AI-Assisted Slicing-Based Resource Management for Two-Tier Radio Access Networks

Conghao Zhou, *Member, IEEE*, Jie Gao, *Senior Member, IEEE*, Mushu Li, *Member, IEEE*, Xuemin Shen, *Fellow, IEEE*, Weihua Zhuang, *Fellow, IEEE*, Xu Li, and Weisen Shi, *Member, IEEE*

***Abstract*—While network slicing has become a prevalent approach to service differentiation, radio access network (RAN) slicing remains challenging due to the need of substantial adaptivity and flexibility to cope with the highly dynamic network environment in RANs. In this paper, we develop a slicing-based resource management framework for a two-tier RAN to support multiple services with different quality of service (QoS) requirements. The developed framework focuses on base station (BS) service coverage (SC) and interference management for multiple slices, each of which corresponds to a service. New designs are introduced in the spatial, temporal, and slice dimensions to cope with spatiotemporal variations in data traffic, balance adaptivity and overhead of resource management, and enhance flexibility in service differentiation. Based on the proposed framework, an energy efficiency maximization problem is formulated, and an artificial intelligence (AI)-assisted approach is proposed to solve the problem. Specifically, a deep unsupervised learning-assisted algorithm is proposed for searching the optimal SC of the BSs, and an optimization-based analytical solution is found for managing interference among BSs. Simulation results under different data traffic distributions demonstrate that our proposed slicing-based resource management framework, empowered by the AI-assisted approach, outperforms the benchmark frameworks and achieves a close-to-optimal performance in energy efficiency.**

***Index Terms*—RAN slicing, service coverage management, interference management, deep unsupervised learning.**

I. Introduction

SINCE the 3rd Generation Partnership Project (3GPP) Release 18 for the advanced fifth generation communication network (5G-advanced) in 2021, academia have commenced their efforts on the development and deployment of next-generation wireless networks (NGWNs) [1]. NGWNs are anticipated to support a diverse set of disruptive new services such as extended reality (XR) and haptic communications [2]. As a result, the research and standardization efforts for NGWNs must address new challenges. First, services in NGWNs will have unprecedentedly stringent quality of service (QoS) requirements since a massive amount of data must be transmitted over networks with extremely low delay and ultra-high reliability [3], [4]. Second, the QoS requirements of services in NGWNs will become highly diverse. Meeting the stringent and diverse QoS requirements to support new services in NGWNs calls for advanced networking and communication techniques [5], [6].

Network slicing, as a key innovation in the fifth generation (5G), can support multiple coexisting virtual networks, i.e., slices, on the same physical network infrastructure [7]. Due to the advantages in QoS guarantee and service differentiation, network slicing lays a foundation for efficient resource management and will continue playing an important role in NGWNs. Some pioneering works have envisioned advanced slicing-based resource management for services in NGWNs with diverse and stringent QoS requirements [8], [9]. In these works, slicing-based resource management can be categorized into two stages, i.e., *planning stage* and *operation stage*. The planning stage focuses on network-wide configuration and proactive network resource reservation for different services, while the operation stage focuses on user-level service provisioning and real-time network resource allocation [10]. A planning period, referred to as the planning window, can be minutes or hours in length, whereas a network operation period, referred to as the operation window, is generally milliseconds in length. While planning and operation stages have different focuses, both play indispensable roles in slicing-based resource managements as they jointly determine QoS satisfaction [11], [12]. However, existing literature and 3GPP standards pay much more attention to the operation stage than to the planning stage.

Compared to the operation stage, slicing-based resource management in the planning stage faces unique challenges. First, real-time information on individual users is unavailable at the beginning of the planning stage when resources are reserved. Consequently, existing slicing-based resource management schemes in the planning stage rely on coarse-grained information such as the aggregated data traffic over a planning window, which may result in an inaccurate estimation of service demands and thus degrade network resource

Manuscript received 2 February 2023; revised 14 June 2023; accepted 12 August 2023. Date of publication 24 August 2023; date of current version 8 December 2023. This work was supported in part by research grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and in part by Huawei Technologies Canada. The associate editor coordinating the review of this article and approving it for publication was J. Liu. *(Corresponding author: Conghao Zhou.)*

Conghao Zhou, Xuemin Shen, and Weihua Zhuang are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: c89zhou@uwaterloo.ca; sshen@uwaterloo.ca; wzhuang@uwaterloo.ca).

Jie Gao was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada. He is now with the School of Information Technology, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: jie.gao6@carleton.ca).

Mushu Li was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada. She is now with the Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada (e-mail: mushu1.li@ryerson.ca).

Xu Li and Weisen Shi are with the Ottawa Advanced Wireless Technology Lab., Huawei Technologies Canada Inc., Ottawa, ON K2K 3J1, Canada (e-mail: xu.lica@huawei.com; weisen.shi1@huawei.com).

Digital Object Identifier 10.1109/TCCN.2023.3307929

utilization [13]. Second, user mobility and time-varying user behaviors result in significant spatiotemporal variations in service demands, which pose a challenge of balancing adaptivity and overhead in the planning stage of slicing-based resource management [14]. Third, differentiating services and satisfying their diverse and stringent QoS requirements further complicate the decision making on network-wide configurations and proactive resource reservation [7].

Following 5G standardization in Releases 15 to 17 as well as commercial 5G deployment, a large number of works have studied slicing-based resource management for supporting diverse services in core networks [3]. Nevertheless, slicing-based resource management for radio access networks (RANs) is still in its infancy [15], [16]. Ensuring service differentiation among multiple slices in RANs is more challenging than in core networks, and the reason is two-fold. First, interference occurs among the data transmissions of different base stations (BSs) within each slice since spectrum reuse takes place among the BSs for improving the spectrum multiplexing gain [9]. Such intra-slice interference causes challenges in accurately estimating the required amount of resources for each slice, thereby adversely affecting their QoS satisfaction [17]. Furthermore, inter-slice interference may occur and result in tightly coupled management (such as coverage management) among different slices in RANs, which hinders efficient slice isolation in RANs [18]. Therefore, slicing-based resource management for RANs that can address the aforementioned challenges needs to be further investigated in NGWNs.

In this paper, we investigate slicing-based resource management for a two-tier RAN, i.e., a single macro-cell in the first tier and multiple small cells in the second tier, to improve resource utilization and achieve service differentiation. Specifically, creating a slice for each service, we determine the service coverage (SC) of BSs for each slice and manage inter-slice and intra-slice interference to support slices with different signal-to-interference-plus-noise ratio (SINR) requirements. Our research objective is to maximize the network energy efficiency by determining the SC and downlink transmission power of BSs for all slices while satisfying their SINR requirements. We propose a RAN slicing framework and formulate an optimization problem based on the proposed framework. Then, we develop an approach to solve the problem for obtaining the optimal solution of SC management (SCM) and interference management (IM). The main contributions of this paper are as follows:

- We develop a novel RAN slicing framework with three designs for the spatial, temporal, and slice dimensions. The proposed grid-based planning and dual time-scale planning can adapt to spatiotemporal variations in data traffic, and the proposed flexible binary slice zooming can enhance the flexibility of service differentiation for satisfying different QoS requirements in a RAN.
- We propose an effective artificial intelligence (AI)-assisted approach to address the challenging RAN slicing problem. By integrating a deep unsupervised learning technique and an optimization-based analytical solution, the proposed approach can cope with the coupling between SCM and IM to balance the adaptivity and overhead of slicing-based resource management.

The remainder of this paper is organized as follows. Section II provides an overview of related studies. Section III describes the network scenario and proposed RAN slicing framework. Section IV presents the system model and problem formulation. Section V introduces the developed AI-assisted approach. Section VI presents the simulation results, followed by the conclusion in Section VII. A list of main symbols is given in Table I.

II. Related Work

Slicing-based resource management for core networks has attracted significant attention since 5G due to its advantage in service differentiation, while research on slicing-based resource management for RANs is still at a nascent stage [3]. Existing works on slicing-based resource management for RANs can be categorized as either a *single-stage* approach or a *two-stage* approach (i.e., with planning and operation stages as mentioned in Section I).

In single-stage approaches, a centralized controller, e.g., a software-defined networking (SDN) controller, is responsible for managing resources in a RAN for each individual user terminal (UT) in each slice [17], [19], [20], [21], [22], [23], [24]. In a single-BS scenario, Korrai et al. focused on the physical-layer RAN slicing and investigated customized physical-layer configurations for UTs in different slices [19], while Yang et al. concentrated on the data link layer and proposed a resource block (RB) scheduling scheme for UTs of enhanced mobile broadband (eMBB) and ultra-reliable and low latency communications (URLLC) slices to satisfy their different latency and reliability requirements [21]. In a multiple-BS scenario, authors in [20] proposed an orthogonal RB allocation scheme for UTs in different slices from the perspective of fairness in data rates of UTs. Moreover, with the consideration of inter-slice and intra-slice interference, a few works presented RB allocation schemes for UTs in different slices to improve their performance in terms of latency, data rate, and RB usage [17], [22], [23], [24]. While single-stage approaches can support service differentiation, their adaptivity is restricted owing to the lack of proactive resource reservation, which poses a challenge to QoS guarantee in highly dynamic network environments [25].

To tackle this problem, lots of researchers recently concentrate on two-stage approaches [7], [9]. Specifically, a centralized controller proactively reserves network resources for slices according to the service demand of each slice in a large time scale, i.e., planning window, whereas each slice allocates the reserved resources to individual UTs based on their real-time status in a short time scale, i.e., operation window. Compared with one-stage approaches, two-stage approaches are capable of achieving high adaptivity by proactively configuring slices and reserving resources in dynamic network environments and offer great flexibility due to having two time scales for different resource management decisions [7], [26]. Focusing on the planning stage, a few existing works investigated proactive resource reservation in RAN slicing by statistically modeling the service demand of each slice [26], [27], [28], [29], e.g., Poisson process-based data packet arrival. Considering vehicular networks with eMBB,

TABLE I
LIST OF MAIN SYMBOLS

Symbols	Definition	Symbols	Definition
$a_{m,n}$	The binary indicator indicating whether the SC of SBS m for slice n is full-size or reduced-size	$p^t_{i,n}$	The total downlink transmission power summarized over all RBs within grid i for slice n in time interval t
$b^t_{i,n,i',n'}$	The indicator indicating whether the downlink transmission to grid i for slice n is interfered by downlink transmission to grid i' for slice n' in time interval t	$\theta^t_{i,n}$	The probability that the downlink transmission to grid i for slice n is interfered by downlink transmission to grid i' for slice n' in time interval t
$E^t_{m,n}$	The energy consumption of slice n at BS m in time interval t	$P^t_{m,n}$	The total transmission power over all grids within the SC of BS m for slice n in time interval t
$h^t_{i,n}$	The average channel gain of downlink transmission of BS $m_{i,n}$ over all UTs within grid i in time interval t	$\mathcal{R}_m$	The set of grids within the ring-shaped area surrounding SBS m
$\mathcal{I}_{m,n}$	The set of grids within the SC of BS m for slice n	$w^t_{i,n}$	The amount of downlink data traffic loads of all UTs within grid i in slice n in time interval t
l^{f}_m, l^{r}_m	The full-size and reduced-size SC radiuses of SBS m, respectively	$\gamma^t_{i,n}$	The SINR of downlink transmission within grid i for slice n in time interval t
$l_{m,n}$	The SC of SBS m for slice n	$m_{i,n}$	The index of BS associated to grid i for slice n
$L_{\max}$	The maximum physical coverage radius of each SBS	Υ^{re}	The set of data records used for solution refinement
r	The diameter of each grid	Υ	The set of data records

URLLC, and massive machine-type communication (mMTC) services, the authors of [27] and [28] proposed two orthogonal radio resource reservation schemes, respectively. Taking into account inter-slice interference, some researchers presented spectrum slicing schemes, e.g., [26], [29], and the authors of [30] analyzed the trade-off between spectrum utilization and inter-slice interference. In addition, joint planning-stage and operation-stage radio resource slicing was studied in various network scenarios, including one-tier [31], [32], two-tier [33], and drone-based RANs [34], where radio resource reservation among slices in the planning stage was conducted based on AI-driven prediction [31], [32] or statistical modeling [33], [34] of the data traffic load in each slice. Existing research on two-stage approaches mainly concentrated on resource reservation for RANs, while SC management for multiple slices with different QoS requirements remains an open issue. Moreover, the existing two-stage approaches rely on coarse-grained information, such as aggregated data traffic within the SC area of a BS, which may degrade network resource utilization.

Different from the existing two-stage approaches, we propose a novel RAN slicing framework for both resource reservation and SC management in the planning stage. With joint resource reservation and SC management, we target fine-grained and flexible resource management for achieving service differentiation in spatiotemporally dynamic network environments.

III. NETWORK SCENARIO AND RAN SLICING FRAMEWORK

In this section, we introduce the considered network scenario and present the proposed RAN slicing framework.

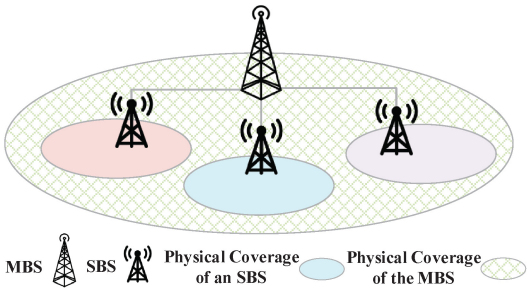
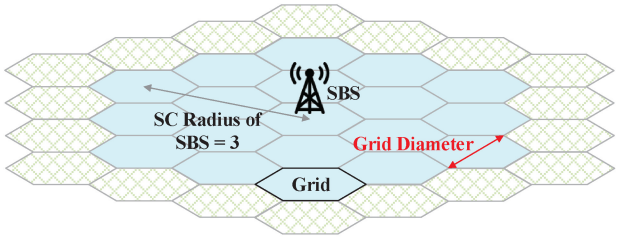


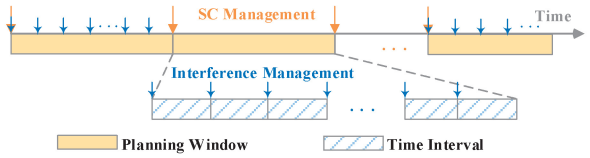
Fig. 1. The physical network scenario.

A. Network Scenario

Consider a two-tier RAN with one macro BS (MBS) in the first tier and M small BSs (SBSs) in the second tier. We show the physical network scenario of the considered two-tier RAN in Fig. 1. All the BSs use the same radio spectrum pool, and each BS orthogonally reserves RBs for downlink transmissions within its coverage area [35]. Using network slicing, N slices (corresponding to N services with different SINR requirements) are created on top of the physical network, and the radio spectrum resource of each BS is shared by all slices. For each slice, the MBS and all SBSs jointly support the corresponding service across the network to ensure that the service is accessible anywhere within the network coverage area. Meanwhile, given any slice, the SC of different SBSs (representing the spatial coverage of these SBSs for the corresponding service) are non-overlapping with each other for mitigating intra-slice interference. Any UT within the SC of an BS is associated to that BS, and each BS solely serves all UTs within its SC for the corresponding service. UTs not within the SC of any SBS are associated to the MBS. A centralized



(a) An illustration of grids and SC.



(b) An illustration of the time scales.

Fig. 2. Grid-based and dual time-scale planning (shown for one slice).

controller located at the MBS determines the SC and total transmission power for each slice at each BS in the planning stage, corresponding to SCM and IM. Then, in the subsequent operation stage, each BS allocates radio resources, such as RBs and transmission power, to individual UTs within its SC for downlink transmissions.

B. RAN Slicing Framework

For the considered scenario, we focus on the planning stage and propose a RAN slicing framework to achieve fine-grained and flexible SCM and IM for services with different SINR requirements. The proposed framework consists of three schemes: 1) grid-based planning in the spatial dimension; 2) dual-time scale planning in the temporal dimension; and 3) flexible binary slice zooming in the slice dimension.

1) Grid-Based Planning: To cope with the uneven spatial distribution of data traffic loads, we propose grid-based planning in the spatial dimension, where the illustration of grid-based planning for one slice is shown in Fig. 2(a). Specifically, the whole network coverage is divided into I hexagon areas, named *grids*, with an identical grid diameter, denoted by r.[1] We assume that each BS is at the center of a grid, and the SC radius of each SBS corresponds to the number of layers of grids within its SC. For each slice, the SC of the MBS includes all the grids that are not in the SC of any SBS. In the example shown in Fig. 2(a), the SC radius of the SBS is 3. The total downlink transmission power for each grid within the SC of a BS can be different.

The benefit of grid-based planning is two-fold. First, the downlink transmissions for UTs within different grids may experience different interference. Customizing the total transmission power for each grid can help mitigate inter-slice and intra-slice interference and thus improve network energy efficiency. Second, adjusting the SC of each SBS in the units of grids is beneficial for balancing data traffic loads among BSs in a fine-grained manner.

[1]In addition to hexagons, some other shapes of grids are also applicable to the proposed grid-based planning.

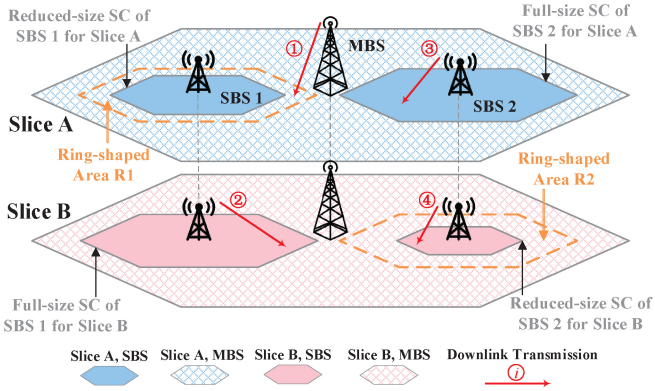


Fig. 3. Flexible binary slice zooming (shown for two slices).

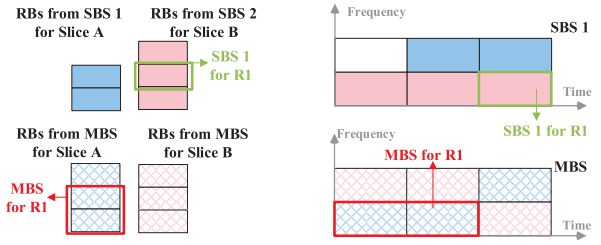
2) Dual Time-Scale Planning: To adapt to the temporal variations of data traffic loads, we propose the scheme of dual time-scale planning, where the illustration of dual time-scale planning for one slice is shown in Fig. 2(b). Each planning window is divided into T $(T > 1)$ *time intervals* with uniform length. The SC of the SBSs is updated at the beginning of each planning window and remains constant till the beginning of the next planning window. By contrast, the total downlink transmission power of the BSs for individual grids is updated at the beginning of each time interval in the planning window if needed.

Dual time-scale planning provides great flexibility in differentiating the time scales of SCM and IM based on the difference in the amount of resource management overhead, i.e., signaling overhead and computation complexity. First, the short planning window for SCM leads to frequent UT association changing during network operations and thus high signaling overhead. Second, SCM has a higher computation complexity than IM since adjusting the SC of BSs results in the update of total transmission power of BSs for individual grids. Dual-time scale planning helps properly balance the adaptivity of resource management and the resource management overhead.

3) Flexible Binary Slice Zooming: To provide the flexibility in service differentiation, we propose a novel scheme called *flexible binary slice zooming* in the slice dimension, including the following two elements. The proposed scheme for two slices is illustrated in Fig. 3.

- *Differentiated IM and SCM across slices:* For IM, the transmission power reserved by each BS for each grid can be different across slices. For SCM, the SC of each SBS can also be different across slices. Specifically, the SC of each SBS for any slice is binary, i.e., either full-size or reduced-size shown in Fig. 3, neither of which can exceed the maximum physical coverage area of the SBS. We refer to the gap between the full-size and the reduced-size SC of each SBS as a *ring-shaped area* surrounding the SBS.[2]
- *Partially non-orthogonal RB reservation among BSs:* All the BSs use the same radio spectrum pool except in the

[2]All SC of an SBS may be identical, i.e., either all SC is reduced-size or all SC is full-size. In this case, there is no ring-shaped area surrounding the SBS.



(a) Virtual slice separation in the planning stage. (b) RB allocation in the operation stage.

Fig. 4. Virtual slice separation at SBS 1.

following case: each SBS and the MBS reserve different sets of RBs for downlink transmissions within the ring-shaped area surrounding the SBS if such an area exists. The partially non-orthogonal RB reservation avoids the interference between the downlink transmissions of each SBS and the MBS within the ring-shaped areas. We highlight the partially non-orthogonal RB reservation for downlink transmissions using red arrows in Fig. 3, which is explained in Section IV-A.

The proposed flexible binary slice zooming has two benefits in facilitating service differentiation in RANs. First, differentiating the downlink transmission power for different slices achieves fine-grained IM in the slice dimension, and thus helps satisfy the diverse and stringent SINR requirements of slices. Second, by customizing the SC of each BS for different slices, the proposed flexible binary slice zooming scheme is more flexible in adapting to the different spatial distributions of data traffic loads than conventional cell-based SCM that uses identical SC for all services.

Based on the aforementioned three schemes, the proposed RAN slicing framework provides great flexibility in enabling isolated SCM and IM for multiple slices, and improves granularity and adaptivity in adapting to the spatiotemporal variations of data traffic loads in RANs.

C. Operation Stage Consideration

The real-time allocation of RBs for individual UTs in the operation stage impacts the interference and thus the SINR of each UT. As a result, making decisions on SCM and IM with the consideration of operation-stage RB allocation is necessary. However, it is impossible to know the future UT-level information, e.g., the locations and data traffic loads of UTs, at the beginning of each planning window, and how RBs will be allocated to individual UTs in the subsequent planning window. To overcome this issue, we adopt *virtual slice separation* to reserve RBs for multiple slices in the planning stage with RB multiplexing in the operation stage. Specifically, only the number of RBs reserved for each slice is determined in the planning stage rather than the specific set of RBs. An example of virtual slice separation is shown in Fig. 4. Fig. 4(a) shows the numbers of RBs reserved to slices A and B by virtual slice separation, and Fig. 4(b) shows the specific sets of RBs that can be flexibly allocated to individual UTs in the operation stage based on the real-time network environment.

IV. System Model and Problem Formulation

In this section, we first present the system model for SCM and IM based on the proposed RAN slicing framework. Then, we formulate an optimization problem to maximize the network energy efficiency.

Denote the set of BSs by $\mathcal{M} = \{0, 1, \ldots, M\}$, and let $m = 0$ and $m \in \mathcal{M}\backslash\{0\}$ be the indexes of the MBS and M SBSs, respectively. Define the sets of slices, grids, and time intervals as $\mathcal{N} = \{1, 2, \ldots, N\}$, $\mathcal{I} = \{1, 2, \ldots, I\}$, and $\mathcal{T} = \{1, 2, \ldots, T\}$, respectively.

A. Model of SCM

We model the SC of BSs in the proposed RAN slicing framework. Denote the SC radius of SBS m for slice n by $l_{m,n}$. We assume that the maximum physical coverage radius of all SBSs are identical, denoted by $L_{\max}$, and define the set of possible SC radius for any slice as $\mathcal{L} = \{1, 2, \ldots, L_{\max}\}$. With flexible binary slice zooming, we determine the full-size or reduced-size SC radiuses of SBS m, denoted by $l_m^{\mathrm{f}} \in \mathcal{L}$ and $l_m^{\mathrm{r}} \in \mathcal{L}$, respectively, where $l_m^{\mathrm{f}} \geq l_m^{\mathrm{r}}$. To indicate whether the SC of SBS m for slice n is full-size or reduced-size, we introduce a binary variable $a_{m,n} \in \{0, 1\}$. Accordingly, $\mathbf{l}^{\mathrm{f}} = [l_m^{\mathrm{f}}]_{\forall m \in \mathcal{M}\backslash\{0\}}$, $\mathbf{l}^{\mathrm{r}} = [l_m^{\mathrm{r}}]_{\forall m \in \mathcal{M}\backslash\{0\}}$, and $\mathbf{a} = [a_{m,n}]_{\forall m \in \mathcal{M}\backslash\{0\}, n \in \mathcal{N}}$ are the variables that determine the SC of BSs during a planning window. The SC radius of SBS $m \in \mathcal{M}\backslash\{0\}$ for slice $n \in \mathcal{N}$ is represented as follows:

$$l_{m,n} = \begin{cases} l_m^{\mathrm{f}}, & \text{if } a_{m,n} = 1; \\ l_m^{\mathrm{r}}, & \text{if } a_{m,n} = 0. \end{cases} \tag{1}$$

Let $d_{m,i}$ denote the distance between BS m and the center of grid i. We define the set of grids within the SC of SBS m for slice n and the set of grids within the ring-shaped area surrounding SBS m as $\mathcal{I}_{m,n} = \{i | d_{m,i} \leq l_{m,n}, i \in \mathcal{I}\}$ and $\mathcal{R}_m = \{i | l_m^{\mathrm{r}} \leq d_{m,i} \leq l_m^{\mathrm{f}}, i \in \mathcal{I}\}$, respectively. We define the set of grids within the SC of the MBS for slice n as $\mathcal{I}_{0,n} = \{i | i \in \mathcal{I}\backslash\mathcal{I}_{m,n}, m \in \mathcal{M}\backslash\{0\}\}$, i.e., for any slice, grids that are not within the SC of any SBS are covered by the MBS.

SCM should consider the spatial distribution of downlink data traffic loads. Denote the amount of downlink data traffic loads (in bits) of all UTs within grid i in time interval t for slice n by $w_{i,n}^t$. Let vectors $\mathbf{w}_n^t = [w_{i,n}^t]_{\forall i \in \mathcal{I}}$ and $\mathbf{W} = [w_{i,n}^t]_{\forall i \in \mathcal{I}, n \in \mathcal{N}, t \in \mathcal{T}}$ be the data traffic distribution (DTD) of slice n in time interval t and the DTD vector of all slices over a planning window, respectively. The DTD vector $\mathbf{W}$ is assumed to be known *a prior* through prediction [36]. The required number of RBs for each BS depends on the data traffic load within the SC of the BS. Let η_n represent the average number of RBs required to support each bit of the downlink data traffic in slice n.[3] To ensure that the number of RBs reserved for the downlink data traffic within the SC of any BS does not exceed the total number of RBs of each BS during a planning window, denoted by C, the condition

[3]The value of η_n can be estimated according to the data rate requirement of slice n and the long-term performance of RB scheduling in the operation stage [37].

$$\sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{I}_{m,n}}\eta_n w_{i,n}^t \le C,\ \forall m\in\mathcal{M}\backslash\{0\}, \tag{2}$$

for each SBS and the condition

$$\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{M}\backslash\{0\}}\sum_{n\in\mathcal{N}}\eta_n\left(\sum_{i\in\mathcal{I}}w_{i,n}^t-\sum_{i\in\mathcal{I}_{m,n}}w_{i,n}^t\right)\le C, \tag{3}$$

for the MBS should be satisfied in SCM.

SCM affects the interference among downlink transmissions of different BSs due to frequency reuse. As the result of partially non-orthogonal RB reservation in flexible binary slice zooming, there are two cases in which downlink transmissions do not interfere with each other: i) between downlink transmissions of the same BS, e.g., communication links 3 and 4 in Fig. 3 (shown as the red arrows with circled numbers 3 and 4 in the figure); and ii) between downlink transmissions of the SBSs and the MBS within a ring-shaped area, e.g., communication links 1 and 2 for the two UTs in the ring-shaped area R1 in Fig. 3 (shown as the red arrows with circled numbers 1 and 2). To achieve non-orthogonal RB reservation as mentioned in Section III-B, the following condition must be satisfied in SCM:

$$\sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{R}_m}w_{i,n}^t\eta_n\le C,\ \forall m\in\mathcal{M}\backslash\{0\}. \tag{4}$$

Constraint (4) ensures that, if a ring-shaped area surrounding an SBS exists, the SBS and the MBS have sufficient RBs for orthogonal RB reservation in the ring-shaped area.

Other than the aforementioned two cases, there exists the interference between downlink transmissions of different BSs. We introduce term $b_{i,n,i',n'}^t\in\{0,1\}$ to indicate whether the downlink transmission to grid i for slice n is interfered by the downlink transmission to grid i' for slice n' in time interval t or not, given by:

$$b_{i,n,i',n'}^t=\begin{cases}0,\ \text{if } i\in\mathcal{I}_{m,n}, i'\in\mathcal{R}_m, a_{m,n}=1, a_{m,n'}=0;\\ 0,\ \text{if } i\in\mathcal{R}_m, i'\in\mathcal{I}_{m,n'}, a_{m,n}=0, a_{m,n'}=1;\\ 0,\ \text{if } m_{i,n}=m_{i',n'}, \forall m_{i,n}, m_{i',n'}\in\mathcal{M};\\ 1,\ \text{otherwise},\end{cases} \tag{5}$$

where $m\in\mathcal{M}\backslash\{0\}$, and $m_{i,n}=\{m|i\in\mathcal{I}_{m,n}, m\in\mathcal{M}\}$ denotes the BS that covers grid i in its SC of slice n. The first and second cases in (5) represent no interference between the downlink transmission of the SBSs and the MBS within the ring-shaped areas. The third case in (5) represents no interference between the downlink transmissions within grid i for slice n and that within gird i' for slice n' if they are from the same BS, i.e., $m_{i,n}=m_{i',n'}$. Otherwise, the downlink transmission within a grid interferes with the downlink transmission within other grids.

B. Model of IM

We model IM based on virtual slice separation mentioned in Section III-C. Denote the total downlink transmission power summarized over all RBs reserved for downlink transmissions in grid i for slice n in time interval t by $p_{i,n}^t$. Define the IM decision in time interval t and in a planning window as $\mathbf{p}^t=[p_{i,n}^t]_{\forall i\in\mathcal{I},n\in\mathcal{N}}$ and $\mathbf{p}=[p_{i,n}^t]_{\forall i\in\mathcal{I},n\in\mathcal{N},t\in\mathcal{T}}$, respectively.

We assume that the maximum downlink transmission power of SBSs are the same, and denote the maximum downlink transmission power of the SBSs and the MBS by p_{SBS} and p_{MBS}, respectively. The following constraint should be satisfied in IM to ensure that the total downlink transmission power of each BS over all slices cannot exceed the maximum downlink transmission power of the BS:

$$\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{I}_{m,n}}p_{i,n}^t\le\begin{cases}p_{\mathrm{MBS}}, & m=0;\\ p_{\mathrm{SBS}}, & m\in\mathcal{M}\backslash\{0\}.\end{cases} \tag{6}$$

Next, we model the interference between downlink transmissions of different BSs. The exact interference depends on real-time RB scheduling during the operation stage and is unknown *a priori* in the planning stage. We define parameter $\theta_{i,n,i',n'}^t\in[0,1]$ to represent the likeliness that the downlink transmission to grid i for slice n is interfered by the downlink transmission to grid i' for slice n' in time interval t and model the planning-stage interference statistically [30], [38].[4] The SCM decisions of the BSs covering grid i and i' for slice n and n' can affect the data traffic loads of the BSs and thus the value of $\theta_{i,n,i',n'}^t$. Given $\theta_{i,n,i',n'}^t$, the total interference to the downlink transmission of BS $m_{i,n}$ to grid i, denoted by $I_{i,n}^t$, is expressed as follows:

$$I_{i,n}^t=\sum_{n'\in\mathcal{N}}\sum_{i'\in\mathcal{I}}b_{i,n,i',n'}^t\theta_{i,n,i',n'}^t p_{i',n'}^t h_{i,n,i',n'}^t \tag{7}$$

where $h_{i,n,i',n'}^t$ denotes the average channel gain of the downlink transmission of BS $m_{i',n'}$ to grid i for slice n in time interval t.

The SINR of the downlink transmission of BS $m_{i,n}$ to grid i for slice n in time interval t, denoted by $\gamma_{i,n}^t$, can be modeled as follows:

$$\gamma_{i,n}^t=\frac{\bar{p}_{i,n}^t h_{i,n,i,n}^t}{N_0+I_{i,n}^t},\quad\forall i\in\mathcal{I},n\in\mathcal{N},t\in\mathcal{T}, \tag{8}$$

where $\bar{p}_{i,n}^t=p_{i,n}^t/(w_{i,n}^t\eta_n)$ represents the average transmission power on a single RB for downlink transmissions in slice n within grid i in time interval t, and N_0 denotes the noise power. IM should satisfy the SINR requirement of each slice, as follows:

$$\gamma_{i,n}^t\ge\rho\gamma_n^{\min},\quad\forall i\in\mathcal{I}_{m,n}, \tag{9}$$

where $\gamma_n^{\min}$ denotes the minimum SINR required by slice n, and ρ is a constant used for flexibly scaling the minimum required SINR level [16].[5]

[4]The value of parameter $\theta_{i,n,i',n'}^t$ can be obtained empirically when the DTD **W**, RB scheduling policy in the operation stage, and RB reservation policy in the planning stage are given [26].

[5]The SINR in the planning stage, i.e., $\gamma_{i,n}^t$ is a reference value over the duration of a time interval, which may not represent the exact SINR level in the operation stage. Thus, we allow a feedback mechanism to change the SINR requirements of slices in the planning stage by adjusting weight ρ based on the real-time power control, RB allocation, and instantaneous SINR in the operation stage.

C. Problem Formulation

In this subsection, we formulate an energy efficiency maximization problem based on the proposed RAN slicing framework. Denote the energy consumption of BS m for serving slice n in time interval t by $E^t_{m,n}$, given by:

$$E^t_{m,n} = \tau P^t_{m,n}, \quad \forall m \in \mathcal{M}, n \in \mathcal{N}, t \in \mathcal{T}, \tag{10}$$

where τ denotes the duration of each time interval. The energy efficiency (measured in the unit of bit/RB/J) of all BSs for serving slice n during a planning window, denoted by ξ_n, is as follows:

$$\xi_n = \frac{w_n}{E_n C_n}, \ \forall n \in \mathcal{N}, \tag{11}$$

where $E_n = \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{M}} E^t_{m,n}$ is the total energy consumption of all BSs in all time intervals of a planning window, $w_n = \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{M}} \sum_{i\in\mathcal{I}_{m,n}} w^t_{i,n}$ represents the total downlink traffic data loads in the planning window, and $C_n = \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{M}} \sum_{i\in\mathcal{I}_{m,n}} w^t_{i,n}\eta_n$ is the total number of RBs reserved in the planning window.

The slicing-based resource management problem with the objective of network energy efficiency maximization is formulated as follows:

$$\text{P1:} \max_{\{\mathbf{p},\mathbf{l}^{\mathrm{f}},\mathbf{l}^{\mathrm{r}},\mathbf{a}\}} \sum_{n\in\mathcal{N}} \lambda_n \xi_n \tag{12a}$$

$$\text{s.t.} \ \ (2),\ (3),\ (4),\ (6),\ (9), \tag{12b}$$

$$D_{m,m'} \geq l^{\mathrm{f}}_m + l^{\mathrm{f}}_{m'}, \ \ \forall m \neq m', \ m, m' \in \mathcal{M}\backslash\{0\}, \tag{12c}$$

$$p^t_{i,n} > 0, \ \ \forall p^t_{i,n} \in \mathbb{R}, \tag{12d}$$

$$l^{\mathrm{f}}_m \geq l^{\mathrm{r}}_m, \ \ \forall l^{\mathrm{r}}_m, l^{\mathrm{f}}_m \in \mathcal{L}, m \in \mathcal{M}\backslash\{0\}, \tag{12e}$$

$$a_{m,n} \in \{0,1\}, \ \ \forall m \in \mathcal{M}, n \in \mathcal{N}, \tag{12f}$$

where λ_n denotes the weight for balancing the energy efficiency for different slices. In Problem P1, the optimization variables include IM decision $\mathbf{p}$ and SCM decisions $\mathbf{l}^{\mathrm{f}}$, $\mathbf{l}^{\mathrm{r}}$, and $\mathbf{a}$. Constraint (12c) ensures that the SC of SBSs does not overlap, in which term $D_{m,m'}$ denotes the physical distance between SBSs m and m'. Constraint (12d) guarantees that the downlink transmission power is positive. Constraints (12e) and (12f) ensure that the selection of the SC of each SBS for each slice is binary and does not exceed the maximum physical coverage of the SBS. Problem P1 is a combinatorial optimization problem, which is difficult to be solved by conventional optimization methods due to two reasons [39]. First, a large number of variables need to be determined. Specifically, the variables for transmission power and SCM are with the dimensions of $N \times I \times T$ and $N \times M$, respectively. Second, the transmission power and SCM decisions are coupled. To solve this problem, we propose an unsupervised-learning-assisted solution in the next section.

V. Unsupervised-Learning-Assisted Solution

We decouple Problem P1 into two sub-problems and solve them in two steps. In the first step, we design an unsupervised-learning-assisted approach to determine the SC of the SBSs. In the second step, given a solution to the SCM sub-problem, we derive the closed-form solution to the IM sub-problem in each time interval. We first discuss the solution to IM in Section V-A, followed by the solution to SCM in Sections V-B and V-C.

A. Optimal Solution of IM

Given the settings of the SC of all SBSs, i.e., $\mathbf{l}_{\mathrm{f}}, \mathbf{l}_{\mathrm{r}}, \mathbf{a}$, we formulate the problem of IM in time interval t as follows:

$$\text{P2:} \max_{\{\mathbf{p}^t\}} \sum_{n\in\mathcal{N}} \lambda_n \xi_n \tag{13a}$$

$$\text{s.t.} \ \ (6),\ (9),\ (12d). \tag{13b}$$

The solution of $\mathbf{p}^t$ in Problem P2 depends on the DTDs of all slices in time interval t. In Theorem 1, we provide the closed-form optimal solution of $\mathbf{p}^t$ in time interval t. Theorem 1 can be applied to all time intervals of a planning window since IM in different time intervals is independent.

Theorem 1: Define $\delta^t_{i,n,i',n'} = b^t_{i,n,i',n'} \theta^t_{i,n,i',n'} h^t_{i,n,i',n'}$. The optimal solution to Problem P2, i.e., $\mathbf{p}^t_*$, is given by (14), as shown at bottom of the next page, where

$$\mathbf{\Omega}^t_{n,n'} = \begin{bmatrix} \delta^t_{1,n,1,n'} & \cdots & \delta^t_{1,n,i',n'} & \cdots & \delta^t_{1,n,I,n'} \\ \vdots & & \vdots & & \vdots \\ \delta^t_{i,n,1,n'} & \cdots & \delta^t_{i,n,i',n'} & \cdots & \delta^t_{i,n,I,n'} \\ \vdots & & \vdots & & \vdots \\ \delta^t_{I,n,1,n'} & \cdots & \delta^t_{I,n,i',n'} & \cdots & \delta^t_{I,n,I,n'} \end{bmatrix}_{I\times I}, \tag{15}$$

and

$$\hat{\mathbf{H}}^t = \mathrm{diag}\left(\frac{h^t_{1,1}}{w^t_{1,1}\eta_1}, \ldots, \frac{h^t_{i,n}}{w^t_{i,n}\eta_n}, \ldots, \frac{h^t_{I,N}}{w^t_{I,N}\eta_N} \right). \tag{16}$$

Proof: See the Appendix. ■

B. Local Optimum SC Search

Given the solution to IM, determining the SC of all BSs for all slices in Problem P1 remains a combinatorial optimization problem. To solve this problem, we propose an unsupervised-learning-assisted approach. The basic idea is to first iteratively find a locally optimal solution to SCM and then use a deep unsupervised learning technique to refine the locally optimal solution obtained by the iterative algorithm. We detail the designed iterative algorithm and the unsupervised-learning-assisted algorithm in this subsection and Section V-C, respectively.

We present the local optimum SC search (LOSCS) algorithm, which iteratively updates the SC of each SBS, searching one SBS at a time, until no further energy efficiency improvement can be achieved by updating the SC of any SBS. Denote the objective function in Problem P1 and the value of the objective function by $\Delta(\mathbf{l}^{\mathrm{f}}, \mathbf{l}^{\mathrm{r}}, \mathbf{a}, \mathbf{p})$ and Δ, respectively. The algorithm is detailed in Algorithm 1. Let set $\mathcal{M}^{\mathrm{s}} \subset \mathcal{M}$ include the SBSs that have not been involved in the iterative search yet. Line 2 initializes set $\mathcal{M}^{\mathrm{s}} = \mathcal{M}\backslash\{0\}$ and the SC of all SBSs, i.e., $\mathbf{l}^{\mathrm{f}}$, $\mathbf{l}^{\mathrm{r}}$, and $\mathbf{a}$, and randomly selects an SBS, i.e., SBS m, to start searching. Given the initialized SC of SBSs, line 3 and line 4 obtain the optimal solution of IM, i.e., $\mathbf{p}$, and the

Algorithm 1: LOSCS Algorithm

1 **Input:** $\mathbf{W}$
2 **Initialize:** Randomly select $m \in \mathcal{M}\backslash\{0\}$, and set $\mathcal{M}^{\mathrm{s}} = \mathcal{M}\backslash\{0\}$, $\mathbf{l}^{\mathrm{f}}$, $\mathbf{l}^{\mathrm{r}}$, $\mathbf{a}$;
3 Obtain $\mathbf{p}$ by Theorem 1 given $\mathbf{W}$, $\mathbf{l}^{\mathrm{f}}$, $\mathbf{l}^{\mathrm{r}}$, and $\mathbf{a}$;
4 Calculate $\Delta(\mathbf{l}^{\mathrm{f}}, \mathbf{l}^{\mathrm{r}}, \mathbf{a}, \mathbf{p})$ given $\mathbf{W}$;
5 **while** $\mathcal{M}^{s} \neq \emptyset$ **do**
6 **for** $\hat{\mathbf{l}}_m \in \mathcal{S}_m$ **do**
7 Obtain $\hat{\mathbf{l}}^{\mathrm{f}}$, $\hat{\mathbf{l}}^{\mathrm{r}}$, $\hat{\mathbf{a}}$ by updating the SC of SBS m with $\hat{\mathbf{l}}_m$;
8 **if** *Constraints (2), (3), (4) are not satisfied* **then**
9 **Continue**;
10 **else**
11 Obtain $\hat{\mathbf{p}}$ by Theorem 1 given $\mathbf{W}$, $\hat{\mathbf{l}}^{\mathrm{f}}$, $\hat{\mathbf{l}}^{\mathrm{r}}$, and $\hat{\mathbf{a}}$;
12 Calculate $\Delta'(\hat{\mathbf{l}}^{\mathrm{f}}, \hat{\mathbf{l}}^{\mathrm{r}}, \hat{\mathbf{a}}, \hat{\mathbf{p}})$ given $\mathbf{W}$;
13 **if** $\Delta' > \Delta$ **then**
14 $\Delta \leftarrow \Delta'$;
15 $\mathbf{l}_{\mathrm{f}}, \mathbf{l}_{\mathrm{r}}, \mathbf{a}, \mathbf{p} \leftarrow \hat{\mathbf{l}}_{\mathrm{f}}, \hat{\mathbf{l}}_{\mathrm{r}}, \hat{\mathbf{a}}, \hat{\mathbf{p}}$;
16 $\mathcal{M}^{\mathrm{s}} \leftarrow \mathcal{M}\backslash\{0\}$;
17 **else**
18 **Continue**;
19 **end**
20 **end**
21 **end**
22 $\mathcal{M}^{\mathrm{s}} \leftarrow \mathcal{M}^{\mathrm{s}}\backslash\{m\}$;
23 Randomly select $m \in \mathcal{M}^{\mathrm{s}}$;
24 **end**
25 **Output:** $\mathbf{l}_{\mathrm{f}}$, $\mathbf{l}_{\mathrm{r}}$, $\mathbf{a}$, $\mathbf{p}$, and Δ

corresponding value of the objective function in Problem P1, i.e., Δ. Line 5 to Line 21 search SCM solution for an SBS, corresponding to one iteration. Denote the SC of SBS m for the slices by vector $\mathbf{l}_m = [l_{m,n}]_{\forall n \in \mathcal{N}}$ which can be obtained by (1). We introduce $\mathcal{S}_m$ to represent the set that includes all possible combinations of the SC of SBS m for all slices, i.e., all possible values of vector $\mathbf{l}_m$ when they satisfy constraints (12c) and (12e). During each iteration, we only search the SC of SBS m for all slices from set $\mathcal{S}_m$ while keeping the SC of other SBSs fixed. If an SC combination yielding a larger value of Δ, is found, the currently best SCM solution is updated, and set $\mathcal{M}^{\mathrm{s}}$ will be reset to the set of all SBSs; Otherwise, no change will be made. At the end of an iteration, another SBS is randomly selected from set $\mathcal{M}^{\mathrm{s}}$ for the next iteration, and the set $\mathcal{M}^{\mathrm{s}}$ is updated. All iterations stop if the set $\mathcal{M}^{\mathrm{s}}$ is an empty set, which means that a solution with a larger value of Δ cannot be found by adjusting the SC of any SBS. The output of Algorithm 1 is an SCM solution with the corresponding optimal solution of IM given by (14).

The computation complexity of Algorithm 1 is $\mathcal{O}((L_{\max}^2 - L_{\max})^N 2^{(M-1)N} I^3 N^3 T)$, where the computation complexity of IM in each time interval is $\mathcal{O}(I^3 N^3)$, and the computation complexity of SCM in each planning window is $\mathcal{O}((L_{\max}^2 - L_{\max})^N 2^{(M-1)N})$. Since the performance of the SCM solution found by Algorithm 1 depends on the initial settings, we design an unsupervised-learning-assisted SC search (ULSCS) algorithm next to reduce the computation complexity of planning-stage resource management while enhancing the performance of the LOSCS algorithm by finding proper initial settings.

C. Unsupervised-Learning-Assisted SC Search

In each planning window, the SCM solution is related to the spatiotemporal service demands of all slices. The amount of downlink data traffic in each grid is continuous, whereas variables of SCM are discrete. As a result, similar $\mathbf{W}$ in different planning windows may lead to the same optimal SCM solution. Thus, we propose a data-driven approach to utilize historical solutions for refining the SCM solution obtained by Algorithm 1 in each planning window. The proposed approach consists of two components: feature extraction and solution refinement. First, we leverage an auto-encoder, a deep unsupervised learning technique, to extract the implicit and low-dimensional features of $\mathbf{W}$ in a planning window. Second, by comparing the extracted features of $\mathbf{W}$ in the historical and the subsequent planning window, we select some historical solutions to use as the initial settings of Algorithm 1. The network energy efficiency, i.e., Δ, is non-decreasing over the iterations of Algorithm 1. As a result, choosing a historical SCM solution as the initial settings results in a relatively high performance compared to Algorithm 1, and the worst-case network energy efficiency equals that obtained by Algorithm 1.

1) Feature Extraction: Considering that the value of $\mathbf{W}$ may vary across planning windows, we name the matrix $\mathbf{W}$ in a planning window as a *DTD instance*. The selection of a solution from a historical planning window is based on whether or not the DTD instance in the historical planning window is similar to that in the upcoming planning window. However, due to the high dimensionality of DTD instances, comparing every element in the two DTD instances is time-consuming. Therefore, reducing the dimensionality of DTD instances while retaining their essential information is important to the comparison. We utilize the deep auto-encoder technique to obtain a low-dimension representation

$$\mathbf{p}_*^t = \rho \left(\hat{\mathbf{H}}^t - \rho \begin{bmatrix} \gamma_1^{\min}\mathbf{\Omega}_{1,1}^t & \cdots & \gamma_1^{\min}\mathbf{\Omega}_{1,n'}^t & \cdots & \gamma_1^{\min}\mathbf{\Omega}_{1,N}^t \\ \vdots & \ddots & \cdots & \ddots & \vdots \\ \gamma_n^{\min}\mathbf{\Omega}_{n,1}^t & \cdots & \gamma_n^{\min}\mathbf{\Omega}_{n,n'}^t & \cdots & \gamma_n^{\min}\mathbf{\Omega}_{n,N}^t \\ \vdots & \ddots & \cdots & \ddots & \vdots \\ \gamma_N^{\min}\mathbf{\Omega}_{N,1}^t & \cdots & \gamma_N^{\min}\mathbf{\Omega}_{N,n'}^t & \cdots & \gamma_N^{\min}\mathbf{\Omega}_{N,N}^t \end{bmatrix}_{IN \times IN} \right)^{-1} \begin{bmatrix} \gamma_1^{\min} N_0 \\ \vdots \\ \gamma_n^{\min} N_0 \\ \vdots \\ \gamma_N^{\min} N_0 \end{bmatrix}_{IN \times 1}, \tag{14}$$

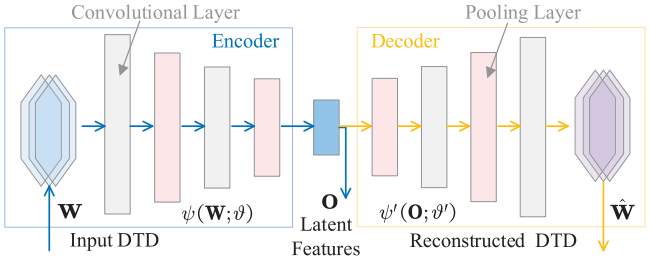


Fig. 5. The designed DNN architecture of the auto-encoder.

of a DTD instance, named *latent features*. Fig. 5 shows our design of deep neural networks (DNNs) for implementing the auto-encoder. The DNNs include two main parts: an encoder and a decoder. The encoder is a non-linear mapping function from a high dimensional space to a low dimensional space, i.e., extracting latent features from a DTD instance, and the decoder is a non-linear mapping function from a low dimensional space to a high dimensional space, i.e., reconstructing a DTD instance based on the latent features. Both parts are implemented by DNNs, and the DNN architecture of the decoder mirrors that of the encoder. In the training phase, the DNNs of both the encoder and the decoder are trained with the goal of minimizing the difference between the input and the reconstructed DTD instances. In the inference phase, only the DNN of the encoder is used for feature extraction [40].

Denote the extracted latent features from a DTD instance by $\mathbf{O}$, and the sets of all possible values of $\mathbf{W}$ and $\mathbf{O}$ by $\mathcal{W}$ and $\mathcal{O}$, respectively. We define the encoder as the function $\psi : \mathcal{W} \to \mathcal{O}$ and the decoder as the function $\psi' : \mathcal{O} \to \mathcal{W}$. Let vectors $\boldsymbol{\vartheta}$ and $\boldsymbol{\vartheta}'$ denote the parameters of DNNs of the encoder and the decoder, respectively. According to the designed DNN architecture for the auto-encoder, $\mathbf{W}$ and $\mathbf{O}$ satisfy the following relations: $\mathbf{O} = \psi(\mathbf{W};\boldsymbol{\vartheta})$ and $\hat{\mathbf{W}} = \psi'(\mathbf{O};\boldsymbol{\vartheta}')$, where $\hat{\mathbf{W}}$ denotes the reconstructed DTD instance from the latent features $\mathbf{O}$. To extract the latent features without neglecting useful information, the input and the reconstructed DTD instances should be as similar as possible. Therefore, the optimal values of parameters $\boldsymbol{\vartheta}$ and $\boldsymbol{\vartheta}'$, denoted by $\boldsymbol{\vartheta}_*$ and $\boldsymbol{\vartheta}'_*$, are obtained by the following equation:

$$\begin{aligned}\{\boldsymbol{\vartheta}_*, \boldsymbol{\vartheta}'_*\} &= \arg\min_{\{\boldsymbol{\vartheta},\boldsymbol{\vartheta}'\}} F\left(\mathbf{W}, \hat{\mathbf{W}}\right) \\ &= \arg\min_{\{\boldsymbol{\vartheta},\boldsymbol{\vartheta}'\}} F\left(\mathbf{W}, \psi'\left(\mathbf{O};\boldsymbol{\vartheta}'\right)\right) \\ &= \arg\min_{\{\boldsymbol{\vartheta},\boldsymbol{\vartheta}'\}} F\left(\mathbf{W}, \psi'\left(\psi(\mathbf{W};\boldsymbol{\vartheta});\boldsymbol{\vartheta}'\right)\right), \quad (17)\end{aligned}$$

where $F(\mathbf{W}, \hat{\mathbf{W}})$ is the cross-entropy loss function [40]. The optimal values of parameters, i.e., $\boldsymbol{\vartheta}_*$ and $\boldsymbol{\vartheta}'_*$ are obtained by using the gradient descent method to minimize the loss function $F(\mathbf{W}, \hat{\mathbf{W}})$. The data regarding DTD instances in the set Υ are utilized to train the DNNs and obtain the optimal parameters offline.

2) Solution Refinement: Using the extracted latent features of DTD instances, we define the similarity of two DTD instances in different planning windows, i.e., $\mathbf{W}$ and $\mathbf{W}'$, as follows:

Algorithm 2: ULSCS Algorithm

1 **Input:** $\boldsymbol{\vartheta}_*$, $\mathbf{W}$, $|\Upsilon^{\text{re}}|$, and Υ
2 Calculate the similarity between $\mathbf{W}$ and each DTD instance, i.e., $\mathbf{W}'$, contained in Υ by (18);
3 $\Upsilon^{\text{re}} \leftarrow$ Select data records containing the $|\Upsilon^{\text{re}}|$ most similar DTD instances from set Υ;
4 Obtain Δ, $\mathbf{l}_{\text{f}}$, $\mathbf{l}_{\text{r}}$, $\mathbf{a}$, $\mathbf{p}$ by Algorithm 1, given $\mathbf{W}$;
5 **for** $\upsilon \in \Upsilon^{\text{re}}$ **do**
6 Obtain $\mathbf{l}^{\text{re}}_{\text{f}}$, $\mathbf{l}^{\text{re}}_{\text{r}}$, $\mathbf{a}^{\text{re}}$ from data record υ;
7 Obtain Δ', $\mathbf{l}'_{\text{f}}$, $\mathbf{l}'_{\text{r}}$, $\mathbf{a}'$, $\mathbf{p}'$ by Algorithm 1 given $\mathbf{W}$ and the initial settings of $\mathbf{l}^{\text{re}}_{\text{f}}$, $\mathbf{l}^{\text{re}}_{\text{r}}$, and $\mathbf{a}^{\text{re}}$;
8 **if** $\Delta' > \Delta$ **then**
9 Δ, $\mathbf{l}_{\text{f}}$, $\mathbf{l}_{\text{r}}$, $\mathbf{a}$, $\mathbf{p} \leftarrow \Delta'$, $\mathbf{l}'_{\text{f}}$, $\mathbf{l}'_{\text{r}}$, $\mathbf{a}'$, $\mathbf{p}'$;
10 **else**
11 **Continue**;
12 **end**
13 **end**
14 Create a data record υ' containing $\mathbf{W}$, $\mathbf{l}_{\text{f}}$, $\mathbf{l}_{\text{r}}$, $\mathbf{a}$, and $\mathbf{p}$;
15 Add υ' to Υ;
16 **Output:** $\mathbf{l}_{\text{f}}$, $\mathbf{l}_{\text{r}}$, $\mathbf{a}$, $\mathbf{p}$, and Δ

$$\begin{aligned}D\left(\mathbf{W}, \mathbf{W}'\right) &= \frac{\psi(\mathbf{W};\boldsymbol{\vartheta}_*)\psi\left(\mathbf{W}';\boldsymbol{\vartheta}_*\right)}{\|\psi(\mathbf{W};\boldsymbol{\vartheta}_*)\|\|\psi(\mathbf{W}';\boldsymbol{\vartheta}_*)\|} \\ &= \frac{\mathbf{O}\cdot\mathbf{O}'}{\|\mathbf{O}\|\|\mathbf{O}'\|}, \quad (18)\end{aligned}$$

where $\mathbf{O} = \psi(\mathbf{W};\boldsymbol{\vartheta}_*)$ and $\mathbf{O}' = \psi(\mathbf{W}';\boldsymbol{\vartheta}_*)$ denote the latent features of DTD instances $\mathbf{W}$ and $\mathbf{W}'$ given the well-trained DNN of the encoder with parameter $\boldsymbol{\vartheta}_*$, respectively.

Algorithm 2 presents the procedure for refining the solutions obtained by Algorithm 1. We refer to the collection of information on the DTD instance, i.e., $\mathbf{W}$, and the corresponding solution obtained by Algorithm 1, i.e., $\mathbf{l}_{\text{f}}$, $\mathbf{l}_{\text{r}}$, $\mathbf{a}$, and $\mathbf{p}$, in a planning window as a data record, denoted by υ. Denote the set of data records and the number of data records in the set by Υ and $|\Upsilon|$, respectively. The value of $|\Upsilon|$ can be determined by balancing the computation complexity and the performance of the ULSCS algorithm. Using (18), Line 2 calculates the similarity between the DTD instance in the upcoming planning window, i.e., $\mathbf{W}$, and each DTD instance in the set Υ. Based on the calculated similarities, a set of data records containing the $|\Upsilon^{\text{re}}|$ most similar DTD instances, denoted by $\Upsilon^{\text{re}} \subseteq \Upsilon$, is selected. Line 4 obtains the solution to Problem P1, i.e., $\mathbf{l}_{\text{f}}$, $\mathbf{l}_{\text{r}}$, $\mathbf{a}$, and $\mathbf{p}$, and the corresponding performance Δ by calling Algorithm 1. From Lines 6 to 12, each historical SCM solution in the set Υ^{re}, i.e., $\mathbf{l}^{\text{re}}_{\text{f}}$, $\mathbf{L}^{\text{re}}_{\text{r}}$, and $\mathbf{a}^{\text{re}}$, is used in the initialization step (Line 2) of Algorithm 1, and the corresponding performance Δ' and solution $\mathbf{l}'_{\text{f}}$, $\mathbf{l}'_{\text{r}}$, $\mathbf{a}'$, $\mathbf{p}'$ are obtained. If $\Delta' > \Delta$, the solution to Problem P1 is updated as $\mathbf{l}'_{\text{f}}$, $\mathbf{l}'_{\text{r}}$, $\mathbf{a}'$, $\mathbf{p}'$; Otherwise, the solution to Problem P1 remains $\mathbf{l}_{\text{f}}$, $\mathbf{l}_{\text{r}}$, $\mathbf{a}$, and $\mathbf{p}$. As a result, the performance of Algorithm 2 is either better than or equal to that of Algorithm 1. When all historical SCM solutions in the set Υ^{re} have been utilized, lines 14 and 15 create a new data record containing the DTD instances and the corresponding solution, i.e., $\mathbf{l}_{\text{f}}$, $\mathbf{l}_{\text{r}}$, $\mathbf{a}$, and $\mathbf{p}$, and add the

data record to the set Υ, which can be useful in subsequent planning windows.

By using deep unsupervised learning, Algorithm 2 can reduce the computation complexity of planning-stage resource management of Algorithm 1 when the set Υ contains extensive historical data records. The computation complexity of the Algorithm 2 is $\mathcal{O}(|\Upsilon|OXI^3N^3)$ for selecting the best solution to Problem P2 from set Υ, where O represents the dimensionality of latent feature $\mathbf{O}$, $X = \sum_{j=1}^{J-1} B_j B_{j+1}$ denotes the computation complexity of the inference of the encoder (i.e., DNN ψ) with J layers, and B_j represents the number of neurons in layer j. Similar to the scheme used for experience replay in reinforcement learning [41], [42], we fix the maximum number of data records in the set Υ, i.e., $|\Upsilon|$, and keep the newly collected data records in Υ. As a result, by collecting and using new data records, Algorithm 2 can enhance the performance of Algorithm 1 while avoiding high computation complexity.

VI. Performance Evaluation

In this section, we first introduce the simulation settings. Then, we evaluate the performance of the proposed RAN slicing framework with the proposed AI-assisted approach.

A. Simulation Settings

The maximum SC and the antenna height of all SBSs are set to identical. The SC radius of the MBS and the maximum SC radius of each SBS are set to 1,500 m and 850 m, respectively. The carrier frequency of each BS is set to 1,500 MHz. The total available bandwidth of each BS and the sub-carrier spacing are set to 100 MHz and 30 kHz, respectively. Based on the COST 231-Hata Model in 3GPP standard [16], the average channel gain of downlink transmission within grid i for slice n in time interval t, i.e., $h^t_{i,n,i',n'}$, is approximated as the following equation:

$$h^t_{i,n} = 46.55 + 33.81 \times \log(f^{\mathrm{c}}_m) - 13.82 \times \log(H_m) + ((44.9 - 6.55 \times \log(H_m))) \times \log\left(d_{m_{i,n},i}\right), \quad (19)$$

where $d_{m_{i,n},i}$ is the distance (in kilometers) between BS $m_{i,n}$ and the center of grid i, f^{c}_m is the carrier frequency (in MHz) of BS m, H_m is the antenna height (in meters) of BS m, and H_{MBS} and H_{SBS} represent the antenna heights of the MBS and each SBS, respectively. UTs within the network coverage area in a time interval are distributed according to a Poisson point distribution (PPP). The rates of the PPP are the same across all time intervals within each planning window but different across planning windows. For each UT, its downlink data traffic load follows a Poisson process during each planning window. The mean values of downlink data traffic loads are different among UTs. We randomize the mean downlink data traffic load for each UT during a planning window within the interval of $[0.1, 1.5]$ Mbits. Other simulation parameters are listed in Table II.

The implementation of the DNNs for the auto-encoder is as follows. The DNN of the encoder contains 3 convolutional layers with channel sizes of 32, 64, and 128 respectively.

TABLE II
Simulation Parameters

Parameter	Value	Parameter	Value
N	2	T	3
$[\gamma_1^{\min}, \gamma_2^{\min}]$	[7, 11] dB	$[\lambda_1, \lambda_2]$	[1, 1]
H_{MBS}	50 m	H_{SBS}	15 m
ρ	1	N_0	-174 dBm/Hz

The kernel size is set as (3, 3) for both convolutional layers, respectively. Each convolutional layer is followed by a max-pooling layer with pool size (2, 2). Two fully-connected layers are then added with 512 and 64 neurons, followed by the output layer. The DNN architecture of the decoder is the reverse of that of the encoder. We adopt the Adam optimizer to train the DNNs. There are 8,000 different DTD instances used for the DNN training.

We compare the proposed RAN slicing framework with the following two benchmark schemes for IM and SCM, respectively:

- *Cell-based IM:* The downlink transmission power of each BS is the same for all grids within the SC of the BS;
- *Cell zooming (CZ):* The SC of each SBS is the same for all slices.

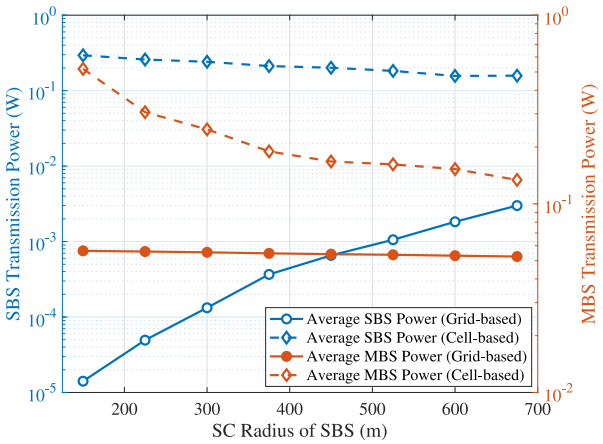
B. Performance of Grid-Based IM

In this subsection, we investigate the performance of the proposed grid-based IM in a simple network scenario with 1 MBS, 1 SBS, and 1 slice.

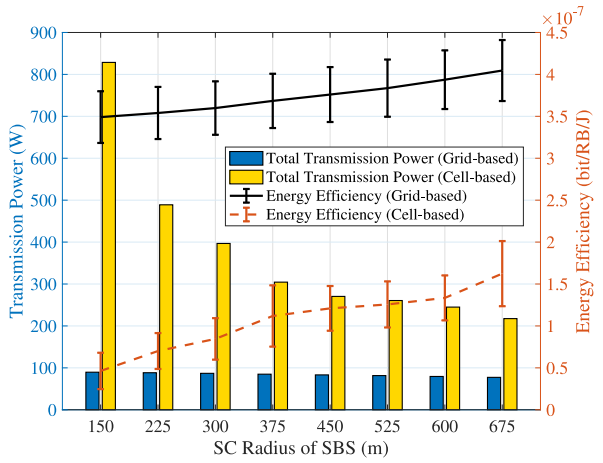
In Fig. 6(a), we compare the performance of transmission power obtained by the proposed grid-based IM with that obtained by cell-based IM. We average the transmission power of all grids within the SC of each BS for comparison. To satisfy the SINR requirement of the slice, the average transmission power of the MBS decreases, and the average transmission power of the SBS increases with the SC radius of the SBS for both grid-based and cell-based IM. This is because the number of grids covered by the MBS and the SBS decreases and increases, respectively. However, the MBS and SBS can achieve lower transmission power with grid-based IM compared to cell-based IM since the proposed grid-based IM can differentiate the transmission power based on their different locations. In addition, the slopes of all curves can vary with the SC radius of the SBS. This is because the uneven spatial distribution of data traffic loads results in non-uniform increments of data traffic loads for both the SBS and the MBS.

As shown in Fig. 6(b), we compare the performance of the two schemes in total transmission power and network energy efficiency. We observe that the proposed grid-based IM achieves higher network energy efficiency and lower total transmission power. The reason is that the proposed grid-based IM has a higher spatial granularity. Thus, the transmission power for each grid can be individually optimized to mitigate the interference among BSs in accordance with the DTD and the BS locations.

Next, we examine the impact of the spatial granularity on the network energy efficiency of grid-based IM. Fig. 7 shows



(a) Average downlink transmission power of the SBS and the MBS versus SC radius.



(b) Total downlink transmission power and network energy efficiency versus SC radius.

Fig. 6. Comparison between the proposed grid-based IM and cell-based IM.

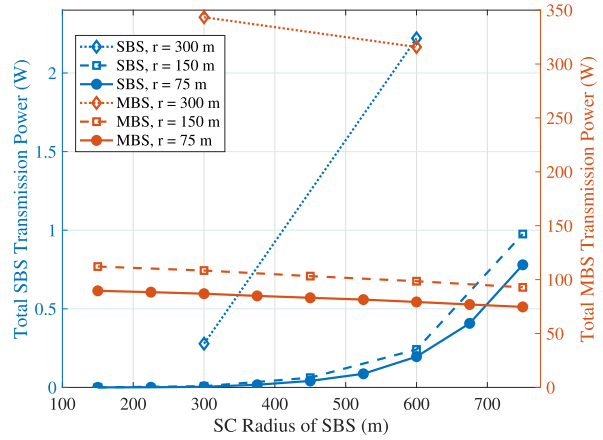


Fig. 7. The impact of spatial granularity on IM.

the total transmission power of the MBS and the SBS of grid-based IM with different grid diameters, i.e., different values of r. From this figure, we can make three observations. First, similar to case in Fig. 6, the total transmission power of the

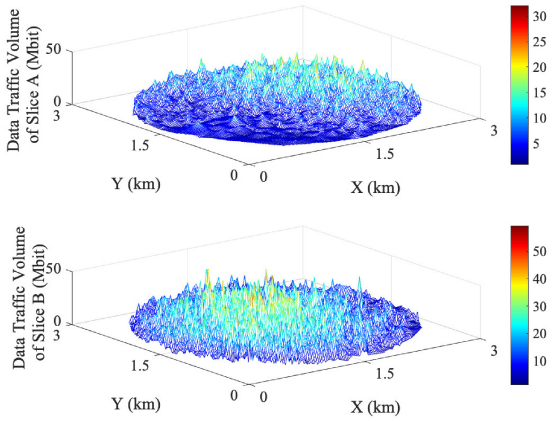


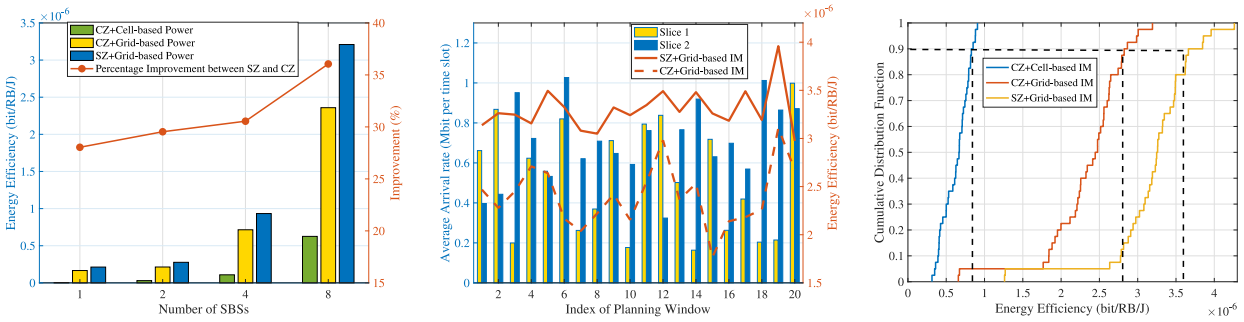
Fig. 8. The DTDs of two slices in a time interval.

MBS of grid-based IM increases with the SC radius of the SBS, while the total transmission power of the SBS of grid-based IM decreases with the SC radius of the SBS. Second, with grid-based IM, the total transmission power of each BS decreases with the grid diameter. This is because, when the grid diameter is smaller, the network can be divided into more grids and IM can be more fine-grained to suit the specific DTD. Third, if the grid diameter is sufficiently small (e.g., below 150 m), the effect of further decreasing the grid diameter on the total transmission power diminishes. This is because the total transmission power of each BS must exceed a threshold to satisfy the SINR requirement of each slice given a DTD.

C. Performance of Slicing-Based Resource Management

In this subsection, we examine the performance of the proposed RAN slicing framework in a network scenario with 1 MBS, 1 to 8 SBSs, and 2 slices. The DTDs of the two slices are different in a planning window. The DTDs of the two slices in a time interval is shown in Fig. 8.

Considering the network scenario with 1 MBS, 8 SBSs, and 2 slices, we compare the performance of the proposed schemes with benchmark schemes as shown in Fig. 9. In Fig. 9(a), we compare the network energy efficiency of the proposed flexible binary slice zooming plus grid-based IM (abbreviated as "SZ+ Grid-based IM") with that of two benchmark schemes, named "CZ + Cell-based IM" and "CZ+ Grid-based IM", averaged over 20 DTD instances. Three observations can be made from this figure. First, the network energy efficiency of all schemes increases with the number of SBSs. This is because, more SBSs can cover more grids, and the downlink transmissions within the grids from SBSs have a higher channel gain than that from the MBS, thereby improving network energy efficiency. Second, the proposed scheme outperforms the benchmark schemes in network energy efficiency in the cases with different number of SBSs. The reason is that the proposed scheme achieves fine-grained IM and SCM in time, space, and slices dimensions based on the different SINR requirements and DTDs of slices. Third, by comparing the "CZ + Cell-based IM" scheme with the "CZ + Grid-based IM" scheme, the performance advantage, i.e., the improvement

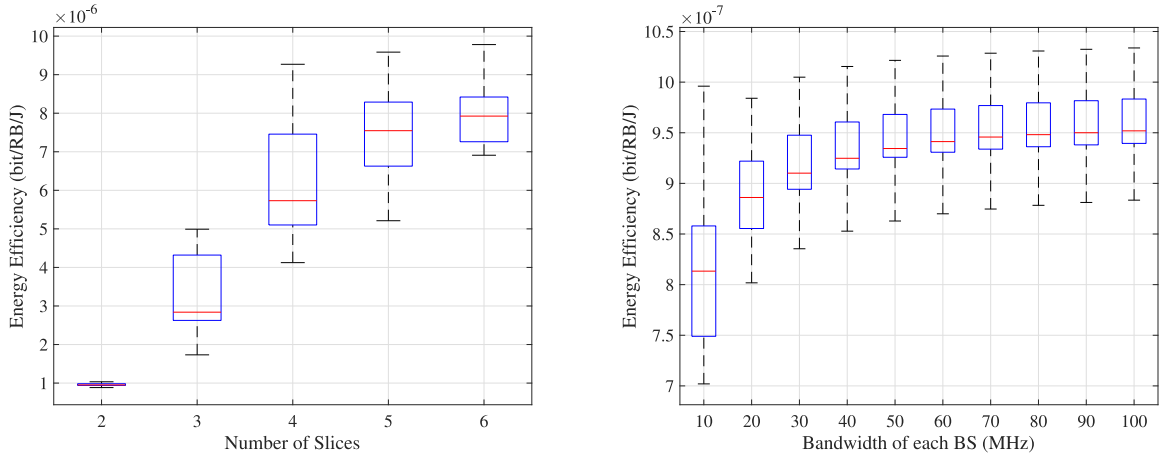


(a) Average network energy efficiency versus the number of SBSs.

(b) Temporal variations in data traffic load and network energy efficiency.

(c) The cumulative distribution function of network energy efficiency.

Fig. 9. Performance comparison between the proposed schemes and benchmark schemes.



(a) Network energy efficiency versus the number of slices.

(b) Network energy efficiency versus the total available bandwidth of each BS.

Fig. 10. The performance of the developed RAN slicing framework in different network scenarios.

(in percentage) of the proposed "SZ+ Grid-based IM" scheme compared to the "CZ+ Grid-based IM" scheme, increases with the number of SBSs. This is because, as more SBSs are deployed, the proposed scheme has more SC options available for selection, resulting in better interference management among BSs. Therefore, the percentage improvement compared to other schemes increases with the number of SBSs.

In Fig. 9(b), we show the temporal variations in network energy efficiency of the proposed scheme across multiple planning windows. The Poisson data arrival rate averaged over all UTs in each slice varies across planning windows, and, accordingly, the network energy efficiency of the proposed scheme temporally varies. Meanwhile, we can observe that the proposed scheme outperforms that of the "CZ + Grid-based IM" scheme in each planning window due to the high adaptivity of the proposed scheme in coping with spatiotemporal network dynamics. Fig. 9(c) shows the cumulative distribution function of the network energy efficiency of the three schemes over the 40 different DTD instances in the same case. We can observe from Fig. 9(c) that the proposed schemes achieve higher network energy efficiency than the benchmark schemes for most DTD instances.

In Fig. 10, we show the performance of the developed RAN slicing framework in different network scenarios. Considering the network with 1 MBS, 4 SBSs, and the grid diameter of 150 m, we show the network energy efficiency versus the number of slices and the total available bandwidth of each BS in Fig. 10(a) and Fig. 10(b), respectively. A box plot representing the range of network energy efficiency over 10 independent simulation runs is shown in Fig. 10(a), in which the number of slices is set from 2 to 6, and the overall data traffic load of all slices is fixed in each simulation run. We can make the following two observations. First, the network energy efficiency of the developed scheme increases with the number of slices. This is because, for the same DTD, the number of decision variables of IM and SCM in the developed scheme increases with the number of slices, thereby improving the granularity of slicing-based resource management. As a result, the developed scheme can achieve higher energy efficiency by balancing the overall data traffic load across BSs due to the refined granularity in the slice dimension. Second, the effect of increasing the number of slices on the network energy efficiency diminishes when the number of slices increases since it becomes more difficult for IM and SCM to satisfy the SINR requirement of each slice.

In Fig. 10(b), varying the total bandwidth of each BS from 10 MHz to 100 MHz, we present the box plot of network energy efficiency over 10 independent simulation runs for

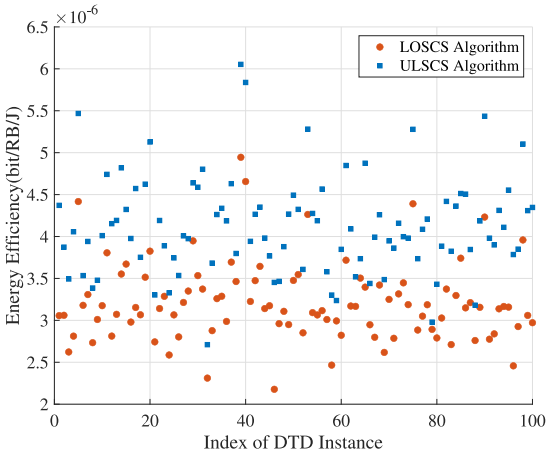


Fig. 11. Network energy efficiency comparison between the LOSCS and ULSCS algorithms.

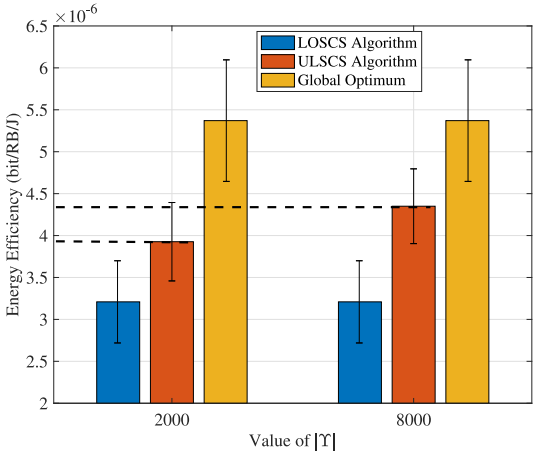
each bandwidth setting. We can observe that network energy efficiency increases with the total available bandwidth of each BS. This is because, for the same data traffic load of each BS, increasing the total bandwidth of each BS can reduce the likeliness of the planning-stage interference among BSs (as discussed in Section IV-B). Consequently, the required transmission power to satisfy the SINR requirement of each slice is reduced, thereby improving the network energy efficiency.

D. Performance of the ULSCS Algorithm

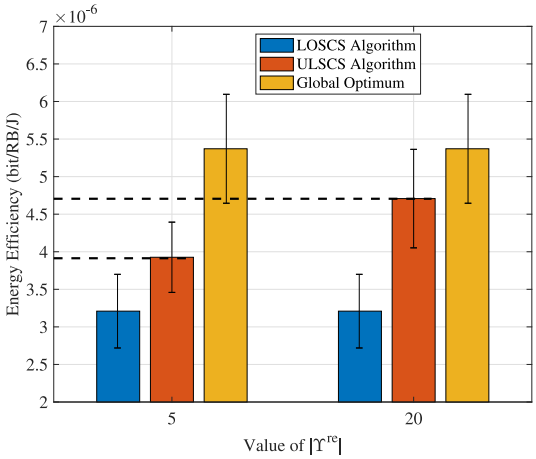
In this subsection, we evaluate the energy efficiency performance of the proposed ULSCS algorithm and the LOSCS algorithm as well as the impact of the number of data records i.e., $|\Upsilon|$, and the number of selected data records, i.e., $|\Upsilon^{\text{re}}|$. We consider a network with 8 SBSs, 1 MBS, and 2 slices.

In Fig. 11, we compare the energy efficiency performance of the ULSCS and LOSCS algorithms for 100 cases with different DTD instances. The network energy efficiency achieved by the ULSCS algorithm is higher than that achieved by the LOSCS algorithm in all cases. The ULSCS algorithm selects some historical solutions to use as the initial settings of the LOSCS algorithm, which results in relatively high performance compared to the LOSCS algorithm. The worst-case network energy efficiency of the ULSCS algorithm equals that obtained by the LOSCS Algorithm.

In Fig. 12(a), we evaluate the network energy efficiency of the ULSCS algorithm, averaged over 40 DTD instances, given different number of data records, i.e., different values of $|\Upsilon|$. Two observations can be made in Fig. 12(a). First, the performance gap between the ULSCS algorithm and the LOSCS algorithm increases when more data records are used. This is because having more data records in Υ can improve the performance of DNN training and provide a large number of historical DTD instances for solution refinement. Second, the performance of the ULSCS algorithm can approach the optimum global value, especially when a large value of $|\Upsilon|$ is used. Moreover, we



(a) The impact of $|\Upsilon|$.



(b) The impact of $|\Upsilon^{\text{re}}|$.

Fig. 12. Network energy efficiency given different values of $|\Upsilon|$ and $|\Upsilon^{\text{re}}|$, respectively.

examine the impact of the number of selected data records, i.e., different values of $|\Upsilon^{\text{re}}|$, in Fig. 12(b). The performance gap between the ULSCS algorithm and the LOSCS algorithm increases with the number of selected data records, and performance of the ULSCS algorithm can approach the global optimum when a larger number of selected historical solutions are used for solution refinement. This is because more data records in $|\Upsilon^{\text{re}}|$ result in more similar DTD instances being selected as the initial settings in the ULSCS algorithm, and thus benefit achieving global optimum. Consequently, Fig. 12 demonstrates the potential of the AI-assisted approach to address the slicing-based resource management problems.

VII. Conclusion and Future Work

In this paper, we have designed a RAN slicing framework for a two-tier RAN to determine the SC and transmission power of BSs. The proposed framework introduces customized SC for different services and improves the granularity of IM to suit service demands in the spatial, temporal, and slice dimensions. Based on the framework, a network

$$\hat{\mathbf{H}}^t \cdot \mathbf{p}^t = \rho \left(\begin{bmatrix} \gamma_1^{\min}\mathbf{\Omega}_{1,1}^t & \cdots & \gamma_1^{\min}\mathbf{\Omega}_{1,n'}^t & \cdots & \gamma_1^{\min}\mathbf{\Omega}_{1,N}^t \\ \vdots & \ddots & \vdots & & \vdots \\ \gamma_n^{\min}\mathbf{\Omega}_{n,1}^t & \cdots & \gamma_n^{\min}\mathbf{\Omega}_{n,n'}^t & \cdots & \gamma_n^{\min}\mathbf{\Omega}_{n,N}^t \\ \vdots & & \vdots & \ddots & \vdots \\ \gamma_N^{\min}\mathbf{\Omega}_{N,1}^t & \cdots & \gamma_N^{\min}\mathbf{\Omega}_{N,n'}^t & \cdots & \gamma_N^{\min}\mathbf{\Omega}_{N,N}^t \end{bmatrix}_{IN\times IN} \mathbf{p}^t + \begin{bmatrix} \gamma_1^{\min}N_0 \\ \vdots \\ \gamma_n^{\min}N_0 \\ \vdots \\ \gamma_N^{\min}N_0 \end{bmatrix}_{IN\times 1} \right). \tag{26}$$

energy efficiency maximization problem has been formulated, which takes into account the inter-slice and intra-slice interference and diverse QoS requirements of slices. The proposed AI-assisted approach decouples the problem into two sub-problems and solve them by incorporating deep unsupervised learning with optimization methods. The results have demonstrated the effectiveness of the proposed RAN slicing framework in improving energy efficiency, and the efficiency of the developed AI-assisted approach. The proposed framework and approach extend the advantages of slicing-based resource management towards supporting diverse services in RANs. In the future, we will investigate slicing-based resource management considering the coupling between the planning and operation stages.

APPENDIX

PROOF OF THEOREM 1

Let Δ^t denote the network energy efficiency in time interval $t \in \mathcal{T}$ and define

$$\varsigma_n^t = \sum_{m\in\mathcal{M}} \sum_{i\in\mathcal{I}_{m,n}} \tau p_{i,n}^t, \quad n \in \mathcal{N}, \tag{20}$$

and

$$\chi_n^t = \frac{\sum_{m\in\mathcal{M}} \sum_{i\in\mathcal{I}_{m,n}} w_{i,n}^t}{\sum_{m\in\mathcal{M}} \sum_{i\in\mathcal{I}_{m,n}} w_{i,n}^t \eta_n}, \quad n \in \mathcal{N}. \tag{21}$$

The network energy efficiency in time interval t is given by:

$$\Delta^t = \sum_{n\in\mathcal{N}} \frac{\lambda_n \chi_n^t}{\varsigma_n^t}. \tag{22}$$

The Hessian matrix of the network energy efficiency Δ^t can be written as the following block matrix:

$$\nabla^2\Delta^t = \frac{\partial^2\Delta^t}{\partial p_{x,y}^t \partial p_{x',y'}^t} = \begin{bmatrix} \mathbf{A}_1^t & & & \mathbf{0} \\ & \ddots & & \\ & & \mathbf{A}_n^t & \\ & & & \ddots \\ \mathbf{0} & & & \mathbf{A}_N^t \end{bmatrix}_{IN\times IN}, \tag{23}$$

where block $\mathbf{A}_n^t$ for any $n \in \mathcal{N}$ is given by:

$$\mathbf{A}_n^t = \frac{2\lambda_n \chi_n^t \tau^2}{(\varsigma_n^t)^3} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}_{I\times I} \tag{24}$$

If constraint (12d) is satisfied, ς_n^t is positive. In this case, the first-order leading principal minor of the Hessian matrix, i.e., $\frac{2\lambda_n\chi_n^t\tau^2}{(\varsigma_n^t)^3}$, is nonnegative. Meanwhile, all the other leading principal minors equal 0. As a result, the Hessian matrix is positive semidefinite when constraint (12d) is satisfied. Thus, when $\forall p_{i,n}^t > 0$, the function Δ^t is convex.

Function Δ^t increases with the decrease of allocated transmission power for all grids, while the allocated transmission power for all grids should satisfy the SINR constraints in (9). Consequently, due to the convexity of function Δ^t, the IM solution must exist on the boundary of the feasible domain. Thus, the optimal IM solution should satisfy (9) with equality, i.e.,

$$\frac{\bar{p}_{i,n}^t h_{i,n,i,n}^t}{N_0 + I_{i,n}^t} = \rho\gamma_n^{\min}. \tag{25}$$

Define $\hat{\mathbf{H}}^t$ and $\mathbf{\Omega}_{n,n'}^t$ in (16) and (15), respectively. We rewrite (25) into the matrix format as (26), as shown at top of the page. Therefore, the optimal downlink transmission power in time interval t can be derived in closed-form as (14).

REFERENCES

[1] W. Chen, J. Montojo, J. Lee, M. Shafi, and Y. Kim, "The standardization of 5G-advanced in 3GPP," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 98–104, Nov. 2022.

[2] X. Shen et al., "Toward immersive communications in 6G," *Front. Comput. Sci.*, vol. 4, Jan. 2023, Art. no. 1068478.

[3] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 3rd Quart., 2018.

[4] J. Wang, J. Liu, J. Li, and N. Kato, "Artificial intelligence-assisted network slicing: Network assurance and service provisioning in 6G," *IEEE Veh. Technol. Mag.*, vol. 18, no. 1, pp. 49–58, Mar. 2023.

[5] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[6] X. Ma, Q. Zeng, H. Chi, and L. Luo, "No more companion Apps hacking but one dongle: Hub-based blackbox fuzzing of loT firmware," in *Proc. ACM MobiSys*, 2023, pp. 205–218.

[7] X. Shen et al., "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, no. 1, pp. 45–66, Jan. 2020.

[8] J. Wang and J. Liu, "Secure and reliable slicing in 5G and beyond vehicular networks," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 126–133, Feb. 2022.

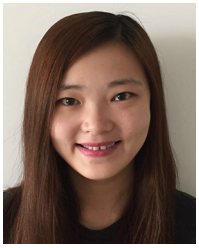
[9] J. Mei, X. Wang, and K. Zheng, "An intelligent self-sustained RAN slicing framework for diverse service provisioning in 5G-beyond and 6G networks," *Intell. Conver. Netw.*, vol. 1, no. 3, pp. 281–294, Dec. 2020.

[10] M. Li, J. Gao, C. Zhou, X. Shen, and W. Zhuang, "Slicing-based artificial intelligence service provisioning on the network edge: Balancing AI service performance and resource consumption of data management," *IEEE Veh. Technol. Mag.*, vol. 16, no. 4, pp. 16–26, Dec. 2021.

[11] W. Zhuang and K. Qu, *Dynamic Resource Management in Service-Oriented Core Networks*. Cham, Switzerland: Springer, 2021.
[12] Z. M. Fadlullah, B. Mao, and N. Kato, "Balancing QoS and security in the edge: Existing practices, challenges, and 6G opportunities with machine learning," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2419–2448, 4th Quart., 2022.
[13] C. Zhou, J. Gao, M. Li, X. Shen, and W. Zhuang, "Digital twin-empowered network planning for multi-tier computing," *J. Commun. Inf. Netw.*, vol. 7, no. 3, pp. 221–238, Sep. 2022.
[14] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 1–30, 1st Quart., 2022.
[15] J. Mei, X. Wang, K. Zheng, G. Boudreau, A. B. Sediq, and H. Abou-Zeid, "Intelligent radio access network slicing for service provisioning in 6G: A hierarchical deep reinforcement learning approach," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6063–6078, Sep. 2021.
[16] "Technical specification group radio access network; Radio network planning aspects (Release 17)," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 43.030 V17.0.0, Mar. 2022.
[17] M. Zambianco and G. Verticale, "Interference minimization in 5G physical-layer network slicing," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4554–4564, Jul. 2020.
[18] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture," in *Proc. ACM MobiCom*, 2017, New York, NY, USA, pp. 127–140.
[19] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks," *IEEE Access*, vol. 8, pp. 45674–45688, 2020.
[20] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Pérez, "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 3044–3058, Oct. 2017.
[21] B. Yang, L. Zhang, O. Onireti, P. Xiao, M. A. Imran, and R. Tafazolli, "Mixed-numerology signals transmission and interference cancellation for radio access network slicing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5132–5147, Aug. 2020.
[22] Y. Sun, S. Qin, G. Feng, L. Zhang, and M. A. Imran, "Service provisioning framework for RAN slicing: User admissibility, slice association and bandwidth allocation," *IEEE Trans. Mobile Comput.*, vol. 20, no. 12, pp. 3409–3422, Dec. 2021.
[23] S. O. Oladejo and O. E. Falowo, "Latency-aware dynamic resource allocation scheme for multi-tier 5G network: A network slicing-multitenancy scenario," *IEEE Access*, vol. 8, pp. 74834–74852, 2020.
[24] H. Xiang, S. Yan, and M. Peng, "A realization of fog-RAN slicing via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2515–2527, Apr. 2020.
[25] A. Filali, Z. Mlika, S. Cherkaoui, and A. Kobbane, "Dynamic SDN-based radio access network slicing with deep reinforcement learning for URLLC and eMBB services," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 4, pp. 2174–2187, Jul./Aug. 2022.
[26] O. Adamuz-Hinojosa, V. Sciancalepore, P. Ameigeiras, J. M. Lopez-Soler, and X. Costa-Pérez, "A stochastic network calculus (SNC)-based model for planning B5G uRLLC RAN slices," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1250–1265, Feb. 2023.
[27] W. Wu et al., "Dynamic RAN slicing for service-oriented vehicular networks via constrained learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2076–2089, Jul. 2021.
[28] J. Janković, Ž. Ilić, A. Oračević, S. A. Kazmi, and R. Hussain, "Effects of differentiated 5G services on computational and radio resource allocation performance," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 2, pp. 2226–2241, Jun. 2021.
[29] S. D'Oro, F. Restuccia, A. Talamonti, and T. Melodia, "The slice is served: Enforcing radio access network slicing in virtualized 5G systems," in *Proc. IEEE INFOCOM*, 2019, Paris, France, pp. 442–450.
[30] P. Munoz, Ñ. Adamuz-Hinojosa, J. Navarro-Ortiz, O. Sallent, and J. Pérez-Romero, "Radio access network slicing strategies at spectrum planning level in 5G and beyond," *IEEE Access*, vol. 8, pp. 79604–79618, 2020.
[31] S. Bakri, P. A. Frangoudis, A. Ksentini, and M. Bouaziz, "Data-driven RAN slicing mechanisms for 5G and beyond," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 4, pp. 4654–4668, Dec. 2021.
[32] M. Yan, G. Feng, J. Zhou, Y. Sun, and Y.-C. Liang, "Intelligent resource scheduling for 5G radio access network slicing," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7691–7703, Aug. 2019.
[33] Q. Ye, W. Zhuang, S. Zhang, A. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896–9910, Oct. 2018.
[34] H. Shen, Q. Ye, W. Zhuang, W. Shi, G. Bai, and G. Yang, "Drone-small-cell-assisted resource slicing for 5G uplink radio access networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 7, pp. 7071–7086, Jul. 2021.
[35] "Technical specification group radio access network; Small cell enhancements for E-UTRA and E-UTRAN-physical layer aspects (Release 12)," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 36.872 V12.1.0, Dec. 2013.
[36] A. Osseiran, J. Monserrat, and P. Marsch, *5G Mobile and Wireless Communications Technology*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
[37] L. Yang, L. Cao, and H. Zheng, "Proactive channel access in dynamic spectrum networks," *Phys. Commun.*, vol. 1, no. 2, pp. 103–111, 2008.
[38] "Technical specification group radio access network; study on Cell-specific Reference Signals (CRS) interference mitigation for homogenous deployments of LTE (Release 12)," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR 36.363 V12.0.0, Dec. 2013.
[39] B. Korte and J. Vygen, *Combinatorial Optimization*. Heidelberg, Germany: Springer, 2011.
[40] K. G. Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *Proc. IEEE ICCV*, 2017, Venice, Italy, pp. 5747–5756.
[41] C. Zhou, W. Wu, H. He, P. Yang, F. Lyu, N. Cheng, and X. Shen, "Deep reinforcement learning for delay-oriented IoT task scheduling in SAGIN," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 911–925, Feb. 2021.
[42] Q. Liu, T. Han, N. Zhang, and Y. Wang, "Deepslicing: Deep reinforcement learning assisted resource allocation for network slicing," in *Proc. IEEE GLOBECOM*, 2020, Taipei, Taiwan, pp. 1–6.

Conghao Zhou (Member, IEEE) received the B.Eng. degree from Northeastern University, Shenyang, China, in 2017, the M.Sc. degree from the University of Illinois at Chicago, Chicago, IL, USA, in 2018, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2022, where he is currently a Postdoctoral Fellow. His research interests include space-air-ground integrated networks, network slicing, and machine learning for wireless networks.

Jie Gao (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from the University of Alberta, Edmonton, AB, Canada, in 2009 and 2014, respectively. He was a Postdoctoral Fellow with Toronto Metropolitan (formerly Ryerson) University, Toronto, ON, Canada, from 2017 to 2019 and a Research Associate with the University of Waterloo, Waterloo, ON, Canada, from 2019 to 2020. He was an Assistant Professor with the Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI, USA, from 2020 to 2022 and is currently an Assistant Professor with the School of Information Technology, Carleton University, Ottawa, ON, Canada. His research interests include machine learning for communications and networking, cloud and multi-access edge computing, Internet of Things and industrial IoT solutions, and B5G/6G networks in general.



Mushu Li (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, ON, Canada, in 2021. She is currently a Postdoctoral Fellow with Toronto Metropolitan (formerly Ryerson) University, ON, Canada. She was a Postdoctoral Fellow with the University of Waterloo from 2021 to 2022. Her research interests include mobile edge computing, the system optimization in wireless networks, and machine learning-assisted network management. She was the recipient of Natural Science and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship in 2022 and a NSERC Canada Graduate Scholarship in 2018.

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990.

He is a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular networks. He received the "West Lake Friendship Award" from Zhejiang Province in 2023, the President's Excellence in Research from the University of Waterloo in 2022, the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory in 2021, the R.A. Fessenden Award from IEEE, Canada, in 2019, the Award of Merit from the Federation of Chinese Canadian Professionals, (Ontario) in 2019, the James Evans Avant Garde Award from the IEEE Vehicular Technology Society in 2018, the Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society (ComSoc), the Technical Recognition Award from Wireless Communications Technical Committee in 2019, and AHSN Technical Committee in 2013. He has also received the Excellent Graduate Supervision Award from the University of Waterloo in 2006 and the Premier's Research Excellence Award in 2003 from the Province of Ontario, Canada. He serves/served as the General Chair for the 6G Global Conference'23 and ACM Mobihoc'15, the Technical Program Committee Chair/Co-Chair for IEEE Globecom'24, 16, and 07, IEEE Infocom'14, IEEE VTC'10 Fall, and the Chair for the IEEE ComSoc Technical Committee on Wireless Communications. He served as the Editor-in-Chief for the IEEE INTERNET OF THINGS JOURNAL, IEEE NETWORK, and *IET Communications*. He is the President of the IEEE ComSoc. He was the Vice President for Technical and Educational Activities, the Vice President for Publications, a Member-at-Large on the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee, and a member of IEEE Fellow Selection Committee of the ComSoc. He is a registered Professional Engineer of Ontario, Canada, the Engineering Institute of Canada Fellow, the Canadian Academy of Engineering Fellow, the Royal Society of Canada Fellow, the Chinese Academy of Engineering Foreign Member, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society.

Weihua Zhuang (Fellow, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Dalian Marine University, China, and the Ph.D. degree in electrical engineering from the University of New Brunswick, Canada. She is a University Professor and a Tier I Canada Research Chair of Wireless Communication Networks with the University of Waterloo, Canada. Her research focuses on network architecture, algorithms and protocols, and service provisioning in future communication systems. She is the recipient of the 2021 Women's Distinguished Career Award from IEEE Vehicular Technology Society, the 2021 Technical Contribution Award in Cognitive Networks from IEEE Communications Society, the 2021 R. A. Fessenden Award from IEEE Canada, and the 2021 Award of Merit from the Federation of Chinese Canadian Professionals in Ontario. She was the Editor-in-Chief of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2007 to 2013, the General Co-Chair of 2021 IEEE/CIC International Conference on Communications in China, the Technical Program Chair/Co-Chair of 2017/2016 IEEE VTC Fall, the Technical Program Symposia Chair of 2011 IEEE Globecom, and an IEEE Communications Society Distinguished Lecturer from 2008 to 2011. She is an Elected Member of the Board of Governors and the President of the IEEE Vehicular Technology Society. She is a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada.

Xu Li received the Ph.D. degree in computer science from Carleton University, Canada, in 2008. He is a Senior Principal Researcher with Huawei Technologies Canada. From 2016 to 2019, he actively participated in 3GPP 5G standardization and contributed extensively through 100+ standard proposals. He has published 100+ refereed scientific papers and is holding 80+ U.S. patents. His work received over 10,000 citations and leads to an H-index of 54. His current research interests are focused in design and development of next-generation wireless networks, computer networks, and especially in AI theory and applications in future networks. He was on the editorial boards of *IEEE Communications Magazine*, the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and the Wiley *Transactions on Emerging Telecommunications Technologies* and a number of other international archive journals.

Weisen Shi (Member, IEEE) received the B.S. degree from Tianjin University, Tianjin, China, in 2013, the M.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, in 2020. He is currently a Senior Engineer with Huawei Technologies Canada, Inc., Ottawa, ON, Canada. His interests include AI in networks, space-air-ground integrated networks, UAV communication and networking, and RAN slicing.