Reconstruction of Missing Data Completely at Random for Trains Based on Improved GAN

He Jing[1][0000-0002-3650-3270], Chen Xin[2] and Zhang Changfan[3][0000-0002-7439-1775]

[1] College of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou, Hunan 412000, China
[2] College of Railway Transportation, Hunan University of Technology, Zhuzhou, Hunan 412000, China
[3] College of Railway Transportation, Hunan University of Technology, Zhuzhou, Hunan 412000, China
Corresponding author, E-mail: zhangchangfan@263.net

Abstract. Reconstruction of missing data for heavy-haul trains is an important factor in ensuring safe train operation. However, the existing methods of generative model require a complete data set for training, and it is very difficult for them to solve the issue of missing data completely at random. For this, this paper proposes a new attention-generative adversarial network to reconstruct missing data. First, a mask matrix is designed to locate the missing data, and the gradient descent algorithm is applied in combination with the output probability matrix of the discriminator, so that the mask matrix can still filling up the data well in the case of incomplete data set. Then, the prompt matrix is derived based on the mask matrix to solve the problem of model overfitting and accelerate the convergence. Finally, an attention mechanism is introduced into the whole GAN to improve the expression of data features by the feature extraction network. The experimental results show that the mean square error and mean absolute error percentage indexes of reconstruction accuracy can be kept below 1.5 for measurement data at different missing rates, and the reconstructed data can also well conform to the distribution law of measurement data.

Keywords: Incomplete data set; Attention-Generative Adversarial Networks; Missing data reconstruction; Reconstruction accuracy.

1 Introduction

High-quality data collection is important for securing a safe operation of the railway and for enabling intelligent operation and maintenance strategies. However, due to the wide operating range and complex operating environments of heavy-haul trains, such as mountainous and continuous tunnels, it is easy to encounter issues of network failure, transmission interruption, harmonic interference, etc., resulting in the missing of a large number of operation and maintenance data. The evaluation of train status and diagnosis of system faults would be impaired if too much data was missing. Therefore,

the key to ensure safe operation of the train is to reconstruction the missing data to the maximum extent, so as to achieve real-time and accurate monitoring of the train.

The missing data can be divided into three categories: missing completely at random, missing at random, and missing non-randomly. Data missing completely at random means that the missing of data is random, does not depend on any incomplete variable or complete variable. And the occurrence of null value is completely unrelated to known or unknown features of the data set. This also means that the feature relationship of the data surface cannot be utilized to fill the missing data. Instead, the internal feature relationship of the data needs to be dug out. The methods of filling up missing data are mainly the traditional methods and the deep learning-based missing data imputation methods.

The traditional missing data filling methods can be further classified into three categories: probability-based[1], interpolation-based[2] and similarity-based[3] methods. The most typical probability-based method is Expectation Maximization (EM). Sun[4] combined an EM interpolation model with the K-Mean algorithm, and obtained an enhanced stability of clustering and a better imputation of the missing data. Smerdon et al. [5] tried to solve the problem of large amounts of missing data in time series, with a data-driven RegEM algorithm. The EM algorithm has been widely used for filling up missing data, however, when the quantity of the missing data in the data set is large, the calculation speed of this algorithm decreases. Interpolation-based methods, such as linear interpolation, spline interpolation, and polynomial interpolation, can infer possible values of missing values based on known data points. When the amount of missing data is relatively small, such methods have fairly high accuracy. However, for large quantity of missing data, the performance of such methods degrade significantly. Similarity-based methods generally include K-Nearest Neighbor (KNN), K-means, and mean interpolation. K-Nearest Neighbor algorithm[6] has a faster calculation speed than EM algorithm, but in the case of extreme sample data missing, the interpolation accuracy will be reduced. Therefore, the traditional data reconstruction method is not superior in filling up missing data of high-speed train measurements.

Missing data imputation methods based on deep learning can be divided into three categories according to training strategies: autoregression[7,8], autoencoding[9,10] and adversarial training[11]. Autoregressive methods generally construct Recurrent Neural Networks (RNNs) and their variants to predict missing time steps by using complete or interpolated time steps. Such methods exploit the advantages of RNN in temporal correlation and interpolate missing values through time series prediction. As the time series prediction lacks a global receptive field[12], error compounding will be caused. Autoencoding methods compress a high-dimensional input into a low-dimensional hidden state with the encoder, and then reconstruct the hidden state with a decoder. The encoder and decoder can be multi-layer perceptrons, convolutional neural network or other neural networks. However, due to the bottleneck structure, the autoencoder inevitably encounters information loss. Adversarial training is an interpolation method based on Generative Adversarial Networks (GAN)[13]. It is an unsupervised generative model that can self-learn data distribution patterns and characteristics, and subsequently generating data that conforms to these patterns and characteristics. However, during the data set training process of traditional GAN, it is difficult to reach Nash equilibrium when

generating samples with the same distribution of the original data from random noise, resulting in gradient vanishing. More fundamentally, the training of these methods is not intuitive and differs from the imputation task. The inconsistency between the training and imputation processes limit the generalization performance of the model.

In this paper, a missing data imputation method based on attention-generative adversarial network (SAGAIN) is proposed. It is devotes to solve the reconstruction problem of data missing completely at random (MCAR) for train data measurements. The main contributions of this paper are summarized as follows:

(1) A new missing data imputation process is proposed. The missing data set is directly used for model training to solve the problem of filling up data that misses completely at random.

(2) A prompt matrix based on mask matrix is designed to locate the position of missing and to be trained data. The missing data would be processed in batches. This not only solves the problem of model overfitting, but also accelerates the model convergence.

(3) The Squeeze-and-Excitation Networks (SE-NET) of attention mechanism is introduced into the new model to increase the expression of data features by feature extraction network, and to explore deep features of data, so as to make the accuracy of filling missing data higher.

2 Attention-generative adversarial networks

The structure of the SAGAIN imputation algorithm designed for the reconstruction of missing data for train measurements is shown in Fig. 1. Its overall architecture includes data set processing, generator network, discriminator network, attention mechanism, output and other modules.

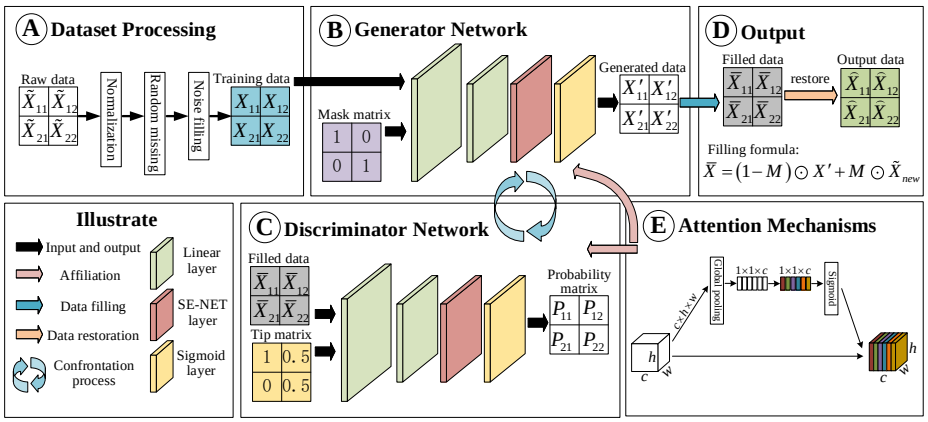


Fig. 1 Overall framework of missing data reconstruction model based attention-generative adversarial network

2.1 Missing data imputation process

The data imputation process flow of the traditional GAN network method is shown in Fig. 2. First, the complete data set for model training is input to the GAN network, and after the generator and discriminator are trained against each other, several sets of data will be created for data imputation. Then, these sets of generated data will be evaluated with missing data in turn, and the set with the best imputation effect will be selected. Finally, the original missing data in this set of data will be located, and used to fill up the positions of missing data void in the original data set.

If there are many sets of missing data to be imputed, the traditional GAN network method needs to perform the operation one by one, and each imputation operation needs to be re-evaluated with the generated data of all sets. Then at last the original data set can be filled up. Therefore, the whole process will be tedious and time consuming. Moreover, in order to learn the potential feature relationship between data, the training data must be very complete. In the actual situation, normally several to-be-processed sets of data are partially complete and partially missing, so it is necessary to select the complete data for training and then fill in the missing data.

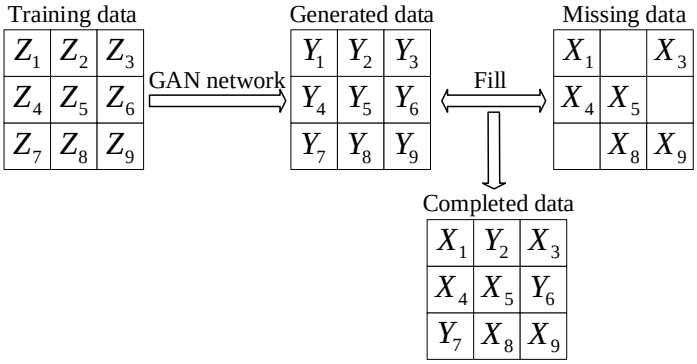


Fig. 2 Missing data imputation process of traditional GAN network

The SAGAIN network proposed in this paper has a simplified process for missing data imputation, as shown in Fig. 3. Specifically, the data mixed with missing data for processing is directly input into the SAGAIN network, and after the adversarial training of the generator and the discriminator, the filled complete data is directly output. Compared with the traditional GAN missing data imputation method, this model does not require to divide the missing data from the complete data. It integrates the data generation process and the data filling process, making the whole missing data imputation process simpler and more efficient.

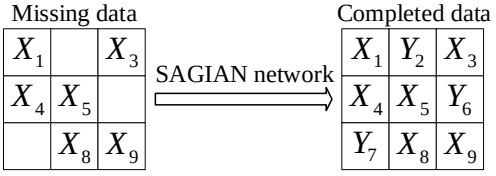


Fig. 3 Missing data imputation process of SAGAIN network proposed in this paper

2.2 Mask matrix and prompt matrix

Mask matrix. To accurately locate the position of missing data, the proposed model utilizes a mask matrix M to reflect which variables are observed data and which are imputed data. The mask matrix M only contains two variables, 0 and 1. As shown in Fig. 4, the blank part of the missing data matrix indicates missing data, the number in the mask matrix for such corresponding position is 0, meanwhile, the corresponding number for intact data is 1.

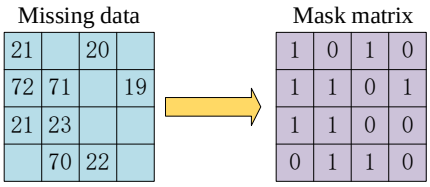


Fig. 4 Schematic diagram of mask matrix

Prompt matrix. If specific values are to be determined for M by training, overfitting is likely to occur. Thus, it is necessary to make some values in M uncertain. Here, an auxiliary variable B is defined first:

$$B = \left(B_1, ..., B_d\right) \in \left\{0,1\right\}^d \tag{2.1}$$

where, the specific value of B is set to 0 for 10% of B at random, and 1 for the rest. The dimension of B is consistent with that of the mask matrix M, then the following calculation is carried out:

$$H = B \odot M + 0.5\left(1 - B\right) \tag{2.2}$$

where, $\odot$ is the Hadamard product matrix operation. The relationship between the mask matrix M, auxiliary variable B, and prompt matrix H is shown in a table:

Table 1. Correlation of M, B and H

M	B	H
1	0	0.5
0	0	0.5
1	1	1
0	1	0

It can be seen in Table 1 that, when B = 0, H = 0.5; When B = 1, H = M. Therefore, when B = 1, the value of M can be inferred accurately from the value of H; When B = 0, the value of M cannot be inferred from H, and this uncertain value of M is the object on which the model is to be trained.

2.3 Generator

In SAGAIN, the goal of the generator is to accurately fill in the missing data. Therefore, the generator shall maximize the classification error rate of the discriminator, so that the two are in a process of opposing each other. The input of the generator is obtained from the mask matrix M and the supplementary data X from noise by Equation (2.3)

$$X = \tilde{X}_{new} \odot M + Z \odot (1 - M) \tag{2.3}$$

where, $\tilde{X}_{new}$ is the normalized raw data, and Z is the random noise with a value between 0 and 1.

The output is provided by Equation (2.4)

$$\bar{X} = (1 - M) \odot X' + M \odot \tilde{X}_{new} \tag{2.4}$$

In which, $\bar{X}$ is the output of the generator, X' is the generated data.

The generator shall interfere with the discrimination results of the discriminator to minimize the probability of the discriminator getting the correct answer. Since only values of B = 0 are trained, the sum of following two losses is to be minimized when B = 0:

(1) When the real data is missing, the discrimination result (the probability of missing the data generated by the generator in the original data) is determined as the missing data:

$$L_G\left(m_j,\ \mathrm{D}\left(\bar{x}_j, h_j, b_j\right)\right) = -\sum_{i \in j} (1 - m_i) \log\left(p_i\right) \tag{2.5}$$

(2) When the real data is not missing, the difference between the generated result and the original data:

$$L_M\left(x_j, x_j'\right) = \sum_{i \in j} m_i \left(x_i' - x_i\right)^2 \tag{2.6}$$

In Equation (2.5) and (2.6): j is any position in the corresponding matrix, i is the position of auxiliary variable B = 0, and P is the output probability matrix of the discriminator.

In summary, the final loss function is:

$$\sum \left[L_G\left(m_j,\ \mathrm{D}\left(\bar{x}_j, h_j, b_j\right)\right) + \alpha L_M\left(x_j, x_j'\right) \right] \tag{2.7}$$

2.4 Discriminator

The input for the discriminator are the output of the generator and the prompt matrix H. The goal of the discriminator is to accurately distinguish whether the data is filled data or real data, so the discriminator shall minimize the classification error rate. At the same time, in order to make the adversarial process obtain more ideal results, a prompt matrix H of partial information about the data is provided for the discriminator, so that the samples generated by the generator are forced to be close to the real data distribution. For this, the discriminator needs to output estimated values of the elements in the mask matrix, that is, the probability that the data is not missing. The estimated value of the discriminator is given by:

$$p_j = D\left(\bar{x}_j, m_j\right) \in (0 \sim 1) \tag{2.8}$$

Regarding the loss function, the discriminator shall maximize: (1) the difference between the discriminant result (the probability that the data generated by the generator is missing from the original data) and 1 when the real data is missing; (2) When the real data is not missing, the difference between the discrimination result and 0.

$$L_D\left(m_j,\ \mathrm{D}\left(\bar{x}_j,h_j,b_j\right)\right)=\sum_{i\in j}\left[m_i\log\left(p_i\right)+\left(1-m_i\right)\log\left(1-p_i\right)\right] \tag{2.9}$$

where, the log terms are all negative, so the difference that really needs to be minimized is (i.e., the discriminator loss function):

$$-L_D\left(m_j,\ \mathrm{D}\left(\bar{x}_j,h_j,b_j\right)\right) \tag{2.10}$$

2.5 SE-NET attention mechanism

Designing a high-performance discriminator D as well as a high-performance generator G is of great significance for learning deep features of data and generating high-quality data. Based on the GAN structure, a SE-NET attention mechanism module is added to the discriminator and the generator respectively, and a weight is used to indicate the importance of each channel in the next stage.

As shown in Fig. 5, the SE-NET attention mechanism is mainly composed of a SE module, a Squeeze operation, an Excitation operation and a feature fusion. The weights of each channel are assigned. After the Squeeze operation, the network obtains a global description. The Excitation operation and the feature fusion enable the fully connected layer to well integrate all the input feature information, and the Sigmoid function can also well map the input to the 0~1 interval.

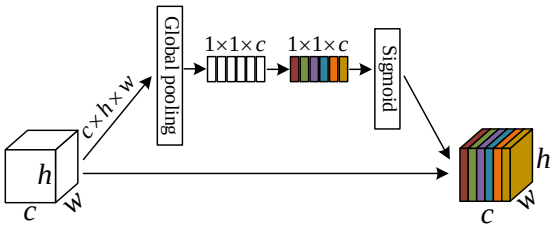


Fig. 5 SE-NET attention mechanism

As can be seen in Fig. 5, this attention mechanism can be divided into the following three steps:

First, the process of Squeeze is to convert the output code of the upper layer into a statistical vector through global pooling compression. The calculation formula is as follows:

$$z_t=F_{sq}(\mathrm{h}_t)=\frac{1}{k}\sum_{t=1}^{k}h_i^{(t)} \tag{2.11}$$

where, $h=\left[h_1,h_2\cdots h_i\right]$ is the output encode of the previous layer, and $z=\left[z_1,\mathrm{z}_{2,}\cdots z_i\right]$ is the statistical vector after compression and conversion. In this process, the performance of average pooling is obviously better than that of maximum pooling, so Z is calculated with the average values.

Second, the process of Excitation is to fuse all input features through the fully connected layer of two nonlinear activation functions, and the Sigmoid function can keep the output within a specific range at last. The assignment calculation is as follows:

$$A = F_{ex}(\mathrm{Z}) = \sigma_2\left(W_2\sigma_1\left(W_2 Z\right)\right) \tag{2.12}$$

where, σ_1 and σ_2 are functions of ReLU and Sigmoid functions, respectively. Through the scaling function and assignment of W_1, W_2, the number of model parameters is reduced, and the calculation speed is increased.

Finally, the fusion process is to fuse the channel weights obtained by the above two operations with the original features, and to assign weights to the features through simple multiplication.

3 Experimental results

3.1 Case preparation and data set processing

The program experiments were implemented by Python code. The hardware environment was CPU processor Intel(R) CoreTM i5-9300H CPU, frequency 2.40GHz; The GPU was NVIDIA GeForce GTX 1650, and the platform version was Python 3.7.7 and torch 1.4.0.

Table 2. Train data analysis

	Max. positioning DC voltage	Min. positioning DC voltage	Average positioning DC voltage	Max. positioning AC voltage	Min. positioning AC voltage
Ranges	21.55-23.0	20.52-22.78	21.53-22.83	70.83-77.09	70.45-73.65
Values	22.2644	22.1262	22.1832	72.2653	71.8572
Variances	0.0937	0.0904	0.0881	0.4837	0.5548

In the experiment, the actual operation and maintenance data of a train in 32 days were selected, and the five characteristics of the same equipment (maximum and minimum values of positioning DC voltage; maximum, minimum and average values of positioning AC voltage) were used. The data selection of the same equipment improves the strong correlation of data characteristics. This group of data had a total of 1300 complete and intact data items, Table 2 shows the specific information of the data. In order to verify the interpolation effect of this model on missing data, random sampling was applied in this paper to generate missing data for the sample data.

Since the data of this model was the complete data set X, the experiment used the binary mask matrix and the complete data to perform Hadamard product operation to represent the missing data. Considering the uncertainty and uncontrollability of the missing position of the measured data during the actual operation of the train, the generated mask matrix was set randomly, in which 1 represented intact and 0 indicated

missing. The number of missing measurements was controlled by controlling the number of 0 in the mask matrix. The specific steps were to generate a mask matrix M (0 set for the missing part and 1 for the complete part) with randomly missing data according to a certain missing rate (proportion of 0), and then the obtained data set was as follows:

$$\bar{X} = M \odot X \tag{3.1}$$

Finally, the data set was divided into training set X_{train} and test set X_{test} at a ratio of 8:2, and the corresponding mask matrices were M_{train} and M_{test}. In actual situation, the mask matrix would be generated from the corresponding missing data set. The dimension of the training set was (1040, 5), while the dimension of the test set was (260, 5). In order to evaluate the ability of the model in handling the imputation of missing data, the training set and test set data must not be duplicated.

3.2 Parameter setting and evaluation indicators

In the experiment, the learning rate of generator G and discriminator D of the training model was set to 0.001. Due to the small number of data samples in this experiment, in order to make the network reach the optimal solution of gradient descent faster and make the model converge to stability quickly, Batch size was set to 128 for 15,000 Epoch cycles. An unsupervised learning effect was achieved in attribution to the model structure.

For the evaluation of missing data reconstruction, two indicators, Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE), were used in this paper. The calculation methods are shown in Equation (3.2) and (3.3):

$$MSE\left(x, \hat{x}\right) = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \hat{x}_i)^2\right) \tag{3.2}$$

$$MAPE\left(x, \hat{x}\right) = \frac{\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right|}{n} \times 100 \tag{3.3}$$

where, x_i represents the original train measurement data, and $\hat{x}_i$ indicates the completed data after reconstruction. The values of above two indicators of the reconstruction results reflect the performance of missing data reconstruction. The smaller the MSE and MAPE values, the better the reconstruction performance.

3.3 Experimental analysis

Comparison of experimental results at different missing rates. The experiment performed reconstruction calculations for random missing data on 260 test data items. The test set X_{test} was input into the trained model. If for a device, data of a feature was missing, the feature values of other devices on the current day and the past days could be learned, and then the missing part could be reconstructed by combining the rules of other non-missing feature values of this device. The MSE and MAPE values shown in Table 3 and 4 were calculated from the measured data and the reconstructed data.

Table 3. MSE of reconstruction results

Missing rate	Max. positioning DC voltage	Min. positioning DC voltage	Average positioning DC voltage	Max. positioning AC voltage	Min. positioning AC voltage
0.1	0.029838	0.014086	0.027521	0.203050	0.066439
0.2	0.022212	0.022171	0.044166	0.231970	0.481436
0.3	0.015801	0.083700	0.021629	0.361021	0.275658
0.4	0.058681	0.116857	0.055213	0.478309	0.411178
0.5	0.147878	0.066751	0.051127	0.642647	1.095744

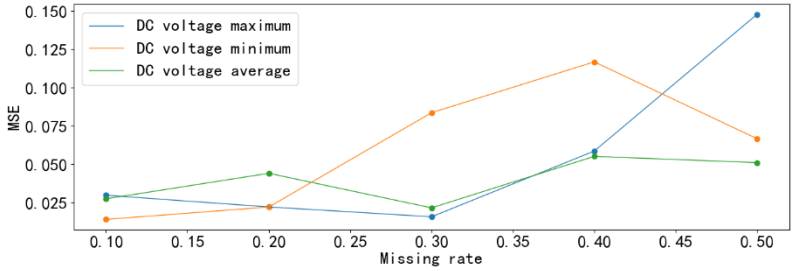


Fig. 6 MSE of DC voltage at different missing rates

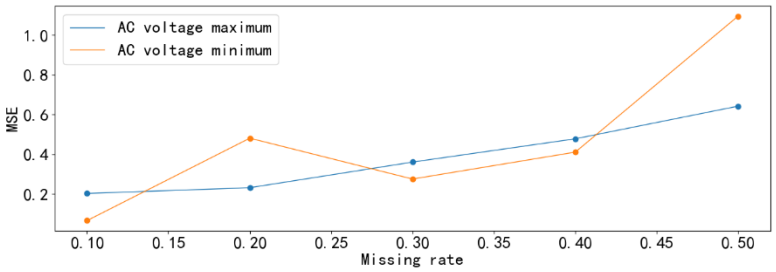


Fig. 7 MSE of AC voltage at different missing rates

The MSE values at different missing rates are shown in Table 3 and Fig. 6-7. As shown in Fig. 6, for the reconstruction of three sets of DC voltage data, MSE has always been at a relatively low value, ranging from 0.01 to 0.15, but with the increase of the missing rate, MSE generally shows a slow upward trend; While Fig. 7 shows that, for the reconstruction of two sets of AC voltage data, the evaluation indicators ranges from 0.1 to 1.1, and shows a uniform upward trend with the increase of the missing rate. This indicates that the model maintains a high reconstruction accuracy, and the reconstruction performance for DC data is better than that for AC data.

Table 4. MAPE of reconstruction results

Missing rate	Max. positioning DC voltage	Min. positioning DC voltage	Average positioning DC voltage	Max. positioning AC voltage	Min. positioning AC voltage
0.1	0.627145	0.415247	0.591765	0.485682	0.266732
0.2	0.442620	0.565620	0.619057	0.511420	0.760844
0.3	0.437926	1.054228	0.532856	0.648620	0.566849
0.4	0.820259	1.289491	0.792290	0.738592	0.712877
0.5	1.447812	0.905875	0.772731	0.871104	1.208223

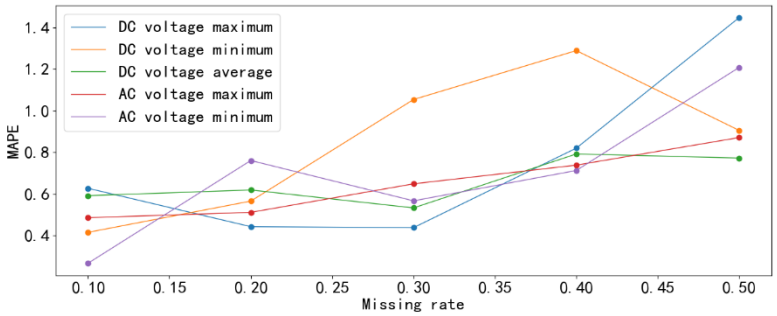


Fig. 8 MAPE at different missing rates

The MAPE values at different missing rates are shown in Table 4 and Fig. 8. It can be seen from Fig. 8 that when the missing rate is less than or equal to 0.3, the evaluation indicators are in a slow rising state and remain at a low value. When the missing rate is greater than 0.3, there are great changes in the evaluation indicators. Although the increased amount of missing data will affect the reconstruction performance, generally, the value of the indicator is always below 1.5. This means the error between the reconstructed data and the original measurement data is very small. In another word, when data is missing at a large quantity in the measured operation and maintenance data of trains, the data reconstruction model SAGAIN proposed in this paper has the ability of reconstructing missing data at high accuracy.

Overall, the data is missing at a rate of 10% to 50%, this model can always reconstruct the missing data with a high accuracy. When the missing rate is lower than 30%, the reconstruction results of the five eigenvalues do not change much and a high accuracy is maintained, the reconstruction effect is still quite good when the missing rate is 50%. It indicates that the model can handle the reconstruction of large amount of missing data for high-speed trains.

Comparison of missing data imputation by different algorithms. As shown in Table 5, in the case of a missing rate of 20%, a comparison of the maximum and

minimum values of positioning DC voltage was carried out. Another four currently widely used generative models (GAN, VAE, VAE-GAN and VAE-FGAN)[14] were applied to obtained corresponding missing imputation results. The parameters were set as follows: the learning rate of encoder E and discriminator D was set to 0.001, the learning rate of generator G was set to 0.0002, and Batch size was set to 128, for 100 Epoch cycles.

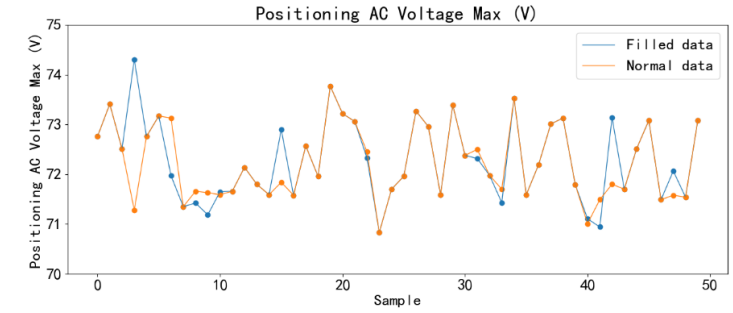
Compared with the three generative models of GAN, VAE and VAE-GAN, the model proposed in this paper has superior performance regarding various indicators. Compared with the VAE-FGAN model, although the evaluation index MAPE of the proposed model is higher for the data missing reconstruction of the maximum positing DC voltage, the increase remains in a low range. While the values of the other three indicators are still significantly better than those of VAE-FFGAN model, especially the MAPE of the minimum DC voltage, which decreases from 0.1623 to 0.0221, indicating a very large improvement. By analysis, the increase of the indicator value could be due to the fact that the data set required by the model SAGAIN in this paper does not need to be a complete data set, but with a certain part of data missing, This is also in line with the actual situation of train measurement data.

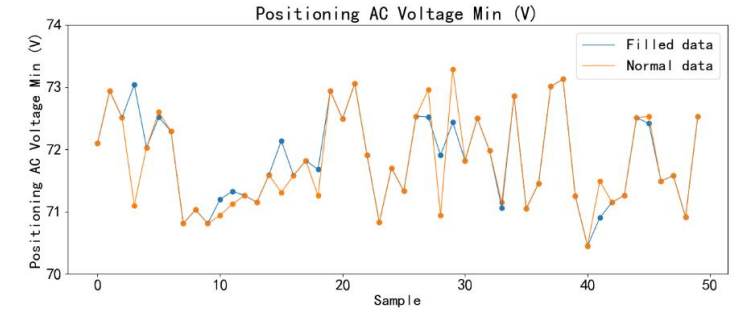
Table 5. Comparison of missing data reconstruction by different models

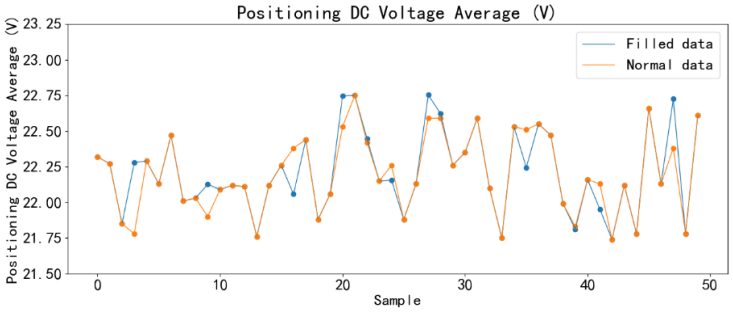
Comparison method	Max. positioning DC voltage		Min. positioning DC voltage	
	MSE	MAPE	MSE	MAPE
GAN	0.1016	1.2674	0.3394	2.2842
VAE	0.0976	1.2234	0.2132	1.2561
VAE-GAN	0.0846	1.1891	0.2098	0.9263
VAE-FGAN	0.0314	0.3505	0.1623	0.6659
SAGAIN (proposed in this paper)	0.0222	0.4426	0.0221	0.5656

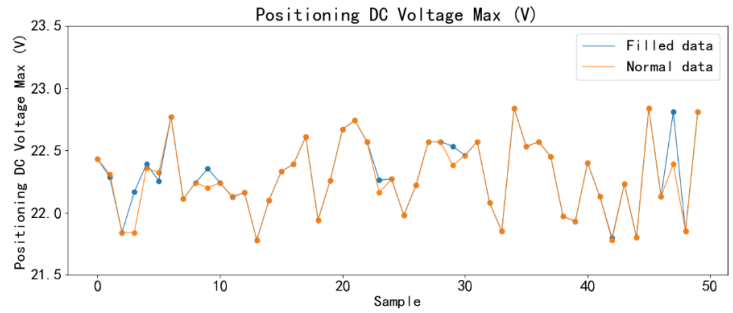
Missing data imputation experiment. For complex railway operation scenarios, the loss of measurement data at certain times is prone to occur, and even a long time data missing is possible due to the maintenance of train sunroof equipment. To evaluate the performance of missing data imputation, the generated missing data shall be filled back into the original contextual data to see if the imputation data fits the distribution characteristics. The test data was randomly deleted with a deletion rate of 20%, and the missing data was reconstructed by the model in this paper, and the first 50 samples of each feature were taken, the results are shown in Fig. 9. The orange curve represents the distribution of the original data, while the blue curve represents the reconstructed data. The non-coincidence of the two curves indicates the positions of missing and filled data. The difference between the blue and orange points can intuitively express the difference of the filled data. It can be seen from the five feature maps in Fig. 9 that the two curves have a high degree of fitting, which indicates that the proposed SAGAIN model

can greatly restore the original data information distribution and achieve a high reconstruction accuracy by learning the feature law of data from the same equipment.









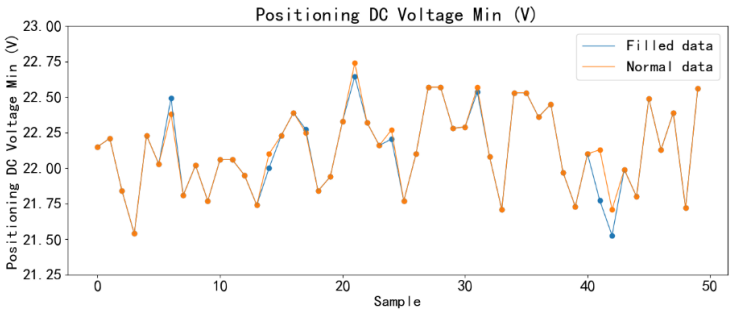


Fig. 9 Visualization of data reconstruction performance

4 Conclusions

In order to ensure the safe operation of trains and solve the problem of measurement data missing at random, this paper proposes a missing data imputation method SAGAIN based on generative adversarial networks. It is verified to be effective by good results of reconstructing missing data. The main conclusions are as follows:

(1) The proposed model in this paper applies the data set with missing directly instead of using only complete data set as the traditional generative models. Based on a new generative adversarial network, the model integrates the generation and imputation of missing data into one process. Data set with missing data as the input for the training, and reconstructed complete data set as the output.

(2) The influence of different data missing rates on reconstruction performance of the proposed model is evaluated. With increasing data missing rate, the value of reconstruction indicators ascend slowly. Fortunately, the MSE and MAPE indicators are always below 1.5 at missing rate of 0.1~0.5, remaining at a relatively low value. It means that the reconstruction model of SAGAIN in this paper has high reconstruction accuracy even when there is a large number of data missing in train operation and maintenance measurements.

(3) The SAGAIN model proposed in this paper is an improved model based on the GAN model. A SE-NET attention mechanism module is introduced into the GAN model. By comparing with other missing data imputation algorithms, it can be seen that after adding the attention mechanism, the four groups of evaluation indicators of the proposed model are improved and significantly superior to those of other generative models, and the reconstructed data also meet the distribution characteristics of measured data, indicating that the attention mechanism has a good effect on channel weighting and improves the imputation accuracy.

For future work, efforts can be devoted to improve the SAGAIN network model. For example, the prompt matrix in the network architecture is to prevent the generator G from failing to obtain effective gradient information and speed up convergence, however, it also complicates the network model to a certain extent and increases the training time. Studies can be carried out to replace the prompt matrix with other methods to further simplify the network model, so as to speed up the convergence of the model on

the premise of ensuring the imputation accuracy. This is also conducive for the network model to use the transfer learning method to solve the problem that the original data is a small sample.

Acknowledgments. This work was supported by the National Key R&D Program of China 2021YFF0501101, the National Natural Science Foundation of China 62173137 and the Project of Hunan Provincial Department of Education 23A0426.

References

1. Z. Ma, H. Li, Y. Weng, E. Blasch and X. Zheng, Hd-Deep-EM: Deep Expectation Maximization for Dynamic Hidden State Recovery Using Heterogeneous Data. IEEE Transactions on Power Systems, 39(2), 3575-3587 (2024).
2. A. Aboutorabi and M. Brockmann, Vehicle Axle Acceleration Prediction: An Interpolation Approach. In: 2024 IEEE 18th International Conference on Advanced Motion Control (AMC), pp. 1-6 (2024).
3. A. B. P. Utama, A. P. Wibawa, A. N. Handayani, W. S. G. Irianto, Aripriharta and A. Nyoto, Improving Time-Series Forecasting Performance Using Imputation Techniques in Deep Learning. In: 2024 International Conference on Smart Computing, IoT and Machine Learning (SIML), pp. 232-238 (2024).
4. S. Hua-Yan, L. Ye-Li, Z. Yun-Fei and H. Xu, Accelerating EM Missing Data Filling Algorithm Based on the K-Means. In: 2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC), pp. 401-406 (2018).
5. Smerdon, J. E., A. Kaplan and D. Chang, On the Origin of the Standardization Sensitivity in RegEM Climate Field Reconstructions. Journal of Climate, 21(24), 6710–6723 (2008).
6. M. Pazhoohesh, Z. Pourmirza and S. Walker, A Comparison of Methods for Missing Data Treatment in Building Sensor Data. In: 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), pp. 255-259 (2019).
7. J. Yoon, W. R. Zame and M. van der Schaar, Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks. IEEE Transactions on Biomedical Engineering, 66(5), 1477-1490 (2019).
8. X. Kong, W. Zhou, G. Shen, W. Zhang, N. Liu and Y. Yang, Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data. Knowledge-Based Systems, 261 (2023).
9. X. Liu and Z. Zhang, A Two-Stage Deep Autoencoder-Based Missing Data Imputation Method for Wind Farm SCADA Data. IEEE Sensors Journal, 21(9), 10933-10945 (2021).
10. Yuwei Fan, Chenlong Feng, Rui Wu, Chao Liu and Dongxiang Jiang, Multiscale-attention masked autoencoder for missing data imputation of wind turbines. Knowledge-Based Systems, 299 (2024).
11. Z. Sun, H. Li, W. Wang, J. Liu and X. Liu, DTIN: dual Transformer-based Imputation Nets for multivariate time series emitter missing data. Knowledge-Based Systems, 284 (2024).
12. W. Du, D. Cote and Y. Liu, SAITS: self-attention-based imputation for time series. Expert Systems with Applications, 219 (2023).
13. R. Shahbazian and S. Greco, Generative Adversarial Networks Assist Missing Data Imputation: A Comprehensive Survey and Evaluation. IEEE Access, 11, 88908-88928 (2023).
14. Changfan Zhang, Hongrun Chen, Jing He and Haonan Yang, Reconstruction method for missing measurement data based on wasserstein generative adversarial network. Journal of Advanced Computational Intelligence and Intelligent Informatics, 25(2), 195-203 (2021).