AI Empowered Wireless Communications: From Bits to Semantics

By Zhijin Qin, *Senior Member IEEE*, Le Liang, *Member IEEE*, Zijing Wang, *Member IEEE*, Shi Jin, *Fellow IEEE*, Xiaoming Tao, *Senior Member IEEE*, Wen Tong, *Fellow IEEE*, and Geoffrey Ye Li, *Fellow IEEE*

ABSTRACT | Artificial intelligence (AI) and machine learning (ML) have shown tremendous potential in reshaping the landscape of wireless communications and are, therefore, widely expected to be an indispensable part of the next-generation wireless network. This article presents an overview of how AI/ML and wireless communications interact synergistically to improve system performance and provides useful tips and tricks on realizing such performance gains when training AI/ML models. In particular, we discuss in detail the use of AI/ML to revolutionize key physical layer and lower medium access control (MAC) layer functionalities in traditional wireless communication systems. In addition, we provide a comprehensive overview of the AI/ML-enabled semantic communication systems, including key techniques from data generation to transmission. We also investigate the role of AI/ML as an optimization tool to facilitate the design of efficient resource allocation algorithms in wireless communication networks at both bit and semantic levels. Finally, we analyze major challenges and roadblocks in applying AI/ML in practical wireless system design and share our thoughts and insights on potential solutions.

KEYWORDS | Artificial intelligence (AI); machine learning (ML); semantic communications; wireless communications.

I. INTRODUCTION

A. Developments of Wireless Communications

The success of wireless communications is one of the biggest achievements in the history of science and technology, changing the way people and machines interact with each other. In 1948, Shannon [1] developed a rigorous mathematical framework, explaining how information is transmitted through a transmission channel and what is the limit of a communication system. Shannon's information theory has been regarded as the foundation of modern communication technology. Since then, wireless communications have experienced significant technical revolutions and have evolved through five generations. The first generation (1G) of cellular communications was introduced in the 1980s, which marked the advent of analog cellular networks and enabled basic voice communication by utilizing frequency-division multiple access (FDMA). The second generation (2G) of cellular communications was developed in the 1990s, which marked the advent of digital cellular networks and enabled both

Manuscript received 2 January 2024; revised 8 April 2024; accepted 22 July 2024. The work of Zhijin Qin was supported in part by the National Key Research and Development Program of China under Grant 2023YFB2904300 and in part by the National Natural Science Foundation of China (NSFC) under Grant 62293484. The work of Le Liang was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20220810 and in part by NSFC under Grant 62201145. The work of Shi Jin was supported in part by NSFC under Grant 62261160576, in part by the Key Technologies Research and Development Program of Jiangsu (Prospective and Key Technologies for Industry) under Grant BE2023022 and Grant BE2023022-1, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242023K5003. The work of Xiaoming Tao was supported by NSFC under Grant 61925105. *(Corresponding author: Le Liang.)*

Zhijin Qin, **Zijing Wang**, and **Xiaoming Tao** are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, with the State Key Laboratory of Space Network and Communications, Beijing 100084, China, and also with Beijing National Research Center for Information Science and Technology, Beijing 100084, China (e-mail: qinzhijin@tsinghua.edu.cn; wangzijing@tsinghua.edu.cn; taoxm@tsinghua.edu.cn).

Le Liang is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China, and also with the Purple Mountain Laboratories, Nanjing 211111, China (e-mail: lliang@seu.edu.cn).

Shi Jin is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: jinshi@seu.edu.cn).

Wen Tong is with the Wireless Technology Labs, Huawei Technologies, Ottawa, ON K2K 3J1, Canada (e-mail: tongwen@huawei.com).

Geoffrey Ye Li is with the School of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K. (e-mail: geoffrey.li@imperial.ac.uk).

Digital Object Identifier 10.1109/JPROC.2024.3437730

voice and text communications. In the 2000s, by utilizing code division multiple access (CDMA) and wideband code division multiple access (WCDMA), the third generation (3G) of cellular communications brought enhanced data transfer capabilities and provided higher data rates to support various services, including accessing the Internet and global positioning system (GPS). The development of orthogonal frequency-division multiplexing (OFDM) and multiple-input multiple-output (MIMO) realized the fourth generation (4G) of cellular communications in the 2010s. The 4G systems can support much faster data speeds and low transmission delay, enabling high-quality video transmission and mobile Internet access. The fifth generation (5G) of cellular communications has been widely deployed since 2020, which are designed to support ultrareliable and low-latency communications (URLLC), massive machine-type communications (mMTC), and enhanced mobile broadband (eMBB) communications. Due to the advance of massive MIMO, millimeter-wave (mmWave) communications, and terahertz (THz) communications, 5G can provide a high quality of experience (QoE) for various applications, such as the Internet of Things (IoT), autonomous driving, and extended reality (XR).

Wireless communication has been making a leap about every ten years. Currently, the scale of wireless networks and the number of wireless devices are enormous, and they are expected to continuously grow in the following years. The significant volume of data traffic poses a huge burden on wireless networks. To address the challenge and fulfill the evolving technological requirement, 5G beyond and the sixth generation (6G) of cellular communications have been studied [2], and many emerging techniques have also been proposed and investigated, such as reconfigurable intelligent surface (RIS) [3] and integrated sensing and communications (ISAC) [4]. Moreover, 6G is expected to go beyond the existing bit-level communication and reach the semantic level. The principle of the traditional Shannon paradigm is to guarantee the accurate reception of every single bit of the transmitted packet regardless of its meaning. Semantic communication in 6G is a new paradigm that focuses on the problem of how transmitted messages convey a desired meaning to the receiver, as well as how effectively the received meaning affects the action in a desired way [5]. By considering data semantics, 6G communication has the potential to make wireless networks significantly more efficient, robust, and sustainable. So far, 6G is still in the conceptualization stage and has not yet begun commercialization. It is important to come up with more intelligent approaches to enable the implementation of the next-generation wireless communication systems.

B. Motivation of Using Artificial Intelligence in Communications

Artificial intelligence (AI) and machine learning (ML) refer to the development of computer systems or software that can execute tasks that typically require human intelligence. These tasks include learning and reasoning, understanding natural language, problem-solving, and decision-making. AI/ML systems are designed to simulate or replicate human cognitive functions, allowing machines to perform complex tasks and adapt to varying environments. AI/ML describes a broad range of technologies and enables numerous services and products in our daily lives. For example, natural language processing (NLP) enables machines to understand, interpret, and generate human language, which plays a vital role in real-time language translation. Computer vision (CV) is developed to teach machines to interpret and make decisions based on visual data, which has been widely used in facial recognition and image analysis.

ML is capable of learning from a large amount of data, enabling computers to program automatically to perform a task and learn from examples to improve their performance over time. ML models can be generally divided into the following three types, supervised learning, unsupervised learning, and reinforcement learning (RL). Algorithms in supervised learning are trained based on labeled example datasets to build a mapping between input and output. For a given input, supervised learning can predict its corresponding output by using the mapping rule. Algorithms in unsupervised learning are trained based on unlabeled example datasets to discover inherent patterns or structures. Algorithms in RL are trained based on their own experience, allowing agents to learn to make decisions by interacting with an environment and receiving feedback which is noted as rewards or penalties. The neural network is a novel structure that involves layers of interconnected neurons (nodes) to model complex relationships and learn from diverse types of data. The neural network with multiple layers is called the deep neural network (DNN) which is enabled by deep learning (DL). The DNN-based architecture includes a variety of well-known models, such as the recurrent neural network (RNN), convolutional neural network (CNN), graph neural network (GNN), and long short-term memory (LSTM) network. DNN is versatile and can be deployed in various ML paradigms based on the specific task and the nature of the data.

Since the data traffic load for wireless communication will increase dramatically in the future, it is difficult for existing wireless communication systems to meet the ever-growing requirements of various intelligent wireless applications. AI/ML technology has been regarded as a powerful approach and brought many benefits in the design of the next-generation wireless communication systems. The first advantage falls in the signal processing at the physical layer. Traditional networks rely on accurate mathematical models for channel parameter estimation, which often fails to find an accurate model when the wireless propagation environment is time-varying and complex. ML is an alternative method to facilitate adaptive channel modeling and estimation by learning from the massive recorded data, relaxing the constraint for

accurate mathematical models. The second benefit is the superior data processing and interpretation capability. The explosion in the number of wireless devices challenges the traditional way of data storage and processing. ML algorithms can be used to identify patterns in raw data and remove redundant information to optimize the storage and processing spaces of data. With the help of NLP and DL, AI technology also allows us to rethink the traditional information theory and investigate post-Shannon wireless communications at the semantic level. Third, integrating AI/ML into wireless communications also benefits resource allocation and network optimization, ensuring the efficiency and scalability of wireless networks. Resource allocation problems can be generalized as optimization problems under some constraints, with the aim of striking a tradeoff between various factors (e.g., energy efficiency, transmission delay, and throughput). Traditional optimization tools can solve the problem efficiently when the objective function satisfies some assumptions, such as convex property, continuous property, and differential property. However, the complexity of the traditional optimization algorithms grows exponentially with the network scale, which is not scalable and acceptable for large-scale real-time applications. In addition, for 6G communication systems, the objective function can be more complex and the number of constraints can be larger; thus, traditional optimization tools may not work smoothly for the next-generation wireless communications. In this regard, AI/ML is seen as an effective tool for solving the challenging optimization and resource management problems in wireless communication systems.

The integration of AI in wireless communications offers numerous benefits across various practical applications [6]. With the help of AI, localization and positioning accuracy in wireless networks could be further improved [7], enabling applications such as indoor navigation and asset tracking. AI algorithms can optimize beamforming in wireless communication systems to dynamically adjust the antennas based on changeable environmental conditions and reduce interference. AI can also facilitate the allocation of various wireless resource [8] such as bandwidth, power, and frequency spectrum in wireless networks, ensuring efficient utilization of available resource and enhancing the network performance. Overall, AI in wireless communications promises improved performance and innovative applications.

Although the integration of AI/ML in wireless networks has witnessed considerable progress, the practical deployment of AI-empowered wireless communications encounters obstacles such as the difficulty in reasoning the meaning of signals, the limitation of computing resources, and the lack of guidance for the selection of proper ML algorithms based on specific tasks. For instance, AI models need substantial data for effective training to achieve near-optimal performance; thus, the scarcity of computing resources emerges as a prominent hurdle and hinders the network robustness. Efficient management of computing resources is as important as the wireless resource management. Moreover, selecting the most suitable ML algorithms based on specific tasks remains an implicit challenge for wireless communications. Compared with data-driven ML methods, classic model-driven methods are still effective in some cases. Therefore, the lack of explicit treatment for algorithmic selection and misuse of ML algorithms pose a challenge to achieving optimal performance tailored to the underlying transmission task.

C. Contributions and Organization

In Sections I-A and I-B, we introduced the development of existing wireless communications and the new vision of semantic communications. We also explained the motivation for integrating AI/ML into wireless communication networks. In light of the above advantages and limitations, there arises a need for a comprehensive survey of AI-empowered wireless communications.

Prior to this work, there were a few survey articles on AI and wireless communications shedding light on the past advancements and future challenges. Chen et al. [6] offered a detailed tutorial on utilizing a broad range of neural networks in DL, including RNN, DNN, and spiking neural network (SNN), to address various wireless networking problems related to unmanned autonomous vehicle (UAV) communications, edge computing and caching, IoT, multiple access, and the XR. Sun et al. [7] reviewed cutting-edge applications of ML across different layers of wireless networks, focusing on resource allocation, mobility management, and localization from the medium access control (MAC) layer to the application layer. Wang et al. [9] gave a comprehensive survey on the development of ML in the past 30 years, all the way from the physical layer to the application layer. Shi et al. [8] provided a systematic review of "learning to optimize" techniques in various areas of large-scale 6G networks, connecting ML algorithms with optimization theory to enhance the interpretability and transparency in the design of AI frameworks from an optimization perspective. These existing works have thoroughly investigated the field of AI-empowered wireless communications, offering perspectives on the design of AI-driven wireless networks and detailing potential use cases and scenarios. However, they mainly focus on how AI can make a significant impact on traditional wireless communications, ignoring the impact of AI on semantic communications. There are also a few survey articles on semantic communications. Gündüz et al. [10] gave a detailed survey to introduce semantic and task-oriented communications from the information-theoretic perspective. Both semantic information theory and ML techniques have been explored in this work. Qin et al. [11] provided an overview of the principles and challenges of semantic communications, explaining the performance gain by considering semantics. In that work, the term semantic communications is equivalent to task-oriented or goal-oriented communications.

Yang et al. [12] further divided semantic communications into three types, i.e., semantic-oriented communications, goal-oriented communications, and semantic awareness communications, and reviewed the related techniques and challenges. In this work, we focus on the transformation of wireless communications from bits to semantics and aim to provide an overview of the impact of AI on the shift of communication paradigms. Besides a fully encapsulated overview, we also summarize the lessons learned from the existing works and provide useful tips to guide the design of AI-empowered next-generation communication systems.

In this article, we give a comprehensive overview and tutorial on AI-empowered wireless communications from the classical bit level to the novel semantic level. Particularly, we provide a detailed overview of a series of well-known ML methods and applications from the physical layer up to the application layer which covers the fields of signal processing, resource allocation, and the state-of-art semantic communication paradigm. We explain the principles of various intelligent algorithms and models, discuss the challenges and open areas, and give useful guidance for researchers in the design of wireless communication systems with the aid of AI/ML. The key contributions of this article are listed as follows.

1) This article gives a detailed overview of the literature applying AI/ML to signal processing on the physical layer, covering the learning-based channel modeling and estimation, channel state information (CSI) feedback, precoding, and signal detection and decoding. The novel data-driven end-to-end communication framework is also discussed. Key insights are derived from comparative analysis, including learning-based CSI feedback versus codebook-based feedback and data-driven framework versus model-driven framework, offering valuable perspectives on the effectiveness of AI-based methods in contrast to classic approaches.
2) AI-empowered semantic communication is comprehensively investigated, encompassing semantic-aware sampling, coding, modulation, and other emerging techniques (e.g., big AI model). Moreover, some critical questions, e.g., whether semantic communication can break the Shannon limit, are explicitly treated and some insights regarding challenges and open areas are provided.
3) Popular ML techniques utilized in resource allocation and network optimization are thoroughly summarized including their basic principles and application scenarios, which are categorized by supervised learning, unsupervised learning, RL, and GNNs. Useful tips and tricks are provided to help researchers choose the most suitable ML algorithm given a specific task, facilitating the decision-making in wireless network optimization.

This article navigates the realm of AI-enabled techniques in both traditional and semantic wireless communications. The structure of this article is shown in Fig. 1. The rest of this article is organized as follows. Section II introduces the learning-based signal processing techniques on the physical layer, together with the comparative analysis of learning-based methods and their traditional counterparts. Section III introduces the semantic-aware intelligent data processing techniques, discussing key issues and open areas in the context of semantic communications. Section IV discusses the advancements in ML-aided resource allocation problems, providing useful tips and tricks on the proper selection and utilization of ML algorithms. Section V concludes this article.

II. LEARNING-BASED PHYSICAL LAYER PROCESSING

The use of AI to revolutionize physical layer processing has received increasing attention over the past few years, and significant performance gain has been observed in learning-enabled design for different communication modules as well as the new paradigm based on end-to-end learning. In this section, we provide an overview of these techniques, including learning-based channel modeling and estimation, CSI feedback, signal detection, and decoding. Finally, we investigate the novel end-to-end learning-based communication architecture.

A. Channel Modeling and Estimation

Conventional methods for channel modeling can be categorized into deterministic approaches, e.g., ray tracing, and stochastic approaches, which usually require extensive calculations and exact descriptions of the environment [13]. In contrast to these conventional approaches, generative models can adopted to capture the wireless channel effects and produce channel parameters due to their capacity to extract underlying properties from observed data. Furthermore, generative model-based channel modeling methods can deal with more complex communication environments and offer higher modeling accuracy [14].

The generative adversarial network (GAN), a typical generative model, has been utilized in [14] and [15] for channel modeling, and the framework is shown in Fig. 2. There are three parts: real channel samples $\mathbf{x}_s$, a channel data generator G, and a channel data discriminator D. Real channel samples comprise the dataset obtained from measurement, which are considered as the training data. The generator, G, tries to generate fake samples using the noise vectors z, and the discriminator, D, attempts to distinguish between real and fake samples. They are trained simultaneously using the following objective:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}_s \sim p_{\text{data}}} \left[\log D\left(\mathbf{x}_s\right)\right] - \mathbb{E}_{z \sim p(z)} \left[\log\left(D\left(G\left(z\right)\right)\right)\right]. \tag{1}$$

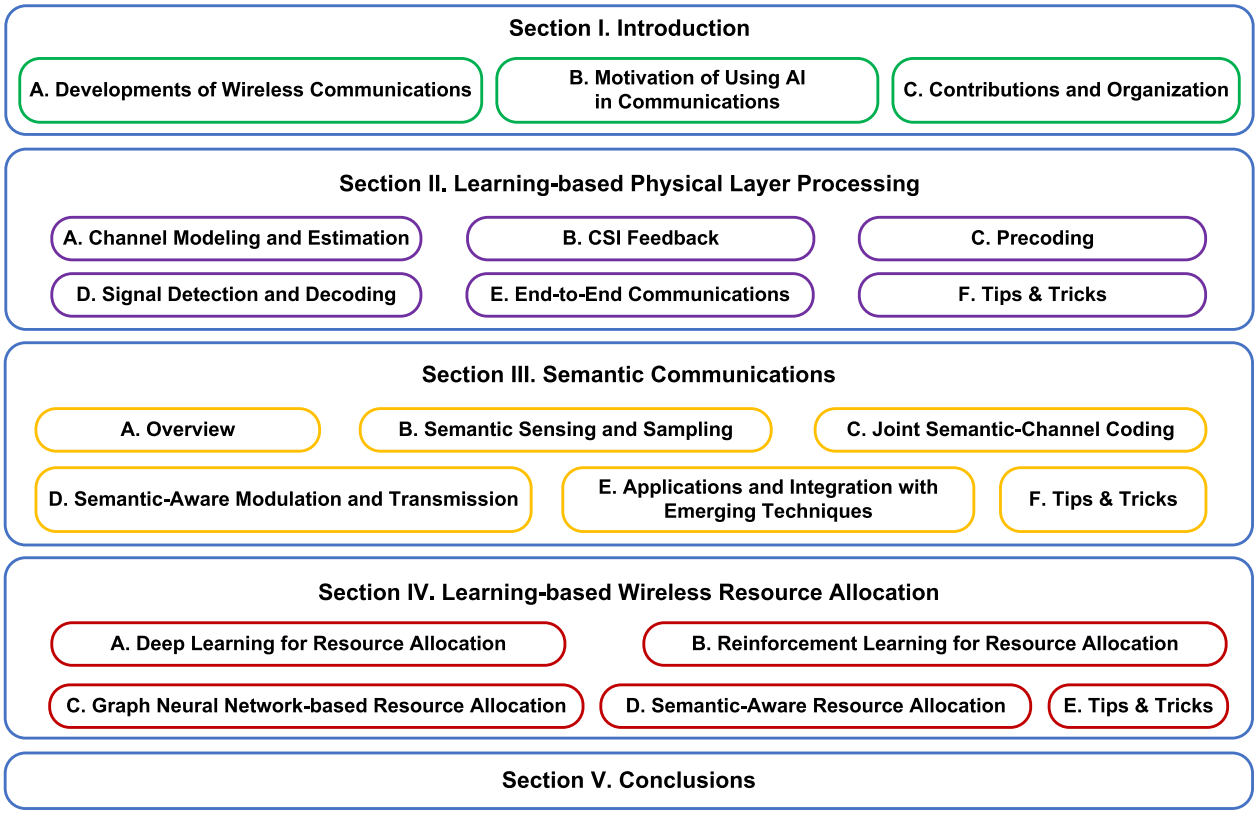

Fig. 1. ***Organization of this article.***

The training stops when the discriminator cannot distinguish the real and fake samples. At this point, the generator has learned the distribution of the real channel samples and it can be directly extracted as the target channel model. For instance, the GAN-based framework is trained with the real samples generated from a Gaussian distribution [14] and a Rayleigh distribution. At the beginning of the training, probability density functions (pdfs) of the generated and real data are quite different. After training, the pdfs of the generated and real samples are shown in Fig. 3, which demonstrates that the GAN-based method provides a commendable approximation of the actual additive white Gaussian noise (AWGN) channel for Rayleigh fading channel without requiring accurate environment information.

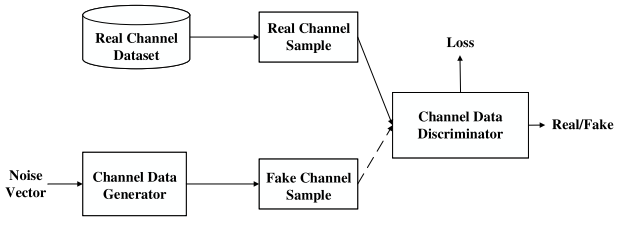

Fig. 2. ***Framework of GAN-based wireless channel modeling.***

This framework can be extended to tackle more complicated propagation channels. In a single-input single-output (SISO) scenario, the channel's 2-D time–frequency response is considered in [16]. The real channel samples are regarded as images, and the deep convolutional GAN is employed for the generator and discriminator. Furthermore, for a typical MIMO system with N_t transmit antennas and N_r receive antennas, the real channel samples in the time domain are represented as $\tilde{\mathbf{H}} \in \mathbb{C}^{N_t \times N_r \times N_d}$, where N_d denotes the number of delay paths. The MIMO channels contain more complex features than SISO channels, so the generator and discriminator should have wider and denser layers. The Wasserstein GAN with gradient penalty is used in [17] to improve training stability in modeling MIMO channels. In practice, the channel impulse response cannot be obtained easily for the MIMO channels. To deal with this issue, a conditional GAN is used in [18] to enable the model to learn the channels from the transmitted and received signal pairs. Utilizing the transmitted signal $\tilde{\mathbf{x}}$ as the conditioning information, the generator can be trained to produce the possible received signal $\tilde{\mathbf{y}}$. In addition to GANs, another efficient generative model called variational autoencoder (VAE) has also been utilized for channel modeling. For instance, it is used to model the channels in UAV communication [19].

Channel estimation is another critical aspect of wireless communication. Pilot-based channel estimation is

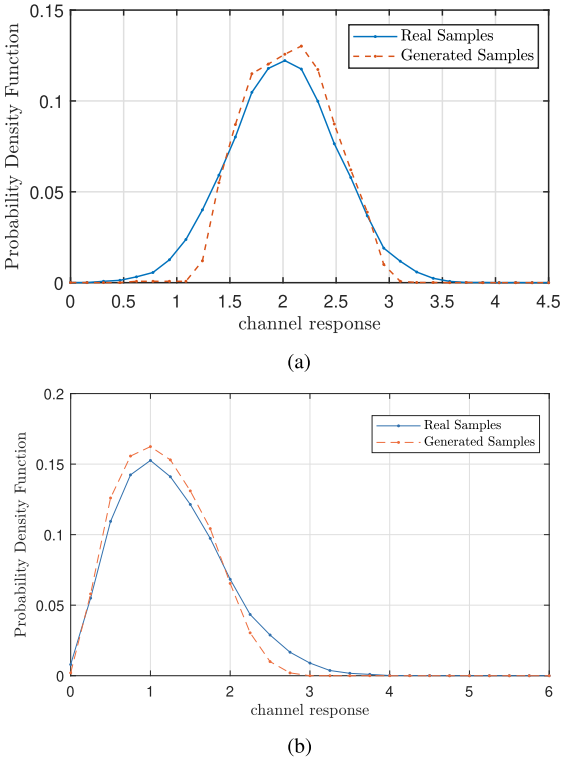

Fig. 3. ***After training with a sufficient number of real channel samples, the channel generator can produce samples with a similar pdf as the real samples. (a) Gaussian distribution. (b) Rayleigh distribution.***

widely used to estimate the unknown channel coefficients in an OFDM system by periodically transmitting known pilots [20], [21]. Least-squares (LS) and minimum mean-squared error (MMSE) methods are two traditional methods to recover the channel vector, $\tilde{\mathbf{h}}$, from the transmitted pilot, $\tilde{\mathbf{x}}_P$, and received signal, $\tilde{\mathbf{y}}_P$. Learning-based channel estimation has recently attracted increasing attention for its capacity to achieve higher estimation accuracy and lower computation complexity than traditional methods. These methods can be categorized into data-driven and model-driven approaches.

The core of data-driven estimation methods is to convert channel estimation to a supervised learning problem by treating $(\tilde{\mathbf{x}}_P, \tilde{\mathbf{y}}_P)$ as the input feature and the channel vector $\tilde{\mathbf{h}}$ as the training label. A DNN-based channel estimator learns a mapping from the input to the channel vector and is optimized by adjusting the parameters of the neural network to fit the dataset. The training stage is conducted offline, while in the online deployment stage, the current channel vector can be estimated immediately once the current transmitted and received pilot signals are obtained. In the simulation of a SISO system, the DNN-based method can achieve similar performance as the MMSE estimator while slightly outperforming the MMSE estimator in a low signal-to-noise ratio (SNR) regime. In MIMO systems, three types of channel correlations are introduced, namely, spatial, frequency, and temporal correlation, which are favorable for channel estimation. These correlations are captured and exploited efficiently in [22] by filtering pilot matrices of adjacent subcarriers and coherence intervals simultaneously using CNN-based channel estimators, improving the estimation accuracy and reducing the spatial pilot overhead caused by massive antennas. The learning-based estimation methods can achieve performance that is close to the ideal MMSE estimator with perfect (yet practically unavailable) correlation information.

Large-scale MIMO channels are known to exhibit sparsity due to limited scattering. Therefore, compressed sensing (CS) is exploited to enhance the learning-based channel estimation as it is an effective model-based approach to deal with sparsity. Typical CS methods, such as the iterative shrinkage thresholding algorithm (ISTA) and approximate message passing (AMP), hinge upon the fine-tuning of numerous hyperparameters. Instead of using a conventional componentwise shrinkage operator in AMP, a learned denoising approximate message passing (LDAMP) method is proposed in [23], which utilizes a denoising CNN (DnCNN) to harness channel correlations and reconstruct the channel more accurately. Furthermore, a learnable iterative shrinkage thresholding algorithm-based channel estimator (LISTA-CE) is designed in [24], which uses DL to obtain sparse transformation matrices and exploits the sparsity of wideband beamspace MIMO-OFDM channels, significantly outperforming conventional CS-based algorithms. The success of LDAMP and LISTA-CE demonstrates the feasibility of combining AI/ML and model-based channel estimation methods to further improve the estimation performance.

Insights and Challenges: Different from conventional methods that require sophisticated theory and high computation complexity, the channel modeling methods based on generative models do not need specific domain knowledge or technical expertise. Utilizing these generative models to approximate more complex and practical channels is an open area for further research. ML-based channel estimation methods can reduce the computation complexity and capture the channel correlations to improve the estimation accuracy. Moreover, exploiting the prior knowledge of the channels, such as sparsity, and designing more efficient and low-complexity estimation approaches deserve further exploration.

B. CSI Feedback

CSI at the transmitter is important for adaptive transmission in time-varying channels and precoding design in MIMO communication systems. When operating in the frequency-division duplex (FDD) mode, reciprocity of uplink and downlink channels faces challenges and the user equipment (UE) needs to report downlink CSI to the base station (BS) for downlink precoding, causing substantial feedback overhead due to the extensive CSI dimensions in massive MIMO systems. Traditional feedback methods,

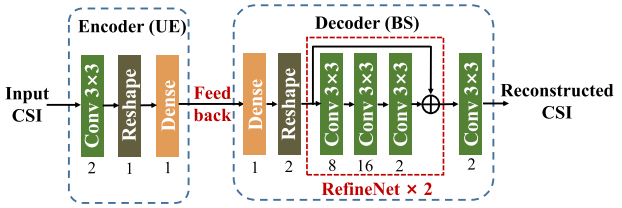

Fig. 4. *Illustration of the CsiNet architecture, where the autoencoder compresses and reconstructs CSI at the transmitter and receiver, respectively.*

including codebook and CS-based approaches, struggle to balance between feedback overhead and accuracy.

To address this issue, CsiNet [25] first introduces AI in CSI feedback and constructs an autoencoder for compressing and decompressing CSI, taking advantage of the CSI sparsity in the angular-delay domain. As shown in Fig. 4, the UE-side encoder extracts CSI features with a convolutional layer. On the BS side, the decoder reconstructs the CSI from codewords through RefineNet, featuring three convolutional layers with residual learning. CsiNet drastically enhances feedback accuracy compared to traditional methods. Subsequent research has extensively explored learning-based CSI feedback, which can be categorized into three main research directions: novel neural network architecture design, multidomain correlation utilization, and practical deployment.

1) Novel Neural Network Architecture Design: Although learning-based algorithms can automatically capture CSI features from extensive training samples, the performance heavily relies on the neural network architecture and warrants thoughtful design. Preliminary methods directly concatenate real and imaginary CSI components with a convolutional layer (e.g., 3×3 kernel), causing performance degradation. In contrast, CLNet [26] maintains the inherent complexity of CSI by replacing it with a 1×1 kernel, thus preserving phase information. Simulation results show significant reconstruction gains over preliminary methods. Another approach, TransNet [27], focuses on prioritizing feature maps that contain richer information. It achieves this by utilizing the attention mechanism in CSI feedback. By replacing feature extraction and reconstruction in the autoencoder with the Transformer, TransNet significantly improves feedback performance. To further streamline the process of designing neural networks, Auto-CsiNet [28] employs the neural architecture search for automating this process for CSI feedback in specific scenarios. By fully exploiting the potential of AI, simulation results showcase Auto-CsiNet's superiority over manually designed models, excelling in both accuracy and complexity.

2) Multidomain Correlation Utilization: Considering the rich wireless propagation information contained within CSI, exploiting multidomain correlation can notably enhance the performance of learning-based CSI feedback, including time correlation, bidirectional channel correlation, and channel correlation among UEs. As an example, CsiNet-LSTM [29] capitalizes on high time correlation across adjacent slots by adopting LSTM modules. It compresses the initial and subsequent CSI images with different compression ratios (CRs). Then, the concatenated codewords are refined by LSTM modules at the decoder. Simulation results demonstrate the superior performance of the proposed method, particularly at lower CRs. Moreover, bidirectional channel magnitude correlation is leveraged by DualNet-MAG [30] due to the shared propagation environment between downlink and uplink. After the encoder compresses the downlink CSI magnitude, the decoder reconstructs it with both the codeword and the uplink CSI magnitude. Simulation results showcase significant accuracy improvement with integrated uplink CSI. Additionally, to cope with the rise in UE density, distributed DeepCMC has been developed in [31] to leverage the CSI correlation among nearby UEs for improving CSI feedback. The BS merges codewords from multiple UEs and reconstructs them through a collective feature decoder. Information shared among nearby UEs no longer requires feedback, thereby reducing overhead.

3) Practical Deployment: To boost the practical deployment of learning-based CSI feedback, many problems have been further considered in the literature. First, CAnet-J [32] integrates pilot design, channel estimation, and CSI feedback. After using uplink CSI for pilot generation and compressing received signals directly, the decoder reconstructs downlink CSI utilizing uplink CSI. This approach reduces pilot overhead and estimation errors by omitting shared bidirectional information in pilot signals. In addition, to mitigate quantization errors, a μ-law nonuniform quantizer is utilized in [33], which adjusts the quantization step sizes according to codeword amplitudes. The encoder's output undergoes μ-law quantization, and then, the BS refines it with an offset neural network before the decoder. Simulation results demonstrate the superior performance of the proposed method compared to the uniform quantization method. Furthermore, in order to automate the choice of CRs in practical systems, AdaL-iCsiNet [34] incorporates a classification module before CSI feedback. Based on a predefined accuracy threshold, this classification model selects an appropriate CR by a neural network, utilizing CSI as the input to predict the desired CR as the output. Additionally, in alignment with existing cellular systems based on implicit feedback, ImCsiNet [35] refines the feedback process by transmitting the precoding matrix rather than the entire CSI matrix. Simulation results indicate that this method mitigates feedback overhead compared to conventional codebook-based methods. Moreover, to address the challenge of resource-intensive high-complexity neural networks, techniques like neural network weight pruning, quantization, and efficient architecture design have been introduced in [36] for learning-based CSI feedback. Simulation results

demonstrate minimal accuracy loss despite the reduced neural network complexity. More details and other related works can be found in [37].

Insights and Challenges: Despite the considerable attention AI-based CSI feedback has received from both academia and industry, it still faces several challenges. A significant challenge arises from the two-sided architecture adopted in AI-based CSI feedback. The intervendor training collaboration of the encoder at the UE side and the decoder at the BS side should be carefully designed. Moreover, the tradeoff between feedback performance and model complexity needs to be investigated to achieve an optimal balance for practical deployment.

C. Precoding

Precoding, colloquially referred to as beamforming, denotes a deliberate and strategic technique employed to purposefully shape transmitted signals. The principal goal of precoding is to optimize the efficiency and reliability of data transmission across communication channels. This shaping process is intricately designed to exploit the inherent spatial characteristics of the communication channel, thereby leading to an enhanced overall performance.

Consider a downlink transmission within a multiuser MIMO system, where a transmitter simultaneously serves K receivers. The transmitter, equipped with N_t antennas, sends N_s data streams to the kth receiver. This transmission is accomplished by using a precoding matrix denoted as $\mathbf{F}_k \in \mathbb{C}^{N_t \times N_s}$. Subsequently, the kth receiver, possessing N_r antennas, employs a combining matrix denoted as $\mathbf{W}_k \in \mathbb{C}^{N_r \times N_s}$ to process the received signal. The achievable rate at the kth receiver is expressed as

$$\mathcal{R}_k = \log\det\left(\mathbf{I} + \mathbf{F}_k^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\mathbf{W}_k\mathbf{R}^{-1}\mathbf{W}_k^{\mathrm{H}}\mathbf{H}_k\mathbf{F}_k\right) \tag{2}$$

where $\mathbf{H}_k \in \mathbb{C}^{N_r \times N_t}$ signifies the channel matrix between the transmitter and the kth receiver. Additionally, $\mathbf{R} = \sum_{m\neq k, m=1}^{K} \mathbf{W}_k^{\mathrm{H}}\mathbf{H}_k\mathbf{F}_m\mathbf{F}_m^{\mathrm{H}}\mathbf{H}_k^{\mathrm{H}}\mathbf{W}_k + \sigma_k^2\mathbf{W}_k^{\mathrm{H}}\mathbf{W}_k$ represents the covariance matrix encompassing the contributions of noise and interference. In the context of precoding design, various objectives may be of interest, such as maximizing the weighted sum rate (WSR), i.e., $\sum_{k=1}^{K}\alpha_k\mathcal{R}_k$, maximizing the minimum achievable rate, i.e., $\min\{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_K\}$, and others.

Preliminary investigations have leveraged data-driven DL methods for the design of precoding and combining matrices [38], [39], [40], [41]. In the context of a single-user multiple-input single-output system, Lin and Zhu [38] introduced a phase-shifter-based analog precoding architecture at the transmitter. Then, a multilayer neural network was constructed to acquire the mapping from the estimated channel to the analog precoding vector. To ensure the output conforms to the unit modulus constraint, the neural network initially generated the phase of the analog precoding vector, which was subsequently utilized to construct the analog precoding vector. However, this design overlooks the potential structural characteristics of the precoding matrix, introducing challenges in network training. In the context of a multiuser multiple-input single-output system, Xia et al. [39] investigated the structural characteristics of optimal precoding vectors and proposed a DL framework. This framework addressed three classical optimization problems in precoding design, i.e., the signal-to-interference-plus-noise ratio (SINR) balancing, the power minimization, and the WSR maximization problem. Each problem was tackled with a meticulously designed neural network. Notwithstanding the efficacy of these approaches, it is important to note that these works predominantly rely on data-driven methods, which are susceptible to issues related to resilience and interpretability.

Inspired by traditional precoding algorithms [42], [43], the utilization of deep unfolding emerges as a promising avenue for enhancing the robustness and interpretability of DL-based methods. By integrating expert knowledge into the design process, such methods demonstrate notable performance and efficiency [44], [45], [46], [47]. In the context of the widely employed hybrid architecture in mmWave systems, the study presented in [44] introduces a hybrid precoding approach based on model-driven DL principles. The optimization objective of WSR maximization is initially reformulated into a weighted minimum mean-squared error (WMMSE) optimization, enabling the direct derivation of digital precoding and combining matrices. For analog precoding and combining matrices, subject to unit modulus constraints, manifold optimization (MO) techniques are employed. To further bolster performance and mitigate complexity, the step sizes of MO are treated as trainable variables. In recognition of the varying channel sparsity and its impact on robustness, an adaptive neural network is introduced to dynamically generate the step sizes of MO. The proposed method exhibits commendable performance and robustness against dynamic system parameter variations. The advent of RIS has introduced a novel dimension to wireless communication design. The phase shifts of RIS elements are considered as passive beamforming, which are jointly designed with active beamforming at the transmitter. In [45], a combination of WMMSE optimization and the power iteration algorithm is proposed. The inclusion of trainable variables expedites the convergence of the proposed algorithm, and a data-driven GNN is incorporated to enhance initialization performance and reduce the required number of iterations. By amalgamating model-driven and data-driven learning methods, the proposed approach attains the advantages of low complexity and strong robustness.

In addition to data-driven and model-driven approaches, recent studies have explored the application of deep RL for precoding design [48], [49], [50]. Treating the communication system as an agent, an RL-based algorithm is developed for hybrid precoding design in [48]. Departing from the conventional strategy of directly learning analog and digital precoding matrices simultaneously, the proposed method employs an

MO-based approach to initially compute a feasible analog precoding matrix. Subsequently, deep RL is employed to determine the digital precoding matrix and analog combining matrix. Finally, the digital combining matrix is computed using the MMSE criterion. Jensen's inequality provides an upper bound of the average WSR with imperfect CSI, serving as the reward for training the agent. It is worth noting that RL-based algorithms often face challenges related to convergence, particularly in scenarios characterized by large state and action dimensions. In the context of an RIS-assisted system, the dimension of the passive beamforming is typically substantial due to the numerous reflective elements. To mitigate convergence difficulties, Wang and Zhang [49] strategically utilize discrete Fourier transform (DFT) vectors as the action set. The selected action involves switching the adopted vector to an adjacent one, effectively reducing the dimensionality of the action space.

In MIMO systems employing a hybrid architecture or RIS-assisted configurations, channel estimation often presents a formidable challenge. Traditional methods, involving both channel estimation and precoding design, encounter substantial difficulties. Several approaches have sought to circumvent the complexities of channel estimation by directly designing beamforming based on received pilot signals [40], [51]. Notably, due to the permutation equivariant and invariant properties, GNNs are employed to model the RIS-assisted wireless communications in [51]. In these methodologies, the received pilot signal of the BS is directly input into the GNN. The GNN then maps this input into the active and passive beamforming vectors, effectively bypassing the conventional channel estimation process.

Insights and Challenges: With the advancement of communication systems, there has been a notable increase in the number of antennas, leading to an expansion in the dimensionality of the precoding matrix. Consequently, data-driven approaches require larger network dimensions, complicating the training process. On the other hand, model-driven methodologies necessitate iterative processes, leading to high complexity. Hence, achieving an optimal tradeoff between precoding performance and computational complexity is of paramount importance. Moreover, as channel dimensions escalate, the channels frequently manifest sparsity and other distinctive characteristics. Investigating strategies to harness these characteristics for design optimization remains an area meriting scholarly attention.

D. Signal Detection and Decoding

Signal detection and channel decoding have long been acknowledged as fundamental tasks in communication receiver design. In this section, we feature several notable instances where AI/ML exhibits significant potential in designing signal detection and channel decoding schemes.

For signal detection, the pioneering work of Ye et al. [52] considers using data-driven learning to improve signal detection in OFDM systems. Specifically, a fully connected deep neural network (FC-DNN) is designed to perform signal detection without the need for explicit channel estimation. This data-driven method replaces the module-by-module design in conventional OFDM receivers by learning a mapping from the received signal to the transmitted data. This approach leads to notable enhancements in handling channel distortions, particularly in complex scenarios, by eliminating the need for analytical modeling. A similar idea is also utilized in molecular communications [53], where a bidirectional RNN is developed for sequence detection. This proposed method does not rely on knowledge of a mathematically complex transmission model and outperforms traditional Viterbi detection in specific scenarios.

Extending beyond the signal detection problem in SISO systems [52], [53], current wireless communication systems generally deploy multiple antennas at both the transmitter and receiver ends, leading to the more challenging MIMO detection problem. Consider a MIMO system that consists of N_t antennas for transmitting data streams independently and N_r antennas for receiving the signal. Let $\mathbf{x} \in \mathcal{A}^{N_t \times 1}$ denote the transmitted data vector, whose elements are drawn from the discrete alphabet $\mathcal{A}$, e.g., quadrature amplitude modulation (QAM) constellation. The relationship between the input and output of this system can be represented by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{3}$$

where $\mathbf{y} \in \mathbb{C}^{N_r \times 1}$ denotes the received signal vector, $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ denotes the channel matrix, and $\mathbf{n} \in \mathbb{C}^{N_r \times 1}$ denotes the Gaussian noise vector. The goal of MIMO detection is to estimate the symbol vector $\mathbf{x}$ based on the observation $\mathbf{y}$ and the knowledge of $\mathbf{H}$. Assuming no a priori information is available, the optimal maximum likelihood detection criterion is given by

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathcal{A}^{N_t \times 1}}{\arg\min} \ \|\mathbf{y} - \mathbf{H}\,\mathbf{x}\|^2. \tag{4}$$

Due to the discrete nature of $\mathbf{x}$, optimal MIMO detection belongs to integer LS problems, known to be NP-hard.

To address the challenge, a DL-based MIMO detector, named DetNet, has been developed in [54] by unfolding the projected gradient descent method and incorporating learnable weights. Under well-conditioned channels, DetNet achieves comparable performance as the AMP [55] and semidefinite relaxation [56] while providing lower complexity. Notably, this study reveals the idea of leveraging model-driven learning [57], which integrates communication domain knowledge (e.g., iterative algorithms) into data-driven pipelines, for the design of MIMO detectors.

To further reduce the training cost and enhance robustness, the orthogonal AMP algorithm is unfolded [58], with only four trainable parameters integrated into each

layer of the unfolded network. The developed model-driven learning-based detector, named OAMPNet, exhibits significant performance gain over DetNet under realistic small-sized MIMO channels or high-order modulation. An extension has been developed in [59] that addresses signal detection in cyclic prefix-free MIMO-OFDM systems, and the model-driven learning-based detector outperforms benchmark algorithms in both numerical simulations and over-the-air experiments. In addition, GNNs have been leveraged to enhance message-passing-based detectors in [60]. Specifically, a GNN-aided expectation propagation (EP) algorithm, referred to as GEPNet, is developed and shows significant gains over the EP detector [61]. Moreover, applying online learning to enhance the environmental adaptability of learning-based MIMO detectors has been studied in [62]. The proposed approach, named MMNet, exploits temporal and spectral correlation of realistic channels to accelerate online learning and demonstrates better performance in handling harsh channel conditions than baselines.

The incorporation of hypernetworks [63] offers an alternative avenue for enhancing the generalization capabilities of model-driven learning-based detectors. Essentially, this approach entails the utilization of an auxiliary network to generate parameters for the primary network, mirroring the relationship akin to that between a genotype (the hypernetwork) and a phenotype (the primary network) [64]. Unlike conventional online training paradigms, hypernetworks facilitate adaptive parameter adjustments in response to evolving scenarios, thereby circumventing retraining. The studies in [65] introduced a hypernetwork-aided MMNet to tailor the MIMO detector to diverse channel realizations and noise levels. Specifically, the hypernetwork can learn the fluctuating trends of the adjustable weights across various scenarios and furnish suitable parameters for the signal detection network. This idea has been further explored in [64] that meticulously designs the inputs of the hypernetwork to remove redundant features and achieves reduced complexity while preserving excellent adaptability.

In another aspect, stochastic sampling-based methods, particularly Markov chain Monte Carlo (MCMC), have shown the potential to approach near-optimal MIMO detection performance with high efficiency. Recently, MCMC has been combined with gradient descent, showing impressive promises in nonconvex optimization problems [66]. The combination of these two powerful ML techniques for MIMO detection has been first investigated in [67], which accelerates MCMC's exploration of the search space utilizing Newton's method [68]. Further enhancements are explored in [69] as they accelerate MCMC sampling by first-order optimal Nesterov's accelerated gradient (NAG) method [70], which avoids the high-complexity matrix inversion required by Newton's method and results in a highly scalable detector. As shown in Fig. 5, the detector proposed in [69], namely, NAG-MCMC, achieves substantial gains over various NN-based detectors and approaches the performance of the optimal maximum likelihood detector.

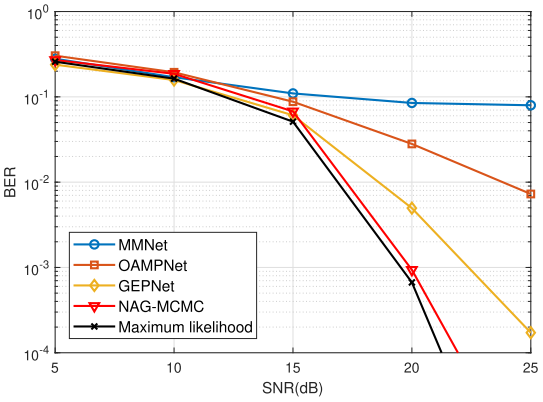

Fig. 5. *BER performance comparison of NAG-MCMC and other baselines in an 8 × 8 MIMO system with 16-QAM and Rayleigh fading channels. The BER curve of DetNet is not presented due to its subpar performance under this high-order modulation.*

Regarding the channel decoding problem, neural networks have long been utilized as alternative approaches to conventional iterative methods, as channel decoding can be framed as a classification problem. In recent years, the remarkable achievements of DL in various domains have sparked a new trend of developing DL-based channel decoders. A data-driven decoding method utilizing an optimized DNN has been proposed in [71] for random and polar codes. It is observed that the DNN decoder can achieve optimal decoding performance for both code families when the code length is short. In particular, the experimental results show great promise in learning robust decoding schemes for polar codes, which are structured as compared to random codes. Notwithstanding, the scalability of neural network-based decoding to long codewords is a fundamental issue to be resolved. To address the curse of dimensionality, the structure of the encoding process and the iterative belief propagation (BP) decoding algorithm are employed as the prior knowledge for constructing neural decoders in [72]. The proposed method assigns weights to the edges of the Tanner graph in standard BP and optimizes these weights via DL techniques, offering improved performance and reduced complexity. Furthermore, RNN-based channel decoding has been investigated in [73] and [74]. Specifically, data-driven RNNs are learned in [73] for decoding convolutional and turbo codes, showcasing enhanced performance and excellent robustness as compared to conventional Viterbi [75] and BCJR [76] algorithms. In [74], the DNN-based BP decoder [72] is improved by using the RNN architecture, which reduces the number of trainable parameters by unifying weights across different iterations/layers without significantly sacrificing decoding performance.

In addition to learning the channel decoder, optimizing the channel encoder has also been investigated in NN-based channel encoding and decoding research. Distinct

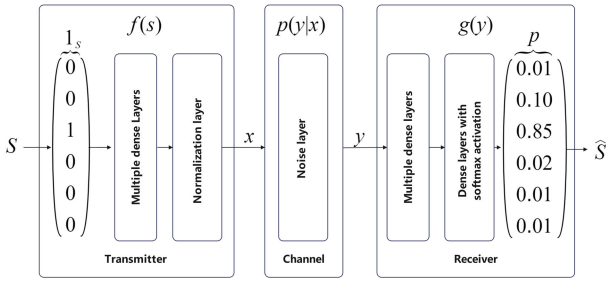

Fig. 6. ***Communications system over an AWGN channel represented as an autoencoder, adapted from [81].***

from end-to-end learning approaches that jointly optimize the encoder and decoder and can be obstructed by a missing channel model [77], it was proposed in [78] to apply a neural estimator [79] to approximate the mutual information between the channel input and output and learn the channel encoder by maximizing the mutual information. This approach decouples the training of the channel encoder with the decoder, thereby circumventing the issue of an unknown channel model.

Insights and Challenges: ML-based signal detection and decoding, despite their potential to surpass traditional methods, still face several challenges that need to be resolved. One notable issue is how to systematically tackle the curse of dimensionality. Developing low-dimensional AI/ML models for processing high-dimensional data is an important topic [80] and has become highly valuable as the problem dimension reaches an unprecedented scale with the evolution of communication systems. Additionally, enhancing the generalization capability of the models remains an open problem that is crucial for practical implementation.

E. End-to-End Communications

In recent years, end-to-end communication systems have attracted considerable attention, which transforms the traditional communication system into a data-driven framework [77], [81]. In this novel end-to-end paradigm, transmitters and receivers are jointly trained based on an end-to-end loss function rather than separate channel coding/decoding and modulation/demodulation. The end-to-end communication architecture can be viewed as an autoencoder, where the transmitter functions as the encoder network, and the receiver serves as the decoder network.

As a pioneering work, a learning-based system over an AWGN channel with a short block size is investigated in [81], as illustrated in Fig. 6. The transmitter uses an encoder network to map the transmitted symbols into $\mathbf{x}$, which is subsequently sent through the AWGN channel. The transfer function of the AWGN channel can be represented as $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ represents the noise. At the receiver side, a decoder network is exploited to recover the transmitted symbols from the received signal $\mathbf{y}$. In this way, the traditional modules at the transmitter and the receiver are replaced with neural networks. Note that since the transfer function of the AWGN channel is differentiable, direct backpropagation can be employed for updating the whole network.

Assuming that the wireless channel model is differentiable, the end-to-end communication system was extended to more general channels, including the Rayleigh fading channel, the multipath fading channel, and the MIMO channel. In [82], an end-to-end communication system for Rayleigh fading channels was constructed based on the autoencoder. Mathematically, the channel coefficient of Rayleigh fading can be represented as a random variable following complex Gaussian distribution, i.e., $h \sim \mathcal{CN}(0, 1)$, and the received signal $\mathbf{y}$ is given by $\mathbf{y} = h\mathbf{x} + \mathbf{n}$. Hierarchical 1-D convolutional layers were then used in [82] at the transmitter and receiver to overcome the distortion caused by Rayleigh fading and noise.

Moreover, the end-to-end communication system was also extended to transmission over more challenging multipath channels [83], [84], [85], [86]. To deal with frequency-selective fading in multipath channels, the end-to-end communication system was combined with the OFDM technique [83], which divides the wideband channel into orthogonal narrowband subcarriers, each exposed to flat fading rather than frequency-selective fading. Moreover, the end-to-end communication system over MIMO channels in [87] demonstrated enhanced performance compared to previous methods. Subsequently, the research in [88] incorporated finite quantization of the CSI, showcasing additional performance improvements. Although O'Shea et al. [87], [88] demonstrated the potential of autoencoder-based MIMO communication systems, the training of these systems was based on the assumption that the CSI is available at the receiver, which can be viewed as the coherent transceiver design. In addition to these studies, [89], [90], and [91] delved into MIMO end-to-end communication in the context of noncoherent transceiver design, where neither the transmitter nor the receiver has access to the CSI. Specifically, in [91], a pilot-free end-to-end paradigm for flat-fading MIMO channels was developed, where the wireless channels are modeled as a stochastic convolutional layer. The end-to-end communication system over the MIMO channel also follows the architecture of the autoencoder, where the transmitter DNN and the receiver DNN correspond to the encoder and the decoder, respectively. Rather than relying on pilots for channel estimation, the receiver utilizes two DNN modules to extract channel information and recover data. A bilinear production operation [92] is used to combine the features extracted from the channel and the received signals, which are further utilized to recover the transmitted data. This approach benefits from the efficient utilization of channel information, thus enhancing the overall process of data reconstruction. Meanwhile, the multiuser interference channel was considered in [93], where the end-to-end multiuser communication exhibits substantial performance

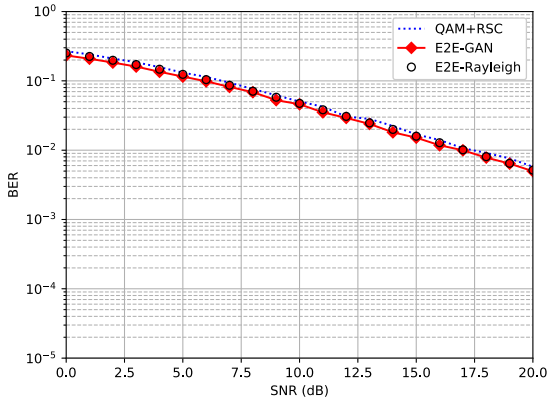

Fig. 7. ***BER under a Rayleigh fading channel [94].***

improvements when compared with the conventional time-sharing baseline scheme.

While assuming known and differentiable channel models has allowed end-to-end systems to demonstrate competitive performance compared to traditional communication systems, certain real-world intricacies cannot be accurately represented in a differentiable manner. The simplified channel models, when applied, may occasionally mislead the trained system due to the mismatch between these models and realistic physical channels. Hence, there is an expectation to train end-to-end communication systems without relying on a specific channel model. However, if the channel transfer function, $\mathbf{y} = f_h(\mathbf{x})$, is unavailable, the backpropagation of gradients through the receiver and transmitter DNNs is obstructed, thus hindering the learning of the end-to-end system.

To tackle this challenge, there are primarily two approaches. The first involves developing a channel-agnostic end-to-end communication system, as explored in [94]. In this approach, the distribution of channel outputs is learned through a conditional GAN [95], eliminating the need for explicit knowledge of the channel transfer function. In this way, a conditional GAN is used to act as a bridge for the gradients to pass through. Fig. 7 shows the comparison of the end-to-end approach with a conventional method that employs QAM and a rate-1/2 recursive systematic convolutional code under a Rayleigh fading channel. The performance of the end-to-end approach utilizing a conditional GAN ("E2E-GAN") is comparable to that of the traditional method ("QAM + RSC") and the end-to-end approach that incorporates a differentiable Rayleigh fading channel ("E2E-Rayleigh") in terms of bit error rate (BER). This consistency proves that the trained conditional GAN can be used as a surrogate for the original Rayleigh channel. The second approach employs an RL-based framework, as investigated in [96], to optimize the end-to-end communication system without relying on the channel transfer function. In this framework, the transmitter is treated as an agent, with both the channel and the receiver regarded as parts of the environment. At each time step, the source bits serve as the state observed by the transmitter, and the transmitted signals represent the actions taken by the transmitter. The end-to-end loss for each sample is calculated by the receiver and subsequently provided as feedback to the transmitter, acting as a reward from the environment. This feedback guides the training of the transmitter. During the training process, obtaining gradients for the receiver is straightforward on the receiver side, while the gradients for the transmitter are approximated using RL.

The autoencoder-based communication architecture was also extended to relay systems, where [97], [98], and [99] have delved into autoencoder-based amplify-and-forward (AF) relay systems, while [100] and [101] focused on autoencoder-based decode-and-forward (DF) relays. In the AF approach, the relay node amplifies and retransmits the received signals in the analog domain, offering low computational complexity and notable performance improvements [99], [102]. In contrast, the DF approach requires more complex neural network-based processing at the relay node to decode and re-encode the signal before retransmission, as detailed in [100] and [101]. Given the DF approach is more complex in various contexts, researchers often prioritize AF relay networks. An autoencoder-based AF relay communication system was introduced in [97], with an emphasis on modulation design. On the other hand, Gupta and Sellathurai [98] advanced the field by developing an autoencoder-based coded modulation design with CSI and a differential coded modulation design without CSI for AF relay systems. Further, Gupta and Sellathurai [99] implemented an autoencoder-based coded modulation design and demonstrated that utilizing a traditional AF relay node within autoencoder-based AF relay networks can reduce implementation complexity while also enhancing BER performance when compared to conventional approaches.

Insights and Challenges: In the end-to-end communication paradigm, the transmitter and receivers are trained in a novel end-to-end manner, which has the potential to surpass the traditional methods with separate channel coding and modulation. Despite that the end-to-end communication shows superiority in simple wireless channels, such as AWGN and Rayleigh fading channels, extension to more complex real-world wireless channels still faces challenges. First of all, enhancing the generalization and scalability of the system in various wireless channels is an open problem. Moreover, how to train the end-to-end communication system with the gradient back-propagated over the air, which is important for online training in practice, is another critical challenge.

F. Tips and Tricks

1) How CSI Feedback Works Conventionally and How Does It Compare to the Learning-Based Method?: In LTE and 5G cellular systems, the codebook-based feedback method (e.g., random vector quantization, RVQ) is the de facto

standard, which requires BS and UE to share and maintain the same codebook. The codebook index is selected by minimizing the distance between the channel matrix and codewords and fed back to the BS. However, high-dimensional CSI significantly increases the difficulty and cost of codebook design, and naive codebook schemes like RVQ cannot guarantee accurate reconstruction. In contrast, learning-based CSI feedback schemes leverage DNNs to conduct offline training using massive CSI data. These DL models efficiently leverage the data fitting capabilities of DNNs to learn and extract underlying features in the channel. This enables fast and accurate reconstruction of CSI during online deployment, thus holding significant potential to surpass traditional methods in terms of reconstruction performance and computational speed [25].

2) Data-Driven Versus Model-Driven Learning Methods for MIMO Detection, Which One Is Better in Practice?: There is no definite answer to this question. However, in practice, we have found that model-driven learning methods are generally more effective than purely data-driven methods in the design of MIMO detectors. This superiority can be attributed to the combinatorial nature of the problem, which plays a crucial role in designing accurate and efficient detection algorithms. Therefore, domain knowledge from established model-based detection methods should be fully absorbed and inherited into the development of learning-based MIMO detectors. Furthermore, studies such as the seminal work [54] have shown that relying solely on a purely data-driven, model-agnostic "black box" approach falls short of achieving the desired performance in MIMO detection. Hence, it is evident that the integration of domain knowledge and data-driven techniques is crucial for MIMO detection. Notably, this perspective on leveraging model-driven learning approaches also applies to various other signal processing tasks within the physical layer [57].

3) Are Pilots Necessary for End-to-End Communication?: Traditional wireless systems heavily depend on pilots for precise channel information, but in end-to-end communication with neural representations at both ends, the need for pilots diminishes. Research explores both pilot-based [94] and pilot-free approaches [91]. Explicit pilots and channel estimation ease the learning process by providing additional channel information. Nevertheless, the design of pilots, including their structure and quantity, demands careful consideration and expert knowledge. The pilot-free designs, on the other hand, enhance resource efficiency by eliminating the need of using dedicated time and frequency resources for pilots. However, the absence of explicit pilot signals requires the end-to-end communication system to autonomously learn and adapt to evolving channel conditions. This introduces challenges, as the encoded data must be reliably recovered with an unknown channel.

4) What Are the Impacts of AI/ML-Based Design on 3GPP Standards?: Integration of AI/ML into existing air interfaces has aroused heated discussions within the 3rd-generation partnership project (3GPP). A notable milestone of this development is the agreement of starting a new study item on AI-native air interface in 3GPP Release 18 [103]. The study item aims to develop a comprehensive framework for using AI/ML to enhance air interface and focuses on specific use cases including CSI feedback, positioning, and beam management. A wide range of challenges and novel perspectives have been identified in this study, such as the collection of datasets for model training, the life cycle management of AI/ML models, and the collaboration between users and BSs for AI/ML operations. Addressing these challenges is an essential step in 3GPP's standards development as it moves toward integrating AI and communications.

III. SEMANTIC COMMUNICATIONS

Semantic communication is a novel wireless communication paradigm that focuses on transmitting task-related data rather than the raw data. Semantic communication involves the exchange of information where the meaning and context of data are necessary for accurate and effective interpretation, enabling the network to understand and process data in a more intelligent and task-oriented manner. The rapid development of AI has brought both opportunities and challenges to semantic communications. In this section, we first briefly introduce semantic communications as well as the differences with traditional communications. Then, we go through the AI-enabled techniques, including semantic-aware sensing, coding, modulation, and other key techniques.

A. Overview of Semantic Communications

The notion of semantic communications is proposed by Shannon and Weaver [104], in which they categorize communications into three levels. The first level is the technical level, focusing on the accurate transmission of bits. The second level is the semantic level, focusing on the precise transmission of the meaning of symbols. The third level is the effectiveness level, focusing on the effectiveness of the semantic meaning that affects conduct in the desired way.

Traditional wireless communication falls into the first level. Inspired by the classical Shannon information theory, traditional communication systems are designed with bit-based metrics to evaluate the network performance, such as channel capacity, spectrum efficiency, and delay, and aim to achieve bit-level reliable and accurate transmission. Driven by the rapid advance in wireless communication techniques, technical-level problems have been widely explored and the system capacity is gradually approaching the Shannon limit. However, the conventional design is independent of the meaning conveyed by the symbol and the underlying transmission task. This means that conventional wireless communication systems are unable to meet

Fig. 8. ***Traditional communication system model.***

the semantic exchange demands for various transmission tasks, especially in the era of data explosion. This motivates people to think about a new paradigm by taking semantics into consideration.

Semantic communication is predicted to be a revolutionary post-Shannon paradigm and falls into the second or even third level. The next-generation communication systems always have a clear task or goal for data transmission, such as emergency detection and image classification. Semantic communication can extract and transmit the most necessary information (i.e., semantic meaning) which is tightly relevant to accomplish a task, instead of fully transmitting bit sequences. Therefore, semantic communication is regarded as a kind of task-oriented or goal-oriented communication.

Figs. 8 and 9 show the system model for traditional communication and semantic communication, respectively. The biggest difference between the two paradigms is the data processing phase. Traditional communication systems are designed to convert the source data into bits and symbols to process. Semantic communication systems are designed at the semantic level instead of the bit level. The semantic encoder is employed to extract semantic features based on the goal of the underlying transmission task. The semantic decoder is employed to reconstruct the corresponding semantics. The knowledge base (KB) is employed on both the transmitter and receiver. KB represents the common knowledge shared by both sides which stores facts, concepts, relationships, and rules in a machine-readable format. KB serves as a foundation for reasoning and plays a crucial role in semantic communications by enabling systems to extract meaning from data, infer new knowledge, and make informed decisions based on the available information. With the help of KB, only the most relevant semantic information is transmitted, and thus, the redundant data can be largely removed and the network efficiency can be further improved. The semantic channel is a virtual channel that focuses on preserving and conveying the meaning or semantics of the data being transmitted. Any discrepancy or distortion in the meaning or interpretation of information exchanged due to differences in KB or reasoning will lead to semantic noise. Semantic channels play a crucial role in facilitating knowledge sharing between the transmitter and receiver.

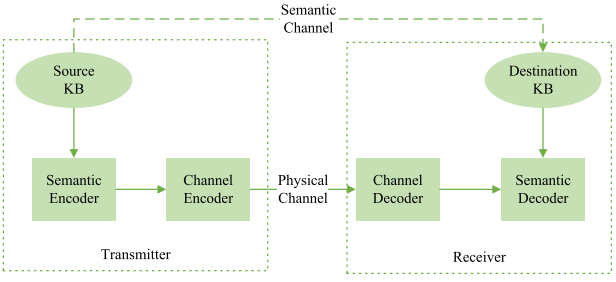

Fig. 9. ***Semantic communication system model.***

The study of semantic communications has attracted much attention in the past few years due to its novelty and utility. Following Shannon's information theory, some research works have studied the semantic information theory [105], [106], [107], [108], [109], [110]. By replacing the statistical probability with logical probability, semantic entropy was first proposed in [105] to measure the amount of semantic information for sentences. The semantic channel coding theorem has been studied, and the corresponding semantic channel capacity for a discrete memoryless channel was defined in [106]. The rate distortion theorem has also been explored in the context of semantic communications [107], [108]. The semantic channel capacity was explored in [109] by taking the semantic KB into consideration. A general mathematical framework of semantic communications was proposed in [110] based on a synonymous mapping between semantic information and syntactic information. Besides these fundamental-level works, there are a few tutorial and survey articles aiming to give a general overview of semantic communications [10], [11], [12], [111].

Inspired by the recent advancements in AI, the semantic communication infrastructure has enabled many intelligent applications, such as immersive reality, industrial control, and live video communications. DL and neural network-aided semantic feature extraction methods have been successfully implemented for various information sources. RL-assisted scheduling and transmission schemes have also been widely investigated in various task-oriented communication systems. In the following, we are going to go through the AI-empowered techniques for semantic communications in detail.

B. Semantic Sensing and Sampling

Most existing cutting-edge information processing techniques share a common framework: the signal is first sensed at a fixed sampling rate, and then, a large amount of raw data is generated and transmitted through the network to the receiver. The raw data contain a large volume of redundant information, and thus, the communication and computing resources are largely wasted for transmitting the unnecessary data. Against this background, the semantic sensing and sampling method is regarded as a possible way to improve the network efficiency by generating a small volume but necessary samples with respect to the underlying task.

Semantic-aware CS technique enables the generation and compression of meaningful content, facilitating the optimization of wireless resources for enhanced efficiency

in semantic processing. CS is a commonly used signal process technique that jointly considers sampling and compression and can use fewer samples than required by the Nyquist–Shannon sampling theorem for signal acquisition and reconstruction [112]. Since camera-based sensing devices have been widely deployed in various intelligent communication systems, image and video streams take up a large amount of IP traffic. CS has been extensively applied to capture high-speed videos and hyperspectral images at low frame rates. Despite their contributions, measurements in existing CS systems are mainly obtained by a fixed sampling matrix which is regardless of the content of images. Therefore, this motivates people to study image-CS in the context of semantic communications. A semantic-aware image CS architecture was studied in [113]. Instead of using the fixed matrix, adaptive measurement matrices have been adopted for different images based on DL to improve the sensing efficiency and restore accuracy. Fig. 10 shows the peak signal-to-noise ratio (PSNR) versus the average CR under two image datasets. PSNR is a classic performance metric for image transmission, which measures the ratio between the maximum power of the desired signal and the power of the noise that corrupts the desired signal. Fig. 10 verifies that the semantic-aware image CS method achieves better performance than the traditional compressed method. Moreover, by jointly considering the sampling rate allocation and model scalability, a content-aware scalable image-CS architecture was proposed in [114]. Specifically, a CNN-aided method has been proposed to recognize the image distribution and a lightweight strategy has been designed to achieve adaptive sampling and ratio allocation. Similarly, an RL-based video CS architecture was built in [115]. In addition, a joint semantic sensing and communication framework has been developed in the scenario of XR [116]. In that work, both spatial and temporal sampling strategies were performed to extract semantic information efficiently.

In the preceding paragraph, semantic-related sensing problems are reviewed in the static scenario, where the data are sampled from a certain signal based on a semantic quality measure. This measure only relates to the transmission task but ignores the changes in the information value over time. For example, in a real-time tracking and monitoring system, the recently received data sample should contain more semantic information than the old one received one hour ago. Therefore, the timeliness of information plays a vital role in such real-time control applications. In this context, timing adds a new domain to the importance of semantic information. In this regard, some researchers define semantics as the timeliness of messages over time rather than the meaning and content of the underlying signal and start to explore the joint sensing-semantic problem in the time domain. The concept of the age of information (AoI) was proposed in [117] as a new performance metric that can be used to measure the timeliness of the semantic information. Specifically, the AoI

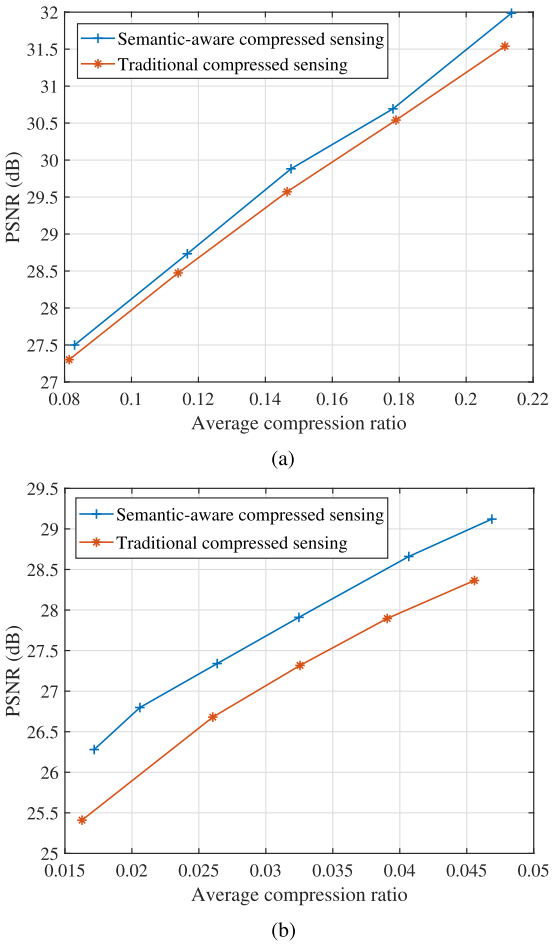

Fig. 10. ***PSNR versus average CR for semantic-aware and traditional CS methods under different datasets [113]. (a) PSNR versus average CR for different methods in the MPI-INF-3DHP dataset. (b) PSNR versus average CR for different methods in the MS-COCO2014 dataset.***

at the given time t is defined as

$$\Delta(t) = t - u(t) \tag{5}$$

where $u(t)$ is a function of time, representing the generation time of the most recently received data sample until time t. It is easy to see that the AoI is not only affected by the transmission process but also by the sampling process. Therefore, AoI and its variants (e.g., peak AoI and average AoI) oriented sampling policies have been widely studied to improve the network performance [118]. The optimal sampling rate has been obtained in different queue models with various additional properties, such as the queues with the packet deadline [119], link capacity constraint [120], and packet priority [121]. However, it is recognized that AoI may not be a perfect semantic metric. The AoI only relates to the duration of time while it is independent of the statistical variations of the data source. In reality, more samples should be generated to

represent the quickly changeable or lessly correlated data (e.g., location of moving target), while a small number of samples can be generated to represent the slowly changeable or highly correlated data (e.g., temperature). Therefore, other approaches are further proposed to quantify the usefulness of semantics, including the AoI penalty metrics [122], [123], information-theoretic metrics [124], [125], and estimation error-based metrics [126]. These metrics can be generalized as a monotonic nonlinear AoI function [127], [128]. The corresponding optimal sampling problem under the maximum sampling rate constraints has been formulated as a Markov decision process (MDP), and the optimal solution is obtained by binary search in [126], [129], [130], and [131].

Although the AoI and nonlinear AoI have created a mathematical framework for capturing the significance of semantics, they are independent of the realization (real value) of the information and may not be suitable for real-time reconstruction in practical applications, such as emergency detection. To address the limitation, more appropriate semantic metrics have been investigated. The semantic metric integrates the AoI and traditional error-based metrics to capture the significance of the packet from both time and content dimensions. The corresponding semantic-aware sampling policies which allow the sampler to generate data adaptive based on the transmission task have also been studied. The age of incorrect information (AoII) was proposed in [132] as an enabler of semantics-empowered communication. AoII is defined as a combined function of AoI and estimation error which can evaluate the semantics through both time and transmission task aspects. The function is chosen according to the goal of the underlying task. The AoII-optimal sampling policy is derived from the video stream transmission, machine overheat monitoring, and fire monitoring systems. The notion of the semantics of information (SoI) is defined in [133] which is similar to AoII. The authors explained that the SoI-based sampling policy can be obtained by deep RL and showcased the potential of applying SoI in timely tracking networks. Moreover, a new concept named goal-oriented tensor (GoT) was proposed to measure the impact of semantic mismatches on task-oriented decision-making utility [134]. A joint sampling and decision-making problem was studied, and an algorithm was designed based on game theory to find the suboptimal solution. Uysal et al. [135] and Kountouris and Pappas [136] generalized the above adaptive semantic-aware sampling policies and developed an end-to-end semantic communication architecture. The performance of semantic-aware sampling was compared with its AoI counterpart, and it was shown that semantic-aware sampling performs better in reducing the reconstruction error.

Insights and Challenges: In this section, we introduce two methods regarding semantic sensing and sampling problems from different perspectives. The semantic-aware CS is more suitable to be applied in static scenarios with sparse signals for efficient image or video reconstruction. The AoI-based semantic sampling is more suitable to be applied in real-time scenarios with correlated data sequences generated by sensing devices. Moreover, developing general metrics is an open area that can help to evaluate the data semantics and guide the process of adaptive sampling.

C. Joint Semantic-Channel Coding

Joint semantic-channel coding is the key component of semantic communication systems, which greatly differs from traditional communications. As shown in Fig. 8, a typical traditional communication system comprises two coding processes. One is the source coding process, in which the source data are compressed into bit sequence with the aim of removing the redundancy; the other is the channel coding process, in which the bit stream adds extra correction code with the aim of resisting errors caused by the imperfect transmission channel and then is modulated into symbols for transmission. This traditional separate architecture can optimize the source and channel coding independently by utilizing the separate modular design method. However, both source and channel coding schemes have approached their respective theoretical limits so far.

In this regard, joint source and channel coding (JSCC) has been widely explored in the context of information theory and coding theory, in which the source and channel are treated simultaneously during the encoding process to improve the system-level performance. Traditional JSCC scheme [137] exploited the statistical probabilities of the source data along with the characteristics of the communication channel, illustrating that the joint design outperforms the separate modular design. However, data semantics have not been taken into consideration in these works. The semantic communication system is designed to transmit semantic information rather than raw data, and the traditional source process is replaced by the semantic coding process (see Fig. 9). Traditional JSCC requires the explicit probability distribution of the source data and this means that it is difficult to model the complex source in the real world. Furthermore, the traditional JSCC is independent of the semantic features with respect to the underlying communication task.

Inspired by the traditional JSCC, semantic-aware JSCC is regarded as a practical technique route to realize semantic communications and improve end-to-end performance in different channel conditions. The advancement of DL and neural networks facilitates the implementation of joint semantic-channel coding techniques. Fig. 11 gives a simple model of joint semantic-channel coding which can be achieved through DNN architecture. DNN acts as the autoencoder and autodecoder, representing various famous and efficient models, such as CNN, GAN, and Transformer. It is pretrained in a joint and end-to-end manner to extract and transmit goal-relevant semantic features. Compared with traditional compression and channel coding methods, DL can utilize any

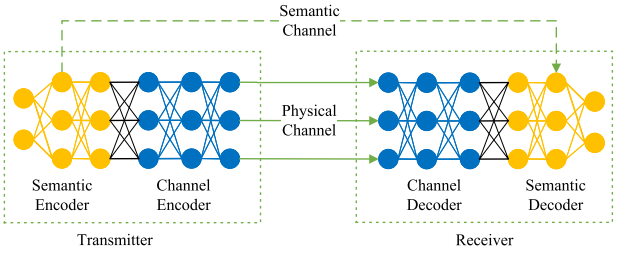

Fig. 11. ***Joint semantic-channel coding model.***

desired fidelity measures for training and is very efficient in enhancing transmission reliability. Therefore, DL-based joint semantic-channel coding has the potential to be applied to various types of data sources.

The fast development of NLP forms the basic groundwork for analyzing and understanding the SoI and facilitates the implementation of joint semantic-channel coding for text [138], [139], [140], [141]. The metric named word error rate (WER) was proposed to reflect the semantic similarity and an LSTM-enabled JSCC scheme named DeepNN was designed to minimize end-to-end distortion induced by semantic noise for text transmission [138]. This early work focuses on the word level without taking synonyms and antonyms into consideration. In [139], a sentence-level metric was proposed to measure the sentence similarity based on the Transformer model, and a DL-based semantic communication (DeepSC) framework was proposed with the aim of achieving robust transmission. A lite variant of DeepSC named L-DeepSC was further proposed in [140] for practical IoT networks. Compared with DeepSC, L-DeepSC considered the capacity of IoT devices and designed a finite-bits constellation with the aim of achieving robust and affordable text transmission. Inspired by these works, Zhou et al. [141] designed a universal Transformer-enabled semantic communication system. Different from the fixed attention structure of classic Transformer-based semantic-channel coding, a flexible circulation mechanism was introduced which enabled adaptive transmission with different semantic information under various channel conditions. In addition, other advanced techniques have also been explored to enhance performance and flexibility, such as the implementation of deep RL [142], knowledge graph [143], [144], and hybrid automatic repeat request (HARQ) scheme [145], [146].

The joint semantic-channel coding system has also been explored for speech signal transmission [147], [148], [149]. Compared with text signals, the speech signal is more complex and difficult to process and understand due to the volume, tone, background noise, and other factors. The signal-to-distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ) are the main metrics to quantify the quality of reconstructed audio signals. Weng and Qin [147] extended the text-based DeepSC framework to speech transmission and proposed a joint semantic-channel coding framework named DeepSC-S. Figs. 12 and 13 show the comparison of the performance of the DeepSC-S with traditional separate source-channel coding and semantic coding schemes under different channel models. The DeepSC-S performs better in SDR and PESQ since the joint semantic-channel coding scheme could deal with both source distortion and channel variation.

With the success of semantic-aware text and audio transmission, joint semantic-channel coding for image/video transmission has also attracted much attention. A novel DL-based JSCC (deep JSCC) scheme was proposed for image transmission over the wireless channel in the presence of AWGN and Rayleigh fading [150]. Particularly, a neural network is deployed to work as the encoder, which maps the pixel values of the input figure to the complex-valued channel input symbols and the corresponding decoder works to recover the image based on a

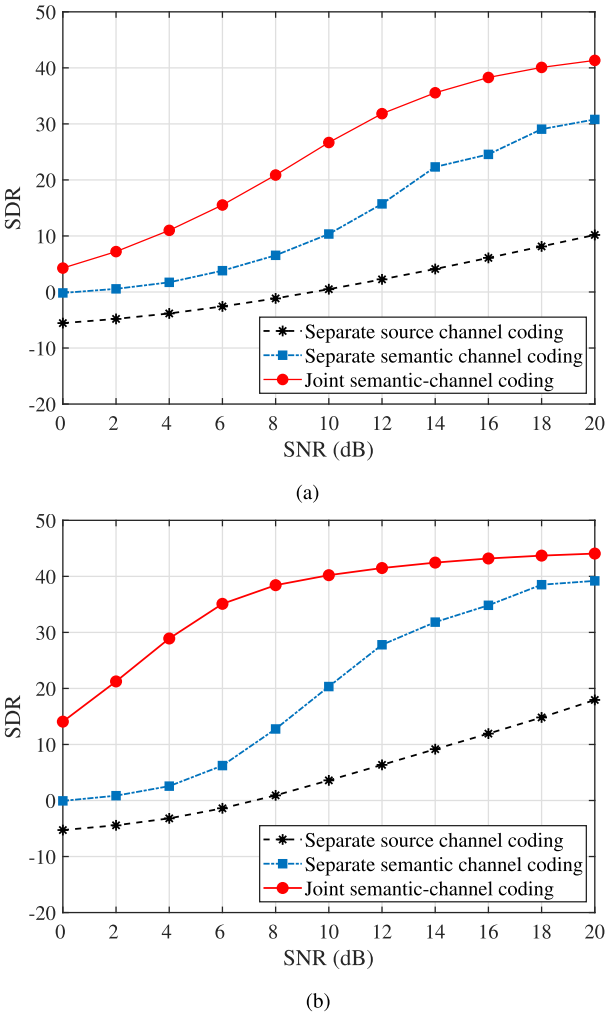

Fig. 12. ***SDR versus SNR for speech transmission with the traditional coding, semantic coding, and joint semantic-channel coding schemes under different channel models [147]. (a) SDR versus SNR in Rayleigh channels. (b) SDR versus SNR in Rician channels.***

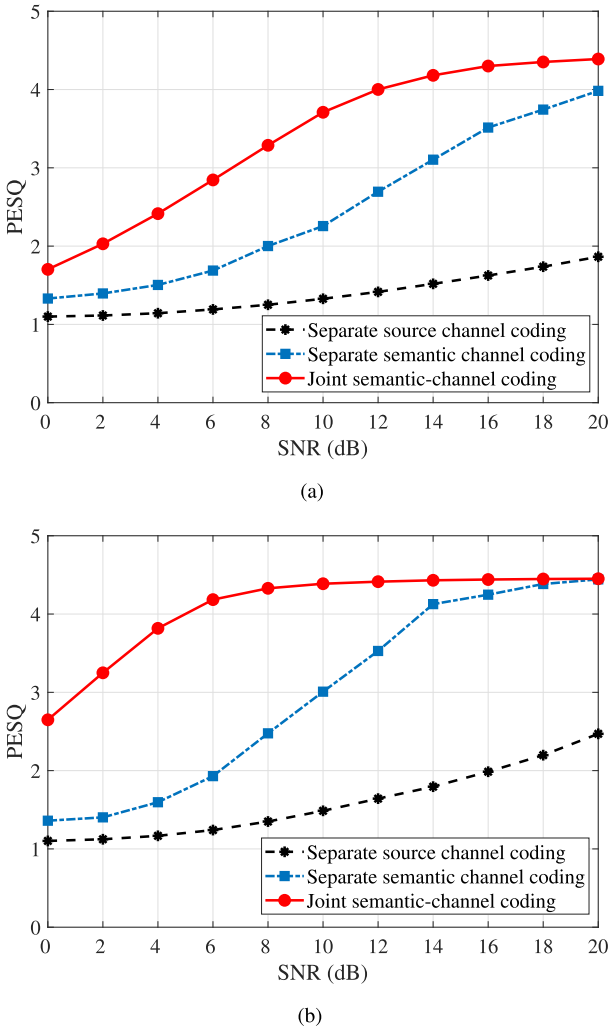

Fig. 13. ***PESQ versus SNR for speech transmission with the traditional coding, semantic coding, and joint semantic-channel coding schemes under different channel models [147]. (a) PESQ versus SNR in Rayleigh channels. (b) PESQ versus SNR in Rician channels.***

classic metric PSNR for semantic similarity. While Bourtsoulatze et al. [150] did not explicitly emphasize the semantics of the image signal, it introduced the notion of semantic extraction and transmission, serving as a significant source of inspiration for subsequent researchers. In [151], a new image reconstruction measurement named rate-semantic-perceptual loss was introduced by jointly considering the figure category and spatial correlation, and a task-oriented semantic coding framework was proposed by DL which can discard the irrelevant data adaptively. Zhang et al. [152] considered a receiver-leading semantic coding scheme to resist the distribution discrepancy between the transmitter and the shared KB for image transmission, in which the receiver works to coordinate the DNN training for semantic coding while the transmitter is not clear with the underlying task. Different from other DL-enabled semantic-channel coding approaches, a nonlinear transform source-channel coding (NTSCC) was proposed [153] where the source data are first mapped to the latent representation based on nonlinear transform and then transmitted through DL-enable JSCC scheme.

In addition to the representative single-modal data, a multimodal data transmission framework has also been developed for semantic communications. For example, the data of different types can still be correlated with each other in the XR-related applications. Semantic-aware multimodal data transmission can further improve the system performance by considering the correlation between different types of data. A DL-enabled multiuser semantic communication system (MU-DeepSC) was proposed in [154] for multimodal data transmission to execute the visual question answering (VQA) task, in which the generated multimodal data are correlated and relevant to the context. In the VQA system, the query is presented in text format while the answer is presented in image format. In that work, LSTM and CNN-based semantic-channel coding schemes were used for text and image transmission, respectively, and then, the receiver predicted the answers by fusing different correlated semantics together [155]. Moreover, by adopting Transformer, MU-DeepSC has been extended to other intelligent tasks, including image retrieval and machine translation tasks [156].

Insights and Challenges: The DL-based JSCC has achieved great success in enhancing the end-to-end performance in semantic communication systems. However, two critical challenges need to be further investigated. The first one is the lack of a general semantic JSCC framework and the high communication and computing costs. If the transmission task changes or the channel condition changes, the DNN needs to be redesigned. The retraining process requires frequent data exchange; thus, the cost of training may be unacceptable. The second challenge is the practical implementation. The existing communication infrastructure is designed and guided by Shannon's theorem; thus, semantic-aware JSCC architecture may be incompatible and the current hardware and software need to be updated accordingly.

D. Semantic-Aware Modulation and Transmission

As aforementioned, DL-based coding schemes have been widely utilized to replace conventional coding modules, in which DNNs are jointly trained to optimize a loss function that is tailored to the desired underlying task. Thanks to the success of DL, existing works have shown that learning-based semantic communications can achieve efficient and reliable transmission for various data sources. Most works focused on the semantic-aware joint coding methods, while the modulation process has not been explicitly treated.

In DeepSC systems, the output of the encoder is continuous real signals. For simplicity, many works [149], [150], [154] considered analog modulation in which the output with continuous signals is transmitted directly without

discretization into constellation symbols and can take any value. This ideal assumption regarding modulation is hard to deploy in practice due to the limitation of hardware components. Therefore, it is important to explore digital modulation in the context of semantic communications, which is more compatible with existing practical systems.

The basic idea of digital modulation is to convert the real value to discrete constellation symbols based on a function. Since this function is nondifferentiable, it is impossible to run the gradient descent algorithm to obtain the optimal mapping. To solve this challenge, quantization-based modulation methods have been explored. Xie and Qin [140] used uniform quantization to equally map the output of DNN into discrete bit sequences and then modulate them into constellation symbols. Jiang et al. [146] and Tung et al. [157] used nonuniform quantization. Specifically, Jiang et al. [146] gave an example of the binary phase-shift keying (BPSK) modulation, in which an additional DNN with the differentiable Sigmoid function (rather than the nondifferentiable step function for classic BPSK modulation) was deployed to generate the likelihood of constellation symbols instead of hard mapping. Li et al. [158] proposed a similar asymmetric quantization method, in which the rectified linear unit (ReLu) function replaced the Sigmoid function to work as the activation function for training in the nonorthogonal multiple access (NOMA)-assisted semantic communication systems with multiple users. Gao et al. [159] proposed a new metric named robustness probability to quantify the robustness degree of inference results, and obtained the optimal modulation scheme and order for robust transmission in semantic communication systems by the bisection search method.

The above quantizer-based modulation methods mainly consider the separate infrastructure in which the modulation module is independent of the coding module. This means that the system is unable to adapt to various wireless environments, and the transmission performance can be largely corrupted by the channel noise. Motivated by the idea of JSCC, the integrated coding and modulation scheme has been explored for digital semantic communication systems. The semantic-aware joint coding-modulation (JCM) scheme with the BPSK modulator was proposed in [160] based on a VAE. This integrated design was implemented by a stochastic coding process and a random coding-based modulation process. An NN was utilized to learn the transfer probability from the source data, and the random encoder was utilized to generate the specific modulated symbols for channel transmission based on probability. The performance of the JCM framework was further analyzed in [161]. Simulation results showed that the JCM outperforms the other semantic-aware digital modulation in both high and low SNR conditions, and outperforms the other semantic-aware analog modulation in the low SNR condition. In addition, channel-adaptive modulation and demodulation techniques were investigated in [162] for image transmission in digital semantic communication systems. The proposed modulation process aims to reduce the transmission delay by altering the order based on the diverse channel conditions. In the demodulation process, a new metric named log-likelihood ratio (LLR) was introduced to measure the uncertainty of the demodulation output and LLR can help to demodulate the data into continuous values rather than binary variables. Therefore, the system's robustness against fading and noise can be further improved.

Insights and Challenges: The semantic-aware modulation-related research is still at an early stage. The advancement of AI has revolutionized and changed the way of coding in semantic communications, while it leads to new challenges to the modulation process at the same time. Existing works have deployed additional NN and utilized various activation functions to map the continuous signal into discrete symbols. However, there are no general and systematic methods regarding how to choose the optimal quantizer. Moreover, the joint semantic-coding-modulation framework is also an open area that can be further explored in the future.

E. Applications and Integration With Emerging Techniques

In Sections III-A–III-D, we have reviewed the leading techniques in the context of semantic communications. In this section, we will further explore the integration of semantic communications and other techniques. More specifically, we first introduce some typical applications and representative cases of task-oriented semantic communication systems. Afterward, we discuss the latest advances toward semantic communications from different technical aspects, including the big AI model-enabled and computing network-enabled semantic communications.

The advance of end-to-end semantic communication framework has enabled many intelligent transmission tasks, such as data reconstruction [154] and image classification [163]. Semantic communication also facilitates the development of many intelligent networks, such as industrial IoT, smart home, and intelligent transport. In autonomous driving systems, a novel cooperative perception semantic communication scheme was proposed in [164]. In addition to text, speech, images, and videos, sensor data (LiDAR point clouds) in such systems also contains essential semantic information [165], [166]. Since the perception field of the autonomous car itself is usually limited by objects, such as buildings and trees, a novel detection scheme with cooperative perception is investigated for a broader perception field. Through the utilization of vehicle-to-vehicle (V2V) communication technology, different connected automated vehicles (CAVs) can exchange and fuse their sensory information. However, previous works [167], [168], [169], [170] typically assume perfect communications between CAVs and ignore underlying channel effects. In response, a cooperative perception semantic communication framework over the air

is introduced [164], which employs an importance map to extract significant semantic information at the transmitter and fuses the intermediate feature through an attention-based mechanism at the receiver. The proposed system is designed based on a JSCC architecture and trained in an end-to-end learning manner, which is optimized to achieve better semantic performance (perception accuracy).

Besides the widely accepted DL-based semantic communication systems, the big AI model or the large language model (LLM) has also been regarded as an effective approach to enable the implementation of semantic communications [171]. The big AI model is a novel framework that creates a unified and fundamental ML system based on a generic class of AI models. The generative pretrained transformer (ChatGPT) and bidirectional encoder representations from transformers (BERT) [172] are two emerging LLMs that are built by OpenAI and Google, respectively. These big models leverage the pretrained vast datasets to process the natural language and have led to remarkable advancements in various language-related tasks. Many semantic metrics are proposed based on the BERT model to measure the similarity on the semantic level, and they have been widely utilized to extract semantic information [139], [146], [173]. For example, a semantic loss function was defined by the BERT model, and a signal shaping method was proposed to minimize the semantic loss in [173]. A semantic importance aware communication (SIAC) scheme was designed in [174] by using the pretrained LLM. Particularly, a metric is defined to quantify the semantic importance of data frames by using LLM, and a semantic importance-optimal power allocation was obtained by minimizing the expected important word errors. The experimental results showed that the proposed SIAC scheme outperforms the DL-based JSCC scheme and is compatible with existing communication infrastructure.

Although DL and LLM demonstrate outstanding performance, they impose a substantial computational burden on communication devices and the computing capacity has been a bottleneck in AI-empowered semantic communication systems. Most AI-driven designs allow offline training and online inference. Although inference demands less computing power than training, the computing power requirements of inference are still substantial comparing with the traditional signal processing methods. This is challenging for battery-powered end devices, such as mobile phones. Moreover, if each part of the communication systems (from coding to transmission) is all implemented with AI designs, the computing power consumption could be extremely high. Therefore, the lack of computing resources is a crucial challenge that needs attention to support the practical applications of semantic communications. Cloud computing and edge computing are the two main approaches to address the challenge. Cloud computing is a paradigm that changes the way computational resources are accessed and managed, providing extra computing resources through the network. However, accessing the cloud network requires additional communication resources and may lead to large transmission delay. Thus, cloud computing may not be appropriate for time-sensitive semantic communication systems. Edge computing is a distributed paradigm that brings computation and data storage closer to the data source, reducing latency and enabling timely communication. Since most edge devices have limited data computing or processing capability, edge computing may be more suitable for transmission tasks with light traffic loads. Against this background, the notion of the computing network has been proposed to solve the problem. A computing network refers to an interconnected infrastructure of the network and various devices that communicate and share resources, facilitating data exchange and processing in a collaborative manner [175]. With the aid of the computing network, semantic communication systems can leverage more computing resources to support data processing and coding processes adaptively with low transmission delay.

Insights and Challenges: The combination with other emerging techniques introduces a new perspective in the design of semantic communication systems. This integrated paradigm still leaves much room for further research. The research of the computing network is in its infancy; thus, the tradeoff between the transmission delay and limited wireless resources still challenges the network performance of semantic communication systems. It is important to develop a formal mathematical framework to evaluate its performance, which can be further used to guide and optimize the design of semantic-aware resource allocation.

F. Tips and Tricks

In previous subsections, we give an overview of semantic communications, including key techniques as well as their advantages and challenges. In this section, we discuss several tricky problems and future directions in the implementation of semantic communications.

1) Will Semantic Communication Break Shannon Channel Capacity?: This is probably the most frequently asked question for semantic communications. The answer is not as they are considering the problem at two different levels. As shown in Fig. 9, there exist two channels in semantic communications. One is the physical wireless channel which is bounded by Shannon channel capacity. The other is the virtual semantic channel. From the technical level, the maximum capacity of the physical channel is the Shannon limit which could not be broken. From the semantic level, the semantic capacity of the virtual channel could be further explored and improved.

2) Can We Develop a General Semantic Information Theory Framework Similar to Shannon's Information Theory?: Shannon's classic information theory gives a comprehensive and general mathematical framework to formally explain what is information and what are the theoretical limits, which has achieved great success in the design of communication systems. Motivated by Shannon's information

theory, many researchers have attempted to develop the semantic information theory. Particularly, semantic entropy is proposed by replacing statistical probability with logical probability. The semantic rate distortion theorem has also been explored by adding the distortion between semantic features. Although great efforts have been made for semantic information theory, a widely accepted definition of semantic entropy or semantic channel capacity is still missing. Different from traditional bit-level communication systems, semantic communication is task- and goal-oriented. Therefore, there may not be a general framework to model various semantic communication systems with multimodal data. We believe semantic entropy or channel capacity should relate to the underlying transmission task and the background KB, and they need to be redesigned according to different contexts.

3) Are Existing Evaluation Metrics Applicable to Semantic Communications?: Suitable performance metrics are the foundation of the implementation of wireless systems. For example, they play a vital role in the design of loss functions and the selection of parameters in training DNN models. Although some works have started to look at semantic metrics for various data sources (e.g., text, image, and speech) as well as various perspectives (e.g., accuracy and timeliness), more appropriate and systematic performance standards are still needed before practical applications of semantic communication systems. Different from traditional communications, both objective and subjective metrics are desired to evaluate the QoE as well as human perception in semantic communication systems. Moreover, existing semantic metrics are developed for specific types of data sources; thus, it is also important to develop general metrics (like BER in traditional wireless communications) which can be applied to evaluate the system-level performance of networks with multimodal data or various tasks.

4) How to Make a Tradeoff Between System Performance and Computing Overhead?: IMT-2030 [176] has proposed that AI and communications to support AI-powered applications is one of the expected usage scenarios for 6G; thus, there is a rising trend in integrating AI into wireless communication systems. Existing semantic communication systems are mostly enabled by DNNs which consume extra computing resources and can lead to demanding computing overhead. Therefore, there exists a tricky tradeoff between network performance and computing capacity in the practical application of semantic communications. Since the computing capacity of end devices is growing, edge intelligence enables semantic communications by leveraging the distribution of computing resources toward the network edge to achieve fast data processing and exchange. Cloud intelligence can also be utilized to provide extra computing resources through the network. In addition, a new framework for computing network-enabled semantic communications [175] has also been proposed, which can integrate various computing resources and share resources among users in a collaborative manner. Moreover, lightweight AI techniques, such as model pruning and quantization, could be utilized to save computing power.

5) How to Ensure Privacy and Security in Semantic Communication Systems?: Information privacy and security is one of the most important issues in various wireless communication systems. Given that semantic communication systems only transmit partial important data (i.e., semantic features) and the decoding process relies on the receiver's KB, semantic communications can be regarded as a prospective approach for ensuring privacy and security to some extent. On the other hand, since only important data are transmitted, it may lead to malicious effects once the semantic information has been eavesdropped. Semantic communication systems include the physical channel and virtual semantic channel; thus, encrypting semantic features during transmission and maintaining the match between KBs are two directions to achieve secure communications. Although privacy can be protected by encrypting the semantic features, the communication overhead may also be increased by introducing extra bits for secure coding. Therefore, the tradeoff between data security and transmission efficiency should be explicitly treated based on the underlying transmission tasks.

6) What Are the Potential Applications of Semantic Communications in 6G? What Is the Expectation for the Standardization?: Semantic communication can be regarded as a powerful enabler for augmented reality (AR)/virtual reality (VR), XR, and immersive communications in 6G, linking the cyber world to the physical world. Immersive communication in 6G has very strict requirements in terms of timeliness and capacity. The semantic communication paradigm can extract the semantics from various signals, such as movements, gestures, and speech. By removing unnecessary information, the volume of transmitted data can be decreased and seamless data exchange can be guaranteed. Semantic communications alleviate downlink pressure and emerge as a facilitator for XR-based immersive communications, providing enhanced and new capabilities. The development of semantic communications is still in the initial phase; thus, there is a long way to go toward standardization. In 2021, an international standard technical report titled "Architectural Framework for Semantic Communication in IoT and Smart City & Community Services" [177] has been formally approved for project initiation by International Telecommunication Union Study Group 20 (ITU SG20). Moreover, IMT-2030 has expanded the capabilities and usage scenarios of IMT-2020. The applicable AI-related capabilities are expected to grow to support immersive communications. In this context, semantic communication is anticipated to work as the key technique to facilitate enhanced and enriched immersive experiences, broaden ubiquitous coverage, and empower new aspects of collaboration.

IV. LEARNING-BASED WIRELESS RESOURCE ALLOCATION

In this section, we examine ML-based techniques for resource allocation in wireless networks. We introduce supervised and unsupervised learning for solving resource optimization problems as formulated in traditional methods and then investigate the new paradigm of using RL to directly optimize toward the system design objective, without explicitly formulating an optimization problem. We further single out the GNN-based techniques since the GNN stands out as an effective tool to capture complicated wireless interference. Finally, we touch upon resource allocation issues in semantic communications.

A. DL for Resource Allocation

Resource allocation involves dynamically assigning communication resources, such as frequency spectrum, transmitted power, and time slots, to multiple users in the wireless network. The goal is to optimize network performance and enhance resource utilization efficiency. Conventionally, addressing resource allocation involves formulating an optimization problem explicitly and employing mathematical programming techniques to solve the problem. Despite its popularity, a great many formulated optimization problems are difficult to resolve. In response, DL has been investigated in recent years in an attempt to find (near-) optimal solutions for optimization problems due to its remarkable potential in addressing complicated decision problems. In particular, DL learns a mapping from the channel or network state to the resource allocation decision, such as power/spectrum/time slot selection. The computation-intensive training can be performed offline, while during online deployment, a simple forward pass through the DNN is sufficient to produce the decision output. There are usually two common ways to leverage DL in solving resource allocation problems, i.e., the supervised and unsupervised learning paradigms.

1) Supervised Learning Approach for Optimization: The supervised learning approach trains a DNN to minimize the discrepancy between the output and the ground truth that is usually generated by traditional optimization algorithms. The following WSR maximization problem for an interference channel with K transmitter–receiver pairs was considered in [178]:

$$\max_{\{p_1,\ldots,p_K\}} \sum_{k=1}^{K} \alpha_k \log\left(1+\frac{|h_{kk}|^2 p_k}{\sum_{j\neq k}|h_{kj}|^2 p_j+\sigma_k^2}\right)$$
$$\text{s.t. } 0\le p_k\le P_{\max} \quad \forall k=1,2,\ldots,K \tag{6}$$

where P_k is the transmit power of the kth transmitter, $P_{\max}$ is the power constraint, σ_k is the noise power of the kth receiver, and α_k is the weight for kth user. h_{kk} and $h_{k,j}$ represent the signal and interference channel, respectively. Specifically, the ground truth was obtained by running the WMMSE algorithm [42], and the loss function is set as the mean-squared error between the ground truth and the output of the network. Simulation results show that the trained DNN can achieve performance close to the WMMSE algorithm over a wide range of settings while reducing the computation time by an order of magnitude. The basic procedure of supervised learning for wireless resource allocation is sketched in Fig. 14. Several enhancements have been developed in the literature to improve upon the architecture proposed in [178]. An approach using the CNN to solve the wireless link scheduling problem was proposed in [179] due to its capability to better capture and exploit spatial correlation information. Specifically, they consider the geographic location information of transmitter–receiver pairs as the density grid matrices, which are constructed by directly counting the total number of transmitters and receivers in each cell. Then, they put the density grid matrices into a CNN to get the interference patterns that each link causes to its neighbors and each link receives from its neighbors. A GNN-based supervised learning framework was proposed in [180] to address the link scheduling problem in similar interference channels. They model the wireless network as a directed graph in which the communication links are modeled as nodes and the interference links are modeled as edges. Significant performance gains have been observed, and the robustness of GNNs to varying system configurations is also demonstrated. Deeper investigation of GNN-based resource allocation design will be relegated to Section IV-C.

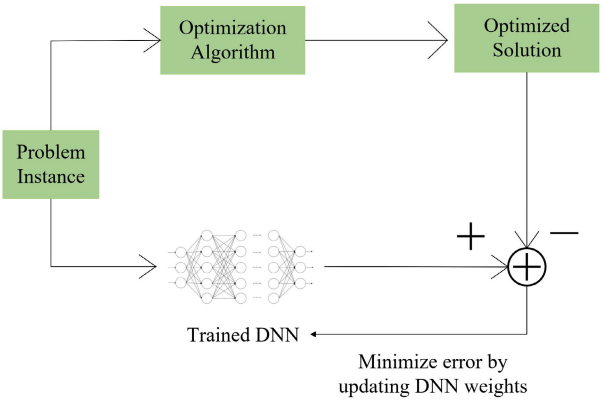

Fig. 14. *Illustration of the supervised learning paradigm of DL for resource optimization.*

2) Unsupervised Learning Approach for Optimization: The performance of supervised learning approaches is bounded by conventional algorithms that are used to produce the training labels. To address the issue and enhance the performance of deep-learning-based optimization, unsupervised learning methods have been proposed. These approaches involve training the neural network directly based on the optimization objective, in contrast to supervised learning approaches that rely on conventional algorithms to provide high-quality labels. For example,

in [181], the training loss function is designed based on the sum rate in order to address the K-user sum rate maximization problem with quality-of-service (QoS) constraints, i.e.,

$$\text{Loss} = \mathbb{E}_{\mathbf{h}}\left[-\sum_{i=1}^{K} R_i\left(\mathbf{h}, \boldsymbol{\theta}\right) + \lambda \sum_{i=1}^{K} \text{ReLU}\left(r_{\min} - R_i\left(\mathbf{h}, \boldsymbol{\theta}\right)\right)\right]$$

where $R_i(\mathbf{h}, \boldsymbol{\theta})$ is the rate of ith receiver, $r_{\min}$ is the per-user minimum rate requirement, and λ is a parameter to balance the two terms. The approach introduces penalty terms to incentivize the network output to meet QoS constraints. If the output does not satisfy the quality of service constraint, i.e., $R_i(\mathbf{h}, \boldsymbol{\theta}) < r_{\min}$ for some i, then we will have $\text{ReLU}(r_{\min} - R_i(\mathbf{h}, \boldsymbol{\theta})) > 0$ and minimizing the loss function will incentivize the network parameters to decrease this term by improving performance of the weak user. In [182], the penalty function was combined with the primal-dual learning approach, introducing a regularized unsupervised learning framework. This framework was devised to improve the resilience of the primal-dual learning-based unsupervised learning framework.

Insights and Challenges: While the DL-based methods have achieved satisfactory results in many resource allocation problems, it is important to note more careful design specialized to wireless resource allocation problems is needed. Special neural network architectures that capture the intricacies of wireless interference shall be developed, and the generalization of such networks to more complicated and time-varying environments needs deep investigation. In supervised learning approaches, determining a minimum-sized neural network architecture with enough capacity to fit the supervised data provided by conventional algorithms is of interest from both theory and practice perspectives. For unsupervised learning, it is critically important to devise an appropriate loss function-based system design objective, which is often highly nonconvex and difficult to train using gradient descent. In addition, a theoretical understanding and comparison of the supervised and unsupervised learning approaches, in terms of performance limit and convergence speed, is also an interesting area worthy of further exploration.

B. RL for Resource Allocation

RL is a promising solution for wireless resource allocation problems, which can be attributed to its capacity to address sequential decision-making under uncertainty. Besides, the hard-to-optimize objective issue can be efficiently tackled by RL because of its flexibility in designing the reward function. In addition, RL provides a feasible distributed learning structure, which is appealing in resource allocation [183]. Mathematically, the RL problem is modeled as an MDP. Fig. 15 demonstrates a general agent-environment interaction of the RL-based resource allocation methods. At each discrete time step t, the agent

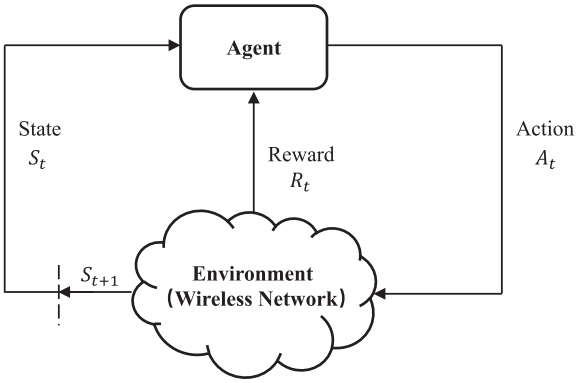

Fig. 15. *Agent–environment interaction of the resource allocation in wireless networks.*

observes its state S_t. Then, the agent takes an action A_t according to its policy network $\pi(A_t|S_t)$. Thereafter, the agents receive a reward R_t and the environment transitions to the next state S_{t+1}, dictated by the probability $p(S_{t+1}, R_t|S_t, A_t)$. The goal of RL is to find an optimal policy π_* that maximizes the expected cumulative discounted rewards G_t, which is denoted as

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k}, \quad 0 \le \gamma \le 1 \tag{7}$$

where γ is the discount factor. A wide variety of deep RL algorithms, including value-based methods, such as deep Q-network (DQN), policy-based methods, such as REINFORCE, and actor-critic algorithms that blend these two categories, such as proximal policy optimization (PPO) and soft actor-critic (SAC), have been utilized in resource allocation problems in wireless networks.

A dynamic spectrum access problem is considered in [184] where the user chooses a good channel among N channels to transmit data. Here, the user is regarded as an agent, and it selects actions using DQN with the history of previous choices and transmission results up to M time slots. Simulation results demonstrate that the RL-based method can closely approach the optimal genie-aided myopic policy. In [185], a joint user association and spectrum access problem is considered, where each user in the heterogeneous cellular network needs to associate with a BS and select the spectrum. RL is leveraged to maximize the users' transmission capacity while satisfying QoS requirements in a distributed way. For power allocation in wireless networks, a problem where multiple communication links share the same spectrum and aim to maximize the WSR is considered in [186]. Each transmitter acts as an agent and explores the environment following the ϵ-greedy policy. A centralized controller is adopted to train the DQN with the experience collected from all agents and broadcast the parameters of the DQN to each agent for distributed execution. This RL-based method can track

channel evolution and outperforms WMMSE and fractional programming algorithms.

Joint spectrum and power allocation is an essential issue in wireless networks, and we discuss its application in vehicular networks as an example. The system model consists of multiple vehicle-to-infrastructure (V2I) and V2V links to provide high data rate service and reliable safety-critical message exchange service. A distributed DQN-based approach for the unicast and broadcast scenarios in vehicle-to-everything (V2X) communications is developed in [187]. Each V2V link is regarded as an agent that selects the spectrum and adjusts its transmit power according to the local observation. The DQN is shared by all agents and is trained using the experience collected by all V2V agents. Besides, their actions are updated asynchronously to stabilize the environment transition. In addition, a multiagent RL (MARL)-based method is utilized in [188], building upon the work in [187]. To overcome the environment nonstationarity problem, the local observation of each agent is augmented by additional information that indicates other agents' behaviors, with which the training process can be stabilized. Numerical results show that the MARL method enables cooperative learning among multiple V2V agents to enhance overall system performance. Further in [189], the architecture is restructured to a centralized decision-making mechanism with very low signaling overhead. Each V2V transmitter learns to compress its local observation using a DNN and feed the compressed information back to the central agent to produce the actions of all V2V links.

Despite the strong capability of RL to solve sequential decision-making problems, it also faces challenges related to scalability. There are already some works trying to improve the scalability of RL and make it more practical. A joint user scheduling and downlink power control problem was investigated in [190], where multiple access points need to select the UE to serve according to the received SINR. Access points act as agents, and their states contain the weights and the SINRs of their associated UEs. As the user number of each access point is different, it may cause the dimension of the network input to vary and disable the trained policy network. To deal with this problem, a fixed number of users are sorted according to the proportional fairness principle to improve the scalability, whose information will be the input of the neural network. Numerical results show that though the RL-based approach schedules the UEs in a decentralized manner, it achieves a similar sum rate with the centralized information-theoretic method.

Another challenge of using RL to address resource allocation is often termed the reality gap, which refers to the difference between the simulation and the real environmental dynamics. RL generally requires retraining to adapt to the new environment, which is quite inefficient. To decrease the number of interactions and enable fast learning in the new environments, meta RL was leveraged in [191]. By pretraining the model across a variety of

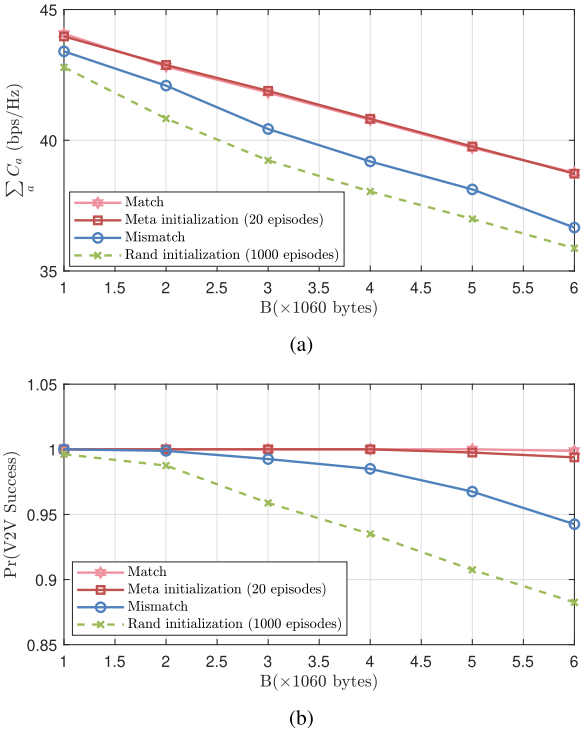

Fig. 16. ***Performance comparison of the meta-RL-based resource allocation approach and other baselines with varying sizes of V2V payload size B. (a) Sum rate of V2I links. (b) Success transmission probability of V2V links.***

environments, the model learns the prior knowledge of these environmental dynamics, which can be combined with a limited amount of data from the new environment to achieve fast adaptation. Fig. 16 shows that when the agent is initialized with the parameters learned from meta-training and rolls out 20 episodes in the new environment, it can achieve a similar performance with the matched policy, which needs to learn from 3000 episodes when trained from scratch. Besides, the randomly initialized agent cannot adapt to this new environment even with 1000 episodes, which demonstrates the efficacy of meta-RL-based approaches.

Insights and Challenges: RL-based resource allocation methods have shown their effectiveness in solving sequential decision-making problems in wireless networks. However, there are still many challenges that call for further research. First, more efforts are needed to better understand the performance limit and stability of MARL problems. Currently, researchers tend to use heuristic methods to tackle nonstationarity issues, which lack theoretical guarantees. Besides, how to make agents communicate with each other to enable better cooperation needs further exploration. Second, RL requires extensive interactions with the environment, which is usually impractical in real scenarios. Therefore, leveraging meta RL, offline RL, or other approaches to reduce the number of online

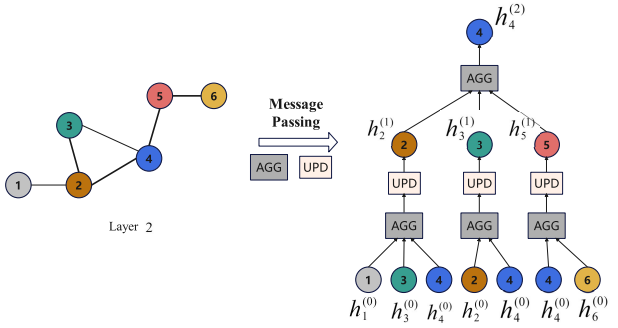

Fig. 17. ***Process of message passing.***

interactions and improve the scalability is also an area that warrants further investigation.

C. GNN-Based Resource Allocation

The graph is an efficient tool to capture wireless interference, where communication links are represented as nodes or vertices in a graph while the edge can be used to describe their mutual interference. As a result, modeling a wireless network as a graph network helps to provide insights into the structure of communication networks and the relationships between communication links. GNNs, neural network structures specialized in processing graph data, thus, serve as an ideal method to improve learning-based wireless resource allocation.

Essentially, GNNs use a framework of message passing where messages are exchanged along edges and updated by neural networks [192]. Fig. 17 illustrates a two-layer message-passing model. The superscripts and the subscripts denote the number of iterations and nodes, respectively. During the lth message-passing iteration, each node $i \in V$ will update its hidden embedding $h_i^{(l)}$ based on the messages from i's neighborhood $N(i)$. The message-passing model in Fig. 17 aggregates messages from node 4's local graph neighbors (i.e., 2, 3, and 5), and in turn, the messages from node 4's local graph neighbors are based on information from their corresponding neighborhoods, and so on. The message passing stage consists of two main functions, UPD$(\cdot)$ and AGG$(\cdot)$ [192], and the update of the hidden embedding can be expressed as

$$\begin{aligned} h_i^{(l+1)} &= \mathrm{UPD}^{(l)}\left(h_i^{(l)}, \mathrm{AGG}^{(l)}\left(\left\{h_j^{(l)} \forall j \in N(i)\right\}\right)\right) \\ &= \mathrm{UPD}^{(l)}\left(h_i^{(l)}, m_{N(i)}^{(l)}\right) \end{aligned} \tag{8}$$

where at the lth iteration, the AGG$(\cdot)$ takes as input the embeddings of i's neighborhood $N(i)$ and computes a message $m_{N(i)}^{(l)}$. Then, the UPD$(\cdot)$ function combines i's previous embedding $h_i^{(l)}$ with the message $m_{N(i)}^{(l)}$ and generates the updated hidden embedding $h_i^{(l+1)}$. Apparently, each iteration aggregates one-hop neighborhood information, and more comprehensive information about the graph structure can be obtained by iterating the message-passing process several times. Furthermore, by designing different UPD$(\cdot)$ and AGG$(\cdot)$ functions, different kinds of GNNs, such as graph convolutional networks (GCNs) [193], graph attention networks (GATs) [194], and graph isomorphism networks (GINs) [195], can be realized.

To utilize GNNs for radio resource management, the graph model for wireless networks is necessary. Commonly, the communication links are viewed as nodes and the interference relationships between communication links are depicted in edges. In [196], radio resource management problems are formulated as an optimization over graphs that enjoy a universal permutation equivariance property, and a resource management scheme based on message-passing graph neural networks (MPNNs) is proposed. The MPNN-based method is amenable to transference to large-scale radio resource management problems subject to peak-power constraints.

Additionally, a random edge graph neural network (REGNN) parameterizing the resource allocation policy is considered in [197]. The REGNN-based method performs convolutions over random graphs in the wireless network and can be applied to different network settings. To address the network utility maximization problem with constraints on the long-term average performance of users, an unsupervised primal-dual approach is proposed in [198]. The state-augmented algorithm takes as input of the instantaneous network state with the dual variables corresponding to the constraints, and simulation results demonstrate that this scheme can achieve near-optimal performance for radio resource management. Further, a resilient radio resource management optimization scheme with per-user minimum-capacity constraints is developed [199]. By introducing learnable slack variables, this method can adaptively set the per-user minimum capacity and automatically match the underlying network conditions.

Fig. 18 shows the performance comparison between several GNN-based methods, including the GCN, the GAT, and the GIN, against the baseline WMMSE method, to address the power control problem in interference channels described in (6). It can be observed that the performances of the three GNN-based methods are stable as the problem size increases and competitive with the WMMSE baseline. This verifies the excellent generalization of GNN-based methods to larger wireless network scales, a desirable characteristic for practical implementation.

Insights and Challenges: Although GNN-based methods have proved efficient for exploiting the spatial structures of wireless networks, the irregularity of large-scale graph structures, the complexity of node features, and the dependence on training samples exert tremendous pressure on computational efficiency, memory management, and communication overheads of GNN models. Further research is needed to accelerate and optimize the GNN programming framework to enable efficient training on large-scale data. Additionally, most of the existing GNN-based resource

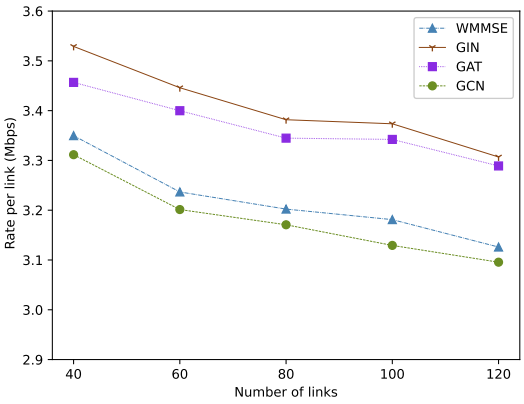

Fig. 18. ***Performance comparison between GNN-based methods and the WMMSE scheme with changing network scales and a fixed density.***

allocation algorithms require accurate CSI, which will cause a huge transmission overhead in a dense wireless network. Some other methods schedule resources solely based on the geographic locations of the transmitters and the receivers while failing in fast-fading communication scenarios. How to reduce CSI requirements while obtaining a satisfactory performance in a fast-fading communication scenario requires further investigation.

D. Semantic-Aware Resource Allocation

Conventional performance metrics that measure bit-based characteristics are no longer applicable to semantic communications. For example, to assess the performance of semantic-aware networks, the achievable data rate of all users is not appropriate for network optimization, and new metrics are desired to facilitate network optimization at the semantic level. Particularly, the achievable data rate should be replaced by the semantic rate, which measures the transmission rate of semantic symbols.

Some recent works have investigated the resource allocation in semantic-aware networks for certain specific tasks. In [200], the semantic information is extracted by the knowledge graph, and the resource blocks in semantic communications are optimized to maximize the semantic similarity between the transmitter and the receiver. Similarly, Liu et al. [201] proposed a resource allocation optimization method based on the importance of extracted semantic features. To further investigate the resource allocation for semantic communications from the transmission perspective, Yan et al. [202], [203] designed two new metrics named semantic spectrum efficiency and semantic QoE, and explored the semantic spectrum efficiency and the semantic QoE maximization problems in text-based semantic communication systems and multimodal semantic communication systems, respectively.

Another challenge for the implementation of semantic communications is the limited computing capacity of end devices. To fully utilize network resources, Ji and Qin [204] proposed a semantic-aware task offloading system by jointly optimizing the allocation of both computation and communication resources. Qin et al. [175] proposed a computing network-enabled semantic communications framework, in which computing resources are integrated and shared collaboratively among multiple end devices.

Generally, the above resource allocation problems can be modeled as mixed integer nonlinear programming (MINLP) optimization problems, which consist of both discrete and continuous variables. Due to the NP-hard nature, traditional mathematical optimization tools require high computational complexity and cannot guarantee performance under fast-varying channels. The learning-based methods, such as RL, can address the unstable issue by making decisions under uncertainty. In [204], a distributed multiagent PPO (MAPPO) algorithm was proposed to jointly optimize the transmit power and offloading policy. As shown in Fig. 19, the proposed MAPPO framework outperforms the benchmarks of conventional RL schemes and achieves near-optimal solutions under various noise power.

E. Tips and Tricks

1) Will Supervised Learning Outperform Unsupervised Learning in Resource Allocation?: From the discussions in Section IV-A, we have seen both supervised and unsupervised learning-based approaches show potential in addressing wireless resource allocation problems. One may be tempted to ask: which method is better in practice? Unfortunately, it is difficult to give a definitive answer to this question in general. Some clues can be found in [205], which focuses on the particular problem of interference management in wireless networks. It is found in the two-user power control problem that unsupervised learning performs much worse than the supervised counterparts

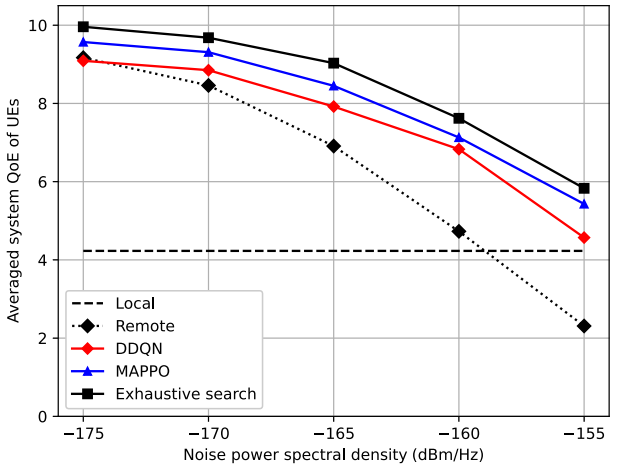

Fig. 19. ***Performance of the MAPPO framework to optimize the transmit power and the offloading policy for the semantic-aware task offloading systems.***

since it is more likely to get stuck at some low-quality local optima. For the more general case, it is established that sufficient conditions can be identified when supervised learning can achieve good performance while none can be ascertained for unsupervised learning. This is particularly true when high-quality labels are provided for supervised learning, in which case it will outperform unsupervised learning-based approaches. In addition, the supervised learning approach converges faster when the labels have better quality.

2) How to Choose the Proper RL Algorithm for a Specific Resource Allocation Problem?: An appropriate RL algorithm can improve the performance of our resource allocation scheme and stabilize the training process. Several factors need to be considered to choose an efficient RL algorithm for a specific resource allocation problem. First, whether the action space of the problem is discrete, continuous, or fixed should be decided. Value-based RL algorithms, e.g., DQN, can only tackle the problems with discrete actions because the policy network outputs the Q value of each action in action space, which is problematic if the action is continuous, such as the continuous transmission power control in the wireless network. Besides, sample efficiency is a crucial factor in RL. On-policy algorithms, e.g., REINFORCE, are inefficient as they discard collected experience data once the policy is updated. Off-policy algorithms reuse the experience data collected before, so they achieve higher sample efficiency and may lead to faster convergence.

Lastly, we find that though several RL algorithms achieve desirable performance in wireless resource allocation, such as deep deterministic policy gradient (DDPG) and twin delayed DDPG (TD3), they depend on sophisticated hyperparameter tunning, which makes it hard to apply to general network settings. Fortunately, some RL algorithms such as PPO and SAC, where the hyperparameters do not affect the ultimate performance significantly or the algorithm can adjust parameters by themselves, maybe better choices when utilizing RL in wireless resource allocation problems.

3) How to Utilize the GNN for the Resource Allocation Problem in Wireless Networks?: In GNNs, the node embedding maps each node to a low-dimensional vector that captures the graph structure and feature information of the node. Most of the resource allocation problems in communication scenarios are node-level tasks, and by learning node embeddings, the GNN can capture the graph structure and obtain the node features, allowing for more accurate node-level prediction and analysis. In each iteration, the GNN model passes and aggregates the information of nodes and edges according to the topology of the graph and gradually updates the node representations to obtain node embeddings. Essentially, the graph topology significantly influences the prediction and analysis of graph data, and how to construct a suitable graph model for the wireless network needs to be carefully considered. Commonly, each communication link can be treated as a node while its interference channels are recognized as links connected to that node. Correspondingly, CSI can be used as node features and edge attributes can incorporate properties of the interference channels. However, depending on the availability of CSI in practice, the node and edge features may change accordingly.

V. CONCLUSION

With innovations in AI, wireless communications have significantly revolutionized the way we interact and communicate. In this article, we explicitly investigated AI-enabled wireless communications, including both traditional bit-related techniques as well as semantic-related techniques. We also studied the challenges and open areas in employing AI/ML in the field of wireless systems. Moreover, we provided useful and insightful tips and tricks to give guidance in the design of intelligent wireless communication systems that can meet the demands for high data rates, low latency, massive connectivity, energy efficiency, and seamless data exchange.

As we move toward the future, there is a growing need to think about standardization which is important for the industry to achieve widespread implementation of AI/ML in wireless communications. When it turns to the semantic level, AI is largely used for coding and decoding at the transceiver sides to serve upper layer applications. However, due to the large volume of data and frequent handover, channel modeling at the semantic level will be more complicated than its traditional counterpart at the bit level. Therefore, in addition to just applying AI to the end-to-end semantic communication model training, AI could also be applied in the physical layer to facilitate channel estimation for semantic communications in the future. AI-empowered traditional wireless communications have achieved great success in various aspects. We believe that AI could also bring new degrees of freedom for future semantic communications.

Acknowledgment

The authors would like to acknowledge the support of Nantong Research Institute for Advanced Communication Technologies. ■

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[2] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020.

[3] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[4] Y. Cui, F. Liu, X. Jing, and J. Mu, "Integrating sensing and communications for ubiquitous IoT: Applications, trends, and challenges," *IEEE Netw.*, vol. 35, no. 5, pp. 158–167, Sep. 2021.

[5] Z. Qin, F. Gao, B. Lin, X. Tao, G. Liu, and C. Pan, "A generalized semantic communication system: From sources to channels," *IEEE Wireless Commun.*, vol. 30, no. 3, pp. 18–26, Jun. 2023.

[6] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based

machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 4th Quart., 2019.

[7] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3072–3108, 4th Quart., 2019.

[8] Y. Shi et al., "Machine learning for large-scale optimization in 6G wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2088–2132, 4th Quart., 2023.

[9] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 3rd Quart., 2020.

[10] D. Gündüz et al., "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.

[11] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Ye Li, "Semantic communications: Principles and challenges," 2021, *arXiv:2201.01389*.

[12] W. Yang et al., "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, 1st Quart., 2023.

[13] C. Huang et al., "Artificial intelligence enabled radio propagation for communications—Part II: Scenario identification and channel modeling," *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 3955–3969, Jun. 2022.

[14] Y. Yang, Y. Li, W. Zhang, F. Qin, P. Zhu, and C.-X. Wang, "Generative-adversarial-network-based wireless channel modeling: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 22–27, Mar. 2019.

[15] T. J. O'Shea, T. Roy, and N. E. West, "Approximating the void: Learning stochastic channel models from observation with variational generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Netw. Commun. (ICNC)*, Feb. 2019, pp. 681–686.

[16] S. Seyedsalehi, V. Pourahmadi, H. Sheikhzadeh, and A. H. G. Foumani, "Propagation channel modeling by deep learning techniques," 2019, *arXiv:1908.06767*.

[17] H. Xiao, W. Tian, W. Liu, and J. Shen, "ChannelGAN: Deep learning-based channel modeling and generating," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 650–654, Mar. 2022.

[18] T. Orekondy, A. Behboodi, and J. B. Soriaga, "MIMO-GAN: Generative MIMO channel modeling," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Seoul, South Korea, May 2022, pp. 5322–5328.

[19] W. Xia et al., "Generative neural network channel modeling for millimeter-wave UAV communication," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 9417–9431, Nov. 2022.

[20] Y. Liu, Z. Tan, H. Hu, L. J. Cimini, and G. Y. Li, "Channel estimation for OFDM," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1891–1908, 4th Quart., 2014.

[21] Y. Li, N. Seshadri, and S. Ariyavisitakul, "Channel estimation for OFDM systems with transmitter diversity in mobile wireless channels," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 3, pp. 461–471, Mar. 1999.

[22] P. Dong, H. Zhang, G. Y. Li, I. S. Gaspar, and N. NaderiAlizadeh, "Deep CNN-based channel estimation for mmWave massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 989–1000, Sep. 2019.

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

[24] W. Jin, H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Adaptive channel estimation based on model-driven deep learning for wideband mmWave systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

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

[26] S. Ji and M. Li, "CLNet: Complex input lightweight neural network designed for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2318–2322, Oct. 2021.

[27] Q. Cai, C. Dong, and K. Niu, "Attention model for massive MIMO CSI compression feedback and recovery," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–5.

[28] X. Li, J. Guo, C.-K. Wen, and S. Jin, "Auto-CsiNet: Scenario-customized automatic neural network architecture generation for massive MIMO CSI feedback," *IEEE Trans. Wireless Commun.*, early access, Jul. 2, 2024, doi: 10.1109/TWC.2024.3418907.

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

[30] Z. Liu, L. Zhang, and Z. Ding, "Exploiting bi-directional channel reciprocity in deep learning for low rate massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 889–892, Jun. 2019.

[31] M. B. Mashhadi, Q. Yang, and D. Gündüz, "Distributed deep convolutional compression for massive MIMO CSI feedback," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2621–2633, Apr. 2021.

[32] J. Guo, C.-K. Wen, and S. Jin, "CANet: Uplink-aided downlink channel acquisition in FDD massive MIMO using deep learning," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 199–214, Jan. 2022.

[33] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2827–2840, Apr. 2020.

[34] S. Jo and J. So, "Adaptive lightweight CNN-based CSI feedback for massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2776–2780, Dec. 2021.

[35] M. Chen, J. Guo, C.-K. Wen, S. Jin, G. Y. Li, and A. Yang, "Deep learning-based implicit CSI feedback in massive MIMO," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 935–950, Feb. 2022.

[36] J. Guo, J. Wang, C.-K. Wen, S. Jin, and G. Y. Li, "Compression and acceleration of neural networks for communications," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 110–117, Aug. 2020.

[37] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Overview of deep learning-based CSI feedback in massive MIMO systems," *IEEE Trans. Commun.*, vol. 70, no. 12, pp. 8017–8045, Dec. 2022.

[38] T. Lin and Y. Zhu, "Beamforming design for large-scale antenna arrays using deep learning," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 103–107, Jan. 2020.

[39] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A deep learning framework for optimization of MISO downlink beamforming," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1866–1880, Mar. 2020.

[40] X. Li and A. Alkhateeb, "Deep learning for direct hybrid precoding in millimeter wave massive MIMO systems," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019, pp. 800–805.

[41] W. Xu, L. Gan, and C. Huang, "A robust deep learning-based beamforming design for RIS-assisted multiuser MISO communications with practical constraints," *IEEE Trans. Cognit. Commun. Netw.*, vol. 8, no. 2, pp. 694–706, Jun. 2022.

[42] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[43] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[44] W. Jin, J. Zhang, C.-K. Wen, and S. Jin, "Model-driven deep learning for hybrid precoding in millimeter wave MU-MIMO system," *IEEE Trans. Commun.*, vol. 71, no. 10, pp. 5862–5876, Oct. 2023.

[45] W. Jin, J. Zhang, C.-K. Wen, S. Jin, X. Li, and S. Han, "Low-complexity joint beamforming for RIS-assisted MU-MISO systems based on model-driven deep learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 6968–6982, Jul. 2024.

[46] Y. He, H. He, C.-K. Wen, and S. Jin, "Model-driven deep learning for massive multiuser MIMO constant envelope precoding," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1835–1839, Nov. 2020.

[47] Q. Hu, Y. Cai, Q. Shi, K. Xu, G. Yu, and Z. Ding, "Iterative algorithm induced deep-unfolding neural networks: Precoding design for multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1394–1410, Feb. 2021.

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

[49] W. Wang and W. Zhang, "Intelligent reflecting surface configurations for smart radio using deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2335–2346, Aug. 2022.

[50] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[51] T. Jiang, H. V. Cheng, and W. Yu, "Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1931–1945, Jul. 2021.

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

[53] N. Farsad and A. Goldsmith, "Neural network detection of data sequences in communication systems," *IEEE Trans. Signal Process.*, vol. 66, no. 21, pp. 5663–5678, Nov. 2018.

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

[55] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 1227–1231.

[56] J. Jalden and B. Ottersten, "The diversity order of the semidefinite relaxation detector," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1406–1422, Apr. 2008.

[57] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proc. IEEE*, vol. 111, no. 5, pp. 465–499, May 2023.

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

[59] X. Zhou, J. Zhang, C.-W. Syu, C.-K. Wen, J. Zhang, and S. Jin, "Model-driven deep learning-based MIMO-OFDM detector: Design, simulation, and experimental results," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5193–5207, Aug. 2022.

[60] A. Kosasih, V. Onasis, V. Miloslavskaya, W. Hardjawana, V. Andrean, and B. Vucetic, "Graph neural network aided MU-MIMO detectors," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2540–2555, Sep. 2022.

[61] 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.

[62] M. Khani, M. Alizadeh, J. Hoydis, and P. Fleming, "Adaptive neural signal detection for massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5635–5648, Aug. 2020.

[63] D. Ha, A. Dai, and Q. V. Le, "HyperNetworks," 2016, *arXiv:1609.09106*.

[64] J. Zhang, C.-K. Wen, and S. Jin, "Adaptive MIMO detector based on hypernetwork: Design, simulation, and experimental test," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 65–81, Jan. 2022.

[65] M. Goutay, F. Ait Aoudia, and J. Hoydis, "Deep hypernetwork-based MIMO detection," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.

[66] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan, "Sampling can be faster than optimization," *Proc. Nat. Acad. Sci.*, vol. 116, no. 42, pp. 20881–20885, Sep. 2019.

[67] N. M. Gowda, S. Krishnamurthy, and A. Belogolovy, "Metropolis–Hastings random walk along the gradient descent direction for MIMO detection," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, QC, Canada, Jun. 2021, pp. 1–7.

[68] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[69] X. Zhou, L. Liang, J. Zhang, C.-K. Wen, and S. Jin, "Gradient-based Markov chain Monte Carlo for MIMO detection," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7566–7581, Jul. 2024.

[70] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87. Cham, Switzerland: Springer, 2003.

[71] T. Gruber, S. Cammerer, J. Hoydis, and S. T. Brink, "On deep learning-based channel decoding," in *Proc. 51st Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2017, pp. 1–6.

[72] E. Nachmani, Y. Be'ery, and D. Burshtein, "Learning to decode linear codes using deep learning," in *Proc. 54th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2016, pp. 341–346.

[73] H. Kim, Y. Jiang, R. Rana, S. Kannan, S. Oh, and P. Viswanath, "Communication algorithms via deep learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2018, pp. 1–17.

[74] 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, Feb. 2018.

[75] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.

[76] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 284–287, Mar. 1974.

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

[78] R. Fritschek, R. F. Schaefer, and G. Wunder, "Deep learning for channel coding via neural mutual information estimation," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[79] M. I. Belghazi et al., "MINE: Mutual information neural estimation," 2018, *arXiv:1801.04062*.

[80] J. Wright and Y. Ma, *High-Dimensional Data Analysis With Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2022.

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

[82] H. Ye, L. Liang, and G. Y. Li, "Circular convolutional auto-encoder for channel coding," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[83] A. Felix, S. Cammerer, S. Dörner, J. Hoydis, and S. ten Brink, "Circular convolutional auto-encoder for channel coding," in *Proc. IEEE SPAWC*, Jul. 2018, pp. 1–5.

[84] T. J. O'Shea, L. Pemula, D. Batra, and T. C. Clancy, "Radio transformer networks: Attention models for learning to synchronize in wireless systems," in *Proc. Conf. Rec. Asilomar Conf. Signals Syst. Comput.*, 2016, pp. 662–666.

[85] A. Felix, S. Cammerer, S. Dörner, J. Hoydis, and S. Ten Brink, "OFDM-autoencoder for end-to-end learning of communications systems," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.

[86] F. Ait Aoudia and J. Hoydis, "End-to-End learning for OFDM: From neural receivers to pilotless communication," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1049–1063, Feb. 2022.

[87] T. J. O'Shea, T. Erpek, and T. C. Clancy, "Physical layer deep learning of encodings for the MIMO fading channel," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017, pp. 76–80.

[88] T. J. O'Shea, T. Erpek, and T. C. Clancy, "Deep learning based MIMO communications," 2017, *arXiv:1707.07980*.

[89] Y. Wang and T. Koike-Akino, "Learning to modulate for non-coherent MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[90] M. A. ElMossallamy, Z. Han, M. Pan, R. Jantti, K. G. Seddik, and G. Y. Li, "Noncoherent MIMO codes construction using autoencoders," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[91] H. Ye, G. Y. Li, and B.-H. Juang, "Deep learning based end-to-end wireless communication systems without pilots," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 3, pp. 702–714, Sep. 2021.

[92] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[93] J. Song, C. Häger, J. Schröder, T. J. O'Shea, E. Agrell, and H. Wymeersch, "Benchmarking and interpreting end-to-end learning of MIMO and multi-user communication," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7287–7298, Sep. 2022.

[94] H. Ye, L. Liang, G. Y. Li, and B.-H. Juang, "Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3133–3143, May 2020.

[95] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[96] F. A. Aoudia and J. Hoydis, "Model-free training of end-to-end communication systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2503–2516, Nov. 2019.

[97] T. Matsumine, T. Koike-Akino, and Y. Wang, "Deep learning-based constellation optimization for physical network coding in two-way relay networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[98] A. Gupta and M. Sellathurai, "End-to-end learning-based amplify-and-forward relay networks using autoencoders," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[99] A. Gupta and M. Sellathurai, "End-to-end learning-based framework for amplify-and-forward relay networks," *IEEE Access*, vol. 9, pp. 81660–81677, 2021.

[100] A. Gupta and M. Sellathurai, "A stacked-autoencoder based end-to-end learning framework for decode-and-forward relay networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 5245–5249.

[101] Y. Lu, P. Cheng, Z. Chen, Y. Li, W. H. Mow, and B. Vucetic, "Deep autoencoder learning for relay-assisted cooperative communication systems," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5471–5488, Sep. 2020.

[102] A. Gupta, K. Singh, and M. Sellathurai, "Time-switching EH-based joint relay selection and resource allocation algorithms for multi-user multi-carrier AF relay networks," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 2, pp. 505–522, Jun. 2019.

[103] *New SI: Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface*, Standard RP-213599, 3rd Generation Partnership Project (3GPP), 2021. [Online]. Available: https://www.3gpp.org/ftp/tsgran/TSGRAN/TSGR94e/Docs/RP-213599.zip

[104] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL, USA: Univ. Illinois Press, 1949.

[105] R. Carnap and Y. Bar-Hillel, *An Outline of a Theory of Semantic Information*. Cambridge, MA, USA: Research Laboratory of Electronics, Massachusetts Institute of Technology, 1952.

[106] J. Bao et al., "Towards a theory of semantic communication," in *Proc. IEEE Netw. Sci. Workshop*, Jun. 2011, pp. 110–117.

[107] J. Liu, W. Zhang, and H. V. Poor, "A rate-distortion framework for characterizing semantic information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 2894–2899.

[108] P. A. Stavrou and M. Kountouris, "The role of fidelity in goal-oriented semantic communication: A rate distortion approach," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 3918–3931, Jul. 2023.

[109] G. Xin and P. Fan, "EXK-SC: A semantic communication model based on information framework expansion and knowledge collision," *Entropy*, vol. 24, no. 12, p. 1842, Dec. 2022.

[110] K. Niu and P. Zhang, "A mathematical theory of semantic communication," 2024, *arXiv:2401.13387*.

[111] Q. Lan et al., "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–371, Dec. 2021.

[112] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[113] B. Zhang, Z. Qin, and G. Y. Li, "Semantic-aware image compressed sensing," in *Proc. IEEE 33rd Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2023, pp. 1–6.

[114] B. Chen and J. Zhang, "Content-aware scalable deep compressed sensing," *IEEE Trans. Image Process.*, vol. 31, pp. 5412–5426, 2022.

[115] B. Zhang, Z. Qin, and G. Y. Li, "Compression ratio learning and semantic communications for video imaging," *IEEE J. Sel. Topics Signal Process.*, early access, May 27, 2024, doi: 10.1109/JSTSP2024.3405853.

[116] B. Zhang, Z. Qin, Y. Guo, and G. Ye Li, "Semantic sensing and communications for ultimate extended reality," 2022, *arXiv:2212.08533*.

[117] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2731–2735.

[118] A. Kosta, N. Pappas, and V. Angelakis, *Age Information: A New Concept, Metric, and Tool*. Norwell, MA, USA: Now Foundations and Trends, 2017.

[119] C. Kam, S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides, "On the age of information with packet deadlines," *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6419–6428, Sep. 2018.

[120] B. Wang, S. Feng, and J. Yang, "To skip or to switch? Minimizing age of information under link capacity constraint," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.

[121] S. K. Kaul and R. D. Yates, "Age of information: Updates with priority," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 2644–2648.

[122] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "The cost of delay in status updates and their value: Non-linear ageing," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4905–4918, Aug. 2020.

[123] Y. Sun, E. Uysal-Biyikoglu, R. D. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7492–7508,

Nov. 2017.
[124] Z. Wang, M.-A. Badiu, and J. P. Coon, "A value of information framework for latent variable models," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.
[125] Y. Sun and B. Cyr, "Information aging through queues: A mutual information perspective," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Kalamata, Greece, Jun. 2018, pp. 1–5.
[126] C. Kam, S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides, "Towards an effective age of information: Remote estimation of a Markov source," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 367–372.
[127] Y. Sun, E. Uysal-Biyikoglu, R. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.
[128] Z. Wang, M.-A. Badiu, and J. P. Coon, "Relationship between age and value of information for a noisy Ornstein–Uhlenbeck process," *Entropy*, vol. 23, no. 8, p. 940, Jul. 2021.
[129] Y. Sun, Y. Polyanskiy, and E. Uysal, "Sampling of the Wiener process for remote estimation over a channel with random delay," *IEEE Trans. Inf. Theory*, vol. 66, no. 2, pp. 1118–1135, Feb. 2020.
[130] T. Z. Ornee and Y. Sun, "Sampling and remote estimation for the ornstein-uhlenbeck process through queues: Age of information and beyond," *IEEE/ACM Trans. Netw.*, vol. 29, no. 5, pp. 1962–1975, Oct. 2021.
[131] Z. Wang, M.-A. Badiu, and J. P. Coon, "A framework for characterizing the value of information in hidden Markov models," *IEEE Trans. Inf. Theory*, vol. 68, no. 8, pp. 5203–5216, Aug. 2022.
[132] A. Maatouk, M. Assaad, and A. Ephremides, "The age of incorrect information: An enabler of semantics-empowered communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2621–2635, Apr. 2023.
[133] N. Pappas and M. Kountouris, "Goal-oriented communication for real-time tracking in autonomous systems," in *Proc. IEEE Int. Conf. Auto. Syst. (ICAS)*, Aug. 2021, pp. 1–5.
[134] A. Li, S. Wu, S. Sun, and J. Cao, "Goal-oriented tensor: Beyond AoI towards semantics-empowered goal-oriented communications," *IEEE Trans. Commun.*, early access, Jun. 19, 2024, doi: 10.1109/TCOMM.2024.3416864.
[135] E. Uysal et al., "Semantic communications in networked systems: A data significance perspective," *IEEE Netw.*, vol. 36, no. 4, pp. 233–240, Jul. 2022.
[136] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 96–102, Jun. 2021.
[137] M. Fresia, F. Peréz-Cruz, H. V. Poor, and S. Verdú, "Joint source and channel coding," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 104–113, Nov. 2010.
[138] N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2326–2330.
[139] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.
[140] H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 142–153, Jan. 2021.
[141] Q. Zhou, R. Li, Z. Zhao, C. Peng, and H. Zhang, "Semantic communication with adaptive universal transformer," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 453–457, Mar. 2022.
[142] K. Lu, R. Li, X. Chen, Z. Zhao, and H. Zhang, "Reinforcement learning-powered semantic communication via semantic similarity," 2021, *arXiv:2108.12121*.
[143] J. Liang, Y. Xiao, Y. Li, G. Shi, and M. Bennis, "Life-long learning for reasoning-based semantic communication," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Seoul, South Korea, May 2022, pp. 271–276.
[144] S. Kadam and D. I. Kim, "Knowledge-aware semantic communication system design," in *Proc. IEEE Int. Conf. Commun.*, May 2023, pp. 6102–6107.
[145] Q. Zhou, R. Li, Z. Zhao, Y. Xiao, and H. Zhang, "Adaptive bit rate control in semantic communication with incremental knowledge-based HARQ," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1076–1089, 2022.
[146] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Deep source-channel coding for sentence semantic transmission with HARQ," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5225–5240, Aug. 2022.
[147] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, Aug. 2021.
[148] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6227–6240, Sep. 2023.
[149] T. Han, Q. Yang, Z. Shi, S. He, and Z. Zhang, "Semantic-preserved communication system for highly efficient speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 245–259, Jan. 2023.
[150] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 4774–4778.
[151] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, "Toward semantic communications: Deep learning-based image semantic coding," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 55–71, Jan. 2023.
[152] H. Zhang, S. Shao, M. Tao, X. Bi, and K. B. Letaief, "Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 170–185, Jan. 2023.
[153] J. Dai et al., "Nonlinear transform source-channel coding for semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2300–2316, Aug. 2022.
[154] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented multi-user semantic communications for VQA," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 553–557, Mar. 2022.
[155] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.
[156] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Sep. 2022.
[157] T.-Y. Tung, D. B. Kurka, M. Jankowski, and D. Gündüz, "DeepJSCC-Q: Constellation constrained deep joint source-channel coding," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 4, pp. 720–731, Dec. 2022.
[158] W. Li, H. Liang, C. Dong, X. Xu, P. Zhang, and K. Liu, "Non-orthogonal multiple access enhanced multi-user semantic communication," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 6, pp. 1438–1453, Dec. 2023.
[159] H. Gao, G. Yu, and Y. Cai, "Adaptive modulation and retransmission scheme for semantic communication systems," *IEEE Trans. Cognit. Commun. Netw.*, vol. 10, no. 1, pp. 150–163, Feb. 2024.
[160] Y. Bo, Y. Duan, S. Shao, and M. Tao, "Learning based joint coding-modulation for digital semantic communication systems," in *Proc. IEEE Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nov. 2022, pp. 1–6.
[161] Y. Bo, Y. Duan, S. Shao, and M. Tao, "Joint coding-modulation for digital semantic communications via variational autoencoder," *IEEE Trans. Commun.*, early access, Apr. 9, 2024, doi: 10.1109/TCOMM.2024.3386577.
[162] J. Park, Y. Oh, S. Kim, and Y.-S. Jeon, "Joint source-channel coding for channel-adaptive digital semantic communications," *IEEE Trans. Cognit. Commun. Netw.*, early access, Apr. 9, 2024, doi: 10.1109/TCOMM.2024.3386577.
[163] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," 2022, *arXiv:2209.07689*.
[164] Y. Sheng, H. Ye, L. Liang, S. Jin, and G. Ye Li, "Semantic communication for cooperative perception based on importance map," 2023, *arXiv:2311.06498*.
[165] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2583–2589.
[166] X. Runsheng, T. Zhengzhong, X. Hao, S. Wei, Z. Bolei, and M. Jiaqi, "CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers," in *Proc. CoRL*, Nov. 2022, pp. 989–1000.
[167] Y. Zhou, J. Xiao, Y. Zhou, and G. Loianno, "Multi-robot collaborative perception with graph neural networks," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2289–2296, Apr. 2022.
[168] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1852–1864, Mar. 2022.
[169] Z. Tu et al., "MAXIM: Multi-axis MLP for image processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5769–5780.
[170] J. Li et al., "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *IEEE Trans. Intell. Vehicles*, vol. 8, pp. 2650–2660, 2023.
[171] H. Xie, Z. Qin, X. Tao, and Z. Han, "Toward intelligent communications: Large model empowered semantic communications," *IEEE Commun. Mag.*, early access, Jul. 15, 2024, doi: 10.1109/MCOM.001.2300807.
[172] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[173] S. Guo, Y. Wang, and P. Zhangz, "Signal shaping for semantic communication systems with a few message candidates," in *Proc. IEEE 96th Veh. Technol. Conf. (VTC-Fall)*, London, U.K., Sep. 2022, pp. 1–5.
[174] S. Guo, Y. Wang, S. Li, and N. Saeed, "Semantic importance-aware communications using pre-trained language models," *IEEE Commun. Lett.*, vol. 27, no. 9, pp. 2328–2332, Sep. 2023.
[175] Z. Qin, J. Ying, D. Yang, H. Wang, and X. Tao, "Computing networks enabled semantic communications," *IEEE Netw.*, vol. 38, no. 2, pp. 122–131, Mar. 2024.
[176] *Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond*, document ITU-R Framework for IMT-2030, Jun. 2023. [Online]. Available: https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2030/Pages/default.aspx
[177] *Proposal of a New Work Item on Architectural Framework for Semantic Communication in IoT and Smart City & Community Service*, document ITU-T SG20 IoT and Smart Cities, Huazhong Univ. Sci. & Technol. (China), 2021. [Online]. Available: https://www.itu.int/md/T17-SG20-C-0902/en
[178] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.
[179] W. Cui, K. Shen, and W. Yu, "Spatial deep learning for wireless scheduling," in *Proc. IEEE Global*

Commun. Conf. (GLOBECOM), Dec. 2018, pp. 1–6.
[180] T. Chen, X. Zhang, M. You, G. Zheng, and S. Lambotharan, "A GNN-based supervised learning framework for resource allocation in wireless IoT networks," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 1712–1724, Feb. 2022.
[181] F. Liang, C. Shen, W. Yu, and F. Wu, "Towards optimal power control via ensembling deep neural networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1760–1776, Mar. 2020.
[182] H. Huang, Y. Lin, G. Gui, H. Gacanin, H. Sari, and F. Adachi, "Regularization strategy aided robust unsupervised learning for wireless resource allocation," *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 9647–9652, Mar. 2023.
[183] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," *Proc. IEEE*, vol. 108, no. 2, pp. 341–356, Feb. 2020.
[184] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, Jun. 2018.
[185] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.
[186] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.
[187] H. Ye, G. Y. Li, and B. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.
[188] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.
[189] L. Wang, H. Ye, L. Liang, and G. Y. Li, "Learn to compress CSI and allocate resources in vehicular networks," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3640–3653, Jun. 2020.
[190] N. Naderializadeh, J. J. Sydir, M. Simsek, and H. Nikopour, "Resource management in wireless networks via multi-agent deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3507–3523, Jun. 2021.
[191] K. Huang, L. Liang, S. Jin, and G. Ye Li, "Meta reinforcement learning for fast spectrum sharing in vehicular networks," 2023, *arXiv:2309.17185*.
[192] W. L. Hamilton, *Graph Representation Learning*. San Rafael, CA, USA: Morgan & Claypool, 2020.
[193] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2017, pp. 1–14.
[194] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–12.
[195] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2019, pp. 1–17.
[196] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "Graph neural networks for scalable radio resource management: Architecture design and theoretical analysis," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 101–115, Jan. 2021.
[197] M. Eisen and A. Ribeiro, "Optimal wireless resource allocation with random edge graph neural networks," *IEEE Trans. Signal Process.*, vol. 68, pp. 2977–2991, 2020.
[198] N. NaderiAlizadeh, M. Eisen, and A. Ribeiro, "State-augmented learnable algorithms for resource management in wireless networks," *IEEE Trans. Signal Process.*, vol. 70, pp. 5898–5912, 2022.
[199] N. NaderiAlizadeh, M. Eisen, and A. Ribeiro, "Learning resilient radio resource management policies with graph neural networks," *IEEE Trans. Signal Process.*, vol. 71, pp. 995–1009, 2023.
[200] Y. Wang et al., "Performance optimization for semantic communications: An attention-based reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2598–2613, Sep. 2022.
[201] T. Liu, C. You, Z. Hu, C. Wu, Y. Gong, and K. Huang, "Semantic-relay-aided text transmission: Placement optimization and bandwidth allocation," in *Proc. IEEE Globecom Workshops*, 2023, pp. 215–220, doi: 10.1109/GCWkshps58843.2023.10464907.
[202] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 7, pp. 1394–1398, Jul. 2022.
[203] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "QoE-aware resource allocation for semantic communication networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2022, pp. 3272–3277.
[204] Z. Ji and Z. Qin, "Energy-efficient task offloading for semantic-aware networks," in *Proc. IEEE Int. Conf. Commun.*, Rome, Italy, Oct. 2023, pp. 3584–3589.
[205] B. Song, H. Sun, W. Pu, S. Liu, and M. Hong, "To supervise or not to supervise: How to effectively learn wireless interference management models?" in *Proc. IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2021, pp. 211–215.

ABOUT THE AUTHORS

Zhijin Qin (Senior Member, IEEE) was with Imperial College London, London, U.K., Lancaster University, Lancaster, U.K., and the Queen Mary University of London, London, from 2016 to 2022. She is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University, Beijing, China. Her research interests include semantic communications and sparse signal processing.

Dr. Qin was a recipient of several awards, such as the 2017 IEEE GLOBECOM Best Paper Award, the 2018 IEEE Signal Processing Society Young Author Best Paper Award, the 2021 IEEE Communications Society Signal Processing for Communications Committee Early Achievement Award, the 2022 IEEE Communications Society Fred W. Ellersick Prize, the 2023 IEEE ICC Best Paper Award, the 2023 IEEE SPCC Best Paper Award, and the 2023 IEEE Signal Processing Society Best Paper Award. She served as a Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) Special Issue on *Semantic Communications* and an Area Editor for IEEE JSAC Series. She also served as the Symposium Co-Chair for IEEE GLOBECOM 2020 and 2021. She serves as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, and an Area Editor for IEEE COMMUNICATIONS LETTERS.

Le Liang (Member, IEEE) received the B.E. degree in information engineering from Southeast University, Nanjing, China, in 2012, the M.A.Sc. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2015, and the Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2018.

From 2019 to 2021, he was a Research Scientist with the Intel Labs, Hillsboro, OR, USA. Since 2021, he has been with the National Mobile Communications Research Laboratory, Southeast University. His main research interests are in wireless communications and machine learning.

Dr. Liang received the Best Paper Award of IEEE/CIC ICCC in 2014. He serves as an Associate Editor for IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING and *China Communications*. He was an Associate Editor of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) Series on *Machine Learning in Communications and Networks* from 2020 to 2022 and an Editor of IEEE COMMUNICATIONS LETTERS from 2019 to 2023. He is a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society.

Zijing Wang (Member, IEEE) received the B.S. and M.S. degrees in information and communication engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2016 and 2019, respectively, and the D.Phil. degree from the University of Oxford, Oxford, U.K., in 2023.

She is currently a Postdoctoral Researcher with Tsinghua University, Beijing. Her research interests fall in the fields of information theory, wireless communications, and semantic communications.

Shi Jin (Fellow, IEEE) received the B.S. degree in communications engineering from Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from Southeast University, Nanjing, in 2007.

From June 2007 to October 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently a Faculty Member with the National Mobile Communications Research Laboratory, Southeast University. His research interests include wireless communications, random matrix theory, and information theory.

Dr. Jin serves as an Area Editor for IEEE TRANSACTIONS ON COMMUNICATIONS and *IET Electronics Letters*. He was an Associate Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and *IET Communications*. He and his coauthors have been awarded the IEEE Communications Society Stephen O. Rice Prize Paper Award in 2011, the IEEE Jack Neubauer Memorial Award in 2023, the IEEE Marconi Prize Paper Award in Wireless Communications in 2024, the IEEE Signal Processing Society Young Author Best Paper Award in 2010, and Best Paper Award in 2022.

Xiaoming Tao (Senior Member, IEEE) received the Ph.D. degree in information and communication systems from Tsinghua University, Beijing, China, in 2008.

She is currently a Professor with the Department of Electronic Engineering, Tsinghua University.

Prof. Tao was a recipient of the National Science Foundation for Outstanding Youth (2017C2019) and many national awards, which include the 2017 China Young Women Scientists Award, 2017 Top Ten Outstanding Scientists and Technologists from China Institute of Electronics, 2017 First Prize of Wu Wen Jun AI Science and Technology Award, 2016 National Award for Technological Invention Progress, and 2015 Science and Technology Award of China Institute of Communications. She was the Workshop General Co-Chair of IEEE INFOCOM 2015 and a Volunteer Leadership of IEEE ICIP 2017. Since 2016, she has been an Editor of the *Journal of Communications and Information Networks* and *China Communications*.

Wen Tong (Fellow, IEEE) was the Nortel Fellow and the Head of the Network Technology Labs, Nortel, Ottawa, ON, Canada. He joined the Wireless Technology Labs, Bell Northern Research, Ottawa, in 1995. He joined Huawei Technologies, Ottawa, in 2009, where he is currently the CTO of the Wireless Technology Labs and the Chief Scientist of Huawei 5G. For the past three decades, he has pioneered fundamental technologies from 1G to 6G wireless and Wi-Fi with more than 500 awarded U.S. patents.

Dr. Tong was a recipient of IEEE ComSoc Industry Innovation Award in 2014, the IEEE ComSoc Distinguished Industry Leader Award, and the R. A. Fessenden Medal. He is a fellow of Canadian Academy of Engineering and Royal Society of Canada. He is a Huawei Fellow.

Geoffrey Ye Li (Fellow, IEEE) was a Professor with Georgia Institute of Technology, Atlanta, GA, USA, for 20 years, and a Principal Technical Staff Member with the AT&T Labs—Research (previous Bell Labs), Middletown, NJ, USA, for five years. He joined Imperial College London, London, U.K., in 2020, where he is currently a Chair Professor. He made fundamental contributions to orthogonal frequency-division multiplexing (OFDM) for wireless communications, established a framework on resource cooperation in wireless networks, and introduced deep learning to communications. In these areas, he has authored around 700 journal articles and conference papers, and holds over 40 granted patents. His publications have been cited over 70 000 times with an H-index of 120.

Dr. Li was elected to IET Fellow for his contributions to signal processing for wireless communications. He won 2024 IEEE Eric E. Sumner Award, 2019 IEEE ComSoc Edwin Howard Armstrong Achievement Award, and several other awards from IEEE Signal Processing, Vehicular Technology, and Communications Societies. He has been listed as a Highly Cited Researcher by Clarivate/Web of Science almost every year.