Edge Intelligence Empowered Metaverse: Architecture, Technologies, and Open Issues

Yanan Xu, Daquan Feng, Mingxiong Zhao, Yao Sun, Xiang-Gen Xia

***Abstract*—Recently, the metaverse has emerged as a focal point of widespread interest, capturing attention across various domains. However, the construction of a pluralistic, realistic, and shared digital world is still in its infancy. Due to the ultra-strict requirements in security, intelligence, and real-time, it is urgent to solve the technical challenges existed in building metaverse ecosystems, such as ensuring the provision of seamless communication and reliable computing services in the face of a dynamic and time-varying complex network environment. In terms of digital infrastructure, edge computing (EC), as a distributed computing paradigm, has the potential to guarantee computing power, bandwidth, and storage. Meanwhile, artificial intelligence (AI) is regarded as a powerful tool to provide technical support for automated and efficient decision-making for metaverse devices. In this context, this paper focuses on integrating EC and AI to facilitate the development of the metaverse, namely, the edge intelligence-empowered metaverse. Specifically, we first outline the metaverse architecture and driving technologies and discuss EC as a key component of the digital infrastructure for metaverse realization. Then, we elaborate on two mainstream classifications of edge intelligence in metaverse scenarios, including AI for edge and AI on edge. Finally, we identify some open issues.**

INTRODUCTION

The origin of the metaverse can be traced back to the 1992 science fiction novel *Snow Disaster* which depicts a virtual world where people interact with others through their virtual avatars. Although there is no unified definition of the metaverse, it can be generally understood that the metaverse is a digital virtual world with a complete economic and social system that maps to (even surpasses) the real world and interacts with it through the integration of various emerging technologies. Over the next two decades, the metaverse is expected to move from unattainable fiction to reality due to the increasing demand for 'digital contact' in socialization, work, and lifestyles, as well as the rapid growth of emerging technologies such as blockchain, artificial intelligence (AI), and the 5th/6th generation of mobile networks (5G/6G).

With the advances of the Internet of Things (IoT) and mobile Internet, the Internet of Everything (IoE) has become a typical application scenario for future information technology. In this context, there are still fundamental technical challenges to building a hyperrealistic metaverse ecology with strict demands on ultra-reliability and ultra-low latency [1]. For example, the high resolution and refresh rate characteristics of metaverse platforms, together with the new interaction mode of IoE, are bound to bring an explosive amount of data, which urgently requires a powerful digital infrastructure that can provide highly scalable computation capacity, reliable communication, large-scale data processing, and storage capabilities. In addition, for digital environment construction, delay-sensitive tasks such as motion capture, real-time rendering, and multivariate data processing need massive computing support to ensure high speed, low latency, and a smooth experience. It is extremely difficult for computation- and energy-constrained mobile devices, such as augmented reality (AR) and virtual reality (VR), to efficiently process such computationally intensive tasks, while under the traditional cloud computing (CC) paradigm, these tasks can be performed on remote data centers (DCs) with powerful computing power and sufficient resources to reduce execution latency, and devices only need to wait for the execution results from the cloud [2]. However, with the exponential growth of mobile communication, mobile services have placed tremendous strain on the backbone network. Relying solely on the centralized paradigm would result in high network delay and fail to meet timely response, thereby significantly degrading the user experience. To cope with this challenge, edge computing (EC) is regarded as a promising distributed computing paradigm that can provide both highly scalable computing power and low network delay [3].

Meanwhile, in order to accelerate the development of the metaverse, automated and intelligent production, efficient decision-making, and accurate execution are the key indicators, which also provide opportunities for applying AI in metaverse construction [4]. Since the debut of AlphaGo in 2016, the era of AI has been ushered in. AI technology has

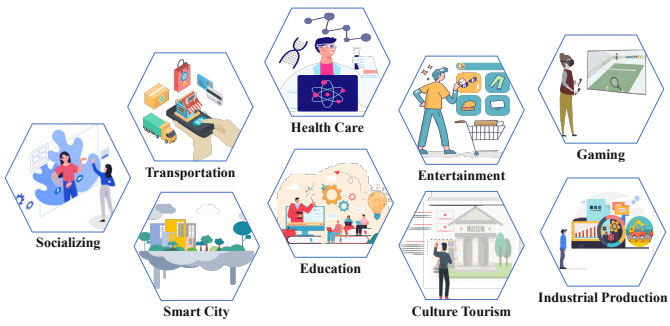

Fig. 1: The application scenarios in the metaverse cover services for individual consumers, specific user groups, and enterprises, such as entertainment, education, industrial production, culture tourism, and so on.

Yanan Xu and Daquan Feng are with Shenzhen University, China.
Mingxiong Zhao is with Yunnan University, China.
Yao Sun is with the University of Glasgow, UK.
Xiang-Gen Xia is with University of Delaware, USA.
Daquan Feng is the corresponding author for this article.

been deeply embedded in a variety of business scenarios, both at the technical level (e.g., computer vision and intelligent speech processing) and the application level (e.g., smart cities and games), especially in decision-making, digital content generation, and rendering, to fully demonstrate its superiority. For instance, reinforcement learning (RL) can be applied to optimization problems in EC scenarios (e.g., efficient offloading decisions, dynamic resource allocation, and caching policies), as well as to improve the performance of synchronization between physical devices and the corresponding digital models in the metaverse (e.g., determining the sampling rate and prediction horizon) [5].

Given the features of massively complex data production on metaverse platforms and the ability of AI to learn from numerous data and adapt well to changing environments, AI is recognized as a viable technical support for metaverse construction, while EC can provide scalable computing capacity and resource support for metaverse applications. In this context, many researchers have focused on how to accelerate the development of the metaverse from the perspective of the fusion of EC and AI (i.e., edge intelligence) and have achieved preliminary research progress. However, the metaverse is still in its infancy and there are still some challenges in metaverse-oriented edge intelligence, such as the applicability of application scenarios, the establishment of accurate evaluation metrics, and privacy security issues. This motivates us to explore how edge intelligence can drive the metaverse ecosystem.

- We introduce the metaverse architecture driven by advanced technologies and discuss the digital infrastructure, such as CC and EC, needed to enable a hyper-immersive metaverse platform.
- We focus on two mainstream classifications (i.e., *AI for edge* and *AI on edge*) of edge intelligence, demonstrate the specific description for each category, and summarize the preliminary progress of research on the edge intelligence-empowered metaverse.
- We discuss the open issues in metaverse-oriented edge intelligence.

The remainder of the paper is organized as follows. First, we provide an overview of the metaverse architecture and the driving techniques. Second, we explore the current research progress in the fusion of EC and AI for building metaverse platforms. Finally, we present open issues.

The Architecture of the Metaverse

The metaverse is a shared, real-time, realistic, and plural digital world that is paralleled to the physical world. With the support of cutting-edge technologies, the development of the metaverse can be expanded to many fields such as industry, culture, education, and military, where Fig. 1 presents typical application scenarios of the metaverse.

- ***The metaverse architecture and backbone technologies***

Building an immersive metaverse ecosystem necessitates a scalable underlying infrastructure for high-performance support, an efficient production platform, and solutions to the technical problems of digital objects and environment content in production, transmission, and interaction. A typical metaverse architecture mainly consists of infrastructure layer, digital engine layer, content generation layer, perceptual interaction layer, and application layer. Fig. 2 shows an illustration of a five-layer metaverse architecture, that is, from bottom to top:

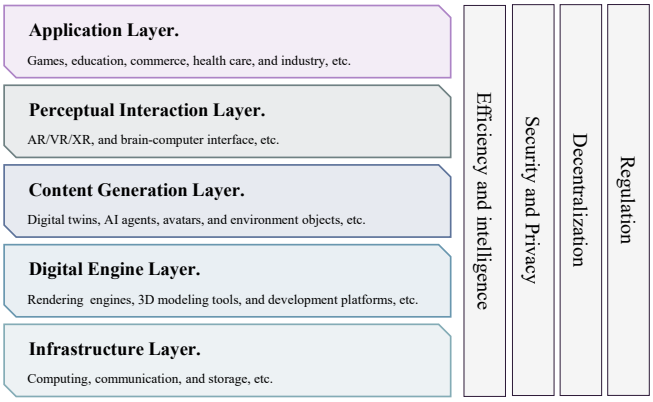

Fig. 2: The metaverse is a large-scale and open platform built by a variety of parties, which requires efficient and intelligent construction of each technical layer, a scientific and efficient digital rule system, and a reliable economic system to assure the sustainable development of the metaverse.

1) The infrastructure layer is the foundation for building the metaverse ecosystem, which mainly provides networking, storage, and computing support.
2) The digital engine layer involves application engines that speed up the development of metaverse applications, such as development platforms and rendering engines.
3) The content generation layer refers to simulating the physical world or innovating virtual environments that do not exist in reality to construct the basic architecture of the metaverse via digital generation technologies.
4) The perceptual interaction layer includes human-computer interaction (HCI) devices that allow people to freely access the virtual world, interact effectively, and obtain an immersive user experience.
5) The application layer provides the ultra-immersive experience of digital life and production.

As stated, the construction of the metaverse technical layers and the creation of a digital ecosystem comparable to the real world cannot be separated from the collaboration of diverse backbone technologies to enable the rapid, intelligent, safe, and reliable development of the metaverse. Fig. 3 shows a general metaverse platform, where the enabling techniques are summarized below:

1) Blockchain has inspired a decentralized creator economy [6] and emerged as one of the most promising technologies to realize the vision of a metaverse of interoperability between the virtual and real worlds.
2) Digital generation covers enabling techniques including digital twins, 3D modeling, computer vision, and real-time rendering to create large-scale, complex immersive

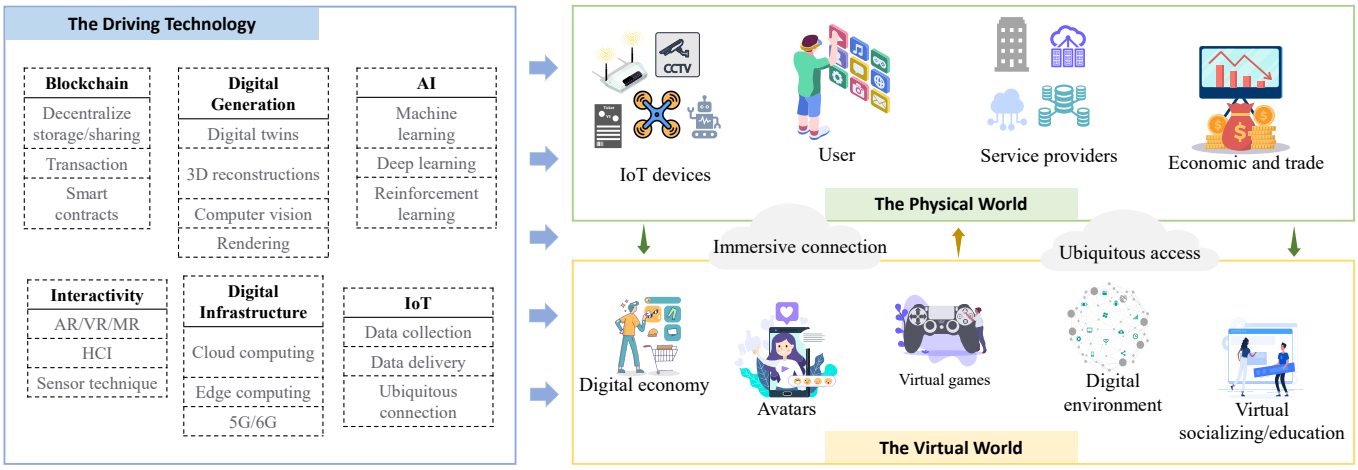

Fig. 3: A general metaverse platform for virtual and real interaction that is driven by leading-edge technologies, such as AI to enable intelligence and cloud/edge computing to provide network environments and IT services.

scenes with high accuracy, including 3D reconstructions of enormous digital objects (e.g., avatars) and environments.

3) Interactivity is a bridge between the physical world and digital space. The presentation of the metaverse mainly relies on VR/AR devices to enable an immersive 3D experience. Particularly, extended reality (XR) deeply integrates AR, VR, and mixed reality (MR) techniques, which has the potential to become the primary access devices. Moreover, the brain-computer interface directly controls external devices by collecting and analyzing brain signals, which is of benefit in enhancing the efficiency of human interaction with the outside world, but it is still in the early stages of research.
4) IoT ensures that everything in the metaverse becomes a part of the network, allowing for human-machine-things interconnection and interaction, including collecting massive amounts of data via various IoT devices (e.g., sensors) and sharing the data over the internet in real time.
5) AI has permeated every aspect of metaverse applications and production. The mainstream AI techniques include machine learning (ML), deep learning (DL), and RL, which leverage their strengths to provide support in terms of reliability, efficiency, and performance in digital content generation, rendering, interaction, and other areas. For example, ML-based methods are utilized for complex data analysis and processing; DL can help speed up large-scale scene rendering; and RL algorithms can handle challenging decision-making problems.
6) Digital infrastructure delivers highly scalable computing and storage services as well as stable and reliable communication guarantees for metaverse applications, such as leveraging distributed computing paradigms (e.g., EC) to apply ultra-low latency and smooth experiences to the metaverse.

- ***The digital infrastructure***

Regarding the digital infrastructure, EC is regarded as a promising distributed computing paradigm that provides both scalable computing power and low network delay, since the resulting high latency under traditional centralized computing may significantly degrade the user experience.

By flexibly deploying edge servers (ESs) and sinking CC resources to the network edge closer to end-users, EC can better provide a distributed computing environment and cloud-like services for key applications, such as AI-related tasks and data analysis and processing from nearby resource-constrained terminals. Therefore, network transmission and the possibility of network congestion on the network core can be reduced, and security concerns in data transmission can be efficiently alleviated. For example, the device can exploit edge-side capabilities to offload expensive foreground rendering or strong interaction tasks to ESs, while ESs can also cache hot content in advance for fast response to requests within its service scope. NVIDIA CloudXR can scale to the edge to deliver immersive and responsive XR experiences via NVIDIA RTX GPU-powered ESs, thereby extending graphics-intensive applications to mobile terminals. In addition, EC can also meet the extremely high requirements for interaction and mirror rendering in VR online games, including strict ultra-low latency and high bandwidth [2]. Under this trend, EC is bound to become an important pillar of future metaverse platforms.

As a crucial component of EC, multi-access edge computing (MEC, formerly known as mobile edge computing) has gained a lot of interest in recent years for improving users' quality of experience (QoE). MEC servers that are positioned in base stations within the radio access network enable moderate-capacity IT service delivery to nearby mobile users, which can ensure lower latency and better bandwidth capacity than remote clouds and give an unparalleled experience [7]. Besides, fog computing (FC) is also a representative computing mode of EC, which reduces communication between DCs and users by migrating tasks from central to edge devices for execution, hence alleviating bandwidth load. FC is similar to MEC except that the focus of FC is communication level requirements, whereas MEC concentrates on computation and network optimization.

In addition, considering that the computing capability and

resources on the edge are relatively limited compared to DCs, it is challenging to meet the rapidly growing service demand caused by massive and heterogeneous smart devices and the ubiquitous connectivity involved in metaverse platforms. Besides, the high deployment cost and long deployment cycle brought by numerous ESs are intolerable. In this regard, a novel computing framework, cloud-edge collaboration, is proposed to overcome this challenge, in which the resources of DCs, ESs, and devices can be well utilized at the same time to guarantee the smooth operation of computation-intensive and delay-sensitive applications (e.g., avatar computing) [1]. Fig. 4 shows a typical three-layer cloud-edge-end architecture.

Although high-quality service support with scalable bandwidth and ultra-low latency is the primary driving factor for metaverse platforms to rely on EC, the characteristics of EC also pose challenges to the performance enhancement of metaverse applications. In general, the unique features of EC refer to five aspects. That is, 1) high heterogeneity: the EC environment comprises multifarious end-users, servers, and edge devices that are heterogeneous in terms of computing power, load conditions, and requirements, etc.; 2) dynamic/time-varying: such as unstable network conditions and bandwidth fluctuations; 3) resource dispersion: ESs are widely distributed and have limited resources, necessitating reasonable management of dispersed edge resources; 4) limited service scope, ESs in MEC can only cover a limited service range through cellular network signals; 5) user mobility: the uncertain movement trajectories of mobile users make it more difficult for EC to provide continuous service. In order to take full advantage of EC to meet the diverse demands of various metaverse applications regarding computing performance and cost (e.g., energy efficiency), the edge needs to comprehensively consider computing requirements, network situations, and user characteristics, which takes improving resource utilization as the main means to make strategic resource management (e.g., the decision of offloading, caching, and resource allocation), so as to provide seamless services for metaverse businesses with a smooth experience in mobile scenarios.

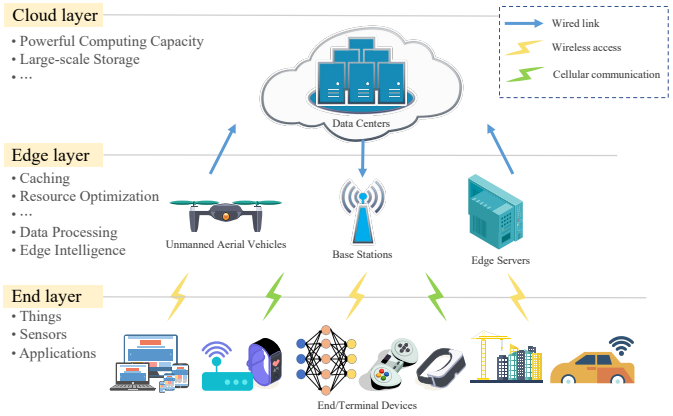

Fig. 4: A typical three-tier collaborative computing and communication framework in the metaverse consists of the cloud layer with powerful computing power and storage centers, the edge layer enabling rapid response, and the end layer with diverse end devices.

Edge Intelligence-Empowered Metaverse

Undoubtedly, the integration of EC with AI (i.e., edge intelligence) is inevitable to deliver stable and reliable technical services and achieve ubiquitous access. It is widely accepted to classify edge intelligence into two categories: 1) *AI for edge* that provides AI-based solutions to optimization problems in EC scenarios (e.g., the optimization of resources, offloading, and caching); 2) *AI on edge* that utilizes the platform and computing power offered by the edge side to solve AI-related problems (e.g., model training and inference), where the specific description is presented as follows:

- ***AI for edge.*** Benefiting from the distributed computing paradigm and deployment scheme, EC is regarded as a promising and highly scalable computing support for immersive experiences and real-time interactions. However, the ultimate form of the metaverse will not only support entertainment consumer applications but will also be widely used in production applications, such as transportation, industrial, medical, and defense. In different application scenarios, the quality of service (QoS) requirements of heterogeneous services, such as reliability, rate, delay, privacy protection, and mobility management, may be very different. Besides, the heterogeneity and dynamics of EC challenge the utilization of edge resources. Therefore, to ensure seamless/continuous service, strategic resource management is necessary, which mainly includes offloading decisions, resource allocation, and caching policies. In addition, since AI technology can learn from massive amounts of complex data and continuously adapt to various environments, AI is regarded as promising technical support to solve the key challenges faced by EC-enabled metaverse platforms, which has the potential to realize intelligent, automated, and efficient decision-making, such as applying deep RL (DRL) to complicated computation offloading problems.
- ***AI on edge***. AI technology has fully demonstrated its advantages in intelligent decision-making, digital content generation, and rendering. In general, the core components of AI applications include training and inference, where the traditional process refers to collecting massive amounts of data from mobile terminals and edge devices and training complex neural network models on the cloud. For the trained model, one is to perform it locally on the devices, and the other is to deploy it on a high-performance cloud platform. The device only needs to send the input to the cloud and wait for the inference results. However, such approaches have certain limitations; that is, local execution is subject to computation capacity and execution costs (e.g., energy consumption and execution latency) due to the large scale of AI models, such as the increasing depth of deep neural networks, while for training/inference on the cloud, the centralized aggregation of mobile network traffic due to the network scale and data transmission involved puts huge pressure on the network core. With the proliferation of smart devices, the development of the immersive

metaverse poses real-time computational requirements for future AI applications. It is obviously not enough to solely rely on the centralized computing paradigm. Applying the EC network architecture to carry out part of the business requirements can adapt well to the above challenges. For instance, AI tasks that are not computation-intensive but have high communication demands can be moved from the cloud to the edge, and it is also encouraged to send necessary data to DCs after complex data acquisition, storage, and preprocessing on the edge. In this way, the delay caused by network transmission can be reduced, and the workloads of the core network can be released.

Research issues in the above two categories mainly involve edge offloading, edge caching, and edge training and inference, where Fig. 5 outlines research issues in these three aspects in edge intelligence metaverse scenarios. For instance, computationally intensive rendering tasks could be offloaded to nearby ESs together with precaching popular content to accelerate execution. Besides, for AI-driven rendering tasks or intelligent decision-making, it is feasible to combine distributed computing paradigms and polynomial coding techniques to ensure performance demands. The edge intelligence-empowered metaverse has become a hot spot in academia, and initial progress has been made, as detailed in the categories below.

- ***Edge Offloading Decision***

The computation offloading technique enables computation-intensive or power-hungry workloads on mobile devices (e.g., AR/VR devices) sent to servers with sufficient resources. In EC, tasks with high computing requirements and execution costs (e.g., model training and foreground rendering) can be offloaded from devices to the edge for remote execution to ensure a high-speed, low-latency, and smooth experience, while devices only need to receive the processing result. Generally, offloading decisions and resource allocation are jointly optimized. In [8], the authors considered a distributed intelligent cloud-edge-end network architecture with the collaboration of ML and proposed a multi-agent DRL framework to obtain offloading and resource allocation schemes to decrease the total energy consumption under the task latency constraint, which applied federated learning (FL) to reduce the training overhead. Sun *et al.* [9] considered digital twin edge networks and proposed a mobile offloading solution based on DRL, aiming to reduce the latency of the offloading process while guaranteeing the migration cost constraint. VR online gaming has extremely high requirements for game interaction and mirror rendering, namely strict latency, high bandwidth, and support for numerous simultaneous players. Considering user mobility, Zhang *et al.* [2] modeled service placement as a Markov decision process and designed a hybrid game framework to dynamically place services on those edge clouds that lead to the best performance, maximizing the game's performance for all players. For example, local view updates and frame rendering can be placed on the edge for timely response and high bandwidth, while global state updates are sent to the cloud.

- ***Edge Caching Scheme***

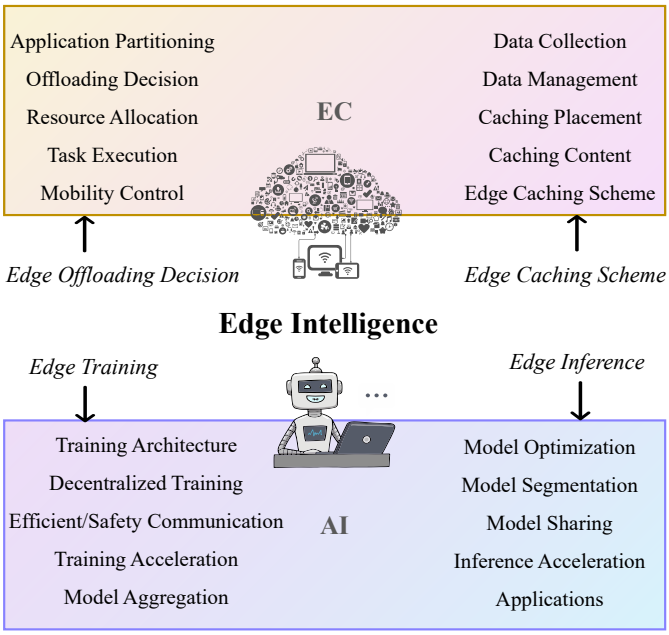

Fig. 5: Research issues of edge intelligence in metaverse scenarios.

As IoE applications grow in complexity, massive mobile services place enormous strain on network architectures, which may result in unnecessary waste of resources caused by repeated access to similar content by different users and high latency that significantly degrades QoE. In this regard, edge caching takes advantage of its geographical layout to pre-cache hot content near the terminals and responds immediately to requests within its service scope, thereby effectively alleviating the problems of data redundancy and transmission overhead network resource waste and improving QoS [10]. For example, the edge can cache data collected by edge devices (e.g., sensors) and provide training data for intelligent applications nearby. In general, edge caching involves cache placement, cache content, and cache policy, which are coupled with many challenges, such as the dynamics of edge networks, the limited and decentralized characteristics of edge resources, and popular content's time-varying nature. In the face of the uncertainty of network conditions, AI-powered edge caching is proposed as a promising solution to deal with the situation of incomplete prior information and an inaccurate environmental model. Guo *et al.* [11] investigated offloading scenarios for real-time VR rendering tasks in MEC, defined functions on energy consumption and latency to represent users' QoE, and proposed effective offloading schemes based on DRL to enhance QoE. Regarding the limited cache space of ESs, Yang *et al.* [12] investigated a novel cloud-edge-end service framework for AI-assisted VR content generation to guarantee VR devices' demands on latency and bandwidth, which allows for independent encoding of background/foreground content. Specifically, the authors proposed a graph neural network (GNN)-based caching policy where the edge content sharing mechanism and content's time-varying features are utilized to cache background content and GNN is used to predict

TABLE I: Summary of Research on Edge Intelligence in the Metaverse

Aspect	Scenarios	Key issues	Viable solutions	Ref.
Edge offloading	- Distributed real-time rendering for AR/VR applications; - Offloading data processing and AI-related computation tasks to the edge side; - Neural network partitioning and layer-based distributed collaborative computing.	- Energy efficiency and resource utilization; - Offloading granularity and application partitioning; - Mobility and network connectivity; - Scheduling and resource management.	- Online algorithms; - Path optimization and migration of virtual machine; - Load balancing; - DRL-based decision-making.	[2], [9]
Edge caching	- Collecting data for realizing AI algorithms; - Caching rendering content for VR/AR content generation/updating; - Computation caching on the edge for AI applications.	- Data collection and management; - Granularity of computation redundancy; - Cache objects determination.	- Path optimization; - Mobility and trajectory prediction; - Model selection and partition; - Load balancing; - Smart cache schemes.	[11]–[13]
Edge training and inference	- Efficiently training AI models required for intelligent metaverse applications (e.g., face and speech recognition, recommender system, and content generation) via utilizing edge resources; - Intelligent decision-making for non-player characters; - Personalized recommender system for metaverse applications.	- Efficiency of training and communication; - Tradeoff between computing overhead and latency; - Security and privacy; - Model optimization and deployment; - Inference acceleration; - Tradeoff between accuracy and latency.	- Offline algorithms; - ML-based distributed cloud-edge-end framework; - FL-enabled collaborative training; - Blockchain-based FL architecture; - Knowledge distillation and parameter pruning for model compression; - Model segmentation and sharing; - Edge caching.	[11], [14]

content, thereby proactively caching the requested content on the edge, and an update algorithm for background content based on GNN was designed to optimize caching. Regarding what content to cache, Kumar *et al.* [13] designed an effective algorithm combining convolutional neural network (CNN) and long short-term memory (LSTM) to obtain the caching scheme for 360° video streaming in the MEC network. To determine which videos needed to be cached, they used LSTM to predict future popularity and then utilized the CNN model to identify tiles suitable for caching instead of caching the entire video, aiming to reduce the pressure on the edge cache space and improve cache efficiency.

- ***Edge Training and Inference***

Typically, edge training mainly refers to two aspects: 1) the training architecture (including solo and collaborative training) needs to be determined according to the capacity of edge devices and servers, and if necessary, DCs may be introduced to cooperate with the edge for model training; 2) training optimization is essential to promote the efficiency of edge training since the model is extremely computationally intensive and the edge resource is relatively limited. As for edge inference, large-scale AI models with high accuracy demands pose challenges to real-time operation. Edge inference needs to make strategic considerations to balance inference accuracy and latency. On one hand, on the metaverse platform, AI applications are complicated and large-scale, requiring a huge amount of computing and storage, and the resource conditions between heterogeneous devices/servers are time-varying. Devices with limited capacity will have a significant impact on the overall model's training efficiency. In this context, it is necessary to make efforts on resource management (e.g., model segmentation and collaborative solutions by different devices), AI model optimization (e.g., parameter pruning, sharing, and model compression), and accelerating inference to achieve high-performance edge training and low-cost and accurate edge inference. In [11], the authors adopted a strategy of multi-agent DRL offline training and online running based on game theory to reduce the delay caused by the training process, while sharing training information between different agents at the same time is supposed to speed up training and improve performance. On the other hand, AI training involves numerous parameters/data transmission and updates, which often rely on co-training between multiple heterogeneous computing nodes, and privacy protection for important input data in the inference stage should be taken into account. To this end, the common solution is to adopt FL which is a distributed learning architecture. As an example, considering an FL-based cloud-edge-end architecture, mobile users in the metaverse that are treated as clients in FL use local data training models without uploading original data to the central cloud and then send updated model parameters to the edge for intermediate model aggregation, in which the central cloud is responsible for global aggregation and ESs can also participate in FL as clients. Such an architecture can make full use of the resources of all parties, alleviate network congestion, and reduce communication costs and the risk of data leakage. To meet the challenges of data security and privacy protection, Li *et al.* [14] considered a blockchain-based FL architecture to eliminate the security problems caused by the central server and allow some nodes to discard historical blocks so as to

alleviate the huge storage consumption for models.

To sum up, the emergence of edge intelligence is expected to further accelerate the construction of the metaverse ecosystem, where EC guarantees computing capability and network speed transmission and AI provides technical solutions to solve the key problems of the EC-enabled metaverse. Table I summarizes the aforementioned aspects of edge intelligence in the metaverse.

Open Issues and Future Directions

The existing research on edge intelligence mainly includes utilizing or optimizing AI algorithms for effective resource management (e.g., resource allocation and offloading strategy), caching schemes, and performing AI applications on terminals based on edge resources. However, there are still some interesting open issues about metaverse-oriented edge intelligence.

- **The quality of data**: The data collected from edge devices directly affects the performance of model training and inference. Compared with DCs, data and labels on edge nodes are scattered and scarce. Besides, the raw data generated by different metaverse applications and collected by devices might be biased and inconsistent. In this regard, the collaborative mode of FL can better adapt to the scenario of decentralized data, while approaches such as transfer learning and incremental learning can be used for personalized model training. Besides, encouraging users to provide more usable data through methods such as incentive mechanisms is also of importance for AI learning performance.
- **Adaptability of models**: Due to the limitations of edge resources, lightweight and high-precision AI models are necessary, and potential solutions to optimize models include knowledge distillation and model compression. It is also critical to design a reasonable coordination mechanism between terminals and servers to achieve compatibility because of the heterogeneity between devices and servers. Furthermore, in order to improve the applicability of AI models to better deal with unfamiliar scenarios, lifelong ML is a feasible solution, but it needs knowledge evaluation to learn and accumulate useful knowledge. Reducing the complexity of the model, rapid deployment, and delivery are also good future research.
- **Applicability of scenarios**: More complex and challenging edge scenarios, such as jointly optimizing resources (e.g., offloading granularity, network slicing, resource allocation), caching, and mobility management, must be considered in order to better serve the variety of applications and demands in the future metaverse. The potential advantages of EC and AI should be fully exploited to deliver personalized and diverse services for users in more complicated cloud edge-end environments consisting of multiple heterogeneous users with cooperative or competing interactions, ESs, and DCs. Furthermore, green EC, such as energy harvesting technology, is an important research direction to assure the sustainability of metaverse development.
- **The establishment of evaluation indexes**: There are various physical and virtual service providers and user accesses on metaverse platforms, and it is difficult to customize the performance assessment criteria for all parties considering the heterogeneous characteristics. For example, AI application metrics refer to inference latency, model overhead, service revenue, and model accuracy, and the user experience includes not only an immediate response but also personalized preferences (e.g., color and visual sensitivity), while service providers need to consider QoE, network overhead, service costs, and customized AI architectures for reasonable resource management and deployment. Thus, evaluation metrics must be optimally defined across disciplines.
- **Security and privacy**: It is inevitable to involve the issues of end-user privacy and data security when performing metaverse applications, such as identity theft of avatars. Although blockchain is regarded as a decentralized solution, it will incur high storage and computing costs while ensuring data security; hence, it is required to incorporate other technologies (e.g., FL) to reduce the impact of cost on performance [15].

Conclusions

This paper has provided an overview of the metaverse architecture and platform driven by cutting-edge technologies. Regarding the digital infrastructure for the metaverse, we have discussed edge computing, as well as research work on the fusion of edge computing and artificial intelligence around the construction of metaverse ecology. Finally, we have pointed out some open issues and potential future research directions.

Acknowledgments

This work was supported in part by the National Science and Technology Major Project under Grant 2020YFB1807601, the Guangdong Key Areas Research and Development Program under Grant 2022B0101010001, the Shenzhen Science and Technology Program under Grants JCYJ20210324095209025, the National Natural Science Foundation of China under Grant No.62361056, and the Applied Basic Research Foundation of Yunnan Province under Grant Nos. 202201AT070203 and 202301AT070422.

References

[1] M. Xu, W. C. Ng, W. Y. B. Lim, J. Kang, Z. Xiong, D. Niyato, Q. Yang, X. Shen, and C. Miao, "A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 656–700, 2023.

[2] W. Zhang, J. Chen, Y. Zhang, and D. Raychaudhuri, "Towards efficient edge cloud augmentation for virtual reality mmogs," in *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, ser. SEC '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3132211.3134463

[3] Y. Fu, C. Li, F. R. Yu, T. H. Luan, P. Zhao, and S. Liu, "A survey of blockchain and intelligent networking for the metaverse," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3587–3610, 2023.

[4] T. Huynh-The, Q.-V. Pham, X.-Q. Pham, T. T. Nguyen, Z. Han, and D.-S. Kim, "Artificial intelligence for the metaverse: A survey," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105581, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197622005711

[5] Z. Meng, C. She, G. Zhao, and D. De Martini, "Sampling, communication, and prediction co-design for synchronizing the real-world device and digital model in metaverse," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 288–300, 2023.
[6] B. Cao, Z. Wang, L. Zhang, D. Feng, M. Peng, L. Zhang, and Z. Han, "Blockchain systems, technologies, and applications: A methodology perspective," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 353–385, 2023.
[7] W. Liu, B. Cao, and M. Peng, "Blockchain based offloading strategy: Incentive, effectiveness and security," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 12, pp. 3533–3546, 2022.
[8] X. Huang, K. Zhang, F. Wu, and S. Leng, "Collaborative machine learning for energy-efficient edge networks in 6g," *IEEE Network*, vol. 35, no. 6, pp. 12–19, 2021.
[9] W. Sun, H. Zhang, R. Wang, and Y. Zhang, "Reducing offloading latency for digital twin edge networks in 6g," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 12 240–12 251, 2020.
[10] Y. Sun, Z. Chen, M. Tao, and H. Liu, "Communications, caching, and computing for mobile virtual reality: Modeling and tradeoff," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7573–7586, 2019.
[11] F. Guo, F. R. Yu, H. Zhang, H. Ji, V. C. M. Leung, and X. Li, "An adaptive wireless virtual reality framework in future wireless networks: A distributed learning approach," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8514–8528, 2020.
[12] J. Yang, Z. Guo, J. Luo, Y. Shen, and K. Yu, "Cloud-edge-end collaborative caching based on graph learning for cyber-physical virtual reality," *IEEE Systems Journal*, pp. 1–12, 2023.
[13] S. Kumar, L. Bhagat, A. A. Franklin, and J. Jin, "Multi-neural network based tiled 360°video caching with mobile edge computing," *Journal of Network and Computer Applications*, vol. 201, p. 103342, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S108480452200011X
[14] Y. Li, C. Chen, N. Liu, H. Huang, Z. Zheng, and Q. Yan, "A blockchain-based decentralized federated learning framework with committee consensus," *IEEE Network*, vol. 35, no. 1, pp. 234–241, 2021.
[15] M. Cao, L. Zhang, and B. Cao, "Toward on-device federated learning: A direct acyclic graph-based blockchain approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 4, pp. 2028–2042, 2023.

BIOGRAPHIES

Yanan Xu (bfg_xyn@163.com) obtained her B.S. and M.S. degrees in Software Engineering from Yunnan University, China, in 2019 and 2023, respectively. She is currently pursuing a Ph.D. degree in College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. Her current research interests include edge computing and energy-efficient computing.

Daquan Feng (fdquan@szu.edu.cn) is an Associate Professor with the Guangdong Province Engineering Laboratory for Digital Creative Technology and Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His research interests include URLLC communications, LTE-U, and massive IoT networks. He is the corresponding author for this article.

Mingxiong Zhao (mx_zhao@ynu.edu.cn) is currently an Associate Professor at the School of Software, Yunnan University, Kunming, China. His current research interests include network security, physical layer security, mobile edge computing, and edge AI techniques.

Yao Sun (Yao.Sun@glasgow.ac.uk) is currently a Lecturer with James Watt School of Engineering, the University of Glasgow, Glasgow, UK. His research interests include intelligent wireless networking, semantic communications, blockchain system, and resource management in next generation mobile networks. He is a senior member of IEEE.

Xiang-Gen Xia (xianggen@udel.edu) is the Charles Black Evans Professor in the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716, USA. His current research interests include space-time coding, MIMO and OFDM systems, digital signal processing, and SAR and ISAR imaging. He is a Fellow of the IEEE and has served as an Associate Editor for numerous international journals, including IEEE Transactions on Signal Processing, IEEE Transactions on Wireless Communications, IEEE Transactions on Mobile Computing, and IEEE Transactions on Vehicular Technolog.