quad \mathbb{E}\Big[\big\|\mathbf{h} - f_\theta(\mathbf{y}; \mathbf{S})\big\|^2 \Big| \hat{\mathbf{C}}_{\mathbf{h}}, \hat{\mathbf{C}}_{\mathbf{z}}\Big] \tag{5}$$

$$\text{subject to} \quad \mathsf{Tr}\big(\mathbf{S}^{\mathsf{H}}\mathbf{S}\big) \leq P.$$

Note that problem (P1) is nonlinear and nonconvex. In particular, it even involves functional optimization that is mathematically intractable. As a result, problem (P1) is generally NP-hard, and thus, it is infeasible to tackle problem (P1) directly.[5] Even for a rather simpler case where the channel estimator is restricted to be linear as well as for a very simplistic case where the LMMSE channel estimator is adopted with an idealistic assumption that the exact knowledge of the channel and noise covariance matrices is available, problem (P1) still remains very difficult to solve even numerically due to the non-convexity. Furthermore, as discussed in Remark 1, the robust channel estimation problem formulated in (P1) is much more challenging than those considered in the previous works [16]–[22]. Consequently, the existing solution approaches in [16]–[22] from the optimization perspective are neither effective nor applicable to solving (P1).

Motivated by breaking through these technical challenges effectively and intelligently, in the following section, we derive a new and innovative solution to problem (P1) based on deep learning via the construction of a novel neural network.

III. Deep Learning-Based Robust MIMO Channel Estimation

In this section, we first elaborate the network structure of the proposed DL model developed for the robust MIMO channel estimation. Then the training methodology with a strategy of the channel and noise covariance compensation is presented.

A. Network Structure

The whole network structure of the proposed DL model is presented in Fig. 2. The covariance compensation is initially carried out to obtain training samples. Then two DL modules, the pilot optimizer and the channel predictor, that are connected in tandem are trained via two different gradient descent methods with the MSE loss function. In what follows, these three components are elaborated and specified.

1) Covariance Compensation: Note that since the actual covariance matrices are not known, these cannot be used for the training. On the other hand, although the estimated covariance matrices are known, only using these for the training results in severely limited performance due to the discrepancies with the actual covariance matrices (as will be demonstrated by the simulation results in Section IV). To properly deal with these issues and to guarantee the robustness against the covariance uncertainties by compromising between the actual and estimated covariance matrices, in the proposed DL model, we compensate the estimated channel and noise covariance matrices, $\hat{\mathbf{C}}_{\mathbf{h}}$ and $\hat{\mathbf{C}}_{\mathbf{z}}$, by intentionally adding some distortions $\bar{\mathbf{E}}_{\mathbf{h}}$ and $\bar{\mathbf{E}}_{\mathbf{z}}$, respectively, as follows:

$$\bar{\mathbf{C}}_{\mathbf{h}} = \hat{\mathbf{C}}_{\mathbf{h}} + \bar{\mathbf{E}}_{\mathbf{h}}, \tag{6}$$

$$\bar{\mathbf{C}}_{\mathbf{z}} = \hat{\mathbf{C}}_{\mathbf{z}} + \bar{\mathbf{E}}_{\mathbf{z}}. \tag{7}$$

[5]Even in certain scenarios such as in the low SNR regime and/or when adopting the LMMSE channel estimator, considering or importing the noise covariance matrix uncertainty particularly poses several major technical challenges beyond just handling more variables.

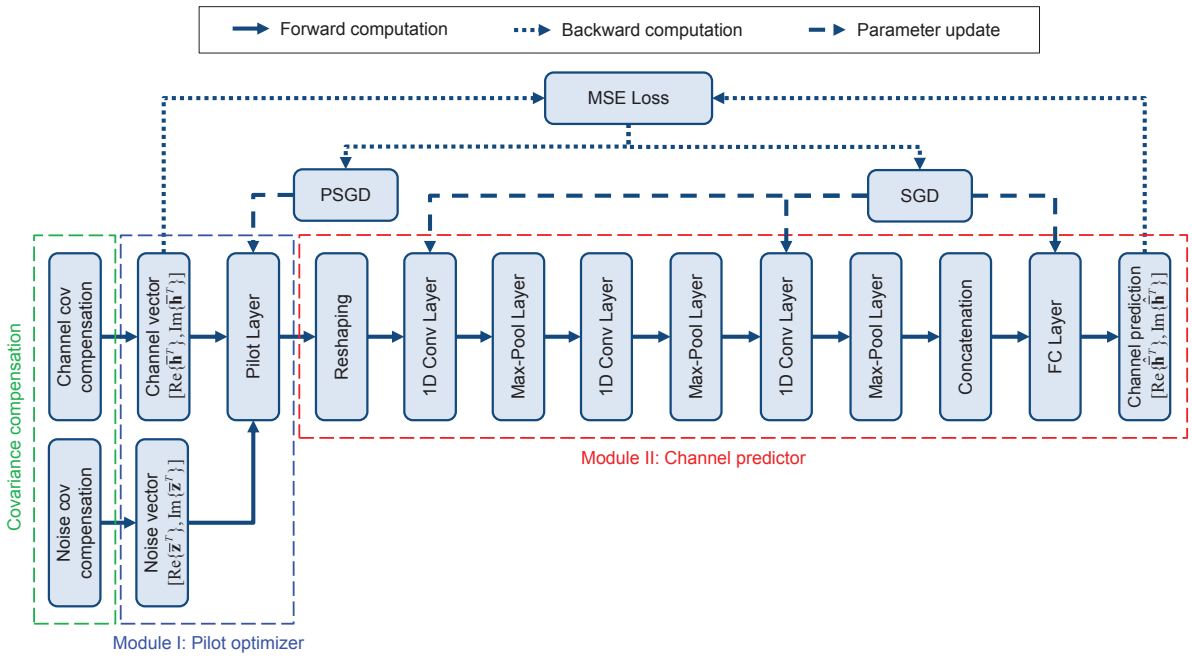

Fig. 2. The whole network structure of the proposed DL model.

Here, $\bar{\mathbf{E}}_{\mathbf{h}}$ and $\bar{\mathbf{E}}_{\mathbf{z}}$ have the same statistics as those of $\mathbf{E}_{\mathbf{h}}$ and $\mathbf{E}_{\mathbf{z}}$, respectively, i.e., $\bar{\mathbf{E}}_{\mathbf{h}} \in \mathcal{E}_{\mathbf{h}}$ and $\bar{\mathbf{E}}_{\mathbf{z}} \in \mathcal{E}_{\mathbf{z}}$. Consequently, $\bar{\mathbf{C}}_{\mathbf{h}}$ and $\bar{\mathbf{C}}_{\mathbf{z}}$ have the same statistics as those of the actual covariance matrices $\mathbf{C}_{\mathbf{h}}$ and $\mathbf{C}_{\mathbf{z}}$, respectively (although the values of $\bar{\mathbf{C}}_{\mathbf{h}}$ and $\bar{\mathbf{C}}_{\mathbf{z}}$ are not exactly the same as those of $\mathbf{C}_{\mathbf{h}}$ and $\mathbf{C}_{\mathbf{z}}$). In the sequel, we will refer to $\bar{\mathbf{C}}_{\mathbf{h}}$ and $\bar{\mathbf{C}}_{\mathbf{z}}$ as the compensated channel and noise covariance matrices, respectively, or simply, the compensated covariance matrices.[6]

Let $\bar{\mathbf{h}}$ and $\bar{\mathbf{z}}$ denote the channel and noise vectors whose covariance matrices are equivalent to the compensated channel and noise covariance matrices $\bar{\mathbf{C}}_{\mathbf{h}}$ and $\bar{\mathbf{C}}_{\mathbf{z}}$, respectively (these will be simply referred to as the compensated channel and noise vectors in the sequel). Then the inputs of the proposed DL model are $\bar{\mathbf{h}}$ and $\bar{\mathbf{z}}$, and the output is the prediction of $\bar{\mathbf{h}}$, denoted by $\hat{\bar{\mathbf{h}}}$. In what follows, we further elaborate and specify the two modules of the proposed DL model.

2) Pilot Optimizer (DL Module I): The goal of employing the pilot optimizer in the proposed DL model is to enable the pilot signal $\mathbf{S}$ optimizable or designable as in problem (P1) by learning the system model of (2) for the transmission and reception of the pilot signal through the noisy MIMO channel.[7] To achieve this design goal, we construct the pilot optimizer using a single layer FNN with a novel local connection and weight sharing between the input and output nodes such that the shared weights between the locally-connected nodes correspond to the pilot signal $\mathbf{S}$ to be optimized. We refer to this specifically designed layer as the *pilot layer*.

The network structure of the pilot optimizer (i.e., DL module I) we construct is shown in Fig. 3, detailed explanations and specifications on which are given as follows:

- *Configuration:* In the constructed pilot optimizer, the number of input nodes (i.e., input size) is $2N(M+L)$ and the number of output nodes (i.e., output size) is $2NL$. For the sake of notational brevity, we denote the first NL output nodes by $\{a_{(l-1)N+n} : l = 1, \cdots, L, n = 1, \cdots, N\}$ and the remaining NL output nodes by $\{a'_{(l-1)N+n} : l = 1, \cdots, L, n = 1, \cdots, N\}$. In a similar fashion, we denote the first $2MN$ input nodes via $\{x_{(m-1)N+n}, x'_{(m-1)N+n} : m = 1, \cdots, M, n = 1, \cdots, N\}$ and the remaining $2NL$ input nodes via $\{b_{(l-1)N+n}, b'_{(l-1)N+n} : l = 1, \cdots, L, n = 1, \cdots, N\}$.
- *Weight connection:* As depicted in Fig. 3(a), in the pilot optimizer, the input and output nodes are locally connected with shared weights such that for each $n \in \{1, \cdots, N\}$ and $l \in \{1, \cdots, L\}$;
  - $\{x_{(m-1)N+n} : m = 1, \cdots, M\}$ are connected to $a_{(l-1)N+n}$ and $a'_{(l-1)N+n}$ with weights $\{w_{l,m} : m = 1, \cdots, M\}$ and $\{w'_{l,m} : m = 1, \cdots, M\}$, respectively.
  - $\{x'_{(m-1)N+n} : m = 1, \cdots, M\}$ are connected to $a_{(l-1)N+n}$ and $a'_{(l-1)N+n}$ with weights $\{-w_{l,m} : m = 1, \cdots, M\}$ and $\{w_{l,m} : m = 1, \cdots, M\}$, respectively.
  - $b_{(l-1)N+n}$ is connected to $a_{(l-1)N+n}$ with a weight fixed to 1.
  - $b'_{(l-1)N+n}$ is connected to $a'_{(l-1)N+n}$ with a weight fixed to 1.

  As an illustrative example, in Fig. 3(b), the network structure of the constructed pilot optimizer is shown when

[6]Overall, according to (6) and (7), $\bar{\mathbf{C}}_{\mathbf{h}}$ and $\bar{\mathbf{C}}_{\mathbf{z}}$ are defined as the intentionally compensated or artificially distorted versions of the estimated covariance matrices $\hat{\mathbf{C}}_{\mathbf{h}}$ and $\hat{\mathbf{C}}_{\mathbf{z}}$, respectively, to mimic the statistics of the actual covariance matrices $\mathbf{C}_{\mathbf{h}}$ and $\mathbf{C}_{\mathbf{z}}$.

[7]It is worth noting that as will be demonstrated by the simulation results in Section IV, a naive DL model solely using the channel predictor without the pilot optimizer turns out to be ineffective, indicating that the pilot optimizer plays a crucial role in the DL-based robust channel estimation.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3324667

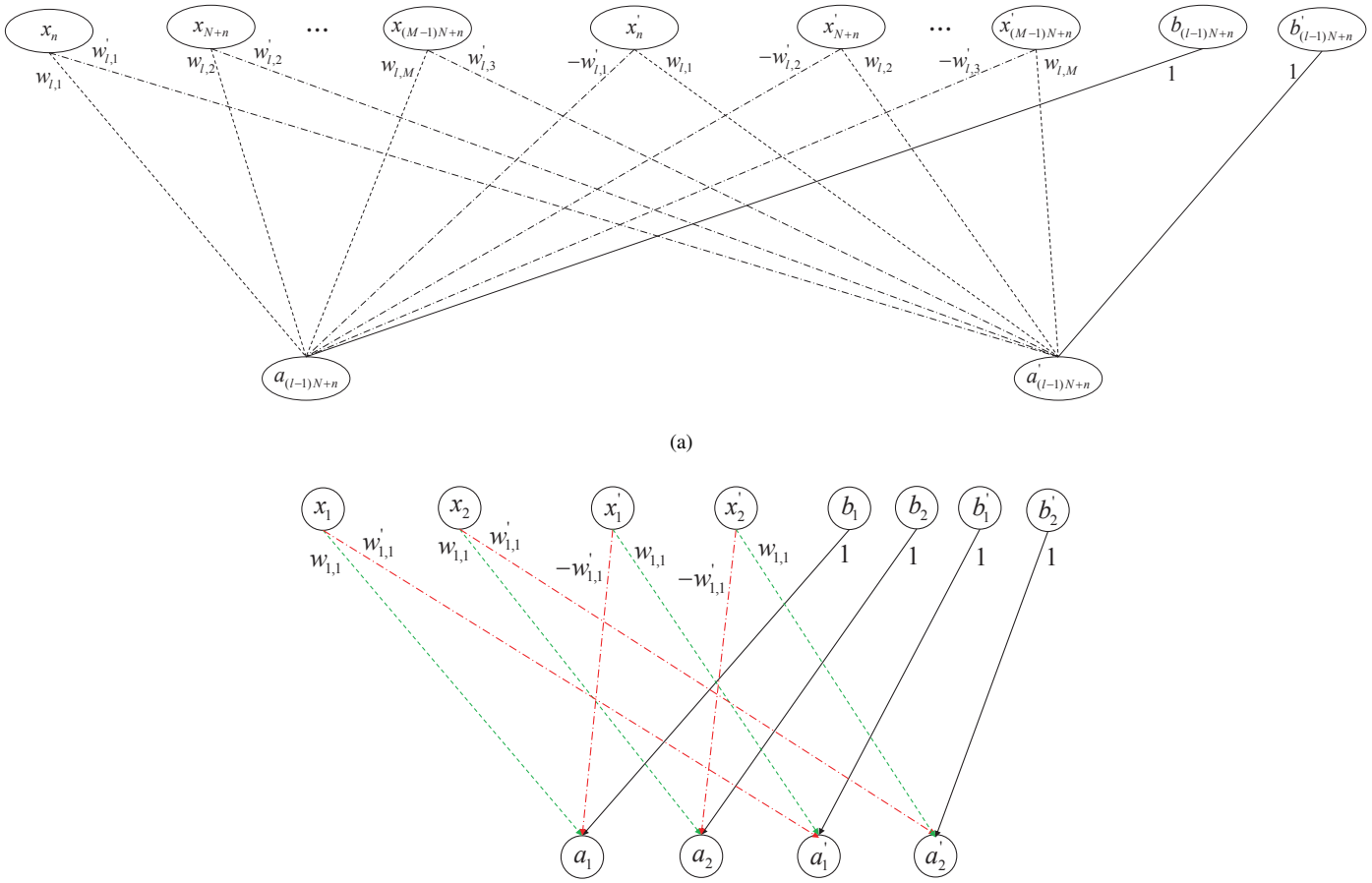

Fig. 3. Network structure of the pilot optimizer in the proposed DL model. (a) Weight connection between the input and output nodes for each $n \in \{1, \cdots, N\}$ and $l \in \{1, \cdots, L\}$, where not all nodes are shown; instead, only the nodes with connections are shown for brevity. The connections with weights $\{w_{l,m} : m = 1, \cdots, M\}$ and $\{w'_{l,m} : m = 1, \cdots, M\}$ are denoted by dashed and dashed-dot lines, respectively. Also, the connections with weights fixed to 1 are denoted by solid lines. (b) Network structure of the pilot optimizer when $M = L = 1$ and $N = 2$.

$M = L = 1$ and $N = 2$.

- *Operation:* At each output node, the weighted sum of the inputs is computed, followed by passing through an activation function $\phi(\cdot)$. Thus, the operation of each output node is given by (8) (shown at the bottom), or equivalently,

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{a}' \end{bmatrix} = \phi\left(\left(\begin{bmatrix} \mathbf{W} & -\mathbf{W}' \\ \mathbf{W}' & \mathbf{W} \end{bmatrix} \otimes \mathbf{I}_N \right) \begin{bmatrix} \mathbf{x} \\ \mathbf{x}' \end{bmatrix} + \begin{bmatrix} \mathbf{b} \\ \mathbf{b}' \end{bmatrix} \right) \in \mathbb{R}^{2NL \times 1} \tag{9}$$

where

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_{NL} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{NL \times 1}, \tag{10}$$

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,M} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ w_{L,1} & w_{L,2} & \cdots & w_{L,M} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{L \times M}, \tag{11}$$

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_{MN} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{MN \times 1}, \tag{12}$$

$$\mathbf{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_{NL} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{NL \times 1}. \tag{13}$$

Also, $\mathbf{a}'$, $\mathbf{W}'$, $\mathbf{x}'$, and $\mathbf{b}'$ are defined similarly as in (10)–(13), respectively.

- *Design inspiration:* The network structure of the pilot optimizer is inspired by the system model of (2). Specifically, we can decompose (2) into real and imaginary parts

$$a_{(l-1)N+n} = \phi\left(\sum_{m=1}^{M} \left[w_{l,m} x_{(m-1)N+n} - w'_{l,m} x'_{(m-1)N+n} \right] + b_{(l-1)N+n} \right), \quad n = 1, \cdots, N, \quad l = 1, \cdots, L, \tag{8a}$$

$$a'_{(l-1)N+n} = \phi\left(\sum_{m=1}^{M} \left[w_{l,m} x'_{(m-1)N+n} + w'_{l,m} x_{(m-1)N+n} \right] + b'_{(l-1)N+n} \right), \quad n = 1, \cdots, N, \quad l = 1, \cdots, L. \tag{8b}$$

as

$$\begin{bmatrix}\mathsf{Re}\{\mathbf{y}\}\\ \mathsf{Im}\{\mathbf{y}\}\end{bmatrix} = \left(\begin{bmatrix}\mathsf{Re}\{\mathbf{S}\} & -\mathsf{Im}\{\mathbf{S}\}\\ \mathsf{Im}\{\mathbf{S}\} & \mathsf{Re}\{\mathbf{S}\}\end{bmatrix} \otimes \mathbf{I}_N\right)\begin{bmatrix}\mathsf{Re}\{\mathbf{h}\}\\ \mathsf{Im}\{\mathbf{h}\}\end{bmatrix} + \begin{bmatrix}\mathsf{Re}\{\mathbf{z}\}\\ \mathsf{Im}\{\mathbf{z}\}\end{bmatrix} \in \mathbb{R}^{2NL\times 1}. \quad (14)$$

Notably and rather intriguingly, we can observe that the mathematical expression of (9) is equivalent to that of (14) provided that $\phi(x) = x$, $\mathbf{a} = \mathsf{Re}\{\mathbf{y}\}$, $\mathbf{a}' = \mathsf{Im}\{\mathbf{y}\}$, $\mathbf{W} = \mathsf{Re}\{\mathbf{S}\}$, $\mathbf{W}' = \mathsf{Im}\{\mathbf{S}\}$ $\mathbf{x} = \mathsf{Re}\{\mathbf{h}\}$, $\mathbf{x}' = \mathsf{Im}\{\mathbf{h}\}$, $\mathbf{b} = \mathsf{Re}\{\mathbf{z}\}$, and $\mathbf{b}' = \mathsf{Im}\{\mathbf{z}\}$. This essentially means that the physical mechanism (as well as relevant important feature) for the transmission and reception of the pilot signal through the noisy MIMO channel can be learned by the constructed pilot optimizer. In particular, the weights of the constructed pilot optimizer correspond to the pilot signal (i.e., $\mathbf{W} = \mathsf{Re}\{\mathbf{S}\}$ and $\mathbf{W}' = \mathsf{Im}\{\mathbf{S}\}$), and thus, the optimization of the pilot signal can be readily carried out through the training procedure, which is clearly a significant design advantage.

- *Parameter setup:* According to the aforementioned design inspiration, in the sequel, we treat the weights of the pilot optimizer as the pilot signal to be optimized, i.e., we set $\mathbf{W} = \mathsf{Re}\{\mathbf{S}\}$ and $\mathbf{W}' = \mathsf{Im}\{\mathbf{S}\}$.
- *Activation function:* Also, the activation function at each output node is set to the linear function, i.e., $\phi(x) = x$.
- *Input:* To enable the optimization of the pilot signal through the training procedure, the pilot optimizer needs to take the actual values of $\mathbf{h}$ and $\mathbf{z}$ as inputs such that $\mathbf{x} = \mathsf{Re}\{\mathbf{h}\}$, $\mathbf{x}' = \mathsf{Im}\{\mathbf{h}\}$, $\mathbf{b} = \mathsf{Re}\{\mathbf{z}\}$, and $\mathbf{b}' = \mathsf{Im}\{\mathbf{z}\}$. Unfortunately, however, this is infeasible since the actual covariance matrices are not known. To address this issue, we instead propose to take the real and imaginary parts of the compensated channel and noise vectors (rather than the actual values that are unknown) as the inputs of the pilot optimizer as follows:

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{x}'\end{bmatrix} = \begin{bmatrix}\mathsf{Re}\{\bar{\mathbf{h}}\}\\ \mathsf{Im}\{\bar{\mathbf{h}}\}\end{bmatrix} \in \mathbb{R}^{2MN\times 1}, \quad (15)$$

$$\begin{bmatrix}\mathbf{b}\\ \mathbf{b}'\end{bmatrix} = \begin{bmatrix}\mathsf{Re}\{\bar{\mathbf{z}}\}\\ \mathsf{Im}\{\bar{\mathbf{z}}\}\end{bmatrix} \in \mathbb{R}^{2NL\times 1}. \quad (16)$$

3) Channel Predictor (DL Module II): In the proposed DL model, we also employ a (deep) neural network subsequent to the pilot optimizer, called the channel predictor, which serves as the channel estimator f_θ in problem (P1), where θ denotes the set of all learnable parameters. This is motivated by the universal function approximation theorem: even a neural network with a single hidden layer has a capability to approximate any nonlinear function within an arbitrarily accuracy [41]–[43]. Particularly, we construct the channel predictor by adopting a CNN structure that is characterized by local connection and weight sharing, in order to pursue a structural matching with the pilot optimizer as well as to efficiently learn and extract the effective/useful features for the robust MIMO channel estimation under the channel and noise covariance uncertainties.

The network structure of the constructed channel predictor (i.e., DL module II) is shown in Fig. 4, which consists of a sequential connection of several convolutional and pooling layers, followed by a fully-connected (FC) layer. Details are given in the following.

- *Reshaping:* The channel predictor in the proposed DL model first reshapes the output of the pilot optimizer for an appropriate processing by the CNN. Specifically, the output $\left[\mathbf{a}^\mathsf{T}, \mathbf{a}'^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{2NL\times 1}$ of the pilot optimizer is divided into $2N$ patches, each of length L, as follows:

$$\begin{bmatrix}\mathbf{A}\\ \mathbf{A}'\end{bmatrix} \in \mathbb{R}^{2N\times L} \quad (17)$$

  where $\mathbf{A} = \mathsf{vec}^{-1}(\mathbf{a})$ and $\mathbf{A}' = \mathsf{vec}^{-1}(\mathbf{a}')$.
- *1D Conv Layers:* In the convolution stage, we employ three one-dimensional (1D) convolutional layers.
  - *Input:* The first convolutional layer takes the reshaped output of the pilot optimizer in (17) as the input. The second and third convolutional layers take the outputs of the first and second max-pooling layers as the inputs, respectively. In each convolutional layer, the input is properly zero padded such that the output has the same size as the input.
  - *Operation:* The kth convolutional layer performs 1D convolution between the zero-padded input and c_k different kernels, each of size (or length) ℓ_k, with unit stride rate. Let $\{\chi_{i,j}\}$ denote the input to the kth convolutional layer. Then the output of the kth convolutional layer of size r_k is given as [42], [43]

$$\alpha_{i,j} = \varphi\left(\sum_{p=1}^{c_{k-1}}\sum_{q=1}^{\ell_{k-1}} \omega_{i,p,q}\chi_{p,j+q-1} + \beta_i\right) \quad (18)$$

    for $i = 1, \cdots, c_k$ and $j = 1, \cdots, r_k$, where $\{\omega_{i,p,q}\}$ is the set of weights of the ith kernel and $\{\beta_i\}$ denotes the set of bias terms. For $k = 1$, we have $c_0 = 2N$, $\ell_0 = L$, and $\{\chi_{i,j}\}$ is properly formed by entries of $\left[\mathbf{A}^\mathsf{T}, \mathbf{A}'^\mathsf{T}\right]^\mathsf{T}$ covered by the c_1 kernels. Also, φ denotes an activation function.
  - *Number and size of kernels:* The number of kernels in the kth convolutional layer is set to $c_k = 8kMN$, $k = 1, 2, 3$. Also, the kernel size is set to $\ell_k = 3$, $\forall k$, which has been demonstrated to be an adequate size to extract sufficient spatial features of the input data [44].
  - *Activation function:* The activation function in each convolutional layer is chosen as the exponential linear unit (ELU), i.e., $\varphi(x) = \exp(x) - 1$ for $x \leq 0$ and $\varphi(x) = x$ for $x > 0$ [45].
- *Max-Pooling Layers:* In the pooling stage, we employ three max-pooling layers, each of which is placed after each convolutional layer, such that the dimensionality of the features extracted by the prior convolutional layers is gradually reduced via 1D down-sampling in order to prevent overfitting as well as make features more robust against noise, shift, distortion, etc.
  - *Input:* The kth max-pooling layer takes the output of the kth convolutional layer as the input.
  - *Operation:* Let ℓ'_k denote the kernel size as well as the stride rate in the kth max-pooling layer. Then the

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3324667

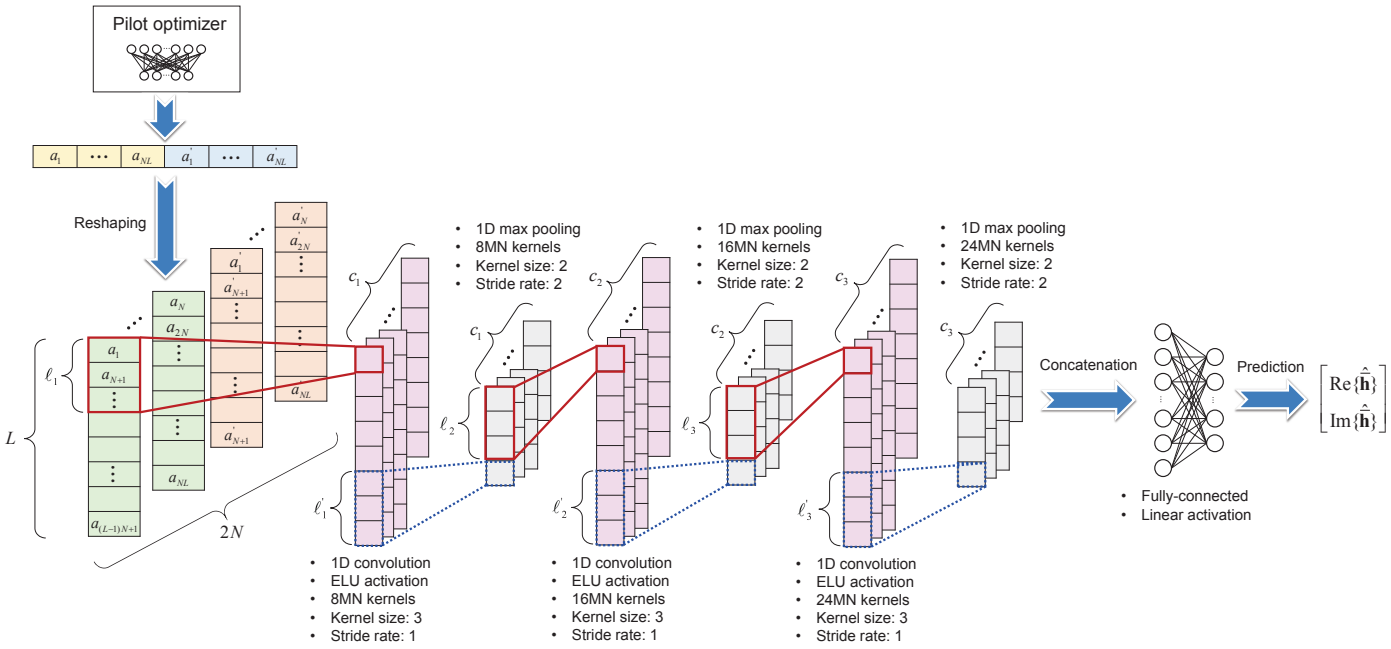

Fig. 4. Network structure of the channel predictor in the proposed DL model.

output of the kth max-pooling layer of size r'_k is given by [43]

$$\alpha'_{i,j} = \max\left\{\alpha_{i,q} : q \in \Omega_j\right\} \tag{19}$$

for $i = 1, \cdots, c_k$ and $j = 1, \cdots, r'_k$, where $\Omega_j = \{q : q \in \{\ell'_k(j-1)+1, \ell'_k(j-1)+2, \cdots, \ell'_k(j-1)+\ell'_k\}\}$.

- *Kernel size:* In each pooling layer of the constructed channel predictor, the kernel size is set to $\ell'_k = 2$, $k = 1, 2, 3$, such that the size of the feature extracted by each convolutional layer is halved after passing through each pooling layer.

- *Concatenation:* The output of the last max-pooling layer is concatenated into a one-dimensional vector.
- *FC Layer:* The final stage of the channel predictor is the FC layer, which is constructed by a single layer FNN with full connection for the purpose of fine-tuning of the features obtained by the convolutional and pooling layers.
  - *Input:* The FC layer takes the concatenated output of the last pooling layer as the input.
  - *Operation:* Let $\mathbf{x}_{\mathsf{F}}$ denote the input of the FC layer, i.e., the concatenated output of the last max-pooling layer. Then the FC layer processes the input by first multiplying a weight matrix $\mathbf{W}_{\mathsf{F}}$ and then adding a bias vector $\mathbf{b}_{\mathsf{F}}$, followed by passing through an activation function $\varphi_{\mathsf{F}}(\cdot)$. Thus, the output of the FC layer is given by $\mathbf{a}_{\mathsf{F}} = \varphi_{\mathsf{F}}\left(\mathbf{W}_{\mathsf{F}}\mathbf{x}_{\mathsf{F}} + \mathbf{b}_{\mathsf{F}}\right)$.
  - *Activation function:* The activation function in the FC layer is set to the linear function, i.e., $\varphi_{\mathsf{F}}(x) = x$.
- *Output:* The channel predictor takes the output of the FC layer as its output such that the real and imaginary parts of the compensated channel vector $\bar{\mathbf{h}}$ are predicted as

$$\begin{bmatrix} \mathsf{Re}\{\hat{\bar{\mathbf{h}}}\} \\ \mathsf{Im}\{\hat{\bar{\mathbf{h}}}\} \end{bmatrix} = \mathbf{W}_{\mathsf{F}}\mathbf{x}_{\mathsf{F}} + \mathbf{b}_{\mathsf{F}}. \tag{20}$$

B. Proposed Training Procedure

Through the training procedure, we jointly train the two modules, the pilot optimizer and channel predictor, of the proposed DL model in the self-supervised manner such that the joint optimization in (P1) can be carried out. The detailed process is explained in the following.

1) Training Data Acquisition: Once the channel and noise covariance matrices are compensated as in (6) and (7), respectively, the samples of the compensated channel and noise vectors, $\bar{\mathbf{h}}$ and $\bar{\mathbf{z}}$, i.e., the inputs of the proposed DL model, are drawn from the compensated covariance matrices $\bar{\mathbf{C}}_{\mathbf{h}}$ and $\bar{\mathbf{C}}_{\mathbf{z}}$, respectively. For example, for the case of Rayleigh fading with additive Gaussian noise (i.e., $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_{\mathbf{h}})$ and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_{\mathbf{z}})$), the samples of $\bar{\mathbf{h}}$ and $\bar{\mathbf{z}}$ can be obtained such that $\bar{\mathbf{h}} \sim \mathcal{CN}(\mathbf{0}, \bar{\mathbf{C}}_{\mathbf{h}})$ and $\bar{\mathbf{z}} \sim \mathcal{CN}(\mathbf{0}, \bar{\mathbf{C}}_{\mathbf{z}})$, respectively.

2) Parameter Update: The parameters of the proposed DL model, i.e., the weights $\mathbf{W} = \mathsf{Re}\{\mathbf{S}\}$ and $\mathbf{W}' = \mathsf{Im}\{\mathbf{S}\}$ in the pilot layer of the pilot optimizer and the set θ of the weights and biases in the convolutional and fully-connected layers of the channel predictor, need to be jointly optimized. For this purpose, the loss function for the training is selected as the empirical MSE between the compensated channel vector $\bar{\mathbf{h}}$ and the predicted value $\hat{\bar{\mathbf{h}}}$ (i.e., the output of the proposed DL model) as follows:

$$\mathcal{L}(\mathbf{S}, \theta) = \frac{1}{|\mathcal{T}|} \sum_{(\bar{\mathbf{h}}, \bar{\mathbf{z}}) \in \mathcal{T}} \left\| \begin{bmatrix} \mathsf{Re}\{\bar{\mathbf{h}}\} \\ \mathsf{Im}\{\bar{\mathbf{h}}\} \end{bmatrix} - \begin{bmatrix} \mathsf{Re}\{\hat{\bar{\mathbf{h}}}\} \\ \mathsf{Im}\{\hat{\bar{\mathbf{h}}}\} \end{bmatrix} \right\|^2 \tag{21}$$

$$\mathbf{\Pi}_{\mathcal{S}}\left(\begin{bmatrix}\mathsf{Re}\{\mathbf{A}\}\\ \mathsf{Im}\{\mathbf{A}\}\end{bmatrix}\right)=\begin{cases}\begin{bmatrix}\mathsf{Re}\{\mathbf{A}\}\\ \mathsf{Im}\{\mathbf{A}\}\end{bmatrix}, & \text{if } \mathsf{Tr}\big(\mathsf{Re}\{\mathbf{A}^{\mathsf{T}}\}\mathsf{Re}\{\mathbf{A}\}+\mathsf{Im}\{\mathbf{A}^{\mathsf{T}}\}\mathsf{Im}\{\mathbf{A}\}\big)\le P\\ \sqrt{\frac{P}{\mathsf{Tr}\big(\mathsf{Re}\{\mathbf{A}^{\mathsf{T}}\}\mathsf{Re}\{\mathbf{A}\}+\mathsf{Im}\{\mathbf{A}^{\mathsf{T}}\}\mathsf{Im}\{\mathbf{A}\}\big)}}\begin{bmatrix}\mathsf{Re}\{\mathbf{A}\}\\ \mathsf{Im}\{\mathbf{A}\}\end{bmatrix}, & \text{if } \mathsf{Tr}\big(\mathsf{Re}\{\mathbf{A}^{\mathsf{T}}\}\mathsf{Re}\{\mathbf{A}\}+\mathsf{Im}\{\mathbf{A}^{\mathsf{T}}\}\mathsf{Im}\{\mathbf{A}\}\big)> P\end{cases}. \quad (24)$$

where $\mathcal{T}$ denotes the set of training samples.

To minimize the loss function in (21), the parameters of the proposed DL model can be updated based on two different gradient descent methods through the backward computations. Specifically, the update of θ can be performed via the stochastic gradient descent (SGD) method as

$$\theta \leftarrow \theta - \gamma \frac{\partial \mathcal{L}(\mathbf{S},\theta)}{\partial \theta} \quad (22)$$

where $\gamma > 0$ is the step size or learning rate. On the other hand, the weight update for the pilot optimizer can be carried out based on the projected SGD (PSGD) method such that the power constraint of the pilot signal in problem (P1) is fulfilled as [38]

$$\begin{bmatrix}\mathsf{Re}\{\mathbf{S}\}\\ \mathsf{Im}\{\mathbf{S}\}\end{bmatrix} \leftarrow \mathbf{\Pi}_{\mathcal{S}}\left(\begin{bmatrix}\mathsf{Re}\{\mathbf{S}\}\\ \mathsf{Im}\{\mathbf{S}\}\end{bmatrix} - \gamma \begin{bmatrix}\frac{\partial \mathcal{L}(\mathbf{S},\theta)}{\partial \mathsf{Re}\{\mathbf{S}\}}\\ \frac{\partial \mathcal{L}(\mathbf{S},\theta)}{\partial \mathsf{Im}\{\mathbf{S}\}}\end{bmatrix}\right) \quad (23)$$

where $\Pi_{\mathcal{S}}$ denotes the projection operator onto the feasible set $\mathcal{S} = \big\{\mathbf{S} \in \mathbb{C}^{L\times M} : \mathsf{Tr}\big(\mathbf{S}^{\mathsf{H}}\mathbf{S}\big) \le P\big\}$, which is given by (24) (shown at the top). That is, if the power constraint is violated after the SGD update, the value of the updated $\mathbf{S}$ is immediately normalized such that $\mathsf{Tr}\big(\mathbf{S}^{\mathsf{H}}\mathbf{S}\big) = P$.

3) Deployment of Proposed DL Model After Training: Note that although the two modules of the proposed DL model are jointly trained in the (offline) training phase, these are used separately at the (online) deployment stage for different purposes, as depicted in Fig. 1. Specifically, the trained channel predictor is used to robustly estimate the MIMO channel $\mathbf{h}$ from the received pilot signal $\mathbf{y}$ via $\hat{\mathbf{h}} = f_\theta(\mathbf{y};\mathbf{S})$. On the other hand, the trained pilot optimizer is not used directly; instead, its learned weights are utilized as the optimized pilot signal.

IV. Simulation Results

In this section, we present the simulation results to validate the performance and effectiveness of the proposed DL model.

A. Simulation Setups

In the simulations, the estimated channel and noise covariance matrices, $\hat{\mathbf{C}}_{\mathbf{h}}$ and $\hat{\mathbf{C}}_{\mathbf{z}}$, are obtained by the well-known exponential model as follows [46]:

$$\big[\hat{\mathbf{C}}_{\mathbf{h}}\big]_{m,n} = \rho_{\mathbf{h}}^{|m-n|}, \quad 0 \le m,n \le MN-1, \quad (25)$$

$$\big[\hat{\mathbf{C}}_{\mathbf{z}}\big]_{n,l} = \sigma^2\rho_{\mathbf{z}}^{|n-l|}, \quad 0 \le n,l \le NL-1, \quad (26)$$

where $0 \le \rho_{\mathbf{h}} \le 1$ and $0 \le \rho_{\mathbf{z}} \le 1$ denote the channel and noise correlation coefficients, respectively. Also, σ^2 denotes a parameter such that the system SNR is defined as $\frac{P}{\sigma^2 NL}$. The spectral norm-bounded covariance uncertainty sets are considered here. Specifically, the values of $\mathbf{E}_{\mathbf{h}}$ and $\mathbf{E}_{\mathbf{z}}$ are randomly generated such that $\mathbf{E}_{\mathbf{h}} \sim \mathcal{W}(MN, \mathbf{I}_{MN})$ and $\mathbf{E}_{\mathbf{z}} \sim \mathcal{W}(NL, \mathbf{I}_{NL})$, respectively,[8] and then, they are normalized such that $\sqrt{\lambda_{\max}(\mathbf{E}_{\mathbf{h}}^{\mathsf{H}}\mathbf{E}_{\mathbf{h}})} = \epsilon_{\mathbf{h}}$ and $\sqrt{\lambda_{\max}(\mathbf{E}_{\mathbf{z}}^{\mathsf{H}}\mathbf{E}_{\mathbf{z}})} = \epsilon_{\mathbf{z}}$, respectively, where the values of $\epsilon_{\mathbf{h}}$ and $\epsilon_{\mathbf{z}}$ are chosen such that $\epsilon_{\mathbf{h}} = \beta_{\mathbf{h}}\sqrt{\lambda_{\max}\big(\hat{\mathbf{C}}_{\mathbf{h}}^{\mathsf{H}}\hat{\mathbf{C}}_{\mathbf{h}}\big)}$ and $\epsilon_{\mathbf{z}} = \beta_{\mathbf{z}}\sqrt{\lambda_{\max}\big(\hat{\mathbf{C}}_{\mathbf{z}}^{\mathsf{H}}\hat{\mathbf{C}}_{\mathbf{z}}\big)}$ for $0 \le \beta_{\mathbf{h}}, \beta_{\mathbf{z}} \le 1$. Throughout the simulations, we use the following parameter settings as default unless specified otherwise: $M = N = 2$, $L = 8$, $P = 2$ W, $\rho_{\mathbf{h}} = \rho_{\mathbf{z}} = 0.7$, $\beta_{\mathbf{h}} = \beta_{\mathbf{z}} = 0.2$, and SNR $= 10$ dB.

In the training phase, the proposed DL model is trained with 10^5 samples of the compensated channel and noise vectors, $\bar{\mathbf{h}}$ and $\bar{\mathbf{z}}$, drawn from $\bar{\mathbf{h}} \sim \mathcal{CN}(\mathbf{0}, \bar{\mathbf{C}}_{\mathbf{h}})$ and $\bar{\mathbf{z}} \sim \mathcal{CN}(\mathbf{0}, \bar{\mathbf{C}}_{\mathbf{z}})$, respectively. During the training (resp. after the training), the performance of the proposed DL model is validated through the validation step (resp. evaluated through the test step) using 3×10^4 samples of the actual channel and noise vectors, $\mathbf{h}$ and $\mathbf{z}$, drawn from $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_{\mathbf{h}})$ and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_{\mathbf{z}})$, respectively.

B. Ablation Studies

We first conduct the ablation studies to examine the effects of parameters and configurations of the proposed DL model and to gain the relevant design insights. Figs. 5(a) and 5(b) show the training and validation performance of the proposed DL model, respectively, for various values of the step size γ when trained over 10^3 epochs with a mini-batch size of 10^3 (i.e., totally over 10^5 iterations). As can be seen from Figs. 5(a) and 5(b), larger values of γ such as $\gamma = 0.1$ and $\gamma = 0.01$ result in unstable learning with poor generalization behaviors, whereas smaller values of γ such as $\gamma = 0.001$ and $\gamma = 0.0001$ yield stable learning with good generalization behaviors. Also, it can be observed that although the training with $\gamma = 0.0001$ exhibits a slower convergence behavior than that with $\gamma = 0.001$, the former has better generalization performance than the latter due to less overfitting. For this reason, in the subsequent simulations, we use $\gamma = 0.0001$ when training the proposed DL model.

In Fig. 6, we compare the performance of the proposed DL model to that of the following variants:

(i) The proposed DL model without (w/o) the pilot optimizer, in which the channel predictor is solely employed

[8]It is practically valid and reasonable to use the Wishart distributions to model the distributions of the (sample) estimation error covariance matrices of the channel and noise when the corresponding estimation errors follow the Gaussian distributions. Meanwhile, the proposed DL model can still be used even when the estimation error covariance matrices follow any other distributions.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3324667

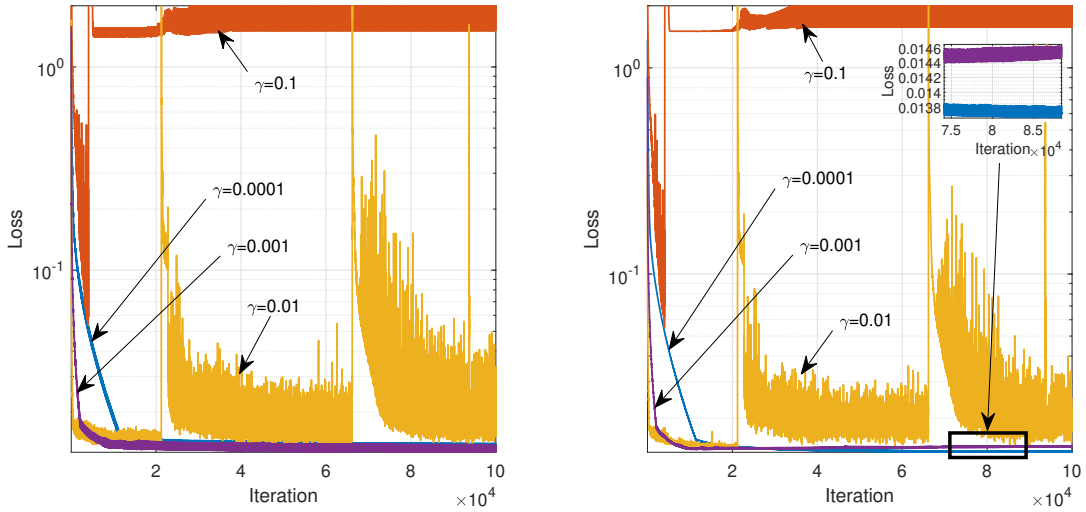

Fig. 5. Learning curves for the proposed DL model with different step sizes. (a) Training performance. (b) Validation performance.

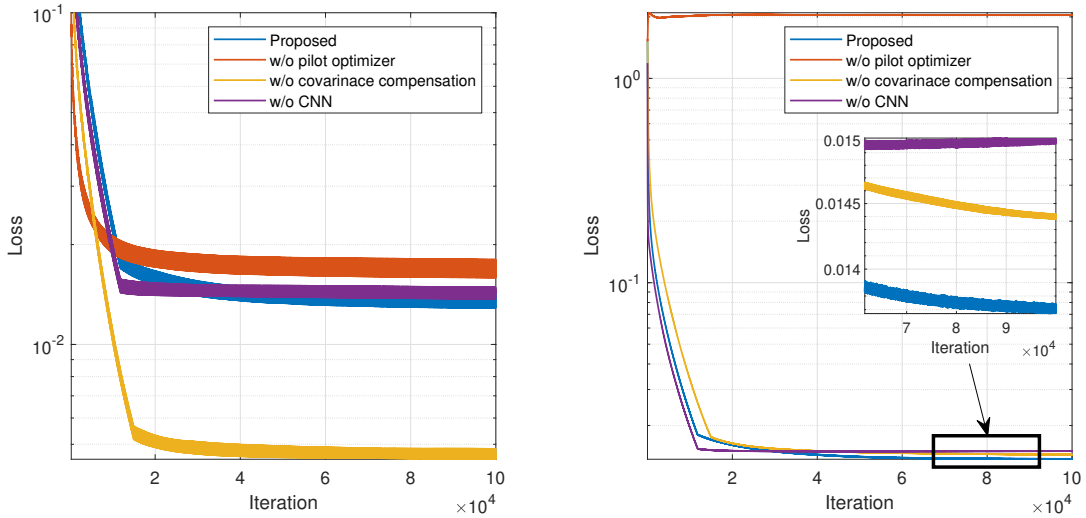

Fig. 6. Learning curves for the proposed DL model with different configurations. (a) Training performance. (b) Validation performance.

by adopting the orthogonal pilot signal with the transmission power equal to P, i.e., $\mathbf{S}^{\mathsf{H}}\mathbf{S} = \frac{P}{M}\mathbf{I}_M$.

(ii) The proposed DL model trained without the covariance compensation, in which the proposed DL model is trained with the samples of the uncompensated channel and noise vectors drawn from $\mathcal{CN}(\mathbf{0}, \hat{\mathbf{C}}_{\mathbf{h}})$ and $\mathcal{CN}(\mathbf{0}, \hat{\mathbf{C}}_{\mathbf{z}})$, respectively.

(iii) The proposed DL model without the CNN in the channel predictor, in which K convolutional and pooling layers are replaced by K fully-connected layers, where the number of learnable parameters in each fully-connected layer is set to be the same as that in each convolutional layer.

From Fig. 6, it can be observed that the proposed DL model provides the best generalization performance (although all of the proposed DL model and its variants are stably trainable), thereby demonstrating the effectiveness and completeness of the proposed network architecture and training strategy. Particularly, solely using the channel predictor without the pilot optimizer is never effective due to severe overfitting with very poor generalization performance.

In Fig. 7, the pilot signal learned by the proposed DL model is visualized. In Fig. 7(a) and 7(b), the weights of the trained pilot optimizer and the Gram matrix $\mathbf{S}^{\mathsf{H}}\mathbf{S}$ of the optimized pilot signal (i.e., weight matrix of the trained pilot optimizer in the proposed DL model) are shown for various SNR values, respectively. From 7(a), it can be seen that the design or optimization pattern of the pilot signal varies depending on the SNR value. Also, from 7(b), we can observe that as the SNR increases, the off-diagonal values of the Gram matrix and

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3324667

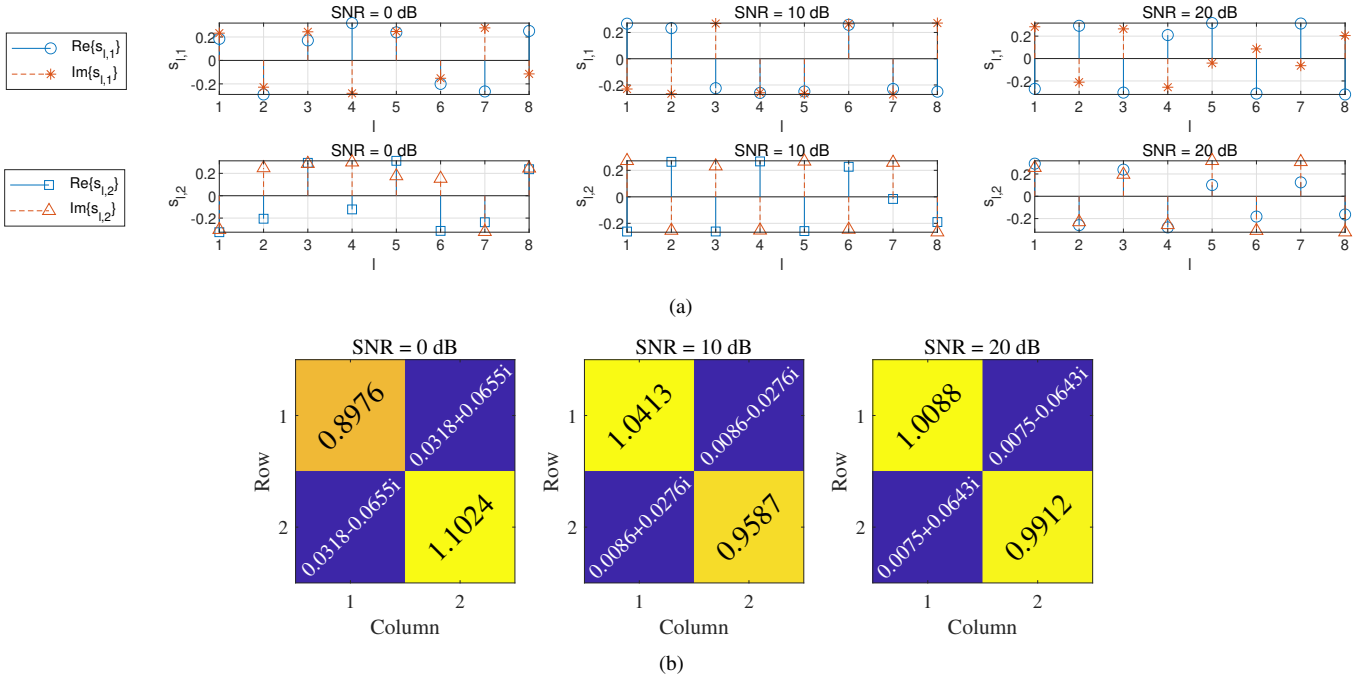

Fig. 7. Visualization of the pilot signal learned by the proposed DL model. (a) Weights of the trained pilot optimizer. (b) Gram matrix of the optimized pilot signal.

the variance of the diagonal values get smaller, meaning that the pilot signal learned by the proposed DL model becomes less correlated with less crosstalk or interference, which agrees well with the intuition.

C. Performance Comparisons

Now, we compare the performance of the proposed DL model with that of the following baseline schemes:

- *Baseline Scheme I:* This scheme corresponds to a non-robust LMMSE channel estimation with the estimated channel and noise covariance matrices, in which the MIMO channel is estimated as

$$\hat{\mathbf{h}} = \hat{\mathbf{C}}_{\mathbf{h}}\big(\mathbf{S}^{\mathsf{H}} \otimes \mathbf{I}_N\big)\Big[\big(\mathbf{S} \otimes \mathbf{I}_N\big)\hat{\mathbf{C}}_{\mathbf{h}}\big(\mathbf{S}^{\mathsf{H}} \otimes \mathbf{I}_N\big) + \hat{\mathbf{C}}_{\mathbf{z}}\Big]^{-1}\mathbf{y}. \tag{27}$$

  Also, the pilot signal is set such that $\mathbf{S}^{\mathsf{H}}\mathbf{S} = \frac{P}{M}\mathbf{I}_M$.
- *Baseline Scheme II:* This scheme is an extension of the existing scheme in [20, Theorem 2] to the case with both the channel and noise covariance uncertainties, in which the MIMO channel is estimated as in (27) with $\hat{\mathbf{C}}_{\mathbf{h}}$ and $\hat{\mathbf{C}}_{\mathbf{z}}$ replaced by $\hat{\mathbf{C}}_{\mathbf{h}} + \epsilon_{\mathbf{h}}\mathbf{I}_{MN}$ and $\hat{\mathbf{C}}_{\mathbf{z}} + \epsilon_{\mathbf{z}}\mathbf{I}_{NL}$ respectively. Also, the pilot signal is set such that $\mathbf{S}^{\mathsf{H}}\mathbf{S} = \begin{bmatrix}\mathbf{\Lambda} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}$, where $\mathbf{\Lambda}$ is a $\nu \times \nu$ diagonal matrix ($\nu \leq M$) such that $\mathbf{\Lambda} = \mu\mathbf{I}_\nu - (\mathbf{\Sigma} + \epsilon_{\mathbf{h}}\mathbf{I}_\nu)^{-1}$. Herein, μ is a smallest value of i such that $\frac{P+\sigma^2\mathrm{Tr}[(\mathbf{\Sigma}+\epsilon_{\mathbf{h}}\mathbf{I}_\nu)^{-1}]}{\sigma^2 M} > \frac{1}{[\mathbf{\Sigma}]_{i,i}+\epsilon_{\mathbf{h}}}$ and $\mathbf{\Sigma}$ is a $\nu \times \nu$ diagonal matrix containing ν largest eigenvalues of $\hat{\mathbf{C}}_{\mathbf{h}}$ in descending order on the diagonal.
- *Baseline Scheme III:* This scheme is the LS channel estimation with no knowledge of the channel and noise covariance matrices, in which the MIMO channel is estimated as

$$\hat{\mathbf{h}} = \big(\mathbf{S}^{\dagger} \otimes \mathbf{I}_N\big)\mathbf{y}. \tag{28}$$

  Also, the pilot signal is set such that $\mathbf{S}^{\mathsf{H}}\mathbf{S} = \frac{P}{M}\mathbf{I}_M$.

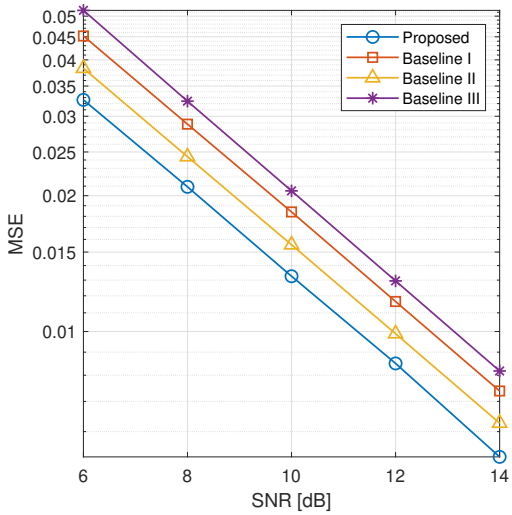

Fig. 8. Channel estimation MSEs of the proposed and baseline schemes over different SNR values.

In Fig. 8, the channel estimation MSEs of the proposed and baseline schemes are shown versus the SNR. From Fig. 8, it can be observed that the proposed scheme consistently outperforms the baseline schemes over the entire SNR range, indicating that the proposed DL model effectively copes with the channel and noise covariance uncertainties via the

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3324667

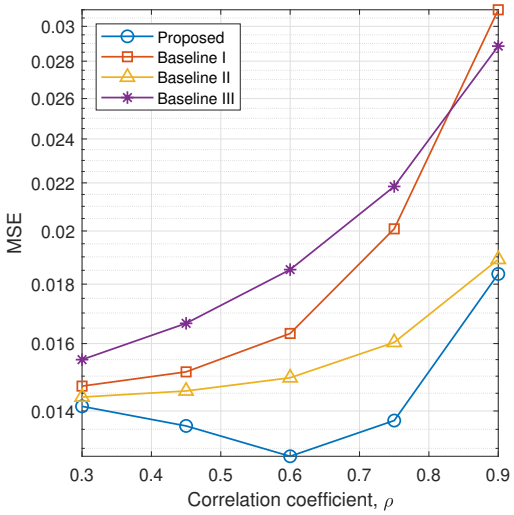

Fig. 9. Channel estimation MSEs of the proposed and baseline schemes over different values of ρ.

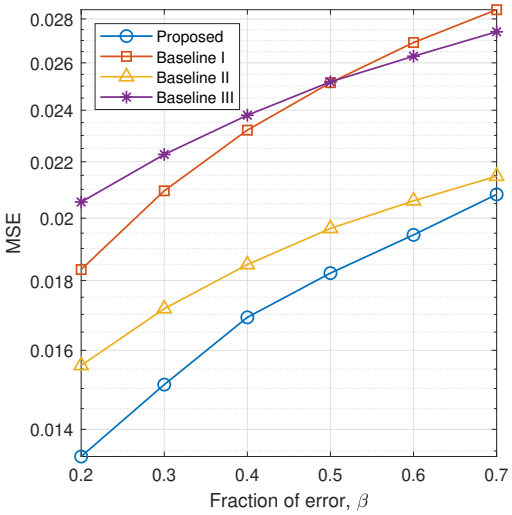

Fig. 10. Channel estimation MSEs of the proposed and baseline schemes over different values of β.

covariance compensation strategy during the training. The baseline scheme III performs worst mainly due to the noise amplification issue and lack of utilizing the channel and noise covariance matrices for the channel estimation. Even though the baseline schemes I and II result in better performance than the baseline scheme III, their performance is still marginal due to the mismatches between the actual and estimated covariance matrices. Overall, the results of Fig. 8 reveal that properly utilizing the channel and noise statistical information and overcoming the uncertainties in such information play crucial roles in improving the performance of the MIMO channel estimation in practice.

In order to investigate the impacts of the strengths of the channel and noise correlations on the channel estimation performance, in Fig. 9, the values of $\rho_{\mathbf{h}}$ and $\rho_{\mathbf{z}}$ are set to be the same as ρ (i.e., $\rho = \rho_{\mathbf{h}} = \rho_{\mathbf{z}}$). Then we depict the channel estimation MSEs of the various schemes as functions of the (common) correlation coefficient ρ. From Fig. 9, we can observe that the performance of all the baseline schemes degrades as ρ increases and very large values of ρ eventually result in the baseline scheme I performing even worse than the baseline scheme III, meaning that the adverse impacts of the channel and noise covariance uncertainties on the channel estimation become (much) more severe in (extremely) strongly correlated environments. On the other hand, the performance of the proposed scheme initially improves until about $\rho = 0.6$ and then degrades, suggesting that the robust channel estimation with the proposed DL model will be most effective in moderately correlated environments.

We further investigate the impacts of the degrees of the channel and noise covariance uncertainties on the channel estimation performance by introducing and controlling a parameter β such that $\beta = \beta_{\mathbf{h}} = \beta_{\mathbf{z}}$. In Fig.10, the channel estimation MSEs of the proposed and baseline schemes are shown for different values of β. It can be seen from Fig. 10 that as β increases, the performance of all the schemes degrades,

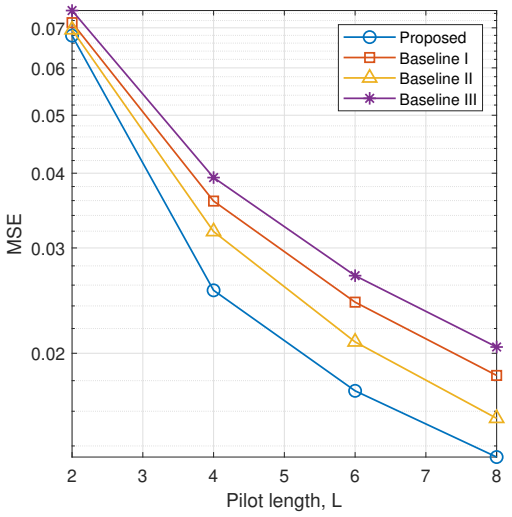

Fig. 11. Channel estimation MSEs of the proposed and baseline schemes over different values of L.

as expected, because there are more uncertainties (or errors) in the channel and noise covariance matrices. Nevertheless, the proposed scheme still performs better than the other schemes and the performance gaps are more pronounced for medium values of β. Therefore, the proposed DL model will be indeed very useful in practical IoT applications where only a coarse (not fine-grained) estimation of the channel and noise covariance matrices is possible with insufficient or erroneous samples.

In Fig. 11, the channel estimation performance of the various schemes is shown versus L to examine the effects of the pilot length. Also, in Fig. 12, setting $\mu = M = N$, we plot the channel estimation performance of the various schemes for different values of μ to investigate the effects of the number of antennas. As can be seen from Fig. 11 (resp. Fig. 12), the

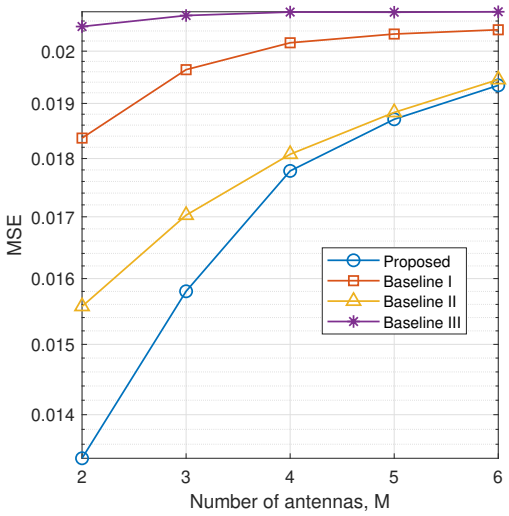

Fig. 12. Channel estimation MSEs of the proposed and baseline schemes over different values of M.

performance of all the schemes improves when L increases (resp. M decreases) since the effect of noise is reduced given the same SNR in our simulation setting (resp. the number of channel coefficients to be estimated increases given the same resources). Nonetheless, the proposed scheme is observed to still surpass the other schemes, where the performance gaps are more pronounced as L increases or M decreases.

D. Computational Complexity Analysis

Lastly, we analyze and compare the computational complexities of the proposed and baseline schemes in terms of the training and inference complexities, which are summarized in Table I.

1) Inference Complexity: The inference complexities of the baseline schemes I and II are all dominated by the computation of the LMMSE channel estimate in (27), which require $\mathcal{O}(LMN^2 + N^3L^3)$ [47]. Similarly, the inference complexity of the baseline scheme III is given by $\mathcal{O}(LMN^2 + L^3)$ [47]. On the other hand, for the interference of the proposed scheme, the feedforward computations of the pilot layer, K convolutional/pooling layers, and the fully-connected layer need to be sequentially performed, of which computational complexities are $\mathcal{O}\left(LMN^2\right)$, $\mathcal{O}\left(LN\sum_{k=1}^{K}\ell_k c_{k-1}c_k\right)$, and $\mathcal{O}\left(MNn_{\mathsf{F}}\right)$, respectively [38], [43], where n_{F} denotes the size of the input in the fully-connected layer of the channel predictor. Thus, the total inference complexity of the proposed scheme is estimated as $\mathcal{O}\left(LMN^2 + LN\sum_{k=1}^{K}\ell_k c_{k-1}c_k + MNn_{\mathsf{F}}\right)$.

2) Training Complexity: The training of the proposed scheme can be done via the backpropagation algorithm, which requires to perform multiple iterations of forward and backward computations [43]. As analyzed just before, one iteration of the forward computation of the proposed scheme requires the computational complexity of $\mathcal{O}\left(LMN^2 + LN\sum_{k=1}^{K}\ell_k c_{k-1}c_k + MNn_{\mathsf{F}}\right)$. Also, the backward computation of the proposed scheme in one iteration has the similar complexity to that of the forward computation [43], which is thus still given by $\mathcal{O}\left(LMN^2 + LN\sum_{k=1}^{K}\ell_k c_{k-1}c_k + MNn_{\mathsf{F}}\right)$. Overall, the training complexity of the proposed scheme is estimated as $\mathcal{O}\left(N_{\text{iter}}\left(LMN^2 + LN\sum_{k=1}^{K}\ell_k c_{k-1}c_k + MNn_{\mathsf{F}}\right)\right)$, where N_{iter} denotes the number of iterations.

Importantly and intriguingly, the above complexity analysis implies that the inference complexity of the proposed DL model can be even lower than those of the baseline schemes if the values of the parameters are properly chosen. For example, when $\frac{\sum_{k=1}^{K}\ell_k c_{k-1}c_k}{N^2L^2} + \frac{Mn_{\mathsf{F}}}{N^2L^3} \leq 1$, the proposed scheme has clearly a lower inference complexity than the baseline schemes I and II.

V. Conclusion

This paper investigated the robust channel estimation problem for the MIMO-aided IoT system in the presence of the channel and noise covariance uncertainties, to solve which the novel DL model composed of the two modules, the pilot optimizer and channel predictor, was proposed. The effective training strategy for the proposed DL model was also devised by properly compensating the channel and noise covariance matrices such that the adverse impacts of the underlying uncertainties were overcome. The extensive simulation results confirmed that the proposed DL model performed better and is more effective than the baseline schemes, rendering it highly useful in practice.

As an interesting and important focus of future research, it is also deserved to investigate the robust channel estimation problem for other wireless communication systems, e.g., with orthogonal frequency division multiplexing (OFDM), large antenna arrays, reconfigurable intelligent surfaces (RIS), and/or satellite-terrestrial links.

References

[1] D. C. Nguyen *et al.*, "6G Internet of Things: A comprehensive survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 359–383, Jan. 2022.

[2] K. Shafique, B. A. Khawaja, F. Sabir, S. Qazi, and M. Mustaqim, "Internet of Things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios," *IEEE Access*, vol. 8, pp. 23022–23040, 2020.

[3] J. Bian, A. Arafat, H. Xiong, J. Li, L. Li, H. Chen, J. Wang, D. Dou, and Z. Guo, "Machine Learning in Real-Time Internet of Things (IoT) Systems: A Survey," *IEEE Internet of Things J.*, vol. 9, no. 11, pp. 8364–8386, Jun. 2022.

[4] D. N. C. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[5] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and V. H. Poor, *MIMO Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[6] J. H. Kotecha and A. M. Sayeed, "Transmit signal design for optimal estimation of correlated MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 546–557, Feb. 2004.

[7] T. F. Wong and B. Park, "Training sequence optimization in MIMO systems with colored interference," *IEEE Trans. Commun.*, vol. 52, no. 11, pp. 1939–1947, Nov. 2004.

[8] M. Biguesh and A. B. Gershman, "Training based MIMO channel estimation: A study of estimator tradeoffs and optimal training signals," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 884–893, Mar. 2006.

[9] Y. Liu, T. Wong, and W. Hager, "Training signal design for estimation of correlated MIMO channels with colored interference," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1486–1497, Apr. 2007.

TABLE I
COMPUTATIONAL COMPLEXITIES OF PROPOSED AND BASELINE SCHEMES

Scheme	Training complexity	Inference complexity
Proposed	$\mathcal{O}\left(N_{\text{iter}}\left(LMN^2 + LN\sum_{k=1}^{K}\ell_k c_{k-1}c_k + MNn_{\mathsf{F}}\right)\right)$	$\mathcal{O}\left(LMN^2 + LN\sum_{k=1}^{K}\ell_k c_{k-1}c_k + MNn_{\mathsf{F}}\right)$
Baseline I	Not required	$\mathcal{O}(LMN^2 + N^3L^3)$
Baseline II	Not required	$\mathcal{O}(LMN^2 + N^3L^3)$
Baseline III	Not required	$\mathcal{O}(LMN^2 + L^3)$

[10] D. Katselis *et al.*, "On training optimization for estimation of correlated MIMO channels in the presence of multiuser interference," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4892–4904, Oct. 2008.

[11] M. Biguesh and M. Gazor, "Optimal training sequence for MIMO wireless systems in colored environments," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3144–3153, Aug. 2009.

[12] E. Björnson and B. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1807–1820, Mar. 2010.

[13] S. Wang, S. Ma, C. Xing, S. Gong, J. An, and H. V. Poor, "Optimal training design for MIMO systems with general power constraints," *IEEE Trans. Signal Process.*, vol. 66, no. 14, pp. 3649–3664, Jul. 2018.

[14] K. Sharma, " Characterization and modeling of MIMO wireless channels based on correlation tensor," *Computers and Mathematics with Applications*, vol. 64, no. 2, pp. 89–101, Jul. 2012.

[15] L. Pradell, A. Comeron, and A. Rarnirez, "A general analysis of errors in noise measurement systems," in *Proc. of the 18th European Microwave Conference*, Sep. 1988, pp. 924–929.

[16] N. Shariati, J. Wang, and M. Bengtsson, "A robust MISO training sequence design," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2013, pp. 4564–4568.

[17] C.-T. Chiang and C. C. Fuang, "Robust training sequence design for spatially correlated MIMO channel estimation," *IEEE Trans. Veh. Technol.*, vol. 60, no. 7, pp. 2882–2894, Sep. 2011.

[18] N. Shariati, J. Wang, and M. Bengtsson, "On robust training sequence design for correlated MIMO channel estimation," in *Proc. IEEE ACSSC*, Nov. 2012, pp. 504–507.

[19] N. Shariati and M. Bengtsson, "Robust training sequence design for spatially correlated MIMO channels and arbitrary colored disturbance," in *Proc. IEEE PIMRC*, Sep. 2011, pp. 1939–1943.

[20] N. Shariati, J. Wang, and M. Bengtsson, "Robust training sequence design for correlated MIMO channel estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 107–120, Jan. 2014.

[21] S. Gong, S. Wang, C. Xing, S. Ma, and T. Q. Quek, "Robust superimposed training optimization for UAV assisted communication systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1704–1721, Mar. 2020.

[22] B. Rong, Z. Zhang, X. Zhao, and X. Yu, "Robust superimposed training designs for MIMO relaying systems under general power constraints," *IEEE Access*, vol. 7, pp. 80404–80420, Jun. 2019.

[23] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.

[24] S. Dörner, S. Cammerer, J. Hoydis, and S. ten Brink, "Deep learning based communication over the air," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 132–143, Feb. 2018.

[25] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.

[26] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

[27] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[28] J.-M. Kang, C.-J. Chun, and I.-M. Kim, "Deep learning based channel estimation for wireless energy transfer," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2310–2313, Nov. 2018.

[29] J. M. Kang, I. M. Kim, and C. J. Chun, "Deep learning-based MIMO-NOMA with imperfect SIC decoding," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3414–3417, Sep. 2020.

[30] J.-M. Kang, C.-J. Chun, I.-M. Kim, and D. I. Kim, "Deep RNN-based channel tracking for wireless energy transfer system," *IEEE Syst. J.*, vol. 14, no. 3, pp. 4340–4343, Sep. 2020.

[31] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, Apr. 2021.

[32] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 852–855, Oct. 2018.

[33] 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," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8549–8560, Sep. 2018.

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

[35] Y. Yang, F. Gao, G. Y. Li, and M. Jian, "Deep learning-based downlink channel prediction for FDD massive MIMO system," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 1994–1998, Nov. 2019.

[36] C. Chun, J.-M. Kang, I.-M. Kim, "Deep learning-based channel estimation for massive MIMO systems," *IEEE Wireless Commun. Letters*, vol. 8, no. 4, pp. 1228–1231, Aug. 2019.

[37] C.-J. Chun, J.-M. Kang, and I.-M. Kim, "Deep learning based joint pilot design and channel estimation for multiuser MIMO channels," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 1999–2003, Nov. 2019.

[38] J.-M. Kang, C.-J. Chun, and I.-M. Kim, "Deep learning based channel estimation for MIMO systems with received SNR feedback," *IEEE Access*, vol. 8, pp. 121162–121181, Jul. 2020.

[39] M. B. Mashhadi and D. Gunduz, "Pruning the pilots: Deep learning-based pilot design and channel estimation for MIMO-OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6315—6328, Oct. 2021.

[40] H. Lutkepohl, *Handbook of Matrices*. New York: Wiley, 1996.

[41] K. Hornik *et al.*, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[42] S. Haykin and K. Elektroingenieur, *Neural Networks and Learning Machines*. Upper Saddle River, NJ, USA: Prentice-Hall, 2009

[43] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, May 2015, pp. 1–14.

[45] D.-A. Clevert *et al.*, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. ICLR*, May. 2016, pp. 1–14.

[46] S. L. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 369–371, Sep. 2001.

[47] R. Hunger, "Floating point operations matrix-vector calculus," Inst. Circuit Theory Signal Process., Munich Univ. Technol., Munich, Germany, Tech. Rep., 2005.

Jae-Mo Kang (M'19) received the Ph.D. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2017. He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada, and was an Assistant Professor with the School of Intelligent Mechatronics Engineering, Sejong University, Seoul, South Korea. He is currently an Associate Professor with the Department of Artificial Intelligence, Kyungpook National University, Daegu, South Korea. His research interests include IoT, 6G, deep learning, training design, and computer vision.