When Digital Twin Meets Generative AI: Intelligent Closed-Loop Network Management

Xinyu Huang, Haojun Yang, Conghao Zhou, Mingcheng He, Xuemin Shen, and Weihua Zhuang

Abstract

Generative artificial intelligence (GAI) and digital twin (DT) are advanced data processing and virtualization technologies to revolutionize communication networks. Thanks to the powerful data processing capabilities of GAI, integrating it into DT is a potential approach to construct an intelligent holistic virtualized network for better network management performance. To this end, we propose a GAI-driven DT (GDT) network architecture to enable intelligent closed-loop network management. In the architecture, various GAI models can empower DT status emulation, feature abstraction, and network decision-making. The interaction between GAI-based and model-based data processing can facilitate intelligent external and internal closed-loop network management. To further enhance network management performance, three potential approaches are proposed, i.e., model light-weighting, adaptive model selection, and data-model-driven network management. We present a case study pertaining to data-model-driven network management for the GDT network, followed by some open research issues.

Introduction

0ver the past decade, digital twin (DT) technology has emerged as a key virtualization technique, experiencing significant growth and development. DT was first introduced to monitor and mitigate anomalous events for flying vehicles via accurately simulating their entire lifecycles [1]. The mobile communication network usually consists of the core network and the radio access network, which are mainly responsible for data routing and wireless transmission, respectively. By applying DT in mobile communication networks, holistic network virtualization for efficient network management can be realized [2], [3]. Specifically, DT consists of three modules to support efficient network management, i.e., 1) network status emulation module to mirror the status of physical mobile communication networks through advanced prediction-based algorithms, which can reduce frequent data collection cost; 2) data feature abstraction module to distill the useful patterns of network traffic and user behaviors, which can simplify network management problems; 3) network decision-making module to make tailored network management strategies, which can be validated in the virtualized environment. Since the above modules require efficient and accurate data processing in the intricate and dynamic network environment, the utilization of advanced data processing techniques is imperative.

As an emerging branch of artificial intelligence (AI), generative AI (GAI) focuses on generating new data instances, analyzing data correlation, and addressing optimization problems [4], which can help improve network status emulation, data feature abstraction, and network decision-making in DT. For instance, generative adversarial network (GAN) can generate high-fidelity network scenario images and expand datasets through its generator and discriminator [5], which can empower DT emulation module. Generative transformer (GT) performs excellently in text understanding and generation due to its attention mechanism [6], which can assist DT in perceiving user intentions and analyzing data correlation. Furthermore, generative diffusion model (GDM) can facilitate conditional generation and decision-making by its state diffusion and inverse dynamics mechanism [7], thereby improving DT network decision-making performance. By integrating GAI and GT into mobile communication networks, an intelligent GAI-driven DT (GDT) network architecture for external and internal closed-loop network management can be realized. Specifically, the external loop emphasizes the interaction between the physical network and DT, where the data quality of generated DT status via GAI is evaluated by an error discriminator in DT for adaptive data collection frequency. The internal loop focuses on the interaction between DT and GAI in the GDT, where the abstracted features are separately fed into the model-based network management module in DT and the GAI-based network management module to evaluate which one can achieve best network performance for adaptive network management policy adjustment.

However, achieving intelligent external and internal closed-loop network management for the GDT network architecture poses technical challenges, including

- *Massive Caching and Computing Overhead:* Since GAI models usually have large model sizes and complex neural network structures, it is challenging to directly deploy GAI models on network edge nodes with limited caching and computing capabilities.
- *Model Scalability and Efficiency:* For the GAI models with different sizes, the larger GAI models usually provide better data

Digital Object Identifier: 10.1109/MNET.2024.3524474
Date of Current Version: 16 September 2025
Date of Publication: 30 December 2024

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada. Xinyu Huang is corresponding author.

processing performance but also consume more network resources. Therefore, how to select an appropriate model size with enough efficiency to adapt to network dynamics is challenging.

- *Reliable Data Processing for Network Robustness:* Since the complexity and "black box" nature of GAI models may lead to unpredictable and biased status emulation, feature abstraction, and network decision-making in DT, it is challenging to design a reliable data processing mechanism to improve network robustness.

In this article, we propose a novel GDT network architecture to realize intelligent external and internal closed-loop network management. Four kinds of typical GAI models, i.e., GAN, GT, variational autoencoder (VAE), and GDM, are selected to assist DT status emulation, feature abstraction, and network decision-making. The intelligent closed-loop network management includes two aspects. Specifically, the interaction between GAI-based status emulation and DT-based error discriminator can facilitate adaptive external closed-loop data collection. The interplay between GAI-based and model-based decision-making can enable adaptive internal closed-loop network decision-making. To address the aforementioned challenges, we first propose a model light-weighting method to reduce model caching and computing overhead. Secondly, we develop an adaptive model selection mechanism to adapt to network dynamics. Thirdly, we design a data-model-driven method to improve network management robustness. A case study pertaining to data-model-driven network management for GDT is presented, followed by a discussion on potential research issues.

The remainder of this article is organized as follows. Firstly, DT and advanced GAI techniques are discussed, followed by the proposed GDT network architecture. Then, we discuss the challenges for GDT network and some potential solutions. Next, a case study about data-model-driven network management for GDT network is presented. Finally, the open research issues are identified, followed by the conclusion.

Generative AI-Driven Digital Twin Network Architecture

In this section, we first introduce DT and advanced GAI techniques, and then propose a GDT network architecture.

Digital Twin

1) Definition: DT is a virtual representation, also termed "black box", of the physical mobile communication network, that reflects the real-time network status and provides efficient management strategies through a variety of embedded data-based models. These models can enable high-fidelity network status emulation, accurate network feature abstraction, and tailored network management strategies to facilitate efficient closed-loop network management.

2) Composition and Functionality: DT mainly consists of three modules, i.e., status emulation module, data feature abstraction module, and network decision-making module.

- Status emulation module is the basis of DT, which is utilized to characterize the real-time status of physical mobile communication network. The input of status emulation module is the collected data from the physical network through access points (APs) and sensors, which can be classified into networking-related data and behavior-related data. The status emulation module relies on the prediction-based algorithms, such as long short-term memory (LSTM) and recurrent neural network (RNN), to predict future network status. The output of status emulation module is the emulated network status, which can provide holistic network information for network management.
- Data feature abstraction module is responsible for distilling useful information from network status. The input of data feature abstraction module includes two aspects, i.e., emulated and realistic network status. The advanced data processing algorithms in DT are responsible for abstracting data features. For instance, the autoencoder can compress the high-dimensional data into a low-dimensional latent representation. The output of data feature abstraction module is distilled network information, such as spatiotemporal traffic distribution and swipe probability distribution, which can help the network controller capture the network dynamics.
- Network decision-making module is responsible for outputting tailored network management decisions. The input of network decision-making module includes network status and abstracted data features. To provide efficient network decisions, the data-based methods in DT can outperform traditional model-based methods. For instance, policy gradient methods can optimize network management policies in an online and incremental way, which can adapt to network dynamics. The output of network decision-making module is the network management decision, which is transferred to the network controller for implementation in the mobile communication network.

3) Benefits: By introducing DT into mobile communication networks, there are following benefits. Firstly, the status emulation module in DT can emulate the real-time status of physical entities to reduce the traffic load. Secondly, the data abstraction module in DT can provide correct and distilled information to the network controller. Thirdly, due to the diversity of users' service demands, such as some users prioritizing service latency and others focusing on the quality of transmitted content, DT can analyze their differences in service requirements to make tailored resource management.

Generative AI

GAI refers to a subset of AI technologies that can autonomously generate novel content, data, and solutions. GAI has developed rapidly due to advancements in machine learning algorithms and the increasing availability of large datasets for training. A key characteristic of GAI is

Model	Functions	Characteristics	Storage Requirement	Computing Requirement	Metrics
GAN	• Image generation • Dataset augmentation	• Generator • Discriminator	3 ~ 25 GB	0.2 ~ 1 PFlops	• Peak signal-to-noise ratio • Structural similarity
GT	• Text understanding & generation • Feature extraction	• Positional encoding • Self & multi-head attention • Residual connection	10 ~ 30 GB	0.5 ~ 2 PFlops	• General language understanding evaluation • Perplexity
VAE	• Feature learning • Denoising	• Probabilistic graphic model • Variational inference	20 ~ 800 MB	1 ~ 100 TFlops	• Mean squared error • Kullback-Leibler divergence
GDM	• Conditional generation • Decision-making	• Reverse diffusion • Markovian transition	0.3 ~ 3 GB	2 ~ 200 TFlops	• Temporal coherence • Interpolation quality

TABLE 1. Part of Emerging GAI Models Applied to Mobile Communication Networks.

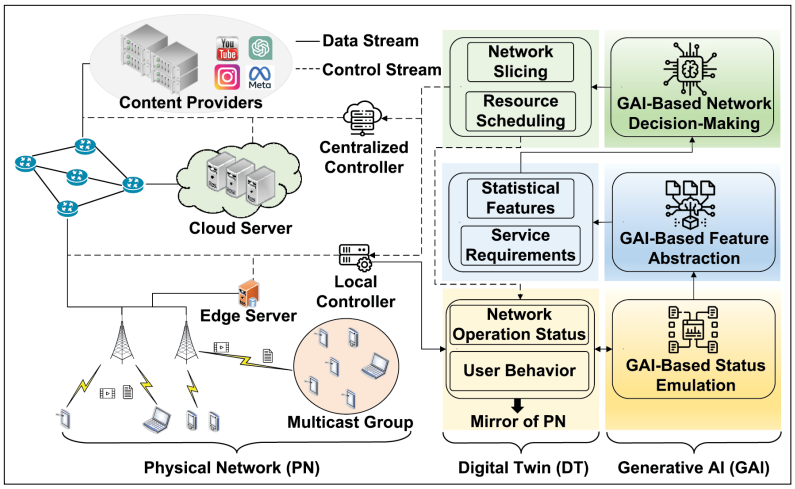

FIGURE 1. The GDT network architecture.

its ability to learn intrinsic characteristics in the data, which can generate new data that closely mimic the original data distribution [6]. Therefore, integrating GAI into mobile communication networks can generate high-fidelity network status and user behaviors, accurate data features, and optimized network configurations [4]. As shown in Table 1, we select four representative GAI models that can be well applied to mobile communication networks for performance enhancement.

- Generative Adversarial Network (GAN): It consists of a dueling generator and discriminator that refine generated data in a unique adversarial process [5]. By generating high-quality network scenario images and expanding datasets, GAN can improve the efficiency of virtual scene construction and network status emulation in DT.
- Generative Transformer (GT): It excels in text understanding and generation as well as feature abstraction through advanced positional encoding, self and multi-head attention mechanisms, and residual connections among neural networks [6]. For instance, a novel multi-modal mutual attention-based sentiment analysis framework was proposed to process complicated contexts and mine the association between unique semantics and common semantics [8]. GT can assist DT in accurately perceiving user intentions and analyzing data correlations, thereby providing tailored network management information.
- Variational Autoencoder (VAE): It is pivotal for feature abstraction and data recovery in mobile communication networks through probabilistic graphic models and variational inference [9]. Specifically, it can compress high-dimensional networking data into a low-dimensional latent representation and reconstruct data based on the captured data distribution.
- Generative Diffusion Model (GDM): It facilitates conditional generation and network decision-making through reverse diffusion and Markovian transition processes [7]. By modeling the network management process as a return-conditional diffusion model, the training efficiency and inference performance can be enhanced.

By leveraging powerful status generation, data analytics, and decision-making capabilities, GAI can effectively enhance DT capabilities.

Generative AI-Driven DT Network Architecture

To seamlessly integrate DT and GAI in mobile communication networks, as shown in Fig. 1, we develop a GDT network architecture, which can effectively improve network management performance. The physical network includes real-world network infrastructures, such as cloud and edge servers for data caching and computing, as well as APs for unicast, multicast, and broadcast transmissions. The virtual network consists of DT and GAI, where they highly collaborate in the network status emulation, data feature abstraction, and network decision-making to improve network management performance. DTs are deployed in a distributed manner at edge and cloud servers to process time-sensitive and time-insensitive tasks. The local controller is responsible for small-timescale data collection and resource scheduling policy implementation, while the centralized controller implements the large-timescale network slicing, i.e., the partitioning of a single physical network infrastructure into multiple and isolated logical networks [10]. The specific module interaction in the GDT part consists of GAI-based status emulation, feature abstraction, and decision-making, as shown in Fig. 2.

1) GAI-Based Status Emulation: The boxes (1) and (2) in Fig. 2 show the status emulation process in GDT part. Specifically, DT status consists of networking-related data (users' channel conditions and service delay, base stations' transmission capabilities

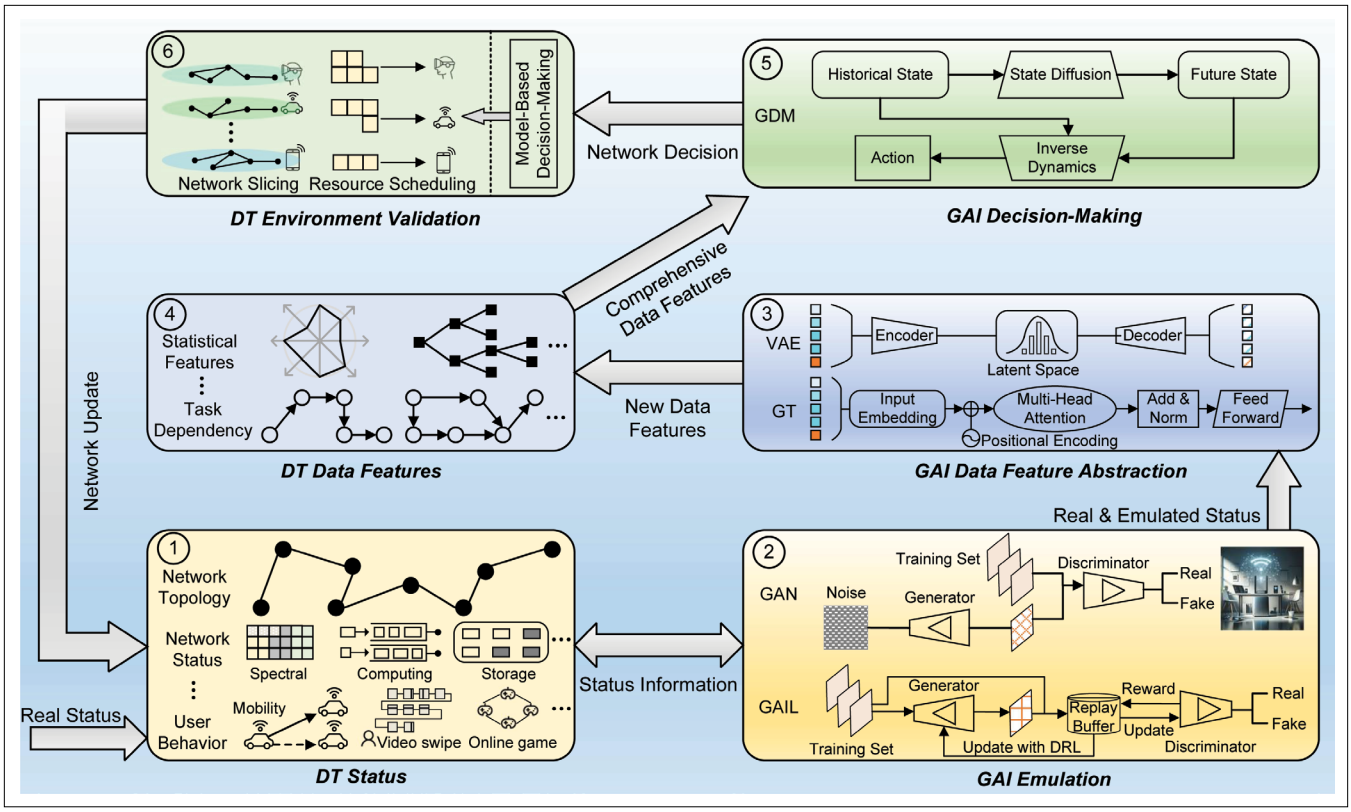

FIGURE 2. The specific module interaction in GDT part.

and traffic load, cloud and edge servers' caching and computing workload, etc.) and behavior-related data (users' interaction frequency, locations, mobility speed, preferences, etc.), to mirror the real-time physical network status. DT data needs to be periodically updated to guarantee freshness. A dual error-based data collection mechanism can utilize the mean squared error (MSE) and entropy to measure the emulated accuracy for adaptive data collection frequency adjustment. To reduce the frequent data collection overhead, advanced GAI techniques are utilized to generate high-fidelity DT status. For instance, GAN can help generate realistic network scenes for network environment construction and update by collaboratively training its generator and discriminator. Furthermore, generative adversarial imitation learning (GAIL) integrates the status generation capability of GAN into deep reinforcement learning (DRL) algorithms, which can accurately emulate users' dynamic behaviors [11]. Through real-time interaction between DT status and GAI emulation, high-fidelity DT status can be effectively supplemented.

2) GAI-Based Feature Abstraction: Based on the realistic and emulated DT status, GAI-based feature abstraction is conducted in the boxes (3) and (4) in Fig. 2. Specifically, since DT status data is usually high-dimensional and time-series, such as varying channel conditions and users' locations in a high-density network, it is hard to directly use them to guide efficient network management. Therefore, accurate and efficient data feature abstraction is necessary. For instance, VAE can encode the high-dimensional DT status data into a low-dimensional representation to capture the network traffic patterns and user behavioral patterns. Furthermore, GT can accurately analyze spatiotemporal correlations from DT status through its advanced attention mechanism. The abstracted features can simplify the network management problem and provide tailored network management information.

3) GAI-Based Decision-Making: The new and previous DT data features are integrated to generate comprehensive inputs for GAI-based network decision-making, as shown in boxes (5) and (6) in Fig. 2. Since the network management problem is usually complex and non-convex, it is difficult to directly use model-based or data-based methods to solve the problem to obtain a near-optimal solution. Therefore, advanced decision-making algorithms are necessary. For instance, GDM utilizes the state diffusion method to generate the future state (network status) based on the learned data distribution. The generated state is integrated with the historical state as an input to the inverse dynamics mechanism for action (network management decision) generation. This can achieve a better convergence performance compared with DRL algorithms [12]. The generated network management decision is fed back to the DT environment for validation and DT status update.

4) Procedure of Closed-Loop Network Management: As shown in Fig. 3, intelligent external and internal closed-loop network management is realized in the GDT network. *In the external closed loop, an adaptive data collection frequency mechanism is realized*. Specifically, user status is first uploaded to AP to update DT status with a prescribed data collection frequency. Then, the updated DT status is sampled to a mini-batch that is transferred to GAI to conduct status emulation. Next, the error discriminator in DT evaluates the quality of newly generated status that is utilized to adjust data collection frequency. Finally, the

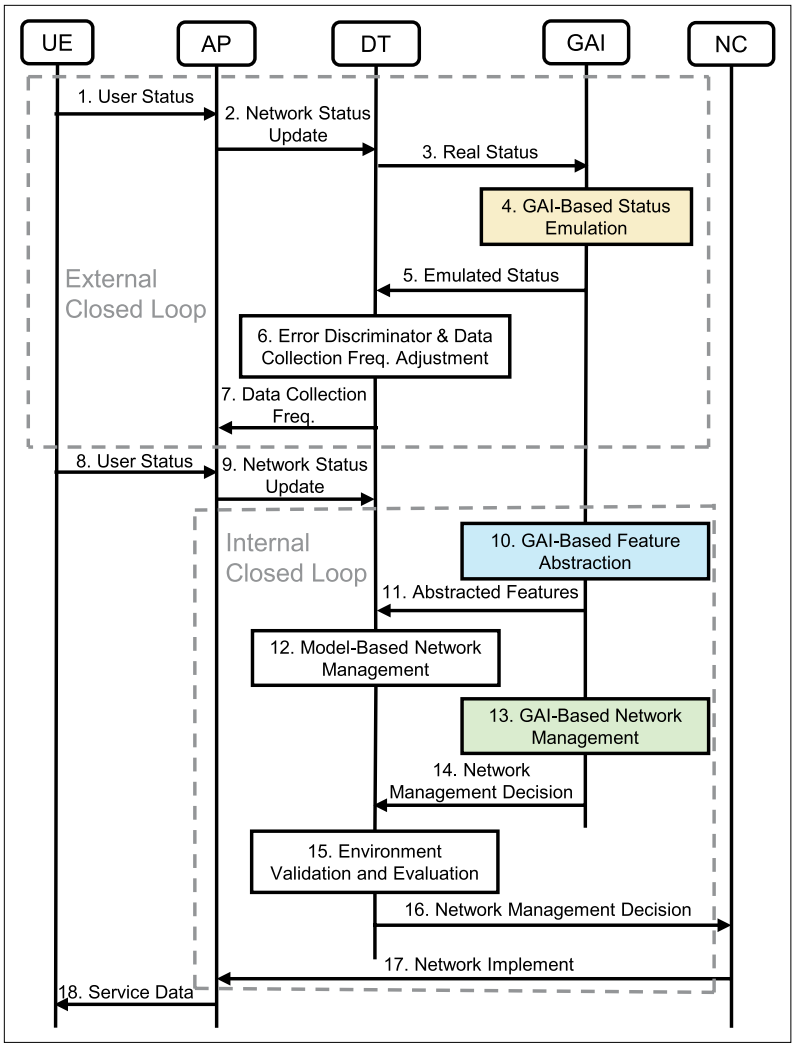

FIGURE 3. Procedure of external and internal closed-loop network management.

adjusted data collection frequency is fed back to the AP to implement a new round of data collection. *In the internal closed loop, an adaptive network decision-making mechanism is realized.* Specifically, the realistic and emulated DT status is first transferred to GAI to abstract data features.

Then, the abstracted features are input to the model-based network management module in DT and the GAI-based network management module, respectively. Next, the generated network management decisions from two modules are validated in DT environment to evaluate which one can achieve better network performance. Finally, the better network management decision is fed back to the network controller for practical implementation. Through the intelligent external and internal closed-loop network management, network performance can be effectively enhanced.

5) Assumptions and Limitations: The proposed GDT network architecture is mainly designed for complex network management problems in dynamic and large-scale networks, but in relatively static or small-scale networks, lightweight AI models or mathematical models could be more efficient with less computational overhead. Moreover, how to access high-quality training data and sufficient edge computational resources for model fine-tuning is challenging.

Challenges and Solutions

To realize the proposed intelligent closed-loop network management in the GDT network architecture, optimizing GAI model deployment, model selection, and data processing in the GDT network is necessary, with some key research challenges.

Challenges

1) Massive Caching and Computing Overhead: From the caching standpoint, the large GAI model size poses a challenge for effective model storage and retrieval. Traditional caching strategies may not be sufficient especially in the network edge nodes, as the voluminous model parameters ranging from millions to billions exceed conventional caching capacities. Therefore, it is hard to directly deploy GAI models on network edge nodes with limited caching capabilities. Moreover, the large sizes and complex neural network structures of GAI models demand substantial computing resources, which leads to increased latency and higher energy consumption. Therefore, it is challenging to improve the computing efficiency of GAI models for computation-intensive data processing.

2) Model Scalability and Efficiency: GAI models usually have various sizes, with larger models typically offering better data processing performance for network management but also requiring more caching and computing resources, whereas smaller models have the opposite characteristics. To efficiently utilize a GAI model under varying network dynamics, an adaptive model selection mechanism is necessary, which can achieve a balance between model scalability and efficiency. For instance, a network might experience varying levels of congestion throughout the day, which affects service latency. A small GAI model that performs well during low-traffic periods might struggle when the network is congested, resulting in inconsistent service quality and degraded user experience. Therefore, how to adaptively select an appropriate GAI model size with enough efficiency for network management presents a significant challenge.

3) Reliable Data Processing for Network Robustness: The utilization of GAI models in DT may introduce instability due to several inherent drawbacks. The complexity and "black box" nature of GAI models can lead to unpredictable and biased decisions. Specifically, if GAI models fail to accurately grasp the operational mechanisms of physical networks, or cannot conduct thorough inferences under conditions of limited network resources, then the emulated network status and abstracted features may be biased. Moreover, the network decision-making process is usually a complex optimization problem, which requires professional theoretics to transform the problem for the contraction of the feasible solution set. Solely relying on the data training without adding any professional optimization theoretics may lead to a sub-optimal solution. These factors combined underscore the challenge of reliable data processing for robust network management.

Solutions

1) Model Light-Weighting: To handle the challenge of massive caching and computing

overhead, model split and knowledge distillation are effective methods to deploy lightweight GAI models. Specifically, model split involves dividing a large GAI model into smaller and more manageable segments for distributed model caching and computing in the mobile communication networks [13]. Through efficient model split, massive data computing can be processed in parallel, which can effectively reduce model inference latency and satisfy caching constraints. Knowledge distillation offers a complementary approach by transferring the knowledge from a large cumbersome teacher model to a small student model without significant performance loss [14]. Since the distilled student model is significantly less resource-intensive, it can be flexibly deployed within mobile communication networks with low caching and computing overhead. Based on the above analysis, the massive caching and computing overhead can be effectively mitigated through model split and knowledge distillation.

2) Adaptive Model Selection: To deal with the challenge of GAI model scalability and efficiency, an adaptive GAI model selection mechanism is necessary. Specifically, the GAI model selection mechanism needs to quantitatively assess the impact of network status, such as bandwidth availability and computing queue congestion, as well as GAI model characteristics including caching and computing demands. By employing a machine learning-based classification method, the GAI model selection mechanism can select the most suitable GAI model size for any given network scenario based on historical performance data, such as quality of service (QoS) and quality of experience (QoE). The GAI model selection mechanism also includes a feedback loop mechanism, where real-time performance data are used to continuously refine the classification algorithms and ensure that GAI models remain accurate and effective in the face of evolving network dynamics and service diversity. By integrating these elements, an adaptive GAI model selection mechanism can be developed.

3) Data-Model-Driven Network Management: To handle the challenge of reliable data processing for network robustness, data-model-driven methods are effective due to the holistic integration of empirical data and theoretical modeling. Unlike data-based methods that rely solely on machine learning algorithms to process historical data, possibly overlooking underlying physical network principles, data-model-driven methods incorporate the principles through mathematical modeling, which can ensure a more comprehensive understanding of network dynamics and user behavior patterns. Specifically, data-based methods usually excel in identifying state transition probability to solve a part of decoupled subproblems, while model-based methods rely on classical optimization techniques to obtain the optimal solution to the remaining decoupled subproblems. Since overfitting could be an essential challenge for GAI models, particularly when the training datasets lack diversity in operational scenarios, training processes can incorporate techniques such as data augmentation, regularization, and cross-validation.

Case Study: Data-Model-Driven Network Management for GDT Network

In this section, a case study is provided on data-model-driven network management for GDT network, aimed at improving QoE.

Considered Scenario

A GDT-assisted multicast short video streaming scenario is considered, which consists of two APs, multiple multicast groups (MGs), and one GDT. In each scheduling slot, bandwidth and computing resources are allocated to each sub-MG (SMG) to receive videos with adaptive bitrate with the objective of maximizing QoE. The GDT consists of a status emulation module, a data feature abstraction module, and a network decision-making module, i.e.,

- In the GDT status emulation module, we first generate users' trajectories within the University of Waterloo campus with the Levy flight model that refers to a random Markovian walk and the probability distribution of step lengths satisfies heavy-tailed distribution. The generated data is used as the label data, where eighty percent of label data is selected as the training sample to train the LSTM model for users' trajectory emulation. The well-trained LSTM model is used to emulate users' trajectories. Based on users' real-time trajectories, the real-time channel conditions are emulated based on PropagationModel at Matlab by analyzing the channel fading between users and base stations. Users' swipe behaviors and preferences are sampled from the real-world video swipe dataset. Through the GDT status emulation module, users' real-time status can be perceived on the network side.
- In the GDT data feature abstraction module, we propose an improved GAI method to update multicast groups and abstract users' swipe probability distribution. Specifically, the improved GAI method consists of three parts, i.e., autoencoder, double deep Q-network (DDQN), and K-means++. The autoencoder is selected as the basic GAI model that can reduce user status dimension by analyzing temporal correlation. Double deep Q-network is further used to mine the compressed user status to find an appropriate clustering number, while K-means++ is responsible for a fast and accurate multicast group update based on a given clustering number and compressed user status. Through the GDT data feature abstraction module, users' intrinsic features can be obtained for tailored network decision-making.
- In the GDT network decision-making module, a data-model-driven method is proposed to realize multicast segment buffering and resource scheduling. Specifically, in the multicast segment buffering, we first utilize the abstracted swipe probability distribution from the GDT data feature abstraction module to analyze the watching probability of segments. Based on the analyzed watching probability and estimated multicast transmission capabilities, the multicast segment

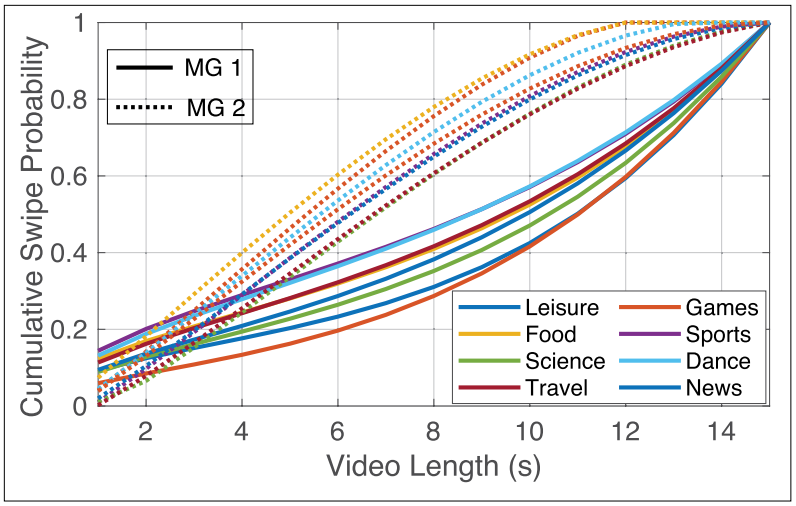

FIGURE 4. Cumulative swipe probability abstracted by the improved autoencoder.

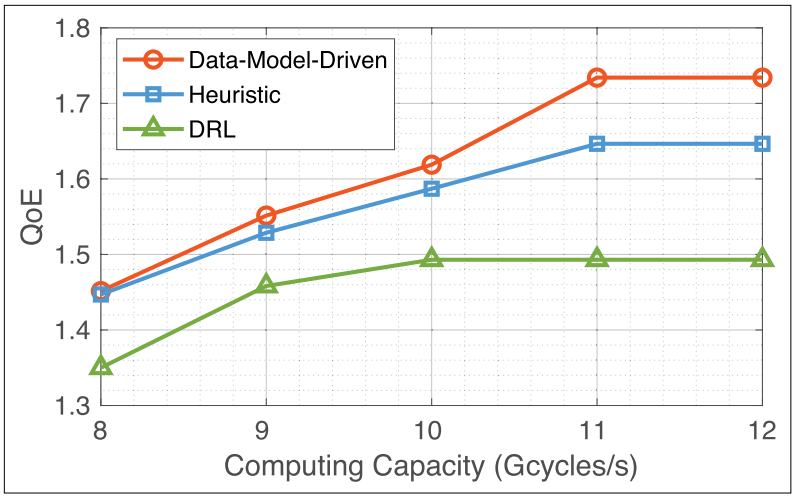

FIGURE 5. QoE versus different computing capacities.

buffering number and order are determined. Then, the branching dueling Q-network is used to determine the segment version selection for each SMG. In the resource scheduling, the sequential least squares quadratic programming method is utilized to determine the occupation time of network resources by each SMG. The joint multicast segment buffering and resource scheduling decision is fed back to the DT environment to obtain the system reward, i.e., QoE, for Q-network update. The detailed simulation setting of network decision-making module can be found in [15].

Simulation Results

As illustrated in Fig. 4, we present the cumulative swipe probability abstracted via GAI, where different line styles and colors represent various MGs and video types, respectively. It can be observed that users in MG 1 have similar cumulative swipe probability as those in MG 2 during the initial phase, but a noticeable divergence ensues over time. Since DT status is high-dimensional and time-series, directly using traditional data processing methods hardly abstracts accurate swipe probability distribution. Therefore, we select GAI to process DT status data, which can effectively differentiate users' swipe behaviors for efficient resource management.

For performance comparison, we select two different schemes to substitute the proposed GDT network decision-making module. Specifically, the heuristic scheme consists of the same GDT-assisted segment buffering strategy and different greedy resource scheduling, while the DRL scheme includes the sequential segment buffering strategy and deep deterministic policy gradient-based resource scheduling. Fig. 5 shows the QoE trend with the increasing computing capacity. It can be observed that the proposed data-model-driven method can always maintain the highest QoE, which reflects its ability to leverage additional computing resources to improve user experience. This is because the proposed data-model-driven method in the GDT decision-making module can effectively abstract segment buffering information for tailored resource scheduling to improve QoE. Furthermore, the heuristic scheme can always achieve better QoE than DRL scheme, because GDT-assisted segment buffering can efficiently abstract buffering information and the greedy resource scheduling can adapt well to different computing capacities.

Open Research Issues

Efficient GDT Module Collaboration

Given the multi-modular composition of GDT and the inherent interactivity among these modules, the efficient collaborative mechanism is critical for network performance enhancement. For instance, GAI-based feature abstraction module usually requires real-time and accurate network status from GAI-based status emulation module. However, the data interaction frequency between the modules and data abstraction level in the GAI-based feature abstraction module can be adaptive and differentiated based on the network dynamics and service requirements. For instance, delay-tolerant tasks with low network dynamics usually require less data interaction between the modules and shallower feature extraction. Deploying GAI models in dynamic network environments is challenging due to potential mismatches between the historical data used for pre-training and real-time data. To address this, online fine-tuning and domain adaptation techniques can be used to continuously adapt GAI models to evolving network conditions.

Specialized Generative Model

Due to the substantial parameterization inherent in GAI models, the associated caching and computing overhead can significantly exacerbate network load. For delay-sensitive and high-reliability tasks, lightweight and specialized models on network edge nodes are necessary. The approach also requires the fine-tuning of GAI models in response to network dynamics and service diversity. To further enhance the responsiveness and adaptability of GAI models, collaborative computing frameworks can be adopted, where edge nodes and cloud infrastructures share computational tasks based on resource availability and task urgency.

Efficient workload balancing strategies, such as distributed task scheduling and dynamic resource allocation, can minimize processing delay and optimize resource utilization.

Efficient Resource Management for GDT Operation

Although GDT can efficiently process networking data to improve network management performance, the maintenance of GDT is equally important, which consumes substantial communication, caching, and computing resource. This is attributed to voluminous networking and user data storage, as well as resource-intensive GAI models for network status emulation, feature abstraction, and network decision-making. Since the limited network resources need to cater to both users' service requests and GDT operation, it is imperative to accurately quantify the impact of GDT operation on QoS and QoE. Based on the quantified result, the network resource scheduling for both GDT operation and users' service requests can be more tailored to further improve QoS and QoE. By integrating real-time user feedback loops and GAI-based predictive analytics, GDT can dynamically optimize network resource management to meet diverse user requirements.

Conclusion

We have proposed a GDT network architecture to achieve intelligent external and internal closed-loop network management. Specifically, advanced GAI models are employed to improve DT status emulation, feature abstraction, and network decision-making modules. In the external closed loop, GAI-based status emulation interacts with DT-based error discriminator to adaptively adjust data collection frequency. In the internal closed loop, GAI-based network decision-making algorithm collaboratives with model-based one in DT to realize adaptive network management. To further optimize GDT network architecture, we have proposed a model light-weighting method, an adaptive model selection mechanism, and a data-model-driven method, respectively. A case study has been presented, and some open research issues have been discussed for accelerating the pace of GDT network development.

References

[1] E. Glaessgen and D. Stargel, "The digital twin paradigm for future NASA and U.S. air force vehicles," in *Proc. Prof. 53rd AIAA Struct. Struct. Dyn. Mater. Conf.*, Honolulu, HI, USA, Apr. 2012, pp. 1–14.

[2] L. Hui et al., "Digital twin for networking: A data-driven performance modeling perspective," *IEEE Netw.*, vol. 37, no. 3, pp. 202–209, May 2023.

[3] X. Shen et al., "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 1–30, 1st Quart., 2022.

[4] M. Xu et al., "Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 1127–1170, 2nd Quart., 2024.

[5] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[6] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 6840–6851.

[8] L. He et al., "Multimodal mutual attention-based sentiment analysis framework adapted to complicated contexts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7131–7143, Dec. 2023.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014, pp. 1–14.

[10] X. Huang et al., "Digital twin-driven network architecture for video streaming," *IEEE Netw.*, vol. 38, no. 6, p. 334, Nov. 2024, Art. no. 341.

[11] R. Bhattacharyya et al., "Modeling human driving behavior through generative adversarial imitation learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 2874–2887, Mar. 2023.

[12] A. Ajay et al., "Is conditional generative modeling all you need for decision-making?" in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2023, pp. 1–24.

[13] W. Wu et al., "Split learning over wireless networks: Parallel design and resource management," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1051–1066, Apr. 2023.

[14] W. Park et al., "Relational knowledge distillation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3967–3976.

[15] X. Huang et al., "Digital twin-based network management for better QoE in multicast short video streaming," *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, pp. 16187–16202, Nov. 2024.

Biographies

Xinyu Huang (Student Member, IEEE) (x357huan@uwaterloo.ca) received the B.E. degree from Xidian University, Xi'an, China, in 2018, and the M.S. degree from Xi'an Jiaotong University, Xi'an, in 2021. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Waterloo, Waterloo, ON, Canada. His research interests include digital twins, generative AI, and network resource management.

Haojun Yang (Member, IEEE) (haojun.yang@uwaterloo.ca) received the B.S. degree in communication engineering and the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014 and 2020, respectively. His research interests include ultra-reliable and low-latency communications, resource management, and vehicular networks.

Conghao Zhou (Member, IEEE) (c89zhou@uwaterloo.ca) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2022. His research interests include space-air-ground integrated networks, network slicing, and machine learning for wireless networks.

Mingcheng He (Student Member, IEEE) (m64he@uwaterloo.ca) received the B.Sc. and M.Eng. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2017 and 2020, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2024. His research interests include network slicing in satellite-terrestrial integrated networks and artificial intelligence for future wireless networks.

Xuemin (Sherman) Shen (Fellow, IEEE) (sshen@uwaterloo.ca) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is an University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, the Internet of Things, AI for networks, and vehicular networks. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Member, an International Fellow of the Engineering Academy of Japan, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society.

Weihua Zhuang (Fellow, IEEE) (wzhuang@uwaterloo.ca) received the B.Sc. and M.Sc. degrees in electrical engineering from Dalian Maritime University, China, and the Ph.D. degree from the University of New Brunswick, Canada. Since 1993, she has been a Faculty Member with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, where she is currently a University Professor and a Tier I Canada Research Chair in wireless communication networks. Her current research focuses on network architecture, algorithms and protocols, and service provisioning in future communication systems. She was a recipient of the Women's Distinguished Career Award from IEEE Vehicular Technology Society in 2021, the R. A. Fessenden Award from IEEE Canada in 2021, the Award of Merit from the Federation of Chinese Canadian Professionals (Ontario), and the Technical Recognition Award in Ad Hoc and Sensor Networks from IEEE Communications Society in 2017.