文章编号：1001-9014(XXXX)XX-0001-07

DOI:10.11972/j.issn.1001-9014.XXXX.XX.001

Millimeter-Wave Modeling based on Transformer model for InP High Electron Mobility Transistor

ZHANG Ya-Xue[1,2], ZHANG Ao[1,3*], GAO Jian-Jun[4]

(1. School of Microelectronics, Nantong University, Nantong 226019, China;
2. State Key Laboratory of Millimeter-waves, Southeast University, Nanjing 210096, China;
3. School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore;
4. School of Physics and Electronic Science, East China Normal University, Shanghai 200241, China)

Abstract: In this paper, the small-signal modeling of the Indium Phosphide High Electron Mobility Transistor (InP HEMT) based on the Transformer neural network model is investigated. The AC S-parameters of the HEMT device are trained and validated using the Transformer model. In the proposed model, the eight layers transformer encoders are connected in series and the encoder layer of each Transformer consists of the multi-head attention layer and the feed-forward neural network layer. The experimental results show that the measured and modeled S-parameters of the HEMT device match well in the frequency range of 0.5-40 GHz, with the errors versus frequency less than 1%. Compared with other models, good accuracy can be achieved to verify the effectiveness of the proposed model.

Key words: Transformer model, Neural Network, high electron mobility transistor (HEMT), small signal model
PACS:

基于Transformer模型的磷化铟高电子迁移率晶体管毫米波建模

张雅雪[1,2], 张 傲[1,3*], 高建军[4]

(1. 南通大学 微电子学院,江苏南通 226019;
2. 东南大学 毫米波国家重点实验室,江苏南京 210096;
3. 新加坡南洋理工大学 电气与电子工程学院,新加坡 639798;
4. 华东师范大学 物理与电子科学学院,上海 200241)

摘要:本文对基于Transformer神经网络模型的磷化铟高电子迁移率晶体管(InP HEMT)小信号建模进行了研究,利用Transformer模型对HEMT器件的交流S参数进行训练和验证。在所提出的模型中,八层Transformer编码器串联,每个Transformer的编码器层由多头注意层和前馈神经网络层组成。实验结果表明,在0.5-40 GHz频率范围内,HEMT器件测量和建模的S参数匹配良好,频率误差小于1%。与其他模型相比,可以达到良好的精度,验证了所提模型的有效性。

关 键 词:Transformer模型;神经网络;高电子迁移率晶体管(HEMT);小信号模型

中图分类号:O43 **文献标识码**:A

Introduction

In recent years, with the rapid development of high-speed communication and RF microwave technologies, high electron mobility transistor (HEMT) devices are increasingly used in microwave and millimeter-wave circuits[1-5]. Among them, indium phosphide (InP) HEMT devices have become an ideal choice for next-generation high-speed and high-frequency electronic devices due to their excellent electron mobility and frequency response characteristics[6-7]. In order to fully utilize the performance of InP HEMT devices, accurate device modeling

Received date: 2024-10-09, **收稿日期**:2024-10-09,
Foundation items: Supported in part by the National Natural Science Foundation of China under Grant 62201293 and Grant 62034003, and in part by the Open-Foundation of State Key Laboratory of Millimeter-Waves under Grant K202313.
Biography: ZHANG Yaxue (1999), female, Qingdao, master. Research area involves microwave device modeling. E-mail: zhangyaxue1207@163.com.
* **Corresponding author**: E-mail: aozhang@ntu.edu.cn

is particularly important. However, traditional small-signal models have limitations in simulating InP HEMT devices, making it difficult to accurately describe their nonlinear characteristics and frequency dependence[8-9].

With the rapid development of machine learning technology, neural networks have been widely used in many fields, such as microwave device modeling, signal processing and RF design[10-11]. In the field of microwave device modeling, neural networks can replace the manual completion of a large number of cumbersome steps, greatly improving the efficiency of scientific research[12-13].

The related techniques have been discussed as follows. The Wiener-type dynamic neural network (DNN) approach for HEMT device modeling was presented in [14]. The analytical formulation of Winer-type DNN structure consists of a cascade of a simplified linear dynamic part. In [15], the approach using decomposed mapping for HEMTs, advancing the Space Mapping technique of neural network for device modeling was discussed. The convolutional neural network (CNN) for HEMT device modeling with various gate and source field plate designs and drain voltages was mentioned in [16]. Moreover, literature [17] presented a modified recurrent neural network (RNN) technique, long-short term memory (LSTM) algorithm-based, small-signal behavioral modeling methodology for HEMTs. Meanwhile, small-signal model based on gated recurrent unit (GRU) neural networks was investigated in [18].

Based on these researches, the small-signal model based on Transformer neural network with multiple transformers for InP HEMT is presented in this paper. In comparison to the previous literature, several novel aspects are shown as follows.

1) The AC S-parameters have been trained and validated by using the transformer neural network with 8-layers transformer encoders in series.

2) The encoder layer of each Transformer consists of two sub-layers: the Multi-head attention layer and the feed forward neural network layer. Residual connection and layer normalization are added after each sub-layer.

3) Higher accuracy compared to other models. The simulated S-parameters perform well on the fitting of the measured S-parameters under normal bias conditions. The errors versus frequency is less than 1%.

The organization of the paper is as follows. Section II gives the details of the proposed neural network with multiple transformers which utilized in the small signal modeling of InP HEMTs. Section III presents the discussion and the analysis of the results. In the end, a conclusion is provided in Section IV.

1 Transformer neural network model

The structure of the proposed neural network model is given in Fig. 1, which uses and improves the encoder part of the conventional transformer model. As seen in Fig. 1, the proposed model can be divided into three parts: the input layer, the hidden layers and the output layer. The input layer is an $n\times3$ second-order matrix, where n represents the number of samples and 3 represents the sample features, which are frequency *freq*, gate-source voltage V_{gs} and drain-source voltage V_{ds}. The hidden layers consist of the transformer encoder layers and the linear layers.

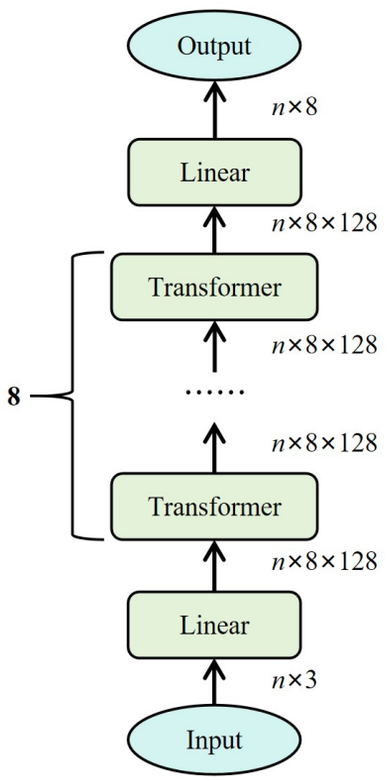


Fig. 1 Model structure
图1 模型结构

To begin with, the data are preprocessed through the first linear layer and the input data are transformed into a third-order matrix of $n\times8\times128$, where 8 denotes the number of samples that are input to the model at one time for training or inference, generally referred to as the batch size. This is immediately followed by eight identical Transformer encoder layers in series, which dimensionally have identical inputs and outputs, all of which are identical to the output of the first linear layer. Finally, through the second linear layer, the data are transformed into an $n\times8$ second-order matrix to be passed to the output layer, at this time, 8 represents the sample characteristics of the output data, which are the magnitude (Mag) and phase (ϕ) of S_{11}, S_{21}, S_{12}, S_{22}, respectively.

The expression for the output matrix S of the model is:

$$\mathbf{S} = \mathrm{f}_{\mathrm{Transformer}}(I), \tag{1}$$

where $\boldsymbol{S}$ is an $n\times8$ order matrix denoted:

$$\mathbf{S} = \begin{bmatrix} \mathrm{Mag}^{11}(S_{11}) & \phi^{12}(S_{11}) & \cdots & \mathrm{Mag}^{17}(S_{22}) & \phi^{18}(S_{22}) \\ \mathrm{Mag}^{21}(S_{11}) & \phi^{22}(S_{11}) & \cdots & \mathrm{Mag}^{27}(S_{22}) & \phi^{28}(S_{22}) \\ \vdots & \ddots & & \vdots & \\ \mathrm{Mag}^{n1}(S_{11}) & \phi^{n2}(S_{11}) & \cdots & \mathrm{Mag}^{n7}(S_{22}) & \phi^{n8}(S_{22}) \end{bmatrix}_{n\times 8}, \quad (2)$$

$\mathbf{I}$ is an n×3 order matrix, denoted

$$\mathbf{I} = \begin{bmatrix} freq^{11} & V_{gs}{}^{12} & V_{ds}{}^{13} \\ freq^{21} & V_{gs}{}^{22} & V_{ds}{}^{23} \\ \vdots & \vdots & \vdots \\ freq^{n1} & V_{gs}{}^{n2} & V_{ds}{}^{n3} \end{bmatrix}_{n\times 3}, \quad (3)$$

The 8-layers transformer encoder in the model is connected in series. By extracting and processing input features layer by layer, the model can capture more complex and abstract feature representations. Increasing the number of layers can improve the expressiveness of the model, capture long-distance dependencies, gradually fuse information, and improve generalization capabilities. Each transformer encoder layer contains two sub-layers, as shown in Fig. 2, namely the Multi-head Attention layer and the Feed Forward Neural Network layer. Residual Connection and Layer Normalization are added after each sub-layer.

The self-attention mechanism is the core of Transformer, which allows each input vector to pay attention to the input vectors at all other positions in the sequence when calculating its own representation. This can help the model capture the global dependencies between input features. The self-attention mechanism is to compute the attention weight $\mathrm{Attention}(Q, K, V)$ by calculating the query matrix, key matrix and value matrix of the input matrix X. The formula is as follows:

$$\begin{aligned} Q &= XW_Q \\ K &= XW_K \\ V &= XW_V \end{aligned}, \quad (4)$$

where W_Q , W_K , W_V is the weight matrix. We have,

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{dk}}\right)V \quad (5)$$

where $\sqrt{dk}$ is the scaling factor.

In multi-head attention, the input is computed in parallel by h independent heads, each with its own query, key, and value matrices. The output of each head is as follows:

$$\mathrm{head}_i = \mathrm{Attention}(QW_Q^i, KW_K^i, VW_V^i) \quad (6)$$

The outputs of all heads are concatenated and subsequently passed through a linear transformation:

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{h} \quad (7)$$

where W_O is the output weight matrix.

Then, through the feed-forward neural network, the proposed model can capture the long-range dependencies in the input sequences, further extract and fuse the feature information at different positions, and enhance the nonlinear and expressive capabilities of the model. The formula is as follows:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (8)$$

where x is the input variable. W_1, W_2 are the weight matrices of the linear transformation, and b_1 , b_2 are the bias vectors.

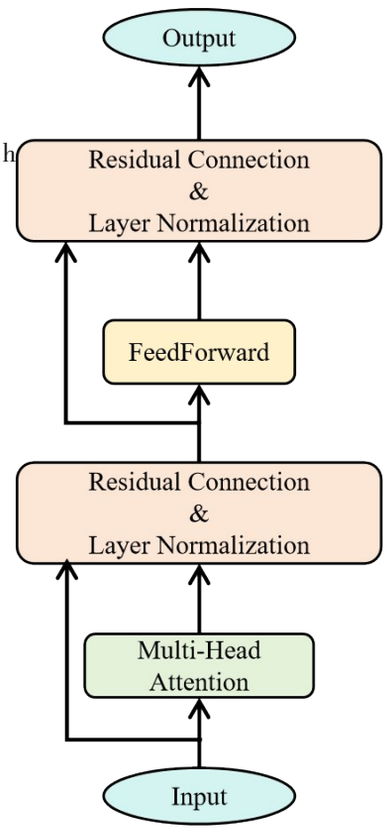


Fig. 2 Structure of the Transformer encoder layer
图2 Transformer编码器层结构

In the proposed model, residual connections and layer normalization are two key components.

The residual connections can alleviate the gradient vanishing problem in neural networks and promote the flow of information. By giving residual connections, the input can bypass one or more layers and be directly added to the output, making the network easier to train. The formula for residual connections is:

$$y = F(x) + x \quad (9)$$

where the input is x and the output after some sublayer (e. g. , a multi-head self-attention layer or a feedforward neural network) is $F(x)$.

Layer normalization can normalize the features of each sample to speed up the training process and improve the stability of the model. Layer normalization is performed independently on each sample. Assuming the input is h, the layer normalization operation is as follows:

Calculate the mean and variance:

$$\begin{aligned} \mu &= \frac{1}{\mathrm{d}}\sum_{i=1}^{d} h_i \\ \sigma^2 &= \frac{1}{\mathrm{d}}\sum_{i=1}^{d}(h_i - \mu)^2 \end{aligned} \quad (10)$$

Standardization:

$$\hat{h}_i = \frac{h_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} \quad (11)$$

Linear transformations:

$$y_i = \gamma \hat{h}_i + \beta \tag{12}$$

where γ and β are trainable parameters and ε is a small constant to prevent division by zero.

The dataset contains 7700 data points. K-Fold cross validation is used as the model evaluation method to provide a more robust assessment of the model. The dataset is divided into 5 subsets, each with 1540 data points. One of the subsets is used as the validation set, and the other subsets are used for training. Finally, the average performance of the model is calculated through multiple rounds of validation.

In order to evaluate the training effect of the model, the Mean Squared Error (MSE) loss is chosen as the loss function of the model. We have:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{13}$$

Where $\hat{y}_i$ is the predicted value of the model and y_i is the true value of the model and n is the number of samples.

The model is trained using the Adam optimizer with a learning rate of 0.001. A batch size of 8 is used, and the training is run for 100 epochs. Dropout is set to 0.1 in each layer to prevent overfitting. The training is conducted on an AMD Radeon (TM) Graphics GPU with 16 GB of memory. The total training time for the 8-layer model is approximately 2.1 hours.

2 Results and discussions

In order to verify the proposed ANN model described in Section 2, the InP-based HEMT devices fabricated using in-house process were characterized. The device with gate width of 2×25μm were investigated in the frequency range of 0.5-40 GHz. The test layout of InP HEMT devices is shown in Fig. 3. The verification was made up to 40 GHz by using Agilent E8363C vector network analyzer, with DC bias supplied by Agilent B1500. All measurements were carried out on wafer using Cascade Microtech's Air-Coplanar Probes ACP50-GSG-100. The measurement setup is illustrated in Fig. 4.

The comparison between modeled and measured S-parameters for InP HEMT devices in the frequency range of 0.5-40 GHz for the bias points at V_{gs}=0 V, V_{ds}=1.0 V, V_{gs}=0 V, V_{ds}=1.2 V and V_{gs}=-0.05 V, V_{ds}=1.2 V are plotted in Fig. 5.

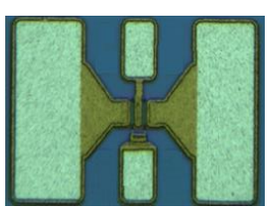


Fig. 3 Test layout of InP HEMT devices.
图3 InP HEMT器件的测试版图

The formula of the absolute error is

$$Error = \left| S_{ij}^{\mathrm{meas}} - S_{ij}^{\mathrm{model}} \right| \times 100\% \tag{14}$$

where S_{ij}^{model} represents the modeled S-parameters, and S_{ij}^{meas} denotes the measured S-parameters. Note that all the S parameters vary with frequency.

Under the condition of bias gate-source voltage V_{gs}= 0 V and drain-source voltage V_{ds}=1.0 V, the absolute error curves of the amplitudes of S_{11}, S_{12}, S_{21}, and S_{22} in the frequency range of 0.5-40 GHz with respect to frequency are shown in Fig. 6. It can be seen that the errors versus frequency is within 1%, which proves the accuracy of the model.

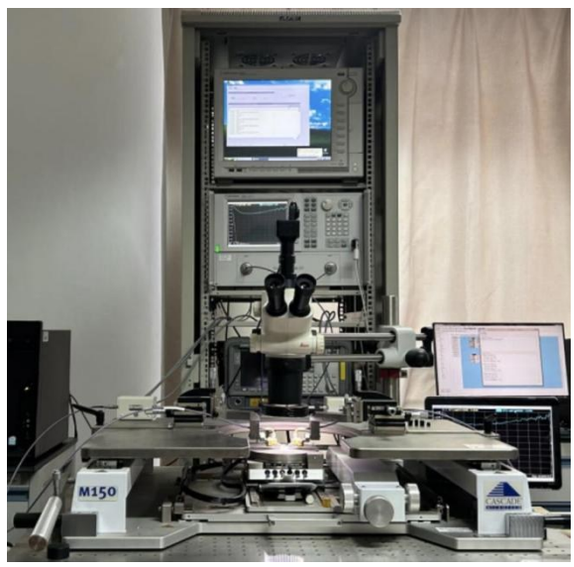
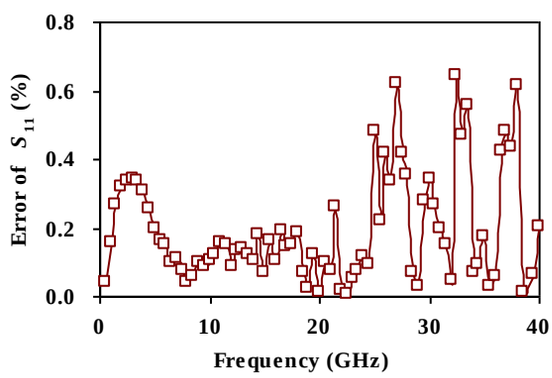
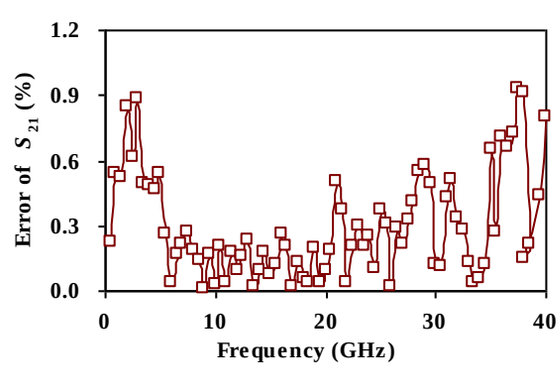


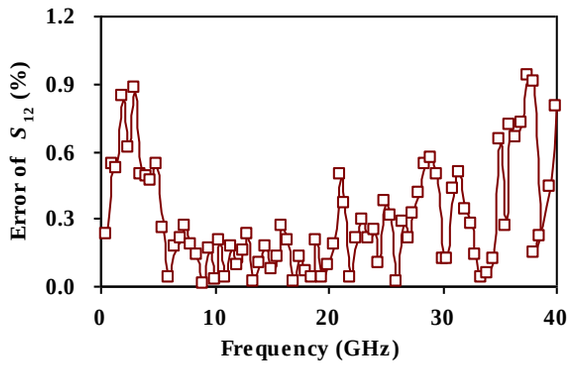
Fig. 4 Measurement setup
图4 测试设备



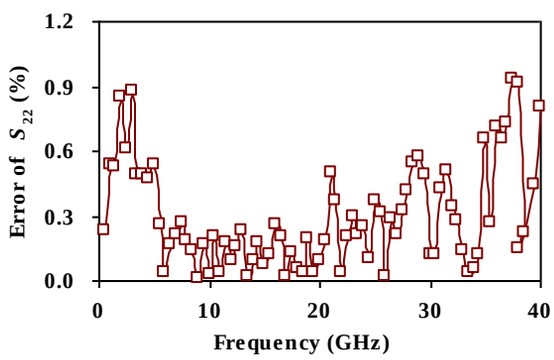
(a) S_{11}



(b) S_{21}



(c) S_{12}



(d) S_{22}

Fig.6 The absolute error of S-parameters versus frequency
图6 S参数绝对误差

Fig. 7 illustrates the training and validation loss curves. The model converged after approximately 80 epochs, with no significant overfitting observed.

For a global evaluation of model accuracy, the Mean Squared Error (MSE) is also calculated and provided in Table 1. In order to further demonstrate the accuracy of the Transformer model, the proposed model based on the transformer is compared with other models, including Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM), and Gate Recurrent Unit (GRU). As can be seen from Table 1, the MSE of proposed model is better than the other models.

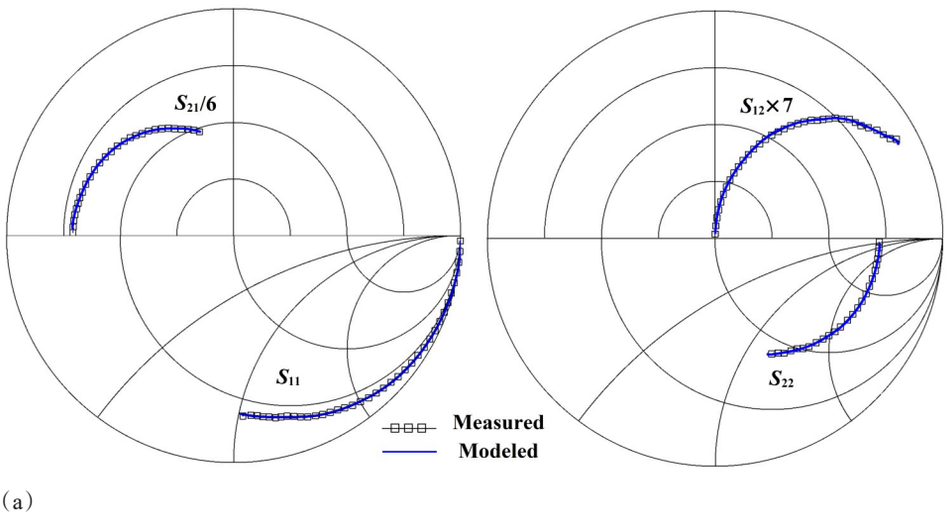
Among them, the input layer of the Transformer model is projected to obtain a 128 dimensional embedding, using 16 attention heads, and each feedforward network has 128 neurons. The convolutional layers of the CNN model use 64, 128, and 256 filters respectively.

Table 1 Comparison of MSE accuracy
表1 MSE精度比较

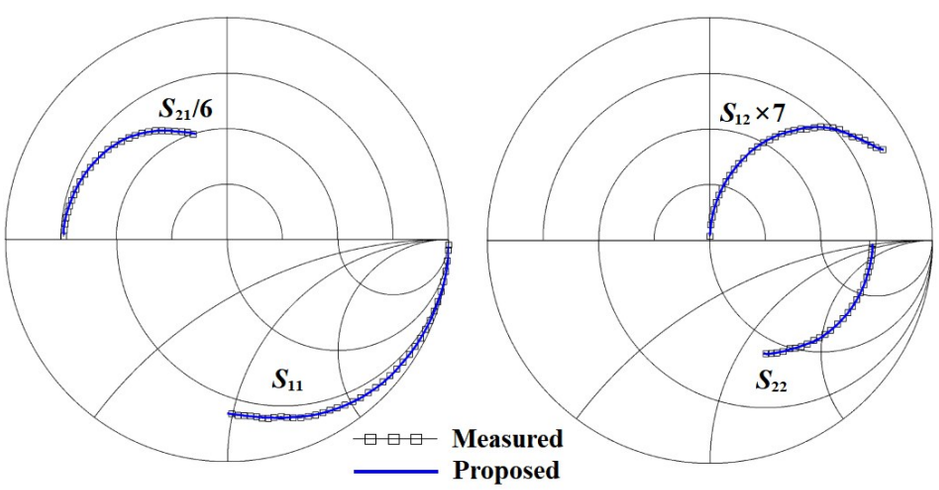
Method	MSE	Number of Neurons	Number of Layers	Training Methods and Parameters
Transformer	0.0012	128-16-128	8 layers	K-Fold cross validation, k=5
CNN[16]	0.0037	64,128,256 per layer	4 convolutional layers	K-Fold cross validation, k=5
LSTM[17]	0.0015	128 per layer	3 layers	K-Fold cross validation, k=5
GRU[18]	0.0017	128 per layer	3 layers	K-Fold cross validation, k=5

3 Conclusions

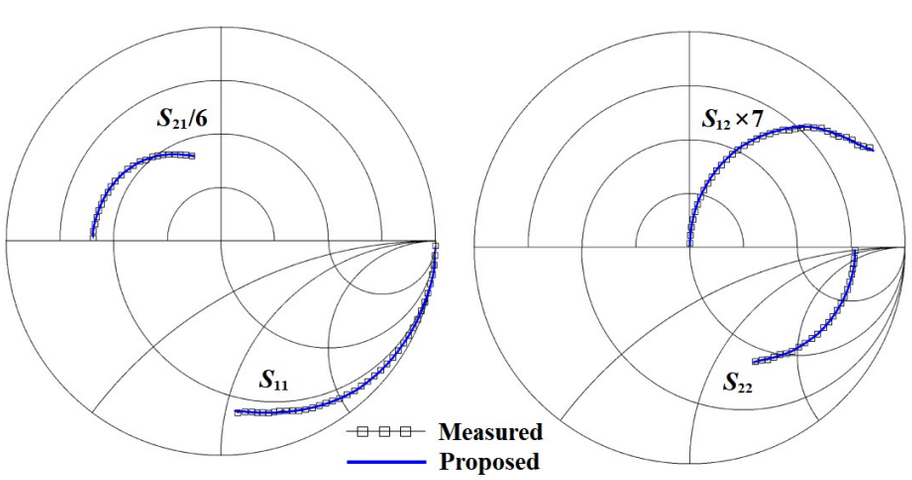
The small-signal modeling of the InP HEMT based on the Transformer model is presented in this paper. For the proposed model, the eight layers transformer model connected in series with multi-head attention layer and the feed-forward neural network layer are utilized to train and validate the S-parameters of the HEMT. Good agreement can be achieved between the simulated and modeled data in the frequency range of 0.5-40 GHz. Compared with other models, higher accuracy can be obtained, with the errors versus frequency within 1%.



（a）



（b）



（c）

Fig.5 Comparison of modeled and measured S-parameters for InP HEMT in 0.5-40 GHz frequency range. Bias: (a) V_{gs}=0 V, V_{ds}=1.0 V (b) V_{gs}=0 V, V_{ds}=1.2 V (c) V_{gs}=-0.05 V, V_{ds}=1.2 V

图5 0.5-40 GHz 频率范围内 InP HEMT 的模拟和测试S参数比较。偏置：(a) Vgs=0 V, Vds=1.0 V (b) Vgs=0 V, Vds=1.2 V (c) Vgs=-0.05 V, Vds=1.2 V

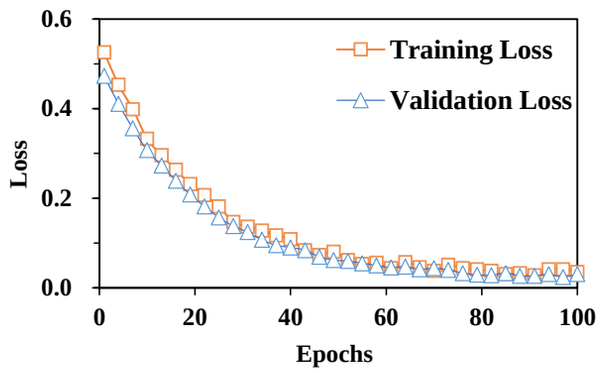


Fig. 7 learning curve
图7 学习曲线

References

[1] LI Ze-Kun, CHEN Ji-Xin, ZHENG Si-Dou and HONG Wei. A 66-112.5 GHz low noise amplifier with minimum NF of 3.9 dB in 0.1-μm GaAs pHEMT technology [J]. Journal of Infrared and Millimeter Waves, 2024, 43(2): 186~190
[2] Liu X, Meng F, Chen Y, Zhang A and Gao J. Design of 230~250 GHz low noise amplifier based on 70 nm InP HEMT process [J]. Journal of Infrared and Millimeter Waves, 2023, 42(1): 37~42.
[3] Chen Y, Meng F, Fang Yuan, Zhang Ao and Gao J. Design of 220GHz power amplifier based on 90nm InP HEMT process [J]. Journal of Infrared and Millimeter Waves, 2022, 41(6): 1037~1041.
[4] Gao J, *RF and Microwave Modeling and Measurement Techniques for Field Effect Transistors* [M]. Raleigh, NC: SciTech Publishing, Inc., 2010.
[5] Leuther A, Merkle T, Weber R, Sommer R and Tessmann A. THz Frequency HEMTs: Future Trends and Applications [C]. 2019 Compound Semiconductor Week (CSW), Nara, Japan, 2019: 1~2
[6] Ruiz D, Saranovac T, Han D, Hambitzer A, Arabhavi A, Ostinelli O and Bolognesi C. InAs channel inset effects on the DC, RF, and noise properties of InP pHEMTs [J]. IEEE Transactions on Electron Devices, 2019, 66(11): 4685~4691.
[7] Jo H, Baek J, Yun D. Lg = 87 nm InAlAs/InGaAs high-electron mobility transistors with a gm_max of 3 S/mm and fT of 559 GHz [J]. IEEE Electron Device Letter, 2018, 39(11):1640~1643.
[8] Schleeh J, Rodilla H, Wadefalk N, Nilsson P and Grahn J. Characterization and Modeling of Cryogenic Ultra low-Noise InP HEMTs [J]. IEEE Transactions on Electron Devices, 2013, 60 (1): 206~212
[9] Liu J, Yu W, Yang S, Hou Y, Cui D and Lyu X. Small signal model and low noise application of InAlAs/InGaAs/InP-based PHEMTs [J]. Journal of Infrared and Millimeter Waves, 2018, 37 (6): 683~687.
[10] Zhang Q and Gupta K, *Neural Networks for RF and Microwave Design* [M]. Norwood, MA, USA: Artech House, 2000.
[11] Zhang Q, Gupta K, and Devabhaktuni V. Artificial neural networks for RF and microwave design-from theory to practice [J]. IEEE Trans. Microw. Theory Techn., 2003, 51(4):1339~1350.
[12] Jin Jet al. Recent advances in neural network-based inverse modeling techniques for microwave applications [J]. Int. J. Numer. Model., Electron. Netw., Devices Fields, 2020, 33(6): 1~18.
[13] Jarndal A. Neural network electrothermal modeling approach for microwave active devices[J]. Int. J. RF Microw. Comput.-Aided Eng., 2019, 29(9): 1~9.
[14] Liu W, Na W, Zhu L, Ma J and Zhang Q. A Wiener-Type Dynamic Neural Network Approach to the Modeling of Nonlinear Microwave Devices [J]. IEEE Transactions on Microwave Theory and Techniques, 2017, 65(6): 2043~2062.
[15] Zhao Z, Zhang L, Feng F, Zhang W and Zhang Q, Space Mapping Technique Using Decomposed Mappings for GaN HEMT Modeling [J]. IEEE Transactions on Microwave Theory and Techniques, 2020, 68(8): 3318-3341.
[16] Chen B, Hsiao Y, Lin WC. et al. Using U-Net convolutional neural network to model pixel-based electrostatic potential distributions in GaN power MIS-HEMTs [J]. Scientific Reports, 2024, 14, 8151.
[17] Geng M, Zhu Z and Cai J. Small-Signal Behavioral Model for GaN HEMTs based on Long-Short Term Memory Networks [C]. 2021 IEEE MTT-S International Wireless Symposium (IWS), Nanjing, China, 2021, pp. 1-3.
[18] Marinković Z. Robustness Validation of a mm-Wave Model based on GRU Neural Networks for a GaN Power HEMT [C]. 2023 IEEE 33rd International Conference on Microelectronics (MIEL), Nis, Serbia, 2023: 1-4