Computer Science and Application 计算机科学与应用, 2026, 16(4), 101-113
Published Online April 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.164113

Hans汉斯

基于潜在空间折叠与纵横交叉注意力的轴承故障诊断模型：面向小样本与强噪声环境的研究

齐嘉泰，吴宗昆，朱诚昕，曹　正，刘小刚*

西京学院计算机学院，陕西　西安

收稿日期：2026年3月1日；录用日期：2026年4月2日；发布日期：2026年4月10日

摘　要

本文在FaultFormer的预训练范式的基础上，提出了一种融合结构化特征重塑与深度特征交互的改进型故障诊断框架——Hybrid CCNet。该模型保留了Transformer强大的序列建模能力的同时，在特征提取上进行针对性的架构改进与融合：(1) 周期启发的潜在空间折叠：参考计算机视觉中的思想，利用旋转机械信号的准周期性特征，构建映射函数将一维时域信号重塑为二维潜在特征网格，将时间维度的周期性冲击转化为空间维度的纹理特征；(2) 强下采样卷积特征分词器：设计了一个三层级联卷积前端替代了简单的线性投影，利用大卷积核进行初步去噪和下采样，结合批归一化与最大池化提取局部的特征；(3) 纵横交叉注意力特征增强：在Transformer编码器输入之前引入稀疏注意力机制，通过计算行与列的亲和度矩阵，聚合二维空间中的跨周期故障特征。实验表明，Hybrid CCNet展现出了优秀的鲁棒性与性能。在极端小样本场景的情况下，模型准确率达到99%，优于CNN和FaultFormer架构。在信噪比低至−4 dB的强噪声环境中，模型保持了95%以上的诊断精度。

关键词

轴承故障诊断，小样本学习，Transformer，潜在空间折叠，纵横交叉注意力

A Bearing Fault Diagnosis Model Based on Latent Space Folding and Criss-Cross Attention: Research on Small-Sample and Heavy-Noise Environments

Jiatai Qi, Zongkun Wu, Chengxin Zhu, Zheng Cao, Xiaogang Liu*

College of Computer Science, Xijing University, Xi'an Shaanxi

Received: March 1, 2026; accepted: April 2, 2026; published: April 10, 2026

*通讯作者。

文章引用: 齐嘉泰, 吴宗昆, 朱诚昕, 曹正, 刘小刚. 基于潜在空间折叠与纵横交叉注意力的轴承故障诊断模型: 面向小样本与强噪声环境的研究[J]. 计算机科学与应用, 2026, 16(4): 101-113. DOI: 10.12677/csa.2026.164113

Abstract

Based on the pre-training paradigm of FaultFormer, this paper proposes an improved fault diagnosis framework—Hybrid CCNet, which integrates structured feature reshaping and deep feature interaction. This model retains the powerful sequence modeling ability of Transformer while making targeted architectural improvements and integration in feature extraction: (1) Periodic-inspired latent space folding: Inspired by the ideas in computer vision, using the quasi-periodic features of rotating mechanical signals, a mapping function is constructed to reshape one-dimensional time-domain signals into two-dimensional latent feature grids, converting the periodic impacts in the time dimension into texture features in the spatial dimension; (2) Strong downsampling convolution feature tokenizer: A three-level cascaded convolution front-end is designed to replace the simple linear projection. Using large convolution kernels for preliminary denoising and downsampling, combined with batch normalization and max pooling to extract local features; (3) Criss-cross attention feature enhancement: A sparse attention mechanism is introduced before the input of the Transformer encoder. By calculating the affinity matrix of rows and columns, cross-periodic fault features in the two-dimensional space are aggregated. Experimental results show that Hybrid CCNet exhibits excellent robustness and performance. In extremely small-sample scenarios, the model accuracy reaches 99%, superior to that of CNN and FaultFormer architectures. In a strong noise environment with a signal-to-noise ratio as low as −4 dB, the model maintains a diagnostic accuracy of over 95%.

Keywords

Bearing Fault Diagnosis, Small-Sample Learning, Transformer, Latent Space Folding, Criss-Cross Attention

1. 引言

1.1. 研究背景

随着现代工业设备向高速化、精密化的方向发展，旋转机械的健康管理变得越来越重要。轴承是这些设备中应用范围最大、工况最恶劣以及最容易受到损坏的部件。据统计得出，大约 45%的电机故障和30%的齿轮箱故障由于轴承失效。一旦轴承早期发生故障，而且未被及时发现，轻则导致设备停机，重则引发安全事故。因此，开发高精度和强鲁棒性的轴承故障诊断技术，对于实现零停机智能制造具有深刻的现实意义。传统的信号处理方法依赖于专家经验进行复杂的特征工程，且难以适应多变工况和复杂噪声环境。近些年来，以深度学习为代表的数据驱动方法实现了端到端的特征提取与分类，极大地降低了对领域知识的依赖性。但是，学术界的研究成果与工业界的实际应用需求仍然存在差距，其主要体现在小样本困境和强噪声干扰。在学术研究中，通常假设拥有海量且类别较平衡的有标签数据。但在真实的工业现场，机器大部分时间中处于健康运行状态，故障数据极其稀缺且故障类型不平衡。对海量振动数据进行逐秒的人工标注成本高昂，而且需要资深专家参与其中，导致在实际部署时，模型往往面临训练数据不足的问题，容易发生过拟合。实验室环境下的数据通常采集自低噪声的测试台架，信号相对干净。但在实际工况中，传感器采集到的信号往往伴随强烈的噪声。如果信噪比极低，故障产生的微弱冲击信

号会被掩盖，导致基于干净数据训练的模型在真实工况应用时性能下降明显。

1.2. 现有方法的局限性与改进空间

卷积神经网络因为其局部特征提取能力和权值共享特性，在轴承故障诊断中得到广泛应用。Zhang 等人[1]提出的 WDCNN 深入分析了卷积核尺寸对信号处理的影响，证实了宽卷积核在抑制高频噪声方面的有效性，其第一层卷积核为 64，能够覆盖比较长的波形片段。针对工业数据的长尾分布特性，Jia 等人[2]引入了深度归一化 CNN 来平衡对于不同类别的损失。CNN 的优势在于强归纳偏置，局部连接和权值共享使得模型拥有平移不变性，对高频噪声具有平滑作用。但 CNN 的局限性在于其感受野受限，所以难以捕捉长序列中相距甚远的故障模式之间的关联，而这正是 Transformer 的特点。Transformer 及自注意力机制被引入预测性维护领域用来解决全局建模问题。Tang 等人[3]提出了 S-Transformer，采用信号嵌入完成一维信号的分割和升维表示，从而丰富高维空间中的信息。Zerveas 等人[4]提出的 TST 模型证明了 Transformer 在多变量时间序列回归和分类中的有效性。然而，自注意力机制缺乏对局部波形的平滑能力，容易被噪声干扰。面对标注数据的匮乏，自监督学习逐渐受到关注。Zhou 和 Barati Farimani [5]提出的 FaultFormer 借鉴了 He 等人[6]的 MAE 思想，设计了掩码信号重建任务，显著提升了模型的诊断能力。本文的工作可以视为对 FaultFormer 架构的一次针对性增强和改进。并未改变其核心的预训练范式，而是重点优化了特征提取网络，通过引入卷积和稀疏注意力机制，来弥补 Transformer 在处理强噪声以及周期性信号时的不足。Huang 等人[7]提出的 CCNet 在计算机视觉领域通过稀疏关注降低了计算复杂度，本文将其引入至折叠后的一维信号处理中，利用其长程依赖捕捉能力。FaultFormer 将 Transformer 的掩码预训练范式系统性地引入轴承故障诊断中，证实了自监督学习能够有效缓解数据稀缺问题，提升了模型的适应性。FaultFormer 在跨工况迁移方面表现良好，但在处理强噪声和长序列时仍然存在一定的优化空间。FaultFormer 以及其他大多数基于 Transformer 的方法，如 TST，直接将振动信号视为一维文本序列。Transformer 采用简单的线性投影将信号切片转换为 Token，这种点对点的线性映射对高频背景噪声较为敏感，且不具备局部平滑能力。在强噪声环境下，输入 Transformer 的 Token 包含了大量噪声成分，会干扰后续的 Attention 机制。

1.3. 本文的工作内容

为了保留 FaultFormer 的预训练优势并针对上述问题进行改进，本文提出了一种名为 Hybrid CCNet 的混合诊断框架。核心思路是架构融合与适配，通过在 Transformer 之前引入基于结构重塑的特征提取与增强模块，使其提高模型在小样本和强噪声环境下的诊断精度。我们构建了卷积分词 + 特征增强 + Transformer 的混合架构。对于前端的优化，设计了三层强下采样卷积分词器，利用大卷积核进行初步去噪，结合 BN 和 MaxPooling 提取鲁棒特征，在特征进入 Transformer 之前进行预处理。对中端进行了增强，引入潜在空间折叠与纵横交叉注意力模块。这一模块将一维时序信号映射为二维特征网格。对于后端的建模，保留标准的 Transformer Encoder，利用其序列建模能力，对增强处理后的特征进行分类。我们进行了鲁棒性评估，在 CWRU 数据集上进行了 Time-split 划分实验。结果显示，Hybrid CCNet 在−4 dB 强噪声下比纯 Transformer 模型准确率有所提升，且推理时间保持在毫秒级，具备一定的工业应用参考价值。

2. 方法论

Hybrid CCNet 的整体架构设计主要由三个部分组成，分别是三层卷积分词器、潜在空间折叠与纵横特征增强模块以及 Transformer 编码器与预训练头，图 1 为总体框架结构。

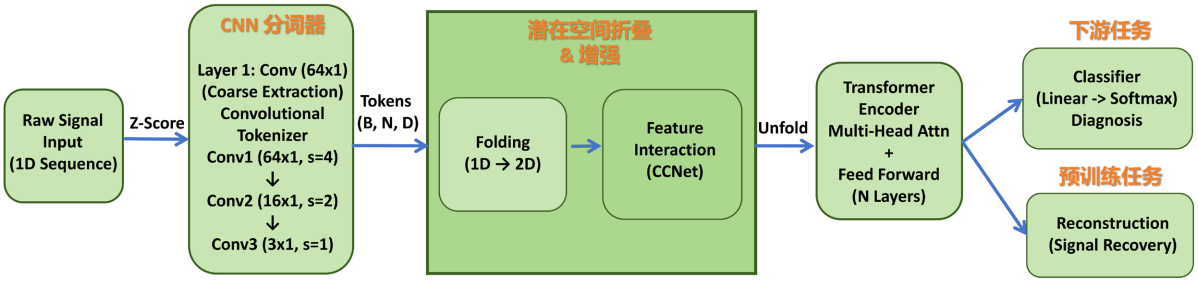

Figure 1. Hybrid CCNet framework structure
图 1. Hybrid CCNet 框架结构

2.1. 信号输入与标准化预处理

采集到的原始振动加速度信号为 $x_{raw} \in \mathbb{R}^L$，其中 L 为采样长度。由于不同负载、转速下信号的幅值差异较大，直接输入网络会导致梯度不稳定，为此我们采用 Z-Score 标准化：

$$\tilde{x} = \frac{x_{raw} - \mu}{\sigma + \varepsilon} \tag{1}$$

其中 μ 和 σ 分别为当前样本的均值和标准差，$\varepsilon = 1e-5$ 为数值稳定项。Z-Score 标准化加速了卷积网络的收敛，且使得不同工况下的信号在统计分布上趋于一致，有助于提高模型的泛化能力。

2.2. 卷积特征分词器

标准的 Transformer 使用线性投影层将标量信号映射为向量。由于考虑到线性映射对高频噪声的敏感性，并参考 WDCNN 的设计思路，我们在 Token 化之前进行去噪和波形提取有助于提升后续模型的鲁棒性。为此设计了一个三层级联卷积网络作为分词器，结合了 Batch Normalization 和 GELU 激活函数以增强训练稳定性。其数学表达如下(令 $h_0 = \tilde{x}$)：

第一层(大核粗提取与下采样)：

$$\mathrm{h}_1 = \mathrm{MaxPool}\left(\mathrm{GELU}\left(\mathrm{BN}\left(\mathrm{Conv1d}\left(\mathrm{h}_0, \mathrm{k}=64, \mathrm{s}=4, \mathrm{c}=\mathrm{d}/4\right)\right)\right)\right) \tag{2}$$

第一层采用 64 × 1 的宽卷积核以及步长 4。在振动信号分析中，大卷积核能够覆盖一个较长的局部时域窗口，这恰好对应于轴承故障冲击的脉冲宽度。该设计不仅充当了低通滤波器，有效平滑高频噪声，还通过大步长降低了时间分辨率，减少了后续的计算量。

第二层(中核特征抽象)：

$$\mathrm{h}_2 = \mathrm{GELU}\left(\mathrm{BN}\left(\mathrm{Conv1d}\left(\mathrm{h}_1, \mathrm{k}=16, \mathrm{s}=2, \mathrm{c}=\mathrm{d}/2\right)\right)\right) \tag{3}$$

采用 16 × 1 卷积核以及步长 2，进一步地提取波形包络特征并进行下采样。这一层负责捕捉信号的局部纹理变化。

第三层(细粒度特征映射)：

$$\mathrm{T} = \mathrm{GELU}\left(\mathrm{BN}\left(\mathrm{Conv1d}\left(\mathrm{h}_2, \mathrm{k}=3, \mathrm{s}=1, \mathrm{c}=\mathrm{d}\right)\right)\right) \tag{4}$$

最后一层使用 3 × 1 小卷积核进行精细化特征调整，最终输出 $\mathrm{T} \in \mathbb{R}^{\mathrm{B} \times \mathrm{d}_{\mathrm{model}} \times \mathrm{N}}$。其确保了输入到 Transformer 的 Token 具备较高的信噪比和丰富的局部语义信息。

2.3. 基于结构化重塑的潜在空间折叠与特征增强

一维时序模型难以捕捉长距离依赖。为了增强特征的结构性，我们在 Token 进入 Transformer 之前，

引入了一个特征精炼阶段，效果见图 2。定义折叠操作，将一维 Token 序列重塑为二维网格：

$$X_{2D} = \mathrm{Reshape}\left(\mathrm{Permute}(T), [B, d, H, W]\right) \tag{5}$$

其中 d 为通道数即特征的维度，$H \times W = N$。这一步将稀疏的时间冲击转化为空间上的特征纹理。若信号具有准周期性，那么在折叠后的二维图上，相邻位置的 Token 在时域上可能相隔较远，但在空间上却属于同一周期的相似相位，从而形成特定的纹理模式。这种折叠操作本质上是一种结构化重塑。通过维度变换将稀疏的时间冲击转化为空间上的特征纹理，从而利用二维注意力机制捕捉跨时间步的关联。利用 CCNet 的稀疏注意力机制，在二维网格上聚合特征：

$$X'_{2D} = \mathrm{CCNet}\left(X_{2D}\right) \tag{6}$$

CCNet 通过计算行与列的亲和度，以 $O\left(N\sqrt{N}\right)$ 的低复杂度实现了长距离信息的交互。它能有效增强在空间上呈现规律分布的纹理特征，并且能够抑制随机分布的背景噪声。最后将增强后的二维特征图展平回一维序列，准备输入 Transformer。展开：

$$T_{\mathrm{refined}} = \mathrm{Permute}\left(\mathrm{Flatten}\left(X'_{2D}\right)\right) \tag{7}$$

在具体实现细节上，尽管卷积分词器采用了总步长为 8 的级联卷积结构，但在分词器的末端，引入了自适应平均池化层，将 Token 序列长度规范化为固定的 64 个。在进入 Transformer 之前，这 64 个 Token 被折叠为 16 × 4 的二维网格($H = 16, W = 4$)。通过 16 × 4 的非方阵结构，在高度方向和宽度方向之间取得了平衡。

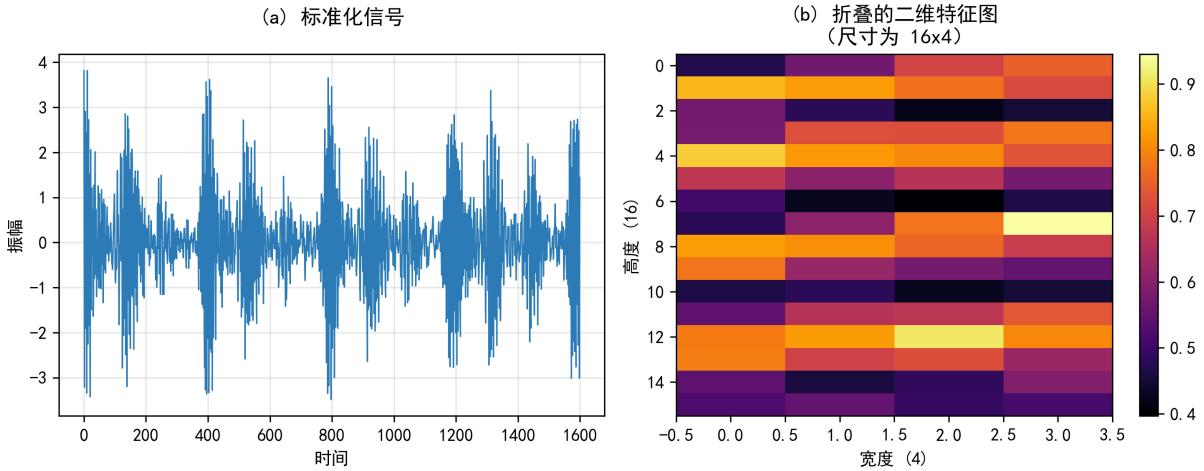

Figure 2. Mapping of latent space characteristics
图 2. 潜在空间特征映射

2.4. Transformer 编码器与端到端信号重建

经过上述去噪和增强后的 Token 序列 T_{refined}，被送入标准的 Transformer Encoder 进行深度建模。为了进一步提高模型的泛化能力，在微调前采用了自监督预训练策略。采用掩码策略，且针对连续信号重建任务进行了优化。随机选择 15%的 Token 位置。采用零值替换策略，即被遮蔽的位置九成用 0 向量替换，一成保持原样。这种遮蔽策略有助于缓解预训练阶段引入的掩码标记在微调阶段缺失所带来的分布偏移，使得模型更专注于利用上下文信息进行重建。Transformer 接收部分被置零的序列，利用 Self-Attention 学习序列上下文。不同于常见的特征重建，我们设计了一个从 LatentSpace 直接映射回

RawSignalSpace 的重建头，即重建目标是原始信号的切片，而不是卷积特征：

$$\hat{x}_{\text{raw}} = \text{Linear}\left(\text{T}_{\text{output}}\right) \tag{8}$$

损失函数计算重建信号与原始信号切片之间的均方误差：

$$\mathcal{L}_{\text{MSE}} = \frac{1}{\sum M_i} \sum_{i=1}^{N} M_i \cdot \left\| x_{\text{raw},i} - \hat{x}_{\text{raw},i} \right\|_2^2 \tag{9}$$

这种端到端的重建目标迫使模型必须深入理解信号的波形结构，而不仅仅是拟合中间特征，从而习得更鲁棒的物理表征。

3. 实验设置

3.1. 数据集：凯斯西储大学

CWRU 数据集是轴承诊断领域的基准数据集。本文实验采用驱动端数据，采样频率为 12 kHz。数据集包含 10 个类别，涵盖正常状态及不同尺寸(0.007 英寸、0.014 英寸、0.021 英寸)的内圈(IR)、外圈(OR)和滚动体故障。

3.2. 数据划分与噪声注入

整个原始信号序列的前 80%用于训练，后 20%用于测试。训练集使用较小的步长进行重叠切片。测试集使用无重叠进行切片，以获得独立且具有代表性的测试样本。在鲁棒性测试实验中，向测试集信号中注入高斯白噪声。信噪比定义为$10\log_{10}\left(\text{P}_{\text{signal}}/\text{P}_{\text{noise}}\right)$。测试范围涵盖 10 dB、5 dB、0 dB 到−4 dB。为了提升模型的抗噪鲁棒性，我们在微调阶段引入了在线对抗噪声增强(On-the-fly Adversarial Noise Augmentation)策略。在每个训练 Batch 中，以 70%的概率向输入信号随机注入−5 dB 到 15 dB 的高斯白噪声，迫使模型在训练过程中适应多变的噪声环境。

3.3. 实验模型与参数设置

为了全面评估模型的性能，本节详细阐述了实验相关的数据预处理、对比模型配置以及下游任务参数的设置。针对 CWRU 数据集的重叠切片比例与归一化等预处理参数细节见表 1；为验证模型架构各组件的有效性，设计了包含传统网络、Transformer 及混合分词器在内的多个基线模型，详细的模型结构与描述见表 2；关于小样本划分比例以及抗噪强度的基准测试设置详见表 3。

Table 1. Preprocessing parameters of the CWRU dataset
表 1. CWRU 数据集预处理参数

参数	设置
数据来源	CWRU 轴承数据
类别	1 类正常 + 9 类故障(内圈/外圈/滚轮 x 007/014/021)
信号长度	1600
训练/测试比例	80%/20%
训练采样重叠步长	200 重叠采样
测试采样重叠步长	1600 无重叠采样
标准化	Z-Score 标准化

Table 2. Comparison models
表 2. 对比模型

模型	分词	描述
MLP	-	多层感知机
CNN	-	一维卷积神经网络
Standard BERT	Constant (200)	Transformer Encoder 配合线性投影
Hybrid BERT	CNN (64)	引入 CNN 前端分词，无注意力增强
Standard CCNet	Constant (200)	标准分词 + 纵横交叉注意力增强
Hybrid CCNet	CNN (64)	CNN 前端 + 纵横交叉注意力增强

Table 3. Downstream tasks and benchmark tests
表 3. 下游任务和基准测试

实验设置	参数详情
任务类型	分类
损失函数	交叉熵损失
微调轮数	10, 20, 30, 40, 50, 100
实验 1：小样本学习	数据比例：0.1, 0.2, 0.3, 0.5, 1.0
实验 2：鲁棒性测试	信噪比水平：10, 5, 0, −4 dB
对抗训练	启用

4. 实验结果与分析

基于 CWRU 数据集，对所提出的 Hybrid CCNet 模型进行了全面的评估。实验分为两个核心部分：极端小样本条件下的性能评估和强噪声环境下的鲁棒性分析。为了确保结果的可靠性，对比了不同训练阶段的表现，以验证模型的稳定性。

4.1. 极端小样本下的性能剖析

在实际工业场景中，获取高质量的故障标注数据往往代价高昂。为了模拟这一现实挑战，我们在训练集中分别抽取了 10%，20%，30%，50%和 100%的样本进行微调，并在统一的独立测试集上进行了评估，可视化见图 3。

4.1.1. 整体趋势与基线对比

从表 4 的数据可以看出，随着训练数据比例的增加，所有模型的准确率均呈现出大体递增的趋势，这验证了数据驱动模型的基本属性。不同模型对数据量的敏感程度存在显著差异。在 10%的数据比例下，MLP 的准确率仅为 50%，这表明在没有先验知识和归纳偏置的情况下，仅靠多层感知机无法从如此稀少的数据中学习到有效的故障模式。CNN 凭借卷积层的参数共享特性，准确率到达 95%，证明了局部特征提取在小样本下的有效性，其性能接近基于注意力机制的模型。引入自监督预训练的 Standard BERT 在 10%数据下达到了 92%，略逊传统 CNN。其线性投影前端在缺乏大量数据微调的情况下，难以完美适应具体的任务分布，导致其性能瓶颈。本文提出的混合模型在极低数据量下展现了较好的效果。Hybrid

CCNet，在 10%数据下准确率达 99%。相比于 Standard BERT，其错误率从 7.67%骤降至 0.81%，实现了近 90%的相对错误率降低。这一结果有力地证明了卷积分词器提取的鲁棒局部特征和潜在空间折叠带来的结构性偏置，极大地降低了对训练样本数量的依赖，使得模型能够在数据匮乏时快速收敛。

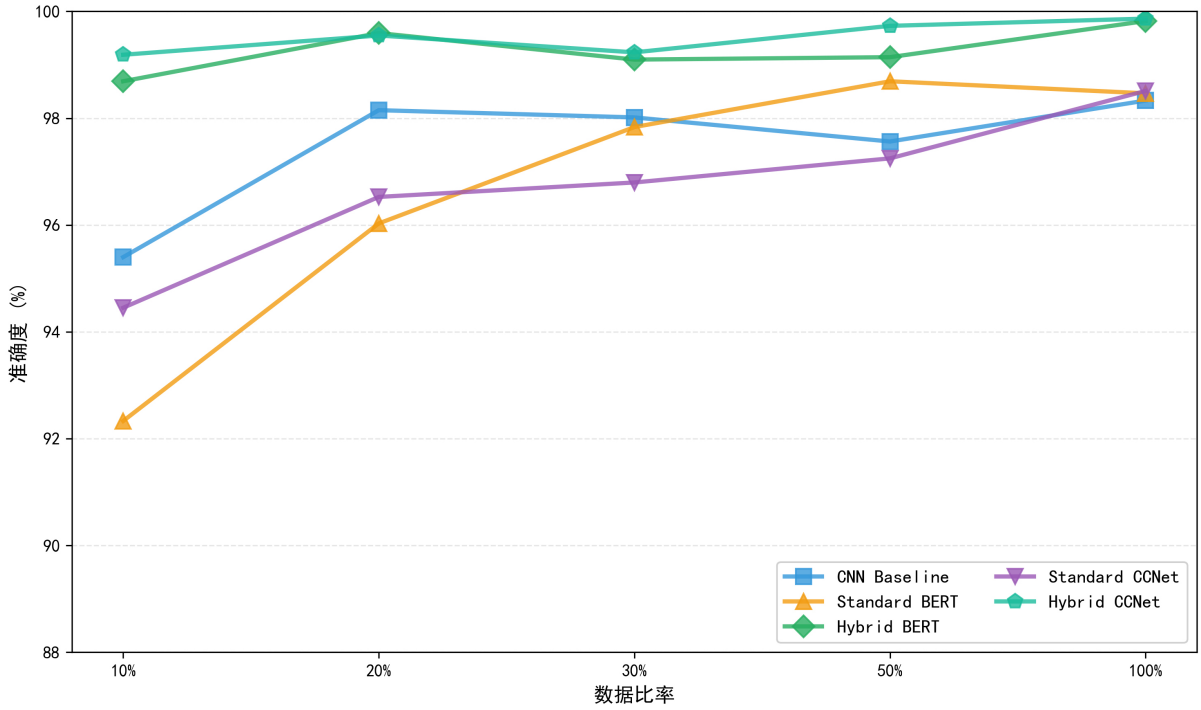

Figure 3. Model performance under different sample proportions (epochs 100; MLP not shown)
图 3. 不同样本比例下的模型性能(epochs 100；未显示 MLP)

Table 4. Performance comparison under different data proportions over 100 epochs
表 4. 不同数据比例下 100 epochs 的性能对比

Ratio	MLP (%)	CNN (%)	Standard BERT (%)	Hybrid BERT (%)	Standard CCNet (%)	Hybrid CCNet (%)
0.1	50.99	95.40	92.33	98.69	94.45	99.19
0.2	58.84	98.15	96.03	99.59	96.53	99.55
0.3	62.05	98.01	97.83	99.10	96.80	99.23
0.5	66.56	97.56	98.69	99.14	97.25	99.73
1.0	68.37	98.33	98.47	99.82	98.51	99.86

4.1.2. 模型变体分析

对比同一类的 Standard 和 Hybrid 变体，可以清晰地看到卷积+折叠策略的贡献。在数据最稀缺的 10%比例下，Hybrid CCNet 相比于 Standard CCNet 提升了 4.74%。单纯的 CCNet 模块虽然能捕捉长程依赖，但在输入特征微弱时效果受限。而引入 CNN 前端和折叠操作后，模型能更好地将一维时间序列中的故障冲击转化为二维空间中的清晰纹理，从而被 CCNet 高效捕捉。随着数据比例的增加，所有 Hybrid 模型的准确率均饱和在 99.8%以上，而 Standard BERT 和 Standard CCNet 则稍显逊色。

4.2. 强噪声环境下的鲁棒性评估

噪声是工业现场最大的阻碍。为了评估模型在极端恶劣环境下的生存能力，在测试集中注入了高斯白噪声，设置信噪比从 10 dB 逐步降至−4 dB，对比见图 4。

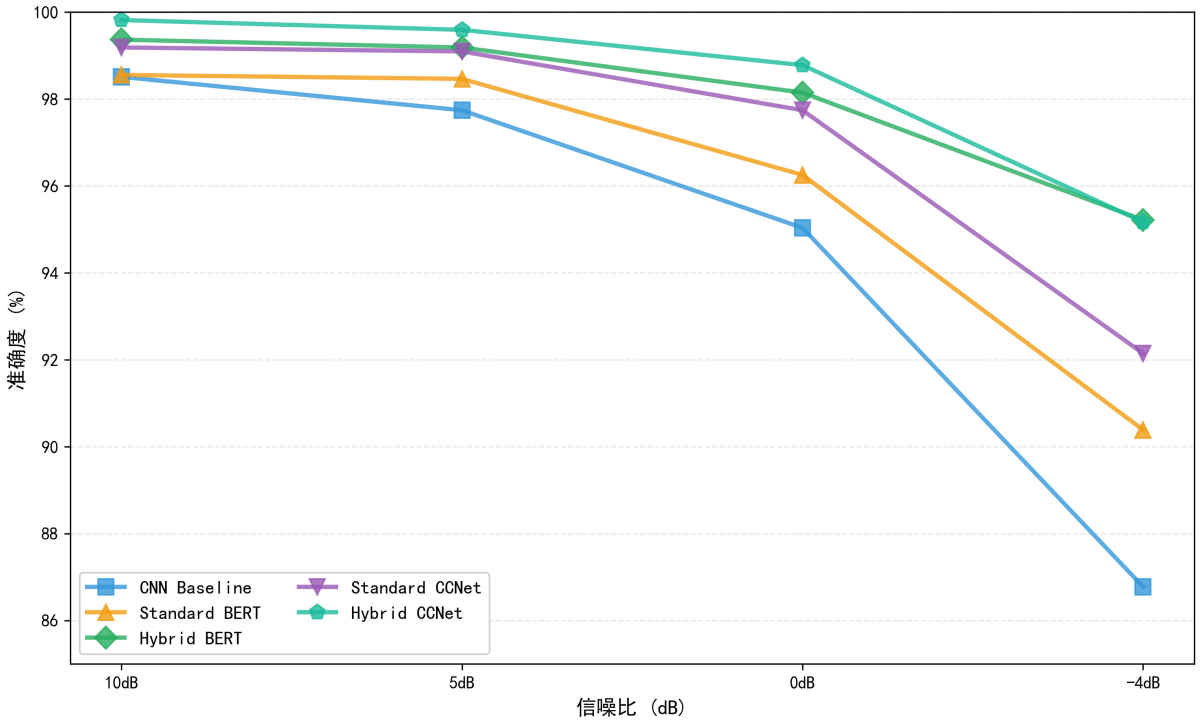

Figure 4. Model performance under different noise levels (epochs 100; MLP not shown)
图 4. 不同噪声水平下模型表现(epochs 100；未显示 MLP)

Table 5. Comparison of robustness over 100 epochs at different signal-to-noise ratios
表 5. 不同信噪比下 100 epochs 的鲁棒性对比

SNR (dB)	MLP (%)	CNN (%)	Standard BERT (%)	Hybrid BERT (%)	Standard CCNet (%)	Hybrid CCNet (%)
10	60.24	98.51	98.56	99.37	99.19	99.82
5	54.87	97.74	98.47	99.19	99.10	99.59
0	46.07	95.04	96.25	98.15	97.74	98.78
-4	34.48	86.78	90.39	95.22	92.15	95.17

从表 5 的数据中，随着信噪比的降低，不同模型的性能衰减模式呈现出显著的差异。总体而言，基于混合架构的模型表现出了最佳的稳定性，而浅层基线模型则遭受了严重的性能滑坡。Standard BERT 在 10 dB 下表现优异，但在−4 dB 下准确率下降至 90%，性能衰减幅度约为 8.2%。在−4 dB 的强噪声下，引入抗噪训练的 Standard BERT 的准确率实际上高于 CNN。相比之下，MLP 和 CNN 在低信噪比下均出现了超过 10%的剧烈性能波动。相比于 Standard 版本，引入卷积分词器的 Hybrid 系列模型表现出了极强的抗噪韧性。以 Hybrid BERT 为例，其在−4 dB 下的准确率为 95.22%，仅比 10 dB 时下降了约 4.15%。

Hybrid CCNet 在−4 dB 下仍保持了 95.17%的高精度。相比于 Standard BERT，Hybrid 模型在−4 dB 下的性能提升约 4.8%。CNN 前端的大卷积核起到了低通滤波器的作用，在数据进入 Transformer 之前就已经滤除了大部分高频随机噪声，保留了低频的故障包络信号，从而极大地提升了信噪比。在−4 dB 这一极限工况下，混合架构完全占据了主导地位。Hybrid BERT 在强噪声下以微弱优势领先。这表明在保留序列一维结构的同时，利用卷积进行降噪和 Transformer 进行全局建模，应对极端噪声更优。Hybrid CCNet 与 Hybrid BERT 的差距极小。尽管受到噪声干扰，CCNet 依然能够从折叠后的二维结构中提取出关键特征。Hybrid CCNet 与 Hybrid BERT 并驾齐驱，即使信号中的周期性纹理被噪声破坏，潜在空间折叠操作依然具备一定的结构化表达能力，能够将含噪信号转化为二维空间中可被捕捉的特征模式。Hybrid BERT 之所以能以极其微弱的优势胜出，可能是因为在极端混沌的噪声环境中，保留原始一维序列的结构信息比将其重构为二维结构更为稳妥。潜在空间折叠旨在通过二维化来聚合周期性特征，但在信噪比低至−4 dB 的极端掩盖下，原始信号的周期性极其微弱。此时，强制的二维重塑可能会将高强度的随机高斯噪声误构建为具有某种空间规律的伪纹理。CCNet 强大的上下文聚合能力可能会通过拟合这些由噪声产生的虚假空间关联。相比之下，Hybrid BERT 保持了一维时序结构，尽管牺牲了部分跨周期特征的捕捉能力，但在信号几乎不可见的情况下，其更简单的拓扑结构减少了对噪声模式的错误归纳，从而在极限边界上表现出略高的鲁棒性。

4.3. 训练收敛与稳定性分析

对比分析从第 10 轮到第 100 轮训练的过程数据。不同架构在收敛速度、初期性能表现以及训练过程中的稳定性方面展现出不同的特性。基于 Ratio 1.0 下的初期表现，混合模型在训练进行到第 10 个 Epoch 时，Hybrid CCNet 和 Hybrid BERT 达到了极高的准确率。Standard BERT 在第 10 个 Epoch 时准确率为 97%，而基线模型 CNN 为 95%。得益于 CNN 前端提取的鲁棒局部特征，混合模型不需要漫长的预训练阶段即可在极短时间内完成对故障模式的拟合。随着训练轮数的增加不同模型的稳定性差异逐渐显现。混合模型以少样本场景为例，Hybrid CCNet 在整个训练周期内保持高位波动上升，虽然中间略有起伏，但整体趋势平稳向上，未出现性能崩塌。Standard BERT 的剧烈波动，在少样本场景下的表现极不稳定。偶尔会出现性能倒退。即线性投影层在处理稀疏数据时的脆弱性，由于缺乏足够的归纳偏置，模型在训练中容易陷入局部最优或受到噪声梯度的误导。CNN 的表现虽然收敛较慢，但呈现出单调递增的趋势，表明了卷积神经网络在特征提取上的稳定性，但其最终性能受限于全局建模能力的缺失。混合架构不仅收敛速度快，而且在长周期训练中保持了极高的稳定性，避免了 Standard BERT 在训练中期的性能剧烈波动问题。

4.4. 消融研究与各组件贡献

基于 epochs10-100 的训练数据，对 Hybrid 架构中各关键组件的贡献进行了定量的消融分析。对比 Standard BERT 与 Hybrid BERT 在少样本场景下的表现，可以清晰地量化 CNN 前端的贡献。在 Ratio 0.1 的极低数据量下，Standard BERT 的准确率为 92%，且训练过程中波动剧烈。引入 CNN 前端的 Hybrid BERT，在 Ratio 0.1 下的准确率为 98%，且在 Epoch 10 即达到 97%。CNN 前端不仅仅是一个下采样层，其还充当了特征滤波器。它利用卷积核的局部连接性，提取出了故障冲击特征，替代了 Standard BERT 中脆弱的线性投影层。这一替换带来了约 6%的精度提升且解决了训练不稳定的问题。通过对比同系列的 BERT 变体与 CCNet 变体，可以分析 CCNet 模块及空间折叠策略的价值。在 Ratio 0.1 下，Standard CCNet 相比于 Standard BERT 提升了约 2%。在数据匮乏时，CCNet 的交叉协方差注意力机制比标准 Transformer 的自注意力机制更能高效地捕捉样本间的长程依赖。在 Ratio 0.1 下，Hybrid CCNet 相比于 Hybrid BERT

仍有约 0.5%的提升。卷积+折叠策略将一维时间序列重构为二维空间纹理，使得模型能够利用计算机视觉中的归纳偏置，如纹理、边缘来辅助诊断。这种结构性的偏置极大地降低了模型对大量标注数据的依赖。综合 epochs10-100 的数据，CNN 前端解决了 Standard BERT 在少样本下无法收敛及波动大的问题。Transformer 全局建模解决了长程依赖捕捉难题。空间折叠与 CCNet 优化，在高性能基础上进一步提供了 0.5%~2%的精细化提升，增强了模型对微弱特征的敏感度。

5. 总结

本研究针对工业轴承故障诊断中面临的极端小样本数据下的模型泛化能力与强噪声环境下的特征提取鲁棒性这两个核心挑战，提出了融合卷积分词器、潜在空间折叠与纵横交叉注意力机制的 Hybrid CCNet 模型，实验表明其在恶劣工况下性能均优于 CNN 与 Standard Transformer 等现有主流方法。该模型的性能提升主要源于混合架构的协同增益，利用三层级联卷积分词器作为特征滤波器，弥补了 Transformer 缺乏平移不变性与局部连接性等归纳偏置的短板，通过大卷积核有效捕捉故障冲击的包络特征并降低优化难度，确保了在强噪声环境下的高精度。同时，潜在空间折叠策略将准周期的时域信号转化为二维纹理，配合纵横交叉注意力机制，这种显式的结构化表征辅助模型高效捕捉跨周期的长程依赖，显著提升了极低数据量下的特征置信度。尽管表现良好，模型仍面临计算复杂度较高导致在边缘设备部署困难以及在复合故障或非平稳工况下验证不足等局限。此外，在信噪比低于−4 dB 的极限环境中，引入复杂的二维空间注意力机制可能会因噪声结构化而产生轻微的性能饱和。这表明在未来的工作中，设计一种能够根据信噪比自适应调节折叠强度或注意力权重的门控机制，是解决极端噪声干扰的关键路径。未来研究将致力于探索自适应折叠机制、通过知识蒸馏等技术实现模型轻量化，并验证模型在更复杂工况下的泛化能力，从而为数据驱动的故障诊断向实际工业场景落地提供可行的技术路径。

基金项目

陕西省自然科学基础研究计划项目(2024JC-YBMS-068)。

参考文献

[1] Zhang, W., Li, C., Peng, G., Chen, Y. and Zhang, Z. (2018) A Deep Convolutional Neural Network with New Training Methods for Bearing Fault Diagnosis under Noisy Environment and Different Working Load. *Mechanical Systems and Signal Processing*, **100**, 439-453. https://doi.org/10.1016/j.ymssp.2017.06.022

[2] Jia, F., Lei, Y., Lu, N. and Xing, S. (2018) Deep Normalized Convolutional Neural Network for Imbalanced Fault Classification of Machinery and Its Understanding via Visualization. *Mechanical Systems and Signal Processing*, **110**, 349-367. https://doi.org/10.1016/j.ymssp.2018.03.025

[3] Tang, J., Zheng, G., Wei, C., Huang, W. and Ding, X. (2022) Signal-Transformer: A Robust and Interpretable Method for Rotating Machinery Intelligent Fault Diagnosis under Variable Operating Conditions. *IEEE Transactions on Instrumentation and Measurement*, **71**, 1-11. https://doi.org/10.1109/tim.2022.3169528

[4] Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A. and Eickhoff, C. (2021) A Transformer-Based Framework for Multivariate Time Series Representation Learning. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2114-2124. https://doi.org/10.1145/3447548.3467401

[5] Zhou, A.Y. and Barati Farimani, A. (2024) Faultformer: Pretraining Transformers for Adaptable Bearing Fault Classification. *IEEE Access*, **12**, 70719-70728. https://doi.org/10.1109/access.2024.3399670

[6] He, K., Chen, X., Xie, S., Li, Y., Dollar, P. and Girshick, R. (2022) Masked Autoencoders Are Scalable Vision Learners. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 16000-16009. https://doi.org/10.1109/cvpr52688.2022.01553

[7] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y. and Liu, W. (2019) CCNet: Criss-Cross Attention for Semantic Segmentation. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 603-612. https://doi.org/10.1109/iccv.2019.00069

附录：详细网络配置

Table A1. CNN baseline
表 A1. CNN 基线模型

模块/层	输入通道	输出通道	卷积核	步长	填充	激活函数
Conv Block 1	1	4	10	2	Valid	GELU
Conv Block 2	4	64	5	1	Valid	GELU
Conv Block 3	64	128	3	1	Valid	GELU
Conv Block 4	128	64	3	1	Valid	GELU
Conv Block 5	64	64	3	2	Valid	None
Pooling	64	64	-	-	-	AdaptiveAvgPool1d
Linear 1	64	128	-	-	-	GELU, Dropout
Linear 2	128	64	-	-	-	GELU
Classifier	64	10	-	-	-	Softmax

Table A2. MLP baseline
表 A2. MLP 基线模型

层名	输入维度	输出维度	缩放因子	组件
Pre-process	$1\times L$	64	-	Transpose → AdaptiveAvgPool1d → Flatten
Hidden 1	64	256	$4\times d_{model}$	GELU, Dropout
Hidden 2	256	256	$4\times d_{model}$	GELU
Hidden 3	256	128	$2\times d_{model}$	GELU
Hidden 4	128	64	$1\times d_{model}$	GELU
Classifier	64	10	-	Linear Output

Table A3. Parameter settings
表 A3. 参数设置

参数	值
Batch Size	64
Optimizer	AdamW
Learning Rate	1e−3
Dropout	0.3
Weight Decay	0.01
Model Dimension (d_{model})	CNN/MLP: 64 Transformer: 128

续表

Attention Heads (n_{head})	4
Encoder Layers (N_{layers})	2
Positional Embedding	Rotary
Attention Mechanism	Flash Attention
Feed Forward	GLU
Mask Probability	0.15
Replace Strategy	90% Mask/10% Original
Pre-training Epochs	200