文章编号：1000-3673（2024）07-2784-11 中图分类号：TM 721 文献标志码：A 学科代码：470·40

基于双向循环插补网络的分布式光伏集群时序数据耦合增强方法

廖若愚[1]，刘友波[1]，沈晓东[1]，高红均[1]，唐冬来[2]，刘俊勇[1]
（1. 四川大学电气工程学院，四川省 成都市 610065；
2. 四川中电启明星信息技术有限公司，四川省 成都市 610067）

Time Series Data Coupling Enhancement Method of Distributed Photovoltaic Cluster Based on Bidirectional Recurrent Imputation Network

LIAO Ruoyu[1], LIU Youbo[1], SHEN Xiaodong[1], GAO Hongjun[1], TANG Donglai[2], LIU Junyong[1]
(1. School of Electrical Engineering, Sichuan University, Chengdu 610065, Sichuan Province, China;
2. Aostar Information Technologies Co., Ltd., Chengdu 610067, Sichuan Province, China)

ABSTRACT: Distributed photovoltaic systems are widely distributed, with high local penetration rates and complex and ever-changing installation environments. Reliable measurement data is the basis for performance analysis, output prediction, and operation and maintenance control. However, factors such as sensor failures and communication blockages can lead to missing measurement values, deteriorating the quality of raw data, and thus affecting the accuracy of distribution network operation decision-making. Traditional data repair methods only consider the distribution characteristics of a single measurement value, ignoring the coupling relationship of multidimensional time series data, resulting in low repair accuracy. A coupled data enhancement method based on a bidirectional multi-stage recurrent imputation network is proposed to address this issue. Experimental results demonstrate that the proposed method exhibits good repair performance even under high levels of missing data, effectively enhancing the quality of fundamental data for distributed photovoltaic clusters, and improving the fine-grained perception capability of grid operators towards photovoltaic clusters.

KEY WORDS: missing data imputation; bidirectional recurrent imputation network; coupled time-series data; distributed photovoltaic cluster

摘要：分布式光伏点多面广、局部渗透率高、安装环境复杂多变，真实可靠的量测数据是其性能分析、出力预测、运维调控的基础。然而，传感器故障和通信堵塞等因素会造成量测值缺失，恶化原始数据质量，进而影响配电网运行决策的准确性。传统数据修复方法只考虑单一量测值的分布特征，忽略了多维时序数据的潜在耦合关系，修复精度有限。为此，该文提出一种基于双向多阶段循环插补网络和 Seq2Seq-Attention 的时序数据耦合增强方法，改进了循环插补网络的结构，并引入衰减机制，能利用少量未缺失数据，潜在地挖掘原始数据的整体分布规律，一次性对多个光伏场站完成高质量数据修复。实验结果表明，所提方法在高比例缺失情况下仍有良好的修复性能，可明显增强分布式光伏集群的基础数据质量，提升电网运营商对光伏集群的细粒度感知能力。

关键词：缺失数据修复；双向循环插补网络；耦合时序数据；分布式光伏集群

DOI：10.13335/j.1000-3673.pst.2023.1681

0 引言

近年来，光伏发电迅速发展，2022 年，全国光伏新增并网容量为 87.4GW，其中分布式光伏新增 51.1GW，占比为 58.5%[1]。高渗透率的分布式光伏会改变配电网单向潮流分布，造成功率倒送、电压越限等问题，影响供电质量[2]。同时，光伏的出力特性也会导致源荷两端存在较大不确定性[3]，给系统运行与规划带来了挑战。

真实可靠的分布式光伏量测数据可为配电网电压控制[4]、分布式光伏消纳交易策略[5]、光伏出力预测以及配网精准态势感知[6]等研究提供数据支撑。先进的控制策略通常需要及时准确的量测数据反映配电网的状态信息。然而，分布式光伏数据采集和传输环节错综复杂，传感器故障、通信阻塞、时间同步异常、数据存储等问题会导致数据缺失[7]。准确、完整的监测数据是电力系统实现信息精准感

基金项目：国家自然科学基金项目(51977133)。
Project Supported by the National Natural Science Foundation of China (51977133).

知、对分布式光伏“可观、可控”的关键[8]。低质量的数据无法准确反映分布式光伏场站的真实状态，会影响配电网运行决策的准确性。因此，对光伏出力数据进行精准修复具有重要意义。

已有缺失值修复方法通常分为两类：数理统计学方法、机器学习方法。前者包括均值法、线性回归法和样条插值法[9]等，这类方法具有局限性，对数据的线性性质以及平滑程度有严格要求。随着人工智能快速发展，基于机器学习的缺失数据插补方法已得到广泛应用。传统机器学习方法包括 BP 神经网络、缺失森林(miss forest，MF)、K-最近邻(K-Nearest Neighbor，KNN)[10]等。然而上述方法均存在数据利用率低和插补精度有限的问题。

在电力系统领域，以生成对抗网络(generative adversarial networks，GAN)为代表的深度生成模型，在数据修复问题上表现出了良好的性能[11]。文献[12]在 GAN 中引入双重语义感知损失函数，有效提升了台区缺失功率数据的插补精度。此外，GAN 还应用于 PMU 量测数据[13]和电动汽车充电负荷数据[14]的修复。传统的 GAN 存在两个缺点：1）生成器只输入随机噪声，忽略了原始量测值中的有效信息；2）鉴别器需要完整的数据样本进行对抗训练，而实际缺失场景中几乎没有连续完整的量测数据可供模型训练[15]。文献[16]引入掩码矩阵与提示矩阵，并对 GAN 的结构进行改进，克服了上述缺点。但该方法只考虑了单一量测值的分布特征，无法提取多维时序数据的耦合关系。

分布式光伏点多面广、局部渗透率高，整体呈现集群产业化发展态势，如整县屋顶光伏推进形态。高密度的分布式光伏并网使得配电网由辐射状无源网转变为大量电源点集中分布的网格有源网[17]。随着分布式光伏的规模化开发与并网，相关建模工作需考虑其集群运营、网格化分布的特点[18]。然而，上述缺失数据修复方法仅适用于单个光伏场站。并且位于特定网格区域内的集群分布式光伏通常具有相似的出力特性[19]，而现有研究未将该特点纳入缺失光伏数据修复中，以提升插补的准确性。

综上分析，提出一种基于双向循环插补网络的分布式光伏集群时序数据耦合增强方法。首先以双向多阶段循环插补网络(Bi-directional multi-stage recurrent imputation network with decay mechanism，BiMSRIN-D)构建编码器，对缺失数据进行初步插补。考虑到光伏周期性、间歇性的出力特点，在编码器中引入了衰减机制(decay mechanism)，以跟踪光伏曲线的波动趋势。BiMSRIN-D 编码器通过捕捉时间序列的动态演变模式以及多维量测数据间的耦合关系，挖掘量测数据整体分布规律，可充分利用相邻光伏场站间的时空关联特性对缺失值进行插补。解码器基于注意力机制对编码器的隐藏状态循环解码，并赋予时序状态不同权重，以实现长时间尺度的信息融合，进一步增强数据质量。本文所提方法无需任何先验假设，无需完整数据作为训练集，能够无监督学习到多维量测数据的时序特性，一次性对多个光伏场站的缺失值完成修复。算例结果表明，在大规模随机缺失、连续片段缺失情况下，所修复数据皆能保持较高的精度。

1　问题建模

1.1　分布式光伏网格化管理

高密度的分布式光伏场站与现有气象资源在地理解析度上是无法匹配的[20]。为实现区域内分布式光伏场站的量测信息全覆盖及其与关键气象信息的匹配，我国河北、福建等地针对分布式光伏密集区域开发了网格化数据建模技术[21]，具体如图 1 所示。

首先根据地理位置、环境条件将原始区域划分

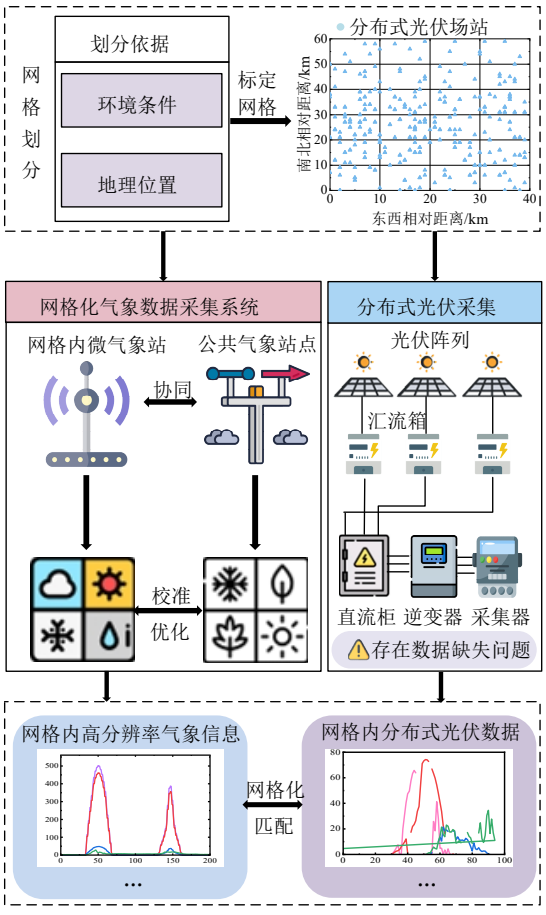


图 1　分布式光伏网格化管理

Fig. 1　Grid management of distributed photovoltaics

为多个网格，网格内涵盖分布式光伏集群；将分布式光伏场站与气象预报数据通过网格进行匹配，并在网格内引入辐照度微气象站进行校准优化，可有效提升数据匹配效率，实现网格化数据增强。然而分布式光伏的运行环境复杂多变，光伏阵列以及微气象监测点的分布式数据采集系统数目众多、安装分散，其运行过程往往会面临传感器故障、通信阻塞等问题，导致关键气象要素以及光伏站点的量测信息采集不全，给分布式光伏的网格化管理带来了困难。因此，研究分布式光伏集群与微气象信息耦合条件下的缺失数据修复方法，以实现网格内数据增强，对分布式光伏的网格化运维具有重要意义。

1.2 数据缺失问题描述

T 个时间窗口内的采集信息为多维时间序列 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T)^{\mathrm{T}}\in\mathbf{R}^{T\times D}$，$\boldsymbol{x}_t$ 表示 D 个变量的第 t 个量测值，x_t^d 表示 $\boldsymbol{x}_t$ 中第 d 个变量的量测值。由于时间序列 $\boldsymbol{X}$ 中存在缺失值，需定义掩码矩阵 $\boldsymbol{M}$ 来描述缺失值所在位置，其中掩码向量 $\boldsymbol{m}_t$ 对应 $\boldsymbol{x}_t$ 中的 D 个变量，$m_t^d=0$ 时表示量测值 x_t^d 出现缺失，否则 $m_t^d=1$。

在实际缺失场景中，一些变量会出现连续片段丢失现象，因此需引入时滞矩阵 $\boldsymbol{\delta}=(\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\ldots,\boldsymbol{\delta}_T)^{\mathrm{T}}\in\mathbf{R}^{T\times D}$ 记录最近一次有效量测值到当前时间戳 S_t 的时间间隔。由于光伏出力数据以及气象数据为均匀采样，假设第一次采样是在时间戳 0 处进行的，时间戳向量即可定义为 $\boldsymbol{S}=[0,1,2,3,\ldots,S_T]$，时滞矩阵的元素 δ_t^d 则通过式(1)进行计算。

$$\delta_t^d=\begin{cases}S_t-S_{t-1}+\delta_{t-1}^d, & t>1,m_{t-1}^d=0\\ S_t-S_{t-1}, & t>1,m_{t-1}^d=1\\ 0, & t=1\end{cases}\tag{1}$$

1.3 数据增强方法总体结构

本文的数据增强模型采用序列到序列(Seq2Seq)架构，其整体框架如附录图 B2 所示，由编码器以及解码器两个模块构成。该模型首先输入多维量测信息、掩码矩阵、时滞矩阵，采用 BiMSRIN-D 编码器初步完成数据的正向重建以及反向重建过程。并输出隐含时间序列动态演变模式的隐藏状态 $\boldsymbol{h}_i$，通过正向重建损失、反向重建损失以及一致性损失约束编码器的训练过程。

正向 MSRIN-D 以及反向 MSRIN-D 作为两个相互独立的模块，分别提取光伏出力数据的历史时序特性以及未来时序特性。为进一步增强数据质量，解码器采用注意力机制对二者的融合隐藏状态进行循环解码并输出结果序列 $\boldsymbol{y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_T\}$，以实现长时间尺度的特征提取与信息融合。并通过上下文相似性损失，促使编码器与解码器协同训练，一次性完成多个光伏场站的数据重建工作。

2 BiMRSIN-D 编码器

2.1 长短期记忆网络

长短期记忆网络(long short-term memory，LSTM)作为一种改进的循环神经网络，在光伏出力预测领域表现出了良好性能，常用于提取光伏数据波动性、间歇性时序特征[22]。本节采用 LSTM 作为 BiMRSIN-D 网络的循环单元，以捕捉时间序列动态演变模式。其具体结构如图 2 所示，其中 $\boldsymbol{x}_t$ 和 $\boldsymbol{h}_t$ 分别为 LSTM 隐含层的输入向量与输出向量，$\boldsymbol{c}_t$ 为记忆单元用于保留历史输入信息的特征。通过式(2)对门控单元的信息进行更新与传递。

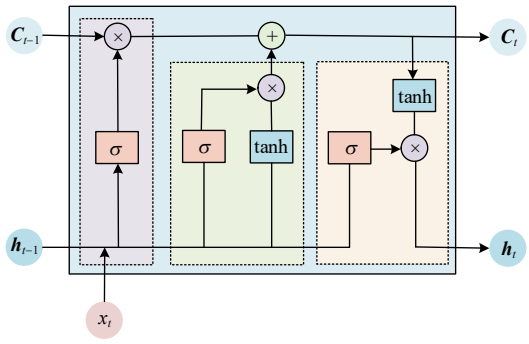


图 2 LSTM 网络结构

Fig. 2 LSTM network structure

$$\begin{cases}\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1};\boldsymbol{x}_t]+\boldsymbol{b}_f)\\ \boldsymbol{i}_t=\sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1};\boldsymbol{x}_t]+\boldsymbol{b}_i)\\ \boldsymbol{o}_t=\sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1};\boldsymbol{x}_t]+\boldsymbol{b}_o)\\ \tilde{\boldsymbol{c}}_t=\tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1};\boldsymbol{x}_t]+\boldsymbol{b}_c)\end{cases}\tag{2}$$

式中：$\boldsymbol{W}_i$、$\boldsymbol{W}_f$、$\boldsymbol{W}_o$、$\boldsymbol{W}_c$ 分别为输入门、遗忘门、输出门和记忆单元的权值矩阵；$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$、$\boldsymbol{b}_c$ 为相应的偏置项；σ为 Sigmoid 激活函数；tanh 为双曲正切激活函数。

记忆单元 $\boldsymbol{c}_t$ 以及隐藏状态 $\boldsymbol{h}_t$ 的更新公式为式(3)。

$$\begin{cases}\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tilde{\boldsymbol{c}}_t\\ \boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t)\end{cases}\tag{3}$$

式中 $\odot$ 为 Hadamard 积。

2.2 多阶段循环插补网络(MSRIN-D)

与循环神经网络类似，多阶段循环插补网络(multi-stage recurrent imputation network with decay mechanism，MSRIN-D)利用隐藏状态 $\boldsymbol{h}_t$ 传递时间序列的动态演变模式，并通过 LSTM 单元不断更新、传递 $\boldsymbol{h}_t$ 的状态信息，以实现循环动态插补过程。此外，MSRIN-D 采用掩码矩阵以及时滞矩阵描述数据缺失模式，并将两者有效地嵌入深度学习算法框

架中，以提取缺失模式信息，增强插补效果。

该算法首先基于历史信息对缺失数据进行估值，采用全连接层将隐含动态时序特性的 $\boldsymbol{h}_{t-1}$ 映射为当前时刻重建数据，完成初步插补。邻近区域内的分布式光伏场站通常具有相似的硬件配置以及相似的气象条件，因而具有相似的出力特性，光伏集群中各站点的空间分布可作为不同地理位置的“环境监测站”，为区域内其他站点提供参考信息[23]，且辐照度等微气象信息与光伏出力具有强关联性。因此第二阶段插补利用当前时刻的关键气象要素、其余场站的出力数据对缺失值进行插补，充分提取多维量测值之间的耦合关系。第三阶段根据缺失模式信息自适应集成前两次插补值，输出最终的重建数据。MSRIN-D 的结构如附录图 B3 所示，主要由插补单元、衰减单元、权值融合单元、长短期记忆单元组成。

2.2.1　第一阶段插补

在第一阶段插补中，利用上一时刻 LSTM 单元输出的隐藏状态 $\boldsymbol{h}_{t-1}$ 对缺失值进行估计与替换，其计算过程如式(4)和式(5)所示。

$$\hat{\boldsymbol{x}}_t = \boldsymbol{W}_x \boldsymbol{h}_{t-1} + \boldsymbol{b}_x \tag{4}$$

$$\boldsymbol{x}_t^c = \boldsymbol{m}_t \odot \boldsymbol{x}_t + (1-\boldsymbol{m}_t) \odot \hat{\boldsymbol{x}}_t \tag{5}$$

式中：$\boldsymbol{W}_x$、$\boldsymbol{b}_x$ 分别为权值矩阵和偏置项；该全连接层将传递历史时序信息的隐藏状态 $\boldsymbol{h}_{t-1}$ 映射为 $\hat{\boldsymbol{x}}_t$；$\hat{\boldsymbol{x}}_t$ 为时序插补向量，其值仅取决于历史观测值，和当前时刻其余量测信息无关，是一种基于时序特性的估计；$\boldsymbol{m}_t$ 为掩码向量，该插补单元使用时序插补向量 $\hat{\boldsymbol{x}}_t$ 中对应的数值来替换 $\boldsymbol{x}_t$ 中缺失值，得到补码向量 $\boldsymbol{x}_t^c$。

2.2.2　第二阶段插补

对于分布式光伏场站而言，相邻站点的出力特性往往比较相似，因此每个光伏场站的量测信息可以代表一个特征量[24]。不仅可以使用目标站点的历史出力数据，还可通过当前时刻相邻站点的出力数据以及关键气象要素增强对目标站点缺失值的修复。如图 3 所示，场站 2 的缺失数据修复除自身历史数据外还可参考其余 3 个相邻场站的采集数据。

第二阶段插补是一种基于特征量的插补，通过式(5)已求取补码向量 $\boldsymbol{x}_t^c$，将基于特征量的估计值定义为特征插补向量 $\hat{\boldsymbol{z}}_t$，其表达式如式(7)所示。

$$\boldsymbol{W}_z = \begin{bmatrix} 0 & W_{12} & \cdots & W_{1n} \\ W_{21} & 0 & \cdots & W_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & 0 \end{bmatrix} \tag{6}$$

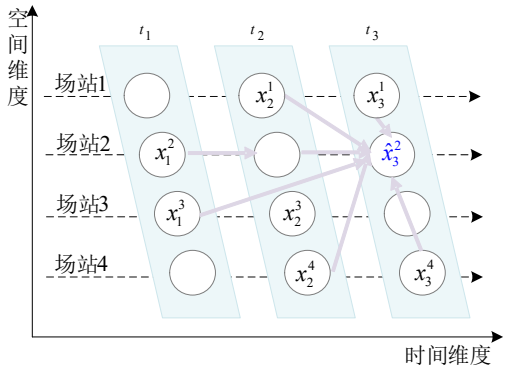


图 3　多场站插补模型

Fig. 3　Multi-station imputation model

$$\hat{\boldsymbol{z}}_t = \boldsymbol{W}_z \boldsymbol{x}_t^c + \boldsymbol{b}_z \tag{7}$$

式中 $\boldsymbol{W}_z$、$\boldsymbol{b}_z$ 为全连接层的相应参数，并且限制权值矩阵 $\boldsymbol{W}_z$ 的对角线元素全为 0，从而保证 $\hat{\boldsymbol{z}}_t$ 中第 d 个变量正好是基于其他量测信息对缺失值 x_t^d 的估计，可充分提取多维量测值之间的耦合关系。

2.2.3　第三阶段插补

1）衰减单元。

光伏出力具有明显的周期性与间歇性，在极端天气和夜晚时分出力会趋于默认值 0，本文在权值融合单元及隐藏状态 $\boldsymbol{h}_t$ 处引入衰减机制以捕捉上述特性。其中衰减因子 $\boldsymbol{\gamma}_t$ 用于控制衰减率，且该模型是在训练过程中学习衰减率，而不是预先固定衰减因子[25]。其计算过程如下：

$$\boldsymbol{\gamma}_t = \exp\{-\max(0, \boldsymbol{W}_\gamma \boldsymbol{\delta}_t + \boldsymbol{b}_\gamma)\} \tag{8}$$

式中：$\boldsymbol{W}_\gamma$、$\boldsymbol{b}_\gamma$ 为相应的模型参数；$\boldsymbol{\delta}_t$ 为时滞向量。选择指数负整流器作为激活函数，以保证衰减率在 0 到 1 的合理范围内单调递减。

2）权值融合单元。

将时序插补向量 $\hat{\boldsymbol{x}}_t$ 与特征插补向量 $\hat{\boldsymbol{z}}_t$ 加权组合，获取全局插补向量 $\hat{\boldsymbol{G}}_t$，具体表达式如下：

$$\boldsymbol{\beta}_t = \sigma(\boldsymbol{W}_\beta[\boldsymbol{\gamma}_t \circ \boldsymbol{m}_t] + \boldsymbol{b}_\beta) \tag{9}$$

$$\hat{\boldsymbol{G}}_t = \boldsymbol{\beta}_t \odot \hat{\boldsymbol{z}}_t + (1-\boldsymbol{\beta}_t) \odot \hat{\boldsymbol{x}}_t \tag{10}$$

式中：$\boldsymbol{\beta}_t$ 为 $\hat{\boldsymbol{x}}_t$ 与 $\hat{\boldsymbol{z}}_t$ 的组合权重值；$\circ$表示数组拼接操作；权值融合单元通过学习时序衰减因子 $\boldsymbol{\gamma}_t$ 及掩码向量 $\boldsymbol{m}_t$ 中蕴含的缺失模式信息以训练参数 $\boldsymbol{W}_\beta$ 与 $\boldsymbol{b}_\beta$，从而实现权重值 $\boldsymbol{\beta}_t$ 依据缺失模式自适应调整。

3）更新隐藏状态。

插补单元将 $\boldsymbol{x}_t$ 中的缺失值替换为全局插补向量 $\hat{\boldsymbol{G}}_t$ 中对应的元素，获取全局补码向量 $\boldsymbol{G}_t^c$，$\boldsymbol{G}_t^c$ 为 MSRIN-D 网络的最终重建数据，并将 $\boldsymbol{G}_t^c$ 馈送至下一个循环单元以更新隐藏状态 $\boldsymbol{h}_t$。

$$\boldsymbol{G}_t^c = \boldsymbol{m}_t \odot \boldsymbol{x}_t + (1-\boldsymbol{m}_t) \odot \hat{\boldsymbol{G}}_t \tag{11}$$

$$\begin{cases} \boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1} \odot \boldsymbol{\gamma}_t; \boldsymbol{G}_t^c \circ \boldsymbol{m}_t] + \boldsymbol{b}_f) \\ \boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1} \odot \boldsymbol{\gamma}_t; \boldsymbol{G}_t^c \circ \boldsymbol{m}_t] + \boldsymbol{b}_i) \\ \boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1} \odot \boldsymbol{\gamma}_t; \boldsymbol{G}_t^c \circ \boldsymbol{m}_t] + \boldsymbol{b}_o) \\ \tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1} \odot \boldsymbol{\gamma}_t; \boldsymbol{G}_t^c \circ \boldsymbol{m}_t] + \boldsymbol{b}_c) \end{cases} \tag{12}$$

$$\begin{cases} \boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \\ \boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \end{cases} \tag{13}$$

式(12)中的 LSTM 更新公式与式(2)中的标准 LSTM 公式主要有两点区别：一是将掩码向量 $\boldsymbol{m}_t$ 通过数组拼接的方式馈送至模型中。二是 $\boldsymbol{h}_{t-1}$ 处引入衰减因子 $\boldsymbol{\gamma}_t$，以控制 LSTM 对历史信息的记忆程度。时间序列波动时，较小的衰减因子使得历史信息对当前时间步影响力较低；当时间序列具有周期性时，较大的衰减因子促使 $\boldsymbol{h}_t$ 记忆更多信息，以有效跟踪历史时序特性[26]。光伏出力具有周期性、间歇性特点，并在某些时刻具有较强的随机性，引入衰减因子有利于拟合光伏曲线的波动趋势，捕捉光伏时序的长期依赖关系。

2.2.4 BiMSRIN-D 网络

单向 MRSIN-D 只能学习光伏时序的正向演变模式，而光伏出力在未来时刻具有较大不确定性。本节通过双向循环插补网络应对上述问题，除正向插补外，时间序列中任一缺失值还可通过反向进行推导。正向 MRSIN-D 与反向 MRSIN-D 同时处理数据，综合提取过去和未来的时序相关信息，具有更强的非线性拟合能力[27]。其更新表达式为

$$\begin{cases} \overrightarrow{\boldsymbol{h}_i} = M(\boldsymbol{W}_1 \vec{x}_t + \boldsymbol{W}_2 \overrightarrow{\boldsymbol{h}_{i-1}} + \vec{\boldsymbol{b}}) \\ \overleftarrow{\boldsymbol{h}_i} = M(\boldsymbol{W}_3 \vec{x}_t + \boldsymbol{W}_4 \overleftarrow{\boldsymbol{h}_{i-1}} + \overleftarrow{\boldsymbol{b}}) \end{cases} \tag{14}$$

式中：$\overrightarrow{\boldsymbol{h}_i}$ 为正向 MRSIN-D 的隐藏状态；$\overleftarrow{\boldsymbol{h}_i}$ 为反向 MRSIN-D 的隐藏状态；M 为单元函数；$\boldsymbol{W}_{1\sim4}$ 以及 $\boldsymbol{b}$ 为相应的模型参数。将双向隐藏状态进行连接即为编码器的输出状态 $\boldsymbol{h}_i$，如式(15)所示。

$$\boldsymbol{h}_i = \text{concat}(\overrightarrow{\boldsymbol{h}_i}; \overleftarrow{\boldsymbol{h}_i}) \tag{15}$$

式中 concat 表示向量拼接操作。

3 Seq2Seq 模型

采集数据涵盖微气象、光伏等多类不确定量，具有高维、时变、波动性强的特点。Seq2Seq 模型具备较强的非线性映射能力，常用于提取多元耦合时间序列的特征信息。本文首先采用 BiMSRIN-D 编码器获取表征光伏出力不确定性的状态向量 $\boldsymbol{h}_i$，通过 LSTM 解码器将其映射为最终重建数据。并引入注意力机制，增强模型对光伏数据复杂时序特征的感知能力。

3.1 基于注意力机制的循环解码

随着时间序列长度增加，LSTM 解码器提取关键信息的难度也会加大。为此，引入全局注意力机制可以自适应为状态向量 $\boldsymbol{h}_i$ 赋予不同权重，从而更精准地提取强相关时刻的特征信息[28]，以提升数据重建质量。该模型具体结构如图 4 所示。

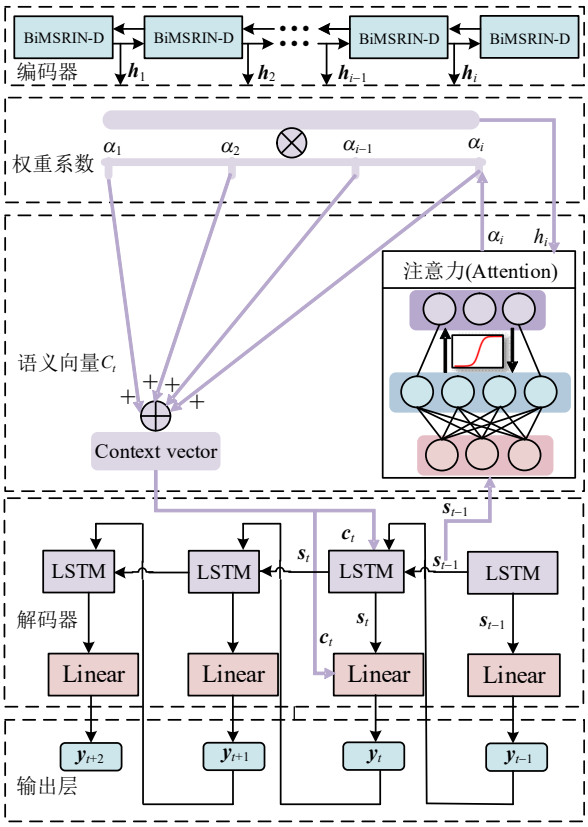


图 4 Seq2Seq 结构

Fig. 4 Seq2Seq network structure

首先根据编码器的状态向量 $\boldsymbol{h}_i$ 以及上一时刻解码器的隐藏状态 $\boldsymbol{s}_{t-1}$，求取当前时刻注意力权重系数 α_{ti}，如下式所示：

$$e_{ti} = \boldsymbol{V}_d \tanh(\boldsymbol{W}_d[\boldsymbol{s}_{t-1}, \boldsymbol{h}_i] + \boldsymbol{U}_d) \tag{16}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{i=1}^{n} \exp(e_{ti})} \tag{17}$$

式中：e_{ti} 反映了状态向量 $\boldsymbol{h}_i$ 与当前时刻输出值之间的相关性[29]；$\boldsymbol{V}_d$、$\boldsymbol{W}_d$ 为全局注意力机制的权值矩阵；$\boldsymbol{U}_d$ 为相应的偏置项。将 softmax 激活函数作用于 e_{ti}，以确保所有注意力权重系数 α_{ti} 的总和被归一化为 1。

在解码器每个时间步 t，将权重系数 α_{ti} 与对应的状态向量 $\boldsymbol{h}_i$ 加权求和得到语义向量 $\boldsymbol{c}_t$。

$$\boldsymbol{c}_t = \sum_{i=1}^{n} \alpha_{ti} \boldsymbol{h}_i \tag{18}$$

将语义向量 $\boldsymbol{c}_t$，解码器上一时刻的输出 $\boldsymbol{y}_{t-1}$ 以

及隐藏状态 $\boldsymbol{s}_{t-1}$ 输入 LSTM 解码器以更新解码器隐藏状态 $\boldsymbol{s}_t$：

$$\boldsymbol{s}_t = f_{\mathrm{de}}(\boldsymbol{y}_{t-1} \circ \boldsymbol{c}_t, \boldsymbol{s}_{t-1}) \tag{19}$$

式中 $f_{\mathrm{de}}(\cdot)$ 为 LSTM 函数；∘为数据拼接操作。在 LSTM 层的下一层，添加全连接层，以输出连续的重建数据。解码器的输出时间序列 $\boldsymbol{y}_t$ 由以下公式计算：

$$\boldsymbol{y}_t = \boldsymbol{W}_l[\boldsymbol{s}_t; \boldsymbol{c}_t] + \boldsymbol{b}_l \tag{20}$$

$$\hat{\boldsymbol{y}}_t = \boldsymbol{m}_t \odot \boldsymbol{x}_t + (1 - \boldsymbol{m}_t) \odot \boldsymbol{y}_t \tag{21}$$

式中：$\boldsymbol{W}_l$、$\boldsymbol{b}_l$ 为相应的参数；$\hat{\boldsymbol{y}}_t$ 保留原始量测值中未缺失数据，并将缺失数据用 $\boldsymbol{y}_t$ 的对应值进行替换，$\hat{\boldsymbol{y}}_t$ 即为最终重建数据。

3.2 模型损失函数

MSRIN-D 算法包含 3 个插补阶段：第 1 阶段输出时序插补向量 $\hat{\boldsymbol{x}}_t$，第 2 阶段输出特征插补向量 $\hat{\boldsymbol{z}}_t$，第 3 阶段对 $\hat{\boldsymbol{x}}_t$ 与 $\hat{\boldsymbol{z}}_t$ 加权组合后得到全局插补值 $\hat{\boldsymbol{G}}_t$。因此，正向 MSRIN-D 编码层的重建损失函数定义如下：

$$L_{\mathrm{MAE}}(x, y, m) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T'} |(x - y) \odot m_t^d|}{\sum_{d=1}^{D}\sum_{t=1}^{T} m_t^d} \tag{22}$$

$$L_t = L_{\mathrm{MAE}}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t, \boldsymbol{m}_t) + L_{\mathrm{MAE}}(\boldsymbol{x}_t, \hat{\boldsymbol{z}}_t, \boldsymbol{m}_t) + L_{\mathrm{MAE}}(\boldsymbol{x}_t, \hat{\boldsymbol{G}}_t, \boldsymbol{m}_t) \tag{23}$$

式中：L_{MAE} 表示求取平均绝对误差；T'、D 分别为时间步长与量测数据的维度；$\boldsymbol{m}_t$ 为掩码向量。与正向 MSRIN-D 类似，反向 MSRIN-D 相应的插补值为 $\hat{\boldsymbol{x}}_t'$、$\hat{\boldsymbol{z}}_t'$、$\hat{\boldsymbol{G}}_t'$，反向 MSRIN-D 编码层的重建损失函数定义如下：

$$L_t' = L_{\mathrm{MAE}}(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t', \boldsymbol{m}_t) + L_{\mathrm{MAE}}(\boldsymbol{x}_t, \hat{\boldsymbol{z}}_t', \boldsymbol{m}_t) + L_{\mathrm{MAE}}(\boldsymbol{x}_t, \hat{\boldsymbol{G}}_t', \boldsymbol{m}_t) \tag{24}$$

重建损失促使补齐的量测数据尽可能拟合真实数据。为促使正向 MSRIN-D 以及反向 MSRIN-D 在每个时间步的插补都尽可能保持一致，并学习到数据的联合分布特性，在 BiMSRIN-D 编码器中引入一致性损失，如下式所示：

$$L_t^{\mathrm{cons}} = \mathrm{dif}(\hat{\boldsymbol{G}}_t, \hat{\boldsymbol{G}}_t') \tag{25}$$

式中 dif 表示求取均方根误差。为促使解码器在循环解码过程中生成与原始量测值相似的样本并优化编码器状态。需定义相似性损失：

$$L_t^{\mathrm{en}} = L_{\mathrm{MAE}}(\boldsymbol{y}_t, \boldsymbol{x}_t) \tag{26}$$

本文的模型需将以上各部分损失函数线性组合，通过控制不同类型损失函数的权重，平衡各个训练目标，以提高模型的泛化能力和准确性。在数据实验过程中，当模型的综合损失函数设置为式(27)时，有较快的收敛速度，可有效避免重建损失主导梯度方向。

$$L_{\mathrm{total}} = L_t / 3 + L_t' / 3 + L_t^{\mathrm{cons}} + L_t^{\mathrm{en}} \tag{27}$$

4 算例分析

本文硬件配置为 Intel Core i5-12400 CPU，NVIDIA GeForce RTX 2060S GPU。编译环境为 Spyder，以 Python3.9+Pytorch2.0+CUDA 11.7 为架构搭建并训练本文模型，以实际光伏数据对该模型的性能进行了测试。

4.1 实验数据集

缺失数据修复是一种无监督学习过程，对于原始量测信息有缺失的数据集，是无法验证插补效果的。为了衡量算法性能，需通过无自然缺失的数据集人为生成缺失数据，将修复的数据与对应的真实量测数据进行对比，计算插补误差。本文数据来源为我国西北地区某光伏站群，分别选取 A 区域内 4 座相邻场站、B 区域内 4 座相邻场站一整年 365 天的实际出力数据进行算例验证。A 区域内 1—3 号分布式光伏场站的装机容量为 4~6MW，4 号场站为集中式光伏电站，装机容量为 15MW；B 区域内 5—8 号分布式光伏场站的装机容量为 6~10MW。数据采样间隔为 15min，每日共计 96 个点。根据文献[23]建立的环境因素与光伏出力的映射关系，本文选择太阳直接辐射系数(direct normal irradiance，DNI)、太阳散射辐射系数(diffuse horizontal irradiance，DHI)、太阳总水平辐射系数(global horizontal irradiance，GHI)、温度 4 个指标作为模型的气象特征。

本文的数据修复模型不需要完整数据作为训练集，模型的输入为带缺失的多维量测数据、掩码矩阵、时滞矩阵，输出为完整的重建数据。用于训练 Seq2Seq 模型的数据即为含缺失的待插补数据集，因此不区分训练集与测试集，每个区域可用于数据实验的样本共计 365 组，对 A 区域以及 B 区域分别进行测试。所提算法的输入特征包括光伏集群的出力数据以及气象信息，将每日 96 个点的量测信息整理为多维向量。

为提升原始量测数据的质量，需首先对异常值进行检测与清洗，本文采用孤立森林(isolation forest)筛选数据集中的离群值并剔除[30]。通过设置掩码矩阵 $\boldsymbol{M}$ 中 0 元素与 1 元素的比例、位置，以生成多种光伏数据的缺失场景，并依据 1.2 节的方法计算相应的时滞矩阵。为避免各物理量因量纲差异

对算法训练过程造成干扰，采用线性归一化函数将不同属性的量测值都映射至[0,1]区间，并构建[样本总数，时序长度，特征量维度]形式的数组。本文最终将(365,96,8)维度的数据输入 Seq2Seq 模型中进行缺失数据修复。

4.2 模型评价指标

本算例采用均方根误差(root mean squared error，RMSE)和平均绝对误差(mean absolute error，MAE)作为模型对缺失数据修复精度的评价指标，如式(28)(29)所示。

$$e_{\mathrm{MAE}}=\frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\left|\left(\tilde{x}_t^d-\hat{y}_t^d\right)\odot\left(1-m_t^d\right)\right|}{\sum_{d=1}^{D}\sum_{t=1}^{T}\left(1-m_t^d\right)} \tag{28}$$

$$e_{\mathrm{RMSE}}=\sqrt{\frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\left[\left(\tilde{x}_t^d-\hat{y}_t^d\right)\odot\left(1-m_t^d\right)\right]^2}{\sum_{d=1}^{D}\sum_{t=1}^{T}\left(1-m_t^d\right)}} \tag{29}$$

式中：$\tilde{x}_t^d$ 为 t 时刻第 d 个量测值的真实数据；$\hat{y}_t^d$ 为相应的插补值；m_t^d 为掩码矩阵中对应位置的元素。

4.3 随机缺失数据修复效果

为模仿实际缺失场景，本节实验中随机设置掩码向量中 0 元素的位置，并保证 0 元素的比例分别为 20%、30%、50%、70%、80%、90%，将不同掩码向量与各场站的出力数据分别相乘，从而生成不同比例的随机缺失数据。

在算法训练过程中，超参数的取值对模型的收敛速度以及插补精度都有显著影响，本模型通过网格搜索法确定超参数的选择，具体如附录表 A1 所示。在随机缺失模式下，采用本文模型进行缺失数据插补的误差情况如附录表 A2 所示。各缺失比例下绝对插补误差的分布情况如附录图 B5 所示。

实验结果表明，所提 Seq2Seq 模型在缺失率不高于 50%时，数据插补效果较为稳定，保持着良好的插补性能，MAE 皆保持在 0.05MW 左右。且绝对误差集中分布于 0~0.15MW 区间，并未出现误差较大的点，只有少数点的误差高于 0.2MW。当缺失率高于 70%时，MAE 随着缺失率的升高有所增长。70%缺失率时，MAE 升至 0.07MW 左右；80%缺失率，MAE 升至 0.1MW 左右；90%缺失率时，MAE 升至 0.15MW 左右，该模型仍保持着较高的精度。

缺失率为 70%、80%时，绝对误差集中分布于 0~0.3MW 区间；缺失率高达 90%时，误差主要分布于 0~0.4MW 区间，部分采样点出现误差偏大的极端情况，但对于大部分缺失点仍能给出理想的插补值。表明该模型能根据少量有效观测值，学习多维量测数据的潜在分布特性，捕捉时间序列的波动特征、周期特征，充分拟合光伏数据的出力特性。

4.4 算法对比

为验证所提多阶段插补算法的必要性，首先进行消融实验，设置 3 种对比算法进行算例验证，分别为单阶段插补、二阶段插补与多阶段插补，详见附录 B。场站 1、场站 4 在不同缺失率下的测试结果见附录表 B1。单阶段插补、二阶段插补在缺失率低于 30%时有较好的插补性能，此时多阶段插补在插补精度上无明显改善。随着缺失率的升高，多阶段插补展现出了明显的优越性。以场站 1 为例，在 80%缺失率下，多阶段插补较单阶段插补的 MAE 降低了 55.1%，较二阶段插补降低 67.2%。多阶段插补算法通过更深层的模型学习数据的潜在分布特性与多维量测值间的互相关信息，分阶段逐步逼近缺失的数据，可有效提升插补准确性。

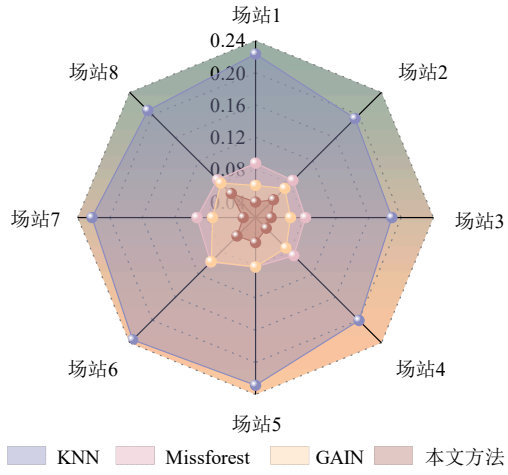
为验证所提多阶段插补算法的性能优势，将本文方法与 KNN、Missforest、生成对抗插补网络(generative adversarial imputation network，GAIN)3 种常见的无监督缺失数据插补方法进行对比测试。

4.4.1 随机缺失

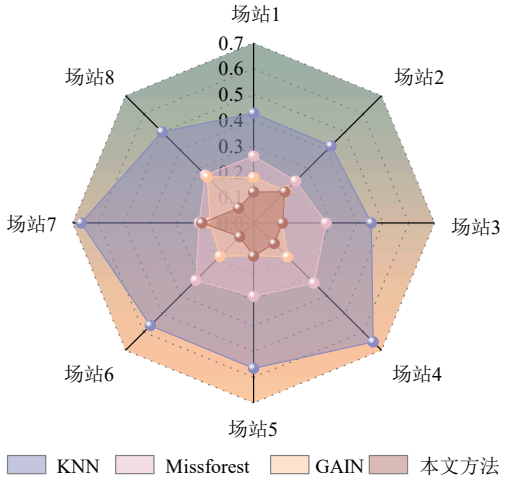
以场站 1 和场站 4 的数据集为例，附录图 B4 展现了不同缺失率下 2 座典型光伏场站在某一天不同插补算法下的重建光伏功率与真实值对比。可以看出本文方法在 50%、80%缺失率下相比于 GAIN，都有着更高的插补精度。且在光伏出力的尖峰时段，本文方法也能更准确地跟随光伏功率的波动趋势。

由图 5 可以看出，由于多光伏场站的属性差异，各模型在不同场站的插补误差有不同的分布特点。在不同数据集中，各模型 e_{MAE} 雷达图的轮廓变化趋势也有所差异，其中场站 4 与场站 7 在 KNN 模型中插补误差明显偏大。而本文方法在多个场站中始终保持较强的稳定性，对 8 个场站的光伏出力曲线均能达到最佳插补效果，优于其他模型。

场站 1、场站 4 在不同比例缺失率下的插补误差对比见附录表 A3 以及附录表 A4，KNN 模型相比于其他 3 种算法，插补效果明显较差。缺失率低于 50%时，Missforest、GAIN 以及本文方法皆能保持较高的插补精度，随着缺失率的不断增加，本文 Seq2Seq 模型展现出了一定优越性。以场站 1 为例，在 80%缺失率下，Seq2Seq 较 Missforest 的数据插补精度提升了 53.6%，较 GAIN 提升了 31.9%；在 90%缺失率下，Seq2Seq 较 Missforest 的数据插补精



(a) 平均绝对误差/MW(缺失率50%)



(b) 平均绝对误差/MW(缺失率80%)

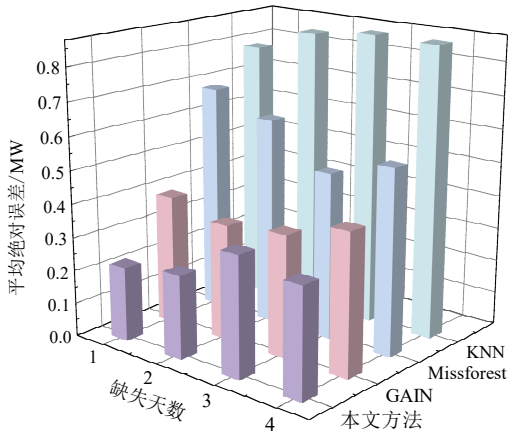
图 5　不同方法下各场站误差对比

Fig. 5　Comparison of errors across different stations

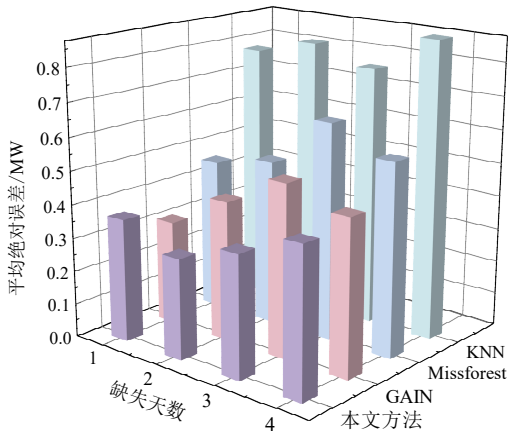
度提升了 62.6%，较 GAIN 提升了 56.3%。通过对比不同缺失率下的插补误差 e_{MAE} 和 e_{RMSE} 值可以得知，本文算法的插补结果与实际值拟合度更高，在插补精度上优于其他 3 种方法。在高缺失率下具有明显优势，可精准完成缺失数据的重建。此外，为进一步模拟量测系统的实际运行状况，向数据集中添加服从正态分布的高斯噪声，并采用本文方法、GAIN、Missforest 算法进行仿真测试，详见附录 B，各场站在 50%缺失率下的仿真结果见附录表 B2。随着噪声扰动的加大，3 种方法的 MAE 皆出现增长，但本文方法的插补性能最好。此外，即使在噪声标准差高达 0.1 时，本文方法的 MAE 仍能保持在 0.1MW 以内，可有效应对测量噪声可能带来的估计错误。

4.4.2　连续片段缺失

在数据采集和传输过程中，传感器、量测表计以及通信设施发生故障时，短时间内无法恢复正常运行，将造成量测数据连续片段丢失。为了验证本文模型对于连续性缺失的适用性，本节在整体缺失率 50%的基础上随机选取总采样时段内连续 1 天、2 天、3 天和 4 天发生量测数据完全缺失。对连续缺失片段的插补误差进行计算。详细数据见附录表 A5、A6。本文算法与 3 种算法的插补性能对比如图 6 所示。



(a) 场站1



(b) 场站4

图 6　连续缺失下误差对比

Fig. 6　Comparison under Continuous Missing Data

由实验结果可知，无论是少量连续缺失还是大量连续缺失，本文模型的插补性能都强于其他 3 种算法。在连续缺失模式下，Missforest 以及 KNN 表现不佳。KNN 基于欧式距离寻找最近邻样本，利用未缺失的近邻样本均值来替代缺失值，在连续片段缺失中，KNN 搜寻近邻样本的难度较大，因此插补性能较差。而 Missforest 是一种基于决策树的算法，采用套袋法抽样多个训练样本构建多棵决策树，从而形成随机森林，作为一种非参数方法，通过不断迭代计算达到收敛条件，发生连续缺失时，收敛过程缓慢，并出现大量袋外数据，导致插补精度较低。而本文方法可有效挖掘时间序列中隐含的动态演变模式，在连续缺失场景中，整体插补效果也优于 GAIN。以场站 4 为例，连续缺失 2 天时，本文方法的精度较 GAIN 提升了 28.2%；连续缺失

3 天时，较 GAIN 提升了 29.6%；连续缺失 4 天时，较 GAIN 提升了 5.4%。表明该模型在整天缺失情况下，仍能根据历史数据及未来数据的时序信息对光伏出力特性进行量化表征，有效拟合缺失部分数据。

4.5 算法的鲁棒性分析

前文只设定了光伏出力数据的缺失，因此本节要进一步研究气象信息缺失对该模型插补精度的影响。每类气象样本的随机缺失比例与光伏数据保持一致，分别设置为 20%、30%、50%、70%、80%、90%，即模型的 8 个输入特征量设定为相同比例的缺失。表 1 为场站 1 在气象数据有无缺失时的插补误差对比。

表 1 气象数据缺失下场站 1 误差情况

Table 1 Imputation error of station 1 under missing meteorological data

缺失率/%	气象无缺失		气象有缺失	
	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}
20	0.0469	0.0765	0.0473	0.0802
30	0.0421	0.0714	0.0535	0.0902
50	0.0393	0.0665	0.0661	0.1155
70	0.0626	0.1121	0.0988	0.1741
80	0.1207	0.2247	0.1894	0.3343
90	0.1483	0.3071	0.3109	0.5287

由上表可以得知，即使在气象信息以及光伏数据缺失比例同时高达 70%时，本文模型仍然有较高的插补精度，MAE 仍保持在 0.1MW 以内。在缺失率低于 80%时，气象数据的缺失对插补误差的影响并不明显。当缺失率高达 90%时,仍保持较为理想的插补性能，但插补误差在气象数据缺失时会显著上升，原因在于气象信息和光伏数据同时缺失 90%时，用于训练的有效量测值非常有限，而本文模型作为一种数据驱动的方法，在训练过程会受到很大干扰，难以精准感知多维数据的整体分布规律。

附录图 B1 展示了 3 个代表性场站在气象信息缺失时的重建效果，在缺失率分别为 70%、80、90%的情况下，本文所提算法仍能有效跟踪光伏出力曲线的波动趋势。附录表 A7 记录了各场站在气象信息缺失时的插补误差情况，可以得知，在多维量测数据缺失比例同时高达 70%时，各场站的 MAE 均能保持在 0.11MW 以内，相较于气象数据无缺失时，MAE 只增加了约 0.04MW。缺失比例为 80%时，各场站的 MAE 均能保持在 0.18MW 以内，相较于气象数据无缺失时，MAE 增加了约 0.08MW，属于可接受范围。以上分析表明，在应对气象数据缺失时本文模型依然能保持良好的插补性能。

图 7 展示了场站 1、场站 4 在不同气象信息缺失率下的插补误差情况。可以得知，气象缺失率在

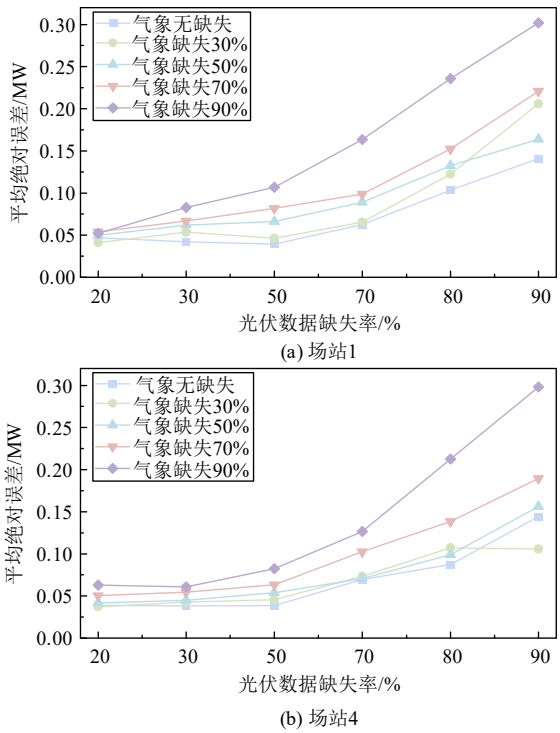


图 7 不同气象缺失率下误差对比

Fig. 7 Error comparison under different meteorological data missing rates

0~70%区间时，光伏数据的插补误差随气象缺失率的上升有所增加，但变化幅度较小。场站 1 缺失 90%时，在气象缺失 70%的条件下相较于气象无缺失时 MAE 增加了 0.0801MW。场站 4 缺失 90%时，在气象缺失 70%的条件下相较于气象无缺失时 MAE 只增加了 0.0453MW。即使在气象与光伏缺失率同时高达 90%时，MAE 仍保持在 0.3MW 左右，说明本文模型有足够的鲁棒性应对低质量的气象数据。

5 结论

针对分布式光伏集群多维量测信息的缺失问题，提出了一种基于双向多阶段循环插补网络的数据增强方法。该模型完全基于数据驱动，通过少量未缺失数据学习原始量测值的潜在分布特征、光伏出力特性等难以量化表征的复杂数值关系。克服了数理方法难以对多元耦合时间序列间的互相关信息显式建模的问题。所提方法无需完整数据作为训练集、无需任何先验分布假设，以无监督学习方式一次性对多个光伏场站的缺失数据完成修复，更适用于实际工程场景。最后通过算例所得结论如下：

1）由于本文模型是通过捕捉时间序列的动态演变模式以及多维量测值之间的耦合关系以修复数据，即使在高比例随机缺失以及整天缺失情况下仍能保持优良修复性能。

2）该模型所修复数据有效拟合了真实光伏曲

线的波动规律，相较于其他方法，在修复精度上有明显优势，可有效增强光伏集群数据质量，提升量测信息可用率，为其网格化运维及边缘智能分析提供数据支撑。

后续研究将结合联邦学习、迁移学习等先进技术，实现模型对多源异构数据的自适应跟踪学习，增强模型泛化能力。

附录见本刊网络版(http://www.dwjs.com.cn/CN/1000-3673/current.shtml)。

参考文献

[1] 国家能源局. 2022 年光伏发电建设运行情况[EB/OL]. (2023-02-17). http://www.nea.gov.cn/2023-02/17/c_1310698128.htm.

[2] ARGÜELLO A，LARA J D，ROJAS J D，et al. Impact of rooftop PV integration in distribution systems considering socioeconomic factors[J]. IEEE Systems Journal，2018，12(4)：3531-3542.

[3] 刘友波，吴浩，刘挺坚，等. 集成经验模态分解与深度学习的用户侧净负荷预测算法[J]. 电力系统自动化，2021，45(24)：57-64.
LIU Youbo，WU Hao，LIU Tingjian，et al. User-side net load forecasting method integrating empirical mode decomposition and deep learning[J]. Automation of Electric Power Systems，2021，45(24)：57-64(in Chinese).

[4] 荆渝，刘友波，邱高，等. 基于区间估计与深度强化学习的有源配电网多智能体电压滚动控制[J]. 电网技术，2023，47(5)：2019-2028.
JING Yu，LIU Youbo，QIU Gao，et al. Multi-agent voltage rolling control of active distribution network based on interval estimation and deep reinforcement learning[J]. Power System Technology，2023，47(5)：2019-2028(in Chinese).

[5] 唐冬来，倪平波，李玉，等. 基于互信共识标识的县域屋顶光伏消纳交易策略[J]. 电力系统自动化，2022，46(22)：41-50.
TANG Donglai，NI Pingbo，LI Yu，et al. Transaction strategy of roof-mounted photovoltaic accommodation for county area based on mutual trust and consensus identification[J]. Automation of Electric Power Systems，2022，46(22)：41-50(in Chinese).

[6] SALEEM B，WENG Yang，GONZALES F M，et. al. Association rule mining for localizing solar power in different distribution grid feeders[J]. IEEE Transactions on Smart Grid，2021，12(3)：2589-2600.

[7] GENES C，ESNAOLA I，PERLAZA S M，et al. Robust recovery of missing data in electricity distribution systems[J]. IEEE Transactions on Smart Grid，2019，10(4)：4057-4067.

[8] 纪德洋，金锋，冬雷，等. 基于皮尔逊相关系数的光伏电站数据修复[J]. 中国电机工程学报，2022，42(4)：1514-1522.
JI Deyang，JIN Feng，DONG Lei，et al. Data repairing of photovoltaic power plant based on Pearson correlation coefficient[J]. Proceedings of the CSEE，2022，42(4)：1514-1522(in Chinese).

[9] 张帅，杨晶显，刘继春，等. 基于多尺度时序建模与估计的电力负荷数据恢复[J]. 电工技术学报，2020，35(13)：2736-2746.
ZHANG Shuai，YANG Jingxian，LIU Jichun，et al. Power load recovery based on multi-scale time-series modeling and estimation[J]. Transactions of China Electrotechnical Society，2020，35(13)：2736-2746(in Chinese).

[10] 冯磊，王石刚，梁庆华. 基于 GAKNN 方法的配电站时间序列缺失数据补全方法[J]. 电力自动化设备，2021，41(12)：187-192.
FENG Lei，WANG Shigang，LIANG Qinghua. Completion method for missing time series data of distribution station based on GAKNN method[J]. Electric Power Automation Equipment，2021，41(12)：187-192(in Chinese).

[11] 王守相，陈海文，潘志新，等. 采用改进生成式对抗网络的电力系统量测缺失数据重建方法[J]. 中国电机工程学报，2019，39(1)：56-64.
WANG Shouxiang，CHEN Haiwen，PAN Zhixin，et al. A reconstruction method for missing data in power system measurement using an improved generative adversarial network[J]. Proceedings of the CSEE，2019，39(1)：56-64(in Chinese).

[12] 杨玉莲，齐林海，王红，等. 基于生成对抗和双重语义感知的配电网量测数据缺失重构[J]. 电力系统自动化，2020，44(18)：46-54.
YANG Yulian，QI Linhai，WANG Hong，et al. Reconstruction of missing measurement data in distribution network based on generative adversarial network and double semantic perception[J]. Automation of Electric Power Systems，2020，44(18)：46-54(in Chinese).

[13] 郭小龙，李子康，刘灏，等. 基于增强生成对抗网络的 PMU 丢失数据恢复方法[J]. 电网技术，2022，46(6)：2114-2121.
GUO Xiaolong，LI Zikang，LIU Hao，et al. PMU missing data recovery algorithm based on enhanced generative adversarial network [J]. Power System Technology，2022，46(6)：2114-2121(in Chinese).

[14] 赵厚翔，沈晓东，吕林，等. 基于 GAN 的负荷数据修复及其在 EV 短期负荷预测中的应用[J]. 电力系统自动化，2021，45(16)：143-151.
ZHAO Houxiang，SHEN Xiaodong，LÜ Lin，et al. Load data restoration based on generative adversarial network and its application in short-term load forecasting of electric vehicle[J]. Automation of Electric Power Systems，2021，45(16)：143-151(in Chinese).

[15] ZHANG Wenjie，LUO Yonghong，ZHANG Ying，et al. SolarGAN：multivariate solar data imputation using generative adversarial network[J]. IEEE Transactions on Sustainable Energy，2021，12(1)：743-746.

[16] 刘科研，周方泽，周晖，等. 基于改进生成对抗网络的台区采集数据修复[J]. 电网技术，2022，46(8)：3231-3239.
LIU Keyan，ZHOU Fangze，ZHOU Hui，et al. Missing data imputation in transformer district based on improved generative adversarial network[J]. Power System Technology，2022，46(8)：3231-3239(in Chinese).

[17] GHADI M J，GHAVIDEL S，RAJABI A，et al. A review on economic and technical operation of active distribution systems[J]. Renewable and Sustainable Energy Reviews，2019，104：38-53.

[18] 栗峰，丁杰，周才期，等. 新型电力系统下分布式光伏规模化并网运行关键技术探讨[J/OL]. 电网技术，2023：1-12[2023-10-24]. https://doi.org/10.13335/j.1000-3673.pst.2023.0771.
LI Feng，DING Jie，ZHOU Caiqi，et al. Discussion on key technologies of large-scale grid-connected operation of distributed photovoltaic under the new-type power system[J/OL]. Power System Technology，2023：1-12[2023-10-24]. https://doi.org/10.13335/j.1000-3673.pst.2023.0771(in Chinese).

[19] HUANG Nantian，ZHAO Xuanyuan，GUO Yu，et al. Distribution network expansion planning considering a distributed hydrogen-thermal storage system based on photovoltaic development of the Whole County of China[J]. Energy，2023，278：127761.

[20] 乔颖，孙荣富，丁然，等. 基于数据增强的分布式光伏电站群短期功率预测(一)：方法框架与数据增强[J]. 电网技术，2021，45(5)：1799-1808.
QIAO Ying，SUN Rongfu，DING Ran，et al. Distributed photovoltaic station cluster gridding short-term power forecasting Part I：methodology and data augmentation[J]. Power System Technology，2021，45(5)：1799-1808(in Chinese).

[21] 林晨翔，王忠平，傅泓源. 提升分布式光伏功率预测精度[N]. 国

家电网报，2022-12-26(03).

[22] 张童彦，廖清芬，唐飞，等. 基于气象资源插值与迁移学习的广域分布式光伏功率预测方法[J/OL]. 中国电机工程学报，2023：1-12[2023-10-24]. https://doi.org/10.13334/j.0258-8013.pcsee.221950.
ZHANG Tongyan，LIAO Qingfen，TANG Fei，et al. Wide-area distributed photovoltaic power forecast method based on meteorological resource interpolation and transfer learning[J/OL]. Proceedings of the CSEE，2023：1-12[2023-10-24]. https://doi.org/10.13334/j.0258-8013.pcsee.221950(in Chinese).

[23] 万灿，宋永华. 新能源电力系统概率预测理论与方法及其应用[J]. 电力系统自动化，2021，45(1)：2-16.
WAN Can，SONG Yonghua. Theories，methodologies and applications of probabilistic forecasting for power systems with renewable energy sources[J]. Automation of Electric Power Systems，2021，45(1)：2-16(in Chinese).

[24] ZHANG Ruiyuan，MA Hui，HUA Wen，et al. Data-driven photovoltaic generation forecasting based on a Bayesian network with spatial‐temporal correlation analysis[J]. IEEE Transactions on Industrial Informatics，2020，16(3)：1635-1644.

[25] CAO Wei，WANG Dong，LI Jian，et al. BRITS：bidirectional recurrent imputation for time series[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal：Curran Associates Inc.，2018：6775-6785.

[26] LUO Yonghong，CAI Xiangrui，ZHANG Ying，et al. Multivariate time series imputation with generative adversarial networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal：Curran Associates Inc.，2018：1596-1607.

[27] 毕贵红，赵鑫，陈臣鹏，等. 基于多通道输入和 PCNN-BiLSTM 的光伏发电功率超短期预测[J]. 电网技术，2022，46(9)：3463-3476.
BI Guihong，ZHAO Xin，CHEN Chenpeng，et al. Ultra-short-term prediction of photovoltaic power generation based on multi-channel input and PCNN-BiLSTM[J]. Power System Technology，2022，46(9)：3463-3476(in Chinese).

[28] 杨晶显，张帅，刘继春，等. 基于 VMD 和双重注意力机制 LSTM 的短期光伏功率预测[J]. 电力系统自动化，2021，45(3)：174-182.
YANG Jingxian，ZHANG Shuai，LIU Jichun，et al. Short-term photovoltaic power prediction based on variational mode decomposition and long shortterm memory with dual-stage attention mechanism[J]. Automation of electric power systems，2021，45(3)：174-182(in Chinese).

[29] 王轲，钟海旺，余南鹏，等. 基于 seq2seq 和 Attention 机制的居民用户非侵入式负荷分解[J]. 中国电机工程学报，2019，39(1)：75-83.
WANG Ke，ZHONG Haiwang，YU Nanpeng，et al. Nonintrusive load monitoring based on sequence-to-sequence model with Attention mechanism[J]. Proceedings of the CSEE，2019，39(1)：75-83(in Chinese).

[30] 杨建，王力，宋冬然，等. 基于孤立森林与稀疏高斯过程回归的风电机组偏航角零点漂移诊断方法[J]. 中国电机工程学报，2021，41(18)：6198-6211.
YANG Jian，WANG Li，SONG Dongran，et al. Diagnostic method of zero-point shifting of wind turbine yaw angle based on isolated forest and sparse Gaussian process regression[J]. Proceedings of the CSEE，2021，41(18)：6198-6211(in Chinese).

[31] 郑欣彤，边婷婷，张德强，等. ARIMA 和 LSTM 方法长时间温度观测数据缺失值插补的比较[J]. 计算机应用，2022，42(S1)：130-135.
ZHENG Xintong，BIAN Tingting，ZHANG Deqiang，et al. Comparison of ARIMA and LSTM methods for interpolation of missing values of long-time temperature observations[J]. Journal of Computer Applications，2022，42(S1)：130-135(in Chinese).

廖若愚

在线出版日期：2023-11-15。
收稿日期：2023-09-18。

作者简介：

廖若愚(2000)，男，硕士研究生，主要研究方向为深度学习、电力大数据处理，E-mail：969386618@qq.com；

刘友波(1983)，男，教授，通信作者，博士生导师，主要研究方向为电力系统人工智能、低碳电力市场、分布式资源边缘控制，E-mail：liuyoubo@scu.edu.cn；

沈晓东(1975)，男，副教授，硕士生导师，主要研究方向为人工智能在电力系统中的应用，E-mail：shengxd@scu.edu.cn。

（责任编辑　徐梅
实习编辑　赵梓含）

附录 A

表 A1　超参数取值

Table A1　Hyperparameter tuning

参数	数值
hidden_size(编码器)	64
hidden_size(解码器)	64
优化器	Adam
训练次数	200
学习率	0.001
batch_size	16
patience	10
weight_decay	0.00001

表 A2　随机缺失下各场站插补误差

Table A2　Imputation error of each station under random missing data

缺失率/%	误差指标	A 区域				B 区域			
		场站 1	场站 2	场站 3	场站 4	场站 5	场站 6	场站 7	场站 8
20	e_{MAE}	0.0469	0.0359	0.0473	0.0389	0.0459	0.0531	0.0426	0.0533
	e_{RMSE}	0.0765	0.0564	0.0785	0.0685	0.0739	0.0829	0.0678	0.0847
30	e_{MAE}	0.0421	0.0455	0.0376	0.0384	0.0497	0.0395	0.0394	0.0410
	e_{RMSE}	0.0714	0.0761	0.0714	0.0684	0.0835	0.0619	0.0726	0.0719
50	e_{MAE}	0.0393	0.0513	0.0393	0.0387	0.0412	0.0523	0.0398	0.0625
	e_{RMSE}	0.0665	0.0919	0.0764	0.0676	0.0689	0.0961	0.0669	0.1153
70	e_{MAE}	0.0626	0.0705	0.0659	0.0696	0.0534	0.0569	0.0689	0.0772
	e_{RMSE}	0.1121	0.1351	0.1349	0.1453	0.0981	0.1057	0.1271	0.1611
80	e_{MAE}	0.1034	0.0831	0.1161	0.0873	0.1123	0.0741	0.0937	0.0961
	e_{RMSE}	0.1954	0.1656	0.2561	0.1936	0.2272	0.1467	0.1766	0.2138
90	e_{MAE}	0.1408	0.1307	0.1341	0.1442	0.1281	0.1612	0.1085	0.1209
	e_{RMSE}	0.2885	0.2674	0.2889	0.3069	0.2761	0.3391	0.2125	0.2509

表 A3　场站 1 随机缺失下不同算法性能对比

Table A3　Performance comparison of different algorithms under random missing data

缺失率/%	KNN		Missforest		GAIN		本文方法	
	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}
20	0.1946	0.4451	0.0564	0.1051	0.0536	0.0895	0.0469	0.0765
30	0.1989	0.4469	0.0664	0.1428	0.0557	0.0961	0.0421	0.0714
50	0.2233	0.5025	0.0876	0.2071	0.0598	0.1041	0.0393	0.0665
70	0.3483	0.6616	0.1597	0.3542	0.1691	0.3334	0.0626	0.1121
80	0.4266	0.8521	0.2602	0.5664	0.1772	0.3451	0.1034	0.1954
90	0.7593	1.3595	0.3964	0.7912	0.3397	0.6301	0.1408	0.2885

表 A4　场站 4 随机缺失下不同算法性能对比

Table A4　Performance comparison of different algorithms under random missing data

缺失率/%	KNN		Missforest		GAIN		本文方法	
	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}
20	0.1657	0.3671	0.0422	0.0992	0.0494	0.0853	0.0389	0.0685
30	0.1754	0.3956	0.0598	0.1471	0.0507	0.0897	0.0384	0.0684
50	0.2011	0.4452	0.0871	0.2043	0.0732	0.1276	0.0387	0.0676
70	0.3237	0.6445	0.1641	0.3827	0.0989	0.2136	0.0696	0.1453
80	0.6553	1.3245	0.2311	0.5271	0.1873	0.3732	0.0873	0.1936
90	0.7412	1.4054	0.3715	0.8019	0.2642	0.4262	0.1442	0.3069

表 A5　场站 1 整天缺失下不同算法性能对比

Table A5　Performance comparison of different algorithms under continuous missing data

缺失天数	KNN		Missforest		GAIN		本文方法	
	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}
1d	0.7851	1.4672	0.6823	1.2826	0.3894	0.6847	0.2224	0.4163
2d	0.8503	1.5431	0.6206	1.1887	0.3478	0.6549	0.2489	0.5263
3d	0.8705	1.5979	0.4967	0.8902	0.3662	0.6475	0.3588	0.6024
4d	0.8673	1.5869	0.5543	0.9923	0.4222	0.7102	0.3264	0.5748

表 A6　场站 4 整天缺失下不同算法性能对比

Table A6　Performance comparison of different algorithms under continuous missing data

缺失天数	KNN		Missforest		GAIN		本文方法	
	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}	e_{MAE}	e_{RMSE}
1d	0.7769	1.4421	0.4592	0.9462	0.3143	0.5637	0.3684	0.6142
2d	0.8237	1.5166	0.4972	0.9986	0.4191	0.7292	0.3009	0.5363
3d	0.7728	1.4495	0.6480	1.2878	0.5159	0.8882	0.3633	0.6377
4d	0.8839	1.6197	0.5747	1.1467	0.4637	0.7508	0.4386	0.7545

表 A7　气象缺失下各场站插补误差

Table A7　Imputation error of each station under missing meteorological data

缺失率/%	误差指标	A 区域				B 区域			
		场站 1	场站 2	场站 3	场站 4	场站 5	场站 6	场站 7	场站 8
20	e_{MAE}	0.0473	0.0364	0.0384	0.0401	0.0411	0.0409	0.0436	0.0394
	e_{RMSE}	0.0801	0.0623	0.0696	0.0702	0.0731	0.0732	0.0761	0.0705
30	e_{MAE}	0.0535	0.0481	0.0417	0.0427	0.0490	0.0362	0.0589	0.0529
	e_{RMSE}	0.0902	0.0837	0.0763	0.0756	0.0827	0.0652	0.0941	0.0902
50	e_{MAE}	0.0661	0.0644	0.0619	0.0539	0.0731	0.0601	0.0687	0.0707
	e_{RMSE}	0.1155	0.1087	0.1096	0.0938	0.1276	0.1043	0.1185	0.1296
70	e_{MAE}	0.0988	0.1155	0.1003	0.1027	0.1058	0.1208	0.0945	0.1148
	e_{RMSE}	0.1741	0.1981	0.1904	0.1763	0.1932	0.2175	0.1702	0.2128
80	e_{MAE}	0.1894	0.1429	0.1608	0.1508	0.1696	0.1712	0.1855	0.1837
	e_{RMSE}	0.3343	0.2529	0.3015	0.2805	0.3024	0.3082	0.3305	0.3204
90	e_{MAE}	0.3109	0.2506	0.3311	0.2981	0.4331	0.3617	0.3193	0.4314
	e_{RMSE}	0.5287	0.4372	0.5926	0.5148	0.7377	0.6241	0.5361	0.6980

附录 B

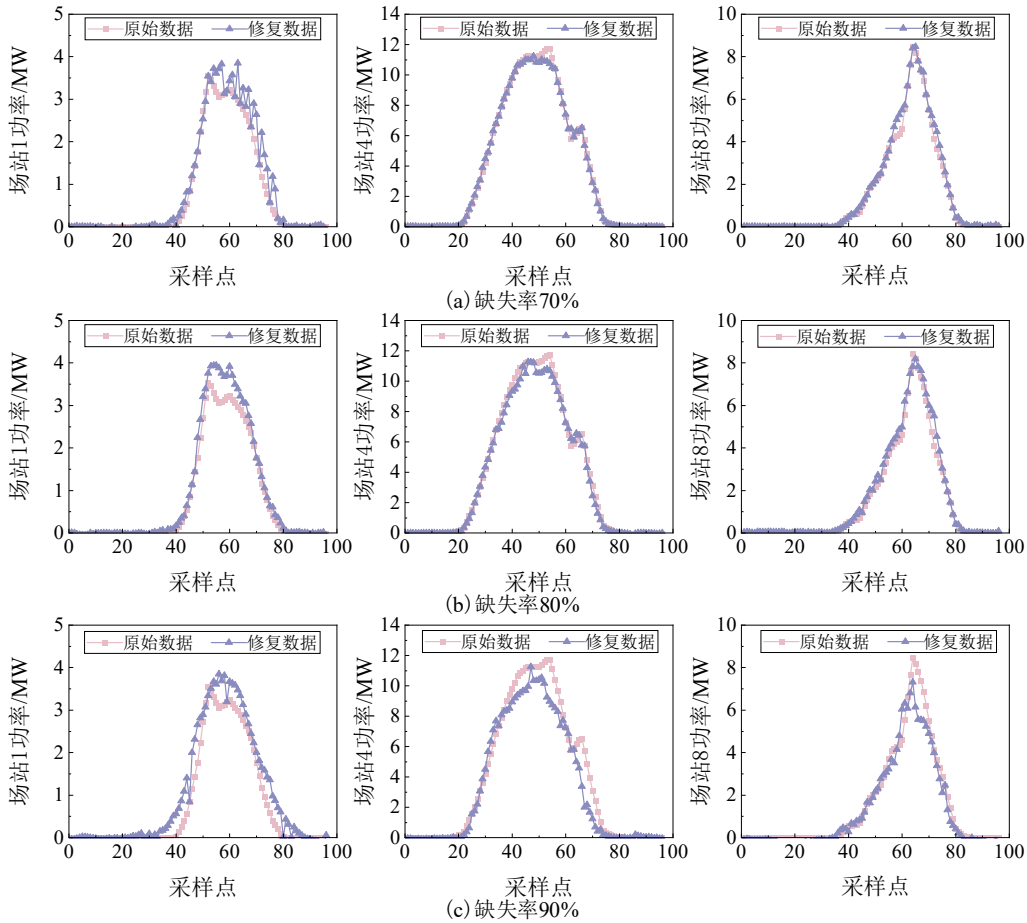


图 B1　气象缺失下各场站插补效果

Fig. B1　Missing data reconstruction results of each station under missing meteorological data

1 消融实验

①单阶段插补：第一阶段插补基于光伏数据的历史时序特性，该方法基于文献[31]提出的用于时间序列数据插补的双向长短期记忆网络。②二阶段插补：第一阶段插补获取补码向量后，第二阶段插补基于特征提取层捕捉多维量测值之间的耦合关系，对缺失数据再次插补。③多阶段插补：为本文所提方法，根据缺失模式信息自适应集成前两阶段插补值，输出最终的重建数据。分别采用三种算法进行仿真实验，场站1、场站 4 在不同缺失率下的测试结果见表 B1。

表 B1　算法的插补误差对比

Table B1　Comparison of imputation errors of different algorithms

场站	算法	平均插补误差/MW					
		20%缺失率	30%缺失率	50%缺失率	70%缺失率	80%缺失率	90%缺失率
场站 1	单阶段插补	0.0694	0.0587	0.0618	0.1432	0.2364	0.4386
	二阶段插补	0.0873	0.0714	0.1063	0.1594	0.3274	0.3587
	多阶段插补	0.0469	0.0421	0.0393	0.0626	0.1034	0.1408
场站 4	单阶段插补	0.0483	0.0597	0.0523	0.1038	0.2014	0.3275
	二阶段插补	0.0432	0.0638	0.1362	0.2337	0.3126	0.3897
	多阶段插补	0.0389	0.0384	0.0387	0.0696	0.0873	0.1442

2 噪声实验

共设置 4 种噪声：①噪声 1：均值为 0，标准差为 0.05；②噪声 2：均值为 0，标准差为 0.08；③噪声 3：均值为 0，标准差为 0.1；④噪声 4：均值为 0.05，标准差为 0.05；分别采用本文方法、GAIN 和 Missforest 进行测试，场站 1、场站 2、场站 3、场站 4 在 50%缺失率下的仿真结果分别如表 B2 所示。

表 B2　算法的插补误差对比

Table B2　Comparison of imputation errors of different algorithms

场站	算法	平均插补误差/MW				
		无噪声添加	噪声 1	噪声 2	噪声 3	噪声 4
场站 1	本文方法	0.0393	0.0405	0.0642	0.0806	0.0407
	GAIN	0.0598	0.0608	0.0664	0.0813	0.0635
	Missforest	0.0876	0.1271	0.1462	0.1616	0.1253
场站 2	本文方法	0.0513	0.0546	0.0783	0.0973	0.0646
	GAIN	0.0569	0.0721	0.1067	0.1098	0.0885
	Missforest	0.0836	0.1213	0.1378	0.1564	0.1193
场站 3	本文方法	0.0393	0.0504	0.0843	0.1034	0.0584
	GAIN	0.0683	0.0801	0.0905	0.1083	0.0751
	Missforest	0.0871	0.1183	0.1415	0.1526	0.1135
场站 4	本文方法	0.0387	0.0486	0.0847	0.0924	0.0563
	GAIN	0.0732	0.0823	0.1032	0.1128	0.0836
	Missforest	0.0871	0.1224	0.1453	0.1573	0.1246

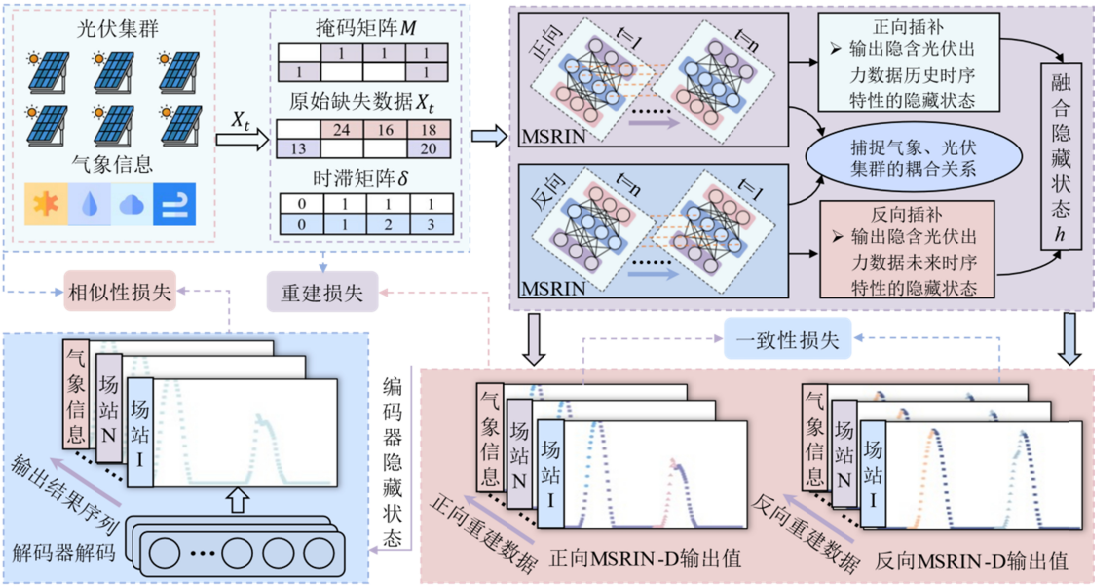


图 B2　数据增强模型整体架构

Fig. B2　Overall architecture of the data augmentation model

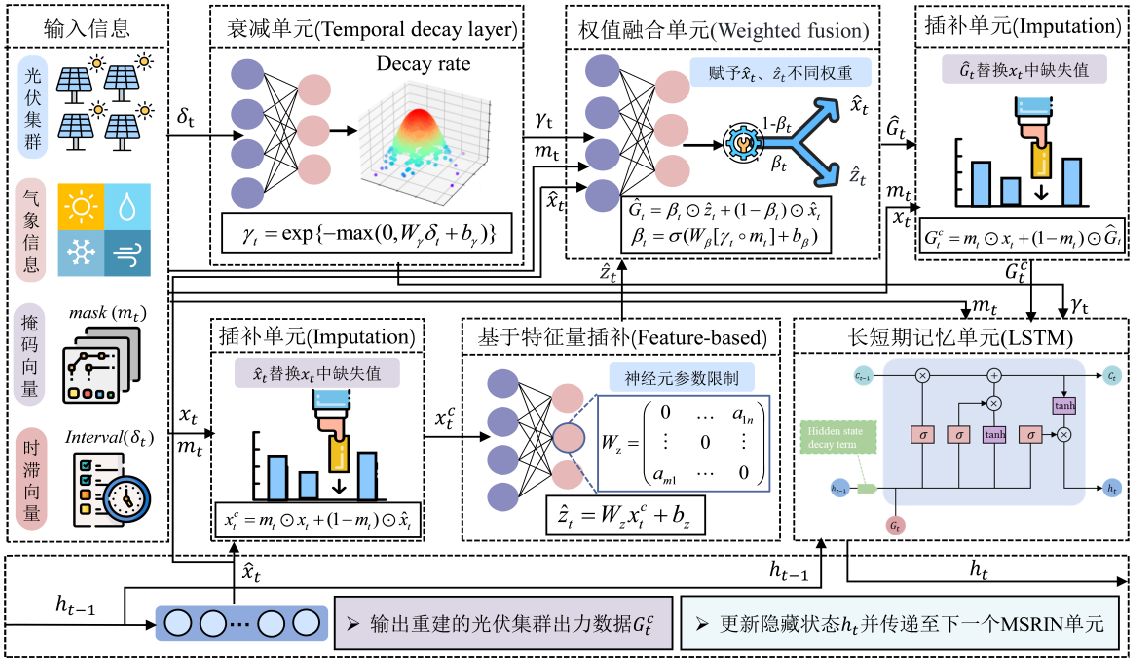


图 B3　MSRIN-D 结构图

Fig. B3　MSRIN-D network structure

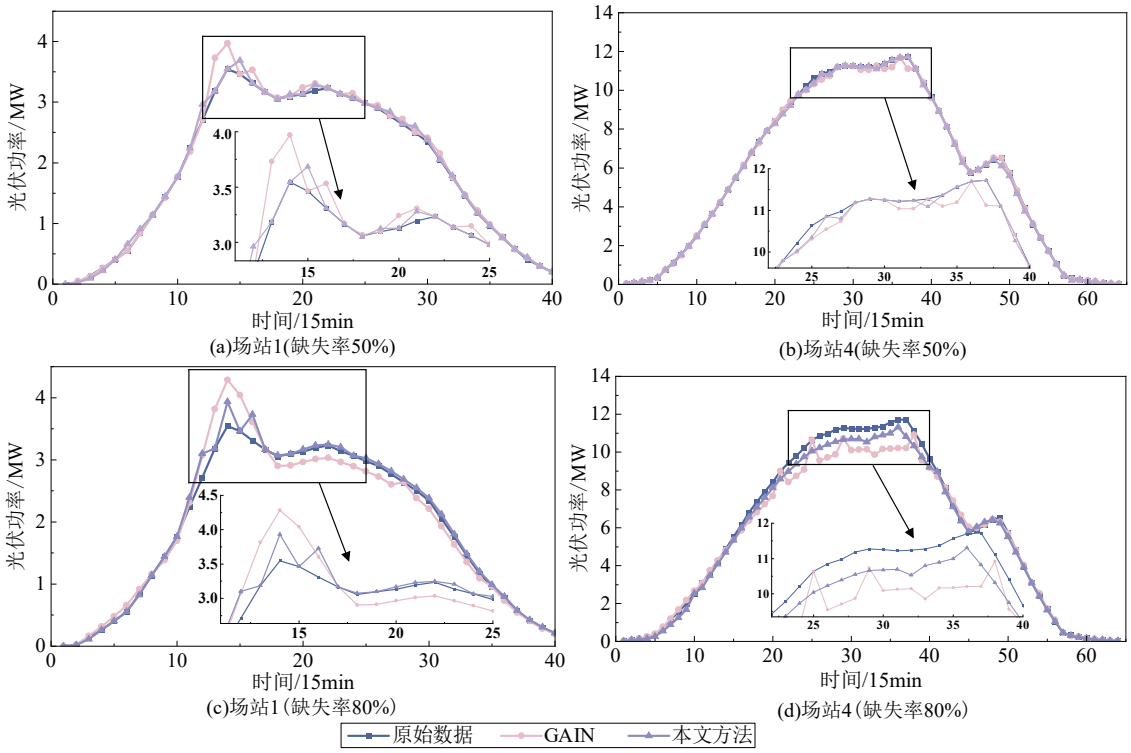


图 B4　随机缺失数据重建效果图

Fig. B4　Random missing data reconstruction results

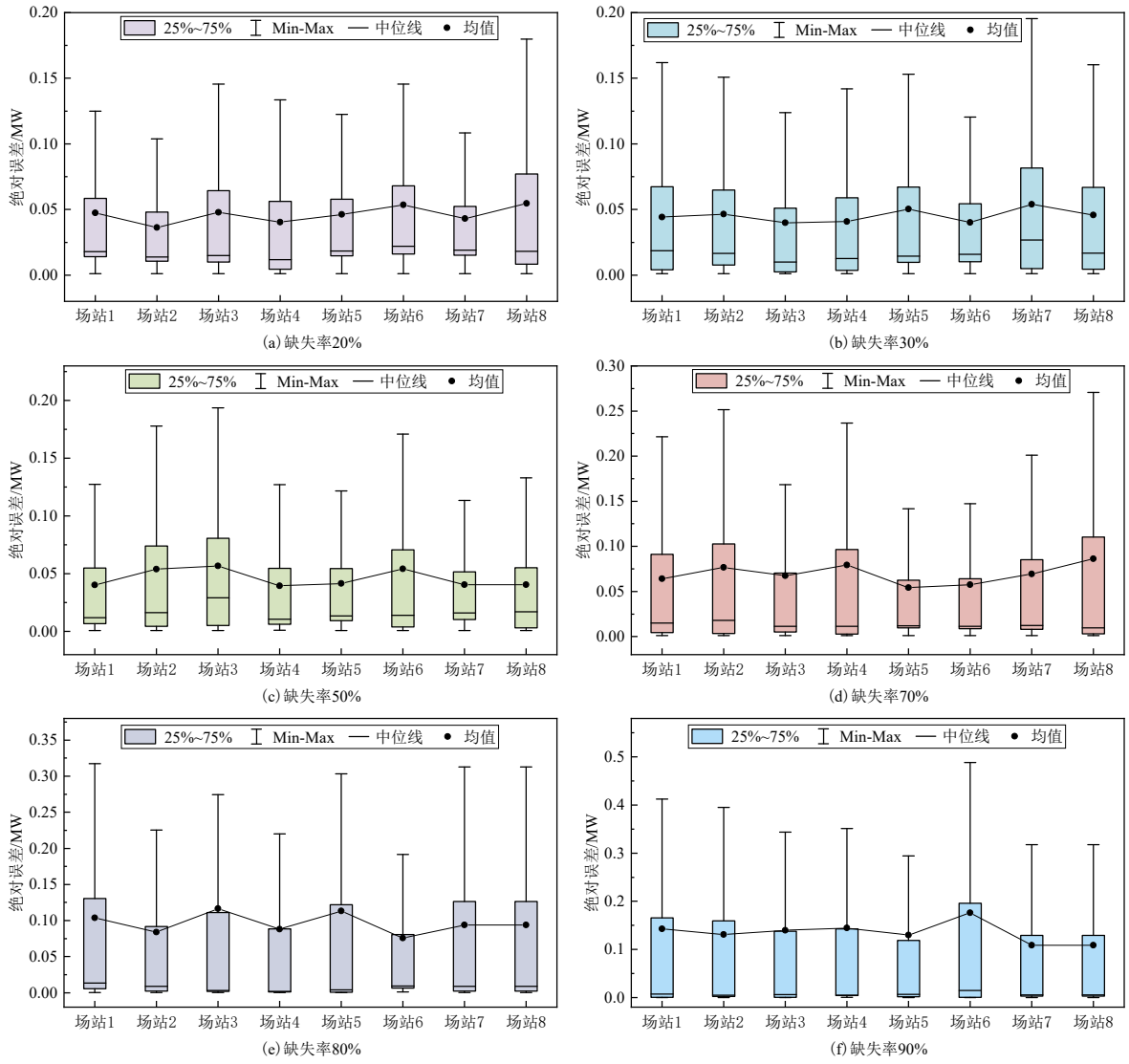


图 B5　数据插补误差分布图

Fig. B5　Imputation error analysis