Transformer-Based Long Distance Fiber Channel Modeling for Optical OFDM Systems

Niuyong Zhang, Hang Yang, Zekun Niu, Lizhuo Zheng, Cao Chen, Shilin Xiao, and Lilin Yi

***Abstract*—The fiber channel model plays an essential role in the simulation and design of optical fiber communication systems. However, it is difficult for conventional model-driven modeling to balance accuracy and efficiency, especially in optical orthogonal frequency division multiplexing (OFDM) systems with complex and long-haul transmission. We introduce the simplified Transformer into optical OFDM systems and combine it with the feature decoupled distributed (FDD) scheme for fast and accurate fiber channel modeling. Unlike the popular Transformer architectures, we remove the Decoder part and cancel the self-attention with quadratic complexity, significantly reducing the computational cost. The modeling performance is investigated from the nonlinear fitting capability, accuracy, and generalization ability. The transmission distance ranges from 80 km to 1600 km. The highly matched four-wave mixing (FWM) power, low error vector magnitudes (EVMs), and similar signal-noise ratios (SNRs) demonstrate the high precision and robustness of the model. Furthermore, the modeling is studied under different transmission rates and is proved to be reliable over a wide transmission bandwidth. Compared to the bidirectional long short-term memory (Bi-LSTM), the Transformer performs better in accuracy and has lower computational and memory costs. For modeling under the same conditions, the required running time of the Transformer is about 60% of Bi-LSTM, less than 1% of the split-step Fourier method (SSFM). The Transformer-based method achieves high precision modeling of the fiber channel in the long-distance and high-rate optical OFDM system and makes a significant breakthrough in complexity.**

***Index Terms*—Split-step fourier method (SSFM), bidirectional long short-term memory (Bi-LSTM), feature decoupled distributed (FDD), fiber channel modeling, optical orthogonal frequency division multiplexing (OFDM), transformer.**

I. Introduction

THE modeling of optical fiber channels is essential for the simulation and design of optical transmission systems. The signal transmission in optical fiber can be described by the nonlinear Schrödinger equation (NLSE) [1], but it is generally impossible to obtain its analytical solution directly. Traditional fiber channel modeling is based on the split-step Fourier method (SSFM) [1]. As a model-driven method, it obtains an approximate solution of the NLSE through iterative calculations, meaning a high computational cost. Especially in the case of long-distance transmission, the number of iterations rises linearly, and its complexity will be unbearable.

Machine learning (ML) as a data-driven approach has been applied to fiber channel modeling to break the barriers of model-driven approaches. ML algorithms such as generative adversarial network (GAN) and bidirectional long short-term memory (Bi-LSTM) show extraordinary potential in fiber channel modeling [2], [3], and the modeling scenarios have evolved from simple on-off keying (OOK) and pulse amplitude modulation 4 (PAM4) to 16 quadrature amplitude modulation (16QAM) single-carrier transmission system. However, GAN has limited accuracy due to its training instability and mode collapse [4]. As a variant of recurrent neural network (RNN), Bi-LSTM has high time complexity due to its sequential computation. Furthermore, as an overall modeling scheme, these works are difficult to extend to more complex scenarios to accurately model all channel effects during long-haul fiber transmission. As a hybrid model-data-driven method, the proposed feature decoupled distributed (FDD) scheme [5] greatly improves the accuracy of machine learning-based fiber channel modeling. This work is based on the Bi-LSTM algorithm, realized by linear-nonlinear decoupling and recursive processing, and the transmission distance in multi-channel wavelength division multiplexing (WDM) systems reaches 1040 km. Nevertheless, the Bi-LSTM used still has high complexity.

With the advantages of high spectrum utilization and dispersion robustness, optical orthogonal frequency division multiplexing (OFDM) [6] is a research hotspot for realizing high-speed and long-haul transmission. Distinguish from other optical fiber communication systems such as WDM, OFDM has the characteristics of high peak to average power ratio (PAPR), dense channels, and narrow channel spacing, which means stronger nonlinearity, especially four-wave mixing (FWM) [7], [8], [9]. The higher the launched optical power and the longer the transmission distance, the nonlinearity of fiber will increase accordingly. In addition, more chromatic dispersion (CD) and amplified spontaneous emission (ASE) noise will accumulate over a long distance. Therefore, it is challenging to model the optical fiber channel with total field effects, especially for optical OFDM systems.

Manuscript received 15 May 2022; revised 28 July 2022; accepted 2 September 2022. Date of publication 8 September 2022; date of current version 20 December 2022. This work was supported by the National Natural Science Foundation of China under Grants 62025503, 62071295, and 61775137. *(Corresponding authors: Shilin Xiao; Lilin Yi.)*

The authors are with the State Key Lab of Advanced Optical Communication Systems and Networks, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: summer146828@sjtu.edu.cn; hangyang@sjtu.edu.cn; zekunniu@sjtu.edu.cn; zlzsjmailbox@sjtu.edu.cn; chen.cao@sjtu.edu.cn; slxiao@sjtu.edu.cn; lilinyi@sjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JLT.2022.3205158.

Digital Object Identifier 10.1109/JLT.2022.3205158

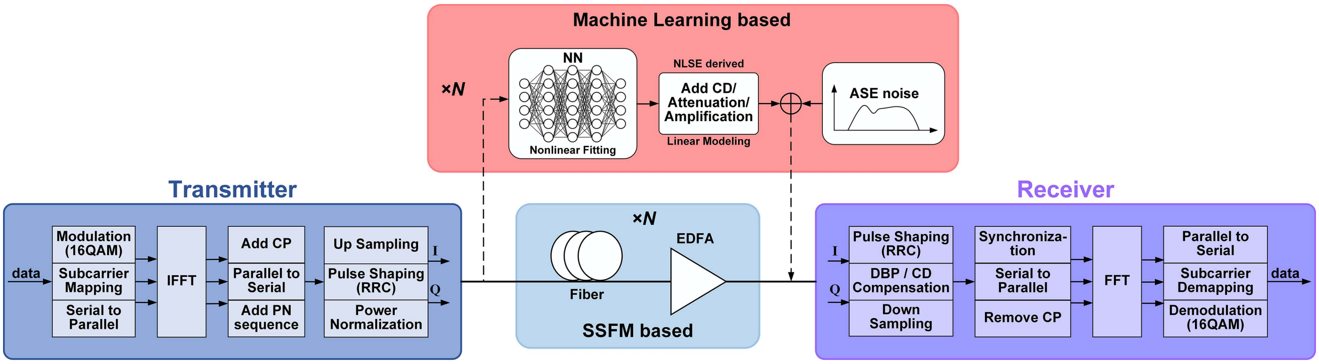

Fig. 1. Architecture of the coherent optical orthogonal frequency division multiplexing (CO-OFDM) system. The system consists of transmitter, fiber channel, and receiver, describing the generation, transmission, and post-processing of the optical OFDM signal. The machine learning-based approach replaces the split-step Fourier method (SSFM)-based for fiber channel modeling.

The Transformer based on self-attention is an ML algorithm that was proposed by Google Brain in 2017 [10]. It is used primarily in the fields of natural language processing and computer vision [11], [12]. Recently, Transformer performed the best in many ML tasks [13], [14], [15], [16], becoming a dominant model in ML. The innovation of using self-attention, which captures global dependencies between input and output, avoids the constraint of sequential computation and allows significantly more parallelization [10]. Moreover, there are more merits to its architecture, including the multi-head mechanism, the feedforward network (FFN), and the use of residual connection [17] and layer normalization [18]. The Transformer has been proven to be abstractable into general architecture [19]. The Transformer is currently well acknowledged as the most powerful ML algorithm for modeling sequential data [20], [21], [22], [23], [24] and has significant advantages over RNN in computational efficiency. Based on these characteristics and advantages of the model, it is attractive to use the Transformer to model the fiber channel more accurately and quickly.

In this paper, the Transformer is introduced into optical OFDM systems for the first time to realize fiber channel modeling, including the channel effects of CD, nonlinearity, and ASE noise. We remove the Decoder part and cancel the self-attention mechanism of the Transformer, and keep the multi-head, residual connection, and FFN to make the model as simple and accurate as possible. The FDD scheme is incorporated to improve accuracy and generalization power for transmission distance. Furthermore, a satisfactory model is obtained through the optimization of the training dataset, the model training improvement, and the design of loss function. The simulation capability for nonlinearity is tested by measuring FWM power, and the modeling accuracy is represented by error vector magnitudes (EVMs) and signal-noise ratios (SNRs). The complexity analysis is carried out from two dimensions, the number of multiplications and the running time. Additionally, the modeling performance is investigated under different transmission rates, including 30 GBd, 160 GBd, and 500 GBd. Results show that the Transformer has high precision in simulating nonlinearity, with the average relative errors of FWM power on 13 sub-channels less than 1.3%. Low EVMs and similar SNRs demonstrate that the Transformer-based model works on different OFDM signals without performance degradation (whether with different subcarrier numbers or different PAPRs) and outperforms the Bi-LSTM in accuracy. In modeling of 1600 km, the multiplication number of the Transformer is about 55% of Bi-LSTM and 30% of SSFM, and the required running time is about 60% of Bi-LSTM and 0.8% of SSFM, which highlights the superiority of the Transformer-based modeling in complexity. Therefore, the Transformer-based method realizes fast and high-precision fiber channel modeling in the long-distance and high-rate optical OFDM system.

II. System Architecture and Optical Fiber Channel Modeling

In this section, we build a coherent optical OFDM (CO-OFDM) communication system for demonstration. The proposed modeling method is also applicable for intensity modulation and direct detection (IM-DD) optical OFDM systems. Our work aims to enable the Transformer-based optical fiber channel to approximate the SSFM-based optical fiber channel regarding its modeling ability.

A. *Optical OFDM Communication System*

To prove the feasibility of the Transformer-based fiber channel modeling method in optical OFDM systems, a digital communication system based on coherent detection is built, as shown in Fig. 1. Notice that all samples and symbols in this system are represented by complex values.

At the transmitter, a binary sequence signal is first modulated by 16QAM. After serial-to-parallel conversion, subcarrier mapping, inverse fast Fourier transform (IFFT), cyclic prefix (CP) insertion, and parallel-to-serial conversion, pseudo-noise (PN) is added before the serial OFDM signal for frame synchronization operation at the receiver. To ensure that the digital signal is equivalent to the analog signal, the OFDM signal is four times up-sampling before entering the optical fiber channel, and the signal shaping is performed through a root raised cosine (RRC) filter to meet the Nyquist criterion of no inter-symbol interference (ISI) [25]. Finally, after adjusting to a certain optical power

TABLE I
PARAMETERS OF SSFM

Parameters	Value
Carrier wavelength	1550 nm
Symbol rate	30 GBd
Attenuation	0.2 dB / km
Core area	80 μm^2
Dispersion	16.75 ps / (nm · km)
Dispersion slope	0.075 ps / (nm^2 · km)
Nonlinear refractive index	2.6e-8 μm^2 / W
Step length	0.01 km
Span length	80 km
Launched optical power	4 dBm
Noise figure	5 dB

by power normalization, the OFDM signal enters the fiber channel for transmission. When the OFDM multi-carrier modulation conforms to Hermitian conjugate mapping, the real OFDM signal required for intensity modulation will be generated, and a photodiode detector (PD) can be used in the receiver to realize direct detection. Then the system changes from CO-OFDM to IM-DD optical OFDM.

At the receiver, the OFDM signal is first passed through a matched RRC filter, and then digital backward propagation (DBP) compensation [26] or CD compensation is used. The DBP algorithm is a fiber channel compensation algorithm based on SSFM, which can be regarded as the inverse process of fiber channel modeling. After quadruple down-sampling, the frame synchronization operation based on the known PN sequence is followed. Then, corresponding to the transmitter, a series of operations are performed to recover the required binary data, including the serial-to-parallel conversion, CP removal, fast Fourier transform (FFT), subcarrier demapping, parallel-to-serial conversion, and 16QAM demodulation.

The entire optical fiber channel is composed of multiple spans with the same structure, and each span contains two parts: standard single-mode fiber (SMF) and erbium-doped fiber amplifier (EDFA) [27]. EDFA is used for optical signal amplification while introducing ASE noise. Transmission of the OFDM signal in an optical fiber channel is simulated by the SSFM at first. The main optical fiber channel parameters are shown in Table I.

The propagation of an optical signal in SMF can be described by the NLSE, which is expressed as [28]

$$\frac{\partial u}{\partial z} + \frac{j\beta_2}{2}\frac{\partial^2 u}{\partial t^2} - \frac{\beta_3}{6}\frac{\partial^3 u}{\partial t^3} = j\gamma |u|^2 u - \frac{\alpha}{2}u, \qquad (1)$$

where u is the complex envelope of the optical field, z is the propagation distance, and t is the time. And α represents the attenuation, β_2 represents the group velocity dispersion parameter, β_3 represents the slope of the group velocity dispersion, γ represents the nonlinear coefficient.

We further adopt the FDD scheme when modeling the fiber channel with Transformer. As shown in Fig. 1, in each span, the nonlinear effects are modeled by a neural network (NN), and the linear effects are modeled by the NLSE-derived method in one step. Linear-nonlinear feature decoupling [5] is implemented on the output signal to eliminate linear effects. This can reduce the ISI length and strengthen the nonlinear characteristics of data, dramatically improving modeling accuracy while maintaining low complexity. The linear decoupling is accomplished by CD compensation [29], and the transfer function can be expressed as

$$H(-L, \omega) = \exp\left[\left(-\frac{j\beta_2}{2}\omega^2 + \frac{\beta_3}{6}\omega^3\right)(-L)\right], \qquad (2)$$

where L is the transmission distance.

The NN model is only used for fitting nonlinearity, so it needs to add CD (the inverse process of CD compensation) through the conventional method. The transfer function of CD modeling can be expressed as $H(L, \ \omega)$. In this work, we set the linear compensation and modeling distance at one span length.

B. The Transformer Settings

The conventional Transformer model consists of two parts: Encoder and Decoder. Each Encoder layer has two sub-layers: the multi-head attention layer and the FFN layer. Residual connection is employed for each sub-layer, followed by layer normalization. The Transformer is originally designed for neural machine translation (NMT) tasks, so the Embedding needs to be used to convert the input word vectors into numeric feature vectors of a particular dimension. The positional encoding is added to the feature vectors to compensate for the absence of positional information in the self-attention mechanism. For each feature vector, self-attention computes a weighted sum of the features by the dot products with all other feature vectors to form an output vector [10], which captures the dependencies among all feature vectors. The structure of the Decoder is similar to the Encoder. The difference is that the Decoder is an autoregressive model that uses Encoder-Decoder attention besides self-attention. The auto-regressive property is preserved by masking, i.e., the information in the Decoder flows unidirectionally in position order.

It is found that a pure self-attention network (SAN), that is, the Transformer without residual connection and multi-layer perceptron (MLP), loses its expressive ability exponentially in terms of network depth [30]. Residual connection facilitates optimization and gradient flow and plays an essential role in preventing network degradation, which is also mitigated by FFN [17], [30], [31]. The work of [32] shows that the multi-head mechanism, residual connection, and FFN are keys to the Transformer model, and there is no additional benefit from source attention on lower encoder layers. [33] replaces self-attention with two cascaded linear layers and two normalization layers in the Transformer architecture and further incorporates the multi-head mechanism into the full MLP model, showing comparable or better performance to the self-attention and some of its variants in extensive experiments. The above researches indicate that self-attention is not necessary for the Transformer, while the vital things are the multi-head mechanism, residual connection, and FFN.

In the nonlinearity modeling work, the self-attention brings no additional benefit but leads to quadratic computational complexity related to the number of input feature vectors. So, the

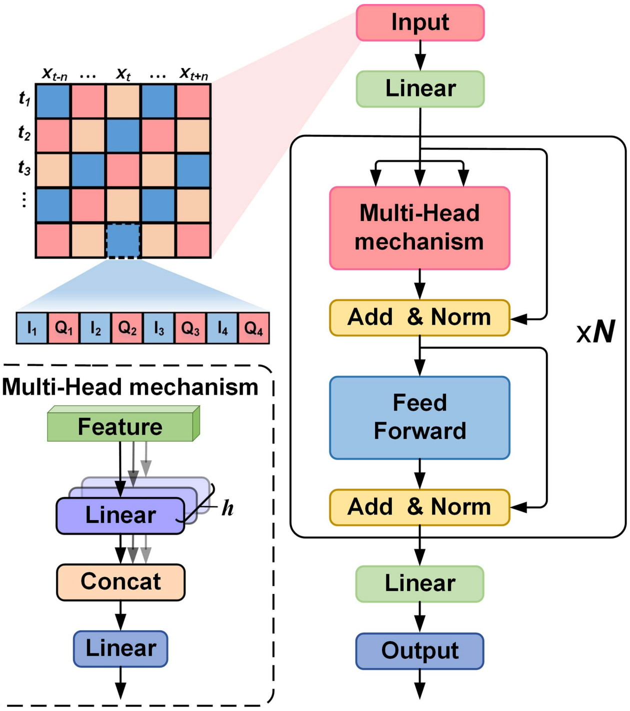

Fig. 2. Structure of the simplified Transformer. Only the Decoder is used here, and the self-attention is canceled. Each model layer has two sub-layers, a multi-head mechanism, and a feedforward network. Residual connection is used around each sub-layer, followed by layer normalization. The input data contain samples of both the current time node and before-after time nodes, and the total time nodes is set to 11.

self-attention mechanism is canceled. Furthermore, to enable the data to be output in parallel to improve efficiency, only the Encoder is used. The modified simplified Transformer architecture is shown in Fig. 2.

The ISI caused by CD effects results in temporal correlations among adjacent samples and can be captured using a sliding window. With linear feature decoupling, a shorter ISI length is preserved in the dataset, which means a smaller sliding window is required. And the smaller the window size, the fewer the NN parameters and the lower the computational complexity. As shown in Fig. 2, the data is arranged into vector structures through a sliding window, and each vector needs to contain both the current time node and before-after time nodes. It is found that when the sliding window contains 11 time nodes (five past time nodes, one current time node, and five future time nodes), the Transformer model achieves the best performance in most cases. There are four samples per time node, and one sample consists of two real numbers representing the real and imaginary parts of a complex number. We can get the initial input dimension as $11 \times 4 \times 2 = 88$. The final output of the model focuses on the samples of the current time node, avoiding the problem of lacking position information. Therefore, there is no need for positional encoding.

The Embedding is designed for the word vector in NMT, which is unsuitable for waveform inputs. Thus, we remove the Embedding layer from the Transformer. The low feature dimension will lead to the underfitting of the model. To improve the fitting ability of the model, we use the Linear layer with more neurons to extract features. This process can be expressed as

TABLE II
PARAMETERS OF TRANSFORMER

Parameters	Value
Input dimension	88
Feature dimension	256
Number of heads	8
Feedforward dimension	1024
Number of layers	1
Output dimension	8

$$X = \text{Linear}\,(X_{in}) = X_{in}W_{in}, \tag{3}$$

where $W_{in} \in \mathrm{R}^{d\_in \times d\_model}$ is the parameter matric, $X_{in} \in \mathrm{R}^{1 \times d\_in}$ is the initial input vector, and $X \in \mathrm{R}^{1 \times d\_model}$ is the feature vector. The *d_in* represents the initial input dimension, and the *d_model* represents the feature dimension.

The multi-head mechanism allows the model to attend to the information from different representation subspaces of the data input [10], as shown in Fig. 2. The multi-head layer performs multiple different linear projections on the input feature vector and concatenates all the results to project once again to generate an output vector. This process can be expressed as

$$\begin{aligned} &V = \text{MultiHead}(X) = \text{Concat}(head_1, \ldots, head_h)W^o, \\ &\text{where } head_i = XW_i, \end{aligned} \tag{4}$$

where the projections are parameter matrices $W_i \in \mathrm{R}^{d\_model \times d\_v}$ and $W^o \in \mathrm{R}^{hd\_v \times d\_model}$, $V \in \mathrm{R}^{1 \times d\_model}$ is the output, *d_v*=*d_mode* / *h*, and *h* is the number of heads. "Concat" means concatenating the *h* vectors of dimension $1 \times d\_v$ into a vector of dimension $1 \times hd\_v$.

The fully connected FFN consists of two linear transformation layers, activated by a ReLU [34]. Expressed as

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{5}$$

where $W_1 \in \mathrm{R}^{d\_modeld \times d\_ff}$ and $W_2 \in \mathrm{R}^{d\_ff \times d\_model}$ are the parameter matrices, and *d_ff* represents the feedforward dimension.

The output of the sub-layer after residual connection and layer normalization can be expressed as

$$q = \text{LayerNorm}(x + \text{Sublayer}(x)), \tag{6}$$

where Sublayer(•) is the function implemented by the sub-layer. It represents the MultiHead(•) in the first sub-layer and the FFN(•) in the second sub-layer. The *q* has the same dimension as the input *x*.

The output of the Encoder is fed into a Linear layer to obtain the final output, i.e., a vector $Y \in \mathrm{R}^{1 \times d\_out}$. And *d_out* represents the final output dimension.

The parameters of the Transformer model are set as shown in Table II.

C. Model Training Method

The preprocessing of the training dataset is shown in Fig. 3. Notice that ASE noise will lead to difficult training because the NN employed cannot directly handle random features. We do

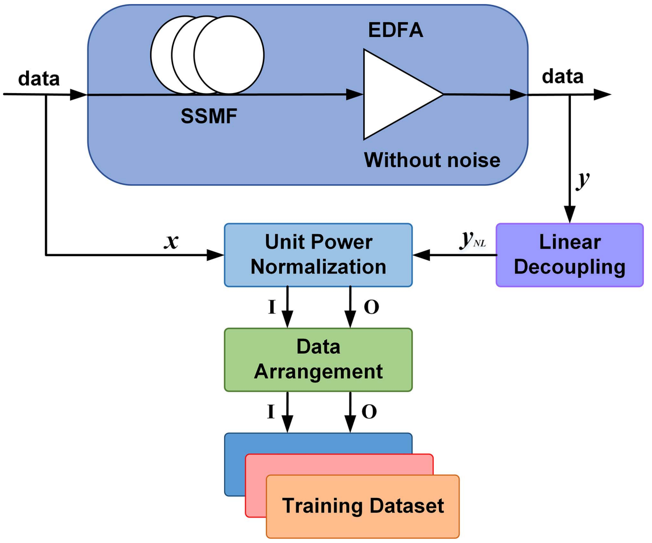

Fig. 3. Preprocessing of the training dataset. This process includes linear-nonlinear decoupling, power normalization, and data arrangement.

not consider ASE noise and collect OFDM signal transmission samples before and after a single span. Linear feature decoupling is performed on the output signal to eliminate linear effects, followed by unit power normalization. Unit power normalization ensures the average absolute value of training data is around 1, which is conducive to faster convergence and higher accuracy of the model [3]. Finally, we arrange the data into vector structures through a sliding window.

The samples collected from the second or later spans are needed when forming the training dataset. The OFDM signal undergoes four times up-sampling and passes through an RRC filter before entering the fiber channel. For the first span of the fiber channel, the high-frequency part (outside the OFDM signal frequency band) of the input signal is almost completely filtered out. After transmission through an optical fiber channel, the signal's frequency spectrum is broadened by nonlinear effects, so its high-frequency part will also contain information. The input signal of the first span is significantly different from that of other spans in spectral distribution. Therefore, it is necessary to avoid collecting samples from the first span.

The OFDM signal is transmitted for 10 spans, and we collect samples from input-output signals of the 2nd, 4th, 6th, 8th, and 10th spans to form five datasets. After these five datasets are preprocessed, including linear feature decoupling, unit power normalization, and data arrangement, a total training dataset is formed, ensuring sample diversity. In the training process, a method of randomly selecting a dataset is adopted. An integer of 1-5 is randomly generated before the beginning of each training epoch, then the dataset with the corresponding number is used for the current training. We set epoch to 200 and batch size to 1000. The optimizer used is Adam [35], and the learning rate (LR) is set to 1e-4 at the beginning and gradually decreases with the training process.

The loss function is designed in the form of relative error. Mean square error (MSE) is generally used for the training loss function. It represents the average of the square of absolute error between generated value and real value [36], which can be expressed as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\left|X_{generated,i} - X_{real,i}\right|\right)^2, \tag{7}$$

where N is the data length, X_{real} is the real data, and $X_{generated}$ is the data generated by the NN model. It may be different for the average power of different signals or different signal parts, so there is no unified reference standard for MSE as an absolute-like-error indicator. To a certain extent, MSE cannot truly reflect the gap between generated and real value. Therefore, we refer to the concept of relative error and design the loss function as

$$Loss = \frac{\sum_{i=1}^{N}\left(\left|X_{generated,i} - X_{real,i}\right|\right)^2}{\sum_{i=1}^{N}\left(\left|X_{real,i}\right|\right)^2}. \tag{8}$$

III. Results and Discussion

In the simulation, we set the length of a single span to 80 km and kept the launched optical power at 4 dBm. We adopt the FDD scheme to model the long-haul optical fiber transmission. In the FDD scheme, the modeling is based on one-span transmission. In this case, the recursive input to the one-span model can realize long-distance transmission. That is, the trained model is iteratively reused across all spans without training the corresponding model for each span. The entire line is composed of multiple spans with the same structure, and ASE noise is added between adjacent spans. For example, a transmission distance of 1600 km consists of 20 spans of 80 km. For the training dataset, the parameters of the OFDM signal are set as follows: the number of subcarriers is 256, the length of one OFDM symbol is 320 after adding the cyclic prefix, the total number of OFDM symbols is 12000, and the length of PN is 128. It can be calculated that the data length of the total training dataset is $(320 \times 12000 + 128) \times 4 \times 2 = 30721024$.

The performance of the Transformer is presented from different dimensions, including the nonlinearity fitting capability, accuracy, and adaptability to different OFDM signals (with different subcarrier numbers and PAPR values). Moreover, the modeling under different transmission rates is studied. Accuracy and complexity comparisons are also performed with the Bi-LSTM. In the setting of Bi-LSTM, the number of LSTM layers is 1, the number of Linear layers is 1, and no nonlinear activation function is used. Each LSTM layer contains only two cells, feeding the real part data into one cell and the imaginary part data into the other. This approach resulted in a model with higher accuracy and lower complexity than arranging the input according to time nodes. Similarly, a sliding window is employed to help Bi-LSTM grasp the correlation among samples. It is found that in most cases, Bi-LSTM can achieve optimal performance when the window contains 21 samples (ten pre-samples, one current sample, and ten post-samples). The final output of Bi-LSTM focuses on one sample. All the parameters of Bi-LSTM are set as shown in Table III. Notice that both Transformer and Bi-LSTM models are trained under the same conditions, including the same training dataset, training

TABLE III
PARAMETERS OF BI-LSTM

Parameters	Value
Input dimension	42
Number of cells per layer	2
Hidden dimension	128
Number of LSTM layers	1
Number of Linear layers	1
Output dimension	2

method, loss function, etc. Linear decoupling is performed for all training datasets during preprocessing.

A. Nonlinear Fitting

The total optical field of the OFDM signal can be represented as [37]:

$$u(z,t) = \sum_{i}^{N} u_i(z,t)\exp(-j\Omega_i t), \tag{9}$$

where$\Omega_i = \omega_i - \omega_0$,$\omega_i$is the frequency of i-th sub-channel, andω_0 is the reference carrier frequency.

After substituting the OFDM signal (9) into the NLSE (1) and simplifying it, the resulting equation for i-th sub-channel takes the form:

$$\frac{\partial u_i}{\partial z} + \Omega_i\beta_2\frac{\partial u_i}{\partial t} + \frac{j\beta_2}{2}\frac{\partial^2 u_i}{\partial t^2} = \frac{j}{2}\beta_2\Omega_i^2 u_i - \frac{\alpha}{2}u_i + j\gamma \times \left(|u_i|^2 + 2\sum_{q\neq i}^{N}|u_q|^2\right)u_i + j\gamma\sum_m\sum_n\sum_k u_m u_n u_k^*. \tag{10}$$

On the right of (10), the third term represents the self-phase modulation (SPM) and cross-phase modulation (XPM) effects related to the signal energy of i-th sub-channel. The last term represents the FWM, and the triple sum is limited to those frequency combinations that satisfy the matching condition [37]: $\omega_m + \omega_n - \omega_k = \omega_i, m \neq i \neq n, i = 1, 2, \ldots, N$.

It is easy to meet the frequency matching condition of FWM for the OFDM signal because of the dense sub-channels and the narrow and equal channel spacing. Therefore, compared with other multi-channel optical fiber transmission systems, the FWM in optical OFDM systems will be more serious, which becomes the main nonlinear damage. In addition, the power of FWM noise in the center of the OFDM band will be higher relative to the edge [7], [38], [39].

It can be seen from (10) that for i-th sub-channel, the generation of SPM and XPM on it must involve the participation of the i-th subcarrier, while the FWM can be generated by the combination of only other sub-carriers when satisfying frequency matching (without the participation of the i-th subcarrier). Assuming that a sub-channel is empty when entering optical fiber, then only FWM noise is generated on this sub-channel during transmission.

We verify the accuracy of the Transformer for simulating nonlinear effects by testing the FWM power. The FWM noise power is obtained by setting a specific sub-channel empty before IFFT at the transmitter and then measuring signal power on the corresponding sub-channel after FFT at the receiver. Notice that DBP compensation will eliminate nonlinear effects, including the FWM noise, so only CD compensation is utilized at the receiver to obtain pure FWM noise.

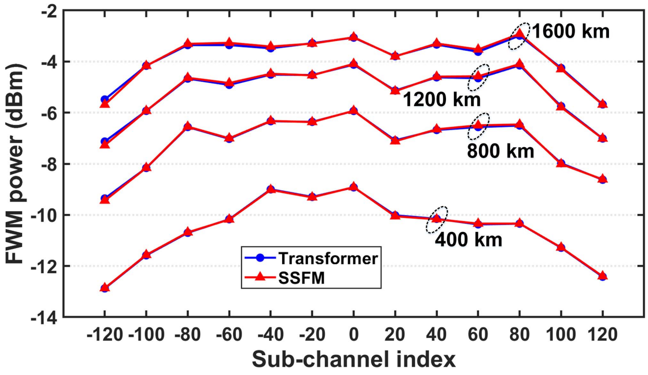

Fig. 4. Four-wave mixing (FWM) power values on each of 13 sub-channels at different transmission distances with ASE noise. The fiber channel models are the Transformer based (blue, circles) and the SSFM based (red, triangles).

We select 13 sub-channels of the OFDM signal (a total of 256 sub-channels) and record the FWM noise power generated thereon when transmitting 400 km, 800 km, 1200 km, and 1600 km, respectively. Fig. 4 shows the matching of FWM noise power simulated by the Transformer and the SSFM with ASE noise. The FWM noise power simulated by the Transformer is highly consistent with the SSFM, and the difference is within a small range. The FWM noise power follows the distribution where the center of the OFDM band is higher than the edge. The average relative errors of 13 sub-channels are 0.38%, 0.64%, 0.97%, and 1.24% at 400 km, 800 km, 1200 km, and 1600 km, respectively. Considering the extreme sensitivity of FWM power (cubic correlation), the results suggest that the Transformer simulates nonlinear effects very well.

B. Accuracy and Generalization

High PAPR makes OFDM signals more sensitive to phase noise and nonlinear effects [40] and has long been considered the main drawback of OFDM. The high PAPR characteristic of the OFDM signal stems from the multi-carrier nature caused by overlapping sub-carriers after the IFFT block. When the phases of multiple sub-carriers are similar, the superimposed power will be much larger than the average power, resulting in a relatively high PAPR [41]. The PAPR depending on the transmitted signal can be calculated by finding the ratio between the peak power and the average power [42]. The PAPR of the time-domain OFDM signal can be defined as

$$PAPR(\mathrm{dB}) = 10\log_{10}\frac{\max\left(|u(t)|^2\right)}{\mathrm{E}\left(|u(t)|^2\right)}, \tag{11}$$

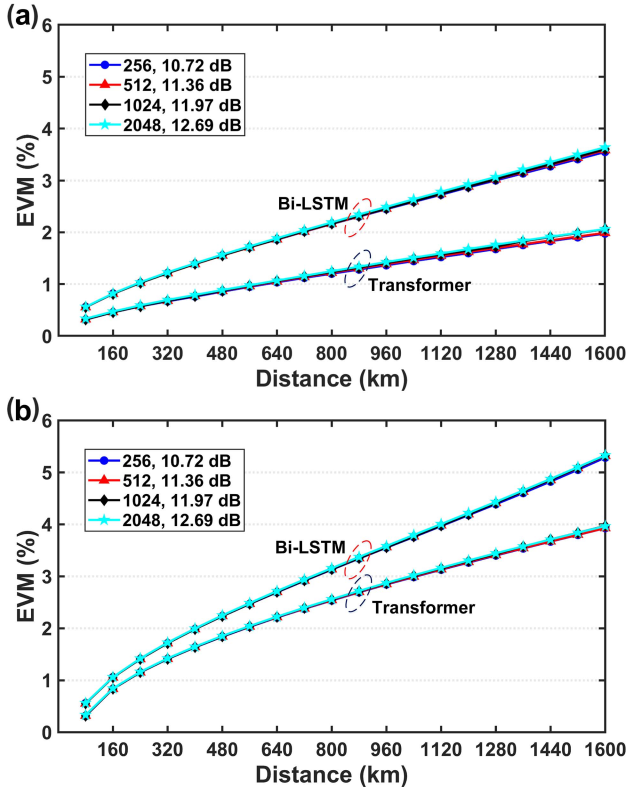

Fig. 5. Error vector magnitudes (EVMs) of optical OFDM signals with various sub-carrier numbers and peak to average power ratios (PAPRs) at different transmission distances. The fiber channel models are the Transformer based (dashed, blue ellipse) and the bidirectional long short-term memory (Bi-LSTM) based (dashed, red ellipse). The sub-carrier numbers and PAPRs are set as 256 and 10.72 dB (blue, circles), 512 and 11.36 dB (red, triangles), 1024 and 11.97 dB (black, diamonds), and 2048 and 12.69 dB (cyan, pentagrams). (a) Without ASE noise and (b) with ASE noise.

where E(•) denotes the average calculation.

An ideal optical OFDM fiber channel model should be highly accurate and suitable for different OFDM signals, including with different subcarrier numbers and different PAPRs. We select OFDM signals with 256, 512, 1024, and 2048 sub-carriers to test the adaptability of the Transformer. Theoretically, the bigger the number of sub-carriers, the higher the probability of high PAPR. Without loss of generality, this correlation of PAPR and subcarrier number is considered when generating OFDM signals.

To evaluate the performance of the Transformer on waveform simulation, EVM [43] as a function of transmission distance is investigated. Here EVM is defined as the relative error between the signal generated by the ML algorithm and the signal simulated by the SSFM, expressed as

$$EVM = \sqrt{\frac{\sum_{i=1}^{N}\left(|X_{generated,i} - X_{real,i}|\right)^2}{\sum_{i=1}^{N}\left(|X_{real,i}|\right)^2}} \times 100\%, \quad (12)$$

where $X_{generated}$ represents the symbols of the signal generated by ML algorithms, and X_{real} represents the symbols of the signal simulated by SSFM.

Fig. 5(a) and (b) show the EVMs of OFDM signals simulated by the Transformer and the Bi-LSTM at different transmission distances, without and with ASE noise. The EVM calculations are for OFDM signals directly from the end of the fiber link without compensation or equalization. Notice that the same ASE noise is added at the same node of the ML algorithm-based model and the SSFM-based model, which will not cause wrong calculation results by random effects.

The fiber channel noise mainly comes from nonlinear noise and ASE noise. From Fig. 5(a), the EVMs are at a low value, below 4%. It shows that the NN model simulates the nonlinear noise of the OFDM signal with high precision on the waveform. Although the NN model is trained on the dataset without ASE noise, it still has high accuracy in the presence of ASE noise. Fig. 5(b) shows that all the EVMs with ASE noise increase by less than 2% compared to those without ASE noise, and the loss of modeling accuracy is within a small range. It indicates that the model has the robustness to ASE noise and accurately simulates the overall fiber channel noise.

The EVM value increases linearly with distance, which is caused by the accumulated error from iteration. The change of subcarrier number or PAPR does not bring significant performance degradation, whether for the Transformer or the Bi-LSTM. The generalization ability of the two models is strong, and they maintain stable performance for different OFDM signals. In terms of accuracy, the Transformer is even better. Whether with or without ASE noise, the OFDM signals simulated by the Transformer have lower EVMs. For the OFDM signals with 2048 sub-carriers at 1600 km, the EVM of the Transformer is 2.06% and 3.97% without and with ASE noise, 3.64% and 5.33% for Bi-LSTM.

We also measure the signal-to-noise ratio (SNR) of the OFDM signal after subcarrier demapping at the receiver. The SNR reflects the communication quality of the optical OFDM system. The SNR is defined as

$$SNR = 10\log_{10}\left(\frac{P_s}{\mathrm{E}|Rx - Tx|^2}\right), \quad (13)$$

where Rx and Tx are the received and transmitted samples, P_s represents the signal power.

Fig. 6 shows the SNRs of OFDM signals transmitted by different fiber channel models for 400 km, 800 km, 1200 km, and 1600 km, with the addition of ASE noise. Notice that DBP compensation is performed here. The SNR difference ranges from 0.60 dB to 1.02 dB between the Transformer and SSFM and 1.43 dB to 2.44 dB between the Bi-LSTM and SSFM. The results show similar communication quality for optical OFDM systems based on fiber channels modeled by the Transformer and the SSFM. In contrast, the SNR difference between the Bi-LSTM and the SSFM is more obvious.

The above results indicate that the Transformer has strong adaptability to different OFDM signals (with different subcarrier numbers and different PAPR values) and outperforms the Bi-LSTM in accuracy.

C. Transmission Rate

Next, the Transformer-based modeling is investigated under different transmission rates. Besides 30 GBd, we add the rates of

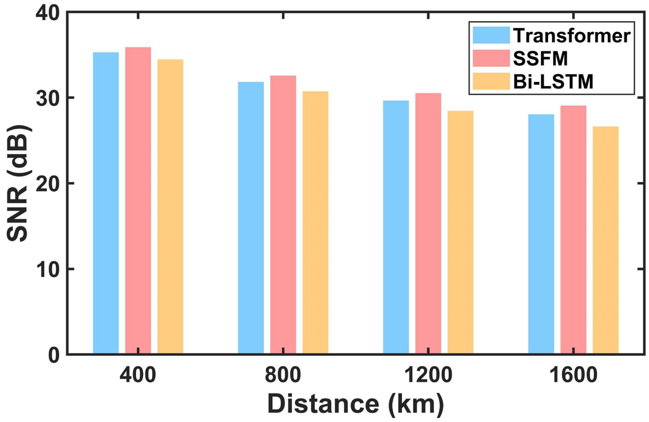

Fig. 6. SNRs of optical OFDM signals transmitted by different fiber channel models for 400 km, 800 km, 1200 km, and 1600 km, with ASE noise. The fiber channel models are the Transformer-based (blue, left), the SSFM-based (red, middle), and the Bi-LSTM-based (orange, right). .

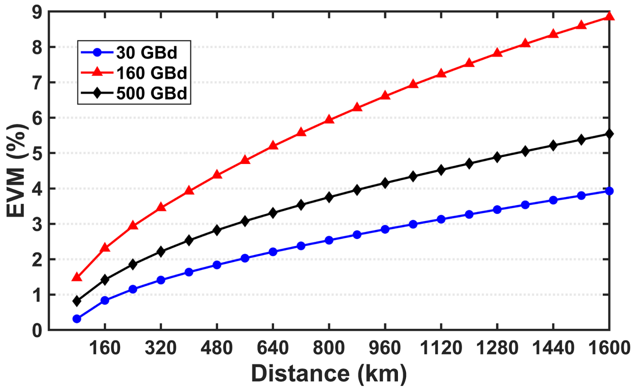

Fig. 7. EVMs at different transmission distances for transmission rates of 30 GBd (blue, circles), 160 GBd (red, triangles), and 500 GBd (black, diamonds), with ASE noise.

160 GBd and 500 GBd. Under the three rates, we build different datasets and train the corresponding Transformer models. With ASE noise, the performance of Transformer models under different rates is shown in Fig. 7, and EVM is still used to measure the accuracy of the Transformer. From the EVMs in Fig. 7, the Transformer model under 30 GBd performs the best, 500 GBd followed and 160 GBd worst.

The modeling performance deteriorates significantly from 30 GBd to 160 GBd, and the EVM increases from 3.93% to 8.84% when transmitting 1600 km. This is caused by the stronger CD effect due to the higher transmission rate. Although linear-nonlinear decoupling is employed in the preprocessing of the training dataset, one-step lumped CD compensation cannot eliminate all linear effects. Some linear effects still exist on the dataset. These legacy linear effects manifest as temporal correlations among adjacent samples, making the learning more difficult for the NN model.

However, from 160 GBd to 500 GBd, there is a noticeable improvement in the modeling accuracy. When transmitting 1600 km, the EVM drops from 8.84% to 5.54%. It seems to be the opposite of the results from 30 GBd to 160 GBd. Actually, this is also related to the CD effect. The existence of CD will cause the phase mismatch of the OFDM signal and reduce the efficiency of the third-order nonlinear interaction [37], thereby reducing the damage caused by FWM to a certain extent. For optical OFDM systems with low transmission rates, the phase mismatch is light due to the closely arranged sub-carriers, and the influence on nonlinear effects is negligible [44]. However, for high rates, CD results in numerous phase mismatches among well-separated sub-carriers, leading to considerable FWM mutual interference effects [45] and significant inhibition of nonlinearity [37]. This simplifies the nonlinear relationship between the input and output datasets and is beneficial for the NN model to learn.

The low EVMs of channel output ranging from 30 GBd and 500 GBd indicate that the Transformer-based fiber channel modeling approach is reliable over a wide range of transmission bandwidths. The modeling accuracy does not decrease monotonically with the increase in transmission rate. Different rate stages have different trends. The modeling accuracy decreases with the increasing rate in the low-rate stage and shows the opposite trend in the high-rate stage. It means that in optical OFDM systems, the ML algorithm-based fiber channel modeling is likely not limited by the high transmission rate.

D. Complexity

To compare the complexity of the Transformer-based model, Bi-LSTM-based model, and SSFM-based model, the two dimensions of multiplication number and running time are considered. For the OFDM signal used as the testing dataset, the parameters are set as follows: the number of subcarriers is 256, the length of one OFDM symbol is 320 after adding the cyclic prefix, the total number of OFDM symbols is 200, and the length of PN is 128. It can be calculated that the data length of the total testing dataset is 513024.

Firstly, the number of multiplications required in the modeling process is theoretically deduced, which generally reflects the hardware complexity of the algorithm [46]. The calculation amount of SSFM mainly derives from FFT and IFFT, and each iterative step contains one FFT and one IFFT [47]. The total multiplication number of SSFM can be expressed as

$$C_{SSFM} = N_{span}\frac{L}{d_z}\left(4N\log_2 N + 2N\right), \tag{14}$$

where N is the number of samples of the transmitted signal, N_{span} is the number of spans, L is the length of each span, and d_z is the step length of the SSFM.

For ML algorithms, ignoring the activation function, the calculation amount mainly comes from matrix multiplication. In the Bi-LSTM model, the LSTM cell is the main component, which consists of cell state and gates [48]. The cell state contains all the useful information. Firstly, the forget gate is used to throw away some old content from the cell state, which is given by

$$f_n = \sigma(W_f \cdot [h_{n-1}, x_n] + b_f), \tag{15}$$

where $h_{n\text{-}1} \in \mathrm{R}^{1\times d\_hidden}$ is the output of the previous LSTM cell, $x_n \in \mathrm{R}^{1\times d\_in}$ is the input of the current LSTM cell, $W_f \in \mathrm{R}^{(d\_hidden+d\_in)\times d\_hidden}$ and b_f are the cell weights and biases of the forget gate, and σ represents the sigmoid function. And *d_in* is the input dimension of each LSTM cell, *d_hidden* is the hidden layer dimension. Next, LSTM decides to add new information to the cell state. The equations of the input gate and new candidate values of the cell state are described by

$$i_n = \sigma(W_i \cdot [h_{n-1}, x_n] + b_i), \tag{16}$$

$$\tilde{C}_n = \tanh(W_c \cdot [h_{n-1}, x_n + b_c]), \tag{17}$$

where $W_i \in \mathrm{R}^{(d\_hidden+d\_in)\times d\_hidden}$ and b_i are the cell weights and biases of the input gate. $W_c \in \mathrm{R}^{(d\_hidden+d\_in)\times d\_hidden}$ and b_c are the weights and biases of the candidate cell state, and the nonlinear activation function is a hyperbolic tangent function. Then the old cell state can be updated by

$$C_n = f_n \times C_{n-1} + i_n \times \tilde{C}_n. \tag{18}$$

With the updated cell state, the output can be obtained by

$$o_n = \sigma(W_o[h_{n-1}, x_n] + b_o), \tag{19}$$

$$h_n = o_n \times \tanh(C_n), \tag{20}$$

where o_n represents the output gate to decide the output parts of the cell state. $W_o \in R^{(d\_hidden+d\_in)\times d\_hidden}$ and b_o are the cell weights and biases of the output gate. $h_n \in R^{1\times d\_hidden}$ is the final output of the current LSTM cell.

For each LSTM cell, (15), (16), (17), and (19) provide the same number of multiplications (*d_hidden*+*d_in*)×*d_hidden*. (18) and (20) provide multiplications of 2*d_hidden* and *d_hidden* respectively. Notice that a Bi-LSTM model contains two LSTM models. The number of multiplications provided by the final Linear layer is 2*d_hidden*×*d_out*, and *d_out* is the final output dimension. The total multiplication number of the Bi-LSTM can be expressed as

$$\begin{aligned} &C_{Bi-LSTM} \\ &= N_{span}\frac{2N}{d_{out}}\left[2N_{layer}d_n\left(4d_{in}d_{hidden} + 4d_{hidden}^2\right.\right. \\ &\quad \left.\left.+3d_{hidden}\right) + 2d_{hidden}d_{out}\right], \end{aligned} \tag{21}$$

where N_{layer} is the number of hidden layers, d_n is the number of cells in one LSTM layer. The parameter settings of the Bi-LSTM are shown in Table III.

For the Transformer model, the number of multiplications provided by the two Linear layers is *d_in*×*d_model* and *d_model*×*d_out*, respectively. The multi-head layer provides 2*d_model*^{2} multiplications, and the FFN layer provides 2*d_model*×*dff* multiplications. The total multiplication number of the Transformer can be expressed as

$$\begin{aligned} &C_{Transformer} \\ &= N_{span}\frac{2N}{d_{out}}\left[d_{in}d_{model} + N_{layer}\left(2d_{model}^2\right.\right. \\ &\quad \left.\left.+2d_{model}d_{ff}\right) + d_{model}d_{out}\right], \end{aligned} \tag{22}$$

where N_{layer} is the number of the Transformer layers. The parameter settings of the Transformer model are shown in Table II.

The relationship between the number of multiplications and the transmission distance in the modeling process is shown in Fig. 8. The number of multiplications increases linearly with the transmission distance. When transmitting 1600 km, the multiplication number is about 3.032·10^{12} for SSFM, 1.576·10^{12} for Bi-LSTM, and 8.701·10^{11} for Transformer. The multiplication number of the Transformer is about 55% of Bi-LSTM and 30% of SSFM. With a small number of multiplications, the Transformer has obvious advantages in hardware complexity.

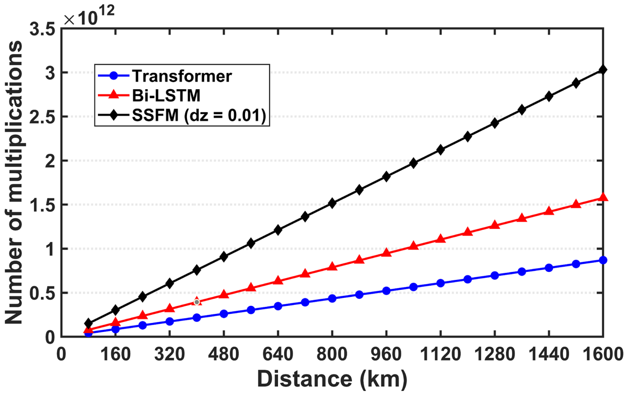

Fig. 8. Number of multiplications vs. distance for different fiber channel models. Transformer based (blue, circles), Bi-LSTM based (red, triangles) and SSFM based (black, diamonds).

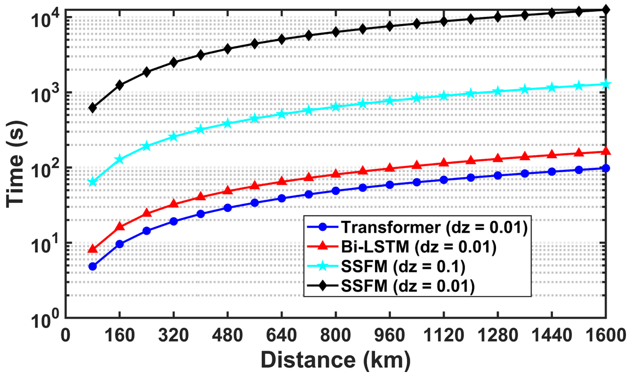

Fig. 9. Running time vs. distance for different fiber channel models. Transformer (dz = 0.01 km) based ((blue, circles), Bi-LSTM (dz = 0.01 km) based (red, triangles), SSFM (dz = 0.1 km) based (cyan, pentagrams) and SSFM (dz = 0.01 km) based (black, diamonds).

The complexity of the three models is also compared from the perspective of running time. The running times on the central processing unit (CPU) are recorded under the same simulation conditions (including hardware conditions and software conditions) when transmitting different distances, as shown in Fig. 9. To avoid contingency, the recorded data is the result averaged from multiple simulations, which has certain statistical significance. Moreover, the setting of the OFDM signal is unchanged.

The ML algorithms can implement parallel computing in the modeling process, so it is far superior to the SSFM in terms of time complexity. As shown in Fig. 9, the running time increases linearly with the transmission distance. When transmitting 1600 km, the required running time is 12502 s for the SSFM (d_z is 0.01 km), 163 s for the Bi-LSTM, and only 98 s for the Transformer, which is about 60% of Bi-LSTM and 0.8% of SSFM. Even if the step length d_z of the SSFM is increased from 0.01 km to 0.1 km, the running time for transmitting 1600 km is 1284 s, which is about 13 times the Transformer. The above results show the superiority of the Transformer-based modeling method in complexity.

IV. Conclusion

This study firstly proposes a Transformer-based fiber channel modeling method for long-haul optical OFDM transmission and achieves high accuracy and low time-consuming simulation. We simplify the Transformer architecture to reduce complexity, i.e., discard the Decoder part and cancel the self-attention mechanism. By combining the FDD scheme, adopting the multi-dataset training method, and redesigning the loss function, the accuracy and generalization ability of the model are improved. Based on the measurements of FWM power, EVM and SNR, we have demonstrated that the Transformer-based fiber channel model has an excellent performance in terms of nonlinear fitting, accuracy, and generalization power. In the transmission of 1600 km, the average relative error of FWM power is less than 1.3%, the EVM is below 4%, and the SNR is within 1 dB. Results show that the model accurately simulates fiber channel effects and is highly adaptable to different OFDM signals. In addition, the modeling is investigated at transmission rates of 30 GBd, 160 GBd and 500 GBd, and proved reliable in wide transmission bandwidth. The multiplication number of the Transformer is about 30% of SSFM, and the running time is less than 1% of SSFM, making a significant breakthrough in complexity while ensuring high accuracy.

The proposed Transformer-based scheme achieves accurate and fast modeling of the fiber channel in optical OFDM systems. This approach can be applied to the simulation of optical OFDM signal transmission and is also beneficial to end-to-end optimization and system design. In addition, it is expected to be extended to other optical systems for modeling work.

References

[1] G. P. Agrawal, *Nonlinear Fiber Optics*. 4th ed. San Diego, CA, USA: Academic, 2007.

[2] D. Wang, Y. Song, J. Li, J. Qin, and A. C. Boucouvalas, “Data-driven optical fiber channel modeling: A deep learning approach,” *J. Lightw. Technol.*, vol. 38, no. 17, pp. 4730–4743, Sep. 2020.

[3] H. Yang et al., “Fast and accurate optical fiber channel modeling using generative adversarial network,” *J. Lightw. Technol.*, vol. 39, no. 5, pp. 1322–1333, Mar. 2021.

[4] M. Arjovsky and L. Bottou, “Towards principled methods for training generative adversarial networks,” in *Proc. Int. Conf. Learn. Representations*, 2017.

[5] H. Yang, Z. Niu, H. Zhao, S. Xiao, W. Hu, and L. Yi, “Fast and accurate waveform modeling of long-haul multi-channel optical fiber transmission using a hybrid model-data driven scheme,” *J. Lightw. Technol.*, vol. 40, no. 14, pp. 4571–4580, Jul. 2022.

[6] A. Lowery and J. Armstrong, “Orthogonal-frequency-division multiplexing for dispersion compensation of long-haul optical systems,” *Opt. Exp.*, vol. 14, pp. 2079–2084, 2006.

[7] K. Inoue, “Phase-mismatching characteristic of four-wave mixing in fiber lines with multistage optical amplifiers,” *Opt. Lett.*, vol. 17, no. 11, pp. 801–803, 1992.

[8] B. Goebel, B. Fesl, L. D. Coelho, and N. Hanik, “On the effect of FWM in coherent optical OFDM systems,” in *Proc. IEEE Opt. Fiber Commun./Nat. Fiber Optic Engineers Conf.*, 2008, pp. 1–3.

[9] A. J. Lowery, S. Wang, and M. Premaratne, “Calculation of power limit due to fiber nonlinearity in optical OFDM systems,” *Opt. Exp.*, vol. 15, no. 20, pp. 13282–13287, 2007.

[10] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[11] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[12] J. Fu et al., “Dual attention network for scene segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3146–3154.

[13] Q. Chen, H. Zhao, W. Li, P. Huang, and W. Ou, “Behavior sequence transformer for e-commerce recommendation in Alibaba,” *Assoc. Comput. Machinery*, vol. 12, pp. 1–4, 2019.

[14] K. Huang, C. Xiao, L. M. Glass, and J. Sun, “MolTrans: Molecular interaction transformer for drug-target interaction prediction,” *Bioinformatics*, vol. 37, no. 6, pp. 830–836, 2021.

[15] H. Wu, J. Xu, J. Wang, and M. Long, “Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 22419–22430, *arXiv:2106.13008*.

[16] K. Lin, L. Wang, and Z. Liu, “Mesh graphormer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12939–12948.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[18] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” 2016, *arXiv:1607.06450*.

[19] W. Yu et al., “MetaFormer is actually what you need for vision,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10819–10829.

[20] T. Brown et al., “Language models are few-shot learners,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.

[21] J. Devlin, M. - W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, vol. 1, pp. 4171–4186.

[22] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” 2020, *arXiv:2005.08100*.

[23] Y. Liu et al., “Roberta: A robustly optimized bert pretraining approach,” 2019, *arXiv:1907.11692*.

[24] C. Ying et al., “Do transformers really perform badly for graph representation?,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 28877–28888.

[25] S. S. Haykin, *Digital Communications*. Hoboken, NJ, USA: Wiley, 1988.

[26] D. Rafique, M. Mussolin, M. Forzati, J. Mrtensson, and A. D. Ellis, “Compensation of intra-channel nonlinear fibre impairments using simplified digital back-propagation algorithm,” *Opt. Exp.*, vol. 19, no. 10, pp. 9453–9460, 2011.

[27] C. R. Giles and E. Desurvire, “Modeling erbium-doped fiber amplifiers,” *J. Lightw. Technol.*, vol. 9, no. 2, pp. 271–283, Feb. 1991.

[28] T. Schneider, *Nonlinear Optics in Telecommunications*. Berlin/Heidelberg, Germany: Springer, 2004.

[29] S. J. Savory, “Digital filters for coherent optical receivers,” *Opt. Exp.*, vol. 16, no. 2, pp. 804–817, 2008.

[30] Y. Dong, J. B. Cordonnier, and A. Loukas, “Attention is not all you need: Pure attention loses rank doubly exponentially with depth,” in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 2793–2803.

[31] D. Balduzzi, M. Frean, L. Leary, J. P. Lewis, K. W.-D. Ma, and B. McWilliams, “The shattered gradients problem: If resnets are the answer, then what is the question?,” in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 342–350.

[32] T. Domhan, “How much attention do you need? A granular analysis of neural machine translation architectures,” in *Proc. 56th Annu. Meeting Assoc. for Comput. Linguistics.*, 2018, vol. 1, pp. 1799–1808.

[33] M. H. Guo, Z. N. Liu, T. J. Mu, and S. M. Hu, “Beyond Self-attention: External attention using two linear layers for visual tasks,” 2021, *arXiv:2105.02358*.

[34] H. Lin, D. Rolnick, and M. Tegmark, “Why does deep and cheap learning work so well?,” *J. Stat. Phys.*, vol. 168, no. 6, pp. 1–25, 2016.

[35] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *Comput. Sci.*, 2014.

[36] O. Köksoy, “Multiresponse robust design: Mean square error (MSE) criterion,” *Appl. Math. Comp.*, vol. 175, no. 2, pp. 1716–1729, 2006.

[37] M. Nazarathy, J. Khurgin, R. Weidenfeld, Y. Meiman, and V. Karagodsky, “Phased-array cancellation of nonlinear FWM in coherent OFDM dispersive multi-span links,” *Opt. Exp.*, vol. 16, no. 20, pp. 15777–15810, 2008.

[38] A. J. Lowery, S. Wang, and M. Premaratne, "Calculation of power limit due to fiber nonlinearity in optical OFDM systems," *Opt. Exp.*, vol. 15, no. 20, pp. 13282–13287, 2007.

[39] V. Pechenkin and I. J. Fair, "On four-wave mixing suppression in dispersion-managed fiber-optic OFDM systems with an optical phase conjugation module," *J. Lightw. Technol.*, vol. 29, no. 11, pp. 1678–1691, Jun. 2011.

[40] A. Ghassemi and T. A. Gulliver, *IEEE Trans. Signal Process.*, vol. 56, 2008, Art. no. 1161.

[41] S. Lin, Y. Chen, and S. Tseng, "Iterative smoothing filtering schemes by using clipping noise-assisted signals for PAPR reduction in OFDM-based carrier aggregation systems," *Commun. IET*, vol. 13, no. 6, pp. 802–808, 2019.

[42] B. K. Kharagpur, U. Wali, and S. Bidwai, "Novel technique to reduce PAPR in OFDM systems by clipping and filtering," in *Proc. IEEE Int. Conf. Adv. Comput., Commun. Inform.*, 2013, pp. 1593–1597.

[43] C. Zhao and R. Baxley, "Error vector magnitude analysis for OFDM systems," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, 2006, pp. 1830–1834.

[44] W. Shieh, X. Yi, and Y. Tang, "Transmission experiment of multi-gigabit coherent optical OFDM systems over 1000 km SSMF fiber," *Electron. Lett.*, vol. 43, pp. 183–185, 2007.

[45] K. Inoue, "Phase mismatching characteristic of four-wave mixing in fiber lines with multistage optical amplifiers," *Opt. Lett.*, vol. 17, 1992, Art. no. 801.

[46] A. Napoli, Z. Maalej, V. A. J. M. Sleiffer, M. Kuschnerov, and D. Rafique, "Reduced complexity digital back-propagation methods for optical communication systems," *J. Lightw. Technol.*, vol. 32, no. 7, pp. 1351–1362, Apr. 2014.

[47] B. Spinnler, "Equalizer design and complexity for digital coherent receivers," *IEEE J. Sel. Top. Quantum Electron.*, vol. 16, no. 5, pp. 1180–1192, Sep./Oct. 2010.

[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.