

FDD Massive MIMO Channel State Information Compression and Reporting Technology Based on AI/ML—Take Reducing Feedback Overhead as the Main Line

Yi Jiang^{1,*}, Xianmeng Wang²

¹*Department of JSNU-SPbPU Institute of Engineering, Jiangsu Normal University, Xuzhou, China*

²*Department of Electronic Information Engineering, Hankou University, Wuhan, China*

**Corresponding Author. Email: 18852606001@163.com*

Abstract. FDD Massive MIMO is an important enabling technology for 5G-Advanced and 6G, but its downlink channel state information (CSI) feedback overhead increases significantly with the increase in the number of antenna array elements and has become the main obstacle affecting system performance. The traditional schemes based on codebooks and compressed sensing rely on preset constraints and a step-by-step processing approach, and their compression performance has approached the optimal compression ratio. In recent years, AI/ML methods represented by deep learning have provided a new idea of "Data-driven end-to-end joint optimization" for CSI compression and feedback. However, these AI/ML methods have several problems such as complex plans, inconsistent assessment criteria, and difficulties in implementing standardization. This article takes reducing feedback overhead as the main line, classifies the existing schemes into two categories: autoencoders and Transformer architectures, and comprehensive overview of methodological principles, strengths, and weaknesses of each type of method. In addition, the article links the latest research results to the standard progress of 3GPP Rel-18/19/20 and points out the gap between the existing work and the standard. Also, this article offers multi-objective optimization suggestions and a deployment plan by combining parameter size, inference latency, energy consumption. Finally, the article proposes generalization ability, lightweight design, and the capability boundaries of large models as future research directions, with the aim of promoting the early practical application of AI/ML-based CSI feedback technology.

Keywords: FDD Massive MIMO, CSI, Autoencoder, Transformer

1. Introduction

In Frequency Division Duplex (FDD) Massive Multiple-Input Multiple-Output (MIMO) systems, the User Equipment (UE) needs to return the complete downlink channel state information (CSI) to the base station so that the base station can generate precise beamforming to complete signal transmission. With the continuous increase in the number of base station antennas, the traditional codebook feedback method is difficult to simultaneously balance the channel reconstruction

accuracy and the occupation of uplink transmission resources. In recent years, end-to-end channel compression feedback schemes based on artificial intelligence and machine learning (AI/ML) have been widely studied. Among them, schemes such as autoencoders rely on excellent effectiveness of nonlinear feature extraction and can achieve higher-quality channel restoration with a small amount of uplink transmission resources [1], which is a key research direction for 5G-Advanced and 6G air interface optimization.

Deep learning technology solves the performance bottlenecks of traditional compressed sensing algorithms. It is currently widely used to optimize CSI feedback in massive MIMO systems. Relevant studies have achieved fruitful results in network structure optimization, lightweight model deployment and dynamic overhead management [2]. Guo et al. [3] concluded the technical framework and standardization development of AI-assisted CSI feedback under 5G-A scenarios. From a macroscopic perspective, Xiao et al. [4] sorted out the application strategies and development prospects of AI technology in CSI feedback, which provides solid theoretical support for relevant research in this field. Liu et al. proposed an innovative wireless foundation model and established the WiFo-CF scheme. This work explores the practical value of large models in channel reconstruction tasks. Radwan et al. [5] combined contrastive predictive coding with compression techniques. They also adopted self-supervised learning to improve the channel feature representation ability. Zhuang et al. [6] proposed the OLAMCF framework. This framework further enhanced the CSI feedback performance in FDD massive MIMO systems by using offline large models. For model lightweighting and low-power deployment, Khan et al. [7] provided a systematic review of AI model compression techniques in 5G wireless networks. These techniques include pruning, quantization, and knowledge distillation. Zhai and Huang [8] designed a communication AI fusion compression algorithm for low-power scenarios. Their work offered a feasible solution for the lightweight deployment of terminals. For dynamic optimization and overhead control, Li et al. [9] achieved dynamic optimization of the CSI compression model based on a digital twin network (DTN), enhancing the system's environmental adaptability. Shehzad et al. [10] utilized AI methods to optimize channel compression loss and signaling overhead, effectively balancing feedback performance and resource occupation. Current studies demonstrate that the intelligent CSI feedback technology still faces several obvious defects, which restrict its standardization process and commercial application. The algorithm simulation conditions and evaluation indicators are not unified, making it difficult to compare mainstream solutions horizontally. Most studies blindly pursue the accuracy of channel reconstruction and complex models, resulting in excessive terminal delay and power consumption, which goes against the original intention of reducing overhead. In addition, few studies closely combine with the 3GPP R18/R19/R20 standards, they ignore the engineering constraints, and have low feasibility.

Focused on the core goal of feedback overhead reduction, this paper analyzes the compression limit of channel data based on information theory. It also explores the balance among compression ratio, channel recovery accuracy, model lightweight performance and scenario adaptation capability. Existing mainstream solutions are classified into two categories, namely autoencoder-based models and Transformer-based models. With a unified evaluation system, this paper compares the compression performance and practical implementation potential of different schemes, and further discusses the research progress combined with the development trend of communication standards. In terms of theoretical research, this paper constructs an analytical framework that considers both compression efficiency and terminal hardware limitations. In practical engineering applications, the comparative results of different schemes in model parameters, running latency and standard adaptability can provide effective references for equipment manufacturers and standard formulation

work. The research findings can promote the commercial deployment of intelligent CSI feedback technology and optimize the overall performance of FDD Massive MIMO systems..

2. Contrastive analysis of AI CSI feedback on mainstream architecture principles and performance architectures

2.1. Autoencoder

Autoencoder is an end-to-end neural network and the most widely used infrastructure in CSI compression. Its encoder is deployed on the user's UE side and is composed of multiple layers of Multi-Layer Fully Connected (Neural Network), convolutional or cyclic networks. It is responsible for reducing the dimension of the high dimensional original channel matrix H to a low-dimensional latent vector, which is fed back to the base station via the uplink, thereby significantly reducing the amount of transmission data volume. The decoder is located on the gNB side of the base station and reconstruct the received latent vector into a channel matrix H with approximate original dimensions via symmetric deconvolution or fully connected layers. During the training phase, the normalized mean square error (NMSE) of the original channel and the reconstructed channel is used as the loss function to synchronously optimize the parameters of the encoder and decoder, achieving end-to-end joint optimization without the need for manually designed codebook or sparsity constraints, which is different from traditional compressive sensing schemes. Representative solutions such as AnciNet optimize the convolutional structure for massive MIMO scenario to enhance the extraction of local channel correlations. The reasoning stage adopts an offline training and online forward propagation mode. The network weights are fixed. In the actual communication scenario, the encoder is compressed in real time and the decoder is reconstructed in real time, without the need for online gradient updates. The advantage of the autoencoder solution lies in that end-to-end data-driven optimization can break through manual constraints, achieve better reconstruction accuracy under the same feedback overhead, and autoencoder has a simple architecture and convenient deployment. It entered the 3GPP standardization discussion earlier and can flexibly control the compression ratio by adjusting the hidden vector dimension. However, its limitations are equally obvious: the model's generalization ability is limited, and its performance drops sharply when facing a channel environment that is significantly different from the training set; The reconstruction accuracy is severely lost under a high compression ratio. The network is unable to dynamically adapt to channel changes online. In the scenario of massive MIMO scenario the number of parameters, delay and energy consumption increase significantly. Moreover, convolutional or cyclic structures can only extract local features, and their global modeling ability is weaker than that of Transformer-based schemes.

2.2. Transformer

The Transformer architecture takes the self-attention mechanism as the core, and can directly model the global dependencies between all elements in the channel matrix, effectively overcoming the limitations of the local receptive field of the autoencoder, especially for large-scale antenna and high-speed mobile scenarios. The system flow includes channel estimation, CSI embedding and position coding, multi-head self-attention coding, quantitative feedback and decoding reconstruction.

Firstly, the real and imaginary parts of the complex CSI are split and serialized, and the hidden feature vector is mapped by the learnable embedding layer, and the spatial and temporal position

coding is introduced to retain the antenna arrangement and time domain change information. The encoder is composed of a multi-head self-attention layer and a feedforward network stack. The global correlation weight is calculated in parallel, and the output low-dimensional hidden vector is quantized to generate feedback bits. The decoder uses the cross-attention mechanism to fuse the compressed hidden vector and the position encoding, and recovers the complete channel matrix layer by layer. In addition, plug-in modules such as adaptive fine-tuning or lightweight operators can be integrated as needed to optimize performance. The advantage of the Transformer scheme lies in its powerful global feature capture ability, which makes its compression efficiency and reconstruction performance better than the autoencoder, and the modular design facilitates flexible expansion. However, its shortcomings cannot be ignored: the computational complexity of self-attention increases quadratically with the channel dimension, resulting in higher parameter scale, terminal delay and power consumption than the autoencoder; the fixed network structure is difficult to adapt to the extreme dynamic environment; the diversification of design schemes lacks uniform standards, which is not conducive to 3GPP standardization. Moreover, the training requires massive multi-scenario channel data, and the cost of data acquisition and calculation is high

3. The latest research results

3.1. 3GPP standardization progress (Rel-18 / 19 / 20 evolution context)

3GPP has officially launched AI / ML air interface standardization research since Rel-18, and listed CSI feedback enhancement as the first batch of representative use cases. Rel-18 completed the feasibility study of AI-CSI compression, outputted TR 38.843 technical report, and established the bilateral architecture of ' UE-side AI encoder + gNB-side AI decoder '. The scheme relies on the autoencoder to realize the joint compression of spatial and frequency domain signals. After several rounds of verification, it can greatly reduce the resource overhead of CSI uplink feedback while ensuring the accuracy of channel information reconstruction. On this basis, the Rel-19 version further deepens the research of related technologies, and the core research direction focuses on balancing the relationship between algorithm performance, model complexity and feedback overhead. It focuses on the lightweight optimization of bilateral AI models and the adaptation optimization of multiple wireless scenarios. At the same time, it conducts in-depth research on the full life cycle management (LCM) mechanism of AI models and the mutual adaptation problems between devices from different manufacturers. In February 2025, Rod and Schwartz jointly completed the industry 's first cross-vendor bilateral AI-CSI interoperability test verification. The test is based on the terminal reference design driven by Qualcomm 's 5G modulation and demodulation RF chip, and the measured environment is built with Rhodes and Schwartz CMX500 5G integrated signaling tester, which successfully realizes the cross-device compatible operation of the machine learning model. In this test, Qualcomm is responsible for developing the AI coding model on the terminal side, Rod and Schwartz develop the decoding model on the base station side, and complete the fusion and docking of the two models through the open neural network exchange format. The measured data show that compared with the traditional CSI feedback mechanism of Type I defined in 3GPP Release 15, the system throughput of the AI optimization scheme can be increased by 51%

In the Rel-20 phase, the AI / ML interface has officially shifted from research projects to normative work projects. The core task is to transform the research results of Rel-18 and Rel-19 into commercially available functional specifications. Rel-20 focuses on the development of three specifications : one is to clarify the functional requirements of UE-side AI encoder and gNB-side AI

decoder, and standardize the whole process of CSI compression, feedback and reconstruction ; the second is to improve the LCM specification of AI model, clarify the signaling process of capability negotiation, model distribution, activation/deactivation, performance monitoring and other links, and solve the core pain points of cross-vendor model interoperability [11]; the third is to develop AI-CSI interoperability test cases led by RAN4, clarify core indicators such as channel reconstruction accuracy, feedback delay, overhead, and establish performance baselines. On the whole, the 3GPP standardization process has moved from the feasibility verification stage to the specification formulation stage, laying a standard foundation for the commercial implementation of AI-CSI compression technology.

3.2. Existing deficiencies and challenges

However, there is still a significant gap between the existing academic research and 3GPP standardization work, which is mainly reflected in the following four aspects. First, the model evaluation criteria are not uniform. A large number of academic studies use non-3GPP standard channel models such as COST 2100 for simulation verification. The performance results obtained cannot be directly adapted to the evaluation system and channel scenarios defined by 3GPP TR 38.843. The 3GPP standardization test requires the use of a consistent channel model, performance benchmarks and test cases. The two evaluation systems of academic research and standard evaluation are parallel, making it difficult for many algorithms with excellent performance to be directly absorbed by the standardization working group [12]. Secondly, the complexity of the model is out of line with the terminal hardware constraints. Although the Transformer-like scheme has significant advantages in reconstruction accuracy, its self-attention computational complexity increases quadratically with the channel dimension. The number of native model parameters, terminal inference delay and energy consumption are much higher than the autoencoder [12]; in contrast, the autoencoder architecture is simple, the number of parameters is small, the reasoning speed is fast, and it is easier to meet the engineering constraints of low power consumption and low delay at the UE side. The research shows that the model volume can be compressed to less than 1 MB or the decoder GFLOPs can be reduced by 32 times through reasonable lightweight design. These achievements provide an implementation idea for AI-driven CSI feedback landing terminals, but most academic papers do not fully combine terminal hardware constraints to carry out research at this stage. Third, the model generalization ability is insufficient. The deep learning model is usually trained based on a fixed wireless environment. When it is migrated to other channel scenarios for actual deployment, it is prone to channel environment mismatch problems, resulting in a sharp decline in performance. Fourth, cross-vendor interoperability has not yet been fully resolved. Although the verification of interoperability between R & S and Qualcomm has taken a key step, the LCM specification is still under development [11].

4. Conclusion

This paper takes reducing the feedback overhead of FDD Massive MIMO CSI as the main line, divides the existing AI compression scheme into two routes : autoencoder and Transformer basic model, establishes a unified quantitative evaluation system, compares the principles, advantages, disadvantages and applicable scenarios of the two types of architectures, and combines the latest 3GPP standards to analyze the gap between academic research and industrial implementation. This paper sorts out the existing bottlenecks in the field and gives multi-dimensional optimization deployment ideas, which makes up for the shortcomings of some existing reviews focusing on

algorithms, ignoring hardware and standardization constraints, and provides theoretical and engineering reference for equipment manufacturers and standard formulation. However, there are still limitations in the research : only focusing on the two mainstream architectures of autoencoder and Transformer, incomplete coverage of new niche schemes such as graph neural network, quantitative analysis only relying on theoretical derivation, lack of multi-scenario measured data support, and insufficient mining of industrialization details such as multi-vendor interoperability and multi-band collaborative feedback. In view of the existing problems such as insufficient generalization performance of CSI feedback in deep learning, large power consumption of terminal computing power, disconnection between academic evaluation and 3GPP standard, and difficulty in approaching the compression limit of information theory, the follow-up research can be improved from six directions: adaptive generalization network, multi-objective model lightweight, collaborative architecture of large and small models, standardized compatibility transformation, information theory-oriented network optimization, and digital twin dynamic model selection, so as to accelerate the commercial landing of technology.

From the perspective of industry development prospects, the short-term lightweight autoencoder will take the lead in completing the standardized landing, the medium-term lightweight Transformer and the size model collaborative architecture will complete the adaptation of the ultra-large-scale antenna scene, and the long-term general wireless basic model will become the core of 6G intelligent air interface. In the future, it is necessary to continue to overcome the pain points of generalization, lightweight and protocol compatibility, and rely on the multi-disciplinary collaboration of communication, AI and hardware to make AI-driven CSI compression technology truly large-scale commercial.

Author contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] Liu, X., Gao, S., Liu, B., et al. (2026). WiFo-CF: Wireless foundation model for CSI feedback. *IEEE Transactions on Wireless Communications*, 25, 15039–15053.
- [2] Zhaohua, L., Chenyang, Y., Jie, W., et al. (2026). Deep CSI compression and feedback for massive MIMO: A survey. *ZTE Communications*, 24(1).
- [3] Guo, J., Wen, C. K., Jin, S., et al. (2022). AI for CSI feedback enhancement in 5G-advanced. *IEEE Wireless Communications*, 31(3), 169–176.
- [4] Xiao, H., Wang, Z., Tian, W., et al. (2021). AI enlightens wireless communication: Analyses, solutions and opportunities on CSI feedback. *China Communications*, 18(11), 104–116.
- [5] Radwan, A. Y., Muhammad, F. S., Baker, M., et al. (2026). Contrastive predictive coding with compression for enhanced channel state feedback in wireless networks. *IEEE Transactions on Neural Networks and Learning Systems*.
- [6] Zhuang, J., Wang, Y., Hou, H., et al. (2025). OLAMCF: Offline large AI models enhanced CSI feedback in FDD massive MIMO systems. In *GLOBECOM 2025 - 2025 IEEE Global Communications Conference* (pp. 3517–3522). IEEE.
- [7] Khan, F. M. A., Hallaq, M., Abou-Zeid, H., et al. (2026). Model compression for sustainable AI in xG wireless networks: Recent advances, challenges, and future directions. *IEEE Communications Surveys & Tutorials*.
- [8] Zhai, Q., & Huang, J. (2026). Design of a low-power constraint communication AI fusion model compression algorithm. *Neural Computing and Applications*, 38(12), 486.
- [9] Li, J., Li, R., Zhao, Z., et al. (2025). Wireless digital twin network enabled model selection for AI based CSI compression. In *2025 IEEE Globecom Workshops (GC Wkshps)* (pp. 1863–1868). IEEE.
- [10] Shehzad, M. K., Rose, L., & Assaad, M. (2021). Dealing with CSI compression to reduce losses and overhead: An artificial intelligence approach. In *2021 IEEE International Conference on Communications Workshops (ICC)*

Workshops) (pp. 1–6). IEEE.

- [11] Huang, C.-H., Wen, C.-K., & Li, G. Y. (2025). AI/ML life cycle management for interoperable AI native RAN (arXiv:2507.18538). *arXiv*.
- [12] Gao, Y., Wu, X., Jiang, J., Hu, B., Du, J., Ye, Q., Zhang, S., Yu, F. R., & Xu, S. (2026). AI/ML for mobile networks: Current status in Rel. 19 and challenges ahead (arXiv:2603.14317). *arXiv*.