Hierarchical Digital Twin for Efficient 6G Network Orchestration via Adaptive Attribute Selection and Scalable Network Modeling

Pengyi Jia, *Member, IEEE*, Xianbin Wang, *Fellow, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—Achieving both a holistic and in-depth understanding of network dynamics through accurate modeling is essential for orchestrating future 6G networks, considering their increasing complexity and service diversity. However, traditional situation-agnostic data collection and network modeling approaches often undermine the efficacy and timeliness of network orchestration in such complex environments. Furthermore, temporal misalignments caused by varying modeling delays across distributed networks further impair centralized decision-making. To address these challenges, this paper proposes a hierarchical digital twin framework with an adaptive layered architecture designed for problem-oriented 6G network modeling and orchestration. At higher layers, we introduce an adaptive attribute selection mechanism that efficiently evaluates network situations and identifies problematic areas. This mechanism prioritizes critical attributes by jointly considering their relevance to current network objectives and modeling complexity. At lower layers, these prioritized attributes and critical users are selectively incorporated into scalable network modeling. More detailed digital twins are then created to deliver targeted solutions for optimizing user association and power allocation. Additionally, we implement a multi-level synchronization mechanism to ensure temporal alignment among the digital twins, thereby enhancing the effectiveness of model-based orchestration. Extensive simulations validate the efficient identification of pressing operational issues and the effective orchestration of complex 6G networks.**

***Index Terms*—Heterogeneous networks, digital twin, user association, resource allocation, decomposition, time synchronization, quality of service.**

I. INTRODUCTION

THE advent of sixth-generation (6G) telecommunications marks a transformative shift in network architecture, characterized by unprecedented heterogeneity and complexity [1]. Anticipated 6G heterogeneous networks (HetNets) are distinguished not only by the proliferation of base station (BS) types, including macro, micro, and femto BSs, but also by the increasing diversification of quality of service (QoS) requirements to support various vertical applications, such as Extended Reality, Metaverse, and Internet of Everything [2]. Amidst this expeditious evolution, managing the surging data traffic from intricate interactions among wireless devices within constrained spectrum resources becomes increasingly challenging, highlighting the pressing necessity for efficient QoS requirement satisfaction [3].

Received 25 April 2024; revised 15 October 2024 and 12 January 2025; accepted 6 March 2025. Date of publication 17 March 2025; date of current version 18 September 2025. This work was supported in part by the Discovery Program of Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2024-05720 and in part by Canada Research Chair Program. The associate editor coordinating the review of this article and approving it for publication was J. Nie. *(Corresponding author: Xianbin Wang.)*

Pengyi Jia and Xianbin Wang are with the Department of Electrical and Computer Engineering, Western University, London, ON N6A 5B9, Canada (e-mail: pjia7@uwo.ca; xianbin.wang@uwo.ca).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Digital Object Identifier 10.1109/TCOMM.2025.3552299

A. Motivations

Given the inherent heterogeneity and complexity of 6G networks, it is imperative to design innovative network orchestration strategies, such as resource allocation and traffic engineering, to efficiently meet the diverse QoS requirements within intricate network architectures [4]. Current approaches to network orchestration typically rely on network-wide modeling and large-scale optimization, such as network congestion control [5] and BS-user association [6]. These methods depend on comprehensive data acquisition and centralized processing to manage all network infrastructure components, ensuring global coordination and optimization. However, as network demands grow, this approach faces increasing challenges in scalability and responsiveness.

The proliferation of communication devices and the associated diversity of network attributes in 6G HetNets further exacerbate the burden on processing centers. Collecting and analyzing data for every attribute across all users to support large-scale network optimization imposes excessive demands on computational and spectrum resources [7]. Consequently, existing unstructured and situation-agnostic orchestration methods often lead to suboptimal performance and delayed decision-making in the highly dynamic environments of 6G HetNets. This highlights the need for scalable and adaptive orchestration strategies that move beyond rigid, one-size-fits-all solutions to efficiently satisfy diverse user QoS demands under varying network conditions.

Attaining both holistic and in-depth insights into network situational dynamics through advanced network modeling becomes crucial. This enables rapid identification of actual needs and emerging issues within complex networks, facilitating targeted and timely decisions to ensure efficient QoS satisfaction. In this context, digital twin technology emerges as a promising solution for real-time and comprehensive modeling of communication networks in remote control centers [8], [9]. By leveraging statistical data on network performance indicators, digital twins enable rapid analysis of large-scale networks while mitigating dependence on highly dynamic physical layer attributes [3]. The real-time and statistical link information provided by digital twins serves as a foundation for informed decision-making across various layers of communication systems [10], [11].

Therefore, digital twins represent indispensable tools for achieving efficient orchestration and pervasive intelligence in 6G networks, enabling timely problem identification and insightful decision-making [12]. However, existing digital twin-based network operation approaches, which typically emphasize fine-grained modeling of the entire network, face significant challenges in efficiently evaluating network situations and identifying critical issues within complex 6G HetNets.

- The significant complexity of creating comprehensive, fine-grained digital twins for large-scale 6G HetNets arises from the vast array of heterogeneous wireless devices, including an enormous number of mobile users and various opportunistically deployed BSs [13]. This complexity is further intensified by dynamic QoS requirements and rapidly fluctuating communication environments among distributed users. Collectively, these factors generate diverse network scenarios characterized by varying QoS provisioning, uneven BS load distributions, and inconsistent resource availability across the network [14]. Applying a uniform approach to digital twin construction and network orchestration in such dynamic and heterogeneous environments can severely reduce responsiveness to real-time network issues, particularly for devices experiencing suboptimal conditions, while risking resource inefficiency for devices in more stable environments. Consequently, accurately capturing essential network dynamics and prioritizing urgent issues among these intricate and interconnected network components become crucial for efficient network orchestration.
- Neglecting the situation-dependent modeling values of heterogeneous network attributes during digital twin construction can result in substantial inefficiencies. The cost of modeling different attributes varies widely, influenced by factors such as the required data volume, inherent data complexity, and stability of the underlying data patterns. Moreover, the relevance of these attributes to current user demands and operational objectives dynamically shifts with evolving network conditions, making each attribute's value highly context-specific. For instance, computationally intensive attributes may retain substantial modeling value if they directly address critical challenges in the current network scenario, despite their complexity. Given limited communication resources and stringent real-time optimization needs, transitioning toward a value-oriented modeling approach is essential for complex and dynamic 6G HetNets. This approach contrasts sharply with traditional methods that indiscriminately model all network attributes, irrespective of their immediate utility.
- Temporal misalignment between wireless devices and their digital twins arises due to intrinsic physical heterogeneity and substantial modeling delays. Variations in local clocks and sampling capabilities across physical devices inevitably lead to discrepancies in the collected modeling data, thereby obscuring important temporal correlations [15], [16]. Additionally, the multiple processes involved in constructing digital twins, from data transmission to centralized processing, introduce accumulative delays that further exacerbate temporal misalignment among digital twins [17]. As a result, interactions between physical devices and their digital twins suffer from significant temporal disparities, causing mismatches between digital twin outputs and the actual real-time network conditions [18]. Traditional approaches often overlook these synchronization challenges, producing insights of limited actionable value and thus undermining their effectiveness in supporting timely and responsive network orchestration.

Based on these observations, it is essential to design a novel digital twin framework that cohesively addresses network modeling complexity, heterogeneous modeling value, and synchronization issues to enable efficient network orchestration.

B. Related Work

Recent advancements in digital twin technology [3], [17], [18], [19], [20], [21], [22], [23], [24] have demonstrated significant potential in accurately representing dynamic network states for network orchestration and real-time decision-making. However, notable challenges arise when these methods are applied to large-scale or highly dynamic networks, primarily due to spatial-temporal complexities and misalignments during the mapping between physical systems to their virtual counterparts.

For instance, a graph neural network (GNN)-based digital twin framework proposed in [19] optimizes power allocation and user association in terahertz bands. By leveraging GNN predictions for topology mining and feature extraction, this framework effectively captures dynamic network characteristics. Nevertheless, although this method performs well in smaller or controlled settings, its scalability to large and more dynamic networks is limited by the increasing complexity of spatial-temporal interactions. Similarly, digital twins in ultra-reliable and low-latency communication (URLLC) networks have been employed for tasks such as user association, task offloading, and resource allocation [20]. Despite their latency-reduction objectives, the variability inherent to spatial and temporal factors within URLLC environments still poses substantial challenges to timely and accurate decision-making. To address these concerns, a digital twin-based mobile network framework introduced in [3] utilizes virtual network topologies to manage load distribution and spatial-temporal dynamics more effectively. Although this approach significantly improves overall network performance and QoS, the complexity of large-scale HetNets continues to create difficulties, especially regarding resource consumption and decision delays.

Additionally, cross-domain synchronization is crucial for digital twins to maintain real-time alignment with their physical counterparts, effectively mitigating misalignments between digital and physical domains [22]. For example, a dynamic synchronization framework is proposed in [18] for digital twins in the Metaverse, aligning virtual models of service providers with real-world data generated by IoT devices. This framework employs a game-theoretic strategy to optimize synchronization frequency and intensity, thereby improving the accuracy of virtual models. Similarly, a learning-based prediction model introduced in [23] aims to achieve accurate synchronization between vehicular digital twins and users in heterogeneous vehicular networks by compensating for data transmission delays. However, both methods insufficiently address temporal misalignment caused by variations in data transmission and collection rates. To further alleviate this issue, a method for synchronizing and resampling multi-attribute data was developed in [17], mitigating temporal misalignments introduced by asynchronous IoT devices and heterogeneous sampling rates. Nonetheless, effectively managing multi-source delays in large-scale networks, particularly those with high data variability, remains an open challenge.

Clearly, existing digital twin-based schemes overlook the complexities and delays associated with network-wide modeling. The substantial communication and computational resources required for digital twin construction, along with temporal misalignments caused by processing and transmission delays, significantly reduce modeling efficiency and accuracy in highly dynamic large-scale HetNets. To address these issues, we previously proposed a dual-layered digital twin paradigm in [25] that identifies hotspots within small 6G HetNets for intelligent network orchestration. Nevertheless, challenges such as inefficient network situation discovery and inadequate cross-domain synchronization remain unsolved for future networks exhibiting increased complexity and heterogeneity.

C. Contributions

Motivated by these observations, we propose a hierarchical digital twin framework to enable efficient network orchestration via adaptive attribute selection and scalable network modeling. By dynamically prioritizing network attributes based on their relevance and impact on current network conditions, our approach optimizes resource utilization to guide the digital twin construction process. Scalable network modeling is realized through a problem-oriented methodology that effectively manages modeling complexity, ensuring cost-effective network orchestration. Furthermore, a multi-level synchronization mechanism is implemented to maintain accurate temporal alignment across digital twins, enhancing their reliability and coherence. The key contributions of this paper are as follows.

- Decomposing complex network modeling into rapid network situation evaluation and problem-oriented model construction. We propose a hierarchical digital twin framework featuring an adaptive layered architecture designed explicitly for efficient network orchestration. At higher layers, coarse-grained digital twins perform rapid system-level network evaluations, allowing scalable selection of target areas and critical network attributes that significantly influence orchestration outcomes. At lower layers, fine-grained digital twins are constructed for the selected elements, providing detailed and predictive strategies for user association and power allocation, thus ensuring precise QoS satisfaction. This hierarchical decomposition approach enables prioritized, adaptive, and efficient management, effectively addressing dynamic service requirements in complex 6G HetNets.
- Prioritizing network attributes based on their modeling benefits and costs in supporting network orchestration at higher layers. An in-depth analysis of modeling values is conducted to categorize network attributes based on their real-time relevance and modeling complexity. Specifically, the modeling benefits of attributes are evaluated using temporal correlation analysis and causality inferences, while their associated costs are assessed based on data complexity quantified by sample entropy. This prioritization approach adapts to real-time network dynamics by allocating more resources to model highly valued attributes at higher layers for rapid network situation assessment and problem identification. Conversely, attributes with a lower value are selectively modeled at lower layers with finer granularity, where additional resources can be allocated for comprehensive analysis. By differentiating and prioritizing attributes, this value-oriented mechanism ensures the efficient inclusion of essential elements in digital twin modeling, dramatically reducing resource wastage during network orchestration.
- Implementing multi-level synchronization for accurate temporal alignment among various digital twins to support efficacious model utilization at the lower layers. This synchronization process begins at the device level, addressing temporal discrepancies caused by clock inaccuracies and diverse sampling capabilities. Subsequently, fine-grained digital twins are developed by leveraging the nonlinear autoregressive exogenous neural network combined with seasonal and trend decomposition techniques to accurately capture network dynamics. Finally, model-level synchronization is employed to compensate for delays between data collection and digital twin utilization, ensuring, the timely and accurate reflection of real-time network conditions.

The remainder of this paper is organized as follows. In Section II, we formulate the network orchestration problem for complex 6G HetNets and introduce the hierarchical digital twin framework. Higher-layered adaptive attribute selection and lower-layered scalable network modeling are discussed in Sections III and IV, respectively. In Section V, simulation results are presented to evaluate the effectiveness of the proposed method, followed by the conclusion in Section VI.

Notations: In this paper, italic letters denote scalars, while bold letters represent vectors. A summary of the primary symbols used can be found in Table I.

II. System Model

In this section, we formulate the problem of maximizing network orchestration efficiency in large-scale 6G HetNets. Given its inherent complexity, we further propose a hierarchical digital twin paradigm that efficiently addresses the problem by decomposing it into manageable components.

A. Heterogeneous Networks Model

Consider the 6G HetNets depicted in Fig. 1(a), where a total of M BSs $\mathcal{M} = \{1, \ldots, M\}$ with distinct locations and transmission powers serve N mobile users $\mathcal{N} = \{1, \ldots, N\}$ with

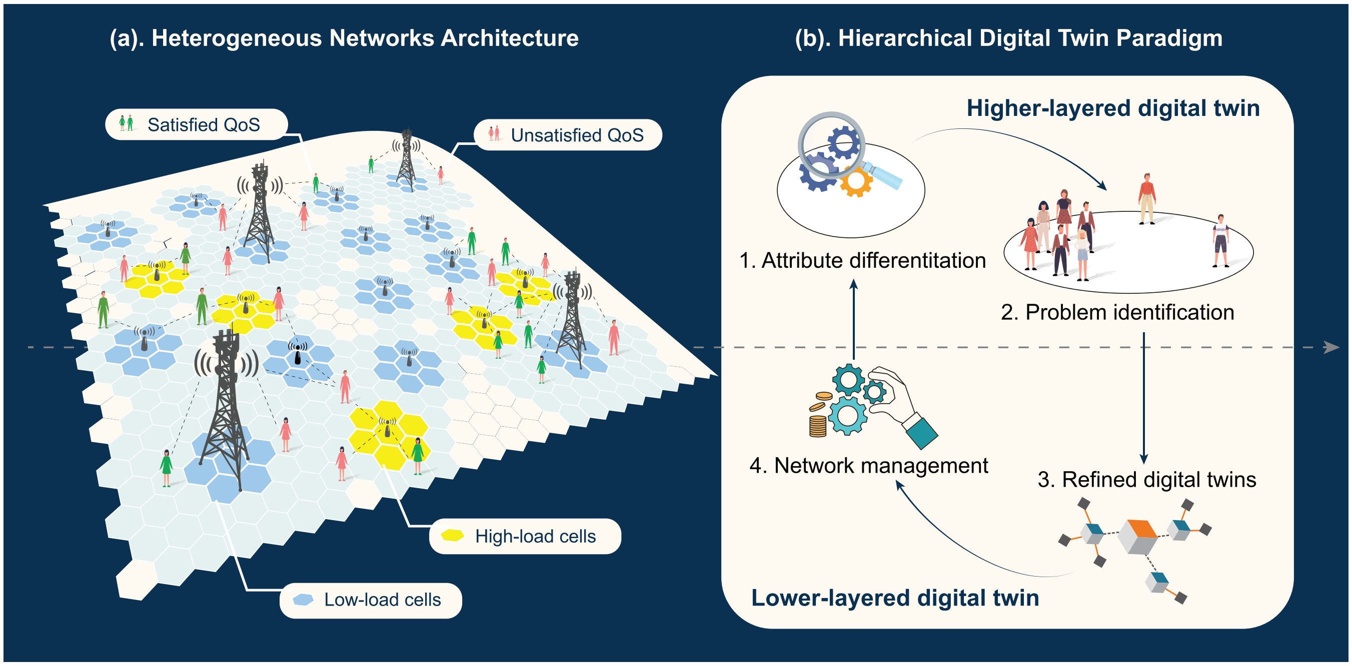

Fig. 1. The overall architecture of the 6G HetNets orchestrated by the proposed hierarchical digital twin paradigm: (a). The overall 6G HetNets consists of macro-cell and small-cell BSs serving distributed users to accommodate their diverse QoS requirements. The main contributors to unsatisfied QoS include unbalanced traffic load distribution and insufficient network capacity for local cells. (b). The proposed hierarchical digital twin paradigm for efficient network orchestration comprises higher layers for attribute differentiation and network problem identification as well as lower layers for problem-focused modeling and network management.

TABLE I
MAJOR NOTATIONS AND DEFINITIONS

Notation	Definition
$a_{i,j}$	Association factor between BS j and user i
$A_{i,j}$, c_i	Award and cost during user assignment
$\mathcal{A}$	Adjacency matrix in graph partitioning
$\mathcal{C}_{i,k}$	Digital twin modeling cost for attribute k
$\mathcal{E}$	Network orchestration efficiency
$\mathcal{L}_j$	Load indicator at BS j
M, N	Number of BSs and users
$\mathcal{N}_s$, $\mathcal{K}_s$	Selected set of users and attributes
$\mathcal{M}$, $\mathcal{N}$, $\mathcal{K}$	Entire set of BSs, users, and attributes
$p_{i,j}$	Transmission power from BS j to user i
Q_i^n, $\hat{Q}_i^n$	The nth QoS condition and QoS requirement
$\mathcal{Q}$	Set of QoS metrics
$r_{i,j}$	Instantaneous data rate
$\mathcal{S}$	Total samples for one attribute
$\mathcal{T}$	Sampling timestamps
$\mathcal{Z}$	Identified target area
$\delta_{i,j}^T$	Traffic between i and j over period T
Φ, Θ	Optimization award and cost using digital twins
$\hat{\rho}$	Digital twin modeling benefit
ξ_i^n	The nth QoS satisfaction indicator

diverse QoS requirements. The network can be represented by a graph $G = (\mathcal{V}, E)$, where the vertices set $\mathcal{V} = \{\mathcal{M}, \mathcal{N}\}$ represents all network entities, and the edge set E denotes communication links between users and their associated BSs.

According to the current user association strategy, each user is served by a single BS at a time. Frequent handover between different BSs can significantly increase network traffic complex. Within this context, highly loaded BSs and QoS-unsatisfied users arising from suboptimal orchestration become bottlenecks for the network. To solve these issues, data traffic and interference must be carefully managed by adapting user association and power allocation strategies according to the network situation to fulfill stringent QoS requirements for all users. The orchestration performance can thus be evaluated by assessing the satisfaction of application-specific QoS requirements for all users, e.g., throughput, packet loss rate, and transmission delay. Specifically, the fulfillment of the *n*th QoS requirement for user *i*, denoted by $Q_{i,j}^n$, is closely linked to the instantaneous downlink data rate from its associated BS *j*, defined as

$$r_{i,j} = B\log_2\left(1 + \frac{p_{i,j}|h_{i,j}|^2}{\sum_{l\in\mathcal{M},l\neq j} p_{i,l}|h_{i,l}|^2 + \sigma_0^2}\right), \quad (1)$$

where B and σ_0 represent bandwidth and background noise power, respectively. The signal strength of user *i*, $p_{i,j}|h_{i,j}|^2$, is determined by the transmission power $p_{i,j}$ from its associated BS *j* and the Rayleigh flat fading channel coefficient $h_{i,j}$, which is further affected by the path loss model and distance. Here, we only consider inter-cell interference $\sum_{l\in\mathcal{M},l\neq j} p_{i,l}|h_{i,l}|^2$ with universal frequency reuse.

Based on the instantaneous data rate obtained in (1), the satisfaction level for different QoS requirements of user *i* can be evaluated by

$$S_i^n = Q_i^n/\widehat{Q}_i^n, \quad (2)$$

where $\widehat{Q}_i^n$ is the *n*th QoS requirement associated to user *i*. The *n*th condition $Q_i^n = \sum_{j\in\mathcal{M}} Q_{i,j}^n$ relates to the quality of link *i*-*j* for each association. QoS conditions can be evaluated based on various network attributes related to link quality and network situations. Taking achievable data rate as an example, the corresponding QoS for user *i* during network orchestration, denoted by Q_i^{dr}, can be written as

$$Q_i^{dr} = \sum_{j \in \mathcal{M}} a_{i,j} r_{i,j}, \tag{3}$$

where $a_{i,j}$ is a binary variable indicating the association between user *i* and BS *j*. Similarly, other QoS metrics, e.g., network delay and packet losses can be derived based on real-time network conditions, indicating the satisfaction of the application requirements. According to the assigned application, different users have unique QoS metrics, thus requiring a scalable and user-specific network orchestration strategy to ensure QoS satisfaction.

B. Problem Formulation

In satisfying diverse user QoS demands, selecting the appropriate BS for data transmission with optimal transmission power is critical. Therefore, in this paper, our network orchestration focuses on user association factor $a_{i,j}$ and transmission power $p_{i,j}$ for each user to ensure overall satisfied QoS requirements. Meanwhile, real-time network orchestration requires frequent data transmission and processing from distributed users for network optimization, leading to excessive consumption of limited network resources. Typically, for the same achievable QoS satisfaction level, an algorithm requiring less data transmission leads to a faster response to the dynamic network situation and efficient utilization of network resources, thus bringing higher benefits for the entire network.

Based on this observation, we formulate the problem to maximize the QoS satisfaction efficiency, defined as the QoS satisfaction level over the resource consumed during network orchestration. For clarity, we define the user association matrix as $\mathbf{a} = \{a_{i,j} | \ \forall i \in \mathcal{N}, \forall j \in \mathcal{M}\}$ and power allocation matrix as $\mathbf{p} = \{p_{i,j} | \ \forall i \in \mathcal{N}, \forall j \in \mathcal{M}\}$, respectively. The problem of network orchestration can be written as

$$\begin{aligned}
\mathcal{P}\mathbf{1} : \max_{\mathbf{a},\mathbf{p}} \ & \mathcal{E}_Q(\mathbf{a}, \mathbf{p}) = \frac{\sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{M}} \sum_{n \in \mathcal{Q}} a_{i,j} \omega_i^n S_i^n(\mathbf{p})}{\sum_{i \in \mathcal{N}} \sum_{n \in \mathcal{Q}} \boldsymbol{\alpha} \mathcal{R}_i^n} \\
\text{s.t.} \quad & \text{C1} : a_{i,j} \in \{0, 1\}, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}, \\
& \text{C2} : \sum_{j \in \mathcal{M}} a_{i,j} \leq 1, \quad \forall i \in \mathcal{N}, \\
& \text{C3} : \sum_{i \in \mathcal{N}} a_{i,j} \leq 1, \quad \forall j \in \mathcal{M}, \\
& \text{C4} : p_{i,j} \geq 0, \quad \forall i \in \mathcal{N}, \forall j \in \mathcal{M}, \\
& \text{C5} : \sum_{i \in \mathcal{N}} a_{i,j} p_{i,j} \leq P_{j,\max}, \quad \forall j \in \mathcal{M}, \\
& \text{C6} : S_i^n \geq 1, \quad \forall i \in \mathcal{N}, \forall n \in \mathcal{Q},
\end{aligned} \tag{4}$$

where the set $\mathcal{Q} = \{1, \ldots, Q^N\}$ is a collection of all QoS metrics weighted by ω_i^n according to the application preference. The scaling vector $\boldsymbol{\alpha}$ is a large number for users with satisfied QoS requirements, thus ensuring prioritized resource allocation for users with $S_i^n < 1$. C1 to C3 are the constraints of the user association basis. C4 and C5 are related to the transmission power of each BS, limited by its maximized power $P_{j,\max}$. Moreover, C6 ensures the QoS satisfaction for all users during network orchestration.

Furthermore, $\mathcal{R}_i^n$ in (4) is the network resource consumed to transmit the network attribute data related to user *i* regarding the *n*th QoS metrics. By adopting different optimization methods, e.g., integer relaxation and problem decomposition [26], the complex network orchestration problem could be solved. However, the overall resource consumption due to the need for all-inclusive collection of current network data from all users and attributes increases exponentially to orchestrate the entire network. In contrast, the improvement of QoS satisfaction for different users and QoS metrics is not identical throughout the network. As a result, the current method that treats all users and attributes equally could significantly degrade the real-time network orchestration efficiency.

C. Digital-Twin-Assisted Predictive Assignment Problem

Substantially, $\mathcal{P}\mathbf{1}$ is a nonlinear mixed integer programming problem that is complex to solve. For clarity, we express $\mathcal{P}\mathbf{1}$ as an assignment problem between *N* users and *M* BSs. Each edge *i*-*j* represents one agent-task assignment between user *i* and BS *j* with achievable QoS Q_i^n under the current network orchestration strategy $\{\mathbf{a}, \mathbf{p}\}$. In other words, different assignment strategies lead to various assignment awards $A_{i,j} = \sum_{n \in \mathcal{Q}} \omega_i^n S_i^n(\mathbf{p})$ and assignment costs $c_i = \sum_{n \in \mathcal{Q}} \boldsymbol{\alpha} \mathcal{R}_i^n$, so that $\mathcal{P}\mathbf{1}$ can be rewritten as

$$\begin{aligned}
\mathcal{P}\mathbf{2} : \quad \max_{\mathbf{a},\mathbf{p}} \ & \mathcal{E}_Q(\mathbf{a}, \mathbf{p}) = \frac{\sum_{(i,j) \in \mathcal{N} \times \mathcal{M}} a_{i,j} A_{i,j}}{\sum_{i \in \mathcal{N}} c_i} \\
\text{s.t.} \quad & \text{C1} - \text{C6}.
\end{aligned} \tag{5}$$

However, due to the combination of binary variable $\mathbf{a}$ and continuous variable $\mathbf{p}$ as well as the joint consideration of assignment reward $A_{i,j}$ and cost c_i, the discrete and nonlinear nature of $\mathcal{P}\mathbf{2}$ limits the available solutions. Existing studies mainly rely on heuristic algorithms or reinforcement learning [27], [28], which are likely to induce significant processing complexity with sub-optimal solutions, thus degrading the timeliness and effectiveness of decision-making for real-time network orchestration.

By contrast, a digital twin-assisted linear mixed-integer programming method was discussed in [3] to maximize QoS satisfaction through user association, where statistical network attributes are modeled to predict the assignment award for different association strategies according to historical user information. The use of prediction can dramatically reduce the need for real-time network data upload, thus improving the network orchestration timeliness and efficiency. Nevertheless, with the increment of network scales and the increasing number of network attributes, the construction of network digital twins could overwhelmingly consume limited network resources and become unaffordable for network orchestration.

Motivated by this observation, the nonlinear problem $\mathcal{P}\mathbf{2}$ can be considered as maximizing the network-wide digital-twin-assisted network orchestration efficiency, with award of using digital twins $\Phi(\mathbf{a}, \mathbf{p}) = \sum_{(i,j) \in \mathcal{N} \times \mathcal{M}} a_{i,j} A_{i,j}$ and corresponding construction cost $\Theta(\mathcal{N}, \mathcal{K}) = \sum_{i \in \mathcal{N}} \sum_{k \in \mathcal{K}} \mathcal{C}_{i,k}$, where $\mathcal{C}_{i,k}$ is the cost for digital twinning the *k*th attribute in user i, while $\mathcal{K}$ is the set of all network attributes. However, the relationship between Φ and Θ is not intuitive, where Φ depends on the various $\{\mathbf{a}, \mathbf{p}\}$ scenarios, while Θ is determined by the orchestration target. Therefore, it is critical to identify *target areas* $\mathcal{Z}$ within the network for optimization, which aligns with Pareto efficiency principles. By focusing on modeling these areas, the orchestration efficiency can be enhanced significantly, ensuring that improvements are made where they yield the greatest benefit without compromising the performance of other unselected areas.

D. Hierarchical Digital Twin for Efficient Orchestration

To identify target areas $\mathcal{Z}$ and achieve efficient network orchestration, we propose a novel hierarchical digital twin paradigm, as illustrated in Fig. 1(b), that decomposes and solves problem $\mathcal{P}\mathbf{2}$ by analyzing and predicting time-series network data. Specifically, two distinct layers are established at the cloud center with different granularities and functionalities.

- ***Higher-layered digital twin*** **(HDT)** rapidly identifies target areas $\mathcal{Z}$ through system-level situation evaluation and differentiated attribute modeling. Instead of creating detailed models, HDT focuses on constructing simplified, coarse-grained digital twins for highly valued attributes across the network, significantly reducing modeling complexity. Real-time data from other attributes are selectively utilized to efficiently evaluate overall network conditions and identify problematic users and areas requiring deeper analysis.
- ***Lower-layered digital twin*** **(LDT)** develops fine-grained digital twin models specifically for the identified target areas $\mathcal{Z}$ and selected attributes. To enable detailed orchestration, advanced data processing and modeling techniques are employed to increase digital twin granularity and precision. LDT also implements multi-level synchronization mechanisms to ensure accurate temporal alignment between virtual and physical realms. By narrowing its scope to intentionally selected elements, LDT substantially enhances orchestration efficiency, offering a problem-centered approach to network management.

More specifically, variations in user activity patterns and application demands result in differing benefits and costs associated with digital twin modeling of various users and network attributes. Therefore, to identify target areas comprising users and attributes that lead to the most efficient orchestration outcome, the objective function for HDT can be written as

$$!\mathcal{P}3 : \min_{\mathcal{N},\mathcal{K}} \Theta(\mathcal{N}, \mathcal{K}). \tag{6}$$

Define $\mathcal{N}_s = \{1, \cdots, N_s\}$ and $\mathcal{K}_s = \{1, \cdots, K_s\}$ as the set of users and attributes selected, i.e., $\{\mathcal{N}_s, \mathcal{K}_s\} = \mathrm{argmin}_{\mathcal{N},\mathcal{K}} \Theta(\mathcal{N}, \mathcal{K})$, the modeling cost $\widehat{\Theta}$ can be given by

$$\widehat{\Theta} = \sum\nolimits_{i\in\mathcal{N}_s} \sum\nolimits_{k\in\mathcal{K}_s} \mathcal{C}_{i,k}. \tag{7}$$

It is worth noting that this selection process is strictly constrained by the QoS requirements and resource availability within the selected areas, as detailed further in Section III.

Then, LDT aims to maximize QoS satisfaction for selected users $\mathcal{N}_s$ based on attributes $\mathcal{K}_s$, given by

$$\begin{aligned}
\mathcal{P}\mathbf{4}: \quad & \max_{\mathbf{a},\mathbf{p}} \ \mathcal{E}_Q(\mathbf{a},\mathbf{p}) = \frac{1}{\widehat{\Theta}} \sum\nolimits_{(i,j)\in\mathcal{N}_s\times\mathcal{M}} a_{i,j} A_{i,j} \\
\text{s.t.} \quad & \mathrm{C1}: a_{i,j} \in \{0,1\}, \quad \forall i \in \mathcal{N}_s, \forall j \in \mathcal{M}, \\
& \mathrm{C2}: \sum_{j\in\mathcal{M}} a_{i,j} \le 1, \quad \forall i \in \mathcal{N}_s, \\
& \mathrm{C3}: \sum_{i\in\mathcal{N}_s} a_{i,j} \le 1, \quad \forall j \in \mathcal{M}, \\
& \mathrm{C4}: p_{i,j} \ge 0, \quad \forall i \in \mathcal{N}_s, \forall j \in \mathcal{M}, \\
& \mathrm{C5}: \sum_{i\in\mathcal{N}_s} a_{i,j} p_{i,j} \le P_{j,\max}, \quad \forall j \in \mathcal{M}, \\
& \mathrm{C6}: S_i^n \ge 1, \quad \forall i \in \mathcal{N}_s, \forall k \in \mathcal{Q},
\end{aligned} \tag{8}$$

which becomes a simplified assignment problem since the minimum modeling cost $\widehat{\Theta}$ is considered constant at LDT. The accurate solution to $\mathcal{P}\mathbf{4}$ is based exclusively on the fine-grained modeling of the network attributes involved $\mathcal{K}_s$ for each user, requiring more advanced data processing and modeling techniques as illustrated in Section IV.

Remark: Problem $\mathcal{P}\mathbf{2}$ prioritizes modeling costs through the scaling vector $\boldsymbol{\alpha}$, effectively regulating resource usage for users already meeting their QoS requirements. Consequently, an optimal solution to $\mathcal{P}\mathbf{2}$ can equivalently be achieved by first identifying unsatisfied users and critical attributes via problem $\mathcal{P}\mathbf{3}$, followed by solving the QoS maximization problem $\mathcal{P}\mathbf{4}$ for these identified components.

III. Problem Identification at the Higher Layers With Adaptive Attribute Selection

The essential role of HDT in identifying $\mathcal{N}_s$ and $\mathcal{K}_s$ among the massive users and heterogeneous attributes within complex 6G HetNets can be achieved by solving $\mathcal{P}\mathbf{3}$. To accommodate the dynamic nature of network situations and user activities during the identification, we develop an adaptive attribute selection algorithm, which enhances the efficiency of network evaluation and problem discovery, as shown in Fig. 2.

A. Value-Oriented Network Attribution Selection

Given the inherent heterogeneity and diversity of network attributes, all-inclusive modeling during digital twin construction is inefficient and impractical due to the varying values of these attributes to the assigned application. Therefore, we propose a value-oriented attribute differentiation mechanism to select the most pertinent attributes $\mathcal{K}_s$ within the network according to its actual demands. Moreover, since HDT is used for rapid problem identification, simple yet effective methods are preferred to enhance efficiency and responsiveness.

As depicted in Fig. 2(a), the value of each network attribute is evaluated by its modeling cost and benefit. Specifically, the modeling cost of an attribute depends on its predictability, which is influenced by the data complexity and irregularity. For a series of X_N data $\mathbf{x_{i,k}} = \{x_{i,k}^l \,|\, l = 1, \ldots, X_N\}$ used

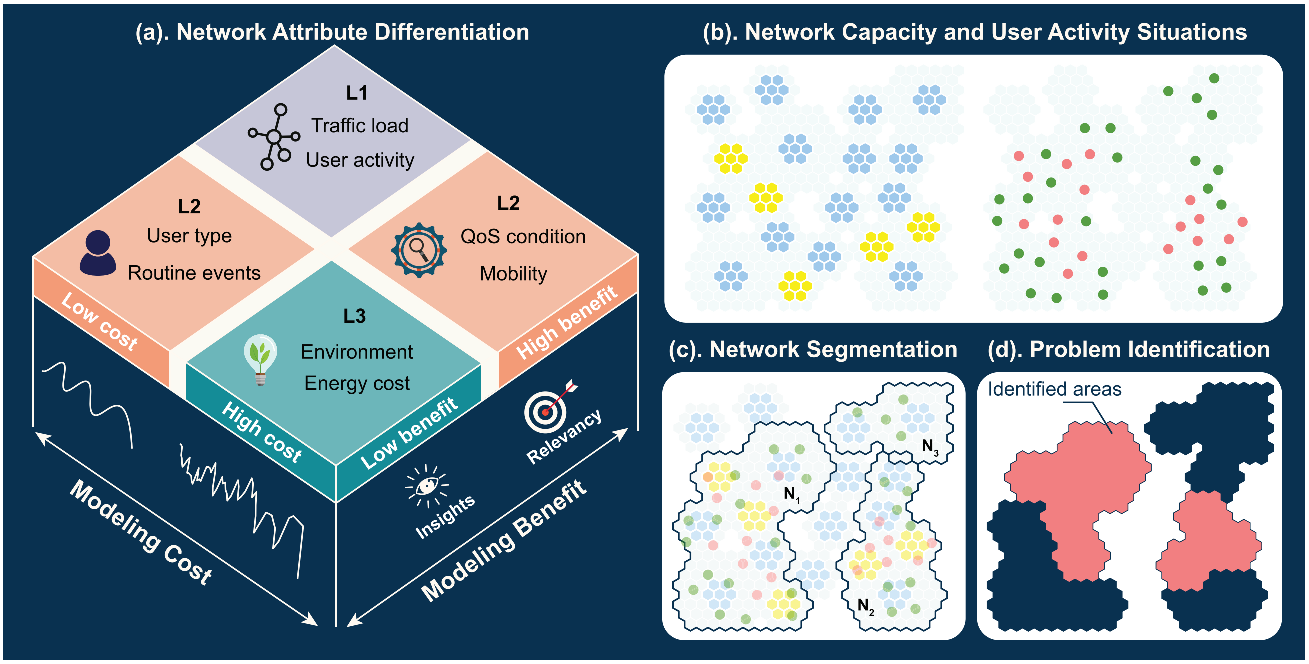

Fig. 2. The main functions of higher-layered digital twins for rapid network situation evaluation: (a). Differentiation and selection of various network attributes, which are classified into three categories based on the cost and benefit of constructing digital twins. (b). The overall network situation, including the network capacity observed by the distributed BSs and the user activities according to associated traffic. (c). Network segmentation by analyzing historical user activity and potential BS-user connections. (d). The identification of target areas in each network segment based on the expected network orchestration efficiency.

to model the kth attribute of user i, the modeling cost can be quantified using sample entropy [29], given by

$$\mathrm{SampEn}_{i,k} = -\ln\left(\frac{\Psi_\gamma^{m+1}(\mathbf{x_{i,k}})}{\Psi_\gamma^{m}(\mathbf{x_{i,k}})}\right), \tag{9}$$

where $\Psi_\gamma^m(\cdot)$ is the conditional probability that a sequence of m consecutive data $\mathbf{x_{i,k}^{m,u}} = \{x_{i,k}^l \,|\, l = u, \ldots, u+m-1\}$ in $\mathbf{x_{i,k}}$ remains similar within a predefined tolerance level γ, which is set as a fixed proportion (e.g., 0.2) of the standard deviation of the data series $\mathbf{x_{i,k}}$. Then, this probability is calculated as

$$\Psi_\gamma^m(\mathbf{x_{i,k}}) = \frac{1}{X_N - m}\sum_{u=1}^{X_N-m}\sum_{v=1,v\neq u}^{X_N-m}\mathbf{1}_{d(\mathbf{x_{i,k}^{m,u}},\mathbf{x_{i,k}^{m,v}})<\gamma}, \tag{10}$$

where $\mathbf{1}$ is a binary indicator that counts the number of sequences in $\mathbf{x_{i,k}}$ similar to $\mathbf{x_{i,k}^{m,u}}$, and $d(\cdot)$ represents the Chebyshev distance, which calculates the maximum absolute difference between elements in two data sequences, defined as

$$d(\mathbf{x_{i,k}^{m,u}},\mathbf{x_{i,k}^{m,v}}) = \max_l(|x_{i,k}^{m,u}(l) - x_{i,k}^{m,v}(l)|). \tag{11}$$

With more sequences of data showing similarity in the kth attribute, i.e., $d < \gamma$, a smaller sample entropy $\mathrm{SampEn}_{i,k}$ is expected, indicating less complex data and a lower modeling cost for that attribute. Similarly, $\mathrm{SampEn}_{i,k}$ increases in the presence of data variabilities, such as outliers, noise, and non-linearities. These factors contribute to higher unpredictability in the data, which, as a direct consequence, leads to an increase in the digital twin construction cost.

In addition to the modeling cost determined by data complexity, it is equally important to evaluate the modeling benefit that each network attribute can bring. This benefit is primarily assessed based on the relevance of the attribute to the operational objective. As the scope of this study is to maximize QoS satisfaction, for the same series of data $\mathbf{x_{i,k}}$, its modeling benefit can be initially reflected by its correlation with the nth QoS situation $\mathbf{Q_i^n} = \{Q_i^{n,l} \,|\, l = 1, \ldots, X_N\}$. To reduce complexity in HDT, we first adopt the Pearson correlation coefficient to rapidly evaluate the modeling benefit, given by

$$\rho_{i,k}^n = \frac{\sum_{l=1}^{X_N}(x_{i,k}^l - \bar{x}_{i,k})(Q_i^{n,l} - \bar{Q}_i^n)}{\sqrt{\sum_{l=1}^{X_N}(x_{i,k}^l - \bar{x}_{i,k})^2\sum_{l=1}^{X_N}(Q_i^{n,l} - \bar{Q}_i^n)^2}}, \tag{12}$$

where $\bar{x}_{i,k}$ and $\bar{Q}_i^n$ represent the mean values of each attribute and QoS situation, respectively. A higher $|\rho_{i,k}^n|$ indicates a stronger relationship between the kth attribute and nth objective, suggesting a higher benefit of modeling this particular attribute in user i.

It is worth noting that $\rho_{i,k}^n$ only provides a basic observation of the data relevance, reflecting primarily linear associations. With a large $|\rho_{i,k}^n|$, e.g., $|\rho_{i,k}^n| > 0.7$ for the kth attribute, it is worthwhile to further validate its relevancy through causality inference [30], which can be applied to obtain a deeper insight into the causal relationship among multiple attributes. For instance, Granger causality can be estimated by comparing the coarse-grained digital twin models for $\mathbf{Q_i^n}$ after adding $\mathbf{x_{i,k}}$, i.e., $\tilde{Q}_i^n(X_N+1) = f_1(\mathbf{Q_i^n})$ and $\check{Q}_i^n(X_N+1) = f_2(\mathbf{Q_i^n}, \mathbf{x_{i,k}})$, where f_1 and f_2 are the modeling of the two cases typically through auto-regression. Based on the predictive power of f_1 and f_2 to compare the modeling accuracy, the causality relationship between $\mathbf{Q_i^n}$ and $\mathbf{x_{i,k}}$ can be obtained to update the modeling benefit $|\hat{\rho}_{i,k}^n|$. Finally, by considering all QoS $\mathcal{Q}$, the modeling benefit for each attribute can be derived as

$|\hat{\rho}_{i,k}| = |\sum_{n\in\mathcal{Q}} \omega_i^n \hat{\rho}_{i,k}^n|$. This combined approach ensures that the preliminary associations identified by $\rho_{i,k}^n$ for each attribute are rigorously validated through causality inference, which enables a more precise and comprehensive assessment of their contributions to operational objectives.

After obtaining the modeling cost and benefit for each attribute, we can differentiate them into distinct groups for inclusion or exclusion during digital twin modeling. For the kth attribute from user i, its modeling value $\mathbf{v}_{i,k} = (\mathrm{SampEn}_{i,k}, |\hat{\rho}_{i,k}|)$ can be used during differentiation to ensure the formation of several distinct groups with similar modeling values. This can be achieved by various methods, e.g., the K-means clustering algorithm, where the objective function of differentiation is given by

$$\min_{\mathcal{G}} \quad \sum\nolimits_{j=1}^{K} \sum\nolimits_{\mathbf{v}_\kappa \in \mathcal{G}_j} |\mathbf{v}_\kappa - \psi_j|^2, \tag{13}$$

which aims to optimally obtain K clusters that contain similar values. In (13), ψ_j is the jth centroid randomly selected from all values during initialization, while $\mathcal{G}_j$ is the collection of the values in cluster j, i.e., $\mathcal{G}_j = \{\mathbf{v}_\kappa | \|\mathbf{v}_\kappa - \psi_j\| \le \|\mathbf{v}_\kappa - \psi_i\|, \forall i < K\}$. After a few iterations, all attributes can be differentiated into a few groups based on their modeling value for adaptive attribute selection during digital twin construction.

Algorithm 1 Value-Oriented Attribute Differentiation

```
Attribute value estimation:
for each user i ∈ N do
    for each attribute k ∈ K do
        Collect data samples x_{i,k}
        Obtain a sequence of m data x_{i,k}^{m,p} from x_{i,k}
        Calculate modeling cost SampEn_{i,k} according to
        Eq. (9) – (11)
        Collect objective data samples Q_i^n
        Calculate modeling benefit ρ_{i,k}^n by Eq. (12)
        if |ρ_{i,k}^n| < 0.3 then
            Test Granger causality for Q_i^n and x_{i,k}
            Update |ρ̂_{i,k}^n| based on causality
        Record modeling value v_{i,k} = (SampEn_{i,k}, |ρ̂_{i,k}|)
Value-oriented differentiation:
for each cluster j = 1 : K do
    repeat
        Update centroid ψ_j as the mean of all modeling
        values in cluster j
        for each attribute value v_κ do
            Assign v_κ to the nearest centroid
    until centroids ψ_j converge
Differentiate attributes into L1 to L3 based on clusters
```

Remark: The proposed Algorithm 1 differentiates heterogeneous attributes by assessing their modeling cost and benefit specific to the applications, enabling value-oriented attribute selection for network modeling. In this study, as an illustrative example, we categorize all attributes into three levels with varying selection priorities, based on their modeling values evaluated through $\mathrm{SampEn}_{i,k}$ and $|\hat{\rho}_{i,k}|$ for different users. Some specific scenarios are outlined below:

- **L1**: High-valued attributes, e.g., traffic load and user activity, which are usually long-term, predictable, and highly relevant to the application. These attributes are prioritized in the digital twin representation for fast problem identification and are modeled in HDT.
- **L2**: Medium-valued attributes with a lower value due to high modeling cost or low modeling benefit, like QoS conditions and user mobility. The real-time data of these dynamic or complex attributes are directly used in HDT for immediate monitoring and analysis. Meanwhile, depending on the available processing capabilities, they are also selectively modeled in LDT for long-term analysis. Typically, $\mathcal{K}_s = \{\mathbf{L_1} \cup \mathbf{L_2}\}$ are attributes selected, adapting to the network situation for the entire hierarchical digital twin paradigm.
- **L3**: Other irrelevant or unpredictable attributes, such as random network outages, are excluded from modeling due to their low values for network orchestration.

Given the limited processing capabilities and varying modeling benefits, the flexibility in handling **L2** attributes becomes crucial in complex scenarios. By selectively modeling the **L2** attributes that provide the highest value, and taking into account user identification as discussed in the following sections, we ensure more reasonable and adaptive attribute selection in different environments. This differentiated digital twin modeling and data collection method enables a more strategic selection of different attributes, relieving the communication and computation burdens in network orchestration.

B. User Activity-Guided Network Analysis and Segmentation

In large-scale HetNets, BSs are ubiquitously deployed to facilitate seamless connectivity, while the inherent spatial distribution of user activities usually results in a non-uniform traffic load distribution. For BS j, any remote users i with $a_{i,j}(t) = 0, \forall t \ge 0$ neither establish connectivity nor consume network resources from j, thereby rendering them non-essential entities for j. Based on this observation, the entire network can be segmented into non-overlapping sub-networks for rapid situation evaluation, where the users in different sub-networks remain insulated.

Given the user-centric traffic distribution, **L1** attributes like historical user activity patterns and user-BS connectivity are paramount to support network segmentation. To analyze the user-BS dynamics, the traffic $\delta_{i,j}^T$ over a period T can be quantified by

$$\delta_{i,j}^T = \sum_{t=0}^{T} a_{i,j}^t r_{i,j}^t, \tag{14}$$

which leads to load analysis in illustrating the demand-supply discrepancy of BS j. A binary load indicator can be given by

$$\mathcal{L}_j = \begin{cases} 0, & \text{if} \quad \sum_{i\in\mathcal{N}} \delta_{i,j}^T < \zeta \mathcal{S}_j, \\ 1, & \text{otherwise}, \end{cases} \tag{15}$$

where $\mathcal{S}_j$ is the theoretical network capacity in BS j while ζ is the threshold level for indicating high loads.

To achieve efficient segmentation, graph partitioning is adopted to divide the entire network graph based on user activity. More specifically, for the entire network G, let $U_l =$

$\{u_1, \ldots, u_s\}$ be the lth segment, which can be obtained by minimizing the ratio between the cut and weight, given by

$$\min_{U_i} \quad \sum_{i\in\mathcal{N}} 1 - \frac{f_c(U_i, U_i)}{f_c(U_i, \mathcal{V})}, \tag{16}$$

where $f_c(\cdot)$ is the cut function within the graph, defined as

$$f_c(U_a, U_b) = \sum_{u\in U_a}\sum_{v\in U_b} \mathcal{A}_{u,v}(\tau). \tag{17}$$

The adjacency matrix $\mathcal{A}(\tau)$ is used to represent the weight for each edge in the graph G at instant τ, given by

$$\mathcal{A}_{i,j}(\tau) = \begin{cases} 0, & \text{if } i = j, \\ \mathcal{X}_{i,j} \quad \underbrace{\sum_l \phi_l \widetilde{\delta^T_{i,j}(\tau - l)}}_{\text{user traffic modeling in HDT}}, & \text{otherwise,} \end{cases} \tag{18}$$

where $\boldsymbol{\mathcal{X}} \in \mathbb{R}_{M\times N}$ is the potential link matrix among all users and BSs, with $\mathcal{X}_{i,j} = 1$ if BS j is available for user i and 0 otherwise. User traffic modeling is achieved through auto-regression, where $\widetilde{(\cdot)}$ denotes the normalization to ensure $0 \le \mathcal{A}_{i,j}(\tau) \le 1$, while ϕ is the digital twin coefficients of user demand $\mathcal{D}_{i,j}(\tau) := \sum_l \phi_l \delta^T_{i,j}(\tau - l)$ obtained in HDT for data traffic $\delta^T_{i,j}$ illustrated in (14).

C. Demand-Supply Analysis for Target Area Identification

By focusing on each sub-network after segmentation, situations of the entire network can be efficiently evaluated. To avoid unnecessary resource consumption and response delays, real-time data of short-term attributes (**L2**), such as instantaneous link quality, QoS satisfaction, and user association, are used to analyze the current network states.

Let $\mathcal{Z}$ denote the *target area* identified within each network segment, which is an area with $N_\mathcal{Z}$ users and $M_\mathcal{Z}$ BSs leading to high orchestration outcomes. First, we use orchestration efficiency to identify users for inclusion, defined as the expected QoS improvement for all users after orchestration over the resource consumed for digital twin, given by

$$\max_{y_i} \quad \eta = \frac{\sum_{i\in\mathcal{N}}\sum_{n\in\mathcal{Q}} y_i \omega^n_i (1-\xi^n_i)\mathcal{B}^n_i}{\sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}_s} y_i \mathcal{C}_{i,k}}. \tag{19}$$

In (19), the numerator denotes the overall expected improvements throughout $\mathcal{Z}$ with y_i as a binary variable indicating the inclusion of user i. $\mathcal{B}^n_i = \max(0, 1 - S^n_i)$ is the expected orchestration benefit, while ξ^n_i is the QoS satisfaction indicator of user i, given by

$$\xi^n_i = \begin{cases} 0, & \text{if } S^n_i < 1, \\ 1, & \text{otherwise,} \end{cases} \tag{20}$$

where the QoS satisfaction condition S^n_i is based on the instantaneous network situations that are directly used in HDT. Moreover, the denominator of (19) represents the resource consumption for establishing the digital twins to support the orchestration, with $\mathcal{C}_{i,k}$ as the resource used for digital twin construction for the kth attribute of user i.

On the other hand, the selection of target areas should guarantee sufficient resources to satisfy the user demands. For BS j, based on the predicted user demand $\mathcal{D}_{i,j}$ and network capacity $\mathcal{S}_j$ from HDT, (19) should be further constrained by

$$\sum_{i\in\mathcal{Z}} y_i \sum_{j\in\mathcal{Z}} \mathcal{D}_{i,j} \le \sum_{j\in\mathcal{Z}} y_j \mathcal{S}_j \sum_{i\in\mathcal{Z}} \mathcal{X}_{i,j}/M_\mathcal{Z} N_\mathcal{Z}, \tag{21}$$

where the scaling factor $\sum_{i\in\mathcal{Z}} \mathcal{X}_{i,j}/M_\mathcal{Z} N_\mathcal{Z}$ aims to ensure sufficient resources within $\mathcal{Z}$ for each BS j.

Problem (19) can be further converted into a graph partitioning problem after adjusting the adjacency matrix $\mathcal{A}_{i,j}$ according to current network situation ξ_i and $\mathcal{L}_i$. Specifically, a higher weight is assigned to unsatisfied users and lower-loaded BSs to ensure their selection during area identification. The weight matrix can be updated as

$$\widehat{\mathcal{A}}_{i,j}(\tau) = (1 - w_{i,j})\mathcal{A}_{i,j}(\tau - 1) + w_{i,j}\Delta\mathcal{A}_{i,j}(\tau), \tag{22}$$

where $w_{i,j}$ is the weighting factor, while $\Delta\boldsymbol{\mathcal{A}}(\boldsymbol{\tau})$ is used to update the adjacency matrix obtained based on real-time network situations evaluated by (15) and (20), written as

$$\Delta\mathcal{A}_{i,j}(\tau) = \begin{cases} \max\left(\dfrac{Q^n_i}{\widehat{Q}^n_i}, 1 - \dfrac{\sum_{i\in\mathcal{N}} \delta^T_{i,j}}{\zeta \mathcal{S}_j}\right), & \text{if } \xi_i \mathcal{L}_j = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

The two graph partitioning problems can be straightforwardly solved by adopting spectral methods [31]. To accommodate for the urgent QoS needs of critical users identified in (19), other relevant BSs and users with a strong connection indicated by $\mathcal{A}$ should also be included, finalizing the target area $\mathcal{Z}$. Unselected areas with sufficient QoS conditions should remain the current orchestration policy with continuous monitoring for future target area identification.

IV. Network Optimization at the Lower Layers Through Scalable Network Modeling

In this section, we focus on the scalable and fine-grained modeling of lower-layered digital twins to address the network orchestration problem $\mathcal{P}4$ for each target area $\mathcal{Z}$ identified from HDT. Additionally, with more attributes $\mathcal{K}_s$ involved in LDT modeling, the misalignment between physical and virtual domains becomes increasingly critical. Therefore, we propose a multi-level synchronization mechanism to enhance temporal consistency between the physical and virtual domains throughout the construction and utilization of digital twins.

A. Data Processing With Device-Level Synchronization

Temporal discrepancies between the physical world and the digital domain are generally tolerable in HDT due to their primary focus on system-level target identification. In contrast, LDT necessitates more detailed modeling of the elements within the identified area $\mathcal{Z}$ to facilitate precise decision-making. The accuracy of the modeling process in LDT is affected by both pre-model data inaccuracies and post-model misalignment of the digital twin, adversely impacting network orchestration performance. In light of these challenges, we propose a multi-level synchronization strategy for LDT, as depicted in Fig. 3, which addresses the issues of data disparity and model misalignment.

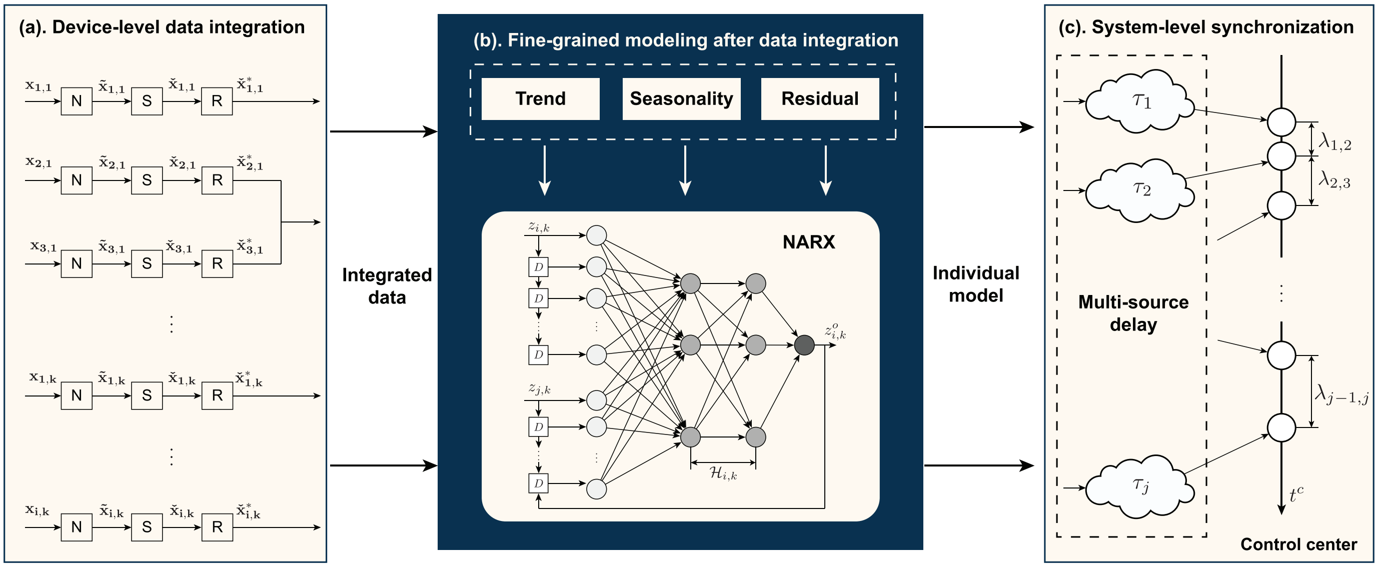

Fig. 3. Illustration of the accurate digital twin construction at the lower layers to support network orchestration: (a). Device-level data integration consists of data normalization, synchronization, and resampling to unify all data in the temporal domain. (b). Fine-grained modeling based on the integrated data using STL-NARX for individual or fused digital twins. (c). Multi-level synchronization for all digital twins to address the model misalignment induced by multi-source modeling delays.

Calculating awards $A_{i,j}$ in (5) relies on accurate temporal alignment among S_i^k for all users and their associated attributes. To maintain a unified understanding of each attribute in the control center, we design a three-step integration process to unify each data stream for device-level data integration before individual model construction.

1) Data Normalization: For the kth attribute of user i, its data $\mathbf{x}_{i,k}$ can be first scaled to a common range for enhanced comparability, given by $\tilde{\mathbf{x}}_{i,k} = (\mathbf{x}_{i,k} - \mu_{\mathbf{x}_{i,k}})/(\sigma_{\mathbf{x}_{i,k}})$, where $\tilde{\mathbf{x}}_{i,k}$ denotes the normalized value, while $\mu_{\mathbf{x}_{i,k}}$ and $\sigma_{\mathbf{x}_{i,k}}$ are the mean and standard deviation of $\mathbf{x}_{i,k}$.

2) Data Synchronization: Then, due to inherent variations in clock quality, normalized data $\tilde{\mathbf{x}}_{i,k}$ is still misaligned in the temporal domain. Therefore, by selecting a sequence of X_N data $\tilde{\mathbf{x}}_{i,k} = \{\tilde{x}_{i,k}^1, \ldots, \tilde{x}_{i,k}^{X_N}\}$ sampled at a fixed interval $\beta_{i,k}$, their sampling timestamps $\mathcal{T}_{\mathbf{i,k}} = \{\mathcal{T}_{i,k}^1, \ldots, \mathcal{T}_{i,k}^{X_N}\}$ can be used to estimate the local time error $e_{i,k}$. The error at a given instant t compared to the control center is derived as [17]

$$e_{i,k}(t) = \left(\underbrace{\frac{\mathcal{T}_{i,k}^{X_N} - \mathcal{T}_{i,k}^1}{\beta_{i,k}(X_N - 1)}}_{\text{clock skew}} - 1 \right) t + \underbrace{\frac{t_{i,k}^R - t_{con}^R + t_{i,k}^T - t_{con}^T}{2}}_{\text{clock offset}}. \tag{24}$$

The clock offset is estimated by exchanging two pairs of timestamped packets with the control center (*con*), where t^T and t^R are the transmitting and receiving instants at the corresponding devices. Therefore, $e_{i,k}$ can be used to compensate each data sample in eliminating clock-induced data inaccuracy, i.e., $\check{x}_{i,k}(t) = \tilde{x}_{i,k}(t - e_{i,k}(t))$.

3) Data Resampling: Furthermore, synchronized data could be still misaligned due to heterogeneous sampling capabilities, time-varying activity cycles, and random missing inputs, which becomes particularly critical for digital twin formation involving multi-attribute data fusion. To address this issue, synchronized data $\check{\mathbf{x}}_{\mathbf{i,k}}$ can be resampled into a series of θ data at instants $\mathcal{T}_{\mathbf{i,k}}^*$, i.e., $\{\check{\mathbf{x}}_{\mathbf{i,k}}^*, \mathcal{T}_{\mathbf{i,k}}^*\} = \{\check{x}_{i,k}^{*,l}, \mathcal{T}_{i,k}^{*,l}\}_{l=1}^{\theta}$, by adopting interpolation techniques such as Gaussian Process Regression (GPR) [32], which can recover the missing values and quantify the uncertainty associated with the resampled data. GPR relies on creating a kernel function $\boldsymbol{\mathcal{F}}$ to calculate covariance, e.g., $\boldsymbol{\mathcal{F}}_{(\mathcal{T}_{\mathbf{i,k}}, \mathcal{T}_{\mathbf{i,k}}^{\mathbf{T}})}$ for the synchronized data. Then, the resampled data at given instants $\mathcal{T}_{\mathbf{i,k}}^*$ can be derived by

$$\check{\mathbf{x}}_{\mathbf{i,k}}^* = \boldsymbol{\mathcal{F}}_{(\mathcal{T}_{\mathbf{i,k}}, \mathcal{T}_{\mathbf{i,k}}^*)}^{\mathbf{T}} \left(\boldsymbol{\mathcal{F}}_{(\mathcal{T}_{\mathbf{i,k}}, \mathcal{T}_{\mathbf{i,k}}^{\mathbf{T}})} + \sigma_f^2 \mathbf{I} \right)^{-1} \check{\mathbf{x}}_{\mathbf{i,k}}, \tag{25}$$

where σ_f^2 is the noise variance. The uncertainty of resampled data can be further evaluated by

$$\boldsymbol{\Gamma}_{\boldsymbol{i,k}}^* = \boldsymbol{\mathcal{F}}_{(\mathcal{T}_{\mathbf{i,k}}^*, \mathcal{T}_{\mathbf{i,k}}^*)} - \check{\mathbf{x}}_{\mathbf{i,k}}^* \check{\mathbf{x}}_{\mathbf{i,k}}^{-1} \boldsymbol{\mathcal{F}}_{(\mathcal{T}_{\mathbf{i,k}}, \mathcal{T}_{\mathbf{i,k}}^*)}. \tag{26}$$

Therefore, $\check{\mathbf{x}}_{\mathbf{i,k}}^*$ can be used in LDT to establish fine-grained models with dramatically enhanced temporal correlation.

B. Fine-Grained Digital Twin Modeling With STL-NARX

Directly modeling the temporally aligned dataset $\check{\mathbf{x}}_{\mathbf{i,k}}^*$ is challenging due to its complexity, driven by inherent nonlinearity, long-term shifts in network demands, and short-term regular patterns such as daily and weekly usage cycles. Motivated by the strong seasonality of network attributes, we propose the STL-NARX algorithm, which establishes digital twins through the nonlinear autoregressive exogenous (NARX) neural network, combined with seasonal and trend decomposition using Loess (STL). The NARX model is well-suited for capturing nonlinear relationships in time-series data, as it predicts future values based on past observations (autoregressive components) and external inputs [33]. Moreover, the integration of STL enables the model to decompose complex temporal patterns into seasonal, trend, and remainder components, thereby enhancing the accuracy of network dynamics modeling. This decomposition allows the model to better isolate and manage different types of variability within the network, improving both short-term prediction accuracy and long-term adaptability to network demand changes.

More specifically, the integrated data $\check{\mathbf{x}}^*_{\mathbf{i},\mathbf{k}}$ can be decomposed into trend $\mathbf{z}^{\mathbf{tr}}_{\mathbf{i},\mathbf{k}}$, seasonality $\mathbf{z}^{\mathbf{sea}}_{\mathbf{i},\mathbf{k}}$, and residual components $\mathbf{z}^{\mathbf{res}}_{\mathbf{i},\mathbf{k}}$, where $\mathbf{z}^{\mathbf{res}}_{\mathbf{i},\mathbf{k}}$ represents the irregularities and noise within the original data, given by

$$\mathbf{z}^{\mathbf{res}}_{\mathbf{i},\mathbf{k}} = \check{\mathbf{x}}^*_{\mathbf{i},\mathbf{k}} - \mathbf{z}^{\mathbf{tr}}_{\mathbf{i},\mathbf{k}} - \mathbf{z}^{\mathbf{sea}}_{\mathbf{i},\mathbf{k}}, \tag{27}$$

where $\mathbf{z}^{\mathbf{tr}}_{\mathbf{i},\mathbf{k}}$ and $\mathbf{z}^{\mathbf{sea}}_{\mathbf{i},\mathbf{k}}$ are obtained by a series of iterative processes, defined by

$$\begin{cases} \mathbf{z}^{\mathbf{sea}}_{\mathbf{i},\mathbf{k}}(\iota) = \text{Loess}\big(\check{\mathbf{x}}^*_{\mathbf{i},\mathbf{k}} - \mathbf{z}^{\mathbf{tr}}_{\mathbf{i},\mathbf{k}}(\iota-1)\big), \\ \mathbf{z}^{\mathbf{tr}}_{\mathbf{i},\mathbf{k}}(\iota) = \text{Loess}\big(\check{\mathbf{x}}^*_{\mathbf{i},\mathbf{k}} - \mathbf{z}^{\mathbf{sea}}_{\mathbf{i},\mathbf{k}}(\iota)\big), \end{cases} \tag{28}$$

where Loess$(\cdot)$ is the Loess operator for local smoothing, while the initial trend estimation $\mathbf{z}^{\mathbf{tr}}_{\mathbf{i},\mathbf{k}}(0)$ can be obtained by averaging the integrated data $\check{\mathbf{x}}^*_{\mathbf{i},\mathbf{k}}$.

As shown in Fig. 3(b), refined terms $\mathbf{z}^{\mathbf{tr}}_{\mathbf{i},\mathbf{k}}$, $\mathbf{z}^{\mathbf{sea}}_{\mathbf{i},\mathbf{k}}$, and $\mathbf{z}^{\mathbf{res}}_{\mathbf{i},\mathbf{k}}$ are modeled separately. For clarity, we only illustrate the modeling of $\mathbf{z}^{\mathbf{res}}_{\mathbf{i},\mathbf{k}}$ by using NARX as an example, which aims to capture the variation of the residual component in $\check{\mathbf{x}}^*_{\mathbf{i},\mathbf{k}}$ while considering potential exogenous inputs from other related attributes. The corresponding NARX model, denoted by a nonlinear mapping function $\mathcal{W}^{res}_{i,k}$, can be formulated as

$$\hat{z}^{res,l}_{i,k} = \mathcal{W}^{res}_{i,k}\big(\mathbf{z}^{\mathbf{res},\mathbf{R_i}}_{\mathbf{i},\mathbf{k}}, \mathbf{z}^{\mathbf{res},\mathbf{R_j}}_{\mathbf{j},\mathbf{k}}, \mathcal{H}^{res}_{i,k}\big) + \epsilon_l, \tag{29}$$

where $\mathbf{z}^{\mathbf{res},\mathbf{R_i}}_{\mathbf{i},\mathbf{k}} = \{z^{res,l-u}_{i,k}\}^{R_i}_{u=1}$ is the internal inputs while $\mathbf{z}^{\mathbf{res},\mathbf{R_j}}_{\mathbf{j},\mathbf{k}} = \{z^{res,l-v}_{j,k}\}^{R_j}_{v=1}$ is one or more exogenous inputs to $\mathcal{W}^{res}_{i,k}$. The optimal performance of NARX training necessitates carefully tuning the number of hidden layers, denoted as $\mathcal{H}^{res}_{i,k}$, alongside the maximum lags R_i and R_j. Additionally, ϵ_l denotes the error term at any given instant l. Consequently, through the integration of NARX models derived from (29) for each decomposed component, STL-NARX can construct a more fine-grained digital twin for the kth attribute of user i, represented as $\hat{\mathbf{x}}_{\mathbf{i},\mathbf{k}} = \hat{\mathbf{z}}^{\mathbf{tr}}_{\mathbf{i},\mathbf{k}} + \hat{\mathbf{z}}^{\mathbf{sea}}_{\mathbf{i},\mathbf{k}} + \hat{\mathbf{z}}^{\mathbf{res}}_{\mathbf{i},\mathbf{k}}$, which enables the prediction of its future behavior for efficient orchestration.

To further enhance the accuracy of the proposed STL-NARX, we adopt an expanded data window to mitigate the boundary effects inherent in STL decomposition. By including additional data points beyond the target period, the local smoothing process is provided with sufficient neighboring data to ensure stability near the boundaries. After the decomposition, only the central segment corresponding to the target period is retained, and the extra points are discarded. This approach preserves the accuracy of the decomposed components within the target period, avoiding significant distortions.

It is worth noting that while the STL-NARX method improves modeling accuracy with data decomposition, it also incurs additional computational costs that may affect network orchestration efficiency. To be specific, the computational complexity of STL decomposition is $\mathcal{O}(k_l N_d)$, where k_l is the span of Loess smoothing and N_d is the number of data points. For the NARX model, its complexity can be estimated by $\mathcal{O}(E_p H^2 N_d)$, where E_p is the number of training epochs, N_d is the number of data points, and H is the number of hidden neurons. Thus, the total complexity of STL-NARX is given by $\mathcal{O}(k_l N_d) + \mathcal{O}(E_p H^2 N_d)$, combining decomposition and nonlinear model training phases. Fortunately, this computational complexity can be effectively addressed by the proposed hierarchical digital twin paradigm, where only the most critical attributes and users are selected for detailed modeling, thus significantly reducing the overall complexity of the network. To further balance accuracy and efficiency, we can also optimize the decomposition frequency, applying it less frequently to more stable attributes, while prioritizing computational resources for attributes with higher variability. By dynamically adjusting the modeling complexity based on network demands, a proper tradeoff between prediction accuracy and real-time performance can be achieved.

C. Digital Twin-Enabled Network Orchestration

Based on the fine-grained digital twins $\hat{\mathbf{x}}_{\mathbf{i},\mathbf{k}}$ for each attribute in all target areas $\mathcal{Z}$, network situations can be predicted to proactively orchestrate the HetNets in terms of user association and power allocation $\{\mathbf{a}, \mathbf{p}\}$. For the problem $\mathcal{P}\mathbf{2}$, optimized results at a given instant t_u can be derived according to (8) as

$$\{\hat{\mathbf{a}}^{\mathbf{t_u}}, \hat{\mathbf{p}}^{\mathbf{t_u}}\} = \underset{\mathbf{a},\mathbf{p}}{\arg\max} \frac{\sum_l \sum_{(i,j)\in\mathcal{Z}_l} a^{t_u}_{i,j} \widehat{A}^{t_u}_{i,j}}{\widehat{\Theta}_{\text{HDT}} + \sum_l \widehat{\Theta}_{\mathcal{Z}_l}}, \tag{30}$$

where $\widehat{A}^{t_u}_{i,j} = \sum_{k\in\mathcal{Q}} \omega^n_i S^{n,t_u}_i(\mathbf{p}^{\mathbf{t_u}})$ is the predicted assignment award between user i and BS j at instant t_u for all selected attributes, while $\widehat{\Theta}_{\text{HDT}} + \sum_l \widehat{\Theta}_{\mathcal{Z}_l}$ is the total resource consumption determined by HDT as a constant. $\widehat{\Theta}_{\text{HDT}} \ll \sum_l \widehat{\Theta}_{\mathcal{Z}_l}$ represents the resource consumed for HDT and LDT modeling, respectively. For the lth identified area, the LDT resource consumption can be estimated by $\widehat{\Theta}_{\mathcal{Z}_l} = \sum_{i\in\mathcal{Z}_l} \sum_{k\in\mathcal{K}_s} \nu_{i,k} \mathcal{C}_{i,k}$, where $\nu_{i,k}$ is a binary indicator to be 1 if attribute $(i,k) \in \{\mathcal{Z}_l \cap \{\mathbf{L1}, \mathbf{L2}\}\}$ and 0 otherwise.

To generate accurate orchestration decisions at t_u, it is essential to ensure all digital twins are coherently synchronized with their physical counterparts at this specific instant. However, the multi-source delay induced during data transmission, processing, and modeling inevitably leads to model-specific temporal disparities, necessitating synchronization across virtual and physical domains when collaboratively supporting the network orchestration in the control center.

Based on this observation, we further adopt a model-level synchronization scheme to address the misalignment between physical devices and virtual models. For the model $\hat{\mathbf{x}}_{\mathbf{i},\mathbf{k}}$, depending on its causality relations, it can be established based on one or more series of integrated data. For each series of data, we use t^{Last}_j to denote its last collecting instant. After adopting STL-NARX, the model can be obtained at instant t_m expressed by $\mathcal{Y}^{t_m}_{i,k} := \hat{\mathbf{x}}^{\mathbf{t_m}}_{\mathbf{i},\mathbf{k}}$, which, however, reflects the information about t^{Last}_j. To ensure a cohesive understanding of the digital twins, each model should be compensated for the multi-source delay before utilization, given by $\hat{\mathcal{Y}}^{t_m}_j = \mathcal{Y}^{t_m+\lambda_{j\to c}}_j$, where $\lambda_{j\to c} = t_u - t^{\text{Last}}_j$ represents the temporal disparity between data collection and model utilization at control center for each digital twin. Therefore, by leveraging these additional timestamps, we can ensure digital twins generated at t_m are effective during their utilization at instant t_u.

After synchronization, the assignment problem can be solved, which is unbalanced given $N \neq M$. By manually inserting a series of virtual users $\hat{\mathcal{N}}$ and BSs $\hat{\mathcal{M}}$, a balanced assignment with $|\widetilde{\mathcal{M}}| = |\widetilde{\mathcal{N}}|$, where $\widetilde{\mathcal{N}} = \{\mathcal{N}, \hat{\mathcal{N}}\}$ and $\widetilde{\mathcal{M}} = \{\mathcal{M}, \hat{\mathcal{M}}\}$, can be derived, which can be efficiently addressed by various techniques like Hungarian algorithm with polynomial-time complexity. By solving the assignment problem, $\{\hat{\mathbf{a}}, \hat{\mathbf{p}}\}$ can be proactively obtained to achieve optimized QoS satisfaction efficiency.

V. Performance Evaluation

Given that we solve the orchestration problem $\mathcal{P}\mathbf{2}$ sequentially in the HDT and LDT via decomposition, evaluating the performance of each layer independently is crucial. Moreover, in the simulation, we focus on thoroughly validating the aspects of precisely identifying target areas and the scalable modeling for fine-grained digital twins, aligning with the core strategies of the proposed hierarchical digital twin paradigm.

A. Simulation Setup

1) Environment of the HetNets: The evaluation of our proposed hierarchical digital twin-enabled network orchestration scheme focuses on the HetNets comprising 5 macro BSs with a maximum transmission power of 40 dBm and a coverage radius of 300 meters, uniformly distributed across the area. In addition, we opportunistically deploy 20 small BSs within this network, each offering a maximum transmission power of 17 dBm and covering a radius of 50 meters. Distinct path loss models are applied to macro cells $\big(15.3 + 37.6 \log_{10}(D)\big)$ and small cells $\big(8.46 + 20 \log_{10}(D) + 0.7D\big)$, with D denoting the distance between each user-BS pair. The total bandwidth of the system is 20 MHz. Moreover, 200 users are densely populated in certain areas to simulate environments like malls or offices. The channel fading model includes log-normal shadowing and small-scale fading, characterized by a background noise power of -104 dBm and a noise figure of 5 dB [34].

2) Parameters for Lower-Layered Digital Twin Modeling: At the lower layers of our simulations, four distinct modeling techniques are compared, namely, auto-regressive integrated moving average (ARIMA) [3], NARX neural network [33], temporal convolutional network (TCN) [35], and our proposed STL-NARX approach. Specifically, ARIMA is a linear time-series model that predicts future values based on the combination of autoregressive terms, moving averages, and differencing to account for the data trend. In our context, we configure the ARIMA models with parameters $(2, 2, 0)$ to capture the linear temporal dependencies of network attributes. Readers interested in further details about ARIMA can refer to [3] and the references therein. Meanwhile, TCN uses dilated causal convolutions to efficiently capture long-term dependencies in time series data, where we configured TCN with a kernel size of 3, dilation factors of 1 to 4, 3 hidden layers of 64 filters each, and trained it using the Adam optimizer.

Moreover, as described earlier, NARX is designed to capture nonlinear relationships between past observations and external inputs. To enhance prediction robustness, we fine-tune the NARX model with a learning rate of 0.001, a delay setting of 10, and 15 hidden layers. These models are used to predict critical network attributes such as throughput, delay, traffic load, and packet loss within a specified area of the HetNets. Additionally, to capture recurring daily and weekly traffic patterns, seasonality periods of 24 and 168 hours are incorporated during STL analysis. This allows the model to effectively account for both short-term and long-term variations in network dynamics.

B. Performance of Higher-Layered Digital Twins

1) Network Attribute Differentiation: As outlined in Section III-A, we assess the modeling value through a detailed cost-benefit analysis of various network elements, determining their impact on network performance. The modeling costs obtained based on sample entropy (9) across a range of user attributes are shown in Fig. 4(a). Attributes such as network throughput (Attribute 1), delay (Attribute 2), and packet loss rate (Attribute 4) exhibit lower modeling costs, attributed to their stable statistical patterns, thereby simplifying prediction efforts. Conversely, dynamically fluctuated attributes, including network traffic load (Attribute 3) and user mobility (Attribute 5), show increased entropy, thus indicating the need for more computational resources for accurate modeling. On the other hand, Fig. 4(b) illustrates the modeling benefits derived from correlation analysis and causality inference relative to the network demands. Specifically, modeling attributes critical to network orchestration, such as network traffic load, exhibit benefits close to one, indicating a substantial return on modeling efforts. In contrast, attributes like instantaneous channel quality and user mobility offer diminished value for long-term network orchestration. It is worth noting that the modeling value varies among users for the same attribute due to differing QoS requirements, resulting in varying modeling benefits even when utilizing the same network information. Additionally, individual users may experience different network conditions, traffic patterns, or usage behaviors, leading to diverse modeling costs across users. These factors, when combined, could cause distinct impacts of each attribute on overall network orchestration, underscoring the necessity for customized attribute differentiation for each user.

Based on the assessed modeling value, we categorize all attributes into three levels, as Fig. 4(c) demonstrates. $\mathbf{L}1$ attributes, which yield high benefits for low costs, are modeled in HDT with high priority to identify target areas in the HetNets. $\mathbf{L}2$ attributes, though still important for network orchestration, are selectively chosen for more granular modeling in the LDT to support specific and problem-centered network orchestration. This ensures that resources are allocated efficiently to attributes that contribute the most to network optimization. In contrast, high-cost attributes can significantly reduce orchestration efficiency if their benefits do not justify the resources consumed, making it crucial to prioritize those attributes that strike the best balance between cost and benefit.

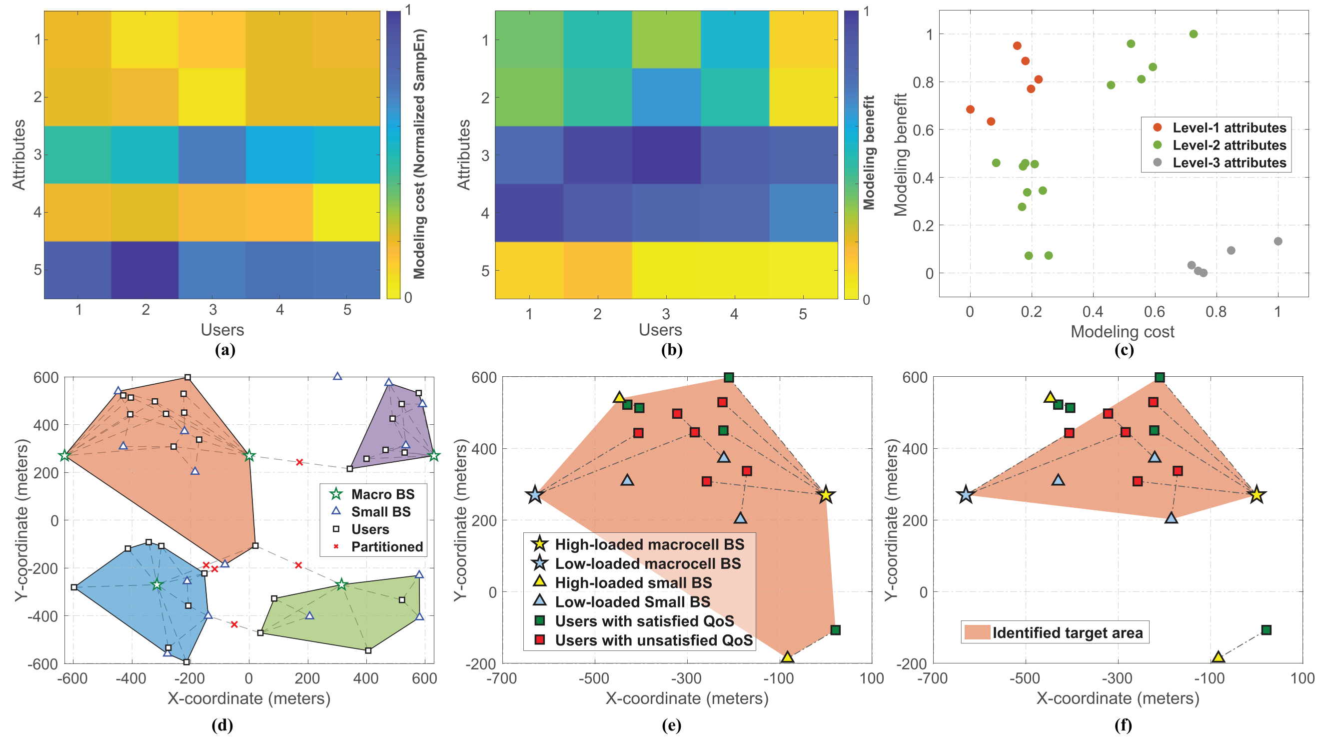

Fig. 4. Identification of the target area at the higher-layered digital twins based on differentiated network attributes and network situation analysis: (a). Modeling cost evaluated by sample entropy for different users and attributes. (b). Modeling benefits evaluated by correlation analysis and causality inference compared to the network objective. (c). Three-level attribute differentiation based on the modeling value. (d). Segmentation of the HetNets into four sub-networks based on user activity, where only a portion of the users are explicitly illustrated to represent user density. (e). Real-time network situation analysis for one of the sub-networks, including traffic load of BSs and user QoS satisfaction. (f). Target area identification based on orchestration outcomes for the selected sub-network.

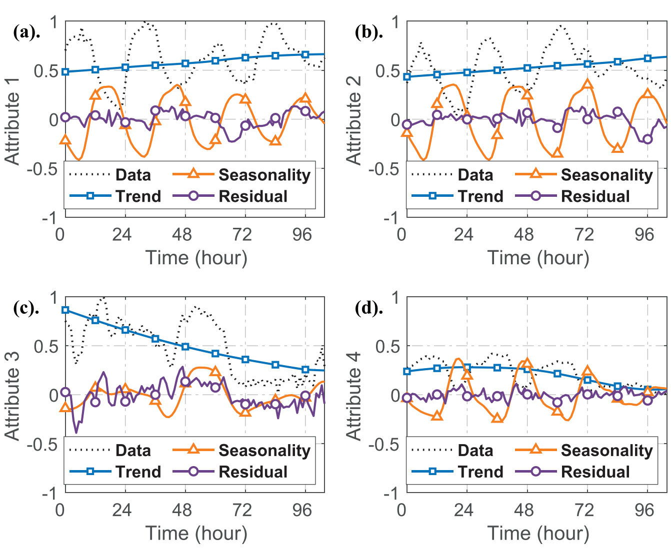

Fig. 5. The analysis of four attributes by applying STL before digital twin modeling, which decomposes data into trend, seasonality, and residual. A clear seasonality is observed for attributes 1, 2, and 4.

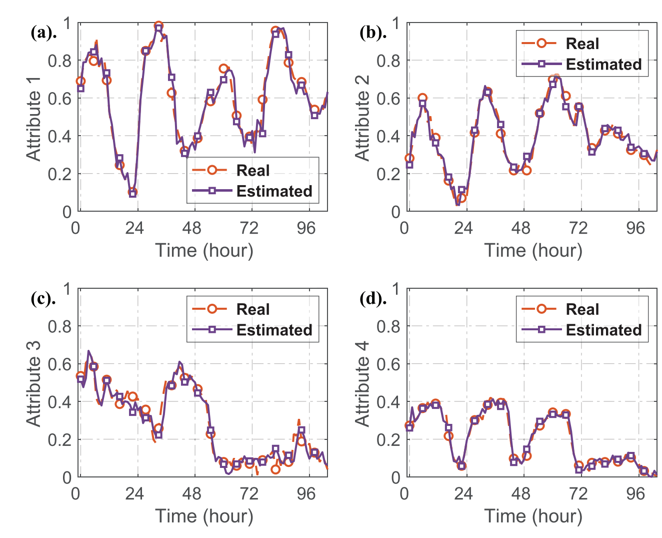

Fig. 6. Comparison of the predicted data with real values by adopting the proposed STL-NARX algorithm, where accurate modeling results are achieved for the selected four attributes.

2) Target Area Identification: The layout of the HetNets is shown in Fig. 4(d), where only a portion of the users are explicitly illustrated to represent user density while maintaining clarity in the visualization. After graph partitioning, the entire HetNets is segmented into four discrete sub-networks. This division is based on the digital twins established in HDT using historical time-series data to identify areas with frequent user-BS interactions and high throughput, grouping them into cohesive sub-networks. In contrast, zones with sparse interactions and low traffic are marked as segments with weak connectivity, thus selected for disconnection due to their minimal impact on partitioning costs.

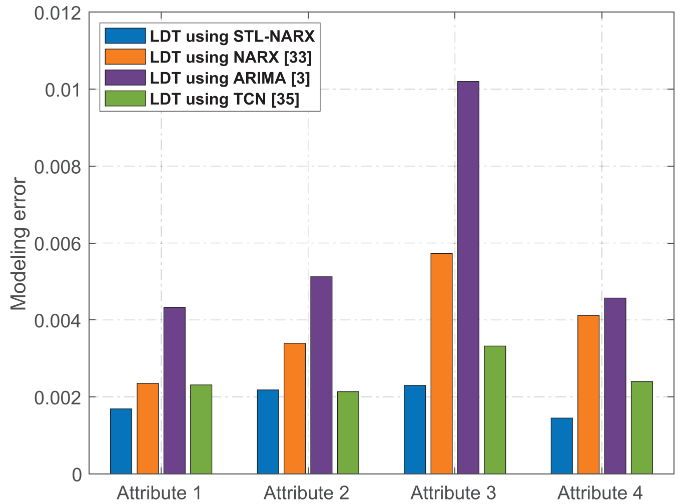

Fig. 7. Comparison of four different modeling methods, including ARIMA, NARX, TCN, and STL-NARX, for the four selected attributes. STL-NARX achieves the highest modeling accuracy for all attributes, particularly for those with a complex data pattern.

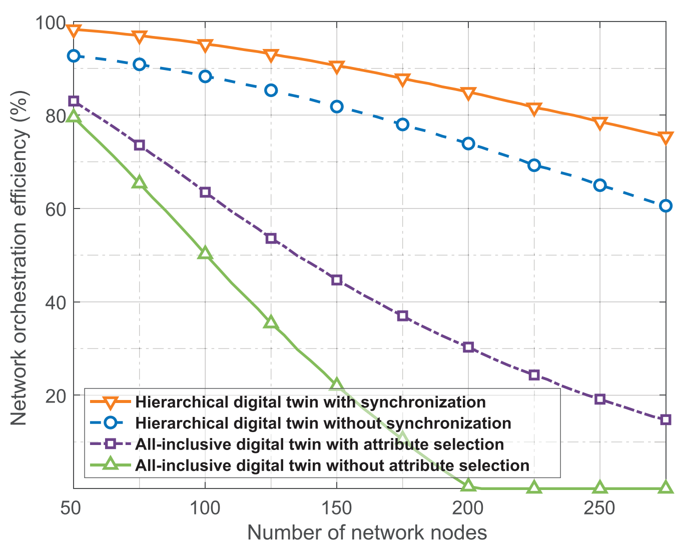

Fig. 8. Comparison of network orchestration efficiency achieved by different digital twin paradigms. With different network scales, the proposed hierarchical digital twin with multi-level synchronization always achieves the most efficient network orchestration.

Based on this network segmentation, a detailed examination of a specific sub-network is given in Fig. 4(e), which shows the interaction between user density and BS capacity in fulfilling various QoS requirements. Clearly, regions with denser users are more likely to associate with lower QoS satisfaction and heavily burdened BSs, which should be prioritized during network orchestration in LDT. It is evident that the load on a BS is not strictly proportional to the number of connected users. Certain BSs, despite having relatively few users, can experience significant load due to the high resource demands of specific services, such as video streaming. Consequently, these BSs may encounter resource constraints, limiting their capacity to participate in the subsequent network optimization.

Target areas within each sub-network are further identified by analyzing their expected cost-benefit for network orchestration, as shown in Fig. 4(f). By adopting a refined graph partitioning process with adjusted cutting weights based on real-time network conditions, areas with the highest management effectiveness can be isolated. Therefore, QoS-satisfied users and high-loaded BSs, if not closely intertwined with other components, can be selectively filtered out. This ensures the prioritization of the most critical elements for detailed modeling in the construction of lower-layered digital twins.

C. Performance of Lower-Layered Digital Twins

According to the target areas identified, elements within the network that need particular management are selected for fine-grained modeling in LDT. Our STL-NARX approach first uses STL to decompose the time-series data generated from the selected users, where trend, seasonality, and residual components are isolated for precise analysis. As shown in Fig. 5, each series of data can be accurately decomposed into three components. Only daily seasonality is shown for clarity, where it clearly shows the daily activity pattern of throughput, delay, and packet loss rate regarding the selected user in Fig. 5(a), (b), and (d), respectively. Meanwhile, Fig. 5(c) illustrates that the network load for a nearby BS lacks a clear daily pattern, suggesting increased challenges in modeling accuracy.

After decomposition, by focusing on each decomposed component, the predictability of attribute data can be improved, resulting in better modeling performance. As shown in Fig. 6, the proposed STL-NARX effectively estimates each attribute over a four-day time window. Notably, the dynamic nature of traffic load shown in Fig. 6(c), which is typically influenced by other users, may lead to reduced modeling precision due to limited modeling scope. This issue is effectively addressed by the STL-NARX algorithm, which sufficiently considers both internal and exogenous inputs to capture the complex interactions within the network.

To validate the advantages of the proposed STL-NARX algorithm, the training accuracy of four different techniques, including ARIMA, NARX, TCN, and STL-NARX, are compared, as shown in Fig. 7. We apply mean squared error (MSE) to evaluate the modeling accuracy for the four attributes. With the integrated data, all four methods can obtain relatively accurate results, where the proposed STL-NARX outperforms the others, especially when pronounced seasonality is shown in the data. This is particularly obvious for attribute 3, which fluctuates over time due to external variations that lead to higher modeling errors. However, by selecting relevant exogenous inputs from nearby users during digital twin modeling using NARX, the modeling error can be dramatically reduced compared to ARIMA. This highlights the significance of tight data integration and multi-level synchronization for multi-attribute digital twin modeling.

The optimal network orchestration strategy is then concurrently achieved for each target area based on the prediction obtained from the digital twin models for efficient network QoS satisfaction. As shown in Fig. 8, the proposed hierarchical digital twin scheme achieves the highest efficiency for different network scales compared to the traditional all-inclusive frameworks, especially when multi-level synchronization is achieved. This improvement can be attributed to the synchronized coordination between different digital twins, allowing for more accurate user behavior analysis and timely decision-making during network orchestration.

Furthermore, adaptive attribute selection and scalable network modeling significantly improve the orchestration efficiency. This adaptability and scalability are evident in Fig. 8, where the proposed method can still efficiently satisfy user QoS demands in larger networks due to the prioritized user of resources. As the number of attributes and users increases exponentially in larger networks, traditional all-inclusive approaches become increasingly inefficient or even impractical. In contrast, by focusing on the most relevant attributes and adjusting the modeling complexity based on current network conditions, the hierarchical paradigm optimizes the utilization of limited resources, thereby avoiding unnecessary computations on less impactful attributes.

VI. Conclusion

This paper has proposed a novel hierarchical digital twin paradigm to achieve efficient network modeling for 6G network orchestration through prioritized decomposition and selective modeling. By assigning higher priority to critical network attributes within the higher-layered digital twin, the approach facilitates the rapid identification of pressing network issues. This prioritization allows for scalable modeling of fine-grained digital twins at the lower layers, focusing on key attributes and users to efficiently optimize network performance. In essence, the proposed scheme enables situation-dependent problem identification in complex networks, leading to problem-centered solutions that significantly reduce resource wastage typically associated with traditional all-inclusive modeling approaches. Simulation results have validated the effectiveness of the scheme in supporting efficient network orchestration in complex environments. Future research directions include developing advanced digital twin synchronization techniques, such as predictive models and context-aware coordination, to enhance the temporal alignment of various digital twin models in supporting large-scale and highly heterogeneous systems.

References

[1] C.-X. Wang et al., "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 905–974, 2nd Quart. 2023.

[2] X. Wang, J. Mei, S. Cui, C.-X. Wang, and X. S. Shen, "Realizing 6G: The operational goals, enabling technologies of future networks, and value-oriented intelligent multi-dimensional multiple access," *IEEE Netw.*, vol. 37, no. 1, pp. 10–17, Jan. 2023.

[3] P. Jia and X. Wang, "A new virtual network topology-based digital twin for spatial–temporal load-balanced user association in 6G HetNets," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 10, pp. 3080–3094, Oct. 2023.

[4] H. F. Alhashimi et al., "A survey on resource management for 6G heterogeneous networks: Current research, future trends, and challenges," *Electronics*, vol. 12, no. 3, p. 647, Jan. 2023.

[5] W. Li, F. Zhou, K. R. Chowdhury, and W. Meleis, "QTCP: Adaptive congestion control with reinforcement learning," *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 3, pp. 445–458, Jul. 2019.

[6] O. Ahmed, F. Ren, A. Hawbani, and Y. Al-Sharabi, "Energy optimized congestion control-based temperature aware routing algorithm for software defined wireless body area networks," *IEEE Access*, vol. 8, pp. 41085–41099, 2020.

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

[8] S. Mihai et al., "Digital twins: A survey on enabling technologies, challenges, trends and future prospects," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2255–2291, 4th Quart., 2022.

[9] L. Yang, Y. Zou, S. Shen, P. Wang, D. Yu, and X. Cheng, "A fault-tolerant communication algorithm for age-of-information optimization in DITENs," *IEEE Trans. Commun.*, vol. 72, no. 5, pp. 2851–2864, May 2024.

[10] A. Alkhateeb, S. Jiang, and G. Charan, "Real-time digital twins: Vision and research directions for 6G and beyond," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 128–134, Nov. 2023.

[11] H. Xu, J. Wu, Q. Pan, X. Guan, and M. Guizani, "A survey on digital twin for Industrial Internet of Things: Applications, technologies and tools," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2569–2598, 4th Quart., 2023.

[12] 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.

[13] B. Agarwal, M. A. Togou, M. Marco, and G.-M. Muntean, "A comprehensive survey on radio resource management in 5G HetNets: Current solutions, future trends and open issues," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2495–2534, 4th Quart., 2022.

[14] A. Alwarafy, M. Abdallah, B. S. Ciftler, A. Al-Fuqaha, and M. Hamdi, "The frontiers of deep reinforcement learning for resource management in future wireless HetNets: Techniques, challenges, and research directions," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 322–365, 2022.

[15] X. Wang, P. Jia, X. S. Shen, and H. V. Poor, "Intelligent and low overhead network synchronization for large-scale industrial IoT systems in the 6G era," *IEEE Netw.*, vol. 37, no. 3, pp. 76–84, Mar. 2023.

[16] X. Huan, H. He, T. Wang, Q. Wu, and H. Hu, "A timestamp-free time synchronization scheme based on reverse asymmetric framework for practical resource-constrained wireless sensor networks," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6109–6121, Sep. 2022.

[17] P. Jia, X. Wang, and X. Shen, "Accurate and efficient digital twin construction using concurrent end-to-end synchronization and multi-attribute data resampling," *IEEE Internet Things J.*, vol. 10, no. 6, pp. 4857–4870, Mar. 2023.

[18] Y. Han et al., "A dynamic hierarchical framework for IoT-assisted digital twin synchronization in the metaverse," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 268–284, Jan. 2023.

[19] H. Zhang, X. Ma, X. Liu, L. Li, and K. Sun, "GNN-based power allocation and user association in digital twin network for the terahertz band," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 10, pp. 3111–3121, Oct. 2023.

[20] D. V. Huynh et al., "URLLC edge networks with joint optimal user association, task offloading and resource allocation: A digital twin approach," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7669–7682, Nov. 2022.

[21] Y. Hao, J. Wang, D. Huo, N. Guizani, L. Hu, and M. Chen, "Digital twin-assisted URLLC-enabled task offloading in mobile edge network via robust combinatorial optimization," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 10, pp. 3022–3033, Oct. 2023.

[22] Y. Wang, Z. Su, S. Guo, M. Dai, T. H. Luan, and Y. Liu, "A survey on digital twins: Architecture, enabling technologies, security and privacy, and future prospects," *IEEE Internet Things J.*, vol. 10, no. 17, pp. 14965–14987, Sep. 2023.

[23] J. Zheng et al., "Data synchronization in vehicular digital twin network: A game theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7635–7647, Nov. 2023.

[24] J. Tang et al., "UAV-assisted digital-twin synchronization with tiny-machine-learning-based semantic communications," *IEEE Internet Things J.*, vol. 11, no. 17, pp. 28437–28451, Sep. 2024.

[25] P. Jia, X. Wang, and X. S. Shen, "Digital twin enabled intelligent network orchestration for 6G: A dual-layered approach," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2023, pp. 1–6.

[26] J. Chen, L. Zhang, Y.-C. Liang, and S. Ma, "Optimal resource allocation for multicarrier NOMA in short packet communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 2141–2156, Feb. 2020.

[27] L.-T. Hsieh, H. Liu, Y. Guo, and R. Gazda, "Deep reinforcement learning-based task assignment for cooperative mobile edge computing," *IEEE Trans. Mobile Comput.*, vol. 23, no. 4, pp. 3156–3171, Apr. 2024.

[28] Y. Xiao, J. Liu, J. Wu, and N. Ansari, "Leveraging deep reinforcement learning for traffic engineering: A survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2064–2097, 4th Quart., 2021.

[29] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, and N. Stergiou, "The appropriate use of approximate entropy and sample entropy with short data sets," *Ann. Biomed. Eng.*, vol. 41, no. 2, pp. 349–365, Feb. 2013.

[30] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Trans. Knowl. Discovery Data*, vol. 15, no. 5, pp. 1–46, 2021.

[31] Z. Wang, Z. Li, R. Wang, F. Nie, and X. Li, "Large graph clustering with simultaneous spectral embedding and discretization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4426–4440, Dec. 2021.

[32] E. Schulz, M. Speekenbrink, and A. Krause, "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions," *J. Math. Psychol.*, vol. 85, pp. 1–16, Aug. 2018.

[33] H. Zarini, M. R. Mili, M. Rasti, S. Andreev, P. H. J. Nardelli, and M. Bennis, "Intelligent analog beam selection and beamspace channel tracking in THz massive MIMO with lens antenna array," *IEEE Trans. Cogn. Commun. Netw.*, vol. 9, no. 3, pp. 629–646, Jun. 2023.

[34] F. Fang, G. Ye, H. Zhang, J. Cheng, and V. C. Leung, "Energy-efficient joint user association and power allocation in a heterogeneous network," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7008–7020, Nov. 2020.

[35] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 156–165.

Pengyi Jia (Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, Western University, London, ON, Canada, in 2021. He is currently a Post-Doctoral Associate with Western University. His research interests include intelligent network synchronization, digital twin, and machine learning and their applications in vertical IoT systems, cellular networks, and advanced manufacturing. His current research is to develop effective digital twin paradigms by investigating the intrinsic temporal correlations behind the massive sampling data to support network optimization and application-oriented heterogeneous service provisioning. He has been involved in organizing IEEE CCECE 2021. He has served as a TPC Member for many IEEE conferences, including INFOCOM, GLOBECOM, VTC, and MECOM. He has served as the Vice-Chair for the ComSoc Chapter in IEEE London Section.

Xianbin Wang (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from the National University of Singapore in 2001.

He has been with Western University, Canada, since 2008, where he is currently a Distinguished University Professor and a Tier-1 Canada Research Chair. Prior to joining Western University, he was with the Communications Research Centre, Canada, as a Research Scientist and later a Senior Research Scientist from 2002 to 2007. From 2001 to 2002, he was a System Designer with STMicroelectronics. He has over 600 highly cited journals and conference papers, over 30 granted and pending patents, and several standard contributions. His current research interests include 5G/6G technologies, the Internet of Things, machine learning, communications security, digital twin, and intelligent communications.

Dr. Wang is a fellow of Canadian Academy of Engineering and the Engineering Institute of Canada. He has received many prestigious awards and recognitions, including the IEEE Canada R. A. Fessenden Award at Canada Research Chair, the Engineering Research Excellence Award at Western University, the Canadian Federal Government Public Service Award, the Ontario Early Researcher Award, and ten best paper awards. He is a member of the Senate, Senate Committee on Academic Policy and Senate Committee on University Planning at Western. He serves on NSERC Discovery Grant Review Panel for Computer Science. He has been involved in many flagship conferences, including IEEE GLOBECOM, ICC, VTC, PIMRC, WCNC, CCECE, and ICNC, in different roles, such as the General Chair, the TPC Chair, the Symposium Chair, the Tutorial Instructor, the Track Chair, the Session Chair, and a Keynote Speaker. He serves/has served as the Editor-in-Chief, the Associate Editor-in-Chief, an area editor, and an editor/an associate editor for over ten journals. He was the Chair of the IEEE ComSoc Signal Processing and Computing for Communications (SPCC) Technical Committee. He is serving as the Central Area Chair for IEEE Canada.

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, the Internet of Things, 5G and beyond, and vehicular networks. He is an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Member, and an Engineering Academy of Japan International Fellow. He received the West Lake Friendship Award from Zhejiang Province in 2023, the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, the R. A. Fessenden Award from IEEE in 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 (2015) and Education Award (2017) from the IEEE Communications Society (ComSoc), and 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 in 2006 from the University of Waterloo; and the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He serves/served as the General Chair for the 6G Global Conference 2023 and ACM Mobihoc 2015; and the Technical Program Committee Chair/Co-Chair for IEEE Globecom 2024, 2016, and 2007, IEEE Infocom 2014, and IEEE VTC 2010 Fall. He is the past President of IEEE ComSoc, the Vice President of technical and educational activities, the Vice President for Publications, the 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 served as the Editor-in-Chief for IEEE INTERNET OF THINGS JOURNAL, *IEEE Network*, and *IET Communications*. He is a registered Professional Engineer of Ontario, Canada.