Semantic Communications for Future Internet: Fundamentals, Applications, and Challenges

Wanting Yang, Hongyang Du, *Graduate Student Member, IEEE*, Zi Qin Liew, Wei Yang Bryan Lim, Zehui Xiong, *Member, IEEE*, Dusit Niyato, *Fellow, IEEE*, Xuefen Chi, Xuemin Shen, *Fellow, IEEE*, and Chunyan Miao, *Senior Member, IEEE*

Abstract—With the increasing demand for intelligent services, the sixth-generation (6G) wireless networks will shift from a traditional architecture that focuses solely on a high transmission rate to a new architecture that is based on the intelligent connection of everything. Semantic communication (SemCom), a revolutionary architecture that integrates user as well as application requirements and the meaning of information into data processing and transmission, is predicted to become a new core paradigm in 6G. While SemCom is expected to progress beyond the classical Shannon paradigm, several obstacles need to be overcome on the way to a SemCom-enabled smart Internet. In this paper, we first highlight the motivations and compelling reasons for SemCom in 6G. Then, we provide an overview of SemCom-related theory development. After that, we introduce three types of SemCom, i.e., semantic-oriented communication, goal-oriented communication, and semantic-aware communication. Following that, we organize the design of the communication system into three dimensions, i.e., semantic information (SI) extraction, SI transmission, and SI metrics. For each dimension, we review existing techniques and discuss their benefits and limitations, as well as the remaining challenges. Then, we introduce the potential applications of SemCom in 6G and portray the vision of future SemCom-empowered network architecture. Finally, we outline future research opportunities. In a nutshell, this paper provides a holistic review of the fundamentals of SemCom, its applications in 6G networks, and the existing challenges and open issues with insights for further in-depth investigations.

Index Terms—Semantic communication, sixth-generation Internet, goal-oriented communication, effectiveness coding, artificial intelligence.

I. INTRODUCTION

A. Motivation

AS WE revisit the development path from the first-generation (1G) to 5G communications, it is evident that the conventional focus has been to optimize *data-oriented* performance metrics, such as communication data rate and bit error probability, while ignoring service-, goal-, or semantic-related metrics. For example, 3G focuses on mobile broadband development. It promises a thousand times the data rate of 2G, whereas 4G unlocks high-speed Internet streaming, delivering a thousand times the data rate of 3G. The motivation for this convention is traced back to the time when Shannon first demonstrated that reliable communication is possible in noisy channels in the classical information theory (CIT) literature [1]. Shannon believed that "the *semantic* aspects of communication should be regarded as *irrelevant* to the engineering problem". The reason is that the meaning of a message can be related to "certain physical and conceptual entities" and that involving the meaning in a mathematical model may affect the generality of the theory [1].

With continuous technological progress following CIT, the advent of the 5G has brought about a breakthrough in communication network design [2], enabling a variety of services from digital twins, edge computing, the Internet of Things (IoT), and more, through the supporting technologies such as ultra-reliable and low-latency communications (URLLC), massive machine type communications (mMTC), and enhanced mobile broadband (eMBB) communications. However, content-centric data-driven communication architecture is increasingly seen as a barrier to providing end-users with services that demand high quality of experience (QoE). This is especially so given that the emerging applications of the 6G will be human-centric, data-, and resource-intensive. One such application is the Metaverse [3], which has been envisioned to be the future Internet. Just as we navigate the Web pages of today's Internet, we will soon explore the virtual worlds of Metaverse through a head-mounted display (HMD) or navigate the augmented physical world through Augmented Reality (AR) glasses. The Metaverse is formed via the synchronization of the virtual and physical worlds,

and the result is that one's actions in the virtual and physical domains will be inextricably linked. Driven by Artificial Intelligence (AI), edge intelligence, virtual and augmented reality, as well as blockchain technology [4], [5], [6], the user-centric QoE metrics required for the successful implementation of Metaverse calls for a rethink of the classical information theory (CIT) driven communication networks, because the massive data from new applications increases significantly the processing latency of conventional communication networks [7]. Specifically, the following difficulties in 6G networks should be addressed:

D1) The emergence of new services, e.g., Metaverse, requires 6G network to support the wireless transmission of massive volumes of data.

D2) 6G applications with a massive number of nodes, e.g., collaborative robots and hyper-intelligent IoT, require fast system responses and reliable, efficient information interaction.

D3) More network resources are consumed for real-time updating of information and analysis of user data to ensure a better service experience.

In response, a novel paradigm known as semantic communication (SemCom) is inspired as a brand new technology in 6G to breakout the "Shannon's trap", which identifies and utilizes the *meaning* of messages during Internet communication. In contrast to conventional data-oriented communication networks, the capacity of which is improved at the cost of system complexity, SemCom enables all communication participants to lighten the network burden via transmitting the most relevant information for the receivers or the goal of communication task after the pre-processing of the data based on the advanced AI technology [8], [9], [10]. The development of SemCom and the advancement of 6G are mutually reinforcing, so as to bring solutions to the three difficulties mentioned above. On the one hand, the availability as well as connectivity of distributed computation and ubiquitous AI networks in 6G will allow SemCom to be feasibly deployed at scale [11]. On the other hand, SemCom overcomes traditional communication constraints and will enable unprecedented improvements in network performance. Thus, with the successful training and development of SemCom, the visions of 6G, e.g., lower latency than 5G and enhanced reliability, can be fully realized. Specifically, the SemCom has the following abilities to address **D1)**, **D2)**, and **D3)**, respectively.

A1) Reduce the wireless data transmission burden of 6G network.

A2) Enhance efficiency of 6G network control and management.

A3) Use the semantics of information to design effective network resource allocation schemes.

However, while the mutually reinforcing convergence properties in 6G and SemCom have attracted the attention of the academic community, there is not yet a comprehensive survey paper that provides a complete overview of the developments, challenges, and future trends for the SemCom-enabled 6G and Beyond networks. As SemCom is a relatively nascent topic, our survey aims to serve as a useful and insightful guide for future studies to researchers and practitioners alike that look

to incorporate SemCom concepts into future communication architectures.

B. Comparisons and Key Contributions

Due to the recent attention in SemCom, some review papers have emerged to address this topic. In [11], the authors state the need to integrate the semantic and effectiveness levels in traditional communications, and suggest an efficient cross-layer design architecture. Moreover, according to the level of communication achieved, they classify the communications beyond Shannon into SemCom and goal-oriented communication. However, for the encoding and decoding in semantic and goal-oriented communications, only highly abstract analytical models are provided. Although they emphasize the importance of machine learning in SemCom, the related technical details, such as the neural networks (NNs) suitable for semantic information (SI) extraction of different data types, are missing. In [8], the authors introduce two SemCom architectures. One is the layer-coupling approach, which is similar to the cross-layer design architecture proposed in [11]. The other is the SplitNet approach, which is used in most existing DL-based SemCom studies. However, similar to [11], most of their work focuses on conceiving the concepts and roadmaps, with relatively less focus on technical details. Moreover, different from [11], they classify SemCom into two types according to the receiver. The encoding process in the two types is called semantic encoding and effectiveness encoding. A few encoding methods for natural language and model/gradient compression are reviewed, while the transmission and decoding processes are not discussed in detail. In [12], the authors in [12] compare the conventional and semantic communication systems and theories, and present SemCom system components, frameworks, and performance metrics. Then, they review recent advancements in DL-enabled SemCom systems for transmitting multimodal data. In their work, they mainly focus on semantic-oriented communication, and the studies about goal-oriented communications are not covered. Moreover, the usage, as well as the benefits and limitations of semantic metrics and the techniques of DL-based SemCom, are not discussed in detail.

In addition, there are several short briefs [9], [10], [13] providing insights from different perspectives with regards to the design of SemCom systems, e.g., FL-enabled SemCom networks [9], goal-oriented SemCom systems [10], and DL-enabled E2E semantic networks [13]. However, these studies review the works from a certain perspective only and do not provide a comprehensive review of the challenges and techniques.

To this end, we aim to provide a comprehensive survey for the implementation of SemCom in 6G, by thoroughly reviewing the existing studies and discussing the 6G applications in potential SemCom-empowered network architecture. In our paper, we classify all communications that take into account the semantic or effectiveness layer into SemCom. Meanwhile, we divide SemCom into three categories: semantic-oriented communication, goal-oriented communication, and semantic-awareness communication. The first two categories belong

TABLE I
A COMPARISON OF CONTRIBUTION BETWEEN RELEVANT SURVEYS AND OUR SURVEY

		Key contributions	Main limitations
[11]	Comprehensive survey	• Highlight 6G use cases, services and related KPIs • Motivate the need of SemCom and suggest an efficient cross-layer design architecture • Motivate a paradigm shift towards semantic and goal-oriented communications • Highlight the importance of learning-based approach in SemCom	• Fail to provide a roadmap for the system design in a cross-layer architecture • Only provide a highly abstract theoretical analytical model for the encoding and decoding in the semantic and goal-oriented communication system, without any practical approaches for specific applications
[8]		• Define semantic and effectiveness encoding according to the destination type • Introduce two SemCom architectures: layer-coupling approach and SplitNet approach • Discuss the potential role of KG technique in SemCom	• Only a few encoding methods for natural language and model/gradient compression are elaborated closely • The communication part of SemCom, such as transmission and decoding processes, is rarely covered
[12]		• Compare the conventional and semantic communication systems and theories • Presents SemCom system components, frameworks, and performance metrics • Discuss recent advancements on DL-enabled SemCom systems for transmitting multimodal data	• Goal-oriented communication studies are not included • Only DL-based SemCom methods are reviewed, and there is no technical discussion about the selection and design of DL models • Simply introduce error-based metrics, without discussing their usage, as well as their advantages and disadvantages

		Key contributions
[9]	Short brief	• Review classical SemCom frameworks • Propose an architecture based on federated edge intelligence for supporting semantic-aware networking
[10]		• Apply SemCom to a communication scenario where the destination is tasked with real-time source reconstruction for the purpose of remote actuation
[13]		• An overview of the latest deep DL-based end-to-end SemCom is given and the open issues that need to be tackled are discussed explicitly
[14]		• Conceive an intelligent semantic communication-empowered ubiquitous-X 6G framework • Present three promising application scenarios for SemCom

		Overlapping Contributions	Distinct Contributions
Our paper	Comprehensive survey	• Classify SemCom into *three* categories, and present their system models and application scenarios • Provide an overview of SemCom related theory development • Discuss the challenges from three design dimensions of SemCom system • Based on the ubiquitous-X 6G framework proposed in [14], we envision the 6G Internet with examples • Identify and outline a series of directions for future research of SemCom	• Review *four* types of SE method in SemCom, and discuss their pros and cons, as well as suitable scenarios • Summary the existing technical DL-based SE models, and discuss their benefits and limitations • Discuss the communication-related techniques and challenges for SemCom • Classify semantic metrics into three basic types, and discuss their limitations and usages • Discuss the potential links between the SemCom and promising 6G applications

to traditional connection-oriented communication, which follows the definitions of two class communication in [11]. The third category of SemCom defined in our paper belongs to task-oriented communication.[1] Meanwhile, to provide a clear roadmap for SemCom implementation, we organize the design of the communication system into three dimensions, i.e., *SI extraction*, *SI transmission*, and *SI metrics*. For each dimension, based on the review of the state-of-the-art methods for the traditional data type, such as text, image, and audio, we summarize the lessons learned about their applicable SemCom categories and scenarios as well as their benefits and limitations. Additionally, we discuss the remaining challenges in each dimension, respectively. Then, we highlight the potential

of SemCom in 6G applications and networks. A series of future research directions are identified. Our discussion aims to shed light on the road ahead for SemCom research.

A more detailed comparison of our paper and existing surveys is listed in Table I. The thorough scope of our survey is presented in Section I-C.

C. Scope of the Survey

The scope of the survey is shown in Fig. 1. In Section II, we first provide a holistic overview of the SemCom-related theory development from semantic information theory to goal-oriented communication theory. Meanwhile, we identify the three categories of SemCom and the corresponding system design. Then, in Sections III–V, we discuss the state-of-the-art techniques and remaining challenges in SI extraction, SI transmission, and SI metrics, respectively. In Section III, we first review four general semantic extraction (SE) methods in Sections III-A–III-D. Among them, DL-based SE and

---

[1]Task-oriented communication here refers to an emerging type of communication in 6G wherein there are multiple explicit or implicit connections between different terminals and network nodes in a proactive or reactive manner [15]. It can be regarded as a counterpart to traditional connection-oriented communication wherein it is easy to tell an explicit pair of source and destination terminals according to the content they intend to communicate [15].
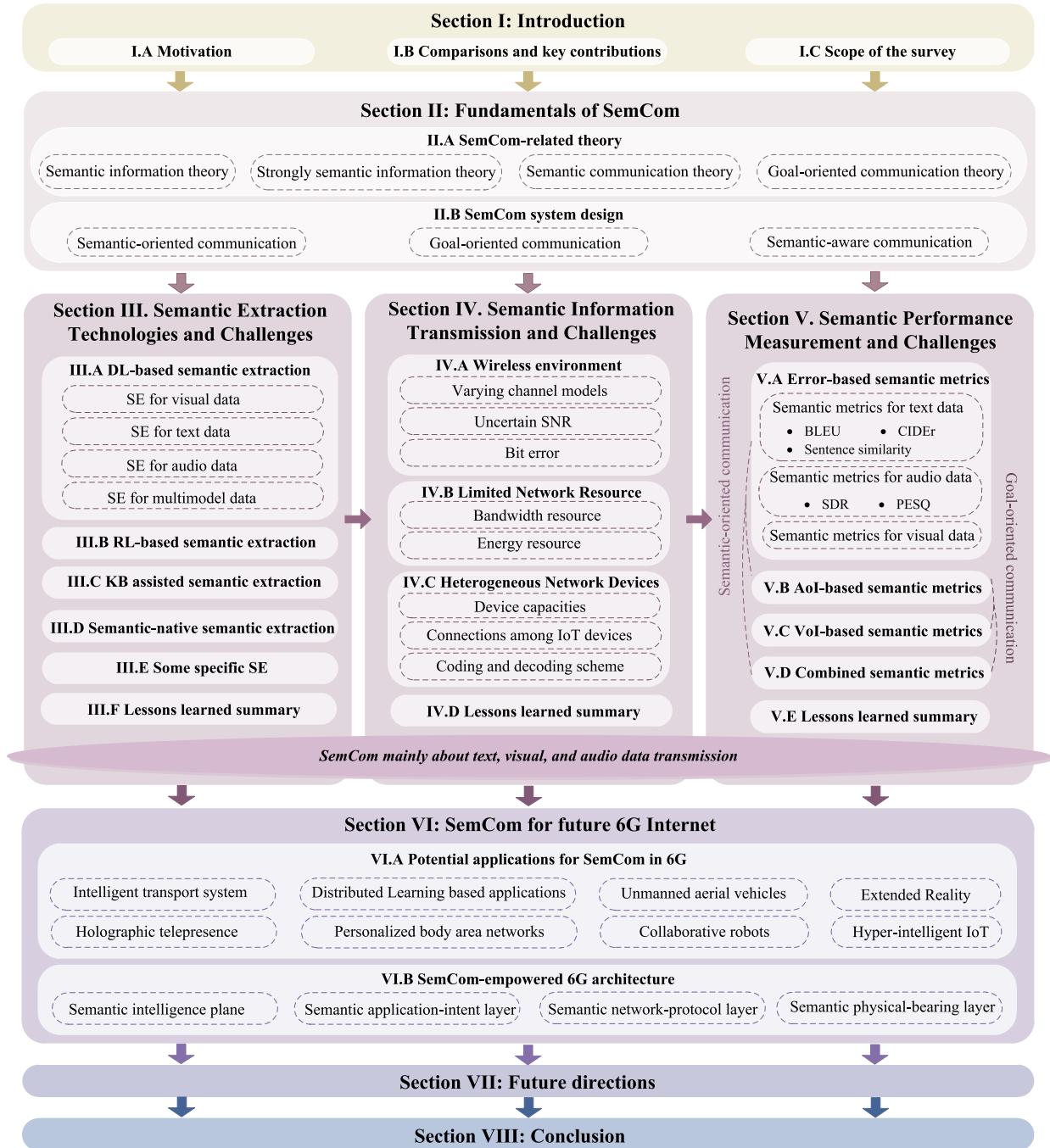
Fig. 1.   Structure of the survey.

RL-based SE mainly apply to semantic-oriented communications, and KB-assisted SE and semantic-native SE can be employed in goal-oriented communications. Meanwhile, in Section III-E, we also take two typical examples to illustrate the role of SE in semantic-aware communications. In Section IV, we focus on the transmission process. We review the techniques and challenges in terms of the wireless environment, limited network resources, and heterogeneous network devices in Sections IV-A–IV-C1, respectively. In Section V, we first discuss three basic types of semantic metrics: error-based semantic metrics, age-of-information (AoI) based semantic

metrics, and value-of-information (VoI) based semantic metrics in Sections V-A–V-C, respectively. Then, in Section V-D, we review existing combined semantic metrics based on the three basic types. In Sections VI-A and VI-B, we aim to highlight the potentials of SemCom in 6G. In Section VI-A, we introduce the potential applications for SemCom in 6G, and discuss the possible roles of SemCom in each application. Furthermore, in Section VI-B, we discuss the implementation of SemCom in 6G Internet with some specific applications based on the ubiquitous-X 6G framework suggested in [14]. At last, we identify and outline a series of directions for future

TABLE II
LIST OF COMMON ABBREVIATIONS

| Abbr.	Description	Abbr.	Description	Abbr.	Description
SI	Semantic information	SE	Semantic extraction	ML	Machine Llearning
DL	Deep learning	RL	Reinforcement learning	KB	knowledge base
KG	Knowledge graph	CR	Compression ratio	CE	Cross entropy
MSE	Mean square error	GAN	Generative adversarial net	DNN	Deep neural network
CNN	Convolutional neural network	CV	Computer vision	NLP	Natural language processing
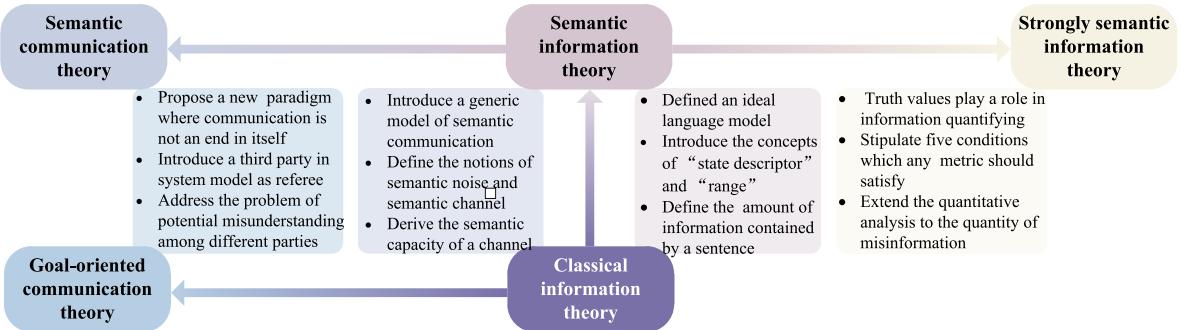
Fig. 2. Development of classical SemCom.

research of SemCom in addition to the challenges ahead in Section VII. Section VIII concludes the survey.

II. FUNDAMENTALS OF SEMCOM

A. SemCom-Related Theory

The concept of *semantics* is initially introduced in the studies on semiotics [16]. In [17], the authors define semiotics as a triple combination of *syntactics*, *semantics*, and *pragmatics*. Syntactics focuses on the interrelation of the formal features for signs (visual and linguistic) without considering the meaning. Semantics specializes in the meaning of the signs at different levels. Pragmatics concentrates on the relationship between the utility of the signs with respect to the user in the sign system [16], [18]. Comparable to the triple-definition for the signs, Weaver [19] identifies three levels of communication as below to further characterize the syntactic, semantic, and pragmatic features of communications [14].

Level A How accurately can the symbols of communication be transmitted? (The technical level.)

Level B How precisely do the transmitted symbols convey the desired meaning? (The semantic level.)

Level C How effectively does the received meaning affect conduct in the desired way? (The effectiveness level.)

Shannon's CIT achieves a big success in deriving a rigorous mathematical theory of communication based on probabilistic models, wherein the concept of *information* is defined as *what can be used to remove uncertainty* and the analysis is based on mutual information in the *entropy* domain. However, CIT focuses only on the technical level. Therefore, some researchers follow Shannon's work and make an attempt to extend it to the semantic level and effectiveness level. The development of classical SemCom has been highlighted in Fig. 2.

1) Semantic Information Theory: The authors in [20], [21] make the first effort in contributing to the *Theory of semantic*

information (TSI) in 1950s. They propose an ideal language model which consists of *n* nouns and *k* adjectives. An arbitrary noun *A* and an arbitrary adjective *a* can be conjugated via a verb. For instance, *Aa* means "*A* is *a*" or "*A* has property *a*". Besides, there exists five connections in the proposed model: $\sim$ (Not), $\vee$ (Or), $\wedge$ (And), $\rightarrow$ (If $\cdots$ then), and $\equiv$ (If and only if). In this sense, many sentences can be generated based on the above conjunctions. Following the definition of *information* in CIT, the amount of SI for a word can be defined as a function of the number of sentences that the word can imply in the considered language model, i.e., the more sentences a word can imply in the language model, the more SI the word contains [16]. Moreover, to quantify the amount of information of a sentence *x*, they further propose a concept named "state descriptor *Z*", which is defined as the conjunction of one noun and one adjective (positive or negative) [16]. The *range* of the valid sentences for sentence *x* is denoted by *R(x)*. By introducing a measurement function for a description denoted by *m(Z)*, where $0 \leq m(Z) \leq 1, \sum m(Z) = 1$ [16], the measurement for a sentences *m(x)* equals the sum of all *m(Z)* within *R(x)* and, similar to Shannon theory of syntactic information, the amount of SI for a sentence can be calculated as its entropy, i.e., $I(i) = -\log_2 m(x)$.

By this point, one significant limitation can be found that it completely ignores the motivation and purpose of the communication at hand. In fact, there is no radical difference between TSI and CIT [16]. In this sense, a false sentence that happens to say much may also be highly informative, as the SI, in such a measurement for SI amount, is not meant to imply truth [20].

2) Strongly Semantic Information Theory: To address the above issue, the study of [22] develops a *Theory of Strongly Semantic Information* (TSSI). Compared to the "weakly" TSI, the truth values play a role in TSSI. Define *f(s)* as the degree of discrepancy of statement *s* from the actual situation. The author in [22] stipulates five conditions that any feasible and satisfactory metric should satisfy as below.

C 1 For a true s that conforms most precisely and accurately to the actual situation, $f(s) = 0$.

C 2 For an s that is made true by every situation, i.e., a tautology, $f(s) = 1$.

C 3 For an s that is made true in no situation, i.e., a contradiction, $f(s) = -1$.

C 4 For a contingently false s, $-1 < f(s) < 0$.

C 5 For a contingently true s that is also made true by situations other than the actual one, $0 < f(s) < 1$.

The details about calculations for the degree of inaccuracy and vacuity in C4 and C5 can be referred to [22]. Based on the degree of discrepancy, the degrees of informativeness for statement s is calculated as $g(s) = 1 - f(s)^2$. Clearly, the more a statement deviates from 0, the more informative it is, which is more in line with the instincts of humans. However, it can only perform the quantitative analysis for the complete class of propositions in logical space and fails to provide rigorous metrics. The authors in [23] further improved this work based on the available works on truthlikeness, which measures the degree of being similar to the truth [24], via extending the quantitative analysis to the semantic concepts of the quantity of misinformation, wherein *SI* is defined as true semantic content and *semantic misinformation* is defined as false semantic content. The related concept of information is rooted in the SI framework using information flow [25]. By employing prior works on truthlikeness for classical systems [26], [27], the proposed SI quantifying method can support a broader range of use cases. However, dealing with non-classical systems is still an open issue. Nevertheless, the transformation of the measurement from uncertainty into message content in [22], [23] has made a milestone step in the development of semantic information theory.

3) Semantic Communication Theory: The authors in [28] initially put forward a *Theory of Semantic Communication* based on SI quantifying method [20], aiming to achieve semantic-level communications. They propose a SemCom model for a basic type source that can just make factual statements in propositional logic. In their model, the source and destination are modeled as a 4-tuple of $<$ world model *W*, background knowledge *K*, inference *I*, and message interpreter $>$. Moreover, the Shannon entropy $H(W)$ is employed to quantify the information amount of the source, i.e., semantic entropy. Furthermore, they consider a finite set of allowed messages *X*, which can be seen as the set of available semantic codes. In this regard, semantic coding is the process of mapping from the observed values of the world model to a specific message. The strategy is a conditional probabilistic distribution $P(X|W)$, and deterministic coding is encoding the observed value *w* into *x* with the highest $P(x|w)$. Furthermore, the relationship between the semantic entropy and message entropy is $H(X) = H(W) + H(X|W) - H(W|X)$, where $H(X|W)$ measures semantic redundancy of coding, and $H(W|X)$ measures semantic ambiguity of the coding [28]. The major difference from CIT is that the SI measure is based on the *logical probabilities* which are determined by the background knowledge and inference, instead of *statistical probabilities*. Secondly, the *side information*, i.e., the destination's prior knowledge about the source, can also be considered in the

coding process to reduce the code length. More detail about encoding method based semantic entropy can be found in [29]. Moreover, by denoting the received message by *y*, the semantic channel can be characterized by the distribution of $p(y|x)$. Furthermore, different from the CIT, the *semantic channel capacity* for the discrete memoryless channel is dependent on three elements. The first one is the mutual information $I(X;Y)$ between *X* and *Y*, which is also the channel capacity for CIT. The second one is the degree of semantic ambiguity introduced in semantic encoding with K_s and I_s, i.e., $H_{K_s,I_s}(W|X)$. The last one is the *average* logical information of the received messages, which is determined by K_d and I_d, i.e., $\overline{H_{K_d,I_d}(Y)}$. If K_s (I_s) and K_d (I_d) do not match, excessive *semantic noise* is generated. For deriving the limit of semantic channel capacity, the authors in [28] simplify the model by assuming $K_s = K_d$ and $I_s = I_d$, and the upper bound is given as $C = \sup_{P(X|W)} \{I(X;Y) - H(W|X) + \overline{H(Y)}\}$.

These works can be seen as an initial but pioneering exploration of SemCom. However, it is only a model-theoretical framework, which could be unrealistic for practical communication scenarios. More relevantly, in the above work, the information amount is merely quantified based on classical Shannon entropy, which has no concerns with the essence of SI, the meaning factor, and thus is inconsistent with the original vision of SemCom.

4) Goal-Oriented Communication Theory: In contrast to the study of [20], which focuses on the extension of Shannon's CIT, the study of [30] focuses on the development of the classical communication system model. In both Shannon's classical system model and the SemCom model [28], all the communication parties need to have a common language or background. Faced with the increasing interaction among the diversified computers at that time, the study of [30] tried to make an attempt to make progress on the *universal SemCom*, wherein the communication parties are expected to obtain a common understanding via learning each other's behavior without any prior common language. In [30], the authors focus on a particular communication model between Alice and Bob, where Bob is a probabilistic polynomial time bounded interactive machine with the goal of solving a hard computational problem, and Alice has unbounded computational power and is willing to help Bob. Meanwhile, they speak different languages and expect to discuss via some binary channel. To solve this problem, the authors introduce a "trusted third party", which knows both languages of Alice and Bob and can give finite encoding rules to translate for this discussion. The results of the theoretical analysis show that Alice can help Bob if and only if the problem that Bob wants to solve is in PSPACE [31], (i.e., the solutions to the problem are verifiable for Bob). Although the above assertions are in a restricted setting, it first highlights that communication is not an end in itself, but rather a means to achieve some general goals among the communicating parties.

Based on the formulated communication model in [30], the authors make an extension to study the general goals of communication and first propose the conception of "*goal-oriented communication*" in [32]. In this work, they clarify two definitions related to the *goals* in communication. One is *meta-goal*,

which captures the intents of communicating agents, and the other is *syntactic goal*, which captures effects that can be observed by an agent. The results show that the meta-goals with different syntactic versions are also achievable, i.e., two communicators do not (necessarily) share a common language under some technical conditions. Based on this, a novel architecture could be enabled for the communication among multiple agents with different protocols, wherein the trusted party called "interpreter" played an essential role. It should be noted that in the above communication model, while the communication parties did not share a common language, they are assumed to be *"sufficiently helpful"*. In [33], the authors further generalize the above work. At this level of generality, misunderstandings might occur between the communication parties. In this work, the third party is renamed as *referee*, which hypothetically monitors the conversation between communication parties and assesses whether or not the goal has been achieved. Moreover, they identify and highlight a new concept called *sensing*, which captures the communication parties' ability to simulate the referee's assessment. Based on the concept, they propose a design principle for communication systems, which could achieve polynomial overhead in the description length of the desired strategy. In [34], the authors claim that the construction of universal users from such sensing functions is equivalent to the design of an *on-line learning* algorithm. However, the above works mostly rely on the "*try and check*" paradigm. They can only provide guidance on the design of the protocol or strategy for the simplistic communication system, such as the conversation between a server and a printer. Although the series of works only focus on a mathematical theory of goal-oriented communication for traditional computer communication models, the system model proposed in their work has laid the foundation for modern goal-oriented communication.

Even though the studies about SemCom theory are still facing simple logic language models or the server-printer communication scenario, the exploration of SemCom systems is not constrained by this. Thanks to the advanced AI technologies, there has been a surge in the research on improving the design of SemCom systems to meet more practical scenarios. In this section, we first focus on three typical generic SemCom system models and the concerns in the performance evaluation that are different compared to traditional communications. Then, in the following three sections, we organize the design of the communication system into three dimensions, i.e., *SI extraction*, *SI transmission*, and *SI metrics*. The available technologies and the remaining challenges are reviewed and discussed in Sections III–V, respectively.

B. SemCom System Design

As mentioned at the beginning of Section II-A, the traditional communication systems only focus on the first level (i.e., technical level) in the three levels of communications identified by Weaver and Shannon. SemCom is proposed to integrate the remaining two higher levels into the design of communication systems. In our survey, we classify the existing works for SemCom into three categories according to the level and the role of the SemCom, i.e., semantic-oriented communication,
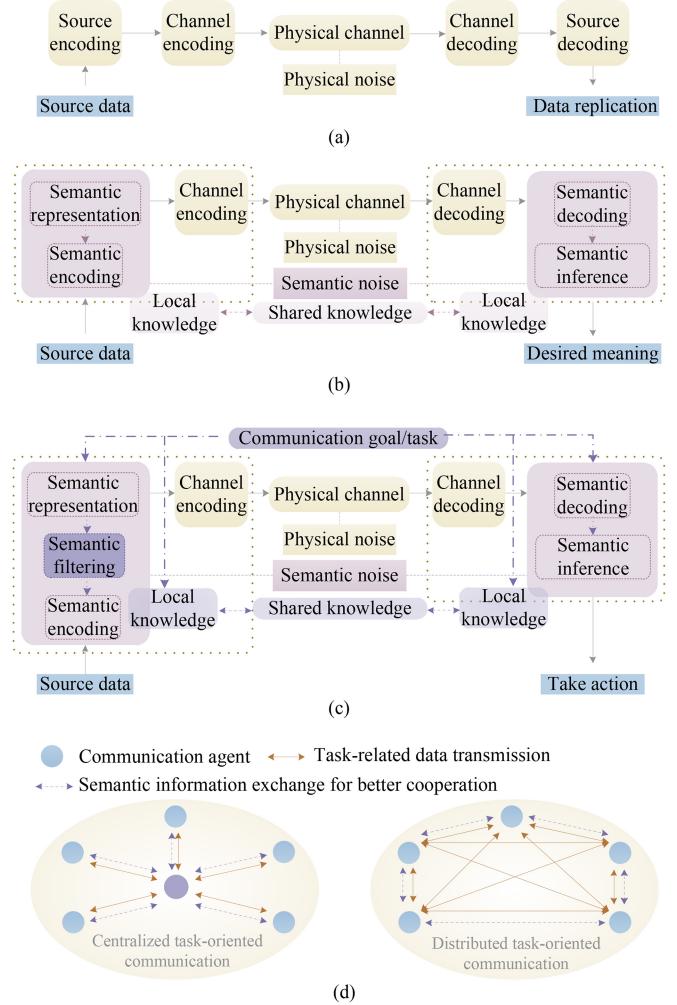

Fig. 3. General communication system models. (a) system model for classical communications; (b) system model for semantic-oriented communications; (c) system model for goal-oriented communications; (d) scenario model for semantic-aware communications.

goal-oriented communication, and semantic-aware communication. The comparison of the communication models can be shown in Fig. 3. Next, we describe the details of the general system models for SemCom.

1) Semantic-Oriented Communications: Different from the content-blind classical communication systems, what matters in semantic-oriented communication design is the accuracy of the semantic content of source data, instead of the average information associated with the possibilities of source data that can be emitted by a source [11]. As such, as shown in Fig. 3(a) and Fig. 3(b), the main changes in the semantic-oriented communication system lie in the data processing phase before sending and after receiving. The traditional source encoding is designed to find a method to convert source data into shortcodes. Meanwhile, since the transmitted message is blind to the underlying meaning, a good source encoding method means that it can cope with more possibilities of source data, which is in line with the information quantification in CIT. However, in SemCom, the definition of "information" needs to be modified.
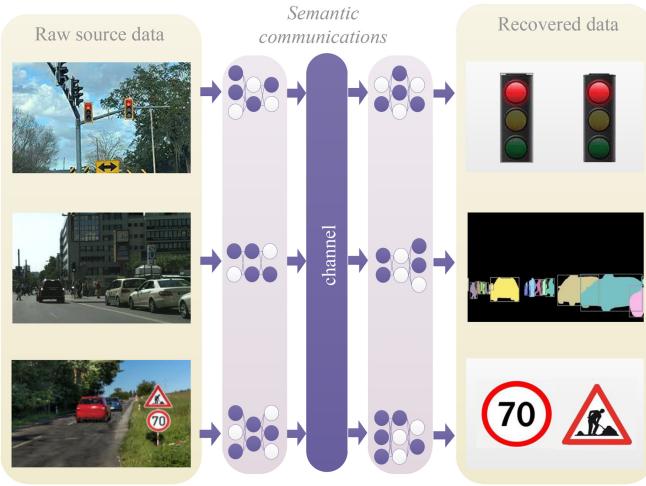
Fig. 4. Example of semantic-oriented communication.

Fig. 5. Relationship among syntactic information, semantic information, and pragmatic information [16].

As stated in [35], information is the commodity capable of yielding knowledge, and the information that a signal carries is what we can learn from it. In this sense, a module of *semantic representation* is introduced before encoding in SemCom, which is responsible for capturing core information embedded in source data and filtering out the unnecessary redundancy information. In many studies, the function of *semantic representation* and *semantic encoding* are integrated into one module called semantic encoding, which jointly plays a similar role to source the shortcodes in traditional communications. Similarly, the combined role of *semantic inference* and *semantic decoding* is equivalent to that of source decoding. In general SemCom scenarios, decoding is the inverse process of encoding, which is based on the AI technologies, such as Transformer and auto-encoder which are powerful with prior knowledge. Since the objective of SemCom is to enable the receiver to successfully infer SI, we regard the joint semantic encoding and decoding process as SE. Take an example of image transmission for the transportation system in Fig. 4. In traditional communications with the goal of image replica, the compressed image based on a content-blind approach is expected to preserve all the details of the original image. In contrast, in semantic-oriented communication, the SE process can filter out irrelevant image details for different tasks before transmission by performing the appropriate image processing techniques, thereby relieving the network burden without compromising the system's performance.

Moreover, as with human conversation, effective conversation requires common knowledge of each other's language and culture. In SemCom, the *local knowledge* of the communication parties needs to be shared in real-time to ensure that the processes of understanding and inference can be well matched for all the source data. If the local knowledge fails to match, *semantic noise* generates, which leads to semantic ambiguity, even in the absence of syntactic errors during the transmission in physical transmission.

2) Goal-Oriented Communication: Recall the triple definition for signs, i.e., syntactics, semantics, and pragmatics. In the above semantic-oriented communications, SE mainly focuses
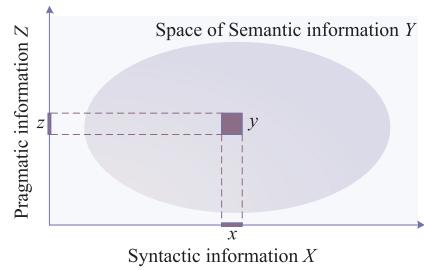
on SI, whereas in goal-oriented communication, it is necessary to capture *pragmatic information*. In [16], the authors illustrate the mutual relationship among the three information types. As shown in Fig. 5, the pragmatic information can be treated as part of all the SI that can be conveyed by syntactic information that can be treated as raw data generated by the sources in communications. It is only relevant to a certain goal of communication. Therefore, we also refer to pragmatic information as SI for the sake of a concise presentation.

Therefore, as shown in Fig. 3(c), the main difference between SE in goal-oriented communication and semantic-oriented communication lies in that the goal of the communication task needs to play an important role in SE as well. Meanwhile, the communication goals also have to be involved in the local knowledge of the communication parties, which helps further filter out the irrelevant SI in each transmission, when the communication goal changes frequently. Take the image transmission as an example. The features (i.e., SI) of the images required for different tasks, such as classification based on different attributes, detection of different targets, or simply replication, are different. Thus, in a transmission system with multiple tasks, perhaps only a local feature of an image needs to be transmitted each time for a certain task in goal-oriented communication. In contrast, in semantic-oriented communication, due to the non-goal-specific SE, the extracted SI should include the features for all the possible tasks, which inevitably results in information redundancy and waste of resources during transmission.

By comparing Fig. 3(b) and Fig. 3(c), another difference is the output of the SemCom system. For semantic-oriented communication, the output of the system is the recovered meaning of the transmitted message. Then, the receiver takes the next step according to the meaning of the received message, but this process is not considered in the design of the communication system. In contrast, the output of the goal-oriented communication system is a direct action to be performed. Recall the semantic-oriented communication example of image transmission in the transportation scenario as shown in Fig. 4, the results inferred by the receiver may be a combination of feature maps similar to the ones on the right of Fig. 4. In contrast, in goal-oriented communication, the output of the inference module is the action execution instruction, such as acceleration, braking, the angle for the steering wheel, and flashing headlights, to respond to pedestrians, roadblocks, and traffic signal status changes. In summary, goal-oriented communication focuses on the effective level and aims to accomplish

the task in the desired way given limited network resources, rather than the SI accuracy focused on the semantic level in semantic-oriented communication.

Moreover, similar to semantic-oriented communication, the local knowledge and communication goal of all the communication parties need to be maintained to be consistent, otherwise, the resulting semantic noise may cause the task to fail.

3) Semantic-Aware Communication: As shown in Fig. 3(b) and Fig. 3(c), both semantic-oriented and goal-oriented communications establish a connection between two specific agents. They belong to the traditional connection-oriented communication, wherein it is easy to tell an explicit pair of source and destination agents according to the content that they intend to communicate [15]. In contrast, semantic-aware communication in this survey refers to the SemCom that plays a role in task-oriented communications, such as automatic driving and unmanned aerial vehicle swarm.

In task-oriented communication, multiple agents cooperate to accomplish a task in a centralized or distributed way as shown in Fig. 3(d). The semantic-aware communication in the task establishes multiple explicit or implicit connections among different terminals in a proactive or reactive manner to enhance the knowledge among agents. In other words, semantic-aware communications can be treated as a kind of "overhead" for the task for better collaboration to facilitate the completion of tasks. In semantic-aware communication, the SI here is obtained by analyzing the agent behavior and the current environment in performing the task, instead of extracting it from a data source. For instance, in autonomous driving, the SI can represent the risk of a collision between two vehicles, which is determined jointly by the vehicle location and kinematic information, traffic density, road conditions, traffic light deployment, etc. Moreover, the SI can also be a description of the views captured by a series of successive cars when they pass the exit of a walled subdivision. After aggregating the SI, a semantic representation of a continuous view over the subdivision exit can be obtained, which can facilitate exit monitoring and activity tracking along the exit lane within the subdivision [36]. In semantic-aware communication, there may be no explicit transceivers or a complete pairwise semantic encoding and decoding processes. Therefore, there is not yet a general system model for semantic-aware communication.

III. SEMANTIC EXTRACTION TECHNOLOGIES AND CHALLENGES

As discussed in Section I, the achieved transmission rates in conventional communication systems are approaching the Shannon limit and the remaining available spectrum resources are becoming increasingly scarce [37], [38]. The key to SemCom being pushed forward to address the bandwidth bottlenecks lies in that it converts the *transmission-before-understanding* communication paradigm to the *understanding-before-transmission* communication paradigm. In this way, SE can be integrated into the communication model to achieve SemCom [8], [14], which allows only the information of interest to the receiver for transmission, rather than raw data, thereby alleviating bandwidth pressure and enhancing privacy

preservation by reducing and hiding the redundant data to be exchanged.

In fact, SE is not a brand-new topic, but it has been evolving [39], [40]. Some comparable works have been explored in other research fields, such as *semantic segmentation* in computer vision, which is used to cluster parts of images together which belong to the same object class [41], *semantic computing*, which addresses the derivation and matching of the semantics of computational content and that of user intentions to retrieve, use, manipulate, or even create the content [42], and *semantic Web*, which can be considered as a knowledge graph formed by combining the linked data with intelligent content and is widely used in recommendation systems to facilitate intelligent and integrated user experience [43], [44]. Compared to [41], [42], [43], [44], *SemCom* is another key field for SE. In this field, all the communication parties have to be highly aligned in semantic representation and interpretation, which imposes challenges for SE. In addition, information in the 6G communication system features a strong time-sensitive nature and is highly demanding in terms of accuracy [45], which is also a stringent requirement not found in other fields.

Hence, in the following, we merely concentrate on SE methods in SemCom. In Sections III-A and III-B, we introduce two general SE methods for semantic-oriented communications. Then, in Sections III-C and III-D, two general SE methods for goal-oriented communications are presented, where the communication goals are integrated into the SE. At last, we take two examples to illustrate the SE in semantic-aware communication in Section III-E.

A. DL-Based SE

Following the success of DL in the individual block optimization in the physical layer [46], [47], [48], [49], [50], DL-based end-to-end communication systems have emerged as another potential direction to outperform the conventional communication structure in block error rate (BLER) and BER performance [51], [52], [53]. Inspired by this, some researchers further introduce the DL-enabled method in fields of computer vision (CV) [54], [55], [56], (natural language processing) NLP [57], [58], [59] and speech processing [60], [61], [62] into end-to-end communication system as SE approaches, which pioneered the *modern* SemCom study [63].

1) SE for Visual Data: Due to the large volume of image data, the authors in [64] first focus on an image transmission scenario, where an IoT device transmits images to the server to perform recognition. The IoT device maintains a direct, point-to-point wireless link to the server. Different from the conventional communication models where multiple modules are cascaded, they propose a DL-constructed *joint transmission-recognition scheme* (JTRS) with the design metric of *recognition accuracy*. In the designed scheme, the ResNet architecture [65] is employed due to its favorable performance and few parameters. In order to complete feature extraction before transmission, the deep neural network (DNN) of ResNet is split into two parts. The first few layers function as a feature extractor (i.e., semantic extractor) at the transmitter, and the rest of the layers serve as a recognizer at

TABLE III
PERFORMANCE COMPARISON FOR JTRS [64](CR = 0.04)

	JTRS	JPEG	CS-DR	CS-R	
runtime	7e-5 s	7e-2 s	1e-3 s	4e-1 s	analog
accuracy	0.9	0.5	0.1	0.14	
SNR_{min}	-4 dB	5 dB	15 dB	15 dB	
runtime	1e-2 s	9e-3 s	1e-3 s	4e-1 s	digital
accuracy	0.9	0.7	0.47	0.14	
$(E_b/N_0)_{min}$	0 dB	5 dB	5 dB	4 dB	

the receiver. Furthermore, to achieve the adaptive semantic extraction in noisy channels, the joint semantic-channel coding (JSCC) is implemented by using the DNN as channel encoders and decoders, which is discussed in detail in Section IV .

To demonstrate the effectiveness of the DNN-constructed JTRS, the authors in [64] compare the scheme with three other cascaded compression-and-recognition schemes given the similar compression ratio (CR) of 0.04. The three baseline schemes are JPEG-compressed scheme (JPEG), compressed sensing with direct recognition (CS-DR), and compressed sensing with reconstruction (CS-R). Table III shows the complexity (in terms of runtime), the highest recognition accuracy, and the corresponding channel condition thresholds of the four schemes for digital and analog transmissions, respectively.[2] From Table III, we can see that, due to the excessively low CR, the schemes of CS-DR and CS-R have almost lost the capability to do recognition even under favorable channel conditions. Among the three baselines, only the scheme of JPEG can achieve the accuracy of more than 50% in digital transmission with LDPC codes. In contrast, the proposed JTRS can achieve the accuracy of up to 0.9 under poor channel conditions in both analog and digital transmission. Surprisingly, JTRS performs better in analog transmission than in digital transmission. More encouragingly, due to the lack of quantification and bit conversion process before modulation and transmission, the runtime of JTRS in the analog transmission is far lower than those in other methods, which means that the DL-based SemCom has an inherent advantage in low–latency communications.

However, this scheme is only designed to operate under a specific SNR level. When channel conditions change, the SE model needs to be retrained or refined, which introduces considerable additional overhead. In the traditional communication system, the general source encoders and channel encoders can achieve an adaptive CR and the channel coding rate according to the SNR to achieve optimal performance given the limited bandwidth. To fill this gap between SemCom and traditional communication, the authors in [66] consider a point-to-point image transmission system with SNR feedback. They integrated the *attention mechanism* [54], [56] that is widely used in CV into SE. The attention mechanism adopts an additional neural network to rigidly select certain features or assign different weights to different features in the original

neural network. In their proposed design, the joint semantic-channel encoding is performed by a single network, which consists of two modules: a feature extraction (FE) module and an attention feature (AF) module. The FE module is used to learn features from the input images. The AF module then takes the output of the FE module and SNR as its input and produces a sequence of scaling parameters. The product of the outputs of the feature learning module and the attention feature module can be seen as a filtered version of the feature learning module output. The decoder is similarly designed. In the simulation, the authors compare the performance of the attention-based DL JSCC scheme trained under the uniform distribution of SNR from 0 dB and 20 dB and five basic DL based JSCC schemes trained at the SNR of 1 dB, 4 dB, 7 dB, 13 dB, and 19 dB, respectively. From the results, the peak signal-to-noise ratio (PSNR) curve achieved by the proposed scheme can be seen as the upper envelope of the other PSNR curves of the baseline scheme trained at different SNRs, which demonstrate the higher robustness, versatility, and adaptability to the wide range of SNR of the attention-based approach.

The above two works focus on image recognition and image recovery, respectively. The authors in [67] focus on the applications of image classification against semantic noise. Exploiting the heavy spatial redundancy of image data, they propose a resource-efficient SE model with an asymmetric encoder-decoder architecture. The encoder employs a masked autoencoder (MAE) with vision Transformer (ViT) architecture [68]. The MAE can reconstruct an image from partial observations. Specifically, in the proposed architecture, a portion of the original image is masked and disregarded first. Then, the unmasked portion is embedded with the information about their position in the original image, which then goes to Transformer blocks to extract the image features [68]. As the encoder only needs to process the portion of unmasked patches, which significantly reduces the memory consumption. On the contrary, the input of the decoder is the full set of tokens consisting of encoded features of unmasked patches and the masked tokens, which is a shared and learned vector suggesting the presence of the patches that are to be predicted [67]. Moreover, different from the above end-to-end SE model, the decoder can be designed independently of the encoder, as the decoder is only used to perform the image reconstruction task, which allows for greater flexibility in the system design.

Meanwhile, MAE can also defend against malicious attackers, i.e., by adding semantic noise to images. Since MAE randomly masks partial patches of the image during the encoding process, the impact of semantic noise added in the patches of the image can be eliminated to some extent [67]. In addition, to further strengthen the resistance to malicious attacks, the authors in [67] propose a codebook for encoded feature representation, which consists of multiple discrete basis vectors trained together with the encoder and decoder parameters. Based on the well-trained codebook, the continuous encoded features output by the encoding neural networks are mapped into the discrete indices of basis vectors by a nearest neighbour search [67]. Hence, the distortion caused by semantic noise can be corrected with a high probability during discrete

---

[2]Here the analog transmission means that the data values are directly used to modulate the signal without going through the steps of quantization. The digital transmission means the data values need to be quantized and converted to bits before modulation and transmission.

TABLE IV
COMPARISON OF THE NUMBER OF SYMBOLS
FOR AN ENCODED IMAGE [67]

	JPEG+LDPC	MAE	Ratio
total symbols	20432	196	0.95%

representation at the transmitter, which greatly enhances the robustness of the communication. In the training process, adversarial learning is employed where the semantic noise is generated by fast gradient sign method.

Since MAE effectively reduces the spatial redundancy of images, the number of symbols of an image to be transmitted is 0.95% of that of an image encoded by the traditional scheme (JPEG + LDPC), as shown in Table IV. Due to the effective SE, the MAE scheme can achieve a classification accuracy of 0.6 even with an SNR of −6 dB. In contrast, the classification accuracy of the traditional scheme (JPEG + LDPC) is close to zero with SNR ranging from −6 dB to 6 dB due to the limited bandwidth. Only when the SNR reaches 14 dB does the traditional scheme achieve a classification accuracy of 0.6. However, the size of the JEPG images used for training and testing is relatively small (5108 bytes). Considering the complexity of the training process, the feasibility and the effectiveness of this scheme have to be further verified. Nevertheless, this result fully demonstrates the importance of SemCom in improving communication performance by reducing the data transmission burden with effective SE.

2) SE for Text Data: Inspired by the success of DL in NLP such as machine translation, the authors in [69] pioneer the implementation of SemCom for text transmission. They consider a simple system model, where a transmitter sends sentences to a receiver using the limited number of bits over an erasure channel. In the proposed scheme, the words are first represented by an embedding vector using GloVe [70], which is the pre-trained lookup table available for extracting SI. Then, motivated by the success of the sequence-to-sequence learning framework in machine translation [57], [71], the long short-term memory (LSTM)-based encoder and decoder are employed, wherein the embedding vector of the previously estimated word is taken as the input for the next step and the beam search algorithm is used to find the most likely sequences of words [71], [72]. In this sense, the SI can be embedded into the sentence recovery. Compared with Gzip and Huffman, LSTM-based SE achieves the lowest word error rate for a given coding length with a high bit-drop rate. Meanwhile, under a certain bit-drop rate, due to the effectiveness of extracted information, the superiority of LSTM-based SE becomes more remarkable as the length of the sentence increases. However, the word representation models like Glove or Word2Vec [73] only capture the relationship among words and fail to describe syntax information [37]. Therefore, the proposed method can only describe the probability of a certain word coming after another in a sentence, which makes it hard to deal with complex sentences.

In the face of the above challenge, a newly proposed architecture called Transformer has attracted a great deal of attention, as it can extract both the SI and syntax from the whole sentences effectively [37]. The Transformer network is combined with *multi-head attention* mechanisms, which allows it to extract multiple characteristics of input sentences in parallel [38]. Therefore, compared with the recurrent neural network (RNN)-based architectures, such as LSTM, the Transformer network achieves lower computational complexity and more parallelizable computations while learning long-range dependencies in input sentences [38], [59]. Hence, in the recent works [37], [74], the Transformer networks replace the RNN networks, and the channel models are extended to additive white Gaussian noise (AWGN) channels and fading channels. In their work, more expert semantic metrics, such as BLEU and sentence similarity, (which are introduced in Section V) are employed to measure the SemCom performance. The superiority of the scheme in terms of semantic metrics under the low SNR region demonstrates the effectiveness of Transformer in SE for text data.

However, the standard Transformer has a fixed attention structure, which makes it treat all inputs indifferently and limits its adaptability in the learning process. In fact, in a sentence processing system, some words or phrases are more likely to cause semantic ambiguity due to polysemy or noise interference. With this in mind, the authors in [75] propose a flexible SE approach based on Universal Transformer (UT) [76], by introducing an adaptive circulation mechanism in the Transformer to break the original fixed structure. Compared to the standard Transformer, UT is integrated with the Adaptive Computation Time (ACT) model [77]. The ACT model dynamically adjusts the number of computational steps required to process each input symbol in the standard RNN, according to a halting probability predicted at each step. Such a dynamic halting mechanism allows UT-based SE to give loop play to its own circulation mechanism for each input symbol (i.e., per-symbol self-attentive RNN) and flexibly respond to different SI and varying physical channels through different cycles.

In [75], the authors compare the performance in terms of BLEU of both the SemCom schemes with UT-based SE approach and classical Transformer-based SE approach, with the traditional source coding and channel coding cascaded schemes with fixed-length coding (5-bit) for source coding and Turbo coding or Reed-Solomon coding for channel coding. For both traditional schemes, the BLEU score keeps staying pretty low over a wide range of SNR, and is only significantly improved when the SNR is increased above 15 dB. In contrast, both SemCom schemes achieve remarkably higher BLEU scores under a variety of changing channel conditions. Specifically, since the adaptive circulation mechanism facilitates a more accurate capture of SI, the UT-based algorithm consistently scores higher than the Transformer-based algorithm over the full SNR region.

3) SE for Audio Data: With the success of E2E SemCom focusing on images and text, the authors in [78] further investigate the SemCom for the audio signal. In [78], the authors design an audio SE based on a DL-based NLP model named Wav2Vec [79]. The semantic encoder consists of two cascaded convolutional neural networks (CNNs), called FE and feature aggregator (FA), respectively. The FE is responsible for extracting the rough audio features from the raw audio vector, and the FA is responsible for combining the rough audio features

into a higher-level latent variable that contains semantic relations among contextual audio features [79]. Accordingly, the semantic decoder is also based on Wav2Vec architecture, which consists of two symmetrical CNNs to the encoder called feature decomposer (FD) audio generator (AG), respectively. This scheme can reduce the MSE to below 2e-4, when the SNR is above 0 dB. However, due to the simplicity of SE model, the extracted SI is somewhat limited. As the SNR increases, there is no obvious downward trend in MSE. Moreover, similar to the LSTM model employed in image SE, the SE model is trained under AWGN channels with a fixed channel coefficient, which makes it challenging to guarantee decent performance under more complicated channel conditions.

At the same time, similar to the evolution of text semantic encoder, the authors in [80], [81] further integrate the *attention mechanism* named SE-ResNet into SE, and the encoder and decoder are constructed by one or multiple sequentially connected SE-ResNet modules. The term "SE" in "SE-ResNet" represents a squeeze-and-excitation network, which is treated as an independent unit and employed to assign high values to the weights corresponding to the essential information during the training phase. In particular, the squeeze operation is to aggregates the 2D spatial dimension of each input feature, and the excitation operation is to learn and output the attention factor of each feature by capturing the inter-dependencies. Meanwhile, the residual network is adopted to alleviate the gradient vanishing issue due to the network depth. With the simulation, it can be shown that the proposed SE approach shows better performance under various fading channels and SNRs compared to the CNN-based methods. However, similar to the CNN-based SE model, the SE-ResNet-based one still fails to implement a dynamic SE that adapts to the channel condition varies.

Later, the authors in [82] further focus on speech recognition tasks for the English language. In [82], the original speech sample sequence is converted into a spectrum before feeding into the transmitter. Moreover, they introduce a transcription of a single speech sample sequence, where each token represents a character in the alphabet or a word boundary. Based on the spectrum and transcription, they design the encoder and decoder. The semantic encoder is constructed by the CNN and the gated recurrent unit-based bidirectional RNN (BiRNN) [83] modules. The CNN is utilized for data compression and the BiRNN is utilized to extract the text-related semantic features before transmission. The channel encoding and decoding are performed by the dense layer, and the semantic decoding is responsible for decoding the recovered text-related semantic features into the text transcriptions. The text-related semantic features are referred to as a probability matrix with the probability that each token corresponds to each letter. Considering the limited number of letters in the English alphabet, the semantic decoder is designed as a greedy decoder, wherein the maximum probability in all the steps is indexed and the corresponding token is employed to construct the final transcription. With the simulation, the SemCom-based speech recognition achieves a much lower character-error-rate and word-error-rate under a low SNR region, compared to the traditional communication systems. In traditional communication systems, the speech

signals are transmitted directly and then transcribed into text at the receiver with automatic speech recognition (ASR) module [84] or the speech signals are first transcribed into text at the transmitter by ASR module and then transmitted. However, as SNR increases, the superiority of the algorithm becomes diminishing due to an unavoidable error floor generated by DL [85].

4) SE for Multimodel Data: In addition to the three representative data above, the authors in [86] take the visual question answering (VQA) task as an example and investigate a SemCom system for multimodal data transmission. In a VQA task, some users transmit images while the others transmit texts to inquire about the information of the images. The answer is obtained at the receiver. In [86], they consider a simple communication scenario with an image transmitter, a text transmitter, and a receiver. Similar to the above works for image and text, the proposed image transmitter employs the ResNet-101 network [65] pre-trained on ImageNet [88] and the proposed text transmitter employs the Bi-LSTM network. Nevertheless, the design of the decoder is not well studied. Since the SI from both users is correlated, the decoder needs to merge the text and image SI as well as answer the vision questions. To address the issue, the authors adopt the memory, attention, and composition (MAC) neural network [89] to deal with the correlated data. Specifically, each MAC cell consists of three units. The *control unit* first generates a query based on the received text SI by an attention module, then the *read unit* receives the query and searches the corresponding key from image SI by another attention module [86]. Finally, the *write unit* integrates the information and outputs the predicted answers to the questions [86]. Compared to the traditional method, where the recovered image and text are input to MAC, the end-to-end scheme achieves significantly higher answer accuracy. However, since the scheme assumes perfect channel state information due to the lack of attention mechanism, it is not robust to channel changes in a real-world environment. Furthermore, in [87], the authors unify the semantic encoding structure for both image transmitter and text transmitter based on Transformer. Meanwhile, they propose a new semantic decoder network that consists of two modules: the query module and the information fusion module. The query module adopts *layer-wise Transformer*, which consists of Transformer encoder layer and Transformer decoder layer. Different from classical Transformer, layer-wise Transformer in [87] takes the output tokens of each encoder layer as the input of each decoder layer, which can exploit more keywords in the text information and the corresponding regions in the image information. The fusion module then fuses both information to get the answer. Compared to [86], the scheme achieves comparable answer accuracy in both perfect and imperfect channel state information and is considerably higher than traditional methods. However, the complex model design introduces extra time consumption and computational resource consumption for training, especially for text encoding. Meanwhile, the size of the images used for training and testing is small. The superiority of this model over traditional methods for VQA with large image sizes is yet to be verified. An instance of a simple VQA task can be found in Fig. 6.

Q: *What family do these animals belong to ?*

A: Canidae ✔ Felidae ✘

Q: *How many dogs are there?*

A: One ✘ Two ✔

Fig. 6. An example of a simple VQA task.

B. RL-Based SE

Intuitively, the learning process guided by sophisticated semantic metrics can facilitate more accurate SE. However, many existing semantic metrics in other fields are *non-differential*. To overcome the stringent requirements of DL for the loss function to be differentiable, RL is treated as a promising alternative.

RL is regarded as a promising paradigm to address the issues with user-defined, task-specific, and non-differentiable task metrics in some other fields [90], [91], [92]. Considering the success of RL in sequence-generation tasks [93], [94], [95], the authors in [96], [97] make the first attempt to integrate RL into the end-to-end SemCom system for text transmission, where the encoder-decoder scheme can be viewed as the agent that interacts with an external "environment", (i.e., sentences). In the general RL framework, the tasks required to be converted into a Markov decision process (MDP), which consists of five elements: state, action, policy, reward, and long-term return [92]. In their proposed encoding-decoding scheme, the LSTM is employed to provide the policy. Similar to the MDPs for other sequence-generation tasks, the state is defined as the recurrent state of the decoder and the previously generated words. In this sense, the transition between two adjacent states is determined by the next generated token. Meanwhile, the action of the RL agent is to generate a new token, and thus the action space is the dictionary dimension. Moreover, the semantic metrics of the whole recovered sentence can be intuitively treated as the long-term return

However, the determination of the immediate reward function form is particularly tricky. Unlike most RL-based strategies with well-defined rewards at each time step, the rewards during decoding cannot be directly measured until the end of a sentence. To overcome this challenge, several methods have been proposed. The first is using the Monte Carlo search to obtain the reward in each time step [95], [98]. The second is training another neural network to estimate the reward or for an incomplete sequence [99]. However, the above methods are more time-consuming and resource-consuming and introduce the risk of divergence in a huge action and state space [96]. Moreover, quantifying the reward value in each time step may be inconsistent with ensuring the semantic meaning of the whole sentence. Thereby, in [96], the authors adopt a newly emerging approach named self-critical sequence training (SCST) [100]. The idea of SCST is to utilize the output of its own test-time inference algorithm to normalize the long-term rewards it experiences, rather than to focus on estimating the reward, or how the reward function should be normalized [100], [101]. In [96], the mean long-term return (i.e., the semantic metric for the whole sentence) from a group

of selected samples is used to normalize the rewards and treated as the baseline term in the objection function, which enables stable and self-supervised training at the cost of nearly no extra computations. Meanwhile, it is necessary to note that the policy network is not updated until the end of a complete transmission of a sentence.

In the simulation, the proposed RL-based scheme is trained with the semantic similarity metric of CIDEr, and the performance is evaluated by BLEU scores from 1-gram to 4-gram. The size of gram means the length of the phrase considered in calculating the similarity between the reference and candidate sentences, which is detailed in Section V. By comparing the proposed scheme with the DL-based SemCom trained with cross-entropy loss, one can be found that, with the increasing size of the gram, the superiority of RL-based algorithms over DL-based scheme becomes more significant. This demonstrates the capability of the proposed scheme to catch the underlying semantics, since the longer phrases generally carry more abundant semantic meanings.

C. KB-Assisted SE

Intuitively, given raw data, the SI may be distinct for different communication goals [8]. ples of image transmission depicted in Fig. 9, the receiver may have to perform different tasks, such as classification based on different attributes, detection of different targets, or simply replication. In this sense, the SI applied to different tasks is different, but is still highly correlative. Therefore, if we employ a general DL-based SE model for multi-tasks, the extracted SI may be redundant for specific tasks. Otherwise, repeatedly performing SE on the same raw data based on multiple DL-based SE models that are discussed in Section III-A may cause much system redundancy. To address this issue, a suitable approach is to extract all the SI units conveyed by the raw data and correspond individual communication goals to the different combinations of SI units. To this end, a shared KB is required to be established at the transmitter and receiver in advance before the task request is sent. Meanwhile, the process of SE can be treated as refining the importance of each SI unit to individual communication goals.

KB is a technology that has been widely used in automated AI systems to store the data with formal representation allowing for inference [105], [106], [107]. In general, a typical KB consists of a computational ontology, facts, rules, and constraints [11]. Particularly for the SemCom system, the KB should be composed of SI, goals of the communication tasks, and the possible ways of reasoning that can be understood, recognized, and learned by all the communication participants [9], [108]. Specifically, the KB can be employed to record the relationship between each SI unit and each task, as well as quantify the level of importance of SI for different tasks, thus instructing the SE under different channel conditions, when the communication task changes.

Following this, in [108], the authors first manage to establish a simple KB based on CNN for an image classification task and accomplish the KB-assisted SE. In their work, CNN is treated as a SI generator, wherein the feature maps for each layer output indicate different aspects of SI of the source
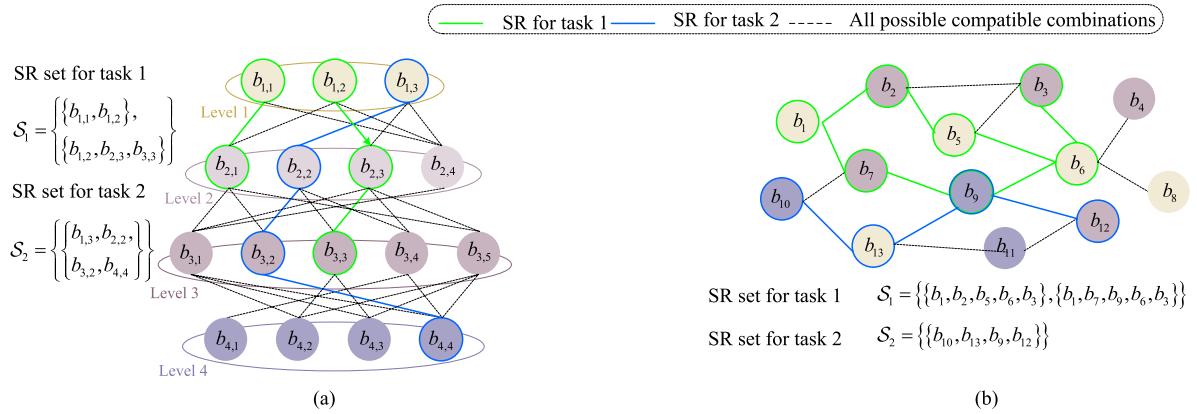
Fig. 7.   Two types of KB models [102], [103], [104].

images, such as the color, the texture. Since the parameters of a well-trained CNN model can identify the optimal form of feature maps that represent the original image (i.e., SI), the gradients of the CNN's output can be treated as the importance weights of the feature map to different classes [108]. Thereby, the KB is established by storing the importance weights of all feature maps for each class. Next, based on the KB, semantic encoding can be accomplished by refining the SI that is relevantly related to the specific task. In addition, since encoding and decoding are mutually reversible processes, the scheme is also implemented in an end-to-end manner. In this sense, the KBs in both the transmitter and receiver should be synchronized by a shared KB at an authoritative third party or a virtual KB. If the two KBs on the two sides mismatch, the *semantic noise* may be generated during the SI inference [29]. Since KB-assisted SE focuses only on the goal-related SI, the KB-assisted SemCom with CR[3] of 98% can still achieve more than 40% classification accuracy gains compared with the conventional communications at 10 dB. However, it still has room for enhancement, such as the optimization of neural network structure and loss function [109], [110], [111].

In addition to the study of the KB establishment of the semantic KB, in works [102], [103], [104] on resource allocation in SemCom scenarios, some ideas are proposed for the KB storage model of the KB. From now, there are two available kinds of KB models as shown in Fig. 7. In [102], a hierarchical structure is proposed for the semantic KB of a task set, wherein the indivisible units of SI are called beliefs. The higher the level that a belief belongs to, the more SI it contains. For a task in the considered task set, there may be multiple feasible semantic representations (SRs) and each SR only includes one belief from each level of the hierarchy [102]. However, such a hierarchical structure is hard to incorporate the relationship between multiple descriptions and a task. Moreover, in the hierarchical structure, a belief in a higher level is completely dependent on a belief in its previous level. Therefore, it is not flexible enough to represent the combination of several beliefs belonging to discontinuous levels. In [9], the authors point out that a graphical structure is one potential solution adopted to model semantic knowledge, where any two SI units can be linked by an edge if necessary. In [103], [104], the authors

firstly employ the graphical structure to model the semantic knowledge for text transmission according to the grammatical structure of sentences. In their works, the tokens encoded with fixed bit length are treated as vertexes, and the relationship between two tokens is reflected by the edge. However, the issue of modeling a generic semantic KB is still open.

D. Semantic-Native SE

All the above three SE methods rely on well-trained neural networks based on a large amount of labeled data, which makes their works only feasible for communication systems with unvarying SI. Hence, they are powerless for scenarios where semantics vary over time or communication context, and such scenarios are more common in real life [112]. Specifically, In this sense, transforming "passive learning" into "active learning" is particularly imperative for SE in communication with varying semantics and context.

Indeed, there have been some primary research studies fitting the above idea called *emergent communication* [114], wherein the semantics and goal-oriented representations are not predefined and are required to be learned during the iterative communication between multiple intelligent agents [112], [114]. However, most of the works merely focus on some simple and specific AI tasks such as image-related referential games [115], where the accomplishment of SE may be spurious owing to the inscrutable patterns of the single transmitted objects [116].

In [112], the authors open up the black box of SE with the focus on a point-to-point communication scenario between two agents who can communicate in both directions. In analyzing the reliability (which is measured by the recognition accuracy in the considered scenario), they introduce the *triangle of meanings* [113] for human communication architectures in linguistics. As shown in Fig. 8, the vertices of the semantic triangle connect the three spaces of the observation of the input, concept (or meaning), and symbol (or representation) [112]. The edge from an input embedding to its concept is termed conceptualization, and the edge from the concept to its symbol is termed symbolization, while their opposite directions represent deconceptualization and desymbolization [112]. Based on this model, they propose two SemCom systems (indexed by System 1 and System 2, respectively). System 1 can be summarized by a multi-triangular model with a shared input embedding. The

---

[3]The value of CR means the percentage of feature maps that are ignored.
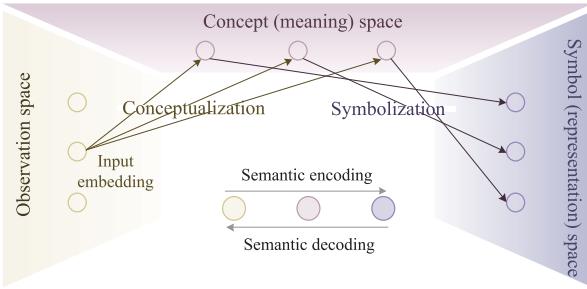
Fig. 8. An illustration of triangle of meanings model [112], [113].

conceptualization process can be interpreted as a stochastic soft decision or the likelihood of a decision in ML, which plays a similar role in unconscious pattern recognition to the ML in the aforementioned methods. In addition, the symbolization process is assumed to be predetermined among the agents. Due to the fact that rational speakers are self-aware of what they say, in System 2, the authors infuse *contextual reasoning* [117] process for each agent. In linguistics, contextual reasoning is often computationally described using the rational speech act model [118], [119], [120], which is rooted in the Gricean view of language use [121]. In the proposed system, contextual reasoning is equivalent to communicating with a virtual agent that mimics and simulates its listener, which allows the agents to communicate effectively and efficiently based on reasoning. To demonstrate the significance of contextual reasoning, the authors abstract both the systems into stochastic models, and derive the bit-length of semantic representation in the two systems with Shannon coding. With the experimental results, it can be seen that the bit-length of semantic representation is significantly reduced with high reliability.

E. Some Specific SE

The above four SE methods shown in Table VI can be generalized to semantic-oriented and goal-oriented communication systems in different scenarios. However, there is no general approach to SE in semantic-aware communication. In this subsection, we first give two typical examples as shown in Fig. 9 to illustrate the motivation of introducing semantic awareness to communication, as well as the role of SE in semantic-aware communication.

In [122], the authors focus on a federated DRL task as shown in Fig. 9(a), where multiple heterogeneous agents participate in model training in a cooperative manner under the coordination of a central controller. The agent requiring training is called the target agent, and the agents that help the target agent with its training are called source agents. In the proposed scheme, SemCom plays a role in the construction of a knowledge graph that records similarities among all the agents. Based on the KG, a subset of source agents with high similarity are selected strategically to contribute to the training of the target agent. In their work, they employ semantic relatedness to measure the similarity of underlying learning tasks of the agents [124]. In the knowledge transfer domain, it can quantify the extent to which the transferred knowledge from a source agent can potentially help the target agent to find its optimal policy. In this sense, semantic relatedness can be regarded as a kind of SI exchanged between the central controller and agents. It is defined as the average return value received by the source agent from a target environment in a limited number of training episodes. Meanwhile, the training process can be treated as its corresponding SE method. However, since the average return value is obtained from limited training steps, the metric may be inaccurate, especially for a complex target environment with a large state space. Therefore, the structural similarity is jointly considered during the similarity KG construction. The structural similarity can be measured by the central controller based on the cosine similarity of the received network parameters of the policy from agents [125]. Besides, since the DNN parameter dimensions of heterogeneous agents should be aligned before comparing the similarity, the principal component analysis (PCA) method and the zero-padding (ZP) method are employed for compression and expansion of the source agents' DNN parameters [126], respectively. With the simulation, the promising performance of the semantic-aware CDRL schemes [122] over bandwidth-constrained wireless networks has been demonstrated by comparing with the uniform or random resource block allocation.

In another example in [123], the authors focus on a non-terrestrial ultra-reliable and low-latency communication (URLLC) system (as shown in Fig. 9(b)), where an unmanned aerial vehicle (UAV) swarm employs the centralized training and decentralized execution (CTDE) multi-agent deep reinforcement learning (MADRL) to serve a moving ground user while avoiding inter-UAV collisions. In their work, SemCom is not employed to enhance the URLLC performance directly, but to be integrated to the differentiable inter-agent learning (DIAL) [127] between UAVs to avoid inter-UAV collisions. DIAL is widely used in CTDE MADRL framework. In the traditional DIAL, the agents exchange their respective observable states, which are treated as an input of their actor models. Then during the training process, the raw state data are progressively translated into meaningful information for better inter-agent collaboration. However, training from scratch could be inefficient. To that end, the authors in [123] perform SE on the observable state before exchanging information. Specifically, each UAV n constructs a local star-topological graph, where the leaf nodes are all its observable UAVs. The SI about UAV m at step t is recorded as the weight of the corresponding edge (i.e., $\bar{\varpi}_{n,m}^{t}$ in Fig. 9(b)), which reflects the level of attention to UAV m, when UAV n takes its action. Considering the fact that UAVs should pay equal attention to each other for collision avoidance, the attention paid to UAV m by UAV n is derived based on an RNN with the two inputs. One is the attention feature extracted by UAV n from its observable state about UAV m based on the self-attention mechanism [59]. The other is the SI about UAV n at last step that is sent by UAV m to UAV n. From the simulation, it can be seen that using SI as the input of actor model instead of raw state data can significantly improve the training efficiency.

F. Lessons Learned Summary

1) Lessons Learned for DL-Based SE: Comparing the DL-based SE methods as summarized in Table V, we can observe that the attention mechanism plays an essential role in
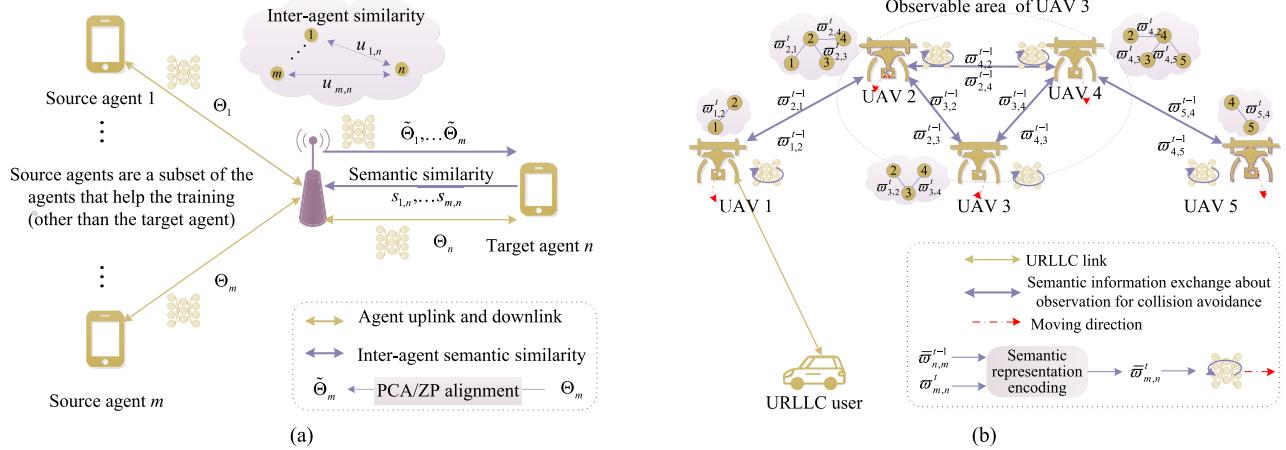
Fig. 9.   Examples for SE in semantic-aware communication [122], [123].

SE performance enhancement due to its excellent performance in capturing long-range dependence of the input. The advantage of DL-based SE is that it can extract important information from the whole raw data, and then re-aggregate and re-extract it at different layers, thus effectively removing redundant information. In this sense, compared with the content-blind traditional encoding and decoding, DL-based SE can achieve much lower CR without losing relevant information. Therefore, under the low SNR region, the superior performance of DL-based SemCom is especially remarkable. Moreover, with the well trained end-to-end SE model, the DL-based SemCom achieves good performance in analog communication. Due to the absence of quantization and complex modulation processes, such as 16-QAM and QPSK, the data processing delay before transmission can be significantly reduced, which shows its potential for low-latency communications.

However, the DL technique has an inherent weakness, i.e., the unavoidable error floor [85]. Therefore, under ideal channel conditions, the DL-based SemCom is often suboptimal compared with traditional communication. Hence, how to overcome the performance bottleneck under the high SNR region is worth future investigation. Moreover, during the end-to-end training, the back-propagation through transceivers requires the loss function in the DL paradigm to be *differentiable* [97]. In this sense, all the above studies still apply the commonly used loss functions in DL (i.e., cross entropy (CE) and mean square error (MSE)) to train neural networks, which leaves the existing works far from the desired SemCom. In other words, all the above-mentioned architectures merely achieve semantic coding for a reliable and efficient transmission [97]. With such end-to-end architectures, as the semantic and channel encoders and decoders need to be trained jointly, the SE and recovering are treated as a black box [112]. Due to the lack of *explainability and interpretability* for the available DL-based SE, the informativeness of the extracted SI is hard to measure and it is also unclear how to make relevant improvements.

2) Lessons Learned for RL-Based SE: Comparing the DL-based SE and RL-based SE, the main difference of RL-based SE lies in that the decoding of the whole sentence is converted into a recurrent procedure. That is, the output of the RL-based decoder is a single word, and the decoded word is the input for the next word to be decoded. Such a recurrent procedure can
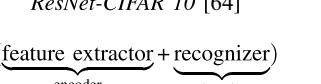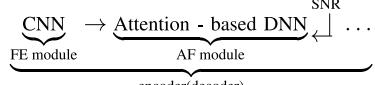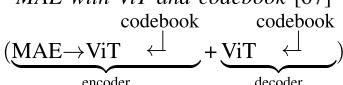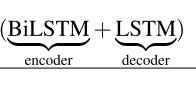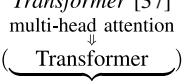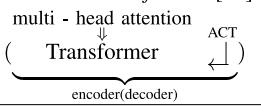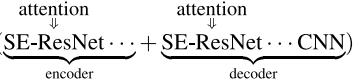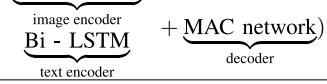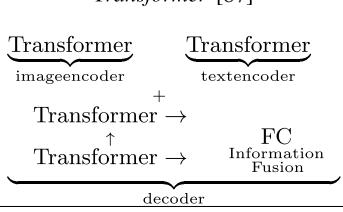
reinforce the learning of correlations among the words within a sentence during the training process, thus allowing the decoding policy to learn relevant features of the non-differentiable semantic metric function. Moreover, as RL can be treated as an online paradigm, in addition to error-based metrics, some other metrics such as AoI-based metrics and transmission delay can also be integrated into the reward, which is another promising advantage for RL-based SE methods [128].

Such a method is naturally applicable to sequence-generation tasks, such as sentence recovery. In [96], [97], the authors also discuss the generalization capability for non-sequence tasks with an example of image transmission. They propose a pixel-level recurrent decoding scheme, where the state of MDP is defined as the intermediate decoded image obtained by increasing or decreasing the pixel values with a small number. However, the effectiveness and necessity of such a conversion for image decoding for the performance enhancement of SemCom are unclear. Moreover, even if such a conversion from the decoding process to a recurrent procedure is feasible to implement in practice, it increases the decoding time.

Moreover, learning the optimal policy through interactions with the environment inevitably increases the training complexity. It is still a critical challenge to train such a complex model from scratch for high-dimensional tasks [97]. In the above works, the initial parameters are utilized the pre-trained model with stochastic gradient descent on the deterministic loss function. Simulation results show that the accuracy is improved by 3% in the middle SNR region with the RL-based SemCom based on the non-differential semantic metric optimization by comparing that with the DL-based counterpart. However, for the more complicated language models like Transformer, whether the RL-based SE is still feasible needs to be further explored and studied.

3) Lessons Learned for KB-Assisted SE: In fact, KB-assisted SE also heavily relies on the DL models. Different from the well-trained DL-based model that performs real-time SE for the new raw data, KB-assisted SE requires the KB for raw data to be well-constructed and synchronized among the parties before communication link establishment. This makes it merely suitable for non-real-time on-demand services. Moreover, compared to the single model training in

TABLE V
SUMMARY OF DL-BASED SE FOR TYPICAL DATA TYPES

	NN Architecture	Benefits	Limitations
image — recognition	*ResNet-CIFAR 10* [64] (feature extractor + recognizer) encoder / decoder	The scheme greatly reduces the compression rate while ensuring recognition accuracy and dramatically reduces system complexity and processing latency.	This scheme is applicable only to a specific SNR range and requires re-training when the channel changes, thus introducing additional overhead.
image — transmission	*Attention-integrated DNN* [66] CNN → Attention - based DNN $\leftarrow$ SNR ⋯ FE module / AF module encoder(decoder)	By integrating SNR into SE, the scheme can operate successfully over a wide range of SNRs with lower computational/storage complexity than that of the basic DNN-based structures.	The effectiveness of the scheme is just demonstrated in AWGN channel. The robustness and adaptability of the algorithm are still to be verified and studied under more general models.
image — classification	*MAE with ViT and codebook* [67] codebook / codebook (MAE→ViT $\leftarrow$ + ViT $\leftarrow$) encoder / decoder	Based on MAE, the scheme achieves high SE efficiency by reducing the image spatial redundancy and resists largely the interference of semantic noise on classification with codebook.	The complexity of the system is high, which poses difficulties in training the SE model. For large image transmission, the feasibility and effectiveness of this scheme are yet to be verified.
text transmission	*LSTM* [69] (BiLSTM + LSTM) encoder / decoder	Compared with Gzip and Huffuman, the scheme can achieve remarkable low word error rate for large-size sentences and high bit-drop rates.	The scheme can only capture the relationship among words and fails to describe syntax information. This makes it hard to deal with complex sentences.
text transmission	*Transformer* [37] multi-head attention ⇓ (Transformer) encoder(decoder)	The multi-head attention module in Transformer can capture long-range dependencies in sentences in parallel with low complexity, thus extracting accurate SI.	The attention structure in Transformer is fixed, which makes it hard to deal with noise interference or polysemy, such as "mouse" has a different meaning in computing and biology.
text transmission	*Universal Transformer* [75] multi - head attention ⇓ ACT (Transformer $\leftarrow$) encoder(decoder)	The scheme can be considered as a per-symbol self-attentive RNN, which can capture more precise SI and flexibly respond to varying channel conditions.	Due to the loop play introduced by the adaptive circulation mechanism, computational complexity increases, which causes extra processing latency and computing resource demand.
speech — transmission	*CNN* [78] (CNN → CNN + CNN → CNN) FE / FA / FD / AG encoder / decoder	The SE model is simple and easy to train. The scheme is remarkably competitive at low SNRs.	The basic SE model fails to extract semantically enriched information and adapt to changing channel conditions.
speech — transmission	*Squeeze-and-excitation network* [80] attention / attention ⇓ ⇓ (SE-ResNet⋯ + SE-ResNet⋯CNN) encoder / decoder	Due to the introduction of the attention mechanism, SE-ResNet-based SE model can achieve higher performance in terms of PESQ and SDR at any given SNR than CNN-based one.	This scheme can only perform the training of SE model under a fixed SNR. A dynamic and flexible SE model that adapts to channel changes remains to be studied.
speech — recognition	*CNN & RNN* [82] (CNN → BiRNN + greedy) encoder / decoder	The scheme achieves a much lower character-error-rate and word-error-rate compared to the traditional communication systems under low SNRs.	The scheme becomes sub-optimal as SNR increases, since the DL-based methods always generate an avoidable error floor.
multi-model data — VQA	*ResNet & LSTM & MAC network* [86] ResNet - 101 image encoder (Bi - LSTM + MAC network) text encoder / decoder	Compared to the traditional method, where the recovered image and text are input to MAC, the end-to-end scheme achieves significantly higher answer accuracy.	Since the scheme assumes perfect channel state information, it is not robust to channel changes in a real-world environment
multi-model data — VQA	*Transformer* [87] Transformer / Transformer imageencoder / textencoder + Transformer → ↑ Transformer → FC Information Fusion decoder	The scheme achieves comparable answer accuracy in both perfect and imperfect channel state information and is considerably higher than traditional methods.	The complex model design introduces extra computing latency and computational resource consumption for training, especially for text encoding. The size of the images used for training and testing is small. The superiority of this model over traditional methods for VQA with large image sizes is yet to be verified.

DL-based SE, the construction of a sophisticated KB is a much more computation-intensive task. This means that the KB cannot be updated frequently, so this method is more suitable for scenarios where the data source is stable.

In addition, as discussed earlier, by introducing communication goals into the SE, KB-assisted SE can improve multi-task system efficiency in two aspects. One is to improve communication efficiency by extracting flexibly the SI related to a

certain task in each transmission. The other is to improve computational efficiency by avoiding repetitive SE of raw data. Beyond these two advantages, the construction of KB also lays the foundation for resource allocation for complex SemCom scenarios with multiple tasks. Since the size and importance of each SI unit can be recorded in the KB model, the SemCom-aware resource allocation for each SI unit can be done with different quality of service (QoS) requirements, such as delay and reliability. This differs from the resource allocation in traditional communication where all the data packets are treated equally, which is discussed in detail in Section IV. Furthermore, it is worth noting that the semantic KB construction relies heavily on the explainability of SE models. However, most of the available SE models have a black-box nature. To this end, enhancing the explainability of SE is the key to breaking the bottleneck of existing SemCom research.

4) Lessons Learned for Semantic-Native SE: Comparing the four SE methods summarized in Table VI, semantic-native SE is the most similar to the way of human conversational communication. The first three SE methods require that the background knowledge of the transmitters and receivers are fully synchronized before SE model training and determination, e.g., the semantic encoder and decoder models are trained based on the same dataset. In contrast, semantic-native SE relaxes this constraint. Contextual reasoning for the agent is more like the process of capturing and inferring the way of thinking of a human during a conversation with a stranger. As communication parties become "familiar" with each other, the background knowledge of the agents converges. Undoubtedly, the semantic-native SemCom system is with a high degree of flexibility and adaptivity, which is more toward the vision of an intelligent and autonomous 6G network [129]. However, the above analysis is again only based on a representative model. It remains a huge challenge to put it into practice.

5) Lessons Learned for Specific SE: From the two examples of a multi-agent collaborative task discussed in Section III-E, we can see that SemCom in the above tasks only serves to facilitate cooperation and indirectly improve task performance. In contrast to semantic-oriented and goal-oriented communications, SE in semantic-aware communication is not performed on the raw data generated directly by the source agent. The relevant SI for cooperation is obtained through analyzing the properties of the task itself and the behavior of the agents. This makes the process of SE have to be more tailored and it is difficult to find a universal and unified approach. However, it is clear that semantic-aware communication will play an essential role in task-oriented communication, as exchanging SI can enhance the knowledge between agents for better collaboration. Moreover, an efficient SE can greatly reduce the communication overhead caused by exchanging information between agents.

IV. SEMANTIC INFORMATION TRANSMISSION AND CHALLENGES

The focus of discussion in this section shifts from semantic-related to communication-related challenges and techniques. While conventional and SemCom systems use different methods to encode and decode information, they both face the
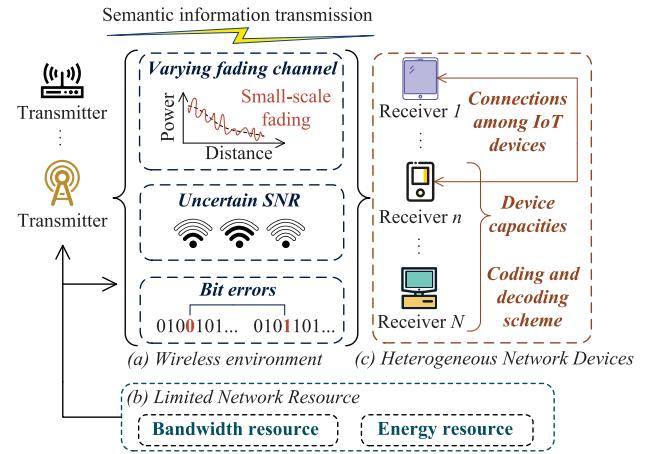

Fig. 10. Semantic information transmission and challenges related to wireless environment, limited network resource, and heterogeneous networks.
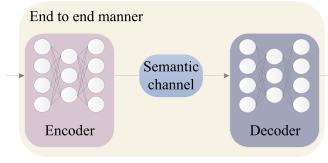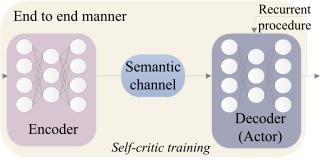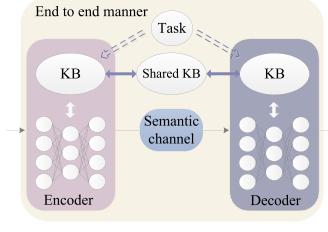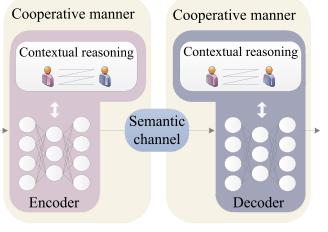
same communication constraints, such as unpredictable channel conditions, limited transmission, and processing resources. However, unlike prior works in conventional communication systems, the solutions in SemCom are required to address new challenges in modern communication systems. In the following, as shown in Fig. 10, we discuss the challenges and techniques related to the wireless environment, limited network resources, and heterogeneous networks in SemCom.

A. Wireless Environment

The fading effect of the wireless channels has a great negative impact on the stability of data transmission, whether in conventional communication or in SemCom. To mitigate the negative effects of fading channels, in conventional communication systems, source and channel coding schemes are carefully designed. Specifically, source coding encodes the data into a sequence of symbols with optimized length, and channel coding adds redundant symbols to the sequence to detect and recover data corruption during wireless transmission. In the SemCom systems, the source and channel coding can be connected more tightly with the help of AI. Jointly designing and training the source and channel coding are shown to benefit data transmission in the DL-based communication systems [69]. However, the AI-based methods cannot currently be explained by explicit mathematical expressions. To overcome this obstacle, SemCom system designers have to consider how to make connections between the complex and changing wireless environment and sophisticated SemCom mechanisms, so as to obtain insights to guide the system design. Thus, we discuss the impact of the varying fading channel, uncertain SNR, and bit error on the SemCom performance.

1) Varying Fading Channel: In conventional wireless communications, several classical channel models are commonly used in performance analysis, e.g., Rician, Rayleigh, and Gaussian channel models. Moreover, to unify system performance with various channel environments, generalized fading distributions are proposed, e.g., $\alpha - \mu$ [130], Fisher-Snedecor $\mathcal{F}$ [131], FTR fading models [132]. These channel models have various parameters to represent the different conditions of the wireless environment, such as the strength of the

TABLE VI
SUMMARY OF GENERIC SEMANTIC EXTRACTION METHODS

Semantic-oriented communication	
DL-based SE	**RL-based SE**
Description: The encoder and decoder are usually modeled as two separate learnable NNs, and linked through a random channel, which are trained jointly. The dataset used for training can be seen as their shared background knowledge.	**Description**: It is developed on the basis of DL-based SE. The decoding process is converted into a recurrent procedure. By employing self-critic training, the non-differentiable metrics, such as BLEU, can guide the learning process directly.
Pros: • Achieve lower CR while preserving the relevant information • Significant superiority in the low SNR region • Reduce processing latency in analog transmission without compromising communication performance **Cons**: • Become sub-optimal in ideal channel conditions due to the error floor of DL • The loss Function for guided learning in training can only be used for differentiable MSE and CE	**Pros**: • Achieve more precise SE guided by the specialized semantic metrics • Time-related metrics, such as AoI can also be integrated into the reward to guide SE due to the online paradigm of RL • Also features the pros of DL-based SE **Cons**: • Frequent interactions with the environment of RL greatly increases the training complexity • Applicable only to sequence-generation tasks, such as sentence recovery
Goal-oriented communication	
KB-assisted SE	**Semantic-native SE**
Description: The KB stores all the SI units conveyed by the raw data and the importance of each SI unit to different tasks, which is well-constructed before communication link establishment. In each transmission, only the task-related SI is transmitted according to the KB and channel states.	**Description**: It is developed based on emergent communication. It converts "passive learning" to "active learning". SI and background knowledge are learned through interaction and feedback between the communicating parties, which does not depend on an existing database.
Pros: • Allow for flexible and more precise task-specific SE • Applicable to complex communication scenarios with multiple goals • Lay the foundation for SemCom-aware resource allocation due to the quantified data size and importance of SI units **Cons**: • Applicable only to the non-real-time on-demand services • The construction of KB is computation-intensive	**Pros**: • Adaptive to changes in the communication context and goal, reducing human intervention • Background knowledge does not need to be shared in real time • Some other features such as channel states and QoS requirements can be considered in the learning process **Cons**: • The training process is time-consuming and computing resource intensive • Convergence of training is hard to be ensured

shadow effect and the multi-path effect. However, in SemCom systems that have an end-to-end structure, modeling the channel layer is a challenging task. Most of the existing works model the channel layer in two ways: fixed channel layers with the fading models used in conventional wireless communications, and generative channel layers with the neural network.

Fixed channel layer modeling scheme: In the fixed channel layer modeling scheme, the channel layer is modeled as a fixed fading model that is used throughout the training process. For example, the erasure channel is used in [69] to model the dropping of data packets. The input of the erasure channel is a binarized bit vector from the encoder. Every element in the bit vector can be -1 or 1, and the dropped element will become 0 at the output of the erasure channel. A drop probability is determined before the training. Eventually, the elements of the output vector after the erasure channel are in $\{-1, 0, 1\}$. This process is similar to the dropout technique used to prevent the over-fitting problem in deep neural networks. Hence a dropout layer can be used to represent the erasure channel of the communication systems. For input that is not quantized or binarized, communication channels such as AWGN, Rayleigh channels, and Rician channels are considered. Recent research works in SemCom for text [37], speech signals [80] and multimodal data [86] consider the channel layer as AWGN, Rayleigh channels, and Rician channels, for the training process. However, the performance evaluation is

done with the same channel conditions in the training, without considering the fact that the changes in the wireless environment can lead to the change of the suitable channel model. Another drawback of using a fixed channel layer modeling scheme is that, if the SemCom model is trained under a certain fading channel, it is impractical to retrain the model for each possible channel condition and load all these models to the transmitter and the receiver. Although the trained model has shown some robustness, e.g., the authors in [80] tested the model trained with Rician channels under AWGN and Rayleigh channels while achieving MSE loss less than 1×10^{-4}, we cannot explain the upper-bound of the robustness, and are not sure whether the model will fail when the environment changes. Instead of training with fixed channel layers, the generative channel layer modeling scheme is used to capture the dynamic behavior of the channel states.

Generative channel layer modeling scheme: The typical generative network adopted in existing works is generative adversarial net (GAN) [133]. There are two main components in GAN, i.e., a generator and a discriminator. The generator aims to generate data samples that are as similar to real data samples as possible. The discriminator will be given real and generated data randomly and it will output a label to indicate whether the given data is real or generated. During the training process, the objective loss function helps the generator to generate more realistic data and the discriminator to output more accurate labels. To generate category-specific data, the conditional GAN [133] is proposed where the extra context information is provided to the GAN to obtain samples of the given context. In [134], a conditional GAN is used to model the channel conditions. The conditional GAN is provided with both the pilot information and encoded signal from the transmitter and asked to generate an output signal that is similar to the real data. To evaluate the performance of the proposed conditional GAN in real channel conditions, the trained model is tested with the WINNER II channels [135]. It is shown that the conditional GAN model outperforms the baseline system in terms of BER and BLER, especially when the SNR is over 10 dB. Although the Generative channel layer modeling scheme improves the adaptability of the training model to the wireless environment, more training overhead is required than the fixed channel layer modeling scheme. This motivates us to think about the tradeoff between the performance and resource overhead.

Overall, fixed channel layers and generative channel layers provide wireless channel modeling during the training of the SemCom systems. The choice of different schemes has an impact on the performance of SemCom. One possible solution is to combine the advantages of both schemes. For example, a two-phase training strategy is proposed in [53] to adapt to the real channels. In Phase I, the model is trained with a suitable channel model to obtain model parameters with reasonable accuracy. In Phase II, the receiver is fine-tuned over the actual channel. The fine-tuned autoencoder constantly achieves lower BLER than the autoencoder without fine-tuning. However, how to choose the optimal training solution in different wireless environments is still an open question.

2) Uncertain SNR: After discussing the effects of the channel environment on the SemCom performance, here we consider the impact of SNR uncertainty on the trained SemCom model. Note that the influence of the wireless environment is mainly from the channel model selection, due to the effects of shadowing or scattering. The SNR uncertainty is from the effect of noise and interference, as well as the variation in transmit power, e.g., when adaptive transmit power schemes are used. Since the fixed SNR approach is typically adopted in the training of SemCom models, e.g., for text [37], speech signals [80] and multi-modal data [86], we need to consider whether the change in SNR will have a negative impact on performance.

Several works tested the robustness of the trained model to SNR. In [80], the model is trained with a fixed SNR of 8 dB, and is then evaluated under SNR from 0 dB to 20 dB. It is found that the model has a higher MSE loss in the lower SNR region in the test under AWGN channels, Rayleigh channels, and Rician channels. However, it is still uncertain whether a model trained at a fixed SNR value can always be applied to a wide range of SNRs. Furthermore, while the SemCom constantly achieves higher performance than the conventional communication systems, both of them suffer a poorer performance in the lower SNR region. Because low SNR environments are common in the cellular edge, shopping malls, or suburban areas, we need to consider the performance of SemCom when the SNR is low and the accuracy of the decoded signal is reduced.

To solve the aforementioned problems by making the model robust to different SNR regions, especially to the low SNR regime, the authors in [136] proposed an SNR adaptive mechanism. In the proposed model, the SNR is estimated by a pilot signal at the receiver. The estimated SNR value is then extended to an SNR map that has the same size as the channel output feature map. Both the SNR and channel output feature maps pass through a CNN layer before they are added together. The result of the element-wise addition is used as the input of a denoising module. A few transposed convolution layers are used to reconstruct the original image. It is found that, compared to the model that is trained with fixed SNR, the model trained with the proposed SNR adaptive mechanism has a smaller gap of PSNR between the high SNR region and low SNR region. By considering the SNR information in the decoding process, the proposed model shows higher adaptability to the SNR.

Instead of adding the SNR values to the channel features, another method to enhance the robustness of the SemCom model is to scale the channel features according to different SNR values [66]. The training method proposed in [66] adopts channel-wise soft attention, where each channel feature is multiplied by a scaling factor. To obtain the scaling factors, the SNR value is concatenated with the context information vector extracted from the input image and fed into two fully connected layers. Each element in the output vector is a scaling factor of a feature channel. It is shown that, with the help of the soft attention mechanism, the model can achieve a higher PSNR compared to baseline models which use basic deep

learning networks without the attention module. In particular, the model with the soft attention mechanism can achieve more than 35 dB PSNR when the SNR is high.

However, both of these solutions are designed to solve specific communication problems. For the generalized SemCom system, the question of how to ensure that the trained semantic model can adapt to the variable SNR is still waiting for a better answer. The boundaries of the generalization capability of semantic models need to be further investigated.

3) Bit errors: To accommodate the changing wireless environment and the uncertainty of SNR, many mechanisms have been designed to improve SemCom performance. Now we focus on the bit error correction mechanism [137] that can further increase the probability that SI is correctly transmitted.

Inspired by error correction algorithms in conventional communication systems, researchers have designed several error correction schemes for SemCom to minimize the transmission errors of SI. For example, in [138], hybrid automatic repeat request (HARQ) is used to reduce the transmission error of semantic text transmission. With the help of HARQ, a re-transmission is requested if the received code block has uncorrectable error. The authors in [138] first develop models with semantic encoder from [37] and Reed Solomon (RS) channel coding [139] with HARQ. Then the performance is further improved by jointly designing source-channel coding and HARQ. Specifically, the RS channel coding is replaced by a dense layer to encode the output of the semantic encoder into a bit vector. To reduce the semantic error, Sim32 encoder and decoder are developed to detect the semantic similarity between original and estimated sentences. It is shown that with the joint source-channel coding and HARQ, the model achieves lower word error rate and sentence error rate when BER is larger than 0.06. In addition, new design schemes that combine correction mechanisms in conventional communication with SemCom require further study.

B. Limited Network Resource

Several resources, such as bandwidth and transmit power, are required for data transmission. Resource allocation frameworks in conventional communication systems aim to minimize metrics such as the bit error rate, packet error rate, and outage probability. However, the SemCom values the importance of the information behind the bit flow. This motivates us to develop the new resource allocation frameworks for the novel SemCom systems. Typically, in the design of a resource allocation scheme, the QoS and the QoE should be taken into consideration to build an effective system. Specifically, the QoS aims to optimize transmission rate, delay, and throughput [140], [141], and the QoE focuses on user satisfaction, clarity, and fluency [142], [143]. In the following, we discuss the methods of resource allocation in terms of bandwidth and energy in SemCom.

1) Bandwidth Resource: Because the bandwidth resources are precious for any communication system, an effective bandwidth allocation is necessary for achieving SemCom to improve the overall system performance. Unlike the allocation of bandwidth resources in conventional communications,

the uneven distribution of SI should be taken into account in SemCom, i.e., more bandwidth should be allocated to data/agents that have more SI.

One possible solution is to jointly perform the bandwidth allocation during the training process. In [122], a CDRL algorithm is designed, where multiple agents can coordinate over a wireless network to share their policies and collaboratively learn the best policy for the respective tasks. However, due to the limited bandwidth, agents that require training (target agents) can only collaborate with limited amount agents (source agents). Therefore, the metrics used to identify the most helpful agents are important for effective resource allocation. Building on previous works that consider the structural similarity of the agent model, the authors in [122] include the semantic relatedness between the agents to construct a KG to aid the task selection of agents. The inter-agent semantic relatedness is defined as the return value of the target agent after a fixed number of training steps under the source agent's policy. After jointly optimizing the training loss and wireless bandwidth allocation, a KG is obtained, in which the values of the edges between the agents capture the structural and semantic relatedness between the connected agents. The KG is then used by the base station to select the most relevant agents for collaboration during the optimization. Simulation results show that the system performance can be improved by 83%, compared to the baseline method that does not consider the semantic relatedness between the agents. However, the issue of combining dynamic allocation of bandwidth to semantic content transmission has not been fully studied. In some SemCom systems that do not require training, the bandwidth allocation schemes need to be designed to allocate more bandwidth for more important transmitted content to ensure information quality.

2) Energy Resource: In addition to the allocation of bandwidth resources, the allocation of energy resources is also an important issue. For the growing number of IoTs with energy harvesting capabilities, it is important to determine the importance of information with the help of semantics. Allocating more energy to transmit data containing richer SI ensures the efficient use of energy. In addition, semantic metrics can be used to determine the quality of the collected energy, which can help build an efficient network market.

Based on the sentence similarity metric proposed in [37], a semantic based valuation function is used in [144] for energy harvesting IoT devices to derive the value of harvested energy. In the proposed system model, there are IoT devices that adopt SemCom systems and a hybrid access point (HAP) that transmits energy to nearby IoT devices. The IoT devices operate by harvesting energy from the HAP [145], [146] to transmit text data to the HAP. However, the HAP is considered to serve only one user at a specific time. To obtain the wireless energy, the IoT devices will submit their bids and the HAP will decide the winner and payment. A truthful auction mechanism is proposed so that the IoT devices will bid according to their true valuation of the energy. Instead of using the performance metrics of conventional communications, a valuation function based on the BLEU score and the similarity score is used by IoT devices to obtain their bid values.

However, at present, the application of semantic metrics in energy resource allocation is still in the early stage. Many SemCom networks that require energy-harvesting devices to work have not been studied, e.g., a UAV-aided network working with simultaneous wireless information and power transfer protocol.

C. Heterogeneous Network Devices

For the SemCom network, the wireless communication layer has a greater impact on the system performance than that of the end-to-end conventional communications. Because many heterogeneous devices work in one SemCom network, the differences in equipment hardware and wireless environments bring challenges to the system construction.

1) Device Capacities: To enable SemCom systems, most of the existing approaches involve installing encoders and decoders into the transmitters and receivers respectively. While the DL-based auto-encoder systems can help to extract meaningful semantic information from raw data effectively, the cost of implementation is not cheap. Particularly more computational power and communication resources are required for the training process. Research shows that scaling up deep neural networks with correct techniques almost always leads to better performance [65], [147]. However, scaling up the model increases the storage requirement to store a higher number of the model parameters. In reality, communication devices have limited computational power, communication resources, and storage capacity. Especially in SemCom networks, it is unrealistic to assume that all devices have sufficient capacity. Therefore, in SemCom networks developing effective methods to balance the performance and cost requirements for heterogeneous devices is one of the important challenges.

To make the model proposed in [37] more affordable to devices in SemCom network with limited computing capability, the authors in [74] experimented with model compression to reduce the size of the model. A joint pruning-quantization scheme [74] is proposed to compress the model effectively with the idea of model pruning [148]. In the proposed method, less significant model weights are zeroed out, and the model weights that are larger than a pruning threshold remain. To determine the pruning threshold, the model weights are first sorted in the ascending order by the weight values, and then the pruning threshold is selected such that the outcome of pruning satisfies a pre-defined sparsity ratio between 0 and 1. The sparsity ratio indicates the desired ratio of zeros of the model weights. The pruned model is then fine-tuned to recover the performance of the model. The size of the model is further reduced by network quantization that converts the model weights from 32-bit float point to m-bits integer, $m < 32$. A calibration process is needed to prevent overflows at the activation layer. In particular, an exponential moving average (EMA) is introduced to dampen the effect of outliers in the output of activation. Similar to model pruning, the model is fine-tuned after the quantization. Remarkably, the compressed model can achieve a similar BLEU score to the uncompressed model after model pruning with a sparsity ratio of 60% and 8-bit integer quantization. However, the performance loss due to further compression of the semantic model needs to be systematically investigated. We need to consider the capacity of different devices in the network and performance requirements.

2) Connections Among IoT Devices: For SemCom networks containing multiple smart devices, we need to design the network according to the different wireless link environments of different devices. One solution is to consider the wireless links as intelligent agents in the training process.

The authors in [149] propose a resource allocation algorithm for semantic video transmission in spectrum multiplexing scenarios in vehicular networks. In the proposed algorithm [149], semantic understanding accuracy of the video transmission is optimized by a multi-agent deep Q-network. In the network, vehicle-to-infrastructure (V2I) links and vehicle-to-vehicle (V2V) links are the agents. Based on the observations about the environment states, such as channel gains and interference power under the resource blocks, the agents choose to reuse spectrum resource blocks. Then, the agents receive the reward based on the V2I average object detection accuracy and V2V average transmission rate. Simulation results show that under the same spectrum and transmit power, the proposed network constantly achieves higher accuracy of video semantic understanding than the QoS and QoE based resource allocation framework, with as high as 70% improvement for the density of correctly detected objects.

However, when the channel model is not available, the feedback links are absent for supervised learning. To solve this problem, the authors in [51] propose a meta-learning approach for the receiver to adapt to the unknown new channel condition. Meta-learning which means "learning to learn" refers to learning the adaptation module in the receiver [51]. The meta-learning method first trains the adaptation rule during the training phase. In particular, during the meta-training phase, the receiver will be meta-trained to update the decoder parameters based on the output of the physical channel. During the testing phase, the receiver will self-optimize the model parameters using the trained adaptation rule. Simulation results show that the model with meta-training can achieve a lower BLER than the model with conventional training, when more than one pilot frame is sent by the transmitter during the testing phase.

3) Coding and Decoding Scheme: The coding and decoding scheme needs to be improved based on the various channel conditions of different users in the SemCom network.

Unlike the two-phase training strategy which fine-tunes the model with supervised data, the authors in [97] propose a self-supervised mechanism in consideration of the varying channel states. In particular, a message is allowed to be encoded and decoded multiple times until a stopping criterion is fulfilled. For every cycle of encoding/decoding, the encoded/decoded information will be evaluated by a confidence mechanism to determine its semantic confidence. If the encoded/decoded information reaches a pre-defined confidence threshold, the encoder/decoder will release the information for the next process. Another stopping criterion is when the cycle length of encoding/decoding reaches a pre-defined maximum cycle length. With the distillation and confidence mechanisms, the encoder and decoder can fine-tune the encoded and decoded information in a self-supervised way, regardless of the channels.

Moreover, in a multi-user scenario, the fluctuation in resources, such as available spectrum and transmit power, can have a non-negligible impact on the SemCom performance [150]. To this end, the variable-length semantic encoding, which is comparable to scalable video coding [151] and multiple description coding [152] in conventional communication, urgently needs to be investigated to cope with the dynamic SemCom network.

D. Lessons Learned Summary

1) Lessons Learned for Wireless Environment: The transmission problems caused by the wireless environment in classical communications should be considered in SemCom. However, due to the particular features of the SemCom, e.g., the differential importance of different bits for the raw data, schemes designed for the conventional communications cannot be used directly. Fortunately, many schemes that are designed for conventional communication systems can inspire the construction of SemCom systems. For example, the end-to-end semantic model training requires system designers to use neural network layers to model wireless fading channels. Compared with using a fixed channel model in the performance analysis of conventional communication systems, a generalized fading channel that includes a variety of classical channels as its special case can bring more insights. Similarly, the introduction of the generative channel layer can offer more freedom to the training of semantic models.

2) Lessons Learned for Limited Network Resources: Different from conventional communications that consider bit transmission only, the purpose of resource allocation in SemCom is to ensure the accurate transmission of semantic information related to tasks. Therefore, the consideration of semantic information introduces a new perspective to resource allocation scheme design in 6G networks. This novel design requires an in-depth analysis of task requirements and joint optimization design, to enhance the system performance. As we discuss above, some literature has attempted to use semantic information to guide the resource allocation process, but this paradigm shift still leaves much room for further research.

3) Lessons Learned for Heterogeneous Network Devices: The heterogeneity of network devices in the network is mainly reflected in two aspects, i.e., the difference in device capacity and the difference in the communication environment of each device. If the heterogeneity in network devices is not considered when training the semantic model, the trained model cannot be directly delivered to each device. Specifically, semantic models trained for high-performance devices may be large and cannot be deployed to small-capacity devices such as mobile phones. In addition, the end-to-end semantic models are influenced by the quality of the wireless channels between the transceiver devices. Therefore, differences in communication channels also affect the deployment of semantic models. Moreover, in the design of the coding and decoding scheme, which is an important component of the semantic communication network, we also need to carefully consider the impact of heterogeneity in devices.
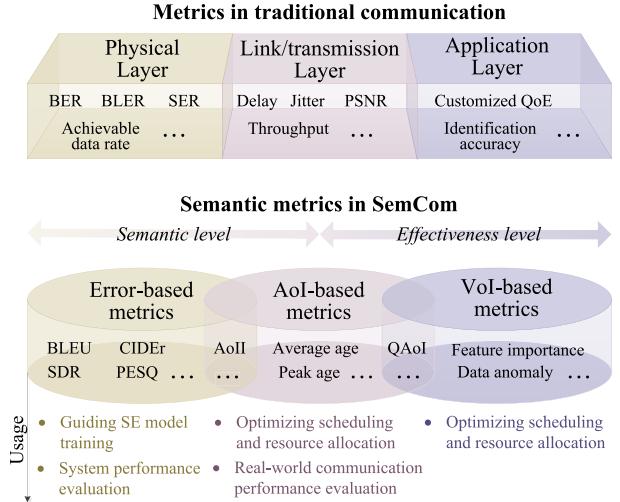

Fig. 11.   Main types of semantic metrics.

V. SEMANTIC PERFORMANCE MEASUREMENT AND CHALLENGES

Network performance measure's choices have always been a nucleus concern in network design and optimization for generations. In the conventional communication system, due to the separation of transmission and data's SI and effectiveness for achieving specific goals, the communication performance tends to be evaluated from different network layers through metrics such as BER, QoS, and QoE, respectively. In contrast, in SemCom, the interlayer coupling is enhanced to a great extent [153]. Hence, the new methods of evaluating communication performance in terms of semantics must be identified before the implementation of SemCom in practice. Existing SemCom evaluations mainly focus on semantic error, AoI, and VoI. Next, we go into the details about the three basic metric types and their related combined forms as shown in Fig. 11, as well as discuss the related remaining issues.

A. Error-Based Semantic Metrics

As aforementioned, different from the metrics of BER and SER for traditional communications, which are concerned with the accuracy of each bit and each symbol and treat all the bits and symbols as equally important, the error-based metrics in the SemCom care about whether the meaning intended by the transmitter is equivalent to the meaning understood by its destination, i.e., *semantic similarity* [11]. Moreover, the available semantic metrics are all task-specific and there is not yet a general metric for different types of embedding [12]. Below we discuss the error-based semantic metrics in terms of specific applications.

1) Semantic Metrics for Text Data: At the moment, text transmission has received the most attention in the study of SemCom. Semantic similarity in text transmission usually refers to the exact degree of meaning conveyed by a whole sentence. To mathematically quantify the similarity, some researchers resort to some pioneering works in NLP, and apply the following metrics in the performance evaluation of their works.

- *Bilingual evaluation understudy (BLEU):* Initially, BLEU is a method for automatic evaluation for machine translation [154], which is in line with what semantic measurement needs to do in the SemCom system. BLEU is used to compare *n*-grams of the candidate with the *n*-grams of the reference translation and count the number of matches, where *n*-grams represents the size of a word group. For example, for sentence "*cat is on mat.*" 1-gram: "*cat*," "*is*," "*on*" and "*mat*," 2-grams: "*cat is*," "*is on*"and "*on mat*". It is first introduced into SemCom in [37], where the *n*-grams precision score denoted by p_n depends on the difference between the minimum frequency of one element in the *n*-th grams. In this sense, the BLEU for the whole sentence is calculated as the product of the sum of the precision scores for the grams of all sizes and a brevity penalty (BP). The BP is determined by the length of the candidate (recovered) and reference (transmitted) sentences. The longer the candidate sentence is compared to the reference sentence, the lower the BLEU score is. Moreover, to make the ranking behavior more noticeable, BLEU is commonly used in its expression in the log domain.
- *Consensus based Image Description Evaluation (CIDEr):* CIDEr is proposed as an automatic consensus metric of image description quality in [155], which is originally used to measure the similarity of a generated sentence against a set of ground truth sentences written by humans. Hence, it can also be used as a semantic metric for text transmission [96]. Similar to BLEU, the similarity between two sentences is calculated based on the set of *n*-grams presented in it. The difference is that, in CIDEr, not just one reference sentence is considered, but a set of reference sentences. When calculating sentence similarity, it takes into account the similarity between the candidate sentences and all semantically similar sentences in the reference set.
- *Sentence similarity:* Sentence similarity is a new metric initialized in [37] and [138] for SemCom based on bidirectional encoder representations from Transformers (BERT). BERT is a state-of-the-art fine-tuned word representation model [147], which employs a huge pre-trained model including billions of parameters used for extracting the SI. Fed by billions of sentences, the performance of SI extraction has been demonstrated in [147]. To this end, the sentence similarity is calculated directly based on the cosine similarity of the semantic features extracted by BERT.

It is to be noted that, although the BLEU and CIDEr have considered some of the linguistic laws, such as that semantically consistent words usually come together in a given corpus [96], they remain at the level of calculating the differences of words between two sentences and have no insight into the meaning of the whole sentence [138]. In this sense, the metric of sentence similarity is much closer to desired SemCom paradigm, as the well-trained BERT model is sensitive to polysemy (e.g., the word "*mouse*" with a different meaning in biology and machine), which allows it to extract information in a sentence level.

On the other hand, the non-differentiability of these metrics reduces their practicality, since they cannot apply to DL-based SE, and the computation complexity of RL-based SE is considerable. Hence, even if both BLEU and sentence similarity have been proposed in [37], the training pipeline in the DL-based SemCom system still adopts CE loss. Moreover, for sentence similarity, the pre-trained BERT network embedding introduces much more resource consumption in the training process and makes it hard to generalize in other tasks.

2) Semantic Metrics for Audio Data: Similar to text data, the audio data is also very close to the human natural language. Thereby, for the SemCom for audio transmission, semantic similarity can be explained by the ease with which the receiver understands the decoded audio signal, i.e., *intelligibility*. Some similar works have been studied in the field of audio signal processing as below [156], [157], [158].

- *Signal-to-distortion ratio (SDR):* SDR is originally defined based on the usual definition of the SNR with a few modifications in [156]. In [81], it is introduced into SemCom the recovered signal as a performance metric, which is expressed by the $\mathcal{L}_2$ error between the transmitted speech signal and $\hat{s}$. Compared to MSE, the ranking behavior of the difference between $\mathbf{s}$ and $\hat{s}$ in SDR is more remarkable. Besides, the numerical precision of this measurement is lower for high-performance values than for low-performance ones, which is more intuitive for the design of the measurement method. However, SDR does not go any further than MSE in terms of semantic awareness.
- *Perceptual evaluation of speech quality (PESQ):* PESQ is a specialized quality assessment model designed for speech use across a wider range of network conditions, which has been standardized as Recommendation ITU-T P.862. It is utilized in [80] and [81] to evaluate the performance. It combines the perceptual speech quality measure (PSQM) and perceptual analysis measurement system (PAMS) [157]. The basic PESQ diagram of PESQ is shown in Fig. 12. Different from the above metric which simply compares the difference between the two signals, PESQ assumes the short memory in human perception, which allows it more similar to the human behavior [80]. However, the method still only looks at the accuracy of the transmission instead of the semantic meaning, and thus it cannot provide effective guidance for semantic compression.

In a nutshell, none of the above approaches evaluates the performance at the level of semantic understanding. In addition, only the MSE metric is used in DL for the SE in the existing works. The SemCom has only reached the semantic encoding level so far. A semantic measurement with semantic understanding like BERT and BLEU in the text transmission is still to be studied in the field of audio SemCom.

3) Semantic Metrics for Visual Data: For the communication for visual data, there are no general semantic metrics that are analogous to human perception yet. The commonly-used metrics in the SemCom for visual data are still shallow functions [159], such as PSNR [66] and structural similarity index (SSIM) [160] employed in conventional communication
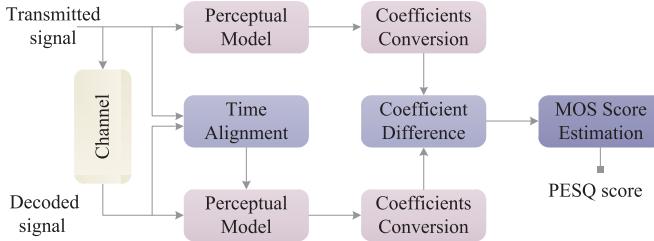
Fig. 12. Basic diagram of PESQ.

systems. Moreover, compared to the text and audio data, the semantic similarity is more context-dependent, that is, it is hard to distinguish different "senses of similarity": is a red circle more similar to a red square or a blue circle [159]? This imposes challenges to visual semantic metric design. Meanwhile, similar to text and audio data, the similarity judgment for visual data must also depend on a high-order structure [161]. To this end, DL-based feature capture can be considered as a potential way to assess the image semantic similarity [12]. In recent years, the internal activations of deep convolutional networks trained on a high-level image classification task have been considered to be often effective as a representational space for a much wider variety of tasks [159]. For instance, features from the Visual Geometry Group (VGG) architecture [162] have been used in other tasks like neural style transfer [163], and conditional image synthesis [164]. However, how to exploit this approach for SemCom performance evaluation needs to be further explored. In addition to image transmission which is only aimed at ensuring the fidelity of the visual data, there are also many emerging visual communications for specific tasks like object recognition and attribute classification, wherein the accuracy of task execution can directly characterize the effectiveness of SemCom.

B. AoI-Based Semantic Metrics

The difference between semantic information in communication and that of other fields such as semantic Web and semantic segmentation lies in its emphasis on time sensitivity. This feature introduces new dimensions to the accuracy of semantic information, i.e., *the right time* [165]. Especially for some applications, such as location tracking, control, and situational awareness, the freshness of information has a significant influence on the action execution at the receivers. In this regard, some metrics focusing on timeliness are required.

In fact, taking AoI into account in performance evaluation can be considered as an initial attempt at SemCom [166]. Different from the metric of delay, which primarily measures the transmission performance without concern for the content of the packets, AoI-based metrics are utilized to quantify the staleness of the information received at the destination. The age of a packet is defined as the difference between the current time and the timestamp of the packet [167], which captures how unfresh the data received by the monitor is. In the traditional content-blind communication paradigm, the systems just pursue to send updates as fast as possible and ensure the minimum transmission delay. Undoubtedly, this requires a lot

of bandwidth resources. Moreover, if the delay-QoS cannot be guaranteed, the backlog of the packets in the communication system throttles the update and leads to a monitor having unnecessarily outdated status information. Fortunately, such issues can be addressed by the scheduling scheme based on AoI minimization. This is attributable to the fact that the fresh data can be given more importance and transmitted with priority during the scheduling process.

Moreover, owing to the stochastic features of environments, an appropriate analysis of AoI can be chosen according to the specific system, such as *time-average age* and *peak age*, which are presented in detail in [168]. However, it should also be noted that there are still inherent flaws for AoI-based metrics, that is, they disregard the validity of the recovered data. For example, in some cases, the monitor is only concerned with the abnormal and abrupt states at sources [169]. Since AoI does not consider the value of current states for its monitor, some useless updates are transmitted to the monitor, which also results in a certain amount of resource waste.

C. VoI-Based Semantic Metrics

VoI is also a newly introduced metric in communication systems, especially for networked control systems. Before that, the concept of VoI is well-known in information analysis wherein it is defined as the price that a decision maker is willing to pay for taking the information into account [170]. For conventional communications, VoI can be defined as a measure of uncertainty reduction from the information set of the source with a successful transmission [171]. In contrast, for communications with specific tasks, the VoI needs to be redefined. In contrast to AoI which just focuses on the timeless and ignores content, VoI is mainly utilized to measure the relevance of a piece of information to the communication tasks. In other words, VoI can be regarded as the quantified contribution of SI to effectiveness.

Take a remote temperature control system [166] as an example. In that setting, the central controller is not concerned with the real-time temperature variation of the sources. The goal of the system is only to make sure the controller reacts swiftly to any abnormal temperature rise. In this sense, the data for the abnormal temperature should be assigned with high VoI. Moreover, in [108] wherein a task of image classification is studied, the VoI here is used to measure the importance of the extracted features to the accurate classification of the images. However, in most cases, the VoI can only be known after the task is completed. Here we revisit the two examples of semantic-aware communications in [122], [123]. The VoI of SI in the federated DRL task with multiple heterogeneous agents [122] is the reduction in convergence time as well as the increase of the long-term reward after convergence. Moreover, the VoI of SI in URLLC-enabled UAV swarm collaboration [123] is an improvement in convergence rate and reduction of inter-UAV collision probability. Therefore, for a general task, the VoI is challenging to be quantified as the AoI before communications, because the VoI is largely determined by a combination of multiple factors in the communication context. Therefore, the available VoI-based metrics
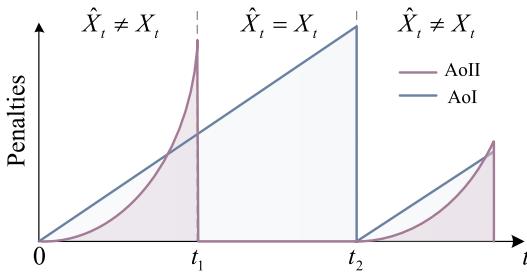
Fig. 13. Comparison of AoI and AoII, where X_t denotes the transmitted information, and $\hat{X}_t$ denotes the estimated information inferred from the transmission [166].


Fig. 14. AoI dynamics of the PQ and QAPA policies, where PQ denotes a permanent query system, which minimizes the traditional AoI, and QAPA denotes a query arrival process aware system, which minimizes the QAoI [165], [173].

are scarce, and the VoI-based scheduling or resource allocation scheme only performs for simple tasks, such as abnormality monitoring.

D. Combined Semantic Metrics

As aforementioned, the above three types of semantic metrics only focus on one attribute of the information conveyed by the recovered data. To address this limitation, burgeoning research efforts have been investigating new semantic metrics that combine multiple attributes to varying degrees [165].

The authors in [166] integrate AoI into error-based metrics, and propose a new metric named age of incorrect information (AoII). AoII characterizes the impact of the prolongation of one inaccurate state on semantic recovering. Compared to both error-based and AoI-based metrics above, AoII is incorporated with more meaningful semantics by jointly considering the content and timeliness. Specifically, AoII considers not only the repercussions of a transient state on the overall communication goal, but also the repercussions of the states lasting for different duration. For instance, the repercussions of a long burst of error are far more severe than an instantaneous burst of error for video transmission [172]. The comparison of the AoI and AoII is shown in Fig. 13. Meanwhile, the authors in [173] integrate VoI into the AoI-based metrics by considering a pull-based system and propose a new metric called age of information at query (QAoI), which reflects the freshness of the instants when the receiver actually needs the data [165]. In a pull-based system, the information is effective at the receiver only at certain query instants. In this sense, the communication is expected to be query-driven, that is, the transmitter knows the query instants and optimizes the transmissions with respect to the timing of the query process. Hence, the query-driven QAoI-based on scheduling scheme is more efficient and effective for such a system than the query-blind AoI-based one to achieve sending just before the query instants. As shown in Fig. 14, the QAoI-based scheme is more likely to provide updates that are fresh when a query arrives, although its average AoI may be worse than the AoI-based scheme [173].

E. Lessons Learned Summary

1) Lessons Learned for Error-Based Semantic Metrics:
As discussed in Section V-A, the research on error-based performance evaluation methods is just beginning and is lagging behind research on SE and SI transmission. Compared to
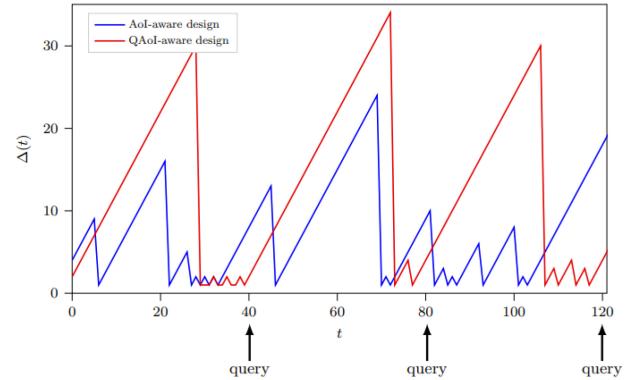
the metrics, such as BER, and SER, used in traditional communications to quantify transmission accuracy, the semantic error metrics summarized in Table VII avoid the shallow bit-by-bit, symbol-by-symbol rigid comparison. When comparing the difference between the received signal and the transmitted signal, these semantic metrics further take into account the characteristics of the data type itself, such as the syntactic structure of the text or the positive impact of the short memory in human perception on speech signal intelligibility. However, most of the above metrics are still comparing the raw signals sent and recovered, albeit from a more personalized and specialized perspective, rather than a direct comparison of the SI those imply. An ideal semantic metric might be similar to sentence similarity proposed in [37] and [138], and also implies an inevitable consumption of computational resources and time for the pre-training process of the relied AI model.

More significantly, it is important to point out that all the existing semantic metrics are *intrusive*, which involves a reference signal to accomplish the performance evaluation. Since a clean reference signal is always unavailable in real-world communications, these metrics can only play a role in the training phase of SE models based on existing databases. Meanwhile, as discussed in Sections III-B and V-A1, the non-differentiable, complex form of these metrics makes the training process more challenging. Hence, the practicality of these metrics is limited. Moreover, a non-intrusive semantic measurement for real-world testing, which can blindly and intelligently estimate the objective scores, such as BLEU, and PESQ, remains to be studied.

2) Lessons Learned for Combined Semantic Metrics:
When conducting performance evaluation based on the three types of metrics in Fig. 11, the data is given different importance. Since the error-based type of semantic metrics aims at comparing the meaning conveyed by the transmitted and recovered messages, it can be treated as the metric designed for semantic-oriented communication. In contrast, VoI-based metric type goes beyond the accuracy comparison from the semantic level, and instead directly evaluates the relevance of the information to task performance from the effectiveness level. In this sense, VoI-based metrics are for goal-oriented communications. Moreover, the AoI-based metrics fall between

TABLE VII
ERROR-BASED SEMANTIC METRICS SUMMARY

	Semantic metrics		Advantages	Drawbacks
text data	BLEU	BLEU is a method for automatic evaluation for machine translation. It is used to compare word groups of different sizes of the candidate with that of the reference translation and count the number of matches.	It considers the linguistic laws, such as that semantically consistent words usually come together in a given corpus.	It only calculates the differences of words between two sentences and has no insight into the meaning of the whole sentence.
	CIDEr	CIDEr is proposed as an automatic consensus metric of image description quality, which is originally used to measure the similarity of a generated sentence against a set of ground truth sentences written by humans.	Compared to BLEU, it does not evaluate semantic similarity on the basis of a reference sentence, but a set of sentences with the same meaning	Similar to BLEU, it is also based on the comparisons between word groups, and the semantic similarity can only be made at the word level.
	Sentence similarity	Sentence similarity is calculated as the cosine similarity of the semantic features extracted by bidirectional encoder representations from transformers (BERT) [37] from different sentences.	The SI considered in this metric is viewed from a sentence level owing to the sensitivity of BERT to polysemy.	BERT is a huge pre-trained model, which introduces much resource consumption in the training process and makes it hard to generalize in other tasks
audio data	SDR	SDR is expressed by the $\mathcal{L}_2$ error between the transmitted audio signal and recovered audio signal.	The numerical precision of SDR is lower for high performance values than for low performance ones.	SDR fails to capture the hidden SI of the speech signal, without any further than MSE in terms of semantic awareness.
	PESQ	PESQ is a specialized quality assessment model designed for speech used across a wider range of network conditions, which has been standardized as Recommendation ITU-T P.862.	Instead of comparing the differences between the two signals directly, PESQ assumes the short memory in human perception.	PESQ still focuses on transmission accuracy, and thus it cannot provide effective guidance for semantic compression.

the error-based metrics and VoI-based metrics, as both the meaning of the message and its impact on task execution may change over time. Combining AoI-based metrics with error-based metrics, the relevance of information can be further differentiated in terms of the time dimension. In an ideal SemCom system, the combined metrics are expected to play a role in resource allocation to guide the filtering of irrelevant information and enhance system efficiency and performance. However, this idea has only been explored in the simplest pull-based system, and the reason is that the quantification of SI has not been achieved, and the most error-based metrics and VoI-based metrics are challenging to estimate before transmission.

VI. SEMCOM FOR FUTURE 6G INTERNET

In this section, we first discuss some potential applications for SemCom in 6G. Then, we motivate a promising SemCom-empowered network architecture in 6G.

A. Potential Applications for SemCom in 6G

1) Intelligent Transport System: In recent years, with the development of hardware in vehicles and vehicular infrastructures, vehicles can be considered as intelligent agents with greater computing, caching and data storage capacities [174], [175]. This paves the way for future 6G intelligent transport systems (ITS), wherein autonomous driving and cooperative vehicle networks can be achieved without the need for human involvement [176]. In most of the existing works, in order to enhance safety, improve assisted driving decisions, or manage vehicles, the basic information about the vehicles and

road, such as locations, braking intensity, potholes, and water puddles, is required to be broadcast periodically [36]. Such a non-differentiated vehicle communication for different situations can affect communication effectiveness and efficiency. To this end, SemCom has a great potential to make ITS more intelligent.

In ITS, the most straightforward application of SemCom is to extract the essential semantic information from the raw sensor data, such as vehicle kinematic information, road conditions, and traffic signs (as shown in Fig. 4). For instance, there are many situations, such as a sudden appearance of a human and a sudden collision happening in front of a vehicle, that have a similar effect on vehicle driving. Therefore, by extracting the SI of a particular situation (i.e., quantifying the impact of each situation on driving), the accuracy and timeliness of the transmission can be greatly improved because of the reduced amount of data. In addition to the compression of the data itself, as discussed in Section V-B, the time point of data sampling is also critical in SE. For example, when a vehicle has a sufficiently long visibility range or sufficient viewing duration along every direction, it can make the driving decision based on its own local. In this case, the driving information about other vehicles is not required to be exchanged frequently. In contrast, if a vehicle's visible duration is very short, then when the vehicle gets the vision, it may not have sufficient information or time to make a correct reaction. In such a case, information exchange needs to be more timely. More specifically, the essential information about safety, such as braking intensity, should be given a higher priority [177]. Moreover, as discussed in Section II-B3, to monitor the traffic flow of walled subdivision, the SI about a continuous view over the

subdivision exit can be shared with the vehicles about to pass the subdivision. To this end, determining when to communicate, and what information to share in a "quantitative" manner is exactly what SemCom is required to do.

2) Distributed Learning Based Applications: With increased computing power on end devices and growing privacy concerns of users, distributed learning, such as federated learning (FL) has become a dominant paradigm for privacy-preserving machine learning [178]. It has penetrated all aspects of human's lives [179], such as medical diagnosis, cyberattack detection, and BS association. As deep neural networks usually contain millions of weight parameters, the frequent exchanges of DNN models or gradients between terminals and servers incur costly communication overheads, which poses challenges for communication networks, especially for the uncertain wireless environment and limited wireless resources.

Fortunately, SemCom can reduce unnecessary communication overhead in two ways, thereby improving performance within limited wireless resources. Firstly, the model parameters and gradients can be compressed in a semantic-aware manner, such as gradient sparsification [180] and model parameter pruning [181], where a subset of the original model parameters is extracted considering the semantics or importance of the parameters for model accuracy and convergence speed. For example, in [180], gradient sparsification is adopted to compress the model at the transmitter by setting all but k elements with the highest magnitudes of entries to zero. Since only the positions of the non-zero elements are to be sent, the receiver can recover the received data in a more reliable manner with advanced noisy measurements. Secondly, as per the two examples of semantic-aware communication in Section III-E, the agents in distributed learning can exchange their own semantic features via SemCom to enhance their knowledge of each other. The semantic features can be extracted from their learning models [122], their observable environments [123], their missions, and so on. Based on the semantic features of small data volume, an optimal subset of semantically-related agents for collaboration can be found. Therefore, the exchange of models with large data volumes between irrelevant agents can be effectively reduced.

3) Unmanned Aerial Vehicles: UAVs have attracted lots of attention especially while being served as aerial base stations (BSs) or as relays [182], [183], [184]. Unlike static ground BSs or relays, UAVs can be flexibly deployed to satisfy various QoS requirements and balance load amongst users. Moreover, UAV swarms collaboratively can complete missions with greater efficiency and economy as compared to single UAV systems. However, the energy constraints of UAVs impede their ability to facilitate long-term communication. Meanwhile, the collision problems in UAV swarm navigation have also been a key concern in the studies of UAV networks.

Fortunately, because SemCom can reduce the amount of information that needs to be transmitted, an efficient communication framework among UAVs can be implemented. For example, when the UAVs are being served as relays, diversity gain can be achieved with the help of cooperative communication protocols, such as decode-and-forward and amplify-and-forward [185]. In this sense, a novel semantic process-and-forward method is motivated to be proposed to cater to SemCom. In addition to the traditional relay function, the UAV can be deployed as a semantic encoder and/or decoder. For example, when one of the sending and receiving parties cannot enable SemCom due to insufficient memory or computing power, UAV can carry out encoding or decoding instead to reduce the data volume of a certain link without compromising communication performance. Certainly, this requires the UAV to know the background knowledge of both sides of the communication. This also poses new challenges for the joint optimization of communication, computing, and caching resources. Moreover, in the case where both parties can enable SemCom, the UAV can perform the semantic decoding for the received signal based on the background knowledge of the transmitter, and then it can re-code the signal based on the background knowledge of the receiver. This greatly reduces the overhead of synchronizing background knowledge for sending and receiving and the semantic noise caused by unsynchronized background knowledge [186]. Moreover, as discussed in Section III-E, SemCom can also play an essential role in UAV swarm navigation. As shown in [123], by introducing SemCom, the UAV swarm navigation based on graph attention exchange network can achieve 6.5x lower latency with the target 10^{-7} error rate compared to the state-of-the-art CTDE based method.

4) Extended Reality: Advances in 6G network technologies provide technical support for next-generation Internet services. In particular, the possibility of synchronization of the physical and virtual world through extended reality (XR) has led to the birth of Metaverse, which has been dubbed as the successor to the Internet. The performance of XR is heavily dependent on the collection and processing of data that reflects or describes human movements and changes of surroundings, e.g., shifting rendered targets, displaying particular videos, and giving the corresponding tactile feedback. To guarantee the ideal immersive Metaverse service experience for users, the end-to-end latency and data rates requirements have to be strictly met [187], [188].

To this end, SemCom can be seen as an enabler for XR-based Metaverse access [189]. In the SemCom paradigm, the data tracked by the end devices, such as head movement, arm swing, gestures, and speeches, need to be extracted semantically first. This allows the end device to transmit the information concerned by the XR server for operation after understanding and filtering out the irrelevant information to save bandwidth and reduce computing latency at the XR server. Meanwhile, the XR server can also extract SI based on the user's preference, ignoring irrelevant details in the face of bandwidth constraints, thus reducing downlink pressure.

5) Holographic Telepresence (HT): As another technique to deliver next-generation services to users, holographic telepresence (HT) can project realistic, full-motion, real-time three-dimensional (3D) images of distant human beings or physical objects with a high level of realism rivaling of the physical presence [176], [190]. It can be applied not only to virtual conferencing, virtual games, and the realm of entertainment, but also to remote repair, and remote

surgery [176], [191]. However, like the immersive XR application, to ensure a real-enough virtual and seamless service experience, HT also requires stringent QoS. Moreover, almost all the human senses, such as smell and taste, are expected to be transmitted through future networks for a fully immersive experience.

For such communication-and-computation intensive services, the traditional content-blind communication paradigm results in a waste of bandwidth resources and computing resources. To this end, the understanding-before-transmission paradigm of SemCom can be regarded as a promising method to alleviate the pressure of bandwidth and the receiver's end processing. This is attributable to the fact that SemCom allows the SI extraction that is adaptive to network conditions and enhances transmission reliability, so that changes in the network conditions cannot be perceived by users, thus ensuring high quality service experience.

6) Personalized Body Area Networks: Personal data management as well as transmission of wearable devices are future trends that will affect how personal services and procedures develop. An important element is the wireless body area network (WBAN). Defined formally by the IEEE 802.15 (Task group 6) as a communication standard optimized for low-power devices, the WBAN can serve a variety of applications such as medical, consumer electronics, and personal entertainment [190]. Because of the energy constrained power supplies of tiny sensor nodes, effective energy consumption is a key challenge in WBAN. SemCom prompts us to think about whether we can save energy and increase the lifetime of wearable devices by reducing the number of actual bits transmitted. In WBAN, it has been shown that the on-board extraction of features on modern low-power wearables is both feasible and beneficial for system lifetime improvement [191]. Although a resource-constrained sensing system needs to strike the balance between the accuracy of the semantic features output and the cost of analyzing the data for extraction, the benefits from reducing the radio duty cycle which is used for transmission, vastly outweigh the cost of increasing the processor duty cycle which is used for semantic features extraction [192]. As knowledge extraction from the raw data can significantly reduce the information that needs to be transmitted, the SemCom-based method increases the lifetime of the wearable device by one order of magnitude, at the cost of approximately 5% degradation of classification accuracy [192]. The development of SemCom and the deeper integration with WBAN will give rise to longer-lasting and more convenient wearable devices.

7) Collaborative Robots: A group of cooperative robots can explore, interact with, and perceive the environments far more efficiently than a single robot working alone [193]. In scenarios such as disaster management, warehouse automation, and surveillance, the application of collaborative robots is rapidly increasing. However, the limited computation capability of each robot limits the widespread deployment in computation intensive tasks [194]. One promising solution is to apply the SemCom techniques to achieve efficient data exchanging and processing. SEMIoTICS, a new SemCom-based control system architecture is proposed in [195], which enables the utilization of logic-based reasoning over declarative language models to reduce the decision-making time. In [195], SEMIoTICS is deployed in a building that consists of 15 IoT components for temperature regulation control. The results show that the overall control processing time can be maintained in 6 minutes, which is only 24% of that of the traditional fuzzy logic control-based method [196]. To further reduce the communication overhead among collaborative robots, a lite distributed SemCom system, named *L*-DeepSC [74], can be used. When the communication environment among collaborative robots experiences low SNR, *L*-DeepSC can enable efficient information exchange. In particular, with *L*-DeepSC, the amount of data needed to interact among robots can be compressed to 2.5% of the information that is needed by the traditional method.

8) Hyper-Intelligent IoT: Hyper-Intelligence (HI) refers to higher- and super-intelligent abilities to accomplish complex tasks. The combination of HI and IoT will lead to a smarter and data-driven society [197], [198]. A general and reasonably predictable trend in the next few years will be a rise in the native intelligence of networks, network nodes, and linked devices. Devices that are formerly utilized merely as sense-and-transmit entities will be endowed with various levels of embedded intelligence that operate directly on the data acquired. The necessity for progressively smarter communication parties opens the door to the creation of "smarter" content to exchange and reason, as well as innovative methods for ensuring that it is done effectively and accurately [199], which coincides with the booming development of SemCom. With the help of SemCom, only the most useful data is transmitted, and therefore the communication effort is optimized. To deploy the SemCom techniques in HT IoT to promote efficiency, a potential solution would be to consider whatever enables the receiver to effectively execute a given task, while relying on SemCom that extracts only the necessary information from the data. By giving HI the ability to process and reason information at the semantic and even effective level, the HI IoT will be more connected, further advancing the construction of an interlinked and connected society.

B. SemCom-Empowered 6G Architecture

In the traditional communication network, the network nodes are not concerned with what the data is trying to convey. The information exchanged between the inter-and intra-layer nodes in the network can be generally seen as homogeneous bit sequences. However, in the SemCom architecture with ubiquitous consciousness, the information representation can achieve a higher level. In line with the concept of "Bit" in Shannon's information theory, the authors in [14] introduce a new concept called "Seb" for SI. A SemCom system dimensioned by Seb is highly modulated compared to traditional communication systems.

In analogy to the construction process, a bit-flow-based network architecture is like a building constructed in a brick-by-brick manner, while a Seb-flow-based network architecture
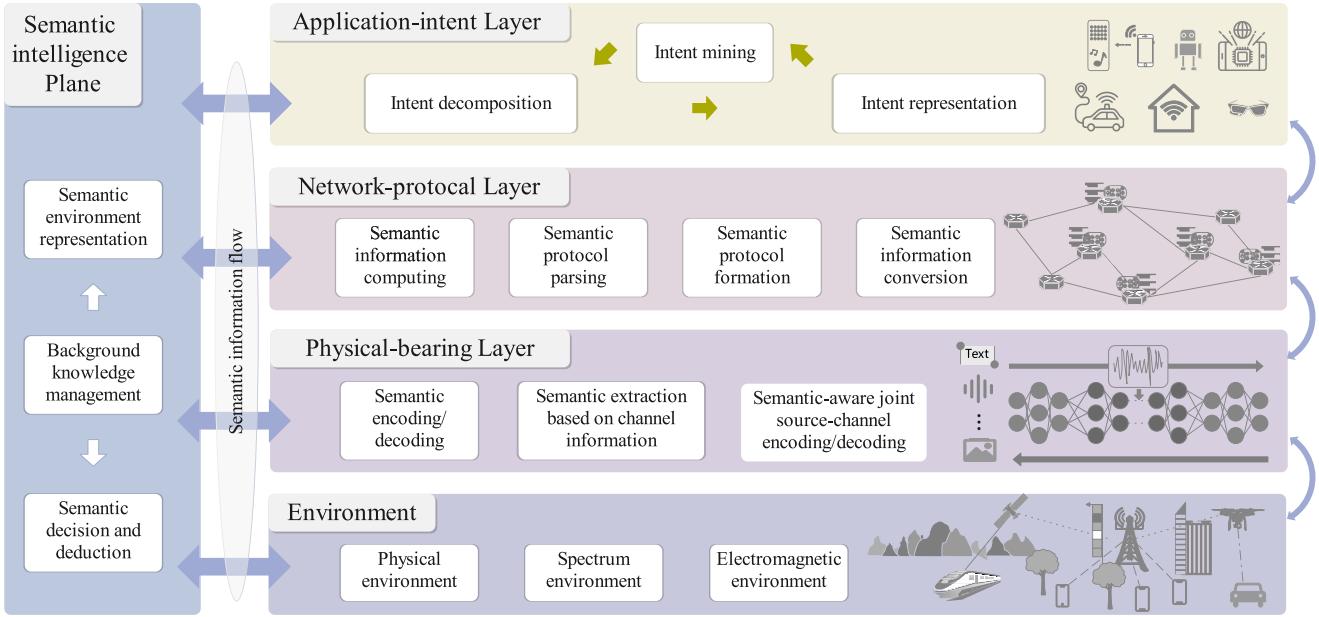
Fig. 15.   Semantic communication-empowered 6G architecture [14].

is similar to the building constructed by the laminboard and integrated window or door [14]. As the integration of raw materials simplifies the construction process, the SE of raw data also calls for a simplified network architecture to efficiently support SemCom.

To this end, the authors in [14] propose a novel intelligent and efficient SemCom (IE-SC) architecture, as shown in Fig. 15. In contrast to the well-known seven-layer open systems interconnection (OSI) model, their proposed IE-SC architecture divides the network into three semantic-empowered layers: semantic application-intent (S-AI) layer, semantic network-protocol (S-NP) layer, and semantic physical-bearing (S-PB) layer. Meanwhile, the three layers, as well as the physical environment, are coordinated by a separated semantic intelligence plane via the semantic information flow (S-IF). In the following, we first introduce the three main functions of the semantic intelligence plane. Then, we present the other three layers with examples in detail.

1) Semantic Intelligence Plane: As a coordinator of the entire IE-SC network, the semantic intelligence plane contains three main functions:

- *Semantic environment representation:* This function deals with the extraction of SI from the raw data and communication context based on the semantic intent information provided by the S-AI layer, which is detailed in Section VI-B2. Meanwhile, the semantic intelligence plane generates the semantic extraction components, such as SemCom protocol design and semantic encoding and decoding, and maps them to the functions of each layer. Meanwhile, SI is embedded into the S-IF to be transferred in the network through the semantic intelligence plane and the interfaces of different layers.
- *Background knowledge management:* Just as human communication requires knowledge of each other's language and culture, SemCom requires that both communicating parties have the same background knowledge, which lays

the foundation of the intent analysis of the S-AI layer. The semantic intelligence plane serves as a coordinator, which performs the synchronization, integration, and storage of background knowledge.

- *Semantic decision and deduction:* The semantic intelligence plane is also responsible for evaluating the achievable performance according to the results of the intent analysis fed by the S-AI layer and performing decision-making for all network layers.

2) Semantic Application-Intent Layer: In the traditional OSI model, the application layer mainly allows users to download data from storage or dataset, or send data to a destination regardless of what and how the data is used. In contrast, the S-AI layer adds the function of intent analysis for a particular communication. It can support decomposing and translating the users' intent into the network's deployment, configuration, or control policies. The three main functions within the S-AI layer are highlighted below.

- *Intent mining:* The S-AI layer is responsible for extracting, analyzing, aggregating, and synthesizing the original intents received from users or applications. For example, for an image, some requesters may use it for target identification, some may be concerned with local details, and some may be interested only in image quality. Different intents may correspond to different communication processes.
- *Intent decomposition:* In order to facilitate the subsequent SE from raw data, the obtained intent after intent mining is decomposed into a set of sub-intents, which may be the attention of the requester to different paragraphs, sentences, words in a text, or the importance of different features of an image. Moreover, recall an application in ITS, the intent of which is to provide vehicles that are about to pass through a walled subdivision with information about the traffic flow of the subdivision exit. During the intent decomposition, the sub-intents can be

a series of continuous views of the subdivision exit. For guiding the implementation of the entire communication process, the sub-intents can be informed to the S-NP layer and S-PB layer and used to design protocols and SE, respectively.

- *Semantic representation:* After obtaining the sub-intent set, the S-AI layer gives the semantic representation of the sub-intents. The semantic representation is passed to the semantic intelligence plane, facilitating SE.

3) Semantic Network-Protocol Layer: The S-NP layer aims to efficiently serve the intents of upper-layer applications with intelligent network protocols. The design of this layer mainly concentrates on the strategies of semantic interaction, which includes experience accumulation for learning (e.g., multiple rounds of dialogues), real-time knowledge sharing, and simplified semantic interaction. In this sense, some key modules should be included in the S-NP layer:

- *SI computation:* In traditional communication, most protocols mainly focus on the destination address and port, and ignore the content of the data. However, in the SemCom-enabled network architecture, if the intermediate forwarding nodes still treat all data equally, it would compromise the analytical significance of the application. To this end, this module is for identifying the intent information on the S-IF and obtaining knowledge from other modules.
- *Semantic protocol parsing:* This module is used to analyze the functions available for the existing protocol.
- *Semantic protocol formation:* This module is responsible for optimizing the existing protocol or forming a new one to adapt to the intent of the application.
- *SI conversion:* In the traditional network architecture, the protocols of the different network layers have individual frame encapsulation, packet encapsulation, segment encapsulation, etc. Similarly, in SemCom-enabled protocols, SI needs to be encapsulated separately. This module is responsible for completing the SI encapsulation based on the generated protocol.

Take the immersive AR application as an example, which is the application with the coexistence of video, audio, and haptic data and requires high bandwidth and ultra-reliable low latency. Suppose that by interacting with the semantic intelligent plane, the S-NP layer can map the sub-intents, such as the attention to different sound sources and different objects in the visual field, as well as the haptic data with highly critical latency requirements, to the data within the S-IF. For the haptic packets, the traditional design of the medium access control frame usually reserves dedicated resources for ensuring that the delay QoS requirement is met. However, the request of the haptic packets is random and intermittent, which inevitably leads to low resource utilization or transmission failures. In order to address the above issues, a semantic protocol should be designed to coordinate the transmission of the data corresponding to different sub-intents. For example, due to the tightest time delay requirement, the haptic data are given the highest priority when being forwarded. Moreover, when the network is under heavy load, the redundant data with less attention from the user can be detected and removed

adaptively, and resources can be freed for delay-sensitive traffic [153].

4) Semantic Physical-Bearing Layer: The S-PB layer is responsible for converting SI from the upper layers into physical signals. Different from conventional communication, SemCom aims to deliver SI instead of raw data, so the traditional source encoding/decoding and channel encoding/decoding are to be replaced. Three possible types of semantic-related modules employed in the S-PB layer are listed below:

- *Semantic encoding/decoding:* Following the modular design method, the modules of semantic encoding and decoding are designed separately from other modules, such as channel coding. For example, as presented in Section III-C, the authors in [108] propose a pair of semantic encoder and decoder based on a well-established KB. Meanwhile, the channel encoder and decoder are still performed in the same way as in traditional communications.
- *Semantic encoding/decoding based on channel information:* Integrate the channel information in terms of fading, SNR, and interference into semantic encoding and decoding. As discussed in Section III-A, in [66], the authors apply SNR to SE by introducing the attention module, which allows the SE method to be adaptive to the changes in channel gain.
- *Semantic-aware joint source-channel encoding/decoding:* Following the integrated design method, channel encoding and decoding can be jointly integrated into the modules of semantic encoding/decoding. In fact, most of the existing end-to-end SE methods [37], [64], [80] discussed in Sections III-A and III-B belong to the type of semantic-aware joint source-channel encoding/decoding.

VII. FUTURE DIRECTIONS

In the previous sections, we review the potential SemCom applications in 6G and the state-of-the-art techniques applied in SemCom. In addition to the remaining issues discussed in Sections III–V, several other SemCom-related directions can be explored further in terms of system effectiveness, sustainability, and trustworthiness.

A. Interpretability and Explainability of SE

The communication environment is always experiencing a variety of uncertainties, such as unexpected changes in the network environment or completely new source information. The black box nature makes the SE model unpredictable for the output corresponding to uncertain inputs in practice, which restricts the SE model's social acceptance and practicality, as well as leaves little basis to use as a guide for SE model optimization. Meanwhile, the available SE models are with little or no understanding of how and why the internal states in the hidden layers and the features contribute to a given example to produce a decision or outcome [200], which fails to give valuable insights into the design of SemCom systems and the

SI transmission. Therefore, the issues related to interpretability and explainability of SE have to be addressed.

As defined in [201], interpretability is used to measure the degree to which a human can consistently predict the model's decisions. Gaining an insight into how and why the SE model arrives at a particular decision or outcome not only builds confidence in the model to deal with unknown situations, thereby reducing the risk of uncertainty, but also helps to understand the overall strengths and weaknesses of the models and guides improvements to the model [202]. In contrast to interpretability, the study of explainable AI focuses on the hidden states in DNN and aims to open up the black box. For example, the contribution of each input semantic feature to the accuracy of the semantic inference can be quantified by analyzing the gradient information of the semantic decoder. Based on this, the radio resource allocation at the sender can be achieved with a more flexible and fine-grained implementation, such as allocating the crucial semantic feature higher transmitting power to ensure its transmission reliability and the accuracy of the semantic inference.

B. Tradeoff Between SE Accuracy and Communication Overhead

Most of the existing works focus on how to perform accurate SE to save radio resources and enhance communication performance, while ignoring the extra communication overhead for SE. In fact, SE model training and updating require significant additional resources. For example, the training of accurate semantic extraction models relies on a complete KB with both senders and receivers, which requires, first of all, adequate storage resources. In addition, as the communication context evolves, each user's local KB is constantly being updated individually. In this sense, ensuring that updates to the local database of all communicating participants can be shared in real time is extremely challenging, especially for the case with a large number of participating users who are geographically distant, which can cause significant communication overhead. Moreover, in an ideal case, retraining or fine-tuning of the SE model needs to be done promptly after the KB update. However, this is unrealistic for practical systems with limited computational resources. Therefore, making a favorable trade-off between SE accuracy and communication overhand is essential for the implementation of SemCom.

For example, we can utilize edge intelligence to train the SE model based on the shared KB of local senders and receivers stored in the MEC server of the local area first. Then, with the help of the distributed learning paradigms, such as federated learning, a generalized SE model can be obtained by aggregating multiple well-trained SE models for different geographical areas. In this way, storage resources scattered around the edge can be efficiently utilized to reduce storage pressure on end devices or the central cloud, and the communication overhead caused by sharing data over long distances can be greatly reduced. However, the reasonable division of geographical areas and the strategic deployment of edge servers are still to be explored. Moreover, the aggregation period and the participants selected in each round are also the essential issues that can be optimized in making a tradeoff between SE accuracy and communication overhead.

C. Combination of SemCom and Semantic Caching

In traditional communications, the implementation of data caching on the router, MEC server, base station, etc., has already shown to be of great benefit in avoiding unnecessary delay and network overhead for [203]. By joint optimizing caching and communication, with a 1% increase in cache hits, the perceived latency is reduced by 35% [204]. However, traditional caching in terms of raw data is no longer ideally suitable for SemCom systems, as the frequent and repetitive semantic extraction of the raw data results in redundancy and inefficiency of the system. Meanwhile, the data volume of SI extracted is much smaller compared to the raw data. In this sense, semantic caching strategies that fit with SemCom cannot only enhance system efficiency, but also save memory resources.

However, there are new issues raised for semantic caching. For example, different from the traditional data caching, such as that mainly focuses on the hit rate of the data content, semantic caching is more concerned with whether the SI in the cache can be accurately inferred by the requester. Since there may be multiple SI for the same data content, which ones to cache demands more prior knowledge, such as the popularity of the specific SI. Moreover, as the context of SemCom is constantly changing, the lifetime of SI is more difficult to determine. To this end, it also requires new estimate refreshing algorithms for semantic caching.

D. Reasoning in Implicit SemCom

The majority of previous SemCom research focused on transferring explicit SI, such as the labels of things that can be directly identified from the source signals, e.g., images, voices, and texts. However, communication between users is not only limited to explicit information, but also contains rich implicit information that is difficult to express, recognize, or recover. For example in [205], a kid sends her father a voice message asking, "What is a Tweety?" The major semantic part of this message, "Tweety", might be interpreted in several ways, such as a smartphone app, a canary bird, or a character from a cartoon television program. Therefore, to deduce the message's exact meaning, the receiver must be able to infer the implicit information from the transmitter's context and background. Thus, it is unrealistic to assume that the destination user has a well-defined analytical expression, such as a reward function or utility function, which is directly optimized to maximize its understanding of the semantic meaning.

A few works have considered this point and tried to propose solutions. A generative adversarial imitation learning-based reasoning mechanism learning (GAML) is designed in [205] for the destination user to learn and imitate the reasoning process of the source user to obtain the implicit semantic meaning. It is shown that GAML can achieve significant error correction performance and offer 20% of accuracy improvement over genetic algorithm (GA)-based reasoning solution. In another work [206], the authors develop a novel inference

function-based approach that can infer hidden information such as incomplete entities and relations that cannot be directly observed from the message, where the solution achieves 76% and 48% of accuracy in recovering missing information when using additive and linear inference functions, respectively. However, both solutions in [205], [206] add additional inference overhead, and there is still room for further performance enhancement. Moreover, because explicit SI is typically dominant, the communication resources should be allocated proportionally between explicit and implicit SI, which inspires us to further design the joint optimization algorithms.

E. Artificial Intelligence in SemCom Channel Management

In SemCom, AI is more often deployed on the transmitter and receiver for coding and decoding to serve upper-layer applications. However, in the 6G wireless communication that has a higher data rate and more frequent handover, channel modeling becomes more and more complex than the traditional stochastic or deterministic approaches [207]. This prompted us to think about whether AI could be brought down to the SemCom channel layer to help model, estimate, and change channel conditions. Unlike simply applying AI to the end-to-end SemCom model training, the development of new intelligent materials gives AI more freedom in wireless channels [208]. It is believed that the radio environment in the future generation of wireless communication networks will become controllable and intelligent by leveraging the emerging technologies of reconfigurable metasurface (RMS) and AI [209]. RMS can effectively control the wavefront, e.g., the phase, amplitude, frequency, and even polarization, of the impinging signals. Through the use of AI-enable programmable intelligent materials, SemCom networks can further surpass the limits predicted by the classical Shannon theory by jointly optimizing the transmitter, the receiver, and the environment.

F. Tradeoff Between SemCom Performance and Security

Data security and privacy issues are always significant topics in the field of wireless communications [210]. Due to the fact that SemCom requires only partial data to be transmitted and the decoding of SI relies on the receiver's background knowledge, it has also been regarded as a potential method for secure communications [29]. In addition, the security of the data can be further enhanced by encrypting the extracted SI. However, this also leads us to consider the tradeoff between computational resource overhead and data security. One possible solution is to use physical layer security technologies. Considering the success of covert communication [211], we can make the data eavesdropper unsure whether the SemCom is ongoing by introduced interference to the physical layer for secure wireless transmission. However, although the computational resources for encrypting the data are reduced, we need to keep the transmitting power not too high to ensure the covertness of the communication. In addition, the interference signals have a negative impact on the transmission of SI, which brings a trade-off between covertness and signal quality.

VIII. Conclusion

In this paper, we have provided a comprehensive survey of SemCom for 6G. First, we have highlighted the mutually reinforcing properties of 6G and SemCom. Then, we introduced the development from SemCom-related theories and identified three types of SemCom. Next, we organized the design of the communication system into three dimensions of SI extraction, SI transmission, and SI metrics, and discussed the state-of-the-art techniques and challenges, respectively. Meanwhile, we have presented the potential applications of SemCom in the 6G network as well as the promising SemCom-empowered network architecture. Moreover, we highlight some future directions with insights for further in-depth investigations.

References

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

[2] R. Li et al., "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 175–183, Oct. 2017.

[3] M. Xu et al., "A full dive into Realizing the edge-enabled Metaverse: Visions, enabling technologies, and challenges," 2022, *arXiv:2203.05471*.

[4] W. Y. B. Lim et al., "Realizing the metaverse with edge intelligence: A match made in heaven," 2022, *arXiv:2201.01634*.

[5] H. Du et al., "Attention-aware resource allocation and QoE analysis for Metaverse xURLLC services," 2022, *arXiv:2208.05438*.

[6] H.-J. Jeon, H.-C. Youn, S.-M. Ko, and T.-H. Kim, "Blockchain and AI meet in the Metaverse," in *Advances in the Convergence of Blockchain and Artificial Intelligence*. London, U.K.: IntechOpen, 2022, p. 73.

[7] H. Du, B. Ma, D. Niyato, and J. Kang, "Rethinking quality of experience for Metaverse services: A consumer-based economics perspective," 2022, *arXiv:2208.01076*.

[8] 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, Apr. 2021.

[9] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Commun. Mag.*, vol. 59, no. 8, pp. 44–50, Aug. 2021.

[10] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 96–102, Jun. 2021.

[11] E. C. Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, May 2021, Art. no. 107930.

[12] Z. Qin, X. Tao, J. Lu, and G. Y. Li, "Semantic communications: Principles and challenges," 2021, *arXiv:2201.01389*.

[13] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 210–219, Feb. 2022.

[14] P. Zhang et al., "Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks," *Engineering*, vol. 8, pp. 60–73, Jan. 2022.

[15] J. Wu et al., "Toward native artificial intelligence in 6G networks: System design, architectures, and paradigms," 2021, *arXiv:2103.02823*.

[16] Y. Zhong, "A theory of semantic information," *China Commun.*, vol. 14, no. 1, pp. 1–17, Jan. 2017.

[17] C. W. Morris, "Foundations of the theory of signs," in *International Encyclopedia of Unified Science*. Chicago, IL, USA: Chicago Univ. Press, 1938, pp. 1–59.

[18] M. Ch, *Writings on the General Theory of Signs*. Hague, The Netherlands: Mouton, 1971.

[19] W. Weaver, "Recent contributions to the mathematical theory of communication," *ETC Rev. Gen. Semant.*, vol. 10, no. 4, pp. 261–281, 1953.

[20] Y. Bar-Hillel and R. Carnap, "Semantic information," *Brit. J. Philos. Sci.*, vol. 4, no. 14, pp. 147–157, 1953.

[21] P. Elias, "An outline of a theory of semantic information," Res. Lab. Electron., Massachusetts Inst. Technol., Rep. 247, 1954.

[22] L. Floridi, "Outline of a theory of strongly semantic information," *Minds. Mach.*, vol. 14, no. 2, pp. 197–221, Feb. 2004.

[23] S. D'Alfonso, "On quantifying semantic information," *Information*, vol. 2, no. 1, pp. 61–101, 2011.

[24] I. Niiniluoto, *Truthlikeness*, vol. 185. Dordrecht, The Netherlands: Springer, 2012.

[25] J. Barwise and J. Seligman, *Information Flow: The Logic of Distributed Systems*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[26] G. Oddie, *Likeness to Truth*. Dordrecht, The Netherlands: Springer, 1986.

[27] I. Niiniluoto, *Truthlikeness*. Dordrecht, The Netherlands: Springer, 1987.

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

[29] P. Basu, J. Bao, M. Dean, and J. Hendler, "Preserving quality of information by using semantic relationships," *Pervasive Mobile Comput.*, vol. 11, pp. 188–202, Apr. 2014.

[30] B. Juba and M. Sudan, "Universal semantic communication I," in *Proc. 40th Annu. ACM Symp. Theory Comput.*, 2008, pp. 123–132.

[31] C. Lund, L. Fortnow, H. Karloff, and N. Nisan, "Algebraic methods for interactive proof systems," *J. ACM*, vol. 39, no. 4, pp. 859–868, Apr. 1992.

[32] B. Juba and M. Sudan, "Universal semantic communication II: A theory of goal-oriented communication," *Electr. Colloq. Comput. Complex.*, vol. 15, no. 95, pp. 1–29, 2008.

[33] O. Goldreich, B. Juba, and M. Sudan, "A theory of goal-oriented communication," *J. ACM*, vol. 59, no. 2, pp. 1–65, Feb. 2012.

[34] B. Juba and S. Vempala, "Semantic communication for simple goals is equivalent to on-line learning," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2011, pp. 277–291.

[35] F. I. Dretske, *Knowledge and The Flow of Information*. Cambridge, MA, USA: MIT Press, 1981.

[36] H. Bista, I.-L. Yen, F. Bastani, M. Mueller, and D. Moore, "Semantic-based information sharing in vehicular networks," in *Proc. IEEE Int. Conf. Web Services (ICWS)*, 2018, pp. 282–289.

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

[38] M. Sana and E. C. Strinati, "Learning semantics: An opportunity for effective 6G communications," 2021, *arXiv:2110.08049*.

[39] A. Meroño-Peñuela et al., "Semantic technologies for historical research: A survey," *Semant. Web*, vol. 6, no. 6, pp. 539–564, 2015.

[40] L. Rachana and S. Shridevi, "A literature survey: Semantic technology approach in machine learning," in *Advances in Smart Grid Technology*. Singapore: Springer, 2021, pp. 467–477.

[41] M. Thoma, "A survey of semantic segmentation," 2016, *arXiv:1602.06541*.

[42] P. C.-Y. Sheu, *Semantic Computing*. Boston, MA, USA: Springer, 2010, pp. 1–9.

[43] P. Hitzler, "A review of the semantic Web field," *Commun. ACM*, vol. 64, no. 2, pp. 76–83, 2021.

[44] W. Chen, "Recommendation system based on semantic Web," in *Proc. Int. Conf. Mach. Learn. Big Data Anal. IoT Security Privacy*, 2021, pp. 511–517.

[45] C. She et al., "Deep learning for ultra-reliable and low-latency communications in 6G networks," *IEEE Netw.*, vol. 34, no. 5, pp. 219–225, Sep./Oct. 2020.

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

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

[48] 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.

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

[50] J. Guo, C.-K. Wen, and S. Jin, "CAnet: Uplink-aided downlink channel acquisition in FDD massive MIMO using deep learning," 2021, *arXiv:2101.04377*.

[51] S. Park, O. Simeone, and J. Kang, "End-to-end fast training of communication links without a channel model via online meta-learning," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2020, pp. 1–5.

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

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

[54] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[55] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[56] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3156–3164.

[57] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[58] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[59] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[60] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 206–219, Mar. 2019.

[61] T. Ogunfunmi, R. P. Ramachandran, R. Togneri, Y. Zhao, and X. Xia, "A primer on deep learning architectures and applications in speech processing," *Circuits, Syst., Signal Process.*, vol. 38, no. 8, pp. 3406–3432, 2019.

[62] R. Haeb-Umbach et al., "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 111–124, Nov. 2019.

[63] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning based semantic communications: An initial investigation," in *Proc. IEEE Global Commun. Conf.*, 2020, pp. 1–6.

[64] C.-H. Lee, J.-W. Lin, P.-H. Chen, and Y.-C. Chang, "Deep learning-constructed joint transmission-recognition for Internet of Things," *IEEE Access*, vol. 7, pp. 76547–76561, 2019.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[66] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, Apr. 2022.

[67] Q. Hu, G. Zhang, Z. Qin, Y. Cai, and G. Yu, "Robust semantic communications against semantic noise," 2022, *arXiv:2202.03338*.

[68] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021, *arXiv:2111.06377*.

[69] 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)*, 2018, pp. 2326–2330.

[70] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[71] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[72] A. Graves, "Sequence transduction with recurrent neural networks," 2012, *arXiv:1211.3711*.

[73] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[74] 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.

[75] 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.

[76] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser, "Universal transformers," 2018, *arXiv:1807.03819*.

[77] A. Graves, "Adaptive computation time for recurrent neural networks," 2016, *arXiv:1603.08983*.

[78] H. Tong et al., "Federated learning for audio semantic communication," *Front. Commun. Netw.*, vol. 2, Sep. 2021, Art. no. 734402.

[79] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," 2019, *arXiv:1904.05862*.

[80] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech signals," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[81] 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.

[82] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech recognition," 2021, *arXiv:2107.11190*.

[83] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[84] D. Amodei et al., "Deep speech 2: End-to-end speech recognition in english and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[85] Y. Jiang, S. Kannan, H. Kim, S. Oh, H. Asnani, and P. Viswanath, "DEEPTURBO: Deep turbo decoder," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2019, pp. 1–5.

[86] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented semantic communications for multimodal data," 2021, *arXiv:2108.07357*.

[87] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," 2021, *arXiv:2112.10255*.

[88] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Mar. 2015.

[89] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," 2018, *arXiv:1803.03067*.

[90] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," 2015, *arXiv:1511.06732*.

[91] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 7008–7024.

[92] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[93] D. Bahdanau et al., "An actor-critic algorithm for sequence prediction," 2016, *arXiv:1607.07086*.

[94] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 290–298.

[95] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 1–7.

[96] K. Lu, R. Li, X. Chen, Z. Zhao, and H. Zhang, "Reinforcement learning-powered semantic communication via semantic similarity," 2021, *arXiv:2108.12121*.

[97] K. Lu et al., "Rethinking modern communication from semantic coding to semantic communication," 2021, *arXiv:2110.08496*.

[98] D. Silver et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[99] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1008–1014.

[100] F. Carpi, C. Häger, M. Martalò, R. Raheli, and H. D. Pfister, "Reinforcement learning for channel coding: Learned bit-flipping decoding," in *Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, 2019, pp. 922–929.

[101] R. Luo, "A better variant of self-critical sequence training," 2020, *arXiv:2003.09971*.

[102] M. K. Farshbafan, W. Saad, and M. Debbah, "Curriculum learning for goal-oriented semantic communications with a common language," 2022, *arXiv:2204.10429*.

[103] Y. Wang, M. Chen, W. Saad, T. Luo, S. Cui, and V. Poor, "Performance optimization for semantic communications: An attention-based learning approach," in *Proc. IEEE Global Commun. Conf.*, 2021, pp. 1–6.

[104] F. Zhou, Y. Li, X. Zhang, Q. Wu, X. Lei, and R. Q. Hu, "Cognitive semantic communication systems driven by knowledge graph," 2022, *arXiv:2202.11958*.

[105] R. L. Rosa, G. M. Schwartz, W. V. Ruggiero, and D. Z. Rodríguez, "A knowledge-based recommendation system that includes sentiment analysis and deep learning," *IEEE Trans. Industr. Inform.*, vol. 15, no. 4, pp. 2124–2135, Apr. 2019.

[106] X. Lin, H. Li, H. Xin, Z. Li, and L. Chen, "KBPearl: A knowledge base population system supported by joint entity and relation linking," *Proc. VLDB Endowment*, vol. 13, no. 7, pp. 1035–1049, Jul. 2020.

[107] W. Zheng, L. Yin, X. Chen, Z. Ma, S. Liu, and B. Yang, "Knowledge base graph embedding module design for visual question answering model," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108153.

[108] Y. Yang et al., "Semantic communications with AI tasks," 2021, *arXiv:2109.14170*.

[109] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "GRAD-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[110] R. Fong and A. Vedaldi, "Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8730–8738.

[111] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[112] H. Seo, J. Park, M. Bennis, and M. Debbah, "Semantics-native communication with contextual reasoning," 2021, *arXiv:2108.05681*.

[113] C. K. Ogden and I. A. Richards, *The Meaning of Meaning: A Study of the Influence of Thought and of the Science of Symbolism*. Mansfield, CT, USA: Martino Publ., 1923.

[114] A. Lazaridou and M. Baroni, "Emergent multi-agent communication in the deep learning era," 2020, *arXiv:2006.02419*.

[115] A. Lazaridou, A. Peysakhovich, and M. Baroni, "Multi-agent cooperation and the emergence of (natural) language," 2016, *arXiv:1612.07182*.

[116] J. Mu and N. Goodman, "Emergent communication of generalizations," 2021, *arXiv:2106.02668*.

[117] J. Bell, *Pragmatic Reasoning–A Model-Based Theory*. Dordrecht, The Netherlands: Springer, 1992.

[118] N. D. Goodman and A. Stuhlmüller, "Knowledge and implicature: Modeling language understanding as social cognition," *Topics Cogn. Sci.*, vol. 5, no. 1, pp. 173–184, Jan. 2013.

[119] J. Kao, L. Bergen, and N. Goodman, "Formalizing the pragmatics of metaphor understanding," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 36, 2014, pp. 1–6.

[120] N. D. Goodman and M. C. Frank, "Pragmatic language interpretation as probabilistic inference," *Trends Cogn. Sci.*, vol. 20, no. 11, pp. 818–829, Nov. 2016.

[121] H. P. Grice, "Logic and conversation," in *Speech Acts*. Leiden, The Netherlands: Brill, 1975, pp. 41–58.

[122] F. Lotfi, O. Semiari, and W. Saad, "Semantic-aware collaborative deep reinforcement learning over wireless cellular networks," 2021, *arXiv:2111.12064*.

[123] W. J. Yun et al., "Attention-based reinforcement learning for real-time UAV semantic communication," in *Proc. 17th Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2021, pp. 1–6.

[124] Y. Tao, S. Genc, J. Chung, T. Sun, and S. Mallya, "REPAINT: Knowledge transfer in deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10141–10152.

[125] Á. Visús, J. García, and F. Fernández, "A taxonomy of similarity metrics for Markov decision processes," 2021, *arXiv:2103.04706*.

[126] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[127] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2145–2153.

[128] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," 2021, *arXiv:2112.10948*.

[129] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Kténas, N. Cassiau, and C. Dehos, "6G: The next frontier," 2019, *arXiv:1901.03239*.

[130] M. D. Yacoub, "The α-μ distribution: A physical fading model for the Stacy distribution," *IEEE Trans. Veh. Technol.*, vol. 56, no. 1, pp. 27–34, Jan. 2007.

[131] H. Du, J. Zhang, J. Cheng, and B. Ai, "Sum of fisher-Snedecor $\mathcal{F}$ random variables and its applications," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 342–356, 2020.

[132] J. Zhang, W. Zeng, X. Li, Q. Sun, and K. P. Peppas, "New results on the fluctuating two-ray model with arbitrary fading parameters and its applications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2766–2770, Mar. 2018.

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

[134] 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.

[135] J. Meinilä, P. Kyösti, T. Jämsä, and L. Hentilä, "WINNER II channel models," in *Radio Technologies and Concepts for IMT-Advanced*. Chichester, U.K.: Wiley, 2009.

[136] M. Ding, J. Li, M. Ma, and X. Fan, "SNR-adaptive deep joint source-channel coding for wireless image transmission," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 1555–1559.

[137] T. K. Moon, *Error Correction Coding: Mathematical Methods and Algorithms*. Hoboken, NJ, USA: Wiley, 2021.

[138] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Deep source-channel coding for sentence semantic transmission with HARQ," 2021, *arXiv:2106.03009*.

[139] P. Shrivastava and U. P. Singh, "Error detection and correction using Reed Solomon codes," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 8, pp. 965–969, Aug. 2013.

[140] Y. Xuan, C. Guo, C. Feng, and Z. Li, "Multi-graph based spectrum sharing scheme in vehicular network with integration of heterogenous spectrum," in *Proc. IEEE Intel. Conf. Commun. Workshops (ICC Workshops)*, 2019, pp. 1–6.

[141] L. Liang, S. Xie, G. Y. Li, Z. Ding, and X. Yu, "Graph-based resource sharing in vehicular communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4579–4592, Jul. 2018.

[142] A. Khan, L. Sun, and E. Ifeachor, "QoE prediction model and its application in video quality adaptation over UMTS networks," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 431–442, Apr. 2012.

[143] C. Xu, F. Zhao, J. Guan, H. Zhang, and G.-M. Muntean, "QoE-driven user-centric VoD services in urban multihomed P2P-based vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2273–2289, Jun. 2013.

[144] Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, and S. Sun, "Economics of semantic communication system in wireless powered Internet of Things," 2021, *arXiv:2110.01423*.

[145] P. Ramezani and A. Jamalipour, "Toward the evolution of wireless powered communication networks for the future Internet of Things," *IEEE Netw.*, vol. 31, no. 6, pp. 62–69, Nov./Dec. 2017.

[146] S. H. Chae, C. Jeong, and S. H. Lim, "Simultaneous wireless information and power transfer for Internet of Things sensor networks," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2829–2843, Aug. 2018.

[147] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[148] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proc. IEEE Intel. Conf. Comput. Vis.*, 2017, pp. 2736–2744.

[149] M. Zhu, C. Feng, J. Chen, C. Guo, and X. Gao, "Video semantics based resource allocation algorithm for spectrum multiplexing scenarios in vehicular networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, 2021, pp. 31–36.

[150] W. Yang et al., "Semantic communication meets edge intelligence," 2022, *arXiv:2202.06471*.

[151] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[152] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.

[153] P. Popovski, O. Simeone, F. Boccardi, D. Gündüz, and O. Sahin, "Semantic-effectiveness filtering and control for post-5G wireless connectivity," *J. Indian Inst. Sci.*, vol. 100, no. 2, pp. 435–443, Feb. 2020.

[154] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, 2002, pp. 311–318.

[155] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4566–4575.

[156] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[157] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Process.*, vol. 2, 2001, pp. 749–752.

[158] R. V. Cox, "Three new speech coders from the ITU cover a range of applications," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 40–47, Sep. 1997.

[159] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 586–595.

[160] M. Yang, C. Bian, and H.-S. Kim, "Deep joint source channel coding for WirelessImage transmission with OFDM," 2021, *arXiv:2101.03909*.

[161] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.

[162] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[163] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2414–2423.

[164] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[165] E. Uysal et al., "Semantic communications in networked systems," 2021, *arXiv:2103.05391*.

[166] A. Maatouk, M. Assaad, and A. Ephremides, "The age of incorrect information: An enabler of semantics-empowered communication," 2020, *arXiv:2012.13214*.

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

[168] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May 2021.

[169] 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.

[170] R. A. Howard, "Information value theory," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 2, no. 1, pp. 22–26, Aug. 1966.

[171] O. Ayan, M. Vilgelm, M. Klügel, S. Hirche, and W. Kellerer, "Age-of-information vs. value-of-information scheduling for cellular networked control systems," in *Proc. 10th ACM/IEEE Int. Conf. Cyber-Phys. Syst.*, 2019, pp. 109–117.

[172] Y. J. Liang, J. G. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Effect of burst losses and correlation between error frames," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 861–874, Jul. 2008.

[173] J. Holm et al., "Freshness on demand: Optimizing age of information for the query process," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[174] M. Li, J. Gao, L. Zhao, and X. Shen, "Adaptive computing scheduling for edge-assisted autonomous driving," *IEEE Trans. Veh. Technol*, vol. 70, no. 6, pp. 5318–5331, Jun. 2021.

[175] Q. Ye, W. Shi, K. Qu, H. He, W. Zhuang, and X. Shen, "Joint RAN slicing and computation offloading for autonomous vehicular networks: A learning-assisted hierarchical approach," *IEEE Open J. Veh. Technol.*, vol. 2, pp. 272–288, Feb. 2021.

[176] C. De Alwis et al., "Survey on 6G frontiers: Trends, applications, requirements, technologies and future research," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 836–886, 2021.

[177] A. Nanda, D. Puthal, J. J. Rodrigues, and S. A. Kozlov, "Internet of autonomous vehicles communications security: Overview, issues, and directions," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 60–65, Aug. 2019.

[178] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed federated learning for ultra-reliable low-latency vehicular communications," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1146–1159, Feb. 2020.

[179] W. Zhang et al., "Optimizing federated learning in distributed industrial IoT: A multi-agent approach," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3688–3703, Dec. 2021.

[180] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, 2020.

[181] Y. Jiang et al., "Model pruning enables efficient federated learning on edge devices," 2019, *arXiv:1909.12326*.

[182] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2624–2661, 4th Quart., 2016.

[183] N. Cheng et al., "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1117–1129, May 2019.

[184] C. Zhou et al., "Deep reinforcement learning for delay-oriented IoT task scheduling in SAGIN," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 911–925, Feb. 2021.

[185] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[186] X. Luo, Z. Chen, B. Xia, and J. Wang, "Autoencoder-based semantic communication systems with relay channels," 2021, *arXiv:2111.10083*.

[187] H. Du et al., "Exploring attention-aware network resource allocation for customized metaverse services," 2022, *arXiv:2208.00369*.

[188] Y.-J. Liu, H. Du, D. Niyato, G. Feng, J. Kang, and Z. Xiong, "Slicing4Meta: An intelligent integration framework with multi-dimensional network resources for Metaverse-as-a-service in Web 3.0," 2022, *arXiv:2208.06081*.

[189] L. Ismail, D. Niyato, S. Sun, D. I. Kim, M. Erol-Kantarci, and C. Miao, "Semantic information market for the Metaverse: An auction based approach," 2022, *arXiv:2204.04878*.

[190] S. Movassaghi, M. Abolhasan, J. Lipman, D. Smith, and A. Jamalipour, "Wireless body area networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1658–1686, 3rd Quart., 2014.

[191] A. Elsts et al., "On-board feature extraction from acceleration data for activity recognition" in *Proc. EWSN*, 2018, pp. 163–168.

[192] P. Zalewski, L. Marchegiani, A. Elsts, R. Piechocki, I. Craddock, and X. Fafoutis, "From bits of data to bits of knowledge—An on-board classification framework for wearable sensing systems," *Sensors*, vol. 20, no. 6, p. 1655, 2020.

[193] S. Bragança, E. Costa, I. Castellucci, and P. M. Arezes, "A brief overview of the use of collaborative robots in industry 4.0: Human role and safety," in *Occupational and Environmental Safety and Health*. Cham, Switzerland: Springer, 2019, pp. 641–650.

[194] Y. Yue, C. Zhao, Z. Wu, C. Yang, Y. Wang, and D. Wang, "Collaborative semantic understanding and mapping framework for autonomous systems," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 2, pp. 978–989, Apr. 2021.

[195] G. M. Milis, C. G. Panayiotou, and M. M. Polycarpou, "SEMIoTICS: Semantically enhanced IoT-enabled intelligent control systems," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 1257–1266, Feb. 2019.

[196] D. Kolokotsa, "Comparison of the performance of fuzzy controllers for the management of the indoor environment," *Build. Environ.*, vol. 38, no. 12, pp. 1439–1450, Dec. 2003.

[197] L. Da Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[198] J. Gao, W. Zhuang, M. Li, X. Shen, and X. Li, "MAC for machine-type communications in industrial IoT—Part I: Protocol design and analysis," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9945–9957, Jun. 2021.

[199] P. Popovski et al., "Internet of Things (IoT) connectivity in 6G: An interplay of time, space, intelligence, and value," 2021, *arXiv:2111.05811*.

[200] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[201] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! Criticism for interpretability," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 29, 2016, pp. 1–9.

[202] Y. Dong, H. Su, J. Zhu, and B. Zhang, "Improving interpretability of deep neural networks with semantic information," in *Proc. IEEE Comput. Soc. Conf. (CVPR)*, 2017, pp. 4306–4314.

[203] M. Sheraz et al., "Artificial intelligence for wireless caching: Schemes, performance, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 631–661, 1st Quart., 2021.

[204] A. Cidon, A. Eisenman, M. Alizadeh, and S. Katti, "Cliffhanger: Scaling performance cliffs in Web memory caches," in *Proc. 13th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, 2016, pp. 379–392.

[205] Y. Xiao, Y. Li, G. Shi, and H. V. Poor, "Reasoning on the air: An implicit semantic communication architecture," 2022, *arXiv:2202.01950*.

[206] J. Liang, Y. Xiao, Y. Li, G. Shi, and M. Bennis, "Life-long learning for reasoning-based semantic communication," 2022, *arXiv:2202.01952*.

[207] C.-X. Wang, J. Huang, H. Wang, X. Gao, X. You, and Y. Hao, "6G wireless channel measurements and models: Trends and challenges," *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 22–32, Dec. 2020.

[208] J. Hu, H. Zhang, K. Bian, Z. Han, H. V. Poor, and L. Song, "MetaSketch: Wireless semantic segmentation by reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 5916–5929, Aug. 2022.

[209] H. Gacanin and M. Di Renzo, "Wireless 2.0: Toward an intelligent radio environment empowered by reconfigurable meta-surfaces and artificial intelligence," *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 74–82, Dec. 2020.

[210] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[211] M. R. Bloch, "Covert communication over noisy channels: A resolvability perspective," *IEEE Trans. Inf. Theory*, vol. 62, no. 5, pp. 2334–2354, May 2016.

Wanting Yang received the B.S. degree from the Department of Communications Engineering, Jilin University, Changchun, China, in 2018, where she is currently pursuing the Ph.D. degree with the College of Communication Engineering. She was a visiting Ph.D. student with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore, from 2021 to 2022, sponsored by the Chinese Scholarship Council. Her research interests include wireless video transmission, predictive resource allocation and semantic communication for vehicular network, learning, ultra-reliable, and low-latency communications.

Hongyang Du (Graduate Student Member, IEEE) received the B.Sc. degree from Beijing Jiaotong University, Beijing, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Energy Research Institute @ NTU, Nanyang Technological University, Singapore, under the Interdisciplinary Graduate Program. His research interests include semantic communications, reconfigurable intelligent surface, and communication theory. He was the recipient of the IEEE Daniel E. Noble Fellowship Award in 2022. He was recognized as an Exemplary Reviewer of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2021.

Zi Qin Liew received the degree (Hons.) in electronic and electrical engineering from Nanyang Technological University in 2018. He is currently pursuing the Ph.D. degree with Alibaba Group and the Alibaba-NTU Joint Research Institute, Nanyang Technological University, Singapore. His research interests include wireless communications and incentive mechanisms.

Wei Yang Bryan Lim received the Ph.D. degree in Nanyang Technological University, Singapore, in 2022, under the Alibaba Ph.D. Talent Programme, where he won the "Most Promising Industrial Postgraduate Programme Student" Award. He is currently Wallenberg-NTU Presidential Postdoctoral Fellow. His works have won Best Paper Awards, including in the IEEE Wireless Communications and Networking Conference, the IEEE Asia Pacifc Board Outstanding Paper Award, and the IEEE SPCC Technical Committee Best Paper Award. He regularly serves as a Technical Programme Committee Member in flagship conferences, is part of the guest editor team of special issues in *IEEE Wireless Communication magazine* and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and is currently a Review Board Member of IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS.

Zehui Xiong (Member, IEEE) received the Ph.D. degree in Nanyang Technological University (NTU), Singapore. He is currently an Assistant Professor with the Singapore University of Technology and Design and also a Honorary Adjunct Senior Research Scientist with Alibaba-NTU Singapore Joint Research Institute, Singapore. He was the Visiting Scholar with Princeton University and the University of Waterloo. He has published more than 200 research papers in leading journals and flagship conferences and many of them are ESI Highly Cited Papers. His research interests include wireless communications, Internet of Things, blockchain, edge intelligence, and Metaverse. He has won over 10 Best Paper Awards in international conferences and is listed in the World's Top 2% Scientists identified by Stanford University. He is the recipient of the IEEE Early Career Researcher Award for Excellence in Scalable Computing, the IEEE Technical Committee on Blockchain and Distributed Ledger Technologies Early Career Award, the IEEE Internet Technical Committee Early Achievement Award, the IEEE TCI Rising Star Award, the IEEE Best Land Transportation Paper Award, the IEEE CSIM Technical Committee Best Journal Paper Award, the IEEE SPCC Technical Committee Best Paper Award, the IEEE VTS Singapore Best Paper Award, Chinese Government Award for Outstanding Students Abroad, and the NTU SCSE Best Ph.D. Thesis Runner-Up Award. He is currently serving as the Editor or Guest Editor for many leading journals, including IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING and the Associate Director of Future Communications Research and Development Programme.

Dusit Niyato (Fellow, IEEE) received the B.E. degree from the King Mongkuk's Institute of Technology Ladkrabang, Thailand, in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Canada, in 2008. He is currently a Professor with the School of Computer Science and Engineering and, by courtesy, the School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore. He has authored or coauthored more than 380 technical papers in the area of wireless and mobile networking and is an Inventor of four U.S. and German Patents. He has authored four books including, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications* (Cambridge University Press). He was the recipient of the Best Young Researcher Award of the IEEE Communications Society Asia Pacific and the 2011 IEEE Communications Society Fred W. Ellersick Prize Paper Award. He is currently a Senior Editor of the IEEE WIRELESS COMMUNICATIONS LETTER, an Area Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (Radio Management and Multiple Access) and the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS (Network and Service Management and Green Communication), an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, an Associate Editor of the IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He was a Guest Editor of IEEE JOURNAL ON SELECTED AREAS ON COMMUNICATIONS. He was a Distinguished Lecturer of the Communications Society from 2016 to 2017. He was named the 2017, 2018, and 2019 Highly Cited Researcher in computer science.

Xuefen Chi received the B.Eng. degree in applied physics from the Beijing University of Posts and Telecommunications, Beijing, China, in 1984, and the M.S. and Ph.D. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 1990 and 2003, respectively. She was a Visiting Scholar with the Department of Computer Science, Loughborough University, Loughborough, U.K., in 2007, and the School of Electronics and Computer Science, University of Southampton, Southampton, U.K., in 2015. She is currently a Professor with the Department of Communications Engineering, Jilin University, Changchun. Her current research interests include machine type communications, indoor visible light communications, random access algorithms, delay-QoS guarantees, and queuing theory and its applications.

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers university, New Brunswick, New Jersey, in 1990. He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular ad hoc and sensor networks. He is a Registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Fellow, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society. He received the AHSN Technical Committee in 2013, the Joseph LoCicero Award in 2015, the Education Award from the IEEE Communications Society in 2017, the James Evans Avant Garde Award from the IEEE Vehicular Technology Society in 2018, the R.A. Fessenden Award from IEEE in 2019, Canada, the Award of Merit from the Federation of Chinese Canadian Professionals, Ontario, in 2019, and the Technical Recognition Award from Wireless Communications Technical Committee in 2019. He has also received the Premier's Research Excellence Award from the Province of Ontario, Canada, in 2003, and the Excellent Graduate Supervision Award from the University of Waterloo, Canada, in 2006. He served as the Technical Program Committee Chair/Co-Chair for the IEEE Globecom'16, the IEEE Infocom'14, IEEE VTC'10 Fall, and IEEE Globecom'07, the Symposia Chair for IEEE ICC'10 and the Chair for the IEEE Communications Society Technical Committee on Wireless Communications. He is the Elected IEEE Communications Society Vice President for Technical and Educational Activities, the Vice President for Publications, Member-at-Large on the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee.

Chunyan Miao (Senior Member, IEEE) is the Director of the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, Nanyang Technological University, Singapore, where she is the Chair with the School of Computer Science and Engineering. She is the Editor-in-Chief of the *International Journal of Information Technology* published by the Singapore Computer Society.