An International ICT Journal
Featuring Industry-University-Institute Cooperation

ISSN 1673-5188
CN 34-1294/TN

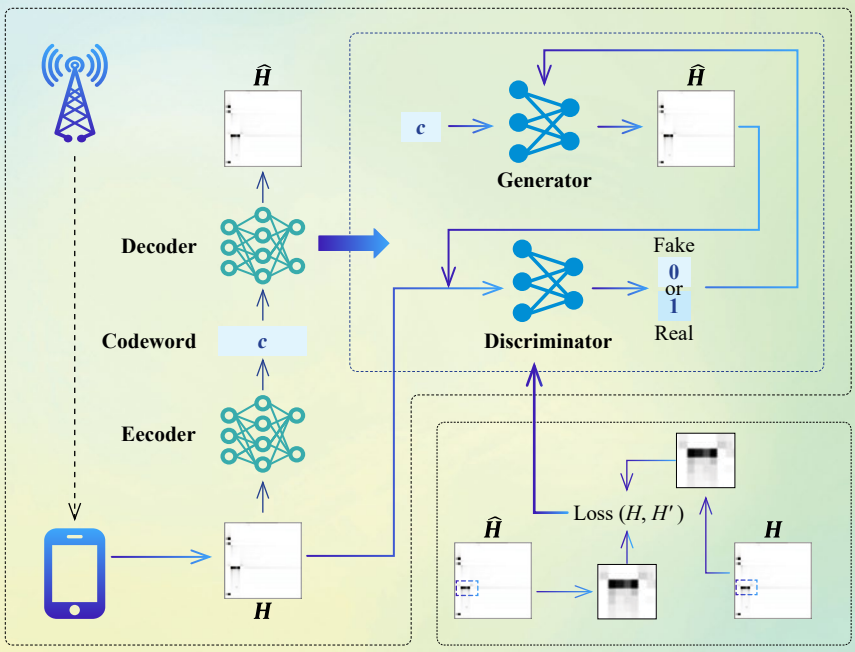
ZTE COMMUNICATIONS

中兴通讯技术(英文版)

https://zte.magtechjournal.com

March 2026, Vol. 24 No. 1

Special Topic: Achievements of ZTE's Industry-University-Institute Cooperation Projects



(See Fig. 2 on P. 7)

ISSN 1673-5188
03>

The 10th Editorial Board of ZTE Communications

CONTENTS

ZTE COMMUNICATIONS
March 2026 Vol. 24 No. 1 (Issue 94)

Serial parameters: CN 34–1294/TN*2003*q*16*106*en*P*¥30.00*2200*13*2026-03

Special Topics for 2026

Special Topic	Leading Guest Editor
1 **Achievements of ZTE's Industry-University-Institute Cooperation Projects**	**Xu Chengcheng (University of Macau, China)**
2 AI-Agent Communication Network (ACN): Architecture, Protocols and Key Technologies	Sun Tao (China Mobile Research Institute, China)
3 Reconfigurable Antenna Systems for Next-Generation Mobile Communications	Yang Kun (Nanjing University, China)
4 Goal-Oriented Joint Semantic and Channel Coding for Future Communications	Yuan Jinhong (University of New South Wales, Australia)

ZTE Communications Guidelines for Authors

Remit of Journal

ZTE Communications publishes original theoretical papers, research findings, and surveys on a broad range of communications topics, including communications and information system design, optical fiber and electro-optical engineering, microwave technology, radio wave propagation, antenna engineering, electromagnetics, signal and image processing, and power engineering. The journal is designed to be an integrated forum for university academics and industry researchers from around the world.

Manuscript Preparation

Manuscripts must be typed in English and submitted electronically in MS Word (or compatible) format. The word length is approximately 3 000 to 8 000, and no more than 8 figures or tables should be included. Authors are requested to submit mathematical material and graphics in an editable format.

Abstract and Keywords

Each manuscript must include an abstract of approximately 150 words written as a single paragraph. The abstract should not include mathematics or references and should not be repeated verbatim in the introduction. The abstract should be a self-contained overview of the aims, methods, experimental results, and significance of research outlined in the paper. Three to eight carefully chosen keywords must be provided with the abstract.

References

Manuscripts must be referenced at a level that conforms to international academic standards. All references must be numbered sequentially intext and listed in corresponding order at the end of the paper. References that are not cited in-text should not be included in the reference list. References must be complete and formatted according to *ZTE Communications* Editorial Style. A minimum of 10 references should be provided. Footnotes should be avoided or kept to a minimum.

Content and Structure

ZTE Communications seeks to publish original content that may build on existing literature in any field of communications. Authors should not dedicate a disproportionate amount of a paper to fundamental background, historical overviews, or chronologies that may be sufficiently dealt with by references. Authors are also requested to avoid the overuse of bullet points when structuring papers. The conclusion should include a commentary on the significance/future implications of the research as well as an overview of the material presented.

Peer Review and Editing

All manuscripts will be subject to a two-stage anonymous peer review as well as copyediting, and formatting. Authors may be asked to revise parts of a manuscript prior to publication.

Biographical Information

All authors are requested to provide a brief biography (approx. 100 words) that includes email address, educational background, career experience, research interests, awards, and publications.

Acknowledgments and Funding

A manuscript based on funded research must clearly state the program name, funding body, and grant number. Individuals who contributed to the manuscript should be acknowledged in a brief statement.

Address for Submission

http://mc03.manuscriptcentral.com/ztecom

To the Communications Community —2026 New Year's Message

Zhang Ping

At this historic juncture of deepening technological revolution and industrial transformation, China's communication sector stands on the eve of another great leap forward. Reflecting on the development of communications over the past two decades, China has forged an innovative path from catching up to keeping pace and then to leading the way. Today, at the new starting point of 6G development and facing the paradigm shift brought about by "AI + communications," China's scientific research community, with the courage to venture into uncharted territory, is advancing original theories such as the new communication paradigm based on a unified theoretical framework of information theory to the global forefront. This fully demonstrates the strategic resolve and institutional advantages of Chinese modernization in the realm of technological innovation.

If the evolution from 3G to 5G represented continuous upgrades in communication technology—innovations along the extension of Western-led classical information theory—its fundamental characteristic was driving generational progress in mobile communications through technological accumulation and resource consumption, particularly energy. In the era of artificial intelligence, big data demands ever-increasing data bandwidth, and bandwidth expansion requires resource compensation. In other words, in the 6G era, it is no longer feasible to continue along the traditional trajectory. There is an urgent need to find an "inflection point" to meet the demand for bandwidth. This inflection point technology will mark an unprecedented paradigm shift in the field of communications.

The traditional paradigm of communication theory can no longer meet the new demands of the "AI +" era. The diverse data generated by massive "device-end" equipment urgently requires more efficient communication paradigms to support it. The proposal of semantic communication has fundamentally altered the core logic of communication: it no longer pursues the mere transfer of information symbols but emphasizes task orientation, intelligent understanding, and efficient collaboration. This is not merely a technical upgrade but a fundamental reconstruction of communication theory.

The core advantage of modern semantic communication lies in its use of end-to-end intelligent learning models, enabling transmission systems to accurately understand task intent rather than mechanically transferring symbols. This approach significantly enhances communication efficiency and markedly reduces network bandwidth and energy resource consumption. This innovative thinking breaks through the traditional boundaries of communication and is widely regarded by the international academic community as the "second communication revolution."

However, such a revolution must address several challenging questions: First, what is the foundational theory supporting this revolution? Second, if this theory overturns traditional theory, how can we explain the success of traditional communication over the past century? Third, how can we explain the generalization of AI's empowerment of communication systems?

It is at such an unprecedented juncture that I hope all communication professionals will possess the unwavering determination to traverse this "uncharted territory." In the new year, let us use our research findings to provide definitive answers that eliminate future uncertainties, thereby laying the foundation for a new edifice of science and technology.

As the Year of the Horse approaches, I wish everyone immediate success, renewed pride, and a relentless pursuit of greater scientific achievements in the journey ahead.

DOI:10.12142/ZTECOM.202601001
Manuscript received: 2025-12-20
Citation: (Format 1): Zhang P. To the communications community—2026 new year's message [J]. ZTE Communications, 2026, 24(1): 1. DOI: 10.12142/ZTECOM.202601001
Citation: (Format 2): P. Zhang, "To the communications community—2026 new year's message," *ZTE Communications*, vol. 24, no. 1, pp. 1, Mar. 2026. doi: 10.12142/ZTECOM.202601001.

Biography

Zhang Ping is a Counsellor of the State Council, an Academician of the Chinese Academy of Engineering, a Professor at Beijing University of Posts and Telecommunications, China and the Director of the State Key Laboratory of Networking and Switching Technology, China. He also serves as the Editor-in-Chief of *Journal on Communications*, an IEEE Fellow, a member of the Expert Group for IMT-2020 (5G), a member of the Advisory Committee of the IMT-2030 (6G) Promotion Group, and the leader of the Innovative Research Group funded by the National Natural Science Foundation of China. He has long been dedicated to theoretical research and technological innovation in mobile communications, and has made fundamental and pioneering contributions to promoting China's independent communication technologies to become mainstream international standards. Previously, he held important academic and research leadership positions, including serving as the Chief Scientist of the National Basic Research Program of China ("973" Program) and an expert of the Theme Expert Group for the National High-Tech Research and Development Program of China ("863" Program). He has been honored with numerous prestigious national and industrial awards in recognition of his outstanding academic achievements and technological contributions, including the Special Grade Award and the First Class Award of the State Science and Technology Progress Award, three Second Class Awards of the State Technological Invention Award, and two Second Class Awards of the Science and Technology Progress Award. In addition, he has received the National Innovation Pioneer Medal, the Guanghua Engineering Science and Technology Award, and the Ho Leung Ho Lee Prize for Scientific and Technological Progress, and the research team led by him was selected into the first batch of Huang Danian-Style Teacher Teams by the Ministry of Education of China.

Guest Editorial >>>

Special Topic on Achievements of ZTE's Industry-University-Institute Cooperation Projects

Guest Editor

Xu Chengzhong

The relentless evolution of Information and Communication Technology (ICT) stands as a testament to the synergistic power of collaboration. It thrives at the dynamic intersection of industrial insight, academic rigor, and dedicated research. This special issue, "Achievements of ZTE's Industry-University-Institute Cooperation Projects," presents a curated collection of cutting-edge research that embodies the fruitful outcomes of deep collaboration, addressing some of the most pressing challenges across wireless communications, artificial intelligence (AI), software engineering, and industrial digitization.

Opening the issue, the paper "Deep CSI Compression and Feedback for Massive MIMO: A Survey" provides a comprehensive overview of deep learning-based Channel State Information (CSI) compression for massive Multiple-Input Multiple-Output (MIMO) in 5G-Advanced and future 6G systems, thereby framing a key research direction for the community.

Pushing the boundaries of physical-layer communication, the second paper "Low-Complexity OTFS Channel Equalization Based on CLU-MMSE" introduces a novel algorithm that significantly reduces the computational complexity of equalization for Orthogonal Time Frequency Space (OTFS) modulation. This work is pivotal for making OTFS, a promising waveform for high-mobility scenarios, more practical and implementable in next-generation wireless systems.

DOI:10.12142/ZTECOM.202601002
Citation: (Format 1): Xu C Z. Editorial: achievements of ZTE's industry-university-institute cooperation projects [J]. ZTE Communications, 2026, 24(1): 2–3. DOI: 10.12142/ZTECOM.202601002
Citation: (Format 2): C. Z. Xu, "Editorial: achievements of ZTE's industry-university-institute cooperation projects," *ZTE Communications*, vol. 24, no. 1, pp. 2–3, Mar. 2026. doi: 10.12142/ZTECOM.202601002.

Enhancing security at the hardware level, the third paper "Carrier Frequency Offset Based Robust Radio Frequency Fingerprint for OFDM Communication in Time-Varying Channels" proposes a robust radio frequency fingerprinting method using Carrier Frequency Offset (CFO), offering a novel layer of physical-layer authentication for securing IoT and other massive device networks.

Addressing a fundamental AI operational challenge, the fourth paper "Key Technologies for AI-Driven Network Traffic Classification Workflow and Data Distribution Shift" delves into the critical issue of model performance degradation when traffic patterns evolve. By proposing a systematic workflow and countermeasures for data distribution shift, this paper provides essential methodologies for deploying sustainable and adaptive AI in dynamic network environments.

Securing collaborative industrial ecosystems, the fifth paper "Efficient and Secure Data Storage in 5G Industrial Internet Collaborative Systems" presents a novel solution that integrates data confidentiality with attribute-based access control. This work tackles the dual challenge of protecting sensitive industrial data while enabling flexible, policy-driven data sharing among multiple entities within a 5G-enabled industrial framework.

Driving the evolution of broadband access, the sixth paper "Complexity-Reduced Equalization for 200 Gbit/s PON Downstream Systems Based on SSB Modulation and Direct Detection" presents a low-complexity equalization scheme for a 200 Gbit/s passive optical network (PON), achieving a 29 dB power budget over 20 km fiber, demonstrating its readiness for metro-access deployment.

Transforming software development practices, the seventh paper "Enhancing Code Quality with LLM in Software Static Analysis" demonstrates a pioneering application of large lan-

guage models. By integrating an AI-powered detection and patching microservice directly into the developer's workflow, this research enables a significant shift-left in software quality and security assurance, showcasing AI's role in augmenting developer productivity and code robustness.

Advancing digital human technology, the eighth paper "AED-NeRF: Audio-Driven and Emotion-Editing Dynamic Neural Radiance Fields for Expressive Talking Face Avatar" advances the state of expressive avatar generation by seamlessly integrating audio-driven lip synchronization with explicit emotion control. This work enables real-time, photorealistic talking faces whose expressions can be intuitively edited, addressing a key limitation in current virtual communication systems.

Empowering intelligent industrial inspection, the ninth paper "Steel Surface Anomaly Detection Using 3D Depth and 2D RGB Features" presents a robust multi-stage visual detection system. By effectively fusing 2D texture and 3D geometric features, this method achieves high accuracy in defect classification and localization, offering a practical and reliable AI solution for quality control in complex industrial manufacturing environments.

Innovating at the component level, the tenth paper "Synthesis and Design of Generalized Strongly Coupled Resonator Quartet Combline Filters with Redundant Resonance" contributes a novel filter synthesis theory and design. This work enables the realization of high-performance filters with desirable transmission zeros using simplified inductive coupling structures, an important advancement for the front-end hardware of compact communication devices.

Finally, rethinking system-level design, the paper "Modern Graphics APIs: Design Principles, A Use Case, and New Perspectives" provides a deep architectural analysis of the evolution of graphics APIs. Through principle elucidation and a concrete rendering engine case study, it offers valuable insights into the design trends that drive efficiency and performance in modern computing systems, a foundational concern for all computationally intensive applications.

Collectively, these contributions exemplify how targeted collaboration turns industrial challenges into research frontiers and translates academic innovation into practical solutions. They move beyond isolated theoretical pursuits, firmly grounding innovation in real-world requirements such as complexity reduction, security hardening, cost efficiency, and adaptability. The works on OTFS equalization, RF fingerprinting, and PON systems directly address the performance and economic constraints of future networks. The explorations into AI for code quality, traffic classification, and industrial inspection tackle scalability and reliability challenges in software and automation. The advancements in filters, graphics APIs, and emotional AI avatars highlight the continuous drive for better performance and more natural interfaces. We hope this special issue inspires continued and deepened collaboration across the industry-academia-research spectrum to meet the exciting challenges that lie ahead.

Biography

Xu Chengzhong received his PhD from the University of Hong Kong, China in 1993. He is Chair Professor of Computer Science at University of Macau, China. Previously, he held faculty positions at Wayne State University, USA and Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (CAS). His recent research focuses on cloud and edge for AI, autonomous driving, and intelligent transportation. Dr. Xu has authored two research monographs and over 600 journal and conference papers, garnering more than 25 000 citations and an H-index of 83. Notably, his work has been cited in 370 international patents, including 240 US patents. He is a co-inventor of more than 200 PCT and China patents and a co-founder of Shenzhen Institute of Beidou Applied Technology. His research has earned him best paper awards or nominations at conferences including SoCC'2021, HPCA'2013, HPDC'2013, and ICPP'2015. He has served on the editorial boards of several journals, including *IEEE TC*, *IEEE TCC*, *IEEE TPDS*, *JPDC*, *Science China*, and *ZTE Communications*. Dr. Xu formerly chaired IEEE Technical Committee on Distributed Processing (2015 to 2020). He is an IEEE Fellow due to contributions in resource management in parallel and distributed computing.

Deep CSI Compression and Feedback for Massive MIMO: A Survey

Lu Zhaohua[1,2], Yi Chenyang[3], Wu Jie[3], Shao Bo[3], Xu Wei[3,4]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China;
4. Purple Mountain Laboratories, Nanjing 211111, China)

DOI: 10.12142/ZTECOM.202601003

https://kns.cnki.net/kcms/detail/34.1294.TN.20260304.1138.002.html, published online March 4, 2026

Manuscript received: 2024-09-20

Abstract: To achieve the potential performance gain of massive multiple-input multiple-output (MIMO) systems, base stations (BS) require downlink channel state information (CSI) fed back by users to execute beamforming design, especially in the frequency division duplex (FDD) systems. However, due to the enormous number of antennas in massive MIMO systems, the feedback overhead of downlink CSI acquisition is extremely large. To address this issue, deep learning (DL) techniques have been introduced to develop high-accuracy feedback strategies under limited backhaul constraints. In this paper, we provide an overview of DL-based CSI compression and feedback approaches in massive MIMO systems. Specifically, we introduce the conventional CSI compression and feedback schemes and the existing problems. Besides, we elaborate on various DL techniques employed in CSI compression from the perspective of network architecture and analyze the advantages of different techniques. We also enumerate the applications of DL-based methods for solving practical challenges in CSI compression and feedback. In addition, we brief the remaining issues in deep CSI compression and indicate potential directions in future wireless networks.

Keywords: deep learning; MIMO; CSI compression; limited feedback; FDD system

Citation (Format 1): Lu Z H, Yi C Y, Wu J, et al. Deep CSI compression and feedback for massive MIMO: a survey [J]. ZTE Communications, 2026, 24(1): 4 - 15. DOI: 10.12142/ZTECOM.202601003
Citation (Format 2): Z. H. Lu, C. Y. Yi, J. Wu, et al., "Deep CSI compression and feedback for massive MIMO: a survey," *ZTE Communications*, vol. 24, no. 1, pp. 4 - 15, Mar. 2026. doi: 10.12142/ZTECOM.202601003.

1 Introduction

With the increasing demand for data traffic and massive connectivity, advanced technologies such as multiple-input multiple-output (MIMO), non-orthogonal multiple access (NOMA), and ultra-dense networks (UDN) have been proposed to meet the high-throughput, high-reliability, and low-latency challenges in 5G and beyond 5G (B5G) wireless communication networks[1-2]. Massive MIMO is considered a key technology in B5G networks since it improves the spectral and energy efficiency of wireless communication networks by simultaneously serving a set of users with multiple antennas at the base station (BS)[3]. To achieve the potential multiplexing gain in massive MIMO systems, the BS requires downlink channel state information (CSI) for transmission design, such as beamforming and power allocation, to enhance the desired signal and eliminate multi-user interference.

Since the performance of massive MIMO systems relies directly on the accuracy of the CSI obtained at the BS, it is significant to develop practical CSI acquisition methods under various scenarios[4]. In time-division duplexing (TDD) systems, downlink CSI can be obtained at the BS from the uplink CSI by utilizing channel reciprocity. In frequency-division duplexing (FDD) systems, uplink and downlink operate in different frequency bands, and the channel reciprocity no longer holds. Consequently, downlink CSI in FDD systems needs to be estimated at the users and then fed back to the BS through a feedback link. However, the feedback overhead in massive MIMO systems is prohibitively large due to the fact that the dimension of CSI increases with the network scale. Thus, there is an urgent need for effective CSI compression and feedback methods to achieve acceptable accuracy under the constraint of limited backhaul.

In recent decades, deep learning (DL) has attracted growing attention in wireless communications due to its exceptional ability in feature extraction and function approximation, which

The corresponding author is Xu Wei.
This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20240319003 and the NSFC under Grant No. 62571112..

makes it a potential methodology to address the intractable nonlinear challenges in signal processing[5]. The DL techniques are also introduced to CSI compression and feedback design in massive MIMO systems to overcome the drawbacks of conventional CSI feedback approaches such as substantial computational complexity and dependence on channel model assumptions. Given the superior performance of DL in image compression, DL-based CSI acquisition methods have the potential to learn the compression of the CSI matrix in a data-driven manner, thereby improving reconstruction accuracy and reducing feedback overhead. Furthermore, based on the domain knowledge of massive MIMO channels, model-driven DL methods can be applied to solve practical problems in CSI compression and feedback, such as network lightweighting and generalization enhancement.

In the rest of this paper, we first review the conventional CSI compression and feedback approaches in Section 2. The DL techniques deployed in CSI compression frameworks are summarized and elaborated in Section 3. Section 4 introduces the applications of DL techniques for solving practical challenges in CSI compression and feedback. The critical challenges and potential directions for CSI compression in future wireless networks are discussed in Section 5. Finally, Section 6 concludes this survey.

2 CSI Compression and Feedback for Massive MIMO

To reduce the tremendous feedback overhead in massive MIMO networks, researchers have proposed CSI compression methods utilizing channel correlations and environmental knowledge. A straightforward approach to address the challenges in feedback overhead is to feed back only statistical CSI[4]. However, this strategy achieves only satisfactory performance in limited scenarios such as slowly changing channels. To achieve the potential multiplexing gain in massive MIMO systems, it is necessary to investigate effective approaches for instantaneous CSI acquisition. In this section, we introduce the conventional CSI compression and feedback schemes based on two popular techniques, i.e., codebook-based methods and compressive sensing (CS).

2.1 Codebook-Based CSI Compression

An effective approach for instantaneous CSI acquisition with limited feedback relies on a pre-defined codebook for channel quantization. By employing a vector quantization codebook designed offline and known to both the BS and users, the users are only required to feed back the quantization index of the selected codeword in the codebook. In Ref. [6], noncoherent trellis-coded quantization (NTCQ) was proposed for channel quantization in massive MIMO. By leveraging the duality between source coding on the Grassmannian manifold and channel coding for noncoherent communication, the complexity of encoding grows linearly with the number of antennas. Moreover, codebook-based channel feedback techniques have been incorporated into wireless standards such as 3GPP LTE and IEEE 802.16m[4].

Codebook-based channel feedback also faces some technical challenges in practical implementation. Since the design of the codebook is closely related to the channel distribution, a specific design is difficult to adapt to different system scenarios. In addition, the codebook size increases exponentially with the number of antennas, and the computational expense for the look-up algorithm at the BS increases accordingly.

2.2 Compressive Sensing-Based CSI Compression

CS is a signal reconstruction framework for recovering sparse signals through sub-Nyquist sampling, which has been widely applied to signal processing in wireless communications[7]. Since the antenna arrays in massive MIMO have strong spatial correlations, the channel matrix is expected to exhibit sparsity in the spatial-frequency domain. Based on the channel sparsity assumption, CS was first applied to the design of the CSI compression and feedback scheme in Ref. [8]. Specifically, the channel matrix was estimated and compressed at the receiver via a predefined measurement matrix, which was randomly generated offline according to Gaussian distributions and known at both transmitter and receiver. Subsequently, the transmitter adopted the orthogonal matching pursuit (OMP) algorithm to recover the channel, leveraging the known measurement matrix and sparsifying bases. The two-dimensional discrete cosine transform and Karhunen-Loeve transform were employed as the sparsifying bases since they can offer a sparser representation of the signal. The dimensionality of the compressed channel was significantly reduced due to the channel sparsity, and the accuracy of recovery was acceptable with the sparsifying bases properly selected.

However, the CS-based channel compression and feedback still have some limitations in practical implementation. On the one hand, the CS-based CSI compression methods demand a channel sparsity assumption in a certain domain; this assumption may not strictly hold in practice, leading to inaccuracies in CSI recovery[9]. On the other hand, the signal reconstruction at the BS is generally solved by employing iterative algorithms such as OMP, linear programming (LP), and basis pursuit (BP)[8]. These iterative algorithms introduce substantial computational complexity and time delay, making the reconstruction process infeasible in practice. Consequently, numerous studies have turned to promising DL techniques to facilitate effective CSI acquisition under limited feedback, which will be introduced in the following sections.

3 Deep Learning Techniques for CSI Compression and Feedback

Due to its strong capability of data processing and function approximation, deep learning-based CSI acquisition is considered a promising approach to addressing the challenges of con-

ventional methods[10]. In this section, we focus on deep learning techniques for CSI compression and feedback. We first introduce the general framework of deep CSI acquisition. Then, we focus on deep CSI acquisition techniques for CSI matrices with specific correlations, i.e., the spatial correlation and temporal correlation. Finally, we analyze the computational complexity of various deep CSI acquisition methods.

3.1 General Framework of Deep CSI Compression

The autoencoder is a common framework adopted in deep CSI compression and feedback. Inspired by image processing, autoencoder-based deep CSI acquisition methods view the CSI matrix as an image. As depicted in Fig. 1, the autoencoder framework consists of an encoder and a decoder. The CSI matrix is first compressed into specific codewords by an encoder on the user side. Subsequently, the BS utilizes a decoder to reconstruct the CSI matrix from the received latent codewords, thereby facilitating efficient information feedback.

Various neural network (NN) architectures can be employed to design the autoencoder, such as convolutional neural networks (CNN)[11], long short-term memory (LSTM) networks[12], and the attention mechanism[13]. Different from images in computer vision, the CSI matrix contains inherent correlations due to the physical propagation environment. Since the performance of deep CSI compression and feedback is significantly affected by the NN architecture of autoencoders, appropriate NNs should be designed based on the specific characteristics of the CSI matrix, which will be discussed in the following two subsections.

The generative adversarial network (GAN)[14] is another effective framework of deep CSI compression and feedback, which learns the latent channel distribution to improve feedback performance. As shown in Fig. 2, GANs consist of two interlinked NNs, i.e., a generator and a discriminator, trained in tandem through an adversarial process. During the training phase, the generator produces samples mirroring the distribution of the training CSI data, whereas the discriminator aims to distinguish between authentic and synthesized samples. During the inference phase, only the generator is deployed as the decoder to reconstruct CSI matrices.

In Ref. [15], a deep convolutional generative adversarial network (DCGAN) framework was proposed to improve feedback accuracy in massive MIMO systems. Specifically, the generator of DCGAN learns to reconstruct high-quality CSI from the compressed vector, and the discriminator network evaluates the recovery quality. The GAN-based framework outperforms CS-based methods and achieves robust performance in outdoor channels. Moreover, a generative network termed PRVNet was proposed in Ref. [16] for CSI acquisition in MIMO-OFDM systems. By utilizing the generative framework of variational autoencoders, the PRVNet achieves robustness against various noise levels.

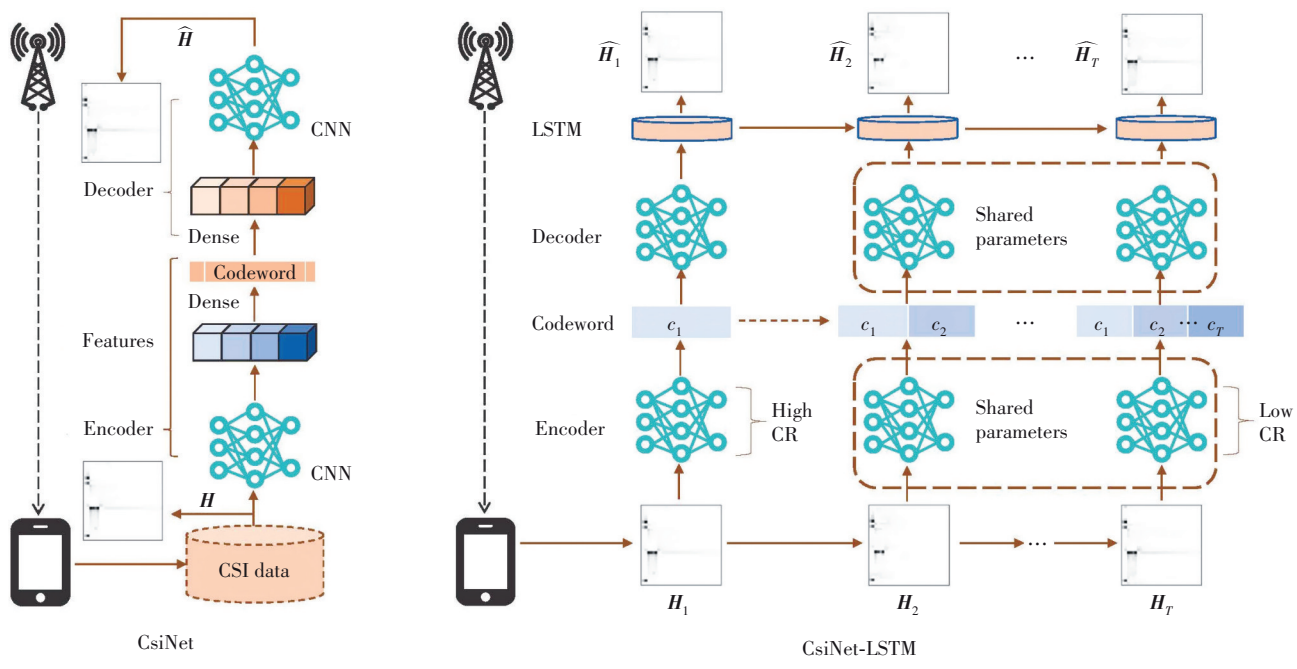


Figure 1. Illustration of autoencoder architecture

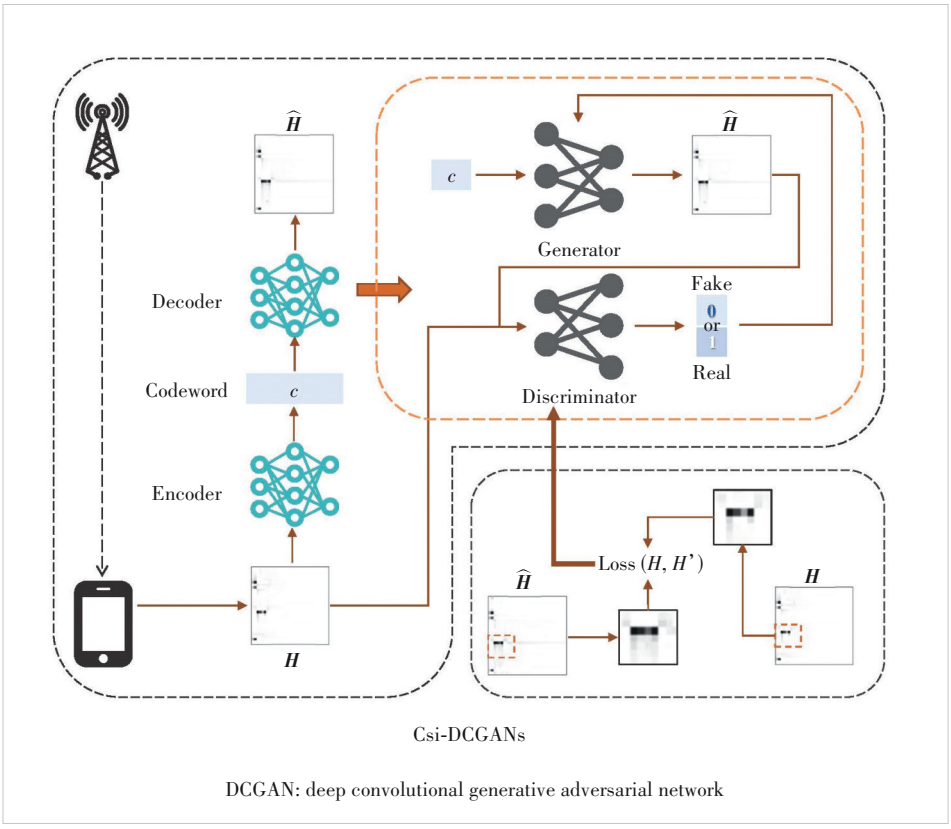


Figure 2. Illustration of generative adversarial network architecture

among transmit antennas. The encoder extracts CSI features with convolutional layers and compresses the CSI with fully connected layers, while the decoder adopts a symmetric structure and adjusts the number of layers and neurons. The reconstruction accuracy of CsiNet is significantly higher than that of the CS-based methods. CsiNet merely considers the CSI feedback in MIMO systems with a single user. However, in multiuser massive MIMO systems, the correlations among CSI matrices of nearby users can be exploited to improve the feedback performance. In Ref. [18], an autoencoder, termed DeepCMC, with fully convolutional layers, was proposed for CSI feedback in multiuser MIMO systems. In DeepCMC, the encoders are distributively deployed across the users, while the decoder at the BS jointly reconstructs the multiuser CSI. The decoder consists of separate decoder branches for users and combining kernels to fuse the side information of users. DeepCMC outperforms CsiNet by exploiting the correlations among users.

3.2 Spatially Correlated CSI Compression

The CSI matrix in the space-frequency domain contains inherent spatial correlations, such as the correlation across BS antennas and the correlation among users. CNN is a suitable NN architecture for CSI acquisition, as it learns the spatial correlation of CSI matrices through convolutional operations. Specifically, in each convolutional layer of CNN, convolutional kernels are used to perform element-wise multiplication and addition with the input data to capture local features. By sliding the convolutional kernels over the input data, an output feature map with detected features is generated. The mathematical representation of the convolutional layer is formulated as:

$$\boldsymbol{Y}_{c,i,j} = \sum_{m=1}^{C_{in}} \sum_{p=1}^{F} \sum_{q=1}^{F} \boldsymbol{X}_{m,i+p,j+q} \cdot \boldsymbol{W}_{c,m,p,q} + b_c \tag{1},$$

where $\boldsymbol{X} \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}$ is the input data, $\boldsymbol{Y} \in \mathbb{R}^{C_{out} \times H_{out} \times W_{out}}$ is the output feature map, $\boldsymbol{W}_c \in \mathbb{R}^{F \times F}$ is the c-th convolutional kernel and $b_c \in \mathbb{R}$ is the corresponding bias. By stacking several convolutional layers, the CNN can adjust the receptive field to learn the spatial local correlation of the CSI matrix.

The CNN architecture was first applied to CSI compression and feedback in Ref. [17], where a CNN-based autoencoder, termed CsiNet, was proposed to learn the spatial correlations

3.3 Temporally Correlated CSI Compression

In time-varying channels, the temporal correlations should be considered in CSI acquisition. LSTM is capable of handling long-term dependencies in sequential data due to its gated mechanisms, making it suitable for extracting temporal correlations to improve CSI feedback. Specifically, LSTM processes sequential data by stacking a series of LSTM cells, which is formulated as:

$$\begin{aligned}
\boldsymbol{i}_t &= \sigma\left(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i\right) \\
\boldsymbol{f}_t &= \sigma\left(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f\right) \\
\boldsymbol{o}_t &= \sigma\left(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o\right) \\
\boldsymbol{C}_t &= \boldsymbol{f}_t \odot \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \odot \tanh\left(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c\right) \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{C}_t\right)
\end{aligned} \tag{2},$$

where $\boldsymbol{i}_t$ is the input gate that decides what information should be added to the cell state, $\boldsymbol{f}_t$ is the forget gate that decides what information should be discarded from the cell state, $\boldsymbol{o}_t$ is the output gate that decides what information should be output from the cell state. Additionally, $\boldsymbol{C}_t$ represents the memory cell, $\boldsymbol{h}_t$ denotes the hidden vector, and t signifies the time

step. By maintaining and updating the cell state through these gates, LSTMs can effectively handle long-term dependencies of input data.

CsiNet-LSTM was proposed in Ref. [19] for CSI acquisition in time-varying massive MIMO systems. In CsiNet-LSTM, CNN-based encoders are deployed at the user side, while the decoders at the BS are composed of CNNs and LSTMs to capture the spatial and temporal features of CSI matrices. The reconstruction accuracy of CsiNet-LSTM is higher than that of CsiNet due to the temporal correlation extraction. Advancing from CsiNet-LSTM, ConvlstmCsiNet was proposed in Ref. [20] for further improvement in reconstruction quality. By adopting pseudo-3D blocks to maintain the independence of temporal and spatial features, ConvlstmCsiNet achieves remarkable feedback accuracy and robustness at low compression ratios.

Although LSTM-based deep CSI feedback methods effectively extract the temporal correlation features of CSI matrices, they ignore the weight assignment of CSI features. An attention mechanism can be deployed in CSI feedback networks to assign more weight to dominant features, thereby enhancing the representation of temporal features and improving the performance of CSI reconstruction. Specifically, the attention mechanism generates a set of weight factors to describe the importance of features, and allocates more weights to the feature maps with more information. The attention weights are calculated as:

$$\alpha_{ij} = \frac{\exp\left(f\left(\boldsymbol{s}_i, \boldsymbol{s}_j\right)\right)}{\sum_k \exp\left(f\left(\boldsymbol{s}_i, \boldsymbol{s}_k\right)\right)} \tag{3},$$

where $\boldsymbol{s}_i \in \mathbb{R}^d$ is the input feature, and f is the alignment model operated by dense layers. Therefore, the output of the attention layer is:

$$\mathbf{c}_i = \sum_j \alpha_{ij} \mathbf{s}_j \tag{4},$$

where $i, j = 1, \cdots, T$ denotes the length of the input sequence.

An LSTM-attention network was proposed in Ref. [21] to improve CSI feedback accuracy by leveraging the attention mechanism to fine-tune the temporal features of CSI data. The autoencoder first deploys LSTM units to exploit the temporal correlation of massive MIMO channels, and subsequently incorporates the attention mechanism to prioritize and weight feature importance. The CSI recovery accuracy is significantly improved by adopting the attention mechanism. Moreover, a CNN-LSTM-A network was proposed in Ref. [22] for CSI feedback in MIMO systems with high user mobility, where the attention mechanism is introduced to assign more weights to dominant features in each time step.

3.4 Computational Complexity Analysis

In this subsection, we evaluate the computational complexity of deep CSI compression and feedback methods in the training and inference stages. Since the complexity of the training stage is mainly influenced by the backpropagation process for updating trainable parameters, we assess the complexity of the training stage based on the parameter size of NN models, which is referred to as space complexity (SC). Meanwhile, the complexity of the inference stage is evaluated by the number of floating-point operations (FLOPs), referred to as time complexity (TC).

We consider the downlink of an FDD massive MIMO-OFDM system with N_t antennas at the BS and a single-antenna user. The system adopts OFDM transmission with N_c subcarriers. Since the complexity of deep CSI feedback models is primarily dominated by the convolutional layers and dense layers, we analyze the time and space complexity of these layers, respectively. According to Refs. [23 – 24], the time and space complexity of convolutional layers are

$$TC_C = O\left(\sum_{l=1}^{L} W_l H_l C_{l-1} C_l F_l^2\right) \tag{5},$$

$$SC_C = O\left(\sum_{l=1}^{L} C_{l-1} C_l F_l^2\right) \tag{6},$$

where C_l is the number of channels in layer l, and F_l is the convolution kernel size in layer l. In deep CSI feedback models, the width and height of input feature satisfies $W_l H_l = 2N_t N_c, \forall l$. Thus, the time and space complexity of convolutional layers in CSI feedback NN models are

$$\mathrm{TC}_C^{\mathrm{CSI}} = 4N_t N_c \sum_{l=1}^{L} C_{l-1} C_l F_l^2 \tag{7},$$

$$\mathrm{SC}_C^{\mathrm{CSI}} = \sum_{l=1}^{L} C_{l-1} C_l F_l^2 \tag{8},$$

where the time complexity consists of the number of multiplication and addition in convolution operations, each occupying $2N_t N_c \sum_{l=1}^{L} C_{l-1} C_l F_l^2$ FLOPs. The space complexity indicates the parameter size of kernels.

Similarly, the time and space complexity of dense layers are

$$\mathrm{TC}_D = O\left(\sum_{l=1}^{L} N_{l-1} N_l\right) \tag{9},$$

$$\mathrm{SC}_D = O\left(\sum_{l=1}^{L} N_{l-1} N_l + N_l\right) \tag{10},$$

where N_l denotes the feature dimension in layer l. In deep CSI feedback models, the input and output features are $N_0 = 2N_t N_c, N_L = 2N_t N_c \gamma$ at the encoder and $N_0 = 2N_t N_c \gamma, N_L = 2N_t N_c$ at the decoder, where γ is the compression ratio. Thus, the time and space complexity of dense layers in CSI feedback NN models are

$$\mathrm{TC}_D^{\mathrm{CSI}} = 4N_tN_c(1+\gamma)(N_1+N_{L-1}) + 4\sum_{l=2}^{L-1}N_{l-1}N_l \quad (11),$$

$$\mathrm{SC}_D^{\mathrm{CSI}} = 2N_tN_c(1+\gamma)(N_1+N_{L-1}+1) + 2\sum_{l=2}^{L-1}(N_{l-1}N_l+N_l) + N_1 + N_{L-1} \quad (12),$$

where the space complexity indicates the parameter size of weights and biases.

The time and space complexity of various feedback NNs are shown in Table 1. Since the feature dimension is related to the dimension of the CSI matrix and the feedback length, the computational complexity increases with the compression ratio. Moreover, the complexity of convolutional layers is generally lower than that of dense layers. Since LSTM contains more dense layers, the complexity of NN in Ref. [17] is lower than that in Refs. [20] and [21].

Table 1. Computational complexity of deep CSI feedback models

Complexity		1/4	1/8	1/16	1/32
Time complexity	Ref. [17]	21 659 648	5 668 864	3 571 712	2 523 136
	Ref. [20]	121 708 544	97 591 296	86 319 104	80 879 616
	Ref. [21]	-	-	-	-
Space complexity	Ref. [17]	2 103 904	1 055 072	530 656	268 448
	Ref. [20]	28 326 904	22 296 312	19 477 624	18 117 432
	Ref. [21]	10 247 148	-	7 484 688	7 024 272

4 Applications of Deep CSI Compression and Feedback

In the 5G R18 standard[25], various methods have been adopted to improve the CSI compression and feedback, including the historical CSI-based prediction methods that utilize historical CSI data to predict future CSI, non-AI/ML prediction methods such as filters and interpolation algorithms, and the advanced prediction models utilizing AI/ML technologies. Current 5G communication protocols focus on enhancing the accuracy and real-time performance of CSI feedback through DL techniques, optimizing the transmission efficiency of 5G networks. DL techniques have been widely applied in 5G communications. In this section, we discuss the applications of DL techniques for solving practical challenges in CSI compression and feedback designs. Specifically, we emphasize the representative research achievements in network lightweighting, reconstruction performance improvement, CSI generalization enhancement, and the joint design of CSI compression, feedback, and precoding. These approaches are significant to the practical implementation of CSI acquisition, as they reduce deployment costs and enhance compression efficiency.

4.1 Network Lightweighting

Network lightweighting is crucial to the practical deployment of DL-based CSI compression and feedback networks, aiming to reduce the network size deployed on both BS and users, thereby saving hardware costs. To date, numerous effective network lightweighting methods have been proposed and applied in the CSI compression and feedback of MIMO systems.

Due to more stringent computational and memory constraints on the users than on the BS, the primary objective of network lightweighting methods is to reduce the size of the encoder network at the user end. The most common approach involves designing innovative convolutional structures to decrease the number of parameters at the encoder[26]. Typical lightweight structures include multi-branch convolutions, dimensionality reduction sampling of CSI feature maps, etc.

On the other hand, the inherent characteristics of CSI are also leveraged for the implementation of network lightweighting. The observed similarity in the probability distributions of the real and imaginary parts of the CSI matrix enabled a method where only the real part of the CSI matrix is inputted into the network for training[27]. Subsequently, the trained network was reused for the compression and feedback of the imaginary part. This approach, without compromising performance, effectively reduces the network parameters by approximately half.

Furthermore,the real and imaginary parts of CSI matrix also carry inherent physical information. This characteristic has been utilized in studies focusing on network lightweighting. One research approach in Ref. [28] involved transforming a real-valued NN designed for lightweight purposes into a complex-valued NN, achieving equivalent network performance with fewer parameters required. Another research in Ref. [29] focused on the design of a pseudo-complex-valued input layer, while retaining the real-valued NN. This approach allows the input CSI matrix to undergo equivalent complex-valued operations, thereby reducing the computational overhead by 24%.

4.2 Performance Improvement

In the context of DL-based CSI compression and feedback, enhancing the efficiency of CSI acquisition represents a paramount research direction. Extracting the inherent features of the CSI matrix to improve the compression performance from a physical perspective is considered a highly promising research avenue.

Researchers initially focused on the sparsity characteristics of the CSI matrix, which vary with different channel scenarios and compression rates. According to existing DL-based theory, dense images are more aptly processed using convolu-

tional kernels of smaller size for feature extraction, while sparse images benefit from larger convolutional kernels. To enable CSI encoders and decoders to effectively extract features of CSI in various scenarios, a multi-path parallel convolutional structure has been proposed and applied to both[30 - 31]. By employing parallel convolutional layers with different kernel sizes to extract CSI features, this architecture significantly enhances the efficiency of CSI compression and feedback across diverse scenarios and compression rates.

In the realm of DL for image compression, a series of efficient techniques have been developed by investigating the structural features of images. In CSI compression and feedback, the CSI matrix is often viewed as an image, thereby enabling the utilization of image characteristics of the CSI matrix to enhance compression efficiency. The CSI image can be divided into many small blocks, where some blocks contain a high level of self-information and the image features within are referred to as shape features; conversely, blocks with less self-information are characterized by their texture information. By preserving blocks with high self-information shape features and discarding those with low self-information texture features during compression, researchers have achieved efficient CSI compression and feedback[32 - 33]. Distinct from black-box neural networks, this method incorporates CSI prior knowledge and significantly reduces the complexity of the encoder network.

The previous researchers designed NNs from the perspective of the image features of CSI. In contrast, other researchers aim to design CSI compression and feedback networks based on the extraction of physical CSI features. In Ref. [34], the authors discovered that the line-of-sight (LoS) propagation path characteristics and non-line-of-sight (NLoS) path features can be effectively extracted by different NNs. Consequently, the authors employed a dual-feature fusion NN that combines a CNN with an attention enhancement network structure to achieve improved compression performance. In Ref. [35], the authors considered the similarity of the CSI matrix across different polarization directions caused by dual-polarized antennas and introduced a decoupled representation learning method. This method reduces the redundant information shared across different polarization directions of CSI, thereby enhancing compression performance.

4.3 Generalization Enhancement

To achieve satisfactory CSI feedback performance, DL-based approaches require a substantial amount of CSI training data, which is exceedingly costly in real-world scenarios. Furthermore, when the channel environment changes, DL networks trained under different channel conditions cannot be applied to new channel environments. This necessitates CSI data re-collection and network re-training, thereby significantly increasing the deployment cost of NN applications. Hence, generalization enhancement is a valuable direction for research.

To address the issues of insufficient training samples and the maladaptation of NNs to new channel environments, direct transfer learning and meta-learning, two methods of deep transfer learning, have been introduced into the CSI compression and feedback domain for generalization enhancement[36 - 38]. Fig. 3 presents a schematic illustration of employing transfer learning methods for CSI compression and feedback. Both direct transfer learning and meta-learning are effective strategies for addressing model generalization and adaptability issues. They achieve this by utilizing the transfer of existing knowledge and adopting a "learning to learn" strategy to rapidly adapt to new tasks, respectively. In Refs. [36 - 38], deep transfer networks employing direct transfer and

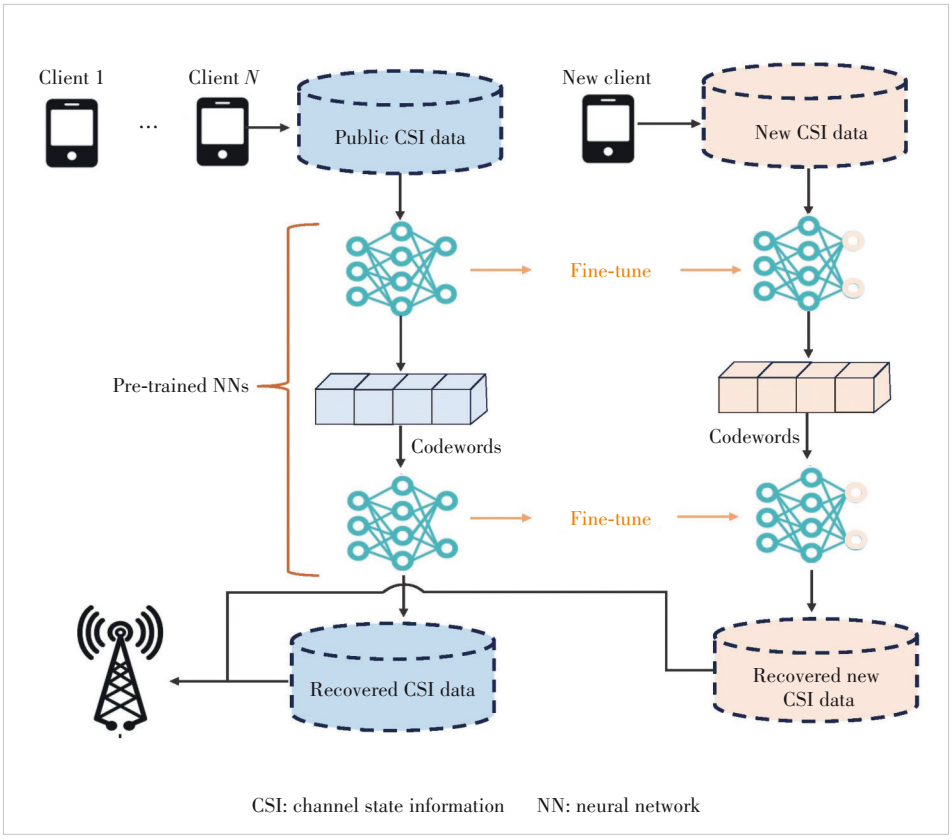


Figure 3. Transfer learning model for CSI feedback

meta-learning methods required only a minimal amount of training data to achieve satisfactory CSI compression and feedback performance. Moreover, through fine-tuning, these networks can quickly adapt to new channel environments, significantly reducing training overhead. Specifically, in Ref. [36], the downlink channel prediction was formulated as a deep transfer learning (DTL) problem, and a direct transfer algorithm based on a fully-connected NN architecture was proposed. The authors also designed a meta-learning algorithm that trains the network by alternately performing intra-task and inter-task updates, which is then adapted to new environments using a small amount of labeled data. Simulation results show that, compared with methods without transfer learning, the direct transfer algorithm and the meta-learning algorithm can improve downlink CSI prediction accuracy by up to 50%. In Refs. [37] and [38], the authors proposed a model-agnostic meta-learning (MAML) approach to address the issue of needing a large number of wireless channel environment samples for training deep neural networks (DNNs) as pre-trained models. By fine-tuning the pre-trained model with a relatively small number of samples, they achieved CSI feedback models for different wireless channel environments at a lower training cost. Simulation results show that the MAML method achieves 7 – 8 dB gains in CSI feedback NMSE compared with the deep transfer learning under different channel models.

Transfer learning methods can also be applied to the reconstruction of MIMO channels, enabling the acquisition of sufficient CSI training data without the need for extensive measurement and labeling of actual channels. Ref. [39] utilized untrained neural networks (UNN) to obtain a large amount of equivalent wireless channel data through minimal measurements (e.g., only a few time snapshots). In this transfer learning training, the UNN acquired prior knowledge about the propagation environment, thereby enabling the reconstruction of wireless channels.

Moreover, transfer learning techniques can also address the inflexibility issue in compression ratios within DL-based CSI compression and feedback methods[40]. In real-world communication scenarios, the channel conditions between the BS and users are constantly changing, necessitating adjustment to the compression ratio. Consequently, the users and BS need to store network parameters for various compression rates, which increases the hardware overhead. By employing transfer learning approaches, this overhead can be saved and the CSI compression performance can be enhanced.

Finally, Ref. [41] addressed the issue of data silos and online training adaptation caused by offline NN training, and proposed a DL approach based on interactive federated and transfer learning (IFTL). This method enables downlink CSI prediction and online update capabilities. The transfer learning approach takes into account various factors, including the asynchrony of different clients, and achieves good performance in the channel environments of different cells. These research findings collectively underscore the potential and promising prospects of transfer learning methods in CSI compression and feedback, making it a promising direction for future research.

4.4 Joint Design of CSI Compression, Feedback, and Precoding

In wireless communication, the purpose of CSI compression and feedback is to facilitate more efficient design of the precoding matrix, thereby enhancing the communication rate in millimeter-wave MIMO (mMIMO) systems. Hence, merely improving the accuracy of CSI compression and feedback without considering precoding matrix design does not guarantee optimal communication performance. To address this, researchers have integrated pilot estimation, CSI compression and feedback, and precoding matrix design into a unified process, optimizing communication performance from the perspective of communication rates. This approach is more practically meaningful than efforts focused solely on achieving higher CSI reconstruction accuracy, as it addresses the overall effectiveness issue in communication systems.

Ref. [42] presented a joint framework for pilot design, CSI compression and feedback, and precoding design in downlink multi-user massive MIMO systems based on DL. The authors observed that this joint design problem could be modeled as a distributed source coding issue. Specifically, the pilot design employed an unbiased single-layer fully connected network for equivalence, with the network's weights serving as the pilot sequences to be optimized. The pilot was input into the CSI compression and feedback network, which output quantized codewords. This network effectively accomplished three tasks: channel estimation, channel compression, and quantization of the compressed codewords. Finally, the quantized codewords were input into a precoding network, which output the precoding matrix optimized by the NN. Simulation results indicate that this approach closely matches the performance of traditional precoding schemes with perfect CSI while requiring significantly less pilot and codeword feedback overhead. Ref. [43] proposed a similar joint training framework while introducing a training strategy distinct from that in Ref. [42]. This approach can circumvent the need for retraining across different network scales intended for scalable designs, thereby reducing training overhead.

Furthermore, to address the need for flexible pilot adaptation due to the limited saturation levels of amplifiers in massive MIMO systems, Ref. [44] devised a DL-based closed-loop massive MIMO system joint optimization scheme. In this scheme, the authors represented the process of generating the beamforming (BF) matrix as a functional optimization problem. The functions of adaptive pilot length, CSI compression, and BF were all substituted by NNs, where each functional block learned the optimal strategy to maximize network utility during the training process.

In real-world communication scenarios, signaling overhead and CSI mismatches can arise due to transmission delays. To address this issue, Ref. [45] developed a dual-timescale DNN architecture composed of long-term and short-term DNNs. The analog precoder, designed by the long-term DNN based on CSI statistical information, was updated once over multiple time slots. In contrast, the digital precoder was optimized at each time slot by the short-term DNN based on a low-dimensional equivalent CSI matrix. Moreover, a corresponding dual-timescale training method was developed, which achieved lower bit error rates and pilot overhead reduction.

To summarize, the main contributions of the above papers are listed in Table 2.

5 Technical Challenges and Future Directions

Although various research has been proposed to solve the practical problems in massive MIMO CSI acquisition, there are still some open issues remaining to be investigated. In the following content, we enumerate the key challenges and future directions of DL-based CSI acquisition in future wireless networks.

5.1 Technical Challenges

The major challenge of DL-based CSI feedback lies in the enormous number of training samples required in the training process. Most existing works utilize the true CSI in massive MIMO downlinks as the training label and learn to minimize the error between true CSI and reconstructed CSI. However, CSI estimation in massive MIMO downlinks incurs overwhelming overhead, as the number of orthogonal pilots increases linearly with the number of antennas. This makes the training phase of autoencoders for CSI compression and reconstruction expensive and time-consuming. Furthermore, the testing dataset in practical implementation often exhibits domain discrepancies with training datasets due to time delay, which may lead to performance degradation[46]. Consequently, it is necessary and challenging to obtain the appropriate training dataset at an acceptable cost.

The cooperation between the BSs and users in DL-based CSI feedback design also brings up various challenges. Since most existing DL-based designs employ end-to-end learning of the encoder and decoder, the users and BSs are required to exchange compressed CSI and calculated gradients, respectively, leading to tremendous signaling overhead. Moreover, the network training process demands huge computational and storage resources, which are unaffordable for the users. Meanwhile, some studies investigate novel DL architectures to deploy end-to-end training on the BS[47]. In this case, the encoder is obtained at the BS and thus brings up extra problems such as the intellectual property among manufacturers.

Another critical issue is enhancing the generalization capability of DL-based CSI feedback models. Most existing designs focus on autoencoder architectures over a specific channel distribution or have limited scalability towards some specific factors, which may suffer from severe performance degradation in real-time implementations, such as high mobility scenarios[48]. In addition, it is infeasible to train diverse models for different channel scenarios due to the high training cost. Therefore, the generalization capability of DL-based CSI feedback models is significant in practical use and remains a crucial challenge for future investigation.

5.2 Future Directions

The artificial intelligence-native air interface (AI-AI) is a disruptive framework that incorporates conventional signal processing modules by deploying AI models to the air interface, which is considered a promising evolution of the network design in the B5G and 6G systems[49 – 50]. Different from the existing systems decoupling the source coding, channel coding, and data transmission into a block-to-block architecture, the goal of the AI-native air interface is to initially build an AI-based communication framework considering the impact of

Table 2. Summary of recent papers on deep CSI feedback

Advantages	Key Techniques	References
Network lightweighting	Design an innovative structure of multi-branch convolutions	Ref. [26]
	Exploit the similarity of real and imaginary parts of CSI	Refs. [27 – 29]
Performance improvement	Exploit the sparse characteristics of CSI	Refs. [30 – 31]
	Exploit the image characteristics of the CSI matrix	Refs. [32 – 33]
	Extract CSI features based on physical propagation environment	Refs. [34 – 35]
Generalization enhancement	Adopt model-agnostic meta-learning approaches	Refs. [36 – 38]
	Adopt deep transfer learning techniques	Refs. [39 – 40]
	Adopt interactive federated and transfer learning	Ref. [41]
End-to-end design	Joint framework of pilot design, CSI feedback, and precoding	Refs. [42 – 43]
	Consider adaptive pilot length for mm-wave MIMO systems	Ref. [44]
	Design a dual-timescale network to reduce signaling overhead	Ref. [45]

CSI: channel state information MIMO: multiple-input multiple-output

hardware imperfections and radio environment. According to 3GPP Release 18, CSI feedback is included as one of the three representative specific use cases in AI-native air interface[51]. The key challenge of designing deep CSI feedback models for the AI-native air interface is the model generalization capability over scenarios and configurations. Training dataset mixing and online learning are two potential approaches to this challenge[52]. By mixing the CSI samples generated with different channel models, a training dataset that covers various channel distributions can be formed, thereby improving the generalization ability of the model. Online learning is required when new channel distribution occurs. Moreover, advanced learning-based techniques such as transfer learning and meta-learning could be deployed to accelerate online learning and reduce training overhead.

Terahertz (THz)-band communication is regarded as a crucial technique to support the increasing demand for communication capacity and bandwidth in future wireless mobile communications. To alleviate the high propagation loss and power limitation in THz communications, densely packed nano-antenna arrays are employed to construct ultra-massive MIMO (UM-MIMO) systems[53]. However, CSI acquisition in UM-MIMO systems is more challenging than that in massive MIMO due to the expanding number of antennas, beam squint, and hybrid-field effects[54]. Moreover, training labels for DL-based CSI feedback design are more difficult to acquire. Model-driven deep learning is a promising approach to these challenges. For example, deep unfolding is a model-driven technique that unfolds iterative algorithms into a layer-wise neural network[55]. Since the CSI reconstruction at the BS is generally achieved via iterative algorithms, deep unfolding can be adopted for CSI feedback design in UM-MIMO systems. Furthermore, unfolding-based DL networks rely on the architecture of the underlying iterative algorithm and incorporate inherent domain knowledge. Thus, the number of trainable parameters in unfolding-based DL networks is considerably lower than that of black-box DNNs, thereby reducing training overhead.

Semantic communication is a novel framework that takes into account the meaning of transmission messages in signal processing designs[56–57]. With the increasing demand for content-based services in 5G and beyond, semantic communication is considered a promising technique to meet the tremendous requirements by exploiting the semantic aspects of communication not included in Shannon's information theory. Semantic communication-based deep CSI feedback is a potential approach to mitigating feedback overhead, while it faces critical challenges in semantic expression modeling. To address this issue, a task-oriented deep CSI feedback framework can be employed. For example, in a data hiding-based CSI feedback system where downlink CSI is hidden within the transmitted images, the autoencoder architecture used for image compression can be adopted to design the CSI feedback network[58]. Moreover, in a precoding-oriented CSI feedback system where the BS aims to design multiuser precoding vectors, an end-to-end DNN architecture can be employed, and a specific loss function can be designed to strike a trade-off between sum-rate performance and feedback overhead[59].

6 Conclusions

In this paper, we provide a comprehensive overview of DL-based CSI compression and feedback techniques in massive MIMO systems. We focus on the critical challenges in conventional CSI acquisition approaches, such as quantized codebooks and compressive sensing. Specifically, we analyze the advantages of various DL techniques applied to CSI compression, including CNN, LSTM, GAN, and attention mechanisms. The applications of DL-based methods for solving practical challenges in CSI compression and feedback such as network lightweighting and generalization enhancement are also discussed. Finally, we emphasize the existing critical challenges and promising future directions.

References

[1] Xu W, Yang Z H, Ng D W K, et al. Edge learning for B5G networks with distributed signal processing: semantic communication, edge computing, and wireless sensing [J]. IEEE journal of selected topics in signal processing, 2023, 17(1): 9–39. DOI: 10.1109/JSTSP.2023.3239189
[2] He M, Li X, Ni J. Physical layer security for mmWave communications: challenges and solutions [J]. ZTE communications, 2022, 20(4): 41–51. DOI: 10.12142/ZTECOM.202204006
[3] Lu L, Li G Y, Swindlehurst A L, et al. An overview of massive MIMO: benefits and challenges [J]. IEEE journal of selected topics in signal processing, 2014, 8(5): 742–758. DOI: 10.1109/JSTSP.2014.2317671
[4] Love D J, Heath R W, Lau V K N, et al. An overview of limited feedback in wireless communication systems [J]. IEEE journal on selected areas in communications, 2008, 26(8): 1341–1365. DOI: 10.1109/JSAC.2008.081002
[5] Zhao M K, Huang Y, Li X. Federated learning for 6G: a survey from perspective of integrated sensing, communication and computation [J]. ZTE communications, 2023, 21(2): 25–33. DOI: 10.12142/ZTECOM.202302005
[6] Choi J, Chance Z, Love D J, et al. Noncoherent trellis coded quantization: a practical limited feedback technique for massive MIMO systems [J]. IEEE transactions on communications, 2013, 61(12): 5016–5029. DOI: 10.1109/TCOMM.2013.111413.130379
[7] Gao Z, Dai L L, Han S F, et al. Compressive sensing techniques for next-generation wireless communications [J]. IEEE wireless communications, 2018, 25(3): 144–153. DOI: 10.1109/MWC.2017.1700147
[8] Kuo P H, Kung H T, Ting P G. Compressive sensing based channel feedback protocols for spatially-correlated massive antenna arrays [C]//IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2012: 492–497. DOI: 10.1109/WCNC.2012.6214417
[9] Liu Z Y, Zhang L, Ding Z. An efficient deep learning framework for low rate massive MIMO CSI reporting [J]. IEEE transactions on communications, 2020, 68(8): 4761–4772. DOI: 10.1109/TCOMM.2020.2993626
[10] Wang T Q, Wen C K, Wang H Q, et al. Deep learning for wireless physical

layer: opportunities and challenges [J]. China communications, 2017, 14 (11): 92 – 111. DOI: 10.1109 /CC.2017.8233654
[11] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278 – 2324. DOI: 10.1109/5.726791
[12] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735 – 1780. DOI: 10.1162/neco.1997.9.8.1735
[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [EB/OL]. [2024-08-12]. https://proceedings. neurips. cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[14] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: an overview [J]. IEEE signal processing magazine, 2018, 35(1): 53 – 65. DOI: 10.1109/MSP.2017.2765202
[15] Tolba B, Elsabrouty M, Abdu-Aguye M G, et al. Massive MIMO CSI feedback based on generative adversarial network [J]. IEEE communications letters, 2020, 24(12): 2805 – 2808. DOI: 10.1109/LCOMM.2020.3017188
[16] Hussien M, Nguyen K K, Cheriet M. PRVNet: a novel partially-regularized variational autoencoders for massive MIMO CSI feedback [C]// Wireless Communications and Networking Conference (WCNC). IEEE, 2022: 2286 – 2291. DOI: 10.1109/WCNC51071.2022.9771642
[17] Wen C K, Shih W T, Jin S. Deep learning for massive MIMO CSI feedback [J]. IEEE wireless communications letters, 2018, 7(5): 748 – 751. DOI: 10.1109/LWC.2018.2818160
[18] Mashhadi M B, Yang Q Q, Gündüz D. Distributed deep convolutional compression for massive MIMO CSI feedback [J]. IEEE transactions on wireless communications, 2021, 20(4): 2621 – 2633. DOI: 10.1109/TWC.2020.3043502
[19] Wang T Q, Wen C K, Jin S, et al. Deep learning-based CSI feedback approach for time-varying massive MIMO channels [J]. IEEE wireless communications letters, 2019, 8(2): 416 – 419. DOI: 10.1109/LWC.2018.2874264
[20] Li X Y, Wu H M. Spatio-temporal representation with deep neural recurrent network in MIMO CSI feedback [J]. IEEE wireless communications letters, 2020, 9(5): 653 – 657. DOI: 10.1109/LWC.2020.2964550
[21] Li Q, Zhang A H, Liu P C, et al. A novel CSI feedback approach for massive MIMO using LSTM-attention CNN [J]. IEEE access, 2020, 8: 7295 – 7302. DOI: 10.1109/ACCESS.2020.2963896
[22] Zhang Z F, Zheng Y, Gan C Q, et al. Massive MIMO CSI reconstruction using CNN-LSTM and attention mechanism [J]. IET communications, 2020, 14(18): 3089 – 3094. DOI: 10.1049/iet-com.2019.1030
[23] Xia W C, Zheng G, Zhu Y X, et al. A deep learning framework for optimization of MISO downlink beamforming [J]. IEEE transactions on communications, 2020, 68(3): 1866 – 1880. DOI: 10.1109/TCOMM.2019.2960361
[24] He K M, Sun J. Convolutional neural networks at constrained time cost [C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 5353 – 5360. DOI: 10.1109/CVPR.2015.7299173
[25] 3GPP TR 38.843. Study on artificial intelligence (AI)/machine learning (ML) for NR air interface (Release 18) [S]. 2023
[26] Cao Z, Shih W T, Guo J J, et al. Lightweight convolutional neural networks for CSI feedback in massive MIMO [J]. IEEE communications letters, 2021, 25(8): 2624 – 2628. DOI: 10.1109/LCOMM.2021.3076504
[27] Sun Y Y, Xu W, Liang L, et al. A lightweight deep network for efficient CSI feedback in massive MIMO systems [J]. IEEE wireless communications letters, 2021, 10(8): 1840 – 1844. DOI: 10.1109/LWC.2021.3083331
[28] Li H Z, Zhang B Y, Chang H R, et al. CVLNet: a complex-valued lightweight network for CSI feedback [J]. IEEE wireless communications letters, 2022, 11(5): 1092 – 1096. DOI: 10.1109/LWC.2022.3157263
[29] Ji S J, Li M. CLNet: complex input lightweight neural network designed for massive MIMO CSI feedback [J]. IEEE wireless communications letters, 2021, 10(10): 2318 – 2322. DOI: 10.1109/LWC.2021.3100493
[30] Lu Z L, Wang J T, Song J. Multi-resolution CSI feedback with deep learning in massive MIMO system [C]//International Conference on Communications (ICC). IEEE, 2020: 1 – 6. DOI: 10.1109/icc40277.2020.9149229
[31] Lu Z L, Zhang X D, He H Y, et al. Binarized aggregated network with quantization: flexible deep learning deployment for CSI feedback in massive MIMO systems [J]. IEEE transactions on wireless communications, 2022, 21(7): 5514 – 5525. DOI: 10.1109/TWC.2022.3141653
[32] Yin Z Q, Xie R J, Xu W, et al. Self-information domain-based neural CSI compression with feature coupling [J]. IEEE transactions on vehicular technology, 2023, 72(10): 13661 – 13665. DOI: 10.1109/TVT.2023.3272560
[33] Yin Z Q, Xu W, Xie R J, et al. Deep CSI compression for massive MIMO: a self-information model-driven neural network [J]. IEEE transactions on wireless communications, 2022, 21(10): 8872 – 8886. DOI: 10.1109/TWC.2022.3170576
[34] Zhang S Q, Xu W, Jin S, et al. Dual-propagation-feature fusion enhanced neural CSI compression for massive MIMO [J]. IEEE transactions on communications, 2023, 71(9): 5182 – 5198. DOI: 10.1109/TCOMM.2023.3282227
[35] Fan S H, Xu W, Xie R J, et al. Deep CSI compression for dual-polarized massive MIMO channels with disentangled representation learning [J]. IEEE transactions on communications, 2024, 72(9): 5564 – 5580. DOI: 10.1109/TCOMM.2024.3384256
[36] Yang Y W, Gao F F, Zhong Z M, et al. Deep transfer learning-based downlink channel prediction for FDD massive MIMO systems [J]. IEEE transactions on communications, 2020, 68(12): 7485 – 7497. DOI: 10.1109/TCOMM.2020.3019077
[37] Zeng J, Sun J L, Gui G, et al. Downlink CSI feedback algorithm with deep transfer learning for FDD massive MIMO systems [J]. IEEE transactions on cognitive communications and networking, 2021, 7(4): 1253 – 1265. DOI: 10.1109/TCCN.2021.3084409
[38] Zeng J, He Z R, Sun J L, et al. Deep transfer learning for 5G massive MIMO downlink CSI feedback [C]//Proceedings of IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2021: 1 – 5. DOI: 10.1109/wcnc49053.2021.9417349
[39] Boas B V, Zirwas W, Haardt M. Transfer learning capabilities of untrained neural networks for MIMO CSI recreation [C]//International Conference on Communications. IEEE, 2022: 1288 – 1293. DOI: 10.1109/ICC45855.2022.9838738
[40] Wang Y T, Sun J L, Wang J, et al. Multi-rate compression for downlink CSI based on transfer learning in FDD massive MIMO systems [C]//The 94th Vehicular Technology Conference (VTC2021-Fall). IEEE, 2021: 1 – 5. DOI: 10.1109/VTC2021-Fall52928.2021.9625585
[41] Sun J L, Zhang Y B, Gui G, et al. Interacting federated and transfer learning-aided CSI prediction for intelligent cellular networks [J]. IEEE transactions on vehicular technology, 2023, 72(12): 15776 – 15787. DOI: 10.1109/TVT.2023.3290660
[42] Sohrabi F, Attiah K M, Yu W. Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO [J]. IEEE transactions on wireless communications, 2021, 20(7): 4044 – 4057. DOI: 10.1109/TWC.2021.3055202
[43] Jang J, Lee H, Kim I M, et al. Deep learning for multi-user MIMO systems: joint design of pilot, limited feedback, and precoding [J]. IEEE transactions on communications, 2022, 70(11): 7279 – 7293. DOI: 10.1109/TCOMM.2022.3209887
[44] Jee J, Park H. Deep learning-based joint optimization of closed-loop FDD mmWave massive MIMO: pilot adaptation, CSI feedback, and beamforming [J]. IEEE transactions on vehicular technology, 2024, 73(3): 4019 – 4034. DOI: 10.1109/TVT.2023.3327276
[45] Hu Q Y, Cai Y L, Kang K, et al. Two-timescale end-to-end learning for channel acquisition and hybrid precoding [J]. IEEE journal on selected areas in communications, 2022, 40(1): 163 – 181. DOI: 10.1109/JSAC.2021.3126050
[46] Zhang H Y, Lu Z L, Zhang X D, et al. Data augmentation for bridging the delay gap in DL-based massive MIMO CSI feedback [J]. IEEE wireless communications letters, 2024, 13(5): 1315 – 1319. DOI: 10.1109/LWC.2024.3368558
[47] Cui Y M, Guo J J, Cao Z, et al. Lightweight neural network with knowledge distillation for CSI feedback [J]. IEEE transactions on communications,

2024, 72(8): 4917 – 4929. DOI: 10.1109/TCOMM.2024.3377724
[48] Zhou T, Liu X P, Xiang Z W, et al. Transformer network based channel prediction for CSI feedback enhancement in AI-native air interface [J]. IEEE transactions on wireless communications, 2024, 23(9): 11154 – 11167. DOI: 10.1109/TWC.2024.3379123
[49] Yan J T, Chen T, Xie B, et al. Hierarchical federated learning: architecture, challenges, and its implementation in vehicular networks [J]. ZTE communications, 2023, 21(1): 38 – 45. DOI: 10.12142/ZTECOM.202301005
[50] Xu W, Huang Y M, Wang W, et al. Toward ubiquitous and intelligent 6G networks: from architecture to technology [J]. Science China information sciences, 2023, 66(3): 130300. DOI: 10.1007/s11432-023-3704-8
[51] Lin X Q. An overview of the 3GPP study on artificial intelligence for 5G new radio [PP/OL]. arXiv (2023-08-10) [2024-08-20]. https://doi. org/10.48550/arXiv.2308.05315
[52] Guo J J, Wen C K, Jin S, et al. AI for CSI feedback enhancement in 5G-advanced [J]. IEEE wireless communications, 2024, 31(3): 169 – 176. DOI: 10.1109/MWC.010.2200304
[53] Sarieddeen H, Alouini M S, Al-Naffouri T Y. Terahertz-band ultra-massive spatial modulation MIMO [J]. IEEE journal on selected areas in communications, 2019, 37(9): 2040 – 2052. DOI: 10.1109/JSAC.2019.2929455
[54] Wang K Y, Gao Z, Chen S, et al. Knowledge and data dual-driven channel estimation and feedback for ultra-massive MIMO systems under hybrid field beam squint effect [J]. IEEE transactions on wireless communications, 2024, 23(9): 11240 – 11259. DOI: 10.1109/TWC.2024.3380638
[55] Balatsoukas-Stimming A, Studer C. Deep unfolding for communications systems: a survey and some new directions [C]//IEEE International Workshop on Signal Processing Systems (SiPS). IEEE, 2019: 266 – 271. DOI: 10.1109/sips47522.2019.9020494
[56] Deng L T, Z Y K. Deep learning-based semantic feature extraction: a literature review and future directions [J]. ZTE communications, 2023, 21 (2): 11 – 17. DOI: 10.12142/ZTECOM.202302003
[57] Shi G M, Xiao Y, Li Y Y, et al. From semantic communication to semantic-aware networking: model, architecture, and open problems [J]. IEEE communications magazine, 2021, 59(8): 44 – 50. DOI: 10.1109/MCOM.001.2001239
[58] Guo J J, Wen C K, Jin S. Deep data hiding-based CSI feedback overhead elimination: An initial investigation [C]//IEEE International Conference on Communications. IEEE, 2022: 5347 – 5352. DOI: 10.1109/ICC45855.2022.9839120
[59] Carpi F, Venkatesan S, Du J F, et al. Precoding-oriented massive MIMO CSI feedback design [C]//International Conference on Communications. IEEE, 2023: 4973 – 4978. DOI: 10.1109/ICC45041.2023.10278955

Biographies

Lu Zhaohua received his BS degree in electrical engineering and PhD degree in signal processing from Tianjin University, China in 2001 and 2006, respectively. Since 2006, he has been engaged in mobile communication physical layer technology at ZTE Corporation, including MIMO, interference control, artificial intelligence, etc. He has published more than 30 papers and held over 200 authorized patents.

Yi Chenyang received her BS degree in electrical engineering from Southeast University, China in 2020. She is currently working toward her PhD degree with the School of Information Science and Engineering, National Mobile Communications Research Laboratory, Southeast University. Her current research interests include massive MIMO, mmWave communications, and artificial intelligence for wireless communications.

Wu Jie received his BS degree in electrical engineering from Southeast University, China in 2022, where he is currently pursuing his MS degree in communication and information engineering. His recent research interests include deep learning for CSI compression and feedback in wireless communications.

Shao Bo received his BS degree in electrical engineering from Xidian University, China in 2023. He is currently pursuing his MS degree in communication and information engineering at Southeast University, China. His recent research interests include deep learning for CSI compression and feedback in wireless communications and massive MIMO systems.

Xu Wei (wxu@seu.edu.cn) received his BS degree in electrical engineering and his MS and PhD degrees in communication and information engineering from Southeast University, China in 2003, 2006, and 2009, respectively. Between 2009 and 2010, he was a post-doctoral research fellow with the Department of Electrical and Computer Engineering, University of Victoria, Canada. He is currently a professor at the National Mobile Communications Research Laboratory, Southeast University. He was an adjunct professor of the University of Victoria, Canada from 2017 to 2020, and a distinguished visiting fellow of the Royal Academy of Engineering, UK in 2019. He has co-authored over 100 refereed journal papers in addition to 36 domestic patents and four US patents granted. His research interests include information theory, signal processing and machine learning for wireless communications. He is currently an editor of *IEEE Transactions on Communications* and a senior editor of *IEEE Communications Letters*. He received the Best Paper Awards from a number of prestigious IEEE conferences including IEEE Globecom/ICCC, etc. He received the Science and Technology Award for Young Scholars of the Chinese Institute of Electronics in 2018. He is an IEEE Fellow and IET Fellow.

New Member of ZTE Communications Editorial Board

Wang Ling received his BSc, MSc, and PhD degrees in electronic engineering from Xidian University, China in 1999, 2002, and 2004, respectively. From 2004 to 2007, he worked at Siemens and Nokia Siemens Networks, Beijing. Since 2007, he has worked at Northwestern Polytechnical University (NPU), China, where he was promoted to Professor in 2012. Currently, he serves as the Vice President of NPU. His research interests include spectrum sensing, array processing, smart antennas, and cognitive radio. He has published over 100 research papers and holds more than 60 authorized invention patents.

Low-Complexity OTFS Channel Equalization Based on CLU-MMSE

Jia Haoxiang[1], Zhao Danfeng[1], Xin Yu[2], Hua Jian[2]
(1. College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202601004
https://kns.cnki.net/kcms/detail/34.1294.TN.20260302.1718.004.html, published online March 3, 2026
Manuscript received: 2024-03-18

Abstract: In view of the high computational complexity of traditional linear equalization algorithms in Orthogonal Time Frequency Space (OTFS) systems, a minimum mean square error (MMSE) channel equalization algorithm based on Matrix Chunking Lower and Upper Triangular Decomposition (CLU) is proposed. The proposed algorithm derives the structural properties of the chunked MMSE equalization matrix by leveraging the block diagonal structure of the Cyclic Prefix OTFS (CP-OTFS) time-domain channel matrix and the quasi-band structure of its constituent block matrices. On this basis, triangular decomposition combined with forward and backward substitution is used to avoid matrix inversion. This approach significantly reduces the complexity of the MMSE algorithm without sacrificing its performance.

Keywords: OTFS; equalization algorithm; MMSE; Lower and Upper Triangular Decomposition

Citation (Format 1): Jia H X, Zhao D F, Xin Y, et al. Low-complexity OTFS channel equalization based on CLU-MMSE [J]. ZTE Communications, 2026, 24(1): 16 - 24. DOI: 10.12142/ZTECOM.202601004
Citation (Format 2): H. X. Jia, D. F. Zhao, Y. Xin, et al., "Low-complexity OTFS channel equalization based on CLU-MMSE," *ZTE Communications*, vol. 24, no. 1, pp. 16 - 24, Mar. 2026. doi: 10.12142/ZTECOM.202601004.

1 Introduction

Orthogonal Time Frequency Space (OTFS) is a novel multicarrier modulation technique that characterizes the time-frequency dual-selective channel as an approximately time-invariant channel in the Delay-Doppler (DD) domain through a two-dimensional time-frequency extension[1], making it suitable for highly dynamic scenarios.

However, in actual communication environments, OTFS systems still suffer from inter-code interference, inter-subcarrier interference, and Doppler interference, which require channel equalization techniques to ensure system reliability[2 - 6]. Since the equivalent channel matrix dimension of OTFS in the DD domain is much higher than that of Orthogonal Frequency Division Multiplexing (OFDM), the complexity of channel equalization increases significantly. Therefore, the study of equalization algorithms with low complexity and high performance is key to the development of OTFS modulation systems.

Existing channel equalization techniques for OTFS can be classified into linear and nonlinear types based on design criteria. Linear equalization offers advantages such as simple structure and easy implementation, making it widely used in communication systems. Among various linear equalization algorithms, the minimum mean square error (MMSE) is the most commonly used. However, for an OTFS system with N symbols and M subcarriers, the size of the equivalent channel matrix in both the time domain and the DD domain is $MN \times MN$, and the complexity of the conventional MMSE with matrix inversion is as high as $O((MN)^3)$. To address this problem, various low-complexity linear equalization algorithms based on the properties of the OTFS channel matrix have been proposed. Under ideal pulse conditions, Surabhi et al. [7] proposed an MMSE equalization algorithm with linear complexity by exploiting the doubly circulant property of the OTFS delay-Doppler domain channel matrix. This algorithm reduces the complexity of MMSE to logarithmic levels, but its performance deteriorates severely under practical rectangular pulse waveforms. To address this issue, Tiwari et al. [8] proposed an MMSE algorithm with logarithmic complexity under rectangular pulse waveforms by utilizing the quasi-banded property of the Reduced Cyclic Prefix (RCP)-OTFS time-domain channel matrix, avoiding large matrix inversion through matrix decomposition. However, this algorithm heavily depends on the channel matrix structure and cannot be directly applied to Cyclic Prefix OTFS (CP-OTFS) systems, which offer better compatibility with OFDM. Based on the above analysis, most existing low-complexity linear equalization algorithms for OTFS are MMSE variants developed in different domains, and most rely on the assumption of ideal pulse or RCP-OTFS. Therefore, it is important to further explore practical and effective low-complexity linear equalization algorithms that can be applied

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. KY10800230005.

to CP-OTFS for the implementation of OTFS technology.

Motivated by this, this paper proposes a low-complexity Matrix Chunking Lower and Upper Triangular Decomposition (CLU)-MMSE equalization algorithm for CP-OTFS systems. The algorithm employs matrix chunking combined with lower-upper (LU) decomposition of banded matrices, leveraging the block-diagonal structure of CP-OTFS time-domain channel matrices.

2 OTFS Fundamentals

2.1 OTFS System Model

Traditional wireless communication signals, such as those in OFDM systems, are typically analyzed and processed in the time-frequency domain. In contrast, OTFS modulation introduces the concept of the DD domain, and realizes the mutual conversion between the DD and time-frequency (TF) domains through the two-dimensional Symplectic Finite Fourier Transform (SFFT), as shown in Fig. 1.

In Fig. 1, M and N denote the number of subcarriers and symbols per OTFS frame, respectively. T is the duration of a single symbol, which is the reciprocal of the subcarrier spacing Δf. The bandwidth occupied by an OTFS frame is $B = M\Delta f$, and the frame duration is $T_f = NT$.

The bitstream signals $\boldsymbol{b} = [b_1, b_2, \cdots, b_k]$ generated at the transmitter side are mapped to a transmit signal vector $\boldsymbol{a} = [a_1, a_2, \cdots, a_{MN}]$ via a modulator. If the constellation size is A, the number of bitstream symbols is $K = MN\log_2 A$. By arranging the transmit signal vector $\boldsymbol{a}$ sequentially on the DD-domain grid, the DD-domain OTFS complex signal matrix $\boldsymbol{x}_{\mathrm{DD}} \in \mathbb{C}^{M\times N}$ is obtained. The OTFS modulation is then performed on this matrix. Fig. 2 illustrates the complete block diagram of the OTFS communication system.

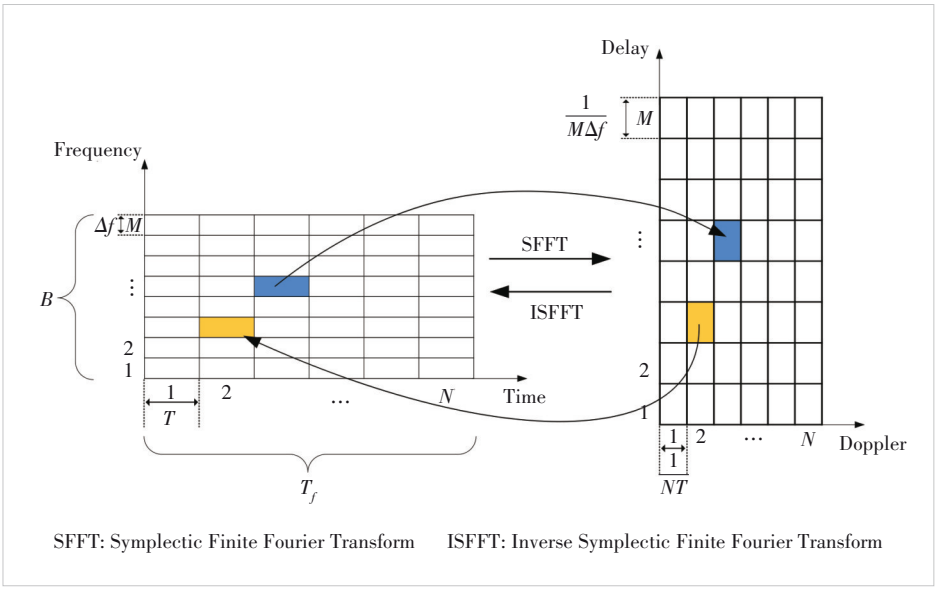


Figure 1. Schematic diagram of the relationship between channel resource transformations in the TF and DD domains

At the transmitter, the transmit symbol $x_{\mathrm{DD}}[k,l]$ at the k-th Doppler and l-th delay grid point in the DD domain undergoes a 2D Inverse Symplectic Finite Fourier Transform (ISFFT) [9], which transforms it into the TF domain symbols $X_{\mathrm{TF}}[n,m]$, where $0 \leqslant n \leqslant N-1$ and $0 \leqslant m \leqslant M-1$. The transformation is given by:

$$X_{\mathrm{TF}}[n,m] = \frac{1}{\sqrt{NM}} \sum_{k=0}^{N-1}\sum_{l=0}^{M-1} x_{\mathrm{DD}}[k,l]\, \mathrm{e}^{\mathrm{j}2\pi\left(\frac{nk}{N}-\frac{ml}{M}\right)} \tag{1}$$

The TF-domain symbol $X_{\mathrm{TF}}[n,m]$ is then converted to the time-domain signal $s(t)$ via the Heisenberg transform, which can be regarded as the process of adding a window to the TF-domain signal after an M-point inverse Fourier transform, typically implemented via the Inverse Fast Fourier Transform (IFFT). The resulting signal is:

$$s(t) = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} X_{\mathrm{TF}}[n,m]\, g_{\mathrm{tx}}(t-nT)\mathrm{e}^{\mathrm{j}2\pi m\Delta f(t-nT)} \tag{2}$$

where g_{tx} denotes the transmit shaping pulse with duration $[0, T]$, repeated N times per frame.

The time domain signal arrives at the receiving end through the wireless channel, yielding the received signal $r(t)$:

$$r(t) = \iint h(\tau,\nu)s(t-\tau)\mathrm{e}^{\mathrm{j}2\pi\nu(t-\tau)}\mathrm{d}\nu\,\mathrm{d}\tau + n(t) \tag{3}$$

where $h(\tau,\nu)$ represents the DD domain channel impulse response, and $n(t)$ is additive white Gaussian noise (AWGN) component. $\mathcal{CN}(\boldsymbol{n};0,\sigma_n^2)$, which is an operator indicating that the variable n obeys a complex Gaussian distribution, is calculated as:

$$\mathcal{CN}(n;\mu,\sigma^2) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{|n-\mu|^2}{\sigma^2}\right) \tag{4}$$

Consider a mobile terminal moving at velocity v with carrier frequency f_c. For the i-th propagation path, let θ_i denote the angle of arrival relative to the moving direction, d_i the path length, and c the speed of light. The corresponding delay and Doppler shift are given by:

$$\tau_i = \frac{d_i}{c} \tag{5}$$

$$\nu_i = f_c\frac{v\cos\theta_i}{c} \tag{6}$$

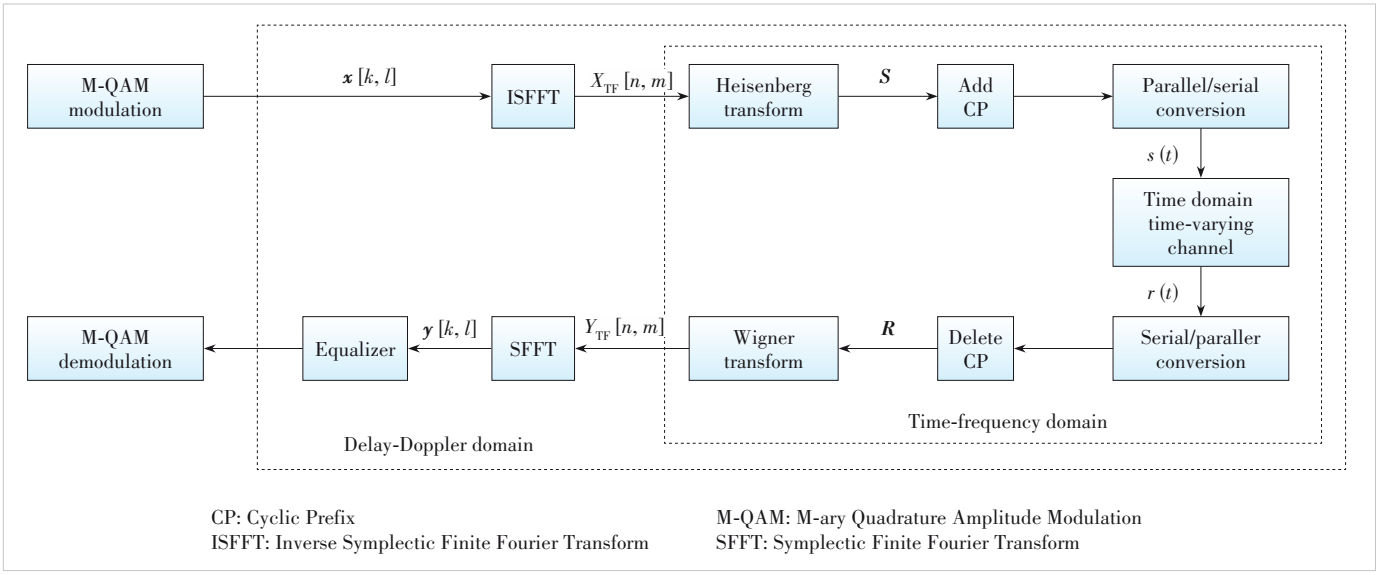


Figure 2. Block diagram of an OTFS communication system

This leads to an expression for the channel impulse response in the DD domain:

$$h(\tau,\nu)=\sum_{i=1}^{P}h_i\delta(\tau-\tau_i)\delta(\nu-\nu_i) \tag{7}$$

where P is the total number of paths and h_i is the complex gain of the i-th path.

The path delays and Doppler shifts in Eqs. (5) and (6) can be obtained by projecting them into the DD domain with resolution of $1/M\Delta f$ and $1/NT$, respectively:

$$\tau_i=\frac{l_{\tau_i}}{M\Delta f},\quad \nu_i=\frac{k_{\nu_i}+\kappa_{\nu_i}}{NT} \tag{8}$$

where l_{τ_i} is the integer delay tap, k_{ν_i} is the integer Doppler taps, and κ_{ν_i} is the fractional Doppler taps with $\kappa_{\nu_i}\in(-0.5,0.5]$. The indices l_{τ_i} and k_{ν_i} represent the path's delay index and Doppler indices in the DD domain, respectively. In general, the delay resolution is sufficient to approximate each path's delay to the nearest sampling point; therefore, there is no need to consider the fractional delay.

Sampling the received signal in Eq. (3) yields the discrete-time signal:

$$r_u=\sum_{i=1}^{P}h_i s(u-l_i)\mathrm{e}^{\mathrm{j}2\pi k_i(u-l_i)}+n_u \tag{9}$$

Unlike time-invariant multipath channels, time-varying multipath channels introduce significant Doppler shifts $(\theta_i)^{k_i-l_i}$, which cause inter-carrier interference (ICI) in the frequency domain and inter-Doppler interference (IDI) in the DD domain. These interference components severely degrade the bit error rate (BER) performance and greatly increase the difficulty of channel estimation and signal detection. For a channel with one direct path and two reflection paths, the received signal r_u is the superposition of these paths, each weighted by its corresponding path gain and shifted by its propagation delay. The direct path arrives first, so its delay can be regarded as zero. The received signal r_u can therefore be expressed as:

$$r_u=h_1s_1(\theta_1)^{k_1}+h_2s_2(\theta_2)^{k_2-l_2}+h_3s_3(\theta_3)^{k_3-l_3}+n_u \tag{10}$$

At the receiver, the received signal is passed through a matched filter to obtain the cross-ambiguity function $A_{g_{rx},y}(t,f)$. Sampling this function at $t=nT$ and $f=m\Delta f$ yields the discrete TF domain received signal $Y_{\mathrm{TF}}[n,m]$, expressed as

$$Y_{\mathrm{TF}}(t,f)=A_{g_{rx},y}(t,f)\triangleq\int g_{rx}^{*}(t-t')r(t')\mathrm{e}^{-\mathrm{j}2\pi f(t-t')}\mathrm{d}t' \tag{11}$$

$$Y_{\mathrm{TF}}[n,m]=Y_{\mathrm{TF}}(t,f)\big|_{t=nT,\ f=m\Delta f} \tag{12}$$

where $g_{rx}(t)$ denotes the receive filter. The operations in Eqs. (11) and (12) together constitute the Wigner transform. Applying the SFFT to the output of the Wigner transform yields the DD domain signal $y_{\mathrm{DD}}[k,l]$ as:

$$y_{\mathrm{DD}}[k,l]=\frac{1}{\sqrt{NM}}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}Y_{\mathrm{TF}}[n,m]\mathrm{e}^{-\mathrm{j}2\pi\left(\frac{nk}{N}-\frac{ml}{M}\right)} \tag{13}$$

The resulting $y_{\mathrm{DD}}[k,l]$ is the demodulated OTFS signal, which serves as the input to the subsequent signal detection stage.

2.2 CP-OTFS

Fig. 3 illustrates the complete implementation of a CP-OTFS

transmitter. As shown, the TF-domain and time-domain processing in CP-OTFS is identical to that in OFDM, except for the addition of a CP to each symbol. This structural similarity renders CP-OTFS highly compatible with OFDM. Consequently, CP-OTFS can be implemented directly on existing OFDM systems by simply adding pre- and post-processing modules, facilitating the rapid deployment of OTFS technology.

The time-domain channel matrix for CP-OTFS can be expressed as[10]:

$$\boldsymbol{H}_t^{\mathrm{CP}} = \begin{bmatrix} \boldsymbol{H}_1 & & & \\ & \boldsymbol{H}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{H}_N \end{bmatrix} \tag{14},$$

where $\boldsymbol{H}_n \in \mathbb{C}^{M\times M}$ is the n-th chunked submatrix, calculated as

$$\boldsymbol{H}_n = \sum_{i=1}^{P} h_i \boldsymbol{\Pi}_M^{l_{\tau_i}} \boldsymbol{\Delta}^{k_{\nu_i},n} \tag{15},$$

where $\boldsymbol{\Pi}_M^{l_{\tau_i}}$ is the forward cyclic shift matrix and $\boldsymbol{\Delta}^{k_{\nu_i},n}$ is the diagonal phase matrix of dimension M, defined as:

$$\boldsymbol{\Pi} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}_{M\times M}$$

$$\boldsymbol{\Delta}^{(k_i)} = \begin{bmatrix} \mathrm{e}^{\mathrm{j}2\pi k_{\nu_i}\frac{[(n-1)M(0)]}{M}} & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \mathrm{e}^{\mathrm{j}2\pi k_{\nu_i}\frac{[(n-1)M-1]}{M}} \end{bmatrix}_{M\times M} \tag{16}.$$

In Eq. (14), the omitted blocks of $\boldsymbol{H}_t^{\mathrm{CP}}$ are zero matrices.

From Eq. (14) and Eq. (16), it can be seen that the time-domain channel matrix of CP-OTFS is a block diagonal matrix with N submatrices. Therefore, the structure of the CP-OTFS time-domain matrix under rectangular pulse conditions can be obtained as shown in Fig. 4.

Consequently, the time-domain channel matrix of CP-OTFS under rectangular pulse shaping assumes the block-diagonal form illustrated in Fig. 4 and given by:

$$\boldsymbol{H}_t^{\mathrm{CP}} = \mathrm{diag}\left\{\boldsymbol{H}_1^{\mathrm{CP}},\cdots,\boldsymbol{H}_N^{\mathrm{CP}}\right\} \tag{17},$$

where each diagonal block $\boldsymbol{H}_i^{\mathrm{CP}} \in \mathbb{C}^{M\times M}$, $1 \leqslant i \leqslant N$, is generally distinct.

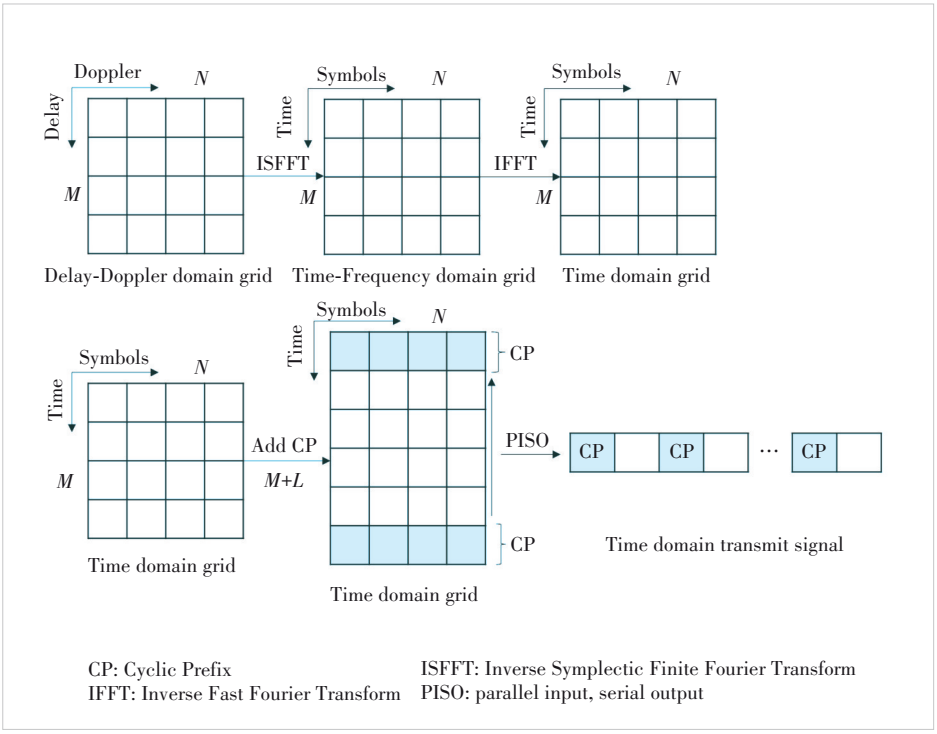


Figure 3. Schematic diagram of a CP-OTFS transmitter

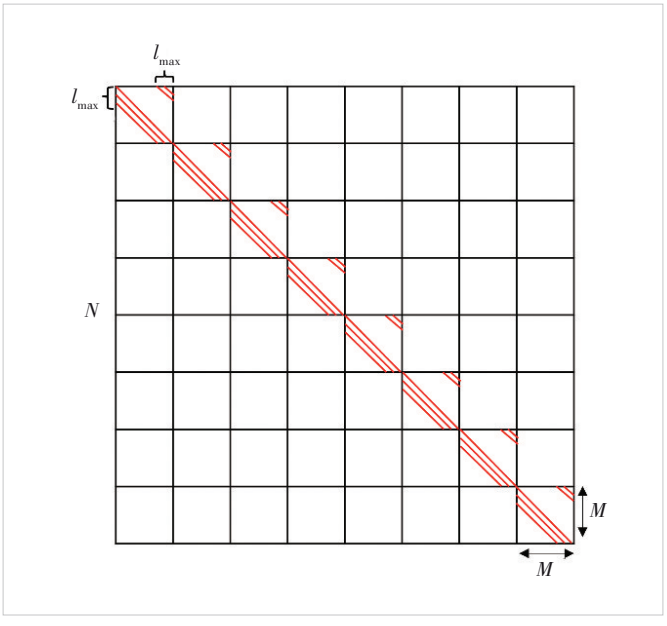


Figure 4. Schematic of CP-OTFS time domain channel matrix structure

3 Channel Equalization Algorithm Based on CLU-MMSE

3.1 MMSE Equalization Algorithm

Let the number of OTFS symbols be N and the number of subcarriers be M. The input-output relationship between the transmitter and receiver can be expressed as:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{18}$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^{MN \times 1}$ denote the transmitted and received symbol vectors in the time or DD domain, $\boldsymbol{H} \in \mathbb{C}^{MN \times MN}$ is the equivalent channel matrix in the corresponding domain, and $\boldsymbol{n} \in \mathbb{C}^{MN \times 1}$ is the additive white Gaussian noise vector.

Equalization aims to recover the transmitted signal by processing the received signal, e.g., through interference cancellation. The equalized output signal can be expressed as:

$$\hat{\boldsymbol{x}} = \boldsymbol{G}\boldsymbol{y} \tag{19}$$

where $\boldsymbol{G} \in \mathbb{C}^{MN \times MN}$ is the linear transformation equalization matrix.

MMSE equalization derives the equalization matrix by minimizing the mean square error between the transmitted signal and the equalized signal. The estimated signal at the MMSE equalization output is given by:

$$\begin{aligned}\hat{\boldsymbol{x}}_{\text{MMSE}} &= \boldsymbol{G}_{\text{MMSE}}\boldsymbol{y} = \left(\boldsymbol{H}^H\boldsymbol{H} + \sigma^2\boldsymbol{I}_{MN}\right)^{-1}\boldsymbol{H}^H\boldsymbol{y} = \\ &\boldsymbol{x} + \left(\boldsymbol{H}^H\boldsymbol{H} + \sigma^2\boldsymbol{I}_{MN}\right)^{-1}\boldsymbol{H}^H\boldsymbol{n}\end{aligned} \tag{20}$$

where σ^2 denotes the noise variance.

Direct matrix inversion in MMSE involves large-scale matrices, resulting in a computational complexity of $\mathrm{O}((MN)^3)$ complex multiplications.

3.2 CLU-MMSE Equalization Algorithm

Substituting Eq. (17) into Eq. (20) yields the time-domain MMSE equalization matrix for CP-OTFS:

$$\begin{aligned}&\boldsymbol{G} = ((\boldsymbol{H}_t^{\text{CP}})^H\boldsymbol{H}_t^{\text{CP}} + \sigma^2\boldsymbol{I}_{MN})^{-1}(\boldsymbol{H}_t^{\text{CP}})^H = \\ &\left(\begin{bmatrix}(\boldsymbol{H}_1^{\text{CP}})^H & & \\ & (\boldsymbol{H}_2^{\text{CP}})^H & \\ & & \ddots \\ & & & (\boldsymbol{H}_N^{\text{CP}})^H\end{bmatrix}\cdot\begin{bmatrix}\boldsymbol{H}_1^{\text{CP}} & & \\ & \boldsymbol{H}_2^{\text{CP}} & \\ & & \ddots \\ & & & \boldsymbol{H}_N^{\text{CP}}\end{bmatrix} + \sigma^2\boldsymbol{I}_{MN}\right)^{-1}\cdot(\boldsymbol{H}_t^{\text{CP}})^H = \\ &\begin{bmatrix}(\boldsymbol{H}_1^{\text{CP}})^H\boldsymbol{H}_1^{\text{CPp}} + \sigma^2\boldsymbol{I}_M & & & \\ & (\boldsymbol{H}_2^{\text{CP}})^H\boldsymbol{H}_2^{\text{CP}} + \sigma^2\boldsymbol{I}_M & & \\ & & \ddots & \\ & & & (\boldsymbol{H}_N^{\text{CP}})^H\boldsymbol{H}_N^{\text{CP}} + \sigma^2\boldsymbol{I}_M\end{bmatrix}^{-1}\cdot(\boldsymbol{H}_t^{\text{CP}})^H = \\ &\begin{bmatrix}(\boldsymbol{G}_1')^{-1} & & \\ & (\boldsymbol{G}_2')^{-1} & \\ & & \ddots \\ & & & (\boldsymbol{G}_N')^{-1}\end{bmatrix}\cdot\begin{bmatrix}(\boldsymbol{H}_1^{\text{CP}})^H & & \\ & (\boldsymbol{H}_2^{\text{CP}})^H & \\ & & \ddots \\ & & & (\boldsymbol{H}_N^{\text{CP}})^H\end{bmatrix}\end{aligned} \tag{21}$$

where $\boldsymbol{G}_i' = (\boldsymbol{H}_i^{\text{CP}})^H\boldsymbol{H}_i^{\text{CP}} + \sigma^2\boldsymbol{I}_M \in \mathbb{C}^{M \times M}$, $i = 1, \cdots, N$.

The result of substituting Eq. (21) into Eq. (19) in a chunked matrix form is given by:

$$\hat{\boldsymbol{x}}_i = (\boldsymbol{G}_i')^{-1}(\boldsymbol{H}_i^{\text{CP}})^H\boldsymbol{y}_i \tag{22}$$

where $\boldsymbol{y}_i, \hat{\boldsymbol{x}}_i \in \mathbb{C}^{M \times 1}$, $i = 1, \cdots, N$, are the i-th subvectors of the time-domain received vector $\boldsymbol{y}_t$ and the equalized estimate vector $\hat{\boldsymbol{x}}$, respectively.

From Eq. (21), the time-domain MMSE equalization matrix $\boldsymbol{G}$ is block diagonal, with each submatrix sharing the same structure. Since the calculation of $\boldsymbol{G}$ requires the inverse of each $\boldsymbol{G}_i'$, the structural properties of the matrix $\boldsymbol{G}_i', i = 1, \cdots, N$ are derived first. Substituting Eq. (14) into Eq. (21), the matrix $\boldsymbol{G}_i'$ can be expressed as:

$$\begin{aligned}&\boldsymbol{G}_i' = (\boldsymbol{H}_i^{\text{CP}})^H\boldsymbol{H}_i^{\text{CP}} + \sigma^2\boldsymbol{I}_M = \\ &\sum_{p=1}^{P} h_p\boldsymbol{\Delta}^{-k_p}\boldsymbol{\Pi}^{-l_p}\sum_{s=1}^{P} h_s'\boldsymbol{\Delta}^{k_s}\boldsymbol{\Pi}^{l_s} + \sigma^2\boldsymbol{I}_M = \\ &\sum_{\substack{p=1 \\ p=s}}^{P}\left(\left|h_p\right|^2 + \sigma^2\right)\boldsymbol{I}_M + \sum_{\substack{p=1 \\ p\neq s}}^{P}\sum_{s=1}^{P} h_p h_s'\boldsymbol{\Pi}^{l_s - l_p}\boldsymbol{\Delta}^{k_s - k_p}\end{aligned} \tag{23}$$

where P is the number of channel paths, h_p and h_s' are the complex channel coefficients of different paths, and l_i and k_i, $i = 1, \cdots, P$, are the delay taps and Doppler taps corresponding to each path, respectively.

Let $l_{\max}$ denote the maximum delay tap and β denote the ceiling of the maximum Doppler tap. Then, the range of $(l_s - l_p)$ is $[-l_{\max}, l_{\max}]$. Therefore, the maximum shifts represented by $\boldsymbol{\Pi}^{l_s - l_p}$ can shift up to $l_{\max}$ positions to the left or right, indicating that the submatrix $\boldsymbol{G}_i'$ is a quasi-banded matrix with a bandwidth of $(2l_{\max} + 1)$, as shown in Fig. 5. The inversion of $\boldsymbol{G}_i'$ can be efficiently performed using matrix factorization algorithms, as described in the following.

First, the LU decomposition of $\boldsymbol{G}_i'$ is illustrated in Fig. 5. Let $Q = M - l_{\max}$; this decomposition process can be expressed in matrix form as:

$$\begin{aligned}&\boldsymbol{G}_i' = \begin{bmatrix}(\boldsymbol{A})_{Q\times Q} & (\boldsymbol{B})_{Q\times l_{\max}} \\ (\boldsymbol{C})_{\alpha\times Q} & (\boldsymbol{D})_{l_{\max}\times l_{\max}}\end{bmatrix} = \\ &\begin{bmatrix}(\boldsymbol{L}_A)_{Q\times Q} & (\boldsymbol{0})_{Q\times l_{\max}} \\ (\boldsymbol{E})_{l_{\max}\times Q} & (\boldsymbol{R})_{l_{\max}\times l_{\max}}\end{bmatrix}\times\begin{bmatrix}(\boldsymbol{U}_A)_{Q\times Q} & (\boldsymbol{F})_{Q\times l_{\max}} \\ (\boldsymbol{0})_{l_{\max}\times Q} & (\boldsymbol{T})_{l_{\max}\times l_{\max}}\end{bmatrix} = \boldsymbol{L}_i\boldsymbol{U}_i\end{aligned} \tag{24}$$

where $\boldsymbol{L}_i$ and $\boldsymbol{U}_i$ are the LU decomposition matrices of $\boldsymbol{G}_i'$, and $\boldsymbol{L}_A$ and $\boldsymbol{U}_A$ are the LU decomposition matrices of the chunking

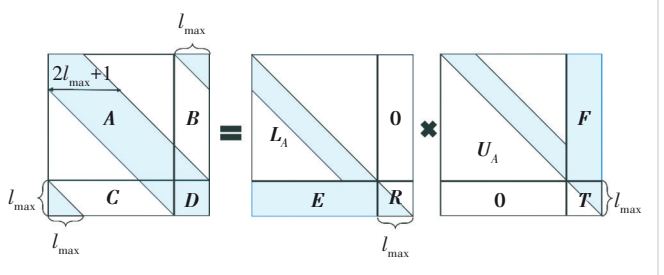


Figure 5. Schematic diagram of the LU decomposition of the chunked matrix

matrix $\boldsymbol{A}$.

The following matrix relationship can be obtained from the above equation:

$$\boldsymbol{A} = \boldsymbol{L}_A \boldsymbol{U}_A \tag{25}$$

$$\boldsymbol{F} = \left(\boldsymbol{L}_A\right)^{-1} \boldsymbol{B} \tag{26}$$

$$\boldsymbol{E} = \boldsymbol{C}\left(\boldsymbol{U}_A\right)^{-1} \tag{27}$$

$$\boldsymbol{RT} = \boldsymbol{D} - \boldsymbol{EF} \tag{28}$$

Therefore, to compute the LU decomposition of $\boldsymbol{G}_i'$, one needs to solve for the matrices $\boldsymbol{L}_A$, $\boldsymbol{U}_A$, $\boldsymbol{E}$, $\boldsymbol{F}$, $\boldsymbol{R}$, and $\boldsymbol{T}$ separately.

From Fig. 5, $\boldsymbol{A} \in \mathbb{C}^{Q \times Q}$ is a standard banded matrix with a non-zero element bandwidth of $2l_{\max} - 1$ and a half-bandwidth of $l_{\max}$. The LU decomposition of the standard banded matrix can be performed using a low-complexity LU decomposition algorithm. The algorithm proceeds in Q steps. For the k-th step ($k = 0,\cdots, Q-1$), only the elements in an $l_{\max} \times l_{\max}$ rectangular window that is active along the diagonal are involved in the operation. The first row and the first column of this window in the k-th step correspond to the k-th row and the k-th column of $\boldsymbol{A}$, respectively. At this point, the k-th column of $\boldsymbol{L}_A$ and the k-th row of $\boldsymbol{U}_A$ can be obtained using Eqs. (29)–(31):

$$\boldsymbol{L}_A(k,k) = 1 \tag{29}$$

$$\boldsymbol{L}_A(k+i,k) = \boldsymbol{A}(k+i,k)/\boldsymbol{A}(k,k) \quad 1 \leqslant i < m \tag{30}$$

$$\boldsymbol{U}_A(k,k+j) = \boldsymbol{A}(k,k+j) \quad 0 \leqslant j < m \tag{31}$$

In the rectangular window of $\boldsymbol{A}$, the elements not directly involved in the current step are updated using Eqs. (30) and (31):

$$\boldsymbol{A}(k+i,k+j) = \boldsymbol{A}(k+i,k+j) - \boldsymbol{L}_A(k+i,k)\boldsymbol{U}_A(k,k+j) \tag{32}$$

where $0 < i < m$, $0 < j < m$.

As k increases, the $l_{\max} \times l_{\max}$ rectangular window moves along the diagonal of $\boldsymbol{A}$ element by element. The corresponding columns and rows of $\boldsymbol{L}_A$ and $\boldsymbol{U}_A$ are computed sequentially, ultimately completing the decomposition.

Using Eq. (28) to solve for $\boldsymbol{R}$ and $\boldsymbol{T}$, we observe that $\boldsymbol{R}$ and $\boldsymbol{T}$ are lower and upper triangular matrices in small dimensions, respectively. They can be directly obtained via LU decomposition of $(\boldsymbol{D} - \boldsymbol{EF})$ using standard Gaussian elimination.

To find the matrix $\boldsymbol{F}$, we first rewrite Eq. (26) as:

$$\boldsymbol{L}_A \boldsymbol{F} = \boldsymbol{B} \tag{33}$$

Since $\boldsymbol{L}_A$ is a known banded lower triangular matrix, Eq. (33) can be transformed into a system of linear equations and solved for $\boldsymbol{F}$ using recursive relations, as shown in Algorithm 1. Similarly, Eq. (27) can be rewritten as $(\boldsymbol{U}_A^T)\boldsymbol{E}^T = \boldsymbol{C}^T$, so $\boldsymbol{E}$ can also be obtained using Algorithm 1.

Algorithm 1. Inversion of standard banded lower triangular matrices and matrix multiplication

Inputs: banded lower triangular matrix $\boldsymbol{L} \in \mathbb{C}^{Q \times Q}$, matrix $\boldsymbol{B} \in \mathbb{C}^{Q \times l_{\tau_{\max}}}$, bandwidth $l = l_{\max}$
Output: $\boldsymbol{F} = (\boldsymbol{L})^{-1}\boldsymbol{B}$
1: for $k = 0$ to $l - 1$ do
2: $\quad \boldsymbol{F}_{0,k} = \boldsymbol{B}_{0,k}/\boldsymbol{L}_{0,0}$
3: for $i = 1$ to $l - 1$ do
4: $\quad \boldsymbol{F}_{i,k} = \boldsymbol{B}_{i,k}/\boldsymbol{L}_{i,i} - \sum_{j=1}^{i} \boldsymbol{B}_{i,i-j}\boldsymbol{F}_{i-j,k}$
5: end for
6: for $i = l$ to $Q - 1$ do
7: $\quad \boldsymbol{F}_{i,k} = \boldsymbol{B}_{i,k}/\boldsymbol{L}_{i,i} - \sum_{j=1}^{l} \boldsymbol{B}_{i,i-j}\boldsymbol{F}_{i-j,k}$
8: end for
9: end for

The LU decomposition matrices $\boldsymbol{L}_i$ and $\boldsymbol{U}_i$ of $\boldsymbol{G}_i'$ can be obtained after computing each chunk submatrix.

Substituting Eq. (24) into Eq. (22) yields:

$$\hat{\boldsymbol{x}}_i = \overbrace{\left(\boldsymbol{U}_i\right)^{-1}\underbrace{\left(\boldsymbol{L}_i\right)^{-1}\underbrace{\left(\boldsymbol{H}_i^{\mathrm{CP}}\right)^H \boldsymbol{y}_i}_{r_1}}_{r_2}}^{r_3}, \quad i = 1,\cdots,N \tag{34}$$

This equation can be computed in three steps. The first step computes $\boldsymbol{r}_1$, exploiting the sparsity of $\boldsymbol{H}_i^{\mathrm{CP}}$ to reduce complexity. The second step computes $\boldsymbol{r}_2$ by leveraging the banded lower triangular matrix structure of $\boldsymbol{L}_i$. In this step, each element of $\boldsymbol{r}_2$ is obtained recursively using forward substitution, as detailed in Algorithm 2. The third step computes $\boldsymbol{r}_3$ using backward substitution, leveraging the upper triangular structure of $\boldsymbol{U}_i$, as shown in Algorithm 3.

Algorithm 2. Forward substitution for banded lower triangular matrices

Inputs: banded lower triangular matrix $\boldsymbol{L} = \boldsymbol{L}_i \in \mathbb{C}^{M \times M}$, vector $\boldsymbol{r}^{(1)} \in \mathbb{C}^{M \times 1}$, bandwidth $l = l_{\max}$, dimension parameter Q
Output: $\boldsymbol{r}^{(2)} = \boldsymbol{L}^{-1}\boldsymbol{r}^{(1)}$
1: $\quad \boldsymbol{r}_0^{(2)} = \boldsymbol{r}_0^{(1)}$
2: for $k = 1$ to $l - 1$ do
3: $\quad \boldsymbol{r}_k^{(2)} = \boldsymbol{r}_k^{(1)} - \sum_{i=1}^{k} \boldsymbol{L}_{k,k-i}\boldsymbol{r}_{k-i}^{(2)}$
4: end for
5: for $k = l$ to $Q - 1$ do
6: $\quad \boldsymbol{r}_k^{(2)} = \boldsymbol{r}_k^{(1)} - \sum_{i=1}^{l} \boldsymbol{L}_{k,k-i}\boldsymbol{r}_{k-i}^{(2)}$
7: end for
8: for Q to $M - 1$ do
9: $\quad \boldsymbol{r}_k^{(2)} = \boldsymbol{r}_k^{(1)} - \sum_{i=1}^{M-1} \boldsymbol{L}_{k,k-i}\boldsymbol{r}_{k-i}^{(2)}$

10: end for

Algorithm 3. Backward substitution for banded upper triangular matrices

Inputs: banded upper triangular matrix $\boldsymbol{U} = \boldsymbol{U}_i \in \mathbb{C}^{M \times M}$, vector $\boldsymbol{r}^{(2)} \in \mathbb{C}^{M \times 1}$, bandwidth $l = l_{\max}$, dimension parameter Q
Output: $\boldsymbol{r}^{(3)} = \boldsymbol{U}^{-1}\boldsymbol{r}^{(2)}$
1: $\boldsymbol{r}^{(3)}_{(M-1)} = \boldsymbol{r}^{(2)}_{(M-1)}/\boldsymbol{U}_{M-1,M-1}$
2: for $k = M-2$ to $M-2l$ do
3: $\boldsymbol{r}^{(3)}_k = \left(\boldsymbol{r}^{(2)}_k/\boldsymbol{U}_{k,k}\right) - \sum_{i=1}^{M-k-1}\boldsymbol{U}_{k,k+i}\boldsymbol{r}^{(3)}_{k+i}$
4: end for
5: for $k = M-2l-1$ to 0 do
6: $\boldsymbol{r}^{(3)}_k = \left(\boldsymbol{r}^{(2)}_k/\boldsymbol{U}_{k,k}\right) - \sum_{i=1}^{l}\boldsymbol{U}_{k,k+i}\boldsymbol{r}^{(3)}_{k+i} - \sum_{r=M-l}^{M-1}\boldsymbol{U}_{k,r}\boldsymbol{r}^{(3)}_r$
7: end for

Both Algorithms 2 and 3 are derived by utilizing the connection between matrix LU decomposition and linear equation systems. This approach reduces the overall complexity of the MMSE algorithm by replacing complex and large matrix inversion operations with simple recursive subtraction and numerical multiplication.

By performing the above operations on N matrices $\boldsymbol{G}'_i$, the N time-domain estimation subvectors $\hat{\boldsymbol{x}}_i$ are computed and merged into the time-domain estimate $\hat{\boldsymbol{x}} = [\hat{\boldsymbol{x}}_1^{\mathrm{T}},\cdots,\hat{\boldsymbol{x}}_N^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{MN\times 1}$ in column-wise order.

According to Eq. (2), the signal matrix in the time-frequency domain is transformed via the Heisenberg transform to obtain the time-domain transmit signal matrix $\boldsymbol{S}$:

$$\boldsymbol{S} = \boldsymbol{G}_{\mathrm{tx}}\boldsymbol{F}_M^H(\boldsymbol{F}_M\boldsymbol{X}_{\mathrm{DD}}\boldsymbol{F}_N^H) = \boldsymbol{G}_{\mathrm{tx}}\boldsymbol{X}_{\mathrm{DD}}\boldsymbol{F}_N^H \tag{35}$$

where $\boldsymbol{S} \in \mathbb{C}^{M\times N}$, and $\boldsymbol{G}_{\mathrm{tx}} \in \mathbb{C}^{M\times M}$ is the matrix representation of the transmit window function $g_{\mathrm{tx}}(t)$:

$$\boldsymbol{G}_{\mathrm{tx}} = \mathrm{diag}\left(\left[g_{\mathrm{tx}}[0], g_{\mathrm{tx}}[T/M],\cdots, g_{\mathrm{tx}}[(M-1)T/M]\right]\right) \tag{36}$$

where $\mathrm{diag}(\cdot)$ forms a diagonal matrix from the given vector.

The matrix form of a rectangular pulse waveform can be expressed as a unit matrix. Under this condition, the modulation process at the transmitter side in Eq. (35) can be simplified as:

$$\boldsymbol{S} = \boldsymbol{G}_{\mathrm{tx}}\boldsymbol{F}_M^{\mathrm{H}}(\boldsymbol{F}_M\boldsymbol{X}_{\mathrm{DD}}\boldsymbol{F}_N^{\mathrm{H}}) = \boldsymbol{G}_{\mathrm{tx}}\boldsymbol{X}_{\mathrm{DD}}\boldsymbol{F}_N^{\mathrm{H}} = \boldsymbol{X}_{\mathrm{DD}}\boldsymbol{F}_N^{\mathrm{H}} \tag{37}$$

From Eq. (37), the equalized time-domain signal is transformed to the DD domain as:

$$\hat{\boldsymbol{x}}_{\mathrm{DD}} = \mathrm{vec}(\mathrm{vec}^{-1}_{M\times N}(\hat{\boldsymbol{x}})\boldsymbol{F}_N) \tag{38}$$

Fig. 6 shows the overall flowchart of the CLU-MMSE algorithm for the CP-OTFS system.

4 Analysis of Simulation Results

4.1 Complexity Analysis

Using the number of complex multiplications as a metric, the computational complexity of the proposed CLU-MMSE algorithm for CP-OTFS under rectangular pulses is summarized as follows:

1) The time-domain channel matrix of CP-OTFS is chunked by Eq. (17) and the N matrices $\mathbf{G}'_i$ are computed using Eq. (23). This step requires $\left((P^2-P)(\beta+1)MN + P^2N\right)$ complex multiplication.

2) According to Eq. (24), the LU decomposition of N matrices $\mathbf{G}'_i$ is performed, and the LU factors of the standard banded matrix $\boldsymbol{A}$ are obtained using Eqs. (29)–(30). This step requires $\left((l_{\max}^2 - l_{\max})MN\right)$ complex multiplication operations.

3) For the N matrices $\mathbf{G}'_i$, Algorithm 1 is used to compute $\boldsymbol{E}$ and $\boldsymbol{F}$ from Eqs. (26) and (27). Standard LU decomposition is then applied to Eq. (28) to obtain $\boldsymbol{R}$ and $\boldsymbol{T}$. This step requires a total of $\left((2l_{\max}^2 + 3l_{\max})MN - (5l_{\max}^3 + 3l_{\max}^2 + 1)N\right)$ complex multiplications.

4) Using Eq. (34), $\boldsymbol{r}_1$ is calculated, and $\boldsymbol{r}_2$ and $\boldsymbol{r}_3$ are computed according to Algorithms 2 and 3 to obtain the time-domain estimate. This step requires $\left((4l_{\max}+1)MN - (7l_{\max}^2 + 5l_{\max} - 2P)N/2\right)$ complex multiplication operations.

5) The DD-domain estimate is obtained from Eq. (38) by performing an $M\times N$-point Fast Fourier Transform, which requires $(MN\log_2 N)/2$ complex multiplications.

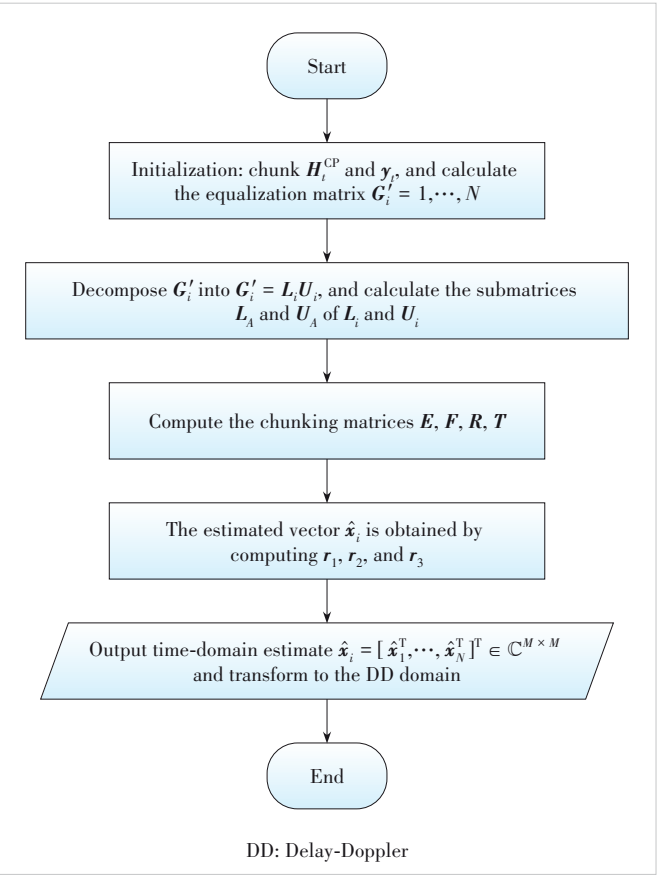


Figure 6. Flowchart of the proposed CLU-MMSE algorithm

In summary, the total complexity of the proposed CLU-MMSE algorithm for CP-OTFS is on the order of $O(MN\log_2 N)$ for large values of MN.

Table 1 compares the computational complexity of several linear equalization algorithms. As shown, the proposed algorithm achieves significantly lower complexity than the traditional matrix-inversion-based MMSE and zero forcing (ZF) algorithms. Moreover, unlike the MMSE equalization algorithm based on the doubly circulant characteristic[11], the proposed algorithm does not rely on the ideal bi-orthogonal pulse assumption, making it more practical.

4.2 Simulation Parameters

To evaluate the bit error rate (BER) performance of the proposed equalization algorithm, simulations are conducted for uncoded OTFS modulation. The Extended Vehicular A (EVA) channel model from LTE is adopted. The detailed simulation parameters are listed in Table 2.

4.3 Simulation Performance

Figs. 7 and 8 compare the BER performance of the proposed algorithm with that of traditional matrix-inversion-based ZF and MMSE algorithms, as well as the MMSE algorithm based on the doubly circulant characteristic. It can be observed that, across different modulation formats and varying values of M and N, the proposed algorithm achieves BER performance nearly identical to that of the traditional matrix-inversion-based MMSE, while significantly outperforming the traditional matrix-inversion-based ZF and the MMSE based on the doubly circulant characteristic.

Table 1. Computational complexity comparison of linear equalization algorithms

Algorithm name	Complexity order
Traditional matrix-inverse ZF	$O((MN)^3)$
Traditional matrix-inverse MMSE	$O((MN)^3)$
MMSE based on doubly circulant characteristic	$O(MN\log_2 MN)$
Proposed CLU-MMSE	$O(MN\log_2 N)$

CLU: Matrix Chunking Lower and Upper Triangular Decomposition
MMSE: minimum mean square error
ZF: zero forcing

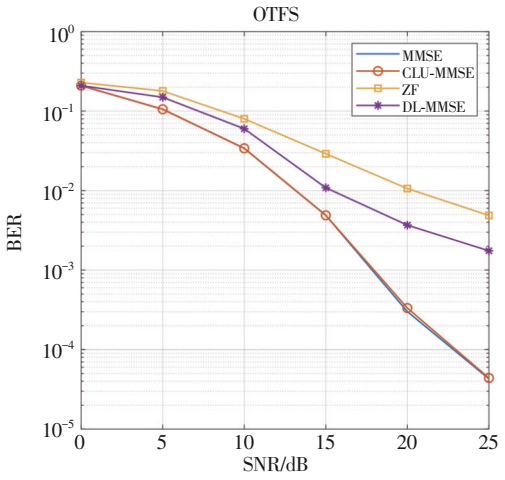
Table 2. Simulation parameters

Parameter	Value
Carrier frequency	4 GHz
Subcarrier spacing (Δf)	15 kHz
Number of subcarriers (M)	32, 16
Number of symbols (N)	16
Modulation type (A)	4QAM, 16QAM
Pulse waveform	Rectangular pulse
Signal path	EVA
Relative movement speed	500 km/h

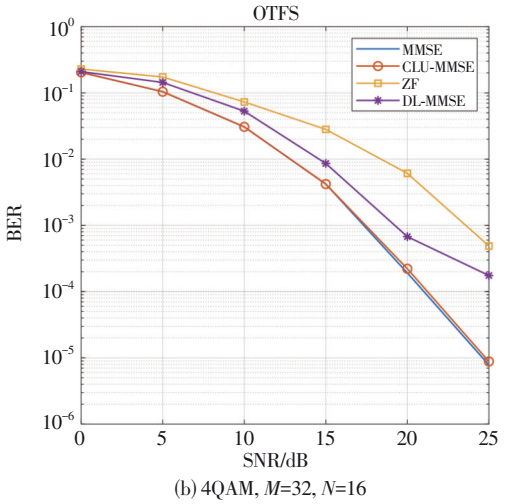
EVA: Extended Vehicular A

5 Conclusions

The traditional OTFS equalization algorithm based on matrix inversion suffers from high computational complexity. In this paper, we address this issue by adopting the time-domain channel matrix with a chunked band structure for equalization. Specifically, the time-domain MMSE equalization matrix is chunked and decomposed multiple times via LU decomposition. Through the factorization into lower and upper triangular matrices, computations involving matrix inversion or matrix-vector/matrix-matrix multiplications can be reformulated as solving linear sys-



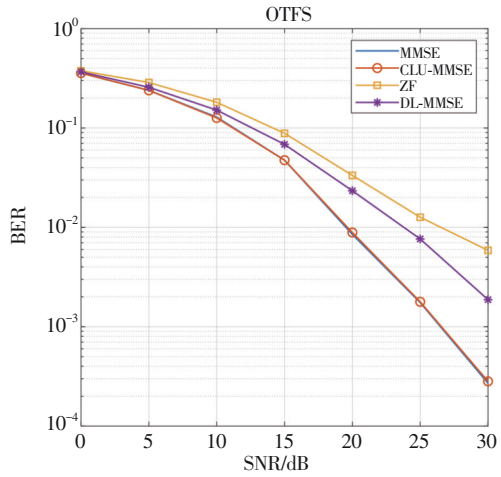
(a) 4QAM, M=N=16



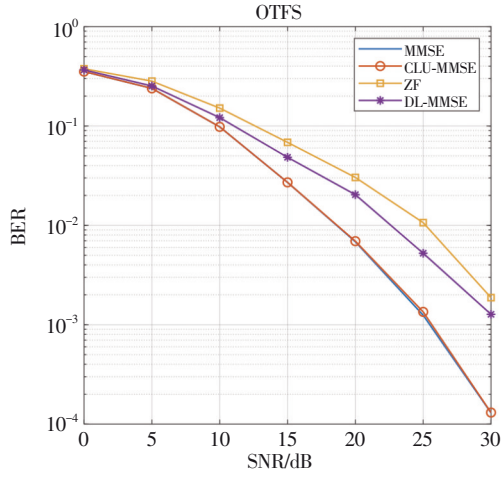
(b) 4QAM, M=32, N=16

BER: bit error rate
CLU: Matrix Chunking Lower and Upper Triangular Decomposition
DL: deep learning
MMSE: minimum mean square error
OTFS: Orthogonal Time Frequency Space
QAM: quadrature amplitude modulation
SNR: signal-to-noise ratio
ZF: zero forcing

Figure 7. BER performance comparison under 4QAM modulation



(a) 16QAM, M=N=16



(b) 16QAM, M=32, N=16

BER: bit error rate
CLU: Matrix Chunking Lower and Upper Triangular Decomposition
DL: deep learning
MMSE: minimum mean square error
OTFS: Orthogonal Time Frequency Space
QAM: quadrature amplitude modulation
SNR: signal-to-noise ratio
ZF: zero forcing

Figure 8. BER performance comparison under 16QAM modulation

tems, thereby avoiding explicit inversion. A recursive algorithm based on forward-backward substitution is then employed to compute each element of the linear equations sequentially. Simulation results demonstrate that the proposed algorithm achieves comparable reliability while reducing the computational complexity of the traditional OTFS-MMSE algorithm, thereby avoiding complex operations such as large-scale matrix multiplication and inversion.

References

[1] He X, Jia H X, Sun Y T, et al. Low-complexity iterative equalization for OTFS based on alternating minimization [J]. Journal of systems engineering and electronics, 2023, 34(4): 851 – 860. DOI: 10.23919/JSEE.2023.000089
[2] Raviteja P, Hong Y, Viterbo E, et al. Effective diversity of OTFS modulation [J]. IEEE wireless communications letters, 2020, 9(2): 249 – 253. DOI: 10.1109/LWC.2019.2951758
[3] Surabhi G D, Augustine R M, Chockalingam A. On the diversity of uncoded OTFS modulation in doubly-dispersive channels [J]. IEEE transactions on wireless communications, 2019, 18(6): 3049 – 3063. DOI: 10.1109/TWC.2019.2909205
[4] Li S Y, Yuan J H, Yuan W J, et al. Performance analysis of coded OTFS systems over high-mobility channels [J]. IEEE transactions on wireless communications, 2021, 20(9): 6033 – 6048. DOI: 10.1109/TWC.2021.3071493
[5] Cheng J Q, Jia C L, Gao H, et al. OTFS based receiver scheme with multi-antennas in high-mobility V2X systems [C]//Proc. IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2020: 1 – 6. DOI: 10.1109/iccworkshops49005.2020.9145313
[6] Shan Y R, Wang F G, Hao Y X, et al. Doppler rate estimation for OTFS via large-scale antenna array [J]. ZTE communications, 2025, 23(1): 115 – 122. doi: 10.12142/ZTECOM.202501015
[7] Surabhi G D, Chockalingam A. Low-complexity linear equalization for OTFS modulation [J]. IEEE communications letters, 2020, 24(2): 330 – 334. DOI: 10.1109/LCOMM.2019.2956709
[8] Tiwari S, Das S S, Rangamgari V. Low complexity LMMSE receiver for OTFS [J]. IEEE communications letters, 2019, 23(12): 2205 – 2209. DOI: 10.1109/LCOMM.2019.2945564
[9] Sun Y T, Jia H X, He X, et al. A low complexity OTFS detection algorithm based on GA-MP [J]. Telecommunication engineering, 2024, 64(2): 288 – 294. DOI: 10.20079/j.issn.1001-893x.220818005
[10] Xiao L X, Li S, Qian Y, et al. An overview of OTFS for Internet of Things: concepts, benefits, and challenges [J]. IEEE Internet of Things journal, 2022, 9(10): 7596 – 7618. DOI: 10.1109/JIOT.2021.3132606
[11] Das S S, Rangamgari V, Tiwari S, et al. Time domain channel estimation and equalization of CP-OTFS under multiple fractional Dopplers and residual synchronization errors [J]. IEEE access, 2021, 9: 10561 – 10576. DOI: 10.1109/ACCESS.2020.3046487

Biographies

Jia Haoxiang (jiahaoxiang@hrbeu.edu.cn) received his BS degree in communication engineering from Shandong University of Science and Technology, China in 2018. He is working toward his PhD degree at the School of Information and Communication Engineering, Harbin Engineering University, China. His main research interests include high-performance coding and modulation schemes.

Zhao Danfeng received his PhD degree in communication systems from Harbin Engineering University, China in 2006, where he is currently a professor. His research interests include network coding, underwater acoustic sensor networks, and high-performance coding and modulation.

Xin Yu graduated from Beijing University of Posts and Telecommunications, China in 2003. He is currently a senior engineer and senior expert in technical pre-research at ZTE Corporation, whose research direction is wireless communication technology. He first proposed the FB-OFDM and GFB-OFDM waveform schemes and has published dozens of papers on waveform technologies. He is now mainly responsible for the pre-research of 6G candidate new waveform technologies.

Hua Jian received his master's degree from Harbin Engineering University, China. He is currently an intermediate engineer at ZTE Corporation. His research interests include phase noise modeling, compensation scheme design, waveform modulation, and other technologies in terahertz communication scenarios.

Carrier Frequency Offset Based Robust Radio Frequency Fingerprint for OFDM Communication in Time-Varying Channels

Liu Gengyi[1], Pan Yijin[1], Wang Junbo[1], Chen Yijian[2], Yu Hongkang[2]

(1. National Mobile Communications Research Laboratory, Southeast University, Nanjing 211111, China;
2. Wireless Product Research and Development Institute, ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202601005

https://kns.cnki.net/kcms/detail/34.1294.TN.20260210.0908.002.html, published online February 10, 2026

Manuscript received: 2025-01-11

Abstract: The radio frequency (RF) fingerprint technique is a robust method for security enhancement of the physical layer by leveraging the unique RF imperfections inherent in various wireless devices. Among these imperfections, the carrier frequency offset (CFO) stands out as a primary RF fingerprint (RFF) of the transmitter, offering the potential to distinguish among different transmitters. However, accurately estimating CFO in time-varying channels poses significant challenges due to multipath effects and Doppler shifts. In this paper, we focus on estimating CFO for wireless device identification in the orthogonal frequency division multiplexing (OFDM) communication system. To achieve precise CFO estimation under time-varying channels, we propose a frequency domain correlation and spline interpolation (FCSI) algorithm. This approach utilizes pilots distributed across different subcarriers to correlate with prior local sequences, facilitating accurate CFO estimation. Classification is then performed based on the Euclidean distance between the prior RFF and the tested RFF dataset. Simulation results demonstrate that the proposed M-consecutive average method effectively reduces the classification error rate in the challenging high-frequency (HF) skywave channel environment.

Keywords: RF fingerprint; RF identification; carrier frequency offset; time-varying channels; OFDM

Citation (Format 1): Liu G Y, Pan Y J, Wang J B, et al. Carrier frequency offset based robust radio frequency fingerprint for OFDM communication in time-varying channels [J]. ZTE Communications, 2026, 24(1): 25 - 33. DOI: 10.12142/ZTECOM.202601005
Citation (Format 2): G. Y. Liu, Y. J. Pan, J. B. Wang, et al., "Carrier frequency offset based robust radio frequency fingerprint for OFDM communication in time-varying channels," *ZTE Communications*, vol. 24, no. 1, pp. 25 - 33, Mar. 2026. doi: 10.12142/ZTECOM.202601005.

1 Introduction

The transmission of wireless communication signals relies on radio frequency (RF) hardware, which is significantly affected by the imperfections of the RF transmitter and receiver. Due to the presence of RF imperfections, the signal received by the receiver inherently carries unique RF fingerprints that can be utilized for transmitter identification. The identification of individual RF transmitters involves extracting and analyzing the unique characteristics of each transmitter through the receiver. This process identifies the different transmitters by analyzing the signals emitted from unknown transmitters. Therefore, the primary focus of RF fingerprint research is to determine which types of RF fingerprints (RFF) are advantageous for the identification of transmitters.

In Ref. [1], RFF extraction methods are categorized into transient-state-based RFF, steady-state-based RFF, and other approaches. Transient-state-based RFF focuses on transition extraction from off to on or the control signal at the transmitter, which occurs before the data transmission of the signal. Ref. [2] extracts the RFF of transmitters from the distance between the start and the end points of the transient signal. Moreover, the normalized amplitude variance, the number of peaks, and the wavelet transform are utilized for the transient RFF. The short-time Fourier transform (STFT) is employed to obtain the spectrum of transient control signals of unmanned aerial vehicles (UAV)[3]. The energy transient is used to extract 15 statistical properties to analyze the control signals of the UAV. Steady-state-based RFF primarily focuses on features extracted from the received modulated signals. Five specific features of RF imperfection are used in a Passive Radiometric Device Identification System (PARADIS) for physical-layer transmitter identification[4]. Differential constellation trace figures (DCTF), carrier frequency offset, modulation offset, and in-phase/quadrature-

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20240723011, National Natural Science Foundation of China under Grant No. 62371123, Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan under Grant No. 20251077, and Research Fund of National Mobile Communications Research Laboratory, Southeast University under Grant No. 2023A03.

phase (IQ) offset extracted from constellation trace figures (CTF) serve as steady-state-based RFFs in Ref. [5]. A hybrid classification method is proposed to classify different transmitters by training and testing these steady-state-based RFFs. Moreover, deep learning significantly impacts RF fingerprint techniques. It is particularly effective for classification tasks due to its adaptability to the features in target signals, making it well-suited for RFF applications. In contrast, methods relying on predefined models or features often struggle to accurately detect sufficient distinguishing characteristics among devices[6]. For instance, Ref. [7] examines two CNN models to assess RF fingerprint effectiveness under various environmental conditions, focusing on factors such as channel influence, the signal-to-noise ratio (SNR), the number of devices, and the training dataset size. Ref. [8] introduces a radio frequency fingerprint identification (RFFI) approach for long range (LoRa) systems, leveraging deep learning to enhance security through the identification of devices' unique hardware features. RFFs can also be extracted from channel-based features or channel fingerprints that rely on location features in static scenarios. The radio signal strength indicator (RSSI), which depends on transmit power and channel attenuation, is used as an RFF feature[9]. The channel impulse response (CIR) [10] and channel frequency response (CFR)[11] represent two kinds of channel fingerprints. Another strategy is the active RFF, which involves the deliberate introduction of controllable imperfections at the transmitter[12] or the insertion of unique "signatures" within IQ signals[13]. This method effectively aids in the differentiation and classification of various transmitters.

However, RFF is highly sensitive to time-varying and multipath channels. Considering high frequency (HF) skywave communication, the extraction of steady-state RFF is significantly challenged by the time-varying feature of wireless channels and the severe multipath effects, adversely affecting extraction precision. HF signals can be transmitted over long distances by the reflection of the ionosphere. The ionosphere moves and changes in density, causing Doppler shifts on the reflected HF signals, which complicates the extraction of RFF features. In time-varying channels, the classification accuracy of the RFF directly related to the received IQ signal will greatly degrade. Apart from the impact of time-varying channels, RFF is highly dependent on the structure and RF components of the RF chain. Different RF chain structures entail different signal processing approaches for baseband signals. For example, a superheterodyne transmitter undergoes more spectrum shift operations than a direct conversion transmitter, leading to more complex scenarios of carrier frequency offset.

In this study, we adopt a direct conversion structure for its simplicity. The baseband signal is modeled to include three types of RF imperfections: CFO, IQ imbalance, and direct current (DC) offset. The contributions of this paper are as follows.

1) We extract carrier frequency offset (CFO) as an RFF in an OFDM-modulated, HF skywave time-varying channel communication system. CFO is a robust RFF in time-varying channels.

2) Traditional CFO estimation algorithms, particularly those based on cyclic prefixes and block pilots, generally perform poorly in time-varying channels[14]. Furthermore, these algorithms have phase ambiguity issues during CFO estimation. To address these challenges, we introduce the frequency domain correlation and spline interpolation (FCSI) algorithm, which estimates CFO in the frequency domain by exploiting the correlation between known local sequences and received pilots across different subcarriers. This approach is designed to be robust against time-varying channels.

3) We utilize the Euclidean distance between the prior RFF and the RFF under test for classification purposes. Moreover, the M-consecutive average method is used for feature post-processing. To validate the proposed estimation and classification method, we simulate a scenario involving the identification of 12 individual transmitters in an HF skywave OFDM communication system.

The rest of this paper is organized as follows: Section 2 introduces the signal model with RF imperfections for HF skywave OFDM systems. Section 3 discusses the proposed FCSI algorithm for estimating CFO. Section 4 details the classification algorithm that utilizes CFO as an RFF to distinguish different devices. Simulation results are presented in Section 5, and the paper concludes with remarks in Section 6.

2 Signal Model

In this section, we analyze an HF skywave OFDM communication system impacted by three types of RF imperfections: CFO, IQ imbalance, and DC offset. We first present the direct conversion RF transmitter structure. Then, we introduce the CFO to evaluate its effect on the baseband signal. Following this, IQ imbalance and DC offset are added to the signal model. Finally, the influence of the HF time-varying channel is considered to accurately simulate the received signal which is used to extract RFF.

The adaptive selection of suitable parameters has enabled the effective implementation of OFDM modulation for wideband HF skywave communication[15]. The OFDM modulation is configured with a cyclic prefix (CP) length of N_{CP}, a set number of subcarriers N_{SC}, and a subcarrier spacing of Δf. Pilots are strategically placed at equal intervals among the subcarriers.

2.1 RF Transmitter Structure

The RF transmitter architecture includes all components, ranging from the digital-to-analog converter (DAC) to the transmitting antenna. As depicted in Fig. 1, within the direct conversion RF chain, the digital signal first undergoes conversion to an analog format by the DAC. It then undergoes low-pass filtering and baseband amplification using an amplifier, which serves to eliminate out-of-band noise and boost the signal's strength. Subsequently, the IQ modulator combines two branches of the IQ signals into one stream. An up-converter is

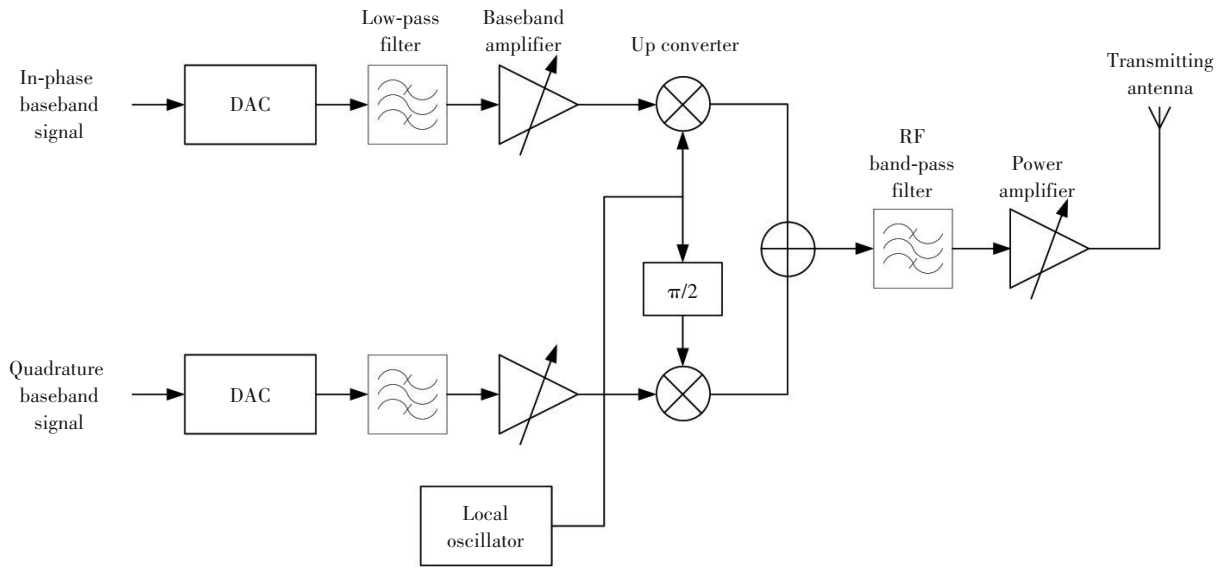


Figure 1. Direct conversion transmitter structure

used for spectrum shift, with the carrier signal produced by a local oscillator. The up-converted signal then passes through an RF band-pass filter and a power amplifier (PA) before being broadcast through the transmitting antenna into the channel.

2.2 Carrier Frequency Offset

CFO typically occurs following the up-conversion process, affected by the characteristics of local oscillators. The natural frequency variation in crystal oscillators causes a persistent difference in the carrier frequencies of the transmitter and receiver. This unique deviation for each transmitter-receiver pair characterizes CFO as a crucial attribute of the transmitter, represented by f_ε. As a result, the up-converted signal x_{CFO}, incorporating the CFO, is expressed as:

$$x_{\mathrm{CFO}} = N\left(x\mathrm{e}^{\mathrm{j}2\pi(f_c+f_\varepsilon)t}\right) \tag{1},$$

where f_c is the carrier frequency and $x = x_r + jx_i$ is the OFDM modulated baseband signal. Considering that the signal transmitted through the channel is a real signal, $N(\cdot)$ signifies taking the real part of a complex number. In the general case, $f_\varepsilon \ll f_c$. Assume that the local oscillator of the down converter in the receiver is perfectly precise, the down-converted signal y is expressed as:

$$y = x_{\mathrm{re}}\mathrm{e}^{-\mathrm{j}2\pi f_c t} \tag{2},$$

where x_{re} is the received signal before the down-conversion. By plugging Eq. (1) into Eq. (2), the down-converted IQ signal y_{IQ} corrupted by carrier frequency offset can be expressed as:

$$y_{\mathrm{IQ}} = \frac{x_r + jx_i}{2}\mathrm{e}^{\mathrm{j}2\pi f_\varepsilon t} + \frac{x_r}{2}\mathrm{e}^{-\mathrm{j}2\pi(2f_c+f_\varepsilon)t} - \frac{jx_i}{2}\mathrm{e}^{-\mathrm{j}2\pi(2f_c+f_\varepsilon)t} \tag{3}.$$

The last two terms of Eq. (3) are filtered out by the low-pass filter. Therefore, Eq. (3) can be modified as:

$$y_{\mathrm{IQ}} = \frac{x_r + jx_i}{2}\mathrm{e}^{\mathrm{j}2\pi f_\varepsilon t} = \frac{x}{2}\mathrm{e}^{\mathrm{j}2\pi f_\varepsilon t} \tag{4}.$$

The spectral shift results in the signal's amplitude being halved and the carrier frequency offset induces a clockwise or counterclockwise rotation, which increases with time in the down-converted IQ signals.

2.3 Other RF Imperfections

Besides CFO, many RF imperfections in the transmitter will corrupt the signal before the signal is transmitted to the channel. Due to the quantization error of DAC and the leakage of the local oscillator, the IQ signals exhibit a notable DC offset denoted by C[16]. Moreover, the gain difference and inexact 90° phase difference may happen to these two signal branches, which will cause IQ imbalance. The IQ parameters are denoted as a_1 and a_2[17]. Therefore, the discrete-time expression of transmitting an IQ signal with DC offset and IQ imbalance can be expressed as follows:

$$X_{\mathrm{IQ}}(n) = \left(a_1 x(n) + a_2 x^*(n)\right) + C \tag{5},$$

where $x(n)$ is the discrete-time expression of the IQ signal from the baseband. Adding the oscillator imperfection CFO, the transmitted signal with RF imperfections after up-conversion

can be modeled as:

$$X_c(n) = N\left(X_{\mathrm{IQ}}(n)e^{\frac{\mathrm{j}2\pi(f_c+f_\varepsilon)n}{F_s}}\right) \tag{6}$$

where $X_c(n)$ is the up-converted signal and F_s is the sample rate.

2.4 HF Skywave Time-Varying Channel

After the up-conversion, the signal goes through the HF skywave channel. In this paper, a wide-sense stationary uncorrelated scattering channel model is considered for the HF skywave channel[18-19]. The intra-path delay can be neglected for the communication signal and the path channel response can be modeled as a time-varying attenuation impulse:

$$c_n(t,\tau) = \alpha_n(t)\delta(\tau) \tag{7}$$

where $\delta(t)$ is the unit-impulse function; τ and $\alpha_n(t)$ are the time delay and impulse of the n-th path, respectively. Assume that the channel is relatively stationary and the relative time delay of each path is independent of time t. The HF skywave channel response can be modeled as:

$$h(t,\tau) = \sum_n c_n(t,\tau-\tau_n) = \sum_n \alpha_n(t)\delta(\tau-\tau_n) \tag{8}$$

We perform the Fourier transform to get the time-varying frequency response:

$$H(f,t) = \int_{-\infty}^{\infty}\sum_n \alpha_n(t)\delta(\tau-\tau_n)\mathrm{e}^{-\mathrm{j}2\pi\tau f}\mathrm{d}\tau = \sum_n \alpha_n(t)\mathrm{e}^{-\mathrm{j}2\pi\tau f} \tag{9}$$

In Eq. (9), the amplitude and phase of the signal are modulated by $\alpha_n(t)$, which is a complex Gaussian random process. The spectrum of $\alpha_n(t)$ follows a Gaussian Doppler spectrum, which can be expressed as:

$$S_G(f) = \frac{1}{\sqrt{2\pi\sigma_G^2}}\mathrm{e}^{-\frac{(f-f_0)^2}{2\sigma_G^2}} \tag{10}$$

where f_0 is the center frequency of the Doppler frequency shift and σ_G is the normalized standard deviation, which is relative to the Doppler spectrum spread f_d.

We can simplify Eq. (8) as:

$$h(t) = \sum_n h_n(t)\delta(t-\tau_n)\mathrm{e}^{\mathrm{j}2\pi f_{dn}t} \tag{11}$$

where $h_n(t)$ and f_{dn} are the n-th path amplitude channel gain and Doppler frequency shift, respectively; f_{dn} is time-varying and Gaussian distributed with f_0 mean and σ_G^2 variance. The discrete-time expression of Eq. (11) is given by:

$$h(n) = \sum_{i=1}^{M_h} h_i(n)\delta(n-D_i)\mathrm{e}^{\frac{\mathrm{j}2\pi f_{di}n}{F_s}} \tag{12}$$

where D_i is the discrete-time form of the time delay and M_h is the number of paths; f_{di} is the Doppler frequency shift corresponding to the i-th path.

Therefore, the received signal $Y(n)$ through the HF skywave channel and additive white Gaussian noise (AWGN) $\omega(n)$ is given by:

$$Y(n) = h(n)*X_c(n) + \omega(n) \tag{13}$$

where * denotes convolution and $\omega(n)\sim N(0,\sigma_N^2)$. Therefore, we can extract RFF from the steady-state received OFDM signal.

3 RFF Extraction Algorithm Design

To precisely estimate the CFO as an RFF in the HF skywave time-varying channel, we propose the FCSI algorithm. The algorithm operates in two stages: first, a frequency domain correlation method is employed to achieve coarse CFO estimation; subsequently, cubic spline interpolation is utilized to refine the estimation for fine CFO correction.

3.1 Frequency Domain Correlation

In OFDM systems, the pilots and the cyclic prefix can be used for the estimation of frequency offsets. In the FCSI algorithm, pilots are employed to correlate with the prior local sequence, thereby extracting the actual frequency offset component. It is assumed that timing synchronization in the OFDM system is perfect. By plugging Eq. (6) and Eq. (13) into Eq. (2), the resultant IQ signal, subjected to a low-pass filtering process, can be written as:

$$Y_{\mathrm{IQ}}(n) = h(n)*\frac{X_{\mathrm{IQ}}(n)}{2}\mathrm{e}^{\frac{\mathrm{j}2\pi f_\varepsilon n}{F_s}} + \omega(n) \tag{14}$$

Without considering the influence of noise, the prior local synchronization sequence, denoted as $x_H(n) = x(n)$, $n = 0,\cdots,L-1$, undergoes a correlation process with the received IQ signal, as specified in Eq. (14). Consequently, the correlation process can be represented as:

$$z(k) = \sum_{n=0}^{L-1} Y_{\mathrm{IQ}}(n+k)x_H^*(n) \tag{15}$$

We assume that the prior local synchronization sequence is the first L symbols of the transmitted IQ signal on the corresponding subcarrier. When $k = 0$, the synchronization is perfect and Eq. (15) can be expanded as follows:

$$z(0) = \sum_{n=0}^{L-1} \frac{h(n)}{2}*X_{\mathrm{IQ}}(n)x^*(n)\mathrm{e}^{\frac{\mathrm{j}2\pi f_\varepsilon n}{F_s}} \tag{16}$$

In the direct conversion communication system, to maintain

communication performance, the IQ amplitude imbalance should be kept below 5 dB, while the phase imbalance should be limited to less than 5°[16], which means that $a_1 \gg a_2$. Moreover, DC offset is calibrated at the receiver and its impact on system performance is smaller than that of IQ imbalance, so that $a_1 \gg C$. Therefore, $X_{\mathrm{IQ}}(n)x^*(n)$ can be expanded as:

$$X_{IQ}(n)x^*(n) = a_1 x(n)x^*(n) + a_2\left[x^*(n)\right]^2 + Cx^*(n) \approx a_1 x(n)x^*(n) = a_1\left|x(n)\right|^2 \quad (17).$$

By plugging Eq. (17) into Eq. (16), $z(0)$ is rewritten as:

$$z(0) = \sum_{n=0}^{L-1}\frac{h(n)}{2} * a_1\left|x(n)\right|^2 \mathrm{e}^{\mathrm{j}2\pi f_\varepsilon \frac{n}{F_s}} \quad (18).$$

In this form, the frequency offset is mainly influenced by the Doppler frequency shift f_{dn} of the HF skywave channel $h(n)$ and the frequency deviation f_ε between the transmitter and receiver. Then we perform N-point fast Fourier transform (FFT) to get the frequency spectrum $Z(f_n)$ of Eq. (18).

$$f_n = \left(n - \frac{N}{2}\right)R_f,\ n = 1,\cdots,N \quad (19),$$

where $R_f = R/N$ is the frequency resolution depending on the symbol rate R and the point number of FFT N. The carrier frequency offset is estimated in the frequency domain, which can be expressed as:

$$F_{\mathrm{coarse}} = \arg\max_{f_n} Z(f_n) \quad (20).$$

Therefore, the coarse estimated frequency offset F_{coarse} of the transmitted signal can be compensated as:

$$Y_{\mathrm{FOcomp}}(n) = Y_{\mathrm{IQ}}(n)\mathrm{e}^{-\mathrm{j}2\pi F_{\mathrm{coarse}}\frac{n}{F_s}} \quad (21).$$

Then we use Eq. (21) to estimate the fine carrier frequency offset.

3.2 Cubic Spline Interpolation

After coarse frequency offset estimation, cubic spline interpolation is used for fine estimation. First, a frequency offset detection range and its step are determined by R_f in Eq. (19). The range is set as $\left[-R_f/2, R_f/2\right]$ and the step is set as R_f/N_R, where N_R is the detection number. Therefore, the detection range can be expressed as:

$$F_{\mathrm{det}}(i) = -\frac{R_f}{2} + (i-1)\frac{R_f}{N_R - 1},\ i = 1,\cdots,N_R \quad (22).$$

Then, we use Eq. (22) to compensate for the frequency offset of the synchronized received IQ sequence in Eq. (21), which can generate N_R IQ signals with different compensated frequency offsets. The inner product of the compensated IQ sequence and the local synchronization sequence is calculated as the correlation peak sequence corresponding to the fine frequency offset detection range. The correlation peak sequence is given by:

$$V_{\mathrm{Peak}}(i) = \sum_{n=0}^{L-1} Y_{\mathrm{FOcomp}}(n)\mathrm{e}^{-\mathrm{j}2\pi F_{\mathrm{det}}(i)\frac{n}{F_s}} x_H^*(n) \quad (23),$$

where $i = 1,\cdots,N_R$. Next, we use the cubic spline interpolation algorithm[20] to fit a finer correlation peak sequence. The detection range $\left[-R_f/2, R_f/2\right]$ is evenly divided into N_S parts and the second derivatives at the boundary $-R_f/2$ and $R_f/2$ are both set to 0. The correlation peak sequence can be interpolated as shown in Fig. 2. We can obtain the finely estimated CFO F_{fine} from the maximum value of the interpolated correlation peak sequence:

$$F_{\mathrm{fine}} = \arg\max_{F_{\mathrm{det}}(i)} V_{\mathrm{Peak}}(i) \quad (24).$$

3.3 FCSI Algorithm

Following the coarse estimation of CFO using frequency domain correlation and subsequent fine CFO estimation through cubic spline interpolation, the accurately estimated CFO F_{est} is obtained by:

$$F_{\mathrm{est}} = F_{\mathrm{coarse}} + F_{\mathrm{fine}} \quad (25),$$

which is the RFF of the transmitter. The flowchart for the FCSI algorithm is depicted in Fig. 3.

As discussed in Section 2, CFO is associated with both the Doppler frequency shift and the frequency deviations in oscillators between the transmitter and the receiver. In HF communication systems, the frequency shift caused by Doppler effects in

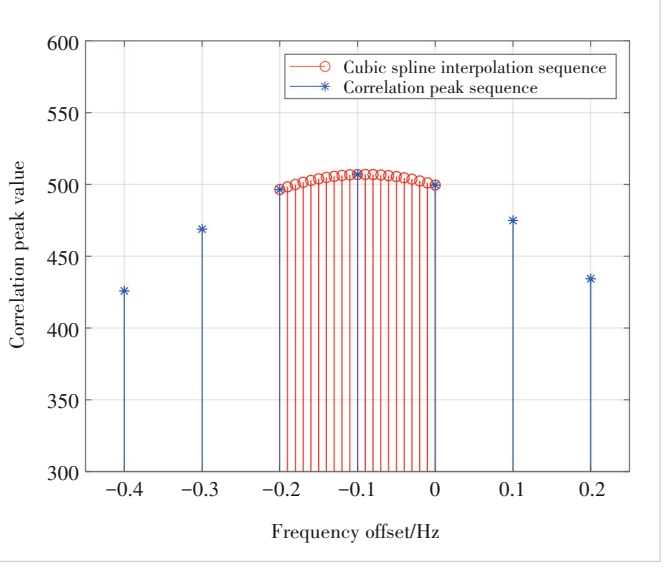


Figure 2. Correlation peak sequence interpolation

time-varying channels is considerably smaller than oscillator frequency deviations. Consequently, frequency domain correlation can accurately determine the primary frequency offset resulting from these deviations. Cubic spline interpolation can estimate fine frequency offsets, thereby making the FCSI algorithm an accurate approach for CFO estimation. Moreover, the CFO is estimated and compensated during the time and frequency synchronization of the received signal, ensuring that this RFF extraction design will not introduce additional computational load. As a result, CFO serves as an effective RFF for distinguishing between different transmitters in HF skywave OFDM communication systems. The complete FCSI algorithm for CFO estimation is summarized as follows.

Algorithm 1. FCSI algorithm for CFO estimation

Input: $Y_{\mathrm{IQ}}(n)$ and $x_H(n)$.

Use Eq. (15) to correlate $Y_{\mathrm{IQ}}(n)$ and $x_H(n)$.

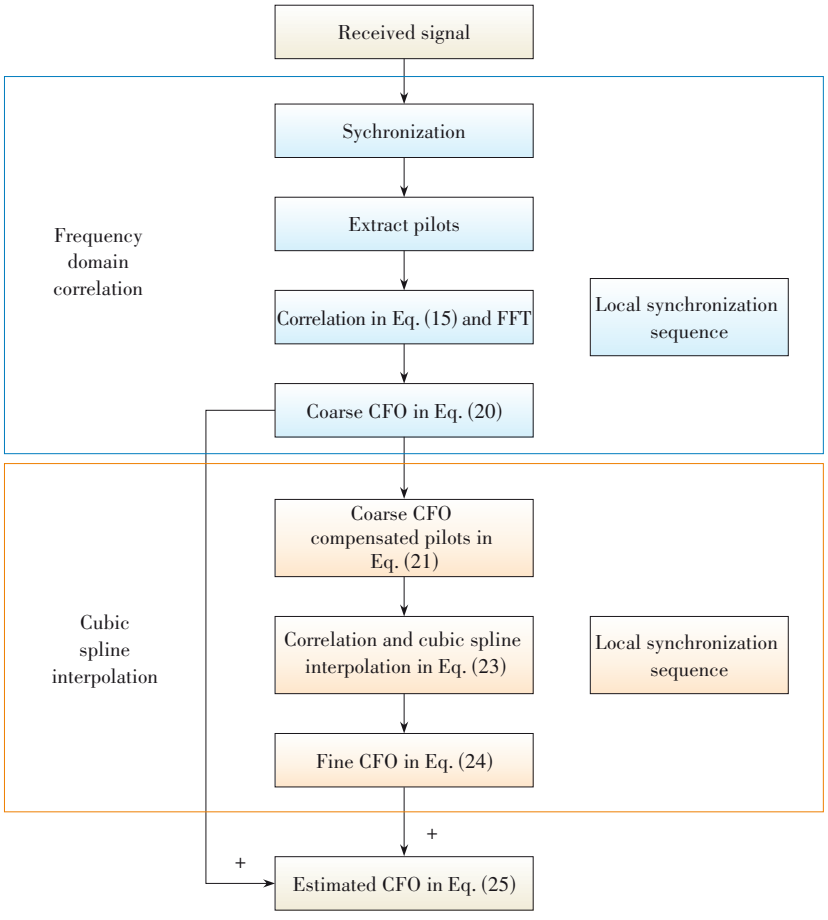
FFT is applied to find the coarse frequency offset by Eq. (18).

Obtain the coarse CFO by Eq. (20).

Obtain the correlation peak sequence in Eq. (23) corresponding to Eq. (22).

Use cubic spline interpolation to estimate fine CFO in Eq. (24) and total estimated CFO in Eq. (25).

Output: Total estimated frequency offset F_{est}.



CFO: carrier frequency offset FFT: fast Fourier transform

Figure 3. Frequency domain correlation and spline interpolation algorithm flowchart

4 Classification Algorithm

In this paper, the classification algorithm employs the Euclidean distance to classify RFF and identify the transmitter of a steady-state received signal. The precise relative deviation ΔR_i, $i = 1,\cdots, N_T$ of the oscillators between the transmitter and the receiver used in the classification test can be measured through the direct connection, where N_T is the number of transmitters. Therefore, the prior RFF of each transmitter can be denoted as $\Phi_i = f_c \Delta R_i$, $i = 1,\cdots, N_T$, which is the precise CFO of the 12 transmitters Tx_i, $i = 1,\cdots, N_T$.

Next, we can use the same receiver to receive the signal from different unidentified devices Tx_j, $j \in 1,\cdots, N_T$, from which the CFO $\Phi_{\mathrm{test},j}$ is extracted and identified through the Euclidean distance d_{ji} between the extracted CFO and the prior RFFs Φ_i, which is given by:

$$d_{ji} = |\Phi_{\mathrm{test},j} - \Phi_i|,\ i = 1,\cdots, N_T \quad (26),$$

where the transmitter number corresponding to the prior RFF with the minimal Euclidean distance is denoted as the judgment result k:

$$k = \arg\min_i d_{ji},\ i = 1,\cdots, N_T \quad (27).$$

A classification error is noted when $k \neq j$, indicating an error associated with the received signal from Tx_j. This framework facilitates the classification of signals from each transmitter via the estimated CFO. The number of classification errors and the total test signals from Tx_j are represented by N_{error} and N_{test}, respectively. Therefore, the classification error rate (CER) can be expressed as:

$$\mathrm{CER} = \frac{N_{\mathrm{error}}}{N_{\mathrm{test}}} \quad (28).$$

In this scheme, a CFO extracted from an OFDM modulated frame constitutes one classification sample. Alternatively, the mean of M CFOs extracted from M consecutive frames in the time domain can be considered as a classification sample, where M denotes the sample period. This M-consecutive average method reduces the influence of estimation deviations on classification. The averaged RFFs are represented as

$\bar{\Phi}_{j,l}$, $j \in 1,\cdots,N_T$, $l = 1,2,\cdots,N_F/M$ and $N_{\text{test}} = N_F N_T/M$, where N_F signifies the total number of transmitted frames. By utilizing averaging to reduce the error in frequency offset estimation, this method enables more accurate classification of different transmitters, proving particularly effective in HF skywave channels affected by Doppler spread.

5 Simulation Results

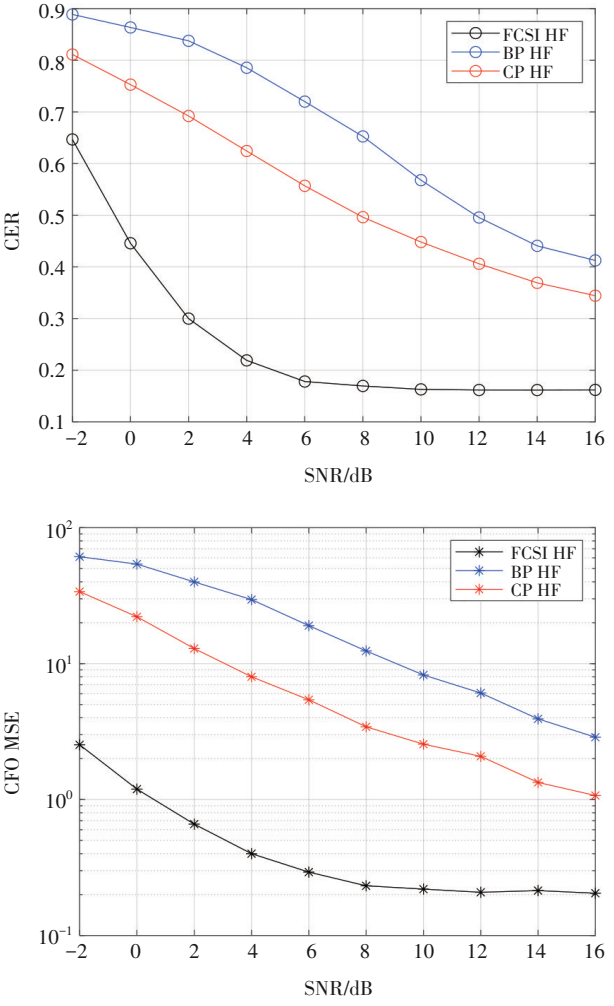
The classification test is conducted with $N_T = 12$ devices, and the simulation parameters are detailed in Table 1. Typically, the frequency tolerance of crystal oscillators varies between −2 and 2 parts per million (ppm). As a result, the corresponding CFO spans from −32 Hz to 32 Hz. Moreover, the ionospheric Doppler shift varies from −1 Hz to 1 Hz in general in the HF skywave channel. The proposed FCSI CFO estimation method is labeled as FCSI. For comparison, the performance of both block pilot (BP)-based and CP-based CFO estimation is simulated, which are labeled as BP and CP, respectively.

A single CFO is estimated from one frame with a synchronization header in the different subcarriers. We conduct CFO estimations using 8 pilots in corresponding subcarriers from 1 000 frames for each transmitter. Following this, we estimate the CFO for $N_F = 1\,000$ frames per transmitter to conduct the classification test. Eqs. (26) and (27) are employed to calculate the number of classification errors, with $N_{\text{test}} = N_F N_T/M$.

Figs. 4 and 5 indicate that the CER is closely related to the mean squared error (MSE) of the CFO estimation algorithm in both HF and AWGN channels. Notably, the CER achieved by the FCSI-based estimation is substantially lower than that obtained using CP-based and BP-based methods, highlighting the superior performance of the FCSI algorithm. Fig. 5 presents the simulated CER in an AWGN channel. The FCSI algorithm demonstrates superior CER performance compared with both BP-based and CP-based algorithms. In the AWGN channel, the absence of Doppler spread means that the estimated frequency offset is free from frequency errors attributable to the channel. The absence of Doppler spread results in an overall CER that is lower than that observed in HF channel, as illustrated in Fig. 4. The CFO estimation in the HF skywave channel is affected by the Gaussian distribution of Doppler shifts, leading to the convergence of the MSE of the FCSI algorithm as the SNR increases, which in turn prevents the CER from decreasing with higher SNR. The MSE of CFO at an SNR of 16 dB is 0.204 7, which is very close to the set Gaussian distributed Doppler shift variance $\sigma_G^2 = 0.25$. This result demonstrates the effectiveness of the FCSI method for CFO estimation. In such scenarios, the actual CFO difference between devices is smaller than the Doppler shift induced by the HF skywave channel, complicating the classification of devices with similar CFOs. Therefore, we need to mitigate the impact of Doppler spread. Considering the distribution of the Doppler

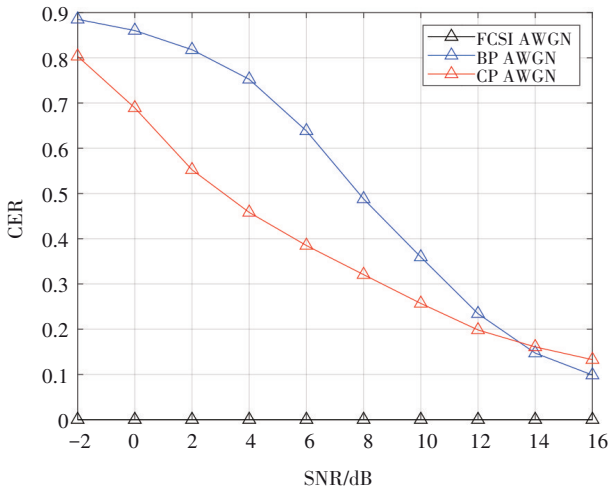
Table 1. Simulation system parameters of the classification test

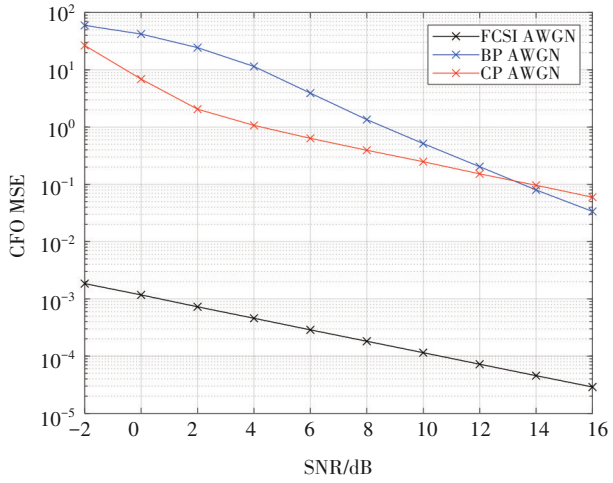
Parameter	Value
Device number N_T	12
HF channel multipath number	2
Multipath delay τ	2 ms
Standard deviation of Doppler shift σ_G	0.5
Carrier frequency f_c	16 MHz
System bandwidth BW	320 kHz
Subcarrier spacing Δf	500 Hz
Subcarrier number N_{SC}	512
CP length N_{CP}	8
Guard band subcarrier number	50



BP: block pilot
CER: classification error rate
CFO: carrier frequency offset
CP: cyclic prefix
FCSI: frequency domain correlation and spline interpolation
HF: high frequency
MSE: mean squared error
SNR: signal-to-noise ratio

Figure 4. CER and MSE of CFO estimation with different SNRs in HF channel





AWGN: additive white Gaussian noise
BP: block pilot
CER: classification error rate
CFO: carrier frequency offset
CP: cyclic prefix
FCSI: frequency domain correlation and spline interpolation
MSE: mean squared error
SNR: signal-to-noise ratio

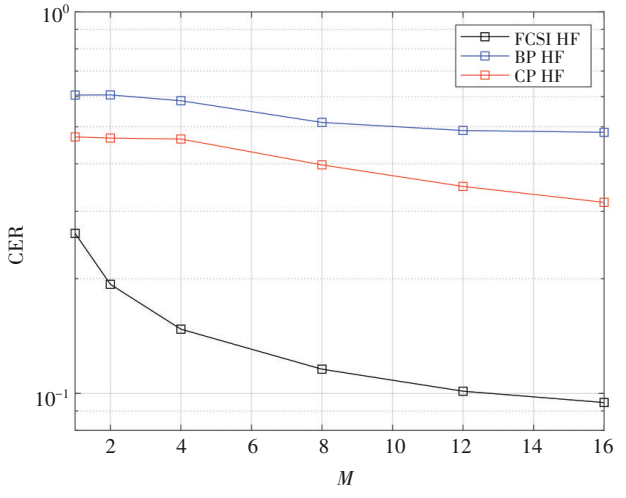
Figure 5. CER and MSE of CFO estimation with different SNRs in AWGN channel

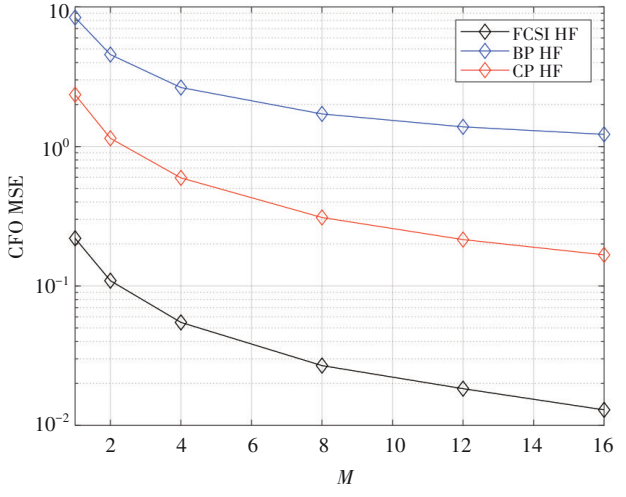
shift, we can use the M-consecutive average method to leverage statistical properties and appropriately reduce the randomness of the Doppler shift.

The M-consecutive average method is applied to the classification in the HF skywave channel. As shown in Figs. 8 and 9, the CER and MSE of the CFO estimation decrease with the increasing M, demonstrating the effectiveness of this method in reducing the CER of classification in the channel. When M = 16, the CER of the FCSI algorithm can reach 9.47% at an SNR of 10 dB. The experimental results indicate that leveraging the statistical properties of CFO effectively mitigates the impact of the Doppler shift on device identification, significantly reducing the CER.

6 Conclusions

We introduce DC offset, IQ imbalance, and CFO into the baseband signal to simulate the impact of RF imperfections on an HF skywave time-varying channel in the OFDM system. We employ an HF skywave channel model characterized by random Doppler frequency shifts and propose the FCSI algorithm to estimate the CFO of the transmitter as the RFF. Subsequently, the Euclidean distance between prior and test RFFs is used to classify 12 transmitters. The simulation results demonstrate that the FCSI method achieves a significantly lower CER than those obtained by CP- and BP-based CFO estimation methods. Furthermore, the M-consecutive average method proves to be an effec-





BP: block pilot
CER: classification error rate
CFO: carrier frequency offset
CP: cyclic prefix
FCSI: frequency domain correlation and spline interpolation
HF: high frequency
MSE: mean squared error

Figure 8. CER and MSE of CFO estimation versus M in HF channel

tive strategy for reducing the CER in the HF skywave channel to resist the influence of Doppler spread on RFF extraction.

References

[1] Soltanieh N, Norouzi Y, Yang Y, et al. A review of radio frequency fingerprinting techniques [J]. IEEE journal of radio frequency identification, 2020, 4(3): 222 – 233. DOI: 10.1109/JRFID.2020.2968369
[2] Bonne R K, Capkun S. Implications of radio fingerprinting on the security of sensor networks [C]//The Third International Conference on Security and Privacy in Communications Networks and the Workshops. IEEE, 2007: 331 – 340. DOI: 10.1109/SECCOM.2007.4550352
[3] Ezuma M, Erden F, Kumar A C, et al. Detection and classification of UAVs using RF fingerprints in the presence of Wi-Fi and bluetooth interference [J]. IEEE open journal of the communications society, 2020, 1: 60 – 76. DOI: 10.1109/OJCOMS.2019.2955889
[4] Brik V, Banerjee S, Gruteser M, et al. Wireless device identification with radiometric signatures [C]//The 14th ACM International Conference on Mobile Computing and Networking. ACM, 2008: 116 – 127. DOI: 10.1145/1409944.1409959
[5] Peng L N, Hu A Q, Zhang J Q, et al. Design of a hybrid RF fingerprint extraction and device classification scheme [J]. IEEE Internet of Things journal, 2019, 6(1): 349 – 360. DOI: 10.1109/JIOT.2018.2838071
[6] Merchant K, Revay S, Stantchev G, et al. Deep learning for RF device fingerprinting in cognitive communication networks [J]. IEEE journal of selected topics in signal processing, 2018, 12(1): 160 – 167. DOI: 10.1109/JSTSP.2018.2796446
[7] Jian T, Rendon B C, Ojuba E, et al. Deep learning for RF fingerprinting: a massive experimental study [J]. IEEE Internet of Things magazine, 2020, 3(1): 50 – 57. DOI: 10.1109/IOTM.0001.1900065
[8] Shen G X, Zhang J Q, Marshall A, et al. Radio frequency fingerprint identification for LoRa using deep learning [J]. IEEE journal on selected areas in communications, 2021, 39(8): 2604 – 2616. DOI: 10.1109/JSAC.2021.3087250
[9] Yiu S, Dashti M, Claussen H, et al. Wireless RSSI fingerprinting localization [J]. Signal processing, 2017, 131: 235 – 244. DOI: 10.1016/j.sigpro.2016.07.005
[10] Liu F J, Wang X B, Primak S L. A two dimensional quantization algorithm for CIR-based physical layer authentication [C]//International Conference on Communications (ICC). IEEE, 2013: 4724 – 4728. DOI: 10.1109/ICC.2013.6655319
[11] Xiao L, Li Y, Han G A, et al. PHY-layer spoofing detection with reinforcement learning in wireless networks [J]. IEEE transactions on vehicular technology, 2016, 65(12): 10037 – 10047. DOI: 10.1109/TVT.2016.2524258
[12] Sankhe K, Belgiovine M, Zhou F, et al. ORACLE: optimized radio classification through convolutional neural networks [C]//IEEE Conference on Computer Communications. IEEE, 2019: 370 – 378. DOI: 10.1109/infocom.2019.8737463
[13] Mohanti S, Soltani N, Sankhe K, et al. AirID: injecting a custom RFF for enhanced UAV identification using deep learning [C]//Global Communications Conference. IEEE, 2020: 1 – 6. DOI: 10.1109/GLOBECOM42002.2020.9322561
[14] Van de Beek J J, Sandell M, Borjesson P O. ML estimation of time and frequency offset in OFDM systems [J]. IEEE transactions on signal processing, 1997, 45(7): 1800 – 1805. DOI: 10.1109/78.599949
[15] Yu X L, Lu A N, Gao X Q, et al. HF skywave massive MIMO communication [J]. IEEE transactions on wireless communications, 2022, 21(4): 2769 – 2785. DOI: 10.1109/TWC.2021.3115820
[16] Gu Q Z. RF system design of transceivers for wireless communications [M]. New York: Springer-Verlag, 2005. DOI: 10.1007/b104642
[17] Zhang W C, de Lamare R C, Pan C H, et al. Widely linear precoding for large-scale MIMO with IQI: algorithms and performance analysis [J]. IEEE transactions on wireless communications, 2017, 16(5): 3298 – 3312. DOI: 10.1109/TWC.2017.2679706
[18] Watterson C, Juroshek J, Bensema W. Experimental confirmation of an HF channel model [J]. IEEE transactions on communication technology, 1970, 18(6): 792 – 803. DOI: 10.1109/TCOM.1970.1090438
[19] Mastrangelo J F, Lemmon J J, Vogler L E, et al. A new wideband high frequency channel simulation system [J]. IEEE transactions on communications, 1997, 45(1): 26 – 34. DOI: 10.1109/26.554283
[20] Mckinley S, Levine M. Cubic spline interpolation [M]. [S.l.]: College of the Redwoods, 1998
[21] Zeng C, Wang J B, Xiao M, et al. Task-oriented semantic communication over rate splitting enabled wireless control systems for URLLC services [J]. IEEE transactions on communications, 2024, 72(2): 722 – 739. DOI: 10.1109/TCOMM.2023.3325901
[22] Gong P Y, Zhang G D, Zhang G. Research on fall detection system based on commercial Wi-Fi devices [J]. ZTE communications, 2023, 21(4): 60 – 68. DOI: 10.12142/ZTECOM.202304008
[23] Wu J Y, Wang C Y, Xie L. Device-free in-air gesture recognition based on RFID tag array [J]. ZTE communications, 2021, 19(3): 13 – 21. DOI: 10.12142/ZTECOM.202103003
[24] Zhu Y T, Li Z, Zhang H T. Robust beamforming under channel prediction errors for time-varying MIMO system [J]. ZTE communications, 2023, 21(3): 77 – 85. DOI: 10.12142/ZTECOM.202303011

Biographies

Liu Gengyi received his BS degree from the School of Information Science and Engineering, Southeast University, China in 2022, where he is pursuing his ME degree. His research interests include reconfigurable intelligent surfaces (RIS), mmWave communications, and radio frequency fingerprint (RFF) identification.

Pan Yijin received her BS and MS degrees in communication engineering from Chongqing University, China in 2011 and 2014, respectively, and PhD degree from the School of Information Science and Engineering, Southeast University, China in 2018, where she is currently an associate professor. She was a recipient of Royal Society Newton International Fellowship (2019 – 2021), UK. Her research focuses on key technologies for future communication networks, specifically for edge intelligence-enhanced communication schemes. She serves as a reviewer and a TPC member for prestigious international journals and conferences.

Wang Junbo (jbwang@seu.edu.cn) received his BS degree in computer science from Hefei University of Technology, China in 2003 and PhD degree in communications engineering from Southeast University, China in 2008. From October 2008 to August 2013, he was with Nanjing University of Aeronautics and Astronautics, China. From February 2011 to February 2013, he was a post-doctoral fellow with the National Laboratory for Information Science and Technology, Tsinghua University, China. Since August 2013, he has been an associate professor with the National Mobile Communications Research Laboratory, Southeast University. From October 2016 to September 2018, he was awarded the Marie Skłodowska-Curie Actions Fellowships and a worked as a research fellow with the University of Kent, UK. His current research interests include cloud radio access networks, mmWave communications, and wireless optical communications.

Chen Yijian graduated from Central South University, China. He is currently working at ZTE Corporation. His research interests include reconfigurable intelligent metasurfaces, extremely large-scale MIMO technology, and electromagnetic information theory.

Yu Hongkang received his BS degree from Beijing Jiaotong University, China in 2016 and PhD degree from Peking University, China in 2021. He is currently an engineer with ZTE Corporation. His current research interests include mmWave, massive MIMO, and channel estimation.

Key Technologies for AI-Driven Network Traffic Classification Workflow and Data Distribution Shift

Zhao Jianchao[1], Geng Zhaosen[1], Li Zeyi[2], Wang Pan[3]

(1. Cable Products Business Department, ZTE Corporation, Shenzhen 518057, China;
2. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
3. School of Modern Posts, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

DOI: 10.12142/ZTECOM.202601006

https://kns.cnki.net/kcms/detail/34.1294.TN.20260112.1759.002.html, published online January 13, 2026

Manuscript received: 2025-01-11

Abstract: With the evolution of next-generation network technologies, the complexity of network management has significantly increased, and the means of network attacks are diversified, bringing new challenges to network traffic classification. This paper presents a general AI-driven network traffic classification workflow and elaborates on a traffic data and feature engineering framework. Most importantly, it analyzes the concept and causes of data distribution shifts in network traffic, proposing detection methods and countermeasures. Experimental results on real traffic collected at different time intervals show that application evolution can induce data distribution shifts, which in turn lead to a noticeable degradation in traffic classification performance. Comparative drift detection experiments further confirm that such shifts are more evident over long-term intervals, while short-term traffic remains relatively stable. These findings demonstrate the necessity of incorporating drift-aware mechanisms into AI-driven network traffic classification systems.

Keywords: traffic classification; traffic identification; deep learning; data distribution shift; concept shifting

Citation (Format 1): Zhao J C, Geng Z S, Li Z Y, et al. Key technologies for AI-driven network traffic classification workflow and data distribution shift [J]. ZTE Communications, 2026, 24(1): 34 - 44. DOI: 10.12142/ZTECOM.202601006

Citation (Format 2): J. C. Zhao, Z. S. Geng, Z. Y. Li, et al., "Key technologies for AI-driven network traffic classification workflow and data distribution shift," *ZTE Communications*, vol. 24, no. 1, pp. 34 - 44, Mar. 2026. doi: 10.12142/ZTECOM.202601006.

1 Introduction

Network traffic classification (TC), as an essential means for network management and security, has received significant attention from academia and industry since the late 1990s. It has been well applied in quality of service/quality of experience (QoS/QoE) management, network resource optimization, congestion control, intrusion detection, etc. [1] With the rapid development of new-generation network technologies (B5G/6G, Internet of Things, celestial and terrestrial integrated networks, etc.), network technology is moving towards high autonomy of self-healing, self-management, self-optimization and self-protection, and the network traffic classification technology plays a key role as one of the decision-making tools for the network service and security management[2]. However, with the ubiquitous access of a large number of heterogeneous terminals, the network shows a high degree of dynamism, heterogeneity and complexity. This brings new challenges to network traffic classification technology. In particular, the frequent upgrading of legacy applications, the continuous emergence of new applications, and the gradual downgrading of silent applications have left network TC technology "one step behind", unable to keep pace with dynamic application change.

The development of TC technology has roughly gone through three stages. The first phase is based on ports or deep packet inspection (DPI) to achieve TC. However, as applications increasingly adopt technologies such as tunnelling, encryption, and random ports, coupled with the security risk of user privacy leakage, these technologies quickly become ineffective. The second phase mainly uses machine learning (ML) based methods to learn the intrinsic laws of different business/application/attack traffic characteristics and implement the delineation of various applications in the data space to achieve traffic classification. However, such methods need to extract high-quality traffic features as the training basis of ML, and

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20220607009.

the extraction and selection of these features are highly dependent on the experience of network experts and time-consuming. With the rapid development of cloud computing, big data, especially deep learning (DL) and high-performance computing technologies, feature learning of massive traffic data has become possible, bringing new extension space to the TC field. DL has three excellent features: automatic feature extraction that can reveal more profound data laws, a large number of mature applications in vision/image/text/voice models, and the ability to address the gaps in ML-based TC methods. In recent years, DL-based TC techniques (hereafter referred to as DL-TC; AI-TC in subsequent text denotes ML/DL-TC) have been proposed and sparked new research enthusiasm. These include methods based on the convolutional neural network (CNN), the auto encoder (AE), the multilayer perceptron (MLP), long short-term memory (LSTM), and the generative adversarial network (GAN), which have achieved better classification performance than ML-TC[3-9].

Many research institutes and personnel are actively researching and developing efficient AI-TC algorithms and models, but there are still many problems for practical deployment and operation. Table 1 shows the comparison between academia and industry in AI-TC.

The data distribution shift has now become a research hotspot in AI, but most studies focus on classical AI fields such as images, vision, speech, and texts, and few studies have ventured into network traffic classification. This paper focuses on the data shift problem in AI-TC, describing its concepts, challenges, and critical techniques. The contributions of this paper are as follows:

1) A generic framework for AI-driven sustainable learning of network traffic classification systems is proposed;

2) The existing research progress on the data distribution shift problem is summarized, and a data distribution shift detection method for AI-TC is proposed;

3) The current advances in continuous learning research are summarized and a continuous learning method for AI-TC is presented;

4) The challenges and future research directions in data distribution shift detection and continuous learning within the context of AI-TC are identified.

2 A Generic Framework for Sustainable Learning of AI-TC System

2.1 AI-TC System Framework

To cope with the three major challenges of dynamic changes in the network environment, rapid evolution of business applications, and continuous upgrading of privacy protection in the new generation of communication networks, the AI-TC system has to be an iterative optimization process with continuous learning. The generic end-to-end AI-TC workflow is illustrated in Fig. 1. From the perspective of an end-to-end machine learning lifecycle, a common framework for sustainable learning AI-TC classification systems includes the definition of AI-TC classification requirements/design principles, traffic data engineering (TDE), feature engineering and model development and evaluation, model interpretation, deployment, model monitoring, and continuous learning phases. The separate phases are detailed below:

• Design principle: It primarily aims to define detailed classification requirements and development principles for AI-TC. These requirements include classification granularity, application scenarios, data sources, and real-time/non-real-time considerations, while the development principles encompass reliability, robustness, security, adaptability, etc.

• TDE: It builds datasets for training and testing. This includes considerations for data sources (baseline data sources, real-time data sources, etc.), traffic collection, traffic sampling, preprocessing (background traffic, redundant data, etc.), and traffic labeling (including both manual and ML-based labeling). Furthermore, after model deployment, it is essential to continuously gather and sample new traffic sample representa-

Table 1. Comparison between academia and industry in AI-TC

	Academia	Industry
Classification requirements	Modeling on training datasets for optimal performance	Comprehensive study of training/classification costs, efficiency, continuous operation, credibility, etc.
End-to-end TC	Focus on the modeling process before the TC model is deployed	Increased focus on monitoring optimization of TC models after deployment
Training data	Static/obsolete, well labeled, noise known	Continuous/changing, no/wrong labeling, noise unknown
Costs	Mostly unconcerned	Great concern
Data distribution shift	Mostly unconcerned	Great concern
Continuous learning	Mostly unconcerned	Great concern
Interpretability	Mostly unconcerned	Consideration
Computational complexity	Focus more on training fast	More concerned with reasoning fast

TC: traffic classification

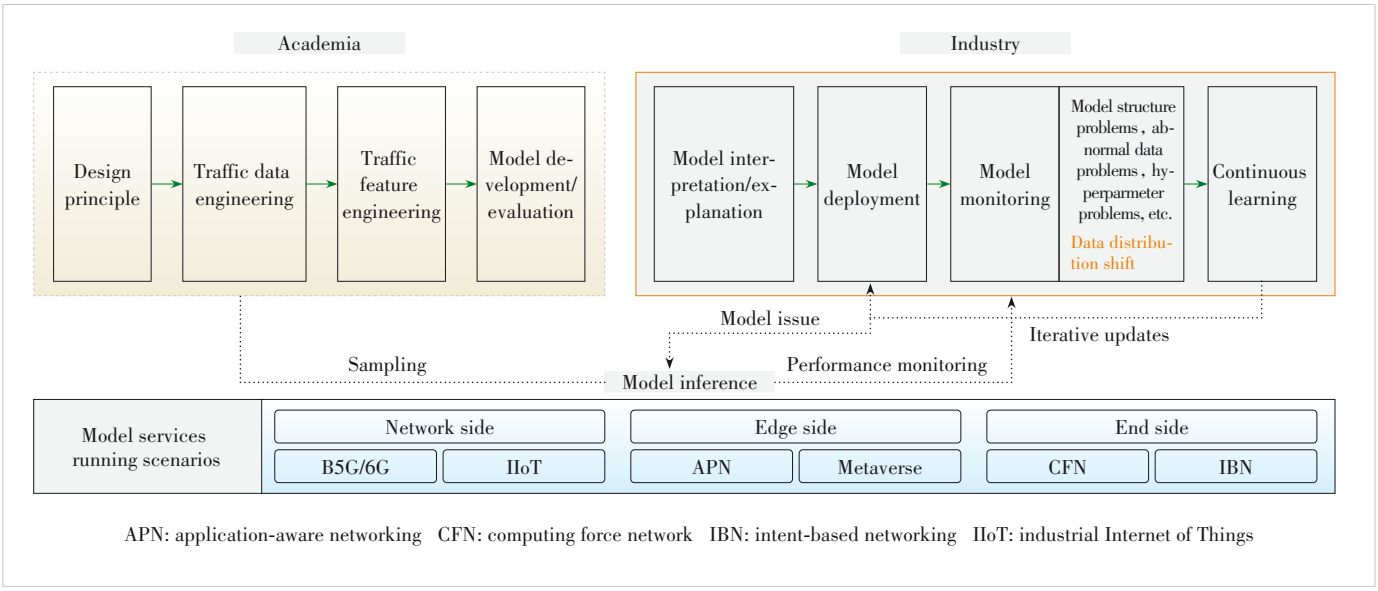


Figure 1. Generic end-to-end AI-TC workflow

tives of typical scenarios (e.g., through active learning) to support ongoing learning.

• Traffic feature engineering (TFE): It performs feature extraction, selection, representation, and reduction for network traffic data to construct an optimal feature subset in accordance with the design principles of the AI-TC classification system.

• Model development/evaluation: This session involves the choice of learning methods (supervised, semi-supervised, unsupervised, and weakly supervised), training methods (centralized/distributed training, federated learning, etc.), whether to pre-train or not, and whether to use the classical model for transfer learning or not. In addition, based on the TC design principles, the corresponding performance evaluation metrics are defined, including accuracy, precision, recall, F1 score, AUC, etc. In addition, the computational complexity, time complexity, computational resources (CPU/memory/flash memory) and time (training/inference) required for training/reasoning of the model are also considered.

• Model interpretation/explanation (XAI): In a narrow sense, model interpretability mainly solves the trustworthiness of classification model users (e.g., operators). Specifically, it focuses on solving the "black box" problem of AI-TC models so that users of the classification model (e.g., operators) can trust the model and have confidence in using the model. In a broader sense, model interpretability also includes transparency and fairness. The former primarily aims to provide model transparency for developers, turning the "black box" into a "white box", which is convenient for model problem diagnosis and iterative optimization. The latter is to detect whether the model is biased, e.g., whether the classification model favors high-value users while neglecting low-value users, especially in the application scenario of bandwidth guarantee based on refined application classification.

• Model deployment: According to different application scenarios, the reasoning environment of classification models can be divided into the network side, the edge side and the terminal side. The model deployment includes model sending methods (pull/push/subscribe, etc.), update strategies, training participation modes (e.g., the federated learning mode), traffic sampling methods (e.g., active learning), and performance parameter reporting.

• Model monitoring: By comprehensively monitoring the status of the classification model and the real-time data flow, this link can detect key issues such as classification system failure and classification model degradation promptly, thereby triggering continuous learning and driving a new round of iterative updates. This is crucial to the robustness of the entire classification system. In addition to general hardware/software failures of the classification system, the scope of monitoring also includes specific machine learning failures such as hyperparameter problems and abnormal data problems. The most important is the data distribution shift, which is particularly severe in network traffic classification.

• Continuous learning: This session is triggered by model monitoring, which initiates a series of iterative optimization and continuous learning from TDE, TFE, as well as model development evaluation, deployment, etc., which is the key to keeping the whole classification system with high adaptability, robustness and reliability.

The following content focuses on an in-depth elaboration of two issues: data distribution shifts and continuous learning. Their concepts, existing technologies, and applications in the context of AI-TC are introduced, and open questions and future research directions are presented.

2.2 Flow Data and Feature Engineering

The traffic data distribution shift problem is inextricably linked to traffic data and feature engineering, and this paper proposes an AI-driven workflow for network traffic classification data and feature engineering, as shown in Fig. 2.

1) Data source: a combination of baseline data and real-time streaming data. Baseline data mainly consists of public and self-built private datasets. In addition, real-time streaming data from the real network being served is also collected to meet the need for continuous learning of new traffic data class distributions.

2) Traffic acquisition and preprocessing: Considering the huge amount of real-time streaming data, it is necessary to design a reasonable traffic sampling strategy to collect representative traffic samples. In addition, pre-processing, i. e., "traffic distilling", is also required because real-time flow data often contains background traffic, redundant groups and other noisy data.

3) Traffic labeling: The traffic is labeled with business types, applications, or attacks, i.e., the ground truth, which is an extremely critical step in dataset construction and will directly affect the performance of the AI-TC model. Developing the automated traffic annotation capability independent of network experts is the future development trend in traffic labelling[10].

4) Feature extraction, selection, and compression: It refers to extracting representative flow features from raw packets, whether using ML or DL. Flow features are mainly classified into raw packet bytes and flow attributes. The former is often used for packet-grained classification in real-time scenarios, while the latter is classified into packet-, flow-, and session-level features according to the granularity of the flow attributes, which are primarily in the form of spatial (packet length, number of packets, flow length, etc.), temporal (inter-packet time interval, flow duration, etc.) and statistical features (expectation, variance, etc.). A typical set of network traffic features is shown in Table 2. To reduce the complexity of the model, it is also necessary to select the most representative characteristics according to certain principles, which involves the feature selection method[11]. The features need to be compressed to further reduce the model parameters and the complexity of model training, usually using AE, principal component analysis (PCA), and other dimensionality reduction methods.

5) Feature representation: After extracting traffic features for model training, it is often necessary to choose a suitable way to express the traffic feature information. Common methods of expression include 2D vectors, images, byte sequences, graphs, etc. This forms a feature set for model training.

6) Data storage and retrieval: Traditional network traffic datasets are often stored as PCAP raw files or stream feature csv files. However, data formats, models, storage, and retrieval methods will face new challenges when building large-scale massive datasets.

3 Data Distribution Shift in AI-TC

3.1 Concept

After the AI-TC classification model is deployed to the real network environment, the data distribution of the application traffic becomes different from that during the training period, which is called the data distribution shift (DDS). One fundamental assumption of ML systems is that the training set and unseen data come from the same static distribution. In the model development and evaluation phase, the testing set represents the unseen data, and the model's performance on the testing set reflects the model's generalization ability. However, in practice, this is often not the case. The unseen data is prone to change (e.g., application updates, emergence of new applications, etc.), i.e., the DDS phenomenon occurs, leading to degraded model performance. The DDS phenomenon arises from two reasons: first, the training set is limited by data collection/sampling, annotation, and preprocessing methods, which are unable to represent real-world data distributions re-

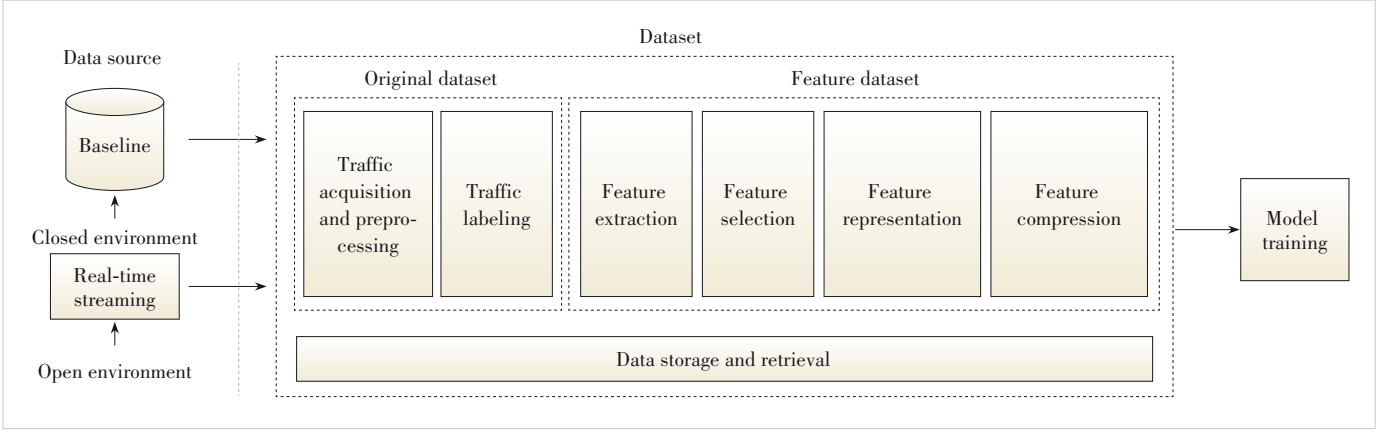


Figure 2. AI-driven network traffic data and feature engineering

Table 2. A typical collection of network traffic features (partial)

Flow Feature	Category	Description	Feature Calculation Method
Flow 5-tuple	Flow index	src/sp/dst/dp/protocol	Serialized regular preprocessing
TCP slide window	TCP window	TCP flow control parameters	Serialized regular preprocessing
TLS handshake packet information	TLS fingerprint	Handshake types, cipher suites, content types, key length, etc	Serialized regular preprocessing
Packet length sequence	Packet-related	A sequence of packet lengths in the stream. It may contain upstream, downstream, and bidirectional sequences as needed.	Packet length variance (max/min/ave/std)
Packet arrival times		A time sequence of packet arrivals in a diversion. Upstream, downstream and bidirectional sequences may be included as needed.	Packet time variance (max/min/ave/std)
Flow length related	Flow-related	Total number of flow bytes per unit of time, which may include upstream, downstream, and bidirectional as needed.	Multi-flow length variance (max/min/ave/std)
Flow duration		TCP flow duration UDP flow duration can be increased if NP resources are sufficient.	Multi-flow duration variance (max/min/ave/std)

NP: network processing TCP: transmission control protocol TLS: transport layer security UDP: user datagram protocol

alistically; second, real-world data is not static, but dynamically changes with new applications and application updates. The solution to the AI-TC data shift problem should achieve three objectives: 1) detecting the drift with the minimum number of traffic samples; 2) accurately characterizing the shift characteristics of the traffic data distribution, and preferably identifying shift samples from test data; 3) quantifying the degree of malignancy of the drift as much as possible.

3.2 Categories

DDS includes three subtypes: the covariate shift (CS), the concept drift (CD), and the label shift (LS). If the input of a classification model is defined as X, its probability distribution is $P(X)$; if the output is Y, its probability distribution is $P(Y)$. Under the supervised learning paradigm, training data can be viewed as a set of samples conforming to the joint probability distribution of $P(X, Y)$, and the ML/DL-based TC classification model aims to build the model of $P(Y|X)$.

CS refers to changes in the distribution of input traffic data $P(X)$ while keeping the conditional probability $P(Y|X)$ unchanged. There are three causes of this drift: 1) The bias in the traffic data collection/sampling process. For example, fewer samples are collected when constructing the traffic dataset due to less Distributed Denial of Service (DDoS) traffic. When DDoS attacks occur in the natural network environment, the traffic distribution differs significantly from that in training; 2) data augmentation operations performed to alleviate the class imbalance problem of the training dataset, which cause the input distribution to be inconsistent with reality, such as over-sampling, SMOTE, or GAN; 3) the model training process causing the traffic data distribution to change. In particular, active learning, which selects the most representative samples based on some assumptions or prior information, making the distribution of training input data differ from the real world.

CD, known as posterior drift, refers to a situation where the input distribution remains unchanged, but given an input, the conditional probability distribution of the output changes, that is, $P(Y|X)$ changes while $P(X)$ remains constant. In network traffic classification, it often refers to the emergence of new applications and iterative updates to old applications. This causes instances previously predicted as Application A to now be predicted as unknown or mistakenly classified as Application B, while classification models use the same input. The CD issue is a relatively concentrated problem in previous research on network traffic classification and intrusion detection involving the drift of concepts.

LS, also known as prior drift or target drift, refers to a situation where $P(Y)$ changes while $P(X|Y)$ remains constant. In other words, when the output distribution changes, the input distribution remains unchanged for a given output. Taking DDoS attack traffic as an example, AI-TC primarily identifies it as a DDoS attack based on the typical characteristic of constant inter-packet time intervals (statistically, this means slight variance). If AI-TC classifies traffic as a DDoS attack flow, the rule based on typical flow characteristics remains unchanged, which means $P(X|Y)$ remains constant. However, due to real network conditions, DDoS attack behavior suddenly surges, causing an increase in this type of attack traffic, which means that $P(Y)$ distribution changes (compared with the DDoS probability distribution in the training dataset). When the input traffic distribution changes, causing a CS phenomenon, the output distribution often changes accordingly, thus leading to the LS phenomenon.

3.3 Causes of Formation

There are many causes of distribution drifts in flow data, which are broadly classified into the following two categories.

3.3.1 Causes of CS/LS

Since CS and LS often occur concurrently, their underlying causes are fundamentally similar.

1) Flaws in data collection methods for the training dataset lead to CS or LS. Due to limitations in data collection, the distribution of some traffic categories in the training set differs from that in real networks. For instance, malicious attacks occur more frequently in real networks, but simulating attack traffic is challenging during data collection for training, resulting in a smaller amount of collected traffic.

2) Feature changes can lead to CS/LS. For example, introducing new features, such as the device type (Apple or Android), for fine-grained mobile application identification in the AI-TC model can result in CS/LS.

3) The use of data augmentation methods to address class imbalance may cause CS/LS.

4) CS/LS arises during the model learning process. For instance, selecting the most representative samples based on heuristic rules for active learning can result in CS/LS.

3.3.2 Causes of CD

CD is often the result of new applications emerging, old applications getting updated, and inactive applications being taken down. This is particularly significant in security scenarios like intrusion detection. For instance, in order to evade detection, hackers continuously enhance and modify their attack methods using various traffic obfuscation techniques.

4 Shift Detection

4.1 Traditional Approaches

Traditional methods infer whether the DDS problem occurs by monitoring the changes in metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. However, these metrics need to be compared with the ground truth. The model is often unable to obtain the real value in real-time or after a considerable delay, so the method is impractical. Lipton et al.[12] proposed a black-box shift estimation method to detect and quantify the degree of shift in the data distribution without the need for actual data labels. This method designs a black-box predictor to reduce data dimensionality and, by utilizing the reversibility of a confusion matrix, predicts whether a data distribution shift has occurred and quantifies the degree of the shift.

4.2 Statistical Methods

The statistical method is to compare the statistical values of the source distribution and the target distribution, such as min, max, mean, median, variance, skewness, and kurtosis. For example, statistics on the expected value and variance of packet lengths in network flow features during model inference are compared with those of the training dataset to determine whether there is any shift. Google's TensorFlow extension plugin currently supports some of these features. However, the statistical values in the training set and those during the inference process may be quite similar, making it difficult to determine whether a data distribution shift has occurred. Currently, shift detection based on statistical methods can be roughly categorized into two types: univariate detection and multivariate detection.

4.3 Two-Sample Hypothesis Test

Another method is using a two-sample hypothesis test (referred to simply as the two-sample test). This test serves to ascertain whether a substantial statistical disparity exists between two datasets. Detecting such a difference indicates a reduced likelihood of random variations within the data. Furthermore, it suggests a greater probability that the two datasets stem from distinct distributions, signifying a shift in the data distribution. A typical two-sample test method is the Kolmogorov-Smirnov test, or the K-S test, which does not require prior knowledge of the data distribution, but detects the difference in the statistical level between the two datasets from the training set and the real-time flow to determine the shift. However, the disadvantage of this method is that it cannot work in a high-dimensional space. Thus, high-dimensional data must first be downscaled before applying the K-S test for statistical level difference analysis. Fig. 3 shows a framework for detecting shifts in network traffic distribution based on the two-sample test. The framework performs a dimensionality reduction operation on data from two distributions—the training

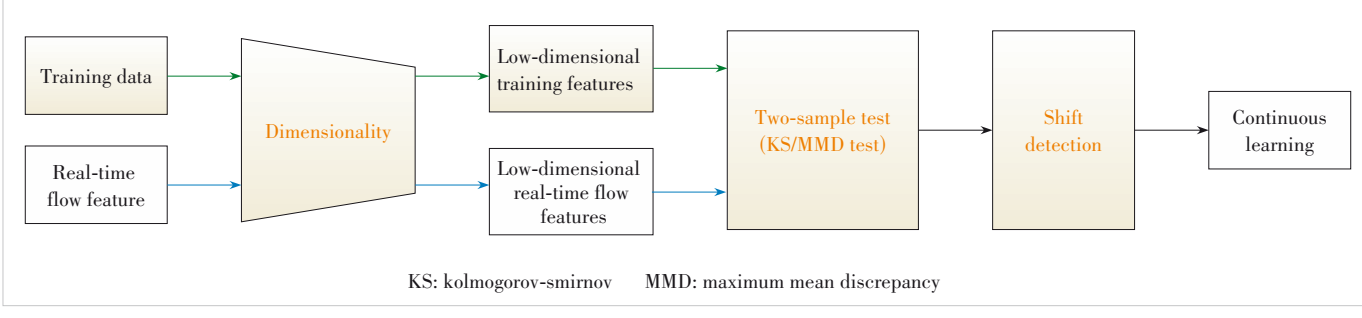


Figure 3. Network traffic distribution shift detection based on a two-sample test

set and the real-time flow—to form low-dimensional training features and real-time flow features, feeds them into the K-S two-sample test, integrates several statistics for shift detection, and finally performs quantitative analyses of the shift degree. These include various dimensionality reduction methods, PCA, AE, etc. There are also many methods for two-sample testing, such as maximum mean discrepancy (MMD)[13] and learned Kernel MMD[14].

Fig. 3 shows the whole shift detection process. First, the training feature set and the real-time flow feature are reduced to low-dimensional features, respectively with dimensionality reduction means like PCA. The low-dimensional training features are referred to $X = x_1, x_2, \cdots, x_n$, while the low-dimensional real-time flow features are $\hat{X} = \widehat{x_1}, \widehat{x_2}, \cdots, \widehat{x_n}$. Then the distance between the two distributions is calculated using MMD. The calculation formula is: $\mathrm{MMD}\left(X, \hat{X}\right) = \left| \frac{1}{n}\sum_{i=1}^{n}\phi\left(x_i\right) - \frac{1}{m}\sum_{j=1}^{m}\phi\left(\widehat{X}_j\right)\right|^2$, which calculates the squared Euclidean distance between the mean of samples mapped to the feature space in X and $\hat{X}$ separately. When the value of MMD is 0, we consider the two distributions to be the same. Thus, MMD measures the difference between the distributions of samples by comparing their means in feature space and thus can be used to detect shift between domains.

4.4 Two-Sample Testing Experiments

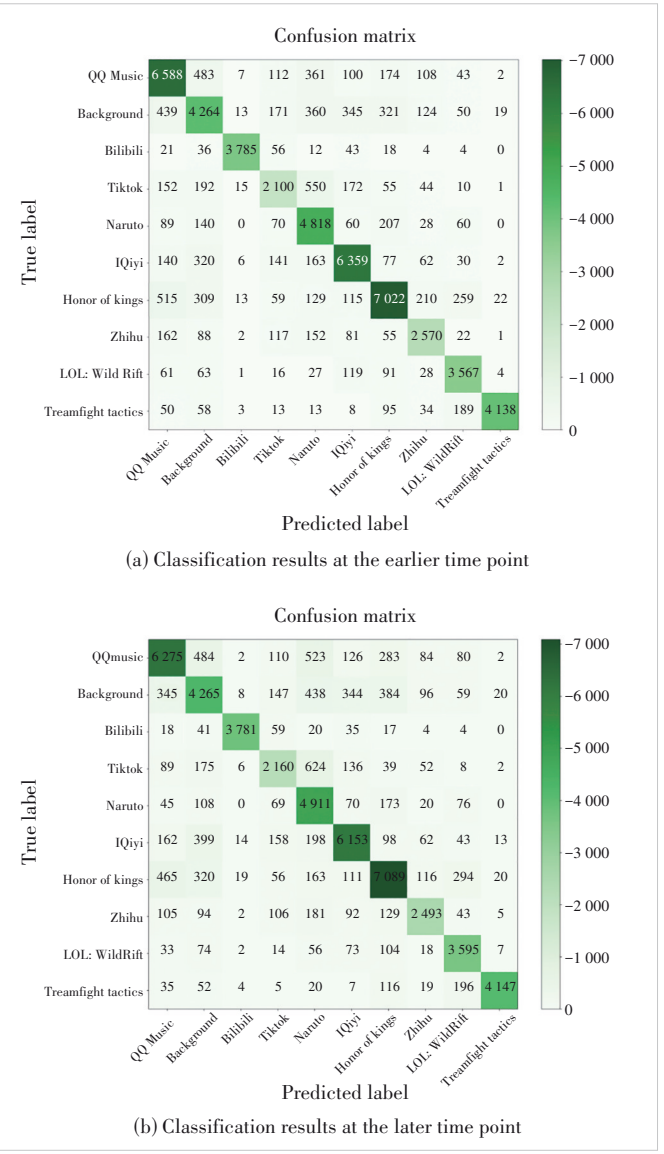
We utilized real traffic collected at different time intervals for the same application as our dataset. On the personal terminal, PCAPdroid is used to mark network traffic, and on the router, Wireshark is used to capture it. Finally, the two are compared and filtered to obtain the traffic data required by the experiment. CICFlowMeter is then used to calculate the network traffic characteristics for each application's PCAP file to get the corresponding csv file. The applications contained in the dataset and the corresponding number of streams are shown in Table 3.

Table 3. Dataset and the corresponding number of streams

App	Flow
QQmusic (Music)	39 465
LOL:Wild Rift (Game)	19 841
Naruto (Game)	27 240
Zhihu (Sociality)	16 643
Bilibili (Video)	20 014
Teamfight Tactics (Game)	23 718
IQiyi (Video)	36 740
Tiktok (Video)	16 640
Honor of Kings (Game)	42 734
Background (Log)	30 000

We employed a consistent traffic classification model, and the results are shown in Fig. 4 and Table 4. Traffic of the same application collected at two different time points with a one-month interval was used for evaluation. Fig. 4a presents the confusion matrix for traffic collected at the earlier time point, while Fig. 4b shows the results for traffic collected at the later time point. As can be observed, the classification performance of QQ Music exhibits a noticeable degradation across the two time points, with the number of correctly recognized samples decreasing from 6 588 to 6 275, whereas the performance of most other applications remains relatively stable. This observation suggests the presence of a potential data distribution shift, which motivates the subsequent drift detection experiments.

The experimental results show that the classification performance of QQ Music degrades noticeably, which is consistent



(a) Classification results at the earlier time point

(b) Classification results at the later time point

Figure 4. Confusion matrices of network traffic classification results at two different time points

Table 4. Classification results of network traffic at different times

	Traffic at Time Point T_1				Traffic at Time Point T_2			
	precision	recall	f1-score	support	precision	recall	f1-score	support
QQ Music	0.802	0.826	0.814	7 978	0.829	0.787	0.807	7 978
Background	0.716	0.698	0.707	6 106	0.709	0.698	0.704	6 106
Bilibili	0.984	0.951	0.968	3 979	0.985	0.950	0.967	3 979
Tiktok	0.736	0.638	0.683	3 291	0.749	0.656	0.700	3 291
Naruto	0.732	0.880	0.799	5 472	0.688	0.897	0.779	5 472
IQiyi	0.859	0.871	0.865	7 300	0.861	0.843	0.852	7 300
Honor of Kings	0.865	0.812	0.838	8 653	0.841	0.819	0.830	8 653
Zhihu	0.800	0.791	0.795	3 250	0.841	0.767	0.802	3 250
LOL: Wild Rift	0.842	0.897	0.869	3 977	0.817	0.904	0.859	3 977
Teamfight Tactics	0.988	0.899	0.942	4 601	0.984	0.901	0.941	4 601
Accuracy	0.828	0.828	0.828	0.828	0.822	0.822	0.822	0.822
Macro avg	0.832	0.826	0.828	54 607	0.830	0.822	0.824	54 607
Weighted avg	0.831	0.828	0.828	54 607	0.827	0.822	0.822	54 607

with the drift detection outcomes. To further investigate this phenomenon, paired-sample drift detection was performed by comparing traffic collected at two different time points. As illustrated in Fig. 4a, the traffic features from two closely spaced time points are highly similar and largely overlap, indicating no observable data drift. In contrast, Fig. 4b demonstrates a clear distribution shift when comparing traffic collected at two time points separated by a longer interval. These results suggest the presence of a data distribution shift, which motivates subsequent drift detection analysis.

The corresponding drift detection results are summarized in Table 5. For the one-day interval data, most methods report no drift, except for Spot-the-diff. For the one-month interval data, the majority of methods successfully detect drift, with the exception of the Least-Squares Density Difference method. In terms of efficiency, Kolmogorov-Smirnov, Cramérvon Mises, and the mixed-type tabular data methods exhibit the lowest execution times, whereas MMD and Spot-the-diff incur the highest computational cost.

The distance for detecting drift in Table 5 denotes a function or metric utilized in measuring the variation or resemblance of data or concepts. For instance, MMD is a representative technique for drift detection based on data distributions. The fundamental concept of MMD is that if the moments of any two data distributions are identical, then these two distributions are consistent. Conversely, the two distributions have

Table 5. Experimental results of data drift in response to different time periods

Data	Methods	Drift Occurs	Distance (unitless)	Execution Time/s
One-month interval	Kolmogorov-Smirnov	√	/	0.062
	Maximum mean discrepancy	√	0.192 245 841	1.385
	Chi-Squared	√	/	0.232
	Cramérvon Mises	√	/	0.030
	Least-squares density difference	√	0.239 282 097	0.398
	Spot-the-diff	×	0.551 293 545	1.293
	Mixed-type tabular data	√	/	0.069
One-day interval	Kolmogorov-Smirnov	×	/	0.064
	Maximum mean discrepancy	×	0.000 526 399	1.449
	Chi-Squared	×	/	0.145
	Cramérvon Mises	×	/	0.025
	Least-squares density difference	√	0.003 137 094	0.391
	Spot-the-diff	×	0.054 495 389	1.286
	Mixed-type tabular data	×	/	0.069

drifted if a particular moment exhibits divergence. If the MMD's value is 0, it demonstrates that the two distributions are identical and have not drifted. The approach employs a Gaussian kernel function to capture changes in the data's higher-order moments. A larger MMD value signifies a greater disparity between the two distributions, implying a higher degree of drift. However, the flow features do not provide sufficient information about QQ Music. For instance, the Spot-the-diff method requires a reference set and a test set. The reference set represents the baseline distribution of the data, while the test set is used to detect any drift. The distance between the histogram vectors of the reference set and the test set is calculated and ranked for each flow feature. The top k-th features with the largest distances are identified as candidates for the occurrence of drift. However, drift detection is often impeded by the inadequate determination of thresholds for k.

Furthermore, a visualization technique was employed to identify instances of drift in QQ Music traffic. Network traffic differs from images and text, making it difficult for human perception. The t-distributed stochastic neighbor embedding (t-SNE) approach was chosen to analyze the network traffic at varying time intervals. Fig. 5a illustrates the daily collected QQ Music network traffic, revealing an overlap of the two data types. In contrast, Fig. 5b displays the monthly collected QQ Music network traffic, where a clear distinction between the left and right sides of the diagram can be observed.

5 Solutions to Network Traffic Data Distribution Shift

5.1 Common Approaches

There are three common ways to cope with the data distribution shift problem: 1) Continuously supplementing new data for model retraining. This requires the training dataset to be large enough for the AI-TC model to learn the distribution as comprehensively as possible. However, this is often difficult to achieve in practical applications because of the enormous size of the network traffic. The massive traffic dataset will take up a great deal of storage resource, which also sharply increases the cost of model training. So, the academic community is also constantly looking for new methods to make the model cope with the network data shift problem at a limited cost. 2) Retraining the model using newly labeled data. The different training methods are divided into two ways: retraining the model on both new and old data (also called stateless training) and using only new data for continuous training on the existing model (known as stateful training or fine-tuning). Regardless of the paradigm selected, two core questions need to be clarified: how to choose between stateful and stateless training and how to start a new training. 3) Domain adaptation. To solve the problems of the previous two approaches, researchers have proposed a new area of research in machine learning. This method allows the model to learn the target distribution without new labels. It often uses representation learning methods combined with unsupervised learning to enable the model to learn invariant features of the data, thereby effectively addressing data distribution shift[15–16].

5.2 Instance-Based Adaptive Methods

Instance-based adaptation is a process in domain adaptation that aims to minimize the bias among data distributions by reweighting source-labeled data according to the target risk. It assumes that the source and target distributions differ only in their marginal distributions, while the posterior distribution remains constant, called covariate bias. The problem is difficult to solve without labeled data in the target domain. To solve this problem, an importance weighting method is used to compensate for the bias by reweighting the samples in the source domain according to the ratio of the densities in the tar-

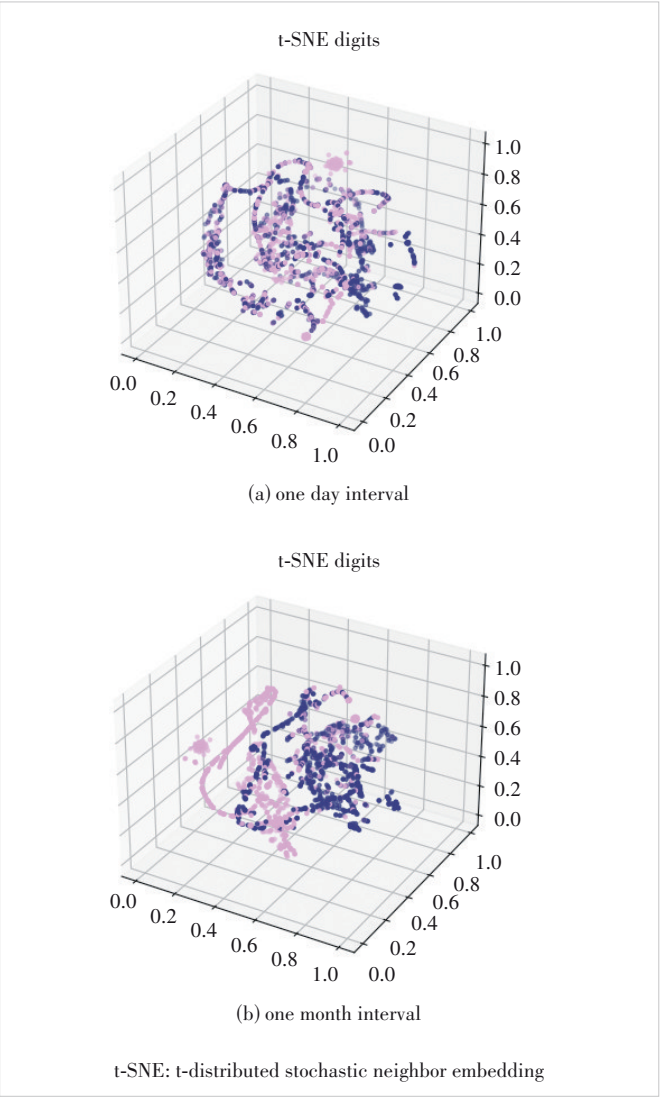


Figure 6. Visualisation of network traffic characteristics for different time intervals

get and source domains. This can be achieved by methods such as kernel mean matching (KMM) and the Kullback-Leibler importance estimation procedure (KLIEP). KMM uses the MMD in the Reproducing Kernel Hilbert Space (RKHS) to estimate the weights by minimizing the difference between the domains. In contrast, the KLIEP uses the KL difference between the target and weighted source distributions to estimate the density ratio directly.

5.3 Feature-Based Adaptive Methods

Feature-based adaptation is a domain adaptation process that aims to map source data to target data by learning transformations that extract invariant feature representations across domains. The method transforms the original features into a new feature space and then minimizes the gap among domains while preserving the underlying structure of the original data. Current feature-based adaptive methods mainly include the following three methods.

1) Subspace-based methods: This approach aims to discover common intermediate representations shared among domains by constructing low-dimensional subspaces for each domain and reducing their differences. Techniques such as PCA can be used to construct subspaces and measure domain offsets.

2) Transformation-based methods: This approach transforms the original features into a new representation to minimize the differences between the marginal and conditional distributions while preserving the structure and characteristics of the original data. Differences among domains are measured using techniques such as MMD and Kullback-Leibler Divergence (KL-Divergence).

3) Reconstruction-based methods: This approach aims to reduce the differences between domain distributions by employing an intermediary feature representation to reconstruct samples. It entails acquiring a linear projection matrix that converts source samples into intermediate representations and subsequently aligns them with the samples present in the target domain.

5.4 Deep Domain Adaptive Methods

Deep domain adaptive methods utilize neural networks (e.g., CNN, VAE, GAN) to close the gap among domains and adjust the distributions between the source and target domains. These methods typically involve training two networks: one for the source data and the other for the target data. These two networks can learn the representation features of the two types of data to minimize the inter-domain differences. Current approaches based on deep domain adaptation are divided into three main directions:

1) The variance adaptive approach uses an adaptive layer in the domain adaptation to match the marginal distribution across domains, aiming to adjust the bias in the distributions.

2) The reconstructive adaptive method adopts an autoencoder to adjust the differences across domains. The autoencoder is trained on samples from one source domain and then adjusts and reconstructs samples from another domain. By minimizing the reconstruction error between the source and target domains, the autoencoder can learn to extract invariant features across domains.

3) Adversarial domain adaptation uses adversarial learning to extract domain-invariant features by minimizing distributional differences among domains.

6 Current Challenges and Future Research Directions

Both continuous learning and domain adaptation have improved network traffic classification in real-world environments to some extent, allowing models to be deployed in such scenarios. However, current techniques still suffer from the following problems:

1) Network data mobility: The characteristics of network traffic change with time, location, network topology, etc. Continuous learning requires timely access to new data, while the domain adaptive model needs to adapt to dynamic changes in domain differences in a timely manner.

2) Application domain shift: Application version updates or business upgrades will most likely lead to significant differences in traffic characteristics. Meanwhile, the features previously found by the domain adaptive models are no longer similar. How can domain shift be detected in open scenarios?

3) Lack of labeled data: The challenge of annotating network data remains persistent and unresolved. Moreover, the presence of imbalanced samples within various categories of network traffic significantly undermines the precision of the continuous learning process.

4) Real-time requirements: Network traffic classification requires rapid response to identify and address real-time cyber threats, so models must be trained and inferred under real-time requirements. When models are then deployed, few studies have considered how network traffic is pre-processed and classified in real time.

5) Model complexity and computational resources: Models used in real-world scenarios are often in a state of failure and retraining due to the extremely large number of web applications available today. This situation greatly consumes computational resources and significantly increases model complexity.

Future researchers should focus on the following directions:

1) Combining continuous learning with dynamic domain adaptive methods to address data trafficability issues;

2) Designing systems that enable end-to-end continuous learning, including model failure feedback, shift detection, model updating, and real-time application, to ensure the sustained effectiveness of the model;

3) Enhancing the rapid response capability in real-time network environments and designing efficient model update and inference strategies to meet the requirements of real-

time scenarios.

4) Investigating how to reduce model complexity and improve computational efficiency so that the model can be applied in practical scenarios;

5) Trying to use small samples or unsupervised learning on continuous learning to improve the quality of data collection and labeling and reduce reliance on labeled data.

7 Conclusions

In this paper, we propose a generic AI-driven network traffic classification workflow and design a traffic data and feature engineering framework in detail. We thoroughly study the concepts and causes of network traffic data distribution shift, summarize the existing research progress on the data distribution shift problem, and then propose corresponding shift detection methods. Considering the current shift detection techniques, we analyze the effective means to address such shift and propose the domain adaptive method applicable to AI-TC to solve the data distribution shift problem. Finally, the current challenges and future research directions of AI-TC in data distribution shift detection and continuous learning are proposed from the requirements of data labeling, model complexity, and real-time performance.

References

[1] Wang H Z, Liu J W. Research status and key technologies of network endogenous security [J]. ZTE technology journal, 2022, 167(6): 2 – 11. DOI: 10.12142/ZTETJ.202206002

[2] Lu H, Chen Y, Lou D. 5G/5G-Advanced/6G access network security technology evolution and endogenous security [J]. ZTE technology journal, 2022, 167(6): 85 – 94. DOI: 10.12142/ZTETJ.202206014

[3] Rezaei S, Liu X. Deep learning for encrypted traffic classification: an overview [J]. IEEE communications magazine, 2019, 57(5): 76 – 81. DOI: 10.1109/MCOM.2019.1800819

[4] Wang P, Chen X J, Ye F, et al. A survey of techniques for mobile service encrypted traffic classification using deep learning [J]. IEEE access, 2019, 7: 54024 – 54033. DOI: 10.1109/ACCESS.2019.2912787

[5] Aceto G, Ciuonzo D, Montieri A, et al. Mobile encrypted traffic classification using deep learning: experimental evaluation, lessons learned, and challenges [J]. IEEE transactions on network and service management, 2019, 16(2): 445 – 458. DOI: 10.1109/TNSM.2019.2899085

[6] Aceto G, Ciuonzo D, Montieri A, et al. Mobile encrypted traffic classification using deep learning [C]//Proceedings of Network Traffic Measurement and Analysis Conference (TMA). IEEE, 2018: 1 – 8. DOI: 10.23919/TMA.2018.8506563

[7] Wang P, Ye F, Chen X J, et al. Datanet: deep learning based encrypted network traffic classification in SDN home gateway [J]. IEEE access, 2018, 6: 55380 – 55391. DOI: 10.1109/ACCESS.2018.2872430

[8] Wang P, Li S H, Ye F, et al. PacketCGAN: exploratory study of class imbalance for encrypted traffic classification using CGAN [C]//International Conference on Communications (ICC). IEEE, 2020: 1 – 7. DOI: 10.1109/icc40277.2020.9148946

[9] Wang P, Wang Z X, Ye F, et al. ByteSGAN: a semi-supervised generative adversarial network for encrypted traffic classification in SDN Edge Gateway [J]. Computer networks, 2021, 200: 108535. DOI: 10.1016/j.comnet.2021.108535

[10] Wang Z X, Wang P, Zhou X K, et al. FLOWGAN: unbalanced network encrypted traffic identification method based on GAN [C]//IEEE International Conference on Big Data and Cloud Computing (BdCloud). IEEE, 2019: 975 – 983. DOI: 10.1109/ispa-bdcloud-sustaincom-socialcom48970.2019.00141

[11] Wang Y, Wang P, Wang Z X, et al. Evaluation of feature selection on network traffic classification [C]//IEEE International Conference on Big Data and Cloud Computing (BdCloud). IEEE, 2021: 813 – 818. DOI: 10.1109/dasc-picom-cbdcom-cyberscitech52372.2021.00135

[12] Lipton Z, Wang Y X, Smola A. Detecting and correcting for label shift with black box predictors [C]//International conference on machine learning. PMLR, 2018: 3122 – 3130. DOI: 10.48550/arXiv.1802.03646

[13] Gretton A, Borgwardt K M, Rasch M J, et al. A kernel two-sample test [J]. The Journal of machine learning research, 2012, 13(1): 723 – 773. DOI: 10.5555/2188385.2188410

[14] Liu F, Xu W, Lu J, et al. Learning deep kernels for non-parametric two-sample tests [C]//International conference on machine learning. PMLR, 2020: 6316 – 6326. DOI: 10.48550/arXiv.2002.09116

[15] Zhang K, Schölkopf B, Muandet K, et al. Domain adaptation under target and conditional shift [C]//International conference on machine learning. PMLR, 2013: 819 – 827. DOI: 10.5555/3042817.3043028

[16] Zhao H, Des Combes R T, Zhang K, et al. On learning invariant representations for domain adaptation [C]//International conference on machine learning. PMLR, 2019: 7523 – 7532. DOI: 10.48550/arXiv.1905.12013

Biographies

Zhao Jianchao is with the Cable Products Business Department, ZTE Corporation. He is engaged in the development and delivery of wireline and cable network solutions, with a focus on product implementation, service integration, and system deployment. His work involves coordinating product requirements with network operations and supporting the deployment of large-scale wireline network services. His professional interests include wireline network solutions, service delivery optimization, and operational support for carrier networks.

Geng Zhaosen is with the Cable Products Business Department, ZTE Corporation. He is a member of the FM Product Team, focusing on the planning, operation, and management of wireline and cable network products. His work involves network service deployment, traffic management, and operational optimization in large-scale wireline networks. His research interests include wireline product planning, network operations, and data-driven network management.

Li Zeyi is currently pursuing his PhD degree in cyberspace security at Nanjing University of Posts and Telecommunications, China. His research interests include network security, anomaly detection, and deep packet inspection.

Wang Pan (wangpan@njupt.edu.cn) received his BS, MS, and PhD degrees in electrical and computer engineering from Nanjing University of Posts and Telecommunications, China in 2001, 2004, and 2013, respectively, where he is currently a full professor. His research interests include AI-powered networking and security in B5G, 6G, IoT, Smart Grid, CFN, and AI-enabled big data analysis. From 2017 to 2018, he was a visiting scholar at the Department of Electrical and Computer Engineering, University of Dayton, USA. He served as a TPC member of IEEE CyberSciTech Congress. He is also a reviewer for several journals, such as *IEEE Transaction on Network and Service Management*, *IEEE Internet of Things Journal*, *Computer and Security*, and *Big Data Research*.

Efficient and Secure Data Storage in 5G Industrial Internet Collaborative Systems

Wang Jigang[1,2], Liu Dong[1,2], Wan Changsheng[3], Lu Ping[1,2]
(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. Southeast University, Nanjing 210096, China)

DOI: 10.12142/ZTECOM.202601007
https://kns.cnki.net/kcms/detail/34.1294.TN.20260306.1414.002.html, published online March 6, 2026
Manuscript received: 2024-06-26

Abstract: Security and access control for data storage in 5G industrial Internet collaborative systems are facing significant challenges. The characteristics of 5G networks, such as low latency and high speed, facilitate data transmission in the industrial Internet but also increase vulnerability to attacks like theft and tampering. Moreover, in 5G industrial Internet collaborative system environments, data flows across multiple entities and links, which necessitates a flexible access control model to meet specific data access requirements. Traditional role-based and attribute-based access control mechanisms are difficult to apply in such dynamic application scenarios. To address these challenges, we propose a novel data storage solution for 5G industrial Internet collaborative systems. Similar to existing approaches, it provides integrity and confidentiality protection for transmitted data. In terms of security, only authenticated data owners and users can obtain file decryption keys, preventing malicious attackers from data forgery. Regarding access control, decryption is permitted only to authorized data users, safeguarding against unauthorized file access. Furthermore, by introducing an attribute-based encryption mechanism, only data users with specific attributes can decrypt files. In terms of efficiency, our approach utilizes bilinear and modular exponentiation operations solely during the authentication process. For handling substantial data loads, lightweight cryptographic algorithms are employed. Consequently, our solution achieves higher efficiency compared with other known methods. Experimental results demonstrate the feasibility of our approach in real-world applications.

Keywords: 5G industrial Internet collaborative systems; data storage; identity-based authentication; access control

Citation (Format 1): Wang J G, Liu D, Wan C S, et al. Efficient and secure data storage in 5G industrial internet collaborative systems [J]. ZTE Communications, 2026, 24(1): 45 - 55. DOI: 10.12142/ZTECOM.202601007
Citation (Format 2): J. G. Wang, D. Liu, C. S. Wan, et al., "Efficient and secure data storage in 5G industrial internet collaborative systems," *ZTE Communications*, vol. 24, no. 1, pp. 45 - 55, Mar. 2026. doi: 10.12142/ZTECOM.202601007.

1 Introduction

In recent years, 5G and beyond technology has been widely adopted and integrated into various fields[1]. As a foundation of the digital economy, 5G industrial Internet collaborative systems play a crucial role. Traditional industrial networks were closed, making it difficult to manage a large number of terminal devices and users and resulting in poor network scalability. With the integration of 5G, industrial networks can efficiently manage numerous industrial terminal devices and users through 5G network elements. Simultaneously, users and devices can easily access the industrial Internet via 5G networks. This transition has opened up industrial networks, allowing users and devices to use the industrial Internet more conveniently, thereby significantly improving production efficiency. Furthermore, by using 5G, industrial networks can seamlessly access external cloud servers and acquire the ability to store large amounts of data at a low cost.

Regardless of the specific technology adopted, a typical data storage scheme for 5G industrial Internet collaborative systems consists of four entities (Fig. 1): the data owner, who stores data in the 5G edge cloud; the data user, who accesses the data; and an authentication server, which provides entity authentication.

Considering the assumptions and requirements for secure data storage in 5G industrial Internet collaborative systems, there is a need to propose a novel data storage scheme based on identity authentication and access control. The design requirements focus on the following five aspects:

1) Industrial data confidentiality[2]. Since the edge cloud is often provided by telecommunications operators, industrial network users may be concerned about data leakage, which could endanger normal production processes, lead to economic losses, and even result in legal disputes. Therefore, the data owner must encrypt the data stored in the cloud.

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20230628015 and the State Key Laboratory of Particle Detection and Electronics under Grant No. SKLPDE-KF-202314.

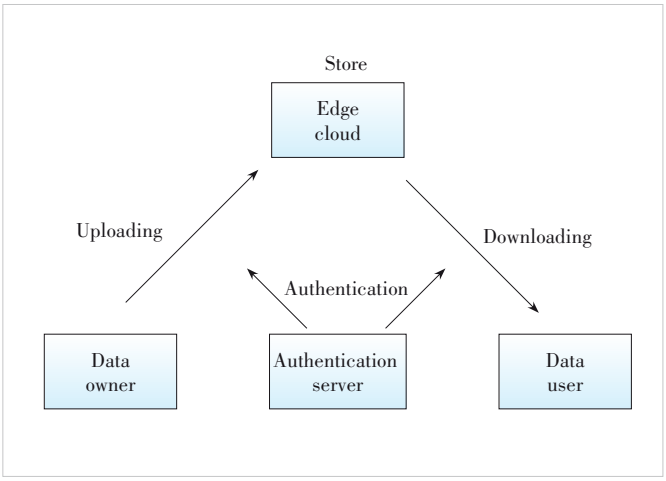


Figure 1. System model and interaction among four entities in a typical data storage scheme for 5G industrial Internet collaborative systems

2) Industrial data integrity[2]. Similar to the confidentiality requirement, the data owner also faces the risk that data stored in the cloud may be tampered with by other users. As a result, the data owner must sign the data stored in the cloud.

3) Authentication of the data owner and the data user[3]. When the data owner authorizes the data user, both parties need to undergo an identity authentication process. Otherwise, malicious attackers may impersonate the data user to illegally access cloud data, or conversely, impersonate the data owner to provide false data. However, since the data stored in the cloud is already encrypted and protected for integrity, this scheme does not require authentication of the edge cloud server.

4) Authorization of the data user and implementation of access control[4]. The data user must obtain authorization from the data owner to access data on the edge cloud. Otherwise, unauthorized malicious attackers may illegally access cloud data, causing economic losses to the enterprise. In addition, only data users with specific attributes should be allowed to access cloud data.

5) Identity-based key management[5]. Since digital certificates require a public key infrastructure and issuing certificates for a large number of terminal devices and users is costly, an identity-based key management mechanism can be used to reduce deployment costs and facilitate the management of a large number of terminal devices and users.

Obviously, designing a secure data storage scheme for 5G industrial Internet collaborative systems is challenging, with diverse and multifaceted design requirements, necessitating careful consideration of each process to be implemented. In this paper, we make the following four contributions:

1) We propose a lightweight and secure data storage system framework for 5G industrial Internet collaborative systems. Compared to other well-known methods, our approach only uses bilinear pairing and modular exponentiation operations during the authentication process, while employing lightweight cryptographic algorithms for handling large amounts of data, resulting in high efficiency.

2) We propose an identity-based authentication scheme, ensuring that only authenticated data owners and data users can obtain the file encryption key, thereby preventing malicious attackers from data forgery.

3) We introduce a lightweight data access control scheme, allowing only authorized data users to decrypt files, preventing illegal attackers from stealing files, while also enabling attribute-based encryption mechanisms for specific data users.

4) We analyze the security of our scheme in the random oracle model, and evaluate the efficiency of newly designed protocols.

The remaining content of this paper is organized as follows. Section 2 reviews relevant literature and research work. Section 3 describes the data storage scheme based on identity authentication and access control for 5G industrial Internet collaborative systems. Subsequently, security analysis is conducted in Section 4, and efficiency evaluation is presented in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

Data storage is a vital component within 5G industrial Internet collaborative systems. Ensuring the security of data storage in the industrial Internet primarily encompasses three aspects, namely confidentiality, integrity, and availability[6]. However, there are still some problems and challenges that affect the security of data storage in these systems. Data leakage risk constitutes a significant issue. Owing to the distinctive characteristics of 5G networks, including low latency and high rates, the massive data generated by the industrial Internet is more vulnerable to attacks such as theft and tampering during transmission, which may seriously affect the confidentiality of core enterprise data and cause huge economic losses. Another challenge lies in the complexity of data access control. In 5G industrial Internet collaborative systems, data flows across multiple entities and links, so a flexible access control model is needed to meet its requirements. However, traditional access control models such as role-based access control (RBAC) and attribute-based access control (ABAC) are difficult to apply in this environment. In addition, the distributed storage of data in such systems poses higher requirements for the efficiency and security of access control mechanisms. Presently, numerous research efforts are dedicated to addressing data storage security issues in cloud storage systems. These papers can be mainly divided into three categories: data encryption, access control, and data integrity.

2.1 Data Encryption

Data encryption serves as the foundational technology for ensuring data storage security, effectively thwarting unauthorized access during transmission, processing, and storage. In 5G industrial Internet collaborative systems, data encryption technology needs to adapt to the characteristics of data, such as scale,

flow, bandwidth, and latency, and meet the encryption requirements of different levels. These papers discuss how to achieve efficient and secure data encryption technology in cloud storage systems. Based on revocable-storage identity-based encryption (RS-IBE) technology, a new method for access control of shared data in the cloud is proposed in Refs. [7] and [8]. This method can achieve forward security and backward security, and can resist attacks of private key leakage. Based on the revocation mechanism, the concept of tree structure is introduced in Ref. [9]. To enhance the resistance of decryption keys against leakage, this scheme[9] divides the attribute set into two disjoint sets, and each set is combined with the master key to generate a key. Integrating direct revocation, partially hidden policy, and outsourced decryption attributes of ciphertext-policy attribute-based encryption (CP-ABE) scheme, an innovative data access control method is proposed in Ref. [10]. The security sharing and dynamic access revocation mechanism of electronic healthcare data in public cloud is deeply studied in Ref. [11], guaranteeing forward and backward security. Using CP-ABE, Ref. [12] designs a linear key-sharing scheme to resist chosen plaintext attack (CPA), and demonstrates good performance in dealing with policy change and file update. In Ref. [13], a privacy protection scheme is also constructed based on CP-ABE, employing concealed access policy to facilitate efficient permission verification. However, this scheme[13] only supports the AND policy, so it belongs to weak security model. In 5G industrial Internet collaborative systems, industrial data privacy concerns should receive greater attention.

2.2 Access Control

To safeguard user data security, cloud storage systems need to implement access control techniques to maintain the rights and interests of legitimate users. However, in 5G industrial Internet collaborative systems, traditional access control faces new challenges. For example, a centralized access control server is easy to become a prime target for attackers; once breached, it may lead to data leakage or service interruption. Attackers may steal or tamper with data and resources, or cause other forms of damage. Moreover, cloud security administrators may abuse their privileges to illegally obtain or modify resources and access permissions, thereby reducing user trust in and dependence on the cloud[14]. The following papers discuss how to achieve efficient, secure and flexible access control technology in cloud data storage systems. In the realm of centralized access control, Ref. [15] addresses the challenge of encrypting multiple files with similar access levels in centralized cloud storage by designing an extended file hierarchy CP-ABE scheme (EFH-CP-ABE). While this scheme enhances security and flexibility for cloud storage users, it does have the drawback of extended encryption and decryption computational times. Aiming at the security problems existing in data management operations in centralized cloud storage, Ref. [16] designs an improved model for data access and sharing based on proxy key protocols. This model can effectively resist various attacks, such as user or cloud impersonation, man-in-the-middle attack, and data confidentiality. However, it has the problem of low searching efficiency. For distributed access control, Ref. [17] stores all nodes' public keys and access matrices in the blockchain and proposes an efficient access control method using access matrices as record access policies. This scheme has superiority in terms of computation and storage consumption. In 5G industrial Internet collaborative systems, the access control technique should reduce the dependence on the centralized server.

2.3 Data Integrity

Data integrity verification and data possession are key security issues in cloud data storage systems. Established techniques primarily encompass mechanisms such as data possession proof, recoverability proof, and storage compliance verification. In 5G industrial Internet collaborative systems, a large amount of data will be generated when a large number of industrial terminals access the network, and such data is characterized by complex types and high mobility, which brings new pressure and challenges to data integrity. The following papers discuss how to achieve efficient, secure, and innovative data integrity in cloud data storage systems. Based on proxy re-encryption technology, a scheme that uses data certificates to achieve duplicate data deletion is proposed in Ref. [18]. A data certificate is a signature mechanism based on ownership proof, which adopts encryption algorithms that allow ciphertext decryption via generated keys. Performance analysis of this scheme shows that it can effectively prevent dictionary attacks and improve data security. Based on a cloud file system and verifier, a user-friendly data auditing scheme is constructed in Ref. [19], which does not rely on the collaboration of third parties. This scheme adopts the data reliability verification method proposed in Ref. [20], which can ensure data security and reduce the resource consumption of cloud storage. It also introduces a low-entropy security mechanism to enhance resistance against malicious data attacks.

Unfortunately, existing solutions such as role-based access control cannot provide fine-grained control, while existing attribute-based solutions are costly since they predominantly rely on computationally expensive bilinear map techniques. To address these issues, this paper designs a secure data storage and access control scheme for the low-latency requirement in 5G industrial Internet collaborative systems, and constructs a data access control system framework. The newly designed scheme can provide fine-grained access control while maintaining high efficiency. Moreover, this paper designs a new authentication and authorization scheme based on identity-based encryption techniques, which has better performance than traditional techniques.

3 Proposed Scheme

3.1 System Model

This paper proposes a secure and efficient data storage and access control scheme based on identity authentication and an encryption algorithm for secure cloud data storage in 5G industrial Internet collaborative systems. To this end, the scheme involves the following entities: the data owner, the data user, the edge cloud, and the authentication server.

The system model, as shown in Fig. 2, comprises five phases as detailed below. The corresponding notations are listed in Table 1.

3.1.1 Initialization Phase

During this phase, the authentication server generates public and private cryptographic parameters for the secure cloud storage system. These cryptographic parameters are used to generate keys for the data user and the data owner in the subsequent authentication phase. The initialization algorithm is defined as $\{sk_v, sk_{\text{prod}}, \text{PUB}\} \leftarrow \text{Init}(\text{ID}_v, \text{ID}_{\text{prod}})$.

After the initialization, the authentication server sends the private key sk_{prod} and the corresponding public keying materials PUB to the data owner via a secure channel. Similarly, it sends the private key sk_v and PUB to the data user via a secure channel.

3.1.2 Data Uploading Phase

When the data owner intends to upload a shared file F, it establishes the data uploading process by generating a file encryption key k_F. Then, it encrypts F using k_F and generates a digital

Table 1. Key notations and definitions in this paper

Notation	Description
G, g, p	The cyclic group, its generator, and prime order
$\text{ID}_{\text{prod}}, \text{ID}_v$	Identities of the data owner and the data user, respectively
sk_{prod}, sk_v	Keying materials for ID_{prod} and ID_v, respectively
$H(), h()$	Hash functions
F	The file to be uploaded and downloaded
$\sigma_F, \sigma_M, \sigma_A, \sigma_{k_F}$	The digital signature of F, M, A and k_F, respectively
C_F, C_A, C_{k_F}	The ciphertext of F, A and k_F
$\text{Dec}_k()$	Decryption function
A	The attribute values of the data user
M	Random number generated by the data owner for authentication
r_1, r_2	Random numbers generated by the data owner and the data user for authentication, respectively
k_F	Key generated by the data owner for encrypting the file F
pk_a, sk_a, pk_b, sk_b	Two sets of public and private keys of the authentication server
R_1, T_1	Parameters used for mutual authentication between the data owner and the data user
$Enc_{k_F}(), Enc_k()$	The symmetric encryption function with keys k_F and k
k	The session key
$sk_{\text{req}}, pk_{\text{req}}$	Secret and public information of the authentication request message
pk_{resp}	Public information of the authentication response message
PUB	The set of public keying materials

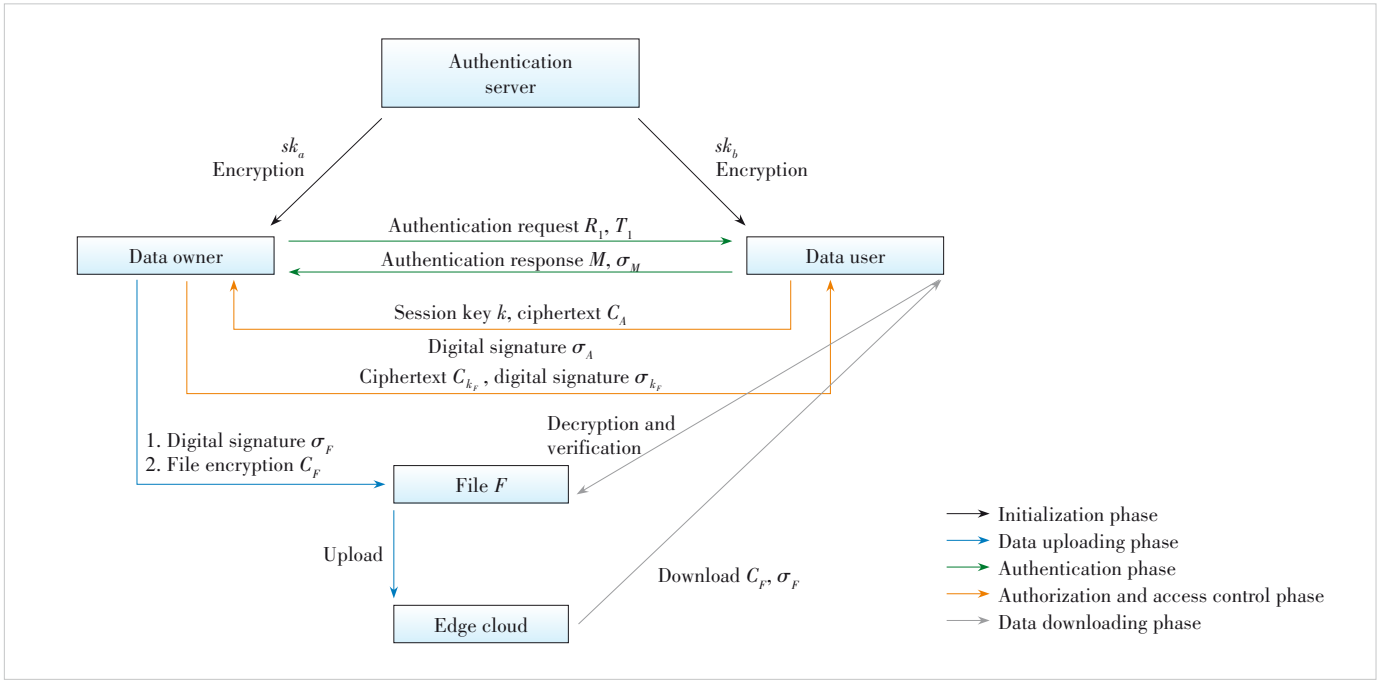


Figure 2. System model and workflow of the proposed scheme

signature σ_F for this file. Finally, the data owner uploads the encrypted file and the digital signature to the edge cloud. The data owner runs the data uploading algorithm to generate the file encryption key, ciphertext, and digital signature as $\{C_F, \sigma_F, k_F\} \leftarrow \mathrm{Dup}(F)$.

After the data uploading, the edge cloud stores the encrypted file C_F and the digital signature σ_F, which will be downloaded by the data user. The file encryption key k_F is held by the data owner, and will be distributed only to the authenticated and authorized data user.

3.1.3 Authentication Phase

Before the data user downloads the shared file F, the data owner initiates mutual authentication by sending an authentication request message pk_{req} to the data user. Upon receiving the authentication request, the data user verifies the request for authenticating the data owner, and then generates and returns an authentication response message (M, σ_M). Finally, the data owner verifies the response message for authenticating the data user. The mutual authentication process comprises the following three steps.

Step 1: The data owner establishes the authentication process by generating a set of secrets sk_{req} and a corresponding authentication request message pk_{req}. Then, the data owner sends pk_{req} to the data user. The authentication request generation algorithm is defined as $\{sk_{\mathrm{req}}, pk_{\mathrm{req}}\} \leftarrow \mathrm{Authreq}(\mathrm{ID}_v, \mathrm{PUB})$.

Step 2: Upon receiving the authentication request message pk_{req} from the data owner, the data user extracts the secret key sk_{req} from pk_{req} and generates the authentication response message pk_{resp} using the set of public keying materials PUB, its secret key sk_v, the data owner's identity $\mathrm{ID}_{\mathrm{prod}}$, and the received authentication request message pk_{req}. It then sends pk_{resp} back to the data owner for authentication. The response generation algorithm is defined as $\{pk_{\mathrm{resp}}\} \leftarrow \mathrm{Authres}(\mathrm{PUB}, sk_v, \mathrm{ID}_{\mathrm{prod}}, pk_{\mathrm{req}})$.

Step 3: Upon receiving the authentication response message pk_{resp} from the data user, the data owner verifies the integrity of pk_{resp} to authenticate the data user. The integrity verification algorithm is defined as $\{\mathrm{True}, \mathrm{False}\} \leftarrow \mathrm{Auth}(\mathrm{PUB}, sk_{\mathrm{req}}, pk_{\mathrm{resp}})$. It takes the set of public keying materials PUB, the set of secrets sk_{req}, and the received authentication response message pk_{resp} as input, and outputs True if the authentication is successful, otherwise False.

Following the authentication, the data owner and the data user mutually verify their identities, and both obtain the set of secrets sk_{req}.

3.1.4 Authorization and Access Control Phase

In this phase, the data user utilizes the newly acquired set of secrets sk_{req}, along with its own attribute values, to request reading permissions from the data owner for extracting the file F. The details of this phase can be described in three steps.

Step 1: The data user employs the set of secrets sk_{req} obtained during the authentication phase to encrypt and sign its attribute values A, generating the ciphertext C_A and digital signature σ_A.

The algorithm for generating the access request message is defined as $\{C_A, \sigma_A\} \leftarrow \mathrm{Accreq}(sk_{\mathrm{req}}, A)$ and run by the data user to generate an access control request. It takes sk_{req} and A as input, and outputs C_A and σ_A.

Step 2: Upon receiving C_A and σ_A, the data owner performs a verification process. If the verification is successful, the data owner will send back the ciphertext of the file encryption key k_F, along with its digital signature σ_{k_F}, to the data user.

The algorithm for generating the access response message is described as $\{C_{k_F}, \sigma_{k_F}\} \leftarrow \mathrm{Accres}(sk_{\mathrm{req}}, k_F)$ and executed by the data owner to generate the ciphertext and digital signature of k_F. It takes the set of secrets sk_{req} and the file encryption key k_F as input, and outputs the ciphertext C_{k_F} and the digital signature σ_{k_F} of k_F.

Step 3: Subsequently, the data user decrypts the received ciphertext C_{k_F} and verifies its signature σ_{k_F}. If the verification is successful, the data user confirms the correctness of the received encryption key and grants access to the file F. The extraction and verification algorithm is defined as $\{\mathrm{False}, k_F\} \leftarrow \mathrm{Acc}(sk_{\mathrm{req}}, C_{k_F}, \sigma_{k_F})$, which outputs the file encryption key k_F for a correct file encryption key, and False otherwise.

Following this phase, the data user gets the file encryption key k_F. Next, in the data downloading phase, the data user will use the newly acquired file encryption key k_F for extracting and verifying the file F.

3.1.5 Data Downloading Phase

In this phase, the data user downloads the ciphertext of the file C_F and its digital signature σ_F from the edge cloud. Then, the data user extracts the file F from C_F and checks the integrity of F using k_F. The extraction and verification algorithm is defined as $\{\mathrm{False}, F\} \leftarrow \mathrm{Ddl}\{k_F, C_F, \sigma_F\}$, which outputs the plaintext of the file F if F is not tampered by an adversary, and False otherwise.

3.2 Construction

The proposed efficient and secure data storage scheme for 5G industrial Internet collaborative systems is defined as a tuple (Init, Dup, Authreq, Authres, Auth, Accreq, Accres, Acc, Ddl) of nine probabilistic polynomial time algorithms. Each algorithm is detailed below.

1) $\{sk_v, sk_{\mathrm{prod}}, \mathrm{PUB}\} \leftarrow \mathrm{Init}(\mathrm{ID}_v, \mathrm{ID}_{\mathrm{prod}})$. The authentication server runs this algorithm to generate system parameters for the secure cloud storage system. This procedure is as follows. First, the authentication server generates a group G with a prime order p and a generator g. Second, the authentication server randomly generates its own private keys, denoted by $sk_a \in Z_p$ and

$sk_b \in Z_p$, respectively. Third, the authentication server generates the corresponding public keys $pk_a = g^{sk_a} \in G$ and $pk_b = g^{sk_b} \in G$. Thus, the public keying materials are defined as PUB={ pk_a, pk_b, G, g, p }. Fourth, Given the data owner's identifier $\mathrm{ID}_{\mathrm{prod}} \in \{0,1\}^n$, the authentication server calculates its secret key as $sk_{\mathrm{prod}} = H(\mathrm{ID}_{\mathrm{prod}})^{sk_a} \in G$, where $H: \{0,1\}^n \rightarrow G$ is a hash function. Finally, for the data user with the identifier $\mathrm{ID}_v \in \{0,1\}^n$, the authentication server computes its secret key as $sk_v = H(\mathrm{ID}_v)^{sk_b} \in G$, using the same hash function H.

2) $\{C_F, \sigma_F, k_F\} \leftarrow \mathrm{Dup}(F)$. The data owner runs this algorithm to generate the file encryption key, the ciphertext and the digital signature. First, for the file F to be uploaded, the data owner randomly generates a number $k_F \in Z_p$, and subsequently computes the digital signature $\sigma_F = h(F, k_F)$, where $h: Z_p \rightarrow Z_p$ is a hash function. Second, the data owner encrypts F as $C_F = \mathrm{Enc}_{k_F}(F)$, where Enc_{k_F} is a symmetric encryption algorithm such as Advanced Encryption Standard (AES), k_F is the file encryption key, and C_F is the ciphertext.

3) $\{sk_{\mathrm{req}}, pk_{\mathrm{req}}\} \leftarrow \mathrm{Authreq}(\mathrm{ID}_v, \mathrm{PUB})$. The data owner runs this algorithm to generate the authentication request message. First, the owner randomly generates $r_1 \in Z_p$ and $k \in Z_p$. Second, the owner computes $R_1 = g^{r_1} \in G$ and calculates $T_1 = h(e(H(\mathrm{ID}_v), pk_b)^{r_1}) \oplus k$, where $H: \{0,1\}^n \rightarrow G$ is a hash function, $e: \{G, G\} \rightarrow G_T$ is a bilinear mapping function, and $h: G \rightarrow Z_p$ is also a hash function. Finally, the owner gets $sk_{\mathrm{req}} = \{k, r_1\}$ and $pk_{\mathrm{req}} = \{R_1, T_1\}$.

4) $\{pk_{\mathrm{resp}}\} \leftarrow \mathrm{Authres}(\mathrm{PUB}, sk_v, pk_{\mathrm{req}})$. The data user runs this algorithm to generate the authentication response. First, the user calculates $k = h(e(sk_v, R_1)) \oplus T_1$, where $e: \{G, G\} \rightarrow G_T$ is a bilinear mapping function and $h: G \rightarrow Z_p$ is a hash function. Then, the user randomly generates $M \in Z_p$, and computes $\sigma_M = h(M, k)$, where $h: Z_p \rightarrow Z_p$ is a hash function. Finally, the user gets $pk_{\mathrm{resp}} = \{M, \sigma_M\}$.

5) $\{\mathrm{True}, \mathrm{False}\} \leftarrow \mathrm{Auth}(\mathrm{PUB}, sk_{\mathrm{req}}, pk_{\mathrm{resp}})$. The data owner runs this algorithm to verify the integrity of the authentication response message by checking whether $\sigma_M = h(M, k)$. If this equation holds, this algorithm outputs True, indicating successful authentication. Otherwise, the data owner outputs False.

6) $\{C_A, \sigma_A\} \leftarrow \mathrm{Accreq}(sk_{\mathrm{req}}, A)$. The data user runs this algorithm to generate an access control request. First, the data user encrypts its attribute value A using the session key $k \in sk_{\mathrm{req}}$ obtained during the authentication phase, producing $C_A = \mathrm{Enc}_k(A)$, where Enc_k is a symmetric encryption algorithm such as AES. Second, the data user generates a digital signature $\sigma_A = h(A, k)$, where $h: Z_p \rightarrow Z_p$ is a hash function.

7) $\{C_{k_F}, \sigma_{k_F}\} \leftarrow \mathrm{Accres}(sk_{\mathrm{req}}, k_F)$. The data owner runs this algorithm to generate an access control response message. First, the owner decrypts C_A using the secret key $k \in sk_{\mathrm{req}}$ to obtain the attribute value $A = \mathrm{Dec}_k(C_A)$, where Dec_k is a symmetric decryption algorithm such as AES. Second, the owner verifies if $\sigma_A \stackrel{?}{=} h(A, k)$ where $h: Z_p \rightarrow Z_p$ is a hash function. If the equation holds, verification succeeds; otherwise, the process aborts. Third, the data owner examines the attribute value A of the data user. If the user satisfies the required access conditions, the owner calculates the ciphertext $C_{k_F} = \mathrm{Enc}_k(k_F)$ and digital signature $\sigma_{k_F} = h(k_F, k)$, where Enc_k is a symmetric encryption algorithm such as AES, and $h: Z_p \rightarrow Z_p$ is a hash function.

8) $\{\mathrm{False}, k_F\} \leftarrow \mathrm{Acc}(sk_{\mathrm{req}}, C_{k_F}, \sigma_{k_F})$. The data user runs this algorithm to extract and verify the file encryption key k_F. First, the data user computes $k_F = \mathrm{Dec}_k(C_{k_F})$, where Dec_k is a symmetric decryption algorithm such as AES. Then, the data user verifies $\sigma_{k_F} \stackrel{?}{=} h(k_F, k)$, where $h: Z_p \rightarrow Z_p$ is a hash function. If this equation holds, verification succeeds, and the data user obtains the file encryption key k_F. Otherwise, the algorithm returns False.

9) $\{\mathrm{False}, F\} \leftarrow \mathrm{Ddl}\{k_F, C_F, \sigma_F\}$. The data user runs this algorithm to retrieve the file F. First, the user decrypts the ciphertext C_F to obtain the plaintext file $F = \mathrm{Dec}_{k_F}(C_F)$, where Dec_{k_F} is a symmetric decryption algorithm such as AES. Then, the user verifies $\sigma_F \stackrel{?}{=} h(F, k_F)$, where $h: Z_p \rightarrow Z_p$ is a hash function. If this equation holds, the file F is not tampered by attackers and this algorithm returns F. Otherwise, it returns False.

In the above construction, the efficient and secure data storage only utilizes bilinear and modular exponentiation operations during the authentication phase, while lightweight cryptographic algorithms are employed for handling large-scale data (i.e., the file F). As a result, these algorithms exhibit high efficiency. We will further evaluate the proposed scheme in Section 5.

4 Security analysis

In this section, we analyze the correctness and security of our proposed solution based on the security requirements outlined in Section 1. These requirements include industrial data integrity and confidentiality, identity authentication, authorization and access control, and identity-based key management. Furthermore, based on these requirements, we extend our analysis to provide proofs for the correctness and security of the secret key $k \in sk_{\mathrm{req}}$, as detailed below.

4.1 Integrity and Confidentiality Requirements

This work employs the symmetric key k_F to encrypt industrial data and generate digital signatures. Therefore, the confidentiality and integrity of industrial data is guaranteed by two factors: 1) the security of the encryption and hashing functions and 2) the security of the key k_F.

Since standard symmetric encryption algorithms and hashing functions are used, their correctness and security are guaranteed by the respective standards. Thus, the primary focus of this paper is on ensuring the security of k_F, as this guarantees the integrity and confidentiality of industrial data. Furthermore, based on the authorization and access control process, k_F is encrypted and digitally signed using the key $k \in sk_{\mathrm{req}}$. Hence, the security of k_F is contingent on the security of $k \in sk_{\mathrm{req}}$, making the protection of $k \in sk_{req}$ the main emphasis of our security design.

4.2 Identity Authentication Requirement

In the authentication phase, the data owner and the data user perform mutual authentication using the secret key $k \in sk_{\text{req}}$ and hash functions. Therefore, similar to the analysis in Section 4.1, the correctness and security of the authentication process rely on the correctness and security of the key $k \in sk_{\text{req}}$.

4.3 Authorization and Access Control Requirement

In the authorization and access control phase, the data owner distributes the file encryption key k_F to the data user for access authorization, and the latter can access the file only upon obtaining k_F. Therefore, the correctness and security of this phase depend on those of k_F.

Furthermore, the authorization and access control phase shows that k_F is encrypted and digitally signed using the secret key $k \in sk_{\text{req}}$. Accordingly, the correctness of authorization and access control relies on the correctness of $k \in sk_{req}$. At the same time, the security of authorization and access control is ensured by the security of $k \in sk_{\text{req}}$.

4.4 Identity-Based Key Management Requirement

The initialization phase indicates that the keys of both the data owner and data user are generated from their respective identities. Therefore, the proposed scheme satisfies the requirement of identity-based key management.

4.5 Secret Key $k \in sk_{\text{req}}$

4.5.1 Correctness

As previously mentioned, we need to prove the correctness of $k \in sk_{\text{req}}$, i.e., to verify that the data owner and data user compute an identical $k \in sk_{\text{req}}$. The proof is presented in four steps as follows.

1) From the Authreq algorithm in Section 3.2, the secret key $k \in sk_{\text{req}}$ is generated by the data owner. Therefore, we only need to ensure that the data user obtains the correct $k \in sk_{\text{req}}$.

2) From the Authres algorithm in Section 3.2, we can see that the data user computes the secret key as $k = h(e(sk_v, R_1)) \oplus T_1$.

3) Based on $sk_v = H(\text{ID}_v)^{sk_b} \in G$ in the Init algorithm in Section 3.2, we derive the key formulation as $k = h(e(H(\text{ID}_v)^{sk_b}, R_1)) \oplus T_1$.

4) Combined with $R_1 = g^{r_1} \in G$ and $T_1 = h(e(H(\text{ID}_v), pk_b)^{r_1}) \oplus k$ in the Authreq algorithm in Section 3.2, we further deduce: $k = h(e(H(\text{ID}_v)^{sk_b}, g^{r_1})) \oplus T_1 = h(e(H(\text{ID}_v), g^{sk_b r_1})) \oplus T_1 = h(e(H(\text{ID}_v), pk_b{}^{r_1})) \oplus T_1 = h(e(H(\text{ID}_v), pk_b)^{r_1}) \oplus T_1 = h(e(H(\text{ID}_v), pk_b)^{r_1}) \oplus h(e(H(\text{ID}_v), pk_b)^{r_1}) \oplus k = k$.

From the above discussion, it is evident that the secret key $k \in sk_{\text{req}}$ obtained by the data owner is the same as that generated by the data user. Therefore, our scheme proposed in this paper is correct.

4.5.2 Security

This subsection primarily demonstrates the security of the secret key $k \in sk_{\text{req}}$. First, we define the Bilinear Computational Diffie-Hellman (BCDH) problem, a well-known mathematical problem hard to solve. Then, we introduce the Indistinguishability under Chosen-Plaintext Attack (IND-CPA) security model to formalize the adversary's attack model. Finally, we prove the security of $k \in sk_{\text{req}}$: If an adversary can obtain $k \in sk_{\text{req}}$ with non-negligible probability, we can solve the BCDH problem with non-negligible probability. Since the BCDH problem is computationally hard, the proposed scheme is provably secure.

Definition 1 (BCDH Problem):

Given $a, b, c \in Z_p$ and $g, g^a, g^b, g^c \in G$, it is hard to compute $e(g,g)^{abc}$ in polynomial time.

Definition 2 (IND-ID-CPA Security Model):

This security model is formally defined by a four-phase game between a Challenger and an Adversary $\mathcal{A}$, as described below:

1) Phase 1: $\mathcal{A}$ adaptively submits an identity $\text{ID}_i \in \{0,1\}^n$, where i=1, 2, ⋯, q_1. The Challenger runs $\{sk_v, sk_{\text{prod}}, \text{PUB}\} \leftarrow \text{Init}(ID_v, ID_{\text{prod}})$ and returns sk_{ID_i} to $\mathcal{A}$.

2) Challenge: $\mathcal{A}$ submits an identity $\text{ID}_v \in \{0,1\}^n$ with $\text{ID}_v \notin \{\text{ID}_1, \text{ID}_2, \cdots, \text{ID}_{q_1}\}$ and two messages k_0 and k_1 with $|k_0| = |k_1|$;

• The Challenger flips an unbiased coin with {0, 1}, and obtains a bit $b \in \{0,1\}$;

• The Challenger runs $\{sk_{\text{req}}, pk_{\text{req}}\} \leftarrow \text{Authreq}(\text{ID}_v, \text{PUB})$ and returns pk_{req} to $\mathcal{A}$.

3) Phase 2: $\mathcal{A}$ adaptively submits an identity $\text{ID}_j \in \{0,1\}^n$ with the limitation $ID_j \neq ID_v$, where j=1, 2, ⋯, q_2. The Challenger runs $\{sk_v, sk_{\text{prod}}, \text{PUB}\} \leftarrow \text{Init}(\text{ID}_v, \text{ID}_{\text{prod}})$ and returns sk_{ID_j} to the Adversary. Let $q_M = q_1 + q_2$.

4) Output: $\mathcal{A}$ outputs its guess b' on b. $\mathcal{A}$ wins the game if $b' = b$.

Theorem 1: Suppose that H and h are random oracles. The proposed scheme is $(\varepsilon(\lambda), q_1, q_2, q_3, t)$-secure in the IND-CPA security model if the $(\varepsilon'(\lambda), t')$ BCDH assumption holds on the bilinear group (e, p, g, G, G_τ), where $\varepsilon'(\lambda) = \dfrac{\varepsilon(\lambda)}{eq_1q_2}$, and q_1, q_2 and q_3 are the numbers of queries made by the Adversary to H, h and the key generation, respectively.

Proof: Suppose there exists $\mathcal{A}$ that can $(\varepsilon(\lambda), q_1, q_2, q_3, t)$-break the IND-CPA security of the proposed scheme, we construct a Simulator $\mathcal{S}$ that uses $\mathcal{A}$ to break the BCDH assumption. Given $(g, g^{sk_a}, g^{sk_b}, g^{sk_c})$, the Simulator aims to output $e(g,g)^{sk_a sk_b sk_c}$.

1) Setup query: $\mathcal{S}$ sets $pk_a = g^{sk_a}$ and sends the system public parameters $(e, p, g, G, G_\tau, pk_a)$ to $\mathcal{A}$. $\mathcal{S}$ implicitly defines the master secret key is sk_a.

2) Random oracle query: It maintains two hash tables T_H and T_h.

• H-query: $\mathcal{S}$ selects a target index $i_v \in [1, q_1]$. $\mathcal{A}$ adaptively submits an identity $\mathrm{ID}_i \in \{0,1\}^n$, i=1, 2, ⋯, q_1. $\mathcal{S}$ first checks whether $H(\mathrm{ID}_i)$ is in T_H. If so, $\mathcal{S}$ returns $H(\mathrm{ID}_i)$ to $\mathcal{A}$; otherwise, $\mathcal{S}$ works as follows:

$$H(\mathrm{ID}_i) = \begin{cases} g^{z_i}(z_i \in Z_p), & i \neq i_v \\ g^{sk_b}, & i = i_v \end{cases} \tag{1}$$

The Simulator adds $(\mathrm{ID}_i, z_i, H(\mathrm{ID}_i))$ into T_H.

• h-query: $\mathcal{A}$ adaptively submits $W_j \in \{0,1\}^n$, j=1,2,⋯,q_2. $\mathcal{S}$ first checks whether $h(W_j)$ is in T_h. If so, $\mathcal{S}$ returns $h(W_j)$ to $\mathcal{A}$; otherwise, $\mathcal{S}$ randomly selects $w_j \in \{0,1\}^n$ sets $w_j = h(W_j)$, returns w_j to $\mathcal{A}$ and adds (W_j, w_i) to T_h.

3) Phase 1: $\mathcal{S}$ submits an identity $\mathrm{ID}_i \in \{0,1\}^n$. If $i = i_v$, $\mathcal{S}$ aborts; otherwise, $\mathcal{S}$ retrieves $(\mathrm{ID}_i, z_i, H(\mathrm{ID}_i))$ from T_H, computes $M_{\mathrm{ID}_i} = g^{sk_a z_i}$, and returns M_{ID_i} to $\mathcal{A}$. $\mathcal{A}$ can adaptively make this query up to q_{M_1} times.

4) Challenge: $\mathcal{A}$ submits two messages $k_0, k_1 \in \{0,1\}^n$ and identity ID_v. If $\mathrm{ID}_v \neq \mathrm{ID}_{i_v}$, $\mathcal{S}$ aborts; otherwise, $\mathcal{S}$ flips an unbiased coin with {0, 1} and obtains a bit $b \in \{0,1\}$. $\mathcal{S}$ randomly choose $\Gamma \in \{0,1\}^n$ and computes $R_1 = g^{sk_c}$ and $T_1 = \Gamma \oplus k$. $\mathcal{S}$ sends the challenged ciphertext pk_{req} to $\mathcal{A}$. When $h(e(g,g)^{sk_a sk_b sk_c}) = \Gamma$, pk_{req} is a correct ciphertext of the message k; otherwise, pk_{req} is a one-time pad of k_0 and k_1.

5) Phase 2: This phase is identical to Phase 1 with the limitation that $\mathrm{ID}_i \neq \mathrm{ID}_v$. $\mathcal{A}$ can adaptively make this query up to q_{M_2} times. Let $q_3 = q_{M_1} + q_{M_2}$.

6) Guess: $\mathcal{A}$ outputs its guess ω' on ω. If $b' \neq b$, $\mathcal{S}$ aborts; if $b' = b$, $\mathcal{S}$ randomly selects (W_j^*, w_j^*) from T_h and outputs W_j^*.

If pk_{req} is a correct ciphertext, $\Gamma = h\left(e(g,g)^{sk_a sk_b sk_c}\right)$ and $e(g,g)^{sk_a sk_b sk_c}$ are selected by $\mathcal{A}$ to query the h oracle. Hence, $\left(e(g,g)^{sk_a sk_b sk_c}, \Gamma\right)$ must exist in T_h. The simulation is computationally indistinguishable from the real scheme for $\mathcal{A}$.

To complete the proof, we calculate the advantage of $\mathcal{S}$ in solving the BCDH problem by defining the following events:

• E_1: $\mathcal{S}$ does not abort in Phases 1 and 2;
• E_2: $\mathcal{S}$ does not abort in the Challenge phase;
• E_3: $\mathcal{S}$ does not abort in the Guess phase;
• E_4: $\mathcal{S}$ outputs $e(g,g)^{sk_a sk_b sk_c}$.

We have:

$$\begin{aligned} &Pr[E_1] = \left(1 - \frac{1}{q_1}\right)^{q_3}, \quad Pr[E_2] = \frac{1}{q_1}, \\ &Pr[E_3] = \frac{1}{2} + \varepsilon(\lambda), \quad Pr[E_4] = \frac{1}{q_2} \end{aligned} \tag{2}$$

Therefore, the advantage of the simulator in breaking the BCDH assumption is:

$$\begin{aligned} &Pr[E_1] \times Pr[E_2] \times (Pr[E_3] - \frac{1}{2}) \times Pr[E_4] = \\ &(1 - \frac{1}{q_1})^{q_3} \times \frac{1}{q_1} \times (\frac{1}{2} + \varepsilon(\lambda) - \frac{1}{2}) \times \frac{1}{q_2} = \\ &(1 - \frac{1}{q_1})^{q_3} \times \frac{\varepsilon(\lambda)}{q_1 q_2} \approx \frac{\varepsilon(\lambda)}{e q_1 q_2} \end{aligned} \tag{3}$$

5 Efficiency Evaluation

Currently, there are a plethora of attribute-based access control systems (e.g., Refs. [15 – 17]). However, a significant proportion of these systems has not adequately accounted for the distinct requirements introduced by the 5G industrial Internet collaborative systems. Consequently, this section aims to compare an identity authentication and access control-based data storage scheme tailored specifically for the 5G industrial Internet collaborative systems context with the investigations presented in Refs. [21] and [22]. The latter two studies expound upon comprehensive encryption algorithms and protocols for attribute-based access control systems. For the 5G industrial Internet collaborative systems, communicational costs emerge as a paramount concern for access control systems, encompassing computational and communicational costs. To this end, Section 5.1 compares the computational costs of the proposed scheme with Refs. [21] and [22]. Subsequently, Section 5.2 analyzes the disparities and similarities in communicational costs between the proposed scheme and those in Refs. [21] and [22]. During the file uploading and downloading phases, we employ lightweight symmetric cryptography for digital signature and encryption, where the computational and communicational overheads are contingent upon the file size. This aspect is discussed in Section 5.3. Note that Sections 5.1 and 5.2 focus exclusively on the authentication, authorization, and access control phases, while Section 5.3 expounds upon the implementation of the proposed system, thereby substantiating its efficacy.

5.1 Comparison of Computational Costs

Given that computational costs are primarily influenced by cryptographic operations, our focus is directed towards the computational costs of fundamental cryptographic operations (such as modular multiplication, hash functions, bilinear pairings, and modular exponentiation). Subsequently, a comprehensive comparison is undertaken between the proposed data storage scheme and the works in Refs. [21] and [22].

To assess the computational costs of fundamental cryptographic operations, we conducted experiments on a computer with an Intel i5 processor and the Ubuntu 22.04 operating system, using OpenSSL[23] and PBC[24] as cryptographic libraries. The cryptographic group (denoted by G) was a 255-bit elliptic curve group[24], the SHA256 hash functions[24] and AES symmetric encryption algorithm were adopted for all experimental evaluations.

The computational costs of fundamental cryptographic opera-

tions are presented in Table 2, with all results averaged over 500 iterations of the basic cryptographic operations. On the basis of these findings, we draw the following conclusions based on the data in Table 2.

1) The computational costs of modular multiplication and hash function are around 10^{-2} – 10^{-3} to those of modular exponentiation and bilinear pairing, because $T_h/T_p = 1.8/689.3 \approx 2.6 \times 10^{-3}$, $T_h/T_{\text{me}} = 1.8/79.5 \approx 2.3 \times 10^{-2}$ and $T_h/T_{\text{mm}} = 1.8/0.5 = 3.6$.

2) The computational cost of modular exponentiation is around 10^{-1} to that of bilinear pairing, since $T_{\text{me}}/T_p = 79.5/689.3 \approx 1.2 \times 10^{-1}$.

3) The computational costs of AES encryption and decryption on a 128-bit block are around 10^{-2} – 10^{-3} to those of modular exponentiation and bilinear pairing, because $T_{\text{ENC}}/T_p = 3.8/689.3 \approx 5.5 \times 10^{-3}$, $T_{\text{ENC}}/T_{\text{me}} = 3.8/79.5 \approx 4.8 \times 10^{-2}$, and $T_{\text{ENC}}/T_{\text{DEC}} = 3.8/1.4 \approx 2.7$.

The foregoing three conclusions indicate that, in terms of computational costs, modular multiplication, hash function, and AES encryption and decryption have negligible time costs compared to modular exponentiation and bilinear pairing. Therefore, in the subsequent evaluation, our focus centers on modular exponentiation and bilinear pairing. Furthermore, the above finding also highlights that the computational cost of bilinear pairings exceeds those of modular exponentiation. Based on this observation, by avoiding bilinear pairings, the proposed data storage scheme achieves a substantial reduction in computational costs.

Furthermore, building upon the data in Table 2, we deduce the composite computational costs for our data storage scheme, as well as those for the methodologies detailed in Refs. [21] and [22]. Our comparative analysis predominantly centers on a cloud environment, assuming uniform data accessibility and cloud-based access policies. The results of this analysis are presented in Table 3, encompassing the cumulative computational costs during the uploading and downloading phases. From Table 3, we arrive at the following conclusions.

1) The computational costs on the authentication server in Refs. [21] and [22] are approximately 5 to 10 times higher than that of our data storage scheme, because $(1.38\eta_t + 1.70)/0.32 > (1.38 \times 1 + 1.70)/0.32 \approx 9.6$ and $(1.38\eta_t + 0.40)/0.32 > (1.38 \times 1 + 0.40)/0.32 \approx 5.6$.

2) The computational expenses pertaining to the data owner and data user in Refs. [21] and [22] are approximately 2 to 3 times higher than that associated with our data storage scheme, because $(0.08\eta_{nln} + 0.69\eta_v + 2.54)/1.54 > (0.08 \times 1 + 0.69 \times 1 + 2.54)/1.54 \approx 2.1$ and $(0.4\eta_\tau + 2.15\eta_v + 1.33)/1.54 > (0.4 \times 1 + 2.15 \times 1 + 1.33)/1.54 \approx 2.5$.

3) The total computational costs of Refs. [21] and [22] are approximately 3 to 4 times higher than that of our data storage scheme, because $(1.38\eta_t + 0.08\eta_{nln} + 0.69\eta_v + 4.24)/1.86 > (1.38 \times 1 + 0.08 \times 1 + 0.69 \times 1 + 4.24)/1.86 \approx 3.4$ and $(1.38\eta_t + 0.4\eta_\tau + 2.15\eta_v + 1.73)/1.86 > (1.38 \times 1 + 0.4 \times 1 + 2.15 \times 1 + 1.73)/1.86 \approx 3.0$.

The aforementioned three research outcomes indicate that the computational costs in Refs. [21] and [22] surpass those of the secure data storage scheme proposed in this paper. This observation leads to the conclusion that the proposed scheme exhibits superior efficiency.

Table 2. Computational costs of basic cryptographic operations (unit: μs)

T_{mm}	T_h	T_p	T_{me}	T_{ENC}	T_{DEC}
0.5	1.8	689.3	79.5	3.8	1.4

Notes: T_{mm} is the computational cost of modular multiplication, T_h is the computational cost of hash function, T_p is the computational cost of bilinear pairing, T_{me} is the computational cost of modular exponentiation, and T_{ENC} and T_{DEC} are the computational costs encrypting and decrypting a 128-bit block using the AES algorithm.

5.2 Comparison of Communicational Costs

Considering that communicational overhead is primarily influenced by message length, we juxtapose the message lengths of the secure data storage scheme proposed in this paper with those of Refs. [21] and [22] in Table 4. For our proposed scheme, owing to variable uploaded and downloaded file sizes, particular emphasis is placed on comparing the communicational overhead during the identity authentication and access control phases while temporarily disregarding the impact of variable-length ciphertexts on communicational costs.

As discerned from Table 4, the message lengths of our data

Table 3. Comparison of computational costs (unit: ms)

Metric	Proposed Data Storage Scheme	Ref. [21]	Ref. [22]
T_S	$4T_{\text{me}} = 0.32$	$(2\eta_t + 2)T_p + 4T_{\text{me}} = 1.38\eta_t + 1.70$	$2\eta_t T_p + 5T_{\text{me}} = 1.38\eta_t + 0.40$
T_U	$2T_p + 2T_{\text{me}} = 1.54$	$(\eta_{nln} + 6)T_{me} + (\eta_v + 3)T_p = 0.08\eta_{nln} + 0.69\eta_v + 2.54$	$(5\eta_\tau + \eta_v + 8)T_{\text{me}} + (3\eta_v + 1)T_p = 0.4\eta_\tau + 2.15\eta_v + 1.33$
T_a	$2T_p + 6T_{\text{me}} = 1.86$	$(\eta_{nln} + 10)T_{\text{me}} + (2\eta_t + \eta_v + 5)T_p = 1.38\eta_t + 0.08\eta_{nln} + 0.69\eta_v + 4.24$	$(5\eta_\tau + \eta_v + 13)T_{\text{me}} + (3\eta_v + 2\eta_t + 1)T_p = 1.38\eta_t + 0.4\eta_\tau + 2.15\eta_v + 1.73$

Notes: T_S is the computational cost of the authentication server; T_U is the computational cost of the data owner and data user; T_a is the total computational cost on the authentication server, data owner and data user; n is the number of attributes in Refs. [21] and [22]; η_t is the number of entries available in the tag-label-policy list in Refs. [21] and [22]; η_{nln} is the number of total non-leaf nodes in the tree access structure in Ref. [21] and [22]; η_v is the total number of user attributes in Refs. [21] and [22]; η_τ is the total number of attributes in the ciphertext in Refs. [21] and [22].

Table 4. Comparison of message lengths (unit: bit)

Metric	Proposed Data Storage Scheme	Ref. [21]	Ref. [22]
L_A	$4G_1 + 172\,4$	$4G_1 + 2\|\tau\|$	$(3\eta_\tau + 9)G_1 + \|\tau\|$
L_U	$512 + 2C_F$	$(\eta_\tau + 3)G_1 + \|\tau\|$	$(6\eta_\tau + 7)G_1 + 2\|\tau\|$
L_{en}	$4G_1 + 233\,6 + 2C_F$	$(\eta_\tau + 7)G_1 + 3\|\tau\|$	$(9\eta_\tau + 16)G_1 + 3\|\tau\|$

Notes: L_A is the communicational overhead during the authentication and authorization access control phase; L_U represents the communicational overhead during the file uploading and downloading phase; L_{en} is the overall communicational overhead incurred between the phases of authentication and authorization access control and the phases of file uploading and downloading; C_F is the ciphertext length; $|\tau|$ is the size of the access policy; G_1 is the multiplicative group of integers modulo a prime number p; η_τ is the total number of attributes in the ciphertext in Refs. [21] and [22].

storage scheme are shorter compared to the message lengths presented in Refs. [21] and [22], because $(9\eta_\tau + 16)G_1 + 3|\tau| > (\eta_\tau + 7)G_1 + 3|\tau| > (1 + 7)G_1 = 8G_1 = 8\,192\,\text{bit} > 4G_1 + 2236 + 2C_F > 6\,432\,\text{bit}$. In order to ensure a robust level of security, we predicate upon the assumption of G_1 possessing a bit-length of 1 024 bit. The results show that the communicational costs of the data security storage scheme proposed in this paper are lower than those of Refs. [21] and [22].

5.3 Application of Proposed Scheme

We conducted further tests on our solution to assess its computational overhead in the uploading and downloading phases under varying file sizes, as depicted in Fig. 3. We conducted separate tests on binary files of sizes 1 MB, 10 MB, 100 MB, 256 MB, 512 MB, 800 MB, and 1 GB (1 024 MB), respectively. Each computation was derived from the average of ten independent runs, and the results were graphically presented. As illustrated in Fig. 3, the computational overhead for both uploading and downloading phases exhibits a linear increase with the file size. The computational costs on the data owner and data user during the uploading and downloading phases are also presented, with these costs primarily stemming from data encryption and decryption. From the figure, it is evident that the data owner incurs significantly greater computational overhead than the data user. This disparity arises from our utilization of AES in the CBC mode for encryption, a process necessitating sequential block-wise encryption, while the decryption process can be parallelized on multi-core CPUs.

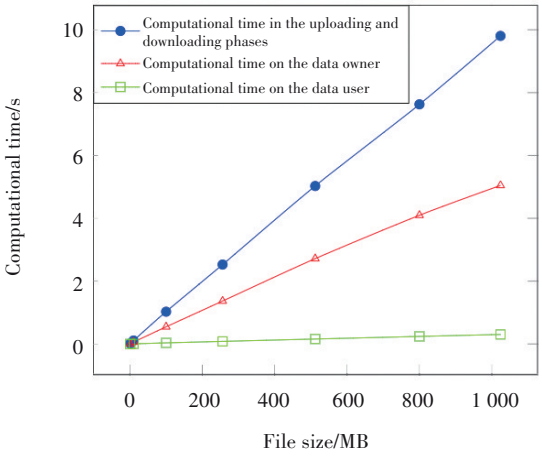


Figure 3. Computational time over different file sizes

To substantiate the efficacy of the proposed data storage scheme, we implemented and tested it in an experimental environment consistent with that described in Section 5.1. Notably, minor adjustments were made to the experimental environment to better align with our research objectives. We also constructed a composite system comprising four distinct computational units, which act as an authentication server, data user, data owner, and edge cloud, respectively. Intercommunication and data interchange amongst these computing entities were facilitated via a 50 Mbit/s Ethernet connection. After a sequence of iterative experiments, we found that the comprehensive execution time of the proposed data storage framework approached 60.71 ms, which is highly consistent with the data in Table 3. Further analysis of the experimental results revealed that the time overhead is mainly consumed by cryptographic operations, thus verifying the practical feasibility of the proposed scheme in real-world scenarios.

6 Conclusions

In this paper, we propose a novel data storage scheme based on identity authentication and access control for 5G industrial Internet collaborative systems. The scheme involves four entities: the data owner, the data user, the edge cloud, and the authentication server. Its distinctive feature is its identity authentication with access control. The data storage scheme includes five phases: system initialization, data uploading, mutual authentication, authorization and access control, and data downloading.

Compared to existing technologies, this approach offers several advantages for 5G industrial Internet collaborative systems. First, it can provide fine-grained access control that traditional schemes such as role-based access control cannot support. Second, it achieves higher efficiency than existing attribute-based access control schemes. Therefore, the proposed scheme provides both high efficiency and fine-grained access control that existing schemes lack. Moreover, this paper presents identity-based authentication and authorization techniques for checking the legitimacy of access requests. Experimental results demonstrate the feasibility of this data storage scheme in practical applications.

In the proposed scheme, the attributes of the data owner and data user are transparent to each other. However, in certain application scenarios, attribute privacy needs to be preserved. Therefore, in future work, we plan to design privacy-preserving

protocols and algorithms that allow the data owner and data user to compare their attributes without disclosing sensitive attribute information.

References

[1] Gao Y, Chen J J, and Li D P. Intelligence driven wireless networks in B5G and 6G era: a survey [J]. ZTE communications, 2024, 22(3): 99 – 105. doi: 10.12142/ZTECOM.202403012.
[2] Wang B Y, Li B C, Li H. Oruta: privacy-preserving public auditing for shared data in the cloud [J]. IEEE transactions on cloud computing, 2014, 2(1): 43 – 56. DOI: 10.1109/TCC.2014.2299807
[3] Shen J, Shen J, Chen X F, et al. An efficient public auditing protocol with novel dynamic structure for cloud data [J]. IEEE transactions on information forensics and security, 2017, 12(10): 2402 – 2415. DOI: 10.1109/TIFS.2017.2705620
[4] Jin H, Jiang H, Zhou K. Dynamic and public auditing with fair arbitration for cloud data [J]. IEEE transactions on cloud computing, 2018, 6(3): 680 – 693. DOI: 10.1109/TCC.2016.2525998
[5] Wang C, Wang Q, Ren K, et al. Privacy-preserving public auditing for data storage security in cloud computing [C]//Proc. IEEE INFOCOM. IEEE, 2010: 1 – 9. DOI: 10.1109/INFCOM.2010.5462173
[6] Yang P, Xiong N X, Ren J L. Data security and privacy protection for cloud storage: a survey [J]. IEEE access, 2020, 8: 131723 – 131740. DOI: 10.1109/ACCESS.2020.3009876
[7] Xu H, Sun B, Ding J W, et al. Analysis of feasible solutions for railway 5G network security assessment [J]. ZTE communications, 2025, 23(3): 59 – 70. doi: 10.12142/ZTECOM.202503007.
[8] Lee K. Comments on "secure data sharing in cloud computing using revocable-storage identity-based encryption" [J]. IEEE transactions on cloud computing, 2020, 8(4): 1299-1300. DOI: 10.1109/TCC.2020.2973623
[9] Xu S M, Yang G M, Mu Y. Revocable attribute-based encryption with decryption key exposure resistance and ciphertext delegation [J]. Information sciences, 2019, 479: 116 – 134. DOI: 10.1016/j.ins.2018.11.031
[10] Xiong H, Zhao Y N, Peng L, et al. Partially policy-hidden attribute-based broadcast encryption with secure delegation in edge computing [J]. Future generation computer systems, 2019, 97: 453 – 461. DOI: 10.1016/j.future.2019.03.008
[11] Wei J H, Chen X F, Huang X Y, et al. RS-HABE: revocable-storage and hierarchical attribute-based access scheme for secure sharing of e-health records in public cloud [J]. IEEE transactions on dependable and secure computing, 2021, 18(5): 2301 – 2315. DOI: 10.1109/TDSC.2019.2947920
[12] Li J Q, Wang S L, Li Y, et al. An efficient attribute-based encryption scheme with policy update and file update in cloud computing [J]. IEEE transactions on industrial informatics, 2019, 15(12): 6500 – 6509. DOI: 10.1109/TII.2019.2931156
[13] Zhang L Y, Cui Y L, Mu Y. Improving security and privacy attribute based data sharing in cloud computing [J]. IEEE systems journal, 2020, 14(1): 387 – 397. DOI: 10.1109/JSYST.2019.2911391
[14] Almutairi S, Alghanmi N, Mostafa M. Survey of centralized and decentralized access control models in cloud computing [J]. International journal of advanced computer science and applications, 2021, 12(2): 1 – 8. DOI: 10.14569/IJACSA.2021.0120243
[15] Li J G, Chen N Y, Zhang Y C. Extended file hierarchy access control scheme with attribute-based encryption in cloud computing [J]. IEEE transactions on emerging topics in computing, 2021, 9(2): 983 – 993. DOI: 10.1109/TETC.2019.2904637
[16] Ghaffar Z, Ahmed S, Mahmood K, et al. An improved authentication scheme for remote data access and sharing over cloud storage in cyber-physical-social-systems [J]. IEEE access, 2020, 8: 47144 – 47160. DOI: 10.1109/ACCESS.2020.2977264
[17] Liu T L, Wu J G, Li J X, et al. Efficient decentralized access control for secure data sharing in cloud computing [J]. Concurrency and computation: practice and experience, 2023, 35(17): e6383. DOI: 10.1002/cpe.6383
[18] Begum B R, Chitra P. SEEDDUP: a three-tier secure data deduplication architecture-based storage and retrieval for cross-domains over cloud [J]. IETE journal of research, 2023, 69(4): 2224 – 2241. DOI: 10.1080/03772063.2021.1886882
[19] Gao X, Yu J, Shen W T, et al. Achieving low-entropy secure cloud data auditing with file and authenticator deduplication [J]. Information sciences, 2021, 546: 177 – 191. DOI: 10.1016/j.ins.2020.08.021
[20] Shen W T, Su Y, Hao R. Lightweight cloud storage auditing with deduplication supporting strong privacy protection [J]. IEEE access, 2020, 8: 44359 – 44372. DOI: 10.1109/ACCESS.2020.2977721
[21] Premkamal P K, Pasupuleti S K, Singh A K, et al. Enhanced attribute based access control with secure deduplication for big data storage in cloud [J]. Peer-to-peer networking and applications, 2021, 14(1): 102 – 120. DOI: 10.1007/s12083-020-00940-3
[22] Cui H, Deng R H, Li Y J, et al. Attribute-based storage supporting secure deduplication of encrypted data in cloud [J]. IEEE transactions on big data, 2019, 5(3): 330 – 342. DOI: 10.1109/TBDATA.2017.2656120
[23] OpenSSL. org. OpenSSL–1.0.1e. tar. gz [EB/OL]. [2023-10-20]. http:// www.openssl.org/source
[24] Lynn B. PBC library manual 0.5.11.2006 [EB/OL]. [2023-10-20]. http://crypto.stanford.edu/pbc/manual

Biographies

Wang Jigang (wang.jigang@zte.com.cn) is General Manager of the Cybersecurity Product Line at ZTE Corporation. His research interests include operating systems, cybersecurity, and cloud computing. Dr. Wang has participated in and supported a number of national key science and technology projects and national science and technology support programs, and has published multiple academic papers.

Liu Dong is Deputy Director of the Central Research Institute at ZTE Corporation. His research interests include operating systems, cybersecurity, and cloud computing. He has participated in and supported a number of national key science and technology projects and national science and technology support programs, and has published multiple academic papers.

Wan Changsheng received his BS degree in applied physics and PhD degree in physical electronics from University of Science and Technology of China in 1999 and 2004, respectively. Since June 2009, he has been with Southeast University, where he is currently a professor with the School of Cyber Science and Engineering. His research interests include network security, wireless communication, and data mining.

Lu Ping is the Vice President and Director of the R&D Project in the Technology Planning Department at ZTE Corporation. He also serves as the Executive Deputy Director of the National Key Laboratory of Mobile Network and Mobile Multimedia Technology. His research directions include cloud computing, big data, augmented reality, and multimedia service-based technologies.

Complexity-Reduced Equalization for 200 Gbit/s PON Downstream Systems Based on SSB Modulation and Direct Detection

Yang Tao[1], Huang Xingang[2], Ma Zhuang[2], Zhong Yiming[2], Huang Xiatao[2], Liu Bo[2]

(1. State Key Lab of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. ZTE Corporation, Shanghai 201203, China)

DOI: 10.12142/ZTECOM.202601008

https://kns.cnki.net/kcms/detail/34.1294.TN.20260309.1357.002.html, published online March 10, 2026

Manuscript received: 2024-06-26

Abstract: The 200 Gbit/s passive optical network (PON) is most likely to be the next-generation scheme following 50G PON. The cost-effective direct detection (DD) system is the economical choice. However, larger-capacity DD systems will face much more serious power fading caused by chromatic dispersion (CD) combined with square-law DD and thereby significantly increases the complexity of equalization algorithms. In this paper, a 200 Gbit/s Nyquist 4-level pulse amplitude modulation (PAM4) single side-band (SSB) modulation-DD downlink scheme is designed, and a low complexity quadratic-nonlinear equalizer is proposed for this system. The computational complexity of the quadratic nonlinear equalizer is about 28% of that of the conventional Volterra nonlinear equalizer, while still exhibiting excellent nonlinear equalization ability. Simulation results for the 200 Gbit/s system with 20 km fiber transmission show that it can achieve a power budget of 29 dB, while a 30.4 dB power budget is obtained in the 50 Gbit/s experimental transmission.

Keywords: 200 Gbit/s passive optical network; single side band modulation; direct detection; equalization

Citation (Format 1): Yang T, Huang X G, Zhong Y M, et al. Complexity-reduced equalization for 200 Gbit/s PON downstream systems based on SSB modulation and direct detection [J]. ZTE Communications, 2026, 24(1): 56 - 64. DOI: 10.12142/ZTECOM.202601008

Citation (Format 2): T. Yang, X. G. Huang, Y. M. Zhong, et al. "Complexity-reduced equalization for 200 Gbit/s PON downstream systems based on SSB modulation and direct detection," *ZTE Communications*, vol. 24, no. 1, pp. 56 - 64, Mar. 2026. doi: 10.12142/ZTECOM.202601008.

1 Introduction

Passive optical networks (PON) based on power splitting provide a cost-effective optical access solution for delivering broadband access to diverse end-users. So far, most commercially deployed PONs are time-division multiplexed (TDM). The optical line terminal (OLT) at the central office is connected through an optical distribution network (ODN) with many optical network units (ONUs) at the user side using passive optical power splitting[1]. Gigabit passive optical networks (GPON) have dominated the early PON deployment, offering 100 Mbit/s broadband access services. To provide gigabit services, network operators have been upgrading the networks to 10G PON systems over the past decade. After 10G PON, ITU-T started to study PON technologies above 10 Gbit/s in 2016, and officially started to formulate the high-speed PON standard based on a 50 Gbit/s line rate in 2018. In September 2021, the first edition of the 50G PON standard was officially released by ITU-T. At present, with the rise of various emerging network applications, such as cloud storage and computing, AR/VR, 3D video, and IPTV, higher requirements are placed on the transmission capacity of PON systems. It can be expected that 50G PON will soon be insufficient to meet the needs of users, and the rapid growth of broadband access demand urges the access network to develop toward a higher line rate exceeding 100 Gbit/s. Generally speaking, the compound annual growth rate of user bandwidth demand is about 20%, increasing four- to fivefold every 8 to 10 years. Given that operators typically upgrade PON generations on an 8- to 10-year cycle, the line rate of the next-generation high-speed PON following 50G PON is very likely to be 200 Gbit/s, which also meets the rhythm and bandwidth requirements of PON network evolution.

Up to now, considering the system cost and energy efficiency, the system based on intensity modulation and direct detection (IM-DD) has been the best choice for PON systems. Unlike coherent detection, which requires a local oscillator

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20230105001 and National Natural Science Foundation of China under Grant No. 62001045.

and a complex receiver structure, IM-DD only needs a photodiode for detection at the receiving end, significantly reducing the system cost. For operators, it is necessary to upgrade the system without changing the existing ODN. To date, research on high-speed IM-DD transmission has been extensively conducted. To cope with the pressure of rate upgrade and device bandwidth improvement, IM-DD technology will combine with a series of advanced multilevel signal modulation technologies to carry more bits per unit bandwidth. Transmissions of 100 Gbit/s and above per wavelength have been reported by using various modulation formats including 4-level pulse amplitude modulation (PAM4), discrete multitone (DMT), and carrier-free amplitude-phase modulation (CAP)[2-6]. Among these, PAM4 is considered the mainstream candidate because of its simple configuration, low cost, and low power consumption[7]. The next-generation IM-DD-based high-speed PON systems are expected to work in the C-band with lower link loss than that of the O-band, which not only alleviates the current spectrum resource shortage[8], but also allows the compensation of link impairments such as fiber dispersion and nonlinear distortion through digital signal processing (DSP). In addition, using a small and easy-to-integrate semiconductor optical amplifier (SOA) can ensure sufficient power budget while further reducing system costs[9].

However, as optical access networks evolve toward a higher line rate exceeding 100 Gbit/s per wavelength, the transmission of IM-DD systems over the C-band optical fiber will be seriously affected by the chromatic dispersion (CD) effect. At high symbol rates, the system faces frequency-selective power fading, which will lead to serious degradation in receiver sensitivity and insufficient power budget[10]. Equalization technology is considered an effective solution to eliminating intersymbol interference (ISI) caused by device bandwidth limitation and CD. Among various techniques, feed-forward equalizers (FFE) and decision feedback equalizers (DFE) are widely used[11]. These equalizers have simple structures and low power consumption, but they cannot compensate for serious nonlinear damage. In contrast, the Volterra nonlinear equalizer (VNLE) can effectively compensate for both linear and nonlinear impairments originating from fiber channels and the square-law direct detection. However, its complex structure and high computational complexity (i.e., the cost and power consumption of hardware circuit implementations) make it difficult to apply the algorithm directly to future ultra-high-speed IM-DD PON systems.

To reduce the bandwidth requirements of key devices and mitigate power fading in high-speed IM-DD systems, a novel 200 Gbit/s Nyquist PAM4 single side-band (SSB)-DD downstream system is proposed in this paper. By analyzing the forms and characteristics of nonlinear damage in SSB-DD systems, we reasonably simplify the VNLE and propose a quadratic nonlinear equalizer with significantly reduced complexity. Simulation results show that the proposed equalizer can effectively compensate for nonlinear impairments in SSB-DD systems, and the computational complexity is equivalent to that of a linear equalizer. Using the proposed equalizer, the 200 Gbit/s simulation results over 20 km fiber transmission show that the system can achieve a power budget of 29 dB, and a 50 Gbit/s transmission experiment over 20 km yields a power budget of 30.4 dB at a bit error ratio (BER) threshold of 10^{-2}.

2 Principle of 200 Gbit/s PON Downstream Scheme

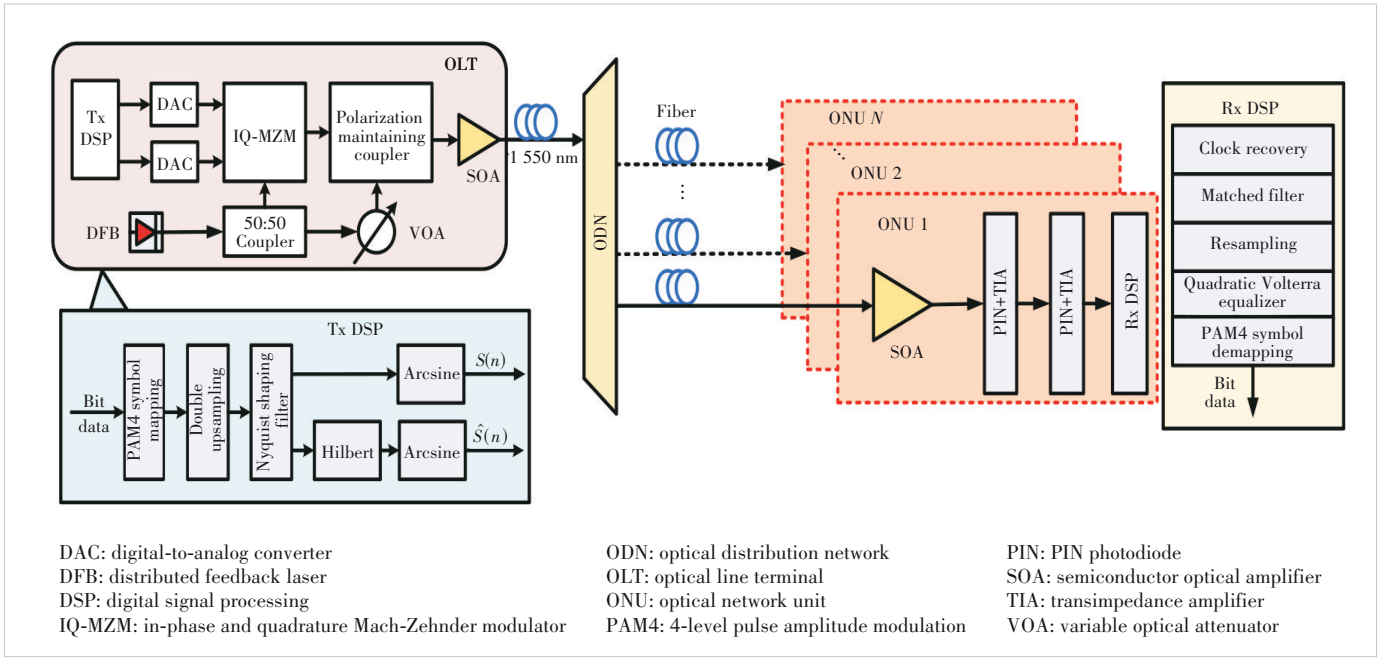
2.1 200 Gbit/s PON Downlink System Scheme

On the OLT side, if the bitrate is increased from 50 Gbit/s to 200 Gbit/s using a conventional non-return-to-zero/ on-off keying (NRZ/OOK) binary modulation format, optoelectronic devices with bandwidths of 100 GHz to 120 GHz are required, which is obviously extremely difficult and costly in practical applications. Therefore, it is necessary to adopt a high-order modulation format for the downlink transmission of the next-generation 200 Gbit/s PON systems to relax the bandwidth requirements of the devices. To keep low cost and achieve a high power budget, a 200 Gbit/s Nyquist PAM4 SSB-DD PON downlink transmission scheme is designed, as shown in Fig. 1.

To ensure an adequate power budget and acceptable device cost, the scheme adopts PAM4 modulation and performs Nyquist pulse shaping with a roll-off factor of 0.1 to further compress the signal spectrum, thereby reducing the bandwidth requirement of the device while minimizing the CD impact. Due to the limitation of high-speed digital-to-analog converters (DAC) in practical applications, only two-times waveform upsampling is considered. Because the direct detection system only detects light intensity information and is not sensitive to phase information, the system has high tolerance to the laser linewidth. As a result, we can choose a distributed feedback laser (DFB) as the light source at the transmitting end. Compared with generating single-sideband modulation signals based on a single dual-drive Mach-Zehnder modulator (MZM) with increased device cost on the OLT side, using IQ-MZM to generate single-sideband modulation signals in the form of optical auxiliary carriers can adjust the carrier-to-signal power ratio (CSPR) more flexibly to ensure optimal overall system performance. Fortunately, due to the point-to-multipoint nature of PON, the overall cost of the system will not increase significantly.

Here, let $s(t)$ be the original zero-mean PAM4 RF drive signal, and the drive voltage applied to the upper and lower arms of each MZM be $v(t) = s(t) + v_{dc}$. In the case of carrier suppression modulation, the v_{dc} bias voltage is set to the minimum transmission point, that is, the zero point of the power transfer function and the field transfer function. Then, the relationship between the input and the output light fields of MZM is:

$$E_{\text{out}}(t) = E_{\text{in}}(t) \times \cos\left(\frac{s(t)}{V_\pi}\pi - \frac{\pi}{2}\right) = E_{\text{in}}(t) \times \sin\left(\frac{s(t)}{V_\pi}\pi\right).$$



DAC: digital-to-analog converter
DFB: distributed feedback laser
DSP: digital signal processing
IQ-MZM: in-phase and quadrature Mach-Zehnder modulator
ODN: optical distribution network
OLT: optical line terminal
ONU: optical network unit
PAM4: 4-level pulse amplitude modulation
PIN: PIN photodiode
SOA: semiconductor optical amplifier
TIA: transimpedance amplifier
VOA: variable optical attenuator

Figure 1. 200 Gbit/s Nyquist PAM4 SSB-DD PON downlink system scheme

It can be seen that the modulation curve of the MZM is in the form of sine/cosine, and the signal will produce nonlinear distortion at both ends of the modulation dynamic range. To fully utilize the dynamic range of the MZM and maximize the system power budget, it is necessary to perform pre-distortion processing in the arcsine form before the OLT side signal enters the IQ-MZM to overcome the nonlinear effect of the MZM. Finally, an SOA is used to amplify the optical signal in the C-band before fiber transmission.

Generally, compared with a PIN photodiode at the same bandwidth, an avalanche photodiode (APD) offers significantly higher receiver sensitivity. However, on the ONU side, as the bitrate increases to 200 Gbit/s, APD also faces problems of insufficient device bandwidth and constrained gain. As the APD bandwidth increases, the gain decreases, and the sensitivity advantage over the PIN photodiode + transimpedance amplifier (PIN+TIA) scheme diminishes. To sum up, in a 200 Gbit/s PON downlink system, the use of PIN+TIA for photoelectric detection is both cost-effective and feasible for practical implementation. Before entering the PD, the optical signal first goes through an SOA for amplification. The sampling rates of the ADC and DAC are the same, which is twice the symbol rate, that is, 200 GSa/s. The receiver-side DSP processing flow on the ONU side includes clock recovery, matched filtering, double down sampling, and the proposed quadratic Volterra equalization.

2.2 Analysis of Nonlinear Damage in SSB-DD systems

The optical field expression of SSB signals generated by optical transmitters can be written as:

$$E_{\mathrm{SSB}}(t) = b + a[s(t) + \mathrm{jHT}[s(t)] = b + a \cdot c(t) \quad (1),$$

where a and b are complex numbers, so that $|a|^2 + |b|^2 = 1$ is used to constrain the average power of optical SSB signals to keep it constant; $s(t)$ is a real baseband signal, $\mathrm{HT}[\cdot]$ represents the Hilbert transform, and $c(t) = s(t) + \mathrm{jHT}[s(t)]$. $\mathrm{HT}[s(t)]$ can be expressed as $-\mathrm{jsign}(\omega)S(\omega)$ in the frequency domain. $s(t)$ is a real signal that satisfies $S(-\omega)^* = S(\omega)$ in the frequency domain. Therefore, all the information of $c(t)$ is contained on one side of the spectrum, such as the positive frequency side, which can be expressed as $C(\omega) = S(\omega)U(\omega)$, and $U(\omega)$ is a unit step function. After the optical SSB signal is transmitted through fibers, the received optical power can be expressed as[12]:

$$\begin{aligned} P_{\mathrm{Rx}}(t) &= \left|(b + a \cdot c(t)) \otimes h_{CD}(t)\right|^2 = \\ &\left|b + a \cdot c(t) \otimes h_{CD}(t)\right|^2 = \\ &|b|^2 + |a||b|\left[c(t) \otimes h_{CD}(t)\mathrm{e}^{\mathrm{j}\gamma} + c^*(t) \otimes h_{CD}^*(t)\mathrm{e}^{-\mathrm{j}\gamma}\right] + \\ &|a|^2\left|c(t) \otimes h_{CD}(t)\right|^2 = A'(t) + B'(t) + C'(t) \end{aligned} \quad (2).$$

In Eq. (2), $\gamma = \phi_b - \phi_a$ denotes the phase difference between b and a; the first term $A'(t) = |a|^2$ denotes the DC component, the second term $B'(t)$ has a useful signal $c(t)$, and the third term $C'(t) = |a|^2\left|c(t) \otimes h_{CD}(t)\right|^2$ denotes the signal-to-signal beat interference (SSBI) introduced by direct detection after transmission through the optical fiber. If $|a||b|$ is omitted for simplicity, the Fourier transform of $B'(t)$ is $B'(\omega) = C(\omega)H_{CD}(\omega)e^{\mathrm{j}\gamma} + C^*(-\omega)H_{CD}^*(-\omega)e^{-\mathrm{j}\gamma}$. Therefore,

$H_{CD}(\omega)e^{j\gamma} = G(\omega)$ and $B'(\omega)$ can be written as:

$$B'(\omega) = S(\omega)U(\omega)G(\omega) + S^*(-\omega)U^*(-\omega)G^*(-\omega) = S(\omega)\left(U(\omega)G(\omega) + U(-\omega)G^*(-\omega)\right) = S(\omega)M(\omega) \quad (3).$$

$H_{CD}(\omega) = H_{CD}(-\omega)$, and $M(\omega)$ can be summed up as:

$$M(\omega) = e^{j\mathrm{sign}(\omega)\left(\frac{1}{2}\beta_2\omega^2 L + \gamma\right)} \quad (4).$$

The inverse Fourier transform of $M(\omega)$ is a real value, which is expressed as $m(t)$. Since $C'(t)$ can be expressed as $|a|^2\left|c(t) \otimes h_{CD}(t)e^{j\gamma}\right|^2$ and $c(t)$ contains only positive spectral components, Eq. (2) can be rewritten as:

$$P_{\mathrm{Rx}}(t) = |b|^2\left(1 + \frac{|a|}{|b|}s(t) \otimes m(t) + \frac{|a|^2}{|b|^2}\left|c(t) \otimes m(t)\right|^2\right) \quad (5).$$

Obviously, the greatest advantage of SSB-DD is the elimination of the power fading effect. No power fading means that $m(t)$ is reversible, expressed as $m^{-1}(t)$. Therefore, transmission impairments can be effectively compensated by equalization algorithms.

In Eq. (5), $|b|^2$ denotes the power of the carrier, $|a|^2$ denotes the power of the signal, and $|b|^2/|a|^2$ denotes the carrier-to-signal power ratio (CSPR). $|a|^2/|b|^2$ can be expressed as 1/CSPR, and then Eq. (5) can be rewritten as:

$$P_{\mathrm{Rx}}(t) = |b|^2\left(1 + \frac{1}{\sqrt{\mathrm{CSPR}}}s(t) \otimes m(t) + \frac{1}{\mathrm{CSPR}}\left|c(t) \otimes m(t)\right|^2\right) \quad (6).$$

As can be seen from Eq. (6), SSBI becomes severe when the CSPR is too low. Although both the effective signal and SSBI decrease with increasing CSPR, the reduction rate of SSBI is significantly higher. When the CSPR is excessively high, the proportion of carrier power will be too large, which leads to a degraded signal-to-noise ratio (SNR). Consequently, the system sensitivity deteriorates. Hence, there is an optimal CSPR.

2.3 Principle of Low Complexity Nonlinear Equalizer

Traditional FFE and DFE feature simple structure and low power consumption, which can effectively compensate for linear distortion but cannot mitigate serious nonlinear impairments. According to Eqs. (1) and (5), both double sideband modulation and single sideband modulation signals transmitted through optical fiber suffer from SSBI, due to the square law detection of DD receivers, which causes signal distortion. VNLE can compensate for the nonlinear distortion, but it significantly increases the complexity of the equalizer, making it unsuitable for cost-sensitive ONU applications. Meanwhile, when the symbol rate is very high, the large number of tap coefficients renders VNLE impractical for real-time implementation[13].

Different from DSB intensity modulation, the SSBI in SSB-DD systems only contains the second-order term and avoids the power fading effect. Thus, the second-order nonlinear distortion becomes the main factor affecting the transmission performance, so the second-order VNLE is sufficient for effective compensation.

The output of the n-th sample of the second-order VNLE can be expressed as[14]:

$$y(n) = \sum_{m=0}^{L_1-1} w_1(m)x(n-m) + \sum_{l=0}^{L_2-1}\sum_{k=0}^{l} w_2(l,k)x(n-l)x(n-k) \quad (7),$$

where $x(n)$ is the n-th input sample, L_p and w_p are the memory length and kernel weight of order p ($p = 1, 2$), respectively. The second term on the right side of Eq. (7) denotes the nonlinear equalizer including the self-beat frequency term and the cross-beat frequency term, and its tap number is $L_2(L_2 + 1)/2$. Obviously, with the increase of L_2, the computational complexity of VNLE grows significantly.

Because of the square law detection of the DD receiver, the beat frequency term between signals is the most important nonlinear damage term. Therefore, by considering only the second-order distortion in Eq. (7), the second-order VNLE can be further simplified to a quadratic nonlinear equalizer (QNLE), which is expressed as:

$$y(n) = \sum_{m=0}^{L_1-1} w_1(m)x(n-m) + \sum_{l=0}^{L_2-1} w_2(l)x^2(n-l) \quad (8).$$

Obviously, the QNLE is much simpler than the second-order VNLE, as the number of taps in the second term is reduced from $L_2(L_2 + 1)/2$ to L_2. Eq. (8) indicates that under low dispersion conditions, the nonlinear term approximately contains only the square term of the signal. It should also be noted that the influence of fiber dispersion on linear and nonlinear terms is different, which needs to be compensated at the receiving end respectively. In short-range transmission, the self-timer frequency is dominant in the nonlinear terms of SSB-DD reception, so it is considered that the second term $\sum_{l=0}^{L_2-1} w_2(l)x^2(n-l)$ in Eq. (8) can compensate for the nonlinear damage. In the single sideband field modulation system, because there is no power fading, $M(\omega) = e^{j\mathit{sign}(\omega)\left(\frac{1}{2}\beta_2\omega^2 L + \gamma\right)}$, whose inverse Fourier transform, expressed as $m(t)$, is a real value; and there is a reversible $m^{-1}(t)$, so the linear term $s(t) \otimes m(t)$ can be compensated by $\sum_{m=0}^{L_1-1} w_1(m)x(n-m)$ in Eq. (8).

When equalization performance is equivalent, computa-

tional complexity becomes a key indicator for evaluating the quality of an equalizer. Because multiplication operations are significantly more computationally expensive than addition operations, the computational complexity can be judged by the number of multiplications in the equalizer. The computational complexity of the complete second-order Volterra equalizer can be expressed as:

$$M = L_1 + L_2(L_2 + 1)/2 \quad (9),$$

where L_1 represents the number of taps of the first-order term, and L_2 represents that of the second-order term. The computational complexity of the proposed QNLE can be expressed as:

$$M' = L_1 + L_2 \quad (10).$$

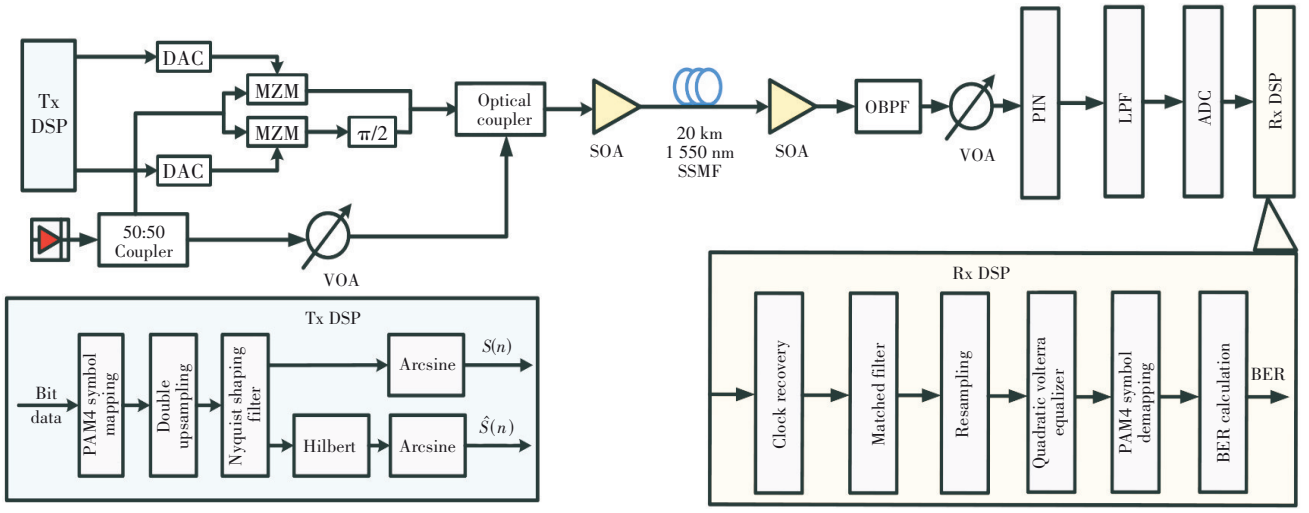
By comparing Eqs. (9) and (10), it can be found that the quadratic Volterra equalizer reduces the computational complexity from quadratic to linear growth, thus greatly reducing the computational complexity.

3 Simulation Setup and Results

The simulation setup is shown in Fig. 2. At the transmitter, a binary pseudo-random bit sequence goes through PAM4 symbol mapping, two-times up-sampling, symbol shaping, and spectrum compression using a root raised cosine filter with a roll-off factor of 0.1. The shaped signal is transformed into its real and imaginary parts via the Hilbert transform. After this processing, the two signals are sent to the IQ modulator following DAC to modulate the optical single sideband signal. The modulated signal is amplified by an SOA and then enters the optical fiber for transmission. The variable optical attenuator (VOA) on the transmitter side is used to control the power of the optical carrier signal and thus adjust the CSPR. The optical signal is further amplified using an SOA as a preamplifier, and the noise outside the signal spectrum is filtered by an optical band-pass filter (OBPF) before PD detection at the receiver. The received optical power is adjusted by another VOA. The electrical signals directly detected by the PD (55 GHz 3 dB bandwidth) are processed offline after passing through the ADC. The receiver DSP includes the Gardner clock recovery algorithm, matched filtering, downsampling, the low-complexity quadratic nonlinear equalizer, symbol de-mapping, and BER calculation. Table 1 shows the main simulation parameters of the system.

CSPR is one of the key parameters of the system. If the CSPR is too high, the proportion of carrier power will be too large, which degrades receiver sensitivity. Conversely, if the CSPR is too low, the SSBI impairment will be too strong to be effectively mitigated at the ONU side, which also degrades receiver sensitivity. Therefore, there is an optimal CSPR value that maximizes receiver sensitivity.

First, the BER varies with CSPR under the condition of 20 km optical fiber transmission. The input optical power is 0 dBm to avoid exciting nonlinear effects in the fiber, while the average received optical power is −10 dBm. A 71-tap FFE equalizer is employed at the receiver. Since the SSBI impair-



ADC: analog-to-digital converter
BER: bit error ratio
DAC: digital-to-analog converter
DSP: digital signal processing
LPF: low-pass filter
MZM: Mach-Zehnder modulator
OBPF: optical band-pass filter
SOA: semiconductor optical amplifier
VOA: variable optical attenuator
PAM4: 4-level pulse amplitude modulation
SSB: single side-band
PD: Photodiode
PIN: PIN photodiode

Figure 2. Simulation setup of 200 Gbit/s Nyquist PAM4 SSB transmission over 20 km SSMF

Table 1. Main parameters of simulation system

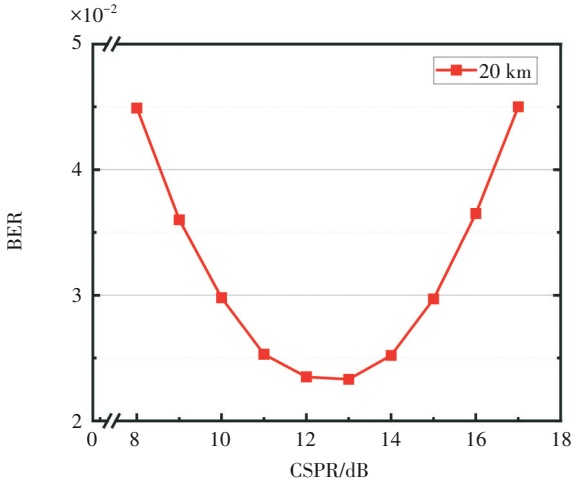
Parameter	Value	Parameter	Value
Bit rate	200 Gbit/s	DAC/ADC sampling rate	200 GSa/s
DAC/ADC ENOB	8	Laser linewidth	10 MHz
MZM extinction ratio	20 dB	SOA noise figure	7.5 dB
Dispersion coefficient	16.5 ps/nm/km	PIN responsiveness	0.8 A/W
PIN thermal noise	1.2e-11 A/Hz^(1/2)	Receiver bandwidth	55 GHz

ADC: analog-to-digital converter
DAC: digital-to-analog converter
ENOB: effective number of bits
MZM: Mach-Zehnder modulator
PIN: PIN photodiode
SOA: semiconductor optical amplifier

ment is also related to dispersion, its impact differs between back-to-back (BtB) and 20 km transmission scenarios. The 20 km transmission results are more representative, as they primarily involve dispersion and SSBI impairments in the system. Since the theoretical sensitivity under BtB conditions is −10 dBm, this paper first optimizes the CSPR around −10 dBm received power. The FFE equalizer can mitigate linear ISI caused by dispersion, but it cannot effectively equalize the nonlinear impairment of SSBI. Therefore, after equalization, there will be residual SSBI impairment, and the size of SSBI impairments is dependent on the system CSPR. In addition, the measured BER curve trend can better reflect the influence of CSPR change on receiver sensitivity, when the received optical power is low. As shown in Fig. 3, for a transmission distance of 20 km, the system achieves the lowest BER and optimal performance when the CSPR is 13 dB.

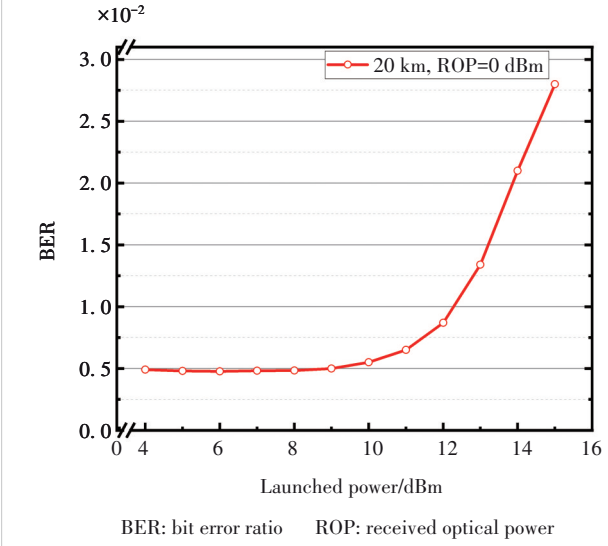
Next, the relationship between launched power and BER is simulated under the transmission condition of 20 km. To ensure that the change of BER is caused only by power variation (excluding receiver noise influence), the received power is set to 0 dBm. As shown in Fig. 4, with the increase of launched power, fiber nonlinearity is gradually excited, which leads to the system performance degradation. When the launched power exceeds 10 dBm, nonlinear effects become significant. Therefore, 10 dBm is selected as the maximum launched power and used as the condition for subsequent simulations.

Fig. 5 depicts the BER versus different first-order and second-order tap lengths at a received optical power (ROP) of −9 dBm, using the proposed low-complexity quadratic nonlinear equalizer. To determine the optimal tap configuration in the vicinity of the sensitivity, the received power is fixed at −9 dBm. It can be seen that the quadratic nonlinear equalizers with 61 first-order tap lengths and 21 second-order tap



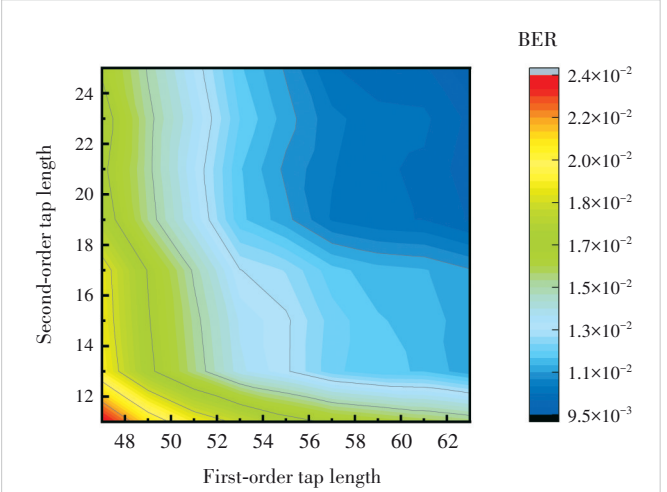
BER: bit error ratio CSPR: carrier-to-signal power ratio

Figure 3. BER versus CSPR under a 20 km transmission simulation



BER: bit error ratio ROP: received optical power

Figure 4. Launched power versus BER



BER: bit error ratio

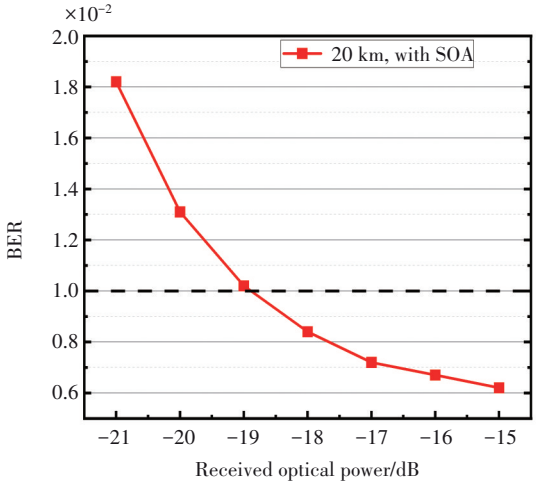
Figure 5. BER versus filter length at −9 dBm received optical power

lengths achieve a bit error rate threshold of 10^{-2}. As discussed in Section 2, the distortion of the SSB-DD system in the C-band mainly comes from inter-symbol interference caused by dispersion and second-order SSBI, so the quadratic nonlinear equalizer can effectively equalize it. Because the first-order tap length is 61 and the second-order tap length is 21, the ratio of the tap numbers of the quadratic nonlinear equalizer to that of the Volterra nonlinear equalizer is 82/292 (approximately 28.1%). This means that the quadratic nonlinear equalizer can significantly reduce the computational complexity. In addition, the results imply that increasing the number of second-order taps can lead to substantial complexity reduction.

Fig. 6 shows the 200 Gbit/s Nyquist PAM4 SSB transmission over a 20 km standard single-mode fiber (SSMF), achieving a 29 dB total link power budget at a BER of 10^{-2}, taking into account 10 dBm launched power.

4 Experiment and results

The experimental platform is shown in Fig. 7. Due to the bandwidth limitations (14 GHz) of the IQ modulator used in the lab, a 200 Gbit/s transmission experiment remains challenging, and thus our experiment was carried out at a symbol rate of 25 GBaud. The OLT part mainly includes an AWG, an optical IQ modulator, a 1 550 nm DFB laser, a 50∶50 2×1 coupler, a polarization controller, and an adjustable attenuator. First, offline Tx DSP is programmed in MATLAB at the originator. Random bit sequences are generated and mapped to PAM4 symbols (normalized from −1 to 1). The optical IQ modulator operates in both linear and nonlinear regions. If the amplitude of the input signal is too large, it will enter the nonlinear region. Consequently, Tx DSP needs to perform nonlinear pre-distortion of the modulator to resist the nonlinear impairment of the IQ modulator. The generated sequence is divided into two parts, and one of the sequences is subjected to Hilbert transformation. Finally, the two sequences are imported into the arbitrary waveform generator (AWG) as I and Q columns at the same time. In the offline experiment, a Keysight M8194A AWG (3 dB bandwidth: 45 GHz, sampling rate: 120 GSa/s with two channels active) is used to generate 2-channel signals, which work with a Coherent Solutions IQ Transmitter-SP-ABC to realize single-polarization carrier suppression optical modulation. To modulate the optical domain single sideband and control the power ratio of the carrier signal, the output light of the laser was split into two paths via a 50∶50 optical coupler. One path is used as the modulation light source of the IQ modulator, and the other serves as the optical carrier. An adjustable optical attenuator is used to adjust the optical power. The CSPR of the system can be obtained by measuring the power of the output optical signal of the IQ modulator first, and then measuring the output optical power of the adjustable optical decay. After that, the optical carrier signal passes through a polarization controller to keep the polarization state of the output optical signal of the IQ modulator consistent. Finally, it is coupled with the output optical signal of the IQ modulator through a 50∶50 2×1 optical coupler, and then enters the optical fiber for transmission with an SOA power amplifier with a transmit power of 10 dBm.

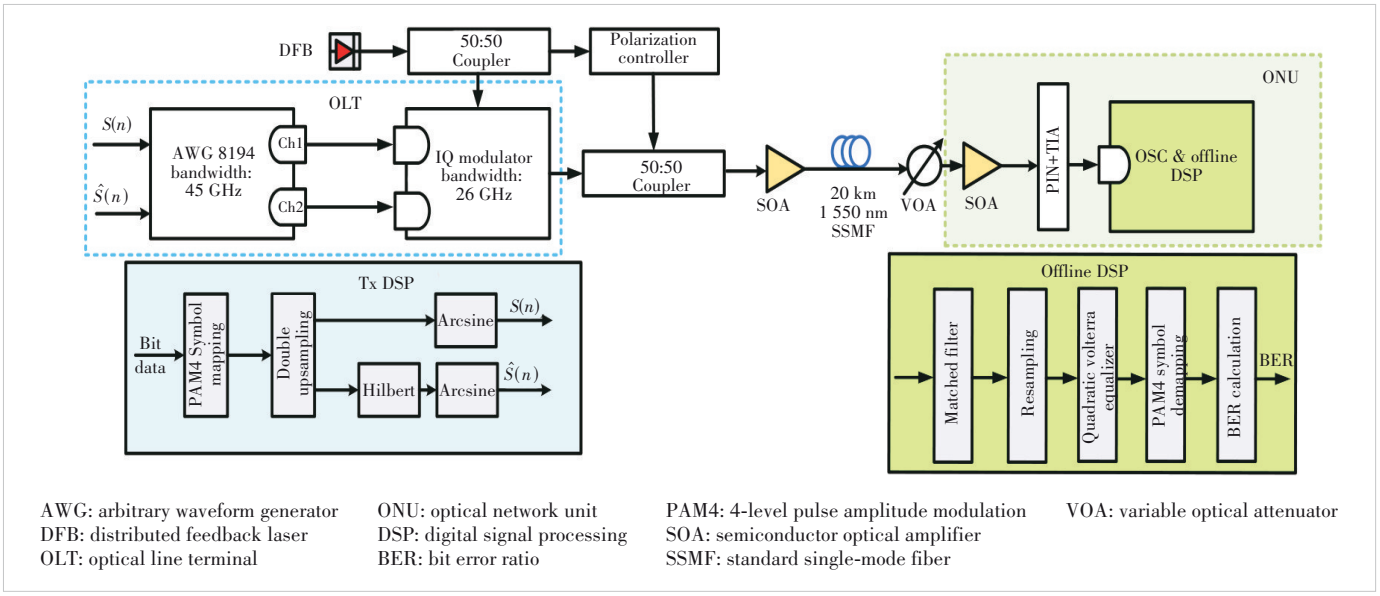


BER: bit error ratio SOA: semiconductor optical amplifier

Figure 6. BER vs ROP under a 20 km transmission simulation

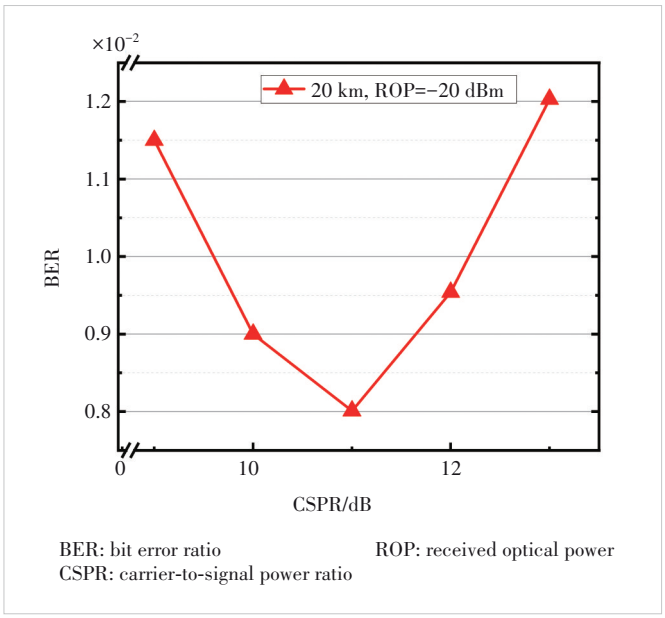
The ONU part mainly includes a PD detection module and a sampling oscilloscope. The optical signal arriving at the ONU first passes through an SOA preamplifier, then through the PD detection module, and is subsequently sampled by the oscilloscope. The responsivity of the PD is 0.65 A/W, and the 3 dB bandwidth is 36 GHz. The signal is sampled by a sampling oscilloscope (LabMaster 10-59Zi-A). When using only one channel, it operates with a sampling rate of 160 GSa/s, a 3 dB bandwidth of 59 GHz, and a resolution of 8 bits. At present, the main DSP algorithm for off-line processing includes matched filtering, double down-sampling, and second-order Volterra equalization. Finally, decision decoding is performed in the data recovery module to restore the original transmitted data, and the error statistics are carried out.

The experimental relationship between BER and CSPR for 20 km transmission is illustrated in Fig. 8, using the same 71-tap FFE configuration as in the simulation. First, the BER varies with CSPR in the case of 20 km fiber transmission with an average received optical power of −20 dBm. As analyzed in Sections 2.3 and 3, while FFE can mitigate linear ISI, the nonlinear SSBI remains a dominant impairment that depends on the CSPR. The BER performance under experimental conditions (20 km transmission) is more representative of these combined effects. As shown in Fig. 8, the experimental system achieves the optimal BER at a CSPR of 11 dB. This is slightly lower than the 13 dB CSPR observed in simulation (see Fig. 3), likely due to the additional bandwidth constraints and noise characteristics of the physical IQ modulator and SOA used in the experimental setup. Therefore, a CSPR of 11 dB



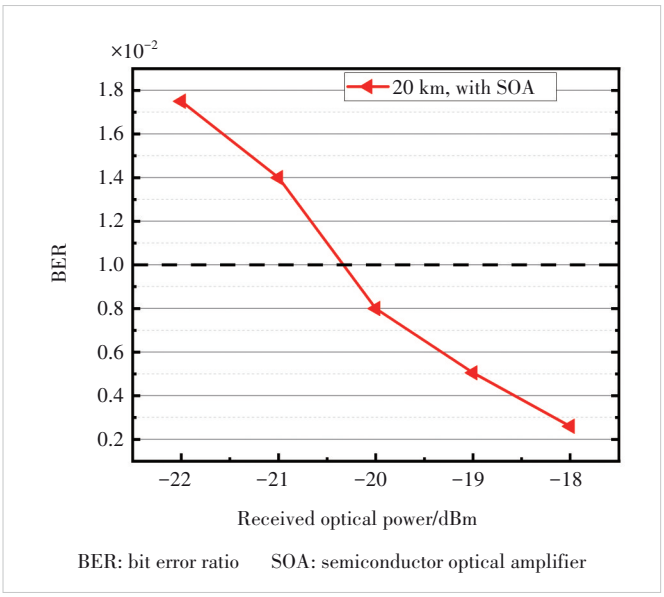
AWG: arbitrary waveform generator
DFB: distributed feedback laser
OLT: optical line terminal
ONU: optical network unit
DSP: digital signal processing
BER: bit error ratio
PAM4: 4-level pulse amplitude modulation
SOA: semiconductor optical amplifier
SSMF: standard single-mode fiber
VOA: variable optical attenuator

Figure 7. 20 km 25 GBaud PAM4 single sideband modulation-direct detection point-to-point downlink experimental platform



BER: bit error ratio
CSPR: carrier-to-signal power ratio
ROP: received optical power

Figure 8. BER vs CSPR for 20 km transmission experiment



BER: bit error ratio
SOA: semiconductor optical amplifier

Figure 9. BER vs ROP under a 20 km transmission experiment

was adopted in subsequent experiments.

Fig. 9 shows how BER varies with ROP. Under the current experimental conditions, the receiver sensitivity of the 20 km transmission experimental system is −20.4 dBm. At this sensitivity, the second-order Volterra equalizer employs 91 linear taps and 21 second-order taps, and the overall power budget of the system is 30.4 dB. At 1550 nm (where dispersion is greater than 300 ps/nm), the quadratic Volterra equalizer is considered effective. Ref. [15] demonstrates a PON system with real-time transmission of 50 Gbit•s^{-1}•λ^{-1} PAM4 using a booster SOA. This system achieves a sensitivity of −22.3 dBm at a BER of 10^{-2}, with a power budget higher than 35 dB across the O-band. The performance of this experiment is essentially comparable to, though slightly lower than, the results reported in Ref. [15]. The gap is attributed to the single sideband modulation method, operating wavelength, bandwidth limitations, and SOA amplifier performance.

5 Conclusions

In this paper, a 200 Gbit/s Nyquist PAM4 SSB-DD downlink system scheme is proposed and demonstrated, and a low-complexity QNLE is proposed for low complexity nonlinear im-

pairment equalization, which can greatly reduce the computational complexity of the conventional VNLE by keeping only the primary linear terms and quadratic nonlinear terms. Simulation results for the 200 Gbit/s SSB-DD downlink system show that the computational complexity of the quadratic nonlinear equalizer is about 28% of that of the conventional Volterra nonlinear equalizer. The proposed QNLE exhibits excellent nonlinear equalization ability, as it achieves a power budget of 29 dB under the 20 km fiber transmission, while a power budget of 30.4 dB is achieved in the 50 Gbit/s experiment. It is concluded that the proposed SSB-DD PON scheme with reduced complexity quadratic Volterra equalization algorithm yields superior system performance, and thus is one of the most promising solutions for future 200 Gbit/s PON applications.

References

[1] Houtsma V, Van Veen D. High speed optical access networks for this decade and the next (invited) [C]//IEEE Photonics Conference (IPC). IEEE, 2022: 1 – 2. DOI: 10.1109/IPC53466.2022.9975573
[2] Gao Y L, Cartledge J C, Kashi A S, et al. Direct modulation of a laser using 112-Gb/s 16-QAM nyquist subcarrier modulation [J]. IEEE photonics technology letters, 2017, 29(1): 35 – 38. DOI: 10.1109/lpt.2016.2627000
[3] Xue L, Lin R, Van Kerrebrouck J, et al. 100G PAM-4 PON with 34 dB power budget using joint nonlinear tomlinson-harashima precoding and Volterra equalization [C]//European Conference on Optical Communication (ECOC). IEEE, 2021: 1 – 4. DOI: 10.1109/ecoc52684.2021.9606041
[4] Wang W Y, Li H L, Zhao P C, et al. Advanced digital signal processing for reach extension and performance enhancement of 112 Gbit/s and beyond direct detected DML-based transmission [J]. Journal of lightwave technology, 2019, 37(1): 163 – 169. DOI: 10.1109/JLT.2018.2885707
[5] Xie C J, Dong P, Randel S, et al. Single-VCSEL 100-Gb/s short-reach system using discrete multi-tone modulation and direct detection [C]//Optical Fiber Communication Conference. OSA, 2015: Tu2H. 2. DOI: 10.1364/ofc.2015.tu2h.2
[6] Sun L, Du J B, He Z Y. Multiband three-dimensional carrierless amplitude phase modulation for short reach optical communications [J]. Journal of lightwave technology, 2016, 34(13): 3103 – 3109. DOI: 10.1109/JLT.2016.2559783
[7] Xiang M, Fu S N, Xu O, et al. Advanced DSP enabled C-band 112 Gbit/s/λ PAM-4 transmissions with severe bandwidth-constraint [J]. Journal of lightwave technology, 2022, 40(4): 987 – 996. DOI: 10.1109/jlt.2021.3125336
[8] Zhong F, Gong P, Zhou Z P, et al. High performance optical modulator and detector for 100 Gbit/s transmission system [J]. ZTE communications, 2017, 15(3): 46 – 51. DOI: 10.3969/j.issn.16735188.2017.03.006
[9] Yan B L, Wu Q, Shi H, et al. Toward low-cost flexible intelligent OAM in optical fiber communication networks [J]. ZTE communications, 2022, 20 (3): 54 – 60. DOI: 10.12142/ZTECOM.202203007
[10] Tang X Z, Qiao Y J, Chang G K. Experimental demonstration of C-band 112-Gb/s PAM4 over 20-km SSMF with joint pre- and post-equalization [C]//Optical Fiber Communication Conference (OFC) 2020. Optica Publishing Group, 2020. DOI: 10.1364/ofc.2020.w2a.46
[11] Tang X Z, Zhou J, Guo M Q, et al. An efficient nonlinear equalizer for 40-Gb/s PAM4-PON systems [C]//Proceedings of Optical Fiber Communication Conference. OSA, 2018: 1 – 3. DOI: 10.1364/ofc.2018.w2a.62
[12] Chagnon M. Optical communications for short reach [C]//European Conference on Optical Communication (ECOC). IEEE, 2018: 1 – 3. DOI: 10.1109/ECOC.2018.8535355
[13] Kaneda N, Lee J, Chen Y K. Nonlinear equalizer for 112-Gb/s SSB-PAM4 in 80-km dispersion uncompensated link [C]//Proceedings of Optical Fiber Communication Conference. OSA, 2017: 1 – 3. DOI: 10.1364/ofc.2017.tu2d.5
[14] Yu Y K, Choi M, Bo T W, et al. Low-complexity quadratic equalizer for DML-based IM/DD systems [C]//Proceedings of 24th OptoElectronics and Communications Conference (OECC) and 2019 International Conference on Photonics in Switching and Computing (PSC). IEEE, 2019: 1 – 3. DOI: 10.23919/PS.2019.8817645
[15] Lee H H, Kim K, Doo K H, et al. Demonstration of high-power budget TDM-PON system with 50 Gb/s PAM4 and saturated SOA [J]. Journal of lightwave technology, 2021, 39(9): 2762 – 2768. DOI: 10.1109/JLT.2021.3059902

Biographies

Yang Tao (yangtao@bupt.edu.cn) is an associate professor at the School of Electronic Engineering, Beijing University of Posts and Telecommunications (BUPT), China. He received his PhD degree in information and communication engineering from BUPT in 2019, followed by a postdoctoral fellowship in Electronic Science and Technology. His research interests include ultra-high-speed optical transmission, digital signal processing algorithms for optical communications, and intelligent optical network monitoring and management. He has authored or co-authored over 70 papers in prestigious journals and conferences like Optics Express and OFC, and holds more than 12 patents. He also serves as a guest editor and reviewer for several international journals.

Huang Xingang is a senior expert in technical pre-research with the Fixed Network Product Line, ZTE Corporation. He has long been engaged in the research and standardization of optical access technologies.

Ma Zhuang is the Head of Fixed Network Product Technology Pre-Research Department at ZTE Corporation. His work focuses on the pre-research, planning, and development of fixed broadband and optical access products. He has published more than 10 technical papers and is the inventor of over 20 granted patents.

Zhong Yiming is a senior engineer with the Fixed Network Pre-research Department, ZTE Corporation. His research centers on the R&D and innovation of PON products, with a primary focus on 50G PON system architecture design and algorithm research. He actively participates in PON standardization in ITU-T SG15, CCSA, and other industrial standards organizations, and has submitted numerous international standard contributions. He holds rich technical expertise in PON systems and has been granted multiple national invention patents.

Huang Xiatao is currently a technology research engineer at ZTE Corporation. He received his PhD degree in information and communication engineering from University of Electronic Science and Technology of China in 2023. His research interests include ultra-high-speed optical transmission, digital signal processing algorithms for optical coherent communications, and optical intelligent access system. He has held more than 6 patents.

Liu Bo is a chief pre-research engineer of OAN Product Line, ZTE Corporation. His research focuses on the pre-research of PON technologies.

Enhancing Code Quality with LLM in Software Static Analysis

Niu Zhi[1,2], Dong Luming[1,2]
(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202601009
https://kns.cnki.net/kcms/detail/34.1294.TN.20240725.1454.002.html, published online July 25, 2024
Manuscript received: 2024-04-18

Abstract: In the modern era of ubiquitous and highly interconnected information technology, cybersecurity threats stemming from software code vulnerabilities have become increasingly severe, posing significant risks to the confidentiality, integrity, and availability of modern information systems. To enhance software code quality, enterprises often integrate static code analysis tools into Continuous Integration (CI) pipelines. However, the high rates of false positives and false negatives remain a challenge. The advent of large language models (LLMs), such as ChatGPT, presents a new opportunity to address these challenges. In this paper, we propose AI-SCDF, a framework that utilizes the custom-built Nebula-Coder AI model for detecting and fixing code security issues in real time during the developer's personal build process. We construct a static code checking rule knowledge base through summarizing and classifying Common Weakness Enumeration (CWE) code security problems identified by security and quality assurance teams. The rule knowledge base is combined with CodeFuse-processed code contexts to serve as input for an AI code security detection microservice, which assists in identifying code quality and security issues. If any abnormalities are detected, they are addressed by an AI code security patching microservice, which alerts the developer and requests confirmation before committing the code into the repository. Experimental results show that our approach effectively improves code quality. We also develop a VSCode plugin for code alert detection and fix based on LLMs, which facilitates test shift-left and lowers the risk of software development.

Keywords: software static analysis; LLM; CWE; knowledge base

Citation (Format 1): Niu Z, Dong L M. Enhancing code quality with LLM in software static analysis [J]. ZTE Communications, 2026, 24(1): 65 - 71. DOI: 10.12142/ZTECOM.202601009
Citation (Format 2): Z. Niu and L. M. Dong. "Enhancing code quality with LLM in software static analysis," *ZTE Communications*, vol. 24, no. 1, pp. 65 - 71, Mar. 2026. doi: 10.12142/ZTECOM.202601009.

1 Introduction

Static Application Security Testing (SAST) is a technique that analyzes source code to detect software vulnerabilities. The general architecture of SAST involves three stages: preprocessing, analysis, and detection. In the first stage, SAST tools use compilers to preprocess source code and generate intermediate files. Based on these files, they then perform lexical analysis and generate an abstract syntax tree (AST). In the second stage, SAST tools analyze the AST to conduct data flow analysis, control flow analysis, dependency analysis, interval analysis, and pointer analysis, thereby obtaining mathematical representations of the program. Finally, in the third stage, SAST tools solve constraints using these mathematical expressions and well-known vulnerability patterns from a database to achieve high-precision detection of code security issues.

State-of-the-art SAST tools, such as Coverity[1], Klocwork[2], Infer[3], SonarQube[4], and CodeQL[5], play a crucial role in code security assessment. Therefore, many enterprises must choose appropriate SAST tools for their Continuous Integration (CI) process. However, for commercial software, achieving high-level static code analysis often requires compiling source code. This adds time overhead to enterprise CI builds. Conversely, open-source tools are mostly rule-based and thus prone to generating false warnings. This increases the cost of eliminating alerts and reduces software productivity.

With the development of machine learning (ML), researchers have begun incorporating it into SAST to detect vulnerabilities in source code. Harer, Kim, et al. proposed a data-driven ML-based approach for automated software vulnerability detection[6]. Li, Zou, et al. developed VulDeePecker, a deep learning-based vulnerability detection system for detecting vulnerabilities in software code[7 - 8]. Wang, Liu, et al. utilized Deep Belief Networks (DBNs) to automatically learn code semantic features from AST to predict code defects[9]. Li, He et al. extracted markings from program ASTs and applied convolutional neural networks to detect software defects[10]. Experimental results demonstrate that pre-trained ML models significantly improve the effectiveness of detecting and fixing program vulnerabilities.

Since OpenAI released GPT-3[11] in 2020, large language models (LLMs) have entered a phase of rapid development. Early and prominent examples include Google's PaLM2[12], Microsoft's Copilot[13], and Meta's open-source LLaMa[14]. These LLMs are widely used in the software development lifecycle (SDLC), laying the foundation for subsequent advances in AI-assisted programming. The superior ability of LLMs to understand complex code makes them capable of providing infinite possibilities in improving code defect detection and self-fix. Berabi, Gronskiy, et al. fine-tuned GPT-4 with data on code contexts, defect reports, and code snippets to achieve automatic software vulnerability fix[15]. Wadhwa and Prdadhan used a pair of LLMs to improve code quality through software code quality improvement and to rank the improvements, respectively[16]. Li, Hao, et al. developed the LLift fully automatic framework by combining static analysis with LLMs to detect use-before-initialization defects[17].

These findings show that combining LLMs (such as GPT-4) with static analysis can significantly improve code quality and provide new ideas for automated code fix. However, these studies also highlight the limitations of LLMs' ability to understand and handle complex code logic, which requires further research and practice.

In light of this, we propose AI-SCDF, a framework that utilizes the Nebula-Coder LLM developed by ZTE to detect and fix code changes during the developers' personal build process. AI-SCDF does not require compiling the source code for defect detection. Instead, it analyzes the contextual information of the code to facilitate LLM learning and reasoning. This significantly enhances code checking efficiency while overcoming the limitations of commercial static code analysis tools. Moreover, it addresses the high false alarm rate issue prevalent in open-source static code analysis tools. First, we summarize the Common Weakness Enumeration (CWE) code security issues based on rules from security and quality teams to build a static code checking rule knowledge base. Second, this knowledge base, combined with processed code context, serves as input to an AI code flaw detection microservice, which intercepts potentially flaw code. Once potential flaw code is detected, we employ the AI code flaw fix microservice to rectify it, and feedback is provided to the developer to confirm the improvement in coding quality and prevent flaw code from being committed to the code repository. We mainly use the Juliet Test Suite[18] to evaluate and verify the flaw detection and fix capabilities of AI-SCDF.

The main contributions of the paper are summarized as follows:

1) Static analysis checker knowledge base: We developed a powerful static analysis checker knowledge base system that covers CWE vulnerability types for various programming languages and includes a large number of violating cases and fixing samples. These cases are derived from open-source code and real-world closed-source code flaws found during continuous integration. The sample data serve as input for LLM learning.

2) Code context analysis: We utilized control flow analysis tools to extract the context of code changed by developers. This enables LLMs to better understand software structure while avoiding the high token cost of processing entire codebases.

3) LLM utilization: We developed AI-SCDF, a code flaw detection and fix framework based on LLMs, and integrated it into personal build processes to improve code quality and reduce software development risks.

4) Results: Through testing on the Juliet Java 1.3 test set and analyzing metrics including true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), accuracy, precision, and recall, we found that AI-SCDF effectively improved code quality and reduced the cost of alert fixing.

2 Overview

Our objective is to develop a fully automated framework that detects and fixes defects during the developer's personal build process to improve code quality and reduce the labor costs of the R&D team. To this end, our approach leverages the context of changed code and a static analysis checker knowledge base to construct prompts that guide LLMs in code flaw detection. Subsequently, information about any detected defective code is combined with these detection prompts to form code flaw fix prompts, which then guide the LLMs to perform automatic code fix.

Consider the Java code snippet shown in Fig. 1, where the resource objects named pstmt and rs are not properly closed. This may cause resource leakage until the resources are exhausted, leading to software quality issues. In actual operations, such resources are typically managed using a try-catch-finally block to ensure they are properly closed.

As shown in Fig. 2, the AI-SCDF workflow involves the following stages:

1) Static analysis checker knowledge base: The knowledge base includes checkers, checker descriptions, severity levels, code flaw violations and corresponding fixed examples for each checker across different programming languages, as well as custom severity ratings. This data is derived from coding best practices, coding standards, commercial and open-source static analysis tools, along with a large number of data sources and cases. By integrating CWE information, a unified static analysis checker knowledge base is formed. The data in this knowledge base are used by subsequent AI microservices to

```
@@ -0,0 +1,11 @@
+ public  List<String> runQuery(Connection conn,String kb) throws SQLException{
+     List<String> examples = new ArrayList<String>();
+     PreparedStatement pstmt = conn.prepareStatement("SELECT * FROM knowledge WHERE kb = ?");
+     pstmt.setString(1,kb);
+     ResultSet rs = pstmt.executeQuery();
+     while(rs.next()){
+       examples.add(rs.getString("example"));
+     }
+     return examples;
+
+  }
```

Figure 1. Flaw code of resource leak

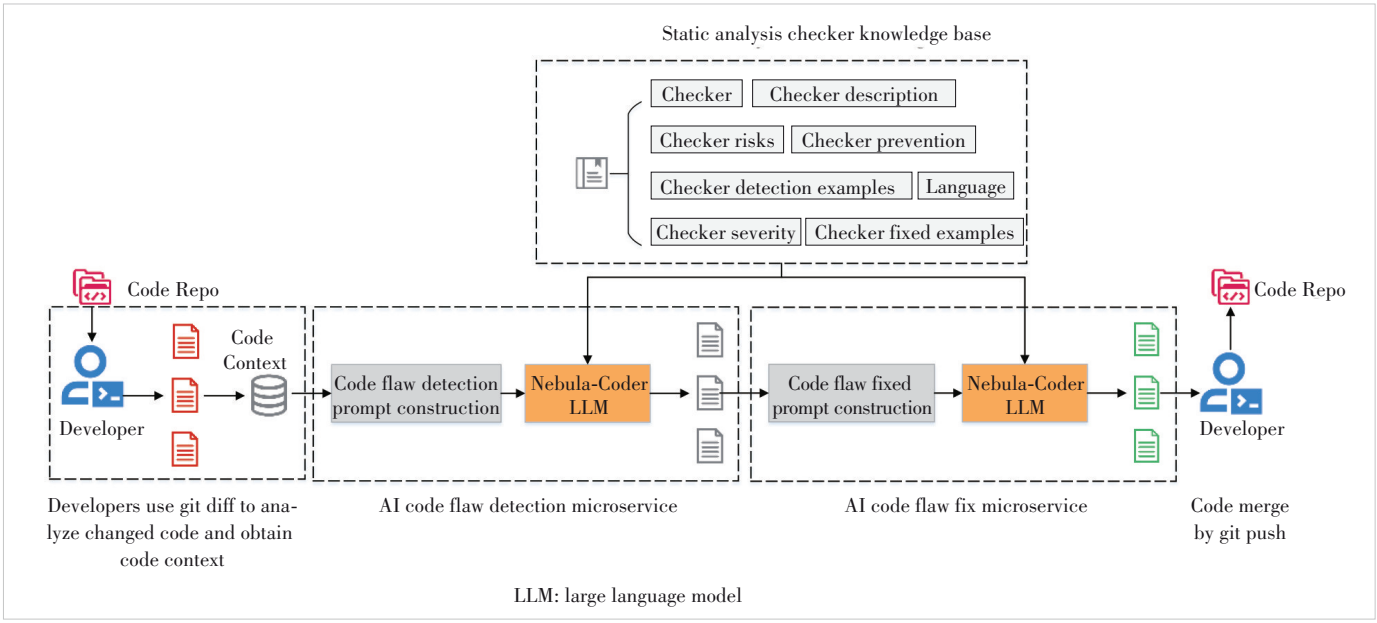


Figure 2. Overview of the AI-SCDF framework and its core workflow

construct prompts, particularly for generating examples of violations and fixes. This process enhances the learning capability of LLMs and reduces the likelihood of generating hallucinations.

2) Code change context analysis: Using git commands and program analysis methods, the context of the changed code is obtained, including all functional code segments in its control flow. This decreases the number of tokens in the LLM prompts, thereby reducing computational costs.

3) AI code flaw detection microservice: Prompts are automatically constructed based on the context of the changed code and the detailed checkers in the static analysis checker knowledge base. The LLM API is then invoked to detect flaws in the changed code. These prompts include code snippets of changed functions or methods, relevant code context, descriptions of static checking rules, and examples of violations. This microservice is also effective in identifying false alarms reported by commercial code checking tools, thereby improving developers' efficiency in addressing alerts.

4) AI code flaw fix microservice: Prompts are automatically constructed based on the context of the changed code, the identified defective code, and the detailed rules in the static analysis checker knowledge base. The LLM API is then invoked to fix defective code. These prompts include code snippets of the changed functions or methods, relevant code context, the flaw code, descriptions of static checking rules, and fix cases.

3 Design

This section describes the detailed design of AI-SCDF based on its architecture and workflow.

3.1 Static Analysis Checker Knowledge Base

The design of the static analysis checker knowledge base involves the following steps. First, CWE-IDs are extracted from the CWE Information Library. CWE is an open-source vulnerability classification system that provides a common language and numeric identifiers for software weaknesses. By parsing the CWE Information Library, we obtain CWE-IDs for various programming languages. Next, these CWE-IDs are fused with data crawled from network resources related to code flaws. These resources include checker descriptions from commercial tools, Common Vulnerabilities and Exposures (CVE) cases related to code, coding standards, the Secure Coding Practices published by the Open Web Application Security Project (OWASP) and publicly available code flaw datasets for machine learning (e.g., from GitHub). These documents may exist in different formats (e.g., HTML, PDF, Word), which are not convenient for data fusion and aggregation. Therefore, all data are converted into a unified format, such as JSON or XML, for subsequent storage and querying. The normalized data are then stored in the Knowledge Base Storage Service, which manages all static analysis checker knowledge, including checking rules, checker descriptions, risk levels, code flaw violations, and corresponding fix cases, as well as custom severity ratings for each checker across different programming languages. The entire process is illustrated in Fig. 3.

The format of a data record in the static analysis knowledge base is shown in Fig. 4. Data are stored using Markdown syntax, as this format is more user-friendly for LLMs when processing prompts that follow popular Markdown styles.

3.2 Code Change Context Analysis

The code change context analysis is designed to obtain the

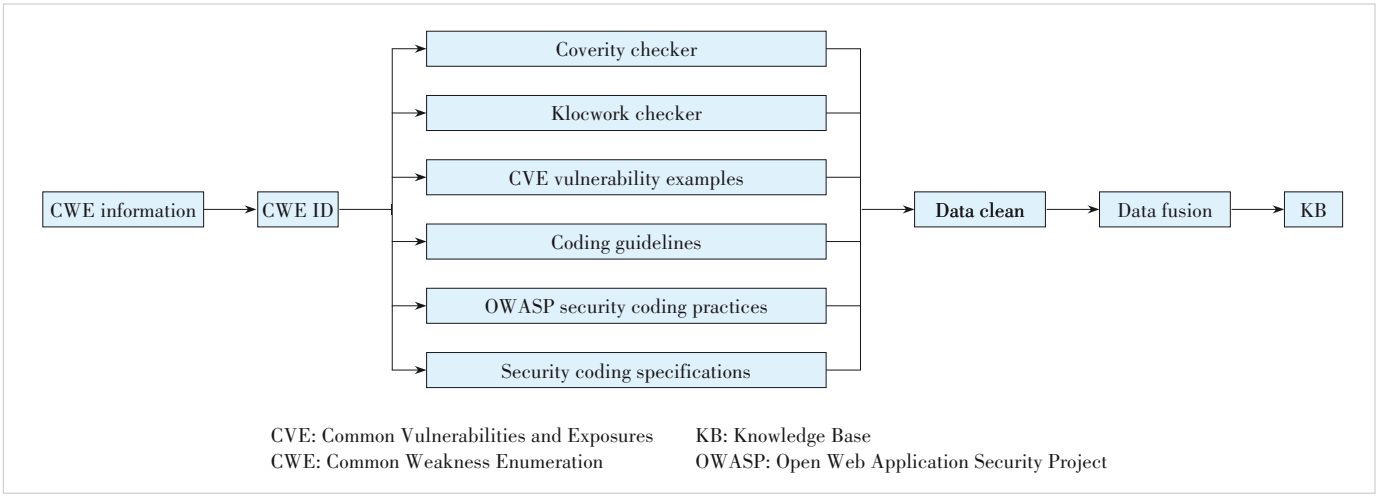


Figure 3. Flow chart of static analysis checker knowledge base

Knowledge Base Example

```
# Checker
   CWE89_SQL_Injection_SV.SQL
# Checker Description
   This Checker detects the situation when unvalidated or tainted data is used directly in an SQL query string.
# Language
   Java
# Risks
   SQL injections put data in the database at risk. Since unvalidated user input is being used in the SQL
   statement, an attacker can inject any SQL statement they wish to execute. This includes deleting, updating
   or creating data. It may also be possible to retrieve sensitive data from the database with this type of
   vulnerability. If the command is used for authentication, it will lead to unauthorized access.
# Prevention
   Prevent SQL injection flaws by validating any and all input from outside the application (user input, file input,
   system parameters, etc.). Validation should include length and content. Typically only alphanumeric
   characters are needed (i.e., A-Za-z, 0-9). Any other accepted characters should be escaped.
# Detect Examples
  ```java
   public ResultSet getUserData(ServletRequest req, Connection con) throws SQLException {
        String accountNumber = req.getParameter("accountNumber");
        String query = "SELECT * FROM user_data WHERE userid = '" + accountNumber + "'";//POTENTIAL FLAW
        Statement statement = con.createStatement(ResultSet.TYPE_SCROLL_INSENSITIVE,
 ResultSet.CONCUR_READ_ONLY);
        ResultSet results = statement.executeQuery(query);
        return results;
    }
  ```
Fixed Examples
   ```java
   public ResultSet getUserData(ServletRequest req, Connection con) throws SQLException {
     String accountNumber = req.getParameter("accountNumber");
     String query = "SELECT * FROM user_data WHERE userid = ?";
     PreparedStatement statement = con.prepareStatement(query);
     statement.setString(1, accountNumber);
     ResultSet results = statement.executeQuery();
     return results;
  }
   ```
```

Figure 4. Example data record illustrating the Markdown format of the static analysis checker knowledge base

contextual information of changed code for LLM learning. Its workflow is shown in Fig. 5. First, we obtain the information about code changes using git diff operations and parse it with a scripting language to extract the absolute file paths and the corresponding source code of the changed functions. Next, we trigger a call graph analysis tool on the local repository to analyze the call chains related to the changed code, and obtain the call chain related to the changed code based on the file paths and function names obtained in the previous step. As shown in this figure (fun_1->fun_8->fun->fun_9->fun_11->fun_12), each function in this call chain is then parsed to obtain the contextual code information relevant to the changed code.

3.3 AI Code Flaw Detection Microservice

The AI code flaw detection microservice constructs detection prompts based on the code context, the changed code functions, and the static analysis checker knowledge base. The format of the AI code flaw detection prompt template is shown in Fig. 6. These prompts include the checker name, checker description, risk ratings, examples of violations, code context, and changed code information. After the prompts are constructed, the LLM API is invoked for analysis. We employ the Nebula-Coder LLM, developed by ZTE using its proprietary AI technology. This model is a fine-tuned version of a foundation model to meet the needs of the code flaw detection service. The output of the LLM model determines whether the changed code violates the current checking rules. This microservice is designed to enhance code quality and proactively identify and rectify potential code flaws, thereby improving software development efficiency and quality.

3.4 AI Code Flaw Fix Microservice

The AI code flaw detection microservice produces a boolean output after modification: a value of 1 indicates that flaws exist in the changed code, while 0 signifies no flaws. If no flaws are detected (0), the fix process is skipped, and the entire AI-SCDF call process ends. If flaws are detected (1), the AI code flaw fix microservice is invoked to rectify them. The prompts for the code flaw fix service are also constructed using the same context and knowledge base as before. The fix prompt template (Fig. 7) includes the checker description, prevention guidance, fixed examples, buggy code, and code context. Finally, the fixed code with the related flaw is output for developers to confirm acceptance. Since LLMs may exhibit hallucination, it is necessary to have the fixed code reaffirmed by developers or quality assurance personnel. If accepted, the fixed code can be directly merged into the code repository.

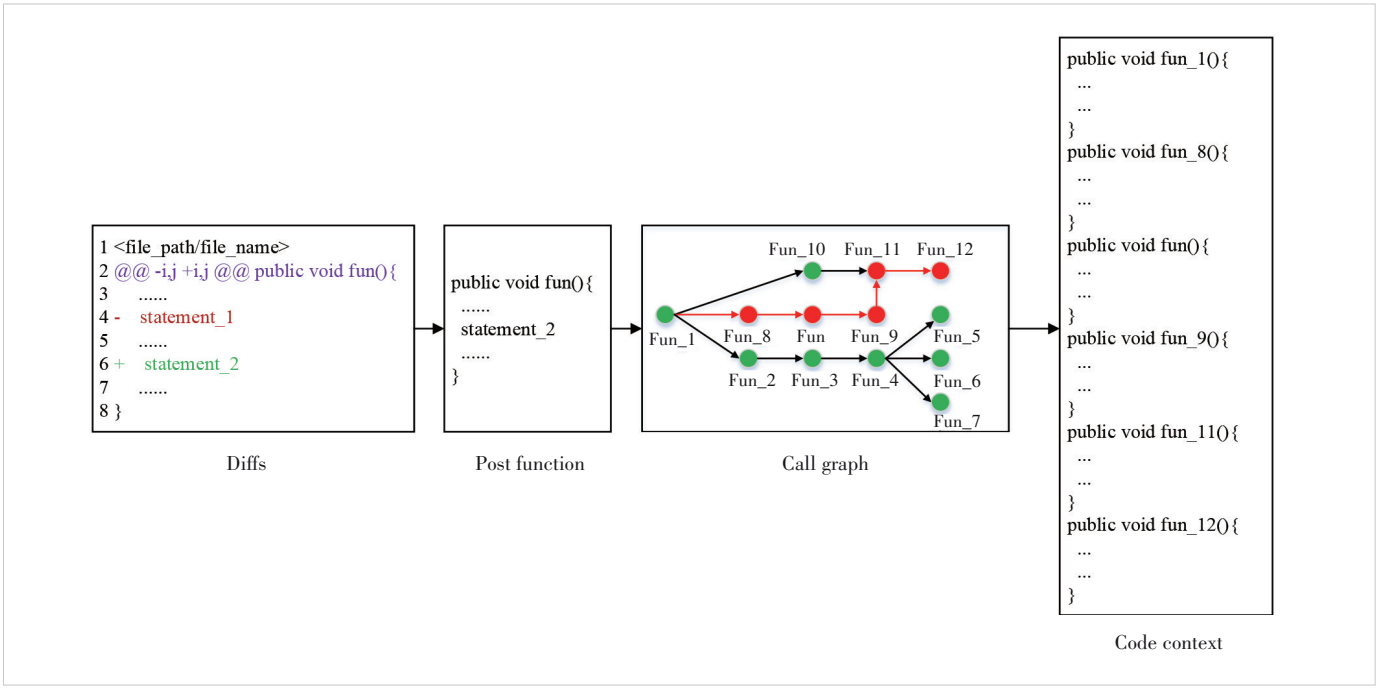


Figure 5. Workflow of code change context analysis

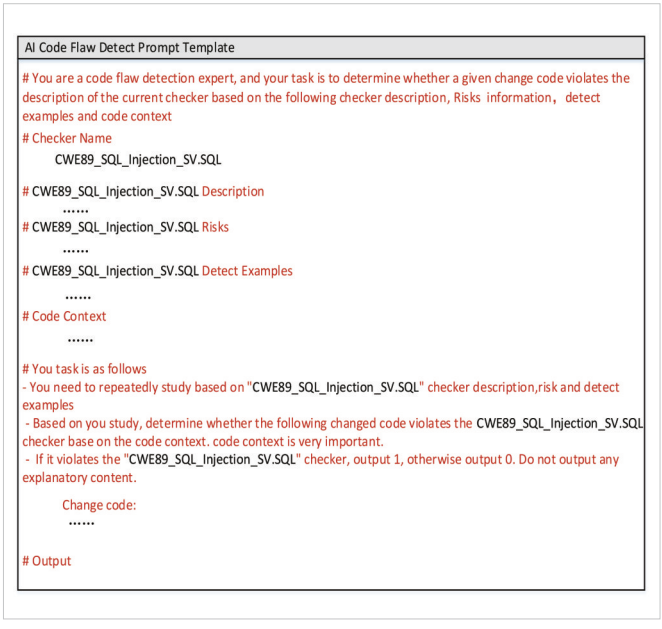


Figure 6. AI code flaw detection prompt template

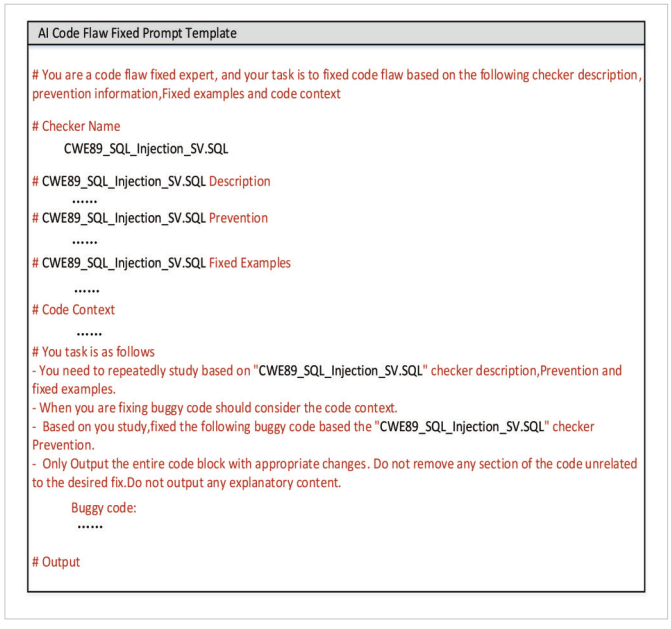


Figure 7. AI code flaw fixed prompt template

3.5 Nebula-Coder LLM

The Nebula-Coder LLM is dedicated to assisting developers in requirement analysis, product design, coding, testing, and deployment. It employs a vast amount of high-quality domain data, knowledge accumulation, and over 100 million technical documents and one trillion tokens of code corpus from ZTE's extensive experience in the telecommunications field. The model is pre-trained incrementally using parallel training frameworks. Since its launch in April 2023, the daily active users of Nebula-Coder have reached 120 000, with a code acceptance rate ranging from 40% to 45%, resulting in a 30% increase in coding efficiency and a 10% overall improvement in development efficiency. AI-SCDF utilizes the Nebula-Coder LLM for code defect detection and fix, enabling seamless integration with the public API provided by Nebula-Coder LLM.

4 Experimental Setup and Evaluation

The experimental dataset was the Juliet Java version 1.3 test suite, created by the NSA's Center for Assured Software (CAS). Juliet covers 112 different CWE code defect cases and is often used to evaluate the effectiveness of static code checking tools. Each test case includes its related code flaws and corresponding fixed examples. We employed the NebulaCoder LLM, which was fine-tuned from a foundation model to meet the needs of experimental verification.

First, we deployed the AI code flaw detection microservice to detect 10 representative CWE categories: CWE-23, CWE-78, CWE-89, CWE-315, CWE-325, CWE-190, CWE-398, CWE-382, CWE-760, and CWE-369. The model performance was evaluated using TP, FN, TN, FP, Accuracy, Precision and Recall as evaluation metrics. The experimental results are shown in Table 1. In this table, TP denotes cases where a defective code sample is correctly identified as flawed, TN represents correct identification of a non-defective code sample, FP indicates non-defective samples incorrectly flagged as flawed, and FN represents defective samples incorrectly classified as non-flawed. Accuracy is the proportion of correct samples predicted by the model, calculated as (TP+TN)/(TP+TN+FP+FN). Precision indicates the correctness of positive sample predictions, computed as TP/(TP + FP). Recall measures the coverage rate of positive sample predictions, calculated as TP/(TP + FN). The experimental results indicate that AI-SCDF achieves higher accuracy in detecting standard coding defect categories, such as CWE-315, CWE-325, CWE-398, CWE-382, and CWE-760, with accuracy rates of 98.4%, 91.2%, 90.5%, 100%, and 100%, respectively. This is because these standard code flaw categories require only attention to coding standards and simple handling, making them easier for LLMs to handle. However, for more complex code flaws like CWE-369, detailed contextual analysis of code snippets is required. This places greater demands on the LLM's learning and reasoning capabilities, resulting in lower accuracy. Further tuning of the LLM is needed to improve the accuracy of code defect detection.

Subsequently, we rectified the code flaw alerted by the AI code flaw detection microservice to verify its defect fix capability. The results are shown in Table 2. In this table, Accept means that the developer confirmed both the presence of a code flaw and its correct fixation, Reject indicates either no code flaw was found or the code flaw was not corrected properly, and Acceptance Rate represents the proportion of accepted fixes among all detected flaws. As shown in Table 2, the AI code flaw fix microservice performs well for standard code defect categories (CWE-315, CWE-325, CWE-398, CWE-382, CWE-760). However, for complex alerts like CWE-369, the LLM's hallucination or reasoning limitations adversely affect the fix capability, resulting in lower fix rates. Future work will focus on optimizing the LLM to improve the code flaw rectification capability and ensure code security.

Table 1. Experimental results of AI code flaw detection microservice on Juliet Java

CWE-ID	CWE Description	Testcase Path	Flaw Sum	Not Flaw Sum	TP	FN	TN	FP	Accuracy/%	Precision/%	Recall/%
CWE-23	Relative path traversal	Java/src/testcases/CWE23_Relative_Path_Traversal	444	276	378	66	202	74	80.6	83.6	85.1
CWE-78	OS command injection	Java/src/testcases/CWE78_OS_Command_Injection	444	276	363	81	171	105	74.2	77.6	81.8
CWE-89	SQL injection	Java/src/testcases/CWE89_SQL_Injection/s01	592	384	451	141	281	103	75	81.4	76.2
CWE-315	Plaintext storage in cookie	Java/src/testcases/CWE315_Plaintext_Storage_in_Cookie	37	24	37	0	23	1	98.4	97.4	100
CWE-325	Missing required cryptographic step	Java/src/testcases/CWE325_Missing_Required_Cryptographic_Step	34	0	31	3	0	0	91.2	100	91.2
CWE-190	Integer overflow	Java/src/testcases/CWE190_Integer_Overflow/s01	592	384	416	176	303	81	73.4	83.7	70.3
CWE-398	Poor code quality	Java/src/testcases/CWE398_Poor_Code_Quality	137	0	124	13	0	0	90.5	100	90.5
CWE-382	Use of system exit	Java/src/testcases/CWE382_Use_of_System_Exit	34	0	34	0	0	0	100	100	100
CWE-760	Predictable salt one way hash	Java/src/testcases/CWE760_Predictable_Salt_One_Way_Hash	17	0	17	0	0	0	100	100	100
CWE-369	Divide by zero	Java/src/testcases/CWE369_Divide_by_Zero/s01	592	384	434	158	276	108	72.3	80.1	73.3

CWE: Common Weakness Enumeration FN: false negative FP: false positive SQL: Structured Query Language TN: true negative TP: true positive

Table 2. Accuracy of the AI code flaw fix microservice for Juliet Java

CWE-ID	CWE Description	Testcase Path	Confirm Flaw	Accept	Reject	Accepted Ratio
CWE-23	Relative path traversal	Java/src/testcases/CWE23_Relative_Path_Traversal	452	271	181	60.0%
CWE-78	OS command injection	Java/src/testcases/CWE78_OS_Command_Injection	468	243	225	51.9%
CWE-89	SQL injection	Java/src/testcases/CWE89_SQL_Injection/s01	554	316	238	57.0%
CWE-315	Plaintext storage in cookie	Java/src/testcases/CWE315_Plaintext_Storage_in_Cookie	38	38	0	100%
CWE-325	Missing required cryptographic step	Java/src/testcases/CWE325_Missing_Required_Cryptographic_Step	31	31	0	100%
CWE-190	Integer overflow	Java/src/testcases/CWE190_Integer_Overflow/s01	497	239	258	48.1%
CWE-398	Poor code quality	Java/src/testcases/CWE398_Poor_Code_Quality	124	120	4	96.7%
CWE-382	Use of system exit	Java/src/testcases/CWE382_Use_of_System_Exit	34	34	0	100%
CWE-760	Predictable salt one way hash	Java/src/testcases/CWE760_Predictable_Salt_One_Way_Hash	17	17	0	100%
CWE-369	Divide by zero	Java/src/testcases/CWE369_Divide_by_Zero/s01	542	233	219	43.0%

CWE: Common Weakness Enumeration

5 Conclusions

This paper introduces the design and implementation of AI-SCDF, a tool designed to enhance software code quality. Through the construction of a static analysis checker knowledge base and the integration of code context, we develop effective LLM prompts to facilitate code flaw detection and fix. The experimental results indicate that our tool has effective flaw detection and rectification capabilities for standard code defect categories such as CWE-315, CWE-325, CWE-398, CWE-382, and CWE-760. However, our tool still faces challenges in detecting and rectifying more complex code defects. Furthermore, due to inherent limitations of LLM technology, its accuracy of code defect rectification still needs to be improved. Future work will focus on collecting high-quality corpora to further fine-tune the LLM model, aiming to better meet the needs of code defect detection and rectification.

References

[1] Coverity static analysis [EB/OL]. [2024-04-12]. https://www.synopsys.com/software-integrity/security-testing/static-analysis-sast.html
[2] Klocwork static analysis [EB/OL]. [2024-04-12]. https://help. klocwork. com/current/en-us/concepts/checkersintro.htm
[3] Infer static analyzer [EB/OL]. [2024-04-12]. https://fbinfer.com
[4] SonarQube [EB/OL]. [2024-04-12]. https://docs.sonarqube.org/latest
[5] CodeQL website [EB/OL]. [2024-04-12]. https://codeql.github.com
[6] Harer J A, Kim L Y, Russell R L, et al. Automated software vulnerability detection with machine learning [PP/OL]. ArXiv (2018-02-14)[2024-04-12]. https://arxiv.org/abs/1803.04497
[7] Li Z, Zou D Q, Xu S H, et al. SySeVR: a framework for using deep learning to detect software vulnerabilities [J]. IEEE transactions on dependable and secure computing, 2022, 19(4): 2244 – 2258. DOI: 10.1109/tdsc.2021.3051525
[8] Li Z, Zou D Q, Xu S H, et al. VulDeePecker: a deep learning-based system for vulnerability detection [C]//Proc. 2018 Network and Distributed System Security Symposium. Internet Society, 2018. DOI: 10.14722/ndss.2018.23158
[9] Wang S, Liu T Y, Tan L. Automatically learning semantic features for defect prediction [C]//Proc. 38th International Conference on Software Engineering. ACM, 2016: 297 – 308. DOI: 10.1145/2884781.2884804
[10] Li J, He P J, Zhu J M, et al. Software defect prediction via convolutional neural network [C]//Proc. IEEE International Conference on Software Quality, Reliability and Security (QRS). IEEE, 2017: 318 – 328. DOI: 10.1109/QRS.2017.42
[11] GPT-3 [EB/OL]. [2024-04-12]. https://openai.com/blog/gpt-3-apps
[12] PaLM2 [EB/OL]. [2024-04-12]. https://ai.google/discover/palm2
[13] Copilot [EB/OL]. [2024-04-12]. https://www. microsoft. com/en-us/microsoft-copilot
[14] Llama [EB/OL]. [2024-04-12]. https://llama.meta.com
[15] Berabi B, Gronskiy A, Raychev V, et al. DeepCode AI Fix: fixing security vulnerabilities with large language models [PP/OL]. ArXiv (2024-02-19)[2024-04-12]. https://arxiv.org/abs/2402.13291
[16] Wadhwa N, Pradhan J, Sonwane A, et al.Frustrated with code quality issues? LLMs can help! [PP/OL]. ArXiv (2023-09-22)[2024-04-12]. https://arxiv.org/abs/2309.12938
[17] Li H N, Hao Y, Zhai Y Z, et al. Enhancing static analysis for practical bug detection: an LLM-integrated approach [J]. Proc. ACM on programming languages, 2024, 8(OOPSLA1): 474 – 499. DOI: 10.1145/3649828
[18] Juliet Java [EB/OL]. [2024-04-12]. https://samate. nist. gov/SARD/test-suites/111

Biographies

Niu Zhi (niu.zhi@zte.com.cn) received his master's degree in control engineering from Chongqing University, China. He is currently working at ZTE Corporation. His research interests include distributed systems, formal verification, and software reliability.

Dong Luming received his master's degree in control theory and control engineering from Huazhong University of Science and Technology, China. He is currently working at ZTE Corporation. His research interests include distributed systems, formal verification, software reliability, and innovative security technologies for wireless communications.

AED-NeRF: Audio-Driven and Emotion-Editing Dynamic Neural Radiance Fields for Expressive Talking Face Avatar

Lu Ping[1,2], Song Li[3], Shi Wenzhe[1,2], Lin Zonghao[3], Ling Jun[3]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. Shanghai Jiao Tong University, Shanghai 200240, China)

DOI: 10.12142/ZTECOM.202601010

https://kns.cnki.net/kcms/detail/34.1294.TN.20260228.1556.002.html, published online February 28, 2025

Manuscript received: 2024-09-06

Abstract: While neural radiance field (NeRF) methods have shown promising results in generating talking faces, existing studies primarily focus on the correlation between avatars and driving sources. However, these studies often overlook emotion modeling, resulting in the generation of emotionless or unnatural facial animations. In response, this paper introduces an audio-driven and emotion-editing dynamic NeRF (AED-NeRF) approach, designed for the real-time generation of expressive talking face avatars driven by audio inputs. Specifically, we integrate audio features into a grid-based NeRF to compensate for the lack of a deformation channel, successfully capturing lip dynamics and enabling end-to-end generation from audio-driven sources to talking face avatars. Emotion labels, comprising emotion categories and intensity levels, guide the proposed NeRF framework to implicitly model visual emotions, allowing for explicit control and editing of facial expressions. Extensive qualitative and quantitative experiments validate the effectiveness and advantages of our proposed method, demonstrating its ability to achieve real-time, photo-realistic talking face avatar generation across different audio and emotion scenarios.

Keywords: talking face avatar; neural radiance fields; AED-NeRF

Citation (Format 1): Lu P, Song L, Shi W Z, et al. AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields for expressive talking face avatar [J]. ZTE Communications, 2026, 24(1): 72 - 80. DOI: 10.12142/ZTECOM.202601010
Citation (Format 2): P. Lu, L. Song, W. Z. Shi, et al., "AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields for expressive talking face avatar," *ZTE Communications*, vol. 24, no. 1, pp. 72 - 80, Mar. 2026. doi: 10.12142/ZTECOM.202601010.

1 Introduction

With the rapid evolution of deep learning[1] and generative modeling[2], talking face avatars are undergoing unprecedented development[3 - 8] and have been progressively integrated into our visual experiences, such as virtual video conferences, film redubbing, and digital human representation. Despite these advancements, talking face generation encounters numerous challenges in practical applications.

Image-based methods generate talking face avatars by employing techniques including image-to-image translation[9 - 11] and generative adversarial networks (GANs) [12 - 14]. Nevertheless, the absence of 3D perception tends to result in flat and unrealistic visual results. Model-based approaches explicitly construct 3D talking faces based on intermediate representations such as facial landmarks[15], coefficients[16] and vertices[4]. While leveraging 3D face modeling produces higher-quality results, cumulative errors and information loss during the intermediate representation prediction can lead to semantic mismatches between lip movements and audio cues.

Recently, the emergence of neural radiance fields (NeRF)[17] has provided a novel framework for talking face generation. NeRF-based methods can render realistic talking face avatars at high resolutions from novel views with reduced training data[5, 18 - 19]. However, the inference speed of the vanilla NeRF is insufficient to meet the real-time requirements of audio-driven talking face avatars in practical applications. Moreover, existing works fail to fully implement emotional modeling for talking face avatars, resulting in emotionless or unnatural human faces[20 - 22].

In this paper, we propose an audio-driven and emotion-editing dynamic NeRF (AED-NeRF) for real-time and expressive talking face avatar generation. Our method consists of three processing modules and two NeRF models. In the processing pipeline, we introduce an audio processing module, an emotion encoder, and a pose estimation module, where the audio encoder extracts features from audio sequences, and the emotion encoder encodes explicit emotion labels based on

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20230921015.

their category and intensity, obtaining emotion features. We employ an off-the-shelf method[23 - 24] to estimate 3D head pose for additional spatial control. Besides, we employ two NeRF models to render the head and torso separately by taking target identity video sequences, synchronized audio sequences, and emotion labels as inputs. In the modeling parts, AED-NeRF utilizes spatial features, audio features, and emotional features as inputs to neural radiance fields, implicitly modeling identity as volume density and RGB color. During the inference stage, given arbitrary head pose sequences, driving audio, and emotion labels, AED-NeRF performs volume rendering[25] according to the predicted volume density and color learned in the training stage, generating an expressive 3D avatar matching the driving source in real time.

Our contributions are summarized as follows:

• We introduce audio and emotional features to compensate for the lack of a deformation channel in the grid NeRF, implicitly modeling head dynamics and enabling end-to-end talking face avatar generation.

• We consider implicit emotion modeling in talking face avatars with emotion labels for diverse emotional expressions, guiding NeRF to implicitly model facial expressions and explicitly control the emotional editing of talking face avatars during the inference stage.

• Extensive experiments demonstrate that AED-NeRF can generate photorealistic, expressive talking face avatars in real time under different audio and emotional conditions.

2 Related Work

1) Image-based talking face generation. Image-based methods generate 2D talking face avatars using image-to-image translation or GANs. Zhou et al.[26] proposed a disentangled audio-visual system to disentangle identity and audio content through adversarial learning, improving lip synchronization for 2D talking face avatars. Das et al.[27] employed cascaded GANs to separately learn general lip motion and identity-specific texture. Zhou et al.[11] animated a single image with an audio clip by predicting landmark displacement from disentangled audio content and identity. Though these methods work well for stylized facial images, they have difficulty in generating realistic human face avatars due to the lack of 3D perception information.

2) Model-based talking face generation. Model-based methods explicitly generate 3D talking faces based on intermediate facial representations such as landmarks, coefficients and vertices. Kumar et al.[6] utilized long short-term memory (LSTM) to learn the mapping from driving audio sources to lip landmarks, and then generated pixel-to-pixel Obama avatars via UNet. Thies et al.[16] proposed a general audio-to-expression network to predict the expression coefficients of a 3D face model based on audio features, and a UNet-based neural rendering network to render talking face avatars from expression coefficients, thus enabling cross-identity audio driving. Richard et al.[28] designed a categorical latent space using 3D face vertices as the intermediate representation based on cross-modality loss. This space disentangles audio-correlated and audio-uncorrelated information and thus results in reasonable movements in audio-uncorrelated facial regions. Though model-based methods can generate high-quality talking face avatars, they suffer from problems including complex pipelines, expensive data labels, and information loss of driving source.

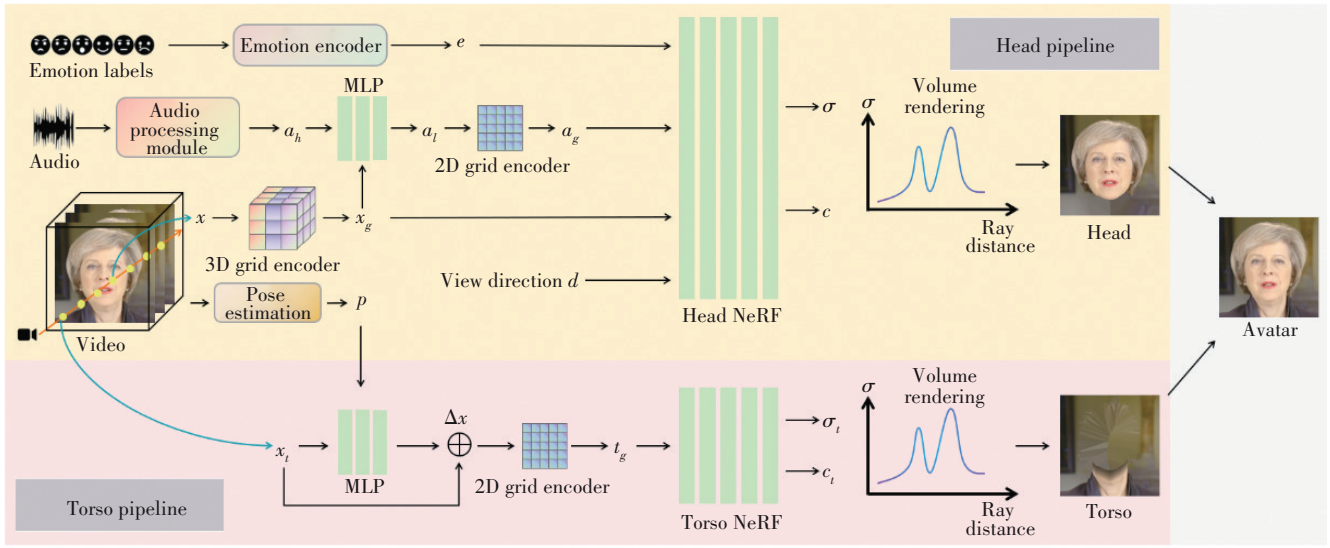
3) NeRF-based talking face generation. Neural radiance fields[17] have achieved great success in natural rendering and provide a new implementation for end-to-end talking face avatar generation. Gafni et al.[29] introduced NeRF to the field of talking face generation and proposed the first dynamic face radiance fields. They trained multilayer perceptron (MLP) conditioned on latent codes based on face coefficients and camera poses reconstructed by a face tracker, realizing face reconstruction and pose control of talking face avatar. Guo et al.[5] proposed audio-driven neural radiance fields which take DeepSpeech[30] audio features as conditional input to MLP and individually model head part and torso part of talking face avatars. Hong et al.[31] designed a parametrized general model for representing faces under different views, expressions and lighting, and significantly improved the rendering speed of NeRF via integrating a 2D neural rendering strategy into it. Shen et al.[18] conditioned NeRF on 2D face images to learn the face prior and fine-tune the face radiance fields with few identity images to generalize to a new identity rapidly. However, aforementioned NeRF-based methods struggle to meet real-time requirements in practical application due to the slow rendering speed.

4) Emotional talking face generation. Though talking face avatar generation has made significant progress in recent years, most works focus on the correlation between digital avatars and driving sources (e.g., audio, text, etc.) while neglecting emotion modeling, which leads to emotionless or unnatural human faces. Therefore, some researchers have turned to the study of emotion modeling for expressive talking face avatars. Eskimez et al.[10] improved emotional expressions by encoding emotion categories and designing an emotion discriminator to supervise the training. Ji et al.[32] proposed an implicit emotion displacement learner to modify facial dynamics for realistic emotion patterns. Tan et al.[33] extracted emotion embeddings from audio as queries and utilized a memory network to retrieve the best-matching expressions for talking face avatars. However, NeRF-based emotion modeling has not been fully explored. We show a simple but effective method to integrate emotion into NeRF modeling in Section 3.

3 Methods

3.1 Overview

In this section, we present our AED-NeRF framework in Fig. 1. The inputs to the network include target identity video sequences, synchronized audio sequences, and emotion la-



MLP: multilayer perceptron NeRF: neural radiance fields

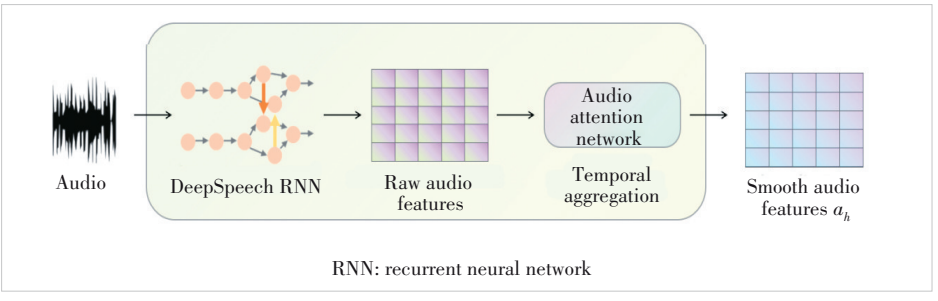
Figure 1. An overview of AED-NeRF framework

bels. We utilize an off-the-shelf method[23 - 24] to estimate 3D human poses for further spatial features. We adopt an audio encoder to extract audio features from the audio sequences and a one-hot encoder to encode explicit emotion labels according to their category and intensity to obtain emotion features. Then, AED-NeRF takes spatial, audio, and emotion features as the inputs to the neural radiance fields and implicitly models the identity, which is represented by volume density and RGB colors. In the inference stage, given an arbitrary reference of head pose sequences, driving audio and emotion labels, AED-NeRF performs volume rendering based on the volume density and color predicted in the training stage, and generates an expressive 3D digital human that matches the driving source in real time.

3.2 Audio Processing Module

The vanilla NeRF[17] is only suitable for static scene modeling. To apply it to dynamic talking faces, we introduce audio features to compensate for the deformation channel of NeRF to model dynamic lip motions. Our audio processing module is illustrated in Fig. 2. The module first extracts the corresponding DeepSpeech[30] audio features from the input audio for each frame using a pre-trained recurrent neural network (RNN) model. Then, an audio attention network aggregates DeepSpeech audio features of neighboring frames in a self-attention manner to obtain smooth high-dimensional audio features a_h.

Previous NeRF-based methods[5, 18] directly concatenate high-dimensional audio features with spatial features and feed them into NeRF. However, this leads to high-dimensional inputs for the MLP, significantly increasing computational cost and resulting in slow training and rendering. To meet the real-time demand of digital human applications, we adopt the grid NeRF[34] to replace a portion of MLP forward propagation to query the spatial and audio features with linear interpolation. It compresses the size of MLP and accelerates the rendering speed effectively. Specifically, for any point x in the dynamic scene, it is firstly encoded as spatial grid features x_g by a 3D spatial grid encoder E^3_{spatial}. Then, high-dimensional audio features a_h are fused with the spatial grid features x_g and compressed to 2D audio features a_l by an MLP. This explicitly conditions audio features on the spatial position to ensure that the effect of audio sequences is constrained to the facial region only, rather than the torso or background. Finally, 2D audio features a_l are encoded as audio grid features a_g by a 2D audio grid encoder E^2_{audio} and then fed into NeRF.



RNN: recurrent neural network

Figure 2. Audio processing module

3.3 Emotion Modeling and Editing

Diverse emotional expressions effectively represent the emotional state of a digital human and contribute to a more realistic and vivid talking face avatars. We propose an explicit method for emotion control and editing based on emotion labels. Specifically, we set five basic emotion categories (i. e., neutral, angry, fearful, happy, and sad) and three levels of emotion intensity (i.e., weak, medium, and strong), and the combination of a specific emotion category and intensity is regarded as an emotion label. As illustrated in Fig. 3, given an emotion label as input, the emotion category and intensity are encoded into category features e_c and intensity features e_i by one-hot encoders, respectively. We concatenate category features e_c and intensity features e_i as the emotion feature $e = (e_c, e_i)$, which is fed into NeRF as the guidance of expression modeling. In the inference stage, facial expressions can be explicitly controlled and edited by combining different category features e_c and intensity features e_i. Experiments demonstrate that emotion labels and the simple but effective emotion encoder are enough for NeRF to model the dynamics of facial expressions through MSE loss and generate expressive talking face avatars.

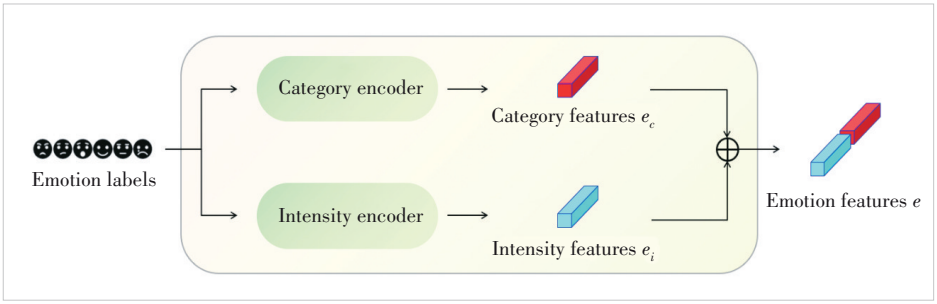


Figure 3. Illustration of emotion encoder

To achieve this, our head NeRF $\mathcal{F}_\Theta^{\text{head}}$ takes spatial grid features x_g , view direction d, audio grid features a_g and emotion features e as inputs to predict density σ and RGB color c of samples in camera rays. This can be formulated as:

$$\mathcal{F}_\Theta^{\text{head}}: (x_g, d, a_g, e) \rightarrow (\sigma, c) \tag{1}$$

3.4 Torso Modeling

Compared with the head part, torso movements are relatively slight and weakly correlated with our driving source. Thus, we follow SSP-NeRF[35] and design a deformation-based NeRF for torso modeling. Specifically, given a point x_t in the 2D image space, we condition it on the head pose p to predict the deformation of torso movements Δx via an MLP. This ensures that torso movements are synchronized with the head to avoid mismatched results caused by independent modeling of the head and torso. Then, the deformation Δx is added to the initial position x_t and fed to the 2D torso grid encoder E^2_{torso} to obtain torso grid features t_g. Finally, we feed grid features t_g into our torso NeRF to predict the density σ_t and RGB color c_t of the torso part. This can be formulated as:

$$\mathcal{F}_\Theta^{\text{torso}}: (t_g) \rightarrow (\sigma_t, c_t) \tag{2}$$

3.5 Implementation Details

1) Volume rendering. Given the density σ and RGB color c, we follow the rendering process of vanilla NeRF[17]. Specifically, we accumulate the density and RGB color of samples along the camera rays cast through each pixel to compute the output color under the specific view direction for talking face avatars. Given the camera center o and view direction d, the camera ray is represented as $r(t) = o + td$. Let near and far bounds be t_n and t_f , and the expected output color $\mathcal{C}$ is:

$$\mathcal{C}(r; \Theta, a_g, e) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), d)\,\mathrm{d}t \tag{3}$$

where $T(t)$ denotes the accumulated transmittance along the ray from t_n to t:

$$T(t) = \exp\left(-\int_{t_n}^{t}\sigma(r(s))\,\mathrm{d}s\right) \tag{4}$$

2) Loss function. We utilize the mean squared error (MSE) loss to minimize pixel-level reconstruction error:

$$L_{\text{MSE}} = \left\| \mathcal{C} - \mathcal{C}_{gt} \right\|_2^2 \tag{5}$$

where $\mathcal{C}$ is the rendered color and $\mathcal{C}_{gt}$ is the ground truth. As lip synchronization is crucial for talking face avatars, the pixel-level loss struggles to learn the complex semantic mapping from audio features to lip movements. Therefore, an additional learned perceptual image patch similarity (LPIPS)[36] loss is introduced for fine-tuning in the lip region:

$$L_{\text{LPIPS}} = \text{LPIPS}(\mathcal{P}, \mathcal{P}_{gt}) \tag{6}$$

where $\mathcal{P}$ is the rendered lip patch and $\mathcal{P}_{gt}$ is the ground truth.

Besides, for more accurate rendered results, we introduce the entropy regularization loss to encourage transmittance to be closer to 0 or 1:

$$L_\alpha = -\sum(\log\alpha + (1-\alpha)\log(1-\alpha)) \tag{7}$$

where α is the transparency of each rendered pixel.

We assume that the driving source only affects the facial region rather than the torso or background. Therefore, we adopt an L_1 regularization loss:

$$L_{\text{aud}} = \sum_{a_l \in \bar{R}_{\text{face}}} |a_l| \quad (8),$$

where $\bar{R}_{\text{face}}$ denotes the non-facial region. This encourages a_l to be 0 and avoids artifacts in the non-facial region.

Finally, the overall loss function can be formulated as:

$$L = L_{\text{MSE}} + \lambda_{\text{LPIPS}} L_{\text{LPIPS}} + \lambda_{\alpha} L_{\alpha} + \lambda_{\text{aud}} L_{\text{aud}} \quad (9).$$

3) Training details. We set the window size to 16 for the DeepSpeech RNN and 8 for the audio attention network. Our NeRF is composed of a 5-layer MLP with 64 hidden dimensions. For a specific identity, we first train the head NeRF for 20 000 epochs and fine-tune lip regions for 5 000 epochs. Then, torso NeRF is trained for 20 000 epochs. At each epoch, we randomly sample 256×256 camera rays and 16 samples for each ray and utilize the Adam optimizer with a learning rate of 0.000 5 to optimize the loss function. The loss coefficients are set to 0.01 for λ_{LPIPS}, 0.001 for λ_{α}, and 0.1 for λ_{aud}. For a 4-minute 25-fps video with resolution 512×512, the training time for the head NeRF and torso NeRF in a single RTX4090 is 5 h and 2 h, respectively.

4 Experiments

4.1 Experimental Settings

1) Datasets. For audio driving, we follow previous studies[5, 18] to collect several public speech videos of different celebrities to construct the celebrity dataset. The average video length is about 5 min, and both the recording camera and background are kept static. For emotion editing, we select MEAD[37] as the experimental data. MEAD is a large-scale audio-visual dataset, including abundant 3 – 5 s video clips of 60 actors speaking with 8 different emotions at 3 different intensity levels. We collect front-view video clips of 5 basic emotion categories (neutral, angry, fearful, happy, and sad) and 3 levels of emotion intensity (weak, medium, and strong) from MEAD for emotion editing experiments.

2) Data preprocessing. We first employ a face detection algorithm to locate the facial regions in all videos. Based on these facial regions, we crop the videos to a resolution of 512×512 with the faces in the center and resample them to 25 fps. Since a single video clip from MEAD is not enough for training, we combine the video clips conditioned on the same emotion category and intensity level from the same identity into a 90 – 120 s video as the data for the corresponding emotion label. Finally, we utilize a face parsing algorithm to annotate the head, torso, and background regions, and extract each part individually for each frame.

3) Metrics. We adopt peak signal-to-noise ratio (PSNR) and LPIPS[36] as image quality metrics. Note that PSNR only takes pixel-level differences into account and cannot faithfully reflect human perception of image quality. In comparison, LPIPS captures semantic information and structural similarity of images and is more consistent with subjective perception. For audio-visual synchronization evaluation, we adopt landmark distance (LMD)[38], SyncNet confidence (Sync-C), and SyncNet distance (Sync-D)[39] as metrics. LMD measures the distance between lip landmarks and the ground truth. Sync-C and Sync-D measure alignment and misalignment between audio and video streams via SyncNet, respectively.

4.2 Quantitative Comparisons

We first evaluate the audio-driving performance of our AED-NeRF under self-driven and cross-driven settings and compare it with non-NeRF-based methods, e. g., Wav2Lip[3] and live speech portraits (LSP)[40], and NeRF-based methods, e.g., audio driven NeRF (AD-NeRF)[5] and Dynamic Facial Radiance Fields (DFRF)[18] baselines. The self-driven results are shown in Table 1. PSNR, LPIPS, and LMD for LSP are not reported since LSP cannot generate the same poses as the ground truth. Our method performs best in most metrics with

Table 1. Quantitative comparison under the self-driven setting

Methods	Image Quality		Audio-Visual Synchronization			Rendering Speed	
	PSNR ↑	LPIPS ↓	LMD ↓	Sync-C ↑	Sync-D ↓	Training time/h ↓	Inference speed/fps ↑
GT	∞	0	0	8.897	6.325	/	/
Wav2Lip	**30.90**	0.139	3.311	**7.898**	**6.694**	/	15
LSP	/	/	/	5.181	8.637	/	25
AD-NeRF	28.79	0.101	3.245	3.944	10.603	36	0.09
DFRF	28.85	0.118	3.815	4.184	10.396	72	0.06
AED-NeRF	28.81	**0.088**	**2.826**	6.786	8.252	**7**	**45**

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields
DFRF: dynamic facial radiance fields
LMD: landmark distance
LPIPS: learned perceptual image patch similarity
LSP: live speech portraits
PSNR: peak signal-to-noise ratio

real-time rendering speed. Specifically, we consider LPIPS as a more informative image quality metric, and AED-NeRF generates higher-quality talking face avatars compared with both existing non-NeRF-based and NeRF-based methods. In terms of audio-visual synchronization, AED-NeRF achieves optimal or sub-optimal scores in all metrics. Since Wav2Lip directly uses SyncNet as a loss term for supervision during training, its SyncNet scores are much better than other methods, even surpassing the ground truth. Our AED-NeRF achieves satisfactory SyncNet scores while significantly outperforming other methods in LMD, indicating that our method can generate synchronized lip movements with the audio source. In addition, AED-NeRF saves 80% – 90% training time and infers 500 – 750 times faster than NeRF-based baselines, enabling real-time applications. The cross-driven results in Table 2 demonstrate that the audio-visual synchronization performance of our AED-NeRF is second only to Wav2Lip but superior to other baselines, indicating that our method can still generate reasonable lip movements under the cross-driven setting.

Table 2. Quantitative comparison under the cross-driven setting

Methods	ID A		ID B	
	Sync-C ↑	Sync-D ↓	Sync-C ↑	Sync-D ↓
Wav2Lip	**8.748**	**7.623**	**8.208**	**7.193**
LSP	3.979	9.656	5.097	8.477
AD-NeRF	3.259	10.123	3.037	10.526
DFRF	4.607	9.235	4.245	10.083
AED-NeRF	6.624	8.799	6.074	8.075

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields
DFRF: dynamic facial radiance fields
LSP: live speech portraits

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields
DFRF: dynamic facial radiance fields

Figure 4. Qualitative comparison under the self-driven setting

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields
DFRF: dynamic facial radiance fields
LSP: live speech portraits

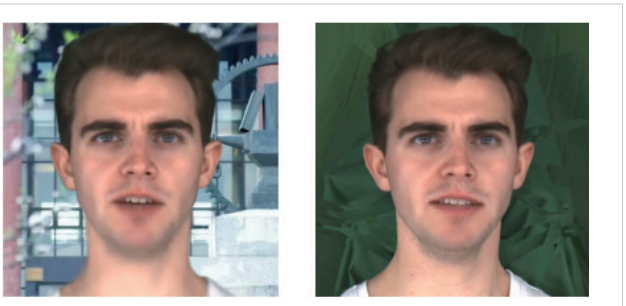
Figure 5. Qualitative comparison under the cross-driven setting

4.3 Qualitative Comparisons

Quantitative metrics have limitations in visual quality assessment and sometimes exhibit inconsistencies with subjective human perception. Therefore, we further conduct a qualitative evaluation of audio driving and emotion editing.

The self-driven results are illustrated in Fig. 4. In terms of image quality, Wav2Lip exhibits skin color distortion and obvious artifacts in the lip region; AD-NeRF loses some high-frequency information of images and suffers from head-torso separation when moving heavily; in contrast, our AED-NeRF faithfully reconstructs the talking face avatar of the reference identity. As for audio-visual synchronization, Wav2Lip, AD-NeRF, and DFRF deviate significantly from the ground truth, while our AED-NeRF synthesizes reasonable and accurate lip movements. The cross-driven results are illustrated in Fig. 5. By analyzing the lip synchronization, our AED-NeRF is capable of robustly synthesizing audio-visual synchronized lip movements even in challenging situations such as the pronunciation of the vowel /o/.

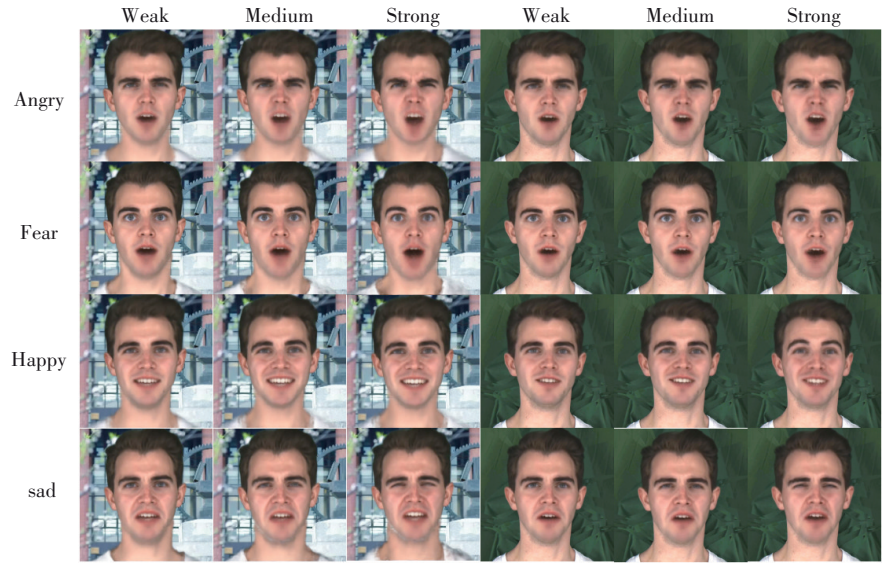
Since none of the baselines can generate corresponding expressions from emotion labels, we train AD-NeRF on each emotion-labeled video individually for comparison with our AED-NeRF. The neutral reference and generated emotion comparison are illustrated in Figs. 6 and 7, respectively. We replace the background with natural scenery in AD-NeRF and keep the green screen in our AED-NeRF for easier visual com-



(a) AD-NeRF (b) AED-NeRF

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields

Figure 6. Neutral reference generated by (a) AD-NeRF and (b) AED-NeRF



(a) AD-NeRF (b) AED-NeRF

AD-NeRF: audio driven neural radiance fields
AED-NeRF: audio-driven and emotion-editing dynamic neural radiance fields

Figure 7. Qualitative comparison of emotion editing

Figure 8. Benefiting from NeRF architecture and background disentanglement, our AED-NeRF can synthesize talking face avatars in novel views and support background editing

parison. Note that our AED-NeRF can generate different emotional expressions by editing emotion labels within a single model. By comparing the facial expressions in the same frames, it can be found that our AED-NeRF warps facial regions (e. g., forehead, eyes, and mouth) corresponding to the emotion labels, and presents the desired expression. The warping becomes more pronounced as the intensity level increases, which reflects the differences among the three intensity levels.

Besides, AED-NeRF can synthesize talking face avatars in novel views and support background editing, which benefits from NeRF architecture and background disentanglement during data preprocessing.

4.4 Ablation Study

We conduct an ablation study to verify the effect of emotion modeling under the self-driven setting in Fig. 9. Without the guidance of emotion labels, the network has to model the visual emotion based only on audio conditions, which leads to a neutral or mismatched face even driven by extremely emotional audio. Our AED-NeRF generates expressive talking face avatars with matched facial expressions and more precise lip movements benefiting from emotion modeling.

5 Limitations

We have demonstrated that our AED-NeRF can generate realistic audio-driven expressive talking face avatars in real time. However, several limitations remain for future work. Lips may be jittering or unsynchronized with audio under the cross-driven setting, as an English automatic speech recognition model is employed to extract audio features which is not accurate for other languages. Therefore, how to extract general audio features across different languages is significant for further audio-visual synchronization improvement. Though our AED-NeRF supports emotion editing to generate expressive talking face avatars, the range of editing is limited to our provided emotion labels; that is, it cannot generalize beyond training data. A possible solution is to design an emotion recognition module to automatically classify emotion categories and intensity lev-

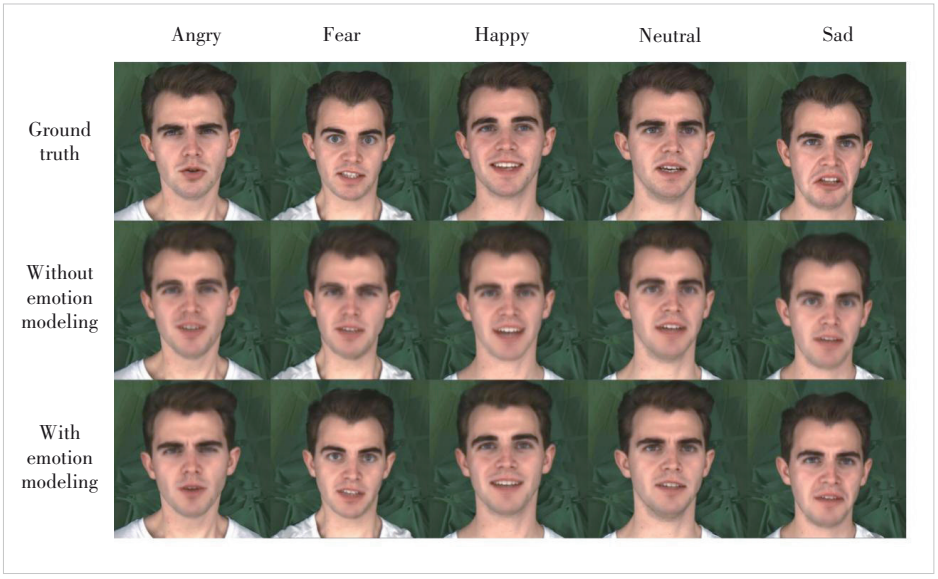


Figure 9. Ablation study on emotion modeling

els from videos, and learn a latent space for emotion embedding. Besides, slight variations, such as camera parameters, lighting and clothing, can affect modeling of one identity and lead to unideal generated results due to the nature of vanilla NeRF. For improved robustness to these variations, we will refer to works like NeRF in the wild[41] to disentangle environmental variations from facial dynamics in the future.

6 Conclusions

We propose AED-NeRF for the real-time generation of audio-driven, expressive talking face avatars. Audio and emotion features are introduced as the deformation channel for NeRF to implicitly model facial dynamics. Emotion labels composed of categories and intensity levels are encoded as guidance for emotion modeling of talking face avatars. Extensive experiments demonstrate that our AED-NeRF can generate photo-realistic expressive talking face avatars under different audio inputs and emotion settings in real time.

Ethical consideration: AED-NeRF can generate photorealistic expressive talking face avatars in real time under different audio and emotional conditions. However, talking face synthesis techniques could be misused. We restrict our AED-NeRF for research purposes only and support the development of deepfake detection methods.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521 (7553): 436 – 444. DOI: 10.1038/nature14539
[2] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: an overview [J]. IEEE signal processing magazine, 2018, 35(1): 53 – 65. DOI: 10.1109/MSP.2017.2765202
[3] Prajwal K R, Mukhopadhyay R, Namboodiri V P, et al. A lip sync expert is all you need for speech to lip generation in the wild [C]//The 28th ACM International Conference on Multimedia. ACM, 2020: 484 – 492. DOI: 10.1145/3394171.3413532
[4] Thies J, Elgharib M, Tewari A, et al. Neural voice puppetry: Audio-driven facial reenactment [PP/OL]. arxiv (2020-07-29) [2024-09-06]. https://arxiv.org/abs/1912.05566
[5] Guo Y D, Chen K Y, Liang S, et al. AD-NeRF: audio driven neural radiance fields for talking head synthesis [C]//International Conference on Computer Vision. IEEE, 2021: 5764 – 5774. DOI: 10.1109/ICCV48922.2021.00573
[6] Kumar R, Sotelo J, Kumar K, et al. Obamanet: photo-realistic lip-sync from text [PP/OL]. arxiv (2017-12-06) [2024-09-06]. https://arxiv.org/abs/1801.01442
[7] Wang J D, Qian X Y, Zhang M L, et al. Seeing what you said: talking face generation guided by a lip reading expert [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 14653 – 14662. DOI: 10.1109/CVPR52729.2023.01408
[8] Zhong W Z, Fang C W, Cai Y Q, et al. Identity-preserving talking face generation with landmark and appearance priors [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 9729 – 9738. DOI: 10.1109/CVPR52729.2023.00938
[9] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 5967 – 5976. DOI: 10.1109/CVPR.2017.632
[10] Eskimez S E, Zhang Y, Duan Z Y. Speech driven talking face generation from a single image and an emotion condition [J]. IEEE transactions on multimedia, 2022, 24: 3480 – 3490. DOI: 10.1109/TMM.2021.3099900
[11] Zhou Y, Han X T, Shechtman E, et al. MakeItTalk: speaker-aware talking-head animation [J]. ACM transactions on graphics, 2020, 39(6): 1 – 15. DOI: 10.1145/3414685.3417774
[12] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C]//The 28th International Conference on Neural Information Processing Systems. ACM, 2014: 2672 – 2680. DOI: 10.5555/2969033.2969125
[13] Yin F, Zhang Y, Cun X D, et al. StyleHEAT: one-shot high-resolution editable talking face generation via Pretrained StyleGAN [C]//European Conference on Computer Vision (ECCV). ECVA, 2022: 85 – 101. DOI: 10.1007/978-3-031-19790-1_6
[14] Doukas M C, Zafeiriou S, Sharmanska V. Headgan: video-and-audio-driven talking head synthesis: Vol. 1 [PP/OL]. arxiv (2021-08-23) [2024-09-06]. https://arxiv.org/abs/2012.08261
[15] Chen L L, Maddox R K, Duan Z Y, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 7824 – 7833. DOI: 10.1109/CVPR.2019.00802
[16] Thies J, Elgharib M, Tewari A, et al. Neural voice puppetry: audio-driven facial reenactment [C]//European conference on computer vision (ECCV). ECVA, 2020: 716 – 731. DOI: 10.1007/978-3-030-58517-4_42
[17] Mildenhall B, Srinivasan P P, Tancik M, et al. NeRF: representing scenes as neural radiance fields for view synthesis [J]. Communications of the ACM, 2021, 65(1): 99 – 106. DOI: 10.1145/3503250
[18] Shen S, Li W H, Zhu Z, et al. Learning dynamic facial radiance fields for Few-shot talking head synthesis [C]//European Conference on Computer Vision (ECCV). ECVA, 2022: 666 – 682. DOI: 10.1007/978-3-031-19775-8_39
[19] Yao S Y, Zhong R Z, Yan Y C, et al. Dfa-NeRF: personalized talking

head generation via disentangled face attributes neural rendering [PP/OL]. arxiv (2022-01-03) [2024-09-16]. https://arxiv.org/abs/2201.00791
[20] Liu X, Xu Y H, Wu Q Y, et al. Semantic-aware implicit neural audio-driven video portrait generation [C]//European Conference on Computer Vision (ECCV). ECVA, 2022: 106 – 125. DOI: 10.1007/978-3-031-19836-6_7
[21] Ye Z H, He J Z, Jiang Z Y, et al. Geneface++: generalized and stable real-time audio-driven 3d talking face generation [PP/OL]. arxiv (2023-05-01) [2024-09-06]. https://arxiv.org/abs/2305.00787
[22] Yu Z T, Yin Z X, Zhou D Y, et al. Talking head generation with probabilistic audio-to-visual diffusion priors [C]//International Conference on Computer Vision (ICCV). IEEE, 2023: 7611 – 7621. DOI: 10.1109/ICCV51070.2023.00703
[23] Garg R, Roussos A, Agapito L. A variational approach to video registration with subspace constraints [J]. International journal of computer vision, 2013, 104(3): 286 – 314. DOI: 10.1007/s11263-012-0607-7
[24] Andrew A M. Multiple view geometry in computer vision [J]. Kybernetes, 2001, 30(9/10): 1333 – 1341. DOI: 10.1108/k.2001.30.9_10.1333.1
[25] Kajiya J T, Von Herzen B P. Ray tracing volume densities [J]. ACM SIGGRAPH computer graphics, 1984, 18(3): 165 – 174. DOI: 10.1145/964965.808594
[26] Zhou H, Liu Y, Liu Z W, et al. Talking face generation by adversarially disentangled audio-visual representation [J]. Proceedings of the AAAI conference on artificial intelligence, 2019, 33(1): 9299 – 9306. DOI: 10.1609/aaai.v33i01.33019299
[27] Das D, Biswas S, Sinha S, et al. Speech-driven facial animation using cascaded GANs for learning of motion and texture [C]//European Conference on Computer Vision. ECVA, 2020: 408 – 424. DOI: 10.1007/978-3-030-58577-8_25
[28] Richard A, Zollhöfer M, Wen Y D, et al. MeshTalk: 3D face animation from speech using cross-modality disentanglement [C]//IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 1153 – 1162. DOI: 10.1109/ICCV48922.2021.00121
[29] Gafni G, Thies J, Zollhofer M, et al. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 8645 – 8654. DOI: 10.1109/cvpr46437.2021.00854
[30] Hannun A, Case C, Casper J, et al. Deep Speech: scaling up end-to-end speech recognition [PP/OL]. arxiv (2014-12-19) [2024-09-06]. https://arxiv.org/abs/1412.5567
[31] Hong Y, Peng B, Xiao H Y, et al. HeadNeRF: a realtime NeRF-based parametric head model [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 20342 – 20352. DOI: 10.1109/CVPR52688.2022.01973
[32] Ji X Y, Zhou H, Wang K, et al. EAMM: one-shot emotional talking face via audio-based emotion-aware motion model [C]//ACM SIGGRAPH 2022 Conference Proceedings. ACM, 2022: 1 – 10. DOI: 10.1145/3528233.3530745
[33] Tan S, Ji B, Pan Y. EMMN: emotional motion memory network for audio-driven emotional talking face generation [C]//International Conference on Computer Vision (ICCV). IEEE, 2023: 22089 – 22099. DOI: 10.1109/ICCV51070.2023.02024
[34] Müller T, Evans A, Schied C, et al. Instant neural graphics primitives with a multiresolution hash encoding [J]. ACM transactions on graphics, 2022, 41(4): 1 – 15. DOI: 10.1145/3528223.3530127
[35] Liu X, Xu Y H, Wu Q Y, et al. Semantic-aware implicit neural audio-driven video portrait generation [C]//European Conference on Computer Vision (ECCV). ECVA, 2022: 106 – 125. DOI: 10.1007/978-3-031-19836-6_7
[36] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C]//Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 586 – 595. DOI: 10.1109/CVPR.2018.00068
[37] Wang K, Wu Q Y, Song L S, et al. MEAD: a large-scale audio-visual dataset for emotional talking-face generation [C]//European conference on computer vision (ECCV). ECVA, 2020: 700 – 717. DOI: 10.1007/978-3-030-58589-1_42
[38] Chen L L, Li Z H, Maddox R K, et al. Lip movements generation at a glance [C]//European conference on computer vision. ECVA, 2018: 538 – 553. DOI: 10.1007/978-3-030-01234-2_32
[39] Chung J S, Zisserman A. Out of time: automated lip sync in the wild [EB/OL]. [2024-09-06]. https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/chung16a.pdf. DOI: 10.1007/978-3-319-54427-4_19
[40] Lu Y X, Chai J X, Cao X. Live speech portraits: real–time photorealistic talking-head animation [J]. ACM transactions on graphics, 2021, 40(6): 1 – 17. DOI: 10.1145/3478513.3480484
[41] Martin-Brualla R, Radwan N, Sajjadi M S M, et al. NeRF in the wild: neural radiance fields for unconstrained photo collections [C]//Computer Vision and Pattern Recognition. IEEE, 2021: 7206 – 7215. DOI: 10.1109/cvpr46437.2021.00713

Biographies

Lu Ping is the Vice President of ZTE Corporation, Director of the R&D Project of the Technology Planning Department, and Deputy Executive Director of the National Key Laboratory of Mobile Network and Mobile Multimedia Technology. His research fields include immersive communication, cloud computing, big data, augmented reality, and multimedia service technologies. He has supported and participated in major national science and technology projects as well as national science and technology support projects, and has published numerous academic papers in related fields.

Song Li (song_li@sjtu.edu.cn) received his BE and MS degrees in engineering in 1997 and 2000, respectively, and his PhD degree in electrical engineering from Shanghai Jiao Tong University (SJTU), China in 2005. He then joined SJTU as a faculty member and is currently a Full Professor at the Department of Electronic Engineering. He was also a Visiting Professor with Santa Clara University, USA from 2011 to 2012. He has more than 200 publications, obtained over 40 granted patents, and proposed 18 standard technical proposals in video coding and image processing. He has been serving as an Associate Editor for *Multidimensional Systems and Signal Processing* since 2012 and a Guest Editor for a special issue on "Quality of Experience for Advanced Broadcast Services" in 2018 in the *IEEE Transactions on Broadcasting*.

Shi Wenzhe is a strategy planning engineer with ZTE Corporation, a member of the National Key Laboratory of Mobile Network and Mobile Multimedia Technology, China, and a planning engineer of XRExplore platform products. His research objects include immersive communication, indoor visual AR navigation, SFM 3D reconstruction, visual SLAM, real-time cloud rendering, VR, and spatial perception.

Lin Zonghao received his BE degree in information engineering from Shanghai Jiao Tong University, China in 2023. He is currently pursuing his master's degree at the Department of Electronic Engineering, Shanghai Jiao Tong University, China. His research interests include image synthesis and talking face generation.

Ling Jun received his master's degree in electronic engineering and information science from University of Science and Technology of China in 2018. He is currently pursuing his PhD degree at the Department of Electronic Engineering, Shanghai Jiao Tong University, China. His research interests include image animation, talking face generation, and deep generative modeling.

Steel Surface Anomaly Detection Using 3D Depth and 2D RGB Features

Zheng Wangguandong[1], Lu Ping[2], Deng Fangwei[2], Huang Shijun[2], Xia Siyu[1]
(1. Southeast University, Nanjing 210096, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202601011
https://kns.cnki.net/kcms/detail/34.1294.TN.20260302.1536.002.html, published online March 02, 2026
Manuscript received: 2024-09-27

Abstract: The detection of steel surface anomalies has become an industrial challenge due to variations in production equipment, processes, and steel characteristics. To alleviate the problem, this paper proposes a detection and localization method combining 3D depth and 2D RGB features. The framework comprises three stages: defect classification, defect location, and warpage judgment. The first stage uses a data-efficient image Transformer model, the second stage utilizes reverse knowledge distillation, and the third stage performs feature fusion using 3D depth and 2D RGB features. Experimental results show that the proposed algorithm achieves relatively high accuracy and feasibility, and can be effectively used in industrial scenarios.

Keywords: anomaly detection; anomaly localization; feature fusion; reverse distillation

Citation (Format 1): Zheng W G D, Lu P, Deng F W, et al. Steel surface anomaly detection using 3D depth and 2D RGB features [J]. ZTE Communications, 2026, 24(1): 81 - 87. DOI: 10.12142/ZTECOM.202601011
Citation (Format 2): W. G. D. Zheng, P. Lu, F. W. Deng, et al., "Steel surface anomaly detection using 3D depth and 2D RGB features," *ZTE Communications*, vol. 24, no. 1, pp. 81 - 87, Mar. 2026. doi: 10.12142/ZTECOM.202601011.

1 Introduction

Steel plates are widely used in various industrial applications. The surface quality of metal products is an important evaluation metric in the metal manufacturing industry[1]. However, metal plates are affected by various factors during manufacturing, such as equipment, processes, and material characteristics. Consequently, surface defects with irregular shapes often emerge[2]. With the development of computer vision, machine vision-based algorithms for detecting such defects have become a research focus[3]. To address challenges in metal surface defect detection—such as scarce defect samples, high variability in shape and type, and the need for precise localization—this paper proposes a steel surface anomaly detection and localization method.

Previous work has introduced various approaches to address challenges in anomaly detection and classification within industrial processes. Zhao et al.[4] proposed a method based on dynamic time warping (DTW) combined with adaptive fuzzy C-means (AFCM), drawing inspiration from similar industrial processes. Wen et al.[5] developed a novel anomaly detection method based on multi-scale knowledge distillation (Ms-KD) and a block domain core information module (BDCI) to quickly screen abnormal images. Yasuno et al.[6] proposed a one-class steel detector using a patch generative adversarial network (GAN) discriminator for visualizing anomalous feature maps.

Fig. 1 shows the pipeline of the proposed method. Given a 2D RGB image, we classify it into two categories (i.e., abnormal and normal) using the data-efficient image Transformer (DEIT) model.

1) For abnormal images, we employ a multi-class DEIT model and a reverse knowledge distillation model to determine the specific defect category and the defect coordinates, respectively.

2) For normal images, we combine the 2D RGB features with the 3D depth map to fuse 2D and 3D features. Furthermore, a new classification head is employed to determine whether the steel plate is warped or flat.

Finally, the defect category is determined by integrating the specific defect type and the warpage status.

2 Proposed Method

In actual production, the limited availability of steel defect samples, coupled with the diversity of defect types, poses a significant challenge to accurate detection and classification. Given this data scarcity, it is necessary to achieve efficient discrimination with minimal data. To tackle this, we implement a data-efficient defect multi-classification method for the abnormal multi-classification, which effectively distinguishes be-

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20221107001.

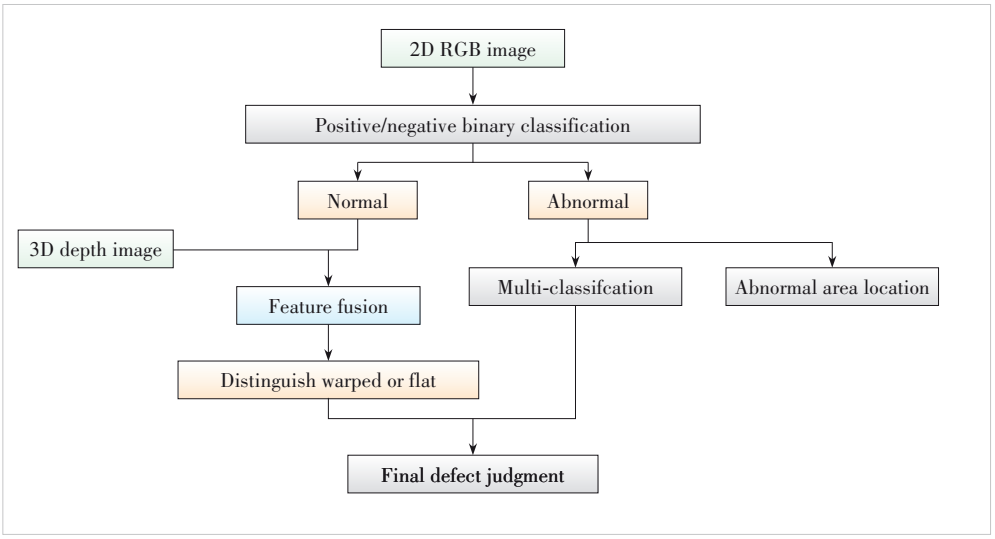


Figure 1. Pipeline of our proposed method

tween different types of steel defects. Considering the difficulties traditional knowledge distillation faces in pinpointing defect areas, we employ the reverse distillation defect location method to accurately identify the defective area. Lastly, due to the spatial unevenness of warpage defects, discerning them using only 2D images is challenging. Therefore, we integrate 3D information. Thus, we utilize feature fusion of both 2D RGB and 3D depth images to determine whether the steel is warped.

2.1 Data-Efficient Defect Multi-Classification Method

The number of steel defect samples in actual production is limited, necessitating a multi-classification model that can operate effectively with minimal data. Existing transformer-based classification models, such as vision Transformer (ViT), need to be pre-trained on large-scale datasets and then fine-tuned on the ImageNet dataset[7], which requires substantial computing resources. DEIT[8] is essentially a ViT model. It uses three methods: better hyper-parameters, data augmentation and distillation, which can achieve better classification performance with a smaller amount of data.

1) Optimized hyper-parameters. The parameter initialization method chosen is a truncated normal distribution. For learning rate adjustment, we employ a decay strategy: the learning rate first increases linearly during the warm-up phase and then decreases via a cosine method.

2) Data augmentation. A variety of data augmentation methods are used, including random erase, MixUp, CutMix, and exponential moving average (EMA). With MixUp, the resulting images are assigned soft labels rather than single labels. In CutMix, labels are given according to the proportion occupied. EMA ensures that the model weight updates are related to the historical values over time. These methods all help to improve the model's efficiency.

3) Distillation through attention. In the training stage, the class token in ViT for classification is equivalent to an additional patch. It learns the relationship with other patches, and then connects the classifier to calculate CELoss. As shown in Fig. 2, for distillation in DEIT, an additional distill token is added. This token also learns the relationship with other tokens, and then connects the teacher model to calculate KLDivLoss. Subsequently, CELoss and KLDivLoss are combined to form a new loss, which guides the student model training (note that the teacher model is not trained during knowledge distillation).

In the prediction stage, class token and distill token generate different results. These results are then weighted (with a weight of 0.5 each) and summed to obtain the final prediction.

2.2 Reverse Distillation Defect Location Method

As shown in Fig. 3, in traditional knowledge distillation, both the teacher and student networks serve as encoders, taking image information as input. The student network learns from the teacher network by reconstructing the representations of the teacher network at different scales[9]. However, in re-

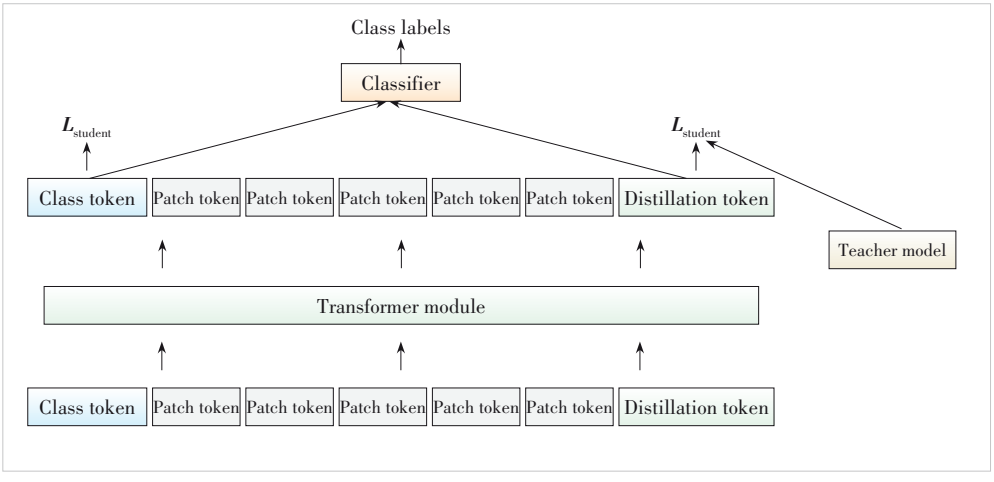


Figure 2. Distillation token in transformer model

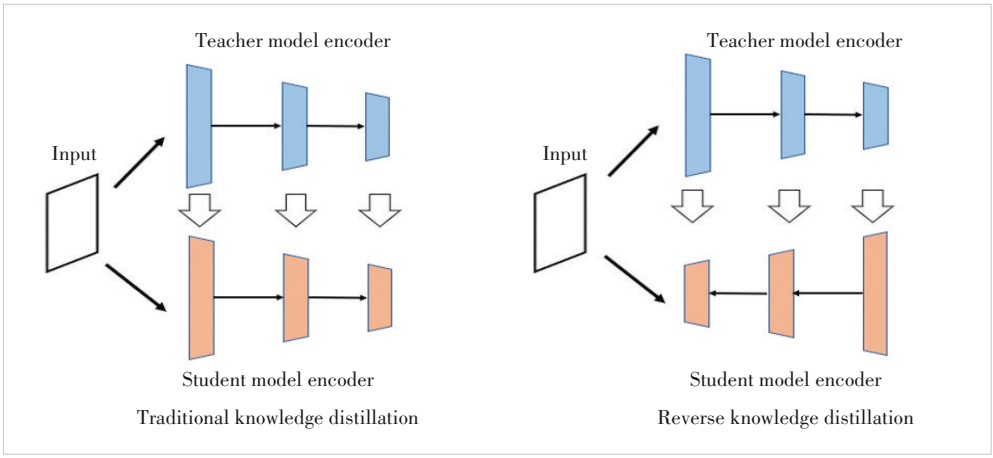


Figure 3. Difference between traditional knowledge distillation and reverse knowledge distillation

verse knowledge distillation, the teacher network still acts as an encoder, while the student network functions as a decoder[10]. The low-dimensional features encoded by the teacher model serve as input, allowing the student network to learn the teacher model's representations at different scales by reconstructing them. This process first extracts high-level representations and then refines low-level features. The teacher encoder functions as a downsampling filter, while the student decoder operates as an upsampling filter, creating a symmetric architecture that addresses the limitations of traditional knowledge distillation.

In the inference stage of traditional knowledge distillation, when abnormal samples are input, the student network may reconstruct results highly similar to those of the teacher network. However, to alleviate the problem, Fig. 4 shows that reverse knowledge distillation adopts the following methods:

1) The encoder module (part of the teacher network) utilizes pretrained models. In our implementation, we employed WideResNet and achieved competitive performance.

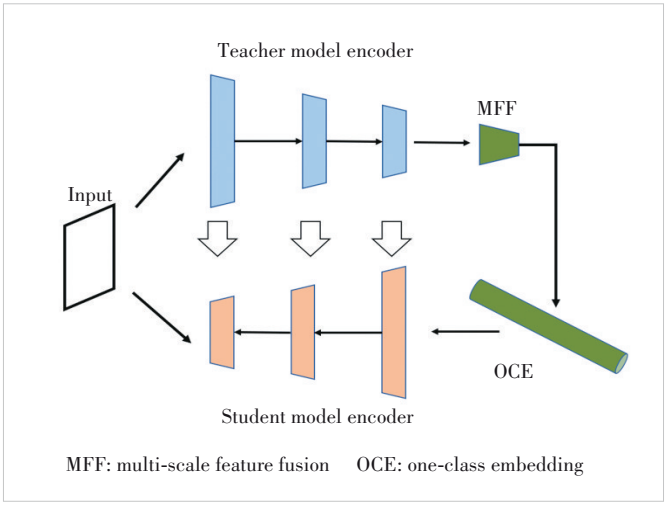


Figure 4. Reverse knowledge distillation network architecture

2) A one-class bottleneck embedding (OCBE) module is incorporated into inverse knowledge distillation. It contains a multi-scale feature fusion (MFF) module and one-class embedding (OCE) module. The motivation for employing multi-scale fusion stems from the distinct characteristics of features at different scales: low-dimensional features are rich in texture and edge details, while high-dimensional features encapsulate semantic information. Using only the activation information from the encoder's final layer as the decoder's input could lead to an excess of redundant semantic details. Therefore, by leveraging multi-scale fusion, redundancy is minimized while preserving details.

3) The decoder module mirrors the teacher network but is not an exact copy. During the inference phase, when abnormal samples are input, the reconstructed shapes exhibit greater differences, making the defect features more apparent.

2.3 Feature Fusion via 2D RGB Images and 3D Depth Images

With the popularization of various sensors, it is easier to obtain multimodal data from different sources, making it increasingly important to use multi-modal information for various classification and regression tasks[11]. According to how multi-modal fusion is performed, it can be divided into the two following types: aggregation-based fusion and alignment-based fusion.

Aggregation-based fusion employs separate sub-networks to process each modality. The outputs are then aggregated into a unified common feature. These features are subsequently mapped to the output dimension to obtain the final result. The specific formula is as follows:

$$\hat{y}^{(i)} = f\left(x^{(i)}\right) = h\left(\mathrm{Agg}\left(f_1\left(x_1^{(i)}\right),\cdots,f_M\left(x_M^{(i)}\right)\right)\right) \quad (1),$$

where h is the global mapping network, f is the feature extractor, $x^{(i)}$ is the input image, and Agg is the aggregation function. There are many ways to implement aggregation functions, such as averaging multiple modal features and concatenating multiple modal features.

The alignment-based fusion method refers to using an alignment loss to align the features of multiple modalities, retaining the outputs of multiple sub-networks for separate prediction, and finally weighting the prediction results of the different modalities[12]. The optimization objective in this case is shown as:

$$\min \frac{1}{N}\sum_{i=1}^{N} L\left(\sum_{m=1}^{M}\alpha_m f_m\left(x_m^{(i)}\right), y^{(i)}\right) + \mathrm{Alig}_{f_{1,M}}\left(x^{(i)}\right), \text{s.t.} \sum_{m=1}^{M}\alpha_m = 1 \quad (2),$$

where $\text{Alig}_{f_{1:M}}$ is a loss that measures the similarity between two distributions, usually using maximum-mean-discrepancy (MMD). The final output $\sum_{m=1}^{M}\alpha_m f_m\left(x_m^{(i)}\right)$ is an ensemble of f_m associated with the decision score α_m, which is learned by an additional softmax output to meet the simplex constraint.

This task focuses on the homogeneous multimodal fusion problem, using 2D RGB images and 3D depth maps for feature fusion. The proposed method belongs to the category of aggregation-based fusion. As shown in Fig. 5, two ResNet18 networks[13] are employed as feature extractors. First, the final fully connected layer of the network is removed; then, 3D depth maps and 2D RGB images are input to extract the two corresponding features. These two features are concatenated, and a new binary classification head is customized to perform the warpage detection task.

3 Experiment

We evaluated the performance of this algorithm to classify and locate defects on steel surfaces. Experimental results show the algorithm yields favorable classification and localization outcomes on iron, aluminum, and stainless steel surfaces.

3.1 Datasets

1) Defect classification dataset. Two-dimensional defects include abrasions, scratches, holes, stripes and flower patterns, totaling five categories. The materials used are aluminum, iron, and stainless steel. Each full-size steel plate image (2 048×2 048 pixels) is divided into 64 small patches (256×256 pixels). As shown in Fig. 6, all images are collected by 2D line array cameras. The aluminum subset contains 452 abnormal and 153 normal samples, the iron subset contains 307 abnormal and 205 normal samples, and the stainless steel subset contains 243 abnormal and 195 normal samples. All samples are divided into training, verification and test sets at a ratio of 6:2:2.

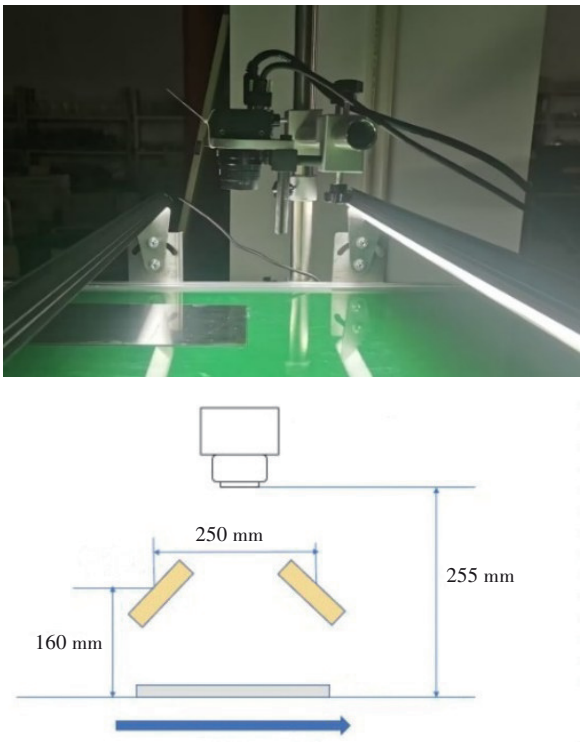


Figure 6. Line array cameras capture real scenes and schematic scenes

2) Defect location dataset. This dataset is constructed similarly to the defect classification dataset, but with the difference that ground truth needs to be incorporated during training to inform the reverse knowledge distillation about the defect locations.

3) Feature fusion dataset. This dataset, comprising 40 2D RGB images and 40 3D depth images, is divided into warping and flattening parts for both modalities. The RGB images are captured of iron plates using a 2D line-scan camera, while the corresponding depth images are acquired using a 3D area-scan camera.

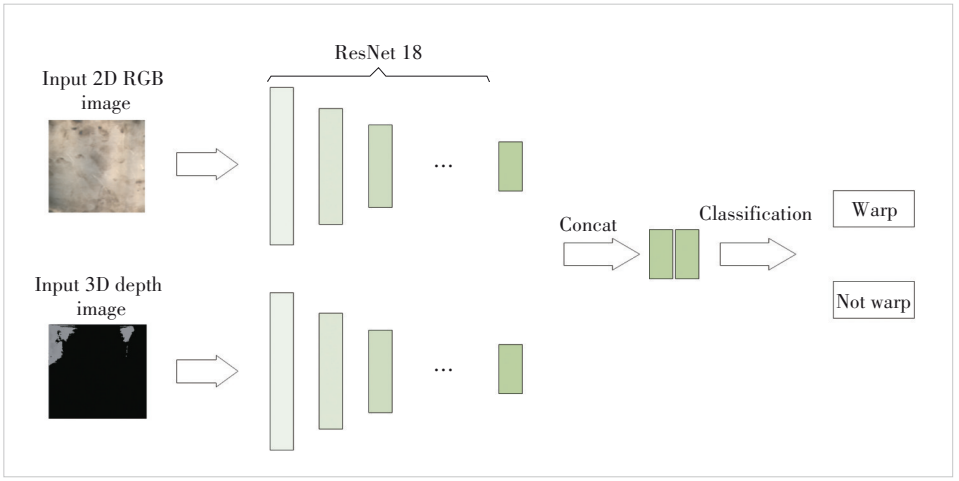


Figure 5. Aggregation-based feature fusion

3.2 Implementation Details

In the defect multi-classification process, we employed the DeiT-Tiny pre-training model, using the AdamW[14] optimizer, and the learning rate was set to 5e-5×4/64. The patch size was set to 16 and the LabelSmoothLoss was adopted as the loss function. In the reverse distillation process, WideResNet50[15] was chosen as the teacher model,

trained with the Adam optimizer and a learning rate scheduler (gamma=0.1, with a step size of 5). For feature fusion, ResNet18 was selected as the feature extraction network with a learning rate of 1e-3.

3.3 Experimental Results and Analysis

1) Standard of evaluation. Given the characteristics of the classification task, precision (P) is adopted as the primary evaluation metric. True Positives (TP) denote the number of positive samples correctly classified by the model, while False Positives (FP) represent the number of negative samples incorrectly classified as positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3).$$

Based on the characteristics of the location results, the pixel-level Area Under the Receiver Operating Characteristics curve (AUROC) and image-level AUROC are selected as the main location evaluation metrics[16].

AUROC assesses the model's ability to distinguish between positive and negative samples by plotting the false positive rate (FPR) against the true positive rate (TPR) at different thresholds. The FPR represents the proportion of negative samples that are incorrectly classified as positive, while the TPR denotes the proportion of actual positive samples that are correctly identified. Typically, we aim for a high TPR while minimizing the FPR to enhance the model's classification ability.

Pixel-level AUROC evaluates the prediction accuracy of the model at the pixel level, treating each pixel as an independent classification problem. However, the image-level AUROC assesses the entire image for binary classification, disregarding the specific positions and types of each pixel.

2) Data-efficient defect multi-classification. Due to the scarcity of defective steel samples, 25 training sets and 9 verification sets are used to achieve better multi-classification effects on aluminum, iron, and stainless steels. Table 1 shows that the classification accuracy is at least 90% and often reaches 100%. Training loss and validation accuracy are shown in Fig. 7, and the classification results on five types of defects are shown in Fig. 8.

3) Reverse distillation defect location. As shown in Table 2, the pixel-level and image-level AUROC scores indicate high defect localization accuracy. Fig. 7 visualizes these results using heatmaps.

4) Feature fusion via 2D RGB and 3D depth images. An iron plate was selected as the experimental sample. The dataset comprises 20 warped and 20 flat samples for both 2D RGB images and 3D depth images. The data was partitioned into training, verification, and test sets according to a ratio of 6:2:2. Due to the large discrimination of feature representation, the classification accuracy for distinguishing warped from flat samples reached 100%.

Table 1. Number of testset and precision of aluminum, iron, and stainless steel samples

Defect Category	Aluminum	Iron	Stainless Steel
Abrasions	8/100%	8/100%	30/96.7%
Scratches	212/97.17%	101/99.1%	18/100%
Holes	28/100%	12/100%	8/100%
Stripes	8/100%	8/100%	8/100%
Flowers	26/100%	8/100%	8/100%

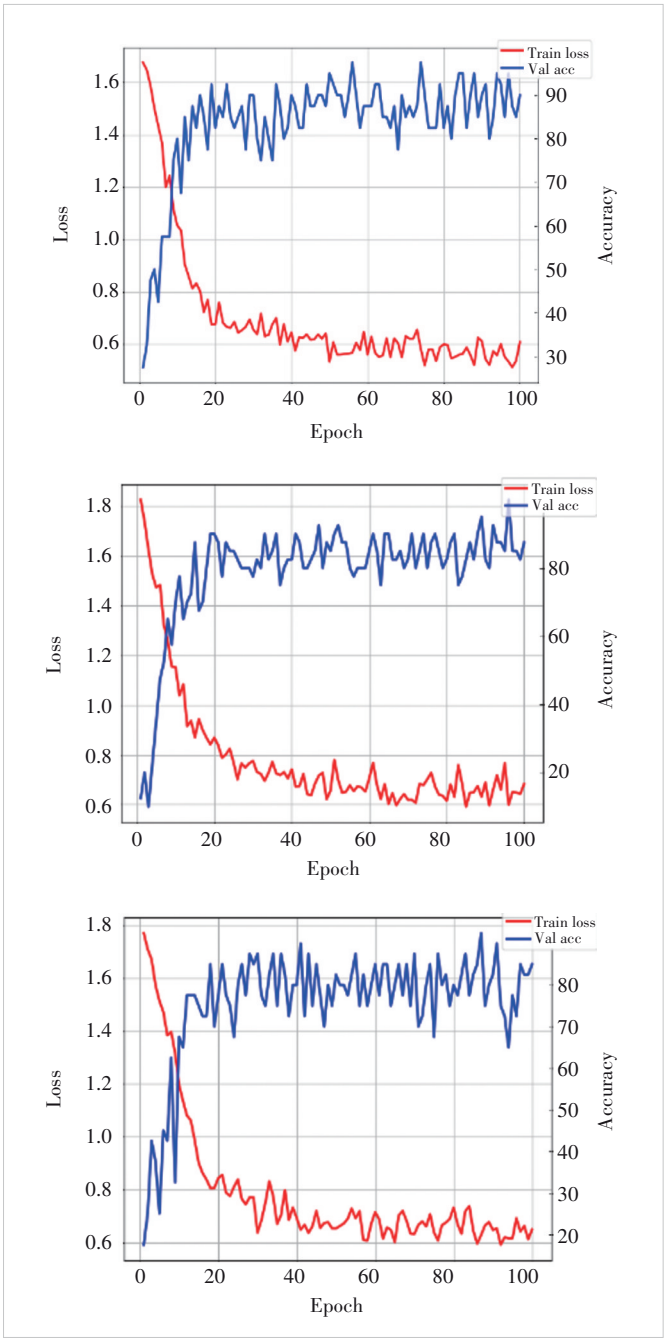


Figure 7. Training loss and validation accuracy of aluminium, iron, and stainless steel samples, where the red curve means training loss and the blue means validation accuracy

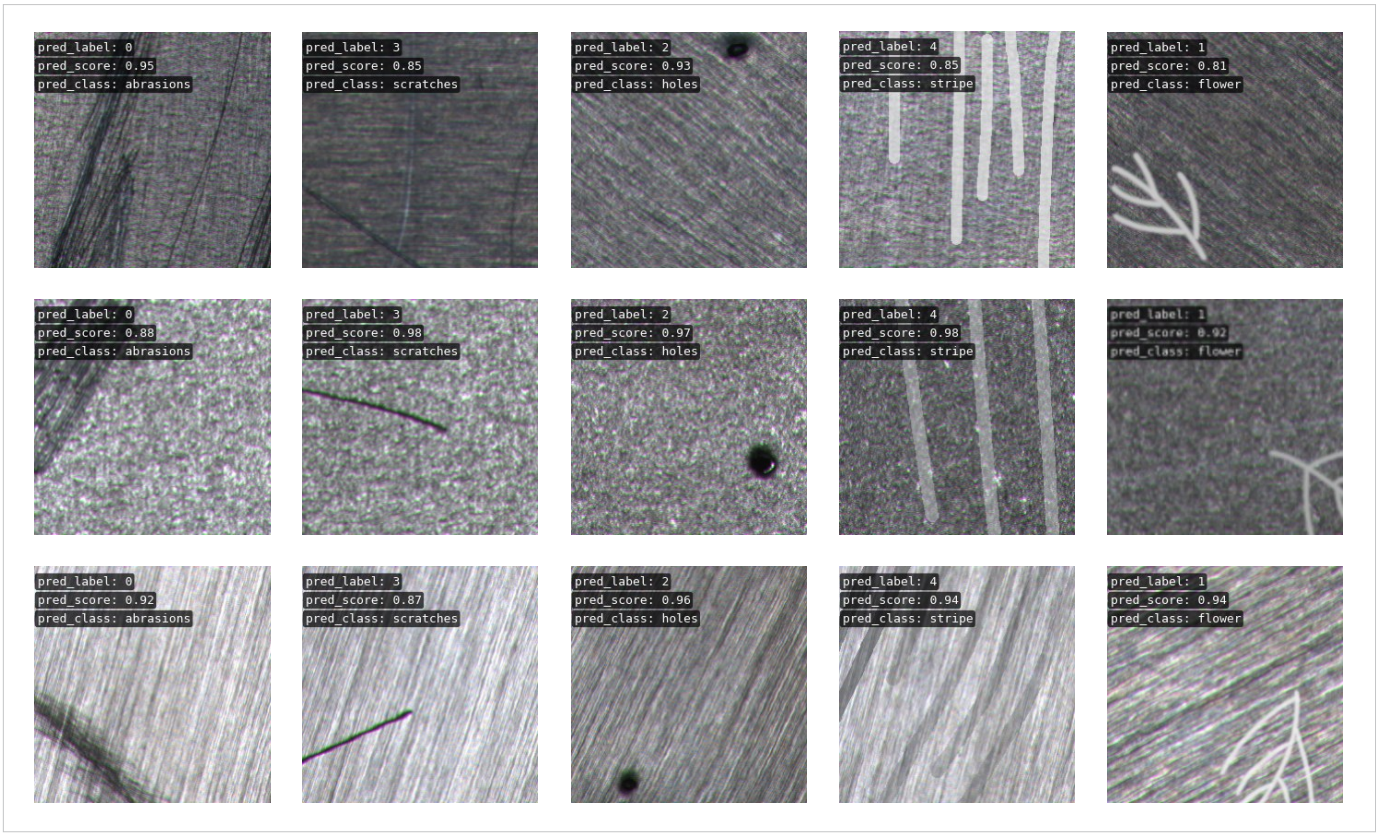


Figure 8. Classification results of aluminum, iron and stainless steel on five types of defects

Table 2. Pixel-level AUROC and image-level AUROC of aluminum, iron, and stainless steel samples

Steel Category	Defect Category	Pixel-level AUROC	Image-level AUROC
Aluminum	Abrasions	97.3%	96.7%
	Holes	99.8%	100%
	Scratches	97.5%	100%
Iron	Abrasions	98.1%	100%
	Holes	97.8%	100%
	Scratches	96.3%	100%
Stainless steel	Abrasions	96.4%	100%
	Holes	99.2%	95.4%
	Scratches	85.8%	98.1%

AUROC: Area Under the Receiver Operating Characteristics curve

4 Conclusions

This paper proposes a method for detecting and locating anomalies of steel surfaces combined with 3D Depth and 2D RGB features, which can be divided into three stages: defect classification, defect location, and warp detection. By leveraging deep learning techniques, the proposed approach minimizes the reliance on manual labor during the inspection process. Experimental results demonstrate that the method achieves the desired accuracy and validates its feasibility.

References

[1] Defard T, Setkov A, Loesch A, et al. PaDiM: a patch distribution modeling framework for anomaly detection and localization [C]//ICPR International Workshops and Challenges. ICPR, 2020: 475 – 489. DOI: 10.1007/978-3-030-68799-1_35
[2] Tao X, Gong X, Zhang X, et al. Deep learning for unsupervised anomaly localization in industrial images: a survey [J]. IEEE Transactions on instrumentation and measurement, 2022, 71(1): 1 – 21. DOI: 10.1109/TIM.2022.3196436
[3] He Y, Song K, Meng Q, et al. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features [J]. IEEE transactions on instrumentation and measurement, 2020, 69(4): 1493 – 1504. DOI: 10.1109/TIM.2019.2915404
[4] Zhao J, Liu K, Wang W, et al. Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry [J]. Information sciences, 2014, 259: 335 – 345. DOI: 10.1016/j.ins.2013.05.018
[5] Wen X, Zhao W, Yu Z, et al. A novel anomaly detection method for strip steel based on multi-scale knowledge distillation and feature information banks network [J]. Coatings, 2023, 13(7): 1171. DOI: 10.3390/coatings13071171
[6] Yasuno T, Fujii J, Fukami S. One-class steel detector using patch GAN discriminator for visualising anomalous feature map [PP/OL]. arXiv (2021-06-30) [2025-08-12]. https://arxiv.org/abs/2107.00143
[7] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale [PP/OL]. arXiv (2021-07-03) [2025-08-12]. https://arxiv.org/abs/2010.11929
[8] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention [PP/OL]. arXiv [2025-08-12]. https://arxiv.org/abs/2012.12877
[9] Salehi M, Sadjadi N, Baselizadeh S, et al. Multiresolution knowledge distil-

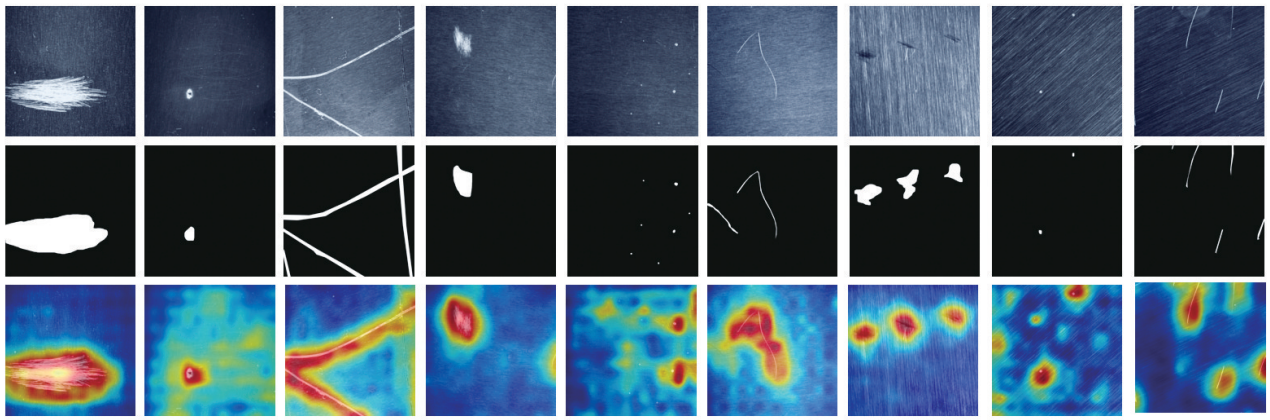


Figure 9. Defect location results by heatmaps

lation for anomaly detection [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 14897 – 14907. DOI: 10.1109/cvpr46437.2021.01466
[10] Deng H Q, Li X Y. Anomaly detection via reverse distillation from one-class embedding [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 9727 – 9736. DOI: 10.1109/cvpr52688.2022.00951
[11] Zhu J G, Tang S X, Chen D P, et al. Complementary relation contrastive distillation [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 9256 – 9265. DOI: 10.1109/cvpr46437.2021.00914
[12] Wang Y K, Huang W B, Sun F C, et al. Deep multimodal fusion by channel exchanging [PP/OL]. arXiv [2025-08-12]. https://arxiv. org/abs/2011.05005
[13] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 770 – 778. DOI: 10.1109/CVPR.2016.90
[14] Zagoruyko S, Komodakis N. Wide residual networks [PP/OL]. arXiv [2025-08-12]. https://arxiv.org/abs/1605.07146
[15] Roth K, Pemula L, Zepeda J, et al. Towards total recall in industrial anomaly detection [C]//Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 14298 – 14308. DOI: 10.1109/CVPR52688.2022.01392
[16] Kingma D P, Ba J. Adam: a method for stochastic optimization [PP/OL]. arXiv (2014-12-22) [2025-08-12]. https://arxiv.org/abs/1412.6980

Biographies

Zheng Wangguandong is pursuing his master's degree at the School of Automation, Southeast University, China. His research interests focus on artificial intelligence and computer vision, with a specific specialization in image and video generation. He has published four CCF-A conference papers and possesses extensive research experience in image segmentation and object detection.

Lu Ping is the executive deputy director of the National Key Laboratory of Mobile Networks and Mobile Multimedia Technology, China. His research areas encompass cloud computing, big data, augmented reality, and multimedia serviceization. He leads and participates in major national science and technology projects and national science and technology support programs. He has published numerous academic papers and is the author of the books *Internet of Things Capability Development and Application* and *Big Data Technology and Application in Cloud Computing*.

Deng Fangwei is a senior strategic planner at ZTE Corporation, specializing in industry-specific digital infrastructure, mobile robots, and supporting products for industrial digital transformation.

Huang Shijun is a senior strategic planner at ZTE Corporation, with research interests encompassing machine vision, artificial intelligence, computer vision, and deep learning.

Xia Siyu (xsy@seu.edu.cn) received his BE and MS degrees in automation engineering from Nanjing University of Aeronautics and Astronautics, China in 2000 and 2003, respectively, and the PhD degree in pattern recognition and intelligence systems from Southeast University, China in 2006. He is currently an associate professor with the School of Automation, Southeast University. His research interests include object detection, applied machine learning, social media analysis, and intelligent vision systems. He was a recipient of the Science Research Famous Achievement Award from the Higher Institution of China in 2015. He has served as a reviewer for many journals including *IEEE T-PAMI*, *T-IP*, *T-SMCB*, *T-IFS*, *T-MM*, and *Neurocomputing*. He received the Outstanding Reviewer Award for Neurocomputing in 2016. He has also served on the PC/SPC for conferences including CVPR, AAAI, ACM MM, and IJCAI. He is a member of the ACM and IEEE.

Synthesis and Design of Generalized Strongly Coupled Resonator Quartet Combline Filters with Redundant Resonance

Xiong Zhi'ang[1], Fan Jiyuan[1], Zhao Ping[1], Zhou Jinzhu[1], Shen Nan[2], Wu Qingqiang[2]

(1. Xidian University, Xi'an 710071, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202601012

https://kns.cnki.net/kcms/detail/34.1294.TN.20260210.1137.004.html, published online February 10, 2026

Manuscript received: 2024-03-20

Abstract: This article proposes a generalized strongly coupled resonator quartet (GSCRQ) filter along with its synthesis approach. By introducing out-of-band reflection zeros (RZs), the proposed GSCRQ can generate a transmission zero on each side of the passband without negative couplings. The coupling coefficients in this coupling structure change with the positions of the out-of-band RZs. Thus, the GSCRQ configuration admits flexible design solutions. For GSCRQ coaxial combline filters, all couplings can be implemented as inductive couplings, simplifying the design and manufacturing process. In this article, a 6-2 filter in the GSCRQ configuration is synthesized and designed. The simulated results of the designed filter agree very well with the theoretical characteristics.

Keywords: filter synthesis; generalized strongly coupled resonator quartets (GSCRQ); out-of-band reflection zero; transmission zero

Citation (Format 1): Xiong Z A, Fan J Y, Zhao P, et al. Synthesis and design of generalized strongly coupled resonator quartet combline filters with redundant resonance [J]. ZTE Communications, 2026, 24(1): 88 - 98. DOI: 10.12142/ZTECOM.202601012

Citation (Format 2): Z. A. Xiong, J. Y. Fan, P. Zhao, et al. "Synthesis and design of generalized strongly coupled resonator quartet combline filters with redundant resonance," *ZTE Communications*, vol. 24, no. 1, pp. 88 - 98, Mar. 2026. doi: 10.12142/ZTECOM.202601012.

1 Introduction

Transmission zeros (TZs) at finite frequencies are commonly seen in modern microwave filters. They can improve frequency selectivity and out-of-band rejection of bandpass filters. Nowadays, cross-couplings are the most commonly used method to realize TZs. Frequently used cross-coupled structures include trisections, quartets, and box sections[1-3]. Cross-coupled configurations generate TZs by multi-path cancellation. TZs are generated when the combined signals interfere destructively at the combining node. A trisection or a box section can realize only one TZ on the imaginary axis. A quartet can introduce two TZs, and the two TZs can be on the imaginary axis or form a para-conjugate pair. Usually, when the generalized Chebyshev filter has TZs on the lower side of the passband, at least one of the couplings in the trisection or quartet has to be negative. Negative couplings in coaxial combline filters are often realized by capacitive probes. The capacitive probe increases complexity, reduces power handling capability, and is inconvenient for post-production tuning.

Macchiarella et al. proposed strongly coupled resonator pairs (SCRP)[4] and strongly coupled resonator triplets (SCRT)[5]. Compared with traditional filters, strongly coupled filter configurations consist solely of positive couplings, even though TZs on the lower stopband are realized. It is also interesting to note that there is a reflection zero (RZ) in the lower stopband. Zeng[6] introduced an out-of-band RZ into the synthesis of equi-ripple filtering functions, enabling the SCRT structure to be synthesized directly by coupling matrix transformations. This synthesis approach addresses the limitation that the out-of-band RZ of the SCRT must be designed far from the passband. The synthesized structure is named the generalized strongly coupled resonator triplet (GSCRT). Recently, a coaxial combline filter design was proposed based on the strongly coupled resonator quadruplet (SCRQ)[7]. The SCRQ consists of an SCRT and a regularly coupled fourth resonator. However, there is no analytical synthesis method

This work was supported by the National Natural Science Foundation of China under Grant No. 62471366.

for SCRQ filters.

This work presents the synthesis and design of a generalized strongly coupled resonator quartet (GSCRQ), where all couplings can be positive, as determined by the synthesis procedure. Compared with SCRQ, GSCRQ can arbitrarily specify the location of the out-of-band RZ to achieve a flexible topology configuration. First, the working mechanism of GSCRQ at the circuit level is analyzed, and the relationship between the TZs and couplings (self-coupling and mutual coupling) is explained. Then, a direct matrix synthesis approach for filters with GSCRQ is presented. In addition, this paper proposes a parameter-extraction method for filters with out-of-band RZs. The method can aid the design and tuning process of filters with GSCRQ. The filter-tuning process is faster than port-tuning[8]. Finally, a sixth-order filter with a GSCRQ is synthesized and the simulation results show good agreement with the synthesized response.

2 Filters with GSCRQ

In order to understand the working mechanism of the GSCRQ configuration, we first analyze the circuit model of the quartet shown in Fig. 1. The TZs are generated by the couplings among resonators 1 – 4. The admittance matrix $\boldsymbol{Y}$ between nodes 1 and 4 is

$$\begin{aligned}\boldsymbol{Y} = &\frac{1}{s+\mathrm{j}B_2}\begin{bmatrix} M_{12}^{\ 2} & M_{12}M_{24} \\ M_{12}M_{24} & M_{24}^2 \end{bmatrix} + \begin{bmatrix} 0 & \mathrm{j}M_{14} \\ \mathrm{j}M_{14} & 0 \end{bmatrix} + \\ &\frac{1}{s+\mathrm{j}B_3}\begin{bmatrix} M_{13}^{\ 2} & M_{13}M_{34} \\ M_{13}M_{34} & M_{34}^2 \end{bmatrix} = \\ &\frac{1}{(s+\mathrm{j}B_2)(s+\mathrm{j}B_3)} \times \\ &\begin{bmatrix} \cdots & (M_{12}M_{24}+\mathrm{j}M_{14}s-B_2M_{14})(s+\mathrm{j}B_3)+(s+\mathrm{j}B_2)(M_{13}M_{34}) \\ \cdots & \cdots \end{bmatrix}\end{aligned} \tag{1}$$

Equating the numerator polynomial of Y_{12} to zero, we can identify the TZs generated by the GSCRQ as:

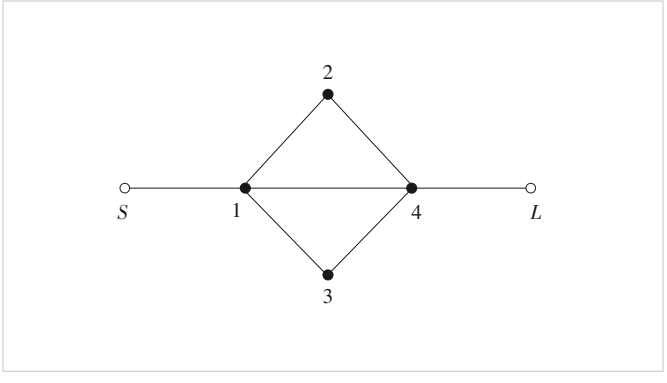


Figure 1. Topology of quartets. Black nodes represent resonators, and solid lines represent couplings

$$\begin{aligned}s_{\mathrm{TZ}} = &\frac{\mathrm{j}}{2M_{14}} \times \\ &\left(\begin{array}{c} \left(M_{12}M_{24}+M_{13}M_{34}-M_{14}\left(B_2+B_3\right)\right) \pm \\ \sqrt{\left(M_{12}M_{24}-B_2M_{14}-B_3M_{14}+M_{13}M_{34}\right)^2 + \\ 4M_{14}\left(B_3M_{12}M_{24}-B_3B_2M_{14}+B_2M_{13}M_{34}\right)} \end{array} \right)\end{aligned} \tag{2}$$

Empirically, for the quartet shown in Fig. 1, whether the TZs are located in the upper or lower stopband, there will exist a negative coupling. According to Eq. (2), it can be seen that the mutual couplings and self-couplings interact with each other to generate two TZs. We consider introducing out-of-band RZs to realize TZs while keeping all the couplings in the quartet positive.

2.1 Remez-Like Algorithm

Lowpass prototype filter responses can be defined as the ratio of polynomials $E(s)$, $F(s)$, $F_{22}(s)$ and $P(s)$ as:

$$\boldsymbol{S} = \frac{1}{E(s)}\begin{bmatrix} F(s)/\varepsilon_R & P(s)/\varepsilon \\ P(s)/\varepsilon_R & F_{22}(s)/\varepsilon_R \end{bmatrix} \tag{3}$$

where ε and ε_R are positive real constants such that the absolute values of the highest-order coefficients in $E(s)$, $F(s)$, $F_{22}(s)$ and $P(s)$ can be normalized.

To synthesize a filter with out-of-band RZs, a Remez-like iterative algorithm can be used. The polynomials $F_{\mathrm{out}}(\omega)$, $F_{\mathrm{in}}(\omega)$, and $P(\omega)$ are defined as:

$$F_{\mathrm{out}}(\omega) = \prod_{i=1}^{nc}\left(\omega - \omega_{\mathrm{out},i}\right) \tag{4a}$$

$$F_{\mathrm{in}}(\omega) = \prod_{i=1}^{np-nc}\left(\omega - \omega_{\mathrm{in},i}\right) \tag{4b}$$

$$P(\omega) = \prod_{i=1}^{nz}\left(\omega - \omega_{nz}\right) \tag{4c}$$

where np is the order of the filter, nc is the number of out-of-band RZs, $\omega_{\mathrm{in},i}$ denotes in-band RZs, ω_{nz} and $\omega_{\mathrm{out},i}$ are the assigned TZs and out-of-band RZs, respectively. In the Remez-like algorithm, $F_{\mathrm{out}}(\omega)$ is completely determined by the assigned out-of-band RZs. We seek the $np-nc$ unknown coefficients $\{a_i\}$ of $F_{\mathrm{in}}(\omega)$ such that the characteristic function $C(\omega)=F_{\mathrm{out}}(\omega)F_{\mathrm{in}}(\omega)/P(\omega)$ is equiripple in the passband. The filtering synthesis procedure then proceeds as follows.

Step 1: Solve the linear system comprising $np-nc+1$ equations with the $np-nc+1$ unknowns $\left[\{a_i\},\Delta\right]$:

$$\begin{cases} C(-1) = \Delta \\ C\left(\Omega_k\right) = (-1)^k\Delta \quad k = 1,\cdots,np-nc-1 \\ C(1) = (-1)^{np-nc}\Delta \end{cases} \tag{5}$$

where Δ is the amplitude of the in-band ripple of $C(\omega)$ and Ω_k are the extreme points. The equations in Eq. (5) can be re-written in a matrix form as:

$$\begin{bmatrix} \Omega_1^{np-nc-1} & \Omega_1^{np-nc-2} & \cdots & 1 & -P(\Omega_1)/F_{out}(\Omega_1) \\ \Omega_2^{np-nc-1} & \Omega_2^{np-nc-2} & \cdots & 1 & P(\Omega_1)/F_{out}(\Omega_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Omega_{np+1}^{np-nc-1} & \Omega_{np+1}^{np-nc-2} & \cdots & 1 & (-1)^{np-nc+1}P(\Omega_1)/F_{out}(\Omega_1) \end{bmatrix} \begin{bmatrix} a_{np-nc-1} \\ a_{np-nc-2} \\ \vdots \\ a_0 \\ \Delta \end{bmatrix} = \begin{bmatrix} -\Omega_1^{np-nc} \\ -\Omega_2^{np-nc} \\ \vdots \\ -\Omega_{np+1}^{np-nc} \end{bmatrix} \quad (6),$$

where a_i ($i = np-nc-1, \cdots, 0$) are the coefficients of the polynomial $F_{in}(\omega)$.

Step 2: After obtaining the coefficients a_i, update the locations of the in-band extrema Ω' by solving $(\partial C(\omega)/\partial\omega) = 0$.

Step 3: Repeat Steps 1 and 2 until convergence is achieved, i.e., $|\Omega - \Omega'|$ is sufficiently close to zero. At convergence, the coefficients $\{a_i\}$ of $F_{in}(\omega)$ are stable. The polynomial $F(\omega)$ is then constructed as:

$$F(\omega) = F_{in}(\omega)F_{out}(\omega) \quad (7).$$

Step 4: Calculate ε and ε_R using

$$\varepsilon_R = 1, \quad \varepsilon = \frac{1}{\sqrt{10^{RL/10}-1}}\left|\frac{P(\omega)}{F(\omega)}\right|_{\omega=\pm1} \quad (8),$$

where RL is the desired return loss level.

The monic polynomial $E(\omega)$ is obtained from $P(\omega)$ and $F(\omega)$ through the Feldtkeller equation:

$$E(\omega)E(\omega)^* = F(\omega)F(\omega)^*/\varepsilon_R^2 + P(\omega)P(\omega)^*/\varepsilon^2 \quad (9).$$

We can solve the right-hand side of Eq. (9) to obtain $2np$ roots, from which those with positive imaginary parts are selected to construct $E(\omega)$. Alternatively, a more robust computational method has been recently proposed in Ref. [9] to accurately solve the Feldtkeller equation for high-order networks. After $E(\omega)$ is obtained, the polynomials $F(s)$, $P(s)$ and $E(s)$ are formed by the variable substitution $\omega = s/j$. If the coefficients of the highest order terms of these polynomials, obtained by variable substitution, are not equal to 1, they should be normalized by dividing the coefficient of the highest order terms. It should be noted that when $np-nc$ is even, $P(s)$ is multiplied by j to satisfy the unitary condition, ensuring its highest-order coefficient is j[10]. Finally, the polynomial $F_{22}(s)$ is given by

$$F_{22}(s) = (-1)^{np}F(s)^* \quad (10),$$

where (*) denotes polynomial para-conjugation.

After the Chebyshev-like polynomials are synthesized, the $np + 2$ transversal coupling matrix can be synthesized using the well-established technique described in Ref. [10]. The GSCRQ coupling topology is then obtained by a sequence of similarity transformations from the transversal coupling matrix.

The synthesis process properties of GSCRQ will be demonstrated and discussed in the following sections.

2.2 A Synthesis Example of GSCRQ Filters

Consider an example with $np = 7$ poles, $nc = 1$ out-of-band RZ, and TZs at $\{s_{TZ}\} = \{-1.36j, 1.25j\}$. The return loss level is RL = 20 dB, and the assigned out-of-band RZ is $s_{out} = -8.26j$.

First, $P(\omega)$ and $F_{out}(\omega)$ are obtained from the prescribed TZs and the out-of-band RZ, respectively:

$$P(\omega) = (\omega + 1.36)(\omega - 1.25) = \omega^2 + 0.11\omega - 1.7 \quad (11a),$$

$$F_{out}(\omega) = \omega + 8.26 \quad (11b).$$

We begin with an initial guess that positions of in-band extrema $\{\Omega_k\}$ are evenly distributed over $[-1, 1]$ rad/s; thus, the initial set is $\Omega = [-1, -2/3, -1/3, 0, 1/3, 2/3, 1]$. In practice, we can arbitrarily choose $N-1$ (where N is the filter order) different extrema within $(-1, 1)$, and our experiments show that the proposed method converges to the correct result. Solving Eq. (6) yields the function $F_{in}(\omega)$ as:

$$F_{in}(\omega) = \omega^6 - 0.028\,5\omega^5 - 1.418\,2\omega^4 + 0.032\,9\omega^3 + 0.442\,5\omega^2 - 0.006\,3\omega - 0.017\,1 \quad (12).$$

By solving $(\partial C(\omega)/\partial\omega) = 0$, we obtain updated positions of in-band extrema: $\Omega' = [-6.595\,2, -1.592\,7, -0.879\,5, -0.444\,8, 0.011\,0, 0.463\,9, 0.893\,3, 1.426\,8]$. Then, the five frequencies lying within the passband are selected and, together with the two band edge frequencies -1 and 1, form the seven extremal points for the next iteration: $\Omega = [-1, -0.879\,5, -0.444\,8, 0.011\,0, 0.463\,9, 0.893\,3, 1]$ rad/s. This process is repeated for 10 iterations, after which the extremal locations converge to stable values. Meanwhile, the coefficients $\{a_i\}$ of $F_{in}(\omega)$ are determined. Fig. 2 shows the results of the first four iterations. It can be found that $C(\omega)$ is stable in the fourth iteration. Then, $F(\omega)$ and ε are calculated using Eqs. (7) and (8), respectively. $E(\omega)$ is solved from the Feldtkeller equation. Finally, $F(\omega)$, $P(\omega)$, and $E(\omega)$ are transformed into $F(s)$, $P(s)$, and $E(s)$. The coefficients of the final polynomials are listed in Table 1, and the filter response is shown in Fig. 3.

A transversal coupling matrix is synthesized from the polynomials. Then, the coupling topology shown in Fig. 4a is obtained through a series of rotation transformations (Table 2). Specifically, Steps 1 – 21 synthesize an "arrow" topology, while Steps 22 – 30 form a quartet among nodes 2 to 5. In

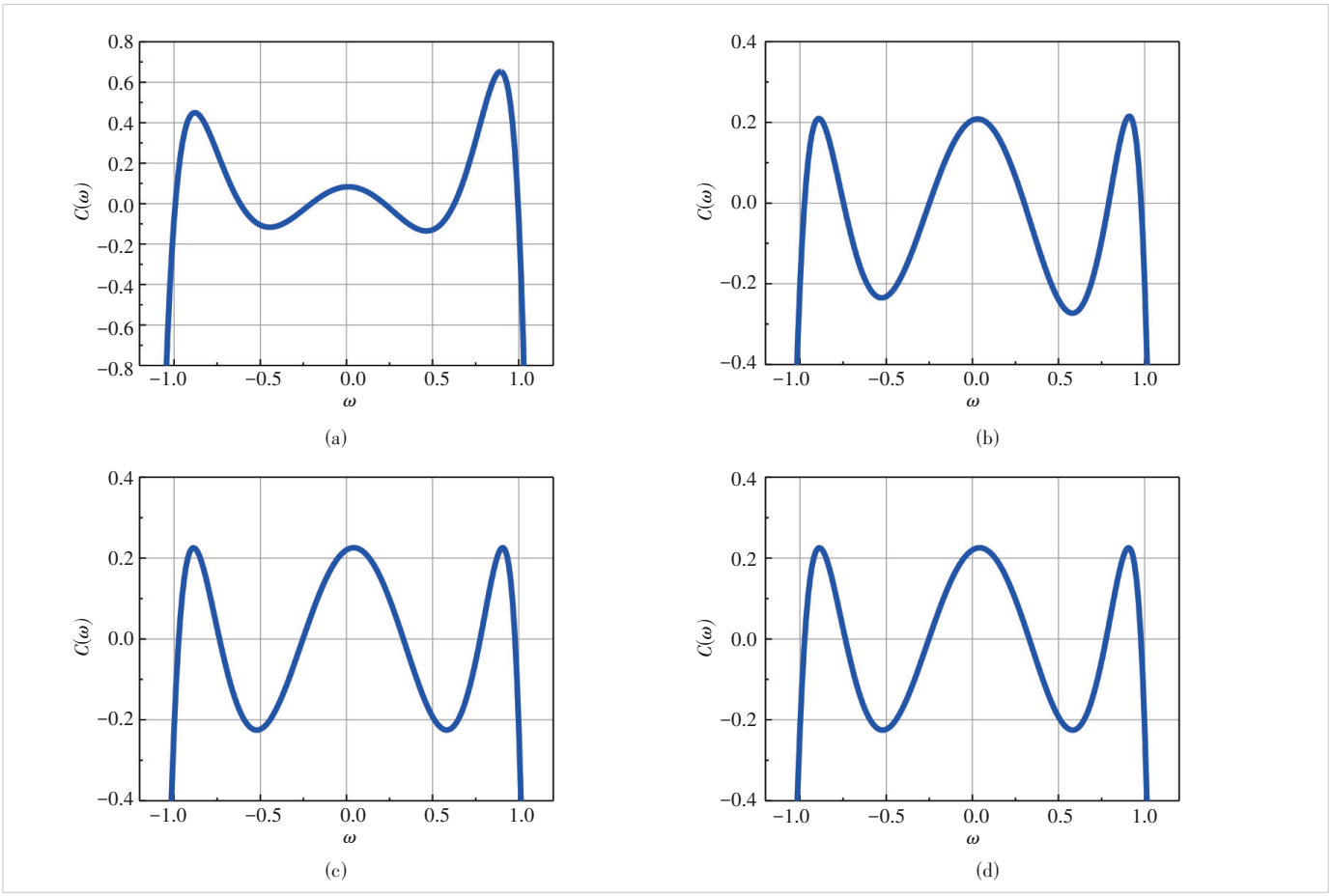


Figure 2. Iterations of Remez-like algorithm in solving the equi-ripple function for the 7th-order filter with an out-of-band RZ: (a) the first iteration; (b) the second iteration; (c) the third iteration; (d) the fourth iteration

Table 1. Coefficients of $F(s)$, $P(s)$, and $E(s)$ of seven-poles with one upper TZ and one lower TZ

s^i, i =	$F(s)$ ($F_{22}(s)$)	$P(s)$	$E(s)$
0	0.373 8j	1.700 0	0.868 3 + 3.731 5j
1	0.422 9	0.110 0j	2.769 1 +12.801 1j
2	5.464 5j	1.000 0	4.956 1 + 24.638 4j
3	2.012 0		6.466 5 + 30.408 0j
4	13.070 5j		5.796 9 + 28.889 2j
5	2.613 9		4.560 8 + 16.045 3j
6	8.137 5j		1.973 3 + 8.137 5j
7	1.000 0		1.000 0
ε_R = 1.000 0		ε = 0.445 8	

TZ: transmission zero

Fig. 4, the numerical values without underscores represent mutual and I/O couplings, whereas those with underscores denote self-couplings. Notably, all mutual coupling values are positive. Therefore, in the physical implementation, all the inter-resonator couplings can be realized as inductive couplings.

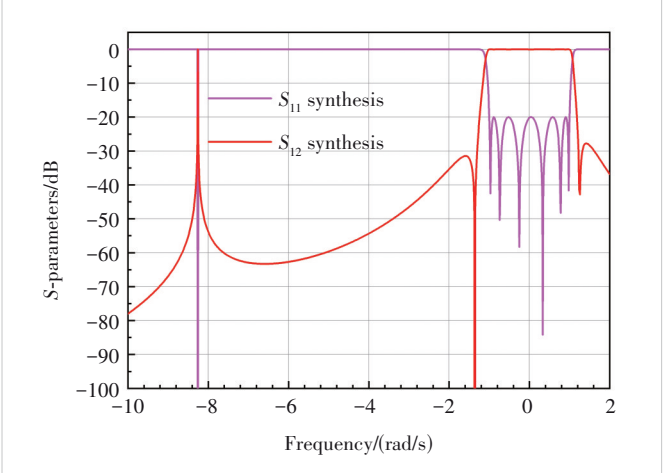


Figure 3. Synthesized response of the 7th-order filter with an out-of-band RZ

2.3 Impact of Out-of-Band RZ

To observe the impact of out-of-band RZ on the filter response, we compare the proposed 7th-order RZ filter (Fig. 4a)

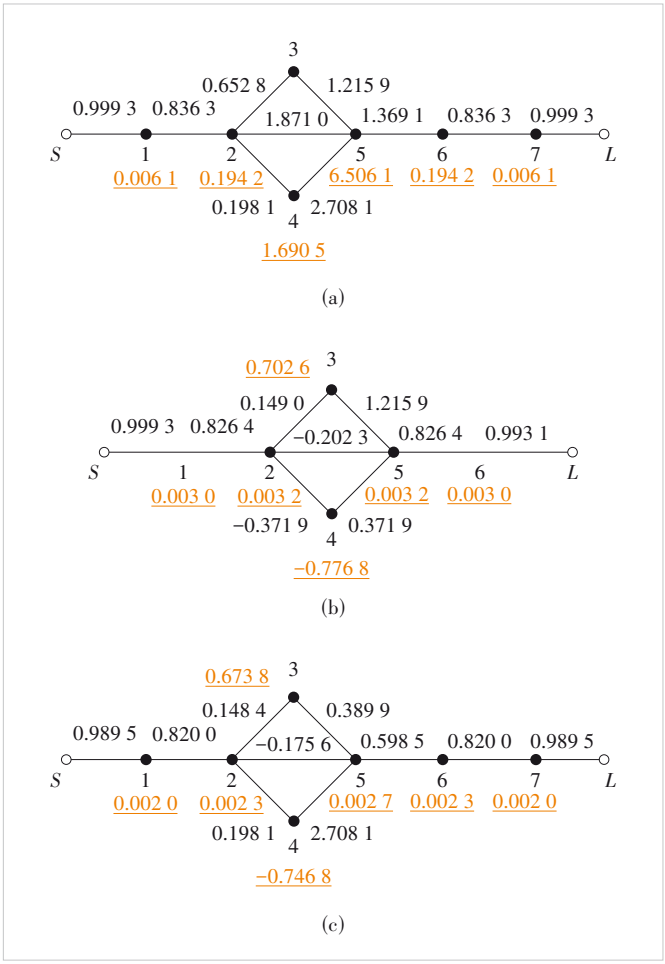


Figure 4. Topology with quartet: (a) 7-2 (7 RZs and 2 TZs) filter with an out-of-band RZ; (b) 6-2 filter; (c) 7-2 filter

Table 2. Rotation sequence to the filter in Fig. 4a

Rotation sequence	Elements to be annihilated	Pivot [i, j]	Rotation sequence	Elements to be annihilated	Pivot [i, j]
1	M_{S7}	[7, 6]	16	M_{37}	[7, 6]
2	M_{S6}	[6, 5]	17	M_{36}	[6, 5]
3	M_{S5}	[5, 4]	18	M_{35}	[5, 4]
4	M_{S4}	[4, 3]	19	M_{47}	[7, 6]
5	M_{S3}	[3, 2]	20	M_{46}	[6, 5]
6	M_{S2}	[2, 1]	21	M_{57}	[7, 6]
7	M_{17}	[7, 6]	22	M_{5L}	[5, 6]
8	M_{16}	[6, 5]	23	M_{6L}	[6, 7]
9	M_{15}	[5, 4]	24	M_{46}	[5, 6]
10	M_{14}	[4, 3]	25	M_{47}	[4, 5]
11	M_{13}	[3, 2]	26	M_{57}	[5, 6]
12	M_{27}	[7, 6]	27	M_{35}	[4, 5]
13	M_{26}	[6, 5]	28	M_{36}	[3, 4]
14	M_{25}	[5, 4]	29	M_{46}	[4, 5]
15	M_{24}	[4, 3]	30	M_{34}	[3, 4]

with two traditional counterparts: a 6th-order filter and a 7th-order filter (Figs. 4b and 4c). The TZs for all three filters are set at $\{s_{TZ}\} = \{-1.36j, 1.25j\}$, with a specified return loss of 20 dB. The key difference lies in their synthesis procedures and the resulting coupling matrices. In the traditional filters, the absolute values of coupling coefficients are all less than 1. In contrast, the proposed filter with an out-of-band RZ has couplings greater than 1. Couplings with normalized coupling coefficients exceeding 1 are defined as strong couplings. The quartet possessing such strong couplings constitutes a GSCRQ. Realizing strong coupling between resonant rods requires reducing the distance between adjacent resonant rods. This close arrangement is beneficial for the miniaturization of coaxial combline filters.

The introduction of out-of-band RZ affects not only the coupling coefficients but also the filter's response. Fig. 5 compares the responses of the three filters in Fig. 4. The 7th-order GSCRQ filter with an out-of-band RZ exhibits in-band characteristics similar to the conventional 6th-order filter, but its lower stopband rejection is worse than that of the 6th-order filter. In addition, its upper stopband rejection is slightly better than that of the 6th-order filter but remains worse than that of the seventh-order filter.

The above analysis confirms that an out-of-band RZ affects both the coupling coefficients and filter characteristics. This introduces additional flexibility into the filter synthesis. To demonstrate this, the location of the out-of-band RZ in the 7th-order filteris varied to −3.5 rad/s or −6.7 rad/s. Fig. 6 compares the frequence responses for the original RZ at −8.26 rad/s, −3.5 rad/s, and −6.7 rad/s with these two new locations. The corresponding coupling matrices are given in Fig. 7. Fig. 6 shows that when the out-of-band RZ is closer to the passband, the out-of-band suppression in the near passband will be

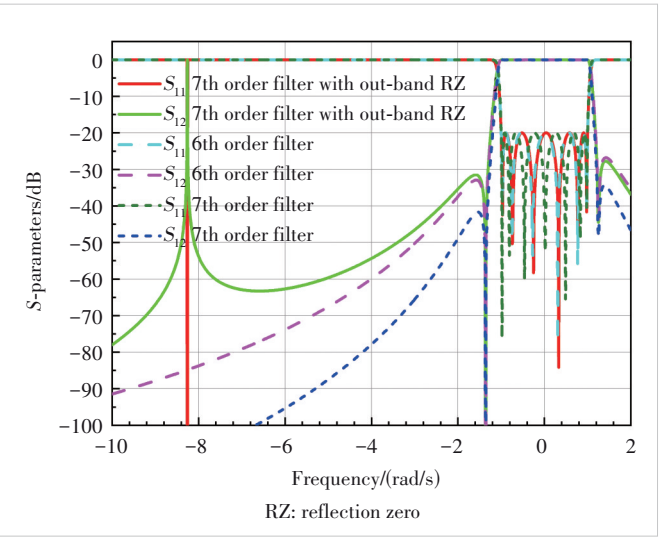


Figure 5. Frequency responses of three filter designs: the 7th-order filter with an out-of-band reflection zero (solid), 6th-order traditional filter (dashed), and 7th-order traditional filter (dotted)

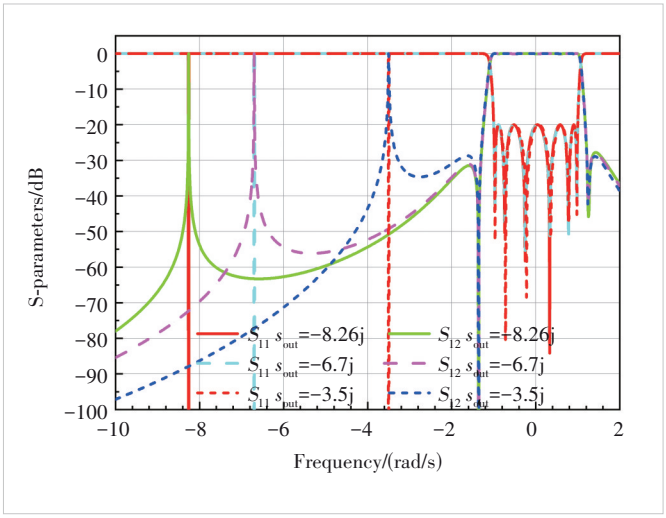


Figure 6. Frequency responses with different out-of-band RZ locations: s_{out} = −8.26j (solid), s_{out} = −6.7j (dashed), and s_{out} = −3.5j (dotted)

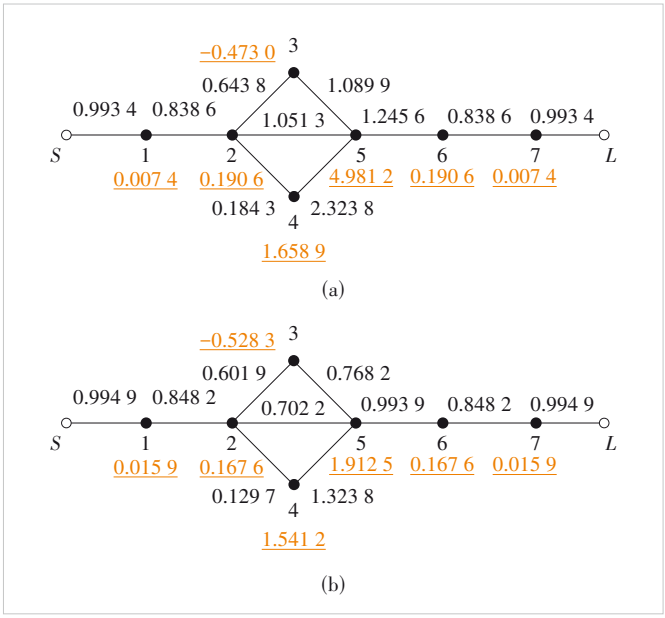


Figure 7. 7-2 filter topology with different out-of-band RZ locations: (a) s_{out} = −6.7j; (b) s_{out} = −3.5j

worse, and the couplings in the quartet become weaker. Therefore, in the actual filter design, we can tune the position of the out-of-band RZ to meet the design specifications. However, it should be noted that placing the out-of-band RZ too far from the passband results in excessively strong mutual coupling in the GSCRQ, making it difficult to realize in practice.

3 Design Example

To validate the synthesis method, a 6th-order bandpass filter using coaxial resonators is designed. Its physical layout and the coupling topology are shown in Figs. 8a and 8b, respectively. The key design specifications include a center frequency f_0 = 2.5 GHz and a bandwidth BW = 200 MHz.

In previous literature, the design of strongly coupled resonator filters is often performed using the "port tuning" technique[11]. This field-circuit co-simulation method uses circuit optimization to tune the filter, which typically requires only a few seconds. In addition, due to dispersion and capacitive loading effects of the resonant rods, a large difference exists between the out-of-band resonance of the simulated response and the ideal response. To make the ideal response consistent with the simulated response, it is generally necessary to optimize a target matrix that takes stray couplings into account. For strongly coupled resonator filters, we propose a novel Model-Based Vector Fitting (MVF)-based method to achieve tuning of strongly coupled filters[12 − 13]. The procedure includes the following five steps:

Step 1: Map the physical frequency to the low-pass domain:

$$\omega = \frac{f_0}{\mathrm{BW}}\left(\frac{f}{f_0} - \frac{f_0}{f}\right) \tag{13},$$

where f_0 and BW are the center frequency and bandwidth of the filter, respectively.

Step 2: De-embed the phase from the simulated *S*-parameters by vector fitting[11].

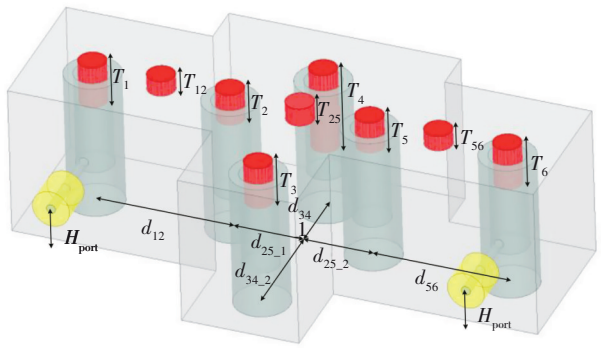
Step 3: Transform phase-corrected *S*-parameter to *Y*-parameter, and use MVF to approximate the *Y*-parameter data[10].

Step 4: Synthesize a transversal coupling matrix from the *Y*-parameter rational functions and transform it to the target form.

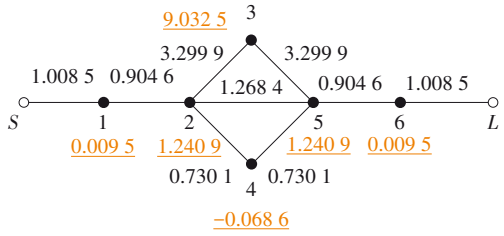
Step 5: Compare extracted and ideal synthesized matrices, and adjust the dimensions of the filter accordingly.

The second step is crucial for identifying the true poles and zeros of the coupled resonator filter. In a traditional single-passband filter, plotting the poles and zeros of the reflection coefficient on the complex s-plane reveals that most of the poles and zeros cluster around the origin, whereas a few are located far away. Those near the origin are contributed by the resonators, while the distant ones result from phase loading and feed lines[13]. Empirically, for traditional single-passband filters, poles and zeros located beyond four units from the origin are caused by phase loading and feed lines. However, this rule does not apply to strongly coupled resonator filters. Let us take the simulated reponses of the EM model in Fig. 8 as an example. Vector fitting (VF) is applied to S_{11} to obtain an optimal rational approximation. The zeros and poles of the fitting rational functions are listed in Table 3.

If zeros and poles more than four units from the origin (i.e., the 6th, 7th, and 8th in Table 3) are selected to construct the phase factor, a spike appears in its amplitude near −11.5 rad/s, as shown in Fig. 9a. This indicates erroneous phase de-embedding, as the amplitude and phase of the phase factor should be smooth. Corresponding errors are also observed in the de-embedded phase of S_{11} or S_{22}. As shown in Fig. 9b, the



(a)



Note: The numerical values with underscores are self-couplings, while the numerical values without underscores are mutual couplings and I/O couplings.

(b)

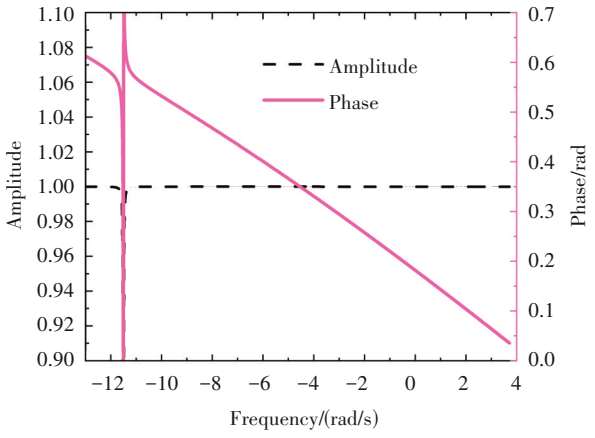
Figure 8. (a) Electromagnetic model of the 6th-order filter with GSCRQ; (b) synthesis results of GSCRQ of the 6-2 filter with one out-of-band RZ

Table 3. Zeros and poles of VF results for S_{11}

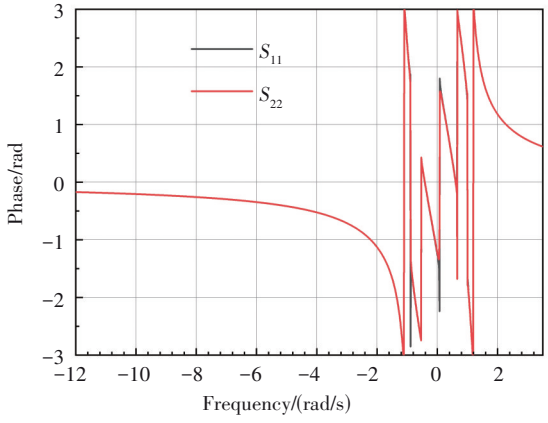
k	poles	zeros
1	−0.692 7 + 0.104 7i	0.001 0 + 0.072 2i
2	−0.556 11−0.675 5i	−0.000 1−0.530 2i
3	−0.481 2 + 0.805 7i	−0.000 5 + 0.657 6i
4	−0.180 53−1.073 5i	−0.000 7−0.889 7i
5	−0.156 41 + 1.149 1i	0.000 3 + 1.000 2i
6	−0.009 1−11.503 2i	0.000 0−11.503 1i
7	1.132 2 + 14.087 9i	1.664 5 + 13.927 9i
8	−26.468 1 + 3.616 4i	26.542 8 + 3.463 7i

phases of S_{11} and S_{22} are very smooth in the lower stopband, which is correct for traditional filters. However, for a strongly coupled resonator filter with an out-of-band RZ, the phases of S_{11} and S_{22} should exhibit abrupt changes, not smoothness. Therefore, we speculate that the wrong phase factor offsets the original mutation, making the phases of S_{11} and S_{22} smooth after phase de-embedding.

To obtain the correct phase factor, the 6th pole-zero pair is attributed to the filter itself rather than to phase loading and transmission lines, yielding the phase factor shown in Fig. 10a. The phase no longer changes suddenly around − 11.5 rad/s,



(a)



(b)

Figure 9. Incorrect phasing de-embedding: (a) amplitude and phase of the phase factor α; (b) S_{11} and S_{22} after phase removal

which proves that this phase factor is physically consistent. In addition, Fig. 10 shows the S-parameter phase after phase de-embedding, where the out-of-band RZ now correctly appears at − 11.5 rad/s. Based on extensive experiments, we recommend that for strongly coupled filters, the threshold for phase de-embedding is set to

$$\text{Thresh} = \min(s_{\text{out}}) - 1 \quad (14),$$

where $\min(s_{\text{out}})$ is the minimum out-of-band RZ.

After phase de-embedding, MVF is used to fit a 7th-order polynomial of the Y-parameters (Step 3). A transversal coupling matrix is then synthesized and transformed to the target form (Step 4). By comparing the extracted and ideal synthesized matrices, the filter dimensions are adjusted accordingly (Step 5), enabling rapid tuning to meet the target specifications. The simulation results with ideal lossless materials are shown in Fig. 11, where dashed lines are simulation data, and solid lines are the ideal synthesis responses. The input

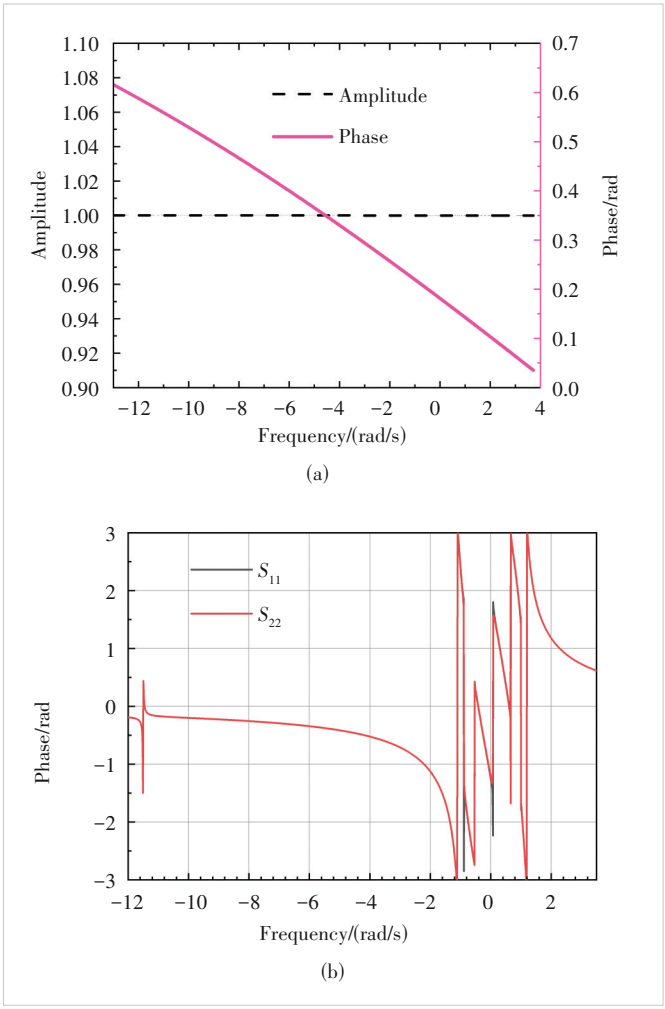


Figure 10. Correct phase de-embedding: (a) amplitude and phase of the phase factor α and (b) S_{11} and S_{22} after phase removal

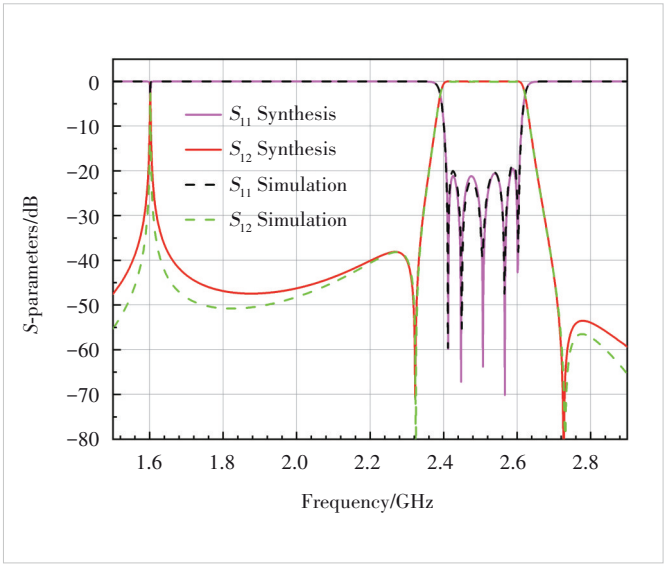


Figure 11. Simulated results using High Frequency Structure Simulator (HFSS) and ideal synthesis responses of the filter

feed comes from a coaxial cable whose inner and outer radii are 0.5 mm and 2 mm, respectively. The coaxial cable is filled with a dielectric with a relative permittivity ε_R = 2.2. The inner and outer radii of the resonant rods are 2 mm and 3 mm and the height of the enclosure box is 16.5 mm. Other detailed dimensions are shown in Table 4.

Table 4. Dimensions of the filter in Fig. 8

Resonator ID	T_1	T_2	T_3	T_4	T_5	T_6	T_{12}	T_{25}	T_{56}
Final/mm	4.70	3.55	9.10	4.7	3.55	4.70	1.70	2.00	2.00
Dimension	d_{12}	d_{25_1}	d_{25_2}	d_{34_1}	d_{34_2}	H_{port}			
Final/mm	1.77	7.55	7.55	6.40	11.43	6.32			

4 Conclusions

In this article, synthesis and design of filters with GSCRQs are presented. The GSCRQ enables the generation of two TZs without requiring negative couplings. Its work mechanism is analyzed based on a circuit model for the first time. A complete filter design procedure is demonstrated, including the coupling matrix synthesis and EM model tuning. The full-wave EM simulation results show excellent agreement with the theoretical predictions, thereby validating the theory of filters with GSCRQ.

References

[1] Levy R. Direct synthesis of cascaded quadruplet (CQ) filters [C]//Proc. 1995 IEEE MTT-S International Microwave Symposium. 1995: 497 - 500. DOI:10.1109/MWSYM.1995.406038
[2] Levy R, Petre P. Design of CT and CQ filters using approximation and optimization [J]. IEEE transactions on microwave theory and techniques, 2001, 49(12): 2350 - 2356. DOI:10.1109/22.971620
[3] Tamiazzo S, Macchiarella G. An analytical technique for the synthesis of cascaded N-tuplets cross-coupled resonators microwave filters using matrix rotations [J]. IEEE transactions on microwave theory and techniques, 2005, 53(5): 1693 - 1698. DOI:10.1109/TMTT.2005.847065
[4] Macchiarella G, Bastioli S, Snyder R V. Design of in-line filters with transmission zeros using strongly coupled resonators pairs [J]. IEEE transactions on microwave theory and techniques, 2018, 66(8): 3836 - 3846. DOI: 10.1109/TMTT.2018.2840981
[5] Bastioli S, Snyder R V, Macchiarella G. Design of in-line filters with strongly coupled resonator triplet [J]. IEEE transactions on microwave theory and techniques, 2018, 66(12): 5585 - 5592. DOI: 10.1109/TMTT.2018.2867004
[6] Zeng Y, Yang Y M, Yu M, et al. Synthesis of generalized strongly coupled resonator triplet filters by regulating redundant resonant modes [J]. IEEE transactions on microwave theory and techniques, 2022, 70(1): 864 - 875. DOI:10.1109/TMTT.2021.3128602
[7] Bastioli S, Snyder R V, Macchiarella G. The strongly coupled resonator quadruplet [J]. IEEE microwave and wireless technology letters, 2023, 33 (8): 1135 - 1138. DOI:10.1109/LMWT.2023.3270099
[8] Otto S, Lauer A, Kassner J, et al. Full wave coupled resonator filter optimization using a multi-port admittance-matrix [C]//2006 Asia-Pacific Microwave Conference. IEEE, 2006: 777 - 780. DOI:10.1109/APMC.2006.4429530
[9] Zhao P, Liu B Z, Oldoni M, et al. Improving accuracy in solving Feldtkeller equation [J]. IEEE microwave and wireless technology letters, 2024,

34(3): 251 – 254. DOI:10.1109/LMWT.2023.3349133
[10] Cameron R J, Kudsia C M, Mansour R R. Microwave filters for communication systems: fundamentals, design and applications [M]. 2nd ed. Hoboken: Wiley, 2018. DOI: 10.1002/9781119292371
[11] Seyfert F, Baratchart L, Marmorat J P, et al. Extraction of coupling parameters for microwave filters: determination of a stable rational model from scattering data [C]//IEEE MTT-S International Microwave Symposium Digest. IEEE, 2003: 25 – 28. DOI: 10.1109/MWSYM.2003.1210875
[12] Zhao P, Wu K L. Model-based vector-fitting method for circuit model extraction of coupled-resonator diplexers [J]. IEEE transactions on microwave theory and techniques, 2016, 64(6): 1787 – 1797. DOI: 10.1109/TMTT.2016.2558639
[13] Zhao P. Phase De-embedding of narrowband coupled-resonator networks by vector fitting [J]. IEEE transactions on microwave theory and techniques, 2023, 71(4): 1439 – 1446. DOI:10.1109/TMTT.2018.2854170

Biographies

Xiong Zhi'ang received his BS degree from Jiangxi University of Science and Technology, China in 2021, and the master's degree from Xidian University, Xi'an, China in 2024. His current research interests include mixed electric and magnetic coupling filters and modeling and optimization of microwave passive filters.

Fan Jiyuan received his BS degree from Nanjing University of Posts and Telecommunications, China in 2022 and his master's degree from Xidian University, China in 2025. His current research interests include mixed electric and magnetic coupling, as well as the synthesis and tuning of multiband filters and multiplexers.

Zhao Ping (aoing56@gmail.com) received his BS degree from Nanjing University, China in 2012, and PhD degree from The Chinese University of Hong Kong, China in 2017. From 2017 to 2019, he was a post-doctoral researcher with the École Polytechnique de Montréal, Canada. He joined the National Key Laboratory of Antennas and Microwave Technology, Xidian University, China in 2020. Since 2024, he has been with the State Key Laboratory of Electromechanical Integrated Manufacturing of High-performance Electronic Equipments, Xidian University, where he is currently an associate professor. His research interests include coupling matrix synthesis techniques for coupled-resonator networks, analytical computer-aided tuning (CAT) algorithms for microwave and millimeter-wave filters, and diplexers with applications in cellular base stations and satellites. He is also interested in modeling and optimization of passive RF components and computer-aided design techniques, such as the homotopy method, artificial neural networks, and machine learning techniques.

Zhou Jinzhu received his PhD degree from Xidian University, China in 2011. He is currently a professor with the State Key Laboratory of Electromechanical Integrated Manufacturing of High-performance Electronic Equipments, Xidian University. He is also the Director of the Department of Electronic Packaging and the Deputy Director of the Xi'an Key Laboratory of Intelligent Instrument and Packaging Test, Xidian University. He received the Outstanding Youth of Shaanxi Province in 2022. His research interests include microwave filter tuning, RF microsystem, smart skin antenna, and machine learning. He has published over 60 papers and holds 40 patents issued. Dr. Zhou received the National Science and Technology Award (2021 First Prize), the Science and Technology Progress Award of China Electronics Society (2022 First Prize), the Shaanxi Province Science and Technology Award (2015 First Prize), and the National Defense Science and Technology Award (2019 Second Prize). He also received four Best Paper Awards for the research of smart skin antenna and automatic microwave filter tuning in the International Conference of the 2010 IEEE International Conference on Mechatronics and Automation (ICMA), the Asia International Symposium on Mechatronics (AISM 2015), the AISM 2017, and the 2021 IEEE International Conference on Electronic Packaging Technology (ICEPT), respectively.

Shen Nan received his BS degree from Northwestern Polytechnical University, China in 2004 and MS degree in electromagnetic field and microwave technology from the same university in 2007. He is currently with ZTE Corporation, where his research focuses on filters, multiplexers, and antennas.

Wu Qingqiang received his BS degree from Xidian University, China in 2019, and MS degree in electromagnetic wave and microwave technology from the National Key Laboratory of Antennas and Microwave Technology, Xidian University in 2021. He is currently with ZTE Corporation, where his research focuses on filters and multiplexers.

New Member of ZTE Communications Editorial Board

Guo Yike is the Provost of the Hong Kong University of Science and Technology (HKUST), China; a Chair Professor in both the Department of Computer Science and Engineering and the Department of Electronic and Computer Engineering at HKUST; and also serves as the Director of the Hong Kong Generative AI Research and Development Center, China. He is a world-renowned computer scientist who has led several large-scale AI and data science research projects in China (including both Mainland and Hong Kong), the UK and other European countries. He was the Founding Director of the Data Science Institute at Imperial College London, UK, one of its seven Global Institutes, as well as the Vice President (Research and Development) at Hong Kong Baptist University, China. He is a Fellow of the Royal Academy of Engineering (FREng), a member of Academia Europaea (MAE), a Fellow of the Hong Kong Academy of Engineering Sciences (FHKEng), a Fellow of the Institute of Electrical and Electronics Engineers (FIEEE), a Fellow of the British Computer Society (FBCS), and a Fellow of Chinese Association for Artificial Intelligence (FCAAI).

Professor Guo was awarded the Outstanding Contribution Award of the 2022 Wu Wenjun Artificial Intelligence Science and Technology Award, China, which is considered to be the highest award for Chinese AI science and technology. In June 2025, Professor Guo was named "Leader of the Year 2024" in the Education/Professional/Technology and Innovation category from Sing Tao News Corporation, Hong Kong, China, for his transformative contributions to education and technological innovation in Hong Kong. In November of the same year, Professor Guo was elected as an international member of the Chinese Academy of Engineering (CAE) in recognition of his exceptional contributions to the Information and Electronics Engineering Division. The CAE is the nation's highest academic institution in engineering science and technology, and membership is considered the highest lifelong academic honor in the field.

Modern Graphics APIs: Design Principles, A Use Case, and New Perspectives

Lu Ping[1,2], Sun Qi[3], Wang Chen[3], Guo Jie[3], Guo Yanwen[3], Shi Wenzhe[1,2]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. Nanjing University, Nanjing 210093, China)

DOI: 10.12142/ZTECOM.202601013

https://kns.cnki.net/kcms/detail/34.1294.TN.20260228.1556.004.html, published online February 28, 2026

Manuscript received: 2024-06-26

Abstract: In this paper, we provide a comprehensive examination of the evolution of graphics Application Programming Interfaces (APIs). We begin by exploring traditional graphics APIs, elucidating their distinct features and inherent challenges. This sets the stage for a detailed exploration of modern graphics APIs, with a focus on four critical design principles. These principles are further analyzed through specific case studies and categorical examinations. The paper then introduces MoerEngine, a bespoke rendering engine, as a practical case to demonstrate the real-world application of these modern principles in software engineering. In conclusion, the study offers insights into the potential future trajectory of graphics APIs, spotlighting emerging design patterns and technological innovations. It also ventures to predict the development trends and capabilities of next-generation graphics APIs.

Keywords: graphics API; rendering; design principle; MoerEngine

Citation (Format 1): Lu P, Sun Q, Wang C, et al. Modern graphics APIs: design principles, a use case, and new perspectives [J]. ZTE Communications, 2026, 24(1): 97 - 106. DOI: 10.12142/ZTECOM.202601013

Citation (Format 2): P. Lu, Q. Sun, C. Wang, et al., "Modern graphics APIs: design principles, a use case, and new perspectives," *ZTE Communications*, vol. 24, no. 1, pp. 97 - 106, Mar. 2026. doi: 10.12142/ZTECOM.202601013.

1 Introduction

Graphics Application Programming Interfaces (APIs) serve as essential toolkits in graphics rendering, offering programming instructions and functions crucial for rendering 3D scenes into images and presenting frames. They bridge software with graphics hardware, enhancing the productivity of rendering and visualization. Historically, APIs like OpenGL[1] and Direct3D[2], built on a state machine model, dominated the scene but struggled with complicated rendering tasks. In contrast, modern APIs such as Vulkan[3], Direct3D 12[4], and Metal[5], are gaining traction due to their high performance and hardware optimization[6], leading to improvements in rendering efficiency and program parallelism. This shift, particularly impactful in gaming and high-performance computing, involves many explorations into general-purpose computing architectures[7 - 8], as advancements in hardware capabilities and rendering complexity push the boundaries of graphics API design and optimization.

This paper delves into the evolution and refinement of modern graphics APIs, tracing their development from early APIs like Direct3D and OpenGL. It demonstrates the limitations of traditional APIs in terms of four aspects: implicit pipeline, complicated driver layer, runtime shader compilation, and implicit task submission. The discussion then pivots to modern graphics APIs, exploring their design based on four key principles: low-level access, abstraction, separation, and parallelism. Utilizing MoerEngine, a self-developed high-performance real-time rendering engine, the paper illustrates the application of modern API design principles in real-world projects. It presents practical examples of modern API implementation in areas such as render hardware interface (RHI), render dependency graph (RDG)[9], and shader compilation. The conclusion forecasts the future trajectory of graphics APIs, exploring potential developments in low-level functionality, cross-platform compatibility, ray tracing, augmented and mixed reality, and neural network integration.

The aim of this paper is to provide a holistic view of how modern graphics APIs are revolutionizing traditional graphics rendering approaches and to speculate on their future roles and advancements in the graphics processing landscape.

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20230921014.

2 Limitations of Traditional Graphics APIs

The design principles of traditional graphics APIs are compatibility and ease of use. These features facilitated the rapid spread and development of traditional graphics APIs in their early stages. However, as hardware performance gradually improves and graphics tasks become more complicated, these design principles lead to numerous efficiency-related issues. This section briefly elaborates on these limitations.

2.1 Implicit Pipeline

Traditional graphics APIs use an implicit pipeline to ensure compatibility and ease of use. They expose only the interfaces for operating the graphics pipeline to software developers, making the entire graphics pipeline transparent to the application layer. This approach was developer-friendly in the early stages, as developers did not need to manage the complex graphics pipeline and could focus solely on graphics programming, objectively promoting the popularization of graphics programming.

However, with the development of GPUs and the increasing complexity of rendering tasks, this implicit pipeline has become a performance bottleneck. Developers are unable to manage the state of the graphics pipeline directly and therefore cannot perform optimization based on its current state. Moreover, they fail to apply advanced design patterns to program architecture design.

2.2 Complicated Driver Layer

The driver layer of traditional graphics APIs is overly bulky and complex. These APIs and their pipelines, exposed to software developers through a state machine model, necessitate the handling of complex tasks at the hardware driver layer. As graphics technology evolved and graphics pipelines and algorithms became more complex, the driver layer of traditional graphics APIs grew increasingly cumbersome (Fig. 1). Specifically, traditional graphics APIs embed state verification within the driver software. During application runtime, these drivers maintain and track all states while performing various validation tasks, such as confirming API markers and ensuring resource data legality.

Runtime state verification introduces additional CPU overhead and fails to maximize the use of the GPU's parallel capabilities, thereby becoming a primary bottleneck that reduces rendering performance.

2.3 Runtime Shader Compilation

Traditional graphics APIs do not provide a profound shader pre-compilation mechanism. In most traditional graphics APIs, shaders need to be compiled at runtime, with the compilation being handled by the driver (Fig. 1). Whether translating from OpenGL Shading Language (GLSL) or High-Level Shading Language (HLSL) to machine code, the process is inefficient, leading to significant shader compilation overhead. Direct3D offers DirectX Bytecode (DXBC) as precompiled code, which improves the efficiency of translating DXBC to machine code and thus enhances the shader loading speed in Direct3D. However, this also introduces the compilation overhead of DXBC. OpenGL/ES, on the other hand, leaves shader compilation entirely to the driver and does not offer a precompilation mechanism. Even though this feature can be enabled through extensions, it cannot be universally supported across all platforms[10].

Modern graphics tasks often require a variety of shaders. On traditional APIs, both the compilation and switching of shaders incur more significant performance overhead than before, which can no longer meet the requirements for high-

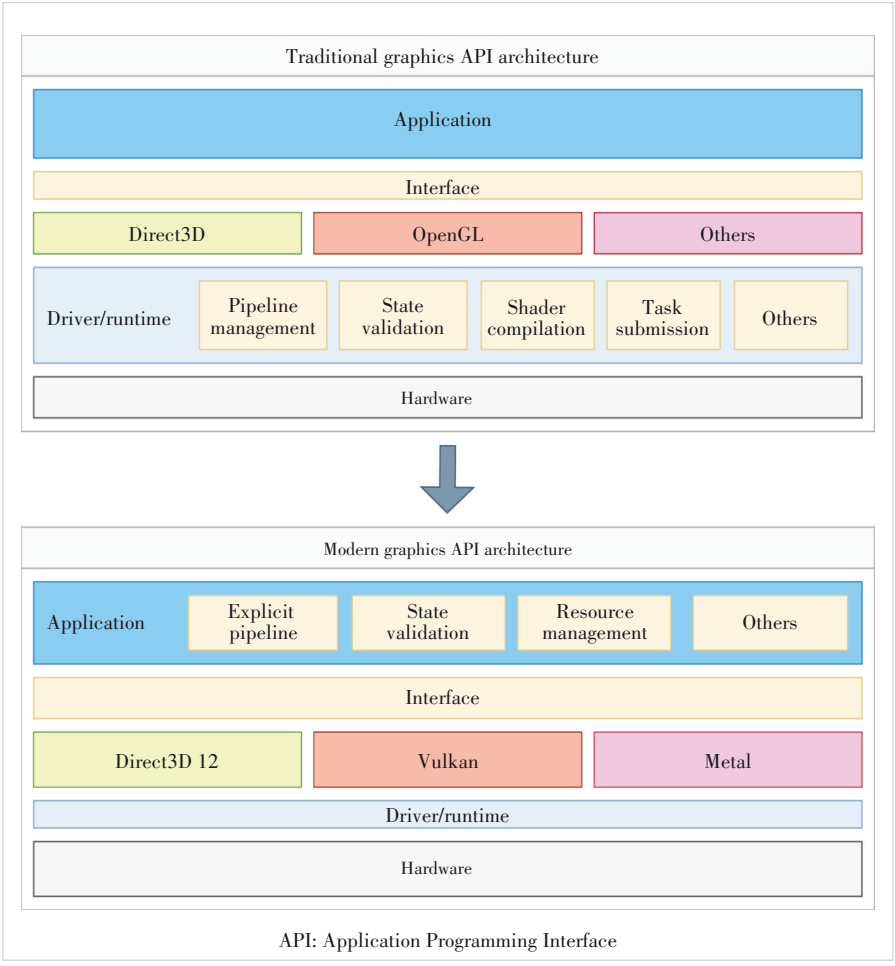


Figure 1. Comparison between traditional and modern graphics APIs

performance real-time rendering.

2.4 Implicit Task Submission

Traditional graphics APIs employ an implicit task submission model. The submission of rendering instructions is executed through an API via an application-layer-transparent runtime layer or driver (Fig. 1). When the internally maintained rendering instruction queue becomes full or when a switch in the graphics pipeline occurs, the internal runtime system or driver automatically submits the current rendering work and waits for the GPU to complete it. This significantly disrupts GPU parallelism, leading to noticeable performance degradation. Moreover, such implicit context management makes it difficult for applications to optimize synchronization between the CPU and GPU.

As hardware evolves and new rendering technologies emerge, traditional APIs are also evolving. However, in the pursuit of maintaining compatibility and ease of use, their interfaces are becoming increasingly complicated. In traditional graphics APIs, the driver manages most tasks, leading to vastly different implementations among various hardware manufacturers. The difficulty in achieving universality with different extensions is growing, as a single functionality might have multiple implementations. This results in an increasingly thick driver layer, making driver maintenance more challenging and the use of traditional APIs more complex for developers. These issues have compelled the industry to develop a new set of graphics APIs suited for modern graphics hardware and capable of meeting diverse rendering requirements, namely the modern graphics APIs.

3 Design Principles of Modern Graphics APIs

Modern graphics APIs feature a more explicit graphics pipeline, resource management strategy and synchronization mechanisms. They possess an extremely lightweight driver layer (Fig. 1), and the API itself is very close to the hardware interface, granting developers significant freedom. This enables more diverse and efficient coding design patterns. The various characteristics of modern graphics APIs have led to a leap in rendering performance.

We summarize the design principles of modern graphics APIs from various perspectives.

3.1 Low-Level Design

The most direct and crucial design principle of modern graphics APIs is their low-level approach. The design of these APIs closely mirrors hardware behavior, transferring a substantial amount of the work traditionally done at the driver layer to the application layer. Consequently, it becomes the responsibility of the application layer to manage most tasks of the modern graphics API (Fig. 1).

3.1.1 Explicit Pipeline

Modern graphics APIs utilize pipeline state objects (PSOs) to explicitly describe and manage the graphics pipeline. Aspects such as resource requirements, usage scenarios, and synchronization are all explicitly defined at the application layer, without the need for runtime-layer or driver involvement. Developers gain full control for performance optimization. If the shift from fixed to programmable pipelines marked the first revolution in graphics, the transition from implicit to explicit pipelines represents the second. Explicit pipelines offer a much larger space for performance optimization.

3.1.2 State Verification

Modern graphics APIs delegate state verification from the driver layer to the application layer. The application layer predefines the usage and format of resources, ensuring that the pipeline synchronizes properly before using resources and promptly releases resources that are no longer needed. These practices are beneficial for the driver layer and hardware to optimize rendering tasks.

3.1.3 Resource Management

Modern graphics APIs employ explicit video memory (VRAM) management. For instance, the Direct3D 12 API provides different types of heaps, allowing the application layer to allocate or release resources to these heaps. The Vulkan API defines resource types and usages, with the application layer determining the required space size and explicitly allocating or releasing various data types. The Metal API is similar to the Direct3D 12 API, where the operations at the application layer are explicit and performed with appropriately granular control.

3.2 Abstraction

Another major design principle of modern graphics APIs is abstraction. These APIs abstract resources and operations that are close to hardware, providing the application layer with a manageable space. This approach ensures the efficiency of the underlying layer while offering a more intuitive usage method. Specifically, abstraction mainly involves two aspects: resources and pipelines.

3.2.1 Resource Abstraction

Modern graphics APIs abstract resource types, storage, and access. Specifically, they classify resources based on access methods and other criteria. In terms of resource storage and access, modern APIs actively adopt an indirect addressing model. They use logical structures, as illustrated in Fig. 2, to describe the arrangement of physical data and represent resources through abstract descriptors, such as index tables. This approach helps to streamline the management and utilization of resources in graphics applications.

For instance, the Direct3D 12 API uses different heaps to store resources and employs Views and Descriptors for resource access. It also provides Root Signatures and Descriptor

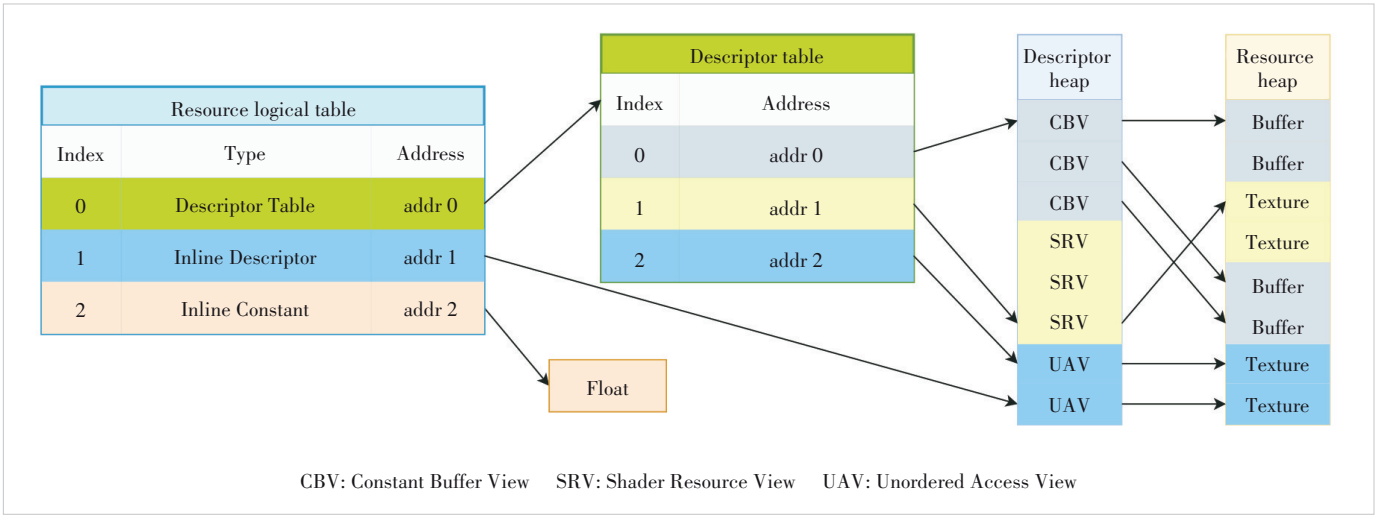


Figure 2. An indirect addressing example of modern graphics APIs

Tables to organize resources. The Vulkan API similarly abstracts resource types and access, using Pipeline Layouts and Descriptor Sets to organize resources. While the Metal API does not abstract resource access, it still follows the design principle of abstraction, organizing resources as efficiently as possible.

The resource abstraction in modern graphics APIs not only provides a flexible resource binding mechanism but also offers performance gains. For example, when switching between different pipelines, only the logical structure of the resources needs to be changed, allowing resources to be quickly adapted.

3.2.2 Pipeline Abstraction

Modern graphics APIs abstract the graphics pipeline. The pipeline is explicitly present and assembled from compiled shader modules, shader resources, and pipeline states (Fig. 3). The pipeline is created before runtime, allowing for preverification of the GPU state to ensure correctness, thereby eliminating the need for runtime checks by the driver or runtime system. The major advantage of this pipeline abstraction is speed, as it avoids runtime state checks and allows for very rapid pipeline switching.

3.3 Separation

A crucial design principle of modern graphics APIs is the separation of construction and execution. Pipeline construction, shader compilation, and resource creation are all completed prior to runtime. During runtime, only instructions are processed without touching the data. This separation enhances efficiency and performance, as it reduces the workload during runtime, allowing the GPU to focus solely on rendering tasks.

3.3.1 Pipeline Separation

The pipeline isolation in modern graphics APIs is reflected in the immutability of the pipeline once it is created. After a pipeline is fully constructed, it cannot be modified; any changes to the pipeline's properties necessitate the creation of a new pipeline. This immutability is advantageous for the runtime system, as it can fully trust the legality of the pipeline and execute it directly without the need for state checks. However, pipelines expose certain easily changeable parameters.

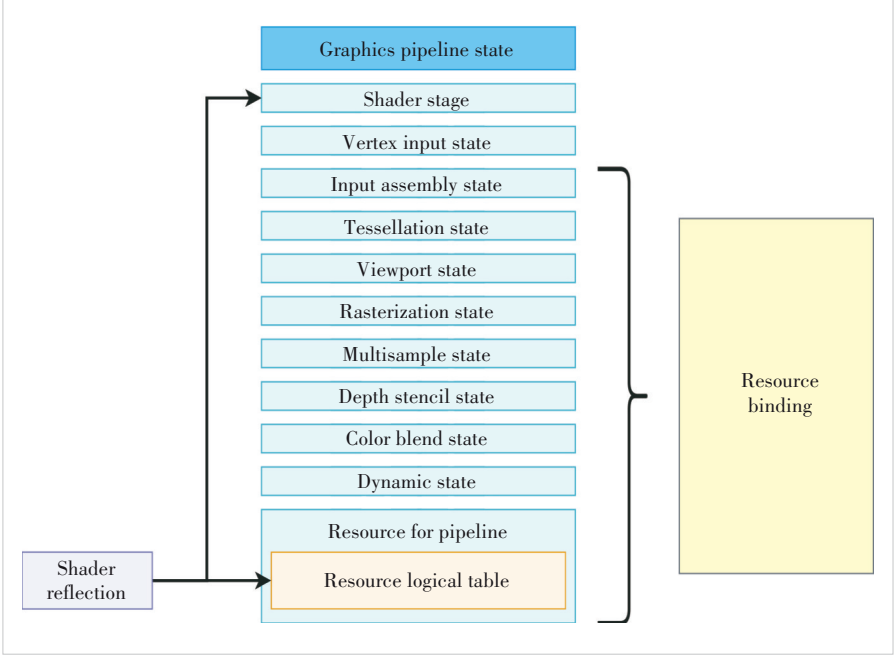


Figure 3. Pipeline layout of modern graphics APIs

For example, in the Vulkan API, the graphics pipeline exposes properties like the viewport and scissor. Similarly, in the Direct3D 12 API and Metal API, various types of pipelines also expose certain properties.

3.3.2 Shader Separation

Shader isolation in modern graphics APIs is exemplified by the shader pre-compilation mechanism and pipeline assembly mechanism. Modern graphics APIs provide shader compilers for precompiling shaders into bytecode, which offers significant performance advantages when loading and assembling pipelines. Additionally, the runtime system does not need to concern itself with shader loading and management, as these tasks are already completed during pipeline assembly. Common shader compilers, such as GLSLang Compiler and DirectX Shader Compiler, offer comprehensive support for various shading languages including GLSL, HLSL, and Slang[11]. This approach streamlines the integration of shaders into the graphics pipeline and optimizes their performance during rendering tasks.

3.3.3 Resource Separation

Resource isolation in modern graphics APIs is manifested in the mechanisms for updating and binding resources. Resources required at runtime are created beforehand and submitted to the pipeline through binding. These APIs provide explicit update mechanisms. When a program is running and using a bound resource, updating that resource is not allowed, ensuring API-level resource correctness. The initial version of the Vulkan API even prohibited updating resources after binding. Direct3D 12, owing to its resource organization, supports updating resources after binding while enforcing validity constraints. Metal also exhibits resource isolation characteristics, with its implementation being more hardware-centric and requiring fewer complex operations at the application layer. This approach ensures stability and efficiency in resource management during the rendering process.

3.4 Parallelism

Modern graphics APIs also focus on parallelism, a feature that significantly distinguishes them from traditional graphics APIs and allows for the full utilization of the powerful capabilities of both CPUs and GPUs. There are three types of parallelism: CPU to GPU parallelism, intra-GPU parallelism, and inter-GPU parallelism.

Modern graphics APIs' emphasis on these parallelism types results in substantial performance improvements[12], especially in applications that require intense graphics processing, such as modern video games and professional graphics design software.

3.4.1 Parallelism Between CPU and GPU

When the CPU calls a modern graphics API, the GPU does not immediately execute the corresponding instructions. Instead, the commands are saved to a list, which is only executed after being submitted to the GPU. Typically, each thread holds a command list, and different threads can record commands simultaneously and submit them in parallel to the GPU (Fig. 4). This process avoids prolonged idle times for the CPU, thereby improving CPU utilization.

3.4.2 Parallelism in GPU

Modern graphics APIs are capable of accessing different engines on the GPU, such as the 3D Engine, Compute Engine, and Copy Engine (Fig. 5). These engines can operate in parallel at the hardware level and are more efficient in handling their respective tasks. The CPU submits tasks to different types of queues, which are eventually dispatched to the corresponding engines. Submitting tasks according to their types significantly enhances the degree of concurrency within the GPU.

3.4.3 Parallelism Between GPUs

Modern graphics APIs explicitly support parallelism across multiple GPUs (Fig. 6). Different resources can be allocated to different GPUs, and different commands can be submitted to separate GPUs. After execution, the results from multiple GPUs are consolidated onto a single GPU used for displaying the results. This inter-GPU parallelism offers significant performance improvements when dealing with large-scale scenes and extensive tasks.

4 A Use Case: MoerEngine

MoerEngine is a high-performance real-time rendering engine developed using modern graphics APIs. Its core render module consists of the RHI, the RDG, and the shader compiler.

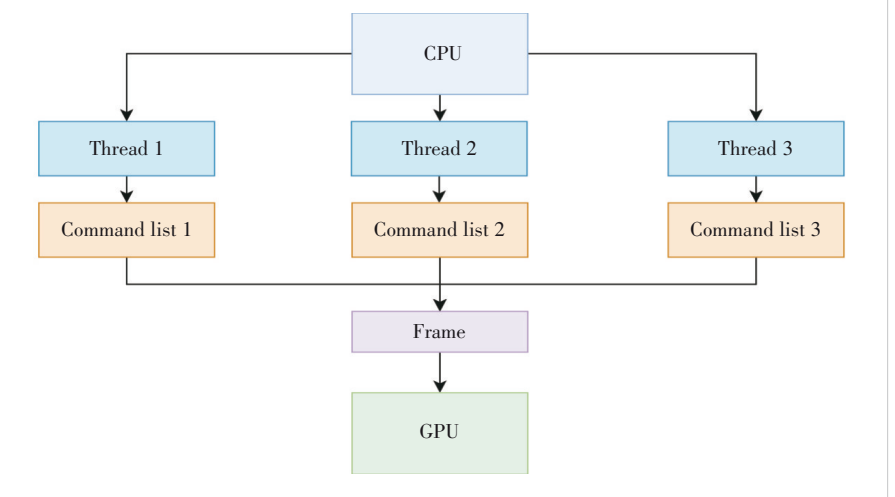


Figure 4. CPU multithreading in rendering

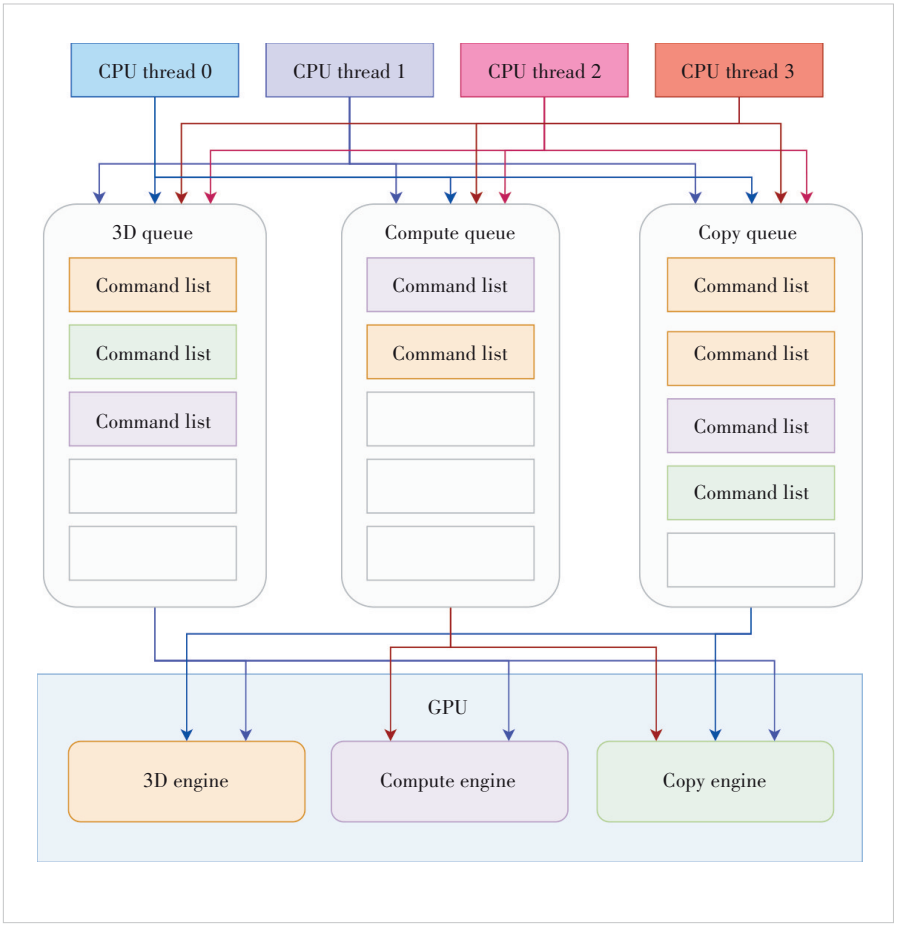


Figure 5. GPU internal parallelism

In this section, we elaborate on the development principles of modern graphics APIs in the context of MoerEngine's render modules.

In the render module of MoerEngine, a single component may reflect multiple design principles of modern graphics APIs (Fig. 7). This is quite common in modern engine development, where good design patterns often comprehensively consider the design principles of modern graphics APIs.

4.1 Challenge

Through the explanations of traditional and modern graphics APIs, it is evident that modern graphics APIs are explicitly designed, approaching the functionality of the driver layer, and require developers to undertake some of the tasks traditionally managed by the driver. This design grants developers greater freedom and, in theory, has the potential to maximize real-time rendering performance. However, achieving this is not straightforward; it requires more management and development work within the rendering engine. Simply porting an existing engine to a modern graphics API without thorough integration may result in performance that is inferior to that of traditional graphics APIs.

This section, based on the self-developed rendering engine MoerEngine, discusses the utilization of modern graphics APIs in high-performance rendering engines. MoerEngine supports multiple graphics APIs, whose designs are largely similar but differ in details. Therefore, the choice of a specific API is not the focus of this article; instead, we concentrate primarily on design approaches rather than specific API code.

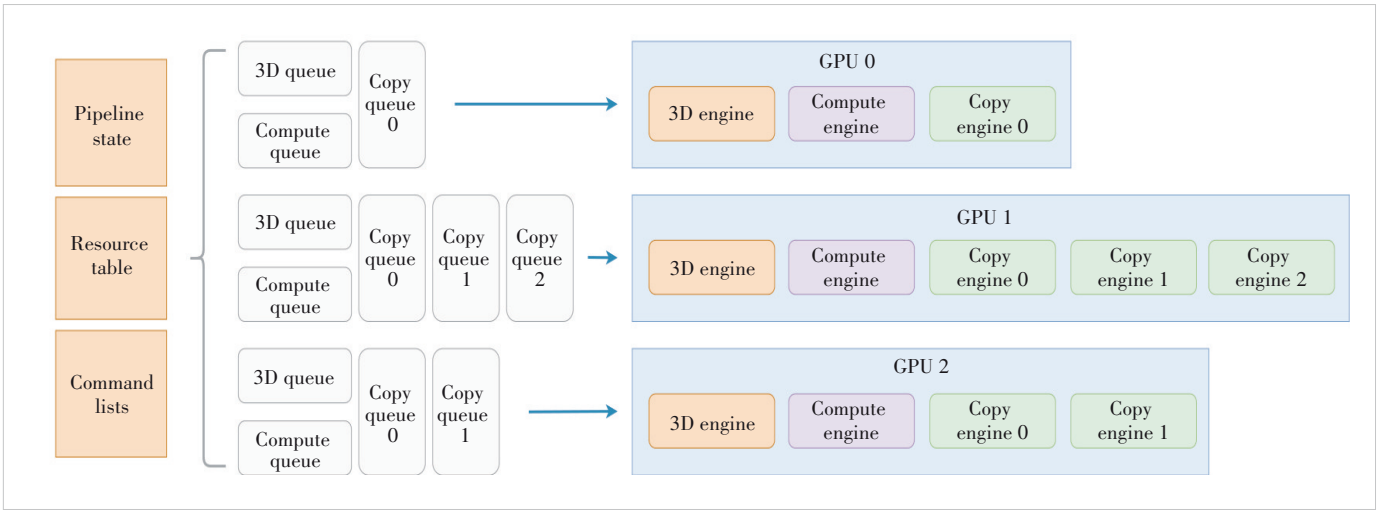


Figure 6. Parallelism across multiple GPUs

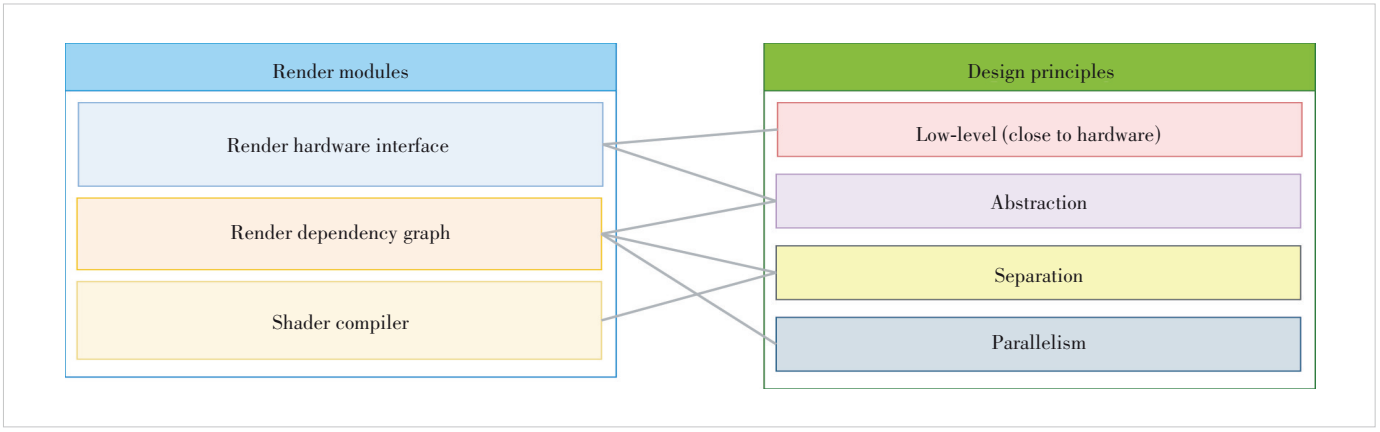


Figure 7. Modern API design principles in MoerEngine render modules

4.2 Render Hardware Interface

Render hardware interface (RHI) is an abstraction layer in the rendering engine for multiple platform-specific graphics APIs. It is designed from the ground up to fully exploit the advantages of various modern graphics APIs across different platforms. RHI embodies the low-level and abstraction principles of modern graphics APIs. The ability of MoerEngine to support multiple graphics APIs is largely attributed to the RHI.

RHI takes on the task of encapsulating low-level modern graphics APIs. It standardizes more generic implementations found in graphics APIs while distinguishing specialized implementations through different interfaces. RHI also conceals the complexities of the original APIs, such as intricate parameters, buffering, and multi-threaded scheduling, allowing higher-level modules to focus solely on the interfaces provided by RHI. Furthermore, RHI can internally perform performance optimizations or platform adaptations without changing its external interfaces. This capability of RHI enables a more efficient and adaptable interaction with the underlying graphics APIs, significantly simplifying the development process for the upper layers of the rendering engine.

4.3 Render Dependency Graph

Render dependency graph (RDG) is a higher-level abstraction within the rendering engine, representing a design pattern for managing complex rendering pipelines. It is a graph-based scheduling system designed to perform whole-frame optimization of the rendering pipeline and is widely used in the industry. RDG leverages modern graphics APIs for automatic asynchronous compute scheduling and more effective memory and barrier management to enhance performance (Fig. 8). It embodies the abstraction, separation, and parallelism design principles of modern graphics APIs. MoerEngine uses RDG to register and schedule rendering tasks for high-performance rendering.

Modern graphics APIs enable advanced rendering pipeline contexts to schedule rendering tasks, thereby improving performance and simplifying the rendering task stack. RDG delays task execution, recording the entire frame's rendering tasks into a dependency graph data structure. Once all passes are collected, the dependency graph is compiled and executed in an order sorted by dependencies. With the whole-frame context of the dependency graph and the powerful features of modern graphics APIs, RDG can perform complex scheduling tasks transparently to the user. Its specific capabilities include:

• automatically scheduling and synchronizing asynchronous compute channels;

• keeping resources' memory active during non-continuous intervals;

• pre-emptively starting barriers and layout transitions to avoid pipeline stalls.

Furthermore, RDG utilizes the dependency graph to provide extensive validation, enabling the automatic capture of functional and performance issues to improve the development process.

4.4 Shader Compiler

In modern graphics APIs, shaders generally support a pre-compilation mechanism and independent compilers, which

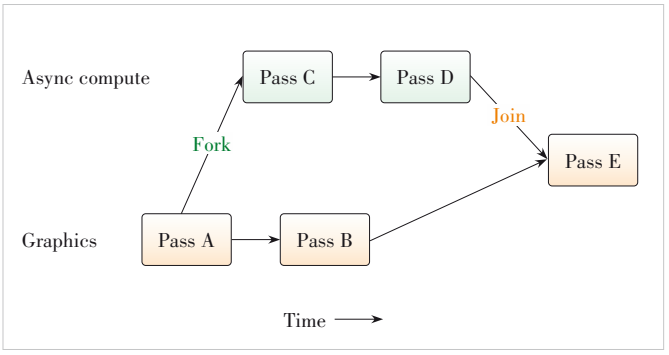


Figure 8. Render Dependency Graph in MoerEngine

has led to the widespread adoption of a general shader reflection mechanism. Shader reflection is not a new feature, as it has been supported since Direct3D 10[13], but widespread adoption has only occurred with modern graphics APIs. The shader reflection mechanism offers distinct advantages: it allows for the acquisition of resource information needed by the shader and automates resource binding to the pipeline during the pipeline construction phase. Shader compilation and reflection exemplify the separation principle of modern graphics APIs.

Specifically, various modern graphics APIs have their own shader pre-compilation formats, such as Direct3D 12's DXIL[14] and Vulkan's SPIR-V[15]. Metal differs from the others in that its intermediate representation resembles LLVM bytecode[16]. There are many shader compilers available that can compile shader languages into intermediate code forms, with reflection being carried out on the intermediate representations. MoerEngine embraces shader pre-compilation and reflection mechanisms, automatically aligning shader resources with pipeline resources. In MoerEngine, the DirectX Shader Compiler is used for shader compilation in Direct3D 12 and Vulkan, and corresponding API reflection is utilized (Fig. 9). This process significantly streamlines shader management, ensuring efficient and accurate shader resource utilization in the rendering pipeline.

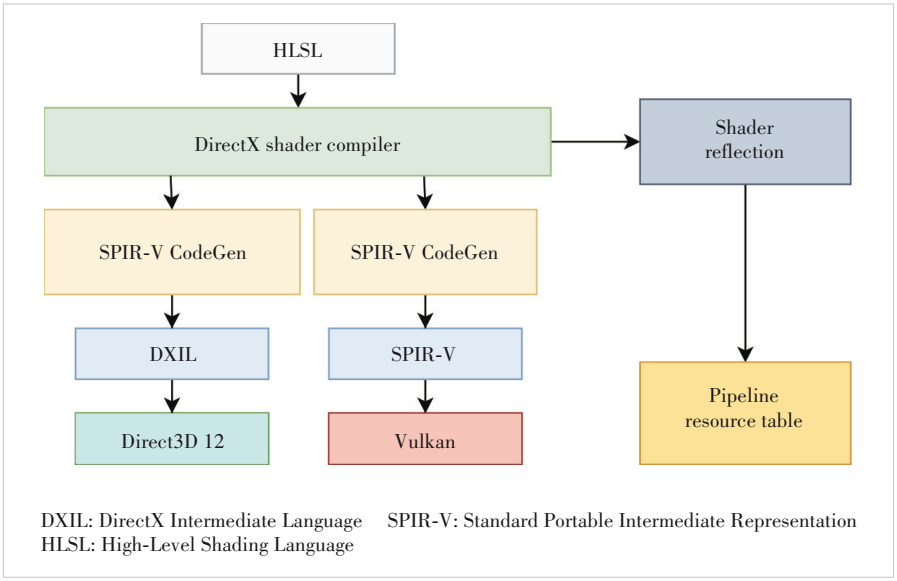


Figure 9. MoerEngine shader pre-compilation and reflection

4.5 Results

MoerEngine, developed based on the design principles of modern graphics APIs, achieves high performance and robust availability. It leverages real-time ray tracing algorithms to render complex scenes at interactive frame rates with high visual fidelity. Fig. 10 presents example scenes rendered in real time using MoerEngine. These scenes contain from hundreds of thousands to millions of triangles, featuring a variety of complex lighting and material types. All scenes are rendered using an NVIDIA 3090 GPU. With real-time ray tracing and denoising, MoerEngine achieves performance of over 90 frames per second (fps).

5 Future of Modern Graphics APIs

Modern graphics APIs are currently widely applied, meeting the rapidly growing demand for efficient graphics rendering. Especially in recent years, we have seen the emergence of new features built upon these APIs, such as Epic Games' GPU-driven Nanite and Lumen[17] technologies, which have elevated real-time rendering to new heights. Additionally, APIs like Direct ML and Vulkan ML have ventured into the field of machine learning, fully leveraging GPU's parallel computing power to accelerate various machine learning tasks. With these new features, modern graphics APIs have become more versatile, capable of handling increasingly complex and diverse tasks. This section explores the future of graphics APIs and graphics rendering based on the evolution of hardware and rendering requirements.

5.1 Evolution of Low-Level APIs

Low-level APIs such as Direct3D 12, Vulkan, and Metal are likely to continue evolving, offering closer control to the hardware layer. This would allow developers to more efficiently utilize the performance of modern multi-core GPUs, such as more direct and finer-grained control of VRAM and

Figure 10. Scenes rendered by MoerEngine

more direct thread control, potentially raising the performance ceiling of these APIs.

5.2 Cross-Platform Rendering

Modern graphics APIs have platform limitations; even Vulkan cannot achieve full hardware performance on every platform with a single codebase. Cross-platform graphics APIs, like WebGPU[18], are expected to be further refined and adopted. These APIs are not limited merely to platform compatibility; they also offer heterogeneous multi-device rendering and the potential for end-to-cloud rendering. Combined with remote calling mechanisms, cross-platform APIs could integrate rendering resources more effectively.

5.3 Native Real-Time Ray Tracing

Some APIs have already begun to support real-time ray tracing, with Microsoft's DirectX providing the DirectX Raytracing (DXR)[19] library for native ray tracing support. Mobile hardware manufacturers are now incorporating ray tracing acceleration units into their GPUs, indicating that real-time ray tracing is the future. Upcoming graphics APIs will inevitably support ray tracing more natively, possibly introducing more specialized ray tracing features or acceleration mechanisms.

5.4 Augmented and Mixed Reality Support

As virtual reality (VR), augmented reality (AR), and mixed reality (MR) technologies evolve, real-time rendering needs to support higher resolutions and faster refresh rates to provide a more immersive user experience[20]. AR and MR technologies, in particular, require rendering results that match real-world information, such as real-world lighting and color tones. Graphics APIs will need to provide access to a richer set of hardware capabilities, enabling applications to read real-world information and accurately reproduce it in virtual environments.

5.5 Neural Network Integration

In modern graphics processing, neural networks are playing an increasingly important role[21]. The trend in graphics hardware development now includes more hardware units optimized for neural network computations. Graphics APIs must also accommodate neural network tasks and integrate them more deeply with graphics pipelines, for example, through more direct input of physical information into neural networks to assist rendering. Müller et al.[22] utilized a neural radiance cache to accelerate global illumination in path tracing, significantly enhancing real-time ray tracing performance. With the advantages of modern hardware capabilities, the overhead for cache updates and queries can be extremely low. Going further, graphics APIs could completely restructure graphics pipelines, relying solely on neural networks for rendering. Such APIs would have a completely different design principle from traditional graphics pipelines, focusing on how to maximize neural network performance and provide the networks with the flexibility to handle diverse scenarios.

6 Conclusions

In conclusion, this paper has thoroughly examined the development and impact of modern graphics APIs such as Direct3D 12, Vulkan, and Metal, highlighting their revolutionary role in enhancing rendering efficiency. It underscores the significant advancements these APIs have brought to the field of graphics rendering, facilitating the transition from traditional to more sophisticated, high-performance techniques. Looking forward, this paper anticipates further innovations in graphics technology that are poised to reshape graphical computing in various applications. This study not only contributes to our understanding of current API capabilities but also paves the way for future research in evolving graphics technologies.

References

[1] Khronos Group, Inc. OpenGL registry [EB/OL]. (1992-06-30) [2022-05-05]. https://registry.khronos.org/OpenGL/index_gl.php
[2] Microsoft. Direct3D [EB/OL]. (1996-06-02) [2021-09-11]. https://learn.microsoft.com/en-us/windows/win32/direct3d
[3] Khronos Group, Inc. Vulkan specification [EB/OL]. (2016-02-16) [2023-12-08]. https://registry.khronos.org/vulkan/specs
[4] Microsoft. Direct3D 12 programming guide [EB/OL]. (2015-07-29) [2021-12-30]. https://learn. microsoft. com/en-us/windows/win32/direct3d12/directx-12-programming-guide
[5] Apple Inc. Metal [EB/OL]. [2023-12-08]. https://developer. apple. com/documentation/metal
[6] Unterguggenberger J, Kerbl B, Wimmer M. Vulkan all the way: transitioning to a modern low-level graphics API in academia [J]. Computers & graphics, 2023, 111: 155 – 165. DOI: 10.1016/j.cag.2023.02.001
[7] Thompson C J, Hahn S, Oskin M. Using modern graphics architectures for general-purpose computing: a framework and analysis [C]//Proc. 35th Annual ACM/IEEE International Symposium on Microarchitecture (MICRO 35. IEEE, 2002: 306 – 317. DOI: 10.1109/MICRO.2002.1176259
[8] Owens J D, Luebke D, Govindaraju N, et al. A survey of general-purpose computation on graphics hardware [J]. Computer graphics forum, 2007, 26 (1): 80 – 113. DOI: 10.1111/j.1467-8659.2007.01012.x
[9] O'Donnell Y. (GDC 2017) FrameGraph: extensible rendering architecture in frostbite [R]. San Francisco, CA: Frostbite/Electronic Arts, 2017
[10] Piñeiro A. ARB_gl_spirv: bringing SPIR-V to Mesa OpenGL: FOSDEM 2018 [R]. Brussels: Igalia, 2018
[11] Bangaru S P, Wu L F, Li T-M, et al. 2023. SLANG.D: fast, modular and differentiable shader programming [J]. ACM trasactions on graphics, 2023, 42(6): 1 – 28. DOI: 10.1145/3618353
[12] John McDonald. (GTC 2016) High performance vulkan: lessons learned from source 2 [R]. San Jose: Valve, 2016
[13] Blythe D. The Direct3D 10 system [J]. ACM transactions on graphics, 2006, 25(3): 724 – 734. DOI: 10.1145/1141911.1141947
[14] Microsoft. DirectX intermediate language [EB/OL]. (2016-12-29) [2023-12-01]. https://github. com/microsoft/DirectXShaderCompiler/blob/main/docs/DXIL.rst
[15] Khronos Group, Inc. SPIR-V specification [EB/OL]. (2015-05-03) [2021-12-16]. https://www.khronos.org/registry/SPIR-V
[16] Aras Pranckevičius. Porting unity to new APIs [C]//Proc. ACM SIGGRAPH 2015 Course Notes: An Overview of Next-generation Graphics APIs. DOI: 10.1145/2776880.2787704
[17] Tatarchuk N, Dupuy J, Deliot T, et al. Advances in real-time rendering in

games: part I [C]//Proc. ACM SIGGRAPH 2022 Courses. ACM, 2022. DOI: 10.1145/3532720.3546895
[18] Kenwright B. Introduction to the WebGPU API [C]//Proc. ACM SIGGRAPH 2022 Courses. ACM, 2022. DOI: 10.1145/3532720.3535625
[19] Matt Sandy. (GDC 2018) DirectX Raytracing [R]. San Francisco: Microsoft, 2018
[20] Billinghurst M, Nebeling M. Rapid prototyping of XR experiences [C]// Proc. ACM SIGGRAPH 2022 Courses. ACM, 2022. DOI: 10.1145/3532720.3535684
[21] Marshall C S. Practical machine learning for rendering: from research to deployment [C]//Proc. ACM SIGGRAPH 2021 Courses. ACM, 2021. DOI: 10.1145/3450508.3464564
[22] Müller T, Rousselle F, Novák J, et al. Real-time neural radiance caching for path tracing [J]. ACM transactions on graphics, 2021, 40(4): 1 – 16. DOI: 10.1145/3450626.3459812

Biographies

Lu Ping is the Vice President and Director of the R&D Project in the Technology Planning Department at ZTE Corporation. He also serves as Executive Deputy Director of the National Key Laboratory of Mobile Network and Mobile Multimedia Technology. His research directions include cloud computing, big data, augmented reality, and multimedia service-based technologies. He has contributed to major national science and technology projects and has published multiple papers and authored two books.

Sun Qi (522023330087@smail.nju.edu.cn) is a master's student in the Department of Computer Science and Technology at Nanjing University, China, affiliated with the Meta Graphics & 3D Vision Lab. His research interests include photorealistic rendering and high-performance real-time rendering.

Wang Chen is a master's student in the Department of Computer Science and Technology at Nanjing University, China, affiliated with the Meta Graphics & 3D Vision Lab. His research interests include real-time ray tracing and rendering software architecture.

Guo Jie is an associate researcher at the Department of Computer Science and Technology, Nanjing University, China. He received his PhD from Nanjing University in 2013. His current research interests include computer graphics, virtual reality, and 3D vision. He has authored over 80 publications in leading international conferences (e.g., SIGGRAPH, SIGGRAPH Asia, CVPR, ICCV, ECCV, IEEE VR) and journals (e.g., *ACM ToG*, *IEEE TVCG*, *IEEE TIP*). He has developed several applications in illumination prediction, material prediction, and real-time rendering, which have been widely used in industry and achieved good economic and social benefits.

Guo Yanwen is a Professor and PhD supervisor at Nanjing University, China. He was appointed as a PhD supervisor in July 2013 and as a Professor in December 2014. He is a recipient of the Jiangsu Province Outstanding Young Scientist Fund and a core member of the Department of Computer Science and Technology and the State Key Laboratory for Novel Software Technology at Nanjing University. He serves as Executive Director of the Nanjing University-iQIYI Joint Innovation Center, a board member of the China Society for Image and Graphics, Chair of the Graphics and Image Committee of the Jiangsu Computer Society, and Chair of the Virtual Reality Committee of the Jiangsu Society of Engineers. He has published nearly 100 high-level papers, including approximately 20 in ACM/IEEE Transactions.

Shi Wenzhe is a strategy planning engineer at Beijing Xingyun Digital Technology Co., Ltd. and a member of the National Key Laboratory for Mobile Network and Mobile Multimedia Technology, China. He is also involved inproduct planning for the XRExplore Platform at Beijing Xingyun Digital Technology Co., Ltd. His research interests include indoor visual AR navigation, SFM 3D reconstruction, visual SLAM, real-time cloud rendering, VR, and spatial perception.

New Member of ZTE Communications Editorial Board

Yu Zhiwen is the Vice President of Harbin Engineering University, China, and a professor at Northwestern Polytechnical University, China. He is a Fellow of the China Computer Federation (CCF), a Distinguished Professor of the Chang Jiang Scholars Program of the Ministry of Education, China, a recipient of the National Science Fund for Distinguished Young Scholars, China, a leader in science and technology innovation under the National Ten Thousand Talents Program, China, and the Chief Scientist of a National Key R&D Program Project, China.

He serves as the Director of the Ministry of Education Key Laboratory of Human-Machine-Object Fusion and Crowd Intelligence Computing, China, the Director of the Ministry of Industry and Information Technology Key Laboratory of Intelligent Perception and Computing, China, and the Leader of the National Innovation Team for Underwater Swarm Intelligence, China.

His research interests include the Internet of Things (IoT), ubiquitous computing, and crowd intelligence sensing and computing. He has published over 300 papers in top international academic journals and conferences, including *IEEE TMC*, *IEEE TKDE*, MobiCom, UbiComp, INFOCOM, and KDD, with 9 papers recognized as ESI Highly Cited Papers.

He is an editorial board member of prestigious international journals such as *IEEE Transactions on Human-Machine Systems*, *IEEE Communications Magazine*, and *ACM IMWUT*. He has also served as conference chair or program committee member for over 50 international conferences, including ACM UbiComp, IEEE PerCom, and IJCAI.

Additionally, he holds positions such as Chair of the ACM Xi'an Chapter, China, Senior Member of IEEE, Executive Council Member of the CCF, and Chair of the Technical Committee on Ubiquitous Computing of the CCF.

He has received numerous awards, including the IEEE HITC Outstanding Technical Achievement Award, the IEEE Smart World Outstanding Research Award, the CCF Excellent Doctoral Dissertation Award, the CCF Young Scientist Award, the Second Prize of National Teaching Achievement, the First Prize of Natural Science of the Ministry of Education, China, the First Prize of Natural Science of Shaanxi Province, and the First Prize of Science and Technology Progress of Heilongjiang Province.

The 2nd Youth Expert Committee
for Promoting Industry-University-Institute Cooperation

Director **Chen Wei,** Beijing Jiaotong University

Deputy Director **Qin Xiaoqi,** Beijing University of Posts and Telecommunications

Lu Dan, ZTE Corporation

Members (Surname in Alphabetical Order)

Cao Jin	Xidian University
Chen Li	University of Science and Technology of China
Chen Qimei	Wuhan University
Chen Shuyi	Harbin Institute of Technology
Chen Siheng	Shanghai Jiao Tong University
Chen Wei	Beijing Jiaotong University
Gao Zhen	Beijing Institute of Technology
Guan Ke	Beijing Jiaotong University
Han Chong	Shanghai Jiao Tong University
Han Kaifeng	China Academy of Information and Communications Technology
He Zi	Nanjing University of Science and Technology
Hou Tianwei	Beijing Jiaotong University
Hu Jie	University of Electronic Science and Technology of China
Huang Chen	Purple Mountain Laboratories
Huo Jiahao	University of Science and Technology Beijing
Li Ang	Xi'an Jiaotong University
Li Li	University of Science and Technology of China
Liu Fan	Southeast University
Liu Junyu	Xidian University
Lu Dan	ZTE Corporation
Lu Youyou	Tsinghua University
Mei Weidong	University of Electronic Science and Technology of China
Ning Zhaolong	Chongqing University of Posts and Telecommunications
Pan Cunhua	Southeast University
Qi Liang	Shanghai Jiao Tong University
Qin Xiaoqi	Beijing University of Posts and Telecommunications
Qin Zhijin	Tsinghua University
Shi Yao	Harbin Institute of Technology
Shi Yinghuan	Nanjing University
Tang Wankai	Southeast University
Wang Jingjing	Beihang University
Wang Xinggang	Huazhong University of Science and Technology
Wang Yongqiang	Tianjin University
Wen Miaowen	South China University of Technology
Wu Qingqing	Shanghai Jiao Tong University
Wu Yongpeng	Shanghai Jiao Tong University
Xia Wenchao	Nanjing University of Posts and Telecommunications
Xiang Luping	Nanjing University
Xu Mengwei	Beijing University of Posts and Telecommunications
Xu Tianheng	Shanghai Advanced Research Institute, Chinese Academy of Sciences
Yang Chuanchuan	Peking University
Ye Yinghui	Xi'an University of Posts and Telecommunications
Yin Haifan	Huazhong University of Science and Technology
You Changsheng	Southern University of Science and Technology
Yu Jihong	Beijing Institute of Technology
Zhang Jiao	Beijing University of Posts and Telecommunications
Zhang Jiayi	Beijing Jiaotong University
Zhang Yuchao	Beijing University of Posts and Telecommunications
Zhao Yizhe	University of Electronic Science and Technology of China
Zhao Yuda	Zhejiang University
Zhao Zhongyuan	Beijing University of Posts and Telecommunications
Zhou Yi	Southwest Jiaotong University
Zhu Bingcheng	Southeast University
Zhu Guangxu	Shenzhen Research Institute of Big Data
Zhu Zhengyu	Zhengzhou University

ZTE COMMUNICATIONS
中兴通讯技术(英文版)

ZTE Communications has been indexed in the following databases:

- Abstract Journal
- China Science and Technology Journal Database
- Chinese Journal Fulltext Databases
- Index Copernicus
- Scopus
- Ulrich's Periodicals Directory
- Wanfang Data
- WJCI 2021–2025

ZTE COMMUNICATIONS
Vol. 24 No. 1 (Issue 94)
Quarterly
First Issue Published in 2003

Supervised by:
Anhui Publishing Group

Sponsored by:
Time Publishing and Media Co., Ltd.
Shenzhen Guangyu Aerospace Industry Co., Ltd.

Published by:
Anhui Science & Technology Publishing House

Edited and Circulated (Home and Abroad) by:
Magazine House of ZTE Communications

Editorial Correspondence:
Add: 12F Kaixuan Building, 329 Jinzhai Road, Hefei 230061, P. R. China
Tel: +86-551-65533356
Email: magazine@zte.com.cn
Website: http://zte.magtechjournal.com

Annual Subscription: RMB 120
Printed by:
Anhui Tianjin Printing Technology Co., Ltd.
Publication Date: March 25, 2026

China Standard Serial Number: ISSN 1673–5188 / CN 34–1294/TN