Resource Allocation in Wireless Semantic Communications: A Comprehensive Survey

Chujun Zhang, Linyu Huang, *Member, IEEE*, and Qian Ning

***Abstract*—With the advent of sixth-generation mobile communication technology (6G) and the emergence of future application scenarios such as Metaverse and digital twin (DT), the existing traditional wireless communication technology based on Shannon's information theory has not been able to meet the increasing demand for data transmission. Semantic communications (SemCom), which greatly reduces the amount of information transmitted and alleviates the burden of communication by transmitting the meaning behind the information, has been considered a promising 6G enabler. SemCom's resource allocation is critical to the system's reliability and effectiveness. Compared to traditional wireless communication systems, the system architecture and performance metrics of SemCom have undergone significant changes, making it difficult for traditional resource allocation strategies to adapt well to this new architecture. However, the issue remains unresolved and inadequately researched. In order to provide researchers with valuable insight to promote follow-up research, this paper reviews the latest research results in recent years and presents an overview of research progress in the field of resource allocation in wireless SemCom.**

***Index Terms*—Performance metrics, resource allocation, semantic communications, semantic similarity.**

I. Introduction

A. Context

In 1949, Weaver expanded Shannon's theory to three levels: technical level, semantic level, and effectiveness level [1]. The lowest level is the technical level, which is mainly responsible for the accurate and effective transmission of information symbols; the middle level is the semantic level, which points to the transmission of information symbols to convey the desired meaning; the upper level is the effectiveness level, which aims at effectively performing intelligent tasks and providing the needed communication efficiency on the lower two levels. Traditional communications operate at the technical level, focusing on accurate bit transmission. However, they transmit all information, including useless and irrelevant data, to the receiver, leading to channel resource waste. As sixth-generation mobile communication technology (6G) emerges, scenarios such as Digital Twin (DT) and Metaverse require wireless communication networks to transmit huge amounts of data. Wireless communication networks must achieve an extremely low transmission delay in scenarios such as autonomous driving and telemedicine. The emergence of these applications presents new challenges to traditional communication systems.

In the face of such a large communication load, how can one go beyond Shannon's limit to the future? Inspired by the three levels of the previous communication problem, a new communication paradigm, semantic communication (SemCom) [2], [3], [4], has been proposed to shift the communication paradigm to the semantic and effectiveness levels.

In traditional communication systems, data is compressed by the source encoder, and redundancy is added to the channel encoder to improve its robustness to interference/noise in the channel. At the destination, a reverse process is performed to recover the original sent data. The transmission and reception of signals do not involve any intelligence and the semantic information is omitted [5].

However, in a SemCom system, the semantic source and destination are intelligent agents that can perform various highly intelligent algorithms. Semantic coding replaces traditional source coding through deep learning (DL) and other technologies to extract semantic information. Unlike traditional communication systems, which are easily affected by channel conditions, SemCom performs well, especially at low signal-to-noise ratios (SNR), because only semantic information is transmitted. Goal-oriented SemCom or task-oriented SemCom is a subset of SemCom that pays more attention to the effectiveness level. Specifically, it focuses on the efficient use of semantic information for the successful execution of tasks at a suitable time [6]. The receiver in a goal-oriented SemCom is interested in the significance and effectiveness (semantics) of the transported source message to achieve a certain task or goal [7]. In summary, SemCom is becoming an excellent solution to the above questions. SemCom is also regarded as a key enabling technology for 6G, and it is an important step towards the future of wireless communication.

B. Resource Allocation

In general, resource allocation refers to a set of methodologies to achieve goals by efficiently allocating resources and using resource allocation methods based on resource availability. The resource allocation problem in wireless communications and SemCom is mapped into a mathematical

Received 25 February 2025; revised 23 April 2025 and 17 June 2025; accepted 30 July 2025. Date of publication 4 August 2025; date of current version 2 January 2026. This work was supported by the National Natural Science Foundation of China under Grant 61801318. *(Corresponding author: Linyu Huang.)*

The authors are with the College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China (e-mail: zhangchujun@stu.scu.edu.cn; lyhuang@scu.edu.cn; ningq@scu.edu.cn).

Digital Object Identifier 10.1109/COMST.2025.3595168

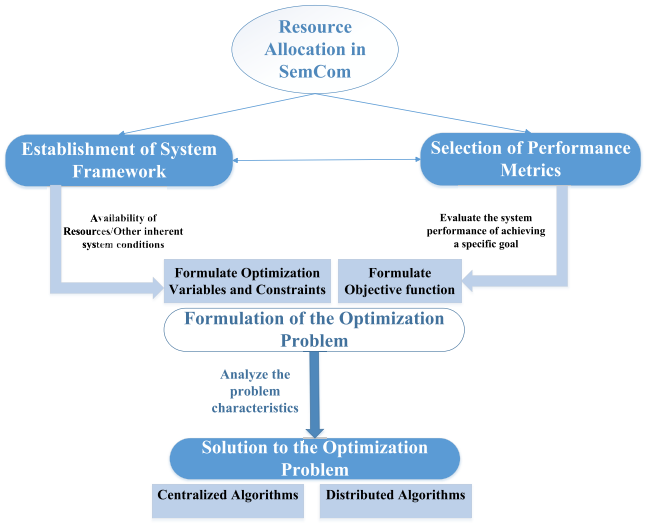

Fig. 1. The integrated framework of resource allocation in SemCom.

optimization problem by modeling the network structure and designing the objective function. In resource allocation, the available resources for allocation are optimization variables; the availability of resources and other inherent conditions are constraints; the objective function is the function to evaluate the system performance of achieving a specific goal; the resource allocation algorithm is a combination of optimization techniques that are used to solve this optimization problem. The resource allocation algorithms in SemCom can be divided into centralized and distributed ways. The centralized algorithm includes techniques based on convex optimization and other mathematical methods, and based on deep reinforcement learning (DRL), etc. The distributed algorithms include techniques based on multi-agent deep reinforcement learning (MADRL) and matching theory, etc. The integrative framework of resource allocation in SemCom is illustrated in Fig. 1.

However, compared to the traditional wireless communication system, the SemCom system architecture and performance metrics have undergone tremendous changes, making it difficult for traditional resource allocation strategies to adapt well to this new architecture. In the next section, we will provide a more detailed description of the difference between traditional communication and SemCom in terms of resource allocation and why it is important.

C. Related Surveys and Motivation

The development of SemCom has led to the publication of numerous surveys in recent years. Existing surveys on SemCom may address resource allocation to some extent, they mostly provide a global perspective of SemCom and often focus on broader aspects such as system architectures, semantic information theory, enable techniques or general applications of SemCom [3], [4], [5], [6], [8], [9], [10], [11], [12], [13], [14], [15]. However, our work is the first to present a dedicated and in-depth review of resource allocation in SemCom systems. To highlight this distinction, we have added a comparative table that outlines the resource allocation aspects covered (or not) by existing surveys, thereby clarifying the unique contribution of this work. Table I provides a comparison between our survey and representative prior works. Although there are some recent surveys on resource allocation in other communication scenarios that provided us with great insights, such as edge computing [16], fifth-generation mobile communication technology (5G)-and-beyond mobile edge computing (MEC) [17], Internet of Things (IoT) enabled vehicular edge computing [18], energy-efficient Orthogonal Frequency Division Multiplexing (OFDM) enabled networks [19], and ultra-dense networks (UDNs) [20]. Resource allocation is a critical and under-explored aspect of SemCom, it significantly differs from traditional communication systems, as it involves unique allocatable resources like semantic fidelity and computation overhead for semantic processing, alongside traditional factors such as bandwidth and power. Moreover, SemCom introduces novel performance metrics that will make the object function more complicated, which we will provide a more explicit description in Section III. These differences highlight the need to focus specifically on resource allocation in SemCom, as existing surveys tend to overlook the unique challenges and optimization strategies required in this domain. By dedicating our review entirely to this topic, our aim is to fill this gap and provide a comprehensive and systematic overview of how resource allocation can be effectively addressed within the context of SemCom. Therefore, we review from multiple perspectives, including SemCom network models, performance metrics, resource allocation optimization algorithms, as well as challenges and future research directions, providing researchers with a new, comprehensive, and rich perspective.

D. Research Methodology

In this subsection, the process followed to collect the references used in this study is described. The methodology includes the selection, inclusion, and exclusion criteria applied to ensure the quality and relevance of the references. The steps followed in the research process are as follows:

- *Literature Search:* The search was performed using databases such as Google Scholar, IEEE Xplore, and ScienceDirect. The primary focus was on peer-reviewed journal articles, conference papers, books, and other reputable sources related to SemCom, resource allocation, optimization, and wireless communication networks.
- *Inclusion Criteria:* To be included in the study, references must meet the following criteria: 1) Published in a peer-reviewed journal or conference proceedings. Directly related to resource allocation in SemCom networks or relevant areas such as optimization, deep learning techniques, and wireless communications. 2) Except for some classic and fundamental literature, references should be published within the last 10 years to ensure that the research is up-to-date and relevant. 3) For research papers, theoretical and empirical studies must be included.
- *Exclusion Criteria:* References that met any of the following conditions were excluded from the review:

TABLE I
COMPARISON OF EXISTING SEMCOM SURVEYS AND THIS WORK

Ref.	Focus Aspects and Brief Description	RA Specific Challenges	System Arch.	Taxonomy, Tables, and In-depth Analysis			RA Future Pros.
				RA Metrics	RA Problem Form.	RA Tech.	
[3]	SemCom theory analysis, system architecture, similarity metrics, DL techniques, frameworks and challenges.	×	✓	△	×	×	×
[4]	SemCom and Task-oriented communications: History, semantic information measures, information theoretic foundations, ML techniques, joint source-channel coding, practical designs, semantic timeliness, effective/pragmatic communications.	×	✓	△	×	×	×
[5]	SemCom background, system architecture, differences from traditional communications, metrics, use cases, open issues and future research directions.	×	✓	×	×	×	×
[6]	SemCom history, theories, metrics, taxonomy, frameworks, enable-techniques and applications.	×	✓	△	×	△	×
[8]	SemCom background, architecture, cross layer interaction, applications, issues and challenges.	△	✓	△	×	×	△
[9]	SemCom background, architecture, enable technologies (semantic extraction, coding, segmentation), performance improvement, applications, issues and challenges.	△	✓	△	×	×	×
[12]	SemCom theory analysis, semantic extraction techniques, information transmission, performance metrics, potential applications, SemCom for future 6G Internet, challenges and future directions.	×	✓	△	×	×	×
[13]	Detailed numerous metrics for SemCom and Goal-oriented SemCom, very comprehensive.	×	✓	✓	×	×	×
[14]	Goal-oriented SemCom: Background, state-of-the-art research landscape, trends, application, mathematical frameworks and theories, challenges and future directions.	△	✓	×	×	△	×
[15]	SemCom resource management (transmission, resource allocation, MEC orchestration), challenges and future directions. Mainly focus on security and privacy issue in SemCom along with their countermeasures. For RA in SemCom, have partially discussed but lack of RA specific in-depth analysis and systematic taxonomy.	△	✓	△	△	△	△
This Work	**All Aspects for RA in SemCom.**	✓	✓	✓	✓	✓	✓

* RA: Resource Allocation (in SemCom); System Arch.: System Architecture; RA Problem Form.: RA Problem Formation; RA Tech.: RA Techniques; RA Future Pros.: RA Future Prospects
* ✓: explicitly covered; ×: not covered or only mentioned some few sentences. △ means partially discussed or moderate analysis.

1) Studies that were completely not related to SemCom or resource allocation; 2) Non-peer-reviewed sources or articles without sufficient methodological rigor. 3) Studies published more than 10 years ago unless they introduced foundational theories or seminal works that remain relevant to current research.

- *Information Extraction and Analysis:* After finalizing the selected references, the key information was extracted and analyzed in resource allocation. This included understanding the research objectives, methodologies used, findings, and how each study contributed to advance the understanding of SemCom resource allocation.

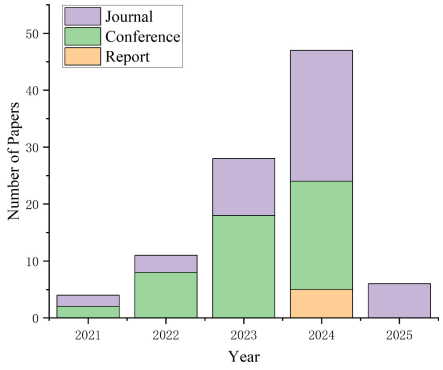

Fig. 2. The distribution of papers surveyed by year and source.

E. Contributions and Organization

This paper reviews the current state of research in the period 2021-2025 (February) on resource allocation in wireless SemCom. Fig. 2 shows the distribution of the articles surveyed by year and source. The report encompasses arXiv articles and website articles, while the conference category includes conference and symposium papers, and the journal category includes journal and magazine articles. The contributions of this paper can be summarized as follows:

- We first explain the basics of resource allocation in SemCom and introduce the network model in the current literature on SemCom resource allocation, which is divided into end-to-end (E2E) and multi-user situations. We also investigated the use of the next generation of multiple access (NGMA) technologies and hybrid semantic/bit communications in SemCom resource allocation. We give the overview of resource allocation in SemCom, thereby explaining the reason why resource allocation in SemCom is important for the theoretical perspective and reality perspective, clarifying the unique specific challenges that inherently exist in the resource allocation of SemCom.

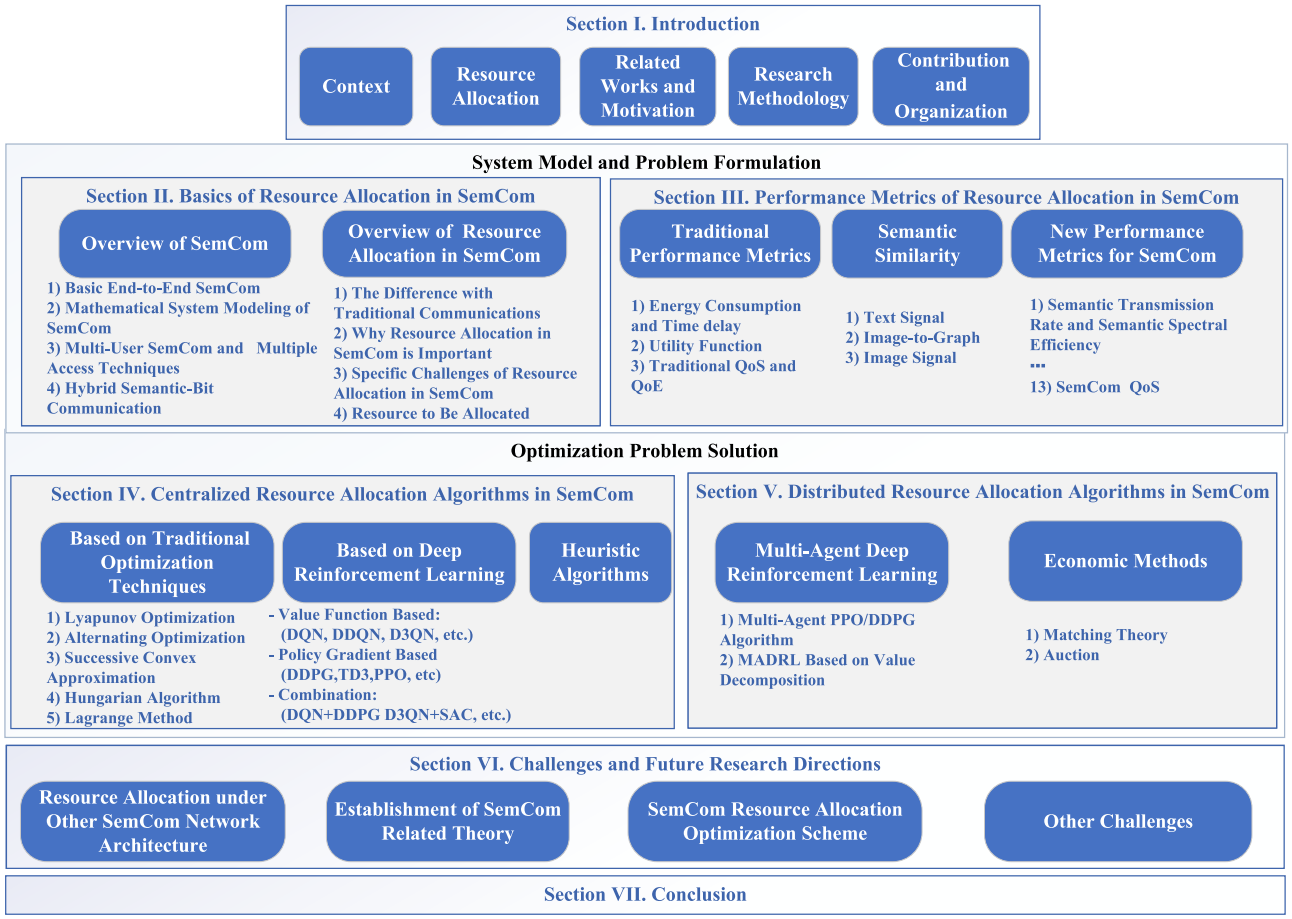

Fig. 3. Road map of the survey.

- The construction of the objective function is the core of the optimization problem modeling, so we introduce the performance metrics in the SemCom resource allocation in detail. We mainly summarized the construction methods into two types. One is utilizing traditional resource allocation performance metrics, such as delay and energy consumption. The other type is based on the semantic similarity, establishing new performance metrics.
- We discuss in detail different optimization algorithms in the allocation of SemCom resources, which are divided into centralized and distributed algorithms. Centralized algorithms include algorithms based on convex optimization and other mathematical methods, algorithms based on DRL, and heuristic algorithms. Distributed algorithms include methods based on MADRL, matching theory, and auction. These methods are summarized in three comprehensive tables for comparison.
- Through the analysis presented above, we propose future research directions and several challenges to be solved in the field of SemCom resource allocation.

The remainder of this paper is organized as follows. Section II introduces the basic architecture of the SemCom resource allocation problem. Section III presents traditional performance metrics, the definition of semantic similarity, and new semantic-based performance metrics. Sections IV and V summarize the different centralized and distributed resource allocation optimization algorithms in detail. Section VI points out the challenges and possible future research directions. Finally, Section VII summarizes this survey. Fig. 3 shows the organization and structure of this survey paper.

II. Basics of Resource Allocation in SemCom

This section will explain the basics of the SemCom resource allocation problem. We provide an overview of SemCom, followed by a review of the fundamental network models found in various SemCom resource allocation studies. Furthermore, we give an explicit contrast between bit-level and semantic-level modeling in Table III, which provides a side-by-side comparison between the two paradigms, highlighting their respective targets, metrics, modeling approaches, and optimization goals. Next, we provide an overview of resource allocation in SemCom. Besides, we give the taxonomy of system framework establishment in Fig. 5. Lastly, we summarize the literature in Table IV.

A. Overview of SemCom

Traditional communications aim to reach the technical level, which means achieving a high data transmission rate and a low symbol error rate. However, the basic idea of SemCom is to extract the "meanings" or "features" of the

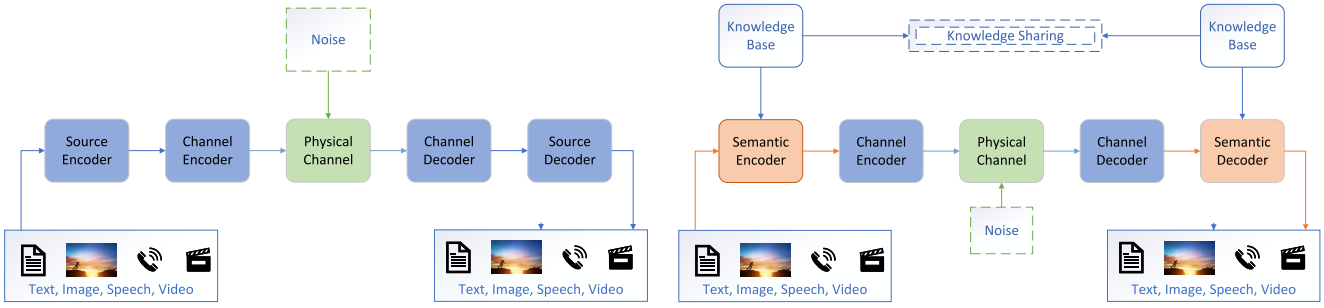

Fig. 4. A comparison between the basic end-to-end network architecture of traditional communication and SemCom.

source and "interpret the semantic information" at a destination. Therefore, SemCom surpasses traditional bit-level transmission to achieve semantic-level transmission, leading to significant changes in the design of the network architecture.

1) Basic End-to-End SemCom: A comparison between the basic end-to-end (E2E) network architecture of traditional communication and SemCom is shown in Fig. 4. Fig. 4a illustrates the typical traditional E2E communication architecture, where the source encoder receives the transmitted data and compresses it initially, partially eliminating redundant information through source encoding. The channel encoder adds redundancy in various coding ways to combat noise and attenuation in the channel, thereby enhancing its anti-interference capability and error correction ability. At the destination, a reverse process is conducted to recover the original sent data. We can see in Fig. 4b that SemCom primarily differs from traditional end-to-end architecture in three key ways.

- Semantic Coding: A SemCom system extracts the semantic information (features) from the original data through semantic coding enabled by technologies such as DL and then encodes these features for channel coding. Due to the implicit meaning inherent in the message under consideration, the amount of redundant data removed is significantly greater than that achieved by source coding. Not like semantic segmentation in computer vision, in SemCom, all communication parties must maintain a high degree of consistency in semantic expression and understanding, which poses a challenge to semantic compression.
- Knowledge Base: Another important feature of SemCom is that it is a knowledge-based system [21]. This means that semantic source and semantic purpose can be like the human brain, through self-learning to establish their own background knowledge bases (KBs) to guide the transmitter to obtain multi-level semantic knowledge description of source data, semantic inference, estimation of transmission environment, and semantic requirements of downstream tasks. The system performs semantic coding and directs the receiver to execute the inverse process, known as semantic decoding.
- Semantic Decoding: Based on technologies such as semantic KBs and DL, the receiver can understand and infer the received information to complete the recovery of the received semantic information.

Currently, most of the research literature is focused on three types of sources: *text signal*, *image signal*, and *speech signal*. Moreover, there is very little literature that points the research direction to multi-modal tasks [22].

Text: Text SemCom systems have been widely studied. Various DL techniques are used to represent the underlying meaning of texts. DL-enabled semantic codecs have been through the early Long Short Term Memory (LSTM)-based models [23], [24], to today's Transformer-based models [25], [26]. In 2018, Farsad et al. [23] proposed a joint source-channel coding (JSCC) scheme for text SemCom, in which the encoder and decoder are implemented by two LSTM networks. Compared to the single source channel coding (SSCC) scheme, the DL-based JSCC scheme performs better [23]. In 2021, Xie et al. [25] proposed the DeepSC framework by fine-tuning the basic structure of Transformer [27]. DeepSC can adapt to different channel environments, perform well under low SNR, and have excellent robustness. The author of [28] proposed a semantic extraction scheme based on the entity recognition model (NER) and LSTM that transforms the transmitted sentence into multiple triplets of semantic importance, and important triplets will be allocated more transmission resources to improve reliability. The authors of [29] introduce a life-long model updating approach in which the receiver can learn from previously received messages and automatically update the rules to reasoning for hidden information when new unknown semantic entities and relations have been discovered.

Image: The image SemCom system is similar to the text SemCom system, and there is much research on it. In contrast to text systems, image SemCom systems extract the original image's features (which, in this context, represent the image's "meaning") and extensively utilize convolutional neural networks (CNNs). In addition, in many task-oriented SemCom systems (such as image classification tasks), the image does not need to be reconstructed at the receiver. In 2019, Bourtsoulatze et al. [30] first proposed an end-to-end image transmission system using CNN's JSCC scheme, which has better performance than traditional image transmission methods. In 2022, Dong et al. [31] proposed a layer-based semantic communication system for images (LSCI), and the concept of semantic slice-models (SeSM) is proposed

to enable flexible model resemblance under the different requirements of the model performance, channel situation, and transmission goals. In 2023, Lokumarambage et al. [32] implemented a semantic communication-based end-to-end image transmission system, where a pre-trained GAN network is used at the receiver as the transmission task to reconstruct the realistic image based on the semantic segmented image at the receiver input. Kadam and Kim [33] proposed a joint CNN-LSTM-based SemCom model in which the semantic encoder of a camera extracts the relevant semantics from the raw images, resulting in a novel approach to the problem of predicting vehicle counts.

Speech: Unlike the previous two modes of the SemCom system, the speech signal possesses more complex performance characteristics, including speech speed, volume, tone, and dialect, all of which can express the same meaning. The general approach is to convert the speech into text for processing. However, the same text information expressed in different intonations will produce different meanings. Therefore, the process of voice semantic transmission is more complex and challenging to manage [34], [35]. The majority of the source modes in SemCom's resource allocation are text and image modes. Currently, there is no relevant research on the allocation of resources for the SemCom speech system. In the following content, we will introduce and compare these papers comprehensively and organize them in tables for reference.

2) Mathematical System Modeling of SemCom: While the previous section has highlighted the core components of semantic communication, it is equally important to understand how these elements integrate into a mathematical framework. We will introduce some essential parts of mathematical modeling in papers, mainly on semantic extraction and semantic metrics (it will be discussed thoroughly in Section III).

- NN features-based semantic extraction: It utilizes deep learning models for end-to-end semantic encoding, offering strong contextual understanding but lacking interpretability and explicit semantic relationships. In such an approach, the encoded symbol stream can be represented by

$$\mathbf{x} = C_\alpha\big(S_\beta(\boldsymbol{s})\big), \tag{1}$$

where, $S_\beta(\cdot)$ is the semantic encoder network with parameter set β and $C_\alpha(\cdot)$ is the channel encoder with parameter set α, the specific networks are various in different systems. In text SemCom systems [25], [36], networks such as Transformer, BERT, or LSTM are utilized for semantic extraction, $\boldsymbol{s} = [w_1, w_2, \ldots, w_L]$ denotes the original sentence, w_l represents the l-th word in each sentence. In speech SemCom systems [34], [37], [38], ResNet, Transformer, CNN and Recurrent Neural Network (RNN) are utilized for semantic extraction in different studies, the input $\boldsymbol{s}$ is the speech sample sequence, $\boldsymbol{s} = [s_1, s_2, \ldots, s_W]$ with W samples, where s_w is the w-th item in $\boldsymbol{s}$ and it is a scalar value. In the DeepSC-ST system [38], text-related semantic features are extracted from the input speech spectrum samples using CNN and the gated recurrent unit (GRU)-based bidirectional RNN (BRNN) modules. In the image SemCom systems [30], [31], [32], [33], [39], networks such as CNN and GAN are often used, and the input is a n-dimensional image, not a sequence like text or speech. Simulation results show that SemCom performs well especially under the low SNR. This is because the extracted semantic features reduce redundancy which will use more channel resources. After semantic extraction, high-level semantic representations are less sensitive to noise, which makes the SemCom system more robust.
- Knowledge Graph-based semantic extraction: It extracts structured information as semantic triples (subject, predicate, object) to form a knowledge graph, which enhances interpretability and enables reasoning but requires high construction and maintenance costs. The semantic information of a knowledge graph is typically expressed as triples in the form of (head, relation, tail). From a piece of text data, multiple triples can be extracted, and these triples can be used to characterize a knowledge graph. The knowledge graph extracted from each sample data T_n is represented as

$$G_n = \left\{\varepsilon_n^1, \varepsilon_n^2, \ldots, \varepsilon_n^m, \ldots, \varepsilon_n^M\right\}, \tag{2}$$

where ε_n^m is the m-th triple in knowledge graph G_n, M is the total number of triples. The triple ε_n^m can be written in the following form:

$$\varepsilon_n^m = (h_n^m, r_n^m, t_n^m), \tag{3}$$

where h_n^m is the head entity of triple ε_n^m, t_n^m is the tail entity, and r_n^m is the relation of head and tail entities. For text, the work in [40] used an information extraction system to extract semantic triples from texts and modeled as KGs, and the receiver used a graph-to-text generative algorithm to recover the original texts based on the received triples. In [41], a cognitive text semantic communication framework is proposed by exploiting knowledge graph. For image, the scene graph (SG) is a visual KG that describes visual relationships between entities, the authors in [42] and [43] used object detection and RE algorithms to extract SG from images.

3) Multi-User SemCom and Multiple Access Techniques: The previous section introduces several end-to-end SemCom systems. However, all the above systems do not involve multi-user transmission. In general, the connection density of 5G is 106 devices per square kilometer, while the connection density of the 6G network will increase to 10 times that of 5G, and the regional traffic density should be 100 times that of 5G, which requires a significant improvement of spectral efficiency [5]. Moreover, the knowledge base within the SemCom system may vary significantly. Therefore, from a more realistic point of view, it is necessary to design a multi-user SemCom system. Notably, we only survey the multi-user SemCom system in papers on resource allocation in SemCom, not all SemCom-related papers.

In the resource allocation problem of multi-user SemCom, the classical multiple access (MA) techniques such as

TABLE II
THE COMPARISON OF DIFFERENT NGMA TECHNIQUES

	SDMA (Spatial Division Multiple Access)	NOMA (Non-orthogonal Multiple Acess)	RSMA (Rate-splitting Multiple Acesss)
Interference Management	Fully treat interference as noise	Fully decode interference	Partially treating interference as noise and partially decoding interference
Rate of user-1:	$B\log_2(1+\frac{\lvert \boldsymbol{h}_1^H \boldsymbol{p}_1\rvert^2}{\lvert \boldsymbol{h}_1^H \boldsymbol{p}_2\rvert^2+N_1})$	$B\log_2(1+\frac{\lvert \boldsymbol{h}_1^H \boldsymbol{p}_1\rvert^2}{\lvert \boldsymbol{h}_1^H \boldsymbol{p}_2\rvert^2+N_1})$	$C_1+B\log_2\left(1+\frac{\lvert \mathbf{h}_1^H \mathbf{p}_1\rvert^2}{\lvert \mathbf{h}_1 \mathbf{p}_2\rvert^2+N_1}\right)$
Rate of user-2:	$B\log_2(1+\frac{\lvert \boldsymbol{h}_2^H \boldsymbol{p}_2\rvert^2}{\lvert \boldsymbol{h}_2^H \boldsymbol{p}_1\rvert^2+N_2})$	$B\log_2(1+\frac{\lvert \boldsymbol{h}_2^H \boldsymbol{p}_2\rvert^2}{N_2})$	$C_2+B\log_2\left(1+\frac{\lvert \mathbf{h}_2^H \mathbf{p}_2\rvert^2}{\lvert \mathbf{h}_2 \mathbf{p}_1\rvert^2+N_2}\right)$

*Here, the BS wants to transmit two messages respectively to user-1 and user-2 in each time frame with bandwidth B, where $\boldsymbol{h}_k$ is the channel gain matrix from user-k to BS, perfectly known at BS. $\boldsymbol{p}_k$ is beamforming vector, and N_k is the noise power of user-k.
*In NOMA, assume the successive interference cancellation (SIC) is deployed at user-2.
*In RSMA, p_k is the power allocated to the private message, it differs from NOMA and SDMA.

frequency division multiple access (FDMA) [36], [44], [45], [46], [47], [48], orthogonal frequency division multiple access (OFDMA) [40], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60] or time division multiple access (TDMA) [61], [62] techniques are mostly used. However, with the continuous development of communication technology, researchers have begun to explore the application of the combination of next-generation multiple access (NGMA) and SemCom in resource allocation. Before comparing different MA techniques in papers on resource allocation in SemCom, we summarized these three key NGMA techniques in Table II.

As in Table II, spatial division multiple access (SDMA) treats the interference of other users fully as noise. Non-orthogonal multiple access (NOMA) will employ successive interference cancellation (SIC) at one user to fully decode the interference. Rate splitting multiple access (RSMA), based on the concept of rate splitting (RS), is considered to be a promising physical layer transmission paradigm for non-orthogonal transmission, interference management, and multiple access strategies in 6G. The main idea of RSMA is to divide user messages into common and private parts (s_c and s_k) and to be able to partially decode interference and partially treat interference as noise, which is in stark contrast to the extreme interference management strategies used in SDMA and NOMA. The flexibility of RSMA makes it perform well at all levels of interference [63]. In RSMA, p_k and p_c are the power allocated to private messages and the common message. The common stream s_c is decoded first by treating the interference from private streams s_1 and s_2 as noise. As s_c contains part of the intended message as well as part of the message of the interferer, it enables the ability to partially decode interference and partially treat interference as noise. The instantaneous rates for decoding the common streams at user-k are

$$R_{\mathrm{c},k}=B\log_2\left(1+\frac{\left|\mathbf{h}_k\mathbf{p}_\mathrm{c}\right|^2}{\left|\mathbf{h}_k\mathbf{p}_1\right|^2+\left|\mathbf{h}_k\mathbf{p}_2\right|^2+N_k}\right). \tag{4}$$

To guarantee that common message s_c is decoded by both users, the common rate shall not exceed

$$R_c=\min\{R_{c,1},R_{c,2}\}. \tag{5}$$

Denote C_k as the common rate portion of user-k: $C_1+C_2=R_c$. Once s_c is decoded and removed from the received signal via SIC, user-k decodes its desired private stream s_k, so the private rate of user k is

$$R_k=B\log_2\left(1+\frac{\left|\mathbf{h}_k^H\mathbf{p}_k\right|^2}{\left|\mathbf{h}_k\mathbf{p}_j\right|^2+N_k}\right), \tag{6}$$

so that the achievable total rate of user k is

$$R_{k,tot}=C_k+R_k. \tag{7}$$

In [64] and [65], the authors used SDMA as the multiple access method and established an SDMA-based multiuser probabilistic SemCom (PSC) framework that considers both transmission and computational consumption. The authors of [61] proposed a new semantic-aware resource allocation scheme in the integration of the radio frequency energy harvesting (EH), cognitive radio (CR), and NOMA scenario. An uplink network consisting of multiple primary users (PU) using TDMA and a secondary user (SU) using NOMA and PU multiplexing spectrum is considered. In the background of PSC, the work in [66] studied the joint communication and computation design in the reconfigurable intelligent surface (RIS)-assisted industrial Internet of Things (IIoT).

Compared to SDMA and NOMA, the research on resource allocation in the combination of RSMA and SemCom is obviously more [67], [68], [69], [70], [71], [72]. In [67], the optimization problem of the energy consumption of the downlink SemCom network with RSMA is studied. The authors of [69], [70] focused on the PSC framework based on RSMA; reference [70] expanded the work of [64], and the multiple access mode was changed from uplink SDMA to downlink RSMA, while the authors of [69] paid more attention to the energy-saving design of the PSC system. The simulations of the above literature [68], [70] compare the SDMA and NOMA-based schemes. The results show that the RSMA based scheme performance is the best in terms of total semantic transmission rate.

4) Hybrid Semantic-Bit Communication: While most research on resource allocation focuses solely on SemCom itself, the coexistence of SemCom and bit communication (BitCom) modes has also received attention [44], [45], [73], [74], [75], [76], [77], [78]. SemCom is more suitable for low signal-to-interference-plus-noise ratio (SINR) and resource-constrained scenarios, while BitCom performs well in high SINR regions. Moreover, it is not possible to completely

TABLE III
COMPARISON OF MODELING AND FRAMEWORKS: TRADITIONAL VS. SEMANTIC COMMUNICATION

Aspect	Traditional Communication	Semantic Communication
Focus of Modeling	Bit-level modeling: focus on modulation, encoding, and physical-layer performance	Semantic-level modeling: focus on semantic extraction, feature representation, and downstream task performance
Modeling Tools	Analytical models (e.g., traditional information theory, various source and channel coding methods)	Deep learning models (e.g., Transformer, CNN, RNN), knowledge graphs
Encoding on Transmitter	Source coding reduces redundancy based on bit-level statistics (will include redundant and useless information)	Semantic coding eliminates redundancy by extracting the meaning or characteristics from the source information
Decoding on Receiver	Decode all information and ignore the true meaning of the intended expression	Intelligent error correction and appropriate recovery on the receiver side based on the knowledge combined with context
Channel Condition	Recovering accurate bits at the receiver to achieve low symbol error rate requires good channel conditions and a high signal-to-noise ratio	High-level semantic features are less sensitive to physical noise which make SemCom are robust to bad channel conditions, limited communication resources or relatively low signal-to-noise ratios.
Noise Type	Physical channel noise	Physical channel noise and semantic noise due to the ambiguity existing in words, sentences or symbols in the messages
Framework Design	Separate source-channel coding (SSCC)	Enables Joint source-channel coding (JSCC)
Performance Metrics	Physical bit-level metrics (e.g. bit error rate and transmission rate)	Semantic-level or task-oriented metrics (e.g., semantic spectral efficiency and success probability of tasks)
Adaptivity to Task	Task-agnostic (same design regardless of task)	Task-oriented/task-aware (customized for classification, translation, etc.)

replace BitCom's current huge infrastructure and user bases at once. In the future, hybrid semantic/bit communication networks will become an inevitable and persistent example of intermediate networks [78]. The authors of [73] proposed a novel multi-carrier E2E system that combines both semantic and Shannon (bit) communications, in which both the BS and the user can communicate by choosing to utilize either bitCom or SemCom on each subcarrier. For resource allocation in the coexistence of semantic and bit communication networks, the focus is how to combine the measurement of the two. In [78], a bit-to-message (B2M) conversion function is used to convert the rate metric into the capacity of the semantic channel (i.e., the achievable message rate in units of messages per unit time, msg/s), let $\mathcal{R}_{ij}(\cdot)$ denote the B2M function of the SemCom link between mobile user (MU) i and BS j, its instantaneous achievable message rate in time slot t should be

$$M_{ij}^{S}(t) = \beta_{ij}(t)\mathcal{R}_{ij}\big(b_{ij}\log_2(1+\gamma_{ij}(t))\big). \tag{8}$$

Here, $\beta_{ij}(t)$, $b_{ij}(t)$, and $\gamma_{ij}(t)$ represents the knowledge-matching degree, bandwidth, and SINR between MU i and its communication counterpart at slot t. Compared to the SemCom link, the instantaneous achievable message rate of the BitCom link in slot t is given by

$$M_{ij}^{B}(t) = \rho_{ij}\mathcal{R}_{ij}\big(b_{ij}\log_2(1+\gamma_{ij}(t))\big). \tag{9}$$

Here, $b_{ij}(t)$, and $\gamma_{ij}(t)$ denotes the same thing as in Eq. (8), and ρ_{ij} is an average B2M transformation ratio to measure network performance with a message-related metric unified with SemCom.

If taking both SemCom and BitCom into account, use y_{ij} to denote the communication mode selection ($y_{ij} = 1$ represents that the SemCom mode is selected for the link between MU i and BS j, and $y_{ij} = 0$ indicates that the BitCom mode is selected), the time-averaged message rate of each link is

$$M_{ij} = \frac{1}{N}\sum_{t=1}^{N}\Big[y_{ij}M_{ij}^{S}(t) + \big(1-y_{ij}\big)M_{ij}^{B}(t)\Big]. \tag{10}$$

In [75], the equivalent transformation method in [36] is used to transform the bit rate into the equivalent semantic rate (suts/s, which will be discussed in the next section), which is unified into the semantic correlation measure to measure the network performance. Compared with the combination mode of SemCom and BitCom in [75], [78], the works in [44], [45], [74] both studied another form of coexistence of SemCom and BitCom separately in the downlink and uplink transmission. A semantic relay (SemRelay)-aided system was proposed. We use the uplink transmission scenario in [74] for explanation: from the users to SemRelay using BitCom, from SemRelay to the BS using SemCom. In the User-SemRelay link, FDMA is adopted, the achievable rate R_n^{us} is:

$$R_n^{us} = B_n^{us}\log_2\left(1+\frac{|h_n^{us}|^2 p_n^u}{B_n^{us}N_0}\right), \tag{11}$$

where N_0 is the power spectral density of the additive white Gaussian noise (AWGN), p_n^u denotes the transmission power of user n, h_n^{us} denotes the channel gain from user n to SemRelay and B_n^{us} denotes the bandwidth allocated to the link. The transmission delay for each user n is given by $t_n^{us} = \frac{D_n}{R_n^{us}}$. Here, D_n is the volume of text data in bits. The computation time cost for semantic compression at SemRelay is t^c, The achievable rate of the SemRelay-BS link is:

$$R^{sb} = B^{sb}\log_2\left(1+\frac{|h^{sb}|^2 p^s}{B^{sb}N_0}\right), \tag{12}$$

where, p_n^s denotes the transmission power of SemRelay, h_n^{sb} denotes the channel gain from SemRelay to BS, and B_n^{sb} denotes the bandwidth allocated to the link. The transmission delay for SemRelay is given by $t^{sb} = \frac{D^{Sem}}{R^{sb}}$. Here, D^{Sem} is the total number of bits for the compressed semantic information. The explicit expression of t^c and D^{Sem} can be found in [74]. So the overall latency t^{all} is

$$t^{all} = \max\{t_n^{us}, \forall n\} + t^c + t^{sb}. \tag{13}$$

Another difference from [75], [78] is that [44] and [45] transform the semantic rate into the bit rate to unify these two rate metrics into a bit-based metric (bit/s).

B. Overview of Resource Allocation in SemCom

We have given a brief description of resource allocation in Section I. We are now giving a more detailed description of the difference between traditional communication and SemCom in terms of resource allocation, as well as why it is important.

1) The Difference With Traditional Communications:

- Optimization Problem: Compared with traditional wireless communication, SemCom's network architecture has changed in many aspects, from codec level to multiple access modes. Due to the inexplicability of neural networks, it is difficult to derive closed-form expressions of some objective functions or variables. Therefore, the constructed optimization problem, from the objective to the constraints and optimization variables, differs significantly from the traditional architecture.
- Optimization Algorithm: As artificial intelligence and machine learning technology continue to advance, an increasing number of intelligent methods have emerged to address resource allocation problems. For example, neural networks are used to approximate the function in which closed-form expressions cannot be obtained, and deep reinforcement learning (DRL) has also become a powerful tool for solving complex resource allocation problems in recent years [79], [80], [81]. Though traditional methods like mathematical and convex optimization-based algorithms are still widely used, resource allocation in SemCom is more applicable to intelligent methods, and many papers tend to use intelligent method-based algorithms. We will give a comprehensive introduction to these algorithms in Section IV.

2) The Reason Why Resource Allocation in SemCom is Important:

- *Theoretical perspective:* Firstly, from the perspective of the network model, SemCom has a lot of new modules to consider, such as the semantic encoder and the knowledge base. Most SemCom systems use DL techniques to adopt semantic extraction. Neural networks will bring about a lot of inexplicability and can result in the lack of a closed form of part of the objective function. Moreover, as it involves unique allocatable resources such as semantic fidelity and computation overhead for semantic processing. Optimization algorithms to optimize these new semantic-related variables directly have a great influence on the whole system performance. Besides, traditional performance metrics do not consider the meaning of information. Using traditional performance metrics for resource allocation may even lead to a decrease in system performance. Therefore, developing new metrics that match the characteristics of SemCom and designing proper optimization algorithms to deal with the new objective functions and constraints caused by these new metrics can also have a positive influence on system performance. Recently, there has been a lot of research on the new performance metrics of SemCom, such as semantic similarity, semantic energy efficiency, and task success rate, of which we will give a detailed description in Section III-C.
- *Reality perspective:* In the context of 6G, the amount of data generated by terminal devices around the world is explosively increasing. Coordinating limited resources to better process these data requires an appropriate resource allocation strategy. Data from different application scenarios may have different service requirements. Vehicles in autonomous driving scenarios need to process data in milliseconds to ensure traffic safety. Therefore, ultra-low latency is its main goal. Semantic sensing systems assisted by uncrewed aerial vehicles (UAVs) usually pay more attention to the long battery life and expect to achieve low energy consumption. In addition, some mobile devices and IoT devices are designed to achieve low data processing costs or achieve the best user satisfaction. Therefore, appropriate resource allocation strategies are needed to meet these diverse needs.

3) Specific Challenges of Resource Allocation in SemCom: SemCom brings fundamental shifts to the modeling, evaluation, and optimization of wireless communication systems. These shifts give rise to several unique challenges that are rare or nonexistent in traditional communications and fundamentally affect how resource allocation must be performed. Although Section VI will discuss open research problems and promising future directions for SemCom, this subsection focuses on the specific and practical challenges that currently arise in existing SemCom system designs and implementations. These challenges reflect the inherent complexity and unique characteristics of SemCom. By clarifying these concrete issues, we lay the foundation for understanding why the optimization techniques in SemCom (which will be introduced in Sections IV and V) are necessary. These specific challenges can be summarized as follows:

- *Tradeoff Caused by Semantic Compression Ratio:* There are many tradeoffs, such as the energy-latency tradeoff and the accuracy-efficiency tradeoff, that already exist in traditional communications. However, SemCom introduces the new resource type, the semantic compression/extraction ratio, which directly affects communication, computation, and semantic fidelity. For instance, a higher compression ratio reduces the data size for transmission and saves transmission delay and energy consumption (communication load reduction), but it lowers the semantic fidelity and task accuracy and needs more computing resources to process the semantic extraction and recover, which results in the local extraction latency and energy consumption at the transmitter, the recover latency and energy consumption at the receiver (computation load increase). Moreover, for intelligent tasks, a higher compression ratio (lower in value) results in higher computing cycles for task processing, thus increasing task computing latency and energy consumption. These tightly coupled tradeoffs of computing, communication, and accuracy make resource allocation in SemCom inherently more complex. The detailed description about how the semantic compression ratio influences latency and energy consumption is in Section III-A.

- *Optimization with Non-differentiable and Implicit Objectives:* Many SemCom key performance metrics rely on semantic similarity, which is difficult to express analytically. These objectives often lack closed-form expressions, are non-differentiable, or even implicitly defined through closed-box models, which make traditional optimization methods hard to apply well.
- *Highly Coupled and Non-convex Optimization Variables:* Unlike conventional systems where resource variables can often be decomposed or linearized, SemCom involves complex coupling between variables such as computation capacity, transmission power, and semantic compression ratio. The resulting optimization problems are typically non-convex and nonlinear, in both objectives and constraints.
- *Task-related Semantic Information Transmission in Task-oriented SemCom:* In task-oriented SemCom systems, the resource allocation is closely tied to the task-related importance of the semantic information. For example, tasks involving safety-critical or context-rich data transmission (e.g., autonomous driving) may need to acquire high semantic fidelity, while other types of tasks may tolerate coarse-grained transmission. This task dependence necessitates adaptive resource allocation schemes that align with task-related semantic information and their utility, to complete the transmission of task-related and high semantic-importance features. At the same time, it ensures the allocation of other resources (bandwidth, power, computing resources) to jointly optimize the overall system performance.

These challenges motivate the development of novel optimization formulations and solution algorithms, as will be discussed in the following sections.

4) Resource to Be Allocated in SemCom: Generally speaking, the current research on resource allocation mainly involves computing, communication, and storage resources, with the following resources typically requiring allocation.

- Computing resources: The computing frequency of CPUs/GPUs on the BS or user side, also known as computing capacity.
- Communication resources: The wireless resources used by BS or clients for data transmission, including bandwidth, power, etc.
- Network parameter resources: The network-parameter resources are the parameter settings in the SemCom system, including the semantic compression ratio, the neural network parameters, and other parameters or policy settings.
- Storage resources: Edge servers or BS use these hardware storage resources to cache computing tasks and popular content (such as road monitoring).

In this paper, we summarize the resources to be allocated in the literature in Tables IX, XI, and XIII, and we need to mention that the storage resources are omitted since only one work [82] considered them. The symbol "–" in the tables indicates that this particular resource type is not allocated.

In this section, we outline the foundational structure of the SemCom resource allocation. It begins with an introduction to SemCom, followed by an examination of key network models used in SemCom resource allocation. We then dive into the core aspects of resource allocation within SemCom and conclude with a preliminary review of the relevant literature, summarized in Table IV, which gives a preliminary summary of the literature on resource allocation in SemCom based on source modal, communication mode, multiple access mode, and resource allocation type. In the table, one(many)-to-many means one(many) BS(s)/edge server(s) to many users/end devices (EDs). Furthermore, we use the symbol "–" to indicate that this property is not presented in the paper.

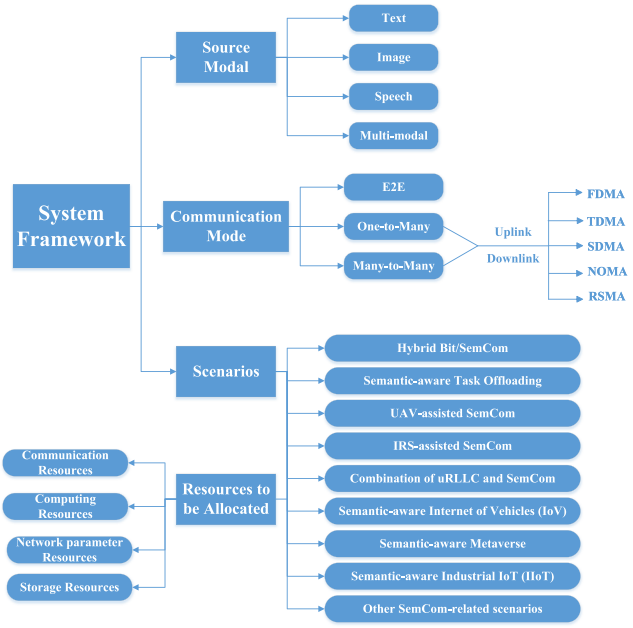

Fig. 5. The taxonomy of system framework establishment.

III. Performance Metrics of Resource Allocation

Building upon the previous section, this section will review the research on performance metrics in SemCom and the formation of optimization objectives in different literature.

Usually, we evaluate a communication system based on its accuracy and effectiveness. The traditional communication method is measured by the bit error rate and the bit transmission rate. For SemCom, accuracy can be measured by task performance and quantified by semantic similarity of text transmission, character error rate of speech recognition, etc. However, the efficiency of SemCom is usually difficult to measure and quantify [121]. As a result, it is critical and challenging to establish new performance metrics for SemCom resource allocation. At present, the research of SemCom resource allocation on constructing optimization objectives is mainly divided into two methods: based on traditional resource allocation performance metrics such as energy consumption, delay, and utility; and establishing new semantic-related performance metrics. See below for details.

In different articles, the symbols for the same variables may be inconsistent. To improve the reader's understanding of the composition of these performance metrics, this paper modifies

TABLE IV
COMPARISON OF PAPERS FOCUSING ON DIFFERENT SOURCE MODAL, COMMUNICATION MODES, AND SCENARIOS

Ref.	Source Modal	Communication Mode		Scenarios
[83]	Text	E2E	Uplink	E2E communication between UAV and BS
[84]			Downlink	No special cases
[73]				Hybrid bitCom/SemCom system
[36]		One-to-many	Uplink-FDMA	No special cases
[74]				SemRelay-aided Hybrid bitCom/SemCom
[44], [45]			Downlink-FDMA	SemRelay-aided Hybrid bitCom/SemCom
[46]				Physical layer security
[49]			Uplink-OFDMA	Fog radio access networks, intelligent computing task, and computation offloading
[50]				Machine translation and task offloading
[51]				SemCom over energy harvesting networks
[52]				Networks with limited resources
[53]			Downlink-OFDMA	Semantic bit quantization
[40], [54]				No special cases
[55]				UAV-assisted and spectral sharing
[56]				Multi-cell network
[62]			Uplink-TDMA	Energy-efficent semantic-aware wireless networks
[67]			Downlink-RSMA	RSMA downlink SemCom system
[68]				RSMA SemCom with uRLLC service
[69], [70]				Combination of PSC and RSMA
[64], [65]			Uplink-SDMA	Combination of PSC and SDMA
[75]			Downlink-OMA+NOMA	Hybrid bitCom/SemCom
[85]			Uplink	Energy harvesting of hybrid access point
[86]				Semantic relevance-based task-oriented communications with query-aware semantic encoder
[76]				Hybrid bitCom/SemCom
[87]–[89]				No special cases
[90]			Downlink	Intelligent reflecting surface (IRS)-assisted SemCom
[91]–[93]				Integrated sensing and semantic communication (ISSC), secure resource allocation
[94]				IRS-assisted semantic spectrum sharing network
[95], [96]		Many-to-many	Uplink-OFDMA	Semantic-aware 5G-Vehicle to everything heterogeneous Networks
[97]			Downlink	No special cases
[98]				Different knowledge matching degree
[99]				Coexistence of textual SemCom service and URLLC service
[100]				Multi-cell downlink PSC system
[101]				Combination of joint processing (JP) and SemCom
[102]	Image	E2E	-	IoV and image transmission in D2D communication
[57]		One-to-many	Downlink-OFDMA	Coexistence of uRLLC and SemCom system
[58]				Image-to-text semantic transmission
[103]				IRS-enhanced secure semantic communication (IRS-SSC)
[104]				UAV-assisted SemCom
[61]			Uplink-TDMA+NOMA	Integration of EH, CR and NOMA
[71]			Downlink-RSMA	Metaverse 3D construction
[105]			Uplink-FDMA	Joint optimization while training an image SemCom network
[106]			Uplink	No special cases
[107]				IoV; target recognition; FL-based SemCom model
[108]				SemCom-enabled IIoT system; workpiece surface defect classification
[109]				Feature importance perception and image classification
[53]				Multi-user IoT system with adaptable semantic compression (ASC)
[110]				Human pose estimation and DT construction
[77]				Hybrid BitCom and SemCom
[111]				Image classification
[112]			Uplink and downlink	Image classification
[113]			UAV to users	UAV-assisted DT construction
[114]			Downlink	Semantic IoT-based image retrieval services, defense to potential adversarial attacks

(Continued)

the expressions in some literature and unifies the mathematical expressions of common variables in different literature, as shown in Table V.

In most of the literature, for the subscript of a single variable, we use n to represent the n-th user, m to represent the m-th subchannel, b to represent the index of BS, $x_{n,m} = 1$ to represent the association of user n and subchannel m, and $x_{n,m} = 0$ to represent disassociation. In some other references, the subscript may refer to a task, a user group in a cellular cell, etc. At this time, we follow the expressions in their work and provide additional descriptions.

A. Traditional Performance Metrics in Resource Allocation

1) Energy Consumption and Time Delay: Energy consumption and delay/latency are two of the most traditional and commonly used performance metrics in resource allocation.

TABLE IV
(Continued.) COMPARISON OF PAPERS FOCUSING ON DIFFERENT SOURCE MODAL, COMMUNICATION MODES, AND SCENARIOS

Ref.	Source Modal	Communication Mode		Scenario
[58]	Image	Many-to-many	Downlink-OFDMA	Image-to-text semantic transmission
[115]			Uplink	Metaverse
[116]				Heterogeneous wireless networks, energy-aware image-based SemCom
[117]				UAV-assisted, target detection, Re-identification (Re-ID), and cloud-edge collaboration
[118]			Downlink	F-RAN and image semantic segmentation
[119]				Multi-cell multi-user MIMO system
[59]	Multi-modal	One-to-many	Uplink-OFDMA	Semantic-aware dynamic long-term MEC systems
[60]				Task offloading
[120]				Multi-modal secure SemCom, VQA
[47]			Downlink-FDMA	Multi-modal semantic transmission of UAV
[22], [121]		Many-to-many	Uplink	Multi-cell/multi-modal task network; VQA
[122]			Downlink	Space-air-ground integrated networks (SAGINs), hybrid bit and semantic communications
[123]		Vehicles to BS/ Vehicle to Vehicle (V2V)	Uplink-OFDMA	Cellular vehicle-to-everything (C-V2X) multi-modal communication platooning systems
[82], [124], [125]	Video	One-to-many	Uplink	IoV and target detection
[126]	-	E2E	Uplink	Semantic coding model
[127]				Integrated sensing, communication and computation (ISCC)
[128]		One-to-many	Uplink	PSC in IIoT
[129]				Edge-assisted SemCom network for IoT devices
[130]			Uplink and downlink	Joint sensing and communication model
[72]			Downlink-RSMA	Satellite-integrated RSMA SemCom downlink system
[48]		Many-to-many	Uplink-FDMA	Adaptive SemCom
[131]			Uplink-OFDMA	Cyber-physical system, data-driven decision making, edge learning, federated learning, distributed optimization
[78]			Uplink and downlink	Hybrid bitCom/SemCom
[66]		BS-RISs-Users	Downlink-SDMA(RIS to users)	RIS-assisted PSC in IIoT
[132]			Downlink	Distributed RISs assisted PSC

TABLE V
VARIABLES DESCRIPTION

Variables	Single	Total
Bandwidth	w	W
Power	p	P
Time delay	t	T
Energy consumption	e	E
Bit rate	r	-
Semantic rate	Γ	Γ
Original text sentence	s	-
Recovered text sentence	$\hat{s}$	-
Expected values of text sentence length	L	-
Expected values of text sentence information	I	-
Average number of semantic symbols for each word	k	-
Semantic similarity	ξ	ξ
Subchannel association	x	-
Channel condition/SNR	γ	-
Utility function	-	U

For applications sensitive to delay, the design of a resource allocation algorithm to reduce latency is one of the main concerns [42], [58], [60], [73]. Delay modeling generally includes the following parts: a) semantic extraction latency at the transmitter (T^1); b) transmission latency (T^2); and c) semantic recovery latency or task process latency at the receiver (T^3).

We previously mentioned in Section I-B3) that the influence of the semantic compression ratio is the computing-transmission tradeoff in latency. For better understanding, we simply model the latency of a single user in the semantic-aware task process scenario, compression latency is

$$T^1 = \frac{F(\rho, D)}{f_e}, \tag{14}$$

ρ is the compression ratio, f_e is the computing capacity at the transmitter, and $F(\rho, D)$ is the required compression CPU cycles, which might be different across the literature. For instance, [60] modeled $F(\rho, D)$ as

$$F(\rho, D) = \frac{\alpha D}{\rho^\beta}, \tag{15}$$

where $\alpha > 0$, $\beta > 0$ are constants relevant to the tasks. Transmission latency is

$$T^2 = \frac{\rho D}{R}, \tag{16}$$

R is the transmission rate. Computing latency is

$$T^3 = \frac{\rho w D G}{f_r}, \tag{17}$$

where w is the required CPU cycles per bit to process the task, f_r is the allocated computing capacity at the receiver. We use G to denote the ratio of computation intensity of semantic data to that of raw data. The increase is caused by computations for processing semantic data and compensations for enhancing accuracy [59]. G can be denoted as

$$G = \frac{1}{\rho^c}, \tag{18}$$

where c is a constant related to specific tasks. Fig. 6 shows the relation of compress ratio and G, where $\rho_{\min}$ is the minimum compression ratio to maintain the integrity of source information or task, which can vary from different tasks/users/information modalities.

So the total latency is

$$T = T^1 + T^2 + T^3 = \frac{F(\rho, D)}{f_e} + \frac{\alpha D}{\rho^\beta} + \frac{\rho w D G}{f_r}. \tag{19}$$

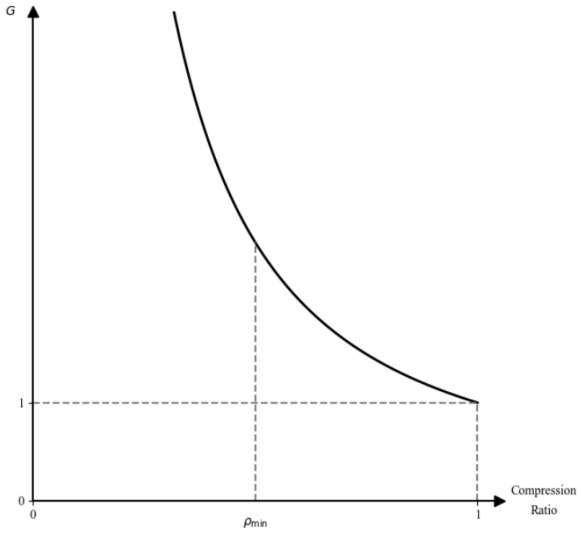

Fig. 6. The relation of G and compression ratio ρ.

It is obvious that compression ratio affect all parts of the total latency, and controls the tradeoff between computing (T^1 and T^3) and transmission (T^2).

When the focus of the article is on "Energy Efficiency", the total energy consumption of the entire system is often used as a performance metric [50], [59], [62], [67], [87], [114]. In most cases, delay and energy consumption are contradictory. The other often becomes a constraint when one is the optimization goal. In [50], the authors proposed a semantic-aware energy-saving task offloading network model. The goal is to extend the battery life of local users, so the sum of local users' energy consumption is used as the objective function. Considering the power shortage of mobile devices, the study in [59] is committed to the allocation of resources for semantic-aware MEC systems to minimize energy consumption. As discussed in [67], the authors modeled the delay and total energy consumption of a single user that consists of these three parts. The goal is to minimize the total energy consumption of the entire system, considering constraints such as delay.

We also previously mentioned in Section I-B3) that the influence of the semantic compression ratio is the computing-transmission tradeoff in energy consumption. Similarly, we also simply model the energy consumption in the semantic-aware task process scenario. The energy consumption of semantic compression can be denoted as

$$E^1 = \kappa F(\rho, D) f_e^2, \tag{20}$$

where κ is a constant coefficient. $F(\rho, D)$ also denotes the CPU cycles required to compress the data D to ρD. The transmission energy is

$$E^2 = pT^2 = p\frac{\rho D}{R}, \tag{21}$$

where p is the transmission power. And the task computing energy can be denoted as

$$E^3 = \kappa(\rho w D G) f_r^2. \tag{22}$$

In many one-to-many uplink wireless communication scenarios, only the transmitter's energy consumption ($E^t = E^1 + E^2$) needs to be considered as a constraint since the user sides (like mobile devices) often have energy budgets. However, there are some other scenarios like energy minimization of energy efficient communication system, which needs to consider the total energy consumption of both transmitter and receiver ($E^1 + E^2 + E^3$).

To better show the relations of compression ratio and latency/energy consumption, we illustrate them in Fig. 7a and Fig. 7b, where ρ_t^* is the optimal compression ratio for the minimum latency of single user and ρ_e^* is the optimal compression ratio for the minimum energy consumption of single user. In the figures, the range of compression ratio ρ is in $[\rho_{\min}, 1]$ due to the ρ below threshold $\rho_{\min}$ can not maintain the integrity of source information or task. $\rho_{\min}$ can vary from different tasks/users/information modalities, here we set it to 0.5 for illustration,

In Fig. 7a and Fig. 7b, we can notice that the transmission delay/energy decreases with decreasing compression ratio, the computation delay/energy increases with the decrease of compression ratio, and the optimal compression ratio to reach the minimum value of latency and energy consumption is different, thus leading to a tradeoff in computing and transmission. (Note: This relation may vary in different system models and with different users. The relation in Fig. 7 is an example illustration of a certain user.)

2) Utility Function: The concept of utility in resource allocation refers mainly to the satisfaction of users under a certain resource allocation scheme. Utility is generally expressed by the utility function. According to various objectives, the utility function is represented and mathematically transformed by different quality of service parameters, such as data transmission rate, delay, energy consumption, and cost, which can achieve a better overall effect. The mathematical transformation mainly includes reciprocal, logarithmic, and weighted summation. Finally, an effective optimization algorithm is designed to maximize the utility [46], [48], [49], [117], [118]. For example, the utility function established in the literature [117] is shown in Eq. (23):

$$U = \beta_1 A - \beta_2 T - \beta_3 E, \tag{23}$$

where A is the total task accuracy, T is the total time delay, E is the total energy consumption, and $\beta_1, \beta_2, \beta_3$ are the weight factors.

3) Traditional QoS and QoE:

- *Quality of Service (QoS):* Defined by the International Telecommunication Union (ITU) as "the totality of characteristics of a telecommunications service that bear on its ability to satisfy the stated and implied needs of the user." It primarily focuses on system performance measured through physical parameters [133].
- *Quality of Experience (QoE):* Refers to users' subjective perception of the system or service performance, influenced by context, culture, expectations, psychological factors, and more [133].

In resource allocation for wireless communications, QoS modeling is often similar to the utility function, but the mathematical complexity is higher than the general utility function. In [47], the QoS modeling based on the transmission

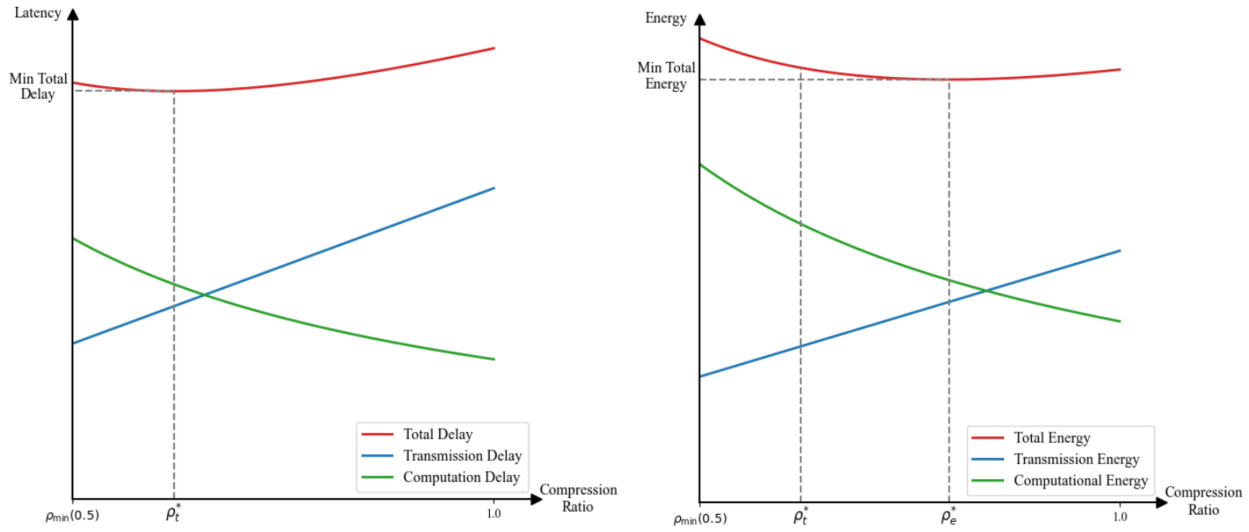

Fig. 7. Latency and energy consumption versus semantic compression ratio.

delay and the number of received semantic information is shown in Eq. (24).

$$QoS_{m,n}(t) = \frac{1}{\left[1+e^{\beta_T(T_{m,n}(t)-T_{th})}\right]\left[1+e^{\beta_H A_m(t)(H_{th}-\tilde{H}_{m,n}(t))}\right]}, \tag{24}$$

The two terms on the right side of the equation represent the transmission delay score and the received semantic information score of the user *n*, respectively, where T_{th} and H_{th} are the transmission delay thresholds and the received semantic information and β_T, β_H are the weight factors of the delay in time and the received semantic information.

The primary goal of wireless communication network services is to provide a user-satisfied quality of experience (QoE) that is more user-centric. QoS does not contain any human-related quality factors, which means that for two different users, the same level of QoS may not guarantee the same level of QoE [134]. Designing QoE and managing it while providing a service is necessary for high-quality experiences. This requires assessment methodologies that can quantify QoE [135]. Reference [71] studied the transmission of image semantic information in the Metaverse 3D construction. Data rate, bit error rate (BER) and interest score (the degree of interest in the image after semantic segmentation, which is related to people) are considered when modeling QoE. In Section III-C of this paper, the QoE in the context of Metaverse and SemCom is introduced.

B. Semantic Similarity

Semantic similarity is defined as the degree of similarity between the sender and the receiver's semantic information under a specific semantic task. For task-oriented SemCom, semantic similarity can be extended to semantic fidelity. The specific representation of semantic fidelity varies with different target tasks. For automated tasks that do not require manual supervision and data reconstruction, such as text sentiment classification, image classification, and target detection, semantic fidelity can be expressed as average classification accuracy or detection accuracy. The establishment of most new performance metrics for SemCom resource allocation must rely on the concept of semantic similarity, so this section will detail the current definition of various types of semantic similarity. The comparison of different types of semantic similarities is presented in Table VI.

1) Semantic Similarity of Text Signal: For text transmission, BER does not reflect the performance well. In machine translation, bilingual evaluation understudy (BLEU) scores are generally used to measure results [136]. However, the BLEU score can only compare the differences between words in two sentences, but cannot compare their semantic information. BLEU outputs a number between 0 and 1, representing the similarity between two sentences, with 1 representing the highest similarity. However, word errors may not alter the meaning of sentences. For example, the two sentences "That car had been deserted" and "That vehicle had been abandoned" have the same meaning, but their BLEU scores are different due to the use of different words to represent "car" and "deserted", which is a flaw in BLEU's recognition of synonyms. A word can have different meanings in different contexts. For example, "bus" can have different meanings in terms of public transportation and a microcomputer. Traditional methods, such as word2vec [137], cannot recognize a polysemy. The problem is how to represent the word with a numerical vector, which is different in different contexts [25].

Therefore, based on the bidirectional encoder representation from transformers (BERT) model [138], Reference [25] proposed a new metric, Sentence Similarity, which describes the similarity of two sentences according to their semantic information, as shown in Eq. (25).

$$\xi = \frac{B_\Phi(s)\cdot B_\Phi(\hat{s})^T}{\|B_\Phi(s)\|\|B_\Phi(\hat{s})\|}, \tag{25}$$

TABLE VI
COMPARISON BETWEEN DIFFERENT SEMANTIC SIMILARITIES

Modal	Metric Name	Semantic Awareness	Description	Strengths(+)/ Weaknesses(-)
Text	Sentence Similarity	✓	Based on BERT, describes the similarity of two sentences according to semantic information	+ Recognize semantically equivalent expressions, more appropriate to SemCom – Not easy to generalize the pre-trained BERT model on other unseen domains
	Semantic Accuracy	✓	Based on token matching, describes the degree of correctness of the information in the recovered text	+ Useful for assessing the correctness of the sentences - Cannot independently assess the overall transmission quality, token matching approach struggles with paraphrasing, synonym usage
	Semantic Completeness	✓	Based on token matching, describes the degree of the original information contained in the recovered text	+ Useful for assessing if key information is missing - Cannot independently assess the overall transmission quality
	Metric of semantic similarity (MSS)	✓	A composite function of semantic accuracy and semantic completeness, which can control the tradeoff between semantic accuracy and semantic completeness	+ Can assess the overall transmission quality while avoiding semantic errors caused by word vectorization, flexible in controlling the tradeoff between accuracy and completeness - May underestimate the semantic similarity
	Bilingual evaluation understudy (BLEU)	✗	Only compare the differences between words in two sentences, ignore the meaning, flaw in recognition of synonyms	+ Easy and fast to compute - Ignores semantic meaning, synonyms, and paraphrases
Image	Image-to-graph semantic similarity (ISS)	✓	The cosine angle between an image vector and its corresponding normalized semantic triple vectors	+ Directly capture the correlation between the original image and its semantic information - Pre-trained DNN have limited generalization in other domains
	Metric for image semantic transmission (MIST)	✓	Combines the importance weight of each semantic information with their respective transmission quality (i.e. SSIM) to obtain the final evaluation results	+ Provides a more comprehensive evaluation than traditional metrics - More computationally complex than standard metrics, need to assign parameters to adjust the importance among different semantic information
	Peak signal-to-noise ratio (PSNR)	✗	Based on the errors between corresponding pixel points	+ Simple and fast to compute - Error-sensitive, do not take the characteristics of human vision into account, and the evaluation results often do not align with human perception
	Structural similarity index measure (SSIM)	✗	Measures the difference between the original and the reconstructed image in terms of brightness, contrast, and structural similarity	+ Easy to implement, aligns more closely with the human visual system compared with PSNR - Particularly sensitive to relative translations, rotations, and scalings of the image, less reliable when evaluating images degraded by blurring or noise

where B_Φ represents the BERT model. The sentence similarity defined in Eq. (25) is a number between 0 and 1, which represents the similarity between the decoded sentence and the transmitted sentence; 1 represents the highest similarity, and 0 represents no similarity.

Currently, to measure text semantic similarity, most of the literature [36], [52], [53], [55], [89], [90], [101] uses sentence similarity based on the BERT model as semantic similarity. However, the authors of [40] proposed a metric of semantic similarity (MSS), which is a function of semantic accuracy and completeness. Based on token matching [139], semantic accuracy is defined as the ratio of the sum of the correct occurrences of each token in the recovered text to the sum of the occurrences of each token in the recovered text. Semantic completeness is defined as the ratio of the sum of the correct occurrences of each token in the recovered text to the sum of the occurrences of each token in the original text. Due to the high complexity of the expressions, we omit the explicit expression of MSS. Reference [40] includes a detailed description of these metrics.

2) Image-to-Graph Semantic Similarity: Although most of the current work in the resource allocation of SemCom is text and image modalities, the work of [58] and [42] combines the two in semantic extraction and establishes an image-to-text semantic information extraction method. The semantic information in the image is extracted into a scene graph (SG) in the form of text, which captures the objects and their relationships in the original image. This interpretable semantic information can not only be directly read and understood by humans but also be used to generate original images and retrieve similar images.

[42] introduced a comprehensive image-to-graph semantic similarity (ISS) metric, which uses a pre-trained deep neural network (DNN) to directly capture the correlation between the original image and its semantic information without any reconstruction of the image. The DNN is trained by Webimagetext [140], a dataset of 400 million image-text pairs collected from the Internet. Compared with the structural similarity index measure (SSIM) [141], which measures the difference between the original image and the reconstructed image on a set of pixels, the DNN can be used directly to obtain the image vector and the semantic information vector of the received SG. The ISS metric is defined as the cosine of the angle between the image vector and its normalized semantic triplet vector, which is calculated by the projection of the image vector on the set of semantic information vectors. The specific calculation steps and formulas are detailed in [42].

3) Semantic Similarity of Image Signal: The semantic similarity of the image signal is used to measure the similarity between the original image and the restored image. The

TABLE VII
MAPPING BETWEEN RESOURCE TYPES AND SEMANTIC PERFORMANCE METRICS

Resource Type	Resource	Influence Metrics	Description
Communi-cation	Bandwidth, Transmit Power	**All metrics related to transmit rate** (e.g. STM, **Transmit Latency**, S-R)	According to Shannon's theorem: $R = W\log_2(1+\text{SNR})$, SNR is influence by bandwidth and transmit power, so higher bandwidth and transmit power allows faster transmission.
	Transmit Power	Energy consumption, S-EE	Higher transmit power will consume more energy.
		All metrics related to semantic accuracy/similarity (e.g. S-SE, Semantic QoE, Semantic Score)	Affects received signal quality and thus recovery accuracy; stronger signals improve semantic recovery.
	Subchannel Allocation	All metrics related to multi-user FDMA transmission (e.g. ES-SE, SC-QoS, STM)	Controls how subchannels are distributed across users/tasks, each subchannel has different channel conditions, thus affecting throughput, delay, and reliability.
Computing	User/MEC computing capacity	Energy	Use f to denote computing capacity. Computation energy can be denoted as: $E = \kappa\gamma f^2$, where c a constant coefficient, γ is the required CPU cycles to process the task.
		Latency	User computing capacity affects the local execution latency, MEC computing capacity affects the offloading execution latency. Use the same expression as above, we have execution latency: $T = \frac{\gamma}{f}$.
		EoSI [108]	Local and MEC computing resources affect the cost function in the expression of EoSI [108].
	Offloading Decision	Energy and Latency	Offloading tasks to the MEC server reduces user-side energy and delay but possibly increasing transmission delay and energy consumption. So it controls the tradeoff between computation and communication cost.
Network Parameters	Semantic Compress Ratio	Latency	Smaller compress ratio can reduce the data volume to be transmitted, thus reduce the transmission latency. But it will result local extraction latency at transmitter and recover latency (or increased task computing cycles, which affect task execution latency). So the allocation of semantic compress ratio is a tradeoff of computing and communication, which is very important in SemCom. Detailed description is in Section III.A.
		Energy Consumption	The compression of data requires extraction CPU cycles, thus generate extraction energy consumption. Reduced data volume will decrease the transmission delay, thus reducing the transmission energy. But it also result in the recover energy consumption, or the additional task computing cycles, which will affect task execution energy consumption. It is also a tradeoff of computing and communication. Detailed description is in Section III.A.
		Task Accuracy, Success Probability of Tasks	Affect the task processing at receiving, low compress ratio will influence the task accuracy. Therefore, how to find the right compression ratio to achieve an optimal tradeoff between transmission costs and semantic correctness is another critical issue in the resource allocation of SemCom. Fig. 9 shows the typical relation between accuracy and compression ratio. (Note: This relation may vary in different system modeling, the relation in Fig. 9 is a typical type.)

more classical method is measured by the peak signal-to-noise ratio (PSNR), which is based on the errors between corresponding pixel points. In the previous section, SSIM is mentioned. It is widely used in the application of image similarity measurement, including the resource allocation of SemCom [47]. These two metrics are used mainly in the image signal similarity evaluation. However, in [113], a metric for image semantic transmission (MIST) is proposed, which combines the importance weight of each semantic information with its respective transmission quality to obtain the final evaluation results. After capturing the image, the UAV sends it to the user and first extracts the semantic information through the target detector. Specifically, a total of U^O objects are detected, where i represents the i-th object and c_i represents its corresponding confidence. The relationship between the importance score Δ_i and the confidence c_i of the object i can be expressed as $\Delta_i = c_i^\sigma$, where σ is a variable that regulates the importance between different semantic information. The final MIST can be expressed as follows:

$$E(A, \Delta_i, Q(p_i)) = A\sum_{i=1}^{U}(\Delta_i \times Q(p_i)), \tag{26}$$

where A represents the accuracy of extracting semantic information, and $Q(p_i)$ represents the SSIM value of target i before and after transmission, which is a function that is positively correlated with the transmission power p_i [43].

C. New Performance Metrics for SemCom

As mentioned above, the traditional resource allocation model is usually modeled based on Shannon capacity, which fails to give full play to the performance advantages of SemCom to ensure the best performance of the SemCom network. SemCom does not require error-free transmission of bits or symbols, so the optimization problem based on Shannon capacity construction may reduce system performance. Therefore, it is essential to reconsider resource utilization from a semantic perspective to develop new performance metrics [142].

Similarly, we will give a systematic summary and comparison of these new metrics in Table VIII, including a critical evaluation of their strengths,limitations and suitability for different modalities and applications. Considering that most of these new metrics are based on the concept of semantic similarity, we illustrate the connections and evolution of the semantic similarity-based metrics in Fig. 8.

1) Semantic Transmission Rate and Semantic Spectral Efficiency: Firstly, reference [36] assumes that the semantic unit (sut), representing the basic unit of semantic information, can measure semantic information in the text transmission scenario. Then, two critical semantic-based performance metrics are defined: *semantic transmission rate* (S-R) and *semantic spectral efficiency* (S-SE).

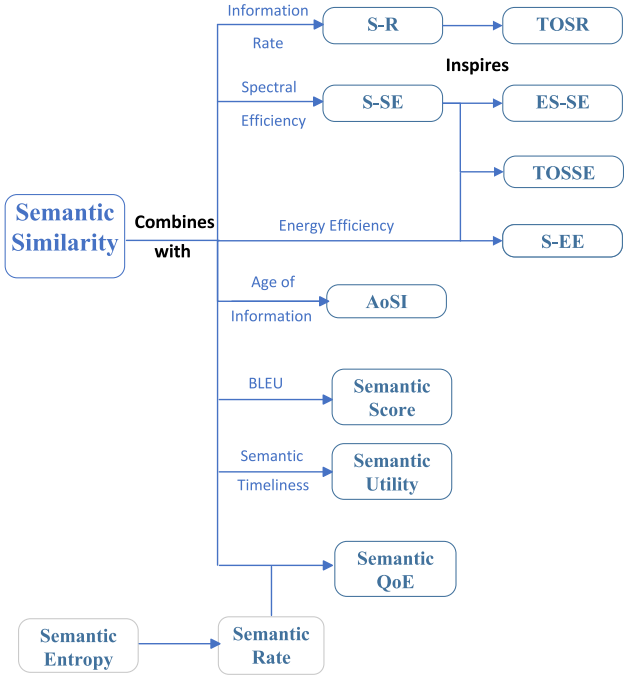

Fig. 8. Metrics that based on semantic similarity, their connections.

- S-R: S-R refers to the effective transmission of semantic information per second, measured by suts/s, as follows:

$$\Gamma_{n,m} = \frac{WI}{k_n L}\xi_{\mathrm{n,m}}, \tag{27}$$

all subchannel bandwidth was allocated equally, using W to represent the subchannel bandwidth. Since the article focuses on long-term text transmission rather than the transmission of a single sentence, *I, L* should take the expected value and not the random value, that is, for each user *n*, *I/L* is a fixed value, so omit the subscript *n*. The unit of I/k_nL is suts/symbol, and the channel bandwidth of the band pass transmission in the ideal state is equal to the symbol rate (unit: symbol/s), so the unit becomes suts/s after multiplying by W. The semantic similarity based on BERT $\xi_{\mathrm{n,m}}$ depends on the structure of the DeepSC neural network k_n and the channel conditions $\gamma_{n,m}$. It can be expressed as $\xi_{n,m} = f(k_n, \gamma_{n,m})$.
- S-SE: S-SE refers to the rate at which semantic information is successfully transmitted within a unit bandwidth, measured by suts/s·Hz, as follows:

$$\Phi_{n,m} = \frac{\Gamma_{n,m}}{W} = \frac{I}{k_n L}\xi_{\mathrm{n,m}}, \tag{28}$$

The proposal of S-R and S-SE provides an important theoretical basis for many subsequent studies such as [88], [90], [109]. Based on these two metrics, they made expansions and cross-domain transformations, and we now continue with our discussion of them.

2) Effective S-SE: The study in [90] considered the requirement of semantic information similarity for downstream semantic tasks, and the concept of *effective semantic spectral efficiency* (ES-SE) is introduced. The serious deviation of semantic similarity will directly lead to inaccurate results. Only SemCom that reaches the semantic similarity threshold ξ_{th} required by the downstream task is considered effective. Let $\eta_{n,m}$ denote whether the user *n* performs an effective semantic transmission on the subchannel *m*. If $\xi_{n,m} > \xi_{th}$, then $\eta_{n,m} = 1$; otherwise, $\eta_{n,m} = 0$. Ψ is called ES-SE, which can be expressed as:

$$\Psi = \sum_{n=1}^{N}\sum_{m=1}^{M} x_{n,m}\eta_{n,m}\Phi_{n,m}, \tag{29}$$

where $\Phi_{n,m}$ is the S-SE of the user *n* in the subchannel *m*.

3) Task-Oriented S-R and S-SE: The authors of [109] integrate S-R and S-SE into the scenario of feature importance-aware image classification, and two performance metrics of the task-oriented SemCom system are defined: *task-oriented semantic transmission rate* (TOSR) and *task-oriented semantic spectral efficiency* (TOSSE). Unlike the definition of [36], which considers long-term text transmission rather than single-sentence transmission, the work of [109] focuses on the performance of each user. When a semantic transmission time slot begins, there are S semantic features after joint source-channel coding (JSCC). BS obtains the feature transmission rate decision vector $\mathbf{r}^f$ based on the channel conditions and the historical data distribution of each user. Then r_n^f is fed back to the feature selection module in each user *n* to determine the number of features that need to be transmitted: $S_n = r_n^f S/2$. Therefore, the average semantic information for each symbol of user *n* is $I_{n,m}/S_n$.

- TOSR: TOSR refers to the amount of semantic information effectively transmitted per second for a specific task. The expression is as follows:

$$\psi_{n,m} = \frac{WI_{n,m}}{S_n}\xi_{n,m}. \tag{30}$$

Compared to [36], it is equivalent to replacing I/k_nL (unit: *sut*/symbol) with $I_{n,m}/S_n$ (unit: *sut*/symbol) of Eq. (27) in this paper, while the other parts remain unchanged.
- TOSSE: TOSSE refers to the rate at which task-related semantic information is successfully transmitted through a single bandwidth unit. The expression is as follows:

$$\phi_{n,m} = \frac{\psi_{n,m}}{W} = \frac{I_{n,m}}{S_n}\xi_{n,m}. \tag{31}$$

4) Semantic Energy Efficiency: Based on the concept of S-R [36], semantic energy efficiency (S-EE) is introduced in [88] as a measure of energy efficiency in the SemCom system, which is quantified by suts/Joule. Traditional communication systems define energy efficiency as the number of bits that the system can transmit per unit of consumed energy. From a semantic point of view, the feature of S-EE is the number of semantic symbols transmitted by unit energy consumption. It is expressed as the S-R ratio that can be achieved by the total power consumed in the SemCom network. The S-EE of user *n* is denoted by:

$$E_n = \frac{\Gamma_n}{p_n + p^c} = \frac{w_n I}{(p_n + p^c)k_n L}\xi_n, \tag{32}$$

p_n is the transmit power of user n, p^c is the electrical power that the circuit consumed, and b_n is the bandwidth. The Γ_n here represents the S-R of user n.

5) Semantic Entropy: Semantic information relies not only on the source data, but also on the specific task, which is significantly different from the information defined by Shannon. Consequently, the same data may contain different amounts of semantic information for different tasks. In this regard, the authors of [121] defined the semantic entropy as follows.

Definition 1: Given semantic source $\mathcal{X}$, semantic entropy is defined as the minimum average number of semantic symbols about data $X \in \mathcal{X}$ that is sufficient to predict task Y, i.e.,

$$\begin{aligned} H(X;Y) &\triangleq \min_{E_S} \mathbb{E}\big(\dim\Big(Code^{E_S}(X)\Big)\big), E_S \in \mathcal{E}_S \\ \text{s.t.} \quad & P\Big(Y|Code^{E_S}(X)\Big) = P(Y|X), \end{aligned} \tag{33}$$

where $Code^{E_S}(X))$ denotes the semantic symbol vector extracted from X with the semantic encoder E_S, $\mathcal{E}_S$ is the set of semantic encoders, and $P(Y|X)$ is the conditional probability of achieving the goal of Y given X.

The constraint in Definition 1 implies that the defined semantic entropy is lossless and that it is actually defined as an expected value throughout the data set $\mathcal{X}$, that is, the semantic entropy is constant for the same task and dataset. However, it is intractable to find an optimal semantic encoder, E_S^*, to derive the semantic entropy [143]. To obtain a measure that is both meaningful and manipulable for semantic communication systems, [121] utilize a well-designed DL model as the encoder to obtain an approximate semantic entropy for a task, which is:

$$\begin{aligned} H(X;Y) &\triangleq \min \mathbb{E}\big(\dim\Big(Code^{E_{DL}}(X)\Big)\big) \\ \text{s.t.} \quad & P(Y|X) - P\Big(Y|Code^{E_{DL}}(X)\Big) < \varepsilon, \end{aligned} \tag{34}$$

where the constraint indicates that the task performance degradation can not exceed ε. From Eq. (34), the defined approximate semantic entropy is lossy. According to the aforementioned method, the approximate semantic entropy of the considered tasks can be derived based the corresponding DL models. Therefore, [121] use semantic entropy to construct the semantic rate and semantic QoE model. We now move on to this semantic entropy-based metric - semantic QoE.

6) Semantic QoE: The accuracy and efficiency of message transmission are different from the user's point of view, and depending on the application, users may have their own preferences for them. For example, some users prefer higher accuracy but have a certain tolerance for delay, while some users may want to get a higher rate but do not need high accuracy [144]. The semantic rate of user based on semantic entropy is given as

$$\varphi_n = \frac{\tilde{H}_{DL}}{k_n / W}, \tag{35}$$

where the meaning of W and k_n is the same as in Table V. $\tilde{H}_{DL}$ is the semantic entropy based on specific DL model. For text modal task, it can be DeepSC [25]. For bi-modal task, it can be DeepSC-VQA [145]. In order to more reasonably reflect the user's QoE requirements, reference [22] established a semantic QoE model, which is expressed as:

$$\begin{aligned} QoE_q^b &= \sum_{n \in \mathcal{G}_n^b} w_n G_n^{\mathrm{R}} + (1 - w_n) G_n^{\mathrm{A}} \\ &= \sum_{n \in \mathcal{G}_q^b} \frac{w_n}{1 + e^{\beta_n(\varphi_n^{\mathrm{req}} - \varphi_n)}} + \frac{(1 - w_n)}{1 + e^{\lambda_n(\xi_n^{\mathrm{req}} - \xi_n)}}. \end{aligned} \tag{36}$$

It should be noted that the authors of [22] modeled the complex situation of multi cell task and user. b denotes the cell index and q denotes the index of the user group in cells. In Eq. (36), $\mathcal{G}_q^b$ denotes the q-th user group in the b-th cell. w_n and $(1 - w_n)$ are the weights of the semantic rate φ_n and the semantic accuracy ξ_n on the user n, respectively. G_n^R and G_n^A are the semantic rate and semantic accuracy for user n, respectively. β_n and λ_n represent the growth rates of G_n^R and G_n^A. In addition, φ_n^{req} and ξ_n^{req} represent the minimum semantic rate and semantic accuracy of 50% scores [22].

7) QoE of Metaverse Service Providers: With the support of virtual reality (VR), augmented reality (AR), and the tactile Internet, Metaverse hardware devices cannot only mobilize all senses of the user and provide an immersive experience [146], but also revolutionize the way people interact with each other and even with objects. Therefore, it is crucial to design the QoE of Metaverse Service Providers (MSPs) as a performance indicator to measure the performance of Metaverse Service [147]. In the proposed framework in reference [71], the authors aim to transmit the semantic information of interest to each MSP. Therefore, the performance metrics of the data rate, the BER, and the interest rating should be considered together. Thus, the QoE of the k-th MSP U_k can be defined as [148]:

$$Q_k = \sum_{i=1}^{N_k} \mathcal{J}_k^i \mathcal{T}\Big(1 - \mathcal{B}_k^i\Big), \tag{37}$$

where N_k is the number of objects that U_k is interested, $\mathcal{J}_k^i$ is the normalized interest rating of U_k for the i-th object recommended to U_k, $\mathcal{T}$ is the normalized time that all MSPs finish the transmission, and $\mathcal{B}_k^i$ is the BER of transmitting the i-th object's semantic information to U_k.

8) System Throughput in Message: System throughput in message (STM) represents the network performance from a semantic point of view, proposed by [97]. In text communication, an entire text sentence ending in a cycle, or in voice communication, a completely emitted voice signal, can be regarded as a message. Taking this into account, the message rate (unit: msg/s) is interpreted as the number of messages transmitted or processed per unit time under the reference of the bit rate (unit: bit/s) definition. Because the system throughput has a very perfect expression of the system framework based on Shannon's theory:

$$S^T = \sum_n \sum_b x_{nb} r_{nb} = \sum_n \sum_b x_{nb} w_{nb} \log_2(1 + \gamma_{nb}). \tag{38}$$

Here, n and b represent the n-th user and the b-th BS, respectively. Among them, w_{nb} and γ_{nb} represent the bandwidth and SNR, respectively. The system throughput represents the number of bits successfully transmitted per unit time in the system, reflecting the network performance. Therefore, the authors of [97] defined a general bit-to-message (B2M) conversion function $S(\cdot)$, which is related to different semantic encoders, knowledge matching, and message properties. Therefore, according to the bit rate r_{nb} given by the Shannon theorem, the message rate $r_{nb}^{M} = S_n(r_{nb})$ can be naturally defined by $S(\cdot)$, and the expression of STM is derived as follows:

$$S^{TM} = \sum_{n}\sum_{b} x_{nb} r_{nb}^{M} = \sum_{n}\sum_{b} x_{nb} S_n(r_{nb}). \tag{39}$$

STM characterizes the number of messages successfully transmitted in the system per unit time, which can well characterize network performance from a semantic perspective.

9) Age of Semantic Information: In traditional communication systems, Age of Information (AoI) [149] is a popular measure of information importance, which is defined as $\Delta^{AoI}(t) = t - u(t)$ by measuring the information delay of the destination. *u(t)* is the generation time of the latest received data packet. In order to capture the freshness of information and semantic loss in the SemCom system, the literature [89] proposed a new measurement method called Age of Semantic Importance (AoSI). Before giving the definition of AoSI, the reference [89] first defined the semantic importance (SI): semantic loss caused by missing or incorrect semantic content [150]. It can be expressed as $\psi = 1 - \xi$. Here, ξ is the semantic similarity, which we discussed in the previous subsection. For example, in a text transmission task, semantic importance can be denoted as

$$\psi = 1 - \xi = 1 - \frac{\boldsymbol{B}(x) \cdot \boldsymbol{B}(\hat{x})^{T}}{\|\boldsymbol{B}(x)\| \cdot \|\boldsymbol{B}(\hat{x})\|}, \tag{40}$$

where $\boldsymbol{B}(\cdot)$ represents the BERT model. The definition of AoSI can be obtained by the definition of SI and AoI:

$$\Delta^{\mathrm{AoSI}}(t) = \Delta^{\mathrm{AoI}}(t) \cdot \psi(u(t)) = (t - u(t)) \cdot \psi(u(t)), \tag{41}$$

where $\psi(u(t))$ is the semantic importance of the last received packet.

10) Utility of Information: Reference [72] introduced a utility of information (UoI) metric. It encompasses multiple contextual attributes to capture the utility grade of the updates transmitted to communication systems or services. From a mathematical perspective, it can be modeled using a composite function:

$$\mathcal{U}(t) = (\Theta \circ U)(\mathcal{D}_t). \tag{42}$$

Here, $\Theta(\cdot) : \mathbb{R}^m \to [0, M]$ is a non-increasing function that converts the penalty into the corresponding utility grade. $U : \mathbb{R}^n \to \mathbb{R}^m, n \geq m$, is a non-decreasing non-linear penalty function with respect to the three attributes as follows:

$$\big(f(t), g(X_t, \hat{X}_t), C(X_t, d_t)\big) \in \mathcal{T} \times \mathcal{X} \times \mathcal{C} \xrightarrow{\mathcal{F}} U(\mathcal{D}_t) \in \mathbb{R}^m. \tag{43}$$

The first attribute is usually represented by a non-decreasing time penalty function $f(t) \in \mathcal{T}$, which includes metrics like AoI. The second attribute is captured by an error detection function $g(X_t, \hat{X}_t) \in \mathcal{X}$, typically encompassing metrics such as mean square error (MSE) or mean percentage error (MPE). The third comes from practical constraints like spectrum limit and energy consumption, denoted by a predefined function $C(X_t, d_t) \in \mathcal{C}$ based on source states X_t and action d_t, where the latter refers to the transmission policies like generation decisions, code rate, and resource allocation.

11) Efficiency of Semantic Information: In the context of the SemCom-Industrial Internet of Things (SemCom-IIoT), traditional performance metrics are no longer the best choice. As reported in [108], a new performance metric was designed at the semantic level, named Efficiency of Semantic Information (EoSI). The scenario is relatively different, and we need to state that *n, m* does not refer to the user and subchannel indexes only here. The intelligent sensing device (ISD) in the scene is divided into m categories, so the subscript of $ISD_{m,n}$ means the n-th ISD in the m-th class. The preliminary expression of EoSI is as follows:

$$EoSI_{m,n}(t) = \frac{UoSI_{m,n}(t)}{cost_{m,n}(t)}. \tag{44}$$

UoSI is semantic information utility: considering both semantic timeliness and task accuracy, the expression is as follows:

$$UoSI_{m,n}(t) = F_{m,n}^{d}(t) F_{m,n}^{a}(t). \tag{45}$$

Among them, task accuracy $F_{m,n}^{a}(t)$ quantifies the impact of semantic information on task accuracy, $F_{m,n}^{d}(t)$ quantifies the impact of the timeliness of semantic information on the timeliness of task results, and the timeliness of task results is also the standard for judging whether the task is successfully completed. $cost_{m,n}(t)$ represents the resource overhead of $ISD_{m,n}$ to complete intelligent tasks, which is a weighting function of bandwidth resources, local computing resources, and MEC computing resources. The complete expression of EoSI is complex. If you are interested in the details and the derivation process, see [108].

12) Success Probability of Tasks: In order to simultaneously evaluate the impact of transmission and adaptive semantic compression (ASC) on the performance of SemCom, a new performance metric is defined in [106]: success probability of tasks. Reference [111] further improved the work in reference [106] and also adapted this performance metric. According to [106], the definition of success transmission probability of users is first introduced, as follows:

$$P(t_n \leq t_0) = 2Q\left(\frac{2^{a_n(1-o_n)} - 1}{b_n \delta}\right), \tag{46}$$

where t_n is the transmission delay of user n, $P(\cdot)$ is the probability, and o_n is the semantic compression ratio. In practical scenarios, such as the Internet of Vehicles (IoV), a large number of tasks are delay sensitive, so there are always strict transmission delay constraints, represented by t_0. Therefore, the user's transmission success probability is $P(t_n \leq t_0)$. $a_n = \frac{d_0}{w_n t_0}$, w_n is the bandwidth of user n,

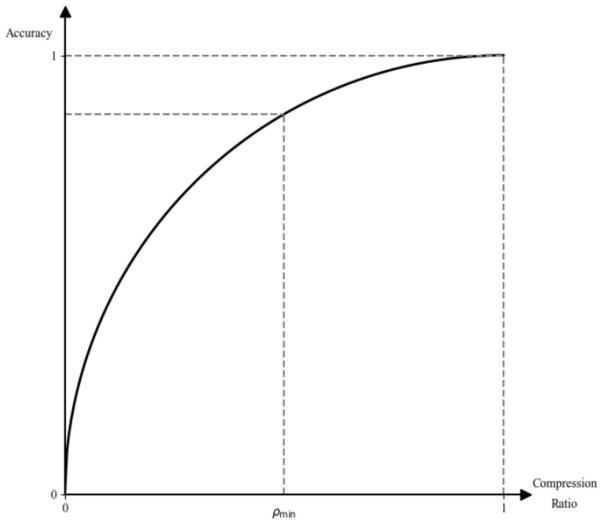

Fig. 9. The relation of task accuracy and semantic compression ratio.

and d_0 is the initial extraction of semantic information for users without semantic compression. $b_n = \frac{p_n}{N_0 B_n}$, p_n is the transmission power of the user n, and N_0 is the spectral density of the noise power. δ^2 is the variance of channel gain. The Q function represents the tail distribution function of a standard normal distribution. So we can obtain the n-th user's success probability of tasks:

$$\Omega_n = \eta(o_n) \times P(t_n \leq t_0), \tag{47}$$

where $\eta(o_n)$ is the probability of which task is successfully executed under successful transmission. It can be seen from (47) that the task success probability proposed by [106] for evaluating SemCom performance can control the tradeoff between semantic transmission and semantic understanding.

13) Transmission Efficiency of Tasks: In [112], the authors modeled the physical channel as a non-trainable fully connected layer to simulate different channel states. With the help of the curve fitting method, the mathematical relationship between compression ratio and task performance under different channel states is explored. Then a new measurement standard is established in [112]: transmission efficiency of tasks.

The transmission efficiency of the task is defined as the weighted sum of the number of packets from each user and the corresponding achievable task accuracy at the receiver. Specifically, the semantic task transmission efficiency v_t in time slot t is defined as follows:

$$v_t = \sum_{j=1}^{J} \sum_{n=1}^{N_j} v_t^{n,j} \times A_t^{n,j}. \tag{48}$$

The subscript n denotes user n and j denotes the intelligent task j corresponding to user n. $A_t^{n,j}$ is the classification accuracy and $v_t^{n,j}$ is the number of data packets that each user can transmit in slot t.

14) Semantic Utility: The reference [107] proposed a semantic utility measurement method that considers semantic timeliness and semantic fidelity.

- Semantic Fidelity: Defined as the fidelity between the original vectorized data $\mathbf{X}$ and the received information $\hat{\mathbf{X}}$. It is expressed as:

$$\mathcal{SF}_{\varepsilon,n}\left(\mathbf{X}, \hat{\mathbf{X}}\right) = f_{sa}\left(\mathbf{X}, \hat{\mathbf{X}}\right), \tag{49}$$

  where the subscript n represents the vehicle n, ε represents the index of the edge server, and $f_{sa}(\cdot)$ is the fidelity mapping function, which varies with the task.
- Semantic Timeliness: Semantics will evolve over time. By modeling and tracking temporal changes, including aggregating new semantic information as much as possible, communication efficiency can be significantly improved, and the probability of errors in semantic transmission can be reduced. The timeliness of the semantic information extracted by the system is defined as:

$$\mathcal{ST}_{\varepsilon,n}(\cdot) = f_{st,\varsigma}\left(\frac{T_{th} - T}{T_{th}}\right), \tag{50}$$

  where $f_{st,\varsigma}(\cdot)$ is a non-linear decreasing function with parameter ς on semantic timeliness. T is the total delay of the system, and T_{th} is the delay constraint. The lower the total delay, the greater the semantic timeliness.

The following formula defines semantic utility:

$$\mathcal{Q}_n^{all} = \zeta_n \mathcal{SF}_{\varepsilon,n} + \chi_n \mathcal{ST}_{\varepsilon,n}. \tag{51}$$

Among them, ζ_n and χ_n are the preferences of semantic fidelity and semantic timeliness, respectively.

15) SemCom QoS: Semantic similarity is further promoted by [53], and SemCom QoS (SC-QoS) based on Semantic Quantization Efficiency (SQE) is created as follows:

- SQE: In order to solve the tradeoff between semantic accuracy and the number of bits consumed, a new metric, SQE, is proposed. This metric quantifies the ratio of the semantic similarity gain of each semantic feature to the bit-related semantic similarity gain. Due to its strong correlation with the novel semantic bit quantization (SBQ) proposed in their work, these contents are not introduced. See [53] for more details.
- SC-QoS: Defined based on SQE and transmission delay, and the effective SC-QoS is expressed as:

$$\Psi = \sum_{n=1}^{N} \left(\widehat{\varpi}_n^{\sharp} - \phi_{\mathcal{G}} \widehat{\mathcal{G}}_n\right), \tag{52}$$

  where the user's index is n, $\widehat{\varpi}_n^{\sharp}$ is the effective SQE (the sum of SQE whose semantic similarity satisfies the minimum threshold), $\widehat{\mathcal{G}}_n$ is the delay, and $\phi_{\mathcal{G}}$ is the balance coefficient.

16) Semantic Score: To measure the overall semantic loss between the original sentence s and the reconstructed sentence $\hat{s}$ at the receiver, the work in [151] defines a new metric named *Semantic Score* (SS). which combines the best of two different quantities, BLEU score and sentence similarity which uses BERT. The BLEU score cannot handle word synonyms, but it is a fast and low-cost algorithm that is language independent and corresponds to human judgment. The sentence similarity score using BERT vectors is slow and has ratings comparable to the BLEU, but it also handles synonyms. Let $\Delta_\lambda(s, \hat{s})$

denote the SS between sentence s and $\hat{s}$, which is a convex combination of BLEU$(s, \hat{s})$ and $\xi(s, \hat{s})$.

$$\Delta_\lambda(s; \hat{s}) = (1 - \lambda)\text{BLEU}(s, \hat{s}) + \lambda\xi(s, \hat{s}), \tag{53}$$

where $\lambda \in [0, 1]$ is a parameter.

In this section, we explore the construction of the objective function in resource allocation of SemCom, which is a key to the modeling of optimization problems. We provide a detailed review of performance metrics, categorizing them into two types. The first type includes traditional metrics such as delay and energy consumption, while the second type focuses on new metrics based on semantic similarity. We give two comprehensive comparative matrices to better synthesize findings across references. To further clarify the influence of different resource types on these performance metrics, we provide a resource–metric mapping summary in Table VII, in which we use some clear examples in different studies to illustrate this influence.

IV. Centralized Resource Allocation Algorithms

In order to realize resource allocation in SemCom and meet the requirements of these performance metrics proposed above, advanced resource allocation strategies and algorithms are essential. The optimization problem constructed is extremely complex and differs significantly from the traditional communication architecture in terms of objectives, constraints, and optimization variables. It is a challenge to construct a well-performing optimization algorithm that can adapt well to SemCom. Currently, there are a variety of centralized algorithms for resource allocation in SemCom, mainly consisting of convex optimization, heuristic algorithms, and DRL. Fig. 10 shows the taxonomy of centralized resource allocation algorithms in SemCom.

In recent years, many researchers have summarized the state-of-the-art resource allocation algorithms of various scenarios in their surveys. In [16], the authors summarized different optimization methods for resource allocation in edge computing. The comparison tables of different papers are designed according to the objective, brief description of the methods, advantages, and disadvantages. Reference [152] summarized different resource allocation schemes for the two dominant vehicular network technologies, e.g., Dedicated Short Range Communications (DSRC) and cellular-based vehicular networks. In this subsection, centralized resource allocation optimization algorithms from different literature in SemCom are reviewed.

A. Algorithms Based on Convex Optimization and Mathematical Techniques

Because resource allocation involves a lot of variables and constraints, the corresponding optimization problems are usually complex, even non-convex or NP-hard. A considerable part of the research transforms the non-convex problems into near-convex or convex optimization problems, which leads to feasible convex optimization methods. The main techniques include Lyapunov optimization techniques, alternating optimization (AO) algorithms, successive convex approximate (SCA) methods, and the interior point method, as well as some other mathematical algorithms based on other mathematical algorithms, such as the Hungarian algorithm [153]. An optimization algorithm based on convex optimization typically combines several of these techniques.

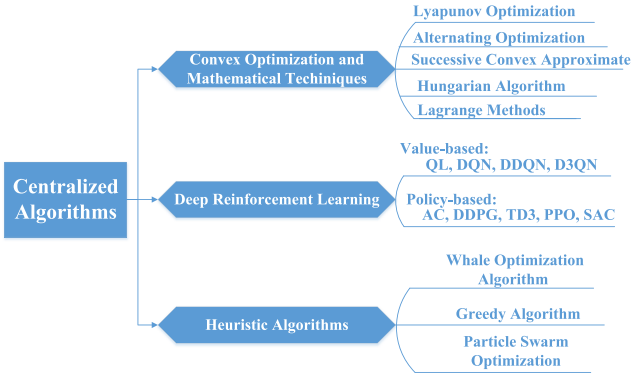

Fig. 10. The taxonomy of centralized resource allocation algorithms in SemCom.

1) Lyapunov Optimization: Lyapunov optimization is a powerful long-term resource optimization scheme to find stability or equilibrium points of dynamical systems with stochastic properties of nonlinear systems. It requires less prior knowledge and has low computational complexity [154]. Lyapunov optimization focuses on analyzing and optimizing stochastic networks (networks characterized by random events, time-varying dynamics, and uncertainties). It is particularly well-suited for applications in communication systems and queueing systems. The authors of [102] adopted the Lyapunov optimization method to solve the problem, which first transforms the long-term constraints into queue stability conditions using the concept of virtual queue and then transforms the long-term objective function and the queue stability conditions into solvable short-term subproblems. Similarly, [59] and [118] also used Lyapunov optimization techniques to transform the original stochastic optimization problem of multiple time slots into a series of deterministic problems in a single time slot. Lyapunov optimization, as a stochastic optimization method, enables online decision making while maintaining sub-optimal performance. Therefore, it applies well in a long-term stochastic scenario in SemCom system, like the semantic-aware dynamic long-term MEC systems using time division duplexing (TDD) in [59]. Lyapunov optimization can also combine with DRL-based method, in [72], expanding on the Lyapunov transformation, the UoI minimization problem is converted into a sequence of deterministic single time-slot optimization problems. Subsequently, the DRL-based method PPO (will be introduced later in Section IV-B) is used to tackle this problem.

2) Alternating Optimization Algorithm: The alternating optimization (AO) algorithm is to decompose the optimization problem into several sub-problems, and then these sub-problems are solved iteratively. Commonly used in the case of multi-variable optimization, which iteratively optimizes each variable while treating other variables as a fixed value. Depending on the specific problem, the complexity of the

TABLE VIII
COMPARISON BETWEEN DIFFERENT NEW SEMANTIC METRICS

Metric	Modality	Suited Applications and Scenarios	Strengths	Limitations
S-SE	Text	Long-term text transmission	Measures the communication efficiency from the semantic perspective.	Relies on the BERT model, computationally expensive, limited generalization and real-time applicability.
ES-SE		Text task-oriented SemCom	Ensures that only transmissions meeting semantic similarity thresholds contribute to efficiency, aligning better with task requirements.	Introduces a binary constraint that may ignore near-threshold transmissions, potentially leading to under-utilization of spectral resources.
S-EE		Energy-constrained text semantic transmission	Evaluate the performance of semantic transmission from a energy perspective, suitable for IoT-devices, industrial wireless communications and battery-powered devices like UAVs.	Relies on the BERT model, computationally expensive, limited generalization, and accurate circuit-level energy consumption may be hard to obtain.
SemCom QoS		Dynamic task-oriented SemCom	Provides a quantitative balance between semantic representation and resource allocation through SQE, granular control over quality and delay.	Relies on semantic-bit quantization (SBQ), limiting generalization and increasing implementation complexity.
Semantic Score		Text transmission	Combines the efficiency of BLEU with the semantic awareness of BERT-based similarity, offering a more comprehensive and flexible assessment.	Sensitive to the choice of the weight parameter λ; BERT-based similarity is computationally expensive and may not generalize well.
TOSSE	Image	Task-related feature importance-aware applications	Captures the transmission performance of task-related semantic features by integrating user-specific feature selection, enabling fine-grained optimization for downstream tasks.	Highly depends on the accuracy and generalization of the feature selection module, which is difficult to design and validate. May be sensitive to dynamic variations in user data distributions
Transmission Efficiency of Tasks		Task-oriented SemCom of intelligent devices	Considering both the number of packets transmitted and the classification accuracy, which accounts for the tradeoff between communication and task performance.	Can only be adopted to communications between intelligent devices, limited generalizability.
QoE of MSPs		Interest-aware Metaverse semantic transmission	Jointly considers interest level, transmission delay, and BER for a more realistic user experience assessment	Relies on the interest rating predicted by the central server, accurate estimation of user interest requires prior behavioral data, which may not always be available.
EoSI	Multi-modal /Generic	SemCom-enabled IIoT systems	Balances semantic task accuracy, timeliness, and resource efficiency for evaluating task-oriented semantic transmission.	The metric structure is application-specific and complex, requiring careful modeling and weight tuning for different resource types and task goals.
Success Probability of Tasks		Task-oriented SemCom; Delay sensitive SemCom applications such as IoV	Captures the joint impact of transmission delay and semantic compression on task execution success, enabling quantitative tradeoff analysis between efficiency and accuracy.	The probability of which task is successfully executed under successful transmission $\eta(o_n)$ is often application-dependent and hard to obtain.
STM		Most scenarios	Characterizes the number of messages successfully transmitted in the system per unit time, which can well characterize network performance from a semantic perspective; flexible for different scenarios.	The B2M conversion function is related to different semantic encoders, knowledge matching, and message properties, though flexible, it is difficult to obtain and apply uniformly across different applications.
Semantic Entropy		Task-oriented SemCom	Captures the minimal task-relevant semantic representation, aligning communication with downstream goals.	Intractable to compute exactly; approximation depends heavily on the quality and architecture of the DL-based encoder.
Semantic QoE		User-centric multi-cell and multi-modal SemCom	Balances semantic rate and accuracy based on individual user preferences, enabling personalized QoE optimization.	Relies on accurate estimation of user-side semantic preferences and DL-based semantic entropy, which are hard to quantify or obtain in real-world systems.
Semantic Utility		Dynamical environments consider both accuracy and timeliness	Balances semantic fidelity and timeliness, enabling flexible adaptation to task-specific preferences.	Requires task-specific fidelity and timeliness functions, which may lack universality and introduce modeling complexity.
AoSI		Time-sensitive applications	Considering both the freshness of data and its semantic accuracy. This is especially useful for applications/scenarios where both the timeliness and semantic accuracy are critical like video streaming and emergency rescue communication.	Depends on the computation of semantic similarity (e.g., using BERT for text), which means the metric may be less meaningful in scenarios where semantic similarity is difficult to define or quantify.
UoI		Applications with multiple constraints (timeliness/energy/accuracy)	Provides a very comprehensive assessment of information value by considering various factors such as delay, error, and resource constraints, allowing for better decision-making in multiple constraints communication systems.	The complexity of modeling and calculating UoI increases with the number of attributes considered. And the sub-metrics used to evaluate each part need to be deeply considered to avoid the conflicts.

decomposed problem varies; the simpler case is decomposed into two to three subproblems, where each subproblem optimizes a single variable in [44], [45], [57], [91], [98], [119], [130], [132]. As the problem and the optimization variables increase, the optimization problem is decomposed into three subproblems in which the subproblem has two or more optimization variables in [59], [66], [67], [74], [93], [100], [105]. A more complicated situation occurs in [68], where the paper employs a nested AO algorithm to divide the optimization problem into two subproblems: the semantic extraction strategy subproblem and the wireless resource allocation subproblem, which will be optimized alternately and iteratively, where each of the two subproblems also employs the AO algorithm to optimize the corresponding parameters. An iteration of the algorithm for the total optimization problem contains the number of iterations L_1 and L_2 of the AO algorithm for the two sub-problems.

3) Successive Convex Approximate: The idea behind successive convex approximate (SCA) is to find a locally optimal solution to the original problem by iteratively solving a series

Algorithm 1 Basic SCA Algorithm for Problem $\mathcal{P}$

Find a feasible solution $\mathbf{x} \in \mathcal{X}$ in $\mathcal{P}$, choose a step size $\theta \in (0, 1]$ and set $k = 0$.

Repeat

1) Compute $\hat{\mathbf{x}}(\mathbf{x}^k)$, the solution of $\mathcal{P}_{\mathbf{x}^k}$;
2) Set $\mathbf{x}^{k+1} = \mathbf{x}^k + \theta(\hat{\mathbf{x}}(\mathbf{x}^k) -)\mathbf{x}^k$;
3) Set $k \leftarrow k + 1$

Until convergence criterion is met.

of convex optimization problems similar to the original non-convex problem. Consider the following optimization:

$$\mathcal{P}: \min_{\mathbf{x}} \quad U(\mathbf{x}) \tag{54}$$

$$\text{s.t.} \quad g_l(\mathbf{x}) \leq 0, \quad \forall l = 1, \ldots, m \tag{54a}$$

$$\mathbf{x} \in \mathcal{K} \tag{54b}$$

where the objective function and constraint (54a) is smooth (possibly nonconvex), the feasible set is denoted as $\mathcal{X}$. The original non-convex or non-concave function is transformed into a series of convex or concave functions. The convex approximation of the original problem can be stated as follows: given $\mathbf{x}^k \in \mathcal{X}$:

$$\mathcal{P}_{\mathbf{x}^k}: \min_{\mathbf{x}} \quad \tilde{U}\left(\mathbf{x}; \mathbf{x}^k\right) \tag{55}$$

$$\text{s.t.} \quad \tilde{g}_l\left(\mathbf{x}; \mathbf{x}^k\right) \leq 0, \quad \forall l = 1, \ldots, m \tag{55a}$$

$$\mathbf{x} \in \mathcal{K} \tag{55b}$$

where $\tilde{U}(\mathbf{x}; \mathbf{x}^k)$ and $\tilde{g}_l(\mathbf{x}; \mathbf{x}^k)$ represent the approximations of $U(\mathbf{x})$ and $g_l(\mathbf{x})$ at current iteration $\mathbf{x}^k$, respectively, the feasible set is denoted as $\mathcal{X}(\mathbf{x}^k)$. We can summarize the basic SCA algorithm in Algorithm 1.

This process is repeated until the stopping criterion is satisfied. It is assumed that at each iteration, some original functions are approximated by their upper bounds, where the same first-order behavior is preserved [155].

Since an approximate solution to the original optimization problem is solved in each iteration, there is no guarantee that the global optimum will be obtained. The convergence of the method is guaranteed due to convexity/concavity [19].

AO algorithms and the SCA algorithm are two methods that work well with each other, and almost all the literature on SCA uses a combination of the two. Decomposing a large non-convex optimization problem into several small non-convex optimization subproblems to solve iteratively reduces the difficulty/complexity of the SCA algorithm, thus allowing the difficulty and complexity of the overall problem to be reduced [44], [45], [49], [59], [67], [68], [70], [91], [92], [106], [111], [119].

4) Hungarian Algorithm: The solution to the maximum matching problem in bipartite graphs is the origin of the Hungarian algorithm. Since a maximum matching of a bipartite graph necessarily exists, e.g., the upper bound is a perfect matching that contains all vertices, it is possible to get a maximum matching of a bipartite graph based on any matching if we have a way to keep searching for augmenting paths until eventually we find no new augmenting paths. The core idea of the Hungarian algorithm is to iteratively search for augmenting paths to get a maximum match.

The Hungarian algorithm can solve the allocation problem in polynomial time, which can significantly reduce the algorithmic complexity. When it comes to the scenario of the resource allocation problem in SemCom, it is usually used for the subproblem of subcarrier pairing/subchannel allocation after the original optimization problem is decomposed by the AO algorithm above. In the literature [36], [109], and [57], the optimization subproblem of channel allocation is regarded as a bipartite graph matching problem, and then the Hungarian algorithm is used to solve this optimization subproblem. Among them, the knowledge-assisted proximal policy optimization (K-PPO) algorithm is proposed in [109], which uses the Hungarian method to determine channel allocation, greatly reducing the complexity of the original proximal policy optimization (PPO) algorithm by introducing the Hungarian algorithm. The details of PPO will be introduced later in Section IV-B.

5) Lagrange Methods: The Lagrange multiplier method is a common method for solving constrained optimization problems. For the optimization problem with only equation constraints, you can directly use the Lagrange multiplier method to list the Lagrange function, which will be transformed into an unconstrained optimization problem to solve. For the optimization problem with inequality constraints, using the Lagrange function to optimize it must satisfy the Karush-Kuhn-Tucker (KKT) condition, which is a necessary condition for taking the optimal parameter values and a sufficient condition for some special convex optimization problems. Problems containing inequality constraints after listing the Lagrangian function still have constraints that are not easy to deal with, then it can be transformed into a Lagrangian dual problem; this dual problem must be a convex optimization problem and therefore easy to solve. But in order to make the dual problem and the original problem have the same solution, it must satisfy the strong duality. The sufficient condition is Slater's condition; the necessary condition is the KKT condition. Lagrangian methods have been employed in many works, where the problem is decomposed into subproblems and then the sub-optimization problem is solved using Lagrangian methods [57], [68], [102], or the problem is transformed directly using the Lagrangian methods to solve [78].

Summary: Traditional optimization algorithms based on convex optimization techniques and other mathematical algorithms are applicable to small-scale solutions and high-reliability demand scenarios. They have the following advantages: a) mature and widely used; b) easy to obtain sub-optimal optimization results; c) not relying on data. However, algorithms based on these techniques are often too complex. As a result, its complexity makes it difficult to implement in practical systems and not suitable for large-scale problems.

Although algorithm complexity may vary due to different problems and scenarios, we can still give a brief summary

of these algorithms. In terms of computational complexity, Lyapunov optimization itself typically has the lowest complexity due to its online and dynamic nature, making it suitable for real-time systems. The complexity of Lyapunov optimization is primarily determined by the per-slot deterministic subproblem, and often falls in the range of $O(n^2)$ to $O(n^3)$ when convex formulations are involved, making it particularly suitable for low-latency real-time systems. Alternating optimization (AO) and Lagrangian methods exhibit moderate complexity, with AO being effective for decomposable nonconvex problems and Lagrangian methods for constrained optimization. The complexity of AO is mainly determined by the complexity of solving each subproblem. For instance, if each subproblem involves convex optimization with complexity $O(n^3)$, and k such subproblems are solved per iteration, the total complexity per iteration becomes approximately $O(k\cdot n^3)$. For Lagrangian-based methods, the overall complexity depends on both the structure of the primal problem and the method used for updating dual variables. If the primal problem admits a closed-form solution, each iteration may involve only dual updates with complexity around $O(n^2)$, leading to a total complexity of $O(K\cdot n^2)$, where K is the number of iterations. However, if the primal problem requires solving a numerical optimization (e.g., quadratic programming), the per-iteration cost may increase to $O(n^3)$, resulting in a total complexity of $O(K\cdot n^3)$. Successive convex approximation (SCA) tends to have higher complexity due to iterative convex approximations. Each subproblem often requires $O(n^3)$ time, and the total complexity $O(T\cdot n^3)$ grows linearly with the number of iterations T. Thus, SCA is suitable for nonconvex problems with a manageable size and structure. The Hungarian algorithm, with a complexity of $O(n^3)$, is efficient for small-scale linear assignment problems but less scalable for larger systems. Table IX reviews the literature using these traditional optimization techniques, which are based on convex optimization techniques and other mathematical algorithms.

B. Algorithms Based on Deep Reinforcement Learning

In the context of SemCom, direct modeling of the relationship between semantic accuracy (or fidelity) and optimization variables, such as the semantic compression ratio, is often infeasible due to the absence of explicit analytical expressions. Then results in the non-differentiable and implicit objectives. To address this challenge, some authors [112], [121] use different curve fitting techniques to approximate this implicit relationship. For instance, in [121], neural networks are adopted to fit the relationship curve of semantic fidelity and optimization variables (power, channel assignment, semantic compression) for each task (single modal and bi-modal). Once this approximation is obtained, the originally implicit objective becomes differentiable or at least numerically tractable. However, after curve fitting, the output fitting function is still complex and non-convex. Traditional mathematical methods are often difficult to model or calculate in the face of these complexities.

With the development of DL and reinforcement learning (RL) techniques, pure data-driven DRL has become a powerful tool to solve complex resource management problems in recent years [81], [156], [157]. By efficiently learning the dynamics of the environment, DRL can provide resource allocation strategies that maximize long-term returns based on pretrained policy networks.

RL and DRL approaches can be mainly distributed in two ways: based on value functions (1-4) and based on policy gradients (5-9). This paper also provides a brief description of the algorithms based on these techniques in various publications.

1) Q-Learning: Q-Learning (QL) [158] is an off-policy control method for finding the optimal policy, mainly used in discrete action space. The core idea is to utilize a Q function that represents the expected reward of taking an action in a particular state. The Q function updating rule satisfies the Bellman equation:

$$Q(s,a)\leftarrow Q(s,a)+\alpha\left[r+\gamma\max_{a'}Q\left(s',a'\right)-Q(s,a)\right]. \tag{56}$$

In [71], the selection of public messages uses QL techniques. The work of [124] compares the QL-based approach with the convex optimization-based approach under the video semantics-driven resource allocation problem. The experimental results prove that the QL-based approach performs better than the convex optimization-based approach.

2) Deep Q Network: Mnih et al. [159] introduced the deep Q network (DQN), which pioneered the field of DRL. In real-world scenarios, the number of states can be large, making the construction of Q-tables computationally intractable. To address this limitation, DQN uses a neural network to estimate the Q-values of each state-action pair. The most important feature of DQN is that it uses experience replay [160] and target networks to stabilize the training of deep neural networks [161]. As mentioned in the previous paper, [89] defined the AoSI metric. In the paper, the long-term average AoSI optimization problem is modeled as an MDP, and a DQN-based algorithm is proposed to find the suboptimal solution for source scheduling and the number of semantic symbols. Compared with the simpler state space case modeled in the literature [71] using QL, most of the resource allocation problems in SemCom have a more complex state space, so DQN is obviously more suitable. In [22], the exhaustive search to solve the semantic compression subproblem will lead to high computational complexity. Because when using exhaustive search to solve this combination optimization problem (for K-users, there are $K!$ permutations), the complexity will grow exponentially with the number of users and cells. Therefore, in the journal version of [22], that is, the reference [121], the authors proposed a solution that combines DQN and matching theory. The exhaustive search is replaced by the DQN-based method to improve overall QoE effectively.

3) Double Deep Q Network: Hasselt et al. [162] proposed the double deep Q network (DDQN) to solve the overestimation problem in QL. The DDQN algorithm borrows from the double-Q learning algorithm [163] and makes

TABLE IX
CENTRALIZED ALGORITHM BASED ON CONVEX OPTIMIZATION METHOD AND MATHEMATICAL TECHNIQUES

Ref.	Optimization objective	Resource to be allocated			Optimization algorithm
		Communication	Computing	Network parameter	
Single-resource Allocation					
[129]	Total semantic rate	-	-	SCM selection	Dynamic programming algorithm
[130]	Semantic spectrum efficiency	Sensing bandwidth, transmission bandwidth	-	-	AO algorithm, Gradient descent method, Projected gradient method
[102]	Long-term satisfaction	Power, application layer access rate			Lyapunov optimization, Bernstein approximation, Lagrange method
Joint Two Resources Allocation					
[45]	Effective bit rate	Bandwidth	-	SemRelay placement	BCD algorithm, SCA
[91], [92]	Secure semantic efficiency	Transmit beamforming		Semantic parameter	AO algorithm, Dinkelbach algorithm, SCA
[57]	Average data reconstruction error	Channel allocation, power, uRLLC scheduling		Network parameters	BCD, Hungarian algorithm, Lagrange method
[46]	Combination of latency and utility function	Power, bandwidth		SI Selection (Seletion of Semantic Information)	SCA, Fractional programming
[87]	Energy consumption				AO algorithm, Graph theory
[84]	Energy consumption	Power			Exhaustive search algorithm, Gradient descent algorithm
[74]	Overall time cost	Bandwidth, power, time cost			BCD, Exhaustive searching Algorithm
[52]	Semantic efficiency	User selection, bandwidth, power			AO algorithm, Linear programming
[36]	S-SE	User association			AO algorithm, Exhaustive search algorithm, Hungarian algorithm
[68]	Weighted sum semantic information transmission rate	Common rate, time delay, transmit beamforming			Lagrange dual method, Fractional programming, BCD, Interior point method, SCA, Low-complexity initial point search
[70]	Total semantic rate	Common rate, transmit beamforming		SC Ratio (Semantic Compression Ratio)	AO algorithm, SCA, Greedy algorithm
[64]	Sum of equivalent rate	Transmit power, receive beamforming			MMSE strategy, AO algorithm, Gradient ascent algorithm
[66]	Semantic-aware sum rate	RIS-user association, transmit power, beamforming, phase shift			AO algorithm, Many-to-many matching, Tensor beamforming, Greedy algorithm
[53]	Probability of task success	Bandwidth, power			Particle swarm optimization algorithm
[119]	Transmission latency and semantic similarity	User association, transceiver beamformer			AO algorithm, SCA, Hungarian algorithm, Exhaustive search
[111]	Task success probability	Bandwidth, power, user selection			AO algorithm, SCA, Branch and bound method
[132]	Sum semantic-aware transmission rate	RIS-user association, transmit power, beamforming vector, phase shift			AO algorithm, many-to-many matching
[78]	STM	User association, bandwidth		Communication mode selection	Lagrange dual method, preference list-based heuristic algorithm, AO algorithm
[73]	Max transmission delay	Power			AO algorithm, Heuristic algorithm
[118]	System utility	System access, user association, bandwidth	CC (Computing Capacity)	-	Lyapunov optimization, AO algorithm, Interior point method, Greedy algorithm
[59]	Long-term total energy consumption	Transmit power, slot division			Lyapunov optimization, BCD algorithm, SCA
[62]	Total energy consumption	Time duration allocation, power			Lagrange method
[49]	Utility function	-	Offloading strategy, CC of MEC	SC Ratio	AO algorithm, SCA
Joint Communication-Computation-Network Parameters Resources Allocation					
[88]	S-EE	Power, bandwidth	CC	Semantic symbol allocation	AO algorithm, Heuristic algorithm: WOARA
[67]	Energy consumption	Common message rate, power, private message transform beamforming	CC	SI Selection	AO algorithm, SCA
[60]	Time delay	Transmit power	CC	SC Ratio	Geometric programming algorithm, Interior point method, AO algorithm
[48]	Utility function	User association, resource blocks, transmit power	CC, CC for compression	Transmit data volume	AO algorithm
[105]	Training delay and energy consumption	Bandwidth, transmission power	CC of (BS and Users)	SC Ratio	AO algorithm, Lagrange method, Modified Newton method

improvements to the DQN algorithm: estimating the policy based on the online Q-network, selecting the action, and estimating the Q-value with the target network. Some experimental results show that DDQN finds a better strategy than DQN in Atari games. The authors of [82] went one step further than the QL-based work [124] that was already mentioned. They wanted to improve the accuracy of video semantic understanding and build a multidimensional resource allocation model that combined communication, computation, and caching. They designed

the DDQN-based algorithm, which is shown to achieve better results than those achieved by the QL-based approach in [124].

4) Dueling Double Deep Q Network: Dueling double deep Q network (D3QN) is a combination of Dueling DQN [164] and DDQN. Dueling DQN separates the computation of Q-values into two components: the value function (V) and the advantage function (A), which enables Dueling DQN to provide more accurate Q-value estimation while needing less discrete action data, thus improving sample efficiency. As in Table X, the authors of [90] and [94] both use D3QN for discrete action in the whole DRL framework.

5) Actor-Critic: The actor-critic (AC) [165] algorithm learns both the policy and the state-value function, using the value function to reduce variance in policy updates. Actor-critic methods tend to be more stable than pure policy gradient methods. In the work of [58], the allocation of transmitted semantic information and resource block (RB) was jointly optimized to minimize the average transmission delay, based on the improved AC algorithm, in which a novel value function is designed to improve the probability of action exploration and finding the optimal solution. In traditional model-free DRL, the value function $V(\boldsymbol{s}_{k+1})$ is approximated by DNN: $\mathbb{E}_{\boldsymbol{s}_{k+1}\sim P}[V(s_{k+1})]$. In the model-based DRL proposed in the article, due to the deterministic nature of the state transitions, there is $\mathbb{E}_{\boldsymbol{s}_{k+1}\sim P}[V(s_{k+1})] = V(s_{k+1})$. Therefore, the proposed algorithm does not need to use DNN to approximate the value function. As a result, the estimation error resulting from the approximation of the value function can be prevented, and the state-action value function can be computed accurately. More details about the algorithm can be found in [58].

6) Deep Deterministic Policy Gradient: Silver et al. [166] proposed the deterministic policy gradient (DPG) algorithm for RL problems with continuous action spaces. The deterministic policy gradient is the expected gradient of the action-valued function, which integrates over the state space and can be estimated more efficiently than the stochastic policy gradient. Lilicrap et al. [167] proposed the deep deterministic policy gradient (DDPG) algorithm in the continuous action space by extending DQN and DPG. DQN can only handle discrete and low-dimensional action spaces, but many cases, especially physical control tasks, have continuous and high-dimensional action spaces, and DQN cannot be directly applied to continuous domains, so DDPG adopts the AC method based on the DPG algorithm.

The authors of [112] developed a joint optimization problem of semantic feature compression rate, transmit power, and bandwidth for each smart device to maximize the long-term transmission efficiency of the task. A DDPG-based wireless resource allocation scheme is proposed to efficiently handle the continuous action space.

7) Twin Delayed Deep Deterministic Policy Gradient: Twin delayed deep deterministic policy gradient (TD3) is proposed by Fujimoto et al. [168] based on the improvement of the DDPG algorithm. The TD3 algorithm incorporates the idea of the double Q-learning algorithm into the DDPG algorithm. A detailed description can refer to [161]. From AC and DDPG to TD3, with the evolution of these RL algorithms, there have also been attempts in the literature to use TD3 instead of DDPG as the base algorithm of the scheme [107], [108], in which [107] proposed a TD3-driven dynamic semantic-aware algorithm: dynamic semantic-aware TD3 (DSATD3) for a federated learning-driven semantic vehicular network to guide agents in adopting accurate semantic extraction and resource allocation strategies. The simulation results showed that DSATD3 has better performance compared to DDPG-based approaches.

In contrast to the previous two papers, the work in [75] improves the TD3 algorithm and proposes the TD3-RNS algorithm (TD3 with reference neuron-enhanced Softmax) to solve a long-term semantic throughput maximization problem. The actor network uses a reference neuron technique and a linearly decreasing Gaussian action noise in the output layer to enhance training efficiency and balance exploration and utilization by the agent.

8) Proximal Policy Optimization: Proximal Policy Optimization (PPO) is proposed by Schulman et al. [169] in 2017. PPO aims to improve and simplify previous policy gradient algorithms, such as Trust Region Policy Optimization (TRPO). The key aspect of the PPO algorithm is that it makes the learning process more stable by limiting the magnitude of policy updates. The authors of [71] designed a power allocation algorithm to maximize the total QoE based on PPO. The algorithm can appropriately allocate the power of public and private messages to maximize the total QoE while guaranteeing individual QoE for each MSP. According to [95], the authors proposed a semantic-aware resource allocation framework with a flexible duty cycle co-existence mechanism (SARADC) algorithm that utilizes PPO to optimize resource allocation in high-speed vehicular networks.

We mentioned in Section II-B3 that in task-oriented SemCom systems, the resource allocation is closely tied to the task-related importance of the semantic information. This task dependence necessitates adaptive resource allocation schemes that align with the utility of semantic content, ensuring the transmission of task-related and semantically important features, while jointly optimizing bandwidth, power, and computing resources for overall system performance. However, traditional PPO methods struggle to handle such cross-layer optimization under semantic-aware constraints. To this end, [109] and [40] made novel improvements to the PPO algorithm. Reference [109] proposes a knowledge-assisted PPO (K-PPO) algorithm, which utilizes a prior model and the Hungarian algorithm to assist PPO in solving the joint optimization problem of importance-aware semantic feature selection and channel assignment within the joint semantic-channel transmission (JSCT) mechanism. Meanwhile, [40] develops an attention-enhanced PPO (APPO) by introducing the attention network [27], enabling the base station to learn the correlation between the semantic importance distribution $\boldsymbol{f}_i(\mathcal{G}_i)$ and the task performance metric MSS, thus optimizing the resource block (RB) allocation and semantic information selection strategies accordingly.

9) Soft Actor-Critic: The soft actor-critic (SAC) algorithm [170] is a model-free DRL algorithm based on maximum entropy, introducing the concept of maximum entropy on

TABLE X
DIFFERENCE IN COMBINATION OF DRL ALGORITHMS

Refs.	Discrete actions	Continuous actions
[61]	**DQN:** Semantic compression ratio	**DDPG:** Time slot division coefficients and transmit power
[55]	**DDQN:** Semantic symbol numbers and subchannels allocation	**DDPG:** UAV trajectory and transmit beamforming
[71]	**DDQN:** Subchannels allocation	**DDPG:** Power allocation
[90]	**D3QN:** DeepSC selection and subchannel assignment	**DDPG:** Transmit beamforming and IRS reflection array
[90]	**D3QN:** DeepSC selection and subchannel assignment	**SAC:** Transmit beamforming and IRS reflection array
[94]	**D3QN:** Semantic symbol and subchannel allocation	**TD3:** Transmit beamforming and IRS reflective elements
[125]	**DDQN:** Variation of travel speed, time interval and transmitting carrier frequency	**A3C:** Bandwidth allocation

top of maximizing future cumulative rewards to enhance the robustness and exploration ability of agents. In reference [53], a dynamic intelligent resource allocation scheme was designed. It is based on SAC and D-SAC to realize real-time decision-making based on perceptual semantic tasks and channel features. Among them, D-SAC is to extend SAC to discrete space to solve the discrete variable allocation problem. The Four-Soft Actor Critical (4-SAC) algorithm is proposed in [83]. It comprises four SAC intelligent agents, which collectively optimize the trajectory of a UAV, number of semantic symbols, and power allocation to strike a balance between data transmission efficiency and energy efficiency, and QL was used to facilitate learning for the optimal policy.

In fact, the challenge of highly coupled and non-convex optimization variables is particularly critical in SemCom. Unlike conventional systems, where optimization variables can often be easily decoupled or approximated linearly. In the resource allocation problem of SemCom, the composition of optimization variables is very complex and hard to decouple, which may be both in the case of discrete action space: semantic symbol number selection, subchannel allocation, communication mode selection, and some discretized variables, etc., and in the case of continuous action space: power allocation, bandwidth allocation, semantic compression rate, etc. To tackle this complexity, recent studies chose to combine two or more of these methods to solve the problem [47], [55], [61], [71], [90], [94], combining the value function-based method and the policy gradient-based methods to form a two-layer DRL framework, which is also succinctly summarized in Table X.

Summary: Centralized optimization algorithms based on DRL are applicable to highly dynamic scenarios. They have the following advantages: a) They can deal with high-dimensional, nonlinear state and action spaces, making them suitable for complex decision problems. b) It can adaptively learn the optimal policy without excessive mathematical derivation and computation. However, it also has the following disadvantages: a) High complexity of training. b) Closed-box process: the learning process is unobservable, and the output results are difficult to interpret, which will affect the credibility and acceptability of the results. c) DRL algorithms are sensitive to the selection of hyperparameters and training data, and the instability is higher. Table XI reviews the literature using DRL-based centralized optimization algorithms.

C. Heuristic Algorithms

A heuristic algorithm is an algorithm based on an intuitive or empirical construction that usually performs well with limited computational resources and is suitable for scenarios with low performance requirements to fulfill engineering needs. They can provide effective approximations, but are not guaranteed to find the global-optimal solution.

As we mentioned earlier, semantic similarity does not have a closed-form expression. This can be regarded as a closed-box optimization problem, which is difficult to solve with traditional optimization algorithms. Heuristic algorithms provide a feasible way to solve the closed-box problem. Reference [88] proposed a variant of the Whale Optimization Algorithm (WOA) [171] that introduces a penalty strategy: the Whale Optimization Algorithm with a Penalty Strategy (WOARA) to solve the optimal resource allocation problem. More details about WOA and WOARA can be seen in [88]. The authors of [53] use the particle swarm optimization (PSO) algorithm to optimize the compression ratio and the allocation of power and bandwidth for each user jointly. In [77], the PSO algorithm is developed to determine the computation resource allocation in each step of the matching game. There are also a few other works that incorporate heuristics into the overall program design, such as [78], which also uses a preference list-based heuristic algorithm for problem solving. Furthermore, [66], [70], [118], [131] have incorporated simple heuristics such as greedy algorithms into their overall program design. Table IX also includes papers that use heuristic algorithms.

Summary: Heuristic algorithms have some advantages in terms of cost and convergence speed, but their performance is relatively poor, are prone to fall into local optimal, and are sensitive to parameters. Therefore, they are applicable for scenarios that only have requirements on low latency and do not have high demands on other performance metrics.

In this section, centralized resource allocation optimization algorithms from different literature in SemCom are reviewed. These algorithms are categorized into several types, including those based on mathematical optimization (Lyapunov optimization, AO algorithm, SCA, etc.), DRL (value-based and policy-based), and heuristic methods. While previous sections have systematically categorized performance metrics and optimization strategies, it is also crucial to understand how these elements interact across various network scenarios. While existing works propose various optimization techniques tailored to SemCom scenarios, there remains a lack of in-depth discussion on how these methods specifically address the unique challenges Table XII provides a challenge-centric synthesis of representative works, their applied optimization techniques, and directions for future hybrid or enhanced methods.

TABLE XI
CENTRALIZED OPTIMIZATION ALGORITHMS BASED ON DRL

Ref.	Optimization objective	Resource to be allocated			Optimization algorithm
		Communication	Computing	Network parameter	
Single-Resource Allocation					
[89]	AoSI	-	-	Scheduling decision, semantic symbols	DQN
Joint Two Resources Allocation					
[40]	MSS	RB allocation	-	SI Selection	APPO algorithm
[58]	Average transmit latency				Actor-critic
[107]	Semantic utility	Bandwidth			TD3-based: DSATD3
[90]	ES-SE	Subchannel allocation, transmit beamforming, IRS's reflection array elements		Semantic symbol selection	D3QN+SAC
[94]	S-SE	Subchannel allocation, beamforming of SBS, IRS's reflective elements			D3QN+TD3
[101]	S-SE	Transmit beamforming			DSMRA algorithm
[95]	HSSE	Power, channel allocation, time slot division			PPO
[96]	HSSE and SRS	Power, Channel selection			SAC
[47]	Combination of QoS and transmit cost	Transmit power, bandwidth			DQN+DDPG
[121]	Semantic QoE	Channel assignment, power			DQN, Matching theory, AO algorithm
[55]	S-SE	Subchannel allocation, UAV's transmit beamforming		Semantic symbol allocation, UAV trajectory	DDQN+DDPG
[56]	Semantic throughput	Power, subchannel allocation		Knowledge match coefficient	DDQN+DDPG
[75]	Long-term equivalent semantic throughput	Power, bandwidth		Mode selection, NOMA user pairing	TD3-RNS algorithm
[72]	Long-term average UoI	Power allocation, rate control		Order of public rates	PPO, Lyapunov Optimization
[109]	TOSSE	Channel assignment		Feature transmission ratio	K-PPO algorithm, Hungarian algorithm
[112]	Long-term task transmit efficency	Transmit power, bandwidth		Compression ratio	DDPG
[53]	SC-QoS	Subchannel allocation, bandwidth, power		SBQ allocation	SAC, D-SAC
[113]	MIST	Transmit power		Semantic extraction accuracy object importance score	Diffusion model, DRL
[125]	System's overall spectrum utilization and object detection accuracy	Channel stability intervals, carrier frequency, bandwidth		Driving speeds	DDQN+A3C
[76]	Sum of short-term semantic transmission rate	Resource unit assignment		Mode selection, semantic symbol allocation	DQN
[103]	Overall secure semantic spectrum efficiency	Subchanne assignment, IRS reflective coefficients		Bit decision set	Dueling DQN+SAC
Joint Communication-Computation-Network Parameters/Storage Resources Allocation					
[108]	EoSI	Bandwidth	CC	Semantic retention ratio	TD3
[82]	Average target detection accuracy			Storage Resource: Cache	DDQN

V. DISTRIBUTED RESOURCE ALLOCATION ALGORITHMS

Nowadays, the network structure of wireless communications is increasingly oriented toward a multilevel heterogeneous network structure, and efficiently managing resource allocation in such a complex environment requires a fundamental shift from traditional centralized mechanisms to self-organizing and self-optimizing approaches [172]. In this context, more and more distributed methods have been utilized to meet the increasingly complex situation. This section will provide an illustration of the distributed optimization algorithms used in the literature on resource allocation in SemCom, among them matching theory and auctions originating from the field of economics, and a portion of reinforcement learning with multi-agents. Fig. 11 shows the taxonomy of centralized resource allocation algorithms in SemCom.

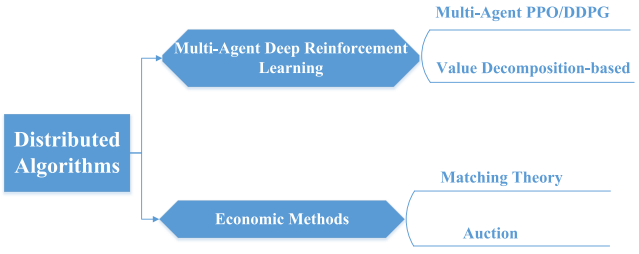

Fig. 11. The taxonomy of distributed resource allocation algorithms in SemCom.

A. Multi-Agent Deep Reinforcement Learning

Multi-agent reinforcement learning (MARL) is the application of reinforcement learning ideas and algorithms to multi-intelligent systems, extending MARL to deep reinforcement learning is multi-agent deep reinforcement learning

TABLE XII
MAPPING OF SEMCOM-SPECIFIC CHALLENGES TO OPTIMIZATION STRATEGIES AND POTENTIAL EXTENSIONS

SemCom-specific Challenges	Representative Works and SemCom Scenario	Metric and Optimization Methods	How They Address the Challenge	Potential Extensions or Hybridization
Tradeoff caused by Semantic Compression Ratio	[60]: Semantic-aware MEC; Tradeoff of transmission and computing	*Latency*: AO, Geometric Programming; Bisection and Interior-Point Method	**Transform then Decouple and Iteratively Solving:** Use auxiliary variable and geometric programming to transform the original non-convex optimization problem to convex. AO decouple it to three subproblems, iteratively solve them by bisection method and interior-point method.	Explore game-theoretic approaches (e.g., non-cooperative or Stackelberg games) to model the interaction among users making semantic compression decisions under limited resources. Hybridize DRL with analytical solvers (e.g., AO, GP) for better convergence and interpretability, improve reward designing.
	[108]: SemCom-IIoT; Trade-off between semantic timeliness, task accuracy and resource cost	*EoSI:* TD3	**DRL for Complex Trade-offs:** Use TD3 to capture the nonlinear tradeoffs between semantic timeliness, task accuracy, and resource cost, adaptively learn how to allocate these coupled resources (bandwidth, computing capacity, compress ratio) in the EoSI objective, which are intractable for traditional methods.	
Optimization with Non-differentiable and Implicit Objectives	[121]:Multi-cell/modal network; No close-from expression on semantic fidelity	*Semantic QoE:* DL-Approximation; DQN + Matching theory	**DL-based Curve Fitting:** Adopt DNN to fit the relationship curve of semantic fidelity and optimization variables for each task (single modal and bimodal); Then use DQN to derive semantic symbol allocation and use matching theory to solve channel assignment and power allocation subproblem.	Future research can explore close-box optimization techniques such as evolutionary strategies or Bayesian optimization to directly optimize non-differentiable objectives without requiring explicit functional forms. These approaches can be further hybridized with DRL frameworks to improve adaptability in complex environments.
	[112]: Image classification task; No close-from expression on task accuracy	*Transmission efficiency of tasks:* Curve fitting method; DDPG	**Parameter-based Curve Fitting:** Use three parameter $(\alpha_1, \alpha_2, \alpha_3)$ to model classification accuracy as $A = \alpha_1(\rho)^{\alpha_2} + \alpha_3$, then MSE of the prediction accuracy and the actual accuracy is used as the loss function, and the Levenberg-Marquardt method is employed to minimize the loss function and solve the three parameters. Then use DDPG to optimize bandwidth, power, and semantic compression ratio.	
Highly Coupled and Non-convex Optimization Variables	[59]: Semantic-aware dynamic long-term MEC systems	*Energy*: Lyapunov Optimization, AO, SCA	**Multiple-decoupling:** Use Lyapunov optimization transforms long-term coupling into per-slot problem; AO decomposed per-slot problem into 3 subproblems and iteratively optimized using SCA and bisection method.	Introduce hierarchical RL to model different levels of variable optimization (e.g., semantic-aware scheduling at upper level, other resource tuning at lower level). Leverage graph-based NN to capture latent coupling structures.
	[90]:IRS-assisted SemCom with subchannels	*ES-SE*: D3QN+SAC	**Double DRL Framework:** Use value-based DRL method **D3QN** to solve subchannel assignment, and policy-based method **SAC** to solve IRS-related actions (IRS reflection array).	
Task-related Semantic Information Transmission	[109]: Importance-aware task-oriented SemCom; Image classification	*TOSSE:* K-PPO; Hungarian Algorithm	**Traditional Method Assisted DRL:** In the system model, an importance-aware feature selection module evaluates semantic importance and decides the feature transmission ratio for each user. To jointly maximize the overall system TOSSE, PPO is adopted to complete importance ratio selection and Hungarian algorithm is used to optimize channel assignment at each step in PPO, thus forming the K-PPO framework for the joint optimization of feature selection and channel assignment.	Use Graph Neural Networks (GNNs) to model correlations between features or tasks, improving semantic importance estimation. Introduce meta-RL to quickly adapt the importance selection policy to new tasks or user demands with few-shot learning.

(MADRL). In [117], direct DQN is generalized to multi-agent DQN, and UAVs in the coverage area of different MEC servers are considered agents in the DQN algorithm. However, other literature utilized more refined and mature multi-intelligent body deep reinforcement learning methods, as follows:

1) Multi-Agent PPO/DDPG Algorithm: The multi-agent PPO (MAPPO) algorithm [173] is a variant of the PPO algorithm applied to multi-agent tasks, the critic can observe the global state, including information about other agents and the environment. The basic idea of the MAPPO algorithm is centralized training and decentralized execution (CTDE). The optimization problem of joint computational resources and bandwidth allocation is established in [126] with the objective of maximizing semantic accuracy. The problem is then transformed into a DRL framework, and MAPPO is utilized to solve the problem.

MADDPG (Multi-Agent Deep Deterministic Policy Gradient) is specifically designed for multi-agent systems, also leveraging a CTDE approach. During training, agents share observations and actions to learn coordinated strategies, while during execution, each agent acts independently based on its own policy. Reference [123], the modified MADDPG method is designed to optimize both global system performance and individual agent behavior in a dynamic semantic communication environment. The simulation results show that the proposed algorithm performs better than centralized DRL methods like DDPG and TD3.

2) MADRL Based on Value Decomposition: MADRL based on Value Decomposition (VD) is one of the many MADRL algorithms. It utilizes some constraints to decompose the joint action-value function of a multi-agent system into a specific combination of individual action-value functions and is able to effectively solve problems such as environmental non-stability and exponential explosion of the action space in multi-agent systems, ensuring the convergence of the algorithm.

In [51], a VD-based DQN is used to allow users and BSs to work together to find a team RB allocation and partial semantic information transmission scheme to optimize the similarity of all users. The work of [42] proposed a VD-based entropy-maximized MARL (VD-ERL) algorithm. The algorithm enables each server to coordinate its work with other servers in the training phase, perform RB allocation in a distributed manner, and approximate the global-optimal performance with fewer training iterations.

Summary: The advantage of MADRL is the ability to solve complex multi-intelligent collaboration problems, which is in line with the trend of increasingly complex real-world network changes. The disadvantages are the complexity of the training process and the difficulty of balancing collaboration and competition.

B. Economic Methods

There are two main economic methods used in distributed optimization algorithms in resource allocation of SemCom: matching theory and auction. They are subfields of economics and are promising concepts in distributed resource management and allocation.

1) Matching Theory: As a powerful tool for studying the dynamics and mutually beneficial relationships formed between different types of agents, the matching theory is particularly well suited to develop practical and high-performance, low-complexity, decentralized solutions in these complex networks. In particular, it can effectively cope with the high dynamics of the network, the selfish, competitive, and distributed nature of the network elements, the limited wireless resources, and the QoS constraints of the different elements [174].

The authors of [66] used many-to-many matching to solve the subproblem of the association between RIS users. However, the authors of [22] utilized matching theory to solve the subproblems of channel association and power allocation. In order to cope with the tight coupling between users in multi-cell user and bimodal user pairs, a matching game pair is constructed for modeling, and a low-complexity matching algorithm is proposed to obtain stable matching in this part. The authors of [77] establish a many-to-one matching game to determine the joint communication mode and the channel selection problem, in which the users and channels act as the game players. The computational resource was allocated by the PSO algorithm in each step of the matching game.

2) Auction: As a subfield of economics and business management, auction theory provides an interdisciplinary technique for the allocation of wireless resources (e.g., subchannels, time slots, and transmit power levels) in wireless systems. Auction methods are widely used in areas such as cognitive radio, cellular networks, and wireless grid networks [175]. In [85], the bids of IoT devices (bidders) for energy and power transmitters (auctioneers) are used to determine the winner and payment by competing for the energy of the hybrid access point (H-AP) through an optimal auction based on DL [176]. The IoT devices will bid for energy based on sentence similarity and BLEU score derived from the BERT-based model. In general, when the BLEU score and similarity score are higher, the device has a greater incentive to pay a higher price for energy.

Summary: The advantages of the economic approaches is: effective in highly dynamic and complex heterogeneous networks, practical in real-world scenarios. However, there are some disadvantages, which include: a) the global optimal solution may not be obtained; b) in the auction, the need for an additional third-party trusted organization for auction management may incur additional costs.

In this section, we explore distributed resource allocation algorithms in SemCom, focusing on multi-agent deep reinforcement learning (MADRL), matching theory and auction methods. MADRL, including the multi-agent PPO and value decomposition-based algorithms, is discussed for its ability to handle complex coordination tasks. Economic methods, such as matching theory and auction theory, are highlighted for their efficiency in decentralized resource allocation in dynamic environments. Table XIII reviews the literature using distributed resource allocation optimization algorithms.

VI. Open Challenges and Future Research Directions

Despite the significant achievements in resource allocation research in SemCom, many key issues remain unexplored. This section discusses several open research challenges and future research directions.

A. Resource Allocation Under Other SemCom Network Architectures

The diversity of SemCom network architectures reflects its adaptability to a wide range of application scenarios, from dynamic environments to multimodal communications. However, without tailored resource allocation strategies, these architectures cannot achieve their full potential. Developing solutions for specific frameworks is critical to maximize performance under real-world constraints. This progress will expand the scope of SemCom's applications, driving innovations in areas such as smart homes, autonomous vehicles, and immersive metaverse applications.

1) Network Architecture Combined With NGMA: Resource allocation problem of multi-user SemCom is particularly critical in scenarios with dense user environments or limited communication resources, most of the previous multiple access methods used FDMA, OFDMA, or TDMA. However, as communication technology continues to develop, researchers have begun to investigate the application of the combination of NGMA and SemCom in resource allocation. Incorporating NGMA into SemCom architectures could lead to transformative advances in scenarios requiring high connectivity, such as smart cities, industrial IoT, and Metaverse. Currently, some scholars have carried out research in this field; see Section II. A 2) of this article.

2) Network Architecture in a Dynamic Environment: Dynamic environments, such as vehicular networks or disaster emergency communications, are highly unpredictable due to factors such as user mobility, interference, and time-varying

TABLE XIII
DISTRIBUTED OPTIMIZATION ALGORITHMS BASED ON MADRL AND ECONOMIC METHODS

Ref.	Optimization objective	Resource to be allocated			Optimization algorithm
		Communication	Computing	Network parameter	
Single-Resource Allocation					
[99]	Averaged service satisfaction level of all users	Sub-channel assignment, power allocation	-	-	Multi-agent DDQN+DDPG with CTDE and Event-Triggered Control (ETC) mechanisms
[126]	Average semantic accuracy	-	-	Semantic coding model (SCM) selection, semantic compression ratio	MAPPO
Joint Two Resources Allocation					
[117]	Utility function	Power, coding rate	Sub-task offloading, GPUs of edge server	-	Multi-agent DQN
[50]	Energy consumption	Transmit power	Offloading decision, local GPU frequency		Self-designed MAPPO algorithm, Federated learning
[51]	Semantic similarity	RB allocation	-	SI Selection	VD-based DQN
[42]	Time delay	User association, RB allocation			VD-ERL
[22]	Semantic QoE	Channel allocation, power		Semantic symbol allocation	Matching theory, AO algorithm, Exhaustive search algorithm
[120]	Secrecy semantic transmission rate	Channel allocation			Matching theory, Stackelberg game
[123]	QoE and success rate of effective semantic information transmission	Channel allocation, power allocation		Semantic symbols allocation, Mode selection	Modified MADDPG
Joint Communication-Computation-Network Parameters Resources Allocation					
[77]	Overall user cost	Channel selection	Computation capacity (BS and users)	Mode selection	Matching theory, AO algorithm, PSO algorithm

channel conditions. Designing resource allocation strategies that account for these dynamic characteristics is critical to ensuring robust and reliable SemCom performance. The ability to adapt to dynamic environments directly affects the network's ability to deliver semantically accurate and timely communication. Moreover, this adaptability is essential for applications like real-time AR/VR, autonomous systems, and telemedicine, where latency and semantic accuracy are paramount.

3) Network Architecture of Speech SemCom: The rise of speech-based interfaces in consumer electronics, smart homes, and healthcare applications highlights the need for optimized resource allocation in speech SemCom. Unlike text and image modalities, speech signals have unique characteristics, such as real-time requirements, continuous data streams, and high sensitivity to latency and noise. Addressing these challenges in resource allocation will be crucial for improving the quality and efficiency of speech communication systems. Improved resource allocation for speech SemCom could enhance applications such as real-time voice recognition, smart home automation, and voice-assisted healthcare, ensuring that these systems operate smoothly with minimal delay.

4) Network Architecture of Multi-Modal SemCom: Most of the existing SemCom network models are developed around a single modality. However, scenarios such as Metaverse require a multi-modal service model that includes multiple types of instant interactions, such as audio, image, video, and tactile services. This requires a multi-modal SemCom network to solve. As a result, resource allocation for multi-modal SemCom networks becomes a critical challenge. Traditional single-modal resource allocation techniques, optimized for simple scenarios, are inadequate when it comes to managing the dynamic and diverse needs of multi-modal data streams. This problem is even more important from a user-centric standpoint, as efficient resource allocation is essential to ensure seamless, high-quality interactions across different types of services. At this time, the resource allocation problem for multimodal SemCom networks will also become a major challenge.

B. Establishment of SemCom Related Theory

Carnap and Bar-Hillel first proposed Classic Semantic Information Theory (CSIT) in 1952, based on logical probability [177]. Inspired by this pioneering work, some theoretical research has been carried out in the past two decades, such as [2] and [178], but it is not sufficient, especially in SemCom based on the DL framework. This gap in foundational theory presents a critical opportunity for advancing SemCom, as developing a solid theoretical framework will enable more effective, robust communication systems in dynamic and evolving environments.

1) Building a Universal Semantic Information Theory Framework: Compared to traditional information theory, which has been studied for many years, the development of semantic information theory is relatively weak. Mainly reflected in three aspects: a) So far, there has been no unified theoretical method for how to represent and measure semantics. b) SemCom lacks a comprehensive mathematical basis. c) It is difficult to extend the theory of traditional information theory to semantic information theory.

2) Building More Advanced Semantic Performance Metrics: The diversity of different scenarios and tasks in which SemCom systems are deployed means that a single static performance metric will not suffice. Moreover, shifting towards a user-centric SemCom paradigm is another critical challenge. Traditional metrics, such as bit error rates, typically focus on technical performance, but do not capture the real-world effectiveness of SemCom systems from the user's

TABLE XIV
CHALLENGES AND FUTURE DIRECTIONS

Exisiting Chanllenges	Proposed Future Directions
Resource Allocation under Other SemCom Network Architectures	**Architecture Combined with NGMA:** Significantly enhance SemCom system performance in high-connectivity scenarios like smart cities, industrial IoT, and the Metaverse.
	Architecture in a Dynamic Environment: Maintaining reliability and accuracy in high mobility SemCom systems, especially for time-sensitive applications like AR/VR, autonomous systems, and telemedicine.
	Architecture of Speech SemCom: Improve the performance of speech applications like international conferences and voice-assisted healthcare.
	Architecture of Multi-modal SemCom: Manage dynamic and diverse data streams across multiple modalities, such as audio, image, video, and tactile services. Enhance seamless, high-quality interactions in applications like the Metaverse.
Establishment of SemCom Related Theory	**Building a Universal Semantic Information Theory Framework:** Prove theoretical support for the critical improvements of SemCom, enable more effective, robust communication systems in dynamic and evolving environments.
	Building More Advanced Semantic Performance Metrics: Accurately evaluate the real-world effectiveness and versatility of SemCom systems across diverse scenarios and tasks. Enhance user satisfaction from the perspective of evaluation.
SemCom Resource Allocation Optimization Scheme	**Combination of Multiple Algorithms:** Improved service quality, reduced energy consumption, and the ability to handle growing user demands.
	FL-enabled and other DL-enabled Techniques: Other under-explored ways, lack of research, have potential advantages.
Other Challenges	**Transformation of Focus:** A shift from system-level to user-centric in metrics and resource allocation schemes, address diverse user needs and preferences, ensuring more user-satisfying and efficient communications.
	Security and Privacy Issue: Failures or attacks can compromise network reliability while safeguarding sensitive data is essential to prevent financial and reputational damage in applications like IoV, IIoT, and telemedicine.

perspective. Establishing such advanced metrics will enable a more accurate evaluation of a system's ability to provide value to users, beyond just raw data transmission efficiency.

C. SemCom Resource Allocation Optimization Scheme

1) Combination of Multiple Algorithms: The resource allocation problem can be solved by combining mathematical optimization and the RL method to save computing resources and achieve the optimal solution. Similarly, combining heuristic algorithms with optimization allows fast, near-optimal solutions in time-sensitive situations. The impact of these combined techniques is significant: improved service quality, reduced energy consumption, and the ability to meet the demands of the growing user.

2) FL-Enabled and Other DL-Enabled Techniques: Federated learning (FL) is a machine learning technology that can train resource scheduling algorithms on multiple distributed edge devices or servers that do not exchange local data samples [179]. FL allows edge devices to collaboratively learn resource scheduling policies without sharing raw semantic data. This decentralized training paradigm naturally protects user privacy, making it well-suited for privacy-sensitive SemCom applications such as telemedicine or autonomous driving. While this approach remains largely unexplored, the following outline presents one possible way to incorporate federated learning into semantic resource allocation:

1) A local model is trained on each device to make resource allocation schemes (e.g., semantic compression ratio, offloading decision).
2) Devices send model updates (e.g., gradients or parameters) to a central aggregator.
3) The server performs model aggregation and updates the global model.
4) The global model is redistributed to devices for the next training round.

Additionally, other deep learning techniques, such as diffusion models and graph neural networks (GNNs), can also play a crucial role in resource allocation optimization. GNNs are particularly useful for modeling complex, structured data, such as network topologies or relationships between devices in a distributed SemCom environment. GNNs can help optimize communication pathways and improve resource allocation by modeling dependencies between nodes in real time, problems partially modeled as combination optimization problems, or other special scenarios. A simple outline below illustrates how GNNs could help address channel assignment as a combinatorial problem in semantic communication systems:

1) Graph modeling: Represent users and interference links as a graph.
2) Problem setup: Frame channel assignment as a combinatorial optimization problem.
3) GNN encoding: Use GNNs to learn node embeddings that capture interference and task demands.
4) Solution generation: Predict channel assignments directly or guide heuristics with learned scores.
5) Training feedback: Optimize GNN using corresponding metrics like S-SE.

Diffusion models, with their generative capabilities, may support joint optimization of semantic compression and resource usage under strict delay or accuracy constraints. Here are some open research issues about utilizing these DL-enabled techniques in resource allocation in SemCom:

- Communication overhead for frequent model updates can be significant in FL.
- Lightweight FL protocols are needed for resource-constrained devices.
- Scalability challenges for large-scale distributed systems and difficulty in integrating multiple objectives (e.g., latency, accuracy, energy) into GNN-based models.
- High computational cost of diffusion models may hinder real-time deployment.

D. Other Challenges

There are still some other challenges that need to be considered in the resource allocation of SemCom.32

1) Transformation of Focus: Resource allocation in SemCom is undergoing a shift from system-level optimization

to a user-centric approach. Traditionally, resource allocation schemes in SemCom primarily aim to maximize system-wide performance metrics, such as throughput, latency, or energy efficiency. However, with the growing demand for personalized services and the increasing diversity of user requirements, future resource allocation schemes must prioritize individual user satisfaction. Despite its importance, transforming resource allocation into a user-centric paradigm poses significant challenges. One of the biggest obstacles is the design of meaningful and measurable metrics that accurately reflect user satisfaction, as it requires accounting for subjective factors such as preferences and context. Additionally, these metrics need to be dynamic and adaptable to varying scenarios, such as real-time changes in user behavior or network conditions. Another challenge lies in the inherent complexity of resource allocation to meet diverse and sometimes conflicting user needs. For example, balancing the satisfaction of multiple users while ensuring the fairness and efficient use of network resources requires advanced algorithms and computationally efficient solutions. Moreover, these resource allocation algorithms must be scalable to accommodate the huge number of devices and users envisioned in 6G.

2) Security and Privacy Issue: In existing research on SemCom resource allocation, security and privacy concerns have not been adequately addressed. However, the failure of a SemCom model or an attack on the system can significantly undermine the reliability and robustness of the entire network, rendering resource allocation ineffective. It is critical to develop mechanisms to enhance the robustness of the system and to formulate intrusion detection strategies to protect against vulnerabilities. In real-world applications such as semantic-aware IoV, SemCom IIoT, 6G-envisioned telemedicine, and the Metaverse, large amounts of transmitted data often involve user privacy and even business-sensitive information. If these data are exposed or leaked during transmission and processing, it could lead to substantial financial and reputational losses. One promising approach to address privacy issues is the application of federated learning (FL) in SemCom. By enabling devices to collaboratively train semantic extraction models without sharing raw data, FL helps preserve user privacy and reduces the risk of sensitive information leakage. This decentralized learning paradigm is especially suitable for privacy-critical scenarios, where traditional centralized training may not be viable. Therefore, this is an important future research direction. This research direction is crucial for the scalability and trustworthiness of SemCom networks in the future, helping them meet the increasing demands of emerging applications without compromising user privacy or system performance.

In this section, we discuss various challenges and future research directions for resource allocation in SemCom networks. These challenges and their future directions are summarized in Table XIV.

VII. Conclusion

In this survey, we provide a systematic and comprehensive overview of the resource allocation problem in SemCom. We also summarize and explain the network structure and resource allocation types in these studies, emphasize the performance indicators and resource allocation optimization algorithms in these studies, and provide detailed tables to summarize these studies. We identify current research bottlenecks and challenges in the allocation of SemCom resources and anticipate further research in the future. We hope that our work can provide references and insight to future researchers, as well as encourage follow-up research.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions, which helped to improve the quality of this article.

References

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL, USA: Univ. Illinois Press, 1949.
[2] J. Bao et al., "Towards a theory of semantic communication," in *Proc. IEEE Netw. Sci. Workshop*, 2011, pp. 110–117.
[3] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," 2022, *arXiv:2201.01389*.
[4] D. Gündüz et al., "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.
[5] 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.
[6] Z. Lu et al., "Semantics-empowered communications: A tutorial-cum-survey," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 1, pp. 41–79, 1st Quart., 2024.
[7] C. Zhang, H. Zou, S. Lasaulce, W. Saad, M. Kountouris, and M. Bennis, "Goal-oriented communications for the IoT and application to data compression," *IEEE Internet Things Mag.*, vol. 5, no. 4, pp. 58–63, Dec. 2022.
[8] S. Iyer et al., "A survey on semantic communications for intelligent wireless networks," *Wireless Pers. Commun.*, vol. 129, no. 1, pp. 569–611, Mar. 2023.
[9] Y. Liu, X. Wang, Z. Ning, M. Zhou, L. Guo, and B. Jedari, "A survey on semantic communications: Technologies, solutions, applications and challenges," *Digit. Commun. Netw.*, vol. 10, no. 3, pp. 528–545, 2024.
[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] 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.
[12] W. Yang et al., "Semantic communications for future Internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, 1st Quart., 2023.
[13] T. M. Getu, G. Kaddoum, and M. Bennis, "Making sense of meaning: A survey on metrics for semantic and goal-oriented communication," *IEEE Access*, vol. 11, pp. 45456–45492, 2023.
[14] T. M. Getu, G. Kaddoum, and M. Bennis, "A survey on goal-oriented semantic communication: Techniques, challenges, and future directions," *IEEE Access*, vol. 12, pp. 51223–51274, 2024.
[15] D. Won et al., "Resource management, security, and privacy issues in semantic communications: A survey," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 3, pp. 1758–1797, Jun. 2025.
[16] Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, "Resource scheduling in edge computing: A survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2131–2165, 4th Quart., 2021.
[17] A. Sarah, G. Nencioni, and M. M. I. Khan, "Resource allocation in multi-access edge computing for 5G-and-beyond networks," *Comput. Netw.*, vol. 227, May 2023, Art. no. 109720.
[18] Naren, A. K. Gaurav, N. Sahu, A. P. Dash, G. Chalapathi, and V. Chamola, "A survey on computation resource allocation in IoT enabled vehicular edge computing," *Complex Intell. Syst.*, vol. 8, pp. 3683–3705, Oct. 2022.

[19] B. Bossy, P. Kryszkiewicz, and H. Bogucka, "Energy-efficient OFDM radio resource allocation optimization with computational awareness: A survey," *IEEE Access*, vol. 10, pp. 94100–94132, 2022.

[20] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2134–2168, 3rd Quart., 2019.

[21] E. C. Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, no. 8, pp. 1–17, May 2021.

[22] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "QoE-aware resource allocation for semantic communication networks," in *Proc. IEEE Global Commun. Conf.*, 2022, pp. 3272–3277.

[23] 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.

[24] M. Rao, N. Farsad, and A. Goldsmith, "Variable length joint source channel coding of text using deep neural networks," in *Proc. 19th Int. Workshop Signal Process Adv. Wireless Commun. (SPAWC), Kalamata*, 2018, pp. 1–5.

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

[26] 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.

[27] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran Assoc., Inc., 2017.

[28] S. Jiang et al., "Reliable semantic communication system enabled by knowledge graph," *Entropy*, vol. 24, no. 6, p. 846, 2022.

[29] J. Liang, Y. Xiao, Y. Li, G. Shi, and M. Bennis, "Life-long learning for reasoning-based semantic communication," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2022, pp. 271–276.

[30] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[31] C. Dong, H. Liang, X. Xu, S. Han, B. Wang, and P. Zhang, "Semantic communication system based on semantic slice models propagation," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 202–213, Jan. 2023.

[32] M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, and A. Fernando, "Wireless end-to-end image transmission system using semantic communications," *IEEE Access*, vol. 11, pp. 37149–37163, 2023.

[33] S. Kadam and D. I. Kim, "Semantic communication-empowered traffic management using vehicle count prediction," 2023, *arXiv:2307.12254*.

[34] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE J Sel Areas Commun*, vol. 39, pp. 2434–2444, Aug. 2021.

[35] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech recognition," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2021, pp. 1–6.

[36] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Commun. Lett.*, vol. 11, no. 7, pp. 1394–1398, Jul. 2022.

[37] Z. Weng, Z. Qin, and X. Tao, "Task-oriented semantic communications for speech transmission," in *Proc. IEEE 98th Veh. Technol. Conf. (VTC-Fall)*, 2023, pp. 1–5.

[38] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6227–6240, Sep. 2023.

[39] D. Huang, X. Tao, F. Gao, and J. Lu, "Deep learning-based image semantic coding for semantic communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2021, pp. 1–6.

[40] Y. Wang et al., "Performance optimization for semantic communications: An attention-based reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2598–2613, Sep. 2022.

[41] F. Zhou, Y. Li, X. Zhang, Q. Wu, X. Lei, and R. Q. Hu, "Cognitive semantic communication systems driven by knowledge graph," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 4860–4865.

[42] W. Zhang, Y. Wang, M. Chen, T. Luo, and D. Niyato, "Optimization of image transmission in cooperative semantic communication networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 2, pp. 861–873, Feb. 2024.

[43] J. Kang et al., "Personalized saliency in task-oriented semantic communications: Image transmission and performance analysis," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 186–201, Jan. 2023.

[44] Z. Hu, T. Liu, C. You, Z. Yang, and M. Chen, "Multiuser resource allocation for semantic-relay-aided text transmissions," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 1273–1278.

[45] T. Liu, C. You, Z. Hu, C. Wu, Y. Gong, and K. Huang, "Semantic-relay-aided text transmission: Placement optimization and bandwidth allocation," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 215–220.

[46] Y. Lil, X. Zhou, and J. Zhao, "Resource allocation for semantic communication under physical-layer security," in *Proc. IEEE Global Commun. Conf.*, 2023, pp. 2063–2068.

[47] H. Hu, X. Zhu, F. Zhou, W. Wu, and R. Q. Hu, "Semantic-oriented resource allocation for multi-modal UAV semantic communication networks," in *Proc. IEEE Global Commun. Conf.*, 2023, pp. 7213–7218.

[48] X. He, C. You, and T. Q. Quek, "Joint user association and resource allocation for multi-cell networks with adaptive semantic communication," 2024, *arXiv:2312.01049*.

[49] Y. Zheng, T. Zhang, R. Huang, and Y. Wang, "Computing offloading and semantic compression for intelligent computing tasks in MEC systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2023, pp. 1–6.

[50] Z. Ji and Z. Qin, "Energy-efficient task offloading for semantic-aware networks," in *Proc. IEEE Int. Conf. Commun.*, 2023, pp. 3584–3589.

[51] M. Chen, Y. Wang, and H. V. Poor, "Performance optimization for wireless semantic communications over energy harvesting networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 8647–8651.

[52] O. Marnissi, H. E. Hammouti, and E. H. Bergou, "Semantic-aware resource allocation in constrained networks with limited user participation," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2024, pp. 1–6.

[53] L. Wang, W. Wu, F. Zhou, Z. Yang, Z. Qin, and Q. Wu, "Adaptive resource allocation for semantic communication networks," *IEEE Trans. Commun.*, vol. 72, no. 11, pp. 6900–6916, Nov. 2024.

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

[55] L. Wang, W. Wu, F. Tian, and H. Hu, "Intelligent resource allocation for UAV-enabled spectrum sharing semantic communication networks," in *Proc. IEEE 23rd Int. Conf. Commun. Technol. (ICCT)*, 2023, pp. 1359–1363.

[56] G. Cheng, X. Wang, D. Li, R. Jiang, and Y. Xu, "Resource allocation for multi-cell semantic communication based on deep reinforcement learning," in *Proc. IEEE 23rd Int. Conf. Communication Technol. (ICCT)*, 2023, pp. 528–533.

[57] G. Ding, S. Liu, J. Yuan, and G. Yu, "Joint URLLC traffic scheduling and resource allocation for semantic communication systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7278–7290, Jul. 2024.

[58] W. Zhang, Y. Wang, M. Chen, T. Luo, and D. Niyato, "Optimization of image transmission in semantic communication networks," in *Proc. IEEE Global Commun. Conf.*, 2022, pp. 5965–5970.

[59] Y. Cang et al., "Online resource allocation for semantic-aware edge computing systems," *IEEE Internet Things J.*, vol. 11, no. 17, pp. 28094–28110, Sep. 2024.

[60] Y. Cang et al., "Resource allocation for semantic-aware mobile edge computing systems," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 1585–1590.

[61] H. Zhang, H. Wang, Y. Li, K. Long, and V. C. Leung, "Toward intelligent resource allocation on task-oriented semantic communication," *IEEE Wireless Commun.*, vol. 30, no. 3, pp. 70–77, Jun. 2023.

[62] M. Poposka, H. A. Suraweera, G. K. Karagiannidis, and Z. Hadzi-Velkov, "Semantic wireless networks with minimal energy consumption," *IEEE Commun. Lett.*, vol. 28, no. 8, pp. 1894–1898, Aug. 2024.

[63] Y. Mao, B. Clerckx, and V. O. K. Li, "Rate-splitting multiple access for downlink communication systems: Bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 133, May 2018.

[64] Z. Zhao, Z. Yang, M. Chen, Z. Zhang, and H. V. Poor, "A joint communication and computation design for probabilistic semantic communications," *Entropy*, vol. 26, no. 5, p. 394, 2024. [Online]. Available: https://www.mdpi.com/1099-4300/26/5/394

[65] Z. Zhao et al., "Multi-user probabilistic semantic communication with semantic compression ratio optimization," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2024, pp. 1647–1652.

[66] Z. Zhao et al., "A joint communication and computation design for distributed RIS-assisted probabilistic semantic communication in IIoT," *IEEE Internet Things J.*, vol. 11, no. 16, pp. 26568–26579, Aug. 2024.

[67] Z. Yang, M. Chen, Z. Zhang, and C. Huang, "Energy efficient semantic communication over wireless networks with rate splitting," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1484–1495, May 2023.

[68] C. Zeng et al., "Task-oriented semantic communication over rate splitting enabled wireless control systems for URLLC services," *IEEE Trans. Commun.*, vol. 72, no. 2, pp. 722–739, Feb. 2024.

[69] R. Xu, Z. Yang, Z. Zhao, Q. Yang, and Z. Zhang, "Resource allocation for green probabilistic semantic communication with rate splitting," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2024, pp. 2017–2022.

[70] Z. Zhao et al., "Spectral efficiency Maximization for probabilistic semantic communication with rate splitting," in *Proc. IEEE 99th Veh. Technol. Conf. (VTC-Spring)*, 2024, pp. 1–5.

[71] Y. Cheng et al., "Resource allocation and common message selection for task-oriented semantic information transmission with RSMA," *IEEE Trans. Wireless Commun.*, vol. 23, no. 6, pp. 5557–5570, Jun. 2024.

[72] M. Lu, J. Huang, T. Yang, Y. Wang, J. Jiao, and Q. Zhang, "Utility loss of information-optimal for semantic empowered RSMA in satellite-integrated Internet," *IEEE Internet Things J.*, vol. 11, no. 24, pp. 40572–40587, Dec. 2024.

[73] N. G. Evgenidis et al., "Delay minimization for hybrid semantic-Shannon communications," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2024, pp. 1–6.

[74] J. Zhao, M. Chen, Z. Yang, C. You, and M. Chen, "Resource allocation for semantic relay aided wireless networks with probability graph," in *Proc. IEEE Int. Conf. Commun.*, 2024, pp. 5317–5322.

[75] M. Zhang, R. Zhong, X. Mu, Y. Chen, and Y. Liu, "Resource management for heterogeneous semantic and bit communication systems," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2023, pp. 1629–1634.

[76] H. Noh, S. Park, and H. J. Yang, "Deep reinforcement learning-based resource allocation and mode selection for semantic communication," in *Proc. 22nd Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, 2024, pp. 1–6.

[77] P. Li, Y. Wang, M. Liu, and H. Liu, "Matching game based resource allocation scheme for adaptive semantic and bit communication networks," in *Proc. IEEE 99th Veh. Technol. Conf. (VTC-Spring)*, 2024, pp. 1–7.

[78] L. Xia, Y. Sun, D. Niyato, L. Zhang, and M. A. Imran, "Wireless resource optimization in hybrid semantic/bit communication networks," *IEEE Trans. Commun.*, vol. 73, no. 5, pp. 3318–3332, May 2025.

[79] J. Li, H. Gao, T. Lv, and Y. Lu, "Deep reinforcement learning based computation offloading and resource allocation for MEC," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2018, pp. 1–6.

[80] Y. Liu, H. Yu, S. Xie, and Y. Zhang, "Deep reinforcement learning for offloading and resource allocation in vehicle edge computing and networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11158–11168, Nov. 2019.

[81] S. Wang, T. Lv, W. Ni, N. C. Beaulieu, and Y. J. Guo, "Joint resource management for MC-NOMA: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5672–5688, Sep. 2021.

[82] R. Lin, C. Guo, J. Chen, and Y. Wang, "Multidimensional resource joint allocation algorithm based on deep double Q network in semantic communication, (in Chinese)," *Mobile Commun.*, vol. 47, no. 4, pp. 45–53, 2023.

[83] H. Wang, L. Wang, and W. Wu, "Resource allocation and intelligent trajectory optimization for UAV-assisted semantic communication system," in *Proc. IEEE 23rd Int. Conf. Commun. Technol. (ICCT)*, 2023, pp. 1370–1374.

[84] Z. Zhao, Z. Yang, Q.-V. Pham, Q. Yang, and Z. Zhang, "Semantic communication with probability graph: A joint communication and computation design," in *Proc. IEEE 98th Veh. Technol. Conf. (VTC-Fall)*, 2023, pp. 1–5.

[85] 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," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 8637–8641.

[86] Q. Cai et al., "Query-aware semantic encoder-based resource allocation in task-oriented communications," *IEEE Trans. Mobile Comput.*, vol. 24, no. 7, pp. 6413–6429, Jul. 2025.

[87] Z. Yang, M. Chen, Z. Zhang, C. Huang, and Q. Yang, "Performance optimization of energy efficient semantic communications over wireless networks," in *Proc. IEEE 96th Veh. Technol. Conf. (VTC-Fall)*, 2022, pp. 1–5.

[88] A. Xiao, K. Zhao, Z. Liu, and C. Liang, "Energy efficiency in semantic networks: A heuristic optimization approach for resource allocation," in *Proc. 28th Asia–Pacific Conf. Commun. (APCC)*, 2023, pp. 219–224.

[89] L. Chen and J. Gong, "Multi-source scheduling and resource allocation for age-of-semantic-importance optimization in status update systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2024, pp. 1–6.

[90] B. Hu, J. Ma, Z. Sun, J. Liu, R. Li, and L. Wang, "DRL-based intelligent resource allocation for physical layer semantic communication with IRS," *Phys. Commun.*, vol. 63, Apr. 2024, Art. no. 102270.

[91] J. Dai, H. Fan, Z. Zhao, Y. Sun, and Z. Yang, "Secure resource allocation for integrated sensing and semantic communication system," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2024, pp. 1225–1230.

[92] J. X. Dai, H. Fan, Z. X. Zhao et al., "Joint communication and computation design for secure integrated sensing and semantic communication system," *Sci. China Inf. Sci.*, vol. 68, no. 3, 2025, Art. no. 132301.

[93] Y. Yang, M. Shikh-Bahaei, Z. Yang, C. Huang, W. Xu, and Z. Zhang, "Joint semantic communication and target sensing for 6G communication system," 2024, *arXiv:2401.17108*.

[94] Y. Zhang, J. Li, G. Mu, and X. Chen, "A DRL-based resource allocation for IRS-enhanced semantic spectrum sharing networks," *EURASIP J. Adv. Signal Process.*, vol. 2024, no. 1, pp. 1–17, 2024.

[95] Z. Shao, Q. Wu, P. Fan, N. Cheng, Q. Fan, and J. Wang, "Semantic-aware resource allocation based on deep reinforcement learning for 5G-V2X HetNets," *IEEE Commun. Lett.*, vol. 28, no. 10, pp. 2452–2456, Oct. 2024.

[96] Z. Shao et al., "Semantic-aware spectrum sharing in Internet of Vehicles based on deep reinforcement learning," *IEEE Internet Things J.*, vol. 11, no. 23, pp. 38521–38536, Dec. 2024.

[97] L. Xia, Y. Sun, X. Li, G. Feng, and M. A. Imran, "Wireless resource management in intelligent semantic communication networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2022, pp. 1–6.

[98] L. Xia, Y. Sun, D. Niyato, X. Li, and M. A. Imran, "Joint user association and bandwidth allocation in semantic communication networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 2, pp. 2699–2711, Feb. 2024.

[99] X. Jia, X. Wang, Y. Zhang, M. Sheng, and G. Cheng, "Resource allocation for multi-cell semantic communication systems based on DRL," in *Proc. 12th Int. Conf. Inf. Syst. Comput. Technol. (ISCTech)*, 2024, pp. 1–6.

[100] L. Li, J. Dai, Z. Yang, Q. Yang, C. Huang, and Z. Zhang, "Joint compression ratio and user association for multi-cell probabilistic semantic communication," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, 2024, pp. 645–650.

[101] X. Pu, T. Lei, W. Wen, and Q. Chen, "Enhancing communication efficiency of semantic transmission via joint processing technique," *IEEE Commun. Lett.*, vol. 28, no. 3, pp. 657–661, Mar. 2024.

[102] J. Su et al., "Semantic communication-based dynamic resource allocation in d2d vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10784–10796, Aug. 2023.

[103] L. Wang, W. Wu, F. Zhou, Z. Qin, and Q. Wu, "IRS-enhanced secure semantic communication networks: Cross-layer and context-awared resource allocation," *IEEE Trans. Wireless Commun.*, vol. 24, no. 1, pp. 494–508, Jan. 2025.

[104] X. Xu, C. He, X. Li, and J. Xu, "Joint optimization trajectory and resource allocation for UAV-assisted semantic communications," *Phys. Commun.*, vol. 68, Feb. 2025, Art. no. 102555.

[105] Y. Li, X. Zhou, and J. Zhao, "Resource allocation for the training of image semantic communication networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 4, pp. 2968–2984, Apr. 2025.

[106] C. Liu, C. Guo, and Y. Yang, "Performance optimization for task-oriented communications," in *Proc. IEEE Int. Conf. Commun.*, 2024, pp. 968–973.

[107] J. Liu, Y. Lu, H. Wu, and Y. Dai, "Efficient resource allocation and semantic extraction for federated learning empowered vehicular semantic communication," in *Proc. IEEE 98th Veh. Technol. Conf. (VTC-Fall)*, 2023, pp. 1–5.

[108] X. Xiang, X. Li, Q. Cui, X. Zhang, and X. Tao, "EoSI-aware resource allocation for semantic communication-enabled industrial IoT system," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2023, pp. 477–483.

[109] Y. Wang et al., "Feature importance-aware task-oriented semantic transmission and optimization," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 4, pp. 1175-1189, Aug. 2024.

[110] G. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, and B. H. Soong, "Vision-based semantic communications for metaverse services: A contest theoretic approach," in *Proc. IEEE Global Commun. Conf.*, 2023, pp. 2426–2432.

[111] C. Liu, C. Guo, Y. Yang, and N. Jiang, "Adaptable semantic compression and resource allocation for task-oriented communications," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 3, pp. 769–782, Jun. 2024.

[112] H. Zhang, H. Wang, Y. Li, K. Long, and A. Nallanathan, "DRL-driven dynamic resource allocation for task-oriented semantic communication," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 3992–4004, Jul. 2023.

[113] B. Du et al., "YOLO-based semantic communication with generative AI-aided resource allocation for digital twins construction," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 7664–7678, Mar. 2024.

[114] J. Zheng, B. Du, H. Du, J. Kang, D. Niyato, and H. Zhang, "Energy-efficient resource allocation in generative AI-aided secure semantic mobile networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 11422–11435, Dec. 2024.

[115] W. C. Ng, H. Du, W. Y. B. Lim, Z. Xiong, D. Niyato, and C. Miao, "Stochastic resource allocation for semantic communication-aided virtual transportation networks in the metaverse," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2024, pp. 1–6.

[116] H. Saadat, A. Albaseer, M. Abdallah, A. Mohamed, and A. Erbad, "Energy-aware service offloading for semantic communications in wireless networks," in *Proc. IEEE Int. Conf. Commun.*, 2024, pp. 5467–5472.

[117] X. Sun, J. Chen, and C. Guo, "Semantic-driven computation offloading and resource allocation for UAV-assisted monitoring system in vehicular networks," in *Proc. 48th Annu. Conf. IEEE Ind. Electron. Soc.*, 2022, pp. 1–6.

[118] Y. Zheng, T. Zhang, and J. Loo, "Dynamic multi-time scale user admission and resource allocation for semantic extraction in MEC systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 12, pp. 16441–16453, Dec. 2023.

[119] X. Han, B. Feng, Y. Shi, Y. Wu, and W. Zhang, "Semantic-aware resource allocation for wireless image transmission," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2024, pp. 2071–2076.

[120] X. Yang, H. Yang, Y. Jiang, A. Alphones, and L. Xiao, "Game-guided matching theory-based resource allocation for secure semantic communications," *IEEE Trans. Veh. Technol.*, vol. 74, no. 5, pp. 8357–8362, May 2025.

[121] L. Yan, Z. Qin, C. Li, R. Zhang, Y. Li, and X. Tao, "QoE-based semantic-aware resource allocation for multi-task networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11958–11971, Sep. 2024.

[122] C. Huang, X. Chen, G. Chen, P. Xiao, G. Y. Li, and W. Huang, "Deep reinforcement learning-based resource allocation for hybrid bit and generative semantic communications in space-air-ground integrated networks," 2024, *arXiv:2412.05647*.

[123] Z. Shao et al., "Semantic-aware resource management for C-V2X platooning via multi-agent reinforcement learning," 2024, *arXiv:2411.04672*.

[124] J. Chen, C. Feng, C. Guo, Y. Yang, Q. Sun, and M. Zhu, "Video semantics-driven resource allocation algorithm in Internet of Vehicles, (in Chinese)," *J. Commun.*, vol. 42, no. 7, pp. 1–11, 2021.

[125] R. Lin, C. Guo, B. Zhang, J. Chen, and H. Li, "Tasks-oriented channel optimization and resource allocation in vehicular networks: A hierarchical reinforcement learning based approach," *IEEE Trans. Veh. Technol.*, vol. 74, no. 5, pp. 7624–7636, May 2025.

[126] F. Zhao, G. Bagwe, E. Mohammed, L. Feng, L. Zhang, and Y. Sun, "Joint computing resource and bandwidth allocation for semantic communication networks," in *Proc. IEEE 98th Veh. Technol. Conf. (VTC-Fall)*, 2023, pp. 1–5.

[127] Y. Zhu, X. Yuan, Y. Hu, and A. Schmeink, "Semantic reliability Maximization: A cooperative perspective in integrated sensing, communication and computation networks," in *Proc. IEEE Global Commun. Conf.*, 2023, pp. 5073–5079.

[128] S. Liang et al., "Fair resource allocation for probabilistic semantic communication in IIoT," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, 2024, pp. 242–247.

[129] H. Chen, F. Fang, and X. Wang, "Semantic extraction model selection for IoT devices in edge-assisted semantic communications," *IEEE Commun. Lett.*, vol. 28, no. 7, pp. 1733–1737, Jul. 2024.

[130] S. Hua et al., "Optimizing spectral efficiency through bandwidth management in semantic communication systems," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2024, pp. 1635–1640.

[131] C. Feng, K. Zheng, Y. Wang, K. Huang, and Q. Chen, "Goal-oriented wireless communication resource allocation for cyber-physical systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, pp. 15768–15783, Nov. 2024.

[132] Z. Zhao, Z. Yang, Q. Yang, C. Huang, M. Shikh-Bahaei, and Z. Zhang, "Sum rate maximization for distributed riss assisted probabilistic semantic communication," in *Proc. IEEE 34th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, 2024, pp. 1–6.

[133] K. Brunnström et al., "QualiNet white paper on definitions of quality of experience," presented at Eur. Netw. Qual. Exp. Multimedia Syst. Services (COST Action IC 1003), 2013.

[134] N. Banović-Ćurguz and D. Ilišević, "Mapping of QoS/QoE in 5G networks," in *Proc. 42nd Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, 2019, pp. 404–408.

[135] A. Takahashi, "Framework and standardization of quality of experience (QoE) design and management for audiovisual communication services," *NTT Tech. Rev.*, vol. 7, no. 4, pp. 1–5, 2009.

[136] M. I. Belghazi et al., "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 531–540.

[137] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 1995, pp. 181–184.

[138] M. E. Peters et al., "Deep contextualized word representations," in *Proc. North Amer. Chapter Assoc. Comput. Linguist. Hum. Lang. Tech.*, Jun. 2018, pp. 2227–2237.

[139] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Meas. Mach. Transl. /Or Summarization*, 2005, pp. 65–72.

[140] A. Radford et al., "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[141] 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.

[142] Y. Ao, Y. Li, S. He, D. Chen, Z. Qin, and X. Tao, "Research on resource allocation in cellular semantic communication systems, (in Chinese)," *Mobile Commun.*, vol. 48, no. 2, pp. 104–110, 2024.

[143] K. Niu et al., "A paradigm shift toward semantic communications," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 113–119, Nov. 2022.

[144] S. Kadam and D. I. Kim, "Knowledge-aware semantic communication system design," in *Proc. IEEE Int. Conf. Commun.*, 2023, pp. 6102–6107.

[145] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Sep. 2022.

[146] Y. Wang et al., "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 319–352, Jan. 2023.

[147] H. Du et al., "Attention-aware resource allocation and QoE analysis for metaverse xURLLC services," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2158–2175, Jul. 2023.

[148] H. Du et al., "Exploring attention-aware network resource allocation for customized metaverse services," *IEEE Netw.*, vol. 37, no. 6, pp. 166–175, Nov. 2023.

[149] A. Kosta, N. Pappas, and V. Angelakis, "Age of information: A new concept, metric, and tool," *Found. Trends® Netw.*, vol. 12, no. 3, pp. 162–259, 2017.

[150] S. Guo, Y. Wang, S. Li, and N. Saeed, "Semantic importance-aware communications using pre-trained language models," *IEEE Commun. Lett.*, vol. 27, no. 9, pp. 2328–2332, Sep. 2023.

[151] S. Kadam and D. I. Kim, "Knowledge-aware semantic communication system design and data allocation," *IEEE Trans. Veh. Technol.*, vol. 73, no. 4, pp. 5755–5769, Apr. 2024.

[152] M. Noor-A-Rahim, Z. Liu, H. Lee, G. G. M. N. Ali, D. Pesch, and P. Xiao, "A survey on resource allocation in vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 701–721, Feb. 2022.

[153] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, 1955.

[154] Z. Zhou, Y. Guo, Y. He, X. Zhao, and W. M. Bazzi, "Access control and resource allocation for M2M communications in industrial automation," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 3093–3103, May 2019.

[155] Z. Yu, Y. Gong, S. Gong, and Y. Guo, "Joint task offloading and resource allocation in UAV-enabled mobile edge computing," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3147–3159, Apr. 2020.

[156] H. Zhang, N. Yang, W. Huangfu, K. Long, and V. C. M. Leung, "Power control based on deep reinforcement learning for spectrum sharing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4209–4219, Jun. 2020.
[157] Z. Ding, R. Schober, and H. V. Poor, "No-pain no-gain: DRL assisted optimization in energy-constrained CR-NOMA networks," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5917–5932, Sep. 2021.
[158] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, pp. 279–292, May 1992.
[159] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[160] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 293–321, 1992.
[161] Y. Li, "Deep reinforcement learning: An overview," 2017, *arXiv:1701.07274*.
[162] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 2094–2100.
[163] H. V. Hasselt, "Double learning," in *Proc. 24th Annu. Conf. Neural Inf. Process. Syst.*, 2010, pp. 2613–2616.
[164] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1995–2003.
[165] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 1–7.
[166] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1–9.
[167] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–10.
[168] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
[169] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
[170] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.
[171] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, May 2016.
[172] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: Fundamentals and applications," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52–59, May 2015.
[173] C. Yu et al., "The surprising effectiveness of PPO in cooperative multi-agent games," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24611–24624.
[174] S. Bayat, Y. Li, L. Song, and Z. Han, "Matching theory: Applications in wireless communications," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 103–122, Nov. 2016.
[175] Y. Zhang, C. Lee, D. Niyato, and P. Wang, "Auction approaches for resource allocation in wireless systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1020–1041, 3rd Quart., 2013.
[176] P. Dütting, Z. Feng, H. Narasimhan, D. Parkes, and S. S. Ravindranath, "Optimal auctions through deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1706–1715.
[177] R. Carnap and Y. Bar-Hillel, "An outline of a theory of semantic information," Res. Lab. Electron., Massachusetts Inst. Technol., Cambridge, MA, USA, Rep. 247, Oct. 1952.
[178] B. Güler, A. Yener, and A. Swami, "The semantic communication game," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 787–802, Dec. 2018.
[179] J. Konečnỳ, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," 2015, *arXiv:1511.03575*.

Chujun Zhang received the B.E. degree in communication engineering from the School of Electrical Engineering and Information, Southwest Petroleum University, Chengdu, China, in 2023. He is currently pursuing the M.Sc. degree in information and communication engineering with the College of Electronics and Information Engineering, Sichuan University, Chengdu.

His current research interests include wireless communications, semantic communications, and resource allocation.

Linyu Huang (Member, IEEE) received the B.E. degree in electronic information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2008, and the Ph.D. degree in electronic engineering from the City University of Hong Kong in 2014. He joined the faculty with the College of Electronics and Information Engineering, Sichuan University, Chengdu, in 2014. His current research interests include wireless communication, signal processing, and machine learning.

Qian Ning received the bachelor's degree from Xidian University in 1990, the master's degree from the University of Electronic Science and Technology of China in 1997, and the Ph.D. degree from Sichuan University, Chengdu, China, in 2006, where she is currently an Associate Professor with the College of Electronics and Information Engineering. Her current research interests include intelligent systems and wireless ad hoc networks.