Generative AI Empowered Network Digital Twins: Architecture, Technologies, and Applications

TONG LI, Department of Electronic Engineering, Tsinghua University, Beijing, China
QINGYUE LONG, Department of Electronic Engineering, Tsinghua University, Beijing, China
HAOYE CHAI, Department of Electronic Engineering, Tsinghua University, Beijing, China
SHIYUAN ZHANG, Department of Electronic Engineering, Tsinghua University, Beijing, China
FENYU JIANG, Department of Electronic Engineering, Tsinghua University, Beijing, China
HAOQIANG LIU, Department of Electronic Engineering, Tsinghua University, Beijing, China
WENZHEN HUANG, Department of Electronic Engineering, Tsinghua University, Beijing, China
DEPENG JIN, Department of Electronic Engineering, Tsinghua University, Beijing, China
YONG LI, Department of Electronic Engineering, Tsinghua University, Beijing, China

The rapid advancement of mobile networks highlights the limitations of traditional network planning and optimization methods, particularly in modeling, evaluation, and application. Network Digital Twins, which simulate networks in the digital domain for evaluation, offer a solution to these challenges. This concept is further enhanced by generative AI technology, which promises more efficient and accurate AI-driven data generation for network simulation and optimization. This survey provides insights into generative AI-empowered network digital twins. We begin by outlining the architecture of a network digital twin, which encompasses both digital and physical domains. This architecture involves four key steps: data processing and network monitoring, digital replication and network simulation, designing and training network optimizers, Sim2Real, and network control. Next, we systematically discuss the related studies in each step and make a detailed taxonomy of the problem studied, the methods used, and the key designs leveraged. Each step is examined with a focus on the role of generative AI, from estimating missing data and simulating network behaviors to designing control strategies and bridging the gap between digital and physical domains. Finally, we discuss the open issues and challenges of generative AI-based network digital twins.

CCS Concepts: • **Networks** → **Mobile networks**; *Network simulations*; *Network management;*

This research has been supported in part by the National Key Research and Development Program of China under Grant 2023YFB2904804, in part by the National Natural Science Foundation of China under Grant 62471277, and in part by the Institute for Guo Qiang, Tsinghua University.

Authors' Contact Information: Tong Li, Department of Electronic Engineering, Tsinghua University, Beijing, Beijing, China; e-mail: tongli@mail.tsinghua.edu.cn; Qingyue Long, Department of Electronic Engineering, Tsinghua University, Beijing, Beijing, China; e-mail: longqingyue@126.com; Haoye Chai, Department of Electronic Engineering, Tsinghua University, Beijing, Beijing, China; e-mail: haoyechai@163.com; Shiyuan Zhang, Department of Electronic Engineering, Tsinghua University, Beijing, Beijing, China; e-mail: zhangshi22@mails.tsinghua.edu.cn; Fenyu Jiang, Department of Electronic Engineering, Tsinghua University, Beijing, China; e-mail: jiangfenyu@tsinghua.edu.cn; Haoqiang Liu, Department of Electronic Engineering, Tsinghua University, Beijing, China; e-mail: lhqbuaa@mail.tsinghua.edu.cn; Wenzhen Huang, Department of Electronic Engineering, Tsinghua University, Beijing, Beijing, China; e-mail: huangwenzhen@tsinghua.edu.cn; Depeng Jin, Department of Electronic Engineering, Tsinghua University, Beijing, China; e-mail: jindp@tsinghua.edu.cn; Yong Li (Corresponding author), Department of Electronic Engineering, Tsinghua University, Beijing, China; e-mail: liyong07@tsinghua.edu.cn.

ACM 0360-0300/2025/02-ART157
https://doi.org/10.1145/3711682

Additional Key Words and Phrases: Generative AI, digital twins, mobile networks, network monitoring, network simulation, network operation

ACM Reference Format:
Tong Li, Qingyue Long, Haoye Chai, Shiyuan Zhang, Fenyu Jiang, Haoqiang Liu, Wenzhen Huang, Depeng Jin, and Yong Li. 2025. Generative AI Empowered Network Digital Twins: Architecture, Technologies, and Applications. *ACM Comput. Surv.* 57, 6, Article 157 (February 2025), 43 pages. https://doi.org/10.1145/3711682

1 Introduction

Over the past two decades, rapid technological advancements have significantly transformed mobile networks. These developments have diversified the services, target users, and devices connected to mobile networks, including innovations such as brain-computer interaction, extended reality, holographic telepresence, global ubiquitous connectivity, and pervasive intelligence. As network environments become increasingly complex, critical players in the telecom industry, such as mobile network operators, equipment vendors, and communication service providers, face significant pressure to manage networks efficiently. Currently, mobile network deployment, optimization, and operation rely on solutions based on mathematical modeling and combinatorial optimization [1, 2]. Operators and researchers first analyze the network problem in-depth, identifying controllable variables and defining optimization objectives. They then utilize mathematical models to explicitly illustrate the functional relationship between these variables and objectives through physical equations and logical rules. Hence, the network problem is modeled into a mathematical optimization problem. Combinatorial optimization or heuristic algorithms are then employed to solve this problem. Such a methodology has been successfully applied to solve various network management problems in relatively simple mobile networks over the past decade.

Based on mathematical modeling and combinatorial optimization, the traditional methodology faces three key challenges that limit its effectiveness in current complex mobile networks [3]. Firstly, the increasing complexity of services makes it difficult for traditional methods to formulate mathematical equations accurately reflecting complex network conditions [4]. Secondly, the necessity of maintaining stable network operations prevents the direct testing of optimization algorithms in real-world mobile networks, hindering accurate evaluation of obtained optimization strategies [5]. Lastly, such offline network optimization approaches, relying on theoretical models, often idealize real-world networks excessively, resulting in suboptimal performance when applied in practical scenarios [6]. To address the challenges faced in network deployment and optimization, an innovative concept called "**Network Digital Twin**" (**NDT**) has been proposed [7–9]. The idea behind NDT is to create a digital replica of the physical mobile network, which aims at simulating the structure, environment, and status of network components or systems with high fidelity [10, 11]. NDT can effectively solve the problems of modeling, evaluation, and application deficits. Firstly, by creating a virtual replica of the mobile network, NDT avoids the need for complex mathematical modeling to explore the relationship between controllable variables and optimization objectives in the mobile network. Secondly, developers and operators can use the NDT for what-if analysis [12, 13], simulating and evaluating the performance of different network optimization algorithms without direct testing in the real-world environment [14]. With such a what-if simulator, **reinforcement learning** (**RL**)-based optimizers can iteratively interact with the simulator to identify the most effective network configurations. Specifically, RL allows an agent to learn by interacting with its environment, using trial and error to gradually develop an optimal policy that guides decision-making to maximize long-term rewards [15]. RL has been particularly successful in dynamic and complex decision-making environments, making it highly suitable for optimizing mobile networks by facilitating adaptive control, efficient resource management, and real-time decision-making in network operations. Finally, the digital twin is supposed to be a high-fidelity replication of the real-world network [16], which ensures the practical application of optimization strategies in the real-world environment. The NDT is revolutionizing the mobile network industry, fundamentally changing how we interact with and manage mobile networks.

Creating virtual simulations of physical mobile networks in the digital space for performance evaluation is an essential research domain in networking [17–19]. To this end, researchers have developed various network simulators, such as NS-3 [20], OPNET [21], and OMNet++ [22]. These simulators are adept at processing virtual packets and communication events under specific network policies. They employ discrete event-driven methods to simulate network element communication behaviors and provide key network performance metrics, like throughput, latency, and user data rates. However, a significant limitation of these simulators is their low execution efficiency, with simulation speeds significantly slower than those of real-world mobile networks [23]. Alternatively, researchers have sought to employ analytical modeling methods to depict the relationships between influencing factors and network performance metrics. Conventional analytical modeling methods, such as stochastic channel models [24], Shannon capacity [25], and queuing theory [26], oversimplified the real-world environment, leading to inaccurate estimation.

In addition to these research-oriented simulators, several commercially available digital twin tools have been introduced for building and managing digital replicas of physical devices. For instance, Azure Digital Twin [27] is a cloud-based platform designed to create digital replicas of real-world systems, where twins are connected into a twin graph by their relationships, and users can query this graph to monitor the states and events of the twinned environment. MATLAB and Simulink [28] provide a predictive maintenance toolbox to support digital twins with physics-based modeling, allowing the electrical components to conduct what-if simulation analyses from first principles. Moreover, Eclipse Ditto [29] focuses on managing the digital representation of connected devices, offering flexibility for IoT and industrial digital twin applications. However, Azure Digital Twin and Eclipse Ditto primarily focus on providing data management and monitoring functions, with limited support for simulation capabilities. Although MATLAB and Simulink tools offer predictive simulation capabilities, their execution efficiency is often low due to the physics-based modeling approach, which involves starting simulations from first principles. For NDT applications, the ability to rapidly and accurately simulate network dynamics to adapt to changes in real-world networks is paramount. The inefficiency of discrete event-driven and first principle-based simulation methods and inaccurate analytical modeling methods fall short of meeting the requirements of such applications. This gap highlights the need for an innovative network virtual simulation technology that can simulate network dynamics and performance efficiently and accurately.

Recent advancements in generative **Artificial Intelligence** (**AI**) have attracted widespread attention from the scientific community and the industry [30–32]. Generative AI leverages machine learning algorithms to learn from input data and generate new data samples that resemble the original input [33]. With its notable success across diverse fields, generative AI has sparked significant interest in mobile networks, especially its integration into NDTs [14, 34, 35]. Specifically, generative AI employs deep neural networks trained on extensive real-world network data to learn the distribution characteristics, which allows for accurate mapping of different factors within the network environment to the behavior and performance data of network elements, enabling the effective generation of network data. As a result, integrating generative AI with NDTs represents a significant evolution from the conventional, inefficient discrete event-driven simulation to a more efficient and accurate AI-driven network data generation approach.

Despite the expanding body of research on NDT and generative AI technologies, there has been a lack of comprehensive surveys that consolidate recent advancements in both fields. This

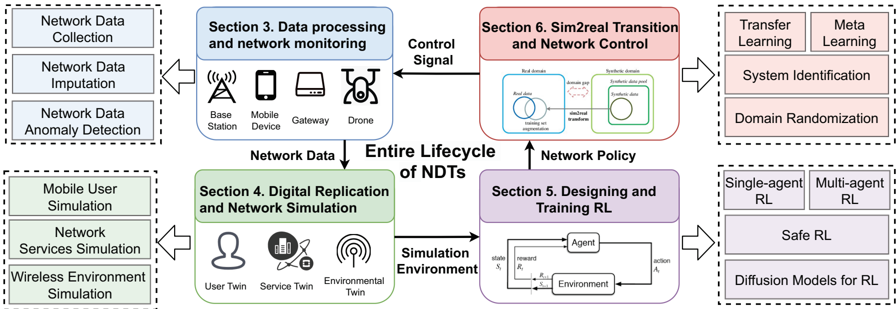

Fig. 1. Taxonomy of existing studies for generative AI-driven NDT.

survey aims at filling that gap by providing a detailed analysis of the architecture, technologies, and applications of generative AI-empowered NDTs. Understanding this integration is crucial for researchers and industry professionals seeking to design and implement more scalable, adaptive, and efficient next-generation AI-empowered network systems. In this survey, we aim at illuminating recent advancements in generative AI as applied to NDTs, offering a historical perspective of this rapidly evolving area. Generative AI technology plays a crucial role in supporting the entire lifecycle of NDTs. To offer a clear understanding, we present a taxonomy that categorizes relevant studies based on their applications across various phases of NDTs. These phases include data processing and network monitoring, digital replication and network simulation, designing and training RL models, and sim-to-real transitions and network control. As illustrated in Figure 1, during the data processing and network monitoring (Section 3), which includes network data collection, imputation, and anomaly detection, generative AI assists by generating realistic data points to fill spatial and temporal gaps, as well as data that represents a network's normal operating state for anomaly detection. The digital replication and network simulation (Section 4) focuses on simulating network behaviors and conducting what-if analyses. Generative AI models can produce high-fidelity data under specific network conditions, covering mobile users, network services, and wireless environment simulations. In the design and training of RL (Section 5), generative AI is a critical tool for network operators, aiding in developing control strategies within the digital domain. During the sim-to-real transition and network control (Section 6), generative AI models the uncertainty distribution of real-world data, bridging the gap between digital and physical domains, which helps correct simulator inaccuracies and ensures a smooth transition of network control strategies from simulation to real-world environments. Additionally, we also provide an overview of the NDT concept and generative AI methods in Section 2. Section 7 discusses the challenges and prospective future developments in generative AI-driven NDTs. Section 8 provides the concluding remarks of the survey.

What sets this survey apart from previous work is its focus on the synergy between generative AI and NDTs. While traditional surveys have covered either NDTs or AI in isolation, this survey emphasizes how generative AI enhances the lifecycle of digital twins–from data processing and network monitoring to network optimization and control. This survey provides a historical perspective and insights into the latest developments and future directions, particularly in dynamic twins, continuous learning, large generative models, and large-scale network twin platforms. The impact of this survey lies in its ability to guide future research and development in the field, highlighting the challenges and opportunities at the intersection of AI and mobile networks. By offering a detailed taxonomy and analysis of the state-of-the-art methods, this work will serve as a valuable reference for academics and practitioners looking to explore new possibilities in NDT applications.

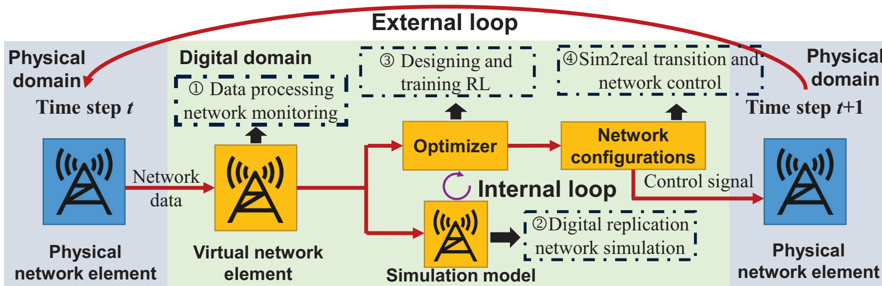

Fig. 2. A "Two-Domain, Four-Step, Dual-Loop" collaborative methodology of NDTs.

The main contributions of this article can be summarized as follows:

— We provide a comprehensive historical overview of the field of generative AI, particularly in its application to the key steps of NDTs, including data processing and network monitoring, digital replication and network simulation, designing and training network optimizer, field trials, and network control. This perspective helps contextualize current developments within a broader historical framework.
— We present an overview of the latest advancements in generative AI for NDTs, highlighting their significant influence on both academia and industry. We bridge theoretical innovation with practical use, showcasing how generative AI reshapes NDTs to enhance management, optimization, and analysis.
— The challenges and future research opportunities are identified, such as dynamic twins and continuous learning, knowledge-informed network data generation, large pre-trained model for NDTs, large language model enhanced network data generation, and construction of large-scale NDT platform.

2 Overview

2.1 Network Digital Twins

Digital twin technology [36, 37] is a virtual model that precisely represents physical devices, systems, or processes in real-time, which is revolutionizing the mobile network industry [7, 8, 10]. The concept behind NDT is to create a digital replica of the mobile network or its components [9, 14], which enables real-time monitoring, analysis, prediction, simulation, and optimization of its behavior. In particular, DTs help evaluate new algorithms, policies, and services in situations where it is too risky to apply them to real-world networks [38], or when opportunities for testing during operation are limited [39]. Effective implementation of DTs for mobile networks can significantly contribute to advancing mobile networks and systems, extending into **Beyond 5G** (**B5G**) [40, 41] or even 6G [42, 43]. Researchers, developers, and operators can conduct "what-if" analyses, emulate and test different configurations, and identify optimal operations using DTs. This advancement will enable the design and implementation of more powerful features and services, fundamentally transforming our interaction with and utilization of mobile networks.

An overview of NDTs is provided in Figure 2. NDTs consist of four steps [8, 42, 44]: data processing and network monitoring, digital replication and network simulation, designing and training RL, Sim2Real transition, and network control.

(1) **Data processing and network monitoring**. This step focuses on tracking and collecting data from a mobile network and its components, including mobile devices, base stations, and network gateways. The main goal is to gather, record, and monitor the real-time operational status of physical network devices. Historical data is collected throughout the lifecycle of

these devices, leading to the creation of data mirrors in the digital domain. This process enables accurate digital descriptions of the devices. Generative AI is vital at this step, as it can generate missing data, thus significantly enhancing the overall data integrity.

(2) **Digital replication and network simulation**. In this step, the data collected previously is used to create digital replicas of the physical network components. The primary aim of these digital twins is to facilitate detailed simulation and modeling of the network. This enables exploring various scenarios and outcomes without affecting the real-world network. This phase is crucial for conducting what-if analyses, allowing for the simulation of different network configurations, strategies, and external influences to assess their impact on network performance and behavior. Here, generative AI can generate network data under specific conditions using network configurations or historical data as input, thus achieving more precise simulations and predictions.

(3) **Designing and training RL**. In this step, attention is turned towards leveraging insights from simulations to design and train network optimization algorithms (i.e., RL). It involves employing advanced AI techniques, such as reinforcement learning, to develop models capable of analyzing complex network scenarios and suggesting optimizations. These models are trained with simulation data, ensuring they can make informed decisions regarding network management and optimization in the digital domain.

(4) **Sim2Real transition and network control**. The final step involves applying the insights and optimizations from previous phases to real-world network operations. Training RL agents in simulated environments presents a major challenge known as the "reality gap"–the discrepancies between simulated and real-world environments. The final phase is crucial for bridging this gap, ensuring that the optimized network configurations derived from simulations can be effectively implemented. This step is essential for validating the practical applicability of the digital twin's insights and confirming the network's adaptability to real-world conditions and demands. Generative AI plays a vital role at this stage by learning and adapting to the dynamic changes of real-world networks, enabling seamless integration between virtual and real-world environments.

The operation of NDTs is characterized by a dual-loop process, comprising an internal loop within the digital domain and an external loop bridging the digital and physical domains. This collaborative interaction of the two loops guarantees both the effectiveness and the timeliness of network operations and optimization.

— Internal Loop: The internal loop comprises an AI-based optimization solver and simulation models of network elements, having simulation capabilities and providing network performance statuses under various network configurations. Simulation models work iteratively with optimizers in the digital domain, continuously conducting what-if analyses and network configuration adjustments to seek optimal mobile network configurations.
— External Loop: The external loop primarily focuses on deploying network configuration parameters derived from the internal loop to the physical domain, dynamically controlling physical network elements, and updating the digital twin models based on feedback from physical network performance. This creates an interactive cycle between the digital and physical domains. This stage includes an in-depth evaluation of differences between real-world network performance and simulation model predictions. Based on these differences, the external loop undertakes steps to update and optimize the internal loop modules, achieving effective integration and coordination between the physical and digital domains.

Thanks to their "Two-Domain, Four-Step, Dual-Loop" collaborative methodology, NDTs offer an adaptable and efficient framework for network system planning, optimization, and operation.

Table 1. The Entire Lifecycle of Digital Twins in Mobile Networks

Steps	Functions	Physical-to-digital	Digital-to-digital	Digital-to-physical
① Data processing and network monitoring	Tracking and monitoring device status, collecting network data	✓	✖	✖
② Digital replication and network simulation	Simulation, conducting what-if analyses in the digital domain	✖	✓	✖
③ Designing and training network optimizer	Network optimization in the digital domain	✖	✓	✖
④ Sim2Real transition and network control	Bridging the gap between digital and physical domains	✖	✖	✓

Table 2. Advantages and Disadvantages of GAN, VAE, Diffusion Models, and AR

Methods	Fast sampling	High quality	Diversity
GAN	✓	✓	✖
VAE	✓	✖	✓
Diffusion	✖	✓	✓
AR	✖	✓	✓

Table 1 delineates and contrasts the four steps of NDTs, providing a clear comparison and overview of each stage within this innovative approach. Notably, NDTs and Digital Twin Networks are distinct concepts within digital twin technology. In Appendix A.1, we discuss their differences and relationships and how NDTs could be extended to related applications such as the metaverse.

2.2 Generative AI Methods

Generative AI models are designed to replicate the data distribution of input data through iterative training, enabling the creation of synthetic data. This section introduces four key generative models: **Variational Autoencoders (VAE)**, **Generative Adversarial Networks (GANs)**, Diffusion Models, and **Autoregressive Models (AR)**. A detailed introduction to these four key generative models can be found in Appendix A.2.

Table 2 presents the advantages and disadvantages of four major generative models: GAN, VAE, Diffusion Models, and AR. Specifically, once a GAN model is trained to convergence, its discriminator often struggles to distinguish between real and generated samples, resulting in high-fidelity generated samples. However, GAN's adversarial loss function may not fully cover the entire data distribution and can sometimes lead to mode collapse [45], where the generator repetitively produces data from a specific subset, reducing sample diversity. For VAE models, the encoder predicts the distribution of latent variables, but the overlap in the latent variable distribution for different inputs may lead to a blurred generation of samples, as the optimal decoding tends to present an average of two inputs [46]. Nonetheless, by maximizing the likelihood function, VAE ensures comprehensive coverage of all patterns in the training dataset, maintaining the diversity of generated samples. Diffusion models excel in generating high-fidelity samples thanks to their gradual noise elimination mechanism. However, the step-by-step construction approach of diffusion models makes the sample generation process relatively slow. AR, due to their serial generation of conditional distribution probabilities, are also significantly less efficient compared to other models that directly generate and sample joint distributions, thus hindering rapid data generation. Notably, no generative models can simultaneously achieve fast sampling, high quality, and diversity. Their limitations stem from each generative model's fundamental neural network design. We should select suitable models based on specific needs. For instance, given their rapid sampling capabilities, GANs and VAEs are preferable for scenarios requiring quick responses. Alternatively, diffusion and AR models are more appropriate when the scenario demands high-quality and precise simulations.

2.3 Generative AI-empowered NDTs

As an innovative solution for NDTs, generative AI technology provides comprehensive technical support across all four key stages: data processing and network monitoring, digital replication and network simulation, the design and training of network optimizers, and Sim2Real transition and network control. This approach encompasses the entire communication cycle between physical entities and their digital twins and interactions between the twins.

In the data processing and network monitoring stage, generative AI is essential for data collection and operational status detection, enabling seamless communication between physical entities and their digital twins. One major challenge in this stage is maintaining real-time synchronization during physical-to-digital communication. Data may be delayed or lost when networks experience high latency or unstable connectivity. In such cases, generative AI can predict future states by analyzing past behaviors and current trends, allowing the digital twin to preemptively update its state based on anticipated changes in the physical entity. Additionally, generative AI models can infer missing information based on observed data, enabling comprehensive data completion across both temporal and spatial dimensions. By analyzing historical patterns and current contexts, these models can generate realistic data points to fill gaps caused by sensor downtime, network interruptions, or incomplete datasets. Moreover, communication between a physical entity and its digital counterpart may occasionally fail due to hardware issues, configuration errors, or environmental disruptions. Here, generative AI plays a crucial role in fault detection and correction by identifying anomalies in communication patterns. By generating data representing a network's normal operating state, these models help identify potential issues; significant deviations between actual network data and AI-generated data may indicate unusual activity or problems requiring attention.

In the digital replication and simulation stage, the focus is on simulating network element behavior and conducting what-if analyses, encompassing digital-to-digital communication between components of the NDT system. Generative AI models can produce various types of network data, including user movement trajectories, network service behaviors, and wireless environment dynamics. By learning the conditional probabilistic distribution of network elements' data, generative AI can synthesize realistic network behavior data that closely mirror real-world scenarios. By shifting from traditional discrete-event-driven network simulations and analytical models to AI-driven network data generation, the efficiency and fidelity of these simulations are significantly enhanced. Notably, unlike conventional network simulators, where digital network elements communicate primarily through discrete event signals, the generative AI framework enables digital twins to exchange state information as conditional signals that inform behavior generation. This shift allows the digital twin to operate with continuous updates and adapt to changing conditions in real time, resulting in more accurate and dynamic simulations. This evolution towards conditional signal-based communication enables the incorporation of complex, probabilistic scenarios that traditional models struggle to replicate. For example, rather than simply reacting to pre-set events, the digital twin can dynamically model user behavior patterns, network traffic fluctuations, and various environmental conditions.

During the design and training of RL, generative AI becomes an invaluable tool for network operators, facilitating the development of control strategies within the digital domain. One of the key benefits of generative AI in this stage is its ability to reduce the communication load between

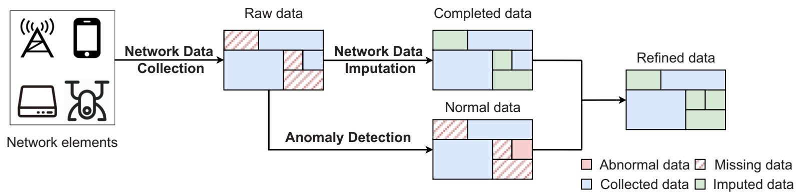

Fig. 3. Three fundamental procedures in data processing and network monitoring.

digital twins and optimizers by generating compact representations of high-dimensional network states. Generative models can transform large-scale network state data into reduced yet highly informative representations by utilizing autoencoders or other advanced compression techniques. This process significantly reduces the bandwidth required for data transmission between twins and optimizers, enabling more efficient and scalable communication. In addition to network state compression, generative AI supports optimization in RL frameworks. For example, diffusion models can act as policy networks within RL frameworks, improving optimization performance. These models are particularly well-suited for capturing complex, multimodal action distributions, thanks to their robust capability to model joint distributions. By effectively handling diverse action distributions, diffusion models facilitate more precise and adaptable control strategies, ultimately enhancing the digital twin's ability to respond to various operational scenarios.

In the Sim2Real transition and network control stage, generative AI is crucial in addressing the inaccuracies inherent in digital domain simulators. The communication gap between the digital twin and its physical counterpart often arises from differences in how the physical system operates versus how the digital model simulates it. Generative AI helps to bridge this gap by correcting discrepancies between real-time data from the physical system and the simulated data within the digital twin. Generative AI enables the digital twin to stay synchronized with the physical entity by modeling uncertainties and accounting for environmental changes in real-world systems, ensuring more accurate and timely control. These models can incorporate probabilistic elements, enabling them to account for variances in system behavior due to factors such as fluctuating environmental conditions, unexpected hardware responses, or latency in data transmission. Moreover, generative AI facilitates the simulation-to-reality transition of network strategies by modeling the differential distribution between digital and physical domains. Through this approach, generative AI can identify and learn the distribution of these differences, creating a more flexible and realistic digital twin better equipped to handle unexpected deviations. This allows operators to refine and validate control strategies in the digital domain before implementing them in real-world scenarios, significantly reducing risks associated with deployment.

In subsequent sections, we will systematically delve into these critical technologies and their applications at each step.

3 Data Processing and Network Monitoring

Data processing and network monitoring is a process that involves the collection and tracking of data from various components of a mobile network, including mobile devices, base stations, and network gateways. Its primary objective is to gather and record real-time operational status about these physical network devices. As illustrated in Figure 3, this process entails three fundamental procedures: network data collection, network data imputation and network data anomaly detection. Generative artificial intelligence is a critical component of this process as it can help generate missing data and conduct anomaly detection, which significantly enhances the overall accuracy and reliability of the data.

3.1 Network Data Collection

Network data collection is a foundational step in constructing digital twins for network elements. In practice, two primary methods are employed for data collection: network-side data measurement and device-side data measurement. For network-side data measurement, state information from network elements is gathered through multiple network interfaces [47, 48], such as the **Serving Gateway** (**SGW**), **Mobility Management Entity** (**MME**), and **Session Management Function** (**SMF**), as defined by the **3rd Generation Partnership Project** (**3GPP**). This type of data typically refers to **Measurement Report** (**MR**) data, which is often collected at intervals of 15 to 30 minutes containing key metrics such as traffic load, the number of users served, throughput, transmission latency, and call drop rates. Such data provides a broad view of the network's operational status and performance. For device-side data measurement, data is collected directly from user devices. The **Minimization of Drive Tests** (**MDT**) [49], introduced by 3GPP in Release 10, allows network operators to gather measurement data directly from user devices. Additionally, device-side measurements can be obtained through crowdsourcing, where dedicated apps or **software development kits** (**SDKs**), provided by third-party mobile analytics companies like OpenSignal[1] or Tutela[2] collect data from user devices. Device-side data is typically more fine-grained than network-side data, providing detailed information such as the **Global Positioning System** (**GPS**) location, **Reference Signal Quality** (**RSRQ**), and **Channel Quality Information** (**CQI**) of mobile users. These metrics offer deeper insights into user experiences and network performance at the device level. Beyond network performance data, device-side measurements can also be extended to gather environmental information. For instance, Yigit et al. [50] proposed a drone-assisted data collection architecture that uses drones to capture environmental data, thereby enhancing the capabilities of digital twins by integrating information about the physical surroundings of network elements.

3.2 Network Data Imputation

In the phase of monitoring mobile networks, a primary challenge we encounter is the limited observability of network data. On the one hand, due to data undersampling or transmission faults, network data often exhibits temporal gaps. Additionally, network data also faces issues of spatial sparsity, especially when data is collected from user devices and other road test equipment, where availability cannot guarantee data collection at all required locations. Fortunately, generative AI models can infer missing information based on observed data, offering completion of cross-temporal and spatial network data by approximating the posterior distribution of observed values. Generative AI models' capability helps enhance network data's completeness and provides robust support for establishing and maintaining network element mirror mappings. For example, Abiri et al. [51] developed a denoising autoencoder capable of reconstructing data by introducing stochastic noise. This method can process various data types and is realized as a stacked denoising autoencoder, which is efficient in computational time. Li et al. [52] proposed a multi-modal deep learning model based on a stacked autoencoder architecture, where two parallel autoencoders simultaneously consider spatial and temporal dependencies to effectively impute spatial and temporal traffic data. In a different approach, Yoon et al. [53] introduced **Generative Adversarial Imputation Nets** (**GAIN**) within the GAN framework to impute missing values. Te generator imputes missing components based on the observed data and outputs a completed vector. The discriminator then evaluates the completed vector to identify which components were originally observed and which were imputed. While effective for non-sequential datasets, GAIN struggles to

[1] https://www.opensignal.com/
[2] https://www.tutela.com/

handle temporal data effectively. To address the challenges with time series data, Tashiro et al. [54] proposed **Conditional Score-based Diffusion models for Imputation** (**CSDI**). CSDI leverages transformers to capture temporal features and generates missing values conditioned on the observed data based on the diffusion framework.

3.3 Network Data Anomaly Detection

Real-time detection of network anomalies is one of the critical tasks in the mirror mapping phase of wireless NDTs. By creating digital twins of network elements, we can achieve real-time, dynamic representations of physical units, allowing for meticulous monitoring and anomaly detection [55]. For instance, Yigit et al. [56] were among the first to propose leveraging digital twins for detecting network service attacks. They implemented a **Yet Another Next Generation** (**YANG**) model and an **automated feature selection** (**AutoFS**) module to handle network data and used online learning to update the model to improve detection accuracy. They then further extended their model to **cyber-physical systems** (**CPS**) in seaports [57] and 6G edge of things networks [58]. In contrast to the online learning approach, Xu et al. [59] introduced a curriculum learning method to address discrepancies between historical and real-time data in digital twin systems, ensuring more reliable anomaly detection across dynamic environments. Bolat-Akça et al. [60] developed an autoencoder-based **eXtreme Gradient Boosting** (**XGBoost**) classifier within digital twins for more accurate predictions. However, these traditional digital-twin-enabled anomaly and attack detection are modeled as supervised learning problems, relying heavily on manual labeling and classification techniques, which can be a significant limitation in dynamic, evolving network environments.

The advantage of generative AI models lies in their ability to generate data representative of a mobile network in a normal operating state. When the deviation between the actual network data and the model-generated data exceeds a certain threshold, the model can identify potential anomalies without the need for labeled data [61]. This approach is particularly effective in detecting anomalies in network data, as it can capture subtle changes that conventional detection methods might overlook. Moreover, this technique can effectively learn without extensive labeled data, which is crucial for identifying new or complex network anomalies. For example, Kong et al. [62] introduced an unsupervised GAN for multivariate time series anomaly detection, featuring a novel **active distortion transformer** (**ADT**) block. This ADT block distinguishes itself from the standard transformer by effectively leveraging prior knowledge of time sequences' overall associations. It actively distorts input sequences during reconstruction, enabling the network to recognize anomalies through sequence associations and reconstruction errors. Additionally, Almodovar et al. [63] presented LogFit, a BERT-based language model fine-tuned to recognize patterns in log data for log anomaly detection. The model demonstrates robustness in handling vocabulary changes within logs and achieves superior performance in anomaly detection.

3.4 Lessons Learned and Discussions

We summarize noteworthy literature on data processing and network monitoring using generative AI in Table 3 (Appendix A.3).

The data collection phase is crucial for building accurate digital twins of network elements. Network-side and device-side data collection methods complement each other, providing a broad and detailed view of the network's status. However, the reliance on fixed intervals (e.g., every 15–30 minutes for network-side data) and the occasional limitations of device-side data (e.g., coverage gaps or sampling irregularities) can lead to data sparsity issues. The key lessons learned are that a multi-faceted data collection approach is essential for capturing a holistic network view, yet

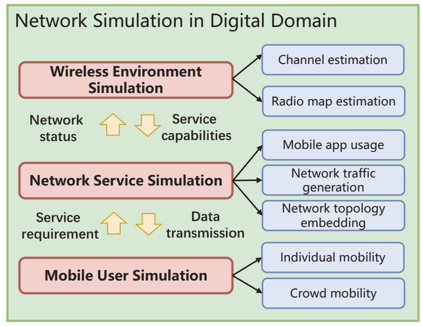

Fig. 4. Construction of digital replicas and network simulation.

addressing spatial and temporal gaps remains challenging. In such cases, generative AI can help mitigate the gaps by predicting network states based on past behaviors and current trends.

Applying generative AI models for data imputation is critical for addressing data sparsity in mobile networks. These models, such as autoencoders, GANs, and diffusion models, have effectively filled temporal and spatial data collection gaps. However, different methods are suited for different contexts. For example, while GAIN is effective for static datasets, it struggles with temporal data, whereas models like CSDI specifically target time series data. The lesson here is that selecting the appropriate AI model for the specific nature of the data (sequential, spatial, etc.) is crucial for effective imputation.

Traditional anomaly detection methods rely on supervised learning, often requiring labeled datasets and struggling to adapt to new or emerging issues. Unsupervised approaches, such as generative models (e.g., GANs and VAE), offer a promising alternative by detecting anomalies without needing labeled data. These methods enhance detection accuracy in dynamic environments and reduce dependence on manual intervention, making them particularly valuable in complex, rapidly evolving networks. However, this also underscores the need to refine these models further to manage the diversity and fast-paced changes in network environments.

In summary, combining multi-source data collection, generative AI for imputation, and unsupervised learning for anomaly detection enhances mobile network monitoring. While network- and device-side data collection complement each other, spatial and temporal gaps remain challenging. Generative AI helps fill these gaps, and unsupervised models improve anomaly detection without needing labeled data. However, ongoing refinement is required to adapt entirely to complex, rapidly changing network environments.

4 Digital Replication and Network Simulation

Constructing digital replicas and network simulations requires considering multiple factors. From the perspective of a network packet, the journey starts with the user's intent, moves to the application level through user interactions, and then traverses the network based on the network services and environments. The subsequent review will examine digital twins from the perspectives of mobile users, network services, and network environments. The organization of this section is illustrated in Figure 4.

4.1 Mobile User Simulation

Mobile users are critical components of mobile networks. Recently, the concept of a **personal digital twin (PDT)** has emerged, aimed at creating a virtual representation of an individual's behavior, preferences, and traits within a given environment [64]. A PDT is designed to encompass a person's external appearance and internal physiological details [65], allowing it to predict health outcomes, behaviors, and emotional states and simulate cognitive processes. In mobile networks, the PDT is crucial in tracking, predicting, and simulating an individual's real-time movements, habits, and needs, enabling enhanced network services such as personalized recommendations, location-based optimizations, and efficient network resource allocation [66].

Accurately modeling user mobility behavior is crucial for optimizing performance in mobile networks, and this has been the focus of extensive research. Specifically, there are two distinct approaches: individual mobility modeling and crowd mobility modeling. Individual mobility modeling represents users' mobility patterns and personalized characteristics to generate their locations and behaviors. Conversely, crowd mobility modeling aims at capturing behavioral trends and patterns within a collective group, revealing the crowd's flow patterns and aggregation behaviors.

4.1.1 Individual Mobility. Individual mobility modeling is used to model and analyze the behavior of individual mobile users. By collecting and analyzing historical data of individual users, such as location information and activity records, models can be constructed to generate users' mobile behavior. The advent of generative AI has been instrumental in generating detailed individual trajectories, utilizing a range of neural network-based models like AR, GAN, VAE, and diffusion models. AR models are adept at sequentially generating mobile user trajectory transitions. For instance, Berke et al. [67] employed a **recurrent neural network** (**RNN**) to generate trajectories, using population distributions (e.g., home and work locations) as inputs to create synthetic mobility traces. Feng et al. [68] utilized **Long Short-Term Memory** (**LSTM**) networks for mobility prediction, while attention has been used to reconstruct user trajectories, enhancing spatiotemporal resolution by leveraging periodic patterns. GANs are particularly effective in generating high-quality, realistic data for individual mobility modeling. The TrajGAN model [69], for instance, combines LSTM and GAN for trajectory generation. Ouyang et al. [70] discretized locations into a matrix, highlighting visit times and duration, and generated sequential paths accordingly. Similarly, Cao et al. [71] separated spatial and temporal data, employing GAN and Seq2Seq models for generating spatial and temporal features separately. Feng et al. [72] proposed MoveSim, which uses a GAN framework with a self-attention-based network to capture complex temporal transitions in human mobility, incorporating urban structure knowledge for realistic trajectory generation. Wang et al. [73] modeled individual movement as a human decision-making process, using generative adversary imitation learning to simulate that process. Yuan et al. [74] further integrated Maslow's need theory into this process for enhanced trajectory generation. VAEs are known for their generative diversity and ability to model complex distributions. Long et al. [75] proposed a two-layer VAE model for modeling user distributions and complex mobility patterns. Wang et al. [76] proposed a VAE-based model that integrates the classical temporal point process to represent continuous temporal distribution effectively. Recently, diffusion models have gained popularity for their precise and controllable generation capabilities. Zhu et al. [77] introduced a diffusion-based trajectory generation framework, effectively integrating diffusion models with spatiotemporal learning. Yuan et al. [78] used a diffusion model with a co-attentive module to capture the interdependence of trajectory time and space, enhancing learning at each step.

4.1.2 Crowd Mobility. Crowd mobility modeling is crucial for understanding large populations' movement patterns and behaviors. This field involves analyzing data such as location trajectories, and pedestrian flows to identify trends, aggregation phenomena, and behavioral regularities in crowds. Generative techniques in this domain span various methods, including mathematical and physics-based models, machine-learning models, and hybrids that incorporate physics and machine learning. Numerous studies have adopted mathematical and physics-based generative methods. For instance, Xu et al. [79] proposed a model highlighting the fractal-like urban morphologies and scaling laws in city growth patterns, focusing on social interactions and long-term memory in human settlement mobility data. Crabtree et al. [80] simulated migration behaviors of Australian ancestors, generating numerous probable migration pathways by considering different mechanisms in migration dynamics. Cornes et al. [81] utilized the **Social Force Model** (**SFM**) to simulate crowds in panic emergency scenarios, incorporating individual panic stress into the desired velocity and modeling the spread and fade of panic using a **Susceptible-Infected-Recovered-Susceptible** (**SIRS**) model. Chen et al. [82] focused on modeling irrational routing decisions of panicky evacuees in emergencies. Wu et al. [83] simulated a high-building evacuation event using a volume control model, treating evacuees as fluid particles in fluid flow. Additionally, machine learning-based generative techniques have gained traction in modeling crowd mobility. Yao et al. [84] proposed a residual network-based model, ResNet-SICS, for scene-independent crowd simulation, using crowd attributes as parameters and learning from real-world data. Rong et al. [85] introduced the **Graph-based Spatial-temporal Embedding with Dynamic Fusion** (**GSTE-DF**) model, which comprises node embedding learning for capturing dynamic spatial-temporal features and flow prediction for inferring population interactions. Shi et al. [86] proposed a novel method under the GNS framework, using a heterogeneous graph to model interactions among people and the environment. Zhang et al. [87] developed a **physics-integrated machine learning** (**PIML**) framework, combining physics and neural network models through iterative learning and discovery processes. These diverse approaches showcase the field's evolution and the increasing sophistication in modeling crowd mobility, leveraging the strengths of both traditional and modern computational methods.

4.2 Network Service Simulation

The concept of network service simulation involves generating data that reflects interactions between user devices and network services. Based on a review of the existing literature, three key aspects currently shape the construction of network service simulations: mobile app usage generation, network traffic generation, and network topology embedding generation.

4.2.1 Mobile App Usage. App usage can be represented as a sequence of Apps, constituting sequential temporal data [88, 89]. The basic idea of temporal-based construction is to explore the temporal correlations between app sequences to generate operation regularity about user preferences for network services. For example, an attention mechanism-based U-net framework was proposed to capture the transaction feature of users for online payment services [90]. Zhang et al. leveraged transformer to model app (un)installation behaviors [91]. The transformer block collectively modeled the installation, uninstallation, and retention embeddings that improved the quality of final user embeddings. Also, He et al. adopted Informer, a transformer variant suitable for long time-series forecasting to predict subsequent app usage in AliPay [92]. Notably, app usage behavior is associated with temporal usage sequences and the aggregation of multidimensional features in various domains, such as user preferences, textual instructions, and social bonds. There is abundant information on the relationship between user-user, app-app, and user-app interactions, which can be utilized to improve generation performance.

We can leverage services and sensor data that users interact with apps and smartphones to build the cyber-representation of users. The type of construction is of great significance for building and understanding user behaviors interacting with apps [91]. Gao et al. [93] combined the phone call and message behaviors with app usage data and designed an attention layer to fuse the multi-view data into a compact feature vector. Zhao et al. [94] designed a dual-DNN framework to extract user-personalized characteristics in app usage behaviors. An attention mechanism was inserted in the framework to depict the mutual impact of multiple apps. Chen et al. [95] further analyzed the

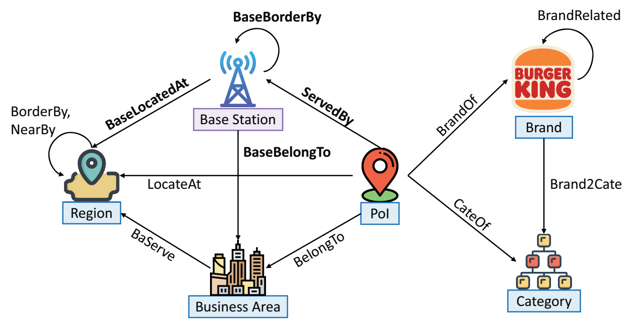

Fig. 5. The urban knowledge graph [98] for traffic generation models base stations, regions, business areas, and points of interest as entities, with their spatial correlations represented as relationships.

inaccurate demographic prediction problem caused by a single data source. The work designed an attention-based DNN network to calculate and fuse the similarities between heterogeneous datasets. Recent research has leveraged GAN to cope with the data sparsity issue. SwipeGan [96] and CAGANet [97] were proposed to generate synthetic user motion data such as touch, swipe, and acceleration. Experiment results demonstrated the feasibility of GAN-based methods for data augmenting both the **False Acceptance Rate** (**FAR**) and **True Acceptance Rate** (**TAR**), which were improved by different degrees.

4.2.2 Network Traffic Generation. Network traffic generation involves two primary tasks: generating traffic volume and generating packet traces. Traffic volume generation aims at predicting network traffic across different areas by leveraging geographical and historical traffic data in a time series format. In contrast, packet trace generation focuses on producing detailed network usage records, including packet size and packet arrival patterns.

Most existing studies working on regional traffic volume generation leverage **convolutional neural networks** (**CNNs**), framing the problem as an image generation task. For example, CartaGenie [99] utilizes population density and points of interest as conditional features to generate regional network traffic using CNNs, while SpectraGAN [100] employs a CNN-based conditional GAN to take into account periodic traffic patterns. APPShot [101], another CNN-based GAN model, further generates city-scale traffic by incorporating data on urban infrastructure, deployment density, and service usage frequency. Beyond image-based approaches, some studies employ knowledge graphs to represent urban information across regions for traffic volume generation. Hui et al. [102] proposed a GAN model that enhances traffic generation by incorporating an urban knowledge graph. As illustrated in Figure 5, in the knowledge graph, urban components, including base stations, regions, business areas, and points of interest, are modeled as entities, with their spatial correlations represented as relationships. Zhang et al. [98] advanced this approach with ADAPTIVE, a deep transfer learning framework for zero-shot city-scale cellular traffic generation, bridging target and source cities through urban knowledge graphs.

Packet trace generation is critical for creating realistic, detailed records of network activity at the packet level. Dowoo et al. [103] introduced the PcapGAN, which generates network packet traces based on time series GAN, achieving realistic representations of network traffic patterns. Building on this, Lin et al. [104] developed DoppelGANger, a model designed to generate packet attributes and feature series simultaneously, thus comprehensively capturing the temporal dependencies between packets. For addressing data imbalance in datasets with diverse traffic types, Wang et al. [105] proposed the PacketCGAN model, which improves the fidelity of generated packet traces across various traffic classes. More recently, diffusion models have introduced significant innovations in packet trace generation. NetDiffusion [106], for instance, transforms time-series data into two-dimensional images, allowing temporal correlations to be captured more effectively. This approach leverages a pre-trained model to generate these images, from which packet traces are then obtained through a decoding process. NetDiff [107] incorporates user app usage intent as a condition within a dual-layer transformer-based diffusion model, enabling it to capture complex relationships between multiple network flow features, such as packet size, inter-arrival time, source, and destination. This advancement highlights diffusion models' potential to improve the accuracy and realism of generated packet traces.

4.2.3 Network Topology Embedding. Networks comprise massive devices and heterogeneous transmission channels, naturally containing a wealth of node-edge information. Network embedding aims at representing the information by decomposing high-dimensional, non-linear network features into low-dimensional representations [108]. Extracting latent and low-dimensional features of node information in the network facilitates downstream analysis, such as spatio-temporal association, network traffic characterization, and resource scheduling.

With the help of generative models, the network embedding could be well designed. It can serve well for downstream tasks like link prediction, network completion, and latency estimation. Gao et al. [109] used GANs to generate network node proximities by training a model on a attributed network graph. Ban et al. [110] further explored structural and content similarity through network homophily, combining features of central nodes and their neighborhoods to reduce distribution bias. He et al. [111] developed a generative network with three modules: a generator, competitor, and discriminator, where the competitor generates fake latent features to guide the generator in creating effective representations for network embedding. Lei et al. [112] investigated **Wasserstein GAN** (**WGAN**) to make network embedding for link prediction. To effectively capture the spatial-temporal characteristics of the network, the generator integrates a **graph convolutional network** (**GCN**) layer and an LSTM layer. Besides link predictions, generative model-empowered network embedding could also tackle network completion, wherein dynamic links and missing nodes could be generated. Tran et al. [113] explored an autoregressive generative model to extract the latent feature of networks, with respective to edge dynamics and node relationships. The generated embedding was utilized to synthesize partial network topology for network completion. Despite topology-oriented applications, Wang et al. [114] investigated network embedding in end-to-end latency estimation for network slicing. The network embedding was generated by a GAN that took account of the service volume of different virtual network functions. The generated embedding was then fed into a GCN for latency estimation.

4.3 Wireless Environment Simulation

Modeling wireless transmission environments is a fundamental aspect of network simulation. During transmission, signals interact with the environment, leading to phenomena such as reflection, diffraction, and scattering, which make accurate modeling challenging. Recently, with the advancements in deep learning, generative AI has emerged as a promising and effective tool for simulating wireless environments, including tasks like channel estimation and radio map estimation.

4.3.1 Channel Estimation. In channel modeling, researchers commonly analyze the relationship between the **transmitter** (**Tx**) and **receiver** (**Rx**) pair, examining various characteristics such as path loss, **channel state information** (**CSI**), **delay spread** (**DS**), **frequency spread** (**FS**), **angle of departure** (**AoD**), and **angle of arrival** (**AoA**). These parameters are crucial for the design of transceivers and antenna systems, which greatly influence transmission performance. GANs have shown significant potential in modeling complex wireless channels. Xiao et al. [115] generated the CSI of a **multiple-input multiple-output** (**MIMO**) system using CNN-based GANs. In their approach, the CSI matrix was organized as an image with dimensions $N_t \times N_r \times 2$, where N_t and N_r represent the number of transmitting and receiving antennas, and the '2' corresponds to the real and imaginary components of the CSI. Hu et al. [116] applied GANs to model multi-frequency channels in mmWave wireless communication. Their method first clustered multi-path transmissions using K-means based on channel characteristics. They then used this cluster information and the receiver's location as input conditions for the GAN to model the channel's AoA and DS. Zou et al. [117] provided a comprehensive overview of the application of GANs to wireless communications, focusing on CSI prediction and signal classification. Xia et al. [118] employed a VAE model for UAV communication channel estimation. Their method involves two stages: in the first stage, a neural network predicts the link status (LoS or NLoS), while in the second stage, conditional generation is performed based on the link status and receiver coordinates, with the mean and variance of the path loss as outputs. Arvinte et al. [119] utilized a diffusion model for MIMO CSI estimation. Their results demonstrated competitive performance within the training distribution and out-of-distribution scenarios, showcasing the diversity and robustness of the diffusion model. Similarly, Sengupta et al. [120] also showed that diffusion models provide more stable training and greater diversity in generating channel characteristics like AoD and AoA compared to GANs. Moreover, after being pre-trained on a comprehensive urban macro-cellular dataset, the diffusion model demonstrated strong generalization ability on a smaller, out-of-distribution urban micro-cellular dataset.

4.3.2 Radio Map Estimation. Radio map estimation is a practical problem that, given the **Base Station** (**BS**) location and configurations and scenario geographic environment information, generates or interpolates to obtain fine-resolution Reference Signal Received Power or path loss distribution within the region of interest. A radio map reveals the coverage area of a BS and supports the network deployment and spectrum planning. Zhang et al. [121] propose a conditional GAN to complement the radio map based on sparse observations. The model consists of two phases: the upsampling for interpolation and the downsampling for refining details. Correspondingly, during the training stage, the model learns to interpolate and later learns to modify the generated radio map by considering the geometric and frequency-domain factors. Vankayala et al. [122] apply GAN to estimate the radio map, given the building layout as a prerequisite. Marey et al. [123] utilize GAN to predict path loss. It trains two GAN models with the same architecture but adopts different graphical information as input: the satellite map and the height map, respectively. The experimental results prove the model can achieve close path loss prediction performance as raytracing. Li et al. [124] rely on VAE to encode the environmental information into latent space and further leverage this hidden environmental representation in a neural network for predicting the RSRP mean and variance. Qiu et al. [125] propose a diffusion model-based approach for indoor radio map interpolation. It regards the indoor layout as the input condition during the denoise procedure. The experimental results prove that the diffusion model can achieve significant interpolation accuracy with only a few measurements.

4.4 Lessons Learned and Discussions

Tables 4–6 (Appendix A.3) present the summary of prominent literature in mobile user simulation, network service simulation, and wireless environment simulation, respectively.

Leveraging generative AI models like AR, GANs, VAEs, and diffusion models reveals the unique strengths each brings to network simulation. For example, GANs are particularly effective for modeling individual mobility and crowd behavior, while diffusion models excel in generating realistic packet traces and handling complex dependencies. However, each model has limitations, particularly regarding data requirements and generalization. Training these models requires large datasets, and they may struggle to adapt to unobserved behaviors or new environments, especially in diverse settings. Additionally, biases in training data can lead to skewed simulations, emphasizing the need for robust data handling and preprocessing.

Effective simulations depend on integrating multiple dimensions, including device density, spatial layout, and temporal factors–especially critical for crowd mobility and network traffic modeling. Urban knowledge graphs, representing data such as base station locations, functional areas, and regional connections, significantly enhance the realism of network traffic and topology simulations. These contextual factors align simulations more closely with real-world conditions, supporting network planning and optimization. Thus, knowledge graphs play a vital role in scaling network simulations for complex, real-world applications.

Accurate wireless channel simulation requires robust models capable of generalizing across diverse environments. Integrating transmitter engineering parameters and geographical and environmental information as preconditions for generative model design can enhance simulation fidelity by tailoring the generation process to specific settings. Among generative models, diffusion models offer superior sample quality and broader mode coverage compared to GANs and VAEs. Conditional diffusion models improve performance by allowing controllable generation and adapting outputs to specific conditions or requirements. Although generative models add significant detail to simulations, they are often computationally intensive. Achieving a balance between realism and efficiency–especially in complex tasks like radio map estimation–calls for a hybrid approach, combining simpler models for general tasks with more advanced models for high-priority areas. This balance is crucial to maintaining accuracy in scalable network simulations without incurring excessive computational costs.

In summary, integrating generative modeling into network simulations requires a balanced, multidimensional approach that leverages the strengths of various model types. Focusing on personalization, contextualization, and predictive capabilities supports the development of resilient, responsive network solutions capable of accurately simulating performance across complex, dynamic environments.

5 Designing and Training Reinforcement Learning

Network optimizers serve as a vital interface for network operators to control digital twin-enabled mobile networks, facilitating adaptive control and system design in the digital domain. RL has become pivotal in mobile network optimizers, particularly for NDT applications [126]. As shown in Figure 6, RL provides a framework for agents to learn optimal actions through trial-and-error experiences under network simulation environment [127]. The primary goal is to develop a policy that maps states to actions for maximum long-term reward, a process underpinned by value estimation–assigning expected rewards to state-action pairs. RL algorithms are typically classified into three categories: **Single-agent RL** (**SARL**), **Multi-agent RL** (**MARL**), and Safe RL. To elucidate how generative AI can enhance the design and training of RL algorithms, we also delve into the most advanced research focusing on the application of diffusion models in RL.

5.1 Single-agent Reinforcement Learning

SARL algorithms in NDTs have become crucial for intelligent decision-making. By learning from interactions with the network simulator, the agent adapts to meet user demands, reduce congestion, and maximize network capacity. For instance, Mseddi et al. [128] applied **Deep Q-Networks** (**DQN**) based RL to tackle resource allocations in dynamic fog computing environments. Liu et al. [129] introduced a **DT-assisted task offloading system** (**DTTOS**) that minimizes power overhead and network delay. DTTOS models task offloading as a **Markov Decision Process** (**MDP**), employing Double DQN and DTA to optimize system performance. Furthermore, Chen

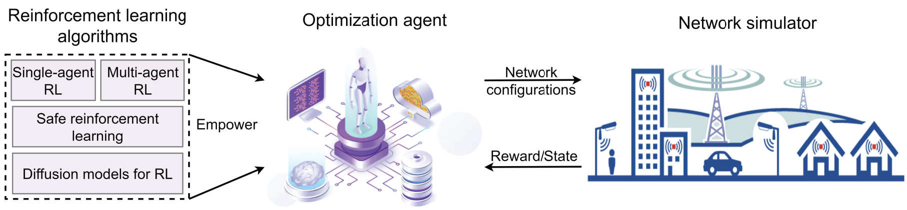

Fig. 6. Reinforcement learning provides a framework for agents to learn optimal actions through trial-and-error experiences under network simulation environment.

et al. [130] presented a resource allocation algorithm for vehicular fog computing networks, utilizing **virtual network embedding** (**VNE**) and **Deep RL** (**DRL**), which incorporates spectral graph theory and GCNs to allocate resources effectively, enhancing revenue and cost efficiency. Also, coverage optimization is a vital application of SARL in mobile NDTs. The RL agent learns to make intelligent decisions about base station placement, antenna configuration, and transmit power settings to enhance network coverage. Vannella et al. [131] optimized antenna tilt control policies using a contextual linear bandit framework, which balances exploiting the current best policy with exploring new policies and establishing theoretical guarantees such as sample complexity bounds and expected regret analysis. SARL algorithms are also widely employed for adapting to evolving network conditions, channel characteristics, and traffic demands, aiming at maximizing throughput and energy efficiency. Li et al. [132] introduced DRL-PPONSA, an intelligent routing algorithm with network situational awareness. DRL-PPONSA gathers network traffic data, predicts future traffic fluctuations, and leverages DRL for data forwarding, enhancing communication quality in dynamic wireless network topologies. Malta et al. [133] proposed using DRL to optimize energy consumption in 5G Ultra-Dense Networks. Their approach develops intelligent sleep mode management policies, dynamically activating or deactivating base station components to minimize energy consumption.

5.2 Multi-agent Reinforcement Learning

In dynamic and uncertain network environments, each network entity must make local decisions to enhance its performance. MARL offers an effective solution by enabling entities to learn optimal policies through interactions with the environment and by observing the behavior of other entities [134]. MARL is particularly well-suited to addressing the challenges of vehicular networks, fostering collaboration among agents to solve complex problems that individual vehicles cannot handle alone. Several studies highlight the effectiveness of MARL in resource allocation, demonstrating its ability to allocate resources–such as bandwidth and power–equitably among vehicles. For example, Fu et al. [135] proposed a DT-assisted framework for vehicle platooning, integrating a tactic-interactive MARL method to address cooperation aging and optimize sub-band and power allocation while meeting **quality-of-service** (**QoS**) requirements. Ji et al. [136] developed a MARL-based approach to maximize the sum rate of **vehicle-to-infrastructure** (**V2I**) links while ensuring cellular transmission rates and meeting the reliability and delay requirements of **vehicle-to-vehicle** (**V2V**) communication. Mafuta et al. [137] introduced a multi-agent **double deep Q-network** (**DDQN**) framework that combines centralized learning with distributed implementation to optimize V2I sum rates while meeting V2V reliability and delay constraints. Similarly, Gui et al. [138] proposed a federated multi-agent DRL strategy that monitors neighboring agents' actions, balances accumulated rewards, and ensures fast convergence to maintain the QoS of V2V links.

5.3 Safe Reinforcement Learning

Safe RL refers to learning policies that maximize the expected return in scenarios where ensuring system performance and safety constraints is crucial during learning and deployment. For example, Lima et al. [139] utilized a Constrained Markov Decision Process to derive an optimal resource allocation policy. They proposed a primal-dual learning algorithm alternating between updating model parameters through RL iterations and a dual variable crucial to ensuring constraint satisfaction. Zhang et al. [140] tackled the challenge of managing the **three-dimensional** (**3D**) dynamic movement of **Unmanned Aerial Vehicles** (**UAVs**) under coverage constraints. They formulated the problem as a **Constrained Markov Decision Process** (**CMDP**) and developed a **Constrained Deep Q-Network** (**cDQN**) algorithm to solve it. A primal-dual method was employed to ensure compliance with the coverage constraints by alternately training the primal and dual variables. Huang et al. [141] introduced Safe-NORA, a Safe RL approach, for dynamic resource allocation to satisfy various users' traffic demands. They utilized the P3O algorithm [142], which employs a particular penalty function to address cost constraints. Assigning a sufficiently large weight to this penalty function makes it possible to guarantee the optimal solution for the original problem.

5.4 Diffusion Models for Reinforcement Learning

Current RL algorithms often utilize Gaussian distributions to parameterize their policies, limiting action generation to exhibit a uni-modal distribution and constraining their capacity to model joint distributions for multi-variable policies effectively [143]. Thus, traditional RL approaches face challenges in managing complex network optimization tasks, such as efficiently optimizing multi-domain network configurations like antenna angles, bandwidth, and power. In contrast, diffusion models offer a more flexible approach by transforming a uni-modal Gaussian distribution into a multi-modal policy distribution through denoising. This enables diffusion models to capture and approximate more complex policy distributions to enhance RL's exploration capabilities. Recent research has explored integrating diffusion models with RL by conceptualizing the RL policy as a reverse denoising diffusion process. For example, Du et al. [144] introduced a tutorial demonstrating how diffusion models can generate optimization trajectories by leveraging expert databases. They further presented an AI-based power allocation framework employing the denoising diffusion model component as the policy network within RL [34]. This framework integrates various environmental factors–such as the wireless channel model and the number of objects in semantic communication–into the denoising conditions. The objective is to maximize the expected cumulative reward over time by optimizing the transmission power weights for each object. Furthermore, Du et al. [145] proposed the AGOD algorithm, which generates optimal discrete decisions from Gaussian noise using a diffusion model. This approach incorporates optimization constraints into the decision-making process, where the denoising component acts as the policy network to produce the best possible decisions for service provider selection.

5.5 Lessons Learned and Discussion

Table 7 (Appendix A.3) provides a summary of key literature on the design and training of reinforcement learning algorithms for NDTs.

SARL has proven effective for network applications with singular objectives, such as mobility management and power control. As network environments grow more complex and interdependencies increase, SARL's limitations become evident. These interconnected scenarios require more dynamic, multi-agent solutions. MARL, in contrast, allows individual network entities to interact and learn cooperative policies, fostering efficient resource sharing and equitable service distribution among agents. This collaborative approach is crucial when independent agents must synchronize their actions, such as resource allocation for vehicular networks, to maintain network quality and QoS requirements. Nevertheless, designing and tuning MARL is complex due to inter-agent dependencies, necessitating sophisticated coordination mechanisms for effective collaboration.

Safe RL methods emphasize prioritizing safety in mission-critical applications, particularly those using CMDPs and primal-dual algorithms. Environments such as UAV operations or constrained-resource systems require strict safety constraints to prevent operational failures. Safe RL enables systems to optimize performance while adhering to these constraints, but achieving this balance demands specialized algorithms that ensure both compliance and optimal operation. This focus on both performance and safety highlights Safe RL's essential role in situations where reliability and user safety are critical.

Traditional RL frameworks often rely on Gaussian distributions to represent policies, which limits their flexibility in handling complex, multi-modal configurations involving multiple variables, such as antenna angles, bandwidth, and power. Diffusion models offer a promising solution by transforming Gaussian-based policies into multi-modal distributions, enhancing RL's exploration capabilities, and providing more refined control over complex configurations. By enabling RL agents to explore a broader solution space, diffusion models prove particularly effective for intricate tasks like dynamic power allocation and bandwidth distribution in NDTs. As such, diffusion models represent a significant advancement, expanding RL's adaptability to complex network optimizations.

RL provides a framework for agents to learn optimal actions through trial-and-error experiences within a network simulation environment. Such trial-and-error methods are not permissible in real-world networks due to stringent safety requirements, mainly since RL necessitates frequent interactions for effective learning. Fortunately, by leveraging simulation environments, RL-based optimizers can be trained online in a digital domain rather than in real-world networks. By conducting online training in the simulation environment and applying the optimized actions to real-world networks through sim-to-real technologies, network performance can be adaptively controlled and improved. Consequently, network elements' future status and corresponding digital twins can also be predicted and continuously evolve.

These lessons highlight the importance of context-aware, collaborative, and safety-focused RL frameworks for optimizing NDTs. Incorporating advanced techniques such as diffusion models and primal-dual optimization methods shows significant promise for future development, enhancing RL's ability to address the complex demands of adaptive network management in modern, large-scale mobile networks.

6 Sim2Real Transition and Network Control

Network simulators offer a controlled and cost-effective environment for training and evaluating RL algorithms. When RL agents are trained in simulation twins, a wide range of behaviors can be explored, and optimal policies can be developed without the risk of damaging real-world systems. However, there is a significant challenge in this area known as the "reality gap" - the differences between simulated and actual environments. As shown in Figure 7, the Sim2Real transition and network control step must effectively bridge this gap to ensure the effectiveness of implementing optimized network configurations derived from network simulations. Several strategies have been proposed to address this challenge, including domain randomization, system identification, meta RL, and transfer learning.

6.1 Domain Randomization

Domain randomization is a powerful technique to tackle the sim2real challenge by systematically introducing variations in simulator parameters. This process aims at capturing the diversity and variability of real-world conditions, enabling the development of adaptable and resilient controllers

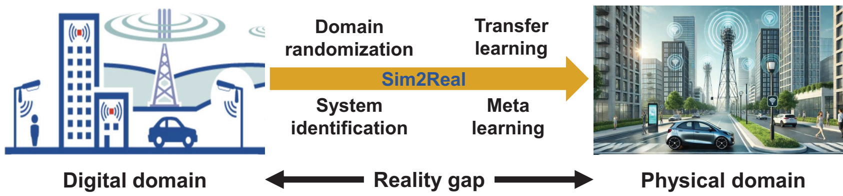

Fig. 7. There are notable differences between simulated environments and actual physical settings, creating a reality gap that must be addressed for effective Sim2Real transitions. Various strategies have been proposed to bridge this gap, including domain randomization, system identification, meta RL, and transfer learning.

in real-world applications. A foundational study by Wang et al. [146] pioneered this approach, using simulator randomization to generate a rich dataset for training. By exploring various initial conditions, external disturbances, goal variations, and actuator noise, their work significantly improved the robustness of controllers, laying the groundwork for subsequent advancements. Building on this foundation, Mordatch et al. [147] applied domain randomization through a finite model ensemble to optimize the trajectories of a humanoid robot, achieving one of the earliest successful sim2real transfers. Their work demonstrated the practical value of domain randomization in real-world scenarios. Further advancing the field, Siekmann et al. [148] implemented model-free RL in settings with uniformly distributed dynamic parameters and randomized simulations. Their study highlighted the critical role of recurrent architectures and terrain randomization in overcoming the sim2real gap, underscoring domain randomization's essential contributions to reliable, real-world performance.

6.2 System Identification

Simulators, though advanced, are not perfect replicas of the real world. Key physical parameters, such as those influenced by temperature, humidity, positioning, and wear and tear, can fluctuate dramatically, adding complexity to the task of network control. System identification seeks to bridge this gap by creating precise mathematical models that closely reflect real-world physical systems [149]. This technique has gained significant traction in aligning simulation parameters with real-world dynamics. Building on foundational research by Gautier et al. [150] and Khosla et al. [151], system identification has established critical principles for parameter tuning to bridge the reality gap. Bayesian inference techniques [152] are widely employed in this alignment process to determine the posterior distribution of simulation parameters. Comparing simulated trajectories with actual real-world trajectories helps identify simulation parameters that accurately align with the dynamics of real-world and simulation environments. In further advancements, Chebotar et al. [153] have formulated this inference challenge as an RL problem, aiming at minimizing trajectory discrepancies between simulations and real-world movements. Despite these RL-based methods, challenges remain when simulations struggle to capture the full complexity and dynamics of the real world. Thus, some researchers [154, 155] have introduced low-dimensional latent representations developed through regression models to align simulation parameters with observed dynamics, offering a more flexible and adaptive approach that can adjust to changing conditions.

6.3 Meta-learning

Meta-learning [156, 157] focuses on developing the ability to adapt to unseen test tasks by leveraging experiences from multiple training tasks. An effective meta-learning model should be trained across a diverse range of learning tasks and optimized to perform well across various task distributions, including those not encountered during training. This field has gained significant attention in recent years, leading to increased research exploring its potential to address the Sim2Real problem. For example, Wang et al. [158] and Nagabandi et al. [159] have successfully integrated meta-learning principles into the framework of RL. During the meta-training phase, a wide array of domain instances is introduced to the model, exposing it to various environmental variations. By confronting the model with these diverse experiences, it becomes better equipped to handle unforeseen situations in real-world settings. As highlighted by Finn et al. [157], the primary goal of meta-learning is to identify a set of initial model parameters–referred to as initial weights–that can efficiently generalize to new tasks with minimal fine-tuning. This capability to rapidly adapt to novel tasks makes meta-learning a powerful tool in dynamic environments, ensuring that the model can effectively address the same task even when environmental conditions change, particularly during the transition of Sim2Real.

6.4 Transfer Learning

Transfer learning aims at improving the performance of target learners in specific domains by leveraging knowledge from different yet related source domains. In the Sim2Real transfer context–where simulated environments are used to develop policies for reliable real-world performance–several innovative approaches address the challenges posed by the reality gap. Christiano et al. [160] introduced a method that trains a deep inverse dynamics model. This model bridges the reality gap by transforming actions generated by a policy into adjusted actions more suitable for real-world execution. The core idea is that applying the original action in the real-world system and the transformed action in the simulator should lead to the same subsequent state. This method minimizes discrepancies between simulated and real-world behavior by aligning outcomes across these two domains. Hanna and Stone [161] proposed an action transformation framework to ensure that actions in the simulator produce outcomes identical to those generated by corresponding actions in the real world. By learning to adjust simulator-generated actions to match real-world dynamics, this framework further aligns the two domains, enhancing the performance of transferred policies. Rusu et al. [162] expanded on these concepts with an end-to-end learning system based on a progressively expanding neural network architecture, demonstrating how perception and control can be integrated for complex control tasks. The network training occurs in two phases: the initial segment is trained entirely in simulation to develop general representations, while a second segment, added during Sim2Real transfer, relies on real-world data to fine-tune the system. This approach ensures the final model effectively adapts to the real-world environments.

6.5 Lessons Learned and Discussion

The transition from simulation to real-world environments in RL for network control presents distinct challenges and opportunities. Table 8 (Appendix A.3) provides a summary of key literature. Through practical experiences, several critical insights have emerged, highlighting key factors that enhance RL model performance and reliability when deployed in complex, unpredictable real-world settings.

Network simulations offer a safe, cost-effective environment for RL training but often introduce a reality gap due to simplified or idealized conditions. This gap varies depending on the application, network dynamics, and environmental factors, requiring adaptive strategies tailored to each use case. Bridging the reality gap demands flexible, context-specific approaches that effectively capture real-world dynamics.

Domain randomization has proven valuable in building robust models capable of generalizing to real-world conditions by introducing controlled variations in simulated environments. However, excessive randomization may lead to inefficiencies or cause models to overfit to unlikely scenarios, potentially reducing performance in practical applications. Striking the right balance between diversity and realism in simulated environments is essential to avoid tradeoffs between robustness and accuracy. Aligning simulation parameters with real-world dynamics through system identification is another critical step in minimizing discrepancies. This process, however, can be computationally intensive, especially for highly dynamic or nonlinear systems. Techniques like Bayesian inference and other statistical methods are promising but often require extensive real-world data, which can be challenging to obtain. Accurate system identification remains essential but must be balanced against practical constraints on resources and data availability.

Meta-learning has shown significant potential for enabling RL models to adapt to new tasks and evolving environmental conditions, making it particularly suited for dynamic network control. However, achieving effective generalization requires training on diverse tasks, which increases computational demands and complexity. As a result, effective meta-learning often depends on access to varied training data and robust computational resources. Transfer learning approaches, including action transformation frameworks, have also shown promise in reducing the reality gap by aligning simulated actions with real-world counterparts. However, the success of transfer learning relies on the similarity between simulated and real-world tasks. Task-sensitive methods that can dynamically adjust to discrepancies between simulation and reality are necessary to ensure reliable performance in real-world applications.

These insights underscore a central principle in Sim2Real research: successful model or policy transfer depends on a balanced combination of robust, simulation-based training and iterative real-world adaptation. By integrating simulated training with real-world feedback, these methods aim at bridging the reality gap, creating theoretically sound and practically effective models for real-world network control. Deploying RL models from simulation to reality is inherently iterative, requiring continuous testing, feedback, and refinement. Field tests, real-time feedback, and incremental adjustments help identify and address residual discrepancies. Establishing feedback loops that update model parameters based on real-world outcomes enhances both effectiveness and stability, reinforcing the model's adaptability.

7 Open Challenges and Future Directions

Generative AI applications in NDTs represent a highly promising and rapidly evolving field, which has garnered substantial attention in recent years and led to significant achievements. However, despite these advances, various challenges and unresolved issues remain that need further exploration. This section will delve into these open challenges and outline potential directions for future research and development within this field.

7.1 Latency and Synchronization in Network Digital Twins

Latency and synchronization are critical challenges in developing effective and reliable NDTs. Data collection from distributed network elements–such as base stations, routers, and mobile devices–often encounters significant delays due to network congestion and bandwidth limitations [163, 164]. For an NDT to function effectively, it must process this data in near real-time to ensure that its state accurately reflects the current conditions of the physical network. Moreover, latency may cause network elements to update their states at different rates [165]. Such inconsistencies can lead to a fragmented or incomplete view of the network, undermining the reliability of the NDT for proactive applications.

AI for predictive synchronization offers significant potential to address these challenges. Machine learning models, particularly generative AI models trained on historical data, can predict and preemptively adjust to fluctuations in network states, effectively smoothing synchronization between the digital twin and the physical network. By estimating future states, these predictive models enable the NDT to anticipate and compensate for delays, enhancing its accuracy and responsiveness [166]. However, some predictive methods are large in size and cannot be executed effectively on edge devices, such as base stations and mobile phones. In such cases, distributed computing and edge-cloud integration provide promising solutions. Computationally intensive tasks can be offloaded from resource-constrained edge devices to more capable cloud servers by combining edge computing with cloud resources. For example, Nan et al. [16] proposed a general architecture selection framework that optimizes the use of local processors, edge servers, and cloud servers for traffic prediction in NDTs. This hybrid approach facilitates efficient data processing while ensuring low-latency, synchronized updates within the NDT.

7.2 Dynamic Twins and Continuous Learning

Dynamic twins are essential in mobile networks due to the constantly evolving nature of network environments. Managing these dynamic twins presents two principal challenges: real-time data uncertainty and the continuous updating of digital models.

Data reliability in the fast-paced context of mobile networks often suffers from factors such as inconsistent data quality and noise. This noise and uncertainty can significantly impact the performance of dynamic twins, making accurate representation and prediction difficult. To address these issues, Kapteyn et al. [167] have suggested employing probabilistic graphical models to represent uncertainty in real-time data. However, the performance of these models is somewhat constrained by the limited capacity of Gaussian distributions to effectively model multi-modal distributions. Thus, enhancing uncertainty modeling within generative AI is an important research direction. This involves developing algorithms capable of comprehensively understanding and adapting to the uncertain nature of network environments and data streams, enabling them to model complex multi-modal distributions more effectively.

Continuous learning is another critical research area for dynamic twins, particularly in updating digital models. This concept entails that digital twins remain constantly learning and adapting, reflecting real-time changes in the physical network. Hashash et al. [168] have introduced an initial edge continual learning framework for simple DNNs in mobile networks. However, a robust continual learning framework designed explicitly for generative AI-driven dynamic twins, especially within large-scale and edge-cloud infrastructures, remains underdeveloped and requires further exploration. Developing such frameworks will improve the adaptability and accuracy of dynamic twins and enhance their ability to operate effectively in the complex, real-time environments typical of modern mobile networks. Future research should focus on creating scalable, efficient algorithms that integrate continuous learning mechanisms while managing uncertainty, ultimately leading to more resilient and reliable NDTs.

7.3 Knowledge-informed Network Data Generation

Generative AI technology primarily employs a data-driven approach for modeling and generation. However, this approach often struggles to account for scenarios not represented in the training dataset, leading to a lack of generalizability. Therefore, there is a growing demand for knowledge-informed generative AI technology research. This method is expected to enhance the generalizability and realism of models in complex environments by incorporating extensive domain knowledge.

Knowledge can be represented in various forms, including algebraic and differential equations, urban spatial relations, logical rules, and expert insights. Generally, there are three methodologies for integrating knowledge into AI models: through input data, model structure, and loss function design. For example, Hui et al. [102] and Zhang et al. [98] utilize urban knowledge graphs as conditional input data, depicting the spatial correlations between base stations, regions, and business areas to improve the simulation of base station traffic. Urban knowledge graphs effectively capture urban spatial dynamics and semantic features. Additionally, Yuan et al. [74] apply Maslow's

Hierarchy of Needs to model human decision-making processes, employing a hierarchical neural network structure to reflect different levels of human needs, thereby adding psychological realism to generative models. Knowledge-informed generative AI technology is precious in wireless environment simulations with established physical models. For instance, channel models can be accurately described using foundational physical equations, such as Maxwell and Helmholtz equations. Jiang et al. [169] have integrated these equations into loss function design for precise modeling of electromagnetic fields and wave propagation. Moreover, incorporating logical rules that govern network protocols into generative AI models presents a promising avenue for enhancing network packet generation. By embedding these rules, generative models can produce more realistic and contextually appropriate network traffic, improving the efficiency and reliability of simulations. This fusion of domain-specific knowledge with generative AI capabilities enhances simulation fidelity and paves the way for various applications, ranging from network design to management.

7.4 Large Pertained Model for Network Digital Twins

Generative Pre-trained Transformers (**GPTs**) have shown impressive generative capabilities in various domains. However, no large-scale generative model is specifically tailored for the mobile network domain. This highlights the need for a domain-specific generative model designed for mobile networks, which is currently an unresolved challenge. While efforts have been made to adapt the GPT framework to train a generative pre-trained transformer for single-model network packet data [170], it is not enough. It is essential to understand that a general, large-scale pre-trained model for mobile networks must effectively explore and model the multimodal characteristics of network data, including spatiotemporal coupling and sequence correlations. Therefore, a large generative model for networking should inherently be multimodal, aiming at generating and predicting network spatiotemporal data accurately. This includes elements such as base station traffic, user trajectories, service behavior, and network performance. Such a multimodal large generative model would play a crucial role in simulating a network's digital twin. One promising approach could be to develop a multimodal diffusion model for joint data generation coupled with graph neural networks and transformers. This combination would be specifically designed to capture the spatial and temporal correlations across different modalities of data.

7.5 Large Language Model Enhanced Network Data Generation

In recent years, **Large Language Models** (**LLMs**) have made remarkable strides, demonstrating the ability to generate high-quality, coherent natural language text by learning from extensive corpora of textual data. For instance, **Bidirectional Encoder Representations from Transformers** (**BERT**) [171], developed by Google, introduced a bidirectional approach that enables the model to understand the context of words from both directions. BERT's success in question answering, sentiment analysis, and natural language inference laid the groundwork for subsequent models. Building on this, OpenAI's GPT [172] employed a unidirectional transformer architecture for text generation, with models like GPT-3 and GPT-4 excelling in content creation, chatbots, and language translation. Following these innovations, Google developed the **Pathways Language Model** (**PaLM**) [173], which powers Bard, their AI-driven conversational agent. PaLM enhances multilingual capabilities and reasoning, making it practical for conversational tasks and complex applications like code generation. Meanwhile, Meta's **Large Language Model Meta AI** (**LLaMA**) [174] focuses on efficiency, designed to be smaller and faster than its predecessors, thus broadening accessibility for research and development. Additionally, Falcon [175], developed by the Technology Innovation Institute, provides a high-performing open-source alternative, delivering scalability and competitive performance in tasks such as text generation and summarization.

Against this backdrop, the idea of applying text-based LLM techniques to network spatiotemporal data, with the aim of facilitating network data generation through LLMs, has emerged as an innovative research avenue. One promising approach is to harness the capabilities of LLMs in natural language processing to manage operational log data within the mobile network domain. This would involve accurately interpreting user commands and intentions. For example, Huang et al. [176] have introduced a framework that utilizes generative AI models to discern user optimization intentions and automate network optimization. By analyzing and learning from a vast array of language data, these models can effectively assist in understanding and predicting user intentions, enabling more personalized and precise digital twin simulations for networks.

7.6 Causal Reasoning in Network Digital Twins

Causal inference techniques offer significant benefits for NDTs by enabling more accurate simulations that capture cause-and-effect relationships among network variables. Understanding these causal links is essential in complex network environments where factors like traffic load, latency, and resource allocation are closely interconnected. For instance, when optimizing network configurations, causal inference helps determine whether performance shifts result from specific adjustments–such as bandwidth changes–or merely the influence of varying user demands. Additionally, causal reasoning allows operators to trace network issues back to their sources, enhancing the generalizability of solutions and models.

Various methodologies support the integration of causal reasoning into NDT frameworks. For example, **structural causal models** (**SCMs**) are commonly used to represent cause-and-effect dynamics mathematically, providing a foundation for inferring causal relationships within network systems [177]. SCMs enable NDTs to simulate interventions and predict outcomes with greater precision. Another valuable tool is do-calculus, which allows NDTs to examine the potential impacts of specific changes–such as traffic flow prioritization–within a controlled virtual environment [178]. Initial studies [179] indicate that incorporating causal AI can reduce the need for re-training by identifying more resilient and generalizable patterns within the data. Integrating causal reasoning into NDTs is essential to achieving autonomous mobile networks.

7.7 NDTs Towards Artificial General Intelligence-Native 6G Networks

Artificial General Intelligence (**AGI**)-native networks are envisioned as a transformative framework for 6G and beyond, designed to incorporate human-like cognitive abilities such as reasoning, planning, perception, and common sense. These networks aim at moving beyond handling predefined tasks, enabling them to adapt to unforeseen scenarios and make context-aware decisions autonomously. Inspired by the **Joint-Embedding Predictive Architecture** (**JEPA**) proposed by Yann LeCun, Walid Saad et al. [180] presented a framework for AGI-native networks centered around a cognitive architecture composed of three key components: the Perception Module, the World Model, and the Action-Planning Module. The Perception Module abstracts real-world elements into generalizable representations, distilling essential features from complex network data. The World Model functions as the network's predictive engine, simulating and reasoning about potential future states based on observed conditions, much like human anticipation of outcomes. Finally, the Action-Planning Module directs the network's responses and enables the network to adapt actions based on overarching goals (such as optimizing resource allocation) or specific objectives (like minimizing latency for critical services).

AGI-native networks are the ultimate evolution of NDTs. Traditional NDTs collect data from real-world networks (analogous to the perception module), simulate network conditions to perform what-if analyses (similar to the world model), and make control decisions to optimize network performance (a precursor to action planning). However, while NDTs support predictive and adaptive capabilities, they lack the cognitive intelligence fundamental to AGI-native networks. By embedding common sense, analogical reasoning, and flexible planning, AGI-native networks extend NDT capabilities, allowing for autonomous adaptation to novel, out-of-domain scenarios, predictive configuration adjustments, and proactive resource management in complex, multi-agent environments. This cognitive evolution positions AGI-native networks as foundational for advanced, autonomous 6G networks, capable of supporting complex applications such as real-time immersive metaverse experiences, holographic teleportation, and cognitive avatars

7.8 Construction of Large-scale Network Digital Twin Platform

Efficient computation is crucial when building a network digital twin platform for large-scale systems. Time is a limiting factor for both network data and decision-making processes. Slow computation and feedback significantly reduce the practical value of the twin platform. For large-scale urban-level mobile network twinning, which involves thousands or even tens of thousands of small cells, traditional single servers are inadequate due to the large number of computational tasks. It may require dozens or even hundreds of servers, so a cloud-based computing platform that can provide high-speed, cost-effective, and scalable computing services is the best solution. Regarding software and hardware design, **Graphics Processing Units (GPUs)** must be introduced in addition to the traditional **Central Processing Units (CPUs)** for parallel acceleration. For instance, the GenNet [14] network twinning platform improved up to 175 times by incorporating GPUs and CPU acceleration. Furthermore, digital twin systems in mobile networks necessitate collecting vast amounts of data. This situation calls for the design of advanced database management techniques tailored for efficient data compression and management. To minimize data interactions and streamline the overall management process, one can study how to collect data based on computational demands and updating frequencies.

7.9 Privacy and Ethical Considerations

The ethical, privacy, and security concerns associated with NDTs and generative AI technology are significant and demand careful attention. Privacy is a primary concern due to the extensive data collection involved in NDTs. These systems monitor real-time user behavior, creating privacy risks if sensitive data is exposed or misused. Compliance with data protection regulations, such as GDPR [181], is essential to ensure users' personal information remains confidential and secure. Techniques like data anonymization and strict access controls are recommended to protect personally identifiable information from unauthorized access. The application of generative AI in NDTs also raises ethical concerns. Issues like bias in AI algorithms, transparency, and accountability are critical, as these systems may inadvertently prioritize certain network configurations, potentially disadvantaging specific user groups or regions. Ensuring fairness and accountability in AI decision-making is essential, especially in applications like public service networks, where the impact on users can be profound. Security risks are also prevalent, as attacks targeting the digital twin infrastructure could lead to unauthorized access to network configurations, resulting in data breaches or network disruptions. Robust security measures such as encryption, multi-factor authentication, and real-time anomaly detection are crucial to mitigate these risks. By addressing these ethical, privacy, and security considerations, NDT developers and operators can promote a more secure and responsible deployment of digital twins in network environments.

8 Conclusion

In this article, we present a comprehensive survey on generative AI-empowered NDTs from various perspectives. We begin with an introduction to the 'Two-Domain, Four-Step, Dual-Loop" architecture of NDT and briefly outline how generative AI can empower NDTs. Following this, we systematically review recent advancements within the four steps of this architecture: data processing and network monitoring, digital replication and network simulation, the design and training of network optimizers, and Sim2Real transition and network control. Specifically, in the data processing and monitoring phase, generative AI plays a crucial role in inferring missing data and detecting anomalies. During the digital replication and simulation stage, the simulation of network behaviors and what-if analyses significantly improves by generating various network data. In the phase of designing and training network optimizers, generative AI aids in the development of control strategies and enhances optimization performance through diffusion models in reinforcement learning. In the Sim2Real transition and network control stage, generative AI effectively narrows the divide between digital simulations and real-world network operations, facilitating a seamless transition of network strategies from simulation to practical implementation. Finally, we conclude the article with a discussion on the pressing challenges and future research directions in the field of generative AI-empowered NDTs.

References

[1] Lusheng Wang and Geng-Sheng G. S. Kuo. 2012. Mathematical modeling for network selection in heterogeneous wireless networks–A tutorial. *IEEE Communications Surveys and Tutorials* 15, 1 (2012), 271–292.

[2] Dmitri Moltchanov, Eduard Sopin, Vyacheslav Begishev, Andrey Samuylov, Yevgeni Koucheryavy, and Konstantin Samouylov. 2022. A tutorial on mathematical modeling of 5G/6G millimeter wave and terahertz cellular systems. *IEEE Communications Surveys and Tutorials* 24, 2 (2022), 1072–1116.

[3] Haoye Chai, Huandong Wang, Tong Li, and Zhaocheng Wang. 2024. Generative ai-driven digital twin for mobile networks. *IEEE Network* 38, 5 (2024), 84–92.

[4] Osvaldo Simeone. 2018. A very brief introduction to machine learning with applications to communication systems. *IEEE Transactions on Cognitive Communications and Networking* 4, 4 (2018), 648–664.

[5] Kan Zheng, Zhe Yang, Kuan Zhang, Periklis Chatzimisios, Kan Yang, and Wei Xiang. 2016. Big data-driven optimization for mobile networks toward 5G. *IEEE Network* 30, 1 (2016), 44–51.

[6] Linglong Dai, Ruicheng Jiao, Fumiyuki Adachi, H. Vincent Poor, and Lajos Hanzo. 2020. Deep learning for wireless communications: An emerging interdisciplinary paradigm. *IEEE Wireless Communications* 27, 4 (2020), 133–139.

[7] Whitepaper. 2019. *Simplifying 5G with the Network Digital Twin.* Sprient.

[8] Fengxiao Tang, Xuehan Chen, Tiago Koketsu Rodrigues, Ming Zhao, and Nei Kato. 2022. Survey on digital twin edge networks (DITEN) toward 6G. *IEEE Open Journal of the Communications Society* 3 (2022), 1360–1381.

[9] Quan Yu, Jing Ren, Yinjin Fu, Ying Li, and Wei Zhang. 2019. Cybertwin: An origin of next generation network architecture. *IEEE Wireless Communications* 26, 6 (2019), 111–117.

[10] Juan Deng, Qingbi Zheng, Guangyi Liu, Jielin Bai, Kaicong Tian, Changhao Sun, Yujie Yan, and Yitong Liu. 2021. A digital twin approach for self-optimization of mobile networks. In *Proceedings of the 2021 IEEE Wireless Communications and Networking Conference Workshops.* IEEE, 1–6.

[11] Paul Almasan, Miquel Ferriol-Galmés, Jordi Paillisse, José Suárez-Varela, Diego Perino, Diego López, Antonio Agustin Pastor Perales, Paul Harvey, Laurent Ciavaglia, Leon Wong, et al. 2022. Network digital twin: Context, enabling technologies, and opportunities. *IEEE Communications Magazine* 60, 11 (2022), 22–27.

[12] Marco Polverini, Ilaria Germini, Antonio Cianfrani, Francesco G. Lavacca, and Marco Listanti. 2023. A digital twin based framework to enable "What-If" analysis in BGP optimization. In *Proceedings of the NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium.* IEEE, 1–6.

[13] Weizheng Wang, Yaoqi Yang, Latif U. Khan, Dusit Niyato, Zhu Han, and Mohsen Guizani. 2023. Digital twin for wireless networks: Security attacks and solutions. *IEEE Wireless Communications* 31, 3 (2023), 278–285.

[14] Jiahui Gong, Qiaohong Yu, Tong Li, Haoqiang Liu, Jun Zhang, Hangyu Fan, Depeng Jin, and Yong Li. 2023. Scalable digital twin system for mobile networks with generative AI. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services.* 610–611.

[15] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* 34, 6 (2017), 26–38.

[16] Haihan Nan, Ruidong Li, Xiaoyan Zhu, Jianfeng Ma, and Dusit Niyato. 2023. An efficient data-driven traffic prediction framework for network digital twin. *IEEE Network* 38, 1 (2023), 22–29.

[17] Dennis Baker, Anthony Ephremides, and Julia Flynn. 1984. The design and simulation of a mobile radio network with distributed control. *IEEE Journal on Selected Areas in Communications* 2, 1 (1984), 226–237.

[18] Joel Short, Rajive Bagrodia, and Leonard Kleinrock. 1995. Mobile wireless network system simulation. In *Proceedings of the 1st Annual International Conference on Mobile Computing and Networking*. 195–209.

[19] Dr S Smys, Dr Haoxiang Wang, and Dr Abul Basar. 2021. 5G network simulation in smart cities using neural network algorithm. *Journal of Artificial Intelligence and Capsule Networks* 3, 1 (2021), 43–52.

[20] Lelio Campanile, Marco Gribaudo, Mauro Iacono, Fiammetta Marulli, and Michele Mastroianni. 2020. Computer network simulation with ns-3: A systematic literature review. *Electronics* 9, 2 (2020), 272.

[21] Min Chen, Yiming Miao, Iztok Humar, Min Chen, Yiming Miao, and Iztok Humar. 2019. Introduction to OPNET network simulation. In *OPNET IoT Simulation*. Springer, Singapore. https://doi.org/10.1007/978-981-32-9170-6_2

[22] Giovanni Nardini, Dario Sabella, Giovanni Stea, Purvi Thakkar, and Antonio Virdis. 2020. Simu5G–An OMNeT++ library for end-to-end performance evaluation of 5G networks. *IEEE Access* 8 (2020), 181176–181191.

[23] Linbo Hui, Mowei Wang, Liang Zhang, Lu Lu, and Yong Cui. 2022. Digital twin for networking: A data-driven performance modeling perspective. *IEEE Network* 37, 3 (2022), 202–209.

[24] Timothy J. O'shea, Tamoghna Roy, and Nathan West. 2019. Approximating the void: Learning stochastic channel models from observation with variational generative adversarial networks. In *Proceedings of the 2019 International Conference on Computing, Networking and Communications* . IEEE, 681–686.

[25] Preben Mogensen, Wei Na, István Z Kovács, Frank Frederiksen, Akhilesh Pokhariyal, Klaus I. Pedersen, Troels Kolding, Klaus Hugl, and Markku Kuusela. 2007. LTE capacity compared to the Shannon bound. In *Proceedings of the 2007 IEEE 65th Vehicular Technology Conference-VTC2007-Spring*. IEEE, 1234–1238.

[26] Giovanni Giambene. 2014. *Queuing Theory and Telecommunications*. Vol. 585. Springer.

[27] Shyam Varan Nath, Pieter Van Schalkwyk, and Dan Isaacs. 2021. *Building Industrial Digital Twins: Design, Develop, and Deploy Digital Twin Solutions for Real-world Industries using Azure Digital Twins*. Packt Publishing Ltd.

[28] Nassim Khaled, Bibin Pattel, and Affan Siddiqui. 2020. *Digital Twin Development and Deployment on the Cloud: Developing Cloud-friendly Dynamic Models using Simulink®/SimscapeTM and Amazon AWS*. Academic Press.

[29] Abdullah Aziz, Olov Schelén, and Ulf Bodin. 2023. Digital twin as a proxy for industrial cyber-physical systems. In *Proceedings of the 2023 10th International Conference on Wireless Communication and Sensor Networks*. 85–92.

[30] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *Computing Surveys* 56, 4 (2023), 39 pages.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[32] Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L. Bockting. 2023. ChatGPT: Five priorities for research. *Nature* 614, 7947 (2023), 224–226.

[33] Mladan Jovanovic and Mark Campbell. 2022. Generative artificial intelligence: Trends and prospects. *Computer* 55, 10 (2022), 107–112.

[34] Baoxia Du, Hongyang Du, Haifeng Liu, Dusit Niyato, Peng Xin, Jun Yu, Mingyang Qi, and You Tang. 2023. YOLO-based semantic communication with generative AI-aided resource allocation for digital twins construction. *IEEE Internet of Things Journal* 11, 5 (2023), 7664–7678.

[35] Jinbo Wen, Jiangtian Nie, Jiawen Kang, Dusit Niyato, Hongyang Du, Yang Zhang, and Mohsen Guizani. 2024. From generative ai to generative internet of things: Fundamentals, framework, and outlooks. *IEEE Internet of Things Magazine* 7, 3 (2024), 30–37.

[36] Fei Tao, He Zhang, Ang Liu, and Andrew YC Nee. 2018. Digital twin in industry: State-of-the-art. *IEEE Transactions on Industrial Informatics* 15, 4 (2018), 2405–2415.

[37] Fei Tao and Qinglin Qi. 2019. Make more digital twins. *Nature* 573, 7775 (2019), 490–491.

[38] Peng Yu, Junye Zhang, Honglin Fang, Wenjing Li, Lei Feng, Fanqin Zhou, Pei Xiao, and Song Guo. 2023. Digital twin driven service self-healing with graph neural networks in 6G edge networks. *IEEE Journal on Selected Areas in Communications* 41, 11 (2023), 3607–3623.

[39] Clement Ruah, Osvaldo Simeone, and Bashir Al-Hashimi. 2023. A Bayesian framework for digital twin-based control, monitoring, and data collection in wireless systems. *IEEE Journal on Selected Areas in Communications* 41, 10 (2023), 3146–3160.

[40] Rui Dong, Changyang She, Wibowo Hardjawana, Yonghui Li, and Branka Vucetic. 2019. Deep learning for hybrid 5G services in mobile edge computing systems: Learn from a digital twin. *IEEE Transactions on Wireless Communications* 18, 10 (2019), 4692–4707.

[41] Huan X. Nguyen, Ramona Trestian, Duc To, and Mallik Tatipamula. 2021. Digital twin for 5G and beyond. *IEEE Communications Magazine* 59, 2 (2021), 10–15.

[42] Latif U. Khan, Walid Saad, Dusit Niyato, Zhu Han, and Choong Seon Hong. 2022. Digital-twin-enabled 6G: Vision, architectural trends, and future directions. *IEEE Communications Magazine* 60, 1 (2022), 74–80.

[43] Huaqiang Gao, Pekka Kyosti, Xiang Zhang, and Wei Fan. 2023. Digital twin enabled 6G radio testing: Concepts, challenges and solutions. *IEEE Communications Magazine* 61, 11 (2023), 88–94.

[44] Yiwen Wu, Ke Zhang, and Yan Zhang. 2021. Digital twin networks: A survey. *IEEE Internet of Things Journal* 8, 18 (2021), 13789–13804.

[45] Zhaoyu Zhang, Mengyan Li, and Jun Yu. 2018. On the convergence and mode collapse of GAN. In *Proceedings of the SIGGRAPH Asia 2018 Technical Briefs*. 1–4.

[46] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. 2021. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 7327–7347.

[47] 3GPP TS 28.552. 2021. Management and orchestration; 5G performance measurements. (2021).

[48] 3GPP TS 32.425. 2022. Performance Management (PM); Performance measurements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN). (2022).

[49] 3GPP TS 37.320. 2017. Radio measurement collection for Minimization of Drive Tests (MDT). (2017).

[50] Yagmur Yigit, Long D. Nguyen, Mehmet Ozdem, Omer Kemal Kinaci, Trang Hoang, Berk Canberk, and Trung Q. Duong. 2023. Twinport: 5g drone-assisted data collection with digital twin for smart seaports. *Scientific Reports* 13, 1 (2023), 12310.

[51] Najmeh Abiri, Björn Linse, Patrik Edén, and Mattias Ohlsson. 2019. Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems. *Neurocomputing* 365 (2019), 137–146.

[52] Linchao Li, Bowen Du, Yonggang Wang, Lingqiao Qin, and Huachun Tan. 2020. Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model. *Knowledge-Based Systems* 194 (2020), 105592.

[53] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *Proceedings of the International Conference on Machine Learning*. PMLR, 5689–5698.

[54] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* 34 (2021), 24804–24816.

[55] Yagmur Yigit, Ioannis Panitsas, Leandros Maglaras, Leandros Tassiulas, and Berk Canberk. 2024. Cyber-Twin: Digital twin-boosted autonomous attack detection for vehicular Ad-Hoc networks. In *Proceedings of the IEEE International Conference on Communications, ICC 2024, Denver, CO, USA, June 9–13, 2024*. IEEE, 2167–2172.

[56] Yagmur Yigit, Bahadir Bal, Aytac Karameseoglu, Trung Q. Duong, and Berk Canberk. 2022. Digital twin-enabled intelligent DDOS detection mechanism for autonomous core networks. *IEEE Communications Standards Magazine* 6, 3 (2022), 38–44.

[57] Yagmur Yigit, Omer Kemal Kinaci, Trung Q. Duong, and Berk Canberk. 2023. TwinPot: Digital twin-assisted honeypot for cyber-secure smart seaports. In *Proceedings of the 2023 IEEE International Conference on Communications Workshops*. IEEE, 740–745.

[58] Yagmur Yigit, Christos Chrysoulas, Gokhan Yurdakul, Leandros Maglaras, and Berk Canberk. 2023. Digital twin-empowered smart attack detection system for 6g edge of things networks. In *Proceedings of the 2023 IEEE Globecom Workshops*. IEEE, 178–183.

[59] Qinghua Xu, Shaukat Ali, and Tao Yue. 2023. Digital twin-based anomaly detection with curriculum learning in cyber-physical systems. *ACM Transactions on Software Engineering and Methodology* 32, 5 (2023), 1–32.

[60] Burcu Bolat-Akça and Elif Bozkaya-Aras. 2024. Digital twin-assisted intelligent anomaly detection system for Internet of Things. *Ad Hoc Networks* 158 (2024), 103484.

[61] Gastón García González, Pedro Casas, Alicia Fernández, and Gabriel Gómez. 2021. On the usage of generative models for network anomaly detection in multivariate time-series. *ACM SIGMETRICS Performance Evaluation Review* 48, 4 (2021), 49–52.

[62] Lingkun Kong, Jinsong Yu, Diyin Tang, Yue Song, and Danyang Han. 2023. Multivariate time series anomaly detection with generative adversarial networks based on active distortion transformer. *IEEE Sensors Journal* 23, 9 (2023), 9658–9668.

[63] Crispin Almodovar, Fariza Sabrina, Sarvnaz Karimi, and Salahuddin Azad. 2022. Can language models help in system security? Investigating log anomaly detection using BERT. In *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*. 139–147.

[64] Chih-Lin I and Zhiming Zheng. 2023. *The Digital Twin in Human Activities: The Personal Digital Twin*. Springer International Publishing, Cham, 1045–1079.

[65] Radhya Sahal, Saeed H. Alsamhi, and Kenneth N. Brown. 2022. Personal digital twin: A close look into the present and a step towards the future of personalised healthcare industry. *Sensors* 22, 15 (2022), 5918.

[66] Aleksi Kopponen, Antti Hahto, Petri Kettunen, Tommi Mikkonen, Niko Mäkitalo, Jarkko Nurmi, and Matti Rossi. 2022. Empowering citizens with digital twins: A blueprint. *IEEE Internet Computing* 26, 5 (2022), 7–16.

[67] Alex Berke, Ronan Doorley, Kent Larson, and Esteban Moro. 2022. Generating synthetic mobility data for a realistic population with RNNs to improve utility and privacy. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 964–967.

[68] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 World Wide Web Conference*. 1459–1468.

[69] Xinyu Chen, Jiajie Xu, Rui Zhou, Wei Chen, Junhua Fang, and Chengfei Liu. 2021. Trajvae: A variational autoencoder model for trajectory generation. *Neurocomputing* 428 (2021), 332–339.

[70] Kun Ouyang, Reza Shokri, David S. Rosenblum, and Wenzhuo Yang. 2018. A non-parametric generative model for human trajectories. In *Proceedings of the IJCAI*. 3812–3817.

[71] Chu Cao and Mo Li. 2021. Generating mobility trajectories with retained Data Utility. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2610–2620.

[72] Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. 2020. Learning to simulate human mobility. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 3426–3433.

[73] Huandong Wang, Changzheng Gao, Yuchen Wu, Depeng Jin, Lina Yao, and Yong Li. 2023. PateGail: A privacy-preserving mobility trajectory generator with imitation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 14539–14547.

[74] Yuan Yuan, Huandong Wang, Jingtao Ding, Depeng Jin, and Yong Li. 2023. Learning to simulate daily activities via modeling dynamic human needs. In *Proceedings of the ACM Web Conference 2023*. 906–916.

[75] Qingyue Long, Huandong Wang, Tong Li, Lisi Huang, Kun Wang, Qiong Wu, Guangyu Li, Yanping Liang, Li Yu, and Yong Li. 2023. Practical synthetic human trajectories generation based on variational point processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[76] Huandong Wang, Qizhong Zhang, Yuchen Wu, Depeng Jin, Xing Wang, Lin Zhu, and Li Yu. 2023. Synthesizing human trajectories based on variational point processes. *IEEE Transactions on Knowledge and Data Engineering* 36, 4 (2023), 1785–1799.

[77] Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. 2024. Controltraj: Controllable trajectory generation with topology-constrained diffusion model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4676–4687.

[78] Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li. 2023. Spatio-temporal diffusion point processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3173–3184.

[79] Fengli Xu, Yong Li, Depeng Jin, Jianhua Lu, and Chaoming Song. 2021. Emergence of urban growth patterns from human mobility behavior. *Nature Computational Science* 1, 12 (2021), 791–800.

[80] Stefani A. Crabtree, Devin A. White, Corey J. A. Bradshaw, Frédérik Saltré, Alan N. Williams, Robin J. Beaman, Michael I. Bird, and Sean Ulm. 2021. Landscape rules predict optimal superhighways for the first peopling of Sahul. *Nature Human Behaviour* 5, 10 (2021), 1303–1313.

[81] F.E. Cornes, G.A. Frank, and C.O. Dorso. 2019. Fear propagation and the evacuation dynamics. *Simulation Modelling Practice and Theory* 95 (2019), 112–133. DOI: http://dx.doi.org/10.1016/j.simpat.2019.04.012

[82] SaiFei Chen, Hui Fu, Yan Qiao, and NaiQi Wu. 2022. Route choice behavior modeling for emergency evacuation and efficiency analysis based on Type-II fuzzy theory. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2022), 6934–6949. DOI: http://dx.doi.org/10.1109/TITS.2021.3064085

[83] Guan-Yuan Wu and Masayuki Mizuno. 2019. The numerical analysis of mass evacuation in Taipei 101 with control volume model. *Simulation Modelling Practice and Theory* 96 (2019), 101937. DOI: http://dx.doi.org/10.1016/j.simpat.2019.101937

[84] Zhenzhen Yao, Guijuan Zhang, Dianjie Lu, and Hong Liu. 2020. Learning crowd behavior from real data: A residual network method for crowd simulation. *Neurocomputing* 404 (2020), 173–185.

[85] Can Rong, Tong Li, Jie Feng, and Yong Li. 2021. Inferring origin-destination flows from population distribution. *IEEE Transactions on Knowledge and Data Engineering* 35, 1 (2021), 603–613.

[86] Hongzhi Shi, Quanming Yao, and Yong Li. 2023. Learning to simulate crowd trajectories with graph networks. In *Proceedings of the ACM Web Conference 2023*. 4200–4209.

[87] Guozhen Zhang, Zihan Yu, Depeng Jin, and Yong Li. 2022. Physics-infused machine learning for crowd simulation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2439–2449.

[88] Tong Li, Tong Xia, Huandong Wang, Zhen Tu, Sasu Tarkoma, Zhu Han, and Pan Hui. 2022. Smartphone app usage analysis: Datasets, methods, and applications. *IEEE Communications Surveys and Tutorials* 24, 2 (2022), 937–966.

[89] Tong Li, Mingyang Zhang, Hancheng Cao, Yong Li, Sasu Tarkoma, and Pan Hui. 2020. "what apps did you use?": Understanding the long-term evolution of mobile app usage. In *Proceedings of the Web Conference 2020*. 66–76.

[90] Cosimo Distante, Laura Fineo, Luca Mainetti, Luigi Manco, Benito Taccardi, and Roberto Vergallo. 2022. HF-SCA: Hands-free strong customer authentication based on a memory-guided attention mechanisms. *Journal of Risk and Financial Management* 15, 8 (2022), 1–24. DOI: http://dx.doi.org/10.3390/jrfm15080342

[91] Junqi Zhang, Bing Bai, Ye Lin, Jian Liang, Kun Bai, and Fei Wang. 2020. General-purpose user embeddings based on mobile app usage. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2831–2840.

[92] Yacheng He, Qianghuai Jia, Lin Yuan, Ruopeng Li, Yixin Ou, and Ningyu Zhang. 2023. A concept knowledge graph for user next intent prediction at alipay. In *Companion Proceedings of the ACM Web Conference 2023*. 45–48.

[93] Songcheng Gao, Wenzhong Li, Lynda J. Song, Xiao Zhang, Mingkai Lin, and Sanglu Lu. 2020. PersonalitySensing: A multi-view multi-task learning approach for personality detection based on smartphone usage. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2862–2870.

[94] Sha Zhao, Zhiling Luo, Ziwen Jiang, Haiyan Wang, Feng Xu, Shijian Li, Jianwei Yin, and Gang Pan. 2019. AppUsage2Vec: Modeling smartphone app usage for prediction. In *Proceedings of the 2019 IEEE 35th International Conference on Data Engineering*. IEEE, 1322–1333.

[95] Xingyu Chen, Ye Guo, Honglei Xu, Hongyan Yan, and Lin Lin. 2022. User demographic prediction based on the fusion of mobile and survey data. *IEEE Access* 10 (2022), 111507–111527.

[96] Attaullah Buriro, Francesco Ricci, and Bruno Crispo. 2021. SWIPEGAN: Swiping data augmentation using generative adversarial networks for smartphone user authentication. In *Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning*. 85–90.

[97] Yantao Li, Jiaxing Luo, Shaojiang Deng, and Gang Zhou. 2021. CNN-based continuous authentication on smartphones with conditional Wasserstein generative adversarial network. *IEEE Internet of Things Journal* 9, 7 (2021), 5447–5460.

[98] Shiyuan Zhang, Tong Li, Shuodi Hui, Guangyu Li, Yanping Liang, Li Yu, Depeng Jin, and Yong Li. 2023. Deep transfer learning for city-scale cellular traffic generation through urban knowledge graph. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4842–4851.

[99] Kai Xu, Rajkarn Singh, Hakan Bilen, Marco Fiore, Mahesh K. Marina, and Yue Wang. 2022. CartaGenie: Context-driven synthesis of city-scale mobile network traffic snapshots. In *Proceedings of the 2022 IEEE International Conference on Pervasive Computing and Communications*. IEEE, 119–129.

[100] Kai Xu, Rajkarn Singh, Marco Fiore, Mahesh K. Marina, Hakan Bilen, Muhammad Usama, Howard Benn, and Cezary Ziemlicki. 2021. Spectragan: Spectrum based generation of city scale spatiotemporal mobile network traffic data. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*. 243–258.

[101] Chuanhao Sun, Kai Xu, Marco Fiore, Mahesh K. Marina, Yue Wang, and Cezary Ziemlicki. 2022. AppShot: A conditional deep generative model for synthesizing service-level mobile traffic snapshots at city scale. *IEEE Transactions on Network and Service Management* 19, 4 (2022), 4136–4150.

[102] Shuodi Hui, Huandong Wang, Tong Li, Xinghao Yang, Xing Wang, Junlan Feng, Lin Zhu, Chao Deng, Pan Hui, Depeng Jin, et al. 2023. Large-scale urban cellular traffic generation via knowledge-enhanced GANs with multi-periodic patterns. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4195–4206.

[103] Baik Dowoo, Yujin Jung, and Changhee Choi. 2019. PcapGAN: Packet capture file generator by style-based generative adversarial networks. In *Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications*. IEEE, 1149–1154.

[104] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. 2020. Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference*. 464–483.

[105] Pan Wang, Shuhang Li, Feng Ye, Zixuan Wang, and Moxuan Zhang. 2020. PacketCGAN: Exploratory study of class imbalance for encrypted traffic classification using CGAN. In *Proceedings of the ICC 2020-2020 IEEE International Conference on Communications*. IEEE, 1–7.

[106] Xi Jiang, Shinan Liu, Aaron Gember-Jacobson, Arjun Nitin Bhagoji, Paul Schmitt, Francesco Bronzino, and Nick Feamster. 2024. Netdiffusion: Network data augmentation through protocol-constrained traffic generation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 8, 1 (2024), 1–32.

[107] Shiyuan Zhang, Tong Li, Depeng Jin, and Yong Li. 2024. NetDiff: A service-guided hierarchical diffusion model for network flow trace generation. *Proceedings of the ACM on Networking* 2, CoNEXT3 (2024), 1–21.

[108] Leonardo F.R. Ribeiro, Pedro H.P. Saverese, and Daniel R. Figueiredo. 2017. struc2vec: Learning node representations from structural identity. In . ACM. DOI: http://dx.doi.org/10.1145/3097983.3098061

[109] Hongchang Gao, Jian Pei, and Heng Huang. 2019. Progan: Network embedding via proximity generative adversarial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1308–1316.

[110] Yinfeng Ban, Juhua Pu, Yujun Chen, and Yuanhong Wang. 2018. NEGAN: Network embedding based on generative adversarial networks. In *Proceedings of the 2018 International Joint Conference on Neural Networks*. 1–8. DOI: http://dx.doi.org/10.1109/IJCNN.2018.8489481

[111] Dongxiao He, Tao Wang, Lu Zhai, Di Jin, Liang Yang, Yuxiao Huang, Zhiyong Feng, and Philip S. Yu. 2023. Adversarial representation mechanism learning for network embedding. *IEEE Transactions on Knowledge and Data Engineering* 35, 2 (2023), 1200–1213. DOI: http://dx.doi.org/10.1109/TKDE.2021.3103193

[112] Kai Lei, Meng Qin, Bo Bai, Gong Zhang, and Min Yang. 2019. GCN-GAN: A non-linear temporal link prediction model for weighted dynamic networks. In *Proceedings of the IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*. 388–396. DOI: http://dx.doi.org/10.1109/INFOCOM.2019.8737631

[113] Cong Tran, Won-Yong Shin, Andreas Spitz, and Michael Gertz. 2022. DeepNCDeepNC: Deep generative network completion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2022), 1837–1852. DOI: http://dx.doi.org/10.1109/TPAMI.2020.3032286

[114] Haozhe Wang, Yulei Wu, Geyong Min, and Wang Miao. 2022. A graph neural network-based digital twin for network slicing management. *IEEE Transactions on Industrial Informatics* 18, 2 (2022), 1367–1376. DOI: http://dx.doi.org/10.1109/TII.2020.3047843

[115] Han Xiao, Wenqiang Tian, Wendong Liu, and Jia Shen. 2022. ChannelGAN: Deep learning-based channel modeling and generating. *IEEE Wireless Communications Letters* 11, 3 (2022), 650–654.

[116] Yaqi Hu, Mingsheng Yin, William Xia, Sundeep Rangan, and Marco Mezzavilla. 2022. Multi-frequency channel modeling for millimeter wave and THz wireless communication via generative adversarial networks. In *Proceedings of the 2022 56th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 670–676.

[117] Cong Zou, Fang Yang, Jian Song, and Zhu Han. 2023. Generative adversarial network for wireless communication: Principle, application, and trends. *IEEE Communications Magazine* 62, 5 (2023), 58–64.

[118] William Xia, Sundeep Rangan, Marco Mezzavilla, Angel Lozano, Giovanni Geraci, Vasilii Semkin, and Giuseppe Loianno. 2022. Generative neural network channel modeling for millimeter-wave UAV communication. *IEEE Transactions on Wireless Communications* 21, 11 (2022), 9417–9431.

[119] Marius Arvinte and Jonathan I. Tamir. 2022. MIMO channel estimation using score-based generative models. *IEEE Transactions on Wireless Communications* 22, 6 (2022), 3698–3713.

[120] Ushnish Sengupta, Chinkuo Jao, Alberto Bernacchia, Sattar Vakili, and Da-shan Shiu. 2023. Generative diffusion models for radio wireless channel modelling and sampling. In *Proceedings of the GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 4779–4784.

[121] Songyang Zhang, Achintha Wijesinghe, and Zhi Ding. 2023. RME-GAN: A learning framework for radio map estimation based on conditional generative adversarial network. *IEEE Internet of Things Journal* 10, 20 (2023), 18016–18027.

[122] Satya Kumar Vankayala, Swaraj Kumar, Ishaan Roy, D. Thirumulanathan, Seungil Yoon, and Ignatius Samuel Kanakaraj. 2021. Radio map estimation using a generative adversarial network and related business aspects. In *Proceedings of the 2021 24th International Symposium on Wireless Personal Multimedia Communications*. IEEE, 1–6.

[123] Ahmed Marey, Mustafa Bal, Hasan F. Ates, and Bahadir K. Gunturk. 2022. PL-GAN: Path loss prediction using generative adversarial networks. *IEEE Access* 10 (2022), 90474–90480.

[124] Peizheng Li, Xiaoyang Wang, Robert Piechocki, Shipra Kapoor, Angela Doufexi, and Arjun Parekh. 2022. Variational autoencoder assisted neural network likelihood RSRP prediction model. In *Proceedings of the 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, 554–559.

[125] Kehai Qiu, Stefanos Bakirtzis, Ian Wassell, Hui Song, Kan Lin, and Jie Zhang. 2023. IRDM: A generative diffusion model for indoor radio map interpolation. In *Proceedings of the GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 01–06.

[126] Filippo Vannella, Jaeseong Jeong, and Alexandre Proutiere. 2022. Off-policy learning in contextual bandits for remote electrical tilt optimization. *IEEE Transactions on Vehicular Technology* 72, 1 (2022), 546–556.

[127] Yawen Chen, Yu Liu, Ming Zeng, Umber Saleem, Zhaoming Lu, Xiangming Wen, Depeng Jin, Zhu Han, Tao Jiang, and Yong Li. 2020. Reinforcement learning meets wireless networks: A layering perspective. *IEEE Internet of Things Journal* 8, 1 (2020), 85–111.

[128] Amina Mseddi, Wael Jaafar, Halima Elbiaze, and Wessam Ajib. 2019. Intelligent resource allocation in dynamic fog computing environments. In *Proceedings of the 2019 IEEE 8th International Conference on Cloud Networking*. IEEE, 1–7.

[129] Tong Liu, Lun Tang, Weili Wang, Qianbin Chen, and Xiaoping Zeng. 2021. Digital-twin-assisted task offloading based on edge collaboration in the digital twin edge network. *IEEE Internet of Things Journal* 9, 2 (2021), 1427–1444.

[130] Ning Chen, Peiying Zhang, Neeraj Kumar, Ching-Hsien Hsu, Laith Abualigah, and Hailong Zhu. 2022. Spectral graph theory-based virtual network embedding for vehicular fog computing: A deep reinforcement learning architecture. *Knowledge-Based Systems* 257 (2022), 109931.

[131] Filippo Vannella, Alexandre Proutiere, Yassir Jedra, and Jaeseong Jeong. 2022. Learning optimal antenna tilt control policies: A contextual linear bandit approach. In *Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 740–749.

[132] Jinqiang Li, Miao Ye, Linqiang Huang, Xiaofang Deng, Hongbing Qiu, Yong Wang, and Qiuxiang Jiang. 2023. An intelligent SDWN routing algorithm based on network situational awareness and deep reinforcement learning. *IEEE Access* 11 (2023), 83322–83342.

[133] Silvestre Malta, Pedro Pinto, and Manuel Fernández-Veiga. 2023. Using reinforcement learning to reduce energy consumption of ultra-dense networks with 5G use cases requirements. *IEEE Access* 11 (2023), 5417–5428.

[134] Qianyue Hao, Wenzhen Huang, Tao Feng, Jian Yuan, and Yong Li. 2023. GAT-MF: Graph attention mean field for very large scale multi-agent reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 685–697.

[135] Xiaoyuan Fu, Quan Yuan, Zirui Zhuang, Yang Li, Jianxin Liao, and Dongmei Zhao. 2024. TacNet: A tactic-interactive resource allocation method for vehicular networks. *IEEE Internet of Things Journal* 11, 8 (2024), 14370–14382.

[136] Baofeng Ji, Bingyi Dong, Da Li, Yi Wang, Lvxi Yang, Charalampos Tsimenidis, and Varun G. Menon. 2023. Optimization of resource allocation for V2X security communication based on multi-agent reinforcement learning. *IEEE Transactions on Vehicular Technology* (2023). Early access. DOI: 10.1109/TVT.2023.3340424

[137] Armeline D. Mafuta, Bodhaswar TJ Maharaj, and Attahiru S. Alfa. 2023. Decentralized resource allocation-based multiagent deep learning in vehicular network. *IEEE Systems Journal* 17, 1 (2023), 87–98.

[138] Jinsong Gui, Liyan Lin, Xiaoheng Deng, and Lin Cai. 2024. Spectrum-energy-efficient mode selection and resource allocation for heterogeneous V2X networks: A federated multi-agent deep reinforcement learning approach. *IEEE/ACM Transactions on Networking* 32, 3 (2024), 2689–2704.

[139] Vinicius Lima, Mark Eisen, and Alejandro Ribeiro. 2020. Learning constrained resource allocation policies in wireless control systems. In *Proceedings of the 2020 59th IEEE Conference on Decision and Control*. IEEE, 2615–2621.

[140] Wenqi Zhang, Qiang Wang, Xiao Liu, Yuanwei Liu, and Yue Chen. 2020. Three-dimension trajectory design for multi-UAV wireless network with deep reinforcement learning. *IEEE Transactions on Vehicular Technology* 70, 1 (2020), 600–612.

[141] Wenzhen Huang, Tong Li, Yuting Cao, Zhe Lyu, Yanping Liang, Li Yu, Depeng Jin, Junge Zhang, and Yong Li. 2023. Safe-NORA: Safe reinforcement learning-based mobile network resource allocation for diverse user demands. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 885–894.

[142] Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Bo Yuan, Xueqian Wang, and Dacheng Tao. 2022. Penalized proximal policy optimization for safe reinforcement learning. arXiv:2205.11814. Retrieved from https://arxiv.org/abs/2205.11814

[143] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research* (2023), 02783649241273668. Early access. https://doi.org/10.1177/02783649241273668

[144] Hongyang Du, Ruichen Zhang, Yinqiu Liu, Jiacheng Wang, Yijing Lin, Zonghang Li, Dusit Niyato, Jiawen Kang, Zehui Xiong, Shuguang Cui, et al. 2024. Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization. *IEEE Communications Surveys and Tutorials* 26, 4 (2024), 2611–2646.

[145] Hongyang Du, Zonghang Li, Dusit Niyato, Jiawen Kang, Zehui Xiong, Huawei Huang, and Shiwen Mao. 2024. Diffusion-based reinforcement learning for edge-enabled AI-Generated content services. *IEEE Transactions on Mobile Computing* 23, 9 (2024), 8902–8918.

[146] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. 2010. Optimizing walking controllers for uncertain inputs and environments. *ACM Transactions on Graphics* 29, 4 (2010), 1–8.

[147] Igor Mordatch, Kendall Lowrey, and Emanuel Todorov. 2015. Ensemble-cio: Full-body dynamic motion planning that transfers to physical humanoids. In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 5307–5314.

[148] Jonah Siekmann, Kevin Green, John Warila, Alan Fern, and Jonathan Hurst. 2021. Blind bipedal stair traversal via sim-to-real reinforcement learning. *Robotics: Science and Systems* XVII (2021), 1–9.

[149] Kristinn Kristinsson and Guy Albert Dumont. 1992. System identification and control using genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics* 22, 5 (1992), 1033–1046.

[150] Maxime Gautier and Wisama Khalil. 1988. On the identification of the inertial parameters of robots. In *Proceedings of the 27th IEEE Conference on Decision and Control*. IEEE Piscataway, NJ, 2264–2269.

[151] Pradeep K. Khosla and Takeo Kanade. 1985. Parameter identification of robot dynamics. In *Proceedings of the 1985 24th IEEE Conference on Decision and Control*. IEEE, 1754–1760.

[152] Fabio Ramos, Rafael Carvalhaes Possas, and Dieter Fox. 2019. Bayessim: Adaptive domain randomization via probabilistic inference for robotics simulators. *Robotics: Science and Systems* XV (2019), 1–10.

[153] Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. 2019. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *Proceedings of the 2019 International Conference on Robotics and Automation*. IEEE, 8973–8979.

[154] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. 2021. RMA: Rapid motor adaptation for legged robots. *Robotics: Science and Systems* XVII (2021), 1–15.

[155] Ben Evans, Abitha Thankaraj, and Lerrel Pinto. 2022. Context is everything: Implicit identification for dynamics adaptation. In *Proceedings of the 2022 International Conference on Robotics and Automation*. IEEE, 2642–2648.

[156] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1842–1850.

[157] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1126–1135.

[158] Jane X. Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. 2016. Learning to reinforcement learn. arXiv:1611.05763. Retrieved from https://arxiv.org/abs/1611.05763

[159] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. 2018. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *Proceedings of the International Conference on Learning Representations*.

[160] Paul Christiano, Zain Shah, Igor Mordatch, Jonas Schneider, Trevor Blackwell, Joshua Tobin, Pieter Abbeel, and Wojciech Zaremba. 2016. Transfer from simulation to real world through learning deep inverse dynamics model. arXiv:1610.03518. Retrieved from https://arxiv.org/abs/1610.03518

[161] Josiah Hanna and Peter Stone. 2017. Grounded action transformation for robot learning in simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[162] Andrei A. Rusu, Matej Večerík, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. 2017. Sim-to-real robot learning from pixels with progressive nets. In *Proceedings of the Conference on Robot Learning*. PMLR, 262–270.

[163] Feng Wang and Jiangchuan Liu. 2010. Networked wireless sensor data collection: Issues, challenges, and approaches. *IEEE Communications Surveys and Tutorials* 13, 4 (2010), 673–687.

[164] Tong Li, Sylvia T. Kouyoumdjieva, Gunnar Karlsson, and Pan Hui. 2019. Data collection and node counting by opportunistic communication. In *Proceedings of the 2019 IFIP Networking Conference*. IEEE, 1–9.

[165] Liang Zhao, Zijia Zhao, Enchao Zhang, Ammar Hawbani, Ahmed Y Al-Dubai, Zhiyuan Tan, and Amir Hussain. 2023. A digital twin-assisted intelligent partial offloading approach for vehicular edge computing. *IEEE Journal on Selected Areas in Communications* 41, 11 (2023), 3386–3400.

[166] Tao Zhang, Changqiao Xu, Yibo Lian, Haijiang Tian, Jiawen Kang, Xiaohui Kuang, and Dusit Niyato. 2023. When moving target defense meets attack prediction in digital twins: A convolutional and hierarchical reinforcement learning approach. *IEEE Journal on Selected Areas in Communications* 41, 10 (2023), 3293–3305.

[167] Michael G. Kapteyn, Jacob VR Pretorius, and Karen E. Willcox. 2021. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science* 1, 5 (2021), 337–347.

[168] Omar Hashash, Christina Chaccour, and Walid Saad. 2022. Edge continual learning for dynamic digital twins over wireless networks. In *Proceedings of the 2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication*. IEEE, 1–5.

[169] Fenyu Jiang, Tong Li, Xingzai Lv, Hua Rui, and Depeng Jin. 2024. Physics-informed neural networks for path loss estimation by solving electromagnetic integral equations. *IEEE Transactions on Wireless Communications* 23, 10 (2024), 15380–15393.

[170] Duo Wu, Xianda Wang, Yaqi Qiao, Zhi Wang, Junchen Jiang, Shuguang Cui, and Fangxin Wang. 2024. NetLLM: Adapting large language models for networking. In *Proceedings of the ACM SIGCOMM 2024 Conference*. 661–678.

[171] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI: http://dx.doi.org/10.18653/v1/N19-1423

[172] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. https://openai.com/research/language-unsupervised

[173] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[174] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv:2302.13971. Retrieved from https://arxiv.org/abs/2302.13971

[175] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM:

Outperforming curated corpora with web data only. In *Proceedings of the 37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. Retrieved from https://openreview.net/forum?id=kM5eGcdCzq
[176] Yudong Huang, Minrui Xu, Xinyuan Zhang, Dusit Niyato, Zehui Xiong, Shuo Wang, and Tao Huang. 2023. AI-generated network design: A diffusion model-based learning approach. *IEEE Network* 38, 3 (2023), 202–209.
[177] M. Sousa, P. Vieira, MP Queluz, and A. Rodrigues. 2024. Enhancing robustness for automated mobile network optimization by uncovering causal relationships. In *Proceedings of the 2024 19th International Symposium on Wireless Communication Systems*. IEEE, 1–6.
[178] Baisakhi Chatterjee, Swapnil Mitra, Rohan Laha, Dattatreyo Roy, Preetarpan Banerjee, Himadri Nath Saha, and Phillip G. Bradford. 2019. Fault diagnosis in vehicular networks using do-calculus. In *Proceedings of the 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference*. IEEE, 1043–1050.
[179] Christo Kurisummoottil Thomas and Walid Saad. 2023. Reliable beamforming at terahertz bands: Are causal representations the way forward?. In *Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1–5.
[180] Walid Saad, Omar Hashash, Christo Kurisummoottil Thomas, Christina Chaccour, Merouane Debbah, Narayan Mandayam, and Zhu Han. 2024. Artificial general intelligence (AGI)-native wireless systems: A journey beyond 6G. arXiv:2405.02336. Retrieved from https://arxiv.org/abs/2405.02336
[181] He Li, Lu Yu, and Wu He. 2019. The Impact of GDPR on Global Technology Development. *Journal of Global Information Technology Management* 22, 1 (2019), 1–6.
[182] Zhihao Wang, Dingde Jiang, and Kaiwen Zhang. 2024. GP-INT: Generating balanced network-wide in-band telemetry path for digital twin networks. In *Proceedings of the ICC 2024-IEEE International Conference on Communications*. IEEE, 1109–1114.
[183] Yuntao Wang, Zhou Su, Ning Zhang, Rui Xing, Dongxiao Liu, Tom H. Luan, and Xuemin Shen. 2022. A survey on metaverse: Fundamentals, security, and privacy. *IEEE Communications Surveys and Tutorials* 25, 1 (2022), 319–352.
[184] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. arXiv:1312.6114. Retrieved from https://arxiv.org/abs/1312.6114
[185] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM* 63, 11 (2020), 139–144.
[186] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.
[187] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.
[188] Tribhuvanesh Orekondy, Arash Behboodi, and Joseph B. Soriaga. 2022. Mimo-gan: Generative mimo channel modeling. In *Proceedings of the ICC 2022-IEEE International Conference on Communications*. IEEE, 5322–5328.

A Appendix

A.1 NDTs, Digital Twin Networks, and Metaverse

NDTs and Digital Twin Networks refer to distinct concepts within digital twin technology. NDTs are virtual replicas of physical communication networks designed to simulate, monitor, and optimize network performance in real time. In contrast, digital twin networks [44, 182] are specifically designed to interconnect multiple digital twins across various domains, such as transportation, healthcare, and manufacturing. This infrastructure supports real-time data exchange, interoperability, and coordinated operations, facilitating integrated industry decision-making. While NDTs focus on optimizing specific communication networks, digital twin networks create the framework for linking and synchronizing multiple digital twins, enabling cross-domain collaboration and functionality.

NDTs can significantly enhance digital twin networks by providing reliable, real-time simulations and optimized network performance–both crucial for interconnected virtual environments. With real-time monitoring and predictive maintenance capabilities, they help sustain seamless interactions for large numbers of users who depend on stable network functionality. Within digital twin networks, NDTs facilitate integrating and synchronizing multiple digital twins across diverse domains, such as smart cities and industrial IoT ecosystems. Simulating complex network interactions ensures efficient, secure, and reliable data exchange between interconnected digital twins. Additionally, NDTs can leverage generative AI to dynamically adapt to evolving digital twin network demands, supporting scalable, flexible, and optimized infrastructures that can expand to accommodate new applications and domains as they arise.

The metaverse [183] is also a virtual, interactive space where users engage in immersive experiences such as gaming, socializing, and learning through VR and AR technologies. Unlike NDTs, which operate behind the scenes to maintain high-quality network performance, the metaverse offers user-centric virtual worlds. NDTs can enhance the metaverse to deliver low-latency, high-bandwidth connectivity for applications like virtual reality, live events, and interactive simulations. Generative AI further amplifies this capability by enabling rapid, adaptive data generation, allowing digital twins to scale efficiently and reflect changing network conditions in real-time. Additionally, AI-driven digital twins facilitate continuous learning, enabling metaverse environments to evolve alongside their physical counterparts. Despite challenges like data privacy and computational demands, integrating NDTs with the metaverse holds significant promise for creating seamless, intelligent, and interactive virtual worlds.

A.2 VAE, GAN, Diffusion, and AR

Variational Autoencoder (VAE) [184]: A VAE comprises two core components: an encoder and a decoder. The encoder transforms input data into latent variables, capturing the mean and standard deviation of these variables. The decoder then utilizes the latent variables to generate new data. The system is trained to minimize the difference between the original and generated data. VAE operates on an unsupervised learning basis and generates data by sampling from its learned probability distribution.

Generative Adversarial Networks (GANs) [185]: A GAN consists of two principal components: a generator and a discriminator. The generator generates data that resembles the original, utilizing random values drawn from a normal distribution. Its objective is to generate realistic data that can deceive the discriminator. Conversely, the discriminator's function is to distinguish between real data and the synthetic data produced by the generator. During the training process, a continuous competitive interaction occurs between these two models. The interplay between these two models enhances their individual capabilities and contributes to the overall effectiveness and reliability of the GAN framework in replicating realistic data.

Diffusion Models [186]: Diffusion models consist of forward and reverse diffusion processes. In the forward diffusion phase, they utilize a Markov chain process to progressively add noise to the input data, transforming it into white noise that aligns with the standard normal distribution. The reverse diffusion process methodically removes the noise to reconstruct the original data. The diffusion model, resembling a specialized multi-layer VAE (Hierarchical VAE) network, features high-dimensional hidden variables. These high-dimensional variables, reflecting the distributions of the original data, are learned using a fixed procedure, facilitating the construction of the desired data sample from noise.

Autoregressive Models (AR) [187]: **Autoregressive** (**AR**) models work by representing the joint probability distribution of a data vector in high-dimensional space as a product of a series of conditional probability distributions. In simpler terms, these models generate data points sequentially, starting with the first component and using the previously generated components to determine the next. This process continues until all components of the sample are generated. Notably, most **large language models** (**LLMs**) use the AR framework as their backbone. Each word in LLMs is generated based on the preceding words, resulting in coherent and contextually accurate text generation.

A.3 Literature Review Tables

Table 3. Literature Review in Data Processing and Network Monitoring with Generative AI

Topic	Ref.	Model Name	Main Modules	Key Design
Network Data Collection	[47, 48]	MR	Network interface	State information from network elements is gathered through multiple network interfaces
	[49]	MDT	Device probe, SDK	Data is collected directly from user devices
	[50]	Twinport	Drone sensor	Use drones to capture environmental data
Network Data Imputation	[51]	SDAi	Denoising autoencoder	Process various data types using a stacked denoising autoencoder for computational efficiency.
	[52]	MMDL	Autoencoder	Use two parallel stacked autoencoders simultaneously considering spatial and temporal dependencies
	[53]	GAIN	GAN, MLP	Generator imputes the missing components and discriminator determines which components were imputed
	[54]	CSDI	Diffusion, Transformer	Missing time series generated to be conditioned on observed data
Network Data Anomaly Detection	[56]	-	YANG, AutoFS	Use an online learning approach to update the model instantly to improve detection accuracy
	[59]	LATTICE	Curriculum learning	Use a curriculum learning method to address discrepancies between historical and real-time data in DT systems
	[61]	Net-GAN, Net-VAE	GAN, VAE, RNN	Discover the underlying distribution of normal operating state and identify potential anomalies without the need for labeled data
	[62]	ADT-GAN	GAN, Transformer	Use the prior knowledge of time sequences' overall associations to recognize anomalies through sequence associations and reconstruction errors
	[63]	LogFiT	BERT	Leverage a pretrained BERT-based language model, fine-tuning it to comprehend the linguistic structure of network system logs

Table 4. Literature Review in Modeling Mobile User Behavior with Generative AI

Topic	Ref.	Model Name	Main Modules	Key Design
Individual Mobility	[67]	TS-RNN	RNN, AR	Consider population distribution by taking <home, work> pairs as input
	[68]	DeepMove	AR, LSTM	Leverage periodic patterns of mobility
	[69]	TrajGAN	GAN, LSTM	Use DeepWalk to pre-train location embedding
	[70]	OuyangGAN	GAN, CNN	Discretize locations into a two-dimensional matrix
	[71]	TrajGen	GAN, Seq2Seq	Separate spatial and temporal information
	[72]	MoveSim	GAN, Attention	Introduce the prior knowledge of urban structure to generate a meaningful trajectory
	[73]	PateGail	GAN, Imitation learning	Model individual movement as a human decision-making process
	[74]	SAND	GAN, Policy function	Integrate Maslow's need theory into human decision-making process
	[75]	VOLUNTEER	VAE, Transformer	Model user distributions
	[76]	TrajSynVAE	VAE, LSTM	Use temporal point process to model continuous temporal distribution
	[77]	ControlTraj	Diffusion Model, U-Net	Integrate Wide and Deep networks to construct conditional information
	[78]	DSTPP	Diffusion Model, Attention	Capture the interdependence of trajectory time and space
Crowd Mobility	[79]	Collective Mobility Model (CMM)	Statistical model	Highlight the fractal-like urban morphologies and scaling laws in city growth patterns
	[80]	Migration Dynamics	Physical model	Consider different mechanisms in migration dynamics
	[81]	SIRS	Physical model	Incorporate panic stress into the desired velocity
	[82]	PanicRoute	Physical model	Model irrational routing decisions
	[83]	Volume control	Physical model	Treat evacuees as fluid particles in fluid flow
	[84]	ResNet-SICS	Residual network	Scene-independent crowd simulation
	[85]	GSTE-DF	GNN	Capture dynamic spatial-temporal features
	[86]	GNS	GNN	Use a heterogeneous graph to model interactions among people and environment
	[87]	PIML	GNN	Combine physics and neural network models through iterative learning

Table 5. Literature Review in Modeling Network Services with Generative AI

Topic	Ref.	Model	Key Design
Mobile App usage	[90]	U-net	Extract latent features from application transactions
	[91]	VAE and Transformer	Generate application sequences using installation, uninstallation and snapshot information
	[92]	Codec and Knowledge graph	Enrich user embedding with timestamp, location and society association for next intention prediction
	[96]	GAN and Knowledge graph	Generate high-quality synthetic swiping modality samples for authentication
	[97]	WGAN and CNN	Utilize WGAN to generate raw motion data
Network traffic generation	[100]	GAN and CNN	Extract traffic features in both time and frequency domains to generate network data.
	[101]	GAN and CNN	Construct city-scale network traffic flow using mobile traffic and spatial context information.
	[102]	GAN and LSTM	Extract multi-periodic traffic patterns and uses urban knowledge graphs to capture urban structure.
	[98]	GAN and GNN	Align base station representations between target and source cities through urban knowledge graphs.
	[103]	GAN	Generate network packet traces trained on packet-level data.
	[104]	Hierarchical GAN	Generate both packet attributes and feature series concurrently.
	[105]	Conditional GAN	Tackle the issue of data imbalance in datasets containing various traffic types using conditional GANs.
	[106]	Diffusion Model	Transform time-series data into 2D images to capture temporal correlations, using pretrained models to generate traffic traces.
	[107]	Transformer-based Diffusion Model	Incorporate users' app usage intent as a condition to model complex relationships between network flow features.
Network topology embedding	[109]	GAN	Extracte latent similarity of nodes in attributed network for link prediction and clustering
	[110]	GAN and GCN	Construct node latent vector via self and neighbour feature in network embedding
	[111]	Hierarchical GAN	Add competitor during latent extraction process to realize precise network embedding
	[112]	WGAN and LSTM	Utilize GCN to incorporate topology latent feature into network embedding for link prediction
	[113]	Autoregressive	Extract node latent features to implement network topology under missing nodes
	[114]	GAN and GCN	Network embedding for end-to-end latency estimation in network slicing.

Table 6. Literature Review of Wireless Environment Simulation

Topic	Ref.	Method	Data	Scenario	Frequency	Bandwidth
Channel modeling/ estimation	[188]	GAN	Simulation, 3GPP TDL, MATLAB	-	4GHz	30.72MHz
	[115]	GAN	Simulation, CDL-C	-	3.5GHz	10MHz
	[116]	GAN	Simulation, Raytracing, Wireless InSite	Urabn region	28GHz, 140GHz	-
	[118]	VAE	Simulation, Raytracing, Wireless InSite	UAV to ground	28GHz	-
	[119]	Diffusion model	Simulation CDL family	-	40GHz	-
	[120]	Diffusion model	Simulation 3GPP TR 38.901	Urban Macro Urban Micro	2GHz	-
Radio map estimation	[121, 122]	GAN	Simulation, Raytracing, WinProp	Urban region	5.9GHz	10MHz
	[123]	GAN	Simulation, Raytracing, Wireless InSite	Urban region	900MHz	-
	[124]	VAE	Real-world dataset	Urban region	-	-
	[125]	Diffusion model	Simulation, Raytracing, Ranplan Professional	Indoor	2.4GHz	-

Table 7. Literature Review in Designing and Training Reinforcement Learning for NDTs

Category	Ref.	Algorithm	Scenario	State	Action	Reward	Agent
SARL	[129]	DDQN	Data offloading	CSI, connection state, channel gain	Resource allocation action	Power consumption and network latency	Network coordinator
	[130]	AC	Data offloading	Environment features	Node and link embedding	LAR, ACCR	Network coordinator
	[131]	CL-MAB	Coverage optimization	Network conditions	Electrical antenna tilt	Coverage and capacity	Base station
	[132]	PPO	Rate control	Location state and network state	Next-hop adjacent node selected	Bandwidth, delay, packet error rate, packet loss	SDN
	[133]	SARSA	Energy management	Sleep mode level, packet buffer load	Whether to be awake	QoS and packet buffer load	Network controller
MARL	[135]	Tactic-interactive MARL	Resource allocation	Channel conditions, interference levels, age of information	Transmission power, sub-band allocation	QoS	Vehicle
	[136]	MARL-DDQN	Resource allocation	Channel state, interferenc	Power levels, spectrum resources	Spectrum efficiency	Vehicle
	[137]	MA-DDQN	Power allocation	Channel state, interference, transmission load	Transmission power, sub-band allocation	V2I link capacity, V2V throughput	V2V link
	[138]	Federated multi-agent DRL	Resource allocation	Channel state, available resources	Frequency band selection, power allocation	QoS	Vehicle
Safety RL	[139]	Primal-dual policy gradient	Resource Allocation	Fading and control state	Power control	Keep operating around a point	Base station
	[140]	DQN	Trajectory design of multi-UAV	Location information and ID	Direction	Downlink capacity	UAV
	[141]	P3O	Resource allocation, power control	Location information, traffic demands	Resource usage, power	Satisfaction ratio	Mobile users
Diffusion for RL	[144]	Diffusion model	Power allocation	Channel state	Power control	Sum rate	Mobile users
	[34]	AI-generated power allocation	Power allocation	Channel state, number of objects in semantic communication	Power	Transmission costs	Objects
	[145]	AGOD	Service provider selection	Availability of resource	Resource usage	User experience	Mobile users

Table 8. Literature Review in Sim2Real Transition and Network Control

Method	Ref.	Application	Simulator	Real Platform	Policy Optimization Method
Domain Randomization	[147]	Dynamic motion planning	MuJoCo	Darwin-OP humanoid	Levenberg-Marquardt
	[148]	Blind bipedal locomotion	MuJoCo	bipedal robot Cassie	PPO
System Identification	[153]	Robotic arm control	Flex	Yumi robot	PPO
	[154]	Legged Locomotion	RaiSim	A1 robot	PPO
	[155]	Robot Sliding	MuJoCo	UR10 robot	CEM
Meta RL	[159]	Locomotion	MuJoCo	legged millirobot	MPC
Transfer Learning	[160]	Arm Swings	MuJoCo	Physical Fetch robot	TRPO
	[161]	Bipedal robot walking	Gazebo, SimSpark	Softbank NAO Robot	ES
	[162]	Robotic arm control	MuJoCo	Jaco	A3C

Received 5 January 2024; revised 31 October 2024; accepted 24 December 2024