中国科学: 技术科学 2026 年 第 56 卷 第 4 期: 1107 ~ 1118
SCIENTIA SINICA Technologica techcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
论 文 西安交通大学建校130周年暨西迁70周年机械工程学科专刊

面向复杂机电系统多维不平衡数据增强的流形感知扩散方法

李江波[1], 高建民[1], 汪臻[2], 张问采[3], 常振元[4], 张珺倩[3], 王荣喜[1*]

1. 西安交通大学机械工程学院, 西安 710049
2. 西安热工研究院有限公司, 西安 710032
3. 中国石油集团渤海石油装备制造有限公司, 天津 300457
4. 陕西省特种设备检验检测研究院, 西安 710049
* E-mail: wangrongxi@xjtu.edu.cn

收稿日期: 2025-12-31; 接受日期: 2026-02-05; 网络版发表日期: 2026-03-27
陕西省重点研发计划项目(编号: 2024QY2-GJHX-02)、中国华能集团有限公司科技项目(编号: HNKJ24-H86)和中国石油集团关键核心技术研发项目(2025ZG77)资助

摘要 工业现场监测数据普遍存在的样本极度不平衡问题, 严重制约了诊断模型的决策可靠性. 针对现有生成模型在处理高维监测数据时, 由于难以有效解耦复杂的观测空间, 导致数据生成模型易陷入模式坍塌且保真度不足, 本文提出一种流形学习赋能的条件扩散生成新范式(MC-CESDM). 该方法通过融合流匹配(flow matching)理论与特征解耦机制, 有效攻克了现有模型的局限性: 流匹配通过构建确定性概率路径避免了冗长的迭代噪声预测, 大幅提升了训练稳定性与采样速度; 特征解耦则确保了生成样本能够精准受控于变工况等连续物理属性. 本研究以典型的风电机组监测数据为案例进行实验验证, 并与主流生成算法进行对比. 结果表明, 该模型在特征关联度与分布距离(Fréchet distance)等关键指标上均取得显著领先, 证明了其优异的生成保真度. 该方法克服了复杂工况下数据生成的效率与质量瓶颈, 为机电系统不平衡数据的样本增强提供了高效、可靠的技术工具, 为提升工业设备智能运维水平奠定了理论基础.

关键词 复杂机电系统, 监测数据, 流形学习, 条件扩散模型, 数据增强

1 引言

风电机组作为典型的复杂机电系统, 长期运行于非平稳、强随机的运行工况中, 监测数据呈现出显著的高维性与非线性耦合特征, 如图1所示, 在实际工业场景中, 监测数据的分布规律极度复杂, 常呈现非高斯分布特性: 数据在常规运行区间高度密集, 而在关键故障区间则表现得极为稀疏.

这种数据分布的不平衡性给当前的数据驱动方法带来了严峻挑战. 观测样本的稀缺性, 特别是故障特征样本的严重匮乏, 制约深度学习模型对高维变量间复杂演化机制的精确捕获, 导致模型在极端工况下的泛

引用格式: 李江波, 高建民, 汪臻, 等. 面向复杂机电系统多维不平衡数据增强的流形感知扩散方法. 中国科学: 技术科学, 2026, 56: 1107–1118
Li J B, Gao J M, Wang Z, et al. Manifold-aware diffusion method for multi-dimensional unbalanced data enhancement of complex electromechanical systems (in Chinese). Sci Sin Tech, 2026, 56: 1107–1118, doi: 10.1360/SST-2025-0459

© 2026 《中国科学》杂志社

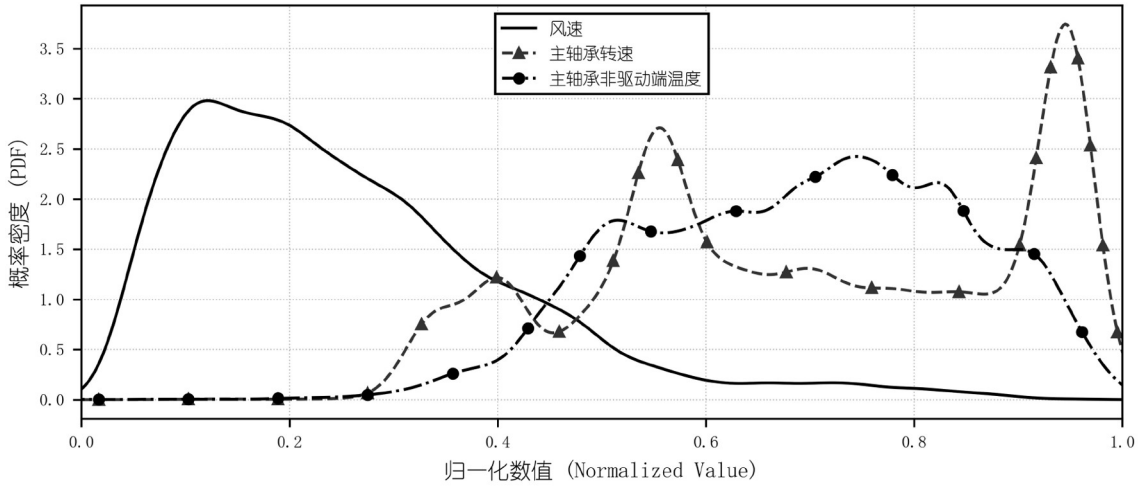

图 1 实际风机SCADA监测系统数据分布示例

Figure 1 A practical example of wind turbine SCADA monitoring system data.

化能力与决策可靠性不足. 因此, 如何克服数据不平衡这一瓶颈、实现高质量样本的增强与生成, 已成为提升机电系统智能化运维水平的关键技术路径.

针对故障样本稀缺问题, 现有的数据增强方法主要可归纳为三类. (1) 基于信号变换的传统增强方法, 主要通过添加随机噪声、时间规整或特征扰动等手段对原始信号进行扩充. 针对跨设备少样本故障诊断中忽视数据分布偏移、诊断效果欠佳的问题, 夏鹏程等人[1]提出一种数据增强判别元学习方法. 然而, 它们的核心局限在于缺乏对数据高阶非线性特征的捕捉能力. (2) 基于统计理论的重采样方法, 典型代表为合成少数类过采样技术(synthetic minority oversampling technique, SMOTE)及其变式. 针对实际工况中滚动轴承正常与故障样本数量严重不均衡, 导致诊断模型泛化能力差的问题, 孟宗等人[2]提出二次增强的数据扩充策略以实现故障样本数据增强. 但是, 这种插值操作忽略了数据的真实流形结构, 极易将合成样本推离真实数据的非线性流形, 从而导致样本的物理不一致性. (3) 基于深度学习的生成式增强方法, 是目前研究的前沿方向. 常用的生成式网络架构包括生成对抗网络类(generative adversarial networks, GANs)、变分自编码器类(variational autoencoders, VAEs)、扩散模型(diffusion model, DM)等.

其中, GANs借助博弈论中的双人极小极大博弈机制, 通过生成器与判别器间的动态对抗演化, 在追求纳什均衡的过程中展现出捕获复杂数据分布及生成高保真样本的卓越能力. Wang等人[3]提出了一种GAN变式, 能够生成紧密模仿真实数据特征的合成数据, 直接缓解多维时间序列数据样本不平衡导致的故障数据稀缺困境. 袁烨等人[4]提出了一种结合门控循环单元的GAN变式, 生成了紧密模仿真实数据特征的合成数据. 张晟斐等人[5]提出了一种利用Wasserstein GAN补全扩充非理想监测时序、改善数据质量的方法. 为了解决传统生成对抗网络无法有效处理时间序列数据固有特性这一问题, Neubürger等人[6]提出了TimeGAN, 并将其应用在高维工业时序数据领域. 为应对多故障类别下的数据严重不平衡和TimeGAN训练不稳定的问题, Deng等人[7]提出了一种带有梯度惩罚的条件TimeGAN, 成功实现对特定故障类别稀缺样本的精确控制和生成. 针对风电机组故障检测中数据不平衡和传统重采样技术导致过拟合的局限性, Chatterjee等人[8]提出了一种基于Wasserstein条件生成对抗网络的数据增强方法, 用于提高生成样本的真实性. 然而, 这种基于博弈的训练机制在面对高维非平稳的机电监测数据时, 往往难以维持稳定的纳什均衡, 极易导致模型陷入模式坍塌[9~11], 在风电机组等典型机电系统中尤为显著. 由于其运行工况受自然环境驱动而表现出高度的非平稳性与多变性, 监测信号在观测空间中的演化路径极度复杂, 进一步加剧了故障特征在全局数据分布中的稀疏性.

相较于生成对抗网络在非平稳博弈下易出现的模式坍塌问题，VAEs通过引入Kullback-Leibler散度作为正则化约束，强制隐空间分布向先验概率对齐具有更强的空间探索能力与概率平滑性，进而提高数据质量. 针对有限故障数据问题，Zhao等人[12]提出了一种改进的Wasserstein自编码器用于故障诊断的合成样本. 为解决机械系统的小样本故障诊断难题，Liu等人[13]开发了一种条件变分自编码器变体以增强生成样本的多样性和类别区分性. Lei等人[14]采用一种基于变分自编码器生成对抗网络的方法，利用VAE和GAN生成具备高度多样性和高度逼真性的样本. 针对零样本故障诊断这一更具挑战性的场景，Fan等人[15]构建了卷积残差SE-注意力VAE模型，实现对陌生故障类别的诊断. 但是，VAEs难以处理高维时间序列数据，在复杂的、连续变化的运行工况上，模型的可控性、生成精度及保证度受到潜在空间建模能力不足的限制[16~18].

尽管以GANs及VAEs为代表的生成模型在时序合成领域取得了显著进展, 但在机电系统复杂工况下, 仍面临着训练失稳导致的模式坍塌、物理耦合逻辑的表征失真以及底层本质流形刻画不足等严峻瓶颈[19,20].

针对机电系统监测数据高噪声、强非平稳导致的生成难题，鉴于扩散模型(DMs)在建模高维数据特征相关性上的卓越能力，学术界正逐步转向对其展开深入研究. 扩散模型凭借其严谨的概率演化框架，克服了传统生成网络难以兼顾样本多样性与保真度的固有局限. 时间序列扩散模型Diff-MTS等的研究表明, 扩散模型在扩充训练数据集、提升诊断模型性能方面已取得令人满意的初步成果，展现出比GANs和VAEs更具优势的高质量数据生成能力[21]. 然而，尽管生成质量卓越，DMs在实际应用中仍面临极高的计算成本与缓慢的采样速度这一核心矛盾. 针对此瓶颈，图像生成领域提出了以低维流形思想为核心的图像生成解决方案[22]，核心逻辑在于：复杂机电系统的时变工况数据的演化并非在观测空间内随机分布，受限于系统内在的物理动力学约束，其本质特征被约束在一个低维嵌入流形之中. 因此，通过在低维空间而非原始高维空间进行去噪和预测，不仅能够大幅降低模型参数量, 还能有效解决扩散模型计算效率低下的诟病.

本研究设计了一种流形赋能下的复杂机电系统监测数据生成方法, 采用低维流形思想, 构建新型时间序列生成方法MC-CESDM(manifold constrained complex electromechanical systems diffusion model), 生成实际运行工况下的运行监测数据. 本文主要的贡献如下.

(1) 提出了一种基于物理引导的条件扩散生成框架MC-CESDM. 针对监测数据的高维耦合特性，本研究并未直接生成全量数据，而是采用7维运行条件作为数据的内在流形坐标，在此基础上建立条件驱动扩散模型生成高维监测变量，能够有效解耦工况变化与故障特征，确保生成数据在物理逻辑上的一致性.

(2) 引入流形正则化约束以缓解小样本过拟合问题. 为了克服传统生成模型在稀缺故障样本上易发生的模式崩塌问题，设计了流形正则化项. 通过约束生成过程遵循数据的局部几何结构，使得模型在有限样本下仍能捕捉到监测变量间复杂的非线性机理关系, 显著提升模型在极不平衡数据集上的泛化能力.

(3) 实验表明，MC-CESDM生成的合成样本具有极高的保真度和多样性, 将其用于扩充训练集, 能够显著改善下游深度诊断模型在复杂时变工况下的识别精度.

2　MC-CESDM理论基础

2.1　扩散模型前向过程原理简述

时间序列扩散模型实现了从随机噪声到目标样本数据的生成，可以分为两个步骤：前向过程和反向过程, 本质上是马尔科夫过程[23].

前向过程是对时间序列逐步加噪的过程，当前时刻的序列x_t只与上一个时刻x_{t-1}有关:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}z_1, \tag{1}$$

其中，α_t为信号保留系数，$1-\alpha_t=\beta_t$为噪声调度参数, 其中, 不同时间步的概率关系如下:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \tag{2}$$

$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{\alpha_t}x_{t-1}, \beta_t I), \tag{3}$$

其中，$q(x_t \mid x_{t-1})$为单步转移分布，N为正态分布，I为单位矩阵, 体现每步扩散过程的独立性.

根据式(1)~(3)，初始时刻和t时刻的关系表达如式(4)所示:

$$q(x_t \mid x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha})I), \tag{4}$$

其中, $q(x_t \mid x_0)$为任意步转移分布, $\bar{\alpha}_t$为累积信号保留系数.

2.2 扩散模型反向过程原理简述

反向过程是一个去噪的过程, 核心在于通过从随机噪声序列x_T中逐步还原x_0, 达到数据生成的目的. 其中, 时刻x_{t-1}与后一时刻的关系x_t如式(5)所示:

$$q(x_{t-1} \mid x_t, x_0) = N(x_{t-1} \mid \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I), \tag{5}$$

其中, 为了获得生成过程中的$\tilde{\mu}_t$和$\tilde{\beta}_t$, 构建式(6)和(7):

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t, \tag{6}$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right). \tag{7}$$

2.3 MC-CESDM流形理论

MC-CESDM理论建立在流形假设的基础上, 即对于监测数据$\mathcal{X} \subset \mathbb{R}^D$嵌入在一个高维空间$\mathbb{R}^D$中, 但监测数据本质上分布在一个维度$d \ll D$的低维子流形$\mathcal{M}$上. 例如, 本文分析的实际运行监测数据$\mathcal{X} \subset \mathbb{R}^D$. 生成对应时变工况下的高维监测数据要求, 时间序列扩散模型通过预测噪声实现图像生成的方法, 需要构建能够观测$\Omega(\Omega >> D)$维空间特征的生成网络模型, 而流形理论下的扩散模型只需构建$\Omega'(\Omega' > d)$维空间特征的网络架构即可实现高维特征数据的生成.

3 MC-CESDM模型设计

3.1 架构设计

为了精确逼近运行监测数据在高维观测空间中的内在低维流形拓扑结构, 本研究构建了一种面向时序特征深度适配的U-DiT(U-shaped diffusion transformer)生成网络架构, 如图2所示. 编码器通过层级化的下采样操作, 将高维、含噪的SCADA信号压缩至紧凑的潜在空间, 强迫模型学习数据的本质低维特征; 解码器负责将潜在特征逐步映射回原始数据空间. 横向的跳跃连接(skip connections)在对应分辨率的特征层之间构建信息高速通路, 能有效缓解深层网络中的梯度消失问题, 并保留重要的高频微弱信号细节. 相比于传统U型架构, 模型在具备多尺度建模能力的基础上, 能协同捕捉局部瞬态冲击与全局演化趋势, 结合Transformer的长程依赖建模优势, 实例验证结果显示能够显著提高在复杂时变工况下的特征解耦效率与计算稳定性.

3.1.1 U型架构设计

由于U型架构为经典的编码解码架构, 其计算复

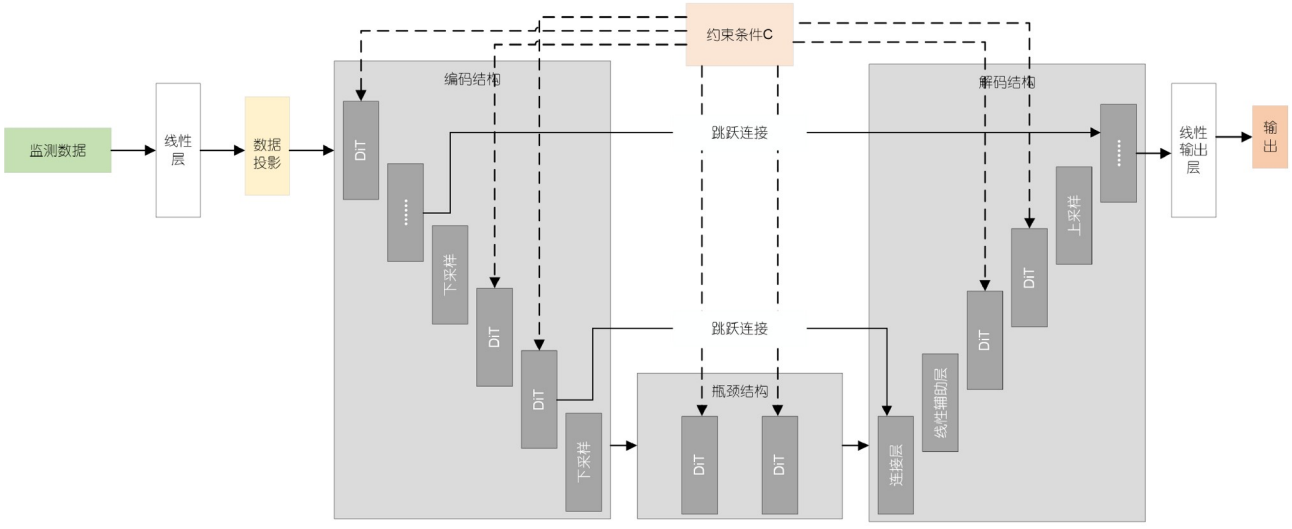

图 2 (网络版彩图)MC-CESDM模型架构示意图(显示主要模块)
Figure 2 (Color online) MC-CESDM model architecture diagram (the main module is shown).

杂度低，本文采用的模型架构基于原有的扩散模型架构，使用Encoder-Decoder架构，在长时间序列中也能用低参数模型训练生成高维运行状态监测数据.

(1) 层级化特征映射: 左侧的Encoder通过多层DiT模块和下采样操作，逐层压缩输入数据的时空分辨率，增加通道维度，本质是一个将高维观测数据映射到低维潜在流形空间的过程.

(2) 跳跃连接: 图中间的横向连接直接将编码器层的特征拼接到对应的解码器层，防止深层网络结构导致的高频特征细节丢失, 提高数据生成的保真度.

(3) 镜像重构: 右侧的Decoder通过对称的DiT模块和上采样操作, 将潜在特征逐步恢复至原始信号维度.

(4) 瓶颈结构: 迫使模型剔除冗余噪声，仅保留决定运行状态的最本质、最致密的低维流形特征.

U型结构使得模型能够同时处理全局语义信息和局部细节信号, 对于时间序列生成至关重要, 既能降低由于长时间序列的输入而导致的模型难以训练，同时也便于提取其中的全局与局部特性.

3.1.2　条件嵌入与DiT模块的集成建模

构建条件嵌入层，如图3所示，作为提取运行条件参数, 结合DiT注意力机制生成时变工况的数据.

(1) 条件嵌入层(约束条件C): 通过AdaLN, 条件信息c，包含时间步和工况参数，被以调节归一化参数的方式注入特征图x中，使模型能够将由x携带的特征内容信息与由c携带的运行工况条件信息有效地解耦, 专注于学习速度场v.

(2) DiT模块: 与传统的Transformer block区别在于使用AdaLN结合门控机制结构处理条件信息c, 而非简单的拼接或者交叉注意力机制[24].

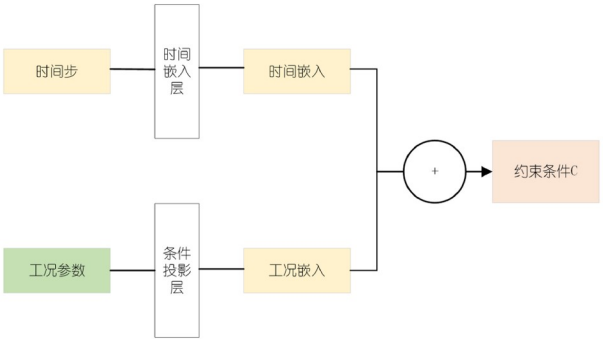

图 3　(网络版彩图)条件嵌入模块

Figure 3　(Color online) Conditional embedding module.

对于机电系统运行监测数据，时间相关性与特征相关性是数据生成的两个重要的考量指标. 时间相关性表示的是在时间序列中，当前时间T对应的监测数据不仅受当前运行条件影响，还与前一时间间隔$T-1$对应的监测状态有关. 特征相关性是指高维监测数据中的多变量关联关系, 这也是高维数据生成的难点.

3.2　以真实数据为目标的扩散生成规则优化

3.2.1　模型预测对象与损失函数设计

在t时刻带噪数据可表示为

$$z_t = tx + (1-t)\epsilon, \tag{8}$$

其中, x为不含噪的时间序列, ϵ为高斯噪声, z_t为t时刻的带噪数据, $t \in [0, 1]$. 在$t = 0$时刻, z_t为纯高斯噪声数据, 在$t = 1$时刻, z_t是真实不含噪数据.

为了降低模型复杂度, 基于流形理论, 改变传统扩散模型在时间序列上预测噪声ϵ的思想, 转而预测真实不含噪数据x, 生成网络架构结果表示为

$$x_\theta = \mathrm{net}(z_t, t). \tag{9}$$

损失函数设计分为三类，x、ϵ及v三类损失，但其本质上可以相互推导得知，其中v是z_t对时间的导数, 可以理解为时间步之间的转移速度. 原理如式(10)所示:

$$v = z_t' = x - \epsilon. \tag{10}$$

结合式(8)可以得到三种损失之间的转换关系.

在模型设计中, 采用以时间序列x为目标, 速度场v为损失函数Loss_v的模型架构, 其中:

$$\mathrm{Loss}_v = \mathrm{MSE}(v_{\mathrm{pred}}, v_{\mathrm{target}}). \tag{11}$$

3.2.2　扩散过程的形式化推导

训练一个模型来预测从噪声(x_1)到数据(x_0)之间线性轨迹上的速度向量, 步骤如下.

(1) 初始化模型和优化器，设置网络参数、AdamW和学习率调度器等.

(2) 采样时间步t设置，$t \in [0, 1]$，且服从logit-Normal分布.

(3) 构造ϵ，并赋值x，使得$x_1 = \epsilon \sim \mathrm{Normal}(0, 1)$, $x_0 = x$，在$t = 0$时刻，z_t为x_1纯噪声数据，在$t = 1$时刻,

z_t为x_0真实数据.

(4) 构造插值轨迹z_t, 同式(8)表示式(12):

$$z_t = tx_0 + (1-t)x_1. \tag{12}$$

(5) 计算目标速度v_{target}, 由式(12)求导得到:

$$v_{\text{target}} = x_0 - x_1. \tag{13}$$

(6) 条件掩码: 使用CFG训练策略, 并在时变工况下引入连续条件Dropout, 使得模型能够学习在有条件约束时的条件与监测数据之间的关联关系, 也能够根据真实的数据分布生成结果.

(7) 前向传播, 预测速度场算法如式(14)所示:

$$v = \text{Model}\left(z_t,\ t,\ c_{\text{Dropout}}\right). \tag{14}$$

计算速度场损失Loss_v, 如式(15)所示:

$$\text{Loss}_v = \text{MSE}(v_{\text{pred}},\ v_{\text{target}}). \tag{15}$$

(8) 反向传播优化, 更新模型、调度器等参数设置.

3.2.3 生成过程的迭代化建模

利用训练好的速度场, 解常微分方程(ODE), 从纯噪声生成符合条件的数据, 步骤如下.

(1) 准备: 生成数据对应的运行工况标签, Model及超参数等.

(2) 生成$t = 0$时刻的z_0纯噪声数据, 如式(16):

$$z_0 \sim \text{Normal}(0,\ 1). \tag{16}$$

生成时间步序列$t \in (0,\ 1)$, 从0趋近1.

(3) 逐步求解$z_0 -> z_1$, 使用ODE数值积分方法计算速度场v, 如式(17)~(19)所示:

$$v_{\text{cond}} = \text{Model}(z, t, c_{\text{test}}), \tag{17}$$

$$v_{\text{uncond}} = \text{Model}(z, t, \text{Zero}_{\text{Vector}}), \tag{18}$$

$$v = (1 + \text{CFG}_{\text{Scale}}) \cdot v_{\text{cond}} - \text{CFG}_{\text{Scale}} \cdot v_{\text{uncond}}, \tag{19}$$

其中, Model为训练好的扩散模型, z为当前含噪数据, c_{test}为条件输入, $\text{Zero}_{\text{Vector}}$为零向量, 用于表示无条件生成条件, $\text{CFG}_{\text{Scale}}$为一个尺度参数, 用于控制条件生成的强度.

(4) 计算下一步的z_t, 由前一刻z_{t-1}得到, 如式(20)和(21)所示:

$$z_t = z_{t-1} + v \cdot \mathrm{d}t \quad \text{(euler)}, \tag{20}$$

$$\begin{aligned} &z_{\text{guess}} = z_{t-1} + v_1 \cdot \mathrm{d}t, \\ &\text{输入步骤(3)} z_{\text{guess}} \oplus t \Rightarrow v_2, \\ &z_t = z_{t-1} + 0.5 \cdot (v_1 + v_2) \cdot \mathrm{d}t, \quad \text{(heun)}. \end{aligned} \tag{21}$$

(5) 反归一化, 生成时变工况数据.

4 结果与讨论

4.1 数据与参数设置

数据来源于华能集团某风场真实风机数据, 数据预处理后共计39070条数据, 如表1所示, 使用其中7维观测条件和16维系统响应特征作为模型的约束条件和生成数据的维度.

MC-CESDM模型参数的设置对模型训练以及生成性能影响显著, 时间步T越大, 生成轨迹越接近连续的真实轨迹, 精度越高, 生成质量越好. 学习率是控制模型参数更新幅度的核心参数, MC-CESDM模型采用Warmup+余弦衰减策略, 能够很好地控制训练过程收敛性能. 嵌入层大小、隐藏层大小、注意力头等超参数设置根据风机数据大小优化调整, 实验发现在瓶颈层, 即图2中的瓶颈结构, 增加注意力头数能够显著提升模型训练及数据生成性能.

设备型号对于模型的训练及生成效果有明显差异, 实验采用RTX 3080 Ti-12GB和RTX 5060-8GB分别训练生成数据, 从数据质量来说, RTX 3080 Ti-12GB布置的模型性能优于RTX 5060-8GB的, 实验对比部分所有模型训练及数据生成都在同一台设备中进行以保持对比结果的严谨性.

4.2 实验对比

MC-CESDM模型对比改进的条件生成对抗网络C-GAN、改进的条件变分自编码器C-VAE以及时间序列扩散生成模型DM-MTS在风机实际数据中训练生成对应工况下的数据.

现有生成模型在风机SCADA监测数据上主要面临两大挑战: 其一, 风机运行属于时变工况, 监测数据受工况影响显著, 而条件生成对抗网络C-GAN等模型因参数量有限, 难以充分表征其复杂特征; 其二, 模型常需从低维观测条件7维学习并生成高维空间特征16维, 存在表征能力不足的困难.

为系统验证本文所提MC-CESDM模型的优势, 将

表 1　数据介绍

Table 1　Dataset introduction

数据类型	监测变量名
观测条件(7维)	平均风速(WS)、平均风向(WD)、平均有功功率(AP)、平均无功功率(RP)、平均环境温度(AT)、平均机舱温度(CT)、桨叶1平均桨距角(P1-PA)
系统响应特征(共计16维)	主轴轴承平均转速(MS)、发电机平均转速(GS)、主轴前轴承平均温度(MS-FBT)、主轴后轴承平均温度(MS-RBT)、发电机驱动端平均轴承温度(G-DBT)、发电机非驱动端平均轴承温度(G-NDBT)、电网AB相线平均电流(Grid-AB)、齿轮箱驱动端平均轴承温度(GB-DBT)、齿轮箱非驱动端平均轴承温度(GB-NDBT)、齿轮箱油池平均温度(GB-OT)、发电机U相绕组平均温度(G-UWT)、发电机V相绕组平均温度(G-VWT)、发电机W相绕组平均温度(G-WWT)、电网BC相线平均电流(Grid-BC)、平均偏航误差(YE)、电网CA相线平均电流(Grid-CA)

从模型生成数据质量层面，与C-GAN、C-VAE和DM-MTS基准模型进行对比. 生成数据的评估需要联系风机SCADA数据特性，考核标准包括：(1) 生成数据各监测维度间及时间序列上的相关性保持; (2) 生成数据与真实测试数据的分布一致性; (3) 生成结果的可视化保真度.

特征相关性矩阵差异度Diff结果如表2所示，从特征相关性矩阵分析中，可以看出MC-CESDM和DM-MTS模型可以很好地识别高维数据生成中特征之间的相互关联，这是扩散模型在高维时间序列生成上的优势.

通过计算自相关系数误差ACF Error、动态时间规整距离DTW及差分分布距离Diff WD衡量生成数据的时间相关性, 结果如表3所示.

从数据波形形态相似度角度(DTW)分析，MC-CESDM模型最好. 虽然它的统计相关性指标(ACF)逊于 C-VAE, 但它的DTW距离大幅领先, 说明它生成的序列在视觉形态和整体趋势上更像真实数据，没有出现 C-VAE存在的“过度平滑”问题.

如图4所示, 从各维度数据与真实数据分布差异性分析可以看出, MC-CESDM数据生成分布最贴合真实数据分布，也验证了基于低维流形理论捕捉高维监测数据所处的流形特征能够提高数据生成质量这一观点.

使用t-SNE将高维监测数据压缩成2D平面图，如图5所示，能明显体现MC-CESDM中的低维流形思想. MC-CESDM具有最佳的生成质量, 合成数据与真实数据在二维空间中完美交织, 精确复现复杂的几何结构, 同时也保持数据的局部密度和多样性，方法在拟合复杂时空数据分布方面具有显著优越性.

特征相关性、数据分布以及t-SNE分析结果都能凸显出MC-CESDM模型的数据生成效果最好, 为了量化MC-CESDM数据生成性能，采用判别器得分(discriminative score, DS), 平均归一化Wasserstein距离(WD)以及FD(Fréchet distance)标准量化MC-CESDM、C-VAE、C-GAN及DM-MTS生成数据质量优劣，结果如表4所示.

量化结果可以看出MC-CESDM性能在所有模型中数据生成效果好, 对于DS项由于C-GAN使用判别器架构对抗训练, 采用DS为零在合理范围内.

同时, MC-CESDM核心思想在于使用低维流形表征高维特征空间, 即模型参数远低于其他模型, 由于训练器和生成器参数规模相同或相近，只对比训练器部分的总参数规模T-Pa和模型大小MS, 如表5所示.

表 2　特征相关性矩阵差异度结果[a)]

Table 2　Difference result of feature correlation matrix

模型	Diff
MC-CESDM	**0.3296**
C-VAE	4.8821
DM-MTS	0.3420
C-GAN	1.4666

a) 粗体表示在此项指标对比上最优

表 3　时间相关性指标结果[a)]

Table 3　Results of time-related indicators

模型	ACF	DTW	Diff WD
MC-CESDM	0.0525	**0.4019**	0.4363
C-VAE	**0.0378**	0.7294	**0.1209**
DM-MTS	0.5507	5.8356	0.1212
C-GAN	0.4107	0.8673	1.4428

a) 粗体表示在对应项指标对比上最优

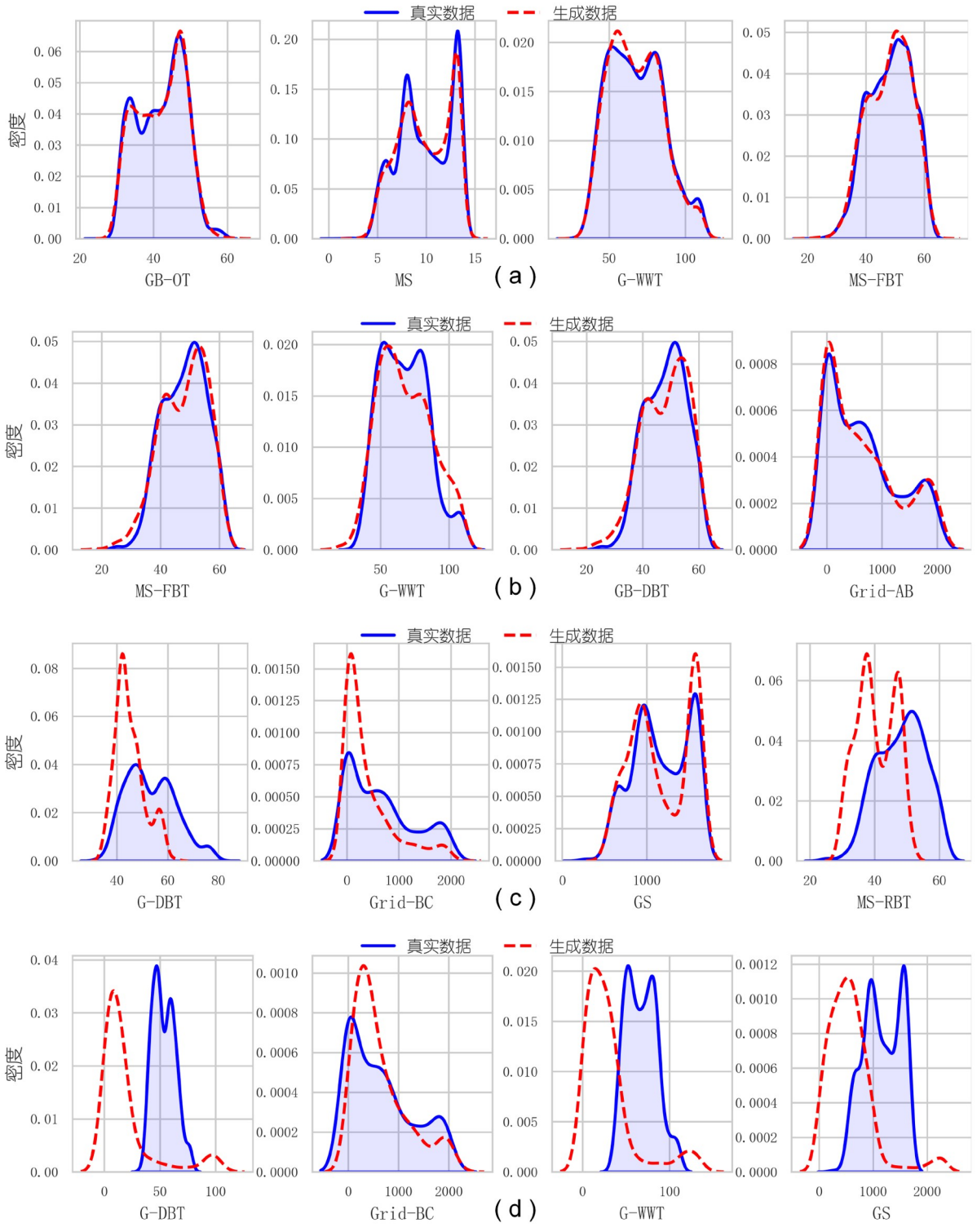

图 4 生成数据与真实数据的对应监测维度数据分布对比(部分). (a) MC-CESDM生成数据分布; (b) C-GAN生成数据分布; (c) C-VAE生成数据分布; (d) DM-MTS生成数据分布

Figure 4 Comparison of data distribution between generated data and real data (part). (a) Data distribution generated by MC-CESDM; (b) data distribution generated by C-GAN; (c) data distribution generated by C-VAE; (d) data distribution generated by DM-MTS.

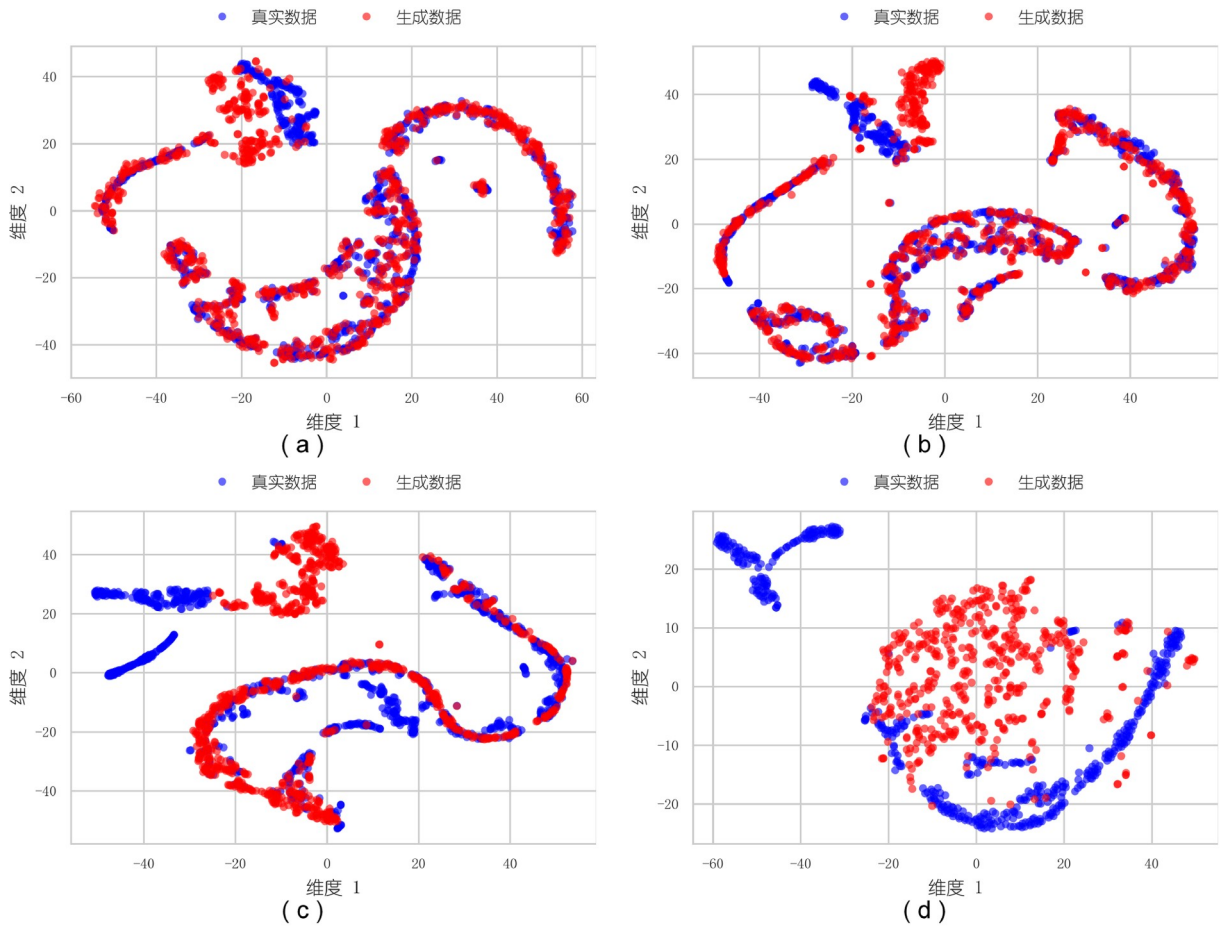

图 5　基于t-SNE的MC-CESDM生成数据与真实数据的数据分布对比. (a) MC-CESDM的t-SNE对比图; (b) C-GAN的t-SNE对比图; (c) C-VAE的t-SNE对比图; (d) DM-MTS的t-SNE对比图

Figure 5　Data distribution comparison between the generated data of MC-CESDM based on t-SNE and the real data. (a) t-SNE comparison plot of MC-CESDM; (b) t-SNE comparison plot of C-GAN; (c) t-SNE comparison plot of C-VAE; (d) t-SNE comparison plot of DM-MTS.

值得一提的是, 在整项研究的过程中, 发现MC-CESDM模型在单一状态下仍具有良好的性能效果. 以此条件约束下性能效果好的模型C-VAE为例, MC-CESDM模型只有t-SNE略逊于C-VAE, 其余各项评判指标均优于C-VAE.

为了探究模型所需的最小样本数量边界, 对于39070条数据, 在训练集划分后, 采用不同比例分配方式[1.0, 0.5, 0.2, 0.1, 0.05, 0.01]随机选取样本, 形成[31216, 15608, 6243, 3121, 1560, 312]样本量训练模型

表 4　模型性能参数对比[a)]

Table 4　Comparison of model performance parameters

模型	DS	WD	FD
MC-CESDM	0.3174	**0.0538**	**0.3720**
C-VAE	0.4842	0.6473	3.4533
DM-MTS	0.4980	3.0673	17.0132
C-GAN	**0**	0.1250	0.7821

a) 粗体表示在对应项指标对比上最优

表 5　模型参数量对比

Table 5　Model parameter count comparison

模型	T-Pa	MS (MB)
MC-CESDM	685461	2.61
C-VAE	1164629	4.44
DM-MTS	2425616	9.25
C-GAN	377872	1.44

并验证数据生成性能.

t-SNE结果如图6所示, 在0.1比率/3121样本量及以上, 模型生成结果与真实数据具有相似的数据分布, 而低于3121样本时整体分布相似, 但是在局部区域发散, 难以学习到真实分布. 对本数据集, 实验验证最佳样本量在3121条数据之上.

5 结论与展望

本研究设计了一种流形赋能下的复杂机电系统监测数据生成方法, 采用低维流形理论, 构建新型时间序列生成方法, 即MC-CESDM, 生成实际运行工况下的监测数据. MC-CESDM不仅可以获得详尽的高维监测变量分布, 还可以通过调节低维运行条件, 如实验对比中风速、变桨角等获得不同工况甚至不同故障模式下的数据, 为下游深度学习故障诊断模型的数据稀缺问题提供可靠的合成数据输入, 也能为机电系统故障诊断、剩余寿命预测、传感器数据缺失补偿的实时数据增强提供可靠的数据支撑.

针对运行数据的高维耦合特性, 本研究并未直接生成全量数据, 而是将7维观测变量作为数据的内在流形坐标, 建立条件驱动扩散模型来生成高维监测变量. 该方法能够有效缓解状态样本稀缺问题, 同时还能确保生成数据在物理逻辑上的一致性, 使得生成的SCADA数据符合风机在特定物理条件下的实际运行规律.

相较于DM-MTS、C-VAE和C-GAN, MC-CESDM模型采用轻量级的模型架构实现高质量数据生成, 有效降低训练复杂度和资源消耗. 从定量指标和定性分析来看, MC-CESDM的数据质量显著优于C-GAN和C-VAE模型, 引入的流形约束和条件扩散机制, 缓解了传统生成模型在小样本故障数据上的模式崩塌问题. t-SNE可视化结果显示, MC-CESDM生成的合成数据与真实数据具有高度的流形结构相似性, 红蓝点具

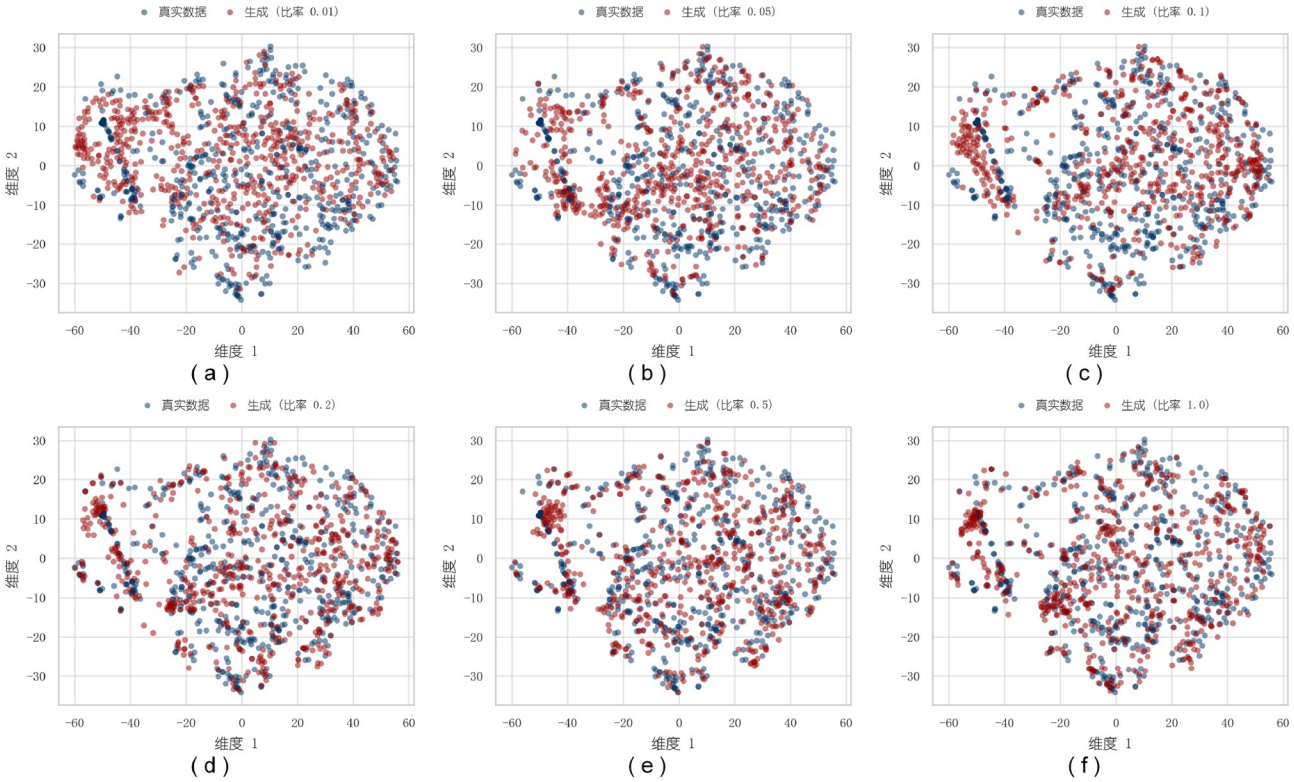

图 6 模型不同数据样本量50周期下的生成效果性能图. (a) 312样本/0.01比率; (b) 1560样本/0.05比率; (c) 3121样本/0.1比率; (d) 6243样本/0.2比率; (e) 15608样本/0.5比率; (f) 31216全样本

Figure 6 Performance graphs of generation results for different models under a data sample size of 50 epochs. (a) 312 samples/0.01 ratio; (b) 1560 samples/0.05 ratio; (c) 3121 samples/0.1 ratio; (d) 6243 samples/0.2 ratio; (e) 15608 samples/0.5 ratio; (f) 31216 full samples.

有较高的重叠度, 证明模型能够捕捉到数据的宏观分布.

然而, 当前模型的t-SNE分析显示, 其预设的物理工况变量未能完全捕捉高维SCADA数据的内在流形结构, 导致生成数据在关键维度与真实数据存在偏差. 同时, 模型依赖人工划分运行状态的策略, 也限制了其在跨风场、跨机型场景下的普适性. 因此, 计划的核心改进是从数据中自动学习最优的低维流形坐标, 以取代预设变量, 构建一个能完全解耦工况变化与故障特征的内在表征空间, 从而最小化生成分布差距, 显著拓展模型的适用性与鲁棒性.

参考文献

1 Xia P C, Huang Y X, Wang Y X, et al. Application of data-enhanced discriminant meta-learning method in cross-device fault diagnosis with few samples (in Chinese). Sci Sin Tech, 2023, 53: 2003–2004 [夏鹏程, 黄亦翔, 王誉翔, 等. 基于数据增强的判别元学习方法在跨设备少样本故障诊断中的应用. 中国科学: 技术科学, 2023, 53: 2003–2004]

2 Meng Z, Guan Y, Pan Z Z, et al. Research on rolling bearing fault diagnosis based on secondary data enhancement and deep convolution (in Chinese). J Mech Eng, 2021, 57: 106–115 [孟宗, 关阳, 潘作舟, 等. 基于二次数据增强和深度卷积的滚动轴承故障诊断研究. 机械工程学报, 2021, 57: 106–115]

3 Wang R, Li T, Gao Z, et al. A generative adversarial networks based methodology for imbalanced multidimensional time-series augmentation of complex electromechanical systems. Appl Soft Comput, 2024, 153: 111301

4 Yuan Y, Huang H, Cheng C, et al. Remaining useful life prediction of the aircraft engine based on the GRU-GAN network with a feature attention mechanism (in Chinese). Sci Sin Tech, 2022, 52: 198–212 [袁烨, 黄虹, 程骋, 等. 基于特征注意力机制的GRU-GAN航空发动机剩余寿命预测. 中国科学: 技术科学, 2022, 52: 198–212]

5 Zhang S F, Li T M, Hu C H, et al. Multidimensional data generation method based on WGAN and its application in RUL prediction (in Chinese). J Mech Eng, 2025, 61: 70–82 [张晟斐, 李天梅, 胡昌华, 等. 基于WGAN的多维数据生成方法及其在RUL预测中的应用. 机械工程学报, 2025, 61: 70–82]

6 Neubürger F, Saeid Y, Kopinski T. TimeGAN for data-driven AI in high-dimensional industrial data. In: Advances in Data-Driven Computing and Intelligent Systems. Singapore: Springer, 2024. 473–484

7 Deng R, Chen D, Yao C, et al. A gradient-penalized conditional TimeGAN combined with multi-scale importance-aware network for fault diagnosis under imbalanced data. Sensors, 2025, 25: 6825

8 Chatterjee S, Byun Y C. Leveraging generative adversarial networks for data augmentation to improve fault detection in wind turbines with imbalanced data. Results Eng, 2025, 25: 103991

9 Zhang Z, Gao H, Sun W, et al. Multivariate time series generation based on dual-channel Transformer conditional GAN for industrial remaining useful life prediction. Knowl-Based Syst, 2025, 308: 112749

10 Li X, Ma J, Wu J, et al. Transformer-based conditional generative transfer learning network for cross domain fault diagnosis under limited data. Sci Rep, 2025, 15: 6836

11 Gao X, Zhang Y, Fu J, et al. Data augmentation using improved conditional GAN under extremely limited fault samples and its application in fault diagnosis of electric submersible pump. J Franklin Inst, 2024, 361: 106629

12 Zhao K, Jiang H, Liu C, et al. A new data generation approach with modified Wasserstein auto-encoder for rotating machinery fault diagnosis with limited fault data. Knowl-Based Syst, 2022, 238: 107892

13 Liu Y, Fu S, Lin L, et al. DECVAE: Data augmentation via conditional variational auto-encoder with distribution enhancement for few-shot fault diagnosis of mechanical system. Meas Sci Technol, 2024, 35: 046104

14 Lei T, Pei Z, Pan F, et al. A novel data augmentation method for steering mechanism fault diagnosis based on variational autoencoding generative adversarial networks with self-attention. Meas Sci Technol, 2024, 35: 055115

15 Fan G, Li L, Zhao Y, et al. Zero-shot rolling bearing fault diagnosis based on attribute description. Electronics, 2025, 14: 452

16 Jin Y, McDaniel R, Tatro N J, et al. Implications of data topology for deep generative models. Front Comput Sci, 2024, 6: 1260604

17 Zhu H, Balsells-Rodas C, Li Y. Markovian Gaussian process variational autoencoders. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR. 2023. 42938–42961

18 Tang X, Yan J, Li Y. Supervised multi-layer conditional variational auto-encoder for process modeling and soft sensor. Sensors, 2023, 23: 9175

19 Yao Q, Chen B, Hu A, et al. A novel diffusion model with Shapley value analysis for anomaly detection and identification of wind turbine. Expert Syst Appl, 2025, 284: 127925
20 Gao Y, Hu S, Chen Y, et al. A novel probabilistic wind power forecasting framework integrating similar curve matching mechanism and an enhanced conditional diffusion model. Appl Energy, 2026, 402: 127031
21 Ren L, Wang H, Laili Y. Diff-MTS: Temporal-augmented conditional diffusion-based AIGC for industrial time series toward the large model era. IEEE Trans Cybern, 2024, 54: 7187–7197
22 Li T, He K. Back to basics: Let denoising generative models denoise. arXiv: 2511.13720
23 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). 2020. 6840–6851
24 Peebles W, Xie S. Scalable Diffusion Models with Transformers. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, 2023. 4172–4182

Manifold-aware diffusion method for multi-dimensional unbalanced data enhancement of complex electromechanical systems

LI JiangBo[1], GAO JianMin[1], WANG Zhen[2], ZHANG WenCai[3], CHANG ZhenYuan[4], ZHANG JunQian[3] & WANG RongXi[1]

[1] *School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China*
[2] *Xi'an Thermal Power Research Institute Co., Ltd., Xi'an 710032, China*
[3] *Bohai Petroleum Equipment Manufacturing Co., Ltd., CNPC, Tianjin 300457, China*
[4] *Shaanxi Special Equipment Inspection and Research Institute, Xi'an 710049, China*

The problem of extreme sample imbalance in industrial field monitoring data seriously restricts the decision-making reliability of diagnostic models. When dealing with high-dimensional monitoring data, the existing generation model is difficult to effectively decouple the complex observation space, resulting in the data generation model easily falling into mode collapse and a lack of fidelity. This paper proposes a new paradigm of conditional diffusion generation with manifold learning (MC-CESDM), which integrates flow matching theory and feature decoupling mechanism. It effectively overcomes the limitations of the existing models: stream matching avoids lengthy iterative noise prediction by constructing a deterministic probability path, which greatly improves the training stability and sampling speed; feature decoupling ensures that the generated samples can be accurately controlled by continuous physical attributes such as variable operating conditions. In this study, a typical wind turbine monitoring data is used as a case for experimental verification, and compared with the mainstream generation algorithm. The results show that the model has achieved a significant lead in key indicators such as feature correlation and distribution distance (Fréchet distance), which proves its excellent generation fidelity. This method overcomes the bottleneck of efficiency and quality of data generation under complex operating conditions, and provides an efficient and reliable technical tool for sample enhancement of unbalanced data of electromechanical systems. It lays a theoretical foundation for improving the level of intelligent operation and maintenance of industrial equipment.

complex electromechanical systems, monitoring data, manifold learning, conditional diffusion model, data enhancement

doi: 10.1360/SST-2025-0459