


Continuous missing data imputation with incomplete dataset by generative adversarial networks–based unsupervised learning for long-term bridge health monitoring

Structural Health Monitoring
2021, Vol. 0(0) 1–17
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14759217211021942
journals.sagepub.com/home/shm


Huachen Jiang¹, Chunfeng Wan¹ , Kang Yang¹,
Youliang Ding¹  and Songtao Xue^{2,3}

Abstract

Wireless sensors are the key components of structural health monitoring systems. During the signal transmission, sensor failure is inevitable, among which, data loss is the most common type. Missing data problem poses a huge challenge to the consequent damage detection and condition assessment, and therefore, great importance should be attached. Conventional missing data imputation basically adopts the correlation-based method, especially for strain monitoring data. However, such methods often require delicate model selection, and the correlations for vehicle-induced strains are much harder to be captured compared with temperature-induced strains. In this article, a novel data-driven generative adversarial network (GAN) for imputing missing strain response is proposed. As opposed to traditional ways where correlations for inter-strains are explicitly modeled, the proposed method directly imputes the missing data considering the spatial–temporal relationships with other strain sensors based on the remaining observed data. Furthermore, the intact and complete dataset is not even necessary during the training process, which shows another great superiority over the model-based imputation method. The proposed method is implemented and verified on a real concrete bridge. In order to demonstrate the applicability and robustness of the GAN, imputation for single and multiple sensors is studied. Results show the proposed method provides an excellent performance of imputation accuracy and efficiency.

Keywords

Structural health monitoring, missing data imputation, generative adversarial network, sensors, deep learning, artificial neural network

Introduction

In recent decades, an increasing number of structural health monitoring (SHM) systems have been deployed on civil infrastructures with the aim of monitoring and evaluating the operational performance of the structures.^{1–5} A comprehensive SHM system always requires a large quantity of sensors. Considering strain responses play a crucial role in tracking down structural conditions,^{6–11} collected strain responses are concerned in this article. Sensor-based SHM system produces a large amount of data every day, which underlies the damage detection and condition assessment technique. However, the abnormal data can undermine the reliability of the overall monitoring and evaluating process. The anomaly often occurs because of improper installation, environmental noise, raw weather, and other factors that are

beyond our control. Among all the anomaly types, data loss is the most pervasive one.

Data loss is a challenging issue which will inevitably impede the analysis of the structure condition. For instance,

¹Southeast University, Key Laboratory of Concrete and Prestressed Concrete Structure of Ministry of Education, Nanjing, China

²Research Institute of Structural Engineering and Disaster Reduction, College of Civil Engineering, Tongji University, Shanghai, China

³Department of Architecture, Tohoku Institute of Technology, Sendai, Japan

Corresponding author:

Chunfeng Wan, School of Civil Engineering, Southeast University, Sipailou 2#, Nanjing 210096, China.
Email: wan@seu.edu.cn

the long-term field static strain measurements can be approximately conceptualized by a linear regression model with a sinusoid component, and such model acts as a baseline for reliability assessment and early warning technique.¹² If the sensor malfunction occurs and results in data loss, then the safety and serviceability will not be correctly estimated and ensured. Furthermore, if there is a vast amount of missing data, the model proposed will not even be workable. One may wonder the substitution of the normally functioning strain sensors for the faulty ones will address such problem; however, it will not be the case we expect. As the sensor fails, there still exists variation of strain (caused by temperature, vehicle, creep, etc.), and the substitution sensor is not able to capture this offset when installed.¹³

Therefore, in order to overcome the missing data problem, efficient missing data imputation or data loss reconstruction techniques are widely adopted, and most of them are correlation-based methods. Chen et al.¹⁴ correlated two strain sensors located at different positions on the bridge using functional data analysis technique. The inter-sensor correlations were composed of three parts: correlations for temperature-induced strain, correlations for vehicle-induced strain, and correlations for stochastic excitation-induced strain. These three correlations were separately modeled in explicit ways and then combined together to impute the continuous loss of data. The imputation result was satisfactory, while the delicate model selection was required. Huang et al.^{15,16} imputed static strain data utilizing correlations between strain sensors and correlations between strain and temperature sensors. Correlations were modeled by LS-SVM and extreme learning machine. In this case, sampling frequency was much lower so that the monitoring dataset was smooth without erratic fluctuation caused by vehicle; thus, many imputation methods are capable of fulfilling the task. Lu et al.¹⁷ successfully reconstructed strain measurements via correlation-based model of partial least square approach. Zhang and Luo¹⁸ interpolated missing stress measurements by analyzing spatial correlations of different strain sensors. They also pointed out that continuous missing data are harder to be imputed compared with discrete missing data, and the ratio of the data loss should not exceed 30%.

With the rapid development of technology, artificial intelligence (AI) flourishes and enjoys the most promising discipline. Deep learning method is one of the emerging AI techniques which can handle complex data in large volume and has been applied in SHM filed in recent years. However, most imputation methods based on deep learning algorithms are used to reconstruct dynamic acceleration or speed measurements^{19–22} and very few of them concern strain measurements.¹³ Fan et al.²⁰ reconstructed dynamic acceleration data using convolutional neural networks (CNNs) even when the loss ratio reaches 90%. Jeong et al.²¹

reconstructed vibration data through bidirectional recurrent neural network (RNN) which implicitly took spatial and temporal correlations among sensors into account. Perez-Ramirez et al.²² presented a novel RNN to accurately predict the vibration response of large buildings under seismic excitations and ambient vibrations. Oh et al.¹³ trained a CNN model to restore the faulty strain sensor only using functioning sensors. In the study, datasets were arranged in 2D grayscale grid to be fed into the CNN and the final input size was properly set as a full square grid format. Fan et al.²³ designed a densely connected CNN, with skip connection and dense block techniques carefully applied, and accurately reconstructed dynamic responses in both time and frequency domain. Wan and Ni²⁴ adopted Bayesian multitask learning methodology for reconstruction of both temperature and acceleration responses, showing its advantages when the dataset is too limited.

It is found that conventional machine learning or statistic methods can hardly capture the correlations for dynamic strain measurements unless the model is well-chosen. Deep learning-based models are qualified for discovering potential dynamic changing relationships among sensors owing to their astonishing learning capacity of neural networks. However, conventional methods mentioned above usually require complete and intact datasets for training, in another word, lacks capacity to impute an incomplete dataset directly. In many cases, data loss is an inherent structure of the system and obtaining complete dataset is challenging and sometimes impossible. In this article, a novel unsupervised leaning imputation method by adapting generative adversarial networks (GANs) is presented. The proposed network only requires incomplete datasets for training, and missing data will be reconstructed directly based on the remaining observed data, that is to say, imputation for each sensor is basically taking the strain distribution upon the bridge into account rather than itself or correlation with another sensor. In this case, the imputation quality is guaranteed and convincible. In this article, ground true data are only demonstrated for comparison with the imputed value and will not be processed through the framework. In the remaining sections, GAN technique will be introduced and the architecture of the network is established. Then, datasets obtained from a real concrete bridge located in China is analyzed. In the analysis, both single-sensor imputation and multiple-sensor imputation results with various loss ratios are exhibited. Considering discrete data loss can be easily imputed using interpolation- or correlation-based methods; continuous missing data are, therefore, only concerned in this article.

GAN-based imputation method

Generative adversarial network is one of the cutting-edge deep learning methods and its astounding capacity of

generating entirely new data has attracted extensive attention. When GAN was first proposed by Goodfellow et al.²⁵ in 2014, many research studies have been devoted to the field, including data generation, style transfer, image enhancement and other applications.²⁶ As for data generation, training whatever algorithms mentioned above or even conventional GANs²⁷ for imputation require real and fully observed data, which is not always an easy case especially for wireless sensor-based SHM systems. Thus, in recent years, more and more incomplete data-based imputation algorithms have emerged, and most of them are implemented by GANs considering its powerful generation capacity.^{28–31} Although with advanced methods being developed, most GAN-based methods are used to detect damage occurrence and evaluate the health condition of the bridge. Generative adversarial network-related imputation methods³² are rarely adopted and mostly based on computer vision techniques which also require complete dataset. In fact, monitoring data tends to be easily affected by the variation of ambient environment and these algorithms will not work stably. In this article, we leverage the idea of GAN imputation and present a novel GAN framework which only requires incomplete data that are applicable for SHM.

GANs for imputation

Generative adversarial network is a type of unsupervised generative model, and a generative model is generally used to summarize the distribution of data. However, unlike many other conventional generative models, including Latent Dirichlet Allocation,³³ restricted Boltzmann machines,³⁴ and deep belief networks,³⁵ which all have serious limitations and require quite a few parameters to represent underlying distributions, GAN-based generative models have a much better generalization ability. Typically, a GAN is composed of two halves: a generator and a discriminator. Generator is trained to generate examples of real data, and discriminator is trained to distinguish the generated examples from true ones. In the GAN-based imputation framework, we propose the generator takes the original observed strain distribution, a random noise matrix, and a mask matrix which indicates which part is real or fake and

then imputes the missing data. The discriminator takes the generated data according to the generator and a hint matrix which reveals partial information of the mask matrix and determines which components are imputed or real. In this framework, the generator's objective is to impute the loss data accurately as possible as it can so as to "confuse" the discriminator, and the discriminator's objective is to accurately distinguish the imputed data from real one as possible as it can. Finally, the missing data are properly generated when the discriminator is incapable of telling the real samples from the fake ones. In another word, the generator has to maximize the misclassification rate of the discriminator and the discriminator aims to minimize the misclassification rate. In such adversarial training process, the generator finally imputes missing data in high accuracy even for highly missing rate. The diagram of the imputation architecture based on GAN is depicted in Figure 1.

Architecture of the presented GAN

The monitoring data imputation experiment conducted in this study is based on a real concrete box girder bridge, located in China. The sensor deployment of the bridge is shown in Figure 2, and 15 strain sensors are selected to conduct the imputation job as plotted in green dots. The main goal of the imputation strategy is to complete the continuous or block missing data, both for single sensor and multiple sensors. Continuous data missing is a very common faulty type in long-term bridge monitoring; continuous missing data are much harder to be imputed compared with discrete missing data because discrete data can be easily imputed by interpolation methods or correlation-based methods. However, as for continuous missing data, conventional imputation methods cease to work because of their limitations in complexity description. Therefore, in this study, continuous data missing will be focused.

Considering 15 strain sensors that form the dataset, we define a 15-dimensional random variable $\mathbf{X} = (X_1, X_2, \dots, X_{15})$, which is a collection of strain measurements from 15 sensors at a same time point. Therefore, the dataset can be arranged as $\mathbf{D} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)^T$, where n is the number of total sampling points. Suppose that $\mathbf{M} = (M_1, M_2, \dots, M_{15})$

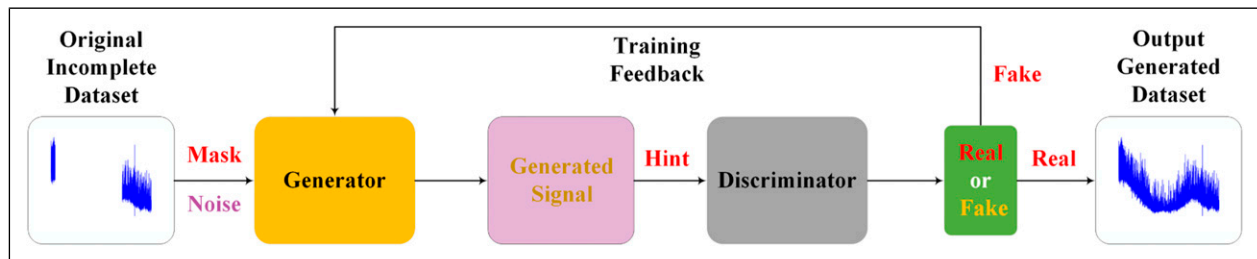


Figure 1. Imputation architecture of the generative adversarial networks.

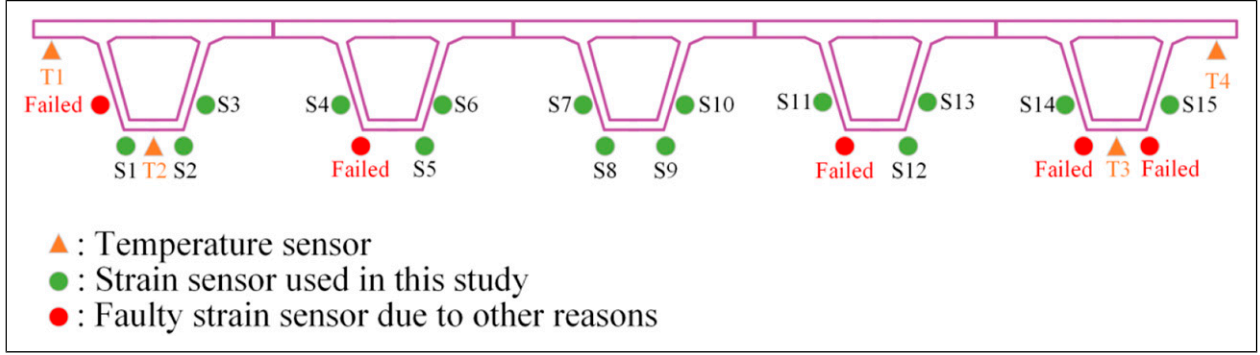


Figure 2. Sensor deployment of the structural health monitoring system.

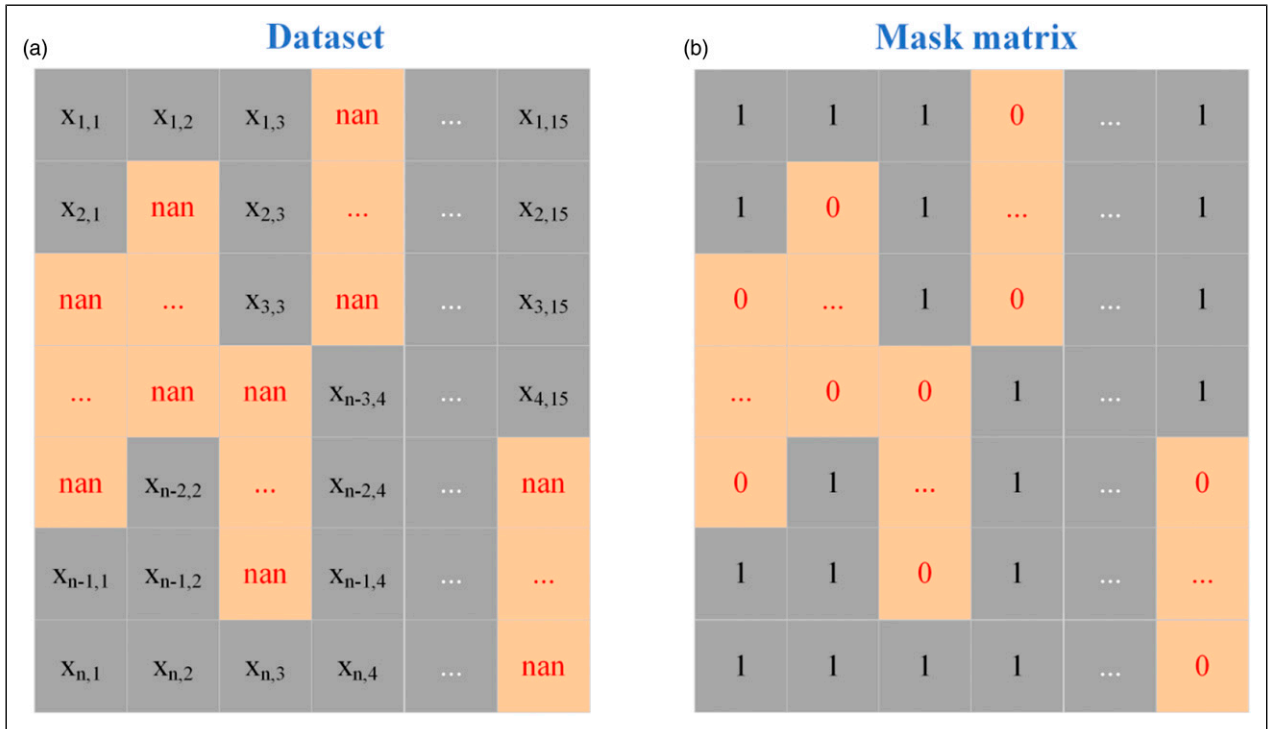


Figure 3. Simple demonstration of the dataset and its corresponding mask matrix: (a) data matrix and (b) mask matrix.

is a 15-dimensional random variable, namely, mask vector which represents the indices of the missing data and takes the value either 0 or 1. Therefore, we define the incomplete dataset $\tilde{\mathbf{D}} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \dots, \tilde{\mathbf{X}}_n)^T$, where $\tilde{\mathbf{X}} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_{15})$ is a random variable and is masked by M

$$\tilde{X}_i = \begin{cases} X_i, & \text{if } M_i = 1 \\ \text{nan}, & \text{if } M_i = 0 \end{cases} \quad (1)$$

where *nan* is the abbreviation of “not a number” which represents the missing data. The form of the dataset is demonstrated in Figure 3. In this study, only incomplete data $\tilde{\mathbf{D}}$ are used to train and then impute the missing data. The original fully observed data \mathbf{D} are only for verifying the

accuracy of the algorithm. The whole process is conducted under unsupervised learning, thus, model based on the entity or the signal itself is not required.

In the imputation framework, our main goal is to impute the missing data for each $\tilde{\mathbf{X}}$ according to $P(\mathbf{X}|\tilde{\mathbf{X}} = \tilde{\mathbf{x}})$, where $\tilde{\mathbf{x}}$ is one realization of $\tilde{\mathbf{X}}$; thus, the imputation strategy is based on the conditional distribution of \mathbf{X} given by $\tilde{\mathbf{X}} = \tilde{\mathbf{x}}$ rather than the expectation of \mathbf{X} only. Therefore, the GAN framework proposed attempts to model the conditional distribution and then generate the missing data.

Generator network. The generator network takes the samples of $\tilde{\mathbf{X}}$, \mathbf{M} , and a random noise variable \mathbf{Z} (same dimension with $\tilde{\mathbf{X}}$ and \mathbf{M} , as shown in Figure 4(a)), $\mathbf{Z} = (Z_1, Z_2, \dots, Z_{15})$.

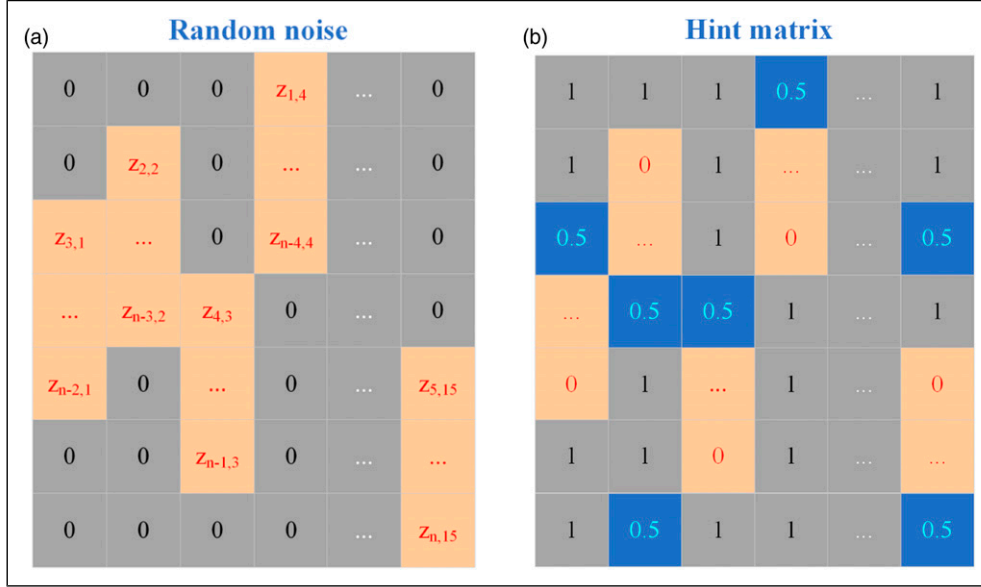


Figure 4. Simple demonstration of the random noise matrix and hint matrix: (a) random noise matrix and (b) hint matrix.

It is noted that the random noise is included in the missing components for initializing and generating complete dataset.

Then, the imputed data and the complete data are defined as follows

$$\bar{\mathbf{X}} = G(\tilde{\mathbf{X}}, \mathbf{M}, (1 - \mathbf{M}) \otimes \mathbf{Z}) \quad (2)$$

$$\hat{\mathbf{X}} = \mathbf{M} \otimes \tilde{\mathbf{X}} + (1 - \mathbf{M}) \otimes \bar{\mathbf{X}} \quad (3)$$

where $\bar{\mathbf{X}}$ is the imputed data vector, and G represents the computation of the generator network. $\hat{\mathbf{X}}$ represents the complete data finally obtained. Specifically, the observed data are directly outputted unchangeably and the missing data are processed during the generator network and then outputted. \otimes denotes element-wise multiplication.

The architecture of the generator neural network is demonstrated in Figure 5 and the generator is composed of two fully connected layers. The activation functions for the hidden layers and the output layer are Rectified Linear Unit (ReLU)³⁶ and sigmoid,³⁷ respectively. 128 represents the batch size for each iteration and 15 represents the number of sensors. ReLU and sigmoid functions are formulated as follows

$$\text{ReLU} : f(x) = \max(0, x) = \begin{cases} x_i, & \text{if } x_i \geq 0 \\ 0, & \text{if } x_i < 0 \end{cases} \quad (4)$$

$$\text{Sigmoid} : f(x) = \frac{1}{1 + \exp^{-x}} \quad (5)$$

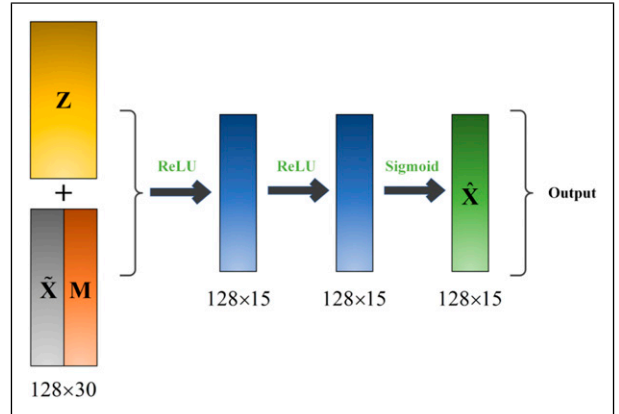


Figure 5. Architecture of the generator network.

Discriminator network. Unlike the discriminator in conventional GAN framework which only outputs the binary label: either fake or real, in this study, partial values are real (actually observed measurements) and partial values are fake (imputed data), so a more flexible classification discriminator is required. In this framework, the discriminator outputs another mask matrix that indicates which part is real or fake; in another word, the discriminator attempts to output the predefined mask matrix mentioned above.

In order to ensure the imputation quality, a hint mechanism²⁴ which delivers partial information of the loss distribution to the discriminator was introduced. The hint matrix is also predefined and dependent on the mask matrix, a simple demonstration of the hint matrix is shown in Figure 4(b). Therefore, the discriminator takes the final imputed data $\hat{\mathbf{X}}$,

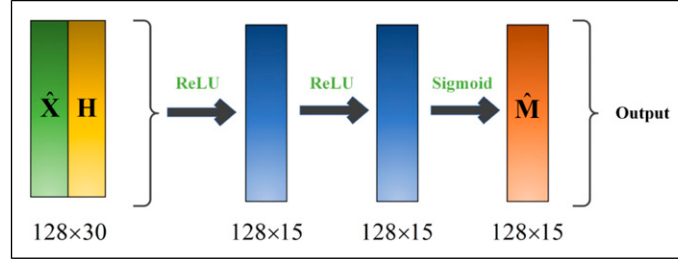


Figure 6. Architecture of the discriminator network.

the hint matrix \mathbf{H} , and then attempts to predict the mask matrix $\hat{\mathbf{M}}$. The discriminator shares the same network architecture with the generator as depicted in Figure 6.

Hint mechanism. In order to ensure the generator actually captures the distribution feature of the real data,²⁴ a hint mechanism \mathbf{H} is introduced. The realization of \mathbf{H} is dependent on the distribution $\mathbf{H}|\mathbf{M}$ and will be fed into the discriminator network. In another word, partial information of the mask matrix is known to the discriminator and provides a “hint” for it. In mask matrix, 0 represents missing data and 1 represents observed data. The discriminator of GAN tries to output the matrix like “mask matrix” in Figure 3(b), which means the discriminator tries to distinguish the missing data (represented by 0) from the observed data (represented by 1). However, a large amount of continuous missing data without any information provided will affect the accuracy of the imputation, and GAN will output different imputation results every time. Therefore, a hint is provided as shown in Figure 4(b). The hint matrix is passed to the discriminator with many 0s and 1s. It is just like the discriminator has been told most answers to the questions: 0 is missing data, 1 is observed data, and should focus on the unknown answers (represented by 0.5, and it could be either missing data or observed one) by learning from known answers. Thus, data represented by 0.5 are what discriminator should distinguish by itself. After iterations, discriminator finally understands the distribution of the data piece by piece and outputs a very satisfactory result no matter how large the amounts of data are missing. \mathbf{H} can be specified with different amount of known information of \mathbf{M} and enables the discriminator to concentrate on the given hints and improves the quality of imputation according to the generator.

Objective function. In the imputation architecture of the GAN, the discriminator is trained to maximize the accuracy of $\hat{\mathbf{M}}$, and the generator is trained to minimize the accuracy of the discriminator predicting $\hat{\mathbf{M}}$. This adversarial training process ensures the imputation quality. Technically, it is still a binary classification problem that should be solved. Therefore, binary cross entropy is adopted to score the distance between the predicted probabilities and actual

labels, which is either 0 or 1. As the discriminator outputs the estimated mask matrix with the aid of hint matrix, the objective function of the GAN architecture can be written as follows

$$\min_G \max_D \mathbb{E}_{\hat{\mathbf{X}}, \mathbf{M}, \mathbf{H}} \left[\mathbf{M}^T \log \hat{\mathbf{M}} + (1 - \mathbf{M})^T \log (1 - \log \hat{\mathbf{M}}) \right] \quad (6)$$

where \log represents the element-wise logarithm operation and $\hat{\mathbf{M}}$ is the predicted mask matrix, which can be denoted as $\hat{\mathbf{M}} = D(\hat{\mathbf{X}}, \mathbf{H})$. Also, considering the primitive binary cross entropy loss function $\mathcal{L}(\mathbf{x}, \mathbf{y})$ is formally defined as the negative expectation of the log of corrected predicted probabilities. It has the form of

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n [x_i \log(y_i) + (1 - x_i) \log(1 - y_i)] \quad (7)$$

Then, equation (6) can be rewritten and simplified as

$$\min_G \max_D \mathbb{E} \left[\mathcal{L}(\mathbf{M}, \hat{\mathbf{M}}) \right] \quad (8)$$

Algorithm implementation and verification

In order to validate the effectiveness of the GAN imputation framework, some numerical experiments are conducted, including single-sensor imputation and multiple-sensor imputation. The target structure is a five-span continuous concrete box girder bridge with 25 m long for each span. The investigated 15 sensors are demonstrated in Figure 2 (sensors are only installed on the mid-span section). In this article, imputation performance for several different loss ratios (ranging from 1 h to 23 h) is investigated. As mentioned before, conventional imputation methods for sensor reconstruction often require complete dataset which is often a troubling issue when selecting complete dataset. In our study, only incomplete dataset is needed and the missing values are imputed directly based on remaining observed data in the framework.

Preparation of dataset

In the modern SHM system, vast amounts of data are generated every day, which accord with the requirements of deep neural networks: the more data are used, the better imputation result will be obtained. In our case, data collected in 1 day from 15 sensors are investigated to implement the imputation and the sampling frequency of the strain sensors is 1 Hz. Considering the fully observed data are only for comparison and validating the effectiveness, we manually appoint the continuous random missing values on the complete dataset by utilizing the mask matrix mentioned above.

Before feeding the incomplete dataset into the GANs, the preprocessing of the data is also necessary for improving the imputation efficiency and evaluating the imputation performance under the same metric. The strain data are first processed by min–max normalization, which maps the data into 0 to 1. The normalization equation is as follows

$$\text{strain}_{\text{normalized}} = \frac{\text{strain} - \min(\text{strain})}{\max(\text{strain}) - \min(\text{strain})} \quad (9)$$

The field monitoring data chosen in 1 day are used to demonstrate the proposed algorithm. For single-sensor imputation, continuous missing data over 1 h, 2 h, 4 h, 8 h, 16 h, and 23 h are, respectively, investigated. For multiple-sensor imputation, random missing data (with different time intervals taken) over 1 hour for 5 sensors, 10 sensors, and 15 sensors are, respectively, investigated. The incomplete dataset will be fed into the neural network after preprocessing. During the network, the missing component is indexed by the mask matrix and automatically detected

by the network; after the adversarial unsupervised training, the network outputs the full dataset with observed data unchanged and missing data imputed. Actually, each sensor represents one dimension in the network, and the network implicitly captures the relationships between sensors and then outputs the missing components. Hence, the GAN-based algorithm is very applicable to imputation task in real time which perfectly meets the requirements of bridge managers and stakeholders.

Single-sensor imputation

The proposed GAN imputation framework is implemented in TensorFlow. The workstation used is configured with two Intel Xeon(R) E5-2696 v4 CPUs, a 256 GB memory, and an NVIDIA TITAN X (Pascal) GPU. The training process is conducted through GPU acceleration.

Only for illustration purpose, strain data collected from Sensor S5 on 9 April 2017 are employed. As can be seen from Figure 7, 1-h data are missing in Figure 7(b) which is represented by the vacancy, and the ground truth is depicted in Figure 7(a). The aim of the study is to impute the 1-h missing data of Sensor S5 based on the remaining data of all 15 sensors. As mentioned above, the complete dataset for training is unnecessary and the ground truth data are only for comparison purpose.

The 1-h missing data imputation result is depicted in Figure 8(b), and the results show an excellent performance of the algorithm, and the imputed data are superimposed on the ground truth except for few peaks. In Figure 8(c), the details about the imputation result are shown: the stationary section represents the temperature-induced strain and the

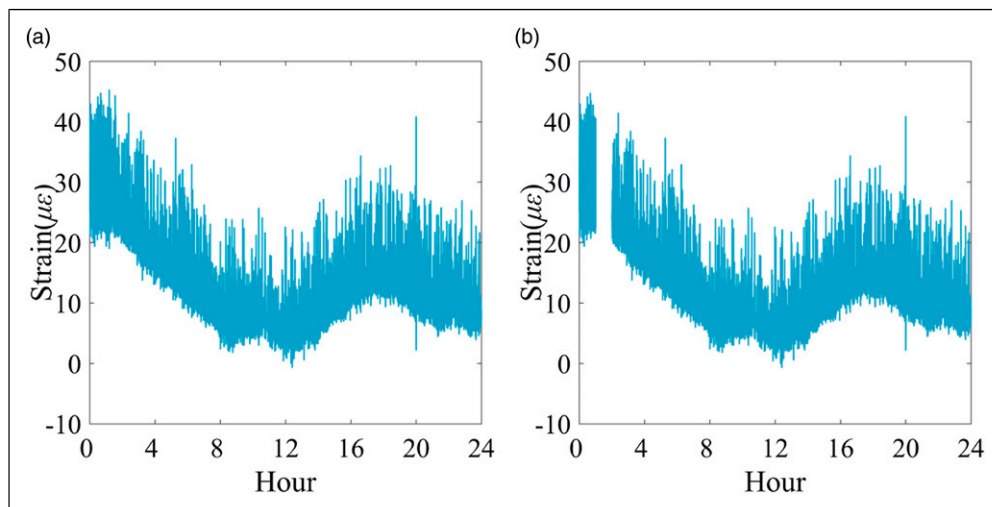


Figure 7. Investigated strain sensor S5 for demonstrating generative adversarial network–based imputation method, and the blank space represents the missing data: (a) fully observed data and (b) 1-h missing data.

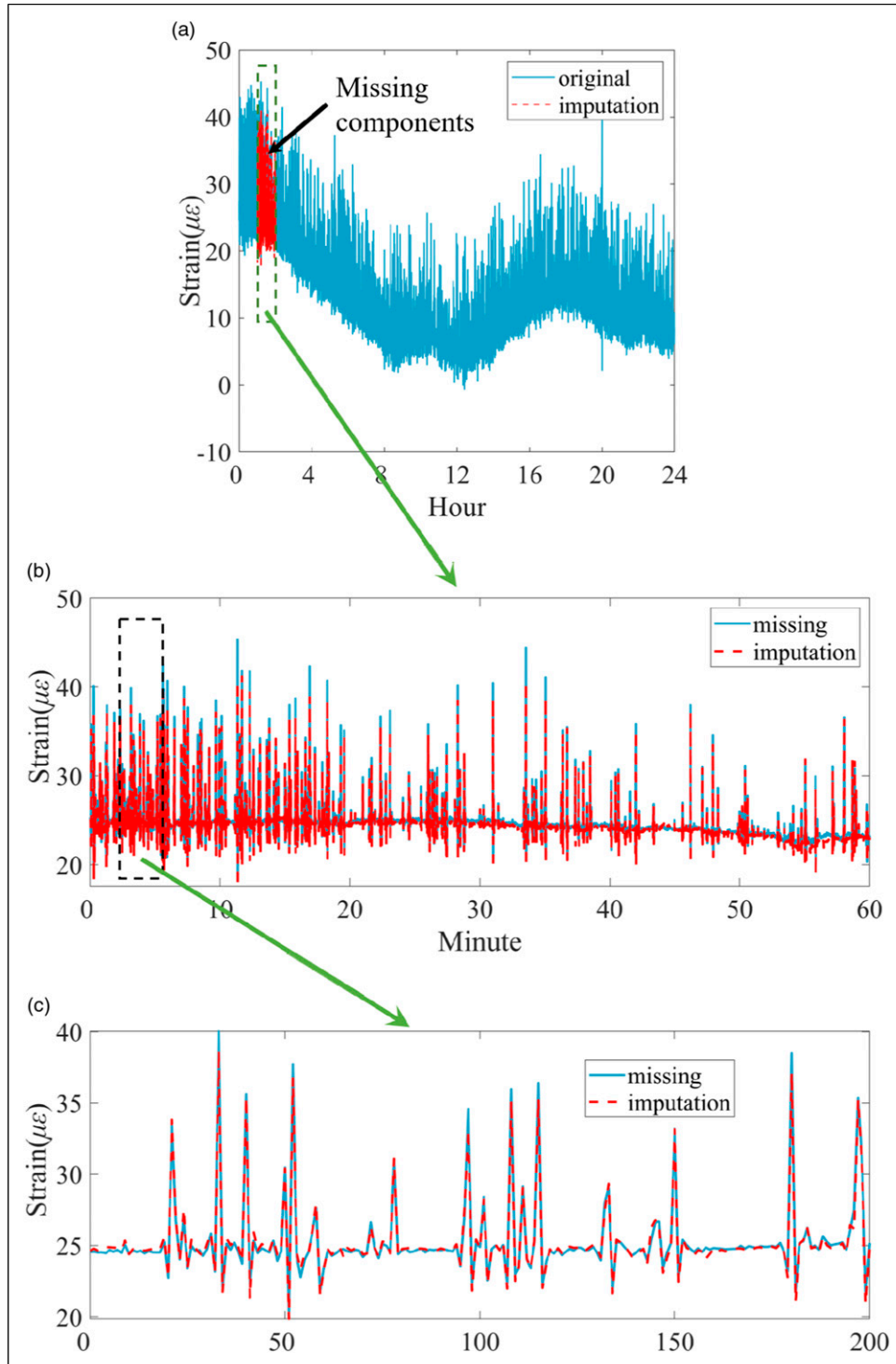


Figure 8. Results of imputation for investigated strain sensor S5. (a) Overall imputation result, (b) I-h imputation result, and (c) details about imputation result.

nonstationary section represents the vehicle-induced strain. As demonstrated, the vehicle-induced strain data are perfectly imputed, which means the GAN framework actually captures the inter-sensor relationships distributed at

different locations even for vehicle-induced response. It is noted that the sampling frequency for strain sensor is 1 Hz, which may not capture the whole vibration characteristics caused by the moving vehicles. However, the dynamic

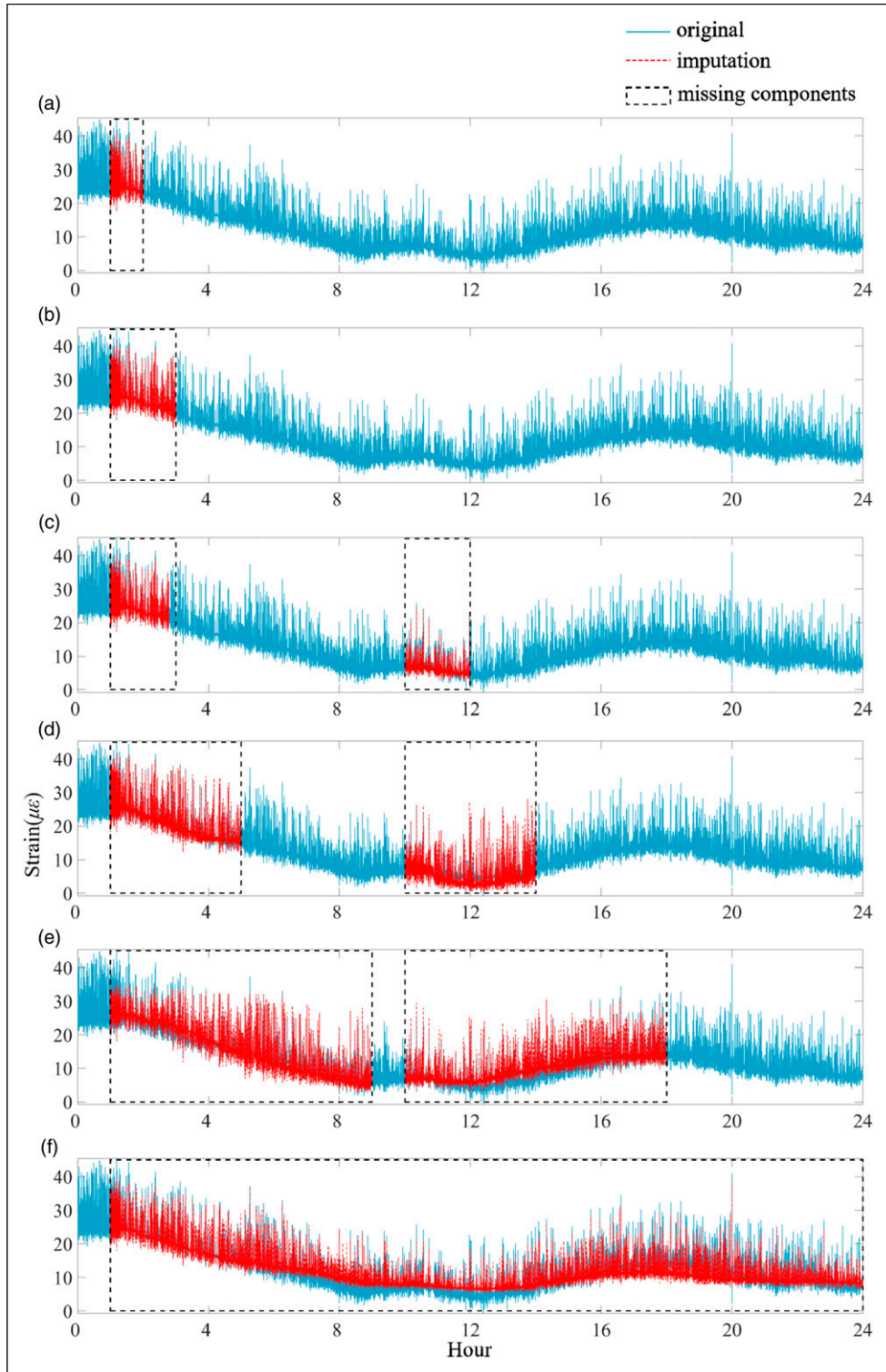


Figure 9. Results of imputation for different loss ratio of sensor S5: (a) 1-h missing, (b) 2-h missing, (c) 4-h missing, (d) 8-h missing, (e) 16-h missing, and (f) 23-h missing.

responses caused by vehicles (represented by spikes) can still be easily distinguished from the static responses (represented by the stationary parts). In this sense, it still can capture the (quasi) dynamic responses both for GAN and physical reality. As mentioned previously, conventional methods often require very delicate model to impute the dynamic response caused by the vehicles. Therefore, GAN-based imputation algorithm is capable of imputing dynamic response in real time. The final outputted data are shown in Figure 8(a) with observed data unchanged and missing data imputed properly.

This article also investigates the imputation result for missing components ranging from 1 h to 23 h. The results are demonstrated in Figure 9, respectively. In the cases of 1-, 2-, and 23-h imputation, the missing data form into one continuous missing part; while in the cases of 4-, 8-, 16-h imputation condition, the missing data are divided into two

separate continuous parts in order to generalize GAN's ability in multiple missing parts imputation. It is noted that from Figures 9(a) to (d), which indicate 1-h missing to 8-h missing, the GAN framework generates quite excellent imputation results. The imputation signals almost overlap the original signals with very limited deviation. When the loss ratio reaches up to 16 h and 23 h (Figures 9(e) and (f)), the error does exist and some peaks and valleys fail to be captured. However, the imputation accuracy is also within in a tolerable extent. The trend and the nonstationary parts are mostly imputed with high accuracy, and the amplitude deviates around the ground truth in a satisfying threshold. The quantitative imputation errors will be analyzed later.

In order to clarify the imputation performance under small time scale, the imputation details of 200 sampling points (200 s) for missing data ranging from 1 h to 23 h are demonstrated in Figure 10. As demonstrated, when the data

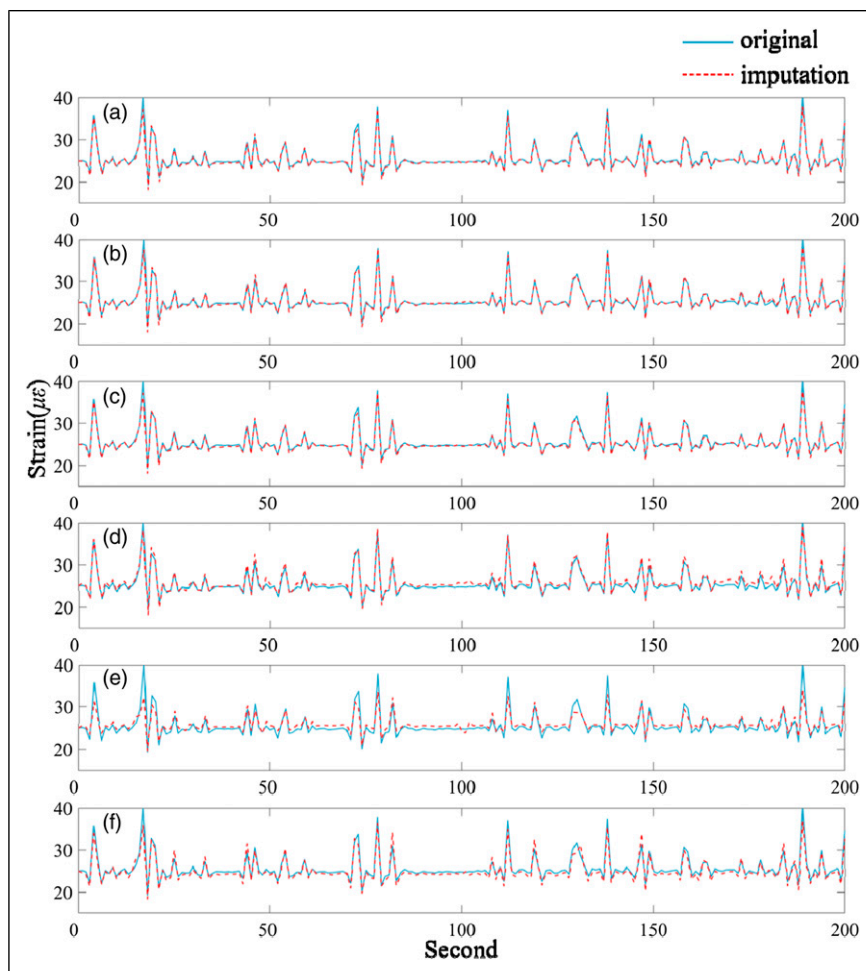


Figure 10. Results of imputation details of sensor S5 for: (a) 1-h missing, (b) 2-h missing, (c) 4-h missing, (d) 8-h missing, (e) 16-h missing, and (f) 23-h missing.

lack less than 8 h, the GAN framework provides outstanding results with both stationary and nonstationary parts carefully recovered. The errors are even too subtle to be identified. For the data under 16- and 23-h missing ratio, errors are relatively much bigger compared with above scenarios, but still are very small. The recovered responses also vary with limited extent and can be tolerated for both academic research and engineering application.

The GAN framework imputes the missing components based on the observed data and trains the network to obtain the mask matrix rather than the measured strain data (see equation (4)). Thus, GAN framework does not need complete and intact dataset to establish a baseline model and is able to impute missing parts with any incomplete data fed into. The training process of six sensors is demonstrated in Figure 11. Apparently, for each sensor, the loss value

decreases dramatically in first 200 epochs, and then stabilizes till the training finishes. The training process is accelerated by GPU: TITAN X, and 1000 epochs are completed within 150 s. In this case, we are able to impute the missing data in real time and acquire higher accurate result for later analysis.

In the field monitoring data, strain sensors are distributed at different locations and data values are often not in the same magnitude. So, in order to evaluate and comprehend the imputation errors, the data are first pre-processed by normalization (see equation (9)). It is noted that the imputation results shown above have been re-normalized to the origin scale. In the following error analysis, the strain data are mapped into $0 - 1 \mu\epsilon$, and two metrics are used to evaluate the imputation errors, namely, root mean square error (RMSE) performance and L2-norm

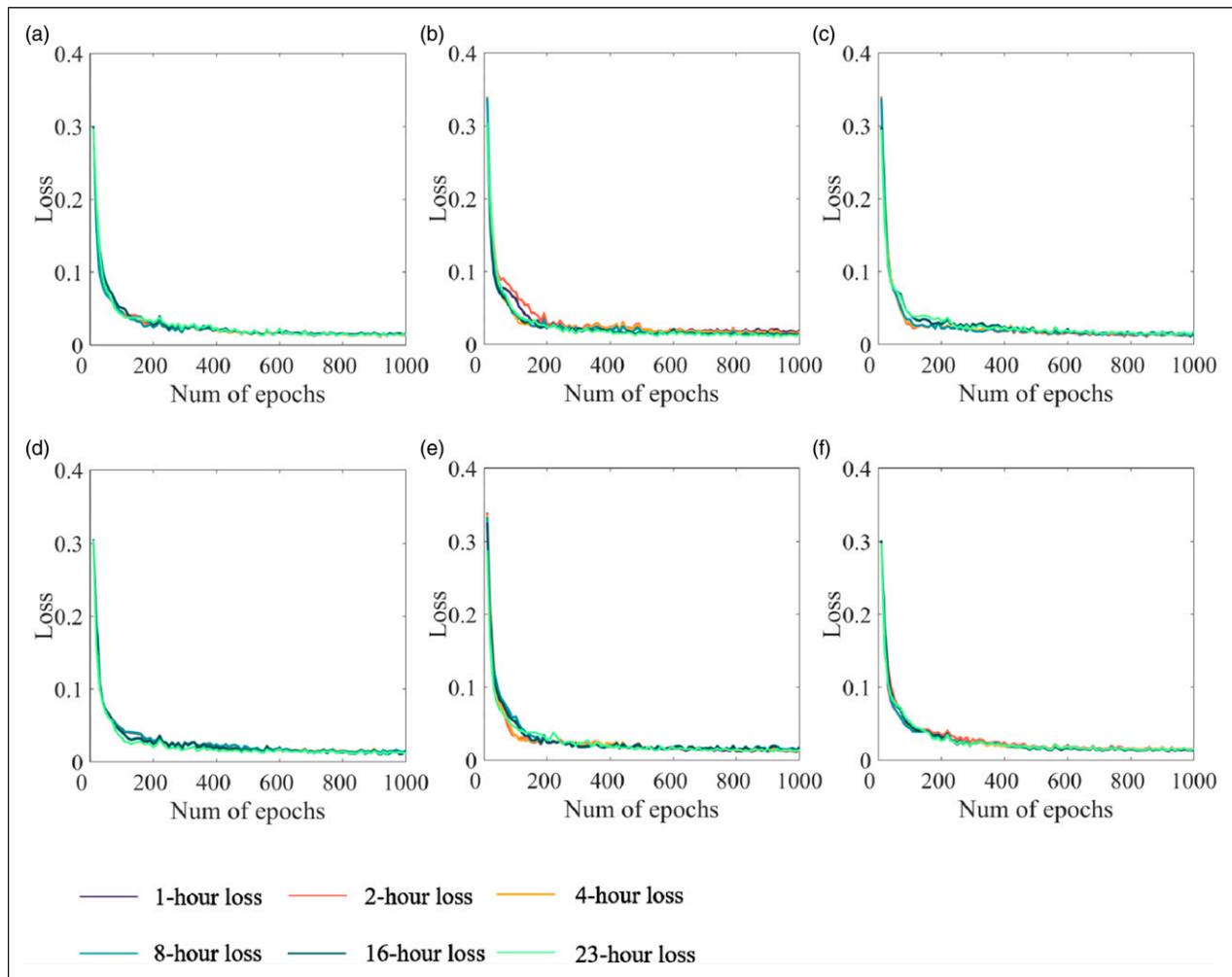


Figure 11. Training loss over 1000 epochs for: (a) sensor S1, (b) sensor S2, (c) sensor S3, (d) sensor S4, (e) sensor S5, and (f) sensor S6.

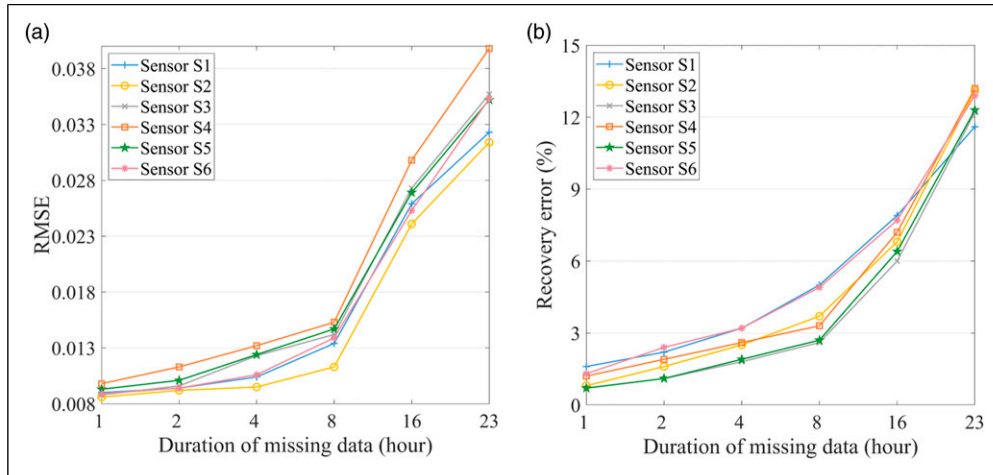


Figure 12. RMSE performance and recovery error between the ground truth data and imputation data: (a) RMSE performance over 23 h and (b) recovery error over 23 h. RMSE: root mean square error.

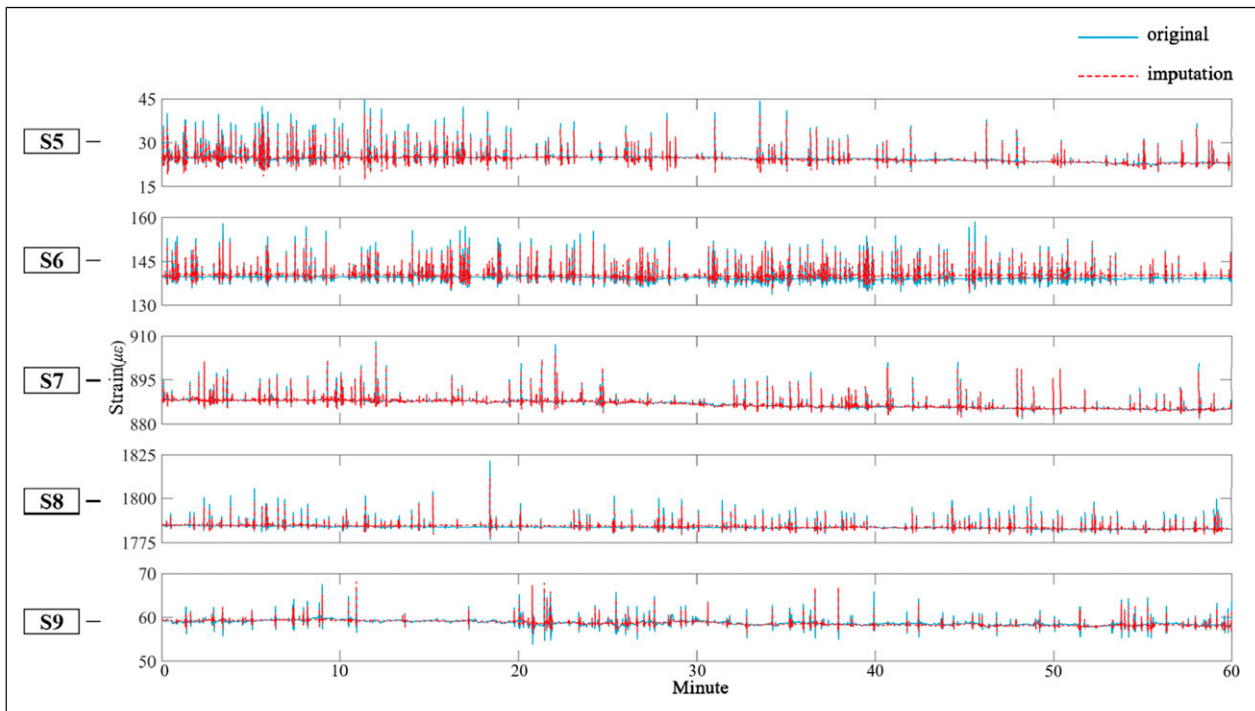


Figure 13. Results of 5-sensor imputation over 1-h continuous missing data.

error. It is noted that because of baseline shifting (which is common in in situ monitoring data, like S7 and S8), the metric “error” will become very small when mapped back and loses its meaning (although is physically meaningful). Also, RMSE is tightly related to the magnitude (amplitude variation) of the original data, while all sensor readings vary in a unique magnitude, which makes it hard to be compared in the original scale. Such metrics are, therefore, calculated in the normalized scale rather than the

Table 1. RMSE performance and recovery error for 5-sensor imputation over 1-h continuous missing data.

Sensor	RMSE	Error (%)	Sensor	RMSE	Error (%)
S5	0.0097	1.7	S8	0.0148	6.5
S6	0.0197	3.1	S9	0.0081	4.3
S7	0.005	2.9			

RMSE: root mean square error.

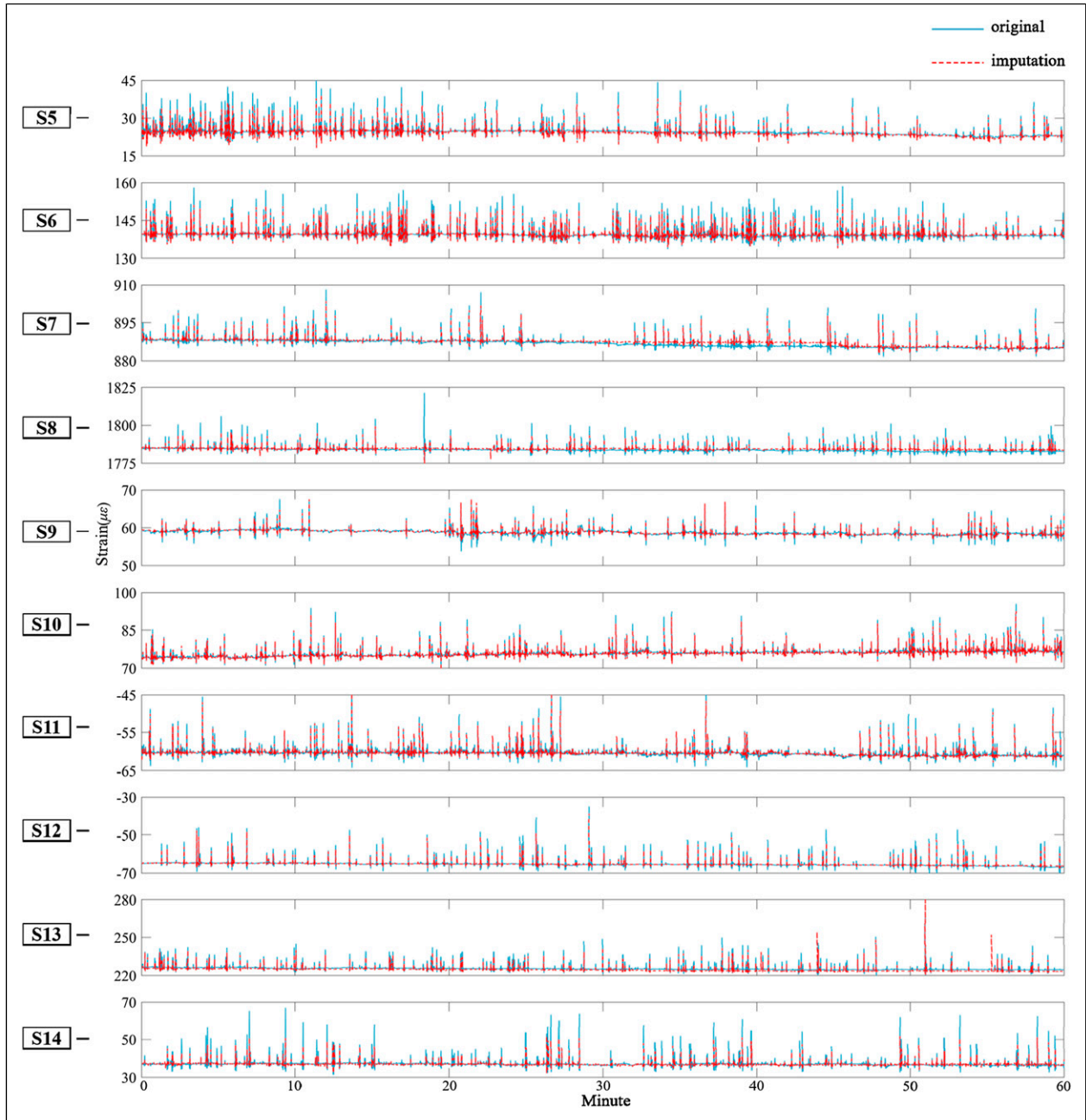


Figure 14. Results of 10-sensor imputation over 1-h continuous missing data.

original scale. The two metrics are, respectively, calculated as follows

$$RMSE = \sqrt{\frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{m}} \quad (10)$$

$$Error = \frac{\|y - \hat{y}\|}{\|y\|} = \frac{\sqrt{\sum_{i=1}^m (y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^m y_i^2}} \quad (11)$$

Table 2. RMSE performance and recovery error for 10-sensor imputation over 1-h continuous missing data.

Sensor	RMSE	Error (%)	Sensor	RMSE	Error (%)
S5	0.0154	6.8	S10	0.0091	3.6
S6	0.0238	7.2	S11	0.0080	5.7
S7	0.0134	7.8	S12	0.0095	8.8
S8	0.0262	12.4	S13	0.0214	13.7
S9	0.0070	3.7	S14	0.0282	16.0

RMSE: root mean square error.

where y_i represents the ground truth data, \hat{y}_i represents the imputed data, and m is the number of missing points. The corresponding RMSE performance and recovery error are illustrated in Figure 12. When the missing data are less than

8 h, the GAN framework generates small RMSE loss and recovery error with only about $0.015 \mu\epsilon$ deviation and 5% imputation error. Although a sharp increasing exists after 8 h, in consistent with above intuitive analysis, the RMSE

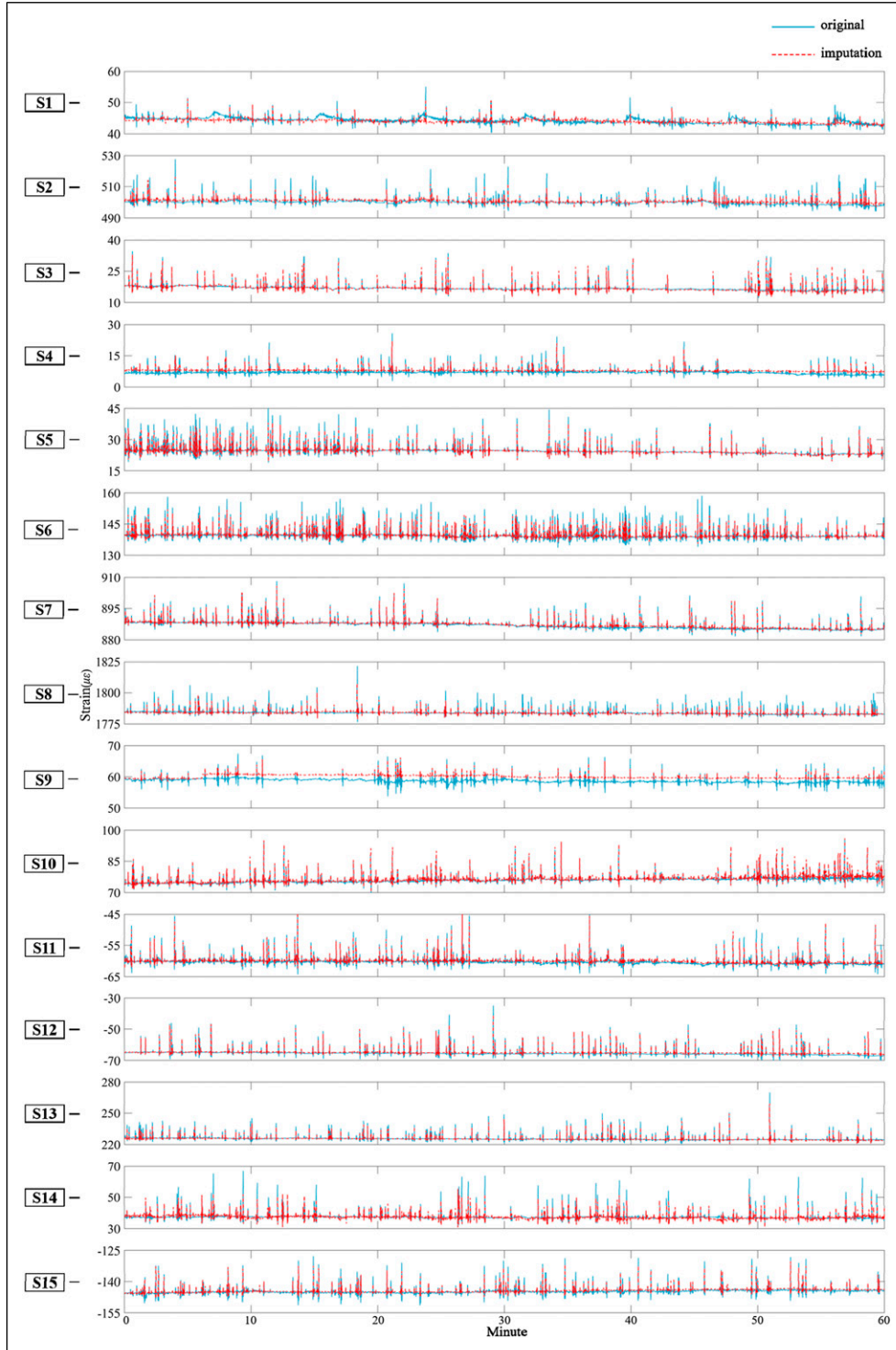


Figure 15. Results of 15-sensor imputation over 1-h continuous missing data.

loss is still within $0.04 \mu\epsilon$ and recovery error is controlled in 15% which is tolerable for such vast missing ratio.

Multiple-sensor imputation

As is often the case in SHM, data loss problem happens among multiple sensors rather than only single one, and sometimes sensors share the same missing components, which makes the imputation much more challenging. In this article, random 1-h missing data for 5, 10, and 15 sensors (with different time intervals taken) are investigated to evaluate the proposed GAN imputation framework.

Results of sensor S5 to S9 imputation performance over 1-h continuous missing data are illustrated in Figure 13: All the sensors share the successful imputation result, and the imputed signals almost overlap the original ones except for some countable peaks that are hard to reach. For sensor S5 and S6, there are dense nonstationary parts caused by ambient excitations like vehicles in 1 h, and the GAN framework successfully captures these excitations and imputes the result with high accuracy.

The RMSE performance and recovery error based on L2-norm are listed in Table 1. Similarly, these metrics are also evaluated after the data being normalized to $0 - 1 \mu\epsilon$. As can be seen from the table, the RMSE losses are under $0.02 \mu\epsilon$, and the recovery error is under 7% for each sensor, a comparatively big increase compared with single-sensor imputation under the same condition (as shown in Figure 12). However, the result will not impede the academic research because it still provides quite satisfactory imputation performance.

Figure 14 demonstrates the imputation results for sensor S5 to S14 over 1-h missing data. Compared with 5-sensor imputation above, the amplitudes of nonstationary parts are imputed with smaller values, especially for sensor S5, S6, S8, and S14. The GAN framework provides very similar waveform and pattern of the vehicle-induced strain data despite with some minor peaks compared with the ground truth ones. The corresponding RMSE and recovery errors are listed in Table 2. Several RMSE losses have exceeded $0.02 \mu\epsilon$ and errors exceeded 10%; however, the majority of the losses are controlled in good limits. Generally, the results for 10-sensor imputation are still workable and within the tolerable extent.

The imputation results for 15 sensors that lack 1-h data (with different time intervals taken) are depicted in Figure 15. Some severe problems occur: for sensor S2, S5, S6, S8, and S14, more amplitudes are hard to be imputed with high accuracy; for S9 and the end part of S4, the stationary parts have been deviated from the baseline; for S9, only the peaks are imputed and the valleys of the signals are ignored; for S1, the periodic bumps of the signal are not able to be imputed due to its own characteristics of the sensor or other abnormal types. The corresponding RMSE

and recovery error performances are demonstrated in Table 3, where the majority of the RMSE losses are within the extent of $0.025 \mu\epsilon$ and the recovery errors within 10%; several RMSE losses and recovery errors exceeds $0.025 \mu\epsilon$ and 10%, respectively; for sensor S9, there is a baseline divergence between the imputed signals and the original ones, so the RMSE loss is over $0.03 \mu\epsilon$ and the error is over 20%. Overall, the imputation results that GAN framework generate are acceptable considering $0.025 \mu\epsilon$ loss and 10% error. Thus, for all sensors that lose 1-h data, the proposed deep learning network successfully meets the requirements of missing data imputation.

Discussion

In the above examples, considering the burden of data storing, transmitting, and computing for the long-term bridge monitoring, sampling frequency of 1 Hz is adopted and data

Table 3. RMSE performance and recovery error for 15-sensor imputation over 1-h continuous missing data.

Sensor	RMSE	Error (%)	Sensor	RMSE	Error (%)
S1	0.0215	9.2	S9	0.0336	21.9
S2	0.0271	8.1	S10	0.0123	4.8
S3	0.0236	3.0	S11	0.0102	7.2
S4	0.0301	9.2	S12	0.0107	9.8
S5	0.0222	7.2	S13	0.0219	11.6
S6	0.0159	12.5	S14	0.0280	14.9
S7	0.0070	4.1	S15	0.0155	6.2
S8	0.0204	11.9			

RMSE: root mean square error.

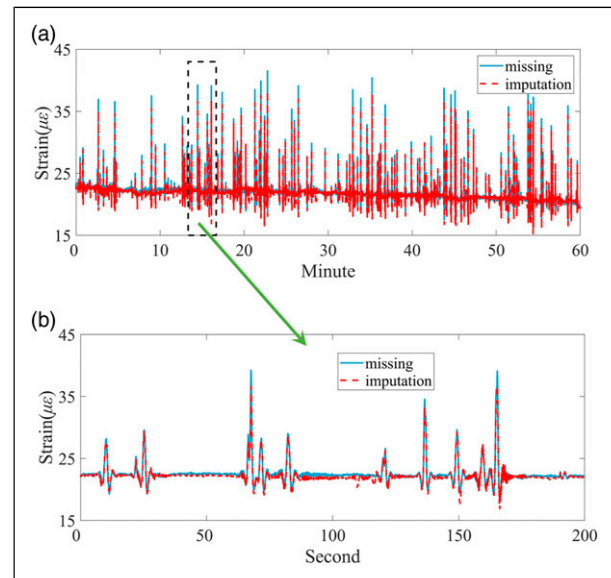


Figure 16. Imputation result for 1-h missing data at frequency of 10 Hz: (a) imputation result for 1-h missing data; (b) details.

imputation has been discussed and demonstrated. Even though the sampling frequency is fairly low, we are still able to distinguish the dynamic responses caused by vehicles (represented by spikes) from the static responses (represented by the stationary parts), which shows that it still could capture the (quasi) dynamic responses both for GAN and physical reality. In this case, it takes only 150 s to train and generate complete datasets, which is quite prompt and efficient. It should be noted that the proposed method can also be applied to higher frequency cases only with more training epochs and time taken. From the view of the signal processing itself, the algorithm just deals with such a series of the data points and higher sampling frequency merely means more data points for a certain period of time. An example of imputation for 1-h missing data with higher sampling frequency of 10 Hz is demonstrated in Figure 16. As is obviously shown, the dynamic responses caused by vehicles and static responses caused by temperature are both imputed with high accuracy. However, the imputation takes about 12 min (4000 epochs) to process and lacks promptness in the practical application albeit it implies that GAN performs well for both lower and higher sampling frequency situation.

Conclusion

This article presents a novel data-driven approach based on GANs for imputing missing data in SHM field. The theory and the architecture of the network are discussed in detail. The main characteristics and advantages of the proposed approach are (1) avoid delicate model selection and parameter settings; (2) only require incomplete dataset; (3) avoid expert knowledge in engineering mechanics or bridge engineering; and (4) imputation in real time with high accuracy and efficiency. The applicability of the method is conducted on a real concrete bridge in China with 15 strain sensors located at different positions considered. Single-sensor imputation and multiple-sensor imputation are both investigated. These imputations are based on the observed data through the GAN framework proposed.

For single-sensor imputation, missing data ranging from 1 h to 23 h are investigated, and the experiments demonstrate the excellent performance of the GAN imputation framework. Although with the increasing of the number of the missing data, the errors increase, the accuracy is controlled in a very satisfactory limit with RMSE loss under $0.04\mu\epsilon$ and recovery error under 15% in the 23-h missing data situation. For multiple-sensor imputation, 1 h missing data for 5, 10, and 15 sensors are investigated. The results show that the presented method also provides acceptable and decent imputation values, although compared with the single-sensor imputation, the errors have increased and other problems occur for some individual sensors. It can be concluded that the GAN framework can be used to impute

the incomplete field monitoring data in real time, especially for single sensor and can also be adopted to impute multiple sensors with satisfactory results.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Fund for Distinguished Young Scientists of Jiangsu Province (BK20190013); Post-graduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_0113).

ORCID iD

Chunfeng Wan  <https://orcid.org/0000-0002-4236-6428>
Youliang Ding  <https://orcid.org/0000-0002-0774-426X>

References

1. Li H and Ou J. The state of the art in structural health monitoring of cable-stayed bridges. *J Civ Struct Health Monit* 2016; 6: 43–67.
2. Li H, Ou J, Zhang X, et al. Research and practice of health monitoring for long-span bridges in the mainland of China. *Smart Struct Syst* 2015; 15: 555–576.
3. Bao Y, Chen Z, Wei S, et al. The state of the art of data science and engineering in structural health monitoring. *Engineering* 2019; 5: 234–242.
4. Azimi M, Eslamlou A and Pekcan G. Data-driven structural health monitoring and damage detection through deep learning: state-of-the-art review. *Sensors* 2020; 20(10): 2778.
5. Wang YW, Ni YQ, Zhang QH, et al. Bayesian approaches for evaluating wind-resistant performance of long-span bridges using structural health monitoring data. *Struct Control Health Monit* 2021; 28(4): e2699.
6. Ni YQ, Xia HW, Wong KY, et al. In-service condition assessment of bridge deck using long-term monitoring data of strain response. *J Bridge Eng* 2012; 17(6): 876–885.
7. Li SL, Li H, Ou JP, et al. Integrity strain response analysis of a long span cable-stayed bridge. *Key Eng Mater* 2009; 413–414: 775–783.
8. Chen C, Wang ZL, Wang YH, et al. Reliability assessment for PSC box-girder bridges based on SHM strain measurements. *J Sensors* 2017; 2017: 8613659.
9. Huang HB, Yi TH, Li HN, et al. Strain-based performance warning method for bridge main girders under variable operating conditions. *J Bridge Eng* 2020; 25(4): 04020013.
10. Wan HP and Ni YQ. Bayesian modeling approach for forecast of structural stress response using structural health monitoring data. *J Struct Eng* 2018; 144(9): 04018130.
11. Jiang H, Wan C, Yang K, et al. Modeling relationships for field strain data under thermal effects using functional data analysis. *Measurement* 2021; 177: 109279.

12. Catbas FN, Susoy M and Frangopol DM. Structural health monitoring and reliability estimation: long span truss bridge application with environmental monitoring data. *Eng Struct* 2008; 30(9): 2347–2359.
13. Oh BK, Glisic B, Kim Y, et al. Convolutional neural network-based data recovery method for structural health monitoring. *Struct Health Monit* 2020; 19(6): 1821–1838.
14. Chen Z, Li H and Bao Y. Analyzing and modeling inter-sensor relationships for strain monitoring data and missing data imputation: a copula and functional data-analytic approach. *Struct Health Monit* 2019; 18(4): 1168–1188.
15. Huang YW, Wu DG, Liu ZH, et al. Lost strain data reconstruction based on least squares support vector machine. *Meas Control Technology* 2010; 29(8): 8–12.
16. Huang YW, Wu DG and Li J. Structural healthy monitoring data recovery based on extreme learning machine. *Comput Eng* 2011; 37(16): 241–243.
17. Lu W, Teng J, Li C, et al. Reconstruction to sensor measurements based on a correlation model of monitoring data. *Appl Sci* 2017; 7(3): 243.
18. Zhang Z and Luo Y. Restoring method for missing data of spatial structural stress monitoring based on correlation. *Mech Syst Signal Process* 2017; 91: 266–277.
19. Wan HP, Dong GS and Luo YZ. Compressive sensing of wind speed data of large-scale spatial structures with dedicated dictionary using time-shift strategy. *Mech Syst Signal Process* 2021; 157: 107685. DOI: [10.1016/j.ymssp.2021.107685](https://doi.org/10.1016/j.ymssp.2021.107685).
20. Fan G, Li J and Hao H. Lost data recovery for structural health monitoring based on convolutional neural networks. *Struct Control Health Monit* 2019; 26(10): e2433.
21. Jeong S, Ferguson M, Hou R, et al. Sensor data reconstruction using bidirectional recurrent neural network with application to bridge monitoring. *Adv Eng Inform* 2019; 42: 100991.
22. Perez-Ramires CA, Amezquita-Sanchez JP, Valtierra-Rodriguez M, et al. Recurrent neural network model with Bayesian training and mutual information for response prediction of large buildings. *Eng Struct* 2019; 178: 603–615.
23. Fan G, Li J and Hao H. Dynamic response reconstruction for structural health monitoring using densely connected convolutional networks. *Struct Health Monit* 2020: 1475921720916881. doi: [10.1177/1475921720916881](https://doi.org/10.1177/1475921720916881)
24. Wan H-P and Ni Y-Q. Bayesian multi-task learning methodology for reconstruction of structural health monitoring data. *Struct Health Monit* 2019; 18(4): 1282–1309.
25. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 27, 2014, pp. 2672–2680.
26. Pan Z, Yu W, Yi X, et al. Recent progress on generative adversarial networks (GANs): a survey. *IEEE Access* 2019; 7: 36322–36333.
27. Allen A and Li W. Generative adversarial denoising autoencoder for face completion, 2016. https://www.cc.gatech.edu/~hays/7476/projects/Avery_Wenchen/
28. Yoon J, Jordan J, Schaar M, et al. GAIN: missing data imputation using generative adversarial nets. In: *Proceedings of the 35th international conference on machine learning*, Stockholmsmässan, Stockholm, Sweden, 2018, pp. 5689–5698.
29. Li S, Jiang B and Marlin B. MisGAN: learning from incomplete data with generative adversarial networks. In: *Proceedings of the 7th international conference on learning representations*, New Orleans, USA, 2019.
30. Yoon S and Sull S. GAMIN: generative adversarial multiple imputation network for highly missing data. In: *Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, 2020, pp. 8453–8561.
31. Kazemi A and Meidani H. IGANI: iterative generative adversarial networks for imputation applied to prediction of traffic data, 2020, <https://arxiv.org/abs/2008.04847>
32. Fan G, Li J, Hao H, et al. Data driven structural dynamic response reconstruction using segment based generative adversarial networks. *Eng Struct* 2021; 234: 111970.
33. Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003; 3(4–5): 993–1022.
34. Smolensky P. *Information processing in dynamical systems: foundations of harmony theory*. Cambridge, MA: MIT Press, 1986.
35. Hinton GE, Osindero S and Teh Y-W. A fast learning algorithm for deep belief nets. *Neural Comput* 2006; 18(7): 1527–1554.
36. Nair V and Hinton GE. Rectified linear units improve restricted boltzmann machines. In: *International conference on machine learning*, Haifa, Israel, 2010.
37. Goodfellow I, Bengio Y and Courville A. *Deep learning*. Cambridge, MA: MIT Press, 2016.