Accurate Channel Prediction Based on Transformer: Making Mobility Negligible

Hao Jiang, *Student Member, IEEE*, Mingyao Cui, *Student Member, IEEE*, Derrick Wing Kwan Ng, *Fellow, IEEE*, and Linglong Dai, *Fellow, IEEE*

***Abstract*— Accurate channel prediction is vital to address the channel aging issue in mobile communications with fast time-varying channels. Existing channel prediction schemes are generally based on the sequential signal processing, i.e., the channel in the next frame can only be sequentially predicted. Thus, the accuracy of channel prediction rapidly degrades with the evolution of frame due to the error propagation problem in the sequential operation. To overcome this challenging problem, we propose a transformer-based parallel channel prediction scheme to predict future channels in parallel. Specifically, we first formulate the channel prediction problem as a parallel channel mapping problem, which predicts the channels in next several frames in parallel. Then, inspired by the recently proposed parallel vector mapping model named transformer, a transformer-based parallel channel prediction scheme is proposed to solve this formulated problem. Relying on the attention mechanism in machine learning, the transformer-based scheme naturally enables parallel signal processing to avoid the error propagation problem. The transformer can also adaptively assign more weights and resources to the more relevant historical channels to facilitate accurate prediction for future channels. Moreover, we propose a pilot-to-precoder (P2P) prediction scheme that incorporates the transformer-based parallel channel prediction as well as pilot-based channel estimation and precoding. In this way, the dedicated channel estimation and precoding can be avoided to reduce the signal processing complexity. Finally, simulation results verify that the proposed schemes are able to achieve a negligible sum-rate performance loss for practical 5G systems in mobile scenarios.**

***Index Terms*— Channel prediction, error propagation, transformer, attention mechanism, machine learning.**

I. Introduction

MILLIMETER-WAVE (mmWave) massive multiple-input multiple-output (MIMO) has been a key technique for the fifth-generation (5G) wireless communications [1]. Equipped with an array with a large number of antennas, massive MIMO can achieve orders of magnitude increase in the achievable sum-rate [2] through different advanced precoding designs [3].

In fact, effective real-time precoding highly depends on the quality of the estimated instantaneous channel state information (CSI). According to the 5G standard [4], each frame in the time-division duplex (TDD) mode contains multiple slots, and the instantaneous CSI is estimated only in the first slot of each frame by using the predefined sounding reference signal (SRS). Then, the subsequent slots within the same frame can only utilize the estimated channel in the first slot for the precoding design.

Since the channel coherence time is inversely proportional to the carrier frequency and user speed, it is possible that the channel coherence time [5] is shorter than the channel estimation period, i.e, the SRS period [4], in mobile scenarios. For example, when the carrier frequency of 28 GHz and the user speed of 60 km/h, the channel coherence time is roughly 0.32 ms, while the smallest SRS period is 0.625 ms according to the 3GPP standard [4]. In such a typical scenario, the actual channels for the second half of the slots in the same frame are likely to have significant changes. This phenomenon is known as channel aging [6], which could result in about 30% achievable sum-rate performance loss with the user speed of 60 km/h [7]. Consequently, channel aging is an essential issue that has to be addressed for mmWave MIMO in mobile scenarios.

A. Prior Works

To alleviate the performance loss caused by channel aging, channel prediction techniques have been extensively studied to predict the future channel by exploiting the temporal correlation between the historical CSI and the future channel [7]–[15]. Specifically, the channel prediction techniques are utilized to predict channels in the next several frames. Since the second half of the slots in the current frame are in the channel coherence time of the predicted channel in the next frame, these slots could perform precoding design according to the predicted channel in the next frame. Furthermore, due to the significant baseband processing delay, which aggravates the channel aging issue, the prediction of the future channels in the next several frames is required. The existing channel prediction methods could be generally divided into two categories, i.e., the model-based methods and the neural network-based methods.

For the model-based methods [7]–[11], several models have been considered to characterize the time-varying channels with

Manuscript received 16 December 2021; revised 17 June 2022; accepted 22 June 2022. Date of publication 19 July 2022; date of current version 19 August 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1807201 and in part by the National Natural Science Foundation of China under Grant 62031019. *(Corresponding author: Linglong Dai.)*

Hao Jiang, Mingyao Cui, and Linglong Dai are with the Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: jiang-h18@mails.tsinghua.edu.cn; cmy20@mails.tsinghua.edu.cn; daill@tsinghua.edu.cn).

Derrick Wing Kwan Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: w.k.ng@unsw.edu.au).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2022.3191334.

Digital Object Identifier 10.1109/JSAC.2022.3191334

0733-8716 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

a set of filter parameters, such as the linear extrapolation model [7], [8], the sum-of-sinusoids model [9], and the autoregressive (AR) model [10], [11]. Unfortunately, due to the multi-path effect and the Doppler effect, practical channels usually evolve over time complicatedly, which makes the existing channel prediction models difficult to match the actual channel.

To this end, neural network-based channel prediction methods have been recently proposed to train a neural network to flexibly match the actual channel in a data-driven way [12]–[15]. Specifically, in [12], the fully-connected neural networks (FCN) were trained to predict the future channel by learning the channel characteristics from the input of historical channels. Whereas, due to the high-dimensional input channels in previous frames, training such an FCN can be intractable. To avoid the high-dimensional input, [13]–[15] employed recurrent neural networks (RNN) to iteratively input historical channels in chronological order.

Nonetheless, the existing neural network-based methods can only predict the channel in the first future frame according to the historical channels in previous frames. As a result, to predict the channel in the second future frame, the predicted channel in the first future frame together with the historical channels have to be jointly served as the input of the trained RNN model for the next prediction. By this means, the future channels in the next several frames can only be predicted one by one over time. This procedure is called sequential channel prediction in this paper, which has been widely considered in existing works [7]–[15].

However, due to the sequential prediction of existing neural network-based methods, the error of channel prediction could be rapidly accumulated and becomes serious with the evolution of frame, which is termed as prediction error propagation. In fact, the prediction error propagation problem caused by sequential prediction seriously degrades the achievable sum-rate performance of existing neural network-based schemes, particularly in mobile mmWave communications. Whereas, to the best of our knowledge, the important channel prediction error propagation problem has not been well studied in the literature.

B. Our Contributions

To avoid the achievable sum-rate loss caused by the error propagation problem, unlike the existing sequential channel prediction schemes, we propose a transformer-based parallel channel prediction scheme by processing historical channels and predicting future channels in parallel[1]. Specifically, the contributions of this paper can be summarized as follows.

- Unlike the classical sequential channel prediction schemes that predict the future channels one by one consecutively, we formulate the channel prediction problem as a parallel channel mapping problem to predict future channels simultaneously. Specifically, the historical channels from past several frames are processed jointly and the future channels in next several frames are predicted at once. In this way, future channels in next several frames can be predicted in parallel, and thus the channel prediction error propagation can be avoided.
- To solve the above parallel channel mapping problem, inspired by the recently proposed parallel vector mapping model named transformer in the field of natural language processing [16], we propose a transformer-based parallel channel prediction scheme in this paper. Specifically, the most important module of the transformer model is the attention mechanism, which is able to establish a parallel mapping between the historical CSI and future channels by simple matrix multiplication. Thus, error propagation does not exist in the proposed problem formulation. Furthermore, as a weighting operation, the attention mechanism can adaptively assign more weights and resources to the historical CSI that are more helpful for predicting future channels. By taking advantage of the above two factors, the transformer-based channel prediction scheme can accurately predict future time-varying channels.
- Finally, we further extend the proposed transformer-based channel prediction scheme to a pilot-to-precoder (P2P) prediction scheme. Specifically, we observe that the inputs of the transformer-based channel prediction model require complicated channel estimation in previous frames, while the output predicted channels have to be utilized for designing future precoders. The channel estimation and precoding design may result in very high computational complexity. As a remedy, P2P prediction can utilize the transformer-based model to jointly perform channel estimation, channel prediction, and precoding by using a single model. This is achieved by replacing the input historical CSI and output predicted channels with the historical received pilots and the predicted precoders in the transformer-based channel prediction model, respectively. In this way, the explicit channel estimation and precoding design do not exist in the P2P prediction, and the associated signal processing complexity can be avoided.

C. Organization and Notation

Organization: The rest of the paper is organized as follows. In Section II, the system model of mmWave massive MIMO is introduced. The problem of parallel channel prediction mapping is formulated in Section III. Then, we elaborate on the proposed transformer-based parallel channel prediction scheme in Section IV, including the framework of the offline training and the online prediction. After that, we further discuss the transformer-based pilot-to-precoder prediction scheme in Section V. Simulation results are provided to verify the advantages of the proposed schemes in Section VI. Finally, conclusions are drawn in Section VII.

Notation: We denote the column vectors and matrices by boldface lower-case and upper-case letters, respectively; $\mathbb{R}^n$ and $\mathbb{C}^n$ denote the n-dimension real number and complex number, respectively; $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^{-1}$, $|\cdot|$, and $||\cdot||$ denote the transpose, conjugate transpose, inverse, determinant, and l_2-norm of a matrix, respectively; $\mathbb{E}\{\cdot\}$ denotes the statistical

[1]Simulation codes are provided to reproduce the results presented in this paper: http://oa.ee.tsinghua.edu.cn/dailinglong/publications/publications.html

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

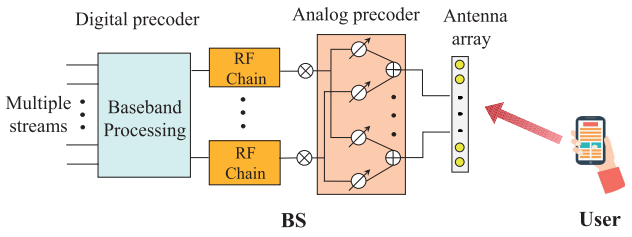


Fig. 1. The architecture of hybrid precoding for mmWave massive MIMO.

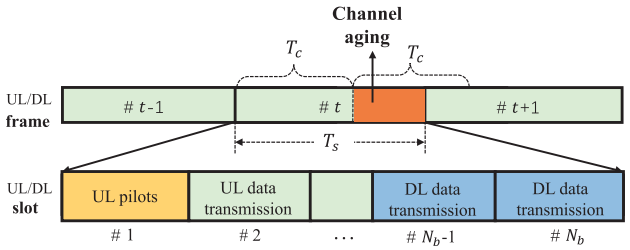


Fig. 2. The frame structure in the TDD mode defined by 5G standard [4].

expectation. x_i denotes the i-th element of the vector $\mathbf{x}$; $\mathbf{X}[i,:]$, $\mathbf{X}[:,j]$, and $\mathbf{X}[i,j]$ denote the i-th row, the j-th column, and the element at i-th row and the j-th column of the matrix $\mathbf{X}$, respectively; $\mathrm{vec}(\mathbf{X})$ denotes the vectorization of the matrix $\mathbf{X}$ and $\mathrm{dvec}(\mathbf{x})$ denotes the inverse process of $\mathrm{vec}(\mathbf{X})$; $\mathbf{I}_N$ denotes an $N \times N$ identity matrix; $\mathbf{X} \otimes \mathbf{Y}$ denotes the Kronecker product of $\mathbf{X}$ and $\mathbf{Y}$; $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary parts of a matrix, respectively; $\mathcal{U}[a,b]$ denotes the probability density function of uniform distribution on $[a,b]$; $\mathbf{tril}(x)$ denotes a lower triangular matrix with each lower triangular element being x.

II. System Model

For simplicity but without loss of generality, we consider a massive MIMO system, where a single base station (BS) with N_{BS} antennas serves a single user with M antennas.[2] To reduce the power consumption, the BS antenna array is realized by a hybrid precoding architecture [17]. In particular, $N_{\mathrm{RF}}N_{\mathrm{BS}}$ analog phase shifters (PSs) with low cost and low power consumption are adopted to reduce the number of radio-frequency (RF) chains from N_{BS} to N_{RF}, as shown in Fig. 1. Moreover, the TDD mode is considered, where the time resource is divided into multiple frames, and each frame contains N_b slots. As shown in Fig. 2, the N_b slots could be divided into three phases, i.e., the uplink channel estimation phase, the uplink data transmission phase, and the downlink data transmission phase. Furthermore, we assume that the channel is block fading, where the channel remains time-invariant in each slot and changes from slot to slot [18], [19][3].

It is well known that the mmWave channel is sparse, since the mmWave propagation environment exhibits a limited number of scattering clusters [20]. Thus, the widely used geometric Saleh-Valenzuela channel model [17] is adopted for describing the mmWave channel. Under this model, the channel from the user to the BS in the t-th frame can be expressed as

$$\mathbf{H}^{(t)} = \sum_{l=1}^{L} \alpha_l e^{-j2\pi f_l t T_s} \mathbf{a}_{\mathrm{BS}}(\phi_{\mathrm{BS},l}) \mathbf{a}_{\mathrm{user}}^H(\phi_{\mathrm{user},l}), \quad (1)$$

where L is the number of propagation paths; T_s is the period of the frame, which equals the SRS period; α_l, f_l, $\phi_{\mathrm{BS},l}$, and $\phi_{\mathrm{user},l}$ are the complex gain, Doppler shift, azimuth angle of arrival (AoA), and azimuth angle of departure (AoD) of the l-th path, respectively. Note that the time-varying channel is mainly caused by the Doppler effect, while the AoA and AoD are time-invariant. This is because the time scale of the frames in the 3GPP standard is in the order of tens of milliseconds [4]. Furthermore, it has been experimentally verified that during such a time scale, the AoA and AoD can be approximately regarded as unchanged [21], [22]. In this paper, we consider the uniform linear arrays (ULAs). Without loss of generality and for ease of presentation, we neglect the subscripts of $\phi_{\mathrm{BS},l}$ and $\phi_{\mathrm{user},l}$. Then, the array steering vector $\mathbf{a}_{\mathrm{BS}}(\phi)$ and $\mathbf{a}_{\mathrm{user}}(\phi)$ could be presented by

$$\mathbf{a}_{\mathrm{BS}}(\phi) = \frac{1}{\sqrt{N_{\mathrm{BS}}}} \left[1, e^{j\frac{2\pi}{\lambda} d\sin(\phi)}, \cdots, e^{j(N_{\mathrm{BS}}-1)\frac{2\pi}{\lambda} d\sin(\phi)}\right]^T, \quad (2)$$

$$\mathbf{a}_{\mathrm{user}}(\phi) = \frac{1}{\sqrt{M}} \left[1, e^{j\frac{2\pi}{\lambda} d\sin(\phi)}, \cdots, e^{j(M-1)\frac{2\pi}{\lambda} d\sin(\phi)}\right]^T, \quad (3)$$

where d is the antenna spacing and λ is the wavelength of the carrier wave.

As shown in Fig. 2, to obtain the channel at the BS, the uplink channel estimation phase is carried out in the first slot by acquiring the uplink training pilot signals sent by the user. Then, the subsequent $N_b - 1$ slots are assigned to support the uplink and the downlink data transmissions.

III. Problem Formulation of Channel Prediction

In this section, we first introduce the signal model of TDD mode. Then, the channel aging issue is illustrated, and the parallel channel mapping problem is formulated to alleviate the channel aging.

A. TDD Signal Model

In TDD mode,[4] for the uplink channel estimation, the user is required to transmit the pilot sequence to the BS. In this paper, the widely used orthogonal pilot transmission strategy is adopted, i.e., the pilot sequence assigned to different antennas are orthogonal to each other, which makes the channel estimation for each user antenna independent [5]. Then, without considering user mobility, the received q-th pilot signal $\mathbf{y}_q \in \mathbb{C}^{N_{\mathrm{RF}} \times 1}$ is written as

$$\mathbf{y}_q = \mathbf{A}_q \mathbf{H} \mathbf{s}_q + \mathbf{A}_q \mathbf{n}_q, \quad q = 1, 2, \cdots, Q, \quad (4)$$

[2]Note that a single-user massive MIMO system is considered in this paper, while the results of this paper could be easily extended to multi-user scenarios by utilizing orthogonal pilot sequences between users.

[3]For the scenario where the channel changes within a single slot, the prediction of the channel in each slot is necessary. This slot-level channel prediction could be regarded as the smaller time scale of the frame-level channel prediction in this paper.

[4]In this paper, we consider a massive MIMO system in the TDD mode, while the methods of this paper could be easily extended to the frequency division duplexing (FDD) mode.

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

where $\mathbf{A}_q \in \mathbb{C}^{N_{\mathrm{RF}} \times N_{\mathrm{BS}}}$ is the BS analog combiner matrix at the q-th pilot signal, $\mathbf{H} \in \mathbb{C}^{N_{\mathrm{BS}} \times M}$ denotes the channel matrix without the Doppler effect, $\mathbf{s}_q \in \mathbb{C}^{M \times 1}$ represents the q-th transmitted pilot signal, $\mathbf{n}_q \in \mathbb{C}^{N_{\mathrm{BS}} \times 1}$ is the zero-mean additive white Gaussian noise (AWGN) with noise power σ_n^2, and Q is the length of the pilot sequence.

For the design of the combiner matrix $\mathbf{A}_q$, we utilize the discrete Fourier transform (DFT) codebook-based analog combiner matrix design method [23], in which each codeword in the DFT codebook is orthogonal to each other and corresponds to a beam directing on the desired AoA. All the codewords in the pre-defined DFT codebook could cover the entire range of AOA. By scanning the entire DFT codebook, the N_{RF} strongest beams can be obtained and then deployed as the analog combiner. Fortunately, as we mentioned before since the changes of the AoA and AoD of a user are negligible in tens of milliseconds, the optimal analog combiner matrix is supposed to remain unchanged in several frames [21], [22]. In this case, we assume $\mathbf{A}_q = \mathbf{A}, \forall q \in \{1, 2, \cdots, Q\}$. After the transmission of the pilot sequence, the BS could form a matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_Q] \in \mathbb{C}^{N_{\mathrm{RF}} \times Q}$ from its observations, that is given by

$$\mathbf{Y} = \mathbf{AHS} + \mathbf{AN} = \mathbf{H}_{\mathrm{e}}\mathbf{S} + \mathbf{AN}, \tag{5}$$

where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_Q] \in \mathbb{C}^{M \times Q}$ is the training pilot matrix for M antennas and each row of $\mathbf{S}$ is orthogonal to other rows, $\mathbf{N} = [\mathbf{n}_1, \mathbf{n}_2, \cdots, \mathbf{n}_Q]$ is the AWGN noise, and $\mathbf{H}_{\mathrm{e}} = \mathbf{AH} \in \mathbb{C}^{N_{\mathrm{RF}} \times M}$ is the effective channel matrix.

According to (5), the effective channel matrix $\mathbf{H}_{\mathrm{e}}$ could be recovered given $\mathbf{Y}$ and $\mathbf{S}$. In the following, we adopt the classical least square (LS) and the minimum mean square error (MMSE) channel estimation methods [1] to estimate the effective channel. First, we vectorize the received signal matrix $\mathbf{Y}$ as follows:

$$\mathrm{vec}(\mathbf{Y}) = (\mathbf{S}^T \otimes \mathbf{I}_{N_{\mathrm{RF}}})\mathrm{vec}(\mathbf{H}_e) + \mathrm{vec}(\mathbf{N}_{\mathrm{e}}). \tag{6}$$

Let $\hat{\mathbf{H}}_{\mathrm{e}}^{\mathrm{LS}}$ and $\hat{\mathbf{H}}_{\mathrm{e}}^{\mathrm{MMSE}}$ denote the LS and MMSE estimators of $\mathbf{H}_{\mathrm{e}}$, respectively. Then, the vectorization of LS and MMSE estimations are given by

$$\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{e}}^{\mathrm{LS}}) = (\mathbf{S}^T \otimes \mathbf{I}_{N_{\mathrm{RF}}})^{-1}\mathrm{vec}(\mathbf{Y}), \tag{7}$$

$$\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{e}}^{\mathrm{MMSE}}) = \mathbf{R}_{\mathbf{H}}(\mathbf{R}_{\mathbf{H}} + \frac{\sigma_n^2}{\sigma_s^2}\mathbf{I}_{N_{\mathrm{RF}}M})^{-1}\mathrm{vec}(\hat{\mathbf{H}}_{\mathrm{e}}^{\mathrm{LS}}), \tag{8}$$

respectively, where $\mathbf{R}_{\mathbf{H}} = \mathbb{E}\{\mathrm{vec}(\mathbf{H}_{\mathrm{e}})\mathrm{vec}(\mathbf{H}_{\mathrm{e}})^H\}$ is the auto-covariance matrix of the effective channel, σ_n^2 and σ_s^2 are the noise power and the pilot power, respectively. According to the estimated $\hat{\mathbf{H}}_{\mathrm{e}}^{\mathrm{LS}}$ or $\hat{\mathbf{H}}_{\mathrm{e}}^{\mathrm{MMSE}}$, the achievable sum-rate for the user in the downlink data transmission is calculated when Gaussian symbols are transmitted over the mmWave channel [24] that is written as

$$R = \mathbb{E}_{\mathbf{H}}\left[\log_2|\mathbf{I}_M + \frac{\rho}{M\sigma_n^2}\mathbf{D}\mathbf{H}_{\mathrm{e}}\mathbf{H}_{\mathrm{e}}^H\mathbf{D}^H|\right], \tag{9}$$

where ρ is the average transmitted power at the BS and we assume $\rho = 1$ without loss of generality; $\mathbf{D}$ is the precoder matrix adopted in the BS, which could be designed by the classical zero-forcing (ZF) algorithm [2] according to the estimated channel $\hat{\mathbf{H}}_{\mathrm{e}}$ using the LS or the MMSE method. Note that, the achievable sum-rate obtained in (9) is actually obtained by assuming that an optimal combiner calculated according to the perfect channel is employed in UE, which decouples the design of the precoding and combining [17] and focuses on the design of the precoding. This is also an important insight into the proposed P2P prediction, which directly predicts the future precoders by maximizing the future achievable sum-rate. The precoder matrix $\mathbf{D}$ could be computed as:

$$\mathbf{D} = (\hat{\mathbf{H}}_{\mathrm{e}}^H\hat{\mathbf{H}}_{\mathrm{e}})^{-1}\hat{\mathbf{H}}_{\mathrm{e}}^H. \tag{10}$$

B. Problem Formulation

In the current 3GPP standard, the precoder matrix $\mathbf{D}$ can achieve a satisfactory achievable sum-rate performance in stationary scenarios. By contrast, in the mobility scenarios, due to the Doppler effect in the time-varying channel (1), except for the first slot, the actual channels of other slots could be significantly different from the acquired channel $\hat{\mathbf{H}}_{\mathrm{e}}$ and thus the sum-rate achieved by precoder $\mathbf{D}$ is severely eroded, especially for the mmWave communications. Specifically, according to [24], the channel coherence time T_c is defined as the time during which the channel can be reasonably well viewed as time-invariant, which is inversely proportional to the frequency and the user speed, i.e.,

$$T_c \approx \frac{0.5c}{vf} = \frac{0.5\lambda}{v}, \tag{11}$$

where v is the user speed and f is the mmWave carrier frequency. Note that the channel coherence time is a rough estimation, which is used to describe the strength of the Doppler effect.[5] The faster the user speed, the stronger the Doppler effect, and the smaller the channel coherence time. As shown in Fig. 2, the length of channel coherence time could be shorter than the period of the frame period (the SRS period) [4] in mmWave scenarios, i.e., $T_c < T_s$. For example, in the typical case with the carrier frequency of 28 GHz in the mmWave frequency band and the user speed of 60 km/h, according to (11) the channel coherence time is roughly $T_c \approx 0.32$ ms, while the least SRS period is $T_s = 0.625$ ms [4][6]. Due to the extremely short channel coherence time, a severe channel aging issue is introduced inevitably. Since the channel is expected to vary rapidly over time due to the user mobility, using the channel estimated at the first slot for the remaining slots would cause a substantial loss of the achievable sum-rate. Besides, the baseband processing delay also aggravates the channel aging issue [6].

Recently, some channel prediction techniques have been proposed to address the aforementioned issues by exploiting the temporal correlation of time-varying channels [7]–[15]. Existing model-based channel prediction could be used to

[5]In particular, for the generally considered Jakes model, the channel coherence time is defined as $T_c = \frac{0.32c}{vf}$ [5].

[6]Since the OFDM symbol period is the inverse of the subcarrier spacing, the larger subcarrier spacing leads to smaller slot lengths. In the 5G standard, The 0.625 ms SRS period includes 5 slots when subcarrier spacing is 120 kHz or 10 slots when subcarrier spacing is 240 kHz.

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

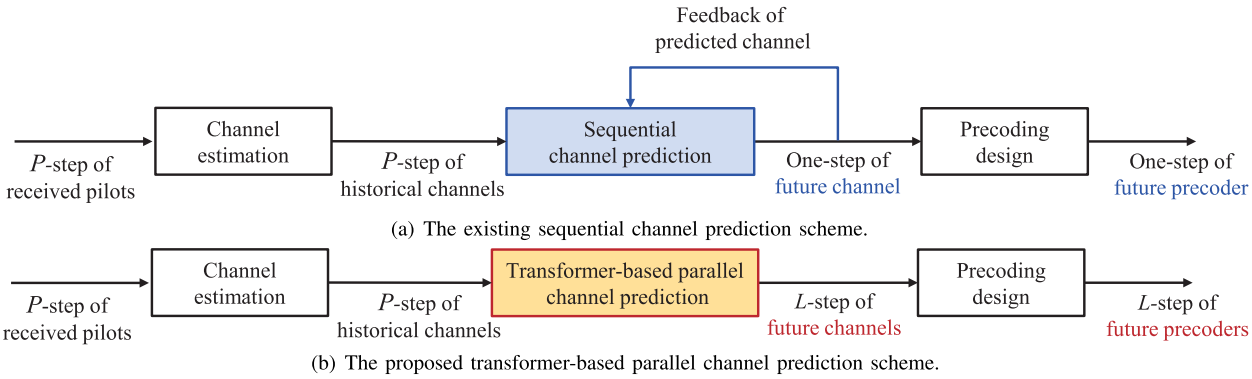


Fig. 3. Comparison between the existing sequential channel prediction scheme and the proposed transformer-based parallel channel prediction scheme.

characterize the time-varying channels. However, the considered channel prediction models in [7]–[11] were only suitable for low-frequency scenarios and difficult to match the complicated and fast changes of the actual channels in high-frequency. To match the fast time-varying channel, neural network-based methods (FCN-based and RNN-based) learn the changes of actual channels by utilizing machine learning [11]–[15]. However, the existing neural network-based methods serially predict the future channels as shown in Fig. 3(a), which suffer from performance loss due to the prediction error propagation problem.

Unlike the existing sequential channel prediction schemes, we formulate the channel prediction problem as a parallel channel mapping problem. Specifically, we adopt the estimated historical channels from the past P frames to predict the future channels in next L consecutive frames simultaneously, as shown in Fig. 3(b). The parallel channel mapping problem can be formulated as

$$\min_{\mathbf{\Theta}} \ \mathcal{L}(\mathbf{\Theta}) = \mathbb{E}\left\{ \frac{\sum_{t=1}^{L} ||\mathbf{H}_{\mathrm{e}}^{(T+t)} - \hat{\mathbf{H}}_{\mathrm{p}}^{(T+t)}||^2}{\sum_{t=1}^{L} ||\mathbf{H}_{\mathrm{e}}^{(T+t)}||^2} \right\}, \tag{12a}$$

$$\text{s.t.} \ \ (\hat{\mathbf{H}}_{\mathrm{p}}^{(T+1)}, \cdots, \hat{\mathbf{H}}_{\mathrm{p}}^{(T+L)}) = f_{\mathbf{\Theta}}(\hat{\mathbf{H}}_{\mathrm{e}}^{(T-P+1)}, \cdots, \hat{\mathbf{H}}_{\mathrm{e}}^{(T)}), \tag{12b}$$

where $f_{\mathbf{\Theta}}(\cdot)$ is the proposed parallel mapping model; $\mathbf{\Theta}$ is the set of model parameters; $\mathbf{H}_{\mathrm{e}}^{(t)}$ is the actual effective channel at the first slot of the t-th frame; $\hat{\mathbf{H}}_{\mathrm{p}}^{(t)}$ is the predicted effective channel; $\hat{\mathbf{H}}_{\mathrm{e}}^{(t)}$ is the estimated effective channel. Note that the normalized MSE (NMSE) loss function is utilized as the minimization target. On the one hand, the normalized loss could produce a steady gradient rather than an oscillating gradient, thereby speeding up model convergence. On the other hand, for the evaluation of channel prediction performance, the NMSE is a better metric than MSE. The MSE can be affected by the amplitude of the channel, while the normalized MSE could eliminate this effect by normalizing. Once the channels in the future frames are accurately predicted, the precoder matrix of the n-th slot at the t-th frame can thus be designed according to the predicted $\hat{\mathbf{H}}_{\mathrm{p}}^{(t)}$ or $\hat{\mathbf{H}}_{\mathrm{p}}^{(t+1)}$. For instance, for the t-th frame, the precoder matrix of the n-th slot can be designed based on the predicted channel $\hat{\mathbf{H}}_{\mathrm{p}}^{(t)}$ if $0 < n < N_b/2$ or be designed based on the predicted channel $\hat{\mathbf{H}}_{\mathrm{p}}^{(t+1)}$ if $N_b/2 < n < N_b$. Since these slots are in the channel coherence time of $\hat{\mathbf{H}}_{\mathrm{p}}^{(t)}$ or $\hat{\mathbf{H}}_{\mathrm{p}}^{(t+1)}$, the loss of achievable sum-rate caused by the channel aging issue is expected to be mitigated.

IV. Transformer-Based Parallel Channel Prediction

In this section, to alleviate the sum-rate performance loss caused by the widely adopted sequential channel prediction, we propose a transformer-based parallel channel prediction scheme to avoid the prediction error propagation. At first, we introduce the framework of the proposed transformer-based scheme. Then, the structure of the transformer model is elaborated. Finally, the computational complexity analysis is provided.

A. Transformer-Based Parallel Channel Prediction Framework

The aim of channel prediction is to design the mapping function $f_{\mathbf{\Theta}}(\cdot)$ in (12b), while in this paper, we propose a transformer-based model to design the mapping function. The framework of the proposed transformer-based parallel channel prediction scheme consists of two phases, i.e., offline training and online prediction. For the offline training phase, a supervised learning algorithm is applied to train the transformer model according to (12a). To start with, the L-step actual channel labels in the next L frames and the past P-step estimated historical channels from the previous P frames are required, which could be obtained by generating a simulation dataset based on some acknowledged channel models, e.g., 5G channel models [25].

According to the generated antenna-domain channel dataset $\mathcal{H}$, we first sample the channels $\{\mathbf{H}^{(T-P+1)}, \cdots, \mathbf{H}^{(T+L)}\}$ as a training sample, which includes contiguous $P+L$ antenna-domain channels and these $P+L$ channels are sampled every other frame in the time domain. Next, by multiplying the DFT matrix $\mathbf{U}_a^H$ with the antenna-domain channel $\mathbf{H}$ and considering the actual noise, each antenna-domain channel $\mathbf{H}$ is transformed into an angle-domain channel $\mathbf{H}_a$, which could be written as

$$\mathbf{H}_a = \mathbf{U}_a^H \mathbf{H} + \mathbf{U}_a^H \mathbf{N}, \tag{13}$$

where $\mathbf{U}_a = [\mathbf{a}_{\mathrm{BS}}(\phi_1), \cdots, \mathbf{a}_{\mathrm{BS}}(\phi_{N_{\mathrm{BS}}})] \in \mathbb{C}^{N_{\mathrm{BS}} \times N_{\mathrm{BS}}}$ is the spatial discrete DFT matrix and $\phi_n = \arcsin(\frac{2}{N_{\mathrm{BS}}}(n - \frac{N_{\mathrm{BS}}+1}{2}))$ with $n = 1, 2, \cdots, N_{\mathrm{BS}}$ are the N_{BS} spatial

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

directions pre-defined by the DFT codebook. Since the limitation of the number of RF chains in the hybrid precoding architecture, only the N_{RF} codewords are selected to perform analog combining. Specifically, based on the angle-domain channel $\mathbf{H}_a$, we could obtain the index of the strongest spatial directions n_i $(i = 1, \cdots, N_{\mathrm{RF}})$ as follows:

$$n_i = \underset{n \in \mathcal{N}/\mathcal{N}_i}{\arg\max} ||\mathbf{H}_a[n,:]||_2^2, \quad i = 1, \cdots, N_{\mathrm{RF}}, \tag{14}$$

where $\mathcal{N} = \{1, \cdots, N_{\mathrm{BS}}\}$ and $\mathcal{N}_i = \{n_1, \cdots, n_{i-1}\}$. The corresponding columns of the DFT matrix constitute the analog combining matrix $\mathbf{A} \in \mathbb{C}^{N_{\mathrm{RF}} \times N_{\mathrm{BS}}}$, i.e.,

$$\mathbf{A} = [a_{\mathrm{BS}}(\phi_{n_1}), \cdots, a_{\mathrm{BS}}(\phi_{n_{N_{\mathrm{RF}}}})]^H. \tag{15}$$

Then, the effective channel $\mathbf{H}_{\mathrm{e}}$ could also be obtained by multiplying the analog combining matrix $\mathbf{A}$ with the antenna-domain channel $\mathbf{H}$ as shown in (5). The input of the transformer-based parallel channel prediction model $\{\hat{\mathbf{H}}_{\mathrm{e}}^{(T-P+1)}, \cdots, \hat{\mathbf{H}}_{\mathrm{e}}^{(T)}\}$ could be obtained by (7) or (8). At the same time, the labels of the outputs of the transformer-based parallel channel prediction model are the actual effective channels $\{\mathbf{H}_{\mathrm{e}}^{(T+1)}, \cdots, \mathbf{H}_{\mathrm{e}}^{(T+L)}\}$. Finally, we aim to minimize the classical normalized mean square error (NMSE) loss function shown in (12a) and the backpropagation (BP) algorithm [26] is used to calculate the gradient of the transformer model.

After training the transformer-based parallel channel prediction model, we deploy the well-trained model and implement the online prediction. For the online prediction, to predict the future effective channels $\{\mathbf{H}_{\mathrm{e}}^{(T+1)}, \cdots, \mathbf{H}_{\mathrm{e}}^{(T+L)}\}$, we need to input the estimated historical channels $\{\hat{\mathbf{H}}_{\mathrm{e}}^{(T-P+1)}, \cdots, \hat{\mathbf{H}}_{\mathrm{e}}^{(T)}\}$ in the previous P frames. To obtain the estimated historical channels, unlike the offline training, which could use the mathematical formula (13)-(15) directly calculate $\mathbf{A}$ according to the full CSI data $\mathbf{H}$, for the online prediction, beam training is preferred due to an unacceptable overhead for estimating $\mathbf{H}$ [27][7]. Specifically, to determine the analog combining matrix $\mathbf{A}$ in BS, we first try to get the N_{RF} spatial directions where the channel is strongest. For this purpose, the BS is required to sweep the entire beam space with the DFT codebook. After sweeping the beams, BS could determine the N_{RF} spatial directions ϕ_{n_i} $(i = 1, \cdots, N_{\mathrm{RF}})$ by choosing the N_{RF} codewords with the strongest received power. Then, the analog combining matrix could be constituted by incorporating these codewords as shown in (15). Fortunately, since the changes of the AoA and AoD are negligible in tens of milliseconds, the strongest beams are supposed to remain unchanged in the contiguous $P + L$ frames. Thus, we can reasonably assume that the analog combining matrix $\mathbf{A}$ holds when predicting the future channels. In the following sections, we elaborate on the details of the utilized mapping function in (12b), i.e., the transformer model.

[7]Actually, the calculation method of the analog combining matrix $\mathbf{A}$ at offline training and online prediction is the same, while the difference is the acquisition method of the angle-domain channel. Specifically, since the full CSI $\mathbf{H}$ is available at offline training, the angle-domain channel could be obtained by (13). In the stage of online prediction, due to the unacceptable overhead for estimating the full CSI $\mathbf{H}$, beam training is preferred to obtain the angle-domain channel $\mathbf{H}_a$.

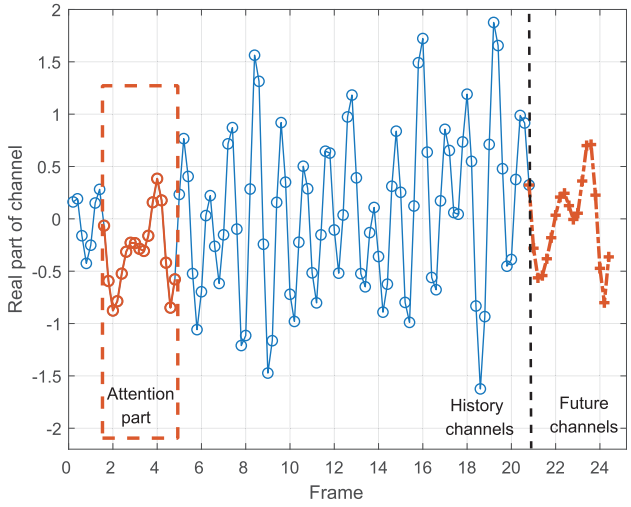


Fig. 4. A sample of the channels in contiguous $P+L$ frames. The Matlab 5G toolbox [28] is used to generate these channel samples, where $v = 60\mathrm{km/h}$ and the delay spread is set as $30\mathrm{ns}$. The period of SRS is $0.625\mathrm{ms}$. The highlighted part of P channels could be used to facilitate channel prediction by utilizing the attention mechanism.

B. Attention Mechanism

Firstly, to make this paper self-contained, we provide a brief review of the most important module of the transformer model, termed the attention mechanism [16]. Essentially, the attention mechanism is a neural network module that can implement parallel signal weighting functions. Using an attention mechanism can greatly improve the performance of channel prediction for two reasons. First of all, different from serial data processing of the conventional RNN, which causes the prediction error propagation due to the sequential channel prediction, the attention mechanism can perform parallel signal processing. Thus, the CSI from the previous frames can be adopted jointly to predict the channels for the next several frames at once, avoiding any prediction error propagation.

On the other hand, there is another disadvantage of serial data processing, that is, it is easy to forget the early input data. By contrast, the attention mechanism can pay attention to the early input in a parallel manner. Specifically, the attention mechanism is a weighting mechanism [29], [30], such that the weights of the model to the historical channels can be adjusted adaptively to the most relevant past data for future channel prediction. At the same time, since the time-varying of the channel is caused by the Doppler effect, the channel itself has a long periodicity. As shown in Fig. 4, for the highlighted early historical channels that are most relevant to the future channels, the attention mechanism can automatically assign more weights to the early part such that the transformer model focuses its learning on them, thereby improving the accuracy of channel prediction. Consequently, as the core module of the transformer, the attention mechanism is well-suited for the problem of parallel channel prediction. Below we will introduce the realization of the attention mechanism in detail.

The basic structure of the attention mechanism is shown in Fig. 5. To realize the attention mechanism, the normalized attention matrix is introduced to characterize different degrees of attention to the input [31], [32]. The more important input parts are given larger weights and the final output is obtained

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

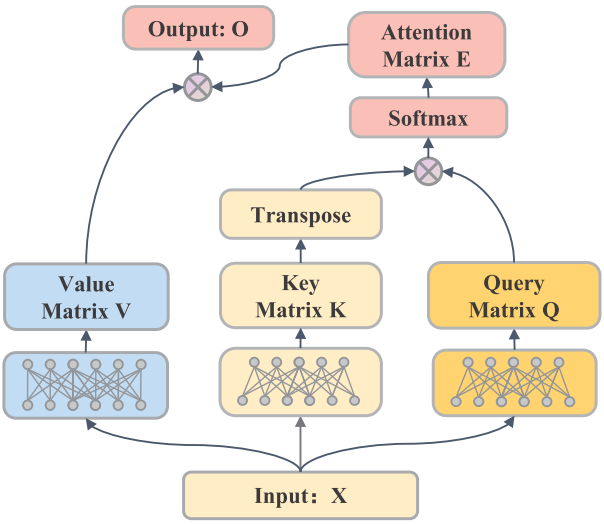


Fig. 5. The structure of the attention mechanism.

by weighting the input according to the attention weights in the attention matrix. By utilizing the attention mechanism, the input could produce a new representation of the feature considering the time sequence information. Specifically, for the input $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$, where n denotes the length of the input historical channels and m denotes the feature dimension of each historical channel, we apply three different linear transformations to $\mathbf{x}_i$ given by:

$$\mathbf{k}_i = \mathbf{W}^k \mathbf{x}_i, \quad i = 1, \cdots, n, \tag{16a}$$

$$\mathbf{q}_i = \mathbf{W}^q \mathbf{x}_i, \quad i = 1, \cdots, n, \tag{16b}$$

$$\mathbf{v}_i = \mathbf{W}^v \mathbf{x}_i, \quad i = 1, \cdots, n, \tag{16c}$$

where $\mathbf{k}_i \in \mathbb{R}^{d\times 1}$, $\mathbf{q}_i \in \mathbb{R}^{d\times 1}$, and $\mathbf{v}_i \in \mathbb{R}^{m\times 1}$ are the key vector, the query vector, and the value vector, respectively; $\mathbf{W}^k \in \mathbb{R}^{d\times m}$, $\mathbf{W}^q \in \mathbb{R}^{d\times m}$, and $\mathbf{W}^v \in \mathbb{R}^{m\times m}$ are the corresponding trainable linear transformation matrices, respectively; d is the feature dimension of the key vector. Specifically, the function for weight allocation is realized by the key vector $\mathbf{k}_i$ and query vector $\mathbf{q}_j$. The correlation between $\mathbf{k}_i$ and $\mathbf{q}_j$ represents the correlation between the i-th input signal and the j-th output signal. The larger correlation $\mathbf{k}_i^T \mathbf{q}_j$ implies the j-th output has to pay more attention to the feature of the i-th input $\mathbf{x}_i$, i.e., the value vector $\mathbf{v}_i$. Generally, this correlation can be adaptively adjusted based on the input $\mathbf{X}$ and the trainable matrix $\mathbf{W}^k$ and $\mathbf{W}^q$. For a better illustration, we show the matrix forms of (16a)-(16c) as follows:

$$\mathbf{K} = \mathbf{W}^k \mathbf{X}, \tag{17a}$$

$$\mathbf{Q} = \mathbf{W}^q \mathbf{X}, \tag{17b}$$

$$\mathbf{V} = \mathbf{W}^v \mathbf{X}, \tag{17c}$$

where $\mathbf{K} = [\mathbf{k}_1, \cdots, \mathbf{k}_n] \in \mathbb{R}^{d\times n}$, $\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_n] \in \mathbb{R}^{d\times n}$, and $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_n] \in \mathbb{R}^{m\times n}$.

Based on $\mathbf{K}$ and $\mathbf{Q}$, the attention matrix $\mathbf{E} \in \mathbb{R}^{n\times n}$ could be obtained by computing the product between key matrix $\mathbf{K}^T$ and query matrix $\mathbf{Q}$, which could be denoted by

$$\mathbf{E} = \mathrm{Softmax}\Big(\frac{\mathbf{K}^T \mathbf{Q}}{\sqrt{d}}\Big). \tag{18}$$

Algorithm 1:: Attention Mechanism

Inputs: $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$;
Outputs: $\mathbf{O} = [\mathbf{o}_1, \cdots, \mathbf{o}_n]$;
1: Key matrix: $\mathbf{K} = \mathbf{W}^k \mathbf{X}$
2: Query matrix: $\mathbf{Q} = \mathbf{W}^q \mathbf{X}$
3: Value matrix: $\mathbf{V} = \mathbf{W}^v \mathbf{X}$
4: Attention matrix: $\mathbf{E} = \mathrm{Softmax}\big(\frac{\mathbf{K}^T \mathbf{Q}}{\sqrt{d}}\big)$
5: $\mathbf{O} = \mathbf{V}\mathbf{E}$

Since the variance of the inner product between the key matrix and query matrix increases as the feature dimension of the key vector d increases, the result of the product is adjusted by $\sqrt{d}$. Then, the softmax operation $\mathrm{Softmax}(\mathbf{x}) = \frac{\exp(x_i)}{\sum \exp(x_i)}$ is applied to normalize each column of the matrix. It is worth noting that each column of the attention matrix is a probability vector whose elements are all non-negative and add up to one. If the key vector $\mathbf{K}[:, i]$ and the query vector $\mathbf{Q}[:, j]$ match better, then the corresponding attention weight $\mathbf{E}[i, j]$ would be larger. Thus, the output of the attention mechanism corresponding to l-th component could be denoted by the weighted sum of all inputs, written as

$$\mathbf{o}_l = \sum_i \mathbf{v}_i \mathbf{E}[i, l] = \mathbf{V}\mathbf{E}[:, l], \tag{19}$$

where $\mathbf{o}_l \in \mathbb{R}^{m\times 1}$ is the l-th output by paying adaptive attention to the historical inputs according to the attention weight $\mathbf{E}[i, l]$. If the attention weight $\mathbf{E}[i, l]$ is larger, the corresponding value vector $\mathbf{v}_i$ will contribute more to the output.

Finally, the overall representation of the attention mechanism could be summarized as follows:

$$\mathbf{O} = \mathbf{W}^v \mathbf{X} \mathrm{Softmax}\Big(\frac{\mathbf{X}^T {\mathbf{W}^k}^T \mathbf{W}^q \mathbf{X}}{\sqrt{d}}\Big), \tag{20}$$

in which $\mathbf{O} = [\mathbf{o}_1, \cdots, \mathbf{o}_n] \in \mathbb{R}^{m\times n}$. Note that the three different linear transformations $\mathbf{W}^k$, $\mathbf{W}^v$, and $\mathbf{W}^q$ could be realized by trainable FCN without applying the activation function. The procedure of the attention mechanism is summarized in **Algorithm 1**.

C. Encoder of Transformer Model

In this subsection, based on the attention mechanism introduced in Subsection IV-B, we present the proposed channel prediction scheme based on the transformer model. The overall architecture of the transformer model is shown in Fig. 6, which consists of two parts, i.e., an encoder and a decoder. Generally, the encoder is to remember and extract the historical CSI and refine the historical CSI as the channel features. Correspondingly, the decoder is to infer and forms the future channels by utilizing and recalling the features. Note that the attention mechanism introduced in the previous subsection is utilized in the encoder and the decoder simultaneously with a little modification, e.g., the mask-attention mechanism and the full-attention mechanism in Fig. 6.

For the encoder, the features of channels in previous frames are extracted to assist the prediction at the decoder.

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

Algorithm 2: : Encoder for Proposed Transformer-Based Model

Inputs:

Channels in previous P frames: $\overline{\mathbf{H}}_{(e)} = [\hat{\mathbf{h}}^{(T-P+1)}, \cdots, \hat{\mathbf{h}}^{(T)}]$;

Outputs:

Extracted feature $\overline{\mathbf{Y}}_{(e)}$;

Key matrix in encoder: $\mathbf{K}_{(e)} = \mathbf{W}^k_{(e)}\overline{\mathbf{H}}_{(e)}$

Query matrix in encoder: $\mathbf{Q}_{(e)} = \mathbf{W}^q_{(e)}\overline{\mathbf{H}}_{(e)}$

Value matrix in encoder: $\mathbf{V}_{(e)} = \mathbf{W}^v_{(e)}\overline{\mathbf{H}}_{(e)}$

Attention matrix in encoder: $\mathbf{E}_{(e)} = \mathrm{Softmax}\big(\frac{\mathbf{K}^T_{(e)}\mathbf{Q}_{(e)}}{\sqrt{d}}\big)$

$\mathbf{Z}_{(e)} = \mathrm{LN}\big(\overline{\mathbf{H}}_{(e)} + \mathbf{V}_{(e)}\mathbf{E}_{(e)}\big)$

$\overline{\mathbf{Y}}_{(e)} = \mathrm{LN}\Big(\mathbf{Z}_{(e)} + \mathrm{FCN}\big(\mathbf{Z}_{(e)}\big)\Big)$

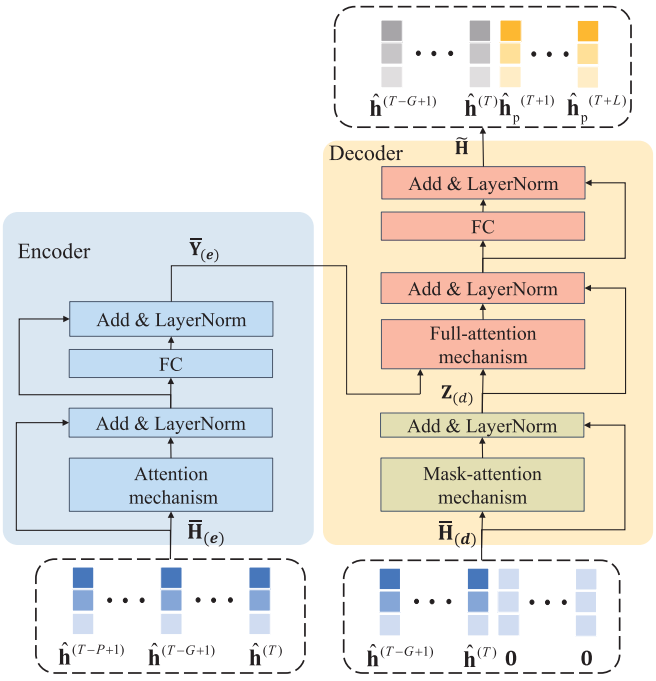


Fig. 6. The proposed transformer-based parallel channel prediction model.

The procedure of the encoder is shown in **Algorithm 2**. For the purpose of applying the attention mechanism, we first extract the real and imaginary parts of each estimated historical channel $\hat{\mathbf{H}}_{\mathrm{e}}^{(t)}$, and then we vectorize and concatenate them into a vector $\hat{\mathbf{h}}^{(t)} \in \mathbb{R}^{2N_{\mathrm{RF}}M\times 1}$ as follows:

$$\hat{\mathbf{h}}^{(t)} = [\mathrm{vec}(\Re[\hat{\mathbf{H}}_{\mathrm{e}}^{(t)}])^T, \mathrm{vec}(\Im[\hat{\mathbf{H}}_{\mathrm{e}}^{(t)}])^T]^T. \tag{21}$$

After that, similar to (17a)-(17c), the attention mechanism of the encoder generates key matrix $\mathbf{K}_{(e)}$, query matrix $\mathbf{Q}_{(e)}$, and value matrix $\mathbf{V}_{(e)}$ by three different linear transformations according to the historical channels in previous P frames $\overline{\mathbf{H}}_{(e)} = [\hat{\mathbf{h}}^{(T-P+1)}, \cdots, \hat{\mathbf{h}}^{(T)}]$. Then, the attention matrix $\mathbf{E}_{(e)} = \mathrm{Softmax}(\frac{\mathbf{K}^T_{(e)}\mathbf{Q}_{(e)}}{\sqrt{d}})$ is calculated.

Next, the output of the attention mechanism could be obtained by summing the value matrix according to the attention weights in the attention matrix. Moreover, to avoid the gradient vanishing problem caused by unstable data distribution [33], several classic methods to mitigate this problem are used in the encoder, such as residual connection [34] and layer normalization [35]. Thus, by taking residual connection and layer normalization into consideration, we can obtain hidden variables of inputs by adding the inputs and normalizing as

$$\mathbf{Z}_{(e)} = \mathrm{LN}\Big(\overline{\mathbf{H}}_{(e)} + \mathbf{V}_{(e)}\mathrm{Softmax}\big(\frac{\mathbf{K}^T_{(e)}\mathbf{Q}_{(e)}}{\sqrt{d}}\big)\Big), \tag{22}$$

where $\mathrm{LN}(\mathbf{X}) = \frac{\mathbf{X}[i,j]-\mu_j}{\sqrt{\delta_j^2+\epsilon}}$ denotes the layer normalization operation, which is used to speed up training by normalizing the data into a standard normal distribution; and μ_j and δ_j^2 present the expectation and variance of the $\mathbf{X}[:,j]$, respectively. Also, a small amount ϵ is added to the denominators to avoid an ill-conditioned problem. Moreover, a two-layer FCN is considered to further extract and synthesize features of each historical input, which could be written as

$$\overline{\mathbf{Y}}_{(e)} = \mathrm{LN}\Big(\mathbf{Z}_{(e)} + \mathrm{FCN}\big(\mathbf{Z}_{(e)}\big)\Big), \tag{23}$$

where $\mathrm{FCN}(\mathbf{X}) = \mathbf{W}^2_{(e)}\mathrm{ReLU}(\mathbf{W}^1_{(e)}\mathbf{X} + \mathbf{b}^1_{(e)}) + \mathbf{b}^2_{(e)}$; $\mathbf{W}^i_{(e)} \in \mathbb{R}^{2N_{\mathrm{RF}}M\times 2N_{\mathrm{RF}}M}$ and $\mathbf{b}^i_{(e)} \in \mathbb{R}^{2N_{\mathrm{RF}}M\times 1}$ are the weight matrix and bias of i-th layer, respectively; and $\mathrm{ReLU}(x) = \max(0, x)$ is a non-linear activation function. The residual connection and layer normalization are considered again.

D. Decoder of Transformer Model

For the decoder, its target is to accurately predict the future channels in next several frames according to the features $\overline{\mathbf{Y}}_{(e)}$ extracted by the encoder. The specific procedure of the decoder is provided in **Algorithm 3**. Specifically, a G-step channel in previous G frames is firstly sampled as a start token of the decoder. The start token, as well as L zeros padding, are used as the decoder input. Then, the decoder aims to generate the predicted L long channels for the next L frames at the zeros padding position by measuring the attention weights with $\overline{\mathbf{Y}}_{(e)}$. After that, the specific generation process of the decoder will be described in detail.

First, similar to the attention mechanism used in encoder, the key matrix $\mathbf{K}_{(d)}$, the query matrix $\mathbf{Q}_{(d)}$, and the value matrix $\mathbf{V}_{(d)}$ in the decoder are produced by three different linear transformations. Note that the input of the decoder $\overline{\mathbf{H}}_{(d)} = [\hat{\mathbf{h}}^{(T-G+1)}, \cdots, \hat{\mathbf{h}}^{(T)}, \mathbf{0}, \cdots, \mathbf{0}] \in \mathbb{R}^{2N_{\mathrm{RF}}M\times(G+L)}$ is consisted of a combination of the historical channels in previous G frames and L long zeros padding. Note that the L long zeros padding input could eventually be replaced by the predicted L-step channels. Limited by this parallel decoder, the input and output must be of the same size. The zero-padding is what we don't expect, and we just fill them in to form a computable structure. Similar to (22), hidden variables of the decoder input could be obtained by considering the time-series information as follows:

$$\mathbf{Z}_{(d)} = \mathrm{LN}\Big(\overline{\mathbf{H}}_{(d)} + \mathbf{V}_{(d)}\mathrm{Softmax}\Big(\mathrm{Mask}\big(\frac{{\mathbf{K}_{(d)}}^T\mathbf{Q}_{(d)}}{\sqrt{d}}\big)\Big)\Big). \tag{24}$$

Note that a slight difference from (22), an additional mask operation is added in (24) to avoid paying attention to the

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

Algorithm 3:: Decoder for Proposed Transformer-Based Model

Inputs:

Extracted feature by encoder $\overline{\mathbf{Y}}_{(e)}$,
$\overline{\mathbf{H}}_{(d)} = [\hat{\mathbf{h}}^{(T-G+1)}, \cdots, \hat{\mathbf{h}}^{(T)}, \mathbf{0}, \cdots, \mathbf{0}]$;

Outputs:

Predicted channels $\overline{\mathbf{H}} = [\hat{\mathbf{h}}^{(T+1)}, \cdots, \hat{\mathbf{h}}^{(T+L)}]$;

Key matrix in decoder: $\mathbf{K}_{(d)} = \mathbf{W}^k_{(d)}\overline{\mathbf{H}}_{(d)}$

Query matrix in decoder: $\mathbf{Q}_{(d)} = \mathbf{W}^q_{(d)}\overline{\mathbf{H}}_{(d)}$

Value matrix in decoder: $\mathbf{V}_{(d)} = \mathbf{W}^v_{(e)}\overline{\mathbf{H}}_{(d)}$

Attention matrix in decoder:
$\mathbf{E}_{(d)} = \mathrm{Softmax}\Big(\mathrm{Mask}\big(\frac{\mathbf{K}_{(d)}{}^T\mathbf{Q}_{(d)}}{\sqrt{d}}\big)\Big)$

$\mathbf{Z}_{(d)} = \mathrm{LN}\big(\overline{\mathbf{H}}_{(d)} + \mathbf{V}_{(d)}\mathbf{E}_{(d)}\big)$

$\mathbf{Q} = \mathbf{W}^q\mathbf{Z}_{(d)}$

$\mathbf{K} = \mathbf{W}^k\overline{\mathbf{Y}}_{(e)}, \mathbf{V} = \mathbf{W}^v\overline{\mathbf{Y}}_{(e)}$

Full-attention matrix between encoder and decoder:
$\mathbf{E} = \mathrm{Softmax}\big(\frac{\mathbf{K}^T\mathbf{Q}}{\sqrt{d}}\big)$

$\mathbf{Z} = \mathrm{LN}\big(\mathbf{Z}_{(d)} + \mathbf{V}\mathbf{E}\big)$

$\widetilde{\mathbf{H}} = \mathrm{LN}\Big(\mathbf{Z} + \mathrm{FCN}\big(\mathbf{Z}\big)\Big)$

$\overline{\mathbf{H}} = [\hat{\mathbf{h}}^{(T+1)}, \cdots, \hat{\mathbf{h}}^{(T+L)}]$ by extracting last L columns of $\widetilde{\mathbf{H}}$.

future channels. For example, when we predict the channel at the $(T+t)$-th frame, the decoder currently only knows the information before the $(T+t)$-th frame and the channel after $(T+t)$-th frames is still an unknown. For this purpose, the mask operation is defined as

$$\mathrm{Mask}(\mathbf{X}) = \mathbf{X} + \mathbf{tril}(-\inf), \tag{25}$$

where $\mathbf{tril}(-\inf)$ denotes a lower triangular matrix with each lower triangular element being negative infinity. After the $\mathrm{Softmax}$ operation, the attention weight of the $(T+t)$-th channel to $(T+t+\tau)$-th channel is zero with $\tau > 0$. Additionally, the masking operation could make the padded zero value have no effect on subsequent calculations, which simplifies the computation in the decoder. Specifically, since the attention weights are zeros for the first G outputs to the last L outputs after masking and Softmax operations, the invalid analysis caused by zero-padding is avoided. Thus, the masking operation is used to shield the effect caused by zero-padding.

Next, based on $\mathbf{Z}_{(d)}$ calculated at the decoder and $\overline{\mathbf{Y}}_{(e)}$ calculated at the encoder, a full-attention mechanism between the encoder and the decoder is carried out, which aims to fully utilize the historical information by paying attention to the features provided by the encoder. For example, a sample obtained by the sampling of $P+L$ channels is shown in Fig. 4. The highlighted part in P historical channels could be observed to be a perfect match with the L channels we aim to predict. To pay more attention to the highlighted part, the larger weights are expected to be assigned to the highlighted part when we predict the L future channels. Specifically, another three linear transformations are applied to $\overline{\mathbf{Y}}_{(e)}$ and $\mathbf{Z}_{(d)}$, respectively. A little different from mask-attention, the key matrix $\mathbf{K}$ and value matrix $\mathbf{V}$ are calculated by feature $\overline{\mathbf{Y}}_{(e)}$ extracted by encoder while the query matrix $\mathbf{Q}$ is calculated by the hidden variable $\mathbf{Z}_{(d)}$ of the decoder.

Following, the attention matrix is acquired by $\mathbf{E} = \mathrm{Softmax}\big(\frac{\mathbf{K}^T\mathbf{Q}}{\sqrt{d}}\big)$, where $\mathbf{E}[i,j]$ denotes the attention weight of the $\mathbf{Z}_{(d)}[:,j]$ to $\overline{\mathbf{Y}}_{(e)}[:,i]$. Each output of the full-attention mechanism could establish the relationship with the feature $\overline{\mathbf{Y}}_{(e)}$ by multiplying the value matrix $\mathbf{V}$ with attention matrix $\mathbf{E}$. Based on the parallel full-attention mechanism, each future channel could naturally correlate with channels in previous frames. Then, the predicted channels in next L frames could be obtained by

$$\mathbf{Z} = \mathrm{LN}\Big(\mathbf{Z}_{(d)} + \mathbf{V}\mathrm{Softmax}\big(\frac{\mathbf{K}^T\mathbf{Q}}{\sqrt{d}}\big)\Big), \tag{26}$$

$$\widetilde{\mathbf{H}} = \mathrm{LN}\Big(\mathbf{Z} + \mathrm{FCN}\big(\mathbf{Z}\big)\Big), \tag{27}$$

where $\widetilde{\mathbf{H}} = [\hat{\mathbf{h}}_{\mathrm{p}}^{(T-G+1)}, \cdots, \hat{\mathbf{h}}_{\mathrm{p}}^{(T)}, \hat{\mathbf{h}}_{\mathrm{p}}^{(T+1)}, \cdots, \hat{\mathbf{h}}_{\mathrm{p}}^{(T+L)}] \in \mathbb{R}^{2N_{\mathrm{RF}}M\times(G+L)}$, $\mathrm{FCN}(\mathbf{X}) = \mathbf{W}^2_{(d)}\mathrm{ReLU}(\mathbf{W}^1_{(d)}\mathbf{X} + \mathbf{b}^1_{(d)}) + \mathbf{b}^2_{(d)}$, and $\mathbf{W}^i_{(d)} \in \mathbb{R}^{2N_{\mathrm{RF}}M\times 2N_{\mathrm{RF}}M}$ and $\mathbf{b}^i_{(d)} \in \mathbb{R}^{2N_{\mathrm{RF}}M\times 1}$ are the weight matrix and the bias of i-th layer, respectively. Compared to the decoder input $\widetilde{\mathbf{Y}}$, the zeros padding position is covered by the generative predicted channels. So, we could extract the last L columns of $\widetilde{\mathbf{H}}$ to obtain the expected channels in next L frames, i.e., $\overline{\mathbf{H}} = [\hat{\mathbf{h}}_{\mathrm{p}}^{(T+1)}, \cdots, \hat{\mathbf{h}}_{\mathrm{p}}^{(T+L)}] \in \mathbb{R}^{2N_{\mathrm{RF}}M\times L}$. The matrix form of the effective channel could be obtained by real-to-complex (R2C) and de-vectorization operations as follows:

$$\hat{\mathbf{H}}_{\mathrm{p}}^{(t)} = \mathrm{dvec}\Big(\mathrm{R2C}\big(\hat{\mathbf{h}}_{\mathrm{p}}^{(t)}\big)\Big), \quad t = T+1, \cdots, T+L. \tag{28}$$

Finally, we use the NMSE loss function in (12a) on prediction w.r.t. the target actual channels in next L frames and use the BP algorithm [26] to update the encoder and decoder of the transformer model. To make the concepts of the proposed transformer-based model clearer, the concepts are summarized in Table I.

TABLE I

THE CONCEPTS OF THE PROPOSED TRANSFORMER-BASED CHANNEL PREDICTION MODEL

Concepts of the transformer model	Notations
Input of the enc/dec	$\overline{\mathbf{H}}_{(e)}/\overline{\mathbf{H}}_{(d)}$
Key matrix in enc/dec	$\mathbf{K}_{(e)}/\mathbf{K}_{(d)}$
Query matrix in enc/dec	$\mathbf{Q}_{(e)}/\mathbf{Q}_{(d)}$
Value matrix in enc/dec	$\mathbf{V}_{(e)}/\mathbf{V}_{(d)}$
Attention matrix in enc/dec	$\mathbf{E}_{(e)}/\mathbf{E}_{(d)}$
Hidden variable in enc/dec	$\mathbf{Z}_{(e)}/\mathbf{Z}_{(d)}$
Key matrix in full-attention	$\mathbf{K}$
Query matrix in full-attention	$\mathbf{Q}$
Value matrix in full-attention	$\mathbf{V}$
Attention matrix in full-attention	$\mathbf{E}$
Output of the enc/dec	$\overline{\mathbf{Y}}_{(e)}/\widetilde{\mathbf{H}}$

E. Complexity Analysis

In this subsection, we evaluate the complexity of the proposed transformer-based parallel channel prediction in the number of multiplication. Since the training process is performed offline and will not affect the online prediction overhead, we mainly consider the computational complexity of the online prediction.

To calculate the complexity of the transformer-based parallel channel prediction scheme, we first compute the complexity of the crucial attention mechanism module. According to

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1, we observe that the complexity of the attention mechanism is dominated by the matrix multiplication from step 1 to step 5. In step 1, we need to calculate the multiplication between the input $\mathbf{X} \in \mathbb{R}^{m\times n}$ and linear transformation matrix $\mathbf{W}^k \in \mathbb{R}^{d\times m}$, which has a complexity in the order of $\mathcal{O}(mnd)$. Similarly, the matrix multiplication from step 2 to step 5 have the complexity in the order of $\mathcal{O}(mnd)$, $\mathcal{O}(m^2n)$, $\mathcal{O}(n^2d)$, and $\mathcal{O}(n^2m)$, respectively. Note that the feature dimension of key vector d is generally set to be consistent with the input dimension m. Thus, the complexity of the attention mechanism can be represented by $\mathcal{O}(m^2n)+\mathcal{O}(n^2m)$.

Next, we calculate the complexity in the encoder of the transformer model. According to **Algorithm 2**, we observe that the complexity of the encoder mainly comes from the attention mechanism from step 1 to 5 and the FCN processing in step 6. Since the input of encoder $\overline{\mathbf{H}}_{(e)} \in \mathbb{R}^{2N_{\mathrm{RF}}M\times P}$, i.e., $m = 2N_{\mathrm{RF}}M$ and $n = P$, the attention mechanism of encoder has a complexity in the order of $\mathcal{O}(N_{\mathrm{RF}}^2M^2P) + \mathcal{O}(P^2N_{\mathrm{RF}}M)$. In step 6, we compute the multiplication between $\mathbf{Z}_{(e)}$ and $\mathbf{W}_{(e)}^i$ at complexity in the order of $\mathcal{O}(N_{\mathrm{RF}}^2M^2P)$. Thus, the encoder of transformer involves complexity in the order of $\mathcal{O}(N_{\mathrm{RF}}^2M^2P) + \mathcal{O}(P^2N_{\mathrm{RF}}M)$. Then, according to **Algorithm 3**, which provides the procedure of decoder, the complexity of the decoder is dominated by the mask-attention mechanism from step 1 to 5, the full-attention mechanism from step 6 to 9, and the FCN processing in step 10. In particular, from step 1 to 5, due to the negligible complexity in the mask operation, the complexity of the mask-attention mechanism is similar to the attention mechanism in the encoder. Therefore, the mask-attention mechanism has the complexity in the order of $\mathcal{O}(N_{\mathrm{RF}}^2M^2(G+L)) + \mathcal{O}((G+L)^2N_{\mathrm{RF}}M)$, based on the decoder input $\overline{\mathbf{H}}_{(d)} \in \mathbb{R}^{2N_{\mathrm{RF}}M\times(G+L)}$, i.e., $m = 2N_{\mathrm{RF}}M$ and $n = G+L$. As for the full-attention mechanism, the full-attention matrix $\mathbf{E} \in \mathbb{R}^{P\times(G+L)}$ is calculated, and the multiplication between $\mathbf{V} \in \mathbb{R}^{2N_{\mathrm{RF}}M\times P}$ and $\mathbf{E}$ are required, which has the complexity in the order of $\mathcal{O}(N_{\mathrm{RF}}^2M^2P) + \mathcal{O}(P(G+L)N_{\mathrm{RF}}M) + \mathcal{O}(N_{\mathrm{RF}}^2M^2(G+L))$. In step 10, the FCN is used to calculate the output of the full-attention mechanism, where the complexity is in the order of $\mathcal{O}(N_{\mathrm{RF}}^2M^2(G+L))$ by multiplying $\mathbf{Z}$ by $\mathbf{W}_{(d)}^i$. As a result, the decoder of transformer model has a complexity in the order of $\mathcal{O}(N_{\mathrm{RF}}^2M^2P) + \mathcal{O}(P(G+L)N_{\mathrm{RF}}M) + \mathcal{O}((G+L)^2N_{\mathrm{RF}}M) + \mathcal{O}(N_{\mathrm{RF}}^2M^2(G+L))$.

Therefore, the overall complexity of the proposed transformer-based parallel prediction model can be summarized as $\mathcal{O}(N_{\mathrm{RF}}^2M^2P) + \mathcal{O}(N_{\mathrm{RF}}^2M^2(G+L)) + \mathcal{O}(P^2N_{\mathrm{RF}}M) + \mathcal{O}((G+L)^2N_{\mathrm{RF}}M) + \mathcal{O}(P(G+L)N_{\mathrm{RF}}M)$, which is closely related to the length of the encoder input P and the length of the decoder input $G+L$. Considering that both the encoder and the decoder use a parallel computing model, the results obtained by each computation based on the previous input can be reused in the next computation, such as the key matrix, query matrix, and value matrix last calculated. Thus, in the next computation, the complexity of the attention mechanism could be reduced from $\mathcal{O}(m^2n) + \mathcal{O}(n^2m)$ to $\mathcal{O}(m^2) + \mathcal{O}(n^2m)$ for the input $\mathbf{X} \in \mathbb{R}^{m\times n}$. Following this, the complexity of the attention mechanism in the encoder and the mask-attention mechanism in the decoder could be reduced from $\mathcal{O}(N_{\mathrm{RF}}^2M^2P) + \mathcal{O}(P^2N_{\mathrm{RF}}M)$ and $\mathcal{O}(N_{\mathrm{RF}}^2M^2(G+L)) + \mathcal{O}((G+L)^2N_{\mathrm{RF}}M)$ to $\mathcal{O}(N_{\mathrm{RF}}^2M^2) + \mathcal{O}(P^2N_{\mathrm{RF}}M)$ and $\mathcal{O}(N_{\mathrm{RF}}^2M^2) + \mathcal{O}((G+L)^2N_{\mathrm{RF}}M)$, respectively. Furthermore, benefiting from parallel hardware architecture [36], the time complexity of the parallel transformer model could be further reduced. Since the different columns of the matrix can be obtained simultaneously, the time complexity of the attention mechanism could be reduced to $\mathcal{O}(m^2) + \mathcal{O}(nm)$. Based on this, the time complexity of the attention mechanism, the mask-attention mechanism, and the full-attention mechanism could be rewritten as $\mathcal{O}(N_{\mathrm{RF}}^2M^2) + \mathcal{O}(PN_{\mathrm{RF}}M)$, $\mathcal{O}(N_{\mathrm{RF}}^2M^2) + \mathcal{O}((G+L)N_{\mathrm{RF}}M)$, and $\mathcal{O}(N_{\mathrm{RF}}^2M^2) + \mathcal{O}(PN_{\mathrm{RF}}M)$. Besides, considering the parallelization, the fully connected networks have a time complexity in the order of $\mathcal{O}(N_{\mathrm{RF}}^2M^2)$. As a result, the encoder and decoder of transformer involve complexities in the orders of $\mathcal{O}(N_{\mathrm{RF}}^2M^2) + \mathcal{O}(PN_{\mathrm{RF}}M)$ and $\mathcal{O}(N_{\mathrm{RF}}^2M^2) + \mathcal{O}((G+L)N_{\mathrm{RF}}M) + \mathcal{O}(PN_{\mathrm{RF}}M)$, respectively. Therefore, by utilizing parallel computing architecture [36], the overall complexity of the proposed transformer-based parallel prediction model can be rewritten as $\mathcal{O}(N_{\mathrm{RF}}^2M^2) + \mathcal{O}(PN_{\mathrm{RF}}M) + \mathcal{O}((G+L)N_{\mathrm{RF}}M)$.

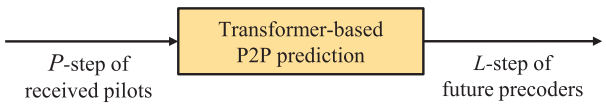


Fig. 7. The proposed transformer-based P2P prediction scheme.

V. Transformer-Based Pilot-to-Precoder (P2P) Prediction

In this section, to reduce the signal processing complexity, we further extend the proposed transformer-based channel prediction scheme to a pilot-to-precoder (P2P) prediction scheme, as shown in Fig. 7. At first, we formulate the P2P prediction problem. Then, the framework of transformer-based P2P prediction is introduced. Finally, the reduced signal processing complexity of P2P is analyzed.

A. Problem Formulation of P2P Prediction

For the transformer-based parallel channel prediction model, the complicated channel estimation is required to estimate the historical channels according to the historical received pilots in previous several frames. Besides, complicated computations are also acquired to the design of the future precoders in next several frames based on predicted channels. As such, the channel estimation and precoding design may cause extremely high computational complexity. To address this issue, we further consider the P2P mapping problem for mmWave massive MIMO. Instead of regarding the channel estimation, channel prediction, and precoding as three independent modules and separately optimizing these modules, we consider the joint design, which regards these three modules as a whole that can be jointly optimized. Unlike the parallel channel prediction problem formulated in Section III, we directly optimize the achievable sum-rate for transformer model, rather than the NMSE of the predicted future channels (12a). Specifically, the same transformer model introduced in Section IV is utilized

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

to directly predict the future precoders in next several frames according to the received pilots in previous frames. Formally, the P2P mapping problem can be formulated as

$$\max_{\mathbf{\Phi}} \ \mathbb{E}\left\{ \sum_{t=T+1}^{T+L} R^{(t)} \right\} \tag{29a}$$

$$\text{s.t.} \quad R^{(t)} = \log_2\left(\left| \mathbf{I}_M + \frac{\rho \mathbf{D}^{(t)} \mathbf{H}_{\mathrm{e}}^{(t)} (\mathbf{D}^{(t)} \mathbf{H}_{\mathrm{e}}^{(t)})^H}{M\sigma_n^2} \right| \right), \tag{29b}$$

$$(\mathbf{D}^{(T+1)}, \cdots, \mathbf{D}^{(T+L)}) = f_{\mathbf{\Phi}}(\mathbf{Y}^{(T-P+1)}, \cdots, \mathbf{Y}^{(T)}), \tag{29c}$$

where $f_{\mathbf{\Phi}}(\cdot)$ is the considered P2P mapping model, $\mathbf{\Phi}$ is model parameters, $\mathbf{D}^{(t)}$ is the predicted precoder matrix at the t-th frame, $\mathbf{H}_{\mathrm{e}}^{(t)}$ is the actual effective channel at the first slot of the t-th frame, and $\mathbf{Y}^{(t)}$ is noisy pilot observation at the t-th frame shown in (5).

A key insight of the considered P2P prediction is that the mutual information between the received pilot and actual channel is larger than the mutual information between the estimated channel and actual channel, i.e., $I(\mathbf{H}_{\mathrm{e}}; \mathbf{Y}) \geq I(\mathbf{H}_{\mathrm{e}}; \hat{\mathbf{H}}_{\mathrm{e}})$. According to the data processing inequality [37], for a Markov process $X \rightarrow Y \rightarrow Z$, then $I(X;Y) \geq I(X;Z)$. This shows that the more processing the data goes through, the less mutual information it contains. For the considered communication system, the effective channel $\mathbf{H}_{\mathrm{e}}$ is regarded as a random variable that follows a distribution, such as the Rayleigh channel following the Rayleigh distribution. Then, through the channel, the transmitted pilot signal $\mathbf{S}$ is affected by channel fading and noise, and noisy pilot observation $\mathbf{Y}$ can also be considered as a random variable. Next, the data processing process of channel estimation is performed according to the received pilot signal $\mathbf{Y}$, and the estimated channel $\hat{\mathbf{H}}_{\mathrm{e}}$ is obtained. Thus, $\mathbf{H}_{\mathrm{e}} \rightarrow \mathbf{Y} \rightarrow \hat{\mathbf{H}}_{\mathrm{e}}$ can be seen as a Markov process. Based on the data processing inequality, the key insight $I(\mathbf{H}_{\mathrm{e}}; \mathbf{Y}) \geq I(\mathbf{H}_{\mathrm{e}}; \hat{\mathbf{H}}_{\mathrm{e}})$ could be derived. By analogy, the precoders calculated from the predicted channel contains generally less information than that of the predicted channel. Thus, it is reasonable to use the historical received pilots to replace the estimated historical channels as the input of the transformer model. For the same reason, the predicted precoders could replace the predicted channel as the output of the transformer model.

B. Proposed Transformer-Based P2P Prediction Framework

Similar to the parallel channel prediction framework proposed in Subsection IV-A, the framework of the proposed P2P prediction scheme consists of two phases, i.e., an offline training and an online prediction. For the offline training phase, the unsupervised learning algorithm is applied to train the transformer model according to (29a). To train the transformer model, the sampling of the $P + L$ effective channels in the contiguous $P + L$ frames is required. The input of the transformer model $\{\mathbf{Y}^{(T-P+1)}, \cdots, \mathbf{Y}^{(T)}\}$ can be obtained by (5) based on the first P effective channels. Then, the last L effective channels and the output of the transformer model $\{\mathbf{D}^{(T+1)}, \cdots, \mathbf{D}^{(T+L)}\}$ construct the optimization objective of achievable sum-rate (29a). Finally, we aim to maximize the achievable sum-rate to obtain the precoding matrix for mmWave massive MIMO.

After training the above transformer-based P2P prediction model, we deploy the well-trained model and realize the precoder prediction online. In the online prediction, we can directly obtain the future precoders $\{\mathbf{D}^{(T+1)}, \cdots, \mathbf{D}^{(T+L)}\}$ in the next L frames, by inputting the historical noisy pilot observations $\{\mathbf{Y}^{(T-P+1)}, \cdots, \mathbf{Y}^{(T)}\}$ from the past P frames into the transformer model. In the process of P2P prediction, by utilizing the powerful transformer model, there is no need for explicit channel requirement, thereby eliminating the needs for channel estimation and precoding modules in proposed the parallel channel prediction scheme.

C. Complexity Analysis

In this subsection, we compare the computational complexity of the proposed P2P prediction scheme with the parallel channel prediction scheme in terms of the number of multiplications. Note that the dedicated channel estimation and precoding are no longer needed in the proposed P2P prediction scheme by directly predicting precoders according to the pilots. Thus, the signal processing complexity of channel estimation and precoding is avoided for the P2P prediction scheme.

To calculate the reduced complexity, we compute the complexity of the dedicated channel estimation and precoding in the parallel channel prediction scheme. Firstly, for the channel estimation, the classical LS channel estimation algorithm (7) and the MMSE channel estimation algorithm (8) are considered, both of which have the complexity in the order of $\mathcal{O}(N_{\mathrm{RF}}^3 M^3)$ due to the high-dimensional matrix inversion. Furthermore, to obtain the estimated channels in the previous P frames, the channel estimation algorithms are executed P times. Therefore, the channel estimation complexity is in the order of $\mathcal{O}(N_{\mathrm{RF}}^3 M^3 P)$. Then, we resort to the zero-forcing algorithm in (10) to solve the precoders, where the pseudo inverse of the predicted channel $\hat{\mathbf{H}}_{\mathrm{e}}^{(t)} \in \mathbb{C}^{N_{\mathrm{RF}} \times M}$ involves complexity in the order of $\mathcal{O}(N_{\mathrm{RF}}^3) + \mathcal{O}(N_{\mathrm{RF}}^2 M)$. To obtain the precoders in the next L frames, the precoding is also executed L times. Thus, the complexity of the precoding design is in the order of $\mathcal{O}(N_{\mathrm{RF}}^3 L) + \mathcal{O}(N_{\mathrm{RF}}^2 ML)$. Therefore, the overall reduced computational complexity of the P2P prediction scheme is in the order of $\mathcal{O}(N_{\mathrm{RF}}^3 M^3 P) + \mathcal{O}(N_{\mathrm{RF}}^3 L) + \mathcal{O}(N_{\mathrm{RF}}^2 ML)$, in which the reduced complexity is dominated by $\mathcal{O}(N_{\mathrm{RF}}^3 M^3 P)$. As a comparison, the complexity of the transformer-based parallel channel estimation in Subsection IV-E is dominated by $\mathcal{O}(P^2 N_{\mathrm{RF}} M)$. According to the 5G standard [38], the number of antenna ports M is set to 2 or 4 or 8 based on different UE categories. The corresponding number of RF chains N_{RF} in BS is usually set to 4. Thus, considering that the length of historical channel P is usually much smaller than $N_{\mathrm{RF}}^2 M^2$, our proposed P2P prediction scheme can greatly reduce the signal processing complexity compared with the parallel channel prediction scheme.

VI. Simulation Results

In this section, we present the performance comparison among the proposed transformer-based parallel channel pre-

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

diction scheme, the P2P prediction scheme, and some existing sequential channel prediction schemes. In order to prove the effectiveness of our work, we provide the simulation results on the CDL-B 3GPP cluster delay line channel model and the 3GPP urban macro (UMa) channel model [25], respectively.

A. Simulation Setup

In our simulations, the parameters of the massive MIMO system are set as: $N_{\mathrm{BS}} = 64$, $N_{\mathrm{RF}} = 4$, $M = 2$, and $f = 28$ GHz. For simplicity, the orthogonal pilot signal matrix $\mathbf{S}$ is considered to be the identity matrix, i.e., $\mathbf{S} = \mathbf{I}_M$. Due to the orthogonality, the length of the training pilot sequence is set to $Q = M = 2$ [39]. The analog combiner is designed based on the first 4 strongest beams in the DFT codebook. In addition, as the displacement of the user is small in tens of milliseconds, the analog combiner is assumed to remain unchanged for considered contiguous $P + L$ frames where $P = 25$ and $L = 5$. Additionally, the length of channels in the previous frames served as the start token of the decoder is set as $G = 10$. The achievable sum-rate performance is calculated by applying (9) for parallel channel prediction and the signal-to-noise ratio (SNR) is defined as $\frac{\rho}{\sigma_n^2}$. According to the 3GPP standard [4], the SRS period is set as 0.625 ms.

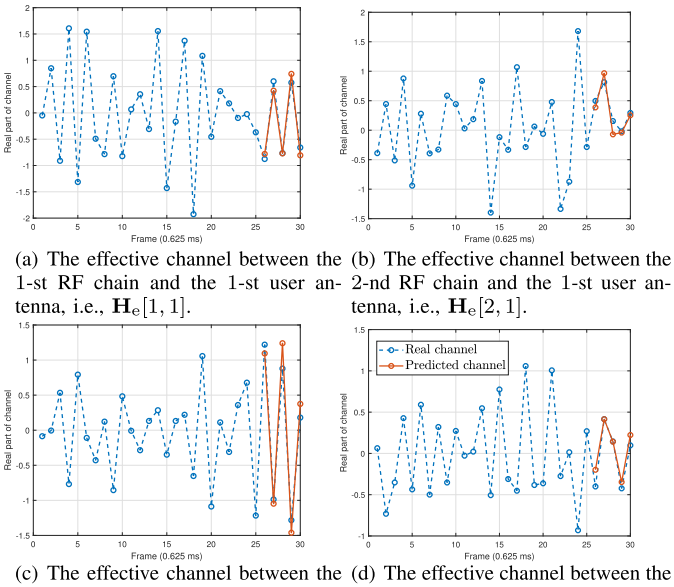
To improve the generalization of our results, 10000 training samples with different user speeds are randomly generated according to the distribution of the user speeds v following $\mathcal{U}[30\mathrm{km/h}, 60\mathrm{km/h}]$. After training, we test the trained transformer model at the speed of 30 km/h and 60 km/h, respectively, and each speed contains 400 test samples. Considering the multi-path effect, the delay spread of the channel is set from 50 to 300 ns randomly, and the SNR is set from 10 dB to 15 dB randomly in the training samples. Note that both the proposed transformer-based prediction schemes and the neural network-based methods use channel parameters introduced above.

B. Simulation Results on the CDL-B Channel Model

In this subsection, based on the CDL-B channel model [25] generated by the Matlab 5G toolbox [28], we provide the performance comparison of the proposed transformer-based parallel channel prediction scheme, the P2P prediction scheme, and some existing sequential channel prediction schemes.

Fig. 8 shows the channel prediction accuracy of the proposed transformer-based channel prediction scheme, where the channel prediction results of $L = 5$ are shown according to $P = 25$ channels in previous frames. The SNR, user speed v, and the delay spread are set as 14 dB, 30 km/h, and 100 ns, respectively. From Fig. 8, we can observe that the proposed transformer-based channel prediction scheme could predict future channels with high accuracy. The gaps between the actual channels and predicted channels are small. We also notice that the prediction accuracy will not degrade with the evolution of the frames. To verify this, we further compare the NMSE performance.

In Fig. 9 and Fig. 10, we compare the NMSE performance versus frame between the proposed transformer-based parallel channel prediction scheme and some existing sequential channel prediction schemes, such as LSTM-based method [14], RNN-based method [13], linear model based prony angular-delay (PAD), and prony vector (PVEC) prediction methods [7]. The SNR is set as 10 dB. The no prediction scheme is also provided by computing the NMSE between the future channel with the current channel. The test scenarios with different user speeds are set as 30 km/h in Fig. 9 and 60 km/h in Fig. 10, while the channel coherence times are roughly equal to 0.643 ms and 0.321 ms, respectively. We can observe from Fig. 9 and Fig. 10 that the proposed transformer-based parallel channel prediction scheme could achieve almost the



(a) The effective channel between the 1-st RF chain and the 1-st user antenna, i.e., $\mathbf{H}_{\mathrm{e}}[1, 1]$.

(b) The effective channel between the 2-nd RF chain and the 1-st user antenna, i.e., $\mathbf{H}_{\mathrm{e}}[2, 1]$.

(c) The effective channel between the 3-rd RF chain and the 1-st user antenna, i.e., $\mathbf{H}_{\mathrm{e}}[3, 1]$.

(d) The effective channel between the 4-th RF chain and the 1-st user antenna, i.e., $\mathbf{H}_{\mathrm{e}}[4, 1]$.

Fig. 8. A parallel channel prediction results of the proposed transformer-based parallel channel prediction scheme, where the $v = 30\mathrm{km/h}$, the delay spread is set as $100\mathrm{ns}$, and $\mathrm{SNR} = 14$ dB. The period of SRS is $0.625\mathrm{ms}$. The blue and orange lines are the real part of the channels and the real part of the predicted channels, respectively. The 4 sub-figures shown above are the effective channels between the 4 RF chains and the first user antenna.

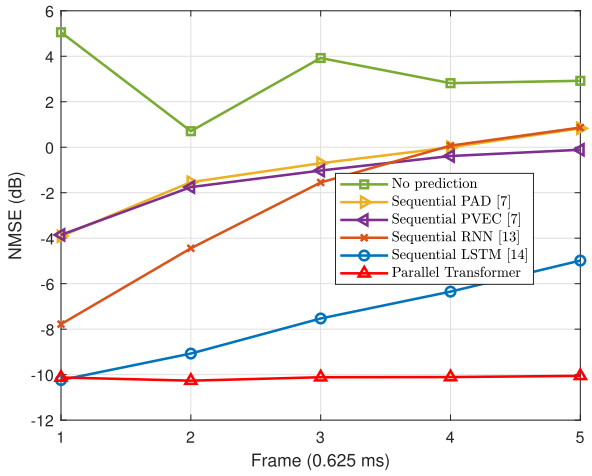


Fig. 9. The NMSE performance versus frame in the CDL-B channel model at $v = 30$ km/h.

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

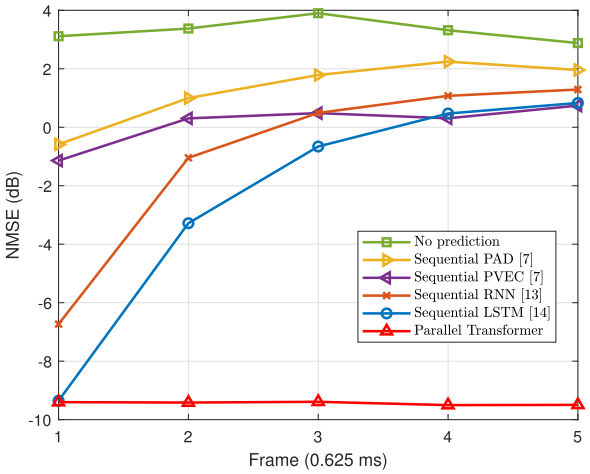


Fig. 10. The NMSE performance versus frame in the CDL-B channel model at $v = 60$ km/h.

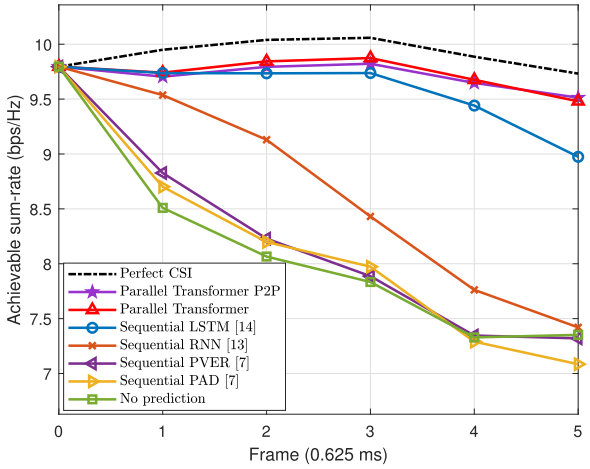


Fig. 11. The achievable sum-rate performance versus frame in the CDL-B channel model at $v = 30$ km/h.

best performance in both $v = 30$ km/h and $v = 60$ km/h, especially the best performance is obtained when predicting the channels from the second to the fifth future frames. Despite the LSTM-based method having comparable NMSE performance when predicting the channel at the first future frame, the transformer-based method shows its stability in performing future channel prediction. For example, for the prediction of the channel at the fifth future frame, the transformer-based scheme could achieve 5 dB and 10 dB NMSE performance gain compared with LSTM-based method at $v = 30$ km/h and $v = 60$ km/h, respectively.

The improvements of the proposed transformer-based parallel channel prediction scheme in Fig. 9 and Fig. 10 come from two aspects. On the one hand, thanks to the parallel signal processing, the future channels in the next several frames could be predicted in parallel, thereby avoiding the prediction error propagation. On the other hand, for the channel with long periodicity, the attention mechanism applied in the transformer model could pay attention to the early parts of historical channels, improving the prediction accuracy. In addition, compared with model-based methods, the NN-based methods have significant performance benefits as the latter can adaptively match the predicted channels with the actual ones.

Fig. 11 and Fig. 12 show the corresponding achievable sum-rate performance versus frame. Additionally, the achievable sum-rate performance of transformer-based P2P prediction is also shown. The upper bound is achieved by the scheme with perfect channel information available. The user speeds are set as 30 km/h in Fig. 11 and 60 km/h in Fig. 12. From Fig. 11 and Fig. 12, we can observe that the proposed transformer model for parallel channel prediction and P2P prediction could approach the near-optimal achievable sum-rate performance achieved by the perfect channel information at two different user speeds scenarios. For example, for the prediction of the channel at the fifth future frame, the channel prediction method with proposed transformer-based can achieve nearly 98% and 97% sum-rate performance of the upper bound exploiting perfect channel information at $v = 30$ km/h and

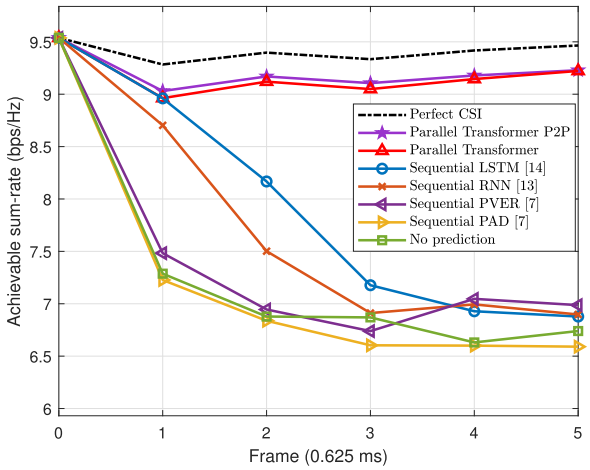


Fig. 12. The achievable sum-rate performance versus frame in the CDL-B channel model at $v = 60$ km/h.

$v = 60$ km/h, respectively. Besides, we can observe that the proposed transformer-based parallel channel prediction and the P2P prediction scheme maintain robustness even in a higher mobility scenario with $v = 60$ km/h, while the LSTM-based method has shown a severe performance degradation.

From Fig. 11, we can observe that the P2P prediction is not always better than the parallel channel prediction. In our view, the parallel transformer P2P prediction mainly adopts unsupervised learning, and there is no precoder label for supervised learning. Thus, the training efficiency of the P2P prediction is slightly worse than that of the supervised channel prediction, and sometimes the performance will not be better than the channel prediction method.

Through the summary of the simulation results in Fig. 8-12, we could conclude that the proposed transformer-based parallel channel prediction scheme can realize higher accuracy when predicting future channels. Besides, both the proposed transformer-based parallel channel prediction and the P2P prediction schemes can effectively alleviate the negative impacts caused by user mobility such that they can approach near-optimal achievable sum-rate performance.

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

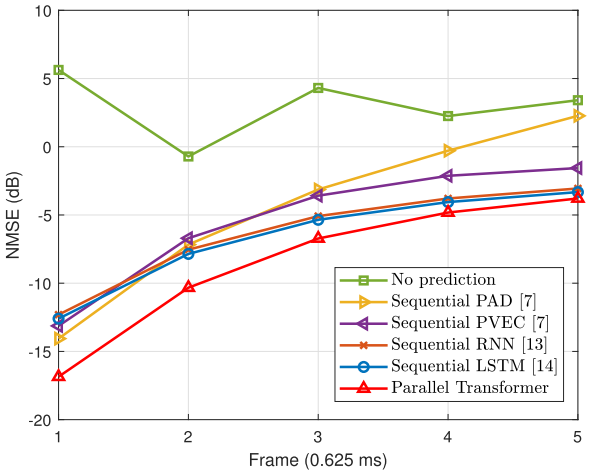


Fig. 13. The NMSE performance versus frame in the UMa channel model at $v = 30$ km/h.

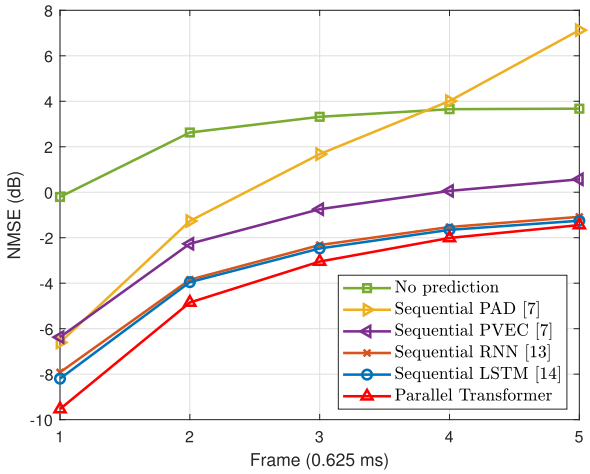


Fig. 14. The NMSE performance versus frame in the UMa channel model at $v = 60$ km/h.

C. Simulation Results on the UMa Channel Model

Since the power of the per-cluster is fixed in the CDL-B model[8]. Thus, we first utilize the QuaDRiGa (QUAsi Deterministic RadIO channel GenerAtor) [40], [41] to generate the 3GPP UMa channel model with random PDP. Then we compare the performance between the proposed parallel channel prediction scheme and the existing sequential channel prediction schemes in the UMa channel model.

The QuaDRiGa channel model [40], [41] follows a geometry-based stochastic channel modeling approach, where the channel parameters such as delay, power, AoA, and AoD, are determined stochastically. Based on statistical distribution for delay spread, delay values, angle spread, shadow fading, etc., specific channel realizations are generated by summing contributions of rays with specific channel parameters. Therefore, the QuaDRiGa channel model can generate the channels with random PDP. Furthermore, the QuaDRiGa channel model also provides a realization of the 3GPP UMa channel model [25].

In Fig. 13 and Fig. 14, we compare the NMSE performance versus frames between the proposed transformer-based parallel channel prediction scheme and some existing sequential channel prediction schemes in the UMa channel model. Testing SNR is set as 20 dB. The no prediction scheme is also provided by computing the NMSE between the future channel with the current channel. The test scenarios with different user speeds are set as 30 km/h in Fig. 13 and 60 km/h in Fig. 14. We can observe from Fig. 13 and Fig. 14 that the proposed transformer-based parallel channel prediction scheme could achieve almost the best performance in both $v = 30$ km/h and $v = 60$ km/h cases.

[8]The CDL-B channel model is suitable for representing the spatial correlation and modelling the channel of analog beamforming architecture which is considered in our paper. Therefore, in our simulation, we first adopted the CDL channel model to show the effectiveness of the proposed transformer-based scheme and further extended it to the UMa channel model., the proposed channel prediction scheme should be verified in the random power delay profile (PDP) channel model to ensure performance improvement.

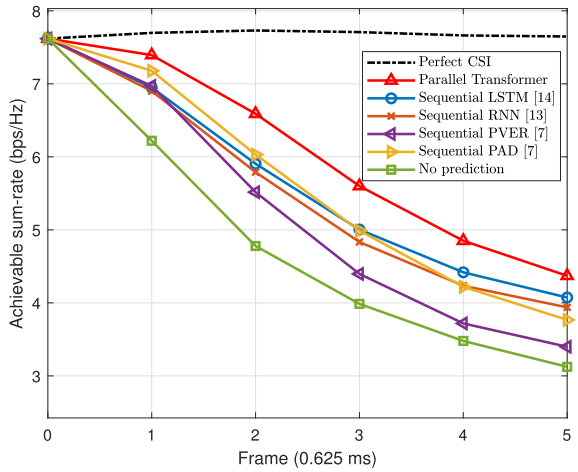


Fig. 15. The achievable sum-rate performance versus frame in the UMa channel model at $v = 30$ km/h.

However, unlike the NMSE performance in the CDL-B channel, which keeps stable when predicting channels in future 5 frames, the NMSE performance in the UMa channel model degrades as the frame increases. One important reason is that the channel parameters such as delay values, angle spread, power, etc. are fixed in the 3GPP CDL-B channel, and thus the trained results converge to the channel model determined by the channel parameters. By contrast, the channel parameters are sampled randomly in the 3GPP UMa channel model, which makes the generated channel more random and less predictable. In this more complex UMa channel model, the proposed transformer-based parallel channel prediction scheme still achieves the best NMSE performance compared with the existing sequential channel prediction schemes. For example, for the prediction of the channel at the second future frames, the transformer-based scheme could achieve 2.5 dB and 1 dB NMSE performance gain compared with LSTM-based method at $v = 30$ km/h and $v = 60$ km/h, respectively.

Furthermore, we show the corresponding achievable sum-rate performance versus frame in Fig. 15 and Fig. 16. The upper bound is achieved by the scheme with perfect channel

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

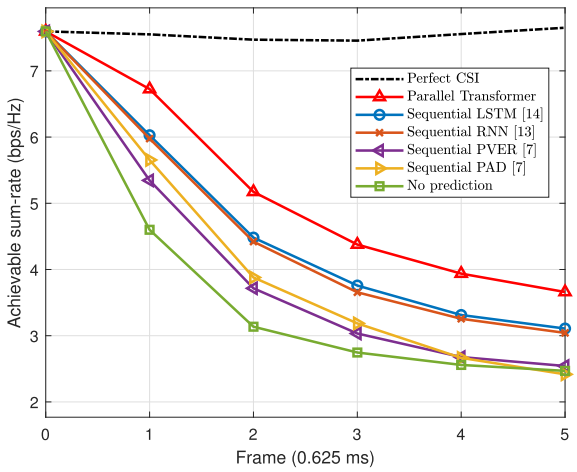


Fig. 16. The achievable sum-rate performance versus frame in the UMa channel model at $v = 60$ km/h.

information available. From Fig. 15 and Fig. 16, we can observe that the proposed transformer model for parallel channel prediction could also achieve the best achievable sum-rate performance. For example, compared with the LSTM-based channel prediction scheme, the channel prediction method with proposed transformer-based at the third future frame can achieve nearly 10% and 15% sum-rate performance improvement at $v = 30$ km/h and $v = 60$ km/h, respectively.

VII. Conclusion

In this paper, we investigated the challenging channel prediction for the mmWave massive MIMO system in the mobility scenarios. Compared to the classical sequential channel prediction methods with severe error propagation problems, we proposed a transformer-based parallel channel prediction scheme to accurately predict the time-varying channels. Moreover, we also proposed a transformer-based P2P prediction scheme to carry out channel estimation, channel prediction, and precoding jointly to significantly reduce the computation complexity for practical communications.

This paper demonstrated that the transformer model, based on the attention mechanism, can process the channel sequence in parallel. In addition, this paper also illustrated that a lower signal processing complexity could be achieved by considering the joint system design of several independent signal processing modules. In future research, we will utilize the transformer-based model to investigate the channel prediction problem in other domains, such as the frequency-domain and the beam-domain channel prediction.

References

[1] S. Mumtaz, J. Rodriquez, and L. Dai, *mmWave Massive MIMO: A Paradigm for 5G*. New York, NY, USA: Academic, 2016.

[2] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[3] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[4] *Radio Resource Control (RCC) Protocol Specification*, document TS 38.331, Version 15.6.0, 3GPP, Jun. 2019.

[5] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[6] K. T. Truong and R. W. Heath, Jr., "Effects of channel aging in massive MIMO systems," *J. Commun. Netw.*, vol. 15, no. 4, pp. 338–351, Sep. 2013.

[7] H. Yin, H. Wang, Y. Liu, and D. Gesbert, "Addressing the curse of mobility in massive MIMO with Prony-based angular-delay domain channel predictions," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2903–2917, Dec. 2020.

[8] H. P. Bui, Y. Ogawa, T. Nishimura, and T. Ohgane, "Performance evaluation of a multi-user MIMO system with prediction of time-varying indoor channels," *IEEE Trans. Antennas Propag.*, vol. 61, no. 1, pp. 371–379, Jan. 2013.

[9] I. C. Wong and B. L. Evans, "Joint channel estimation and prediction for OFDM systems," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Sep. 2005, pp. 2255–2259.

[10] K. E. Baddour and N. C. Beaulieu, "Autoregressive modeling for fading channel simulation," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1650–1662, Jul. 2005.

[11] C. Wu, X. Yi, Y. Zhu, W. Wang, L. You, and X. Gao, "Channel prediction in high-mobility massive MIMO: From spatio-temporal autoregression to deep learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1915–1930, Jul. 2021.

[12] H. Kim, S. Kim, H. Lee, C. Jang, Y. Choi, and J. Choi, "Massive MIMO channel prediction: Kalman filtering vs. Machine learning," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 518–528, Jan. 2021.

[13] W. Jiang and H. D. Schotten, "Neural network-based fading channel prediction: A comprehensive overview," *IEEE Access*, vol. 7, pp. 118112–118124, 2019.

[14] W. Jiang and H. D. Schotten, "Deep learning for fading channel prediction," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 320–332, 2020.

[15] C. Liu, X. Liu, Z. Wei, S. Hu, D. W. K. Ng, and J. Yuan, "Deep learning-empowered predictive beamforming for IRS-assisted multi-user communications," 2021, *arXiv:2104.12309*.

[16] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Jun. 2017, pp. 6000–6010.

[17] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Jan. 2014.

[18] Z. Zhou, J. Fang, L. Yang, H. Li, Z. Chen, and S. Li, "Channel estimation for millimeter-wave multiuser MIMO systems via PARAFAC decomposition," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7501–7516, Nov. 2016.

[19] H. Xie, F. Gao, S. Zhang, and S. Jin, "A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3170–3184, Apr. 2017.

[20] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[21] S. H. Lim, S. Kim, B. Shim, and J. W. Choi, "Deep learning-based beam tracking for millimeter-wave communications under mobility," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7458–7469, Nov. 2021.

[22] C. Liu *et al.*, "Robust adaptive beam tracking for mobile millimetre wave communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1918–1934, Mar. 2021.

[23] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, Jul. 2013.

[24] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[25] *Study on Channel Model for Frequencies From 0.5 to 100 GHz*, document TR 38.901, Version 16.1.0, 3GPP, Dec. 2019.

[26] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[27] J. Tan and L. Dai, "Wideband beam tracking in THz massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1693–1710, Jun. 2021.

[28] MATLAB and MathWorks. *5G Toolbox*. Accessed: Sep. 15, 2020. [Online]. Available: https://www.mathworks.com/products/5g.html

[29] S. Treue and S. Katzner, "Visual attention: Of features and transparent surfaces," *Trends Cogn. Sci.*, vol. 11, no. 11, pp. 451–453, Nov. 2007.

[30] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 37, Jul. 2015, pp. 2048–2057.

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.

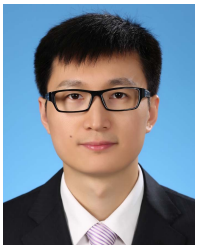
[31] D. B. Ba, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Sep. 2014, pp. 1–15.
[32] D. Xu, C. Ruan, S. Kumar, E. Korpeoglu, and K. Achan, "Self-attention with functional time representation learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Jun. 2019, pp. 15889–15899.
[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2016, pp. 770–778.
[35] R. Xiong *et al.*, "On layer normalization in the transformer architecture," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, vol. 119, Jul. 2020, pp. 10524–10533.
[36] J. Sanders and E. Kandrot, *CUDA by Example: An Introduction to General Purpose GPU Programming*. Trenton, NJ, USA: Addison-Wesley, 2010.
[37] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.
[38] *Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Access Capabilities*, document TS 36.306, Version 16.7.0, 3GPP, Dec. 2021.
[39] X. Gao, L. Dai, S. Zhou, A. M. Sayeed, and L. Hanzo, "Wideband beamspace channel estimation for millimeter-wave MIMO systems relying on lens antenna arrays," *IEEE Trans. Signal Process.*, vol. 67, no. 18, pp. 4809–4824, Sep. 2019.
[40] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, Jun. 2014.
[41] S. Jaeckel, L. Raschkowski, K. Börner, L. Thiele, F. Burkhardt, and E. Eberlein, "QuaDRiGa-quasi deterministic radio channel generator, user manual and documentation," Fraunhofer Heinrich Hertz Inst., Berlin, Germany, Tech. Rep., Version 2.0.0, 2017. [Online]. Available: https://quadriga-channel-model.de/wp-content/uploads/2015/02/quadriga_documentation_v1.2.3.pdf

Hao Jiang (Student Member, IEEE) received the B.S. degree in physics from Tsinghua University, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His research interests include mmWave communications, machine learning for wireless communications, and reconfigurable intelligent surface (RIS).

Mingyao Cui (Student Member, IEEE) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2020, where he is currently pursuing the M.S. degree in electronic engineering. His research interests include massive MIMO, millimeter-wave communications, and near field communications.

Derrick Wing Kwan Ng (Fellow, IEEE) received the bachelor's degree (Hons.) and the M.Phil. degree in electronic engineering from the Hong Kong University of Science and Technology (HKUST) in 2006 and 2008, respectively, and the Ph.D. degree from the University of British Columbia (UBC) in November 2012.

He was a Senior Post-Doctoral Fellow at the Institute for Digital Communications, Friedrich-Alexander-University ErlangenNurnberg (FAU), Germany. He is a Scientia Associate Professor with the University of New South Wales, Sydney, Australia. His research interests include convex and non-convex optimization, physical layer security, IRS-assisted communication, UAV-assisted communication, wireless information and power transfer, and green (energy-efficient) wireless communications. He has been listed as a Highly Cited Researcher by Clarivate Analytics (Web of Science), since 2018. He received the Australian Research Council (ARC) Discovery Early Career Researcher Award 2017, the IEEE Communications Society Stephen O. Rice Prize 2022, the Best Paper Awards at the WCSP 2020, 2021, the IEEE TCGCC Best Journal Paper Award 2018, the INISCOM 2018, the IEEE International Conference on Communications (ICC) 2018, 2021, the IEEE International Conference on Computing, Networking and Communications (ICNC) 2016, the IEEE Wireless Communications and Networking Conference (WCNC) 2012, the IEEE Global Telecommunication Conference (Globecom) 2011, 2021, and the IEEE Third International Conference on Communications and Networking in China 2008. He was an Editorial Assistant to the Editor-in-Chief of the IEEE TRANSACTIONS ON COMMUNICATIONS, from January 2012 to December 2019. He is now serving as the Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and an Area Editor of the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.



Linglong Dai (Fellow, IEEE) received the B.S. degree from Zhejiang University, Hangzhou, China, in 2003, the M.S. degree (Hons.) from the China Academy of Telecommunications Technology, Beijing, China, in 2006, and the Ph.D. degree (Hons.) from Tsinghua University, Beijing, in 2011.

From 2011 to 2013, he was a Post-Doctoral Research Fellow at the Department of Electronic Engineering, Tsinghua University, where he was an Assistant Professor from 2013 to 2016, and has been an Associate Professor, since 2016. His current research interests include reconfigurable intelligent surface (RIS), massive MIMO, millimeter-wave and terahertz communications, and machine learning for wireless communications.

He has coauthored the book *MmWave Massive MIMO: A Paradigm for 5G* (Academic Press, 2016). He has authored or coauthored over 70 IEEE journal papers and over 40 IEEE conference papers. He also holds 19 granted patents. He has received five IEEE Best Paper Awards at the IEEE International Conference on Communications (ICC) 2013, the IEEE ICC 2014, the IEEE ICC 2017, the IEEE Vehicular Technology Conference (VTC) 2017-Fall, and the IEEE ICC 2018. He has also received the Tsinghua University Outstanding Ph.D. Graduate Award in 2011, the Beijing Excellent Doctoral Dissertation Award in 2012, the China National Excellent Doctoral Dissertation Nomination Award in 2013, the URSI Young Scientist Award in 2014, the IEEE TRANSACTIONS ON BROADCASTING Best Paper Award in 2015, the Electronics Letters Best Paper Award in 2016, the National Natural Science Foundation of China for Outstanding Young Scholars in 2017, the IEEE Communications Society (ComSoc) Asia-Pacific Outstanding Young Researcher Award in 2017, the IEEE Communications Society Asia-Pacific Outstanding Paper Award in 2018, the China Communications Best Paper Award in 2019, the IEEE ACCESS Best Multimedia Award in 2020, the IEEE Communications Society Leonard G. Abraham Prize in 2020, the IEEE Communications Society Stephen O. Rice Prize in 2022, and the IEEE ICC Best Demo Award in 2022. He was listed as a Highly Cited Researcher by Clarivate Analytics in 2020 and 2021. He is an Area Editor of IEEE COMMUNICATIONS LETTERS. Particularly, he is dedicated to reproducible research and has made a large amount of simulation codes publicly available.

Authorized licensed use limited to: Chungnam National University. Downloaded on March 21,2025 at 08:30:37 UTC from IEEE Xplore. Restrictions apply.