Deep Learning for Massive MIMO Uplink Detectors

Mahmoud A. Albreem, *Senior Member, IEEE,* Alaa Alhabbash, *Member, IEEE,*
Shahriar Shahabuddin, *Member, IEEE,* and Markku Juntti, *Fellow, IEEE*

Abstract*—Detection techniques for massive multiple-input multiple-output (MIMO) have gained a lot of attention in both academia and industry. Detection techniques have a significant impact on the massive MIMO receivers' performance and complexity. Although a plethora of research is conducted using the classical detection theory and techniques, the performance is deteriorated when the ratio between the numbers of antennas and users is relatively small. In addition, most of classical detection techniques are suffering from severe performance loss and/or high computational complexity in real channel scenarios. Therefore, there is a significant room for fundamental research contributions in data detection based on the deep learning (DL) approach. DL architectures can be exploited to provide optimal performance with similar complexity of conventional detection techniques. This paper aims to provide insights on DL based detectors to a generalist of wireless communications. We garner the DL based massive MIMO detectors and classify them so that a reader can find the differences between various architectures with a wider range of potential solutions and variations. In this paper, we discuss the performance-complexity profile, pros and cons, and implementation stiffness of each DL based detector's architecture. Detection in cell-free massive MIMO is also presented. Challenges and our perspectives for future research directions are also discussed. This article is not meant to be a survey of a mature-subject, but rather serve as a catalyst to encourage more DL research in massive MIMO.

***Index Terms*—Massive MIMO, detection, deep learning, detection networks, message passing, sphere decoding, cell-free massive MIMO, deep convolutional neural networks**

I. Introduction

communications systems are developed from the first to the fifth generation (5G) and have propelled to the sixth generation (6G) to offer advanced wireless communications services, such as the autonomous driving and massive internet of things (IoT). The total of 2.8 billion subscriptions are expected by the end of 2025 [1]. By 2022, the business IP traffic is expected to reach more than 63 exabytes/month [2]. With the tremendous traffic growth, systems with massive multiple-input multiple-output (MIMO) have gained a lot of attention in both academia and industry where a large number of antennas at the base station (BS) is utilized to serve dozens of user terminals [3]. The classical massive MIMO is an extension of the conventional small-scale MIMO technology that was implemented since the third generation (3G) wireless systems where multiple antennas are used at the transmitter and receivers to enhance the spectral efficiency, the range, and the link reliability. In the massive MIMO system operating below 6 GHz carrier frequency, the total number of user terminals within the service area is clearly smaller than the number of antennas at the BS. The early massive MIMO technology assumed that the scattering and multipath propagation in radio channels is rich. The conventional matched filter (MF) based receivers were considered to be approximately optimal due the interference averaging. Recent works show that interference mitigation in the receiver can improve the performance and enhance the system capacity, in particular, when the massive MIMO concept is generalized to the cell-free massive MIMO system [4].

The receiver's design is not a trivial task where advanced signal processing is required when a large number of antennas is utilized. Detection techniques for the massive MIMO technology have gained a lot of attention in last years. Detectors based on linear/non-linear techniques, message passing, local search, and other techniques have been proposed in literature. The performance-complexity profile of each detection technique is highly affected by the number of antennas, the propagation and environment, the modulation scheme, and the initial solution. However, in any physical layer of communications systems, there are always several challenges [5], [6] such as:

- Channel modeling in realistic scenarios: The efficiency of wireless communications system depends significantly on a channel model that characterize the complex scenarios, imperfections, and non-linearities. It is also difficult to characterize all the channel features by simple models.
- Fast and effective signal processing: Nonlinear imperfections could appear when low-cost hardware, such as low-resolution analog-to-digital converters with low energy consumption, is utilized. Increasing carrier frequencies toward the millimeter wave (mmWave) and THz band make the hardware (HW) imperfections even more complicated to model.
- Block-structure of communications systems: The conventional communications system includes a series of blocks such as the coding, modulation, and detection. The optimal performance of the entire communications system cannot be guaranteed if each block is optimized independently. In practice, rigid joint optimization is too complex.

The optimal balance between the performance and computational complexity of the entire communications system can be achieved with a channel modeling in realistic scenarios,

M. Albreem is with Department of Electrical Engineering, University of Sharjah, Sharjah 27272, UAE, e-mail: malbreem@sharjah.ac.ae

A. Al Habbash is with Palestinian ICT Research Agency, Gaza, Palestine; email: alaa.alhabbash@asu.edu.om

S. Shahabuddin is with Mobile Networks, Nokia, Oulu 90620, Finland and Centre for Wireless Communications, University of Oulu, Oulu 90014, Finland; e-mail: firstname.lastname@nokia.com and firstname.lastname@oulu.fi

M. Juntti is with Centre for Wireless Communications, University of Oulu, Oulu 90014, Finland; e-mail: firstname.lastname@oulu.fi.

This research has been financially supported in part by the University of Sharjah under Seed Research Grant, and in part by the Research Council (TRC) of the Sultanate of Oman (agreement No. TRC/BFP/ASU/01/2018).

fast and effective signal processing, and block-structure of communications systems. However, with increasing network complexity and growing diversity, the need for automated design and optimization of the processing is clear. Therefore, embedding machine intelligence into future communications systems is drawing unparalleled research interest.

Machine learning has been a key approach to minimize costs and errors, and to increase the efficiency in many disciplines. It is a sub-field of the artificial intelligence (AI) and is described as ´´the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning´´ [7], [8]. Based on learning aspects, machine learning is classified as a supervised learning [9], unsupervised learning [10], semi-supervised learning [11], and reinforcement learning [12]. In the last few decades, machine learning has provided a significant improvement in many fields, such as image recognition [13], speech recognition [14], drug discovery [15], biomedical sciences [16], computer vision [17], signal processing, and wireless communications [5], [6], [18]–[23]. In MIMO detection, machine learning was successfully applied. For instance, the MIMO detection problem is converted into a clustering problem which is then solved by expectation-maximization algorithm [24]. In [25], the MIMO detection problem was reformulated as a least absolute shrinkage and selection operator (LASSO) problem. Then, it is optimized by a two-stage alternating direction method of multipliers (ADMM). Simulation results showed that this detector outperforms the classical detectors. However, due to a limited learning capabilities of the conventional machine learning algorithms and a high computational complexity of handling physical channels, conventional machine learning algorithms are not exploited commercially in massive MIMO [6].

In the last decade, the availability of a large amount of data, technological progress, the development of optimization techniques, the availability of powerful graphical processing units (GPUs), and huge amount of available memory have jointly laid the basis to the deep learning (DL) revolution [26] to achieve further improvements to the physical layer. The DL is a ”flagship approach” of the machine learning to improve the learning process where machines/computers are learned from experience and understand the world in terms of a hierarchy of concepts [23], [27], [28]. In other words, DL can learn features from raw data automatically and adjust the model structures flexibly to improve the performance. In massive MIMO detection, DL can be a promising approach because:

- It has a superior algorithmic learning ability despite the complex channel conditions [29]. Instead of rigid mathematical models, the communications systems are represented by learned weights in the deep networks through training methods.
- It depends on distributed and parallel computing architectures. Therefore, deep networks have a superior ability to handle explosive growth of data volume and ensure computation speed and processing capacity. In recent years, GPUs have been shown to be energy efficient with a high performance when leveraged by concurrent algorithms. Therefore, DL structures are suitable for running on GPUs [5].
- To deploy various DL architectures in wide applications, various libraries/frameworks (TensorFlow, Theano, and Caffe) have been established that accelerate experiments.

Therefore, DL architectures can be exploited for data detection in realistic scenarios by unfolding specific iterative detection techniques and for obtaining a balance between accuracy and complexity. By leveraging flexible layer structures, data detection becomes a simple forward pass through the deep networks.

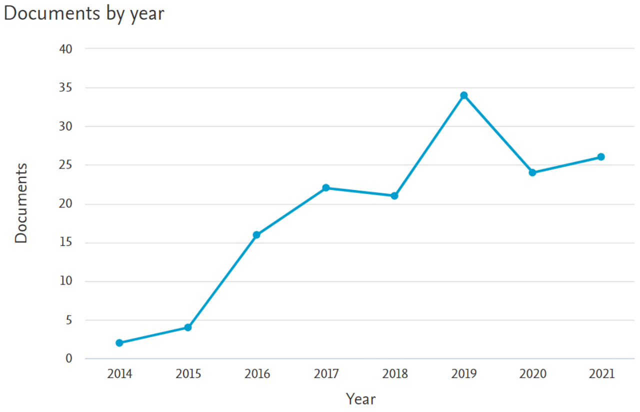

Figure 1. Number of survey articles published concerning the massive MIMO (data extracted from Scopus on October 2021).

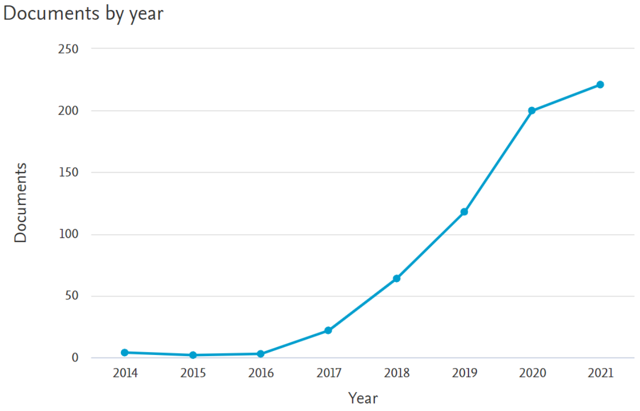

Figure 2. Number of survey articles published concerning the DL (data extracted from Scopus on October 2021).

A. Relevant Prior Art

Figure 1 shows that the literature is rich with survey papers related to massive MIMO technology [30], [31], [33]–[44]. While these papers review the work in a number of significant research areas of the massive MIMO, none of them has extensively reviewed the role of DL in massive MIMO detection techniques. In [30], the half-a-century history of MIMO detection is recited and MIMO detection fundamentals and concepts are presented. A comprehensive review of the tree-search detectors, lattice reduction, probabilistic data association, and semi-definite relaxation (SDR) detections is illustrated. Ideas and lessons behind the design of complexity-scalable MIMO systems are also presented. In addition, the

Table I
PRIOR RELEVANT SURVEYS

Reference	Linear Detectors	Nonlinear Detectors	Cell-free	FullyCon	DetNet	OAMP-Net	WeSNet	MMNet	ScNet	DCNN	CNNLASS
[30]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
[31]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
[32]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
[8]	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗
This work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

paper has concluded that the conventional MIMO detection algorithms might be infeasible with a certain massive MIMO scenario. Although the article is extensive, the primary focus was not in massive MIMO systems. We [31] garner the massive MIMO detection algorithms and classify them so that the reader can find the differences between various linear/nonlinear detection algorithms. The performance and complexity trade-off and the practical implementation of detection algorithms are comprehensively reviewed. Detectors based on the local search, belief propagation (BP), BOX-detection, and approximate/avoid matrix inversion methods are presented. The article has also reviewed the pros and cons of each detector based on the performance-complexity profile and the implementation stiffness. Although the paper has mentioned that the machine learning is a promising approach in the design of massive MIMO detectors, it is not well-investigated and only few research papers are cited. The paper has concluded that massive MIMO detection techniques will remain a hot research area in coming years and the DL approach will play a crucial role in developing efficient massive MIMO receivers.

In [33], the measurement and modeling of massive MIMO channels are classified and analyzed. The paper has surveyed major techniques of both physical and network layers in massive MIMO systems. The performance of massive MIMO in real propagation environments is investigated in [45]. A survey in [36] has summarized the requirements of channel modeling and provided a review of channel measurement and models. An extensive review of massive MIMO propagation channels, channel modeling approaches, and characteristics has been presented in [37]. The paper has concluded that the massive MIMO propagation channels will remain a hot research area in coming years and they are essential for developing efficient massive MIMO technologies. In [38], an overview of massive MIMO propagation characteristics is discussed on the context of 5G channel modeling. A comprehensive survey on the sources of pilot contamination in massive MIMO is presented in [39]. In addition, the effect of pilot contamination on the performance and complexity is extensively discussed. Techniques to mitigate the pilot contamination are categorized in pilot-based approach and subspace-based approach. Various aspects of pilot contamination such as advantages and limitations of different mitigation methods are discussed in [41]. Moreover, numerical evaluation methods of different decontamination methods are presented. In [40], hybrid beamforming structures using instantaneous or average channel state information (CSI) are comprehensively reviewed. Constraints and limitations of mmWave bands due to propagation scenarios and hardware impairments are also discussed. The paper has concluded that there is no single structure that obtains the ”best” balance between the performance and the complexity in all applications. In [34], design issues and practical implementations of linear precoding algorithms are comprehensively discussed for downlink (DL) transmission under both single-cell (SC) and multi-cell (MC) scenarios. In [35], a frequency synchronisation in massive MIMO is reviewed. The article studied the adjustment of the clock frequency of local nodes to the clock frequency offset. Authors provide an extensive classification of frequency synchronisation where different antenna architectures and modulation schemes are considered. It is concluded that frequency synchronisation techniques are urgently required to make the practical implementation of massive MIMO feasible.

Massive MIMO is currently a mature technology and has made its way into the 5G standards. In order to improve the user experience, guarantee a high connectivity, and reduce inter-cell interference, the concept of massive MIMO is generalized to the cell-free massive MIMO system where a large number of access points (APs) is distributed over a large geographical area to jointly and coherently serve a much smaller number of user equipments (UEs) on the same time-frequency resource [46]–[48]. It enables coherent user-centric transmission realized by distributed massive MIMO systems [38], [49]. Therefore, cell-free massive MIMO is considered as a promising technology for beyond 5G (B5G) systems with its features such as higher spectral efficiency and superior spatial diversity as compared to traditional MIMO systems [50]. In [51], the phenomenology associated with precoding techniques and power allocation algorithms in cell-free massive MIMO is discussed. In [4], a framework for scalable cell-free massive MIMO systems is proposed where the complexity and signaling at each AP is finite even when the number of UEs goes to infinity. In [52], the performance of cell-free massive MIMO systems with the minimum mean square error (MMSE) algorithm and large scale fading decoding (LSFD) receivers is investigated. A joint coherent signal processing based on the MMSE algorithm for channel estimation and data detection in cell-free massive MIMO systems is presented in [53]. However, cell-free massive MIMO has several challenges posed by several design aspects such as the detection complexity, hardware implementation, channel estimation, channel hardening, and security aspect [49], [54], [55].

In communications systems, machine learning is utilized for a systematic mining and extraction of valuable information from traffic data to automatically find the correlations that would otherwise be too complex for human experts [8]. Insights on current research on the machine learning in communications networks with future research challenges are comprehensively presented in [8]. In [23], a study of DL applications for mobile networks is illustrated. DL architectures to enhance the performance of wireless networks are presented in [21]. In addition, utilization of the DL to improve network functions like network security and sensing data compression is comprehensively discussed. Since 2017, it is shown that the machine learning can also be exploited to enhance the massive MIMO performance and computational complexity [8]. Recently, DL models have attracted tremendous attention from researchers in various fields such as medical imaging [56], pattern recognition [57], biological data classification [58], intelligent transportation [59], and industrial processes [60]. It can also play a significant role in B5G networks such as cell-free massive MIMO, beamspace massive MIMO, and intelligent reflecting surfaces [61]. The concepts of DL models in mobile networks are comprehensively discussed in [23]. In [62], neural network is utilized in the physical layer to classify the signals based on cyclic spectral analysis and pattern recognition with and without prior knowledge of the bandwidth and the carrier. In [63]–[65], neural network is exploited in the channel modeling and identification. Recently, deep neural network approaches are utilized in channel estimation for the massive MIMO [66]–[78]. In [42], a broad survey on security for 5G and Beyond 5G (B5G) networks is presented. This survey has discussed the security vulnerabilities and provided solutions to specific threats. In [43], opportunities and advantages to utilize the artificial intelligence and machine learning in 5G network security are comprehensively discussed. It is shown that most of the machine learning concepts are taken from mature technologies such as robotics and computer vision as it is and utilized in 5G networks, and hence, many challenges are risen [44]. In [79], a fast and flexible denoising convolutional neural network (FFDNet) for channel estimation of cell-free massive MIMO is proposed to deal with multiple noise levels. A DL architecture for limited-fronthaul cell-free massive MIMO using a heuristic sub-optimal scheme is exploited to convert the power allocation problem into a standard geometric programme [80]. In [81], a cascade of two DL networks is utilized for reciprocity calibration and obtaining the complete channel estimate for precoding purposes. In [82] [83], a DL network is exploited to perform power allocation in the UL of a cell-free massive MIMO.

Figure 2 shows that there is a tremendous increase in the number of published survey articles in DL approach. Despite growing interest in the DL for 5G and B5G, the existing contributions are scattered across different research areas and a comprehensive survey in the DL for detection techniques in 5G/B5G is lacking. Up to our knowledge, this is the first article to comprehensively study the DL approach in massive MIMO networks. Table I presents the differences between the current paper and other relevant prior surveys in data detection for massive MIMO and DL architectures. Unlike other surveys/tutorials, this paper is illustrating the detection based on DL architectures. In addition, the role of conventional linear and nonlinear techniques with DL architectures is also illustrated. Data detection based on deep convolutional neural networks (DCNN) is also considered.

B. Contribution and Outline

This paper presents a comprehensive survey on the DL approach for massive MIMO data detection techniques. Our particular focus is on performance, computational complexity, and the potential of realization of data detection techniques based on the DL. Although detection techniques for massive MIMO are comprehensively illustrated in [31], there is a paucity of reviews on advanced detectors based on the DL. To our best knowledge, this is the first survey to explore the DL approach for massive MIMO detectors. Although a plethora of DL based massive MIMO detectors has been proposed in the literature since 2017, it is not considered as a mature-subject yet. This paper aims to serve as a catalyst to encourage more DL research efforts in massive MIMO. The major contributions of this article are summarized as:

- This paper presents the limitations of conventional massive MIMO detectors. It starts off with a dive into a history of massive MIMO detection techniques and gives an overview of conventional massive MIMO detectors such as linear detectors, the BP, the conjugate gradient (CG), and the sphere decoding (SD).
- This paper provides insights on massive MIMO detection based on DL architectures to a generalist of wireless communications. In this paper, we garner the massive MIMO detectors based on the DL and their variations. We present their performance-complexity profile, pros and cons, and implementation stiffness so that a reader can find a distinction between different detection techniques from a wider range of possible DL architectures.
- This paper presents detectors based on multi-layer neural networks. It also demonstrates detectors based on approximate message passing. The role of weights and scaling parameters of neural networks is comprehensively illustrated. In addition, sparse and convolutional neural networks are presented. This paper reviews the alternating direction method of multipliers (ADMM) deep networks for massive MIMO detectors. Detectors based on the trainable projected gradient are also discussed.
- The design of an efficient detector based on SD algorithm and the selection of radius in SD algorithm based on the DL are also presented.
- This paper also presents the role of deep convolutioanl neural networks based ML detection (DCNN-MLD) and likelihood ascent search (CNNLAS) in data detection. Initial estimation of the DCNN architecture based on ZF, ZF-SIC, MMSE, and MMSE-SIC is also illustrated.
- Cell-free massive MIMO (or distributed massive MIMO) is considered as a promising technology in 5G and B5G. Thus, this paper surveys the detection techniques for cell-free massive MIMO in respect to the DL architectures.

For smooth readability, the outline of the article is depicted in Fig. 3. The most used acronyms are presented in full form in Table II. Furthermore, Table IV presents the chronology of DNN for MIMO detectors. In addition, to provide readers with a big picture, significance, limitations, and the computational complexity formulas of each architecture are summarized in Table V and Table VI, respectively.

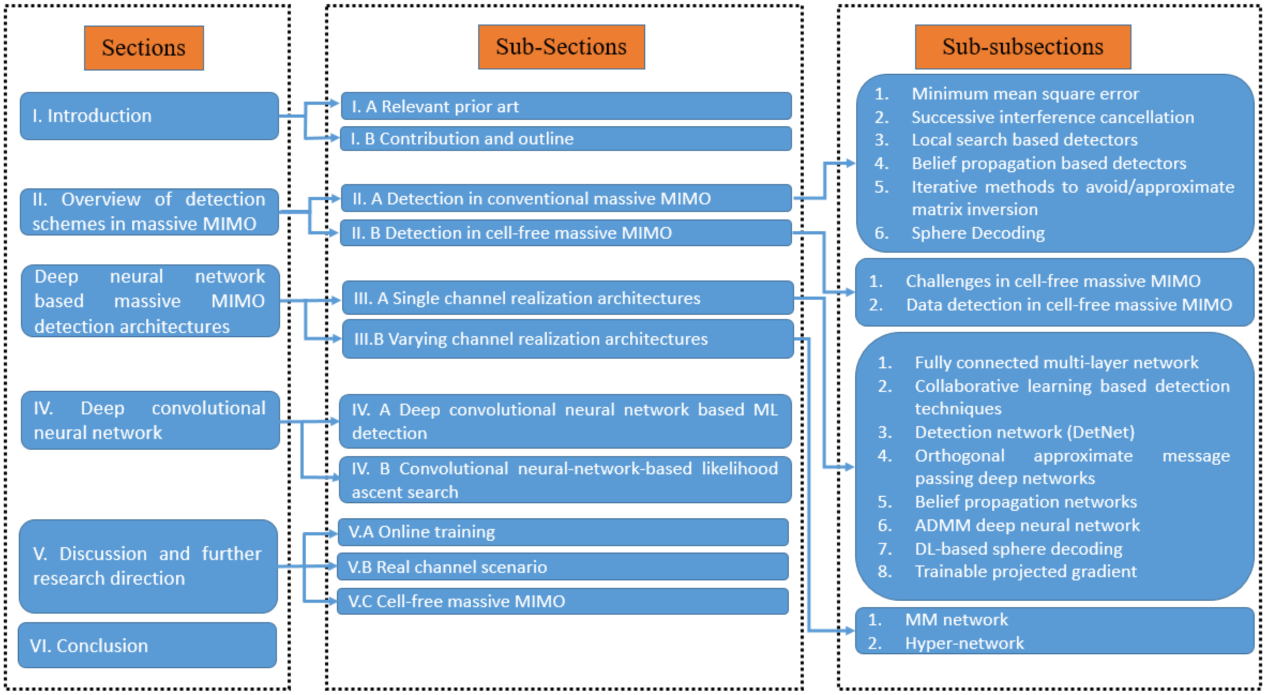
Sections
Sub-Sections
Sub-subsections
I. Introduction
II. Overview of detection schemes in massive MIMO
Deep neural network based massive MIMO detection architectures
IV. Deep convolutional neural network
V. Discussion and further research direction
VI. Conclusion
I. A Relevant prior art
I. B Contribution and outline
II. A Detection in conventional massive MIMO
II. B Detection in cell-free massive MIMO
III. A Single channel realization architectures
III.B Varying channel realization architectures
IV. A Deep convolutional neural network based ML detection
IV. B Convolutional neural-network-based likelihood ascent search
V.A Online training
V.B Real channel scenario
V.C Cell-free massive MIMO
1. Minimum mean square error
2. Successive interference cancellation
3. Local search based detectors
4. Belief propagation based detectors
5. Iterative methods to avoid/approximate matrix inversion
6. Sphere Decoding
1. Challenges in cell-free massive MIMO
2. Data detection in cell-free massive MIMO
1. Fully connected multi-layer network
2. Collaborative learning based detection techniques
3. Detection network (DetNet)
4. Orthogonal approximate message passing deep networks
5. Belief propagation networks
6. ADMM deep neural network
7. DL-based sphere decoding
8. Trainable projected gradient
1. MM network
2. Hyper-network

Figure 3. Outline of the article

Section II presents the overview of conventional detection techniques for massive MIMO receivers. It also illustrates the detection techniques for cell-free massive MIMO networks. Section III describes comprehensively massive MIMO detectors based on the DNN approach. Section IV describes comprehensively massive MIMO detectors based on the DCNN approach. Section V illustrates challenges and open research directions in DL based detectors. Section VI concludes the paper.

II. Overview of Detection Schemes in Massive MIMO

This section aims to provide readers with a brief of data detection concepts, overview of conventional algorithms, and challenges. In addition, these algorithms are exploited later with many DL architectures to achieve the target performance and computational complexity.

Detection techniques for MIMO have gained a lot of attention over the past decades [30]. The first massive MIMO detector date back to 2008 where it has been realized that the full potential of a small-scale MIMO (in terms of achieving high capacity using large number of antennas) is not achieved [84]. The main issue with utilizing a large number of antennas is the high detection complexity involved. Researchers put tremendous efforts in massive MIMO receivers' design, thus, several algorithms have been developed and adopted for massive MIMO detectors and a comprehensive survey is presented in [31]. The purpose of signal detection methods is to estimate the transmitted signal $\mathbf{x}$ from the received vector $\mathbf{y}$ at the BS antennas and their relationship is described as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{H}$ and $\mathbf{n}$ are the channel matrix and the Gaussian noise, respectively. Although the concepts of detection algorithms are comprehensively explained in [31], we briefly present them here to provide readers with a big picture view before we move to DL-based massive MIMO detectors.

A. Detection in conventional massive MIMO

1) Minimum mean square error: The mean-square error (MSE) between the transmitted $\mathbf{x}$ and the estimated signal $\mathbf{H}^H\mathbf{y}$ is minimized by the minimum mean square error (MMSE) detector as

$$\mathbf{A}^H_{MMSE} = \arg \min_{\mathbf{H}\in\mathbb{C}^{N_r\times N_t}} \mathbb{E}\|\mathbf{x}-\mathbf{H}^H\mathbf{y}\|^2, \tag{2}$$

where $\mathbb{E}$ is the expectation operator. The MMSE detector takes the noise effect into consideration as $\mathbf{A}^H_{MMSE} = \left(\mathbf{H}^H\mathbf{H}+\frac{N_t}{SNR}\mathbf{I}\right)^{-1}\mathbf{H}^H$, where $\mathbf{I}$ is the identity matrix. The output of the MMSE detector is obtained by $\hat{\mathbf{x}}_{MMSE} = \mathcal{S}(\mathbf{A}^H_{MMSE}\mathbf{y})$. However, the MMSE based detector requires a high-order matrix inversion which increases the computational complexity. In addition, the inversion is computational unstable for high-order matrix [85].

2) Successive interference cancellation: The successive interference cancellation (SIC) is a nonlinear detector where a signal is selected and detected using a linear detector (i.e., zero-forcing (ZF) or MMSE) [86]. The signal is detected and canceled from the remaining signals set, and so on. The detection and cancellation process will be repeated until all signals are detected [87]. Performance of the SIC based detectors is highly influenced by the first detected signal. Therefore, the signal with highest signal-to-noise-plus-interference (SINR) has to be detected first to obtain the best possible error rate performance [88]. After that, the second strongest signal will be detected and canceled from the remaining signals set. The process is repeated until all signals are estimated.

3) Local search based detectors: The first massive MIMO detector used likelihood ascent search (LAS) because of its linear average per-bit complexity in number of users and its ability to achieve near-maximum likelihood (ML) performance [84]. A detector based on LAS starts with an initial solution and iteratively searches the neighborhood for a better estimation. Unfortunately, the bit-error-rate (BER) is significantly deteriorated in the scenario of high modulation order. In addition, computation of the initial solution includes a matrix inversion which increases the computational complexity. However, a concatenated LAS detector with turbo codes is implemented in [89] where a satisfactory balance between the performance and the complexity is obtained. Reactive tabu search (RTS) is another local search method where more restrictions are introduced to avoid an early termination. It includes stopping criteria parameters, initial tabu period, and maximum number of iterations. Therefore, it usually outperforms the LAS detector. Unfortunately, computational complexity of the RTS is high and it also suffers from a high performance loss when a high modulation order is used.

4) Belief propagation based detectors: In order to reduce the complexity, most of proposed detectors during 2008–2013 had used local search algorithms and BP algorithms. The BP algorithms, such as the message passing and the Bayesian belief networks, iteratively search the optimal solution in a space where the damping factor (DF) has to be carefully optimized. In other words, the performance is remarkably decreased when the DF is not appropriately selected. BP algorithms are very sensitive to both the message update rules and prior information. However, the BP based detector achieves a high performance when the correlation between channel elements is relatively small.

BP is an iterative search based algorithm where the optimal solution is obtained in a reduced search space. It works by passing the message in a graphical model. Popular examples of exploiting the BP in MIMO detectors are Baysian belief networks and Markov random fields, the turbo codes, the low density parity check (LDPC) codes [90], [91] and the message passing [92]–[94]. In last years, the BP has gained a lot of attention in the research community to detect the transmitted signal as shown in [93], [95], [96]. In [97], it is exploited to recover the transmitted signal where MIMO channel is presented as a factor graph. The factor graph includes observation nodes correspond to $\mathbf{y}$ and variable nodes correspond to $\mathbf{x}$. In order to recover the transmitted data using the BP, messages are iteratively passed between the observation nodes and the variable nodes. Unfortunately, the BP is not optimal and may fail to converge when the graph is fully connected and contains many cycles. Detailed description of the BP in massive MIMO can be found in [93], [97], [98].

5) Iterative methods to avoid/approximate matrix inversion: In years after, due to not guaranteed convergence and implementation difficulties, a research on linear detectors based on iterative methods to avoid/approximate the matrix inversion is conducted. For instance, detectors based on the Neumann series (NS) [99], Newton iterations (NI) [100], successive over

Table II
ACRONYMS AND CORRESPONDING FULL MEANING

Acronym	Full Form
AP	Access point
ADMM	Alternating direction method of multipliers
AMP	Approximate message passing
BS	Base station
B5G	Beyond fifth generation
BER	Bit-error-rate
BP	Belief propagation
CSI	Channel state information
CG	Conjugate gradient
CNNLAS	Convolutional neural-network-based likelihood ascent search
CHEMP	Channel hardening-exploiting message passing
DL	Deep learning
DF	Damping factor
DetNet	Detection network
DNN	Deep neural network
DLNet	Deep learning based network
DNN-sMPD	Deep neural network-simplified message passing detector
DNN-dBP	Deep neural networks-damping belief propagation
DNN-MS	Deep neural networks-max-sum
DLBP	Deep learning based on the BP algorithm
DCNN	Deep convolutional neural network
DCNN-MLD	Deep convolutional neural network-based maximum likelihood detection
EP	Expectation propagation
FullyCon	Fully connected multi-layer network
FDL-SD	Fast deep learning aided sphere decoding
FDL-MSD	Fast deep learning aided M-best sphere decoding
FS-Net	Fast-convergence sparsely connected detection network
GS	Gauss-Seidel
GD	Gradient descent
G-DCNN	Generic deep convolutional neural network
HyperMIMO	Deep hyper-network-based uplink massive MIMO detection
IoT	Internet of things
IW-SOA	Iterative weighted sum-of-absolute value
JA	Jacobi
LAS	Likelihood ascent search
LA	Lanczos
LDPC	Low density parity check
LcgNet	Learned conjugate gradient descent network
massive MIMO	Massive multiple-input multiple-output
MC	Multi-cell
mmWave	Millimeter wave
MSE	Mean-square error
MMSE	Minimum mean square error
MPD	Message passing detector
ML	Maximum likelihood
MEPD	Modified expectation propagation-based MIMO detector
MEPNet	Modified expectation propagation network
MMNet	MM network
MLSD	Maximum likelihood sequential detector
NI	Newton iterations
OAMP-Net	Orthogonal approximate message passing networks
QLcgNet	Quantized Learned conjugate gradient descent network
QuaDRiGa	QUAsi Deterministic RadIo channel GenerAtor
RTS	Reactive tabu search
RI	Richardson
RE	Residual
ReLU	Rectified linear activation function
ResNet	Residual neural network
SD	Sphere decoding
SDR	Semi-definite relaxation
SC	Single-cell
SOR	Successive over relaxation
ScNet	Sparsely connected neural network
SIC	Successive interference cancellation
SD-DL	Sphere decoding based on deep learning
SD-IRS	Sphere decoding with increasing radius search
SNR	Signal to noise ratio
TPG	Trainable projected gradient
TISTA	Trainable iterative soft thresholding algorithm
WeSNet	Weight-scaling neural network
ZF	Zero-forcing
3G	Third generation
5G	Fifth generation
6G	Sixth generation

relaxation (SOR) [101], Gauss-Seidel (GS) [102], Jacobi (JA) [103], Richardson (RI) [104], CG [105], Lanczos (LA) [106], and residual (RE) methods are proposed in the context of massive MIMO receivers. The CG algorithm is an effective iterative method to solve the linear equations through kth iterations [105], [107]. In the CG method, the solution depends on scalar parameters. It is also refined iteratively where the search is performed in the conjugate direction with a movement towards the best solution [32], [108].In [109], a detector based on the CG algorithm is implemented in Xilinx Virtex-7 FPGA for a 128 × 8. It is also implemented using a GPU platform [110]. The CG detector outperforms the NS based detector in both the performance and complexity [107]. In [111], a revised incomplete Cholesky factorization precondition method is utilized with the CG detector to reduce the number of iterations, and hence, the computational complexity is reduced. In [112], a recursive CG detector is proposed to reduce the computational complexity and obtain a parallelism for hardware implementation. Unfortunately, these detectors suffer from a high performance loss and high computational complexity when the massive MIMO size is large or the ratio between the BS antennas and user antennas is close to 1. Other detectors require a decomposition which increases the computational complexity [113], [114]. Therefore, most of proposed detectors are not feasible in implementation due to a high computational complexity.

6) Sphere decoding: The SD algorithm searches only through the constellation points that are restricted within a sphere with a predefined radius "d" [115], [116]. The optimum performance can be obtained and the computational complexity is reduced by eliminating lattice points inside the sphere as long as d is properly selected [117]. In addition, $\mathbf{H}$ is decomposed by the QR decomposition to a unitary matrix ($\mathbf{Q}$) and an upper triangular matrix ($\mathbf{R}$). However, the radius can be selected based on the Babai estimation to guarantee the existence of at least one lattice point inside the sphere [118]. In literature, many variations of SD algorithms are proposed to reduce the computational complexity [119]–[123]. For instance, SD with increasing radius search (SD-IRS) is proposed in [124], [125].

B. Detection in cell-free massive MIMO

Most of existing signal processing algorithms are designed for centralized massive MIMO systems. However, cell-free massive MIMO network (or distributed massive MIMO) has recently gained a lot of attention due to its potential to improve the energy efficiency and spectral efficiency of wireless communications systems. Although this paper provides insights on DL based massive MIMO detectors, we present the detection schemes for cell-free massive MIMO networks. While centralized massive MIMO has already been adopted for 5G, cell-free massive MIMO is considered as a promising technology for 5G and B5G systems where a large number of individually controllable antennas distributed over a large geographical area are simultaneously serving much smaller number of user equipments. Thus, the existing DL architectures for centralized massive MIMO could be extended for cell-free massive MIMO networks.

In cell-free massive MIMO, data detection is performed locally at each AP, centrally at the central processing unit (CPU), or partially first at each AP and then at the CPU. However, it is not a trivial task to conduct the distribution of such signal processing tasks [126]. It enables coherent user-centric transmission realized by a distributed massive MIMO systems [38], [49].

1) Challenges in cell-free massive MIMO: However, the cell-free massive MIMO systems have several challenges related to the network synchronization, high data rate, detection complexity, hardware implementation, and low-latency [49]. There are many obstacles to offer an optimal channel hardening and favorable propagation conditions that owing to some spatial channel correlation between cell-free massive MIMO system's antennas [49], [54]. In [46], [127], utilization of a maximal-ratio detector is locally advocated at each AP in cell-free massive MIMO systems where effects of spatially correlated channels, imperfect CSI, and hardware impairments are considered. Unfortunately, the spectral efficiency of a maximal-ratio detector in cell-free massive MIMO is much lower than the large-scale fading decoding in centralized massive MIMO systems [128]. In [129], the user data rate is significantly improved by optimizing receiver filter coefficients at the CPU in cell-free massive MIMO systems. However, this optimization is a non-convex problem which causes a high computational complexity. In [130], a low-complexity iterative soft-input soft-output (SISO) detection algorithm in a distributed large-scale MIMO system based on an improved MMSE iterative soft decision interference cancellation is proposed and experimentally tested. The prototype system has shown that a data rate of 10 Gbps could be achieved by a 128×128 cell-free massive MIMO with 100 MHz bandwidth. In [131], a daisy chain topology exploiting the channel property of asymptotic orthogonality is utilized to obtain a low complexity distributed detection algorithm. In order to reduce the system control overhead, a distributed sparse activity detection algorithm is proposed in [132]. In [133], a joint channel estimation and data detection algorithm for cell-free massive MU MIMO is proposed to minimize the overhead of pilot-based channel estimation for cell-free systems. In [134], a MU cell free massive MIMO system with linear decoders based on approximate matrix inversion methods is proposed. In [135], polynomial expansion detectors and multistage Wiener filters are exploited to propose a low complexity detector. The only available utilization of neural network in cell-free data detection is shown in [136] where ZF and DCNN algorithms are used at the CPU.

2) Data detection in cell-free massive MIMO: Many testbeds, such as the Argos, the LuMaMi, and the BigStation are available to support the decentralized channel estimation and data detection at antenna elements [137]. Unfortunately, they rely on the maximum ratio combining (MRC) that reduces significantly the spectral efficiency, and hence, prevents the use of high-rate modulation and coding schemes. Therefore, alternative BS architectures based on a decentralized approach are proposed. A decentralized data detection method based on the CG is proposed where the BS antenna array is partitioned into clusters and each cluster is associated with independent local radio-frequency (RF) elements and computing circuitry [138]. It is demonstrated by mapping the detection method to a Xeon Phi cluster. It is shown that the design with hundreds or even thousands of BS antennas could be supported. Another decentralized data detection based on alternating direction method of multipliers (ADMM) [137], partially decentralized (PD) and fully decentralized (FD) data detectors based on the AMP are proposed [139]. The DBP is studied based on free-matrix-inversion methods in different channel conditions [140]. The FD architectures based on the coordinate descent (CD) method and FD data detector based on recursive lease square (RLS), stochastic gradient descent (SGD), and averaged stochastic gradient descent (ASGD) have also been proposed [141], [142]. The decentralized baseband processing (DBP) architecture splits the BS antennas into different individual clusters. Each cluster has an independent radio frequency (RF) chain, analog-to-digital converters (ADC), and computing hardware. The implementation was demonstrated on a GPU cluster. The results show that it has a scalability in massive MU-MIMO systems with thousands antennas. Unfortunately, the proposed DBP is not tested in different system configurations and realistic channel conditions. In [143], a VLSI architecture based on approximate message passing (AMP) of decentralized feed-forward and parallel equalization is proposed. It depends on a high-level synthesis (HLS). The results show that the proposed VLSI architecture can achieve a competitive balance between the performance and the computational complexity. In [144], a DBP based on expectation propagation algorithm (EPA) is proposed for signal detection in UL massive MIMO. The BS is partitioned into multiple independent antenna clusters, each associated with analog and digital modulation circuitry, RF chains, and computing hardware where local channel estimation and signal detection are executed.

Table III shows the pros and cons of several detection methods. All of them are not achieving a satisfactory balance between the performance and complexity when the ratio between the number of antennas at the BS and user terminals is small. They are also not achieving a good performance in ill-conditioned environment as well as realistic channel scenarios. Therefore, a significant room for fundamental research in data detection based on a DL approach is introduced to achieve a satisfactory balance between the performance and complexity in realistic channel scenarios and different MIMO configurations.

Owing to a powerful learning ability from the data and the development of advanced optimization techniques and fast-growing of computing power, there is a remarkable trend to utilize the DL approaches in massive MIMO receiver's design where the most expensive process is the ”learning” which can be completed off-line. However, it is noticed that there are many attempts to do it on-line. Therefore, the DL is incorporated in many existing algorithms by adding some adjustable and trainable parameters to improve the detector's performance and the computational complexity in real channel scenarios. For instance, it is incorporated into the BP (i.e., orthogonal approximate message passing (OAMP) and message passing detector (MPD)), the SD, the SIC, and CG techniques. In addition, DL approaches are still in its infancy for cell-free massive MIMO systems. However, proposed DL architectures for conventional massive MIMO can be considered as a base-stone for a future research in DL for cell-free massive MIMO networks.

III. Deep Neural Network Based Massive MIMO Detection Architectures

As shown in Section II, conventional detectors with their variations suffer from mediocre performance under certain circumstances. In addition, physical channels are playing a significant role in achieving a high performance and low computational complexity. Instead of the classical detection theory, the DL is exploited to achieve the best performance-complexity profile. It depends on the number of parameters, training samples, initial solutions, and the architecture. The earliest massive MIMO detector based on the DL date back to 2017 where a projected GD is utilized in a single training phase [145]. In literature, some architectures of DNN based massive MIMO detectors have an effective online training for varying channel realization and have a satisfactory performance in realistic channel scenarios [146]. While other architectures have a single training shot and have a satisfactory performance over constant and Rayleigh fading channels only [22], [145], [146], [147]. In this section, these two types are described where pros and cons of each type are presented.

A. Single Channel Realization Architectures

1) Fully Connected Multi-Layer Network:: The fully connected multi-layer network (FullyCon) architecture is one of the earliest trials to employ the DL in massive MIMO detectors [22]. It is a type of neural networks where all neurons in current layer are connected to neurons in the next layer. The FullyCon architecture consists of sequential L layers where the output of the current layer is the input of the subsequent layer and it only depends on the input $\mathbf{y}$, and does not employ the channel $\mathbf{H}$ [22]. It is presented as

$$\begin{gathered}\mathbf{q}_1 = \mathbf{y}\\ \mathbf{q}_{k+1} = \rho(\Theta_k^{(1)}\mathbf{q}_k + \theta_k^{(2)}) \quad \text{for} \quad k = 1, ..., L\\ \hat{\mathbf{x}} = \Theta_L^{(1)}\mathbf{q}_L + \theta_L^{(2)},\end{gathered} \tag{3}$$

where $\mathbf{q}_1$ is the initial input of the first layer and equals to $\mathbf{y}$, $\mathbf{q}_k$ is the iterative input of the next layers, and $\hat{\mathbf{x}}$ is an estimated vector of the unknown vector $\mathbf{x}$. ρ is an activation function and can be adopted as a rectified linear activation function (ReLU) as $\rho[x] = \max(x, 0)$ [22], where θ is an optimized parameter during the learning stage and consists of $\theta = \left\{\Theta_k^{(1)}, \theta_k^{(2)}\right\}_{k=1}^{L}$. To obtain a robust detector, a loss function l is presented as ($\mathbf{x}$; $\hat{\mathbf{x}}$ ($\mathbf{H}$, $\mathbf{y}$; θ)) and determines the distance between the real vectors $\mathbf{x}$ and estimated vector $\hat{\mathbf{x}}$. The parameter θ is optimized to minimize the loss function over the massive MIMO model distribution as [22]

$$\min_{\theta} \mathbb{E}\left\{l(\mathbf{x}; \hat{\mathbf{x}}(\mathbf{H}, \mathbf{y}; \theta))\right\} = \min_{\hat{\mathbf{x}}} \mathbb{E}\|\mathbf{x} - \hat{\mathbf{x}}\|^2. \tag{4}$$

In general, the FullyCon architecture is simple and has a small number of optimized parameters. It is perfect for detection

Table III
CONVENTIONAL DETECTION METHODS FOR MASSIVE MIMO SYSTEMS

Method	Pros	Cons
Linear detectors (i.e., ZF and MMSE)	• Work properly if columns of the propagation matrix are nearly orthogonal. • Relatively simple to implement.	• Suffer from a sever performance degradation and high computational complexity in ill-conditioned environment. • MMSE obtains a significant performance loss in highly loaded systems.
Approximate matrix inversions (i.e., NS, NI, SOR, GS, JA, RI, CG, LA, and RE)	• The optimal performance can be achieved. • If the initial solution is properly selected, the optimal performance can be achieved within few iterations.	• When the ratio between the BS antenna and the user antennas is close to 1, approximate matrix inversion methods suffer from a performance deterioration. • The convergence rate is highly affected by the initial solution which could lead to a wrong estimation.
Sphere decoder (SD)	• A satisfactory balance between the computational complexity and the performance can be achieved	• High computational complexity if the sphere radius is not properly selected.
Successive interference cancellation (SIC)	• A good performance can be achieved when the number of BS antennas is greater than the number of user terminals.	• The performance is highly affected by the initial solution. • High computational complexity.
Belief propagation (BP)	• When the channel correlation is low, the maximum-likelihood performance is obtained.	• Optimal damping factor is not easy to obtain. • Performance is significantly degraded if a bad conditioned factor graph is utilized. • Not easy to guarantee the convergence.

over fixed channel where $\mathbf{H}$ is deterministic and fixed and the channel is known within the training phase [22]. Numerical results in [22] show that the FullyCon architecture over fixed channel $\mathbf{H}$ can achieve a near optimal accuracy with a low complexity. Unfortunately, the performance of FullyCon is significantly deteriorated over real channels. The total error of a DL based detection consists of variance, bias, and irreducible error [148], [149]. The irreducible error is intractable to remove, but the bias and the variance have an optimized trade-off between them.

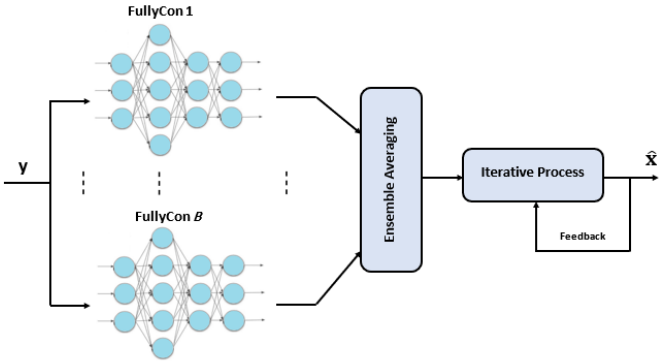

Figure 4. Model of the ELRID architecture

2) Collaborative learning based detection techniques:: In [150], [151], a machine learning method is proposed which is known as a collaborative learning. It combines, in a strategic manner, multiple models with their predictions and leanings, to produce more optimal results. In addition, the collaborative learning has an ensembling operation which leads to a near-optimal performance by smoothing the variance from multiple models with roughly fixed bias. Inspired by the collaborative learning, authors in [152] proposed a collaborative learning based detection technique with iterative process (ELRID). It employs a collaborative learning with iterative prediction and offers a robust and low complexity detector. As shown in Fig. 4, the ELRID architecture ensembles B fully connected architectures and employs iterative meta-predictor to reduce the final estimation. It has significantly improved the performance with a low computational complexity. In [153], a fully-connected multi-layer DNNs for joint channel estimation and data detection of spatial modulation (SM) MIMO system, called Deep spatial modulation (DeepSM), is proposed. The DeepSM architecture operates in a data-driven approach and obtains a BER performance close to that of the conventional SM MIMO system over time-invariant channels with low detection complexity.

3) Detection Network:: The detection network (DetNet) for massive MIMO systems is proposed by Samuel *et al.* [22], [145]. It has more expressive architecture and designed specifically to address the challenges of varying channels in FullyCon based detectors [22]. The DetNet architecture performs well over both constant and Rayleigh fading channels with a single training shot [22], [154]. Furthermore, It offers a near-optimal performance with low modulation orders (i.e., BPSK and QPSK) [146]. The DetNet architecture is a data-driven method that overcomes the performance of the approximate message passing (AMP) and SDR algorithms [155], [156]. It also offers the performance of SDR detection algorithm over independently and identically distributed (i.i.d.) Gaussian channels while working 30× faster [146]. Furthermore, the DetNet architecture addresses challenges of the vanishing gradients, initialization sensitivity, and saturation of the activation functions [22] [157]. Unlike the FullyCon architecture, the DetNet does not work directly with $\mathbf{y}$. It uses the compressed sufficient statistic as [22]

$$\mathbf{H}^T\mathbf{y} = \mathbf{H}^T\mathbf{H}\mathbf{x} + \mathbf{H}^T\mathbf{n}. \tag{5}$$

The DetNet architecture unfolds a projected GD algorithm for the ML optimization that leads to an iterative form as

$$\hat{\mathbf{x}}_{k+1} = \prod[\hat{\mathbf{x}}_k - \theta_k \mathbf{H}^T\mathbf{y} + \theta_k \mathbf{H}^T\mathbf{H}\hat{\mathbf{x}}_k], \tag{6}$$

where $\hat{\mathbf{x}}_k$ is the estimated vector at the k^{th} iteration, θ_k is a gradient step size, and $\prod[.]$ is a nonlinear projection function. Each iteration in the DetNet architecture is carried out by a single layer which composed of a linear combination of $\hat{\mathbf{x}}_k$, $\mathbf{H}^T\mathbf{y}$, and $\mathbf{H}^T\hat{\mathbf{x}}_k$ and a non-linear projection [22], [158]. The performance can be enhanced when the step size θ_k at each step is considered as a learned parameter within the training stage. Thus, the DetNet architecture is described as [22]

$$\begin{gathered}
\mathbf{q}_k = \hat{\mathbf{x}}_{k-1} - \theta_k^{(1)}\mathbf{H}^T\mathbf{y} + \theta_k^{(2)}\mathbf{H}^T\mathbf{H}\hat{\mathbf{x}}_{k-1} \\
\mathbf{z}_k = \rho(\Theta_k^{(3)}\mathbf{q}_k + \Theta_k^{(4)}\mathbf{v}_{k-1} + \theta_k^{(5)}) \\
\hat{\mathbf{x}}_k = \psi_{t\,k}(\Theta_k^{(6)}\mathbf{z}_k + \theta_k^{(7)}) \\
\hat{\mathbf{v}}_k = \Theta_k^{(8)}\mathbf{z}_k + \theta_k^{(9)} \\
\hat{\mathbf{x}}_0 = \mathbf{0} \\
\hat{\mathbf{v}}_0 = \mathbf{0},
\end{gathered} \tag{7}$$

where θ presents the optimized parameters during the learning stage and ψ_t is a piece-wise linear soft sign operator [145], [159]. Where ψ_t is used to approximate $\prod[.]$ of the projected GD algorithm [145], [159]. Parameter matrices $\Theta_k^{(3)}$ and $\Theta_k^{(4)}$ in (7) are $m \times n$ matrices where $m > n$, and employed in the DetNet architecture to map the projection input to an even larger dimension vector before mapping it again to a vector with N_t dimension [22] [146]. For example, the DetNet architecture connection with 2×2 MIMO structure is depicted in Fig. 5. Inspired by the aided classifiers in GoogLeNet architecture [160], the DetNet architecture espouses a loss function that considers the outputs of all of layers [22]. In order to decrease the loss function values, it constrains the layer output $\hat{\mathbf{x}}_k$ to be near to $\mathbf{x}$. These constraints reduce the ability of the deep network to compute sophisticated features where the only information passed between layers is estimated as $\mathbf{x}_k$. Thus, $\hat{\mathbf{v}}_k$ vector in (7) is added to the network to allow passing unconstrained information between layers [22]. Unfortunately, the DetNet based detector depends heavily on a large amount of parameters, approximately 1-10 million parameters. Therefore, offline training is conducted [146], [161]. It is also a data-driven method and requires a large-training data and training procedure for three days on a standard Intel i7-6700 processor [161] [162]. In addition, it experiences a performance loss in a large-scale MIMO systems with $N_t \approx N_r$. The heuristic behavior of the DetNet makes it difficult to employ over realistic channel like QUAsi Deterministic RadIo channel GenerAtor (QuaDRiGa) [146]. It also does not employ known features of iterative algorithms, and hence, incurs unnecessarily complexity. Accordingly, the DetNet architecture, in contrast to DL based detectors that utilize the straightforward non-linear denoiser, employs a non-linear projection which is a fully-connected 2-layer neural network [146].

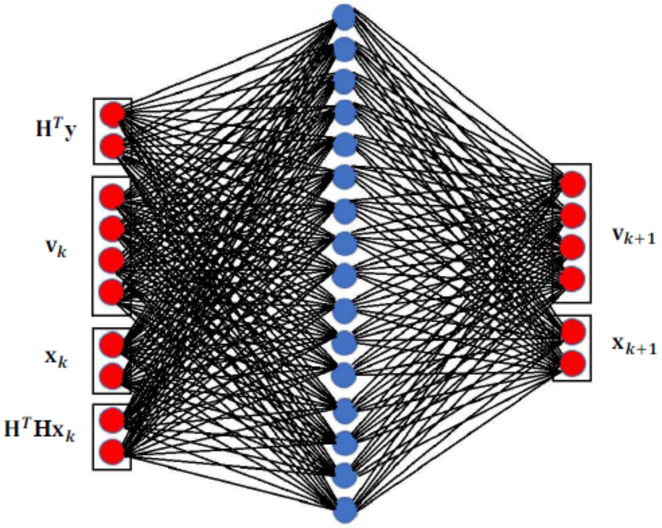

Figure 5. The DetNet architecture Connection

Many research efforts are conducted to enhance the DetNet architecture. In [163], a new architecture is proposed to improve the activation function in the DetNet by using a multi-level-plateau sigmoid activation function. The modified DetNet architecture employs twin DL networks with various initial values to simultaneously detect the transmitted signals. The modified DetNet architecture achieves a near-optimal performance with a reasonable number of parameters. In [158], authors proposed a simplified version of the DetNet architecture called as sparsely connected neural network (ScNet) where the number of inputs and network layers are significantly reduced by converting the DetNet architecture to a sparse connectivity instead of a full connectivity. It is well known that the square matrix is not reversible and matrix inversion is a very complicated operation. Therefore, the loss function of the network is optimized to avoid irreversible problems with the matrix [158]. With these simplifications, complexity of the DetNet architecture is reduced from $O(64N_t^2)$ to $O(3N_t)$. The form of the ScNet architecture connection with 2×2 MIMO structure are illustrated in Fig. 6. ScNet architecture is not only reducing the computational complexity of the DetNet but also has obtained a performance gain, especially with a large scale antenna.

Another architecture known as a fast-convergence sparsely connected detection network (FS-Net) is proposed in [164]. It is acquired by optimizing the DetNet and ScNet architectures. The FS-Net simplifies the network connections to reduce the complexity. It improves the loss function by considering the correlation between the output of each layer and the desired solution to achieve faster convergence rate. The FS-Net architecture has a significant performance improvement with lower complexity in contrast to the DetNet and ScNet architectures.

In [165], a DL based network (DLNet) for signal decoding in massive MIMO systems is proposed. It considers time-varying Gaussian random channel with a perfect CSI at the receiver. The DLNet architecture has 50-layers DNN and is based on the projected GD algorithm. The DLNet architecture

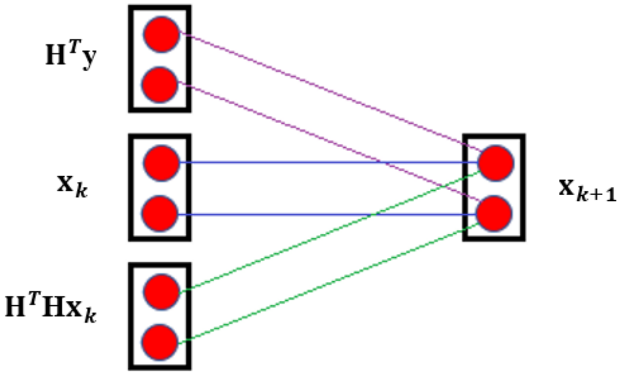

Figure 6. The ScNet architecture Connection

is implemented in the TensorFlow [166] with the Adam optimizer [167]. It is trained with 50000 iterations on a standard intel Xeon 4114 10C/20T 2.2G 13.75M 9.6GT UPI processors that took about 14 hours with 30×60 massive MIMO system and about 8 hours with 20×30 massive MIMO system. However, the DLNet offers a better BER performance with 164× faster in running speed and 9× less computational complexity than the DetNet. Furthermore, the DLNet achieves a comparable BER performance with the SDR algorithm and 28200× faster.

Inspired by the DetNet architecture, authors in [162] proposed a novel learned conjugate GD network (LcgNet) and quantized LcgNetV (QLcgNetV) network. The LcgNet architecture is implemented by unfolding iterative conjugate GD detector. The distinctions lie in the step-sizes which are found to be universal, instead of calculating the accurate values of the scalar step-sizes, and can be learned within offline training. Furthermore, the LcgNet can be enhanced by augmenting the dimensions of these step-sizes. In addition, the matrix-vector multiplication and division operations are replaced by some prestored parameters which are fixed within online detections. The QLcgNetV architecture is introduced in order to minimize the memory costs. It quantizes the learned parameters by using a low-resolution nonuniform quantizer (i.e., 3-bits or 4-bits). The quantizer in the QLcgNetV architecture is based on an adaptive designed soft staircase function which is structured from a series of tanh(·) functions with adjustable parameters to minimize the MSE loss function. The LcgNet and the QLcgNetV architectures are implemented in Python using the TensorFlow library with the Adam optimizer. The number of learnable parameters in the LcgNet and the QLcgNetV architectures is very limited compared with the FullyCon and DetNet architectures. The LcgNet and the QLcgNetV architectures are trained offline within 2 hours with Intel (i3-6100) CPU running at 3.7GHz and 8GB RAM. Furthermore, the LcgNet and the QLcgNetV architectures obtain a good performance in realistic channel models (i.e., a spatial correlated channel model) and low-order modulation scheme e.g., binary phase-shift keying (BPSK), quadrature PSK (QPSK) or 16-level quadrature-amplitude modulation (QAM).

A modified version of the DetNet architecture termed as weight-scaling neural-network based MIMO detector (WeSNet) is proposed to reduce the computational complexity [168]. The WeSNet architecture is realized by adjusting layer weights within monotonic profile functions. It is mainly based on the DetNet architecture by unfolding a projected GD algorithm. The WeSNet architecture imposes constraints on the layer weights in order to permit for entire layers to be repealed in a controllable method within inference that leads to a promising reduction in the model size and the computational complexity with reasonable degradation in the accuracy. Performance of the WeSNet architecture is improved by dealing with weight profile functions themselves as trainable parameters in order to prohibit vanishing gradients. However, this improvement leads to a satisfactory performance at the cost of increased memory due to an increment in the number of parameters. It is implemented in TensorFlow 1.12.0 and evaluated under both 30×60 and 16×16 massive MIMO systems with low-order modulation schemes (i.e., BPSK and QPSK) under Rayleigh fading channel. It outperforms the DetNet architecture and offers 51.43% reduction in complexity and about 50% reduction in model size.

4) Orthogonal Approximate Massage Passing Deep Network:: The OAMP-Net architecture is inspired by the OAMP algorithm which is an extension of the AMP iterative algorithm to be able to serve in correlated channels [169] [147]. The OAMP-Net architecture aims to address a large amount of trained parameters in the DetNet architecture [147]. It also aims to achieve a satisfactory performance of the massive MIMO detection over varying channels. Like the DetNet based detectors, the OAMP-Net architecture is trained offline and learns a single detector within a training stage [146] [147]. It is a model-driven method and adds just two trainable parameters ($\theta_k^{(1)}$,$\theta_k^{(2)}$) to each iteration in the OAMP algorithm as [146], [147]

$$\begin{aligned} \mathbf{z}_k &= \hat{\mathbf{x}}_k - \theta_k^{(1)}\mathbf{H}^H(v_k^2\mathbf{H}\mathbf{H}^H + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}_k) \\ \hat{\mathbf{x}}_{k+1} &= \eta_k(\mathbf{z}_k;\sigma_k^2), \end{aligned} \tag{8}$$

where v_k^2 is proportional to the average noise power of the denoiser output at k^{th} iteration and can work out based on the given signal-to-noise ratio (SNR) and system sizes. $\theta_k^{(1)}$ is a first learning parameter used instead of a normalizing factor γ_k in the OAMP algorithm. $\theta_k^{(2)}$ is a second learning parameter used to balance the estimate of denoisers input noise variance σ_k^2. $\eta_k(.)$ is the straightforward denoiser function and can be computed based on the prior distribution of the original signal $\mathbf{x}$ and is chosen to minimize the MSE $\mathbb{E}_{\mathbf{x}}\left[\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \mid \mathbf{z}\right]$ by assuming the noise is independent and Gaussian distributed. The number of learning variables is independent of the system size. The OAMP-Net depends on the number of layers. However, with only two learning parameters, the training process in the OAMP-Net architecture is significantly improved in the stability and convergence rate. It also can be trained during a short time where few trainable parameters are needed to be optimized [146]. In addition, it can also offer soft decision, which is more appropriate in modern wireless communications systems [146], [147]. The

OAMP-Net based detector outperforms the OAMP algorithm and is considered as one of the best learning algorithms [146]. Although the OAMP-Net architecture offers a good performance under i.i.d. Gaussian channels, it suffers from a significant performance loss under a realistic channel model such as the QuaDRiGa [146], [147]. In addition, a matrix inversion has to be computed, as in (8), which increases the computational complexity.

Inspired by the trainable iterative soft thresholding algorithm (TISTA) for sparse signal recovery [170], a modified version of the OAMP-Net architecture, known as OAMP-Net2, is proposed in [171]. Most of other existing DL-based massive MIMO detectors [74], [145], [146], [162], [172], [173], [174] just deal with an accurate CSI where the channel estimation error is ignored. The OAMP-Net2 architecture is mainly proposed to enhance the detection performance of the OAMP-Net. It considers the characteristics of channel estimation error and channel statistics. It uses the estimated payload data to elucidate the channel estimation. The OAMP-Net2 architecture unfolds the existing iterative OAMP algorithm with additional learnable parameters. The OAMP-Net2 architecture needs only four trainable parameters in each layer to be learned which significantly reduces the training time and the computing resources. The OAMP-Net2 obtains a considerable performance gain.

5) Belief Propagation Networks:: Similarities between the BP factor graph and DNN have encouraged interested researchers to conduct an extensive research on the DL based on BP detection. In [74], [175], a DNN is employed to enhance the BP detection algorithms for i.i.d. Rayleigh and correlated fading massive MIMO channels with different antenna configurations. The correction factors in the MPD detectors, including damped BP (dBP) [95], [95], [176], max-sum BF (MS) [177], and channel hardening-exploiting message passing (CHEMP) detectors are optimized within DL architectures [178] where three DNNs are proposed to deal with a loopy factor graph and a high complexity of BP algorithms. These DNN detectors are deep neural network-simplified message passing detector (DNN-sMPD), deep neural networks-damping belief propagation (DNN-dBP), and deep neural networks-max-sum (DNN-MS) [74]. They optimize the damping factors of the BP algorithms to enhance the convergence rate, and optimize the normalized and offset factors of the BP algorithms to further minimize the computational complexity. The training is implemented in the TensorFlow library with the Adam optimizer. They use one offline training for each antenna configuration and can be used for multiple online detections [74]. Compared to other message passing detectors and linear detectors, DNN-sMPD, DNN-dBP, DNN-MS detectors can offer lower BER with improved robustness in different antenna configurations and channel conditions with the same computational complexity. However, DNN detectors did not consider high-order modulation schemes and realistic channel scenarios. In addition, the DNN is trained for a small and moderate-size MIMO system (i.e., 16×16 and 8×32). However, the offline training needs a huge amount of data and requires powerful computational and storage devices to store the trained networks for multiple online uses. In addition, DNN detectors depend on the range of the training data. Detection performance of the DNN-dBP network deteriorates due to differences in the channel correlation between training and test where the optimal damping factors are changed with the channel correlation [98], [179]. In [98], [179], a node selection method is proposed to tackle the detection performance deterioration in the DNN-dBP network.

In [180], a DL MIMO detector based on the BP algorithm (DLBP) is proposed to address the convergence challenge of the BP algorithm due to a fully connected factor graph. The DLBP architecture unfolds the BP algorithm by employing four-layers neural network. The DLBP achieved a low complexity and a good BER performance in contrast to the BP detector and the dBP detector. The training is implemented in the TensorFlow library with the Adam optimizer. However, the DLBP architecture is tested with an equal number of transmitter and receiver antennas (i.e., 8×8 and 16×16), low-order modulation scheme (i.e., BPSK), and through the Rayleigh channel.

In [181], a modified expectation propagation (EP)-based MIMO detector (MEPD) is proposed to tackle challenges of the conventional EP detector which are incurred due to the empirical parameter selection such as damping factors and initial variance [182]. Furthermore, a modified EP network (MEPNet) is proposed to offer the optimal damping factors and initial variance by adopting a DL scheme and unfolding the proposed iterative MEPD detector. The MEPNet architecture is a model-based data-driven neural network. It is implemented on the Google Tensorflow platform and is trained offline with 5 layers using Adam optimizer for various antenna scales and modulation orders. The MEPNet architecture with two learnable variables outperforms the MMSE-SIC, OAMPNet, OAMPNet2, and conventional EP detector algorithms under i.i.d. Rayleigh MIMO channels. In addition, the MEPNet architecture is more robust than the conventional EP detector under correlated channels.

6) ADMM Deep Neural Network:: Inspired by the ADMM algorithm [183]–[185] and by the DetNet architecture, ADMM-Net based detector is proposed in [159] to tackle the massive MIMO detection problem. In the ADMM-Net architecture, the first layer is excluded, and hence, the arithmetic process in each layer requires matrix-vector multiplications and/or simple element-wise multiplications. The number of required learnable parameters in the ADMM-Net architecture is smaller than that of the DetNet architecture [159]. It unfolds the conventional iterative ADMM algorithm by untying its parameters and then maps them into a DNN architecture. Accordingly, the ADMM-Net architecture replaces the non-negative penalty parameter λ in the ADMM algorithm by $\lambda = \nu \circ \omega$, where ν is channel power vector, $\nu = \left[\|\mathbf{h}_1\|_2^2, \ldots, \|\mathbf{h}_{N_t}\|_2^2 \right]$, and ω is a learnable parameter [159]. In addition, the ADMM-Net architecture unties the projection function $\prod_{\{\pm 1\}^{N_r}}$ in the ADMM algorithm by replacing it with

$$\Upsilon_{\beta,t}(x) = \Pi_{[-\beta,\beta]}(x/t) = -\beta + \frac{\rho(x+\beta \circ t)}{t} - \frac{\rho(x-\beta \circ t)}{t}, \quad (9)$$

where $\Upsilon_{\beta,t}(x)$ is the projection operator, β is a learnable parameter which is used to a piece-wise linear soft sign operator $\psi_t(.)$ and to learn the clipping level of the projection function. t is also a learnable parameter of the linear soft sign function [159]. The ADMM-Net architecture is presented as [159]

$$\begin{gathered}
\mathbf{z}_k = \Upsilon_{\beta_k,t_k}(\hat{\mathbf{x}}_{k-1} - \mathbf{u}_{k-1}) \\
\mathbf{u}_k = \mathbf{z}_k - \hat{\mathbf{x}}_{k-1} + \mathbf{u}_{k-1} \\
\hat{\mathbf{x}}_k = \mathbf{D}(\mathbf{H}^T\mathbf{y} + (\omega_k \circ \nu) \circ (\mathbf{z}_k + \mathbf{u}_k)) \\
\hat{\mathbf{x}}(\theta) = \Upsilon_{1,t_L}(\hat{\mathbf{x}}_L) \\
\text{subject to} \quad \hat{\mathbf{x}}_0 = \mathbf{0}, \mathbf{u}_0 = \mathbf{0}, \nu = \operatorname{diag}(\mathbf{H}^T\mathbf{H}), \\
\mathbf{D} = (\operatorname{diag}(\omega_0 \circ \nu) + \mathbf{H}^T\mathbf{H})^{-1},
\end{gathered} \tag{10}$$

where $\hat{\mathbf{x}}(\theta)$ is the output of network and based on θ which includes the learnable parameters, $\theta = \left\{\{\omega_l\}_{l=0}^{L}, \{\beta_k\}_{k=1}^{L}\right\}$. The ADMM-Net architecture outperforms the ZF, SDR, and DetNet based detectors with a small number of layers. It also obtains a good performance in low-order modulation schemes (i.e., BPSK and QPSK) over i.i.d. Gaussian channels. Particularly, the ADMM-Net architecture for a 160×160 MIMO system and with 40 layers can achieve a quasi optimal performance [159]. However, the ADMM-Net architecture does not consider high-order modulation schemes and realistic channel scenarios. In addition, it has an average run time higher than that in the SD algorithm for a small-size massive MIMO system.

7) DL-Based Sphere Decoding:: A SD algorithm based on the DL (SD-DL) is proposed in [174], where the radius of a sphere is learned through a DNN. It depends on both the structure of $\mathbf{H}$ and the noise statistics. The DNN used in the SD-DL architecture is a fully connected feedforward neural network. Learnable radiuses lead to a remarkable reduction of lattice points inside the sphere and hence, the computational complexity is significantly reduced. Furthermore, the probability of failing to find a solution in the SD-DL architecture is close to zero. The main idea of the SD-DL architecture is to implement the SD-IRS algorithm with a small number of learnable radiuses. The SD-DL architecture obtains $q \times 1$ radius vector $\mathbf{r}$, i.e., the q closest lattice points to vector $\mathbf{y}$, through the trained DNN as

$$\hat{\mathbf{r}} = \Phi(\mathbf{z};\Theta) = \left[\hat{r}_{i_1}, \hat{r}_{i_2}, ..., \hat{r}_{i_q}\right]^T, \tag{11}$$

with

$$\mathbf{z} \triangleq \left[\Re\{\mathbf{y}\}, \Im\{\mathbf{y}\}, \Re\{h_{11}\}, \Im\{h_{11}\}, ..., \Re\{h_{N_rN_t}\}, \Im\{h_{N_rN_t}\}\right], \tag{12}$$

and

$$\Theta \triangleq [\theta_1, \theta_1, ..., \theta_K], \tag{13}$$

where $\mathbf{z}$ is the input vector of the DNN and Θ is the vector of all parameters of the SD-DL architecture. The SD-DL architecture consists of two phases. The first phase is an offline training phase where the DNN is trained independently for each SNR value. The DNN which has three layers with one hidden layer is considered for a 10×10 MIMO system with 16-QAM and 64-QAM. Clipped ReLu [174] is used as the activation function in these hidden layers. In the second phase, the estimated transmitted vector $\hat{\mathbf{x}}$ is obtained through the DNN [174]. The SD-DL architecture is implemented by a DL Toolbox of MATLAB 2019a with the Adam optimizer. The SD-DL architecture offers a significant performance enhancement in i.i.d. Gaussian channels and a high-order modulation scheme. Furthermore, it has a significant performance gain in contrast to the MMSE algorithm and has a comparable performance with the SD-IRS algorithm with a remarkable complexity reduction.

In [186], a fast DL aided SD (FDL-SD) and a fast DL-aided M-best SD (FDL-MSD) architectures are proposed to accelerate the searching process in the SD and M-best SD algorithms, respectively. Furthermore, the FDL-SD and FDL-MSD architectures have more beneficial in both offline training and online decoding phases in contrast to the SD-DL architecture. The main idea of the FDL-SD and FDL-MSD architectures is to leverage the FS-Net architecture to produce a highly dependable initial solution with a low computational complexity. The initial solution generated by the FS-Net is employed in layer ordering within the FDL-SD architecture and in layer ordering with early rejection within the FDL-MSD architecture. FDL-SD and FDL-MSD architectures are not utilizing the conventional SD in the training phase. Consequently, the FDL-SD and FDL-MSD architectures can be trained with remarkably lower time and computational resources in contrast to the SD-DL architecture. The FDL-SD architecture, for a 24 × 24 MIMO system with QPSK, can offer lower computational complexity by about 90% without any performance loss in contrast to conventional SD schemes. In addition, the FDL-MSD architecture achieves the same performance of the conventional M-best SD algorithm, with M= 256 survival paths, with just M=32 survival paths in 32×32 MIMO system with QPSK modulation scheme [186].

8) Trainable Projected Gradient:: A DL-aided iterative decoder known as a trainable projected gradient (TPG) detector is proposed in [173], [187] for overloaded massive MIMO systems. The TPG architecture is based on the data-driven concept and on a projected GD for a total of L iterations or layers $\{k = 0, ..., L-1\}$. It includes the GD and the soft projection step. The TPG architecture has three learnable parameters in each layer as

$$\begin{gathered}
\mathbf{z}_k = \hat{\mathbf{x}}_k + \gamma_k \mathbf{W}\left(\mathbf{y} - \mathbf{H}\hat{\mathbf{x}}_k\right) \\
\hat{\mathbf{x}}_{k+1} = \tanh\left(\frac{\mathbf{z}_k}{|\theta_k|}\right) \\
\text{where} \quad \mathbf{W} = \mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T + \alpha\mathbf{I}\right)^{-1} \\
\hat{\mathbf{x}}_0 = \mathbf{0}.
\end{gathered} \tag{14}$$

The first parameter γ_k controls the size of the GD step, the second parameter θ_k controls the softness of the soft projection, and the third parameter α in the linear MMSE matrix $\mathbf{W}$ is optimized within the training phase. The vanishing gradient problems are avoided in the TPG architecture by using an incremental training. It is more scalable for massive MIMO systems in terms of the computational complexity than the FullyCon, DetNet, and OAMP-Net architectures. The TPG architecture has a smaller number of trainable parameters compared to FullyCon and DetNet architectures. In addition, these parameters are not the MIMO size dependent which makes the TPG architecture promoting fast and has a stable training. In addition, the TPG architecture does not need to initialize the MMSE matrix in each layer as that in the OAMP-Net architecture. It also offers a comparable performance with an iterative weighted sum-of-absolute value (IW-SOAV) algorithm [188] which is known as one of the most effective iterative algorithms for overloaded massive MIMO systems with a low computational complexity. However, the TPG based detector is dominated by a matrix inverse in each layer which increases the computational complexity. In addition, it has been designed for low modulation schemes and for a perfect CSI.

B. Varying Channel Realization Architecture

1) MM Network:: The MM network (MMNet) is proposed by Khani *et al.* in [146] to tackle a poor performance of the DetNet and OAMP-Net based detectors under realistic channels and online training issues. The main concept of the MMNet architecture is to offer a balance between the complexity and flexibility in the linear and denoising ingredients within each layer of the neural network [146]. It is designed based on the theory of iterative soft-thresholding algorithms [189]–[191] and leverages spectral and temporal correlation in realistic channels to speed up training procedure. It has an effective online training for varying channel realization in contrast to the DetNet and OAMP-Net architectures that employ a single detection model for all channels [146]. Furthermore, MMNet based detector can offer better performance by 4–8 dB than a traditional linear algorithms such as the MMSE detector [146]. MMNet attains a near-optimal performance with a small number of operations under i.i.d. Gaussian channels. It can achieve the same performance of the DetNet and OAMP-Net based detectors with low computational complexity. Under a realistic channel model, the MMNet based detector obtains the same performance of the OAMP-Net with 10× fewer computational complexity and lower SNR by 2.5dB [146].

The MMNet architecture offers more freedom degrees and has significantly more efficacious than the constrained OAMP-Net algorithm [146]. The linear stages within the MMNet architecture create appropriate conditions for the non-linear denoisers by shaping the distribution of noise at the input of denoisers to approach a Gaussian distribution. Furthermore, it does not include any matrix inverse which reduces the computational complexity. Compared to the DetNet architecture, it has a simple architecture with two learnable parameter. Unfortunately, the sequential online training within the MMNet architecture incurs latency [146]. MMNet requires to be retrained on each channel realization which leads impractical implementation scenario [192].

2) Hyper-Network:: In [192], a deep hyper-network-based uplink massive MIMO detection (HyperMIMO) is proposed to address the computationally demanding of the retraining for each channel realization in the MMNet architecture. The HyperMIMO architecture takes $\mathbf{H}$ as an input and generates the weights of the MMNet architecture [193]. The HyperMIMO architecture substitutes the training process that would be needed for each channel realization by one inference of a hypernetwork. It modifies the original MMNet architecture by decomposing $\mathbf{H}$ [1], and hence, the number of parameters are reduced as

$$\mathbf{H} = \mathbf{Q}^T \mathbf{R}, \tag{15}$$

where $\mathbf{Q}$ is $N_r \times N_r$ orthogonal matrix and $\mathbf{R}$ is $N_r \times N_t$ upper triangular matrix. Assume that $N_r > N_t$, then $\mathbf{R} = \begin{bmatrix} \mathbf{R_A} \\ \mathbf{0} \end{bmatrix}$ and $\mathbf{Q} = \begin{bmatrix} \mathbf{Q_A} \\ \mathbf{Q_B} \end{bmatrix}$, where $\mathbf{R_A}$ is a $N_t \times N_t$ matrix, and $\mathbf{Q_A}$ is a $N_r \times N_t$ matrix By leveraging the QR-decomposition in the MMNet architecture, the $N_t \times N_r$ trainable matrix $\Theta_k^{(1)}$, in [146], is become $N_t \times N_t$ instead of $N_t \times N_r$. In order to obtain a high performance, the $\mathbf{R_A}$ and the channel noise standard deviation σ are considered as inputs of the hypernetwork. To further decrease the number of hypernetwork outputs and inspired by [194], the HyperMIMO architecture employs a relaxed form of weight sharing. Figure 7 shows the model of HyperMIMO architecture which composed of the hypernetwork and the modified MMNet architecture. R2C is a layer to convert the elements of $\mathbf{R_A}$ from complex-valued to real-valued. C2R is a layer to make a reverse process of R2C layer. In addition, $\|.\|$ activation function is employed in the last layer to guarantee the positive values of the $\theta_k^{(1)}$ parameters. The hypernetwork has three dense layers where a number of units in the first dense layer is matching the number of inputs. The second dense layer has 75 units, and a number of units in the third dense layer depends on the number of parameters needed to the detector. The exponential linear unit (ELU) activation functions are used in first two layers and linear activation functions are used in the third dense layer.

The HyperMIMO based MIMO detector offers a performance near to that of the MMNet based detector and outperforms the OAMP-Net based detector. In addition, the HyperMIMO architecture is robust to user mobility within a certain range which is desired in a practical implementation.

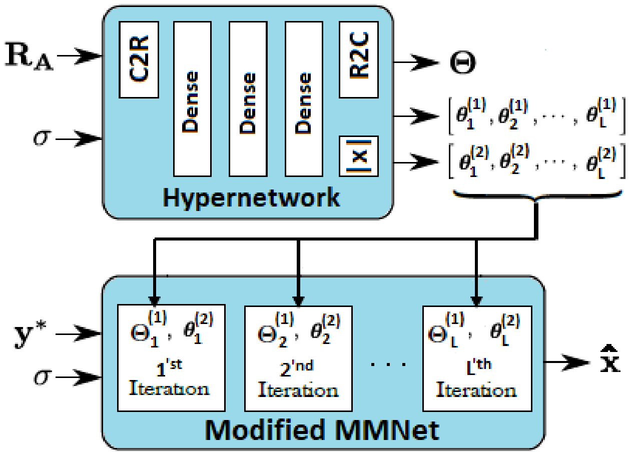
$\mathbf{R_A}$
σ
C2R
Dense
Dense
Dense
R2C
$|x|$
Hypernetwork
Θ
$[\theta_1^{(1)}, \theta_2^{(1)}, \cdots, \theta_L^{(1)}]$
$[\theta_1^{(2)}, \theta_2^{(2)}, \cdots, \theta_L^{(2)}]$
$\mathbf{y}^*$
σ
$\Theta_1^{(1)}, \theta_1^{(2)}$
1'st Iteration
$\Theta_2^{(1)}, \theta_2^{(2)}$
2'nd Iteration
. . .
$\Theta_L^{(1)}, \theta_L^{(2)}$
L'th Iteration
$\hat{\mathbf{x}}$
Modified MMNet

Figure 7. The design model of the HyperMIMO architecture

[1] In the HyperMIMO [192], the notation of N_r is used to denote the number of antennas at the BS and $N_t \backslash N_u$ to denote the number of antennas at user terminals.

IV. Deep Convolutional Neural Network

The DNN based massive MIMO detectors incur a performance loss in implementation and real scenarios. Hence, the correlation features between symbols over frequency or time domain cannot be exploited especially in vehicle MIMO systems. However, the DCNN has the ability to invest the correlation features between symbols and performs well in vehicle MIMO systems. Motivated by this, the DCNN is employed for detection in massive MIMO systems.

A. Deep Convolutional Neural Network Based ML Detection

A joint framework of DCNN-based ML detection (DCNN-MLD) is proposed by X. Junjuan *et al.* in [195] to investigate the conventional detection problem for vehicle networks with MIMO systems and to suppress the interference by exploiting the correlation features. Practical communications scenarios are considered where interfering signals, which may arise because of aggressive reuse of frequency resources, are correlated over time or frequency. In general, the DCNN-MLD tackles the inability of the ML detection, DetNet, OAMP-Net, and MMNet, to invest the correlation between different symbols over frequency or time domain, where these architectures fail to perform well with a practical correlated interference. The DCNN-MLD architecture is used in order to tackle limitations of the optimal ML sequential detector MLSD [195], where the symbol-by-symbol ML detector and the DCNN network are jointly used with K iterations between ML detector and the DCNN network. The structure of the DCNN-MLD architecture is depicted in Fig. 8. At each iteration, the ML detector firstly produces the estimation of the transmitted signal $\hat{\mathbf{x}}(n)$. Then, $\hat{\mathbf{x}}(n)$ is used to obtain the initial estimate of the interference $\hat{\mathbf{w}}(n)$ as

$$\hat{\mathbf{w}}(n) = \mathbf{y}(n)\,\mathbf{H}(n)\,\hat{\mathbf{x}}(n). \tag{16}$$

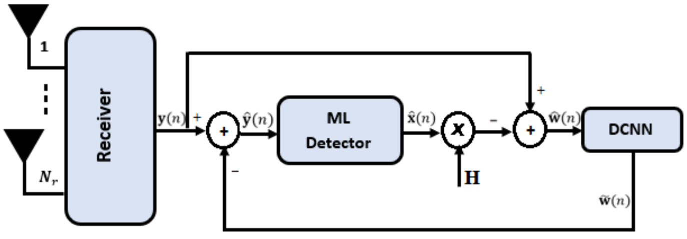

Figure 8. The structure of the DCNN-MLD architecture

The DCNN network, based on $\hat{\mathbf{w}}(n)$, produces a more precise estimation of the interference $\tilde{\mathbf{w}}(n)$ by exploiting the characteristics inherent in the interfering signals, essentially about the local correlation through different signals. $\tilde{\mathbf{w}}(n)$ is fed back into ML detector, and input signal of the ML detector for next iteration $\hat{\mathbf{y}}(n)$ can be obtained as

$$\begin{aligned}\hat{\mathbf{y}}(n) &= \mathbf{y}(n) - \tilde{\mathbf{w}}(n)\\ &= \mathbf{H}(n)\,\hat{\mathbf{x}}(n) + \mathbf{w}(n) - \tilde{\mathbf{w}}(n)\\ &\triangleq \mathbf{H}(n)\,\hat{\mathbf{x}}(n) + \mathbf{z}(n),\end{aligned} \tag{17}$$

where $\mathbf{w}(n)$ is the additive interfering signal and $\mathbf{z}(n)$ is the effective residual interference.

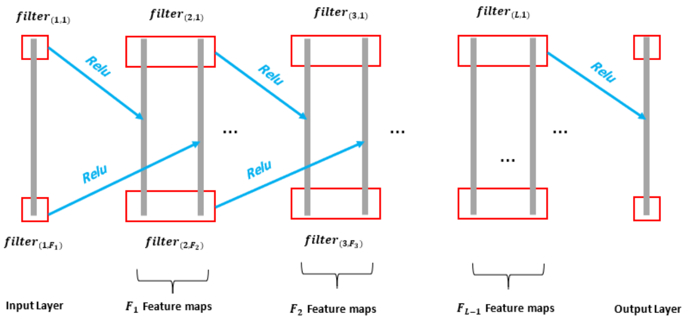

Figure 9. The structure of the DCNN network

Figure 9 shows the structure of DCNN network which is composed of L convolutional layers: the input convolutional layer, $(L-2)$ hidden convolutional layers, and the output convolutional layer. The input layer is convolved by F_1 filters, with filter size of R_1 denoted by $filter(1,f_1)|f_1 \in [1,F_1]$. Consequently, the input layer produces F_1 feature maps within the convolutional operation, and send it to a second convolutional layer which is composed of F_2 filters with filter size of R_2 and denoted by $filter(2,f_2)|f_1 \in [1,F_2]$. Generally, l^{th} layer $(1 \leqslant l \leqslant L)$ is convolved by F_l filters with a filter size of R_l and denoted by $filter(l,f_l)|f_1 \in [1,F_l]$. Consequently, the l^{th} layer obtains the F_{l-1} feature maps which are produced by the previous layer and then produce F_l feature maps for the next layer. This procedure is repeated until the final output layer is applied. The output of the convolutional layer is convolved by just one filter and the output feature maps have the dimension of the input data. Overall, the structure of the DCNN network has the following parameters $\{L;R_1,R_2,...,R_L;F_1,F_2,...,F_L\}$. The DCNN-MLD attains a good performance in the presence of correlated interference (i.e., Jakes model) over time or frequency and naturally in i.i.d. Gaussian channels. It outperforms the OAMP-Net based detector. Unfortunately, DCNN-MLD is just tested with an equal number of transmitter and receiver antennas, symmetric MIMO system, and with low-order modulation scheme. It also considers only suppressing the interfering signals with ignoring the distribution of the residual interference within the training procedure. Subsequently, performance of the DCNN-MLD architecture is deteriorated with a non-Gaussian interference.

However, authors in [195] have proposed an improved DCNN network to deal with non-Gaussian distribution case by forcing the residual interference to be near to the Gaussian distribution by devising the loss function through the cross-entropy of the detection. In [196], a generic DCNN-based linear detectors (G-DCNN) for MIMO systems over correlated noise environments is proposed. Instead of using the ML, the G-DCNN architecture is employed with the ZF, ZF-SIC, MMSE, and MMSE-SIC, to generate an initial estimate of transmitted signals. The G-DCNN architecture has substantially improved the performance in comparison with the conventional linear detectors.

B. Convolutional Neural-Network-Based Likelihood Ascent Search

In [197], a convolutional neural-network-based likelihood ascent search (CNNLAS) detection architecture and a graphical detection model are proposed for uplink multiuser (MU) massive MIMO systems. The proposed detector depends on the Vertical Bell Layered Space-Time (VBLAST) system where hundreds of centralized BS antennas received the signals from tens of users through the uplink channel. Compared with other competitive algorithms (i.e., MMSE, SDR, and DetNet), the CNNLAS architecture has a solid robustness against the channel estimation errors. In particular, the CNNLAS architecture, in presence of channel estimation errors, needs a significantly lower average received SNRs to acquire better BER performance. Furthermore, it offers a high spectral efficiency with a low computational complexity and with a significant low average received SNRs, both for low and high order modulation schemes (i.e., 16QAM and 64QAM). Particularly, performance of the CNNLAS based detector is evaluated in $N_t = N_r$ and $N_t < N_r$. For instance, the BER of 10^{-5} can be attained at SNRs of 5 dB, 12 dB and 14 dB for QPSK, 16-QAM and 64-QAM modulation schemes, respectively, with $N_t = N_r = 288$. However, the CNNLAS architecture is just tested over i.i.d. Gaussian channels [197].

In [198], a DL-based MIMO receiver architecture known as DeepRX for massive MIMO detection is proposed to improve the BER performance of the conventional linear MMSE receivers with perfect and imperfect CSI and over tapped delay line channel model. The DeepRx architecture consists of a residual neural network (ResNet)-based convolutional neural network [199]. The DeepRx architecture facilitates a detection mechanism in MIMO systems, which needs a separation of various overlapping spatial streams within the equalization and symbol detection stage. To reduce the complexity of DeepRx architecture, two novel transformation layers are proposed. First transformation layer is the maximal ratio combining-based transformation and depends on expert knowledge. Second transformation layer is a fully learned transformation that employs learned multiplicative layers.

The chronology of the DNNs for MIMO detectors is presented in Table IV where the efforts to exploit the DL in massive MIMO detection was started in 2017 when DetNet was proposed. Furthermore, the significance and limitations of the DNN detection algorithms in massive MIMO systems are comprehensively reviewed in Table V. The computational complexity of such detectors is illustrated in Table VI. Most of proposed algorithms suffer from a significant performance loss when high modulation scheme is used. In addition, the performance is deteriorated in realistic channel scenarios.

As shown in Tables VI and V, the detection process in the DNN-single channel realization architectures expect some of simplified versions of the DetNet architecture (i.e., LcgNet and QLcgNetV architectures) is not affected by instantaneous channel realization. In other words, these architectures experience a performance loss with a realistic channel (e.g., the QuaDRiGa channel). In addition, the DNN-single channel realization architectures cannot invest the correlation between different symbols over frequency or time domain. The OAMP-Net or the OAMP-Net2 architecture of the DNN-single channel realization architectures have a high computational complexity (i.e., $O(N_t^3)$). While, the DetNet, FS-Net, LcgNet architectures have a relatively high computational complexity (i.e., $O(N_t^2)$). The remaining architectures of the DNN-single channel realization architectures have lower computational complexity (i.e., $O(N_t)$) such as the ELRID, ScNet, BP and TPG architectures. Although the BP architectures has a low order of computational complexity, its computational complexity has linearly increasing with the order of modulation. The DNN-varying channel realization architectures come to overcome the inability of the DNN-single channel realization architectures to offer a well performance with a realistic channel. In addition, the DNN-varying channel realization architectures like MMNet architecture has $O(N_t^2)$) computational complexity and offers 10-15× less computational complexity than the OAMP-Net and DetNet architectures. Nevertheless, the DNN-varying channel realization architectures still have inability to invest the correlation between different symbols over frequency or time domain. The DCNN architectures (i.e., DCNN-MLD, G-DCNN-MMSE, and CNNLAS architectures) are mainly proposed to deal with an inability of the DNN architectures to invest the correlation between different symbols over frequency or time domain. Unfortunately, the DCNN suffers from a high computational complexity. The DCNN-MLD architecture has a very high computational complexity and its computational complexity increases exponentially with the size of MIMO system and the order of modulation. The DCNN-MLD architecture has $O(N_t^3)$) computational complexity. The CNNLAS architecture has $O(N_t^2)$) computational complexity which is comparable with the computational complexity of the DetNet architecture. It is worth to note that most of the DNN and DCNN architectures suffer from a performance loss when a high order of modulation is utilized.

V. Discussion and Further Research Direction

The research in DL for massive MIMO and cell-free massive MIMO detectors is still in its infancy and there is a significant room for fundamental research contributions in data detection and channel estimation such as:

A. Online Training

Instead of the classical detection theory to obtain the best estimate of unknown vectors, DL architectures could be exploited to choose the best algorithm to be applied. As shown in Section III, most of the DL based detection architectures were learned offline to exhibit a satisfactory performance over unfolding detection algorithms (i.e., the GD and AMP based detectors). Unfortunately, the training process is very complex and time consuming. Although the offline training was conduced to reduce the time consumption, it still consumes several days in some computing architectures. The time-consuming overhead depends on two factors: (i) the total number of required training samples (i.e., batch size), and (ii) the size of the model. On other hand, employing of the online training in DL based detection architectures need to be optimized for every

Table IV
CHRONOLOGY OF DEEP NEURAL NETWORK FOR MIMO DETECTORS

Year	Summary of work performed	Channels	MIMO size
2017	Introduction of deep network (DetNet) where unfolding a projected GD method is utilized. It is designed to handle multiple channels simultaneously with a single training phase. The training is implemented in TensorFlow [145].	Random i.i.d. Gaussian channels	30×60
2018	An architecture is proposed to improve the activation function in the DetNet architecture by using a multi-level-plateau sigmoid activation function. The modified DetNet architecture employed twin DL networks with various initial values to simultaneously detect the transmitted signals [163].	Random i.i.d. Gaussian channels	8×8
2018	A simplified version of the DetNet architecture called sparsely connected neural network (ScNet) is proposed. The simplification is done by reducing the number of inputs, reducing the number of network layers, and optimizing the loss function [158].	Random i.i.d. Gaussian channels	20×30 & 40×80
2018	An OAMP-Net, a model-driven DL network, is proposed. It adds some adjustable parameters to the existing iterative method, OAMP. It is designed for a perfect CSI. The training is implemented in TensorFlow [166] [147].	Rayleigh channels	4×4 & 8×8 & 64×64
2018	A DL based BP detector is proposed. The structure of proposed detector contains a four-layers neural network to minimize the loss function. Simulations are implemented in TensorFlow [166] [180].	Rayleigh channels	8×8 & 16×16
2019	A joint detector based on linear and non-linear algorithms is utilized. The structure of neural networks is proposed to decrease the number of mapping between inputs and outputs [200].	Typical channel model as proposed in [201]	8×8 & 32×32 & 64×64
2019	The fully connected multi-layer network (FullyCon) and modified DetNet architecture are proposed, where a relatively small number of parameters are required to optimize. In FullyCon, the output of each layer is the input of the next layer. the DetNet architecture applies a projected GD method for detecting the signal. The training is implemented in TensorFlow [166] [22].	Channels with covariance matrices of a uniform linear array based on the one-ring model in [202]	20×30 & 30×60 & 15×25
2019	A DL based sphere decoding (DL-SD) detector is proposed. Prior to decoding, the hypershpere radius is learned by the DNN. A sequence of learned radiuses at its output layer is mapped into a sequence of the fading channel matrix entries by the DNN. The training is implemented using the DL Toolbox of MATLAB 2019a and Adam optimizer [174].	Rayleigh block-fading channels	10×10
2019	A trainable project gradient detector (TPG-detector) is proposed based on the GD step and the soft projection step. Internal parameters are optimized by DL techniques. The training is implemented in PyTorch 0.4.0 [203] [173], [187].	Flat Rayleigh fading	32×50 & 64×100 & 96×150
2019	A comprehensive summary of unfolded learned algorithms for massive MIMO detection algorithms is presented. In this paper, several future research directions have been mentioned such as the need of acceleration methods, the convergence of the training method, on-line and off-line training, and hardware implementation [204].	Rayleigh and correlated channels	8×8
2019	The recent advancements in model-driven DL approaches in physical layer communications (i.e., DL based detectors, transmission schemes, receiver design, and channel information recovery). In addition, various open research areas are highlighted [154].	Random i.i.d. Gaussian channels	32×32
2019	A DNN is employed to enhance message passing detectors (MPDs) for MIMO systems is proposed. It is based on modified MPDs including damped BP (dBP), max-sum (MS) BP, and simplified channel hardening-exploiting message passing (CHEMP). The training is implemented in TensorFlow [166] and Adam optimizer [167] [175].	i.i.d. Rayleigh and correlated fading channels	8×32 & 64×128
2019	A detector based on alternating direction method of multipliers network (ADMM-Net) is proposed for the BPSK and QPSK constellation cases. The training is implemented in TensorFlow [159].	Random Gaussian channels	30×60 & 40×60 & 160×160
2020	A fast-convergence sparsely connected detection network (FS-Net) is proposed. It approximates the initial solution of a DL-aided tabu search (TS) algorithm and it is acquired by optimizing the DetNet and ScNet architecture. The training is implemented in TensorFlow [166] and Adam optimizer [167] [164].	Random i.i.d. Gaussian channels	32×32 & 32×64
2020	A DL network (DLNet) is proposed where exact knowledge of channel parameters is assumed. The DNN layers depend on projected GD. The training is implemented in the TensorFlow library with the Adam optimizer [165].	Random i.i.d. Gaussian channels	30×60 & 20×30
2020	Two model-driven networks, namely learned CG network (LcgNet) and quantized LcqNet (QL-cqNetV), based on the learned CG descent have been proposed where each layer is considered as one iteration with additional parameters. The training is implemented in Python using the TensorFlow library with the Adam optimizer [162].	Rayleigh fading channel	32×32 & 32×64 & 32×128
2020	This paper proposes a framework for systematic complexity scaling of DNN, called as weight-scaling neural-network (WeSNet), for massive MIMO detection. This work introduced the concept of monotonic non-increasing profile function to allow the network to dynamically learn the best attenuation strategy for its own weights during the training. Training was performed using the TensorFlow [168].	Rayleigh fading channels	30×60 & 16×16
2020	An OAMP-Net2 is proposed. It is a development of OAMP-Net architecture in [147] where some new trainable parameters are considered. Unlike the OAMP-Net architecture, it is designed for imperfect CSI and channel estimation is improved by a data-aided scheme. It contains a linear and a nonlinear estimators. The training is implemented in TensorFlow [166] [171].	Rayleigh fading channels	8×8 & 16×16 & 32×32
2020	A low complexity symbol detection technique based on iterative meta-predictor aided collaborative learning is proposed for symbol detection in massive MIMO with large number of users. The training is implemented in Adam optimizer [152].	Random i.i.d. Gaussian channels	50×256
2020	A DNN based on unfolding MPD is proposed. The DNN structure is exploited to optimize the damping and correction factors. The training is implemented in TensorFlow [166] and Adam optimizer [74], [167].	i.i.d. Rayleigh and correlated fading MIMO channels	8×128
2020	A model-driven DL method is proposed by unfolding an iterative algorithm in [205]. Two auxiliary parameters at each layer are introduced. The first parameter generates the residual error vector while the second parameter adjusts the relationship among previous layers. The training is implemented in Adam optimizer [167] [161].	i.i.d. Gaussian channels	8×128 & 8×64

2020	A DNN is exploited to calculate the optimal damping factor (DF) where it is trained off-line for each antenna configuration. The training is implemented in Adam optimizer [167] [179].	i.i.d. Rayleigh-fading channels	16×16
2020	A DNN is exploited to improve the detection performance deterioration due to the mismatches of the channel correlations between training and test in the deep neural network-based damped BP (DNN-dBP). In addition, the convergence property of the BP algorithm is enhanced by applying the node selection method. It is trained off-line for each antenna configuration and implemented in Adam optimizer [167] [98].	i.i.d. Rayleigh-fading channels	16×16
2020	MMNet is proposed to overcome the challenges in DetNet and OAMP-Net architectures. It is based on iterative soft thresholding algorithm and is implemented in QuaDRiGa channel simulator. The training is implemented in Adam optimizer [167] [146].	3GPP 3D MIMO channels as implemented in QuaDRiGA	16×32 & 16×64
2020	A HyperMIMO-based detector replaces the training process required by MMNet architecture for each channel realization by a single inference through a trained hypernetwork. It also reduces the number of parameters of MMNet architecture. Training was performed using the Adam optimizer [192].	Channel spatial correlation matrices	6×12
2020	A convolutional-neural-network based likelihood ascent search (CNNLAS) based on a graphical detection model is proposed for uplink MU massive MIMO systems is proposed. It is trained off-line [197].	i.i.d. Gaussian channels	288×288 & 32×48
2021	A modified expectation propagation (EP)-based MIMO detector (MEPD) is proposed to tackle the challenges of the conventional EP detector [182] which are incurred due to the empirical parameter selection such as damping factors and initial variance. Furthermore, an modified EP network (MEPNet) is proposed to offer the optimal damping factors and initial variance by adopting a DL scheme and unfolding the proposed iterative MEPD detector. It is trained off-line with 5 layers and implemented in TensorFlow [166] and Adam optimizer. [167] [181].	i.i.d. Rayleigh-fading channels	32×32 & 16×16 & 64×64 & 128×128
2021	A fast DL aided SD (FDL-SD) and a fast DL-aided M-best SD (FDL-MSD) architectures are proposed to accelerate the searching process in the SD and M-best SD algorithms. Furthermore, these proposed architectures have more beneficial in both off-line training and on-line decoding phases in contrast to the SD-DL architecture [174] [186].	Highly correlated channels and i.i.d. Rayleigh channels	32×32 & 16×16
2021	A fully convolutional neural network based receiver MIMO receiver architecture known as DeepRX for MIMO detection is proposed. The DeepRx architecture consists of a ResNet-based convolutional neural network and has a significant higher BER performance than convention linear MMSE receivers with perfect and imperfect CSI and over tapped delay line channel model [198].	A 5G physical uplink shared channel (PUSCH) scenario	4×16
2021	A DNN based SDR detection algorithm is proposed on the basis of the graphical detection model [206].	Flat fading channels	128×256 & 128×128 & 256×256
2021	A model-driven DL based massive MIMO detector is proposed by improving the approximate expectation propagation algorithm. It is constructed by adding learnable parameters to enhance the performance and the convergence robustness through the DL approach. The proposed detector is is trained by Adam optimizer with 150 iterations [207].	i.i.d. Rayleigh fading channels	32×64 & 32×48

realization of the channel instead of using a fixed detector for a wide variety of channels. However, online training is mostly dominated by the cost of computational complexity for each new realization of the channel which depends on the channel coherence time. Furthermore, the majority of the DL based detection architectures have a significant difficulty to be trained online due to the stringent performance requirements (i.e., the number of learnable parameters, wide SNR range, and the number of iterations). Therefore, an online training should be carefully considered in the future work.

One of the major challenges in applying machine learning, AI and DL in communications systems design relates to the training complexity and the generalization capabilities of the trained models. The initial results in the literature suggest that deep unfolding methods, which efficiently capitalize the known structure of the problem, tend to be more often practically feasible than purely data based approaches. This is of no surprise, because the efficient use of model structure, when known, outperform the purely data based ones. The strength of machine learning based approaches is in solving inference problems for which precise models are not available. The other key strength is in approximating computationally complex operations by a DNN. Further work is needed in understanding this general trade-off in practical MIMO and cell-free networks with realistic channel and hardware models.

Sections II & III show that most existing detectors were initialized using the linear MMSE which is diagonally dominant and has a high complexity because of the matrix inverse. Therefore, approximate/avoid matrix inversion methods, such as the GS, the SOR, the RI, the JA, and others, could be considered in the initialization to reduce the computational complexity. In addition, the MMSE is diagonally dominant, and hence, most of the existing linear MMSE detectors are using the diagonal matrix which may not converge in some scenarios. In our recent work [208], it is shown that the utilization of a stair matrix, instead of a diagonal matrix, impacts greatly the convergence rate, the performance, and the computational complexity. Therefore, the work in [208] can be extended to test the efficiency of a stair matrix in DL based detectors. On the other hand, instead of doing a straightforward matrix inversion, several matrix decomposition methods can be utilized which are more numerically stable. Matrix decomposition algorithms have been extensively utilized for the matrix inversion procedure of a small-scale MIMO detection [114], [209]. They provide better numerical stability over straightforward inversion methods. As shown in Section III, QR decomposition was exploited in few works. However, the LDL and Cholesky decomposition algorithms were not well investigated in the context of DL based detectors.

B. Real Channel Scenario

As shown in Section I, an optimal balance between the performance and computational complexity of the entire communications system can be obtained with channel modeling in realistic scenarios and effective signal processing. Although most DL based detection architectures demonstrated a strong

Table V
PROS AND CONS OF DEEP NEURAL NETWORK FOR DETECTION ARCHITECTURE IN MASSIVE MIMO SYSTEMS

Architecture	Significance	Limitation
FullyCon	• It has has a small number of optimized parameters [22]. • It achieves a near-optimal accuracy over the fixed channels [22]. • It is fast [22].	• Over varying channels, FullyCon has poor performance and did not manage to learn how to detect properly [22].
DetNet	• It performs very well in case of i.i.d. complex Gaussian channel matrices and low-order modulation schemes (i.e., BPSK and 4-QAM) [146]. • It performs well over both constant and Rayleigh fading channels with a single training shot [22], [154]. • It overcomes the performance of the AMP and SDR algorithms [155], [156]. • The BER performance is comparable with the $M-$best SD [22]. • It offers the performance of the SDR detection algorithm, with 30× faster [155], [156].	• The training is unstable for realistic channels (i.e., QuaDRiGa channel simulator) [146]. • DetNet's performance on correlated channels is not satisfactory [195]. • Due to a large number of tuning parameters, it has a poor scalability (in large MIMO size and high modulation scheme). In other words, it is prohibitively expensive to train on-line [161], [162]. • It performs poorly for a larg-scale MIMO systems with $N_t \approx N_r$ [161]. • It does not employ known features of iterative algorithms [22]. • It employs a single detection model for all channels [146].
DLNet	• It achieves a good performance and a low complexity in i.i.d. channels [165]. • It offers a better BER performance with 164× faster in running speed and 9× less computational complexity than the DetNet architecture [165]. • It achieves a comparable BER performance with the SDR algorithm, with 28200× faster [165].	• CSI has to be perfectly known [165]. • It is not tested in a realistic channel environment [165].
LcgNet	• It has a very limited number of learnable parameters compared with the Fullycon and DetNet architectures [162]. • The matrix-vector multiplication and division operations, are replaced by some prestored parameters which are fixed during on-line detections [162]. • It has an universal step-sizes [162]. • It has a good performance in realistic channel models (i.e., a spatial correlated channel model) and low-order modulation scheme [162].	• It suffers from a considerable performance loss when higher order modulation is used [162].
QLcgNetV	• It has a very limited number of learnable parameters compared with the Fullycon and DetNet architectures [162]. • It reduces the memory costs brought up by the storage of the step-sizes with minor performance loss [162]. • It has an universal step-sizes [162]. • It quantizes the learned parameters by using a low-resolution nonuniform (i.e., 3-bits or 4-bits) quantizer [162]. • It has a good performance in realistic channel models (i.e., a spatial correlated channel model) and low-order modulation scheme [162].	• It suffers from a considerable performance loss when higher order modulation is used [162].
WeSNet	• The neural network architecture is self-adjustable to the detection complexity [168]. • It deals with the weight profile functions themselves as trainable parameters in order to prohibit vanishing gradients [168]. • It achieves a good performance with low-order modulation schemes (i.e., BPSK and 4-QAM) [168]. • The detector is evaluated under asymmetric and symmetric channels [168]. • It outperforms the DetNet and OAMP-Net architectures [168]. • It outperforms the DetNet architecture with offering 51.43% reduction in complexity and about 50% reduction in model size [168]. • It outperforms the OAMP-Net architecture and offers detection accuracy similar to the SDR algorithm with about 10× lower computational complexity [168].	• It is designed for low-order modulation schemes [168]. • It is not tested in a realistic channel scenario (i.e., QuaDRiGa simulator) [168].
OAMP-Net	• It addresses the large amount of trained parameters in the DetNet architecture [147]. • It has just two trainable parameters in each layer. • It offers a soft decision [146], [147]. • It significantly overcomes the OAMP algorithm [146]. • It is considered as the next-best learning algorithm [146]. • It has a very good performance in i.i.d. Gaussian channels and low-order modulation scheme [146], [147].	• It does not generalize to realistic channels with a spatial correlation [146], [147]. • It is very restrictive where a strict assumption has to exist (i.e., unitarily-invariant channel matrices). The performance degrades significantly when the channel matrices do not conform to this assumption [146]. • It has a high computational complexity and it is dominant by the matrix inverse in each layer [146], [147]. • It employs a single detection model for all channels [146]. • It cannot exploit the correlation among different symbols over time or frequency domain [146], [195].
OAMP-Net2	• It enhances the performance of the OAMP-Net architecture [171]. • It considers the characteristics of channel estimation error and statistics of channel [171]. • It is designed for imperfect CSI [171]. • It has strong robustness to SNR, channel correlation, modulation scheme, and MIMO configuration mismatches [171]. • It needs only four trainable parameters in each layer [171]. • It outperforms the OAMP and the LMMSE algorithms [171].	• It suffers from a large performance loss over the realistic channel [171]. • It has a high computational complexity and it is dominant by the matrix inverse in each layer [171].
MMNet	• It tackles a poor performance of the DetNet and OAMP-Net architectures under realistic channels [146]. • It is designed to be trained on-line for each realization of **H**. In other words, the receiver parameters are continually adapted as new **H** is observed [146]. • It offers a balance between complexity and flexibility in the linear and denoising ingredients within each layer [146]. • It does not require any matrix inverse operation [146]. • It achieves a near-ML performance with 10-15x less computational complexity than the OAMP-Net and DetNet architectures [146]. • It offers a better BER performance by about 4–8 dB than a traditional linear algorithm such as MMSE algorithm [146].	• It incurs latency due to the sequential on-line training [146]. • Its performance degrades of high modulation scheme (i.e., 64QAM) [146]. • It performs the symbol-by-symbol detection, and cannot exploit the correlation among different symbols over time or frequency domain [146]. • It needs to be retrained on each channel realization, which makes its practical implementation challenging [146], [192].
HyperMIMO	• It reduces the number of parameters of the MMNet architecture and weight sharing and hence, a low computational complexity is required [192]. • It substitutes the training process that would be needed for each channel realization by one inference of the hypernetwork neural network [192]. • It outperforms the OAMP-Net architecture and LMMSE based detector [192]. • It has a practical implementation where it is robust against user mobility [192].	• It achieves SER slightly worse than the MMNet architecture [192]. • It needs to be re-trained when the channel statistics change significantly [192].

DCNN-MLD	• It has a good performance in i.i.d. Gaussian channels, and low-order modulation scheme (i.e., QPSK) [195]. • It has a good performance in the presence of correlated interference (i.e.,Jakes model) over time or frequency [195]. • It outperforms the performance of the conventional ML detector and has significant better performance than the OAMP-Net architecture [195]. • It considers practical communications scenarios and deals with an aggressive reuse of frequency resources [195]. • It tackles the inability of the ML detection, DetNet, OAMP-Net, and MMNet to invest the correlation between different symbols over frequency or time domain [195].	• It is just tested with an equal number of transmitter and receiver antennas, symmetric MIMO system, and with low-order modulation scheme (i.e., QPSK) [195]. • Its performance is deteriorated for non-Gaussian interference [195]. • It experiences a high computational complexity of the ML detection [195], [196]. • It ignores the distribution of the residual interference within the training procedure [195].
G-DCNN	• It is designed for correlated noise environments (i.e., Jakes model) [196]. • It has a good performance in low-order modulation scheme (i.e., BPSK) [196]. • It employs a low complexity detectors (i.e., ZF, ZF-SIC, MMSE, and MMSE-SIC) instead of ML detector [196].	• The initialization has a matrix inverse component, and hence, the computational complexity is increased [196]. • It is just tested with an equal number of transmitter and receiver antennas, symmetric MIMO system, and with low-order modulation scheme (i.e., QPSK) [196].
CNNLAS	• It has a solid robustness with the channel estimation errors [197]. • It is based on a graphical detection model [197]. • It outperforms the performance of the MMSE, and SDR algorithms [197]. • It outperforms the performance of the DetNet architecture [197]. • It has a very good performance in i.i.d. Gaussian channels, and low-order and high-order modulation scheme (i.e., 16QAM and 64QAM) [197].	• It just tested over i.i.d. Gaussian channels [197].
ADMM-Net	• It needs just matrix-vector multiplications and/or simple element-wise multiplications in each layer [159]. • It has a lesser number of trainable parameters than that in the DetNet architecture [159]. • It outperforms the performance of the ZF, SDR, and DetNet based detectors with a small number of layers [159]. • It can achieve a good performance in low-order modulation scheme (i.e., BPSK and QPSK) [159]. • It has a good performance in i.i.d. Gaussian channels [159].	• It has a relatively high computational complexity [159]. • It is designed for low-order modulation schemes [159]. • It is not taking into consideration the realistic channel scenario [159]. • Its average run time is higher than the SD algorithm [159].
DL-SD	• Its learnable radiuses lead to a remarkable reduction of lattice points inside the sphere [174]. • It significantly reduces the computational complexity of the SD algorithm [174]. • Its probability of failing to find a solution is close to zero [174]. • It has a very good performance in i.i.d. Gaussian channels and high-order modulation scheme (i.e., 64QAM) [174]. • It has a significant performance gain in contrast to the MMSE algorithm and has a comparable performance with the SD-IRS algorithm with a remarkable complexity reduction [174].	• It is designed based on a fully connected feedforward neural network [174]. • It includes a matrix inversion [174]. • It needs an off-line training for each SNR value [174].
TPG	• It is suitable for overloaded massive MIMO systems [173], [187]. • It requires just three trainable parameters in each layer [173], [187]. • The number of trainable parameters does not depend on the MIMO size [173], [187]. • It can avoid the vanishing gradient problems by using an incremental training [173], [187]. • It has a more scalability in terms of the computational complexity than the FullyCon, DetNet, and OAMP-Net architectures [173], [187]. • It does not need to initialize the LMMSE matrix in each layer as that in the OAMP-Net architecture [173], [187]. • It offers a comparable performance with IW-SOAV algorithm [173], [187], [188]. • It has a very good performance in i.i.d. Gaussian channels and low-order modulation scheme (i.e., QPSK) [173], [187].	• It is designed for a perfect CSI [173], [187]. • It is dominant by the matrix inverse in each layer [173], [187]. • It considers just an i.i.d. Gaussian channels [173], [187].

performance in both i.i.d. Gaussian and small-sized correlated with perfect CSI, the detection performance of these architectures deteriorates with a realistic channel (e.g., the QuaDRiGa channel) or with an imperfect CSI. In other words, most existing detectors are not adaptable to changes in channel statistics and realizations. They were designed based on simple channel models. In addition, most existing detectors are designed for low modulation orders and they are not performing well when a high modulation order is used. Therefore, realistic channel scenarios and high modulation orders should be considered in future proposals.

C. Cell-free Massive MIMO

Cell-free massive MIMO has a potential to play a crucial role in 5G and B5G systems. It is not a trivial task to efficiently conduct distributed signal processing tasks. Therefore, DL architectures could be a solution to reduce the complexity. Although the DL has been utilized to address the power allocation and channel estimation in cell-free massive MIMO, it is not yet exploited in detection algorithms [81], [210], [211]. The literature has shown a paucity of employing DL architectures for detection in decentralized and cell-free massive MIMO systems. DL architectures could be exploited in cell-free massive MIMO networks to obtain the best estimate of unknown vectors. The high traffic load on the fronthaul and backhaul is one core problem in the cell-free network. DL with federated learning could reduce that for the control and learning side, while the actual data sharing may still be needed.

We would also like to note that the approximate inversion-based detectors might not be suitable for cell-free massive MIMO systems. The approximate inversion-based detectors utilize the channel hardening property where the diagonal

Table VI
THE COMPUTATIONAL COMPLEXITY OF DL BASED DETECTOR ARCHITECTURES

Architecture	Computational Complexity	Notes	Reference
DNN-Single Channel Realization Architectures			
ELRID	$O(N_rN_t+N_t)N_t)$	-	[152]
DetNet	$O(36N_t^2+4N_t)L)$	-	[22], [162]
ScNet	$O(3N_tL)$	-	[158]
FS-Net	$O((8N_t^2+8N_{t)}L)$	-	[164]
LcgNet	$O(4N_t^2+6N_t)L)$	-	[162]
OAMP-net	$O(N_t^3L)$	-	[171]
OAMP-net2	$O(N_t^3L)$	-	[171]
DNN-dBP	$O(MN_tN_rL)$	M is the order of the modulation scheme.	[74]
DNN-MS	$O(MN_tN_rL)$	M is the order of the modulation scheme.	[74]
DNN-sMPD	$O(MN_tN_rL)$	M is the order of the modulation scheme	[74]
TPG	$O(N_tN_rL)$	-	[173], [187]
DNN-Varying Channel Realization Architectures			
MMNet	$O(bN_t^2L)$	b is the batch size	[146]
DCNN Architectures			
DCNN-MLD	$O((K+1)M^{N_t}+\sum_{l=1}^{L}(F_{l-1}R_lN_tF_l))$	F is the number of feature maps, R is the filter size and M is the order of the modulation scheme	[195]
G-DCNN-MMSE	$O((K+1)N_t^3+\sum_{l=1}^{L}(F_{l-1}R_lN_tF_l))$	F is the number of feature maps and R is the filter size	[195]
CNNLAS	$O(N_t^2+N_c^2+N^2)$	N_c is the total number of convolutional filters and N is the total number of the 3-D graphical signal matrices.	[197]

terms of the Gram matrix $\mathbf{H}^H\mathbf{H}$ become significantly more dominant than the off-diagonal terms. The APs of a cell-free system is distributed over a large geographical area. The AP antennas are not co-located like conventional massive MIMO systems, but rather distributed over the geographical area of a cell. Therefore, it is still unclear whether the cell-free massive MIMO systems will inherit the channel hardening property from a conventional massive MIMO system. In [212], the authors used stochastic geometry to investigate this issue. The authors concluded that having many distributed antennas does not necessarily lead to channel hardening. The channel hardening criterion is strongly affected by the number of antennas per AP and the propagation environment. This result is significant in the MIMO detection context due to the fact that most low complexity detectors utilize the channel hardening property. If channel hardening is not reliable, then the highly complex non-linear detectors, such as sphere decoder or successive interference cancellation, could be the appropriate solution in cell-free systems. The DL detectors cannot compete with approximate-inversion based detectors in terms of complexity. However, the DL detectors can be a competitive solution in comparison to sphere decoders or SIC. The DL solutions are also not inferior to any nonlinear solutions in terms of error-rate performance.

There has been a tremendous and reinvigorating interest on DL techniques, circuits and platforms during this decade. Major semiconductor industries have invested heavily to develop platforms supporting DL algorithms. It is highly likely that the future base stations will have generic hardware accelerators to support different DL algorithms. Therefore, the complexity of DL based MIMO detectors will be less relevant as it is currently, but the focus will shift on utilizing these DL hardware accelerators by configuring them for different computation intensive applications, such as uplink MIMO detection. Regardless of that the learning approaches utilizing the data as efficiently as possible reduce the training complexity and cost.

VI. CONCLUSION

A remarkable research dedicated to the massive MIMO receiver's design was conducted. In this paper, a review of various detection techniques based on DL architectures was provided to achieve optimal and quasi-optimal performance. Unfortunately, optimal performance was achieved in expenses of a high computational complexity. We covered detectors based on FullyCon, DetNet, and OAMP-Net and their variations. In addition, deep hyber networks, WeSNet, and ScNet were comprehensively illustrated. The architecture and impact of DCNN, generic DCNN architecture, and CNNLAS were also discussed. This paper also reviewed the ADMM networks, SD with radius selection and TPG detectors based on the DL are presented. The HyperMIMO based detector can achieve a satisfactory performance in real scenarios. Although most of proposed detectors were not tested in real scenarios and high order modulation schemes, there is a room for contributions to develop the DL based detectors in centralized, decentralized, and cell-free massive MIMO systems.

REFERENCES

[1] P. Cerwall, A. Lundvall, P. Jonsson, S. Carson, R. Moller, P. Lindberg, K. Ohman, J. Travers, F. Pedersen, P. Linder, J. Sethi, P. Rinderud, J. Rubio, J. Garcia, H. Hemmer, C. Ashraf, and A. Powell, "Ericsson

mobility report june 2020," *Ericsson Mobility Report*, pp. 1–36, June 2020.

[2] "Global networking trends report," *Report*, pp. 1–95, 2020.

[3] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, February 2014.

[4] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," vol. 68, no. 7, pp. 4247–4261, 2020.

[5] H. Huang, S. Guo, G. Gui, Z. Yang, J. Zhang, H. Sari, and F. Adachi, "Deep learning for physical-layer 5G wireless techniques: Opportunities, challenges and solutions," *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 214–222, 2020.

[6] T. Wang, C. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," *China Commun.*, vol. 14, no. 11, pp. 92–111, 2017.

[7] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 3, no. 3, pp. 210–229, 1959.

[8] I. Ahmad, S. Shahabuddin, H. Malik, E. Harjula, T. Leppänen, L. Lovén, A. Anttonen, A. H. Sodhro, M. M. Alam, M. Juntti, A. Yla-Jaaski, T. Sauter, A. Gurtov, M. Ylianttila, and J. Riekki, "Machine learning meets communication networks: Current trends and future challenges," *IEEE Access*, pp. 1–1, 2020.

[9] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 2002.

[10] H. B. Barlow, "Unsupervised learning," *Neural computation*, vol. 1, no. 3, pp. 295–311, 1989.

[11] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2005.

[12] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.

[13] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, 2013.

[14] B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S. M. Siniscalchi, and C. Lee, "An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1289–1300, 2017.

[15] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure–activity relationships," *Journal of chemical information and modeling*, vol. 55, no. 2, pp. 263–274, 2015.

[16] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk, "Connectomic reconstruction of the inner plexiform layer in the mouse retina," *Nature*, vol. 500, no. 7461, pp. 168–174, 2013.

[17] C. Zhang, P. Zhou, C. Li, and L. Liu, "A convolutional neural network for leaves recognition using data augmentation," in *IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 2015, pp. 2143–2150.

[18] H. Ye, G. Y. Li, and B. 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, 2018.

[19] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, and Y. Be'ery, "Deep learning methods for improved decoding of linear codes," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 119–131, 2018.

[20] M. Kim, N. Kim, W. Lee, and D. Cho, "Deep learning-aided SCMA," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 720–723, 2018.

[21] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2595–2621, 2018.

[22] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2554–2564, 2019.

[23] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224–2287, 2019.

[24] Y.-D. Huang, P. P. Liang, Q. Zhang, and Y.-C. Liang, "A machine learning approach to MIMO communications," in *Proc. IEEE Int. Conf. Commun.*, 2018, pp. 1–6.

[25] A. Elgabli, A. Elghariani, A. O. Al-Abbasi, and M. Bell, "Two-stage LASSO ADMM signal detection algorithm for large scale MIMO," in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, 2017, pp. 1660–1664.

[26] V. Carletti, A. Greco, G. Percannella, and M. Vento, "Age from faces in the deep learning revolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2113–2132, 2020.

[27] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[28] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

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

[30] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1941–1988, 2015.

[31] M. A. Albreem, M. Juntti, and S. Shahabuddin, "Massive MIMO detection techniques: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3109–3132, 2019.

[32] M. A. Albreem, W. Salah, A. Kumar, M. H. Alsharif, A. H. Rambe, M. Jusoh, and A. N. Uwaechia, "Low complexity linear detectors for massive MIMO: A comparative study," *IEEE Access*, vol. 9, pp. 45 740–45 753, 2021.

[33] K. Zheng, L. Zhao, J. Mei, B. Shao, W. Xiang, and L. Hanzo, "Survey of large-scale MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1738–1760, 2015.

[34] N. Fatema, G. Hua, Y. Xiang, D. Peng, and I. Natgunanathan, "Massive MIMO linear precoding: A survey," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3920–3931, Dec 2018.

[35] S. Jeong, A. Farhang, F. Gao, and M. F. Flanagan, "Frequency synchronisation for massive MIMO: a survey," *IET Communications*, vol. 14, no. 16, pp. 2639–2645, 2020.

[36] C. Wang, J. Bian, J. Sun, W. Zhang, and M. Zhang, "A survey of 5G channel measurements and models," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3142–3168, 2018.

[37] P. Zhang, J. Chen, X. Yang, N. Ma, and Z. Zhang, "Recent research on massive MIMO propagation channels: A survey," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 22–29, December 2018.

[38] A. N. Uwaechia and N. M. Mahyuddin, "A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges," *IEEE Access*, vol. 8, pp. 62 367–62 414, 2020.

[39] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive MIMO—5G system," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 905–923, Secondquarter 2016.

[40] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017.

[41] Z. Gong, C. Li, and F. Jiang, "Pilot contamination mitigation strategies in massive MIMO systems," *IET Communications*, vol. 11, no. 16, pp. 2403–2409, 2017.

[42] I. Ahmad, S. Shahabuddin, T. Kumar, J. Okwuibe, A. Gurtov, and M. Ylianttila, "Security for 5G and beyond," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3682–3722, 2019.

[43] N. Haider, M. Z. Baig, and M. Imran, "Artificial intelligence and machine learning in 5G network security: Opportunities, advantages, and future research trends," *arXiv preprint arXiv:2007.04490*, 2020.

[44] J. Suomalainen, A. Juhola, S. Shahabuddin, A. Mämmelä, and I. Ahmad, "Machine learning threatens 5G security," *IEEE Access*, vol. 8, pp. 190 822–190 842, 2020.

[45] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data," vol. 14, no. 7, pp. 3899–3911, 2015.

[46] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar 2017.

[47] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. IEEE Int. Conf. Commun.*, 2019, pp. 1–6.

[48] S. Elhoushy, M. Ibrahim, and W. Hamouda, "Cell-free massive MIMO: A survey," *IEEE Commun. Surveys Tuts.*, pp. 1–1, 2021.

[49] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99 878–99 888, 2019.

[50] Z. H. Shaik, E. Björnson, and E. G. Larsson, "MMSE-optimal sequential processing for cell-free massive MIMO with radio stripes," *arXiv preprint arXiv:2012.13928*, 2020.

[51] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, 2017.

[52] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and B. D. Rao, "Performance of cell-free massive MIMO systems with MMSE and LSFD receivers," in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, 2016, pp. 203–207.

[53] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, 2020.

[54] Z. Chen and E. Björnson, "Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5205–5219, Nov 2018.

[55] T. Q. Duong, X. Chu, and H. A. Suraweera, *Security for Cell-free Massive MIMO Networks*, 2019, pp. 135–150.

[56] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli, E. Peschiera, R. Trevisan, G. Maschietto, E. Torri, R. Inchingolo, A. Smargiassi, G. Soldati, P. Rota, A. Passerini, R. J. G. van Sloun, E. Ricci, and L. Demi, "Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2676–2687, 2020.

[57] T. Würfl, M. Hoffmann, V. Christlein, K. Breininger, Y. Huang, M. Unberath, and A. K. Maier, "Deep learning computed tomography: Learning projection-domain weights from image domain in limited angle problems," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1454–1463, 2018.

[58] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2063–2079, 2018.

[59] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3152–3168, 2020.

[60] X. Yuan, Y. Gu, Y. Wang, C. Yang, and W. Gui, "A deep supervised learning framework for data-driven soft sensor modeling of industrial processes," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4737–4746, 2020.

[61] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, 2020.

[62] A. Fehske, J. Gaeddert, and J. H. Reed, "A new approach to signal classification using spectral correlation and neural networks," in *Proc. IEEE Int. Symp. Dynamic Spectrum Access Networks*, 2005, pp. 144–150.

[63] M. Ibukahla, J. Sombria, F. Castanie, and N. J. Bershad, "Neural networks for modeling nonlinear memoryless communication channels," *IEEE Trans. Commun.*, vol. 45, no. 7, pp. 768–771, 1997.

[64] M. Ibnkahla, N. J. Bershad, J. Sombrin, and F. Castanie, "Neural network modeling and identification of nonlinear channels with memory: algorithms, applications, and analytic models," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1208–1220, 1998.

[65] M. Ibnkahla, "Neural network modeling and identification of nonlinear MIMO channels," in *Proc. IEEE Int. Symp. on Signal Process. and Info. Technol.*, 2007, pp. 1–4.

[66] 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 Veh. Technol. Mag.*, vol. 67, no. 9, pp. 8549–8560, Sep. 2018.

[67] Y. Jin, J. Zhang, B. Ai, and X. Zhang, "Channel estimation for mmWave massive MIMO with convolutional blind denoising network," *IEEE Commun. Lett.*, vol. 24, no. 1, pp. 95–98, 2020.

[68] E. Balevi, A. Doshi, and J. G. Andrews, "Massive MIMO channel estimation with an untrained deep neural network," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2079–2090, 2020.

[69] 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, 2020.

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

[71] H. He, C. 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, 2018.

[72] Z. Gao, Y. Wang, X. Liu, F. Zhou, and K. Wong, "FFDNet-based channel estimation for massive MIMO visible light communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 340–343, 2020.

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

[74] X. Tan, W. Xu, K. Sun, Y. Xu, Y. Be'ery, X. You, and C. Zhang, "Improving massive MIMO message passing detectors with deep neural network," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1267–1280, Feb 2020.

[75] T. Demir and E. Björnson, "Channel estimation in massive MIMO under hardware non-linearities: Bayesian methods versus deep learning," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 109–124, 2020.

[76] J. Yuan, H. Q. Ngo, and M. Matthaiou, "Machine learning-based channel prediction in massive MIMO with channel aging," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 2960–2973, 2020.

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

[78] J. Li, Q. Zhang, X. Xin, Y. Tao, Q. Tian, F. Tian, D. Chen, Y. Shen, G. Cao, Z. Gao, and J. Qian, "Deep learning-based massive MIMO CSI feedback," in *IEEE Int. Conf. on Opt. Commun. Net.*, 2019, pp. 1–3.

[79] Y. Jin, J. Zhang, S. Jin, and B. Ai, "Channel estimation for cell-free mmWave massive MIMO through deep learning," *Proc. IEEE Veh. Technol. Conf.*, vol. 68, no. 10, pp. 10 325–10 329, 2019.

[80] M. Bashar, A. Akbari, K. Cumanan, H. Q. Ngo, A. G. Burr, P. Xiao, M. Debbah, and J. Kittler, "Exploiting deep learning in limited-fronthaul cell-free massive MIMO uplink," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1678–1697, 2020.

[81] N. Athreya, V. Raj, and S. Kalyani, "Beyond 5G: Leveraging cell free TDD massive MIMO using cascaded deep learning," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1533–1537, 2020.

[82] C. D'Andrea, A. Zappone, S. Buzzi, and M. Debbah, "Uplink power control in cell-free massive MIMO via deep learning," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019, pp. 554–558.

[83] S. Chakraborty, E. Björnson, and L. Sanguinetti, "Centralized and distributed power allocation for max-min fairness in cell-free massive MIMO," in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, 2019, pp. 576–580.

[84] K. V. Vardhan, S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low-complexity detector for large MIMO systems and multicarrier CDMA systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 473–485, April 2008.

[85] J. Sun, Y. Zhang, J. Xue, and Z. Xu, "Learning to search for MIMO detection," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7571–7584, 2020.

[86] B. Hassibi, "An efficient square-root algorithm for BLAST," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.*, vol. 2, 2000, pp. II737–II740 vol.2.

[87] S. Loyka and F. Gagnon, "Analytical framework for outage and BER analysis of the V-BLAST algorithm," in *Proc. Int. Zurich Seminar Broadband Commun.*, 2004, pp. 120–123.

[88] Y. Jiang, M. K. Varanasi, and J. Li, "Performance analysis of ZF and MMSE equalizers for MIMO systems: An in-depth study of the high SNR regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2008–2026, April 2011.

[89] B. Cerato and E. Viterbo, "Hardware implementation of a low-complexity detector for large MIMO," in *Proc. IEEE Int. Symp. on Circuits and Systems*, May 2009, pp. 593–596.

[90] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's (belief propagation) algorithm," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 2, pp. 140–152, Feb 1998.

[91] D. J. C. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 399–431, Mar 1999.

[92] K. Takeuchi, T. Tanaka, and T. Kawabata, "Performance improvement of iterative multiuser detection for large sparsely spread CDMA systems by spatial coupling," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1768–1794, April 2015.

[93] Y. Gao, H. Niu, and T. Kaiser, "Massive MIMO detection based on belief propagation in spatially correlated channels," in *Proc. ITG Workshop Smart Antennas*, Feb 2017, pp. 1–6.

[94] A. Mezghani and J. A. Nossek, "Belief propagation based MIMO detection operating on quantized channel output," in *Proc. IEEE Int. Symp. Inform. Theory*, June 2010, pp. 2113–2117.

[95] J. Yang, C. Zhang, X. Liang, S. Xu, and X. You, “Improved symbol-based belief propagation detection for large-scale MIMO,” in *2015 IEEE Workshop on Signal Processing Systems (SiPS)*, Oct 2015, pp. 1–6.

[96] F. Long, T. Lv, R. Cao, and H. Gao, “Single edge based belief propagation algorithms for MIMO detection,” in *34^{t}h IEEE Sarnoff Symposium*, 2011, pp. 1–5.

[97] J. Hu and T. M. Duman, “Graph-based detection algorithms for layered space-time architectures,” *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 2, pp. 269–280, 2008.

[98] J. Tachibana and T. Ohtsuki, “Learning and analysis of damping factor in massive MIMO detection using BP algorithm with node selection,” *IEEE Access*, vol. 8, pp. 96 859–96 866, 2020.

[99] B. Kang, J. H. Yoon, and J. Park, “Low complexity massive MIMO detection architecture based on Neumann method,” in *Proc. Int. SoC Design Conf.*, Nov 2015, pp. 293–294.

[100] C. Tang, C. Liu, L. Yuan, and Z. Xing, “High precision low complexity matrix inversion based on Newton iteration for data detection in the massive MIMO,” *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 490–493, March 2016.

[101] P. Zhang, L. Liu, G. Peng, and S. Wei, “Large-scale MIMO detection design and FPGA implementations using SOR method,” in *Proc. IEEE Int. Conf. Commun. Software and Net.*, June 2016, pp. 206–210.

[102] Z. Wu, L. Ge, X. You, and C. Zhang, “Efficient near-MMSE detector for large-scale MIMO systems,” in *Proc. IEEE Workshop on Signal Proess. Syst.*, Oct 2017, pp. 1–6.

[103] J. Minango and A. C. Flores, “Low-complexity MMSE detector based on refinement Jacobi method for massive MIMO uplink,” *Physical Communication*, vol. 26, pp. 128 – 133, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1874490717302550

[104] H. Costa and V. Roda, “A scalable soft Richardson method for detection in a massive MIMO system,” *Przeglad Elektrotechniczny*, vol. 92, no. 5, pp. 199–203, August 2016.

[105] J. Jin, Y. Xue, Y. L. Ueng, X. You, and C. Zhang, “A split pre-conditioned conjugate gradient method for massive MIMO detection,” in *Proc. IEEE Workshop on Signal Proess. Syst.*, Oct 2017, pp. 1–6.

[106] C. Xiao, X. Su, J. Zeng, L. Rong, X. Xu, and J. Wang, “Low-complexity soft-output detection for massive MIMO using SCBiCG and Lanczos methods,” *China Commun.*, vol. 12, no. Supplement, pp. 9–17, December 2015.

[107] B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, “Conjugate gradient-based soft-output detection and precoding in massive MIMO systems,” in *Proc. IEEE Global Telecommun. Conf.*, Dec 2014, pp. 3696–3701.

[108] Y. Hu, Z. Wang, X. Gaol, and J. Ning, “Low-complexity signal detection using CG method for uplink large-scale MIMO systems,” in *Proc. IEEE Int. Conf. Commun. Syst.*, Nov 2014, pp. 477–481.

[109] B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, “VLSI design of large-scale soft-output MIMO detection using conjugate gradients,” in *Proc. IEEE Int. Symp. on Circuits and Systems*, May 2015, pp. 1498–1501.

[110] K. Li, B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, “Accelerating massive MIMO uplink detection on GPU for SDR systems,” in *Proc. IEEE Dallas Circuits and Syst. Conf.*, Oct 2015, pp. 1–4.

[111] Y. Xue, C. Zhang, S. Zhang, and X. You, “A fast-convergent pre-conditioned conjugate gradient detection for massive MIMO uplink,” in *2016 IEEE International Conference on Digital Signal Processing (DSP)*, 2016, pp. 331–335.

[112] L. Liu, G. Peng, P. Wang, S. Zhou, Q. Wei, S. Yin, and S. Wei, “Energy- and area-efficient recursive-conjugate-gradient-based MMSE detector for massive MIMO systems,” *IEEE Trans. Signal Process.*, vol. 68, pp. 573–588, 2020.

[113] I. Al-Nahhal, M. Alghoniemy, O. Muta, and A. B. A. El-Rahman, “Reduced complexity K-best sphere decoding algorithms for ill-conditioned MIMO channels,” in *IEEE Annual Cons. Commun. Netw. Conf.*, Jan 2016, pp. 183–187.

[114] S. Shahabuddin, M. H. Islam, M. S. Shahabuddin, M. A. Albreem, and M. Juntti, “Matrix decomposition for massive MIMO detection,” in *2020 IEEE Nordic Circuits and Systems Conference (NorCAS)*, 2020, pp. 1–6.

[115] M. A. M. Albreem and M. F. M. Salleh, “Near-A$_n$-lattice sphere decoding technique assisted optimum detection for block data transmission systems,” *IEICE Trans. Commun.*, vol. E96-B, no. 1, pp. 365–359, Jan 2013.

[116] M. A. M. Albreem, “An efficient lattice sphere decoding technique for multi-carrier systems,” *Wireless Personal Communications*, vol. 82, no. 1, pp. 1825–1831, Jan 2015.

[117] M. El-Khamy, H. Vikalo, B. Hassibi, and R. J. McEliece, “On the performance of sphere decoding of block codes,” in *Proc. IEEE Int. Symp. Inform. Theory*, July 2006, pp. 1964–1968.

[118] R. S. Mozos and M. J. Fernandez-Getino Garcia, “Efficient complex sphere decoding for MC-CDMA systems,” *IEEE Transactions on Wireless Communications*, vol. 5, no. 11, pp. 2992–2996, 2006.

[119] C. A. Shen and A. M. Eltawil, “A radius adaptive K-best decoder with early termination: Algorithm and VLSI architecture,” *IEEE Trans. Circuits Syst. I*, vol. 57, no. 9, pp. 2476–2486, Sept 2010.

[120] T. H. Kim and I. C. Park, “Small-area and low-energy K-best MIMO detector using relaxed tree expansion and early forwarding,” in *Proc. IEEE Int. Symp. on Low-Power Electron. and Design*, Aug 2010, pp. 231–236.

[121] Y. H. Wu, Y. T. Liu, H. C. Chang, and Y. C. Liao, “Early-pruned K-best sphere decoding algorithm based on radius constraints,” in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 4496–4500.

[122] Q. Li and Z. Wang, “Early-pruning K-best sphere decoder for MIMO systems,” in *Proc. IEEE Workshop on Signal Proess. Syst.*, Oct 2007, pp. 40–44.

[123] K. C. Lai, C. C. Huang, and J. J. Jia, “Variation of the fixed-complexity sphere decoder,” *IEEE Commun. Lett.*, vol. 15, no. 9, pp. 1001–1003, September 2011.

[124] B. Hassibi and H. Vikalo, “On the expected complexity of sphere decoding,” in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, vol. 2, Nov 2001, pp. 1051–1055 vol.2.

[125] B. M. Hochwald and S. ten Brink, “Achieving near-capacity on a multiple-antenna channel,” *IEEE Trans. Commun.*, vol. 51, no. 3, pp. 389–399, June 2003.

[126] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, “Ubiquitous cell-free massive MIMO communications,” *EURASIP J. Wireless Comm. and Netw.*, vol. 2019, no. 1, p. 197, 2019.

[127] J. Zhang, Y. Wei, E. Björnson, Y. Han, and S. Jin, “Performance analysis and power control of cell-free massive MIMO systems with hardware impairments,” *IEEE Access*, vol. 6, pp. 55 302–55 314, Sep 2018.

[128] A. Adhikary, A. Ashikhmin, and T. L. Marzetta, “Uplink interference reduction in large-scale antenna systems,” *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2194–2206, May 2017.

[129] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, “On the uplink max–min SINR of cell-free massive MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2021–2036, Apr 2019.

[130] Y. Feng, M. Wang, D. Wang, and X. You, “Low complexity iterative detection for a large-scale distributed MIMO prototyping system,” in *Proc. IEEE Int. Conf. Commun.*, May 2019, pp. 1–6.

[131] Q. Liu, H. Liu, Y. Yan, and P. Wu, “A distributed detection algorithm for uplink massive MIMO systems,” in *Proc. IEEE Workshop on Signal Proess. Syst.*, 2019, pp. 213–217.

[132] M. Guo and M. C. Gursoy, “Distributed sparse activity detection in cell-free massive MIMO systems,” in *Proc. IEEE Global Conf. Sign. Inf. Proc.* Institute of Electrical and Electronics Engineers Inc., 2019, p. 8969500.

[133] H. Song, X. You, C. Zhang, O. Tirkkonen, and C. Studer, “Minimizing pilot overhead in cell-free massive MIMO systems via joint estimation and detection,” in *Proc. IEEE Works. on Sign. Proc. Adv. in Wirel. Comms.*, 2020, pp. 1–5.

[134] A. M. Saray, J. Pourrostam, S. H. Mousavi, and M. M. Feghhi, “A low-complexity space time block codes detection for cell free massive MIMO systems,” in *2020 28th Iranian Conference on Electrical Engineering (ICEE)*, 2020, pp. 1–5.

[135] R. Gholami, L. Cottatellucci, and D. Slock, “Favorable propagation and linear multiuser detection for distributed antenna systems,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.*, 2020, pp. 5190–5194.

[136] M. Bashar, A. Akbari, K. Cumanan, H. Q. Ngo, A. G. Burr, P. Xiao, and M. Debbah, “Deep learning-aided finite-capacity fronthaul cell-free massive MIMO with zero forcing,” in *Proc. IEEE Int. Conf. Commun.*, 2020, pp. 1–6.

[137] K. Li, R. R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, “Decentralized baseband processing for massive MU-MIMO systems,” *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 7, no. 4, pp. 491–507, 2017.

[138] K. Li, Y. Chen, R. Sharan, T. Goldstein, J. R. Cavallaro, and C. Studer, “Decentralized data detection for massive MU-MIMO on a Xeon Phi cluster,” in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, 2016, pp. 468–472.

[139] C. Jeon, K. Li, J. R. Cavallaro, and C. Studer, “Decentralized equalization with feedforward architectures for massive MU-MIMO,” *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4418–4432, 2019.

[140] K. Li, J. McNaney, C. Tarver, O. Castañeda, C. Jeon, J. R. Cavallaro, and C. Studer, "Design trade-offs for decentralized baseband processing in massive MU-MIMO systems," in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, 2019, pp. 906–912.

[141] J. Rodríguez Sánchez, F. Rusek, O. Edfors, M. Sarajlić, and L. Liu, "Decentralized massive MIMO processing exploring daisy-chain architecture and recursive algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 687–700, 2020.

[142] J. R. Sanchez, F. Rusek, M. Sarajlic, O. Edfors, and L. Liu, "Fully decentralized massive MIMO detection based on recursive methods," in *Proc. IEEE Workshop on Signal Proess. Syst.*, 2018, pp. 53–58.

[143] K. Li, C. Jeon, J. R. Cavallaro, and C. Studer, "Decentralized equalization for massive MU-MIMO on FPGA," in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, Nov 2017, pp. 1532–1536.

[144] Z. Zhang, H. Li, Y. Dong, X. Wang, and X. Dai, "Decentralized signal detection via expectation propagation algorithm for uplink massive MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 11 233–11 240, Oct 2020.

[145] N. Samuel, T. Diskin, and A. Wiesel, "Deep MIMO detection," in *Proc. IEEE Works. on Sign. Proc. Adv. in Wirel. Comms.*, July 2017, pp. 1–5.

[146] M. Khani, M. Alizadeh, J. Hoydis, and P. Fleming, "Adaptive neural signal detection for massive MIMO," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2020.

[147] H. He, C. Wen, S. Jin, and G. Y. Li, "A model-driven deep learning network for MIMO detection," in *Proc. IEEE Global Conf. Sign. Inf. Proc.*, 2018, pp. 584–588.

[148] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, vol. 4, no. 1, pp. 1–58, Jan 1992.

[149] Jan Kozak, *Decision Tree and Ensemble Learning Based on Ant Colony Optimization*, 2019.

[150] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct 2019.

[151] R. Polikar, "Ensemble learning," in *Ensemble machine learning*. Springer, 2012, pp. 1–34.

[152] A. Datta, M. T. Deo, and V. Bhatia, "Collaborative learning based symbol detection in massive MIMO," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1678–1682.

[153] L. Xiang, Y. Liu, T. Van Luong, R. G. Maunder, L. L. Yang, and L. Hanzo, "Deep-learning-aided joint channel estimation and data detection for spatial modulation," *IEEE Access*, vol. 8, pp. 191 910–191 919, Oct 2020.

[154] H. He, S. Jin, C. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 77–83, Oct 2019.

[155] Z. Luo, W. Ma, A. M. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, May 2010.

[156] Wing-Kin Ma, T. N. Davidson, Kon Max Wong, Zhi-Quan Luo, and Pak-Chung Ching, "Quasi-maximum-likelihood multiuser detection using semi-definite relaxation with application to synchronous CDMA," *IEEE Trans. Signal Process.*, vol. 50, no. 4, pp. 912–922, Apr 2002.

[157] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[158] G. Gao, C. Dong, and K. Niu, "Sparsely connected neural network for massive MIMO detection," in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, Dec 2018, pp. 397–402.

[159] M. Un, M. Shao, W. Ma, and P. C. Ching, "Deep MIMO detection using ADMM unfolding," in *IEEE Data Scei. Workshop*, 2019, pp. 333–337.

[160] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Oct 2015, pp. 1–9.

[161] J. Liao, J. Zhao, F. Gao, and G. Y. Li, "A model-driven deep learning method for massive MIMO detection," *IEEE Commun. Lett.*, pp. 1–1, 2020.

[162] Y. Wei, M. M. Zhao, M. Hong, M. J. Zhao, and M. Lei, "Learned conjugate gradient descent network for massive MIMO detection," *IEEE Trans. Signal Process.*, vol. 68, pp. 6336–6349, 2020.

[163] V. Corlay, J. J. Boutros, P. Ciblat, and L. Brunel, "Multilevel MIMO detection with deep learning," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Oct 2018, pp. 1805–1809.

[164] N. T. Nguyen and K. Lee, "Deep learning-aided tabu search detection for large MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4262–4275, June 2020.

[165] S. Kumar, A. Singh, and R. Mahapatra, "Deep learning based massive-MIMO decoder," in *Proc. Int. Conf. on adv. Networks and Telecommun. Syst.*, 2019, pp. 1–6.

[166] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.

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

[168] A. Mohammad, C. Masouros, and Y. Andreopoulos, "Complexity-scalable neural network based MIMO detection with learnable weight scaling," *IEEE Trans. Commun.*, pp. 1–1, 2020.

[169] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020–2033, Jan 2017.

[170] D. Ito, S. Takabe, and T. Wadayama, "Trainable ISTA for sparse signal recovery," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3113–3125, June 2019.

[171] H. He, C. Wen, S. Jin, and G. Y. Li, "Model-driven deep learning for MIMO detection," *IEEE Trans. Signal Process.*, vol. 68, pp. 1702–1715, 2020.

[172] X. Tan, Z. Zhong, Z. Zhang, X. You, and C. Zhang, "Low-complexity message passing MIMO detection algorithm with deep neural network," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov 2018, pp. 559–563.

[173] S. Takabe, M. Imanishi, T. Wadayama, R. Hayakawa, and K. Hayashi, "Trainable projected gradient detector for massive overloaded MIMO channels: Data-driven tuning approach," *IEEE Access*, vol. 7, pp. 93 326–93 338, 2019.

[174] M. Mohammadkarimi, M. Mehrabi, M. Ardakani, and Y. Jing, "Deep learning-based sphere decoding," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4368–4378, 2019.

[175] X. Tan, W. Xu, K. Sun, Y. Xu, Y. Be'ery, X. You, and C. Zhang, "Improving massive MIMO message passing detectors with deep neural network," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1267–1280, 2020.

[176] J. Yang, W. Song, S. Zhang, X. You, and C. Zhang, "Low-complexity belief propagation detection for correlated large-scale MIMO systems," *Journal of Signal Processing Systems*, vol. 90, no. 4, pp. 585–599, Apr 2018.

[177] Y. Zhang, L. Ge, X. You, and C. Zhang, "Belief propagation detection based on max-sum algorithm for massive MIMO systems," in *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct 2017, pp. 1–6.

[178] T. L. Narasimhan and A. Chockalingam, "Channel hardening-exploiting message passing (chemp) receiver in large MIMO systems," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr 2014, pp. 815–820.

[179] J. Tachibana and T. Ohtsuki, "Damping factor learning of BP detection with node selection in massive MIMO using neural network," in *Proc. IEEE Veh. Technol. Conf.*, 2020, pp. 1–6.

[180] X. Liu and Y. Li, "Deep MIMO detection based on belief propagation," in *Proc. IEEE Inform. Theory Workshop*, Nov 2018, pp. 1–5.

[181] H. Chen, G. Yao, and J. Hu, "Algorithm parameters selection method with deep learning for EP MIMO detector," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 146–10 156, 2021.

[182] J. Céspedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, "Expectation propagation detection for high-order high-dimensional MIMO systems," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2840–2849, Aug 2014.

[183] H. Lopes and N. Souto, "Iterative signal detection for large-scale GSM-MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7734–7738, Aug 2018.

[184] A. Elgabli, A. Elghariani, V. Aggarwal, and M. R. Bell, "A low-complexity detection algorithm for uplink massive MIMO systems based on alternating minimization," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 917–920, June 2019.

[185] A. Elgabli, A. Elghariani, V. Aggarwal, M. Bennis, and M. R. Bell, "A proximal Jacobian ADMM approach for fast massive MIMO signal detection in low-latency communications," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–6.

[186] N. T. Nguyen, K. Lee, and H. DaiIEEE, "Application of deep learning to sphere decoding for large MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6787–6803, 2021.

[187] S. Takabe, M. Imanishi, T. Wadayama, and K. Hayashi, "Deep learning-aided projected gradient detector for massive overloaded MIMO channels," in *Proc. IEEE Int. Conf. Commun.*, 2019, pp. 1–6.

[188] R. Hayakawa and K. Hayashi, "Convex optimization-based signal detection for massive overloaded MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7080–7091, Nov 2017.

[189] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, November 2004.

[190] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, March 2009.

[191] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1227–1231.

[192] M. Goutay, F. Ait Aoudia, and J. Hoydis, "Deep hypernetwork-based MIMO detection," pp. 1–5, 2020.

[193] Q. V. L. David Ha, Andrew Dai, "HyperNetworks," *arXiv*, vol. 4, Sep 2016.

[194] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Advances in neural information processing systems*, 2016, pp. 523–531.

[195] J. Xia, K. He, W. Xu, S. Zhang, L. Fan, and G. K. Karagiannidis, "A MIMO detector with deep learning in the presence of correlated interference," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4492–4497, 2020.

[196] K. He, Z. Wang, W. Huang, D. Deng, J. Xia, and L. Fan, "Generic deep learning-based linear detectors for MIMO systems over correlated noise environments," *IEEE Access*, vol. 8, pp. 29 922–29 929, 2020.

[197] L. Li, H. Hou, and W. Meng, "Convolutional-neural-network-based detection algorithm for uplink multiuser massive MIMO systems," *IEEE Access*, vol. 8, pp. 64 250–64 265, 2020.

[198] D. Korpi, M. Honkala, J. M. Huttunen, and V. Starck, "Deeprx MIMO: Convolutional MIMO detection with learned multiplicative transformations," in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–7.

[199] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[200] Z. Jia, W. Cheng, and H. Zhang, "A partial learning-based detection scheme for massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1137–1140, 2019.

[201] C.-K. Wen, S. Jin, K.-K. Wong, J.-C. Chen, and P. Ting, "Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1356–1368, 2015.

[202] D.-S. Shiu, G. Foschini, M. Gans, and J. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 502–513, 2000.

[203] P. Team, "Pytorch," *URL: https://pytorch.org*, 2019.

[204] A. Balatsoukas-Stimming and C. Studer, "Deep unfolding for communications systems: A survey and some new directions," in *Proc. IEEE Workshop on Signal Proess. Syst.*, 2019, pp. 266–271.

[205] M. Mandloi and V. Bhatia, "Low-complexity near-optimal iterative sequential detection for uplink massive MIMO systems," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 568–571, March 2017.

[206] H. Hou, L. Li, and W. Meng, "Deep neural network based detection algorithm for high-order modulation in uplink massive MIMO," in *2021 International Wireless Communications and Mobile Computing (IWCMC)*, 2021, pp. 1326–1331.

[207] Y. Ge, X. Tan, Z. Ji, Z. Zhang, X. You, and C. Zhang, "Improving approximate expectation propagation massive MIMO detector with deep learning," *IEEE Wireless Communications Letters*, vol. 10, no. 10, pp. 2145–2149, 2021.

[208] M. Albreem, M. Juntti, and S. Shahabuddin, "Efficient initialisation of iterative linear massive MIMO detectors using a stair matrix," *IEE Electron. Lett.*, vol. 56, no. 1, pp. 50–52, 2020.

[209] S. Shahabuddin, M. Juntti, and C. Studer, "ADMM-based infinity norm detection for large MU-MIMO: Algorithm and VLSI architecture," in *Proc. IEEE Int. Symp. on Circuits and Systems*, May 2017, pp. 1–4.

[210] Y. Zhao, I. G. Niemegeers, and S. H. De Groot, "Power allocation in cell-free massive MIMO: A deep learning method," *IEEE Access*, vol. 8, pp. 87 185–87 200, 2020.

[211] M. Alrabeiah and A. Alkhateeb, "Deep learning for TDD and FDD massive MIMO: Mapping channels in space and frequency," in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, 2019, pp. 1465–1470.

[212] Z. Chen and E. Björnson, "Can we rely on channel hardening in cell-free massive MIMO?" in *2017 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2017, pp. 1–6.

Mahmoud A. Albreem (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from Islamic University of Gaza, Palestine, in 2008, and the M.Sc. (EE) and Ph.D. (EE) degrees from University Sains Malaysia (USM), Malaysia, in 2010 and 2013, respectively.

From 2014 till 2016, Dr. Albreem was a Senior Lecturer with University Malaysia Perlis. In 2016–2021, he chaired Department of Electronics and Communications Engineering, A'Sharqiyah University, Oman. Currently, he is an Assistant Professor with Department of Electrical Engineering, University of Sharjah, UAE. He is also a Visiting Assistant Professor with Centre for Wireless Communications (CWC), University of Oulu, Finland. Dr. Albreem is the author of more than 70 journal and conference papers. He received several scholarships and grants, such as Nokia Foundation Centennial Grant (2018), USM Fellowship (2011-2013), and Best Master's Thesis Award of the School of Electrical and Electronics Engineering, USM (2010). Dr. Albreem served on the Editorial Board for Journal of Wireless Communications and Mobile Computing. His research interests include multiple-input multiple-output detection and precoding techniques, machine learning applications for wireless communication systems, and green communications.

Alaa H. Al Habbash was born in Riyadh, Saudi Arabia, in 1985. He received the B.Sc. degree and the M.Sc. degree in Telecommunication Engineering from Islamic University Gaza, Palestine, in 2008 and 2013, respectively. In 2020-2021, he was a research assistant at A'Sharqiyah University, Oman. He is currently working as a junior researcher in wireless communications at the Palestinian ICT research agency (P-ICTRA), Gaza, Palestine, and as a communication engineer at Ministry of Telecommunication, Gaza, Palestine. Al Habbash is co-recipient of 2018 IEEE ICEPT Best Paper Awards. His current research interests are space-time coding, turbo codes, spatial modulation, and deep learning.

Shahriar Shahabuddin received his M.Sc. and Ph.D. degrees from the Centre for Wireless Communications, University of Oulu, Finland. During Spring 2015, he worked at the Computer Systems Laboratory of Cornell University, NY, USA as a Visiting Researcher. He received distinction in M.Sc. and several scholarships and grants such as Nokia Foundation Scholarship, University of Oulu Scholarship Foundation Grant, Tauno Tönning Foundation Grant during his Doctoral studies. Since 2017 and 2020, he has been with Nokia Networks, Oulu, Finland and Nokia of America Corporation, Dallas, TX, USA, respectively. During 2021, Shahriar has received the title of Adjunct Professor from Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland. His research interest includes VLSI signal processing, 6G communication systems and wireless network security. He is a member of IEEE Communication Society and IEEE Circuits and Systems Society.

Markku Juntti (S'93–M'98–SM'04–F'20) received his M.Sc. (EE) and Dr.Sc. (EE) degrees from University of Oulu, Oulu, Finland in 1993 and 1997, respectively.

Dr. Juntti was with University of Oulu in 1992–98. In academic year 1994–95, he was a Visiting Scholar at Rice University, Houston, Texas. In 1999–2000, he was a Senior Specialist with Nokia Networks in Oulu, Finland. Dr. Juntti has been a professor of communications engineering since 2000 at University of Oulu, Centre for Wireless Communications (CWC), where he leads the Communications Signal Processing (CSP) Research Group. He also serves as Head of CWC – Radio Technologies (RT) Research Unit. His research interests include signal processing for wireless networks as well as communication and information theory. He is an author or co-author in almost 500 papers published in international journals and conference records as well as in books *Wideband CDMA for UMTS* in 2000–2010, *Handbook of Signal Processing Systems* in 2013 and 2018 and *5G Wireless Technologies* in 2017. Dr. Juntti is also an Adjunct Professor at Department of Electrical and Computer Engineering, Rice University, Houston, Texas, USA.

Dr. Juntti is an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and served previously in similar role in IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He was Secretary of IEEE Communication Society Finland Chapter in 1996–97 and the Chairman for years 2000–01. He has been Secretary of the Technical Program Committee (TPC) of the 2001 IEEE International Conference on Communications (ICC), and the Chair or Co-Chair of the Technical Program Committee of several conferences including 2006 and 2021 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), the Signal Processing for Communications Symposium of IEEE Globecom 2014, Symposium on Transceivers and Signal Processing for 5G Wireless and mm-Wave Systems of IEEE GlobalSIP 2016, ACM NanoCom 2018, and 2019 International Symposium on Wireless Communication Systems (ISWCS). He has also served as the General Chair of 2011 IEEE Communication Theory Workshop (CTW 2011) and 2022 IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC).