AI for CSI Feedback Enhancement in 5G-Advanced

Jiajia Guo, *Graduate Student Member, IEEE*, Chao-Kai Wen, *Senior Member, IEEE*,
Shi Jin, *Senior Member, IEEE*, and Xiao Li, *Member, IEEE*

Abstract—The 3rd Generation Partnership Project started the study of Release 18 in 2021. Artificial intelligence (AI)-native air interface is one of the key features of Release 18, where AI for channel state information (CSI) feedback enhancement is selected as the representative use case. This article provides an overview of AI for CSI feedback enhancement in 5G-Advanced. Several representative non-AI and AI-enabled CSI feedback frameworks are first introduced and compared. Then, the standardization of AI for CSI feedback enhancement in 5G-advanced is presented in detail. First, the scope of the AI for CSI feedback enhancement in 5G-Advanced is presented and discussed. Then, the main challenges and open problems in the standardization of AI for CSI feedback enhancement, especially focusing on performance evaluation and the design of new protocols for AI-enabled CSI feedback, are identified and discussed. This article provides a guideline for the standardization study of AI-based CSI feedback enhancement.

I. INTRODUCTION

IN June 2022, the 3rd Generation Partnership Project (3GPP) completed the latest release of the fifth generation (5G) cellular networks, namely, Release 17, which is the last version of the first stage of 5G evolution. With the functional freeze of Release 17, the 3GPP started planning Release 18, the first release of the second stage of 5G evolution, namely, 5G-Advanced [1], [2]. Artificial intelligence (AI) is regarded as one of the key features of the 5G-Advanced and sixth generation (6G) cellular networks, and it will trigger a paradigm shift and lay a strong foundation for 5G-Advanced and 6G [1]–[3].

Unlike the conventional methods supported by domain knowledge and rigorous theoretical proofs, AI-based methods, including deep learning and reinforcement learning, can automatically learn or extract features from a training dataset via neural networks (NNs). To efficiently learn the features from the dataset, considerable NN architectures, such as dense, convolutional, and recurrent NNs, are developed. Inspired by AI's great success in computer vision, it has been introduced to wireless communications in the past few years [3]. AI-enabled radio access networks were investigated in Release 17 [4]. However, the AI-native air interface is excluded from the study scope of Release 17. In December 2021, a new study item on AI-native air interface was proposed in Release 18 [5]. This study item aims to explore the benefits of integrating AI into the air interface, including performance improvement and complexity reduction. In particular, this study item focuses

Corresponding author: Shi Jin and Xiao Li

Jiajia Guo, Shi Jin, and Xiao Li are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, 210096, P. R. China.

Chao-Kai Wen is with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan.

on assessing the performance of the AI-native air interface compared with conventional methods and its specification impacts, establishing a common AI framework for the AI-native air interface, and identifying the requirement of an AI-native air interface. The use-case-driven approach is employed for this item, and three representative use cases, including channel state information (CSI) feedback enhancement, beam management, and position accuracy enhancement, are selected to investigate the performance, complexity, and potential specification impacts of the AI-native air interface [5].

Herein, we focus on the first use case, that is, AI for CSI feedback enhancement. Massive multiple-input and multiple-output (MIMO), which equips the base station (BS) with many antennas, can considerably improve the system performance. However, the benefit of massive MIMO is based on the knowledge of downlink CSI. In frequency division duplexing (FDD) massive MIMO systems, the user equipment (UE) must feedback the downlink CSI to the BS through the uplink because of the lack of channel reciprocity. The feedback overhead is substantial due to the high dimension of the CSI in massive MIMO systems. Recently, AI has been introduced to CSI feedback by [6]. AI-enabled CSI feedback learns to automatedly compress and reconstruct CSI and considerably improves the feedback accuracy compared with codebook- and compressive sensing (CS)-based feedback algorithms [7]. AI-enabled CSI feedback has been selected as a use case in [5] because of its excellent performance, as supported by approximately 50 companies. However, many problems should be clarified and solved before deploying AI-enabled CSI feedback to practical systems. For example, most of the existing works are based on simulation and cannot be fairly compared with conventional feedback methods [8] in accuracy, complexity, and generalization. Moreover, the standardization of AI-native air interface technology has not been considered and discussed by the 3GPP, and developing a protocol for autoencoder-based CSI feedback is a challenge.

For an overview of AI-enabled CSI feedback, we refer the interested readers to [7]. In contrast to [7], this paper introduces AI-enabled CSI feedback enhancement from the industrial perspective and the challenges and open problems in the deployment of AI-enabled CSI feedback, especially in the standardization of 5G-Advanced. The rest of the article is organized as follows. Section II introduces and compares the representative non-AI and AI-enabled CSI feedback frameworks. Then, the scope of AI for CSI feedback enhancement in 5G-Advanced is briefly presented and discussed in Section III. The main considerations in the standardization of AI for CSI feedback enhancement in 5G-Advanced, including performance evaluation under 5G new radio (NR) simulation assumptions and new standardization challenges posed by AI,

are then provided and discussed in Section III. Finally, Section IV concludes this article.

II. REPRESENTATIVE NON-AI AND AI-ENABLED CSI FEEDBACK FRAMEWORKS

In this section, non-AI feedback frameworks (including codebook- and CS-based feedback) and their main shortcomings are introduced. Then, three representative AI-enabled CSI feedback frameworks are presented, respectively. Finally, the non-AI and AI-enabled frameworks are compared in terms of feedback accuracy and overhead.

A. Non-AI CSI Feedback Frameworks

The existing non-AI feedback strategies can be divided into two main kinds: codebook- and CS-based frameworks.

1) Codebook-based CSI Feedback: Codebook-based feedback is adopted in existing communication systems [8]. In this framework, the UE searches the nearest codeword in a predefined codebook, which is shared by the UE and the BS. Then, the UE feedbacks the index of the selected codeword to the BS. The BS obtains the corresponding codeword by looking up the codebook.

This feedback faces two main challenges in massive MIMO systems. On the one hand, the codebook resolution and feedback accuracy are improved with the codebook size, that is, the feedback bit number. In massive MIMO systems, the channel dimension is high, and many bits are needed to improve the feedback accuracy. On the other hand, the algorithm complexity increases with the codebook size.

2) CS-based CSI Feedback: In CS theory, the original signal can be exactly reconstructed from the compressed signal if sparsity exists. CS-based CSI feedback has been widely studied in the past 10 years. The basic assumption of this type of feedback is the CSI sparsity in a certain domain, such as the angular and delay domain. On the basis of this assumption, the downlink CSI is compressed by a sensing matrix and then reconstructed by some CS algorithms.

The main concerns of the CS-based feedback are the sparsity assumption and the complexity of the reconstruction algorithms. First, the sparsity characteristic cannot fully describe the structure of the CSI matrix. Second, the reconstruction problem in CS is usually solved by some iterative algorithms. The time complexity of these reconstruction algorithms is extremely high to meet the requirement of practical systems.

B. AI-enabled CSI Feedback Frameworks

Many novel AI-enabled technologies have been introduced and studied in the past few years. However, these technologies cannot be immediately deployed in practical systems. The technologies that need no/few modifications for application in existing systems are prioritized in 5G-Advanced. Meanwhile, the technologies that need to totally change the existing systems, such as autoencoder-based end-to-end communications, will be studied in 6G and beyond. Reference [3] divided the technologies in AI-native air interface into three types: replacing single communication modules, designing several modules
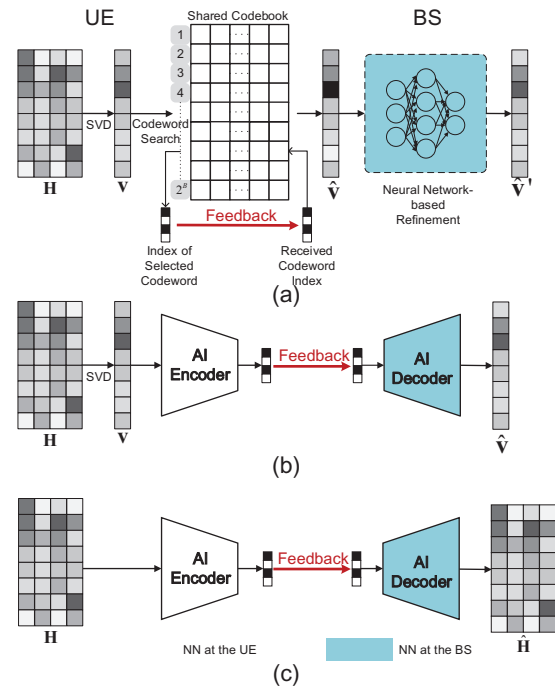

Fig. 1. Three representative frameworks of AI for CSI feedback, including (a) one-sided refinement for implicit CSI feedback, (b) two-sided enhancement for implicit CSI feedback, and (c) two-sided enhancement for explicit CSI feedback.

jointly with AI, and totally changing the existing block-by-block communication framework. On the basis of [3], the existing frameworks of AI for CSI feedback enhancement are divided into three types, namely, one-sided refinement and two-sided enhancement for implicit CSI feedback and two-sided enhancement for explicit CSI feedback, as shown in Fig. 1.

1) AI-enabled One-sided Refinement for Implicit CSI Feedback: Fig. 1(a) shows the framework of the AI-enabled one-sided refinement for CSI feedback, which need not change the existing feedback framework and is easy to deploy [9]. Once the BS is installed, the environment around the BS remains stable for a long time. Given that wireless channels greatly depend on the propagation environment, the channel of a certain cell shows certain characteristics that can be regarded as environmental knowledge. AI is a good tool that extracts and utilizes environmental knowledge to help CSI feedback via end-to-end learning.

Specifically, the UE first performs singular value decomposition on the downlink CSI ($\mathbf{H}$) to produce the eigenvectors, that is, the precoding matrix ($\mathbf{V}$) [10]. The UE feedbacks the precoding matrix with the codebook shared by the BS. Upon receiving the index of the selected precoding codeword, namely, the precoding matrix indicator (PMI), the BS selects the corresponding codeword through the shared codebook. Then, the pretrained NN refines the obtained codeword. The accuracy of the refined precoding matrix is considerably improved with the help of environmental knowledge. The training and inference operations of NNs are conducted at the BS without the participation of the UE. In the training phase, the input and output of the NN are the selected channel codeword ($\hat{\mathbf{v}}$) and the perfect precoding matrix ($\mathbf{v}$), respectively.

In addition, the NN module can be also added on the UE side. As mentioned in Section II-A1, the complexity of the codeword search is the major challenge in codebook-based CSI feedback. Therefore, adding an NN module at the UE side to help codeword search is a promising approach that is worth exploring.

2) Autoencoder-based Two-sided Enhancement for Implicit CSI Feedback: Fig. 1(b) shows the autoencoder-based two-sided framework for implicit CSI feedback enhancement [10]. Unlike that in Fig. 1(a), this framework must change the existing feedback strategy. Inspired by autoencoder-based image compression, an NN-based encoder is adopted at the UE to compress and quantize the generated precoding matrix. We can regard the generated bitstream in this framework as the PMI in the existing codebook-based feedback strategy. After obtaining the feedback bitstream, the NN-based decoder reconstructs the original precoding matrix.

3) Autoencoder-based Two-sided Enhancement for Explicit CSI Feedback: Fig. 1(c) shows the autoencoder-based two-sided framework for explicit CSI feedback enhancement [6]. In contrast with the above-mentioned frameworks, this third framework feedbacks the entire downlink CSI, $\mathbf{H}$, to the BS via an NN-based encoder. Then, the decoder at the BS reconstructs the original CSI on the basis of the received bitstream. Given that all information of the downlink CSI is fed back, the feedback overhead of this framework is larger than that of the second framework. This framework is the most widely studied in academia.

4) Comparison among Three Frameworks: The three frameworks presented above have been introduced to CSI feedback enhancement, and their main differences include their feedback information form and effects on the air interface. Implicit feedback is adopted in the first two frameworks, while the last one adopts an explicit feedback strategy. Implicit feedback is used in 5G NR.

Explicit feedback may bring more performance gains than implicit feedback. However, the utilization of the downlink CSI in the entire system should be changed. The CSI utilization in practical systems is based on partial CSI. In addition, AI-enabled explicit feedback cannot be directly compared with the feedback strategy adopted in practical systems. AI-based implicit feedback enhancement frameworks can be evaluated and compared with the Type II codebook.

The first framework, which refines the obtained implicit CSI at the BS, adopts the existing feedback strategy. The last two frameworks need to replace the original codebook-based coding and decoding with the NN-based encoder and decoder, respectively. Specifically, the existing feedback standard should be completely changed. Accordingly, we believe that these three frameworks will be deployed in practical systems at different stages. The first framework is plug-and-play and can be embedded into the existing BS without standardization. The second framework, which feedbacks the implicit CSI with an autoencoder, will be introduced in 5G-Advanced. The last framework, which completely changes the CSI feedback and utilization strategy, will be deployed in 6G and beyond.

TABLE I
GCS COMPARISON BETWEEN THE SECOND AI-ENABLED FRAMEWORK
AND TYPE II CODEBOOK [10]

Methods	Feedback Bit Number	GCS
Type II Codebook	300	0.9042
AI-enabled feedback	78	0.8507
	104	0.8721
	156	**0.9044**
	208	0.9221
	312	0.9448
	416	0.9574

C. Performance Comparison between Non-AI and AI-enabled CSI Feedback Frameworks

Implicit feedback is adopted in 5G NR. Therefore, we first compare the AI-enabled implicit CSI feedback with the Type II codebook adopted in 5G NR. Table I shows the generalized cosine similarity (GCS) performance comparison between the second AI-enabled framework and the Type II codebook. The simulation settings satisfy the simulation requirement of the 3GPP, and the simulation details can be found in [10]. The table indicates that the AI-enabled implicit feedback still outperforms the Type II codebook when the feedback overhead is approximately reduced by 50%.

As mentioned in Section II-B3, the third framework, that is, the autoencoder-based explicit CSI feedback, is the most widely studied in academia. Therefore, this type of feedback has been compared with conventional CS-based feedback algorithms, as shown in [6]. Table II shows the normalized mean-squared error (NMSE) performance of the CS-based algorithms and the AI-enabled methods, including CsiNet [6] and TransNet [11]. TransNet adopts more advanced NNs than the vanilla CsiNet. Table II shows that the conventional algorithms perform much worse than the AI-enabled methods under all compression ratios. The AI-enabled methods still perform well when the conventional algorithms do not work at all. For example, if the compression ratio is set to 1/32 and the indoor scenario is considered, the best NMSE achieved by the conventional algorithms is only -0.27 dB, while that of TransNet reaches -10.49 dB.

The high feedback accuracy achieved by AI is due to its ability of learning CSI distribution from data samples. The CSI distribution is highly dependent on the propagation environment and can be regarded as a kind of environmental knowledge [9], which is extracted by NNs and utilized during compression and reconstruction as shown in Fig. 2. The non-AI methods design the feedback strategy on the basis of an assumed distribution, which cannot well match the practical systems and leads to poor performance.

Except for the advantage in feedback accuracy, the inference speed of AI-enabled CSI feedback is higher than that of the conventional algorithms. For example, CsiNet performed over 50 times faster than the CS-based algorithms in [6]. The inference time of the AI-enabled implicit CSI feedback in [10] is below 1 ms.

III. STANDARDIZATION OF AI FOR CSI FEEDBACK ENHANCEMENT IN 5G-ADVANCED

The transceivers of different companies can use different technologies and need not be shared. However, the downlink

TABLE II
NMSE dB PERFORMANCE COMPARISON OF DIFFERENT EXPLICIT CSI
FEEDBACK ALGORITHMS [6], [11]

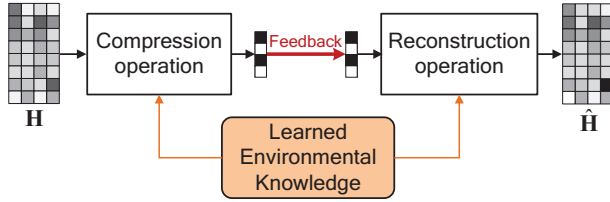| compression ratio	Methods	Indoor	Outdoor
1/4	LASSO	−7.59	−5.08
	BM3D-AMP	−4.33	−1.33
	TVAL3	−14.87	−6.90
	CsiNet	−17.36	−8.75
	TransNet	**−32.38**	**−14.86**
1/16	LASSO	−2.72	−1.01
	BM3D-AMP	0.26	0.55
	TVAL3	−2.61	−0.43
	CsiNet	−8.65	−4.51
	TransNet	**−15.00**	**−7.82**
1/32	LASSO	−1.03	−0.24
	BM3D-AMP	24.72	22.66
	TVAL3	−0.27	0.46
	CsiNet	−6.24	−2.81
	TransNet	**−10.49**	**−4.13**
1/64	LASSO	−0.14	−0.06
	BM3D-AMP	0.22	25.45
	TVAL3	0.63	0.76
	CsiNet	−5.84	−1.93
	TransNet	**−6.08**	**−2.62**

Fig. 2. Illustration of CSI compression and reconstruction with the aid of learned environmental knowledge

CSI feedback in the air interface should be kept the same for all UE and BSs, and the key modules in the CSI feedback must be standardized. Therefore, many problems should be considered and clarified. In this section, the scope of AI for CSI feedback enhancement is presented. Then, the main challenges and open problems during the standardization are presented. The 3GPP focuses on standardizing key frameworks, which all companies must follow. The detailed techniques need not be standardized, and companies can use their own techniques. Therefore, on the basis of the scope and goal of this paper, we focus on standardization and do not discuss the techniques in detail.

A. Scope of AI for CSI Feedback Enhancement in 5G-Advanced

The main scope of AI for CSI feedback enhancement in the AI-native air interface study item of the 3GPP Release 18 [5] involves overhead reduction, accuracy improvement, and channel prediction.

1) Overhead Reduction: The feedback accuracy of codebook-based CSI feedback improves but the feedback overheads increase. In a 5G NR, the Type II codebook is adopted in the multi-UE scenario, where a highly accurate CSI is needed to reduce the interference among different UE compared with the Type I codebook adopted in the single-UE scenario [8]. For example, the feedback bit number is

approximately 300 when the respective numbers of antennas of the BS and the UE are 32 and 4, and the sub-band number is 13 (additional details can be found in [10]). This feedback overhead consumes substantial resources of the uplink control channel. Therefore, AI-enabled CSI feedback is expected to reduce the feedback overhead without a decline in accuracy.

2) Accuracy Improvement: The downlink performance greatly depends on the quality of the downlink CSI obtained at the BS. However, the feedback accuracy in practical systems does not meet the requirement of future communication systems. GCS is widely used to evaluate CSI feedback performance. According to [10], the GCS of the Type I codebook adopted in the single-UE scenario is only 0.76. The accuracy of the codebook-based feedback is low, unavoidably leading to a throughput drop in downlink transmission. Therefore, an AI-enabled feedback method that considerably improves the feedback accuracy without overhead increase must be developed.

3) Channel Prediction: The speed of the UE mobility is high in some cases. For example, the speed of a high-speed train is up to 350 km/h, where the propagation environment rapidly changes. In these high-speed cases, channel aging seriously negatively affects the downlink transmission performance even when the downlink CSI is perfectly fed back to the BS. Moreover, feedback is no longer needed if the difference between the estimated CSI at the UE and the predicted CSI at the BS is acceptable. Therefore, channel prediction should be considered when AI is introduced to CSI feedback enhancement.

B. Main Challenges and Open Problems during Standardization

Many challenges and problems must be articulated and solved during the standardization phase. First, the performance of AI-enabled CSI feedback must be evaluated and fairly compared with that of the existing methods. Sections III-B-(1-3) introduce the performance evaluation of AI for CSI feedback enhancement in 5G-Advanced. Then, AI-enabled CSI feedback creates new requirements for standardization, and the corresponding protocols must be carefully designed for deployment. Sections III-B-(4-7) introduce the collaboration between the UE and the BS, the shared information between different companies, the joint design with channel prediction, and the combination with reciprocity-based feedback.

1) Accuracy Evaluation of the AI-enabled CSI Feedback: Most of the existing studies on AI for CSI feedback enhancement were conducted on the COST 2100 channel dataset shared by [6]. This channel dataset is simple and far from the practical systems, and thus cannot be used to compare the AI-based feedback with the existing systems. The accuracy evaluation of CSI feedback should be fair and easily show the potential of AI-based CSI feedback enhancement.

First, system-level simulation is preferred over link-level simulation. To the best of our knowledge, link-level simulation was adopted in nearly all of the existing work. Consequently, the performance of AI for CSI feedback enhancement in systems is unclear. System-level simulation is beneficial in

achieving a comprehensive performance in throughput in a single- or multi-cell perspective. Moreover, system-level performance is widely adopted in 5G NR MIMO enhancement and near the complicated practical systems. The performance of AI-enabled CSI feedback can be easily compared with the existing 5G NR technology. Given that system-level simulation is time-consuming, link-level simulation can be optionally adopted to evaluate the novel works on CSI feedback.

Second, as previously mentioned, the COST 2100 channel dataset is used to evaluate the CSI feedback. However, this channel model is not adopted in 5G NR and cannot be used to compare the AI-enabled methods with the Type I and Type II codebooks in 5G NR. Consequently, the agreed 3GPP channel models should be used, such as TR 38.901. Moreover, the channel settings adopted in the CSI feedback enhancements of Releases 16 and 17 can be directly used, facilitating the comparison of AI-enabled CSI feedback with the latest codebook in 5G NR, that is, the enhanced Type II codebook. In addition, the downlink channel at the UE was assumed perfect, that is, an ideal downlink channel estimation, in most works. The realistic channel estimation must be considered during feedback to evaluate its effects on AI-based feedback.

Finally, NMSE and GCS are the most widely used key performance indicators (KPIs) in AI-enabled CSI feedback work [6], [7]. Given that system-level simulation is preferred, the KPIs consist of two types: intermediate and eventual KPIs. The first type is adopted to compare the feedback accuracy of different NN models. NMSE, GCS, and square GCS can be used in this part. The second type evaluates the performance of the entire system and is used to compare the AI methods with the conventional methods. The throughput and the overhead can be used in this case.

Overall, the accuracy evaluation of AI-enabled CSI feedback should be fair to show the potential of AI methods compared with non-AI methods under the agreed 3GPP assumption [12].

2) Complexity of the AI-enabled CSI Feedback: Communication algorithms have a high requirement in inference latency and computational complexity [13]. Although AI-enabled CSI feedback promises to outperform conventional algorithms, its computational complexity is a major challenge because of the following reasons. First, the performance of AI-enabled methods is improved by stacking many NN layers or widening the NN layers. Performance improvement is achieved with a considerable increase in NN complexity. Second, all simulations are conducted on a computer/workstation equipped with several powerful graphics processing units (GPUs), which have more computational power than the computing units of the UE. Finally, the NN weights must be stored, thereby occupying memory storage in the UE. Therefore, the complexity should be evaluated and reduced during the deployment of AI-enabled CSI feedback.

A major problem is evaluating the complexity of AI-based CSI feedback. The number of floating-point operations (FLOPs) is widely adopted in many domains, including CSI feedback. The FLOP numbers of the encoder at the UE and the decoder at the BS for the two-sided NN model must be evaluated. The BS has better computational power than

the UE. Accordingly, the NN model at the UE must be lightweight. Although the number of FLOPs is the main KPI of the computational complexity of NNs, it does not fully reflect the inference speed of NNs in computing units. The memory access cost and the degree of parallelism are not considered in the FLOP number. In terms of the NNs with the same FLOP number, the one with less memory access and a higher degree of parallelism can be much faster than the one with more memory cost and a lower degree of parallelism. Consequently, the FLOP number is the indirect KPI of the NN inference speed, and the discrepancy between the FLOP number and the NN inference speed is unavoidable. The memory access cost and the degree of parallelism must also be considered in the NN speed evaluation.

The memory storage of the NN model must also be evaluated. Two parts must be considered when assessing the memory storage occupy, namely, the NN model size and the NN weights. The quantization bit of NN weights also affects the storage requirement. For example, the binary weights need less storage than the 32-bit floating point ones.

3) Generalization of the AI-enabled CSI Feedback: As mentioned in Section II-C, the excellent performance of AI-enabled CSI feedback can be attributed to the ability of learning knowledge from the data through an end-to-end approach. The NN model can be regarded as overfitting in a certain distribution of the training dataset. If the distribution of the training dataset is the same as (or near) that of the deployment scenario, then the trained NN model efficiently works. However, the performance unavoidably drops if the model is deployed to the systems with a different distribution, resulting in an NN generalization problem. The key problem that must be clarified and solved in NN generalization is assessing and improving the generalization ability of the feedback NN model.

The common evaluation method is to assess the NN model with the test dataset, where the channel model (such as CDL-A) is different from that (such as CDL-C) adopted in the training dataset. The generalization performance heavily depends on the generated test dataset. An NN model may perform rather differently in a different test dataset. For example, if the distributions of two different channel models are similar, then the performance drop may be very small. Therefore, a public and general test dataset is essential.

Two strategies are expected to improve the generalization of the CSI feedback NN model, namely, training dataset mix and online learning. The generalization errors in the feedback NN model result from the mismatch between the distributions of the training and test datasets. If the training dataset covers the distribution of the test dataset, then the generalization errors are small. Therefore, generating a training dataset that covers the entire distribution can considerably improve the generalization ability of the feedback NN model. The mixed dataset-based strategy is developed based on this observation. The CSI samples generated with different channel models are mixed to form a large training dataset. However, the mixed training dataset cannot cover all distributions. If a new distribution appears, then online learning is required.

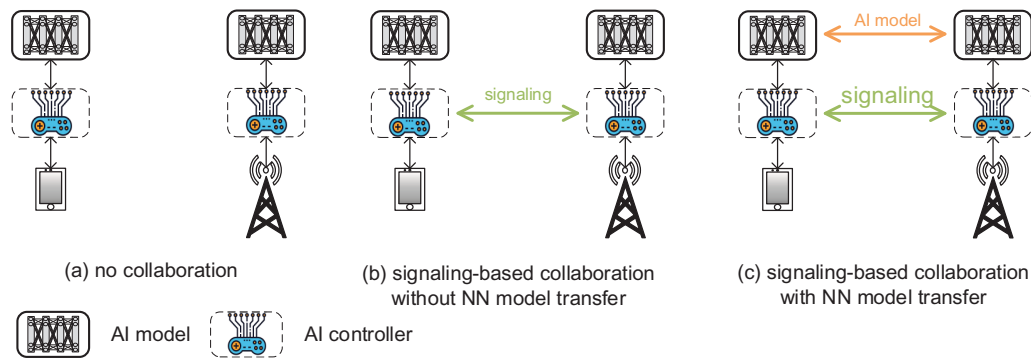Online learning involves not only the air interface but also

Fig. 3. Three collaborations between the UE and the BS of the AI-enabled CSI feedback.

the upper layer. The protocol of online learning for the one-sided or two-sided feedback NN models should be carefully designed. The protocol needs to define when to conduct online learning, how to collect CSI samples, which UE must participate in the online learning, how to allocate the computational power, and so on. The protocol for online learning is a general challenge in an AI-native air interface. In addition, more advanced learning tools, such as transfer learning and meta-learning, should be adopted to accelerate the online learning and reduce the training overhead. In transfer learning, the NNs are finetuned with new CSI samples rather than trained from scratch. Meta-learning provides a better initialization of the NN weights and helps the NNs quickly adapt to a new task (distribution), thereby considerably reducing the training epochs and the requirement for online training samples [14].

4) Collaboration between the UE and the BS of the AI-enabled CSI Feedback: In 5G NR, all UE and BSs share the same codebook, and extra collaboration between the BS and the UE is not needed. However, collaboration is indispensable in AI-enabled CSI feedback. We divide the collaboration into three types, as shown in Fig. 3.

- *No collaboration:* In this case, the AI model is deployed on one side (only UE or BS), and few standardization impacts are observed. The AI-enabled one-sided refinement-based framework in Section II-B1 is a representative use case of this type. However, in all AI-enabled feedback algorithms, the NN model is trained with the collected CSI samples. Some studies proposed to train the feedback NNs with the uplink CSI samples on the basis of the assumption that the bidirectional channels share the same distribution. However, this training strategy has not been verified in practical systems. If the distribution reciprocity does not hold, then the CSI dataset should be transmitted from the UE to the BS, where collaboration is necessary.
- *Signaling-based collaboration without NN model transfer:* This type of collaboration mainly includes training dataset collection and signaling-based NN model management. As mentioned in the first type, signaling-based collaboration is required if the UE needs to transmit the stored CSI samples to the BS. In addition, the UE may store several NN models for different scenarios, and the BS needs to send some signals to the UE for NN model switching or activation.
- *Signaling-based collaboration with NN model transfer:* This collaboration includes all collaboration operations

in the second mechanism. The difference between them is that the NN model in the last type must be transmitted among the BS and the UE. Given the dataset privacy, the NN model for the CSI feedback may need to be trained with federated learning or model splitting. Federated learning is a decentralized learning framework, where the UE upload the local models to the server at the BS, and the BS aggregates the received NN model and then transmits the aggregated NN model to the UE. In model splitting, the encoder and the decoder are trained at the UE and the BS, respectively, and the gradients are transmitted between the BS and the UE. In addition, the BS may train a better NN model because more CSI samples are collected after deployment. The new NN model must be transmitted to all the UE in the cell.

The above-mentioned three collaboration mechanisms can be mixed and used according to actual communication scenarios.

5) Shared Information of the AI-enabled CSI Feedback: Intellectual property is essential in standardization. The company, whose technology is agreed in the 3GPP and deployed to practical systems, can earn huge economic profits from the technology. Therefore, the 3GPP must clarify the content that should be standardized in AI-enabled CSI feedback. As agreed by most companies, the NN architecture does not need to be standardized, and each company can design and train a feedback NN model with their own dataset.

The one-sided NN-based CSI feedback at the BS (UE) can work well even if the other side, that is, the UE (BS), has no knowledge of the NN model. However, the encoder at the UE and the decoder at the BS should be trained together for the two-sided autoencoder-based CSI feedback. If the encoder and the decoder do not match, then the feedback accuracy will considerably drop, resulting in a great decrease in system throughput. In image compression, some recent studies have attempted to improve the compression performance by designing the encoder (compression module) without changing the original JPEG-2000 decoder. Their experimental results show that the performance can be considerably improved only when the encoder can be edited. On this basis, the manufacturers of the BSs can release their decoder, and the manufacturers of the UE can design the encoder on the basis of the released decoder. The UE requires storing several encoders and switching the encoder according to the BS types. However, the model download is needed if the decoder at the BS changes, as
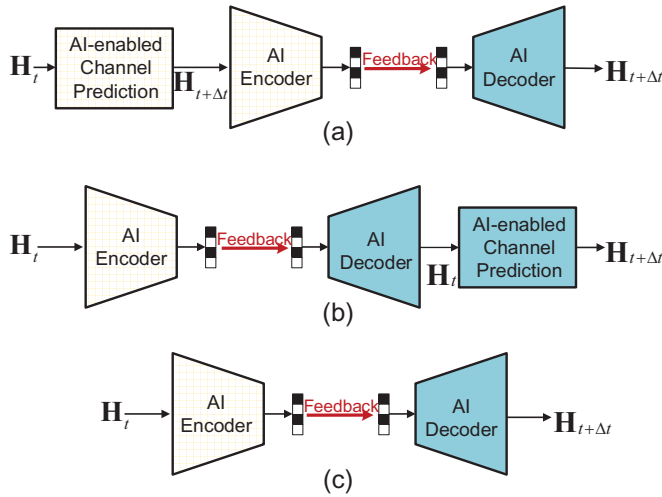
Fig. 4. Three frameworks for joint AI-enabled CSI feedback and prediction. The first two NN-based prediction modules work at the UE and the BS, and the last one implicitly realizes channel prediction.

mentioned in Section III-B4.

6) Joint Design with Channel Prediction: As mentioned in Section III-A3, channel prediction is being studied in Release 18 because of the channel aging in the high-speed scenario. Although the channel rapidly changes, the correlation among the adjacent CSI still holds. The change within a short time can be predicted. Recently, AI-based channel prediction has been explored in some studies. If the difference between the predicted and estimated CSI at the UE is tolerable, then feedback is no longer needed. The BS uses the predicted CSI to design precoding vectors. In this case, the BS and the UE share the same CSI predictor. If the difference is large, three types of design will emerge when the feedback and prediction are jointly built as shown in Fig. 4. We use the AI-enabled two-sided explicit CSI feedback as an example. In the first joint design frameworks, the UE first predicts the future channel, $\mathbf{H}_{t+\Delta t}$, based on the current channel, $\mathbf{H}_t$. Then, the predicted channel is fed back to the BS by an autoencoder. In the second one, the current channel, $\mathbf{H}_t$, is directly fed back to the BS by an autoencoder. The reconstructed channel, $\hat{\mathbf{H}}_t$, is then used as input to a channel prediction module to reduce the effects of channel aging. The above-mentioned frameworks jointly design the feedback and prediction explicitly, where two modules, namely, channel feedback and prediction modules, are adopted. The last framework jointly designs the channel feedback and the prediction implicitly. This framework seems to be the same as that in Fig. 1(c). The main difference is the training phase, where the input is the current channel, and the output of the autoencoder (decoder) is the future channel. The framework embeds channel prediction into the feedback process.

The main advantage of the first framework is that the UE also obtains the future CSI, which can be used in the design of the receiver. However, the extra module increases the computational complexity at the UE. In the second framework, the increase in complexity can be ignored because of the large computational power at the BS. No extra modules are added in the third framework, and the main concern is the lack of interpretability. The practical system must select the most suitable framework according to the application scenario.

7) Combination with the Enhanced Reciprocity-based Feedback: In 3GPP Release 17, the exploitation of the angle-delay reciprocity between the uplink and downlink channels is considered and discussed to improve the performance (including the accuracy and overhead) of the enhanced Type II codebook in 3GPP Release 16 [15]. Given that the angle-delay information of the bidirectional channels is reciprocal, the UE need not feedback this information to the BS, thereby reducing feedback overhead and computational complexity.

Accordingly, AI-enabled CSI feedback enhancement also requires studying the angular-delay reciprocity. The key problem is introducing the uplink channel information to the downlink CSI reconstruction at the BS, thereby reducing the feedback of the information shared by the bidirectional channels. Meanwhile, channel estimation is the basis of the CSI feedback. If less information about the downlink channel must be fed back, then the channel estimation does not estimate the information that is not fed back. Therefore, the pilot overhead can be considerably reduced if the channel estimation and feedback are jointly designed on the observation of the angular-delay reciprocity.

8) Others : During deployment, more challenges and problems must be addressed. First, some new use cases, such as reconfigurable intelligent surfaces, are expected to play an important role in the future. Some new characteristics may appear in these cases. The combination of AI-enabled CSI feedback with these new cases is worth exploring. Second, the computational powers of different UE or a specific UE at different times vary greatly. Therefore, feedback NNs should be scalable with different complexities instead of training and storing many NNs with various complexities. Third, the performance of AI-enabled CSI feedback should be evaluated with the CSI samples collected from the practical systems before deployment.

IV. CONCLUSIONS

In this article, an overview of the AI for CSI feedback enhanced in 5G-Advanced is provided. First, non-AI CSI feedback frameworks and three representative frameworks of AI-enabled CSI feedback enhancement are introduced and compared. Then, the scope of AI for CSI feedback enhancement is introduced and discussed. Finally, the considerations in the standardization of AI for CSI feedback enhancement, especially focusing on accuracy, complexity, collaboration, generalization, information sharing, joint design with channel prediction, and reciprocity, are elaborated.

As pointed out in [5], the goal of studying three representative use cases is to lay the foundation for future air interface use cases leveraging AI techniques. The above-mentioned challenges will be addressed by introducing advanced AI techniques and communication expert knowledge. The thorough study of AI for CSI feedback enhancement will guide the development of AI-native air interface in 5G-Advanced and 6G.

REFERENCES

[1] W. Chen, J. Montojo, J. Lee, M. Shafi, and Y. Kim, "The standardization of 5G-Advanced in 3GPP," *IEEE Commun. Mag.*, vol. 60, no. 11, pp. 98–104, Nov. 2022.

[2] X. Lin, "An overview of 5G Advanced evolution in 3GPP Release 18," *IEEE Commun. Stds. Mag.*, vol. 6, no. 3, pp. 77–83, Sep. 2022.

[3] J. Hoydis, F. A. Aoudia, A. Valcarce, and H. Viswanathan, "Toward a 6G AI-native air interface," *IEEE Commun. Mag.*, vol. 59, no. 5, pp. 76–81, May 2021.

[4] X. Lin, M. Chen, H. Rydén *et al.*, "Fueling the next quantum leap in cellular networks: Embracing AI in 5G evolution towards 6G," *arXiv preprint arXiv:2111.10663*, 2021.

[5] 3GPP RP-213599, "New SI: Study on artificial intelligence (AI)/Machine Learning (ML) for NR air interface," Moderator (Qualcomm), Tech. Rep., Dec. 2021, Accessed on Aug. 29, 2022. [Online]. Available: https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_94e/Docs/RP-213599.zip

[6] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, Oct. 2018.

[7] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Overview of deep learning-based CSI feedback in massive MIMO systems," *IEEE Trans. Commun.*, 2022, Early access.

[8] 3GPP TS 38.214, "Physical layer procedures for data (Release 16)," Tech. Rep., Jul. 2020, Accessed on Aug. 29, 2022. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/16.02.00_60/ts_138214v160200p.pdf

[9] J. Guo, C.-K. Wen, M. Chen, and S. Jin, "Environment knowledge-aided massive MIMO feedback codebook enhancement using artificial intelligence," *IEEE Trans. Commun.*, vol. 70, no. 7, pp. 4527–4542, Jul. 2022.

[10] M. Chen, J. Guo, C.-K. Wen, S. Jin, G. Y. Li, and A. Yang, "Deep learning-based implicit CSI feedback in massive MIMO," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 935–950, Feb. 2022.

[11] Y. Cui, A. Guo, and C. Song, "TransNet: Full attention network for CSI feedback in FDD massive MIMO system," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 903–907, May 2022.

[12] 3GPP, "Session notes for 9.2 (Study on AI/ ML for NR air interface)," Ad-hoc Chair (CMCC), Tech. Rep., Jun. 2022, Accessed on Aug. 29, 2022. [Online]. Available: https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_109-e/Inbox/Xiaodong_sessions/Xiaodong's%20Session%20Notes%20RAN1%23109-e%20(9.2%20AI%26ML)%20v11.zip

[13] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6G," *IEEE Wireless Commun.*, vol. 29, no. 4, pp. 140–145, Aug. 2022.

[14] M. Jung and W. Saad, "Meta-learning for 6G communication networks with reconfigurable intelligent surfaces," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 8082–8086.

[15] 3GPP TR 21.917, "Release 17 description; Summary of Rel-17 work items (v0.6.0)," Tech. Rep., Jun. 2022, Accessed on Aug. 29, 2022. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3937

Jiajia Guo [S'21] (jiacloud2020@outlook.com) received the B.S. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2016, and the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2019. He is currently pursuing the Ph.D. degree in information and communications engineering with Southeast University, Nanjing, China. His current research interests include deep learning for physical layer, distributed learning, and neural network compression.

Chao-Kai Wen [SM'20] (chaokai.wen@mail.nsysu.edu.tw) received the Ph.D. degree from the Institute of Communications Engineering, National Tsing Hua University, Taiwan, in 2004. He was with Industrial Technology Research Institute, Hsinchu, Taiwan, and MediaTek Inc., Hsinchu, Taiwan, from 2004 to 2009, where he was engaged in broadband digital transceiver design. In 2009, he joined the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, where he is currently Professor. His research interests center around the optimization of wireless multimedia networks.

Shi Jin [SM'17] (jinshi@seu.edu.cn) received his Ph.D. degree in communications and information systems from Southeast University in 2007. From June 2007 to October 2009, he was a research fellow with the Adastral Park Research Campus, University College London, United Kingdom. He is currently with the faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space-time wireless communications, information theory, intelligent communications, and RIS. He and his coauthors received the 2010 Young Author Best Paper Award from the IEEE Signal Processing Society and the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory.

Xiao Li [M'10] (li_xiao@seu.edu.cn) received the Ph.D. degree in communication and information systems from Southeast University, Nanjing, China, in 2010. She then joined the School of Information Science and Engineering, Southeast University, where she has been a Professor in information systems and communications since July 2020. From January 2013 to January 2014, she was a Postdoctoral Fellow at The University of Texas at Austin, Austin, TX, USA. Her current research interests include Reconfigurable intelligent surface assisted communications and intelligent communications. She serves as an Associate Editor for the IEEE Wireless Communications Letters and Electronics Letters, and a Guest Editor for the Digital Communications and Networks.