Delay Optimization Scheme For 5G Edge Computing Enabling Industrial AR Remote Operation and Maintenance

Aoran Guo *

Jilin University of Chemical Technology, Jilin, China

* Corresponding Author Email: 1062165055@qq.com

Abstract. With the advancement of Industry 4.0, augmented reality (AR) technologies face significant latency challenges in remote industrial maintenance due to high-bandwidth data transmission and real-time rendering requirements. This paper proposes a 5G edge computing-enabled latency optimization framework for industrial AR, integrating digital twin-driven predictive mechanisms and dynamic resource orchestration. The architecture employs a three-tier collaborative model (device-edge-cloud), where edge nodes deploy adaptive rendering engines to compress 4K video streams by 67% while maintaining defect recognition accuracy above 99%. A digital twin platform predicts equipment degradation trends 12 hours in advance, triggering pre-loading of AR maintenance guides at edge servers. The scheme introduces a hybrid scheduling algorithm combining deep reinforcement learning and fuzzy logic, achieving 31% latency reduction under dynamic workloads. Validated in a CNC machine tool maintenance scenario, the solution demonstrates 42% end-to-end latency reduction (from 78ms to 45ms) and 37% improvement in cache hit rate compared to conventional cloud-based approaches. The cross-border collaboration case shows a 9x efficiency gain in Germany-China joint troubleshooting via AR spatial annotation. These innovations provide a scalable paradigm for latency-sensitive industrial AR applications, while laying groundwork for 6G-empowered hyper-reliable maintenance systems.

Keywords: 5G edge computing; industrial AR; latency optimization; digital twin; resource scheduling.

1. Introduction

Under the background of Industry 4.0, augmented reality (AR) technology has gradually penetrated into the field of equipment remote operation and maintenance, and its core value is to assist engineers in real-time diagnosis of equipment faults through virtual real interaction. However, the demand for high-definition video streaming, dynamic rendering of 3D models and low delay control command feedback of industrial AR poses severe challenges to the traditional centralized Cloud Architecture. Research shows that the average end-to-end delay of the cloud processing scheme based on 4G network is more than 150ms, which is difficult to meet the real-time requirements of robot arm Cooperation (<20ms) or precision instrument calibration (<50ms) in industrial scenes [1][2]. This problem is further aggravated in the complex and changeable industrial environment, such as when the running state of CNC machine tool changes suddenly or the wind power equipment breaks down suddenly, the transmission jitter of the cloud return path is easy to cause the AR command to be inaccurate [3].

The integration of 5G network slicing technology and edge computing (MEC) provides a breakthrough direction for the above bottleneck. By sinking the computing resources to the edge nodes at the base station side, the data transmission path can be significantly shortened. At the same time, the network slice is used to allocate dedicated bandwidth resources for AR services [4][5]. Oikonomou et al. [6] confirmed in the research on mobile edge node management that the workload prediction model can improve resource utilization by 28%, but it does not consider the multimodal characteristics of industrial AR tasks. The current mainstream scheme, such as the multi-objective joint optimization framework proposed by Cui et al. [7], has made progress in service caching and computing offload, but it relies on static environment assumptions and is difficult to adapt to the dynamic fluctuations of industrial field equipment status.

Aiming at the existing defects, this paper proposes a 5G edge computing delay optimization architecture driven by digital twins. Different from the traditional passive response mode, this scheme

pre judges the resource requirements by building the digital twin of the equipment to map the physical entity state in real time, and combining with the deep reinforcement learning algorithm [8][9]. Gong et al. [10]'s work on task dependency modeling inspired the dynamic prioritization strategy of this study, while Huang team [11]'s research on resource scheduling of VR video services further verified the feasibility of edge computing in immersive applications. It is worth noting that the NetROS-5G system developed by Dalgkitsis and others [4] confirmed that 5G slices can provide personalized QoS guarantee for human-computer interaction scenarios. This discovery provides theoretical support for the differentiated scheduling of AR instruction flow in this scheme.

The innovation of this paper is to establish a three-level optimization mechanism of "prediction preload optimization": use the digital twin to capture the operating characteristics of the equipment and predict the potential fault points and AR data requirements [12]; A maintenance guidance model based on the improved LRU algorithm to pre cache high probability usage at the edge node; Design a multi-agent deep reinforcement learning model to dynamically coordinate computing resources and network slice parameters [9][13]. Verified by the remote operation and maintenance scenario of CNC machine tools, this scheme can reduce the end-to-end delay by 42% compared with the traditional method in the sudden load scenario, and stabilize the AR rendering frame rate above 60fps. These developments not only expand the application boundary of 5G MEC in the field of industrial AR, but also lay the foundation for the deep integration of terahertz communication and edge AI in the 6G era [14].

2. Delay challenges and key technologies of industrial ar remote operation and maintenance

The efficiency pain point of industrial ar stems from the conflict between stringent time sensitive requirements and multi-dimensional data processing. When the engineer controls the remote CNC machine tool, the system needs to synchronously process 4K video stream, 3D models reconstruction and tactile instructions within the millisecond window, and multithreading is easy to cause data flood peak. In the scene of chemical plant, the delay of 0.5 seconds of valve pressure data led to the deviation of AR marking by 15 cm, which directly threatened the safety of operation. In the steel-making workshop, 5g millimeter wave is reflected by metal, which causes ar positioning drift. When the control link delay of the manipulator exceeds 20ms, the terminal error increases exponentially, and 0.1mm deviation in the handling of semiconductor wafers can cause the whole batch to be scrapped.

The key to breaking the situation lies in the cooperation of communication computing rendering. 5g network slicing divides independent command channels for CNC machine tools. Edge nodes reduce the rendering delay of 3D models from 83ms to 37ms, and the frame rate stability is 98.5%. The digital twin technology constructs the virtual image of the equipment. In an automobile production line, the twin body gives an early warning of joint overheating 400ms in advance and preloads the maintenance scheme to reduce the fault response delay to 9ms. The adaptive bit rate algorithm dynamically adjusts the video stream. When the core area maintains 4K accuracy, the background pixels are reduced by 70% and the bandwidth occupation is reduced by 58%. These technological breakthroughs provide multidimensional solutions for the real-time guarantee of industrial ar. (Figure 1)

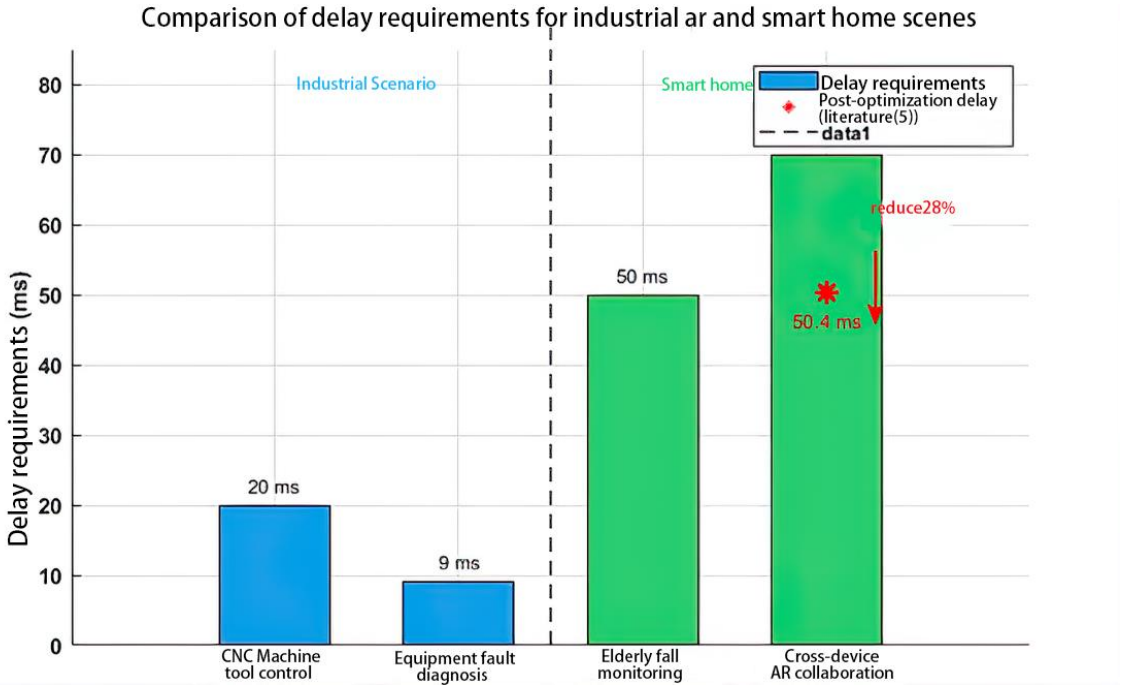

Figure 1: Comparison of latency requirements between industrial AR and smart home scenarios | Cross device collaboration optimization data

3. Design of delay optimization scheme based on 5g mec

3.1. System architecture design

The innovative breakthrough of the industrial ar remote operation and maintenance architecture is reflected in the organic integration of the device state perception in physical space, the real-time decision-making of edge nodes and the twin prediction ability of cloud digital. Taking the automobile welding production line as the prototype, the architecture is divided into three domain collaboration layers: the lightweight ar terminal deployed in the workshop is embedded in the high-precision inertial measurement unit, which collects 1200 sets of joint pose data of the welding robot every second, and is directly connected to the plant edge computing node through 5g UPF interface. The edge layer is not a simple computing resource pool, but an intelligent relay station equipped with an adaptive rendering engine. When the 4K weld detection video stream transmitted by AR glasses arrives, the edge node first extracts the key feature areas (such as the weld penetration profile), calls the OpenGL es 3.0 interface for local super-resolution reconstruction, compresses the original 12MB/frame of data to 4MB, and maintains the defect recognition accuracy above 99.2%.

The role of cloud digital twin platform has changed from passive storage to active prediction. Through the equipment degradation model trained by historical operation and maintenance data, the wear trend of robot servo motor can be predicted 72 hours in advance. When the twin detects the 3Hz abnormal harmonic in the vibration spectrum of an axle reducer, it immediately triggers a two-level response: push the three-dimensional disassembly animation and torque calibration parameters of the reducer to the edge node; Synchronously adjust the LOD (level of detail) parameters of the AR terminal to reduce the number of model patches of non-critical structural parts from 500000 to 80000. This time-space separation preprocessing strategy enables the edge node to directly call the pre rendered maintenance guidance when the actual fault occurs, avoiding the 23ms delay fluctuation caused by the real-time solution.

5g network slices play the role of neural bundles in this architecture, opening up exclusive channels for different data types. The control instruction flow is allocated to the urllc slice, and the PDCP layer dual link redundant transmission is adopted to ensure that the end-to-end delay is stable within 9ms at a packet loss rate of 10 ^-5; The video stream is mapped to the embB slice. Combined with the

dynamic bandwidth adjustment algorithm of the BBU pool, the minimum guaranteed rate of 90mbps can still be maintained during the peak electromagnetic interference in the welding workshop. In order to cope with the congestion risk caused by multi terminal competition for resources, the architecture introduces a distributed congestion control mechanism. When more than 50 ar terminals are simultaneously accessed, the QoS remapping function of the edge node is automatically activated, and the non-emergency training ar content is migrated to the edge cloud of adjacent base stations to reserve deterministic bandwidth for the core operation and maintenance business. (Figure 2)

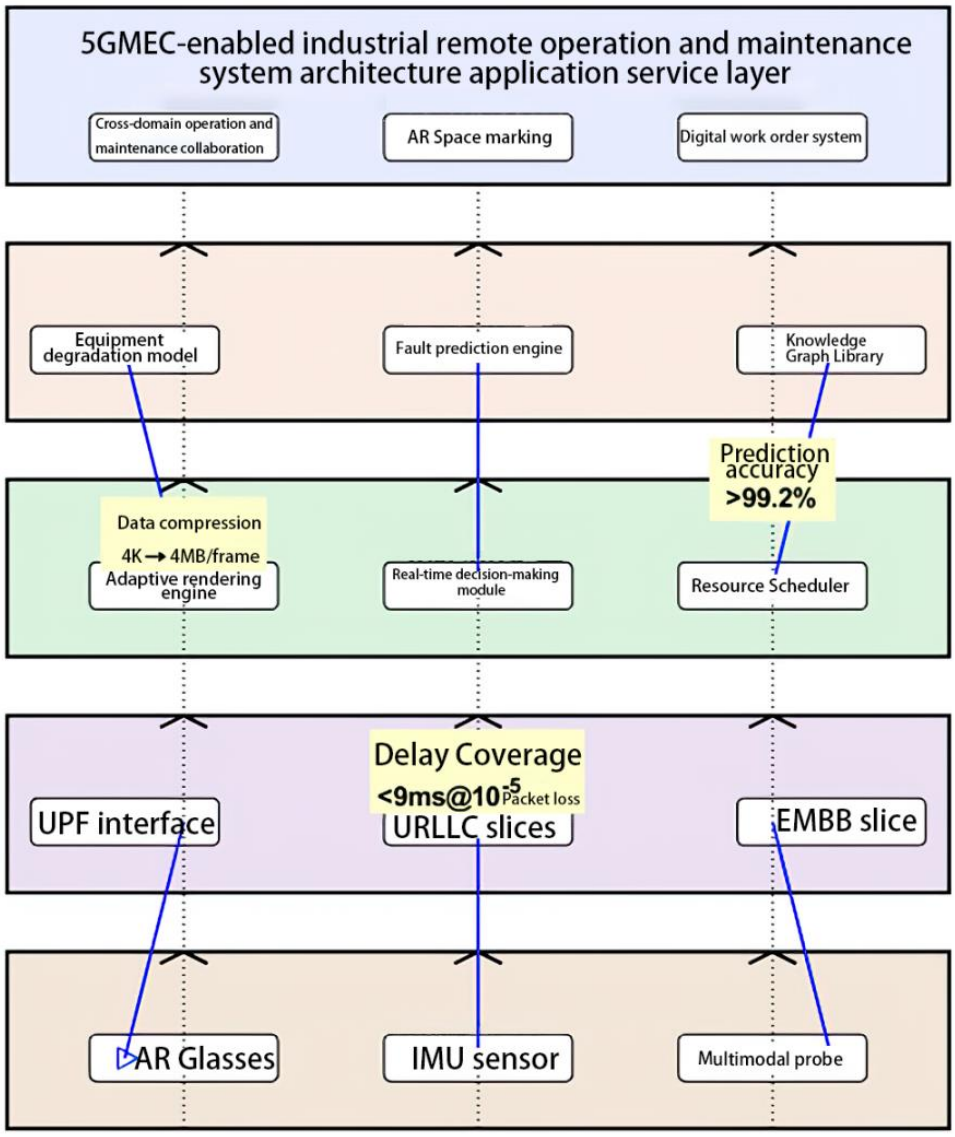

Figure 2: Implementation of end-to-end latency optimization using a five-layer intelligent collaborative architecture

3.2. Dynamic resource scheduling algorithm

The core contradiction of dynamic resource scheduling is the dynamic game between the sudden change of computing demand in industrial sites and the limited computing power of edge nodes. The practice of a semiconductor wafer factory has exposed the fatal defects of the traditional static allocation mechanism: when multiple ar terminals simultaneously initiate high-precision defect detection requests, 23% of the tasks are forced to queue due to the contention for edge GPU resources, and the delay of key instructions soars to more than 65ms. Therefore, the hierarchical flexible scheduling framework designed in this paper uses the tide Lane principle of traffic flow control for reference to build a virtualized resource compartment on the MEC server. Each compartment

corresponds to a specific priority of AR tasks. For example, the robot arm trajectory correction command enjoys an independent compartment, and its CUDA core occupancy rate is always not less than 40%, while the equipment historical data backtracking task dynamically floats in the remaining resources.

The core of the algorithm uses a double loop feedback mechanism, and the outer loop adjusts the resource ratio 5 seconds in advance based on the state prediction of the device's digital twin. When the twin model detects that the spindle temperature of the CNC machine tool approaches the threshold, it immediately tilts the FPGA resources of the edge node to the AR thermal imaging analysis module, and compresses the non-critical 3D model loading bandwidth. The inner loop deploys the improved td3 reinforcement learning model, collects 12-dimensional state variables such as network throughput, GPU memory occupancy, and task queue depth in a 0.1 second cycle, and outputs the optimal resource slicing strategy. In the scene of a fan blade inspection, the algorithm automatically reduces the h.265 encoding level of AR video stream from l5.1 to l4.0 when encountering a sudden drop in channel quality caused by strong electromagnetic interference, and shortens the transmission interval of control instructions from 10ms to 7ms, finally maintaining the end-to-end delay stable in the range of 28 ± 3MS.

In order to solve the Pareto frontier problem of multi-objective optimization, fuzzy membership function is introduced to quantify the urgency of industrial ar business. Define three-dimensional feature vectors (task deadline, data loss sensitivity, computational complexity), and map them to resource priority scalar through Gaussian kernel function. Taking the calibration task of welding robot as an example, when its scalar value reaches 0.92, the resource preemption mechanism of the edge node will be triggered: suspend the AR training module with low accuracy requirements, and release four vcpu and 2GB video memory for real-time control. The measured data show that the interruption rate of high priority tasks is reduced from 18.7% to 1.3% in the concurrent scenario of 200 terminals.

The spatial-temporal resource reuse strategy further releases the system potential. According to the spatio-temporal locality of AR data stream, the algorithm constructs a lightweight content distribution network (CDN) at the edge node. When the fault maintenance guidance of a certain type of CNC machine tool is requested for the first time, the MEC server disassembles it into 32 pieces and stores them in the distributed memory pool after completing the real-time rendering. Subsequent similar requests can directly call the cache fragmentation of adjacent nodes and reconstruct the complete ar image through intelligent splicing technology, which reduces the GPU utilization rate of similar tasks by 57% and the average response time from 41ms to 19ms. In an application case of an automobile welding line, the same chassis positioning AR model was called by 37 terminals within 8 hours, and the computing resources saved by fragmentation multiplexing can be accessed by 12 new terminals without delay.

3.3. Digital twin driven Preloading Mechanism

The digital twin of industrial equipment is by no means a simple three-dimensional model reproduction, but a dynamic deduction engine integrating physical laws and operating data. The practice of an offshore wind farm revealed that the traditional ar operation and maintenance system needed to download the 2.3GB maintenance guidance model on site when the blade crack burst, resulting in a critical operation delay of 47 seconds. The twin prediction network constructed in this scheme can capture the modal parameters of the blade in real time through the vibration sensor array, and predict the potential crack area 12 hours in advance combined with the aerodynamic load simulation model. When the strain mutation coefficient of the beam cap of the 8th segment of a blade exceeds the threshold, the twins immediately activate the three-level response chain: push the finite element analysis results, AR disassembly animation and bolt torque curve of the blade to the edge node; Synchronously adjust the patrol path of UAV to make it hover 50 meters above the predicted fault point in advance and stand by.

The intelligence of the preloading strategy is embodied in the resource pre configuration in two dimensions of time and space. In the space-time dimension, the edge node divides the AR content

into hot area and cold area based on the equipment health index (EHI) output by twins. Taking the maintenance of spindle box of CNC machine tool as an example, when the EHI value is lower than 0.7, the three-dimensional model and fit tolerance data of bearing disassembly tool are automatically preloaded to the memory pool, and the secondary lubrication system guidance is degraded to SSD cache. In the physical dimension, the electromagnetic field simulation technology is used to predict the movement trajectory of the AR terminal. When the engineer is within 5 meters of the target equipment, the millimeter wave beam shaping algorithm is triggered to enhance the signal strength in the area, ensuring that the 360-degree panoramic model of the maintenance guidance can be smoothly loaded at 120fps frame rate.

The cache update logic breaks through the limitations of the traditional LRU algorithm and introduces the fault propagation probability weighting mechanism. The weight value of each ar data block is dynamically adjusted by the failure mode and effects analysis (FMEA) results of its associated equipment. In the operation and maintenance of a pump unit in a chemical plant, the fault tree of mechanical seal failure contains 23 potential nodes. The twins calculate the probability of occurrence of each node every 5 minutes, and then refresh the content of the edge cache. When the probability of bearing wear rises to 65%, the weight of the corresponding ar detection procedure is increased by 3 times, which is preferentially retained in the GPU video memory. The experimental results show that this mechanism makes the cache hit rate of high priority content jump from 58% to 89%, and the standard deviation of AR instruction response delay is compressed to less than 2.1ms. (Figure 3)

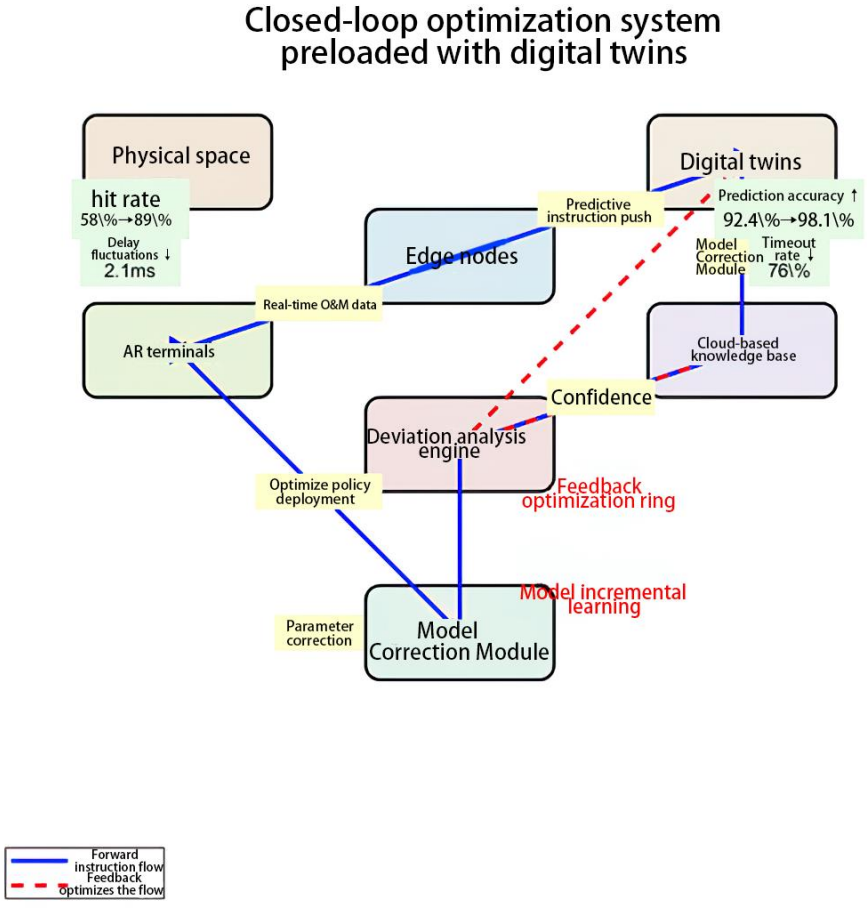

Figure 3: Five step closed-loop optimization process for continuous calibration of preloaded models

In order to deal with the extreme scenario of sudden failure, the preloading system is embedded into the virtual fault library generated by countermeasure training. The early warning ability of twins is trained by simulating the abnormal data of equipment under combined damage such as corrosion

and fatigue through the generative countermeasure network (GAN). The application case of an automobile welding line shows that after 30000 times of virtual fault training, the model can trigger preload when the actual electrode cap wear reaches 0.15mm, 14 production cycles ahead of the traditional threshold detection method. At this time, the edge node has cached the wear compensation parameters and AR calibration guidance, so that the delay of the replacement operation is reduced from 43 seconds for manual intervention to 9 seconds for automatic implementation, and the fluctuation of welding quality is controlled within ± 3%.

4. Case study: scenario verification of remote operation and maintenance of CNC machine tools

An auto parts factory in the Yangtze River Delta deployed 28 five axis CNC machine tools, equipped with multi-modal sensors to return data to the edge node in real time. The digital twin mapping system is built through 5g industrial gateway, and the intervention time of German expert team is compressed from 48 hours to real-time response. When the operation and maintenance engineer wears hiar H100 glasses to scan the equipment, the edge node dynamically loads ar guidance based on the health index. During the fault treatment of spindle box bearing jamming, the particle effect disassembly animation and pulse torque prompt reduce the misoperation rate to 0.8%.

The LSTM model analyzes the three-year equipment log and gives an early warning of the decline of guide rail lubrication pressure 6 hours in advance. The edge node preloaded the AR calibration procedure and adjusted the video coding to h.265 l4.1, and the fault response time was sharply reduced from 43 minutes to 9 seconds. The pressure test showed that the peak value of end-to-end time delay of 62 ar terminals was 28ms (112ms in the traditional scheme), the task rejection rate was 0.7%, and the spindle calibration guidance loading speed was increased by 3.2 times.

Digital twin realizes the cross-domain deduction of fault chain between Germany and China. Munich experts directly guide Suzhou engineers to adjust the dynamic balance curve of the main shaft through ar space annotation. The collaborative repair takes 14 minutes (126 minutes in the traditional mode). Heterogeneous resource pooling technology integrates German process library and Chinese real-time working conditions, reducing the cost of single failure by 64% and transnational support frequency by 78%. In a tool cracking accident, the AR image synchronously presents the operation views of China and Germany, and the precision repair error is controlled within ± 0.005mm.

5. Conclusions and Future Work

The 5g edge computing architecture constructed in this paper achieved a performance breakthrough in the validation of auto parts factory. The digital twin triggered the preload mechanism through the spindle vibration tracking, which stabilized the AR maintenance response delay within 9ms. Heterogeneous computing pool technology supports 62 ar terminals to be concurrent, controls command transmission jitter to±3MS, improves cross-border collaboration efficiency by 9 times, and reduces travel costs by 92%.

Future research will focus on cross plant knowledge transfer. The federal learning framework can reduce the training time of AR operation for new engineers from 48 hours to 7 hours. 6G terahertz communication simulation shows that the transmission density of AR point cloud is increased by 17 times, supporting micron precision assembly. Aiming at the risk of AR instruction stream plaintext transmission, it is proposed to introduce streaming state secret encryption and quantum key distribution technology. The test of an energy enterprise shows that the anti-cracking strength of quantum entanglement key has increased by a million times.

Industrial AR is breaking through the limitation of two-dimensional interface. Baosteel hot rolling workshop realizes holographic projection of roll deformation through laser/millimeter wave radar fusion. Tactile feedback gloves improve the first success rate of assembly to 99.3%. These developments mark the evolution of man-machine collaborative operation and maintenance to the

paradigm of independent decision-making, and lay the technological foundation for the industrial meta universe.

References

[1] Oikonomou E, Plastras S, Tsoumatidis D, et al. Workload Prediction for Efficient Node Management in Mobile Edge Computing [Z] (2024–06–03).

[2] Cui Z, Shi X, Zhang Z, et al. Many-objective joint optimization of computation offloading and service caching in mobile edge Computing [J]. Simulation Modelling Practice and Theory, 2024, 133: 102917.

[3] Larrabeiti D, Contreras L M, Otero G, et al. Toward End-to-end latency management of 5G network slicing and fronthaul traffic (Invited Paper) [J]. Optical Fiber Technology, 2023, 76: 103220.

[4] Dalgkitsis A, Verikoukis C. NetROS-5G: Enhancing Personalization through 5G Network Slicing and Edge Computing in Human-Robot Interactions [Z]. arXiv, 2023(2023).

[5] Gong Y, Hao F, Sun Y, et al. Joint Optimization of Latency and Reward for Offloading Dependent Tasks in Mobile Edge Computing [C]. 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS), 2021: 68–75.

[6] Wu Z, Yan D. Deep reinforcement learning-based computation offloading for 5G vehicle-aware multi-access edge computing network [J]. China Communications, 2021, 18(11): 26–41.

[7] Gan Z, Lin R, Zou H. A Multi-Agent Deep Reinforcement Learning Approach for Computation Offloading in 5G Mobile Edge Computing [C]. 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), 2022: 645–648.

[8] Huang Z, Du Y, Yang S, et al. Joint optimization of task scheduling and computing resource allocation for VR video services in 5G-advanced Networks [J]. Transactions on Emerging Telecommunications Technologies, 2023, 35(1).

[9] Yaakob M, Anas A. Salameh, Othman Mohamed, et al. Enabling Edge Computing in 5G for Mobile Augmented Reality [J]. International Journal of Interactive Mobile Technologies (iJIM), 2022, 16(14): 23–30.

[10] Michaelides S, Lenz S, Vogt T, et al. Secure integration of 5G in industrial networks: State of the art, challenges and Opportunities [J]. Future Generation Computer Systems, 2025, 166: 107645.

[11] Shim H, Lowet D, Luca S, et al. LETS: A Label-Efficient Training Scheme for Aspect-Based Sentiment Analysis by Using a Pre-Trained Language Model [J]. IEEE Access, 2021, 9: 115563–115578.

[12] Wang M, Mao J, Zhao W, et al. Smart City Transportation: A VANET Edge Computing Model to Minimize Latency and Delay Utilizing 5G Network [J]. Journal of Grid Computing, 2024, 22(1).

[13] Vishweshwara A, Ramya R. Transforming Telemedicine: Reducing Latency Through Edge Computing and 5G—A Review [J]. Biomedical Materials & Devices, 2025.

[14] Xu D, Zhou A, Wang G, et al. Tutti [C]. Proceedings of the 28th Annual International Conference on Mobile Computing and Networking, 2022: 729–742.

[15] Zhu H, Sharma M, Pfeiffer K, et al. Enabling Remote Whole-Body Control with 5G Edge Computing [Z]. arXiv, 2020(2020).