上海交通大学博士学位论文

面向旋转机械智能诊断的深度学习可解释性研究

姓　　名：陈　钱

学　　号：020020910003

导　　师：董兴建　副教授

程长明　副教授

院　　系：机械与动力工程学院

学 科/专 业：机械工程

申 请 学 位：工学博士

2025 年 6 月

A Dissertation Submitted to

Shanghai Jiao Tong University for the Degree of Doctor

RESEARCH ON INTERPRETABILITY OF DEEP LEARNING FOR INTELLIGENT DIAGNOSIS OF ROTATING MACHINERY

Author: CHEN Qian

Supervisor: Associate Prof. DONG Xingjian
Associate Prof. CHENG Changming

School of Mechanical Engineering

Shanghai Jiao Tong University

Shanghai, P.R. China

June, 2025

上海交通大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：陈钱

日期：2025 年 5 月 6 日

上海交通大学

学位论文使用授权书

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于：

☑**公开论文**

☐**内部论文**，保密☐ 1 年/☐ 2 年/☐ 3 年，过保密期后适用本授权书。

☐**秘密论文**，保密____ 年（不超过 10 年），过保密期后适用本授权书。

☐**机密论文**，保密____ 年（不超过 20 年），过保密期后适用本授权书。

（请在以上方框内选择打“✓”）

学位论文作者签名：陈钱　　指导教师签名：程长明 茅海建

日期：2025 年 5 月 6 日　　日期：2025年 5月 6 日

摘 要

旋转机械是现代工业装备的核心部件，针对其开展的故障诊断对提升重大机械装备的可靠性和安全性具有重要战略意义。在大容量、低密度、多样性和时效性的大数据背景下，基于深度学习的智能诊断凭借高精度、端到端和高效率的优势，逐渐成为旋转机械故障诊断领域的研究热点。然而，深度学习作为典型的黑箱模型，其诊断依据、诊断逻辑和适用范围均不明晰，在可信性、性能优化与缺陷分析等方面也缺少科学指导。可解释性的不足，已成为深度学习智能诊断从学术研究迈向工业应用的巨大阻碍。

为提升诊断模型可解释性，主动解释方法通过增加额外约束以提高透明度，但也导致网络可拓展性不足且缺乏明确的解释结果；被动解释方法虽对现有模型进行分析，却存在解释形式不直观、计算耗时高的问题。因此，本文以高兼容主动解释与高性能被动解释为目标，构建面向旋转机械智能诊断模型的综合解释框架，为实现可解释智能诊断提供系统性和实用性解决方案。本文的主要研究内容包括：

(1) 针对主动解释效果、可拓展性和诊断性能的共赢问题，聚焦输入层，提出基于时频变换的诊断模型输入层主动解释方法。基于时频变换与卷积层在内积运算上的等价性，在传统卷积层中引入实虚部机制和核函数约束，从而建立等效于可学习时频变换的时频卷积层，并通过训练后频响揭示模型关注的关键频带。在此基础上，将时频卷积层和现有基准模型相结合以建立时频卷积网络。多个实测数据集表明，时频卷积网络在诊断精度、少样本学习能力和收敛速度等方面相较于同类方法具有显著优势，并且具有揭示数据集关键频带的解释能力。

(2) 在提高输入层可解释性后，聚焦决策层，提出基于原型匹配的诊断模型决策层主动解释方法。首先建立能够显式构建原型向量、并基于样本与各原型相似程度进行故障分类的原型匹配层，然后将原型匹配层与自编码器相结合以形成原型匹配网络，并通过分类损失、重构损失和原型匹配距离损失对网络进行联合训练。最后，梳理原型匹配网络在分类逻辑、类别原型、相似性来源三方面的解释能力。传统故障诊断和领域泛化两类任务表明，原型匹配网络不仅具有优异的故障诊断表现，并且可以通过重构原型来刻画模型视角下的典型故障特征。

(3) 针对被动解释的形式不直观问题，提出结合域变换的诊断模型被动解释形式优化方法。首先基于传统的循环谱分析，对确定性信号的二阶自相关函数进行近

似估计，进而建立确定性时域信号到循环域的域变换以及逆变换方法。最后，经由样本预处理与模型集成，构建将 SHAP（SHapley Additive exPlanations）归因从时域扩展至循环域的新型被动解释方法。仿真数据集及两个实测数据集表明，所提方法不仅能获得清晰直观的解释效果，而且具有对近邻信号成分的强大区分能力。

(4) 针对被动解释的计算高耗时问题，提出组合块归因及复杂度简化的被动解释效率优化方法。首先基于 SHAP 计算过程的理论分析，提出将相邻特征绑定以求解联合贡献的组合块归因策略，然后简化耗时的子集枚举过程，并通过两个典型实例构建可近似原始 SHAP 的新型归因方法，从而将计算复杂度从指数级降至线性级。最后，仿真数据集与两个实测数据集表明，所提解释方法在不同域、不同组合块尺寸下具有与原始 SHAP 解释结果高度一致性，以及显著的计算效率优势。

(5) 在上述研究基础上，针对当前可解释性研究零散化、表面化的问题，提出面向旋转机械智能诊断模型的综合解释框架。通过整合两类主动解释方法，建立可同时对模型输入层和决策层进行主动解释的联合解释网络，并构建能指导用户依据任务需求选择合适方法的综合解释框架。行星齿轮传动系统的实验结果表明，三类主动解释网络均具有优异诊断性能，无须考虑额外约束带来的潜在负面影响；全体解释方法的解释效果显著并能够交叉印证，有效证实所提综合解释框架的有效性和可信性。

本文的研究工作从主动解释和被动解释两部分出发，一方面聚焦于网络局部，在保证模型性能和可拓展性的同时提高主动解释效果；另一方面则结合振动信号特点对被动解释进行优化，在实现清晰直观解释效果的同时降低计算成本。两者结合下的综合解释框架，为解释诊断模型的决策逻辑、决策依据和提高模型可信性提供了可行方案和技术源泉。

关键词：旋转机械，智能诊断，可解释性，主动解释，被动解释

Abstract

Rotating machinery serves as the core component of modern industrial equipment, and its fault diagnosis holds strategic significance in enhancing the reliability and safety of critical mechanical systems. In the contemporary big-data landscape characterized by large volume, low density, diversity, and time sensitivity, deep-learning-based intelligent fault diagnosis (IFD) has emerged as a pivotal research focus in rotating machinery fault diagnosis, offering advantages in accuracy, end-to-end modeling, and computational efficiency. However, the inherent "black-box" nature of deep learning models obscures their diagnostic rationale, logical processes, and applicable scope, providing insufficient scientific guidance regarding trustworthiness, performance optimization, and defect analysis. This lack of interpretability constitutes a significant barrier to the transition of IFD from academic research into industrial applications.

To enhance model interpretability, current approaches follow two principal paradigms: ante-hoc methods that incorporate additional constraints to improve transparency (though often at the expense of network scalability and interpretive clarity), and post-hoc methods that explain existing models (but frequently produce non-intuitive explanations with high computational demands). This dissertation addresses these limitations by proposing a comprehensive interpretability framework for rotating machinery IFD models, designed to achieve both high compatibility in ante-hoc interpretation and superior performance in post-hoc interpretation. The principal contributions of this research are as follows:

(1) To address the challenge of simultaneously achieving ante-hoc interpretability, network scalability, and diagnostic performance at the input layer, this study proposes a diagnostic model input-layer approach based on time-frequency transforms. Specifically, by leveraging the equivalence between time-frequency transform and convolutional layer under inner-product operation, real and imaginary components are introduced into the conventional convolutional layer under kernel function constraints. This establishes a time-frequency convolutional layer that is equivalent to a learnable time-frequency transform. Post-training frequency responses reveal the key frequency bands the model focuses on. Integrating this time-frequency convolutional layer with

existing baseline models forms the time-frequency transform network. Experiments on multiple real-world datasets show that the proposed time-frequency transform “Huanglong” achieves notable advantages in diagnostic accuracy, few-sample learning capability, and convergence speed relative to comparable methods, and further provides interpretability by identifying critical frequency bands in the data.

(2) After enhancing interpretability at the input layer, attention shifts to the decision layer, where an ante-hoc interpretability approach based on prototype matching is put forward for classification layer. First, the prototype matching layer is constructed to explicitly learn prototype vectors and classify faults based on each sample’s similarity to these prototypes. The prototype matching layer is then integrated with an autoencoder to form a prototype matching network, which is jointly trained using classification loss, reconstruction loss, and prototype matching distance loss. Subsequently, the explanatory capacity of the prototype matching network is analyzed from the perspectives of classification logic, class prototypes, and source of similarity. Experiments on traditional fault diagnosis and domain generalization tasks indicate that the prototype matching network not only demonstrates excellent diagnostic performance but also reconstructs prototypes to depict the model’s view of typical fault characteristics.

(3) To address the lack of clarity in post-hoc interpretability, a diagnostic model post-hoc interpretability approach is introduced by combining domain transforms. Drawing on conventional cyclic spectral analysis, an approximate estimation of the second-order autocorrelation function for deterministic signals is provided, leading to domain transforms and inverse transforms between deterministic time-domain signals and the cyclostationary domain. Building on this, by means of sample preprocessing and model integration, the proposed method extends SHAP (SHapley Additive exPlanations) from the time domain to the cyclostationary domain, producing a novel post-hoc interpretability approach. Experiments on simulation and two real-world datasets confirm that the proposed method delivers clear, intuitive explanations and robustly distinguishes neighboring signal components.

(4) To address the high computational overhead of post-hoc interpretability, a patch-wise attribution method and complexity simplification strategy are introduced to enhance the efficiency of post-hoc interpretability. Based on theoretical analysis of SHAP’s

computational process, adjacent features are bound together to solve for their joint contributions using a patch-wise attribution strategy, and the time-consuming subset enumeration process is simplified. Two representative examples illustrate novel attribution methods that approximate the original SHAP results, thus reducing the complexity from exponential to linear. Finally, experiments on simulation and two real-world datasets show that the proposed approach achieves highly consistent interpretability results comparable to the original SHAP across different domains and block sizes, with significant computational efficiency gains.

(5) Based on the above research, to counter the fragmented and superficial nature of current interpretability studies, a comprehensive interpretability framework for rotating machinery intelligent diagnostic models is established. By integrating the two types of ante-hoc interpretability approaches, a joint interpretability network is developed that simultaneously interprets the input and decision layers. In addition, a comprehensive interpretability framework is presented to guide users in choosing appropriate methods according to task requirements. Experiments on a planetary gear transmission system demonstrate that all three ante-hoc interpretability networks offer outstanding diagnostic performance without concern for potential negative impacts of additional constraints. Furthermore, the interpretability of all proposed methods is significant, mutually corroborative, and effectively validates the efficacy and trustworthiness of the comprehensive interpretability framework.

This dissertation addresses interpretability from both ante-hoc and post-hoc perspectives. On the one hand, it concentrates on local network regions to enhance ante-hoc interpretability while preserving model performance and scalability; on the other hand, it incorporates signal characteristics to refine post-hoc interpretability, thus achieving clear and intuitive explanations while reducing computational overhead. The integration of these two approaches yields a comprehensive interpretability framework, providing feasible solutions and technological resources for elucidating model decision logic, diagnostic foundations, and enhancing model trustworthiness.

Key words: Rotating machinery, Intelligent fault diagnosis, Interpretability, Ante-hoc interpretability, Post-hoc interpretability

目 录

插 图

表 格

缩略语对照表

缩写词	英文全称	中文全称	页码
AE	Auto-Encoder	自编码器	4
RBM	Restricted Boltzmann Machine	受限玻尔兹曼机	4
CNN	Convolutional Neural Network	卷积神经网络	4
RNN	Recurrent Neural Networks	循环神经网络	4
DBN	Deep Belief Network	深度置信网络	4
LSTM	Long Short-Term Memory	长短期记忆网络	5
GRU	Gated Recurrent Unit	门控循环单元网络	5
BiLSTM	Bidirectional Long Short-Term Memory	双向长短期记忆网络	5
AM	Activation Maximization	激活最大化	10
CAM	Class Activation Mapping	类别激活映射	11
Grad-CAM	Gradient - Class Activation Mapping	梯度类别激活映射	11
DeepLIFT	Deep Learning Important FeaTures	深度学习重要特征	11
LRP	Layer-wise Relevance Propagation	层级相关性传播	11
LIME	Local Interpretable Model-agnostic Explanations	局部可知模型无关解释	11
MAPLE	Model Agnostic suPervised Local Explanations	模型无关监督局部解释	11
CWT	Continuous Wavelet Transform	连续小波变换	13
DWT	Discrete Wavelet Transform	离散小波变换	13
STFT	Short-Time Fourier Transform	短时傅里叶变换	16
SHAP	SHapley Additive exPlanations	夏普利可加解释	17
SHEP	SHapley Estimated exPlanation	夏普利估计解释	22

TFT	Time-Frequency Transform	时频变换	24
CT	Chirplet Transform	线性调频变换	24
WT	Wavelet Transform	小波变换	24
STTF	Short-Time Trigonometric Function	短时正弦函数	27
FIR	Finite Impulse Response	有限冲击响应	28
TFconv	Time-Frequency Convolutional	时频卷积层	29
SGD	Stochastic Gradient Descent	随机梯度下降	31
FR	Amplitude-Frequency Response	幅频响应	31
C-FR	Channel-wise FR	通道层级幅频响应	31
O-FR	Overall FR	整体幅频响应	31
TFN	Time-Frequency convolutional Network	时频卷积网络	33
CWRU	Case Western Reserve University	凯斯西储大学	34
Adam	Adaptive Moment Estimation	自适应动量估计	35
PMN	Prototype-Matching Network	原型匹配网络	55
MSE	Mean Squared Error	均方误差	59
PML	Prototype-Matching Layer	原型匹配层	61
MDE	Mixture Density Estimation	混合密度估计	61
MLP	Multi-Layer Perceptron	多层感知机	68
t-SNE	t-distributed Stochastic Neighbor Embedding	t 分布随机近邻嵌入	69
CS-SHAP	Cyclic Spectral - SHapley Additive exPlanations	循环谱夏普利可加解释	81
CSC	Cyclic-Spectral Correlation	循环谱相关	84
CAF	Cyclic Autocorrelation Function	循环自相关函数	85

第一章 绪论

1.1 研究背景和意义

以齿轮、轴承、转子为代表的旋转机械是现代工业装备的核心部件，在风力发电机、航空发动机、燃气轮机等重大技术装备中发挥着关键作用。上至航空航天、远洋航海、国防事业等国家战略性产业，下到高速列车、汽车交通、全自动生产线等日常民用领域，旋转机械都发挥着举足轻重的作用，其动态服役性能关系到国民经济和国防安全的方方面面。随着科技的快速发展和生产力提升的迫切需求，现代旋转机械已呈现出大型化、复杂化、集成化的发展趋势，而这也增加了机械系统安全的不确定性和潜在风险。在高速、重载、长期持续运行等恶劣工况条件下，旋转机械极易发生材料疲劳、零部件磨损和结构损伤等性能退化现象，继而引发振动异常、效率下降甚至突发故障等一系列严重问题，造成生产停滞、设备损毁，甚至导致灾难性人员伤亡事故。

国内外由于旋转机械故障引起的严重事故不胜枚举，其所造成的经济损失和社会影响极为深远。2004 年 7 月，我国 19015 次列车由于加工缺陷和疲劳累积损伤发生车轴断裂，造成十余节车厢脱轨损坏，导致京广线中断近 23 小时[1]。2016 年 4 月，挪威卑尔根市一架 EC225 超级美洲豹直升机由于主减速器行星齿轮碎裂导致发动机失效，机上 13 人全部遇难[2]。2021 年 2 月，美国联合航空公司 UA328 航班的 PW4077 发动机发生叶片断裂，使得发动机失效并燃烧，类似故障还发生在 2020 年 12 月的日本 JL904 航班和 2018 年 2 月的美国 UA1175 航班[3]。因此，开展对旋转机械及关键零部件的故障诊断研究，识别和诊断故障发生的部位、机理和严重程度，并依据诊断结果制订维修方案实现早发现和早维修，对于保障机械系统安全运行至关重要。当前，我国的《中国制造 2025》[4]、《机械工程学科发展战略报告（2021-2035）》[5] 等重要战略规划，均将提高重大装备可靠性和安全性列为优先发展的重点研究方向。

随着传感器和物联网技术的进步以及故障诊断需求的增加，旋转机械设备群规模逐渐扩大，测点数目逐渐增多，数据采集时长覆盖更广，从而获得了海量的监测数据，推动故障诊断领域进入“大数据”时代[6]。法国空客公司的新型飞机状态监控系统（Aircraft Condition Monitoring System，ACMS）在每架空客 A350 上每天收集的数据超过 1.8 TB，收集的参数接近 60 万个[7]。我国三一重工通过构建工业大数据平台，实现对 132 类工程机械装备的 6143 种状态信息（包括位置、油温、油位、压力、温

度、工作时长等关键参数）的低成本实时采集，覆盖全球分布的 21 万余台工程机械设备，累积采集了超过 1000 亿条的工业大数据记录[8]。当前，旋转机械的监测数据呈现大容量、低密度、多样性和时效性的特点，给故障诊断领域带来了前所未有的挑战，依赖于人工特征提取与专家经验的传统诊断方法逐渐难以满足现代机械设备的诊断需求。因此，以大数据为驱动的智能诊断逐渐进入人们视野。智能诊断从多源监测数据中提取故障特征，利用各类智能模型进行故障识别与分类，有效解决了传统故障诊断方法中 “数据多而专家少” 的结构性矛盾。在智能诊断中，基于深度学习的新一代人工智能技术表现优异，以端到端的方式，实现了对原始监测数据的自动特征提取，获得了更为完备且包含丰富故障信息的特征表达，并能够自主地进行知识学习和故障模式识别。相比传统基于信号处理或机器学习的诊断方法，深度学习智能诊断在诊断准确率、抗噪声能力和泛化性能等关键指标上具有显著优势。

尽管深度学习模型能够有效处理大数据环境下的故障诊断问题并展现出卓越的诊断性能，但作为典型的 “黑箱” 模型，深度学习模型先天缺乏解释性，其诊断依据、诊断逻辑和适用范围均不明晰。机器学习模型的复杂度与可解释性呈现反比关系，神经网络模型为追求更优诊断性能而持续复杂化、规模化，导致深度学习模型的解释难度不断攀升。图灵奖得主姚期智院士在 2020 年浦江创新论坛中明确指出，缺乏可解释性已成为人工智能领域三大瓶颈之一[9]。世界工程组织联合会主席龚克也在 2021 世界人工智能大会上强调，下一阶段 AI 基础研究应主攻 “可解释性”[10]。

可解释性研究对于旋转机械深度学习故障诊断模型迈向实际工业应用具有决定性作用。从用户层面，缺乏可解释性的深度学习模型难以获得机械设备运维人员的充分信任，迫使引入冗余诊断方案进行交叉验证，从而增加实际运维成本；从开发者层面，诊断模型的开发人员无法理解和追踪模型诊断的内在逻辑与依据，当错误诊断发生时，难以对模型实施针对性调试与改进，因而限制了模型诊断性能的有效修正和持续提升；从潜在风险方面，实际故障诊断场景的复杂性和不可预见性远超实验室环境，欠缺可解释性的模型难以保证其在实际工作条件下的稳定表现，存在不可忽视的误判错判风险。尽管基于深度学习的故障诊断模型已在工业领域取得一定应用成果，但缺乏可解释性这一关键缺陷，仍然制约了其在航空航天等高可靠性要求场景的深入应用。近年来，深度学习可解释性研究已引起人工智能、数学和统计等领域的广泛关注，如图 1-1 所示，相关论文数量呈指数级增长。然而，这些研究工作主要以图像和语言文字为研究对象，虽然建立了多种模型解释方法和理论，但并不直接适用于旋转机械故障诊断场景中的振动信号分析。旋转机械智能诊断的可解释研究具有其独

特性，简单地移植现有图像领域解释方法难以获得理想的解释效果。

旋转机械故障诊断的深度学习模型可解释性研究尚处于初始阶段，相关研究文献和成果相对匮乏，构成了一个既具前瞻性又充满挑战的新兴研究方向。其挑战可总结为：(1) 旋转机械振动信号与图像或文本数据存在本质差异，其信号特征难以直观展现，导致现有可解释方法无法直接迁移至旋转机械智能诊断模型；(2) 主动解释方法虽能提升模型透明度，但引入的特定可解释结构往往对模型施加了严格约束，在增强解释能力的同时也限制了模型的拓展性，甚至削弱其诊断性能；(3) 振动信号天然具有不直观性和高维的特性，使得被动解释方法在实际应用中面临解释形式难以理解及计算代价高昂的双重困境；(4) 旋转机械智能诊断领域的可解释性研究呈现零散性和表面性，无法为主动解释、事后解释等多场景应用提供全面而深入的解释方案，阻碍了可解释性技术在实际工业环境中的有效落地。因此，立足于旋转机械的特定领域特点，因地制宜，开展针对性的深度学习故障诊断可解释性研究，实现主动解释方法与诊断性能和模型拓展性的共赢，并优化被动解释方法的表达形式和计算效率，是具有学术理论价值和工业应用价值的重要研究课题。

1.2 基于深度学习的旋转机械智能诊断研究现状

从方法论角度，旋转机械故障诊断主要依靠传统信号分析理论和智能诊断两大范式，其中智能诊断又可进一步细分为基于机器学习和基于深度学习的故障诊断方法。“智能” 的核心内涵在于系统能够从监测数据中自动学习特征与故障模式间的映射关系，无需人工干预即可完成诊断决策过程。在这三类方法中，基于传统信号分析理论的故障诊断以信号处理技术为核心，通过揭示设备的故障特征频率，结合专家分析得出诊断结论。基于机器学习的故障诊断则以特征提取和模式识别为核心，从监测信号中提取故障特征，并对这些特征进行映射学习和分类，其显著特点是可解释性强但依赖于手工设计的特征。

随着计算机技术与图像处理硬件的飞速发展，计算能力的显著提升使以神经网络为代表的深度学习方法的大规模应用成为可能。这一技术进步同时催生了深度学习领域的繁荣发展，深度学习方法在多个应用领域中超越了传统机器学习方法，取得了卓越性能，引发了全球范围内的深度学习研究热潮[11]。以深度学习为理论基础的旋转机械故障诊断提供了一种端到端的诊断范式，此类方法既具备出色的故障特征自动提取能力，又能通过经验风险最小化原则学习故障特征至故障类别的精确映射关系，从而在无需人工干预的条件下完成旋转机械的智能故障诊断[12,13]。近十年来，基

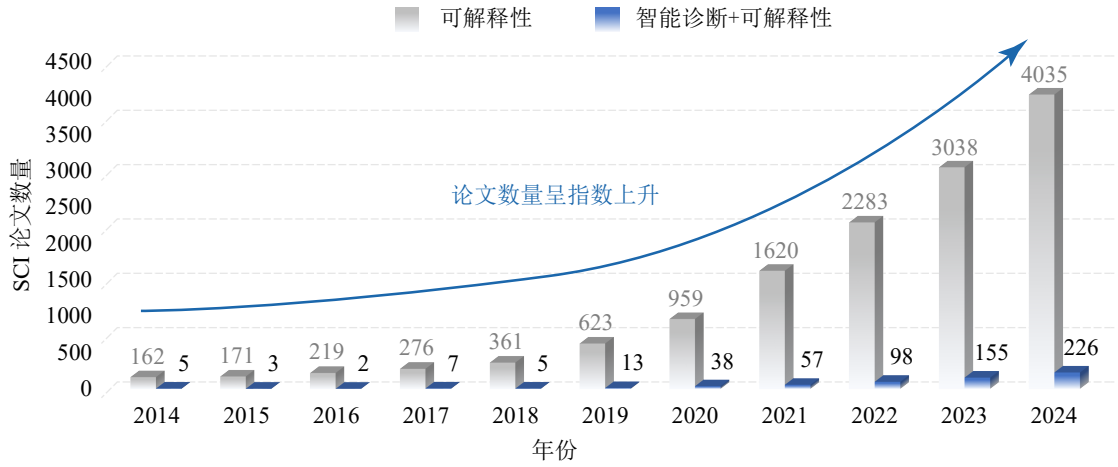


图 1-1　可解释性研究①及其在旋转机械智能诊断领域②的 SCI 论文数量统计

Fig. 1-1　Intrerpretability research and the number of SCI papers in the field of IFD

于深度学习的故障诊断方法呈现出蓬勃发展的态势，按照所采用的深度网络模型架构可系统地分类为：基于自编码器（Auto-Encoder，AE）或受限玻尔兹曼机（Restricted Boltzmann Machine，RBM）的方法、基于卷积神经网络（Convolutional Neural Network，CNN）的方法、基于循环神经网络（Recurrent Neural Networks，RNN）的方法和基于 Transformer 网络的方法。

AE 和 RBM 作为无监督学习的经典方法，能够在不依赖标签信息的情况下对深度网络进行逐层预训练，从旋转机械振动信号中自主提取潜在故障特征。Jia 等[14] 以振动信号频谱为输入，通过 AE 提取判别性故障特征，成功实现了对滚动轴承与齿轮的精确故障诊断。在此基础上，他们提出规范化稀疏自编码模型（Normalized Sparse Autoencoder），进一步提升模型对故障特征的表征能力[15]。Lu 等[16] 和 Shao 等[17] 也分别从自编码器结构优化和训练策略改进方面进行了系统性探索，进一步验证了自编码器类模型在故障特征自动提取方面的有效性。李等[18] 通过叠加多层 RBM 构建深层网络结构，建立基于深度置信网络（Deep Belief Network, DBN），成功实现了轴承多类故障的特征提取和高精度诊断。

CNN 是由 LeCun 最早提出的深度学习方法，其架构设计初衷主要针对图像处理任务。CNN 的基本结构由三类关键组件构成：卷积层负责通过共享卷积核对输入数据实施滑动卷积操作，从而有效提取图像的空间特征；激活层引入非线性变换，显著增强模型的复杂映射能力；池化层则对特征图进行降维压缩，有效减少参数数量并

① 智能诊断：[“*explain**” OR “*interpre**” (Title)] AND [“*AI*” OR “*network*” (Topic)] 在 WOS 核心库的检索结果。

② 智能诊断 + 可解释性：在上述检索式基础上，增加 [“*mechan**” OR “*bearing*” OR “*gear**” (Topic)] AND [“*diagnosis*” OR “*RUL*” OR “*monitor**” (Topic)]。

控制模型规模。CNN 的这种结构设计不仅能够充分捕获数据的空间层次特征，其参数共享机制和特征降维策略还有效抑制了过拟合风险，提高了模型泛化能力。鉴于 CNN 在二维数据处理上的天然优势，研究者们首先探索了将一维振动信号转换为二维表征的预处理策略，以便充分利用 CNN 的特征提取能力进行旋转机械故障诊断。这些预处理方法主要包括：简单拼接、时频变换[19]、谱分析[20] 等。与此同时，研究者们也发现 CNN 同样适用于直接处理一维时序信号，原始时域振动信号、频域变换后的频谱以及经过解调处理的包络谱均可作为 CNN 的有效输入形式。例如，Jing 等[21] 针对复杂齿轮箱的故障诊断问题，提出了以频域数据作为 CNN 输入的诊断方法，取得了良好效果。Li 等[22] 则将原始时域信号直接作为 CNN 的输入，并通过振动信号数据增强，显著提升 CNN 在旋转机械故障诊断中的识别准确率与鲁棒性。

与 CNN 擅长提取空间特征的特性不同，RNN 专注于捕捉数据的时序依赖关系，具备出色的时序信息编码能力。理论上，RNN 能够将历史输入序列的完整信息映射至目标向量，通过在网络内部状态中保持对先前输入的记忆实现时序建模。在传统 RNN 架构基础上，研究者进一步发展了长短时记忆网络（Long Short-Term Memory, LSTM）和门控循环单元网络（Gated Recurrent Unit, GRU）等高级变体。在旋转机械故障诊断领域，Yuan 等[23] 通过对比实验系统评估了三种循环神经网络架构在航空发动机故障诊断任务中的性能表现，实验结果充分证实了 LSTM 和 GRU 相较于传统 RNN 的显著性能优势。Zhao 等[24] 提出将人工特征提取与增强型双向 GRU 架构相结合的混合诊断范式，该方法在变速箱故障诊断和早期轴承故障检测等多种场景下展现出优异的通用适应能力。Shi 等[25] 设计了双向长短时记忆网络（BiLSTM）结构，实现了对多传感器数据的时空特征联合提取，成功应用于变速箱故障类型识别和故障位置定位的精确诊断任务。

Transformer 网络是一种基于注意力机制的深度学习模型，由 Vaswani 等[26] 提出，其在自然语言处理领域取得了巨大成功，被 OpenAI 用于构建 GPT 系列模型[27]，并逐渐引入计算机视觉领域[28] 及其他学科应用中。Transformer 网络通过自注意力机制实现了对序列数据的全局依赖性建模，有效规避了传统 RNN 存在的梯度消失和梯度爆炸问题，同时具备更优的并行计算能力和更高的训练效率。在旋转机械故障诊断领域，研究者们开始探索将 Transformer 网络应用于振动信号的特征提取和故障诊断任务。Hou 等[29] 和 Ding 等[30] 分别以振动信号的频谱和时频谱为输入，借助 Transformer 网络实现旋转机械故障特征提取与高精度诊断。Xiao 等[31] 将注意力机制的权重视为潜在的随机变量，构建贝叶斯变分 Transformer 网络，以增加模型在旋转机械故障诊

断任务下的泛化能力。

相较于基于传统信号分析和机器学习的故障诊断方法，深度学习方法能够充分挖掘历史数据价值，提供了端到端的故障诊断范式，是解决大数据背景下旋转机械故障诊断难题的关键技术路径[6]。一方面，神经网络强大的非线性表征能力使其能够直接从原始信号学习具有高区分性的故障特征，显著降低了对故障机理和信号处理技术先验知识的依赖程度；另一方面，神经网络端到端的诊断模式将特征提取与故障识别作为统一整体进行优化，更具发现全局最优解的潜力。但当前深度学习故障诊断方法主要局限于实验室场景，距离真正的工业落地应用仍存在以下挑战：

(1) 数据获取困难：尽管机械部件多为批量化工业产品，但对于具体机械系统而言，可获取的有效数据样本通常十分有限，针对此问题，研究者将少样本学习方法引入旋转机械故障诊断领域，旨在降低数据需求量的同时，最大程度保证深度学习模型的故障诊断性能。

(2) 类别分布不均衡：在旋转机械完整服役周期中，系统历经正常运行、性能退化和故障诊断维护多个阶段，导致采集的数据中正常样本显著多于故障样本，对此，研究者开展了一系列针对旋转机械类别不平衡的深度学习方法研究。

(3) 模型泛化能力不足：用于训练的旋转机械振动数据在数量规模和多样性上均受到限制，而实际应用环境则更为复杂多变，这使得基于有限数据训练的深度学习模型难以保证在多样化实际工况下的诊断表现，为提高模型泛化能力，研究者积极开展了基于迁移学习的变工况、变设备旋转机械故障诊断方法研究。

上述问题已引起旋转机械故障诊断领域研究者的广泛关注，相关研究工作不断深入，成果日益丰富，使得这些挑战在一定程度上得到了缓解。然而，当前旋转机械深度学习故障诊断方法距离实际工业应用仍面临一个亟待突破的关键瓶颈——可解释性不足。深度学习故障诊断的端到端特性实质上是一把双刃剑：积极层面上，深度学习有效规避了人工干预的主观局限，满足了大数据环境下故障诊断的客观需求，并通过特征提取与模式识别的整体优化实现了卓越的诊断性能；消极层面上，深度学习的黑箱特性也导致诊断决策过程不透明，使得使用者无法理解模型的推理逻辑和判断依据，从而制约了其在高可靠性要求场景中的实际应用。

总结而言，与少样本学习和迁移学习等直接影响诊断性能的方向相比，可解释性研究虽然作用机制较为间接，却在推动深度学习走向旋转机械故障诊断实际应用的进程中扮演着不可替代的基础性角色。可解释性研究的价值在于：(1) 可解释性能够提升运维人员对诊断模型的信任度，避免因信任缺失而引入冗余诊断方案进行交叉

验证，从而有效降低实际运维成本；(2) 可解释性能够深化开发者对诊断模型的内在机理理解，为提高模型诊断性能和纠正误诊提供精准的优化方向；(3) 可解释性能够从模型视角对振动数据进行深入解读，借助模型的高诊断性能归纳提炼出先进的故障诊断知识，推动故障诊断理论发展；(4) 深度学习故障诊断面临的诸多应用难题本质上是一个有机整体，可解释性研究通过增强模型透明度，能够系统性地指导模型小样本学习能力和泛化能力的提升，为旋转机械故障诊断领域的其他技术难题提供间接但有效的解决思路。因此，可解释性研究已成为当前旋转机械深度学习故障诊断领域的一项严峻挑战，是推动该技术从实验室研究迈向工业实践应用中亟待攻克的关键科学问题。

1.3 人工智能领域的可解释性研究现状

近十年来，深度学习模型在计算机视觉[32]、语音识别[33] 和自然语言处理[34] 等领域取得突破性进展，其应用范围正持续扩展至医学诊断[35]、生物信息学[36] 和天文学[37] 等前沿科学领域。然而，深度学习模型在实际应用中仍面临诸多根本性挑战。Szegedy 等[38] 的研究表明，通过对输入图像施加人眼难以察觉的微小扰动，可以导致深度神经网络的预测结果发生显著变化，这类经过精心设计的输入被定义为“对抗样本”。此外，Nguyen 等[39] 通过实验证实，对于完全不具可识别特征的图像（如纯粹的白噪声图像），深度神经网络仍会以接近 100% 的高置信度将其归类为特定对象。这些实验现象表明，尽管深度神经网络在多种任务上展现出超越人类的性能表现，但其内部决策机制与人类认知过程存在本质差异，且这种差异尚未被科学界充分理解和解释。

为深入揭示深度学“黑箱”模型的内在决策机制，学术界展开了广泛而深入的可解释性研究。联合国《人工智能伦理建议书》[40] 将可解释性定义为使人工智能系统的结果可被理解并能够提供清晰阐释的能力。在学术领域，可解释性是指以人类可理解的形式提供模型解释（Explanation）的能力[41]。其中，“解释” 是指可被表述为逻辑规则、可转换为逻辑规则或用于解释的关键元素；而 “可理解的形式” 则是由特定领域知识体系和任务特性所决定的表达方式。如图 1-2 所示，文献[42] 基于解释的类型将现有可解释性研究划分为：规则（Rule）方式、语义（Semantic）方式、归因（Attribution）方式和案例（Example）方式。其中，规则方式将模型的复杂决策过程提炼为一系列显式的逻辑规则，使决策过程更易能被人理解；语义方式探究输入样本所含语义与特定神经元激活模式之间的对应关系，揭示神经元背后的语义表征能力；

归因方式将模型决策量化归因至输入数据的各个组成部分，精确衡量各部分对最终决策的贡献度，从而明晰模型决策的数据依据；案例方式则通过识别与当前输入高度相似的典型案例，借助案例间的类比关系间接地解释模型的决策逻辑。

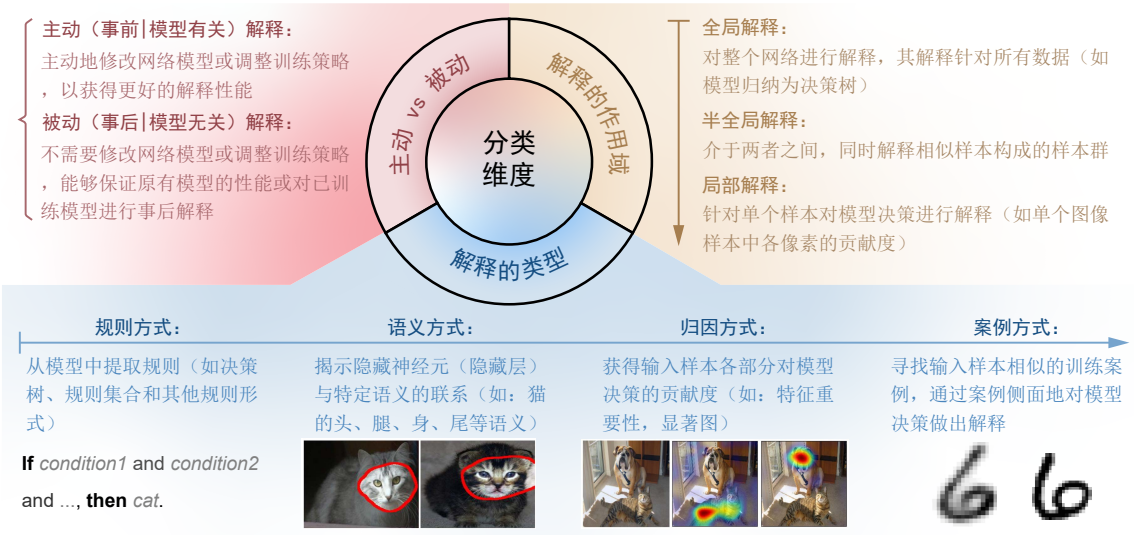


图 1-2　深度学习可解释性研究的三种分类维度[42]

Fig. 1-2　Three classification dimensions of deep learning interpretability research

在深入探讨可解释性方法的研究现状之前，有必要对当前的可解释方法进行系统分类与界定。如图 1-2 所示，可解释性研究可从三个维度进行分类[42]：解释的类型、解释的作用域、以及主动与被动。其中解释的类型已在前文详细阐述，此处不再赘述。就解释的作用域而言，其差异主要体现在解释结果的适用范围上。全局可解释性旨在揭示模型的整体决策逻辑和推理规律，其解释结果适用于所有可能的输入样本，能够提供对模型行为的宏观理解。局部可解释性则聚焦于特定输入样本的单次预测解释，通过分析目标输入的特征重要性或决策路径，提供针对个体样本的精细解释。半全局解释则处于全局与局部之间的过渡地带，既保留了一定程度的全局适用性，又能针对特定样本子集提供更为精准的解释。就主动与被动解释而言，其核心在于是否需要更改神经网络架构或优化过程。主动解释也叫事前（Ante-hoc）解释，要求在模型设计或训练阶段进行干预，如引入特定网络结构或修改训练目标函数，从而在模型构建过程中显性地增强其固有解释性。例如，通过添加正则化项引导深度学习模型向决策树等高度可解释的结构靠拢，或者设计专门的可解释网络层。相比之下，被动解释方法，即事后（Post-hoc）解释，则无须修改原有模型架构与训练过程，而是对已训练完成的模型进行事后解释，这种方法能够最大程度保持原模型的性能特性和拓展潜力。

鉴于模型修改权限是可解释性方法在实际应用的决定性因素，本节以主动解释与被动解释作为主，以解释的类型为辅，对深度学习可解释性研究现状进行全面梳理与深入分析。

1.3.1 可解释性研究中的主动解释

规则方式的解释方法旨在对神经网络模型进行规则提取，从而获得诸如决策树与规则集合等高度可解释的模型表征。此类研究虽以被动解释为主流，但仍有部分方法积极介入网络模型的设计与训练过程，使模型主动向可解释规则结构靠拢。例如，Wu 等[43] 通过决策树构建正则化损失函数，促使神经网络模型更易被浅层决策树有效近似。这种树正则化方法的核心理念在于通过显式约束模型行为，使其决策过程能够被简单决策树精确表达，从而实现全局层面的模型可解释性。该方法还进一步扩展至半局部层面，能够针对特定数据子区域提供更为精细的解释能力[44]。然而，值得注意的是，规则方式的可解释性方法主要集中于神经网络早期发展阶段，随着神经网络规模与复杂度的持续增长，所提取的规则日益难以准确反映神经网络的内部决策机制，因此规则方式的解释研究逐渐从全局解释转向局部或半局部解释领域。

语义方式的解释方法源于神经科学中“祖母细胞假说”，即特定的语义概念会选择性地激活特定的隐藏神经元。基于此理论基础，此类方法致力于揭示神经网络中特定神经元与具体语义概念之间的内在对应关系，例如在动物识别任务中发现负责识别马的身体、头部、尾部等特定语义部位的神经元，这些方法普遍具有全局解释的特性。在语义方式的可解释方法中，主动解释旨在通过结构设计和训练策略引导卷积神经网络学习更为优化且解耦的隐藏语义表征。Zhang 等[45] 对损失函数设计，用以有效促使高层滤波器对应于单一且明确的语义概念。其方法核心在于预先定义一组理想的、类似于高斯分布的特征滤波核作为参考基准，并引入基于互信息构建的正则化损失项，使得每个滤波器要么呈现高度一致的激活模式，要么保持完全不激活状态，从而显著增强了网络内部表征的语义透明性和可解释性。

归因方式的解释方法旨在通过将决策结果归因至输入数据的各个组成部分，量化输入特征对模型决策的贡献度。值得注意的是，该类方法仅能在线性模型等简单结构上实现全局解释（例如线性模型中的权重系数直接反映特征重要性），而在深度学习等复杂模型中则主要提供局部解释。在主动解释范畴内，研究者通过引入特定的正则化项或先验知识，有效引导模型形成更为精确的归因解释结果。例如，Plumb 等[46] 在训练过程中融入了局部归因高保真度和稳定性的可解释性正则化项，显著提升了局部归因解释的质量。而 Weinberger 等[47] 则在此基础上进一步整合领域先验知

识，实现了更具说服力的归因解释。除了针对单个输入样本的归因解释外，Dual-net[48]被提出用于评估群体层面的特征重要性，该方法致力于为输入样本群体集体确定“最优”特征子集，从全局视角对特征重要性进行系统性排序与评估。

案例方式的解释方法通过寻找输入样本相关的支撑案例或反支撑案例，从而侧面地对输入样本做出解释。在主动解释领域，Li 等[49]将原型层整合入神经网络架构，基于编码后的输入与学习到的原型之间的相似度关系进行预测决策，并通过引入距离正则化项，使所学习的原型具备清晰的案例解释能力。Chen 等[50]在此基础上进一步将原型层技术拓展至标准卷积神经网络结构中，实现了输入图像局部区域与原型案例的精细化相似性匹配，从而达成更为细粒度的案例式解释机制。

1.3.2 可解释性研究中的被动解释

规则方式的被动解释方法大多集中于全局解释，仅部分方法提供局部解释。全局解释方法可进一步细分为分解法和教学法两大类。分解法立足于模型内部结构分析，通过对神经网络进行系统性分解和解耦，提取出明确的决策规则，代表性方法包括 KT 法[51]、“M-of-N”法[52]和 NeuroRule 法[53]。教学法则将模型视为完全的“黑箱”，通过构建与目标模型的持续交互机制，逐步重构出等效的决策规则或决策树，其核心研究重点涵盖规则提取策略优化[54,55]以及模糊规则构建[56]等方向。在局部解释领域，基于特征扰动的支撑解释[57]和反事实解释（Counterfactual Explanation）[58]是一种典型的方法，其解释形式为：“由于样本中特征 A 存在且特征 B 不存在，导致样本 x 被归类为 y 类”以及“样本的特征 A 使预测结果为该类，而非其他类”。此外，CDRPs（Critical Data Routing Paths）方法[59]通过揭示模型处理特定样本的关键信息流动通路，为局部解释提供了另一种有效视角。

语义方式的被动解释方法主要基于激活最大化（Activation Maximization，AM）原理，即寻找能够最大程度激活特定神经元、通道或网络层的输入样本。由于理论上需要在整个样本空间进行搜索，AM 方法面临高维搜索空间的挑战。针对此问题，Erhan 等[60]率先提出将先验信息融入正则化项的优化框架，随后 L2 正则化、变分正则化[61]和裁切正则化[62]等被陆续引入以约束搜索空间，提高搜索效率。近年来，生成对抗网络也被应用于激活最大化样本的高效生成过程，进一步拓展了 AM 方法的应用边界[63]。除激活最大化外，研究者还深入探索了卷积核与图像材质等语义概念间的对应关系。Bau 等[64]提出的 Network Dissection 方法对卷积核与特定视觉概念的关联程度进行系统量化；在此基础上，Fong 等[65]进一步研究了语义概念在多个卷积核组合中的嵌入模式，揭示了更为复杂的语义编码机制。此外，语义方式的可解释研

究已扩展至自然语言处理领域，如 Dalvi 等[66] 对自然语言处理任务中神经元与特定词汇间的映射关系展开了深入分析，为语义解释方法在多模态领域的应用提供了重要参考。

归因方式的被动解释方法是整个可解释领域研究的主流方向之一，主要可分为基于模型梯度和模型无关的两大类方法。基于模型梯度的归因方法通过利用模型反向传播机制，计算样本相对预测结果的梯度分布，从而量化样本各组成部分对模型决策的重要性贡献。这一方法的理论基础在于梯度表征了损失函数中增长最快的“方向”与速率，揭示输入样本需要沿哪个“方向”变化才能使其更接近（或远离）目标类别。Baehrens 等[67] 率先将梯度归因方法应用于高斯分类过程、k-NN 和 SVM 等传统机器学习模型的解释中。随着研究深入，基于模型梯度的归因解释方法逐步拓展至计算机视觉领域，形成普通反向传播[68]、引导反向传播（Guided Backpropagation）[69]、类别激活映射（Class Activation Mapping，CAM）[70] 和梯度类别激活映射（Grad-CAM）[71] 等一系列经典方法。然而，神经网络中非线性层普遍存在梯度饱和现象，严重制约了梯度信息的解释效力。针对此问题，DeepLIFT[72]（Deep Learning Important FeaTures）和 LRP[73]（Layer-wise Relevance Propagation）等方法提出计算“离散梯度”的策略，本质上是对梯度函数的反向传播过程进行有针对性的修正。但值得注意的是，离散梯度并不完全符合链式求导法则，这一理论缺陷限制了其在复杂模型中的应用。为解决此问题，Sundararajan 等[74] 提出了积分梯度法，该方法通过对输入样本与参考输入之间直线路径上所有点梯度的积分运算，在解决梯度饱和问题的同时保证了链式求导法则的适用性。对于模型无关的归因方法而言，其主要可细分为敏感性扰动法[75]、以 LIME[76]（Local Interpretable Model-agnostic Explanations）和 MAPLE[77]（Model Agnostic suPervised Local Explanations）为代表的局部模型简化法，以及以 Shapley 值[78] 和互信息[79] 为代表的博弈论与信息论方法。敏感性扰动法的核心思路是通过对样本实施系统性修改（如扰动[80]、遮挡[81] 等操作），观察模型输出的变化响应，进而衡量输入样本各部分对模型决策的敏感程度。LIME 和 MAPLE 则采取局部模型简化的策略，在数据的局部邻域内将复杂模型近似为高度可解释的线性模型，从而实现半全局范围的归因解释，两者主要区别在于局部邻域的定义方式。SHapley 值和互信息方法则是借助相关理论框架，定量估计样本各特征对模型决策的贡献度或信息关联程度，为归因解释提供了坚实的理论支撑。

案例方式的被动解释方法主要聚焦于从已有训练样本中寻找对模型决策具有关键影响的案例。Koh 等[82] 通过估算训练样本对模型参数的影响，继而量化这种参数

变化对测试点预测结果的作用，从而识别对特定测试样本预测最具影响力的训练案例。与此相关，Yeh 等[83] 证明了分类前的 logit 层可被分解为训练样本在 pre-logit 层激活值的线性组合，其中训练点系数清晰指示了测试样本与这些训练样本之间的相似性关系，为模型行为提供了基于案例的直观解释。

1.3.3 可解释性研究的总结

深度学习可解释性研究主要聚焦于计算机视觉与自然语言处理领域, 呈现出蓬勃发展态势。这些研究从主动和被动两种场景出发，形成规则提取、语义解析、特征归因和案例匹配四种解释类型。可解释性研究有效增强了深度学习模型的透明度与可信度，并提高可信度和接受度，拓展了深度学习技术在医疗诊断、自动驾驶和金融预测等高风险决策场景的应用。

然而，计算机视觉领域的可解释性研究与旋转机械故障诊断领域存在本质差异。(1) 数据理解难度：图像具有丰富且明确的语义信息，人类可直接理解，而旋转机械振动信号属于典型的非过程数据，需经专业信号处理才能揭示其物理含义，这使得振动信号模型的解释难度显著增加；(2) 模型机理不透明：深度学习处理图像的理论基础已相对完善，如卷积神经网络的滤波匹配、语义提取和分类机制，而深度学习对振动信号的处理逻辑尚未形成清晰理论框架，所提取特征缺乏明确的语义解释；(3) 解释形式受限：图像的解释结果可被人类直观评估，但振动信号的直接解释结果难以被非专业人员理解，这要求旋转机械智能诊断模型必须提供更适合人类认知的解释形式。因此，将计算机视觉领域的可解释方法直接移植至旋转机械智能诊断场景中，不仅难以获得理想的解释效果，更可能导致解释结果的偏差与误导。

1.4 旋转机械智能诊断领域的可解释性研究现状

旋转机械智能诊断领域的可解释性研究既至关重要又极具挑战性。可解释性研究既能增强用户信任、指导模型改进与归纳诊断知识，也能在一定程度上辅助应对其他应用难题。然而，图像、文本与旋转机械振动信号在数据形态和特征上存在显著差异，使得主流解释方法难以直接迁移至旋转机械故障诊断场景。因此，结合旋转机械振动信号自身特点开展针对性的可解释性研究，具有关键的重要价值与研究意义。

在旋转机械故障诊断领域中，现有可解释性研究主要围绕“先验赋能主动解释”[84] 与“归因被动解释”[85] 两方面进行探索。前者利用丰富的故障诊断先验知识，对诊断模型的局部或整体进行针对性改造，以揭示故障诊断黑箱模型的潜在机理；后

者则在既有诊断模型的基础上，通过梯度传播、输入扰动和权重可视化等手段，量化输入各部分对模型决策的贡献度，从而解释模型的决策依据。

1.4.1 旋转机械智能诊断领域的先验赋能主动解释

虽然振动信号不够直观、理解难度更大，但这也促使研究者们不断对微弱特征提取、故障模式识别等方法方面进行迭代，积累了包括信号处理、专家经验等的多方面先验知识。利用这些先验知识来赋能神经网络，便成为了实现黑箱模型主动解释的可行方案。

(1) 基于信号处理的主动解释

将信号处理方法融入神经网络是一种常见的主动解释策略，即以信号处理为出发点，主动设计网络的局部模块或完整结构，在引入信号处理先验知识的同时赋予神经网络相应的解释能力。从局部模块角度来看，已有研究大多关注基于信号处理的可解释网络第一层，例如，Abid 等[86] 将 SincNet[87] 应用于电机故障诊断，通过约束 CNN 输入层使其等价于带通滤波器从而提高诊断精度，但该工作尚未明确强调模型解释性，后续研究针对 SincNet 进行了进一步改进[88]。Li 等[89] 则在此基础上迈进一步，将 CNN 输入层约束为连续小波变换（Continuous Wavelet Transform，CWT）以提高模型诊断能力，并借助输出分析凸显其捕捉冲击特征的优势。He 等[90,91] 则持续聚焦小波变换与卷积层的融合，包括对初始化方式[92] 及卷积核调控机制[93] 等的探索。Yuan 等[94] 参考可提取冲击特征的提升小波核构建了提升小波网络，并验证小波核可随数据形状收敛，从而印证其对数据的适配能力。Han 等[95] 则基于离散小波变换（Discrete Wavelet Transform，DWT）设计了新型卷积层，结合自注意力机制深度融合所提取特征，并通过分析网络各层输出来解释模型行为。在 CNN 框架之外，Li 等[96,97] 进一步将小波包分解引入图神经网络的特征提取环节，以应对噪声工况下的故障诊断问题，并通过信号包络谱验证其在去噪和保留故障相关分量方面的有效性。Zhu 等[98] 则将可解释局部模块与数字孪生技术结合，用于发掘潜在故障特征。这些方法在网络局部模块中融合信号处理先验知识，既能在降低数据需求和缓解过拟合的同时提升准确率及少样本诊断性能，又因局部约束而仍具备即插即用等优势。不过，这类方法的可解释性仍相对有限，尽管研究者可通过逻辑梳理或输出特征可视化的方式来理解网络工作机制，但对专业知识有较高要求，且缺少直观明晰的解释结果作为输出。

除局部模块外，信号处理方法也可用于指导网络整体结构的设计。Michau 等[99]

借鉴 DWT 提出了可学习的降噪稀疏小波网络，通过层层级联将输入信号可逆地分解为一系列系数，并设置合适阈值使系数稀疏化，再利用重组信号实现降噪。从小波分解角度出发，Wang 等[100] 设计了可模拟小波包分解的新型模块，通过堆叠实施信号的分解与重构，并借助自注意力机制完成特征加权。Wang 等[101] 则提出可微分、频率可学习的离散小波模块，并通过堆叠方式构建小波引导网络，借助对不同离散小波模块的输出分析来验证模型对关键故障频率的提取能力。Li 等[102] 在可学习的小波包分解网络基础上，将系数降噪阈值设计为与噪声偏差相关的自适应学习结果，实验表明该方法在轴承故障特征频率提取方面成效显著。在小波分解之外，Liu 等[103] 基于小波散射变换理论构建小波散射网络，通过参数固定的卷积层、模量非线性激活函数及池化层的组合，实现小波散射变换的端到端可微分处理，并利用全连接层将提取的散射特征映射至故障类别空间。Liu 等[104] 还进一步提出归一化小波散射网络，从理论上严格证明了归一化散射特征对线性时不变系统的不变性特征。Li 等[105] 引入可学习的 Morlet 小波作为信号运算层并进行多层堆叠，在末端结合统计特征分类器构建完整的透明操作网络，最终通过小波核的收敛结果及各层输出的可视化来解释模型决策过程。Wang 等[106] 则提出了全可解释的机器状态监测网络框架，将小波变换、平方包络和 Fourier 变换融合为网络的预处理层，再借助稀疏测度与极限学习机相结合，完成从信号处理到故障分类的完整流程可解释性，以及对轴承关键频带的精准定位。西交严如强教授团队[107] 参考传统信号处理诊断的完整流程，构建包含小波卷积、动态硬阈值、基于指标的自适应滤波以及分类在内的信号处理信息网络框架，并通过阈值化后的特征可视化与小波核累积频带以验证网络解释能力。网络整体结构的统一设计不仅能提升故障诊断能力，也使模型具备更强的解释特性，但全局约束同时限制了网络的灵活性和扩展空间，削弱了神经网络所特有的自由拓展和定制化能力。此外，此类方法仍缺乏直观的解释结果输出。

(2) 基于稀疏理论的主动解释

信号处理之外，稀疏理论也是信号特征提取的重要方法，不少研究者将稀疏理论引入可解释网络的结构设计之中。An 等[108] 将嵌套迭代软阈值算法与稀疏编码相结合，构建了具备完全可解释性的嵌套迭代软阈值网络，它通过迭代方式将输入信号分解为字典基的线性组合，并以字典基稀疏的稀疏程度为目标展开优化。其实验结果表明，基于此分解系数重构出的信号在有效保留关键故障特征的同时，可显著衰减噪声成分。为进一步增强网络的抗噪鲁棒性，Zhao 等[109] 对多层稀疏编码模型展开深入研究，创立了跨层面的 Generalized Sparse Coding 算法，并在此基础上推导出层级理论

框架，将该框架展开为可训练的层级稀疏编码网络结构。借助高效的稀疏特征提取机制，该方法大幅提升了网络对故障特征的表达能力与抗噪性能。An 等[110] 又进一步将此思路推广至异常监测任务，并利用所学习出的字典对关键故障的局部冲击特征进行解释。除此之外，Wu 等[111] 则提出了一种基于稀疏约束的可解释卷积网络，利用信号频谱作为输入，并通过抗锯齿约束与 L1 正则化约束让乘法滤波核参数实现稀疏化，从而增强网络对关键故障特征的提取能力。上述方法均通过稀疏理论主动构建完整网络，以便更精准地提取关键故障特征。不过，稀疏性的严格约束也限制了网络结构本身的灵活调节空间，且其解释结果缺乏更直观的输出形式。

(3) 基于注意力机制或网络改造的主动解释

注意力机制作为深度学习中的重要创新方法，能够自适应地为输入数据中的关键特征赋予更高权重，实现对诊断决策具有价值的信号成分的精准筛选，天然地具备一定的可解释能力。在此基础上，部分研究者通过主动引入额外约束进一步提升注意力机制的解释性。Liu 等[112] 将变分模态分解与 Transformer 相结合，并将传统的位置编码替换为信号中心频率，最终通过注意力权重可视化实现了频域归因。Li 等[113] 则将变分推断理论嵌入注意力机制，构建了变分注意力 Transformer，通过将注意力权重转换为 Dirichlet 分布并施加稀疏约束，显著增强了时域归因结果的清晰度与先验知识一致性。除注意力机制外，直接对网络结构进行改造也是实现主动解释的另一途径。Kim 等[114] 将网络激活与池化分别替换为 Softsign 激活和全局能量池化，从而更好地保留输入信号中蕴含的频率特征，获得更理想的频域归因可视化效果。Guo 等[115] 则利用因果学习思想设计了由诊断器与解释器组成的新型网络，其中解释器以条件互信息为指标对因果特征进行辨识，并将其输出的掩码结果纳入诊断网络的优化过程，最终生成的掩码可有效解释输入频谱中的关键成分。

(4) 智能诊断领域主动解释研究的总结

智能诊断领域的主动解释研究仍处于探索阶段，与之关联的实现路径较为多样，且所获得的解释结果也存在差异。总体而言，利用信号处理、稀疏约束等先验知识对网络的局部或整体进行设计，一方面能够将先验知识融入网络之中，从而提升诊断性能；另一方面也使网络结构相对透明，可通过输出特征的可视化实现一定程度的解释与验证。然而，该类方法同样面临两方面局限：首先，该类方法往往对网络结构施加过强约束，导致模型虽在已知实验环境中表现良好，但在未知应用场景中缺乏足够的灵活性与拓展性；其次，该类方法缺少明确的可视化解释结果输出，大多依赖特征可视化来表达模型可解释性，这不仅对专业知识有较高依赖，而且只能在有限程度上揭

示模型的决策依据和决策逻辑。

1.4.2　旋转机械智能诊断领域的归因被动解释

故障诊断的应用场景往往十分复杂，诊断需求也各不相同，并非在所有情形下都能为了增强解释能力而主动调整网络结构。实际上，可解释性所带来的收益更为间接，大多作为诊断系统中的辅助模块，用于支撑现有模型的决策结果。由此，对既有网络进行被动解释更加直接而实用。现有被动解释研究主要集中在归因分析上，通过梯度传播、输入扰动以及权重可视化等手段，揭示输入中各部分对模型决策的贡献度，从而解释模型的决策依据。

(1) 基于梯度传播的被动解释

以 CAM 和 LRP 为代表的梯度传播类方法是其中主要方式，其核心思路在于通过对梯度（或贡献度）进行逐层反向传播，将网络输出的类别激活映射回输入空间，从而实现各输入部分的归因。在特征层面，Wu 等[116] 将时域、频域和时频域共 35 个常用特征输入至 LSTM 模型，并借助 LRP 获取不同类别样本中各特征的贡献度。

在现有研究中，通过数据预处理将一维振动信号转换为二维表征谱，进而沿用图像领域的梯度传播解释方法是一种常见的做法。Grezmak 等[117] 使用 CWT 将感应电机振动信号转换至时频域，并通过 LRP 对不同频段进行归因，实验结果表明 CNN 决策主要依赖特定频带。Chen 等[118] 采用短时傅里叶变换（Short-Time Fourier Transform，STFT）将滚动轴承振动信号转换为时频图，再利用 Grad-CAM 揭示时频图各成分对 CNN 决策的贡献度，并将归因结果与决策树进行交叉验证。Wu 等[19] 则将水轮机振动信号经 STFT 预处理后输入 AE 模型，通过 Grad-CAM 对时频域样本进行归因解释分析，利用贡献度热力图定位轴承的故障频率。Liu 等[119] 采用小波包分解作预处理，结合 Grad-CAM 对磨削齿轮诊断的 CNN 模型进行事后归因分析，定位关键频率的贡献度。在预处理方式优化方面，Kim 等[120] 在预处理阶段引入转速归一化和相位采样等解调方法，将含有频域阶次轴和时间轴的二维解调谱作为输入，并借助 LRP 进行被动解释归因。为改进解释方式，Sun 等[121] 通过梯度平滑、去除 ReLU 和全局平均等操作对 Grad-CAM 进行改进，并利用频率切片小波变换作为数据预处理，为 CNN 模型提供时频域归因。其实验表明，改进后的 SGG-CAM 所得到的类激活图对故障成分的贡献度更加集中。此外，Brito 等[122] 将简单频域变换的预处理与 Grad-CAM 相结合，估计旋转机械中频域成分的贡献度，Mey[123] 也进行了类似探索。然而，此类方法虽然实现简便，但由于信号表征谱与自然图像在理解上存在较大差异，解释效果

有限；且预处理的引入破坏了神经网络的端到端优势。

为保持诊断模型的端到端优势，梯度传播解释方法也被直接用于时域信号的归因分析。Yu 等[124] 基于轴承时域信号训练了残差网络模型，并分别采用 Grad-CAM 和 Eigen-CAM 来衡量时域信号各组成部分的贡献度。Chen 等[125] 将滚动轴承采集的时域信号输入 BiLSTM 进行训练，并借助 Grad-CAM++ 可视化网络所关注的时域片段。其实验结果表明，模型决策主要依赖于信号中的周期性成分。随后，Chen 等[126] 针对梯度传播解释不稳定的问题，通过扰动策略与得分加权来构建梯度得分 CAM，实验证明该方法能够更集中地聚焦于时域信号中的关键故障片段，提升视觉解释效果。Miettinen 等[127] 则利用扭转振动信号信噪比更高的优势，将其时域信号用于 CNN 的训练和诊断，并借助 Grad-CAM 评估各时域分量的贡献度。结果显示，对扭转振动信号起主导作用的低频率成分对故障类别具有更大影响。尽管针对时域信号的梯度传播归因方法能够很好地保留诊断网络的端到端特性，但由于所得时域解释难以直接反映机械系统的内在机理，上述方法仅能揭示冲击出现的时刻或周期性信息，解释效果并不直观。此外，梯度归因本身也存在结果不稳定的问题，常在同类样本之间出现差异，并且易关注到无关区域。

考虑到时域解释结果往往不够直观，一些研究者在此基础上引入后处理操作，以更清晰地揭示时域解释背后的物理意义。Li 等[128] 在完成轴承故障诊断网络训练后，利用积分梯度法获取不同类别样本的时域解释结果，并进一步通过快速 Fourier 变换进行后处理，将其转换到频域加以分析，最终确认了轴承外圈和内圈特征频率对于模型决策的重要作用及错误诊断案例的成因。Li[129] 则以网络各层的数据特征与反向梯度的相关性系数作为加权依据，对不同层级的 CAM 结果完成加权融合，从而构建多层 Grad-CAM 被动解释方法。在解释性分析中，他将所获得的时域归因结果通过后处理转换至频域，并与原始 Grad-CAM、LRP 等作对比，验证所提方法在捕捉转频特征方面的优势。尽管这类方法借助后处理可间接获取其他域的归因结果，但其本质仍基于时域贡献度分析，解释效果依旧难以令人满意。

(2) 基于扰动的被动解释

在梯度传播类型之外，以 SHAP（SHapley Additive exPlanations）为代表的扰动解释类型同样应用广泛。SHAP 将模型视为完全黑箱，仅通过与模型进行交互来确定输入各部分对模型决策的影响关系。从特征层级来看，Brito 等[130] 分别针对轴承与齿轮故障诊断任务，构建了峭度、均方根值和特征频率幅值等多种特征，并利用 SHAP 进行特征归因，以衡量不同特征对模型决策的影响程度。Lee 等[131] 和 Jia 等[132] 则分

别基于电机电流信号与搅拌釜监测信号构造多个特征，随后使用 SHAP 评估各特征的贡献度。在特征层级之外，直接对时域信号实行扰动往往难以展现物理意义，由此研究者尝试通过其他域对时域样本进行扰动解释。Gwak 等[133] 提出的基于能量扰动的决策边界分析法（Power-Perturbation-Based Decision Boundary Analysis），先对时域信号进行关键频带的迭代式筛选，再在所选频带中对带宽与强度实施扰动，以观测扰动对模型输出的影响，从而发掘样本的决策边界。另有一些工作将 SHAP 进一步扩展到其他域，从而在端到端框架下获得更具物理意义的解释结果。Decker 等[134] 将包络谱分析和 SHAP 方法相结合，通过样本预处理与构建包含逆变换的组合模型，使端到端诊断模型的 SHAP 归因结果可以落脚到更能反映调制频率的包络谱域，获得更清晰的解释形式。Herwig 等[135] 同样将 SHAP 引入频域或时频域，通过对频谱或时频谱的扰动来阐明不同信号成分对于端到端模型的边际贡献。总的来看，扰动解释类型完全不依赖模型结构，但解释对象局限在特征层级，对时域样本的解释效果不佳、且扰动的方式不够明确。尽管一些方法将扰动解释结果转换至其他域，在端到端框架下获得更具物理意义的解释结果，但往往需要大量的交互、较为耗时，且解释结果的清晰性和深入性还有待提高。

(3) 基于注意力机制的被动解释

注意力机制通过模仿人类视觉过程，为输入数据的各部分赋予不同权重，这些权重与所承载信息的重要程度呈正相关。由此，利用注意力机制权重便可直接揭示模型对输入数据的关注区域，从而实现归因解释[136,137]。Li 等[138] 将注意力机制应用于轴承故障诊断网络，并分别以时域信号和包络谱信号作为模型输入，进而根据注意力权重在时域和包络谱域进行归因。Wang 等[139] 则将注意力机制与一维 CNN 融合以实现滚动轴承故障诊断，通过注意力权重可视化揭示模型对不同时域片段的关注程度差异。Tang 等[140] 将 Transformer 用于变工况故障诊断，并通过注意力权重发现，健康样本的贡献度相对均匀，而故障样本的贡献度则集中于少数冲击片段。Cui 等[141] 将 Transformer 用于自监督轴承故障诊断，借助注意力权重可视化发现故障样本的贡献度相对集中，不同长度的同类信号在贡献度分布上也保持相似。Wang 等[142] 将 Transformer 应用于少样本滚珠轴承故障诊断，通过注意力权重可视化表明，包含故障脉冲的片段贡献度通常更高，而趋于随机噪声的片段贡献较低。尽管注意力机制能够直接揭示模型对输入的关注程度，但其前提是网络具备注意力结构，同时其解释结果仅局限于时域信息，仅仅能够揭示关键时域片段，而缺乏更为深入的物理内涵。

(4) 其他类的被动解释

在此之外，亦有少量研究从语义角度对模型进行被动解释[143]。Yang 等[144] 将激活最大化思想引入滚动轴承故障诊断任务，使随机生成的频域输入沿着梯度方向逐步迭代，最终得到可使各类神经元最大激活的虚拟样本。其实验结果表明，转速和故障特征频率以及它们的谐波与边频带，均是促使类别神经元达到最大激活的重要频率。

此外，还有部分工作结合故障诊断任务对网络结构各层的作用展开深入分析。Borghesani 等[145] 将卷积神经网络的典型架构与滤波、下采样和包络等信号处理技术相联系，解析了卷积层、激活层、池化层在信号处理中的映射关系，从而将典型卷积神经网络与类似金字塔形的频率分析方式相对应，并在仿真信号和实测轴承信号上通过特征可视化验证了相关理论。与之相似，Pang 等[146] 则基于信号处理视角对不同激活函数和池化函数的作用机理进行剖析，进而展开针对性设计，提出了更具抗噪性和可解释性的轻量网络。此类研究可从信号处理或语义层面对网络结构展开剖析，获得独特的解释结果。然而，其分析对象往往局限于经典网络或浅层网络，不具备更广泛的适用性，且解释结果理解门槛较高，缺乏直观而有效的呈现形式。

(5) 智能诊断领域被动解释研究的总结

总结而言，归因被动解释方法无需介入模型结构或训练过程，约束较少，兼容性与实用性更佳。其中，梯度传播类型通过逐层传递梯度或贡献度实现归因，尽管计算效率较高，却常面临解释结果不够稳定的问题。其归因解释形式受输入样本所在域的影响，端到端模型往往只能在不够直观的时域进行分析，若要获得其他域上的解释结果，则需要借助前处理或后处理技术。扰动解释类型完全不依赖网络结构，其扰动可与信号处理方法相结合，从而将端到端模型的解释结果拓展至频域、时频域、包络谱域，以获取更具物理意义的解释形式。但其解释结果在清晰度与深入性方面仍相对不足，且需要与模型多次交互，造成较高的计算开销。注意力机制天然具备归因解释能力，但需网络具备对应结构，适用性并不广泛，且解释结果仅限于时域，缺乏更深层的物理信息。语义类型和分析类型的解释方法则能提供独特的解释结果，但分析对象较为局限，且对专业知识依赖度高，理解门槛随之提升。

1.5 现有研究存在的问题

从研究现状可知，故障诊断对于保障旋转机械系统安全运行至关重要，而深度学习凭借强大的非线性映射能力、高度自动化与数据驱动等特性，能够在大容量、低密

度、多样性和时效性的大数据背景下，为旋转机械故障诊断提供了有力解决方案。尽管深度学习智能诊断在走向实际应用的过程中仍面临诸多有待解决的难题，但深度学习可解释性研究却是其中一项虽被低估却极为关键的核心问题，在提升模型可信度、指导针对性修改、获取先进诊断知识以及推动其他相关问题的解决方面，具有重要作用。旋转机械深度学习故障诊断的可解释性研究整体尚处于起步阶段，众多学者围绕先验赋能主动解释和归因被动解释两方面展开了一系列研究，其中存在以下关键科学问题：

(1) 智能诊断主动解释与诊断性能、模型可拓展性的多方共赢问题

主动解释方法通过引入先验知识加以约束，既能提升网络的诊断性能，又可利用特征可视化帮助专家理解网络的决策过程。然而，现有方法在可拓展性和解释效果两个方面仍然存在不足：(1) 在可拓展性方面，先验知识所带来的约束虽增强了可解释性，但也对网络结构施加了限制，导致其在应对复杂应用场景时缺乏足够的灵活性与拓展性。(2) 在解释效果方面，此类方法的解释结果多侧重于特征可视化，不仅直观性较弱，且需要专业知识配合才能准确理解，导致其在实际应用中效益相对有限，仅能在一定程度上增进对模型内部机理的认知。因此，如何在保证网络诊断性能与可拓展性的同时，进一步增强主动解释能力并形成更直观、实用的解释结果输出，是当前故障诊断主动解释方法的研究重点。

(2) 智能诊断被动解释中形式不直观和计算高耗时的优化问题

被动解释方法无需参与模型的设计或训练过程，能够作为独立模块对现有模型开展分析，兼容性与实用性更强。然而，现有被动解释方法在解释形式和计算效率两个方面仍然存在不足：(1) 在解释形式上，梯度类与注意力机制类方法的可视化结果依赖输入样本所在域，对于端到端模型只能在并不直观的时域进行解释，若要获得其他域的解释结果则需要配合前处理或后处理技术，从而使解释效果受限；扰动类方法可以将解释形式扩展至频域、时频域以及包络谱域，但其解释结果在清晰度与深入性方面仍有待增强。(2) 在解释效率上，扰动类方法需多次与模型进行交互，计算开销显著，而振动信号的高维特性愈发加剧了这一耗时问题。因此，结合故障诊断这一任务特点，进一步探索更为清晰、准确的解释形式，并着力优化高耗时归因解释的计算效率，是现阶段故障诊断被动解释方法的研究重点。

1.6 本文主要研究内容

针对 1.5 节提出的两个关键问题，本文围绕旋转机械深度学习故障诊断的可解释性需求展开系统性研究，分别从主动解释与被动解释两方面出发，致力于建立约束程度低且解释效果明确的高兼容性主动解释网络，同时形成解释形式清晰、计算效率高效的高性能被动解释方法，从而构建面向旋转机械智能诊断的系统性解释框架，为提升模型可信度、揭示模型决策依据和逻辑提供切实可行的解决方案。本文的章节结构及其内在逻辑关联如图 1-3 所示，各章节具体安排如下：

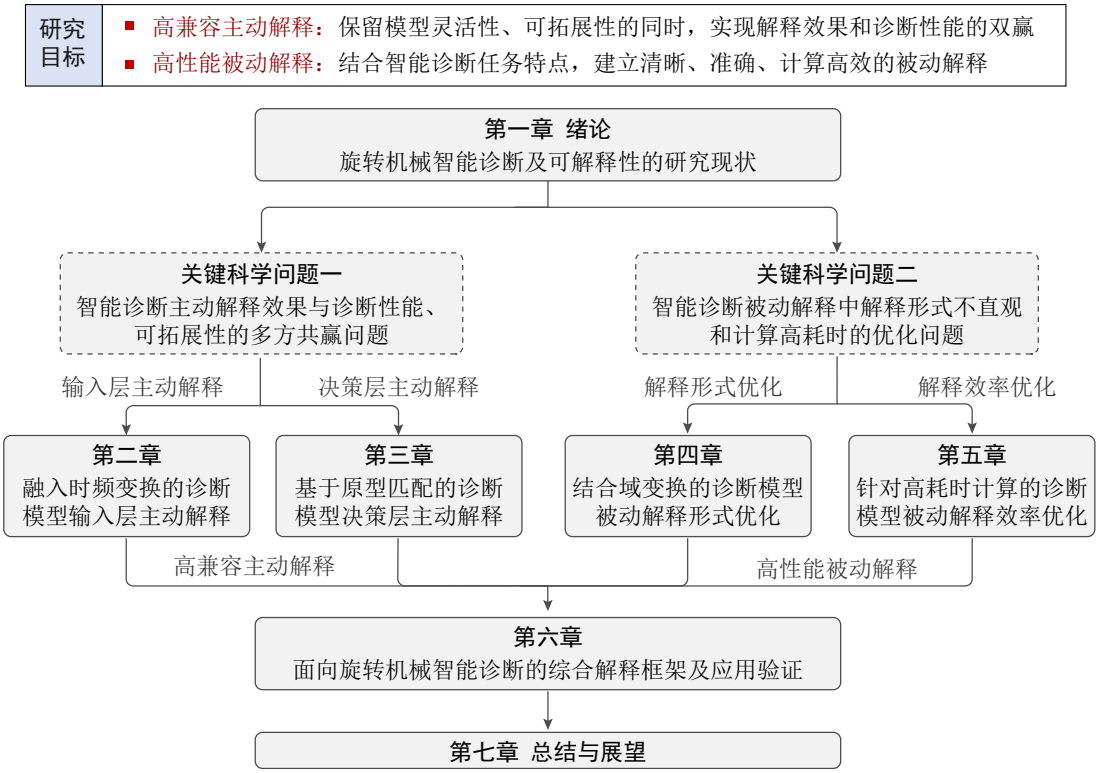


图 1-3 本文的章节结构

Fig. 1-3 The structure of this thesis

第一章，绪论。本章系统梳理旋转机械智能诊断、深度学习可解释性以及二者交叉研究领域的国内外研究现状，深入分析智能诊断可解释性研究在主动解释与被动解释两方面的关键挑战，进而明确本文的研究思路与框架结构。

第二章，融入时频变换的智能诊断模型输入层主动解释。本章立足于智能诊断模型的输入层设计，基于时频变换与卷积层在内积运算上的理论等价性，构建具有物理可解释性的时频变换预处理层，实现模型主动解释、诊断性能和可拓展性的多方共赢。具体而言，首先阐述时频变换与卷积层的等价性并设计时频变换核函数，随后详

细介绍时频变换卷积层的结构设计及可解释性分析方法，最后通过三个典型数据集验证所提方法在综合诊断性能与可解释性方面的优势。

第三章，基于原型匹配的智能诊断模型决策层主动解释。本章聚焦智能诊断模型的决策层设计，将人类认知中的原型匹配概念融入模型决策层以构建原型匹配分类层，并与自编码器结合形成完整的原型匹配网络，旨在阐明模型决策的显式逻辑与模型视角下的典型故障样本特征。本章首先概述原型匹配逻辑与自编码器的理论基础，继而详细阐述原型匹配网络的结构设计、损失函数优化与解释策略，最后通过传统故障诊断任务与领域泛化任务全面评估网络的诊断性能与可解释性。

第四章，结合域变换的智能诊断模型被动解释形式优化。针对智能诊断被动解释形式不直观的问题，本章将信号处理中的循环谱相关与 SHAP 分析方法相结合，在不破坏模型端到端特性的前提下，将归因结果拓展至故障区分能力更强、物理意义更为直观的循环域。本章首先介绍 SHAP 归因与循环谱分析的理论基础，随后系统性阐述确定性信号的循环域变换及其与 SHAP 分析的融合方法，最后通过三类具有特定特性的数据集验证所提方法在解释形式方面的优势，并深入探讨模型差异与噪声因素对解释效果的影响。

第五章，针对振动信号高耗时计算的诊断模型被动解释效率优化。考虑到 SHAP 方法在处理高维振动信号时面临的计算效率瓶颈，本章以解释效率优化为核心目标，提出能够有效缩减数据维度的组合块归因策略，以及降低计算复杂度的 SHEP（SHapley Estimated exPlanation）解释方法。本章首先对 SHAP 的计算复杂度进行定量分析，进而详细阐述组合块归因策略与 SHEP 解释方法的设计原则与实现机制，最后通过仿真与实测实验全面验证所提方法在解释效率方面的优势与解释效果方面的可行性。

第六章，面向旋转机械智能诊断模型的综合解释框架及应用验证。本章将前文所提主动解释与被动解释方法进行系统性整合，构建面向旋转机械智能诊断的综合解释框架，旨在提升智能诊断可解释性研究的系统化水平与实践应用价值。具体而言，首先通过融合输入层主动解释与决策层主动解释技术构建联合解释网络，随后依据模型设计参与程度进行分类讨论，明确阐述综合解释框架在不同解释需求场景下的具体应用流程，最后基于实测高速重载行星齿轮传动系统，对主动解释网络的诊断性能及综合解释框架的完整解释能力展开全面验证与评估。

第七章，总结与展望。总结本文的主要研究内容和成果，提炼工作创新点，并针智能诊断可解释性研究的未来方向进行展望。

第二章　融入时频变换的智能诊断模型输入层主动解释

2.1　引言

随着科学研究的深入，众多学者尝试将传统信号处理和神经网络相结合[87,89,99]。这类方法将信号处理的先验知识融入神经网络，并将信号处理方法中的关键参数转化为网络的可训练变量，从而解决了传统诊断方法参数优化难的问题，在旋转机械故障诊断任务中通常能够取得更为优异的诊断表现。但当前研究往往侧重于网络的降噪性能、诊断精度和泛化能力，而忽略了其在可解释性方面的潜力。信号处理方法具有有明确的物理意义，这类网络通过梯度传播对融入其中的信号处理方法的重要变量进行优化。这些优化后的变量一方面提高神经网络的故障诊断能力，另一方面也能够借助信号处理理论进行分析，获取优化变量背后对应的物理意义，从而实现对网络的解释。

时频变换作为一种常见的信号处理方法，能够将信号从时域转换到时频域，被广泛应用于旋转机械故障诊断中的特征提取。一方面，时频变换拥有坚实的理论基础，其计算过程和参数的物理含义清晰明确，具有良好的物理可解释性；另一方面，时频变换与神经网络的卷积层均基于内积运算，在数值计算上具有等价性。因此，借助时频变换和卷积过程的等价性，本章提出了一种融入时频变换的旋转机械故障诊断输入层主动解释方法。它将具有物理意义的时频变换方法融入到神经网络的传统卷积层中，获得能够揭示网络关注频带的时频卷积层。与传统卷积层不同，时频卷积层的权重由时频变换的核函数控制，并参与网络训练过程，从而自适应地提取更有效的时频特征。将时频卷积层作为可解释预处理层和现有神经网络相结合的时频卷积网络，不仅能够提高模型在诊断精度、收敛速度和少样本学习方面的性能，更能够通过频响分析，从频域视角主动解释神经网络故障诊断的决策依据。

本章首先分析时频变换和卷积层的等价性并构建时频变换核函数，然后介绍融入时频变换的旋转机械故障诊断输入层主动解释方法，包括时频卷积层的结构设计和可解释性分析、以及时频卷积网络的构建和应用流程。最后，通过一个经典的开源数据集和两个实测的旋转机械数据集验证所提时频卷积网络在诊断精度和可解释性等方面的优越性。为了更好地传播该工作，本章的方法代码已开源在 *https://github.com/ChenQian0618/TFN*。

2.2　时频变换和卷积层的等价性分析及时频变换核函数设计

2.2.1　基于内积运算的信号处理时频变换方法

时频变换（Time-Frequency Transform, TFT）是一种基于内积的信号处理方法，广泛应用于机械故障诊断中。从数学角度而言，内积的本质是衡量两个对象的相似程度，给定一组正交基，任何向量都可以通过内积运算被分解为这组正交基的组合。同理，一维振动信号也视为一组向量，进而可将信号进行分解为多个基的组合。Fourier变换作为最基础的信号处理方法，以不同频率的正弦函数作为正交基，进而通过内积运算获得振动信号在不同频率下的幅值，即频谱：

$$X(f)=\int_{-\infty}^{+\infty}x(t)\mathrm{e}^{-\mathrm{i}2\pi ft}\,\mathrm{d}t=\left\langle x(t),\,\mathrm{e}^{\mathrm{i}2\pi ft}\right\rangle, \tag{2-1}$$

式中 $X(f)$ 代表信号频谱，$x(t)$ 代表输入信号，$\mathrm{e}^{\mathrm{i}2\pi ft}$ 代表以 f 为频率的正弦函数，$\langle *,\,*\rangle$ 代表内积运算.

通过与正弦函数作内积获得的频谱能够有效揭示振动信号的频域信息，但却忽略了时域信息，仅适用于平稳信号的分析。为了处理非平稳信号，还需获取信号在不同时间和不同频率下的能量，即时频谱。为此，时频变换方法中设计了与频率和时间均相关的核函数，再将输入信号和该函数作内积从而获得时频谱，其数学表达式为

$$TF(\tau,f)=\left|\int_{-\infty}^{+\infty}x(t){\psi_f}^*(t-\tau)\,\mathrm{d}t\right|=\left|\langle x(t),\psi_f(t-\tau)\rangle\right|, \tag{2-2}$$

式中 $TF(\tau,f)$ 代表时频谱，$\psi_f(t-\tau)$ 代表与频率 f 和时间 τ 均相关的核函数，而 $\psi_f(t)$ 代表在时域和频域都具有紧支撑的内积窗函数。

基于内积的时频变换过程如图 2-1 所示。输入信号 $x(t)$ 与具有不同频带的内积窗函数 $\psi_f(t)$ 进行卷积，以获得信号在特定时间点 τ 处的频谱。逐渐移动时间点 τ，可以获得完整的时频分布 $TF(\tau,f)$。在此过程中，参数 τ 和 f 分别用于调整内积窗函数在时域和频域中的聚焦区域。需要注意的是，内积窗函数 $\psi_f(t)$ 是一个复函数，因此获得的时频分布也是复数形式，即频谱既包含能量信息，也包含相位信息。为了分离出能量信息，还需对时频分布进行复数求模运算。

时频变换方法包括 STFT、线性调频变换（Chirplet Transform，CT）[147,148] 以及小波变换（Wavelet Transform，WT）[149] 等。这些方法均遵循上述计算过程，其区别在于内积窗函数的设计。STFT 通过对正弦函数加窗来构造内积窗函数。以高斯窗为例，高斯窗可表示为

$$w(t)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}, \tag{2-3}$$

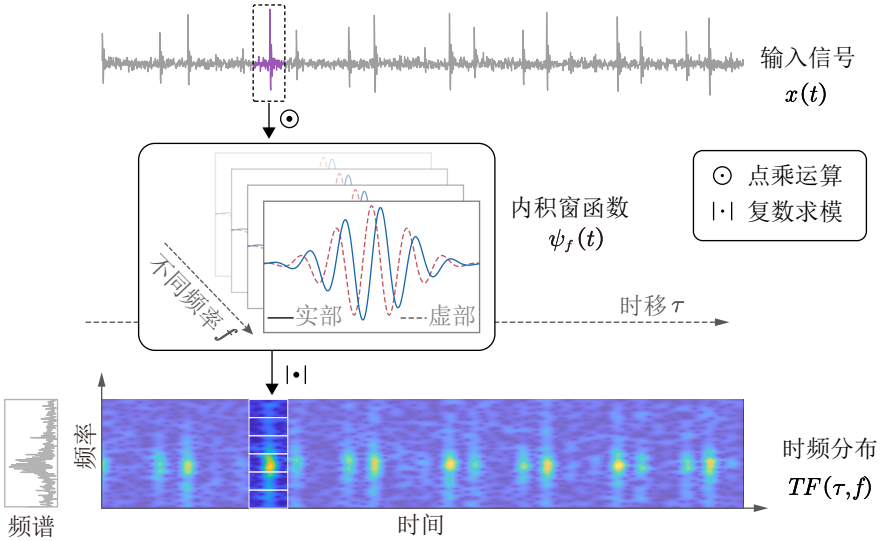


图 2-1　时频变换计算过程

Fig. 2-1　The calculation process of time-frequency transform

式中 σ 代表窗函数的标准差，用于调整窗函数的时域尺度。进而，STFT 的内积窗函数可代表为高斯窗和正弦函数的乘积：

$$\psi_f(t) = w(t) \cdot \mathrm{e}^{-\mathrm{i}2\pi ft}. \tag{2-4}$$

为了更好地应对变转速工况，线性调频变换在 STFT 的基础上引入了线性调频因子，从而提高了变转速工况下所得时频谱的准确性和稳定性。线性调频变换的内积窗函数可表示为

$$\psi_{f,\alpha}(t) = w(t) \cdot \mathrm{e}^{-\mathrm{i}2\pi\left[\frac{\alpha}{2}t^2+ft\right]}, \tag{2-5}$$

式中 α 代表线性调频因子。小波变换则使用小波族作为内积窗函数，小波族通过母小波的缩放和平移生成，从而获得随频率变化的时频窗。小波变换的内积窗函数可表示为

$$\psi_s(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t}{s}\right), \quad \Psi(t) = \mathrm{e}^{-\beta\frac{t^2}{2}}\mathrm{e}^{\mathrm{i}2\pi f_0 t}, \tag{2-6}$$

式中 s 代表尺度因子，$\Psi(t)$ 代表母小波，β 和 f_0 分别代表母小波的衰减系数和基准频率。小波变换在低频段具有较低的时域分辨率和较高的频域分辨率，而在高频段则相反。上述三种时频变换方法的内积窗函数如表 2-1 所示。

时频变换方法能够将非平稳信号从时域投影到时频域，从而获取其时频联合信息。如图 2-1 所示，输入信号为滚动轴承内圈故障的模拟信号，而时频分布能够有效揭示其轴承内圈的故障频率和双冲击特性。时频变换方法具有强大的分析能力，在旋转机械故障诊断的特征提取中发挥着重要作用。

2.2.2　基于内积运算的神经网络卷积层

卷积层是 CNN 的核心组件，通过卷积运算来提取输入信号的特征。由于输入样本通常为一维振动信号，因此旋转机械故障诊断任务常常使用一维卷积层。卷积层的核心操作是卷积运算，一维卷积层的卷积过程如图 2-2 所示。每个随机初始化的卷积核沿一维输入信号进行卷积操作，多个卷积核的结果拼接后形成特征图。第 l 个卷积层中第 k 个通道的输出可代表为

$$\boldsymbol{h}_k^l = \boldsymbol{w}_k^l * \boldsymbol{x}^l + \boldsymbol{b}_k^l, \tag{2-7}$$

式中 $\boldsymbol{x}^l$ 代表着第 l 个卷积层的输入，$\boldsymbol{w}_k^l$ 和 $\boldsymbol{b}_k^l$ 分别代表着该卷积层第 k 个通道的权重和偏置，符号 $*$ 代表着卷积运算。

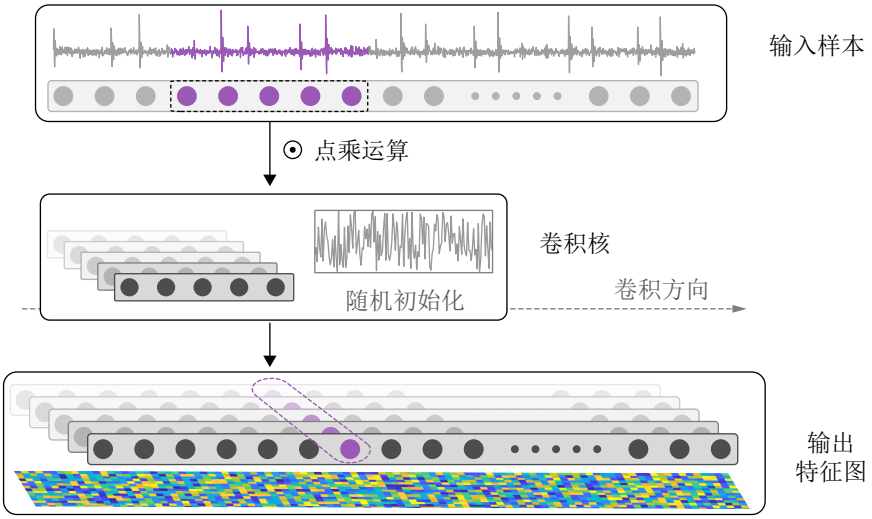


图 2-2　传统卷积层的计算过程

Fig. 2-2　The illustration of calculation process of convolutional layer

将图 2-1 所示的时频变换和图 2-2 所示的传统卷积层进行对比，可以发现两者本质上均是输入信号与特定对象的内积过程，时频变换采用精心设计的内积窗函数而卷积层采用可学习的随机卷积核，这也为本章将时频变换融入到卷积层中提供了理论基础。两者的具体区别在于：(1) 时频变换的内积窗函数的参数由专家根据先验知识设定，具有物理意义；而传统卷积层的卷积核是随机初始化的，并无特定含义，需要通过网络训练来优化。(2) 时频变换的内积窗函数采用复数形式，能够有效考虑相位影响，而卷积层的卷积核采用实数形式。

2.2.3　时频变换核函数的设计

基于时频变换和卷积层在内积运算的相似性，可以从时频变换的内积窗函数出发，提取出在时域、频域同样具有紧支撑性质的时频变换核函数，为后续对卷积核的显式约束提供基础。

时频变换核函数虽然来源于时频变换的内积窗函数，但考虑到两者的差异，还需进行必要的离散化和针对性修改。此外，时频变换的参数在具有物理意义的同时，也受到对应的物理约束，而由此设计的时频变换核函数也应满足这些约束，即核函数的可学习参数 $\boldsymbol{\theta}$ 有一定的限制。以 STFT 为例，由于 Nyquist 采样定理，具有物理意义的归一化频率是 [0, 0.5]，而超出该范围的频率则会受到频率混叠的影响。因此，STFT 卷积核函数的中心频率参数 f 也相应地需要限制在 [0, 0.5] 之间。

根据上述讨论，以 STFT、Chirplet 变换和 Morlet 小波变换为参考，基于这三种典型时频变换的内积窗函数来设计时频变换核函数。三种时频变换的内积窗函数、对应时频变换核函数及其学习参数和约束如表 2-1 所示，包含短时正弦函数（Short-Time Trigonometric Function，STTF）、Chirplet 函数和 Morlet 小波三类核函数。与 STFT 和 Chirplet 核函数不同，Morlet 小波可以通过尺度因子 s 在时域上伸缩。由此，Morlet 小波核函数的核长度设计得比 STTF 和 Chirplet 核函数更长，以避免发生时域截断。

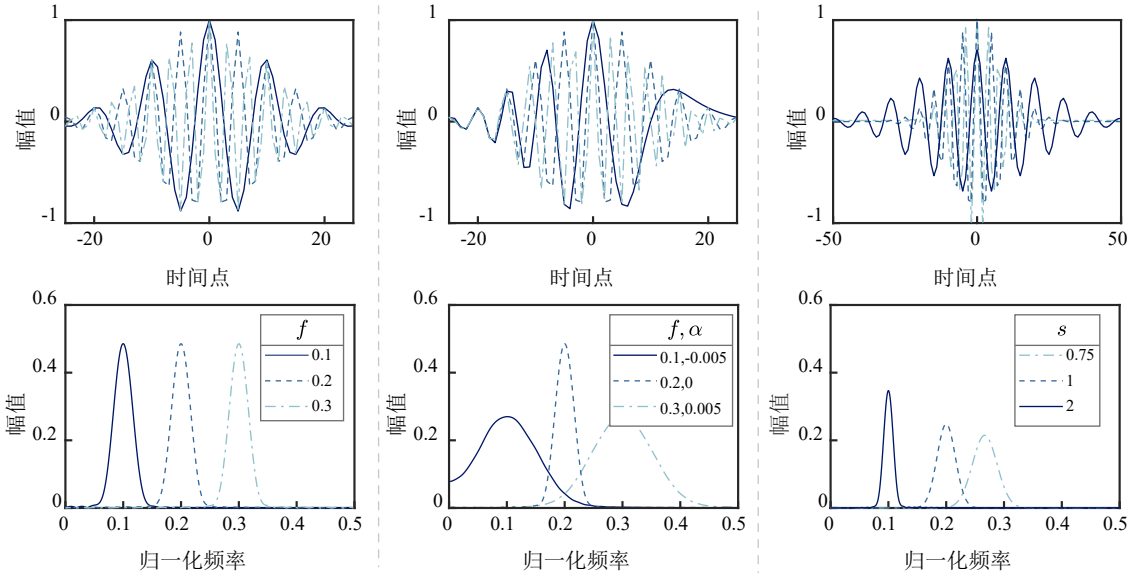
表 2-1　时频变换的内积窗函数、时频变换核函数及其训练参数和约束

Table 2-1　The inner product functions of TFT, the corresponding time-frequency kernel functions and trainable parameter with their constraints

核函数	时频变换的内积窗函数	时频变换核函数	训练参数及约束
STTF	$\psi_f(t)=w(t)\cdot \mathrm{e}^{-\mathrm{i}2\pi ft},$ $w(t)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}$	$\psi_f[n]=\mathrm{e}^{-\frac{1}{2}\left(\frac{n}{\sigma N_c}\right)^2}\mathrm{e}^{\mathrm{i}2\pi fn},$ $\sigma=0.52,$ $n=[-(N_c-1),\ldots,(N_c-1)]$	$f_0\in[0,0.5]$
Chirplet	$\psi_{f,\alpha}(t)=w(t)\cdot \mathrm{e}^{-\mathrm{i}2\pi\left[\frac{\alpha}{2}t^2+ft\right]},$ $w(t)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}$	$\psi_{f,\alpha}[n]=\mathrm{e}^{-\frac{1}{2}\left(\frac{n}{\sigma N_c}\right)^2}\mathrm{e}^{-\mathrm{i}2\pi\left[\frac{\alpha}{2}n^2+fn\right]},$ $\sigma=0.52,$ $n=[-(N_c-1),\ldots,(N_c-1)]$	$f\in[0,0.5],$ $\lvert\alpha\rvert<0.1/N_c$
Morlet 小波	$\psi_s(t)=\frac{1}{\sqrt{s}}\Psi\left(\frac{t}{s}\right),$ $\Psi(t)=\mathrm{e}^{-\beta\frac{t^2}{2}}\mathrm{e}^{\mathrm{i}2\pi f_0t}$	$\psi_s[n]=\frac{1}{\sqrt{s}}\Psi\left(\frac{n}{s}\right),$ $\Psi(n)=\mathrm{e}^{-\frac{1}{2}\left(\frac{n}{\sigma N_c}\right)^2}\mathrm{e}^{\mathrm{i}2\pi f_0n},$ $\sigma=0.6, f_0=0.2$ $n=[-10(N_c-1),\ldots,10(N_c-1)]$	$s\in[0.4,10]$

为加深对所提时频变换核函数的理解，现将不同参数下 STTF、Chirplet 和 Morlet 小波三种核函数的时域和频域表征展示在图 2-3 中。由信号处理的基础知识可知，卷

积过程可视为对输入信号施加有限冲击响应（Finite Impulse Response，FIR）滤波器，而卷积核的频谱则能够反映该滤波器的幅频响应[150]。从频谱上看，三类核函数均可视为带通滤波器。其中 STTF 核函数具有固定的滤波带宽，滤波中心频率由频率参数 f 决定。Chirplet 核函数在 STTF 核函数的基础上引入了线性调频因子 α，在通过频率参数 f 控制滤波中心频率的同时，也能通过线性调频因子 α 动态地调整滤波带宽。Morlet 小波核函数则通过尺度因子 s 对母小波进行缩放，进而以调整其频率特性。随着尺度因子的增加，Morlet 小波核函数的滤波中心频率会随之升高，且滤波带宽也会同步增加。



(a) STTF核函数　　(b) Chirplet核函数　　(c) Morlet核函数

图 2-3　三种时频卷积层核函数的时域和频域表征

Fig. 2-3　The time-domain and frequency-domain diagrams of three kernel functions of TFconv layer

信号处理时频变换和神经网络卷积层虽然在应用上存在迥然差异，但在内积运算上仍存在等价性，其差异在于时频变换使用精心设计的复数内积窗函数，而卷积层使用随机初始化的可学习实数卷积核。为了将时频变换融入到卷积层中，一方面需要对卷积核进行约束，使其具备时域、频域的紧支撑特性，从而获得刻画时频信息的能力；另一方面，需要构建复数形式的卷积核，从而实现分离能量信息和相位信息的能力。本节聚焦卷积核约束方面，通过参考时频变换内积窗函数，构建 STTF、Chirplet 和 Morlet 三类时频变换核函数，为后续时频卷积层的设计提供理论基础。

2.3　基于时频卷积网络的输入层主动解释

2.3.1　时频卷积层的结构设计

时频变换具有物理可解释性，但需要人为设定关键参数，无法根据旋转机械的特点自适应地提取时频信息。相反地，卷积神经网络能够从原始样本中自动提取高维特征并高效地进行准确分类，但其决策依据不够清晰，缺乏可解释性。为了结合这两种方法的优势，本章基于内积运算这一共同点，将时频变换融入到传统卷积层中，从而建立能够提取时频信息的时频卷积层（Time-Frequency Convolutional Layer, TFconv layer）。

时频变换和传统卷积层的关键差别在于时频变换使用复数形式的内积窗函数，而同类信号处理融入的网络均使用实数卷积核[87,89,151]。为了更好地模拟具有物理可解释性的时频变换，所提出的时频卷积层将通用的实数卷积核替换为复数卷积核。但考虑到大多数 CNN 模型使用实数变量，并实现与现有模型的良好兼容，时频卷积层仍旧采用实数变量来对复数卷积核进行等效，其结构设计如图 2-4 所示。

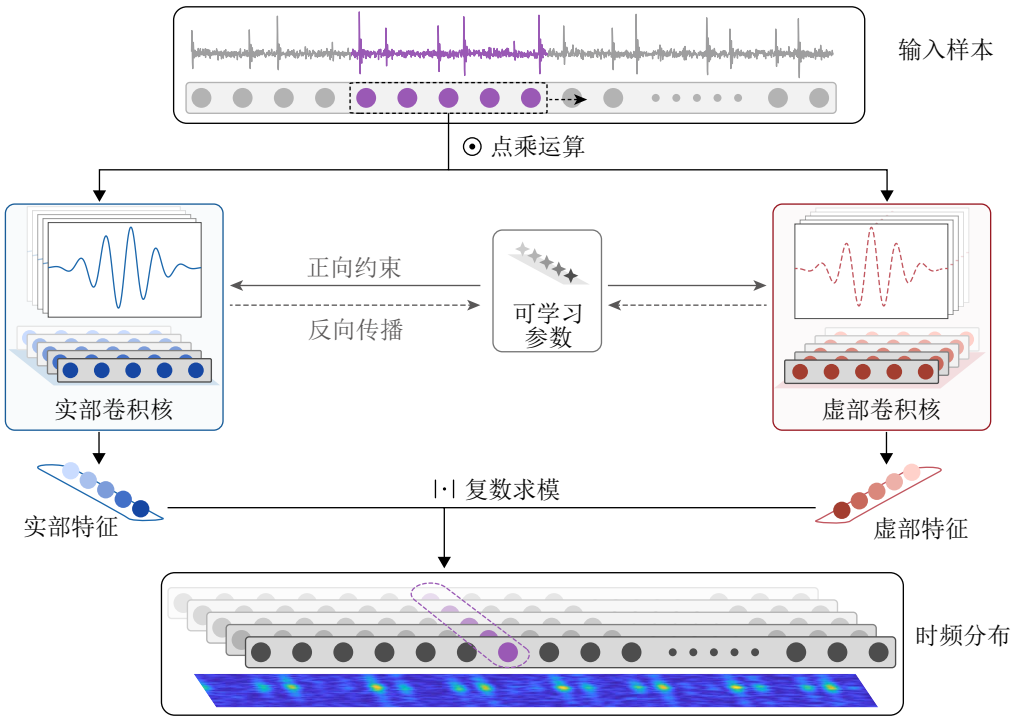


图 2-4　时频卷积层的计算过程
Fig. 2-4　The calculation process of TFconv layer

具体而言，时频卷积层由实部卷积核和虚部卷积核组成，它们分别沿输入振动信号的长度方向进行卷积，并独立地获取输入信号的实部特征和虚部特征。随后，对实

部和虚部特征进行复数求模运算，在舍弃信号相位信息的同时，获得信号的时频能量分布来作为时频卷积层的输出。和传统的卷积层一致，上述过程中每个通道均独立处理。但不同的是，传统卷积层的卷积核是随机初始化的，而时频卷积层的实部和虚部卷积核的权重由核函数控制。卷积核权重和核函数的关系可表示为

$$\begin{cases} \boldsymbol{\psi}_\theta = \psi(\theta) \in \mathbb{C}^N \\ \boldsymbol{w}_{\theta,\mathrm{Re}} = \mathrm{Re}(\boldsymbol{\psi}_\theta) \in \mathbb{R}^N \\ \boldsymbol{w}_{\theta,\mathrm{Im}} = \mathrm{Im}(\boldsymbol{\psi}_\theta) \in \mathbb{R}^N \end{cases} \tag{2-8}$$

式中 $\boldsymbol{\psi}_\theta$ 代表核函数，$\boldsymbol{w}_{\theta,\mathrm{Re}}$ 和 $\boldsymbol{w}_{\theta,\mathrm{Im}}$ 分别代表实部核和虚部核的权重，N 代表卷积核的长度，$\mathrm{Re}(\cdot)$ 和 $\mathrm{Im}(\cdot)$ 分别代表获取复数值的实部和虚部的运算符。控制参数 θ 用于调整核函数的频率特性，并在反向传播过程中作为可学习参数进行更新，从而实现时频卷积层对时频信息的自适应提取。时频卷积层的核函数等效于时频变换中的内积窗函数，将时频卷积层的输入记为 x，其输出可表示为

$$\begin{cases} \boldsymbol{h}_{k,\mathrm{Re}} = \boldsymbol{w}^k_{\boldsymbol{\theta},\mathrm{Re}} * \boldsymbol{x} \\ \boldsymbol{h}_{k,\mathrm{Im}} = \boldsymbol{w}^k_{\boldsymbol{\theta},\mathrm{Im}} * \boldsymbol{x} \\ \boldsymbol{h}_k = \sqrt{\boldsymbol{h}^2_{k,\mathrm{Re}} + \boldsymbol{h}^2_{k,\mathrm{Im}}} \end{cases} \tag{2-9}$$

式中 k 代表卷积核的第 k 个通道，$\boldsymbol{\theta}$ 代表核函数的可训练控制参数，$\boldsymbol{h}_{\mathrm{Re}}$ 和 $\boldsymbol{h}_{\mathrm{Im}}$ 分别代表实部和虚部特征图，h 代表最终输出（即时频能量分布）。

总结而言，相比于传统卷积层，所提出的时频卷积层具有如下三个特点：

(1) 实虚部机制：时频卷积层包含实部卷积核和虚部卷积核的两个卷积过程来模拟复数卷积核，并通过复数求模运算将两部分特征进行合并。

(2) 核函数约束：时频卷积层的卷积核权重由精心设计的核函数决定，而不是传统卷积层采用的随机初始化。

(3) 可学习参数：时频卷积层的可学习参数是核函数的控制参数 $\boldsymbol{\theta}$（例如 STTF 核中的频率因子 f），而不是传统的卷积核权重。

由于时频卷积层的可学习参数与传统卷积层不同，其反向传播过程也有所不同。具体而言，时频卷积层计算核函数参数 $\boldsymbol{\theta}$ 的梯度，并在每次训练步骤中对其更新，其计算过程可表示为

$$\begin{cases} \boldsymbol{\delta}_{\boldsymbol{\theta}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{h}} \left(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{w}_{\boldsymbol{\theta},\mathrm{Re}}} \frac{\partial \boldsymbol{w}_{\boldsymbol{\theta},\mathrm{Re}}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{w}_{\boldsymbol{\theta},\mathrm{Im}}} \frac{\partial \boldsymbol{w}_{\boldsymbol{\theta},\mathrm{Im}}}{\partial \boldsymbol{\theta}} \right) \\ \boldsymbol{\theta} \leftarrow \mathrm{Optimizer}\left(\boldsymbol{\theta}, \delta_{\boldsymbol{\theta}}, \eta\right) \end{cases} \tag{2-10}$$

式中 $\boldsymbol{\theta}$ 代表时频变换核函数的可学习参数，$\boldsymbol{\delta}_{\boldsymbol{\theta}}$ 代表 $\boldsymbol{\theta}$ 的梯度，∂ 代表偏导数运算符，$\mathcal{L}$ 代表分类损失，$\boldsymbol{h}$ 代表时频卷积层的输出，Optimizer 和 η 分别代表神经网络优化

器及设置参数。通过式 (2-10) 所示的链式法则获得梯度后，可学习参数 $\boldsymbol{\theta}$ 可以通过随机梯度下降（Stochastic Gradient Descent，SGD）等优化算法进行更新。

2.3.2　时频卷积层的可解释性分析

通过融入时频变换，时频卷积层不仅能够从旋转机械振动信号中自适应提取时频信息，还可以将优化后的核函数参数和背后的物理含义相联系，从而实现对黑箱神经网络的解释。具体而言，将时频卷积层视为一系列的 FIR 滤波器组，便可以对训练后的时频卷积层进行幅频响应分析以获得其幅频响应（Amplitude-Frequency Response，FR），进而可以揭示 CNN 对不同频率的关注程度。时频卷积层作为时频卷积网络的预处理层，只有幅频响应中具有足够幅值的频率，才能够通过时频卷积层并参与后续预测过程，这些频率便是黑箱网络的决策依据。

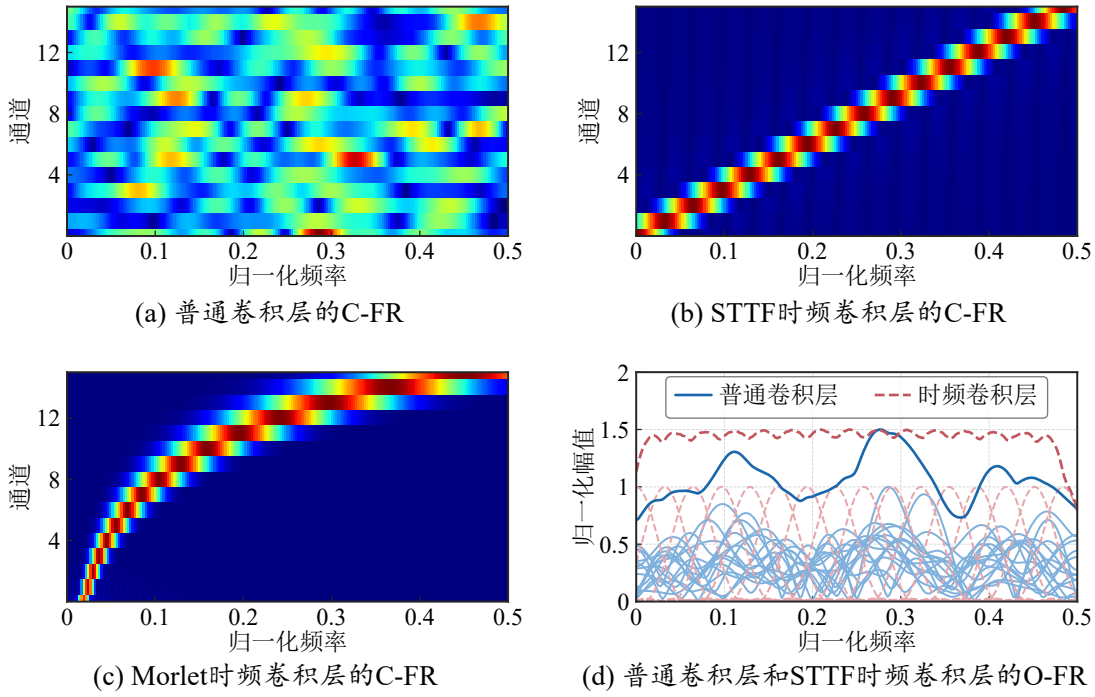
卷积层的频率响应分析方法源于信号处理领域的滤波器理论。卷积层中的卷积过程与 FIR 滤波器的滤波过程完全等价[152]，卷积层的卷积核即是 FIR 滤波器，而卷积层的输入则是待滤波的机械振动信号。在幅频响应分析中，对卷积核进行快速 Fourier 变换，便能获得卷积层的幅频响应[150]。考虑到卷积层包含多个通道，需要首先计算通道层级幅频响应（Channel-wise Amplitude-Frequency Response，C-FR），然后对其进行平均以获得整体幅频响应（Overall Amplitude-Frequency Response，O-FR）。C-FR 和 O-FR 的计算过程可以表示为

$$\begin{aligned} \boldsymbol{H}_i(f) &= |\mathrm{FFT}(\boldsymbol{w}_i)| = \left| \sum_{n=0}^{N} \boldsymbol{w}_i[n] \cdot \mathrm{e}^{-\mathrm{i}2\pi f n/N} \right|, \\ \boldsymbol{H}(f) &= \frac{1}{n_c} \sum_{i}^{n_c} \boldsymbol{H}_i(f), \end{aligned} \tag{2-11}$$

式中 $\boldsymbol{w}_i$ 代表长度为 N 的第 i 个通道卷积核，$\boldsymbol{H}_i(f)$ 和 $\boldsymbol{H}(f)$ 分别代表第 i 个通道的 C-FR 和 O-FR，n_c 代表卷积层的通道数目。由此，上述频率响应分析尽管是基于传统的实数卷积层，但也同样适用于包含实部、虚部两部分的时频卷积层，仅需将两部分卷积核视为一个复数卷积核，便能同样获得卷积层的幅频响应。此外，时频卷积层的两部分卷积核分别对应卷积核函数的实部和虚部，这使得两部分具有相位相差 90 度但幅值相同的幅频响应。

尽管传统卷积层和时频卷积层都可以通过频率响应分析进行研究，但它们在幅频响应上存在显著差异。如图 2-5(a) 所示的传统卷积层，其 C-FR 呈现出随机分布，难以通过 FIR 滤波器的幅频响应，很好地确认其关注的频率区域。相反地，如图 2-5(b) 和 (c) 所示的时频卷积层，由于引入卷积核函数的约束等价于多个带通滤波器，而滤

波器的带通频率则能能明确地解释神经网络的频率偏好。最后，传统卷积层和 STTF 时频卷积层的 O-FR 如图 2-5(d) 所示。总结而言，相比于传统卷积层，时频卷积层具有带通滤波器的特性，其关注频率区域更容易识别，为解释 CNN 模型的频率依据奠定基础。



(a) 普通卷积层的C-FR　(b) STTF时频卷积层的C-FR

(c) Morlet时频卷积层的C-FR　(d) 普通卷积层和STTF时频卷积层的O-FR

图 2-5　初始状态下传统卷积层和时频卷积层在 C-FR 和 O-FR 上的对比

Fig. 2-5　The comparison between C-FR and O-FR of initialized traditional convolutional layer and that of initialized TFconv layers

此外，有必要进一步阐述时频卷积层的解释逻辑。时频卷积层可以看作是一系列可训练的带通 FIR 滤波器，不同通道提取不同频段的能量和相位信息。后续的复数求模运算，便是则在保留与故障更为相关的能量信息的基础上，舍弃可视为噪声、带有随机性的相位信息，来实现振动信号的更优特征提取。当时频卷积层作为预处理层与基准 CNN 结合形成时频卷积网络时，时频卷积层便成为整个网络的数据入口。这意味着只有在幅频响应（O-FR）中具有足够幅值的频率才能通过时频卷积层，并被后续的基准 CNN 用于故障诊断预测。在网络训练过程，时频卷积层的参数将会根据训练数据的特性进行调整，从而使得 O-FR 中的峰值逐渐向携带大部分信息的数据集信息频带靠拢，以实现更好的诊断性能。因此，时频卷积层训练后的 O-FR，可以用来解释 CNN 模型对不同频率的关注，从而揭示 CNN 模型的决策依据。在后续的实验中，若时频卷积层训练后的 O-FR 能够与数据集频谱的信息频带较好吻合，则能够有

效证实上述解释逻辑的正确性。

2.3.3　时频卷积网络的构建及其故障诊断应用流程

时频卷积层可以作为预处理层与基准 CNN 结合，来提高基准网络的诊断性能。两者结合得到的新型网络称之为时频卷积网络（Time-Frequency Convolutional Network, TFN）。借助融入时频变换先验知识的时频卷积层，TFN 可以从原始振动信号中提取与故障相关的时频信息，从而实现故障状态的高精度诊断，其在智能机械故障诊断中的应用全过程如图 2-6 所示。

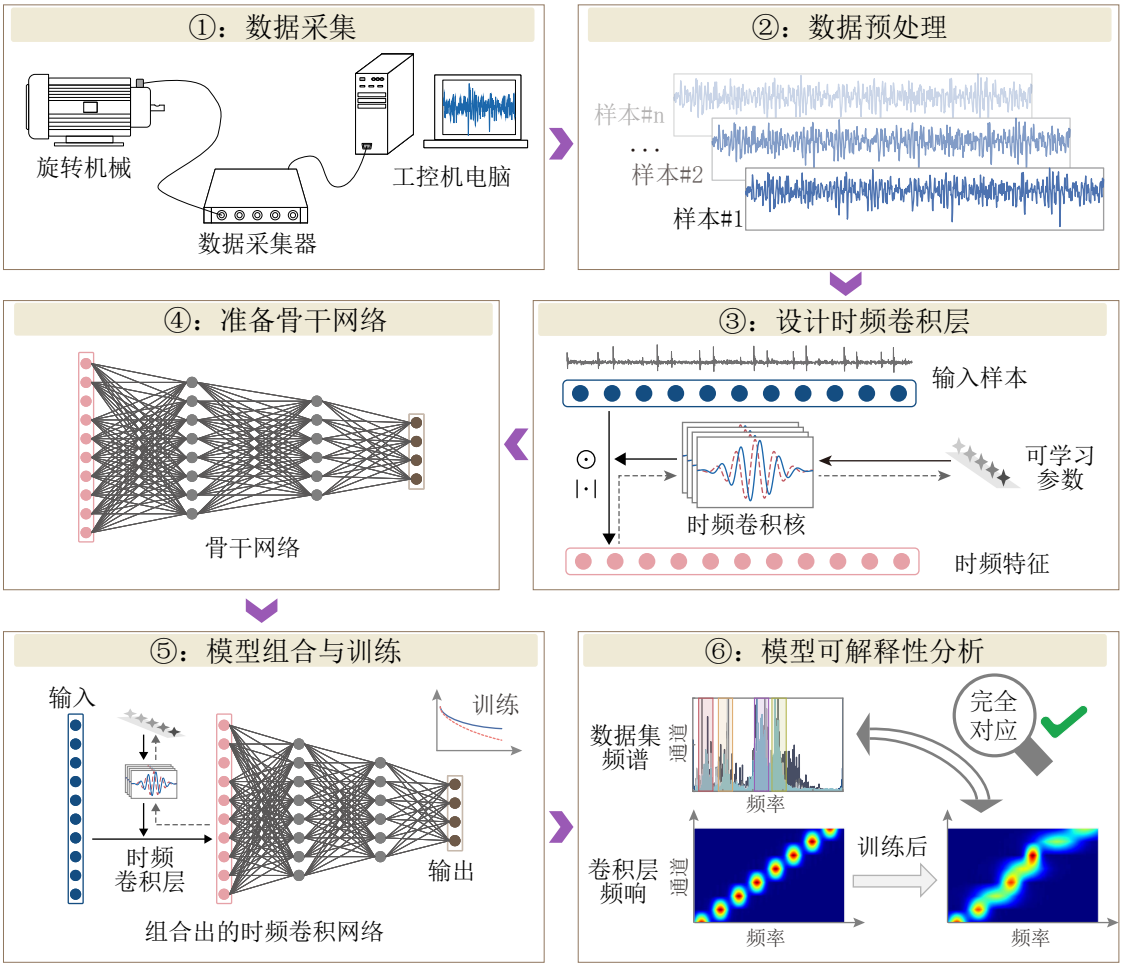


图 2-6　时频卷积网络应用于智能机械故障诊断的全过程

Fig. 2-6　The entire process of applying TFN to intelligent mechanical fault diagnosis

首先，通过安装在旋转机械设备上的加速度传感器，采集恒定工况下的振动信号。其次，以固定长度的滑动窗口对采集的振动信号进行分割，生成一系列振动信号样本作为 TFN 的输入。然后，参考表 2-1 从时频变换方法中构建卷积核函数并融入

到时频卷积层中。之后，可选择任意的现有 CNN 模型作为基准网络，并将时频卷积层作为预处理层与基准网络结合以得到 TFN。最后，通过故障诊断任务中的训练样本和测试样本对获得的 TFN 进行模型训练和性能测试。在完成训练后，可根据式 (2-11) 获取时频卷积层训练后的幅频频响，从而从模型视角解释数据集的信息频带。

2.4　输入层主动解释方法的故障诊断性能和解释效果实验验证

实验部分将使用三个实验数据集来验证 TFN 的诊断性能和可解释性，包括一个公开数据集和两个私有数据集。在诊断性能部分，考虑到预处理层通道数量对诊断精度具有显著影响，本节将对比不同通道数量下，TFN 模型和同类信号处理模型的诊断精度和少样本学习能力。在可解释性部分，本节将对比数据集的频谱与时频卷积层训练后的 O-FR，从而验证通过时频卷积层幅频响应来解释网络的关注频带的可行性。

实验过程均采用一个较为简单的 CNN 作为基准网络，以更直接地验证时频卷积层的有效性，所构建的 TFN 结构如表 2-2 所示。其中，n_c、N 和 K 分别表示预处理层的通道数、时频卷积层的长度和分类的类别数。n_c 和 N 在时频卷积层的设计过程中确定，K 由具体的数据集确定。

表 2-2　实验中所采用的时频卷积网络架构
Table 2-2　The architecture of TFN used in the experiment

网络部分	序号	网络层参数	输出尺寸
预处理层	-	模型输入	1*1024
	1	TFconv(n_c@N*1)	n_c*1024
基准 CNN	2	Conv(16@15*1)-BN-ReLU	16*1010
	3	Conv(32@3*1)-BN-ReLU-MaxPool(2)	32*504
	4	Conv(64@3*1)-BN-ReLU	64*502
	5	Conv(128@3*1)-BN-ReLU-AdaptivePool(4)	128*4
	6	Flatten	512
	7	FC(256)-ReLU-FC(64)-ReLU-FC(K)	K

2.4.1　可复现的 CWRU 轴承开源数据集

凯斯西储大学（Case Western Reserve University，CWRU）轴承数据集是机械故障诊断中倍受欢迎、广泛使用的开源数据集之一[153]。这个开源数据集被作为基准对象来测试 TFN 的诊断性能和可解释性，以保证本章方法的实测性和可复现性。CWRU

轴承数据集的试验台如图 2-7 所示。在驱动端和风扇端的轴承外壳处均装有加速度传感器，用以采集不同负载、不同故障位置和不同故障程度下的振动信号。CWRU 数据集包含四种负载：0、1、2 和 3 HP，采样频率为 12 kHz 和 48 kHz。除了健康状态（Health, H），该数据集还包含三种不同的轴承故障类型：内圈故障（Inner race fault, I）、滚动体故障（Rolling ball fault, B）和外圈故障（Outer race fault, O）。对于每种故障类型，分别考虑了不同的故障尺寸，即 0.007、0.014 和 0.021 英寸。因此，该数据集包含九种故障状态和一种正常状态，共十个类别。CWRU 轴承数据集的故障诊断可以视为一个十类别的分类任务。

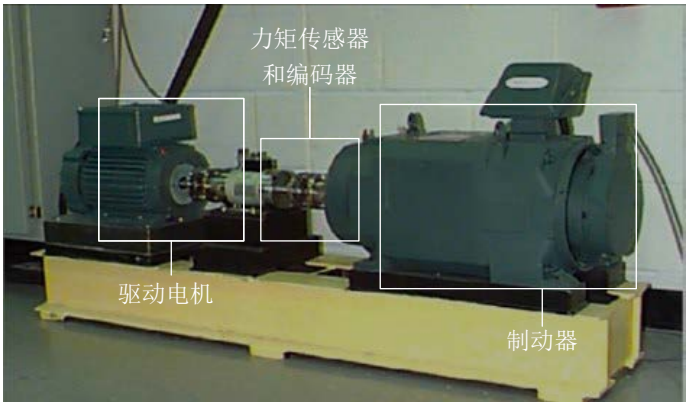


图 2-7　CWRU 轴承故障试验台[154]
Fig. 2-7　CWRU bearing fault test rig[154]

CWRU 轴承数据集中，大多数模型在 12 kHz 振动信号上的诊断准确率均接近 100%，难以区分不同模型的诊断性能。相反地，48 kHz 的振动信号则具有更高的诊断难度，选择 48 kHz 的振动信号进行后续的诊断任务。振动信号被截断为长度为 1024 的样本，每种状态有 450 个样本，总共 4500 个样本。各个类别样本的 60% 被用于训练，其余样本用于测试。分类的损失函数为交叉熵，训练优化器为 Adam (Adaptive Moment Estimation)，动量设置为 0.9，初始学习率为 0.001，学习率的衰减比例设置为每个周期 0.99，训练周期设置为 50。每个模型重复 10 次以消除随机性影响，来保证诊断准确率的可信性。

实验过程共采用如表 2-3 所示的三类模型：第一类是基准模型，包括表 2-2 所示的基准 CNN (CNN-Backbone)，以及将传统随机初始化卷积层作为预处理层与基准 CNN 结合的网络（CNN-Random）。第二类是对比模型，包括 SincNet[87]，使用 Morlet 小波核和 Laplace 小波核的 WKN (WKN-Morlet 和 WKN-Laplace)[89]，W-CNN[151]。第三类是 TFN 模型，包括使用 STTF、Chirplet 和 Morlet 卷积核的 TFN (TFN-STTF、TFN-

Chirplet 和 TFN-Morlet)。除了不包含预处理层的 CNN-Backbone 外，每个模型都考虑了 4 种不同通道数的预处理层：16、32、64 和 128。

表 2-3　时频卷积网络验证实验中使用的网络及含义

Table 2-3　The networks used in the validation experiment of TFN and their meanings

模型类别	具体网络	含义
基准模型	CNN-Backbone	如表 2-2 所示的基准 CNN
	CNN-Random	基准 CNN + 传统卷积层
对比模型	SincNet[87]	基准 CNN + 受特定函数调控的卷积层
	WKN-Morlet[89]	基准 CNN + 受 Morlet 小波调控的卷积层
	WKN-Laplace[89]	基准 CNN + 受 Laplace 小波调控的卷积层
	W-CNN[151]	基准 CNN + 受 Morlet 小波初始化的卷积层
TFN 模型	TFN-STTF	基准 CNN + STTF 时频卷积层
	TFN-Chirplet	基准 CNN + Chirplet 时频卷积层
	TFN-Morlet	基准 CNN + Morlet 时频卷积层

由于预处理层通道数对模型性能具有较大影响，所以将其设置为实验变量。不同预处理层通道数下，各类模型在凯斯西储轴承数据集的故障诊断准确率如图 2-8 所示，由结果可知：

(1) 从模型角度来看，CNN-Backbone 和 CNN-Random 的诊断准确率最低，引入各类定制预处理层的 SincNet、WKN-Morlet、WKN-Laplace 和 W-CNN 相比基准 CNN 在诊断准确率方面有着显著提高。而 TFN-STTF、TFN-Chirplet 和 TFN-Morlet 的诊断表现显著优于其他所有模型，并在 64 或 128 通道数下的诊断准确率接近 100%，充分证明了 TFN 在故障诊断性能方面的卓越优势。其原因在于，TFN 借助时频卷积层，可以将原始振动信号转换为与故障相关的时频特征，进而有效提高了模型的故障诊断表现。

(2) 从通道数目来看，同种模型的通道数越多，其诊断准确率越高，并且这种现象在三种 TFN 模型中更为显著。当通道数为 16 时，TFN 模型和其他模型的诊断准确率相似。随着通道数的增加，TFN 模型的诊断准确率提升得越高，显著优于其他模型。更多通道能够使时频卷积层提取更精细的时频信息，从而获得更高的故障诊断准确率。但通道数增加的效果具有饱和效应，64 通道的 TFN 与 128 通道的 TFN 的诊断准确率相似。这表明当时频卷积层具有足够的通道来提取时频信息时，增加通道数并不会带来相应的诊断准确率提升。

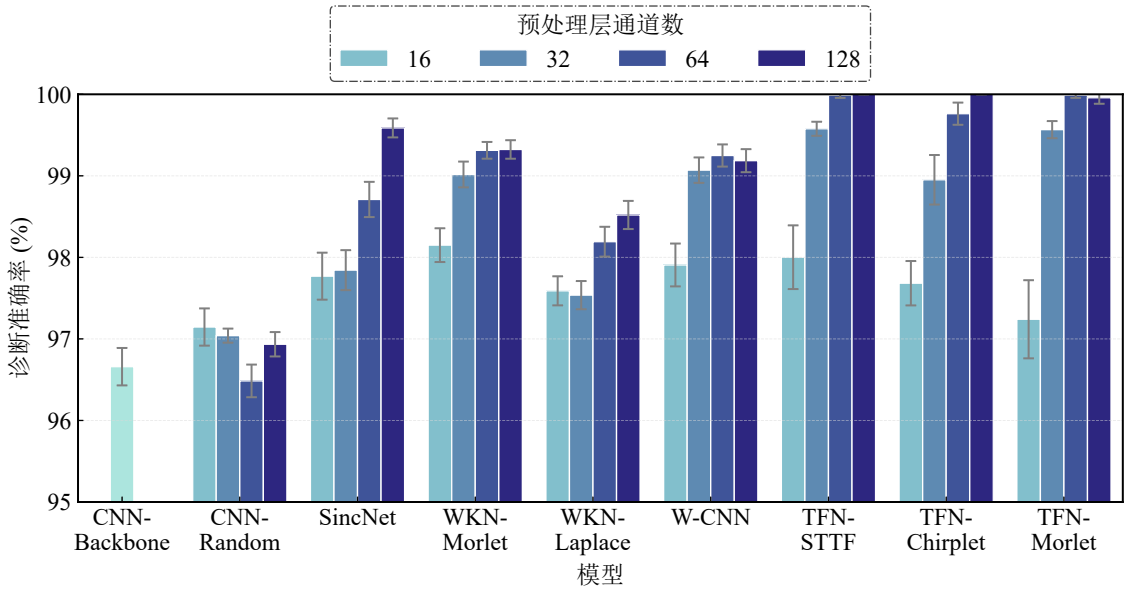


图 2-8　不同预处理层通道数下各类模型在凯斯西储轴承数据集的诊断准确率

Fig. 2-8　Diagnostic accuracies of various models under different numbers of preprocessing channels in the CWRU dataset

(3) 从卷积核函数来看，三种卷积核函数的诊断性能相似，STTF 卷积核总体上比 Chirplet 卷积核和 Morlet 卷积核具有微弱优势。

在测试完 TFN 的诊断性能后，现对 TFN 的可解释性进行分析。其中，CWRU 数据集的负载选择为 3 HP，将 12 kHz 采样频率的个类别振动信号作为输入样本，对具有不同核函数的 TFN 进行训练。为了获得更清晰的可解释性结果，预处理层的通道数设置为 8，其他实验设置和准确率实验保持一致。

根据式 (2-11) 所示的频率响应分析，可以获得不同模型预处理层训练前后的 O-FR。CWRU 数据集的频谱和不同模型预处理层的 O-FR 结果如图 2-9 所示。CWRU 数据集的频率幅值主要存在于图中用不同颜色标注出的四个信息频带，即频带 #1-#2-#3-#4。这些频带携带了数据集的大部分信息，而一个好的 CNN 模型应关注这些频带以获得良好的故障诊断能力。从图 2-9(b)-(i) 所示的 O-FR 结果可知：

(1) 训练后的 CNN-Backbone 和 CNN-Random 的 O-FR 大体上呈现出均匀分布，仅在频带 #2 或频带 #4 处有幅值增加，表明这两类模型不具有倾向性，难以有效捕捉包含故障类别关键特征的信息频带。

(2) 对比模型（SincNet、WKN-Morlet 和 W-CNN）的 O-FR 如图 2-9(d)-(f) 所示。SincNet 和 WKN-Morlet 的 O-FR 在训练后变化不大，其结果与图 2-9(a) 所示的 CWRU 数据集频谱不完全对应。与前两种方法不同，W-CNN 仅通过卷积核

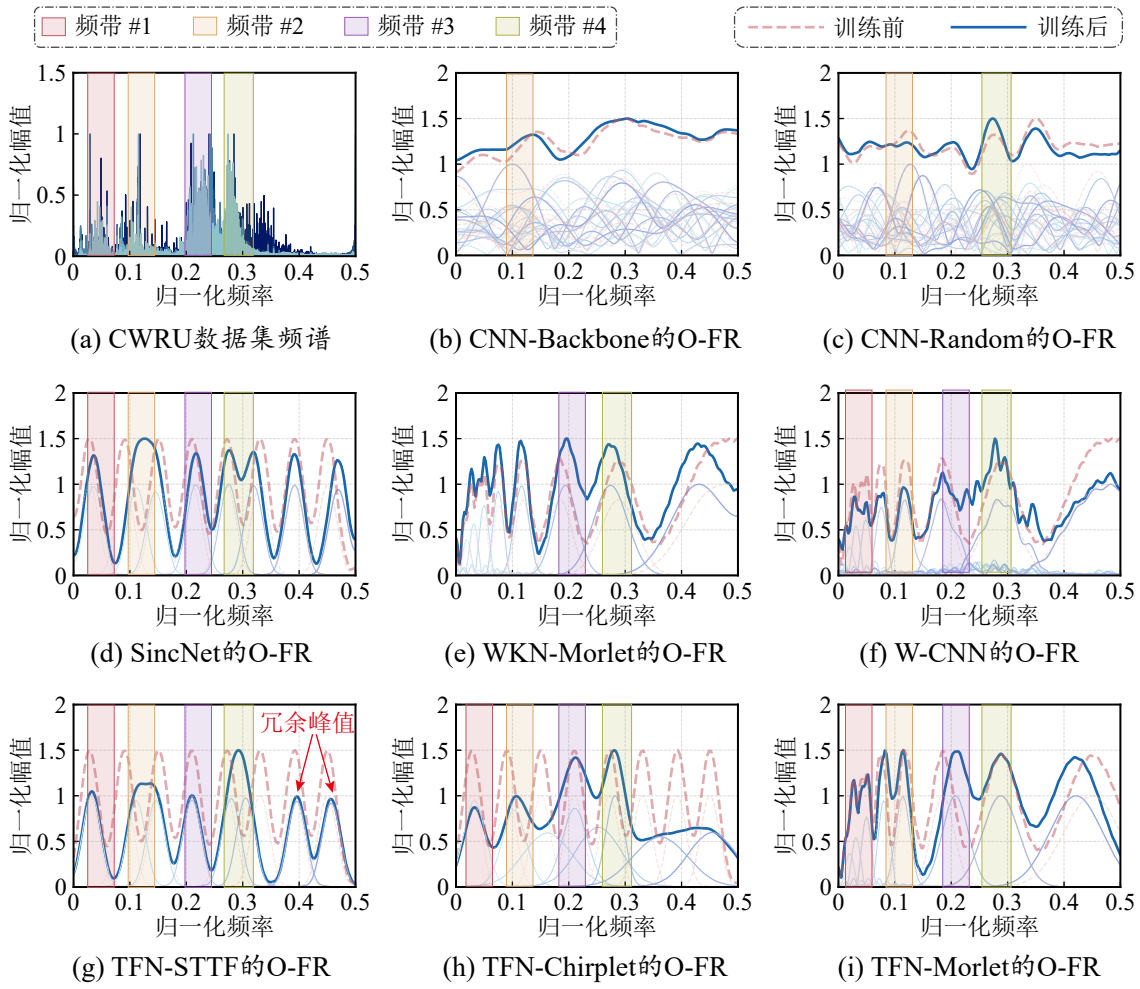


图 2-9　CWRU 数据集的频谱以及不同模型预处理层的综合幅频响应

Fig. 2-9　Frequency spectrum of CWRU dataset and the O-FRs of different model preprocessing layers

函数来初始化卷积权重而非完全控制，因此其 O-FR 在训练过程不受卷积核函数约束。W-CNN 的 O-FR 在训练后变化显著，并且与数据集频谱很好地匹配。但也正是由于 W-CNN 不受核函数约束且仅为实数核，其 O-FR 相对混乱，并且在高频区域存在冗余峰值，因此 W-CNN 并不是解释模型关注故障相关频率的理想方法。

(3) TFN-STTF 训练后的 O-FR 在所有信息频带处都有峰值，这表明 TFN-STTF 在训练过程正确地关注了这些信息频带，这也一定程度上对应着 TFN-STTF 的优异故障诊断能力。然而，由于 STTF 核函数仅能通过参数 f 改变滤波中心频率，不能改变滤波带宽，因此 TFN-STTF 的 O-FR 在信息频带之外仍然存在一些峰值，即图 2-9(g) 中标出的两个冗余峰值。

(4) 训练后的 TFN-Chirplet 同样能够捕捉所有信息频带，其 O-FR 也在这些信息频带内均有有幅频峰值。然而，与 STTF 核函数不同，Chirplet 核函数具有额外的线性调频因子 α 来改变其滤波带宽，当该通道无法获得有用信息时，它会增加其带宽以搜索更宽的频带，因此在 TFN-Chirplet 的 O-FR 中不存在信息频带之外的峰值。TFN-Chirplet 的 O-FR 与 CWRU 数据集的频谱完全一致，TFN-Chirplet 在物理可解释性方面优于其他模型。

(5) 训练后的 TFN-Morlet 的 O-FR 在训练后变化不大，仅能勉强识别一个峰值（对应于频带 #3）。尽管小波变换核函数的自适应频率带宽有助于提取故障相关信息，但它也使幅频响应的频带缺少灵活性，从而阻碍了 O-FR 向信息频带的收敛。这导致 TFN-Morlet 与数据集关键频带的一致性不如 TFN-STTF 和 TFN-Chirplet。

通过分析时频卷积层训练后的 O-FR，有效验证了 TFN 在可解释性方面的优异表现。训练后的 TFN-Chirplet 的 O-FR 与 CWRU 数据集的频谱很好地匹配，同时也证明了 2.3.2 节中关于 TFN 可解释性的假设，即模型在训练过程中更倾向于关注与故障相关的频率，从而能够借助 O-FR 反映了模型对不同频率的关注程度。

现基于 CWRU 数据集上开展少样本实验，以进一步分析 TFN 模型的少样本学习能力。预处理层的通道数均设置为 64，实验设置与 CWRU 数据集上的诊断性能实验相同，每个类别有 450 个样本，总共 4500 个样本。为了测试 TFN 的少样本学习能力，将训练样本的数量作为实验变量，在每个类别中随机选择一定数量（即 5、10、20、50、100、150、200、250 和 300）的样本作为训练数据，其余样本用于测试。考虑到训练样本数量的差异，将训练周期进行相应调整，以确保训练批次数尽可能相等但不超过 300，具体的训练周期数目可表示为

$$N_{\text{epoch}} = \min\left(\frac{50 \times 300}{N_{\text{training\_sample}}}, 300\right), \tag{2-12}$$

式中 N_{epoch} 代表训练周期数目，$N_{\text{training\_sample}}$ 代表每个类别的训练样本数目。

少样本实验的结果如图 2-10 所示。当训练样本数为 5 时，各模型的诊断准确率均低于 80%。具体而言，其他模型的准确率均低于 65%，而 TFN 模型的准确率接近 75%，两类模型诊断准确率的差距超过 10%，体现出 TFN 优异的少样本学习能力。当训练样本数增加到 50 时，TFN 的诊断准确率接近 100%，TFN 与其他模型的诊断准确率差距缩小到约 5%。之后，随着训练样本数量的增加，TFN 的表现保持在约 100%，而其他模型的诊断准确率逐渐提高，最终准确率差距约为 1∼2%。总之，所提出的 TFN 在少样本诊断任务中表现远优于其他模型，这种出色的表现归功于时频卷积层所具有的时频特征提取能力。

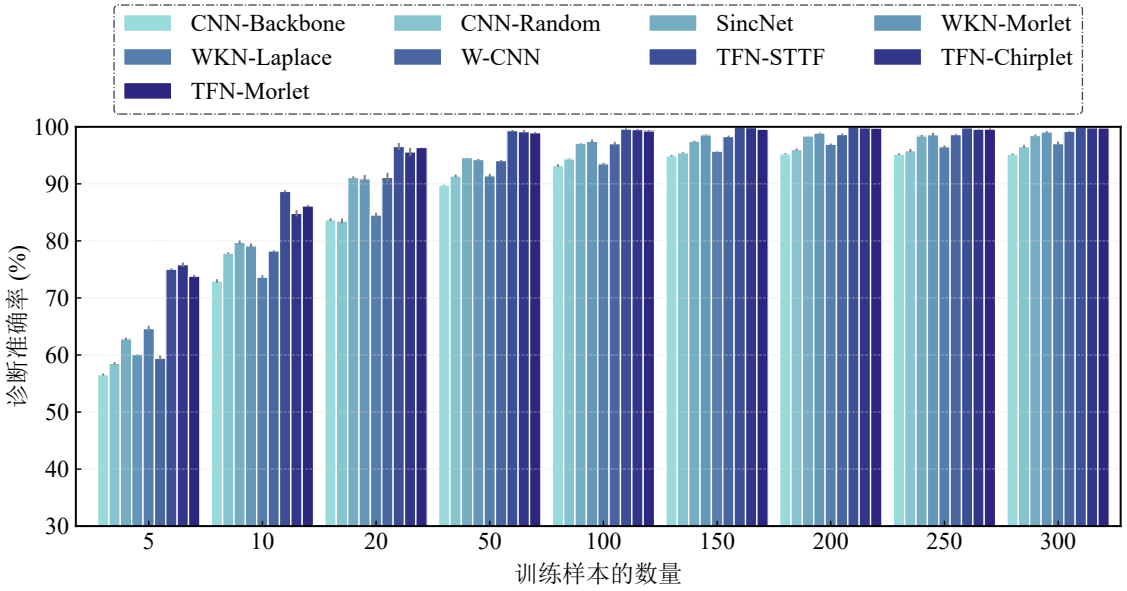


图 2-10　不同训练样本数量下各类模型在 CWRU 轴承数据集的诊断准确率

Fig. 2-10　The diagnostic accuracies of different models under different training sample numbers on the CWRU dataset

表 2-4　行星齿轮数据集的故障工况类别

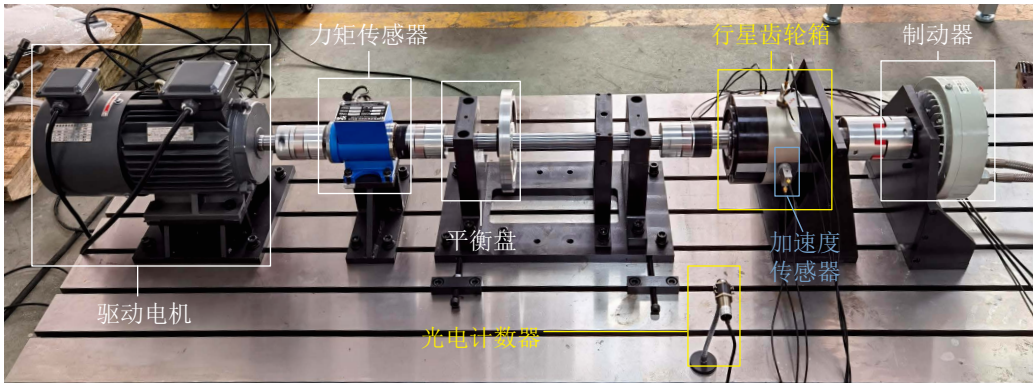
Table 2-4　The working conditions of planetary gearbox dataset

故障类别	标签	故障部件	训练/测试样本数
健康	N	无	264 / 176
单点故障	S	齿圈	264 / 176
双点故障	D	齿圈和太阳轮	264 / 176
三点故障	T	齿圈、太阳轮和行星轮	264 / 176
复合故障	C	齿圈和滚动轴承	264 / 176

2.4.2　实验室场景下的行星齿轮箱数据集

行星齿轮试验台如图 2-11(a) 所示，包括电动机、传动轴、扭矩传感器、行星齿轮箱、磁粉制动器和一系列传感器。振动信号由位于行星齿轮箱外壳上的加速度计采集，并以 10.2 kHz 的采样频率传输到信号采集卡。

该数据集考虑了四种类型的部件故障，包括三个不同齿轮的断齿故障和滚动轴承外圈的点蚀故障，如图 2-11(b)-(e) 所示。基于这些故障，行星齿轮数据集包含如表 2-4 中所示的五种故障工况类别，具体为正常（N）、单点故障（S）、双点故障（D）、三点故障（T）和复合故障（C）。行星齿轮箱的故障诊断可以视为一个五分类任务。



(a) 行星齿轮试验台

(b) 齿圈断齿故障　(c) 行星轮断齿故障　(d) 太阳轮断齿故障　(e) 轴承外圈点蚀故障

图 2-11　行星齿轮数据集的试验台和故障件

Fig. 2-11　Test bench and faulty components of planetary gear dataset

在数据准备过程中，通过滑动窗口对原始振动信号进行无重叠截断，以获得输入样本。每个类别包含 440 个样本，每个样本的长度设置为 1024。之后，随机将 60% 的样本划分为训练集，其余样本作为测试集。此外，为了增加诊断难度，在原始信号中加入信噪比为 0 的高斯白噪声。其余实验设置与 CWRU 数据集上的诊断实验一致。

不同预处理层通道数下各类模型在行星齿轮数据集的诊断准确率如图 2-12 所示。该数据集的诊断难度相对较低，不同模型的诊断准确率差距不如之前 CWRU 诊断实验明显。CNN-Backbone 的诊断准确率为 97.8%，CNN-Random 的表现优于 CNN-Backbone，而 SincNet、WKN-Morlet、WKN-Laplace 和 W-CNN 四种模型的表现略优于前两种模型。TFN-STTF、TFN-Chirplet 和 TFN-Morlet 取得了总体上最佳的诊断性能，证明了所提出方法在诊断精度方面的有效性。此外，具有 32 通道的 TFN 表现明显优于具有 16 通道的 TFN，但具有更多通道（即 64、128）的 TFN 准确率相比 18 通道的 TFN 在准确率上并没有显著优势。这表明 32 通道足以使 TFN 在该行星齿轮数据集上提取足够的时频信息。

对于行星齿轮数据集的可解释性分析，考虑到基准模型、对比模型和 TFN-Morlet 在可解释性方面的表现较差，为简洁起见，现仅展示 TFN-STTF 和 TFN-Chirplet 的可

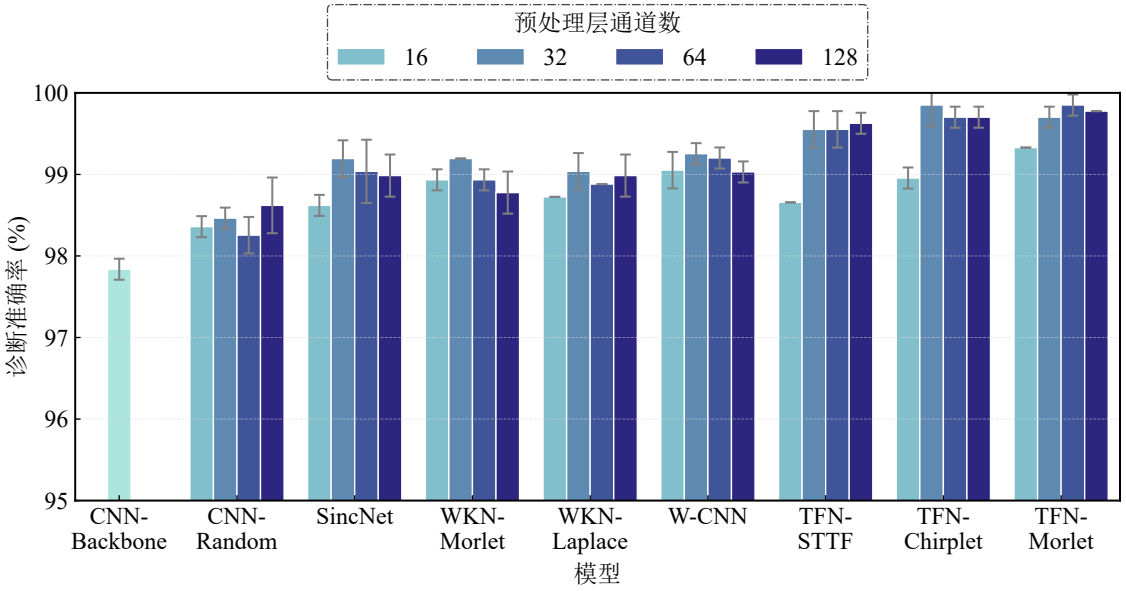
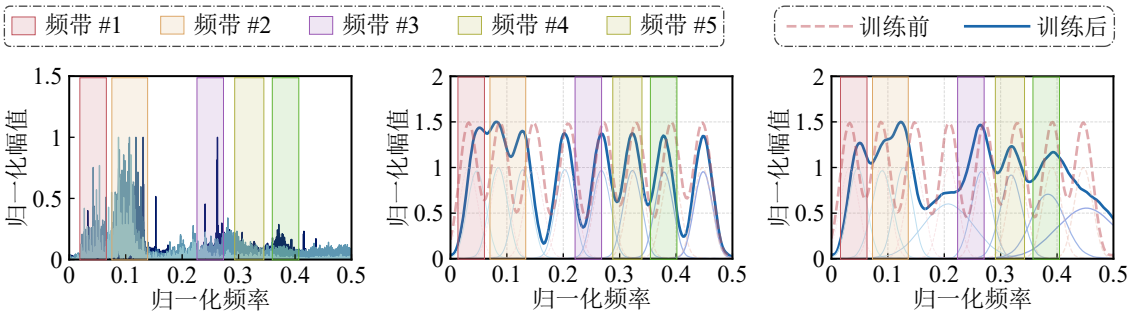


图 2-12　不同预处理层通道数下各类模型在行星齿轮数据集的诊断准确率

Fig. 2-12　Diagnostic accuracies of various models under different numbers of preprocessing channels in the planetary gearbox dataset

解释性结果。预处理层的通道数设置为 8，其他实验设置与行星齿轮数据集上的诊断实验保持一致。行星齿轮数据集的频谱和 TFN-STTF 及 TFN-Chirplet 的综合幅频响应（O-FR）如图 2-13 所示。



(a) 行星齿轮数据集频谱　(b) TFN-STTF的O-FR　(c) TFN-Chirplet的O-FR

图 2-13　行星齿轮数据集的频谱以及不同模型预处理层的综合幅频响应

Fig. 2-13　Frequency spectrum ofplanetary gearbox dataset and the O-FRs of different model preprocessing layers

由图 2-13(a) 可知，行星齿轮数据集大部分信息主要存在于所标注出的五个信息频带中，即频带 #1-#2-#3-#4-#5。图 2-13(b)-(c) 显示，训练后的 TFN-STTF 和 TFN-Chirplet 的 O-FR 在这五个信息频带处均有峰值，表明 TFN-STTF 和 TFN-Chirplet 正确地关注了行星齿轮数据集的信息频带。然而，与 STTF 核函数相比，Chirplet 核函

数具有额外的线性调频因子 α 用于调整其滤波带宽，因此 TFN-STTF 的 O-FR 在信息频带之外有两个峰值（一个频率接近 0.2，另一个频率接近 0.45），而 TFN-Chirplet 的 O-FR 没有无关的峰值，更加符合行星齿轮数据集的频谱。这一现象与 CWRU 轴承数据集的结合相吻合，表明了 TFN-Chirplet 在揭示 CNN 模型关注频率区域方面的优异可解释性。

在少样本学习方面，行星齿轮数据集中每个类别包含 440 个样本，总共 2200 个样本。将训练样本的数量作为实验变量以测试 TFN 的少样本学习能力，并使用如式 (2-12) 所示的策略来控制训练周期。少样本实验的结果如图 2-14 所示。包含 TFN-STTF、TFN-Chirplet 和 TFN-Morlet 的 TFN 模型，其在少样本场景下的诊断准确率明显优于其他模型。当训练样本数量为 5 到 50 时，TFN 模型与对比模型的诊断准确率差距保持在 10% 左右。随着训练样本数量的继续增加，诊断准确率差距逐渐缩小，但 TFN 模型的准确率仍然高于对比模型。

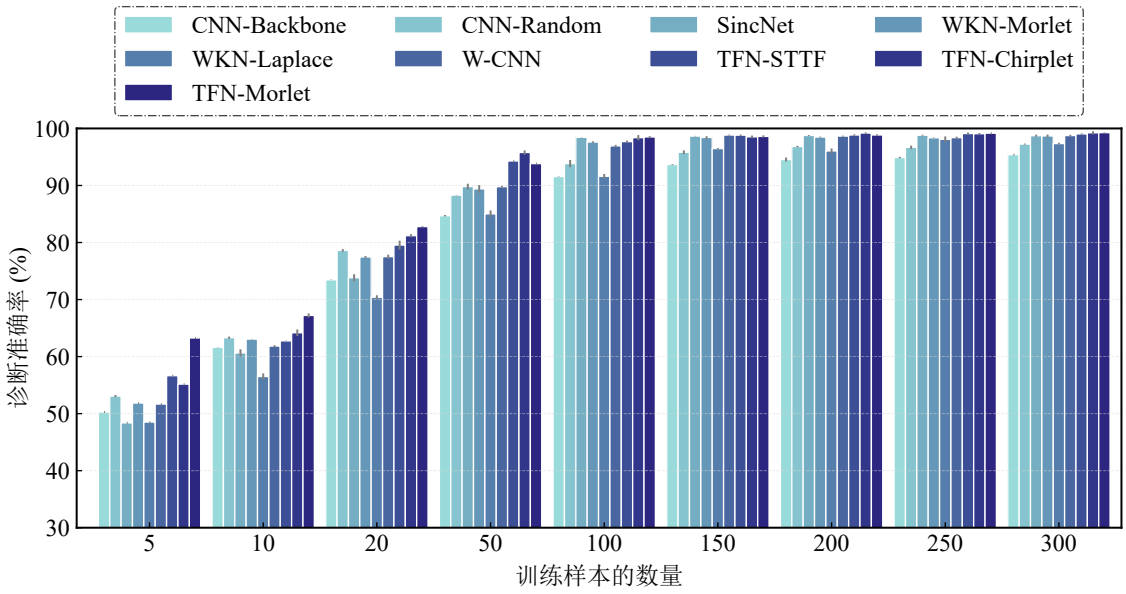


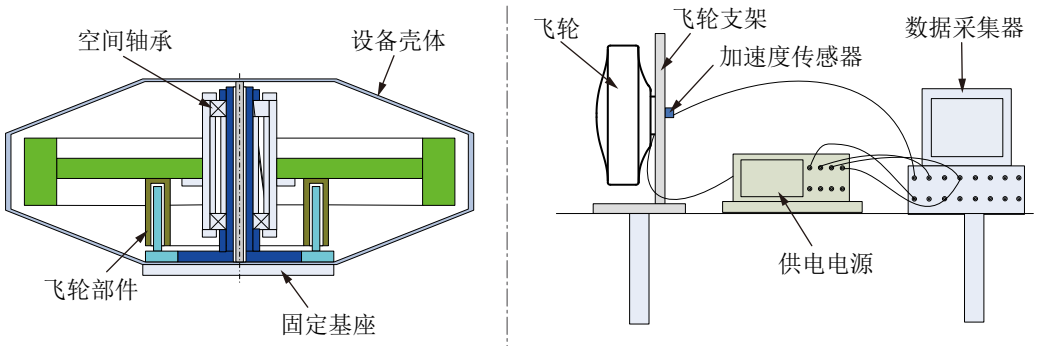
图 2-14　不同训练样本数量下各类模型在行星齿轮数据集的诊断准确率

Fig. 2-14　The diagnostic accuracies of different models under different training sample numbers on the planetary gearbox dataset

2.4.3　工业应用场景下的空间轴承数据集

前两个数据集是在实验室场景中采集的，而该空间轴承数据集则来源于工业场景。如图 2-15 所示，空间轴承是飞轮试验台的核心部件，该测试台包括电机组件、空间轴承、飞轮组件、外壳和安装底座，如图 2-15(a) 所示。飞轮由电动机驱动，然后使

空间轴承工作。数据采集设备如图 2-15(b) 所示，包括加速度计、电源、信号采集和分析系统。飞轮固定在一个直立的支架上，安装在支架上的加速度传感器以 25.6 kHz 的采样频率采集空间轴承的三向振动信号。



(a) 飞轮结构示意图　　(b) 振动信号采集过程示意图

图 2-15　空间轴承数据集的飞轮结构和试验台示意图

Fig. 2-15　Schematic diagram of flywheel structure and test bench for aerospace bearing dataset

空间轴承数据集包含五种健康状态：健康状态（H）、引导面划伤（S）、保持架故障（C）、滚动体故障（B）和内圈故障（I）。对于每种状态，径向振动信号被截断成样本，各类别包含 1000 个样本，总共 5000 个样本。实验中，各类别中 60% 的样本用于训练，其余样本用于测试。空间轴承数据集的故障诊断可以视为一个五分类任务。与行星齿轮数据集的处理类似，原始信号中加入了信噪比为 0 的高斯白噪声，以增加诊断难度。其余实验设置与之前行星齿轮数据集的实验一致。

实验结果如图 2-16 所示。CNN-Backbone 的诊断准确率最低，为 87.3%；而 CNN-Random 的准确率约为 93%，其诊断精度有着明显提高。这可能是由于新添加的传统卷积层增加了模型深度，使得模型的学习能力有所提高。包含 SincNet、WKN-Morlet、WKN-Laplace、W-CNN 的对比模型具有和 CNN-Random 相近的诊断准确率，其中 128 通道的 SincNet 表现最优，其诊断准确率达到 96.7%。对于 TFN 模型，TFN-STTF、TFN-Chirplet 和 TFN-Morlet 在诊断准确率上表现最佳，其中 128 通道的 STTF-TFN 的平均准确率最高，为 98.3%。具有更多通道的时频卷积层可以提取更精细的时频信息，在获得更高诊断准确率的同时，也导致更长训练时间。由此，预处理层通道数目和训练时间的关系，也是后续分析部分的研究对象之一。总结而言，所提出的 TFN 模型在诊断准确率上远优于基准模型和对比模型，并且 TFN 模型的诊断准确率随着通道数量的增加显著提高，证明了通道数量选取对模型的诊断表现具有重要影响。

在可解释性分析方面，选择预处理层通道数为 8 的 TFN-STTF 和 TFN-Morlet 在

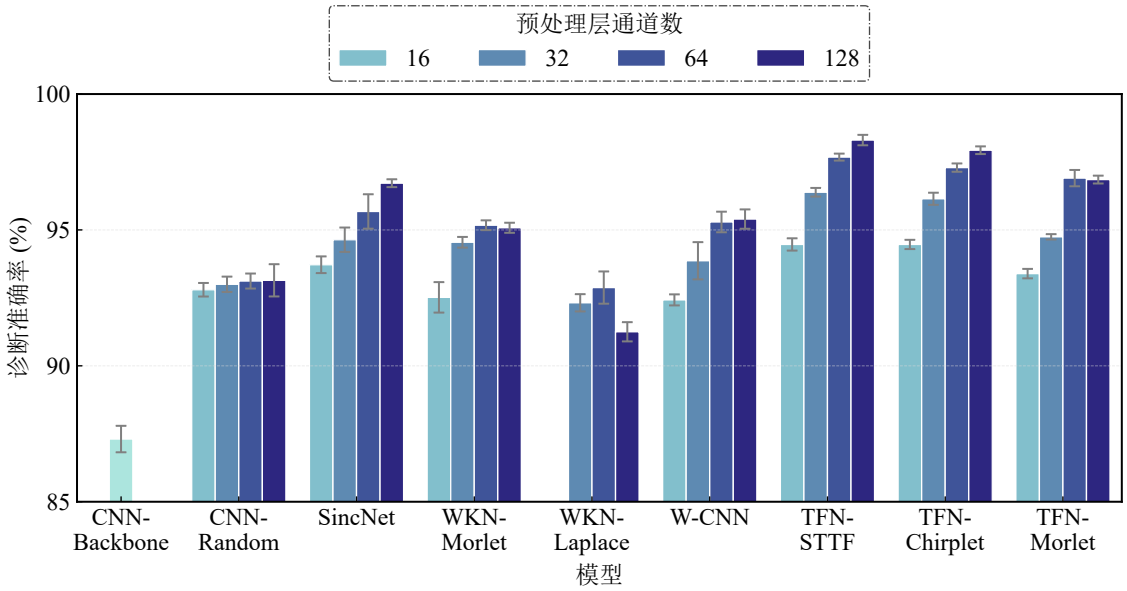
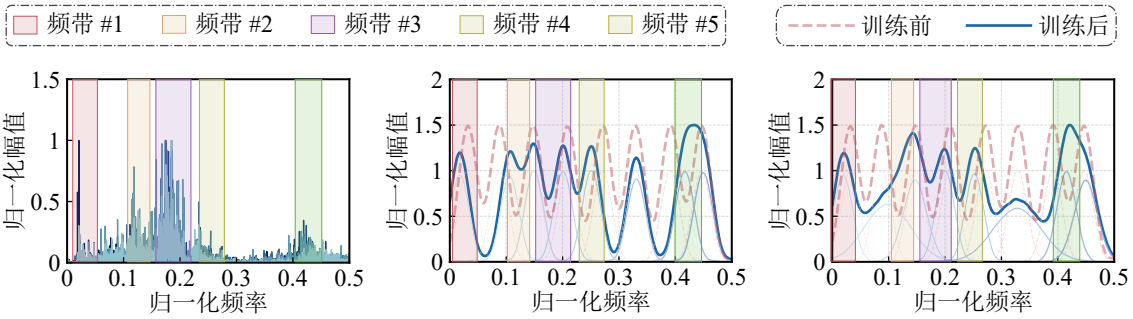


图 2-16　不同预处理层通道数下各类模型在空间轴承数据集的诊断准确率

Fig. 2-16　Diagnostic accuracies of various models under different numbers of preprocessing channels in the aeropsace bearing dataset

空间轴承数据集上进行训练，实验设置与之前的诊断实验相同。空间轴承数据集的频谱和 TFN-STTF 及 TFN-Morlet 的综合幅频响应（O-FR）如图 2-17 所示。



(a) 空间轴承数据集频谱　(b) TFN-STTF的O-FR　(c) TFN-Chirplet的O-FR

图 2-17　空间轴承数据集的频谱以及不同模型预处理层的综合幅频响应

Fig. 2-17　Frequency spectrum of aeropsace bearing dataset and the O-FRs of different model preprocessing layers

如图 2-17(a) 所示，五个信息频带包含空间轴承数据集大部分信息，即频带 #1-#2-#3-#4-#5。图 2-17(b)-(c) 显示，结果与前两个数据集的可解释性分析一致。两个 TFN 模型都正确地关注了空间轴承数据集的信息频带，但 TFN-STTF 的 O-FR 在信息频带之外有一个峰值（频率接近 0.35），而 TFN-Chirplet 能够通过线性调频因子 α 调整滤

波带宽，使得其 O-FR 不具有额外的峰值。总结而言，训练后的 TFN-Chirplet 的 O-FR 与空间轴承数据集的频谱有很好的对应关系，再次证明了 TFN-Chirplet 的卓越可解释性。

少样本学习方面，与之前的处理一致，选择不同数量（即 5、10、20、50、100、200、300、500 和 700）的各类样本作为训练数据，其余样本用于测试，空间轴承数据集下的少样本学习实验结果如图 2-18 所示。

当训练样本数量较少时，TFN 模型，尤其是 TFN-Morlet，其表现显著优于对比模型。具体而言，当训练样本数量为 5 时，对比模型的测试准确率接近 50%，而 TFN-STTF 和 TFN-Chirplet 超过 60%，TFN-Morlet 达到 71%。在训练样本数量增加到 200 之前，TFN 模型与对比模型的诊断准确率差异始终保持在 10% 以上，特别是 TFN-Morlet 的表现更为突出。在所有情况下，TFN 模型的准确率均高于对比模型。上述结果有效地展示了 TFN 模型在少样本场景中的优越性。

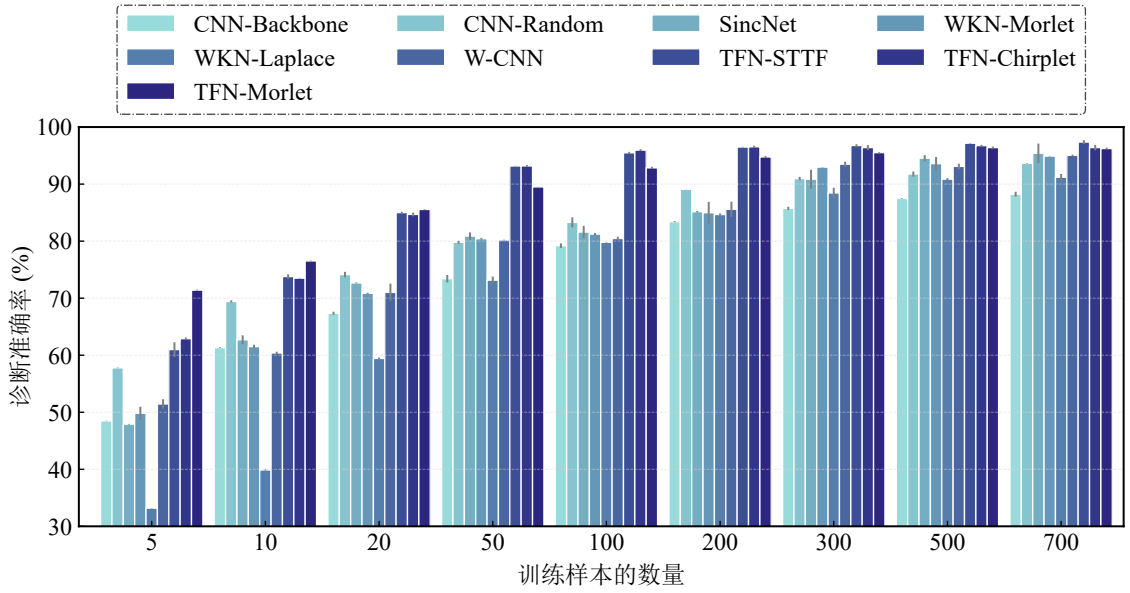


图 2-18　不同训练样本数量下各类模型在空间轴承数据集的诊断准确率

Fig. 2-18　The diagnostic accuracies of different models under different training sample numbers on the aeropsace bearing dataset

基于上述诊断性能和可解释性的实验结果，现对这三种卷积核函数进行了总体评价。总结而言，尽管 Morlet 卷积核在诊断性能和少样本学习方面表现良好，但其在可解释性方面则表现较差。因此，建议在实际应用中使用 STTF 卷积核或 Chirplet 卷积核，其中 STTF 卷积核在诊断准确率上具有优势，而 Chirplet 卷积核更适合用于可解释性分析。

2.5　输入层主动解释方法的本质剖析及优势验证

为进一步分析 TFN 的特性，现基于 CWRU 数据集开展对 TFN 的本质剖析，包括与其他方法的本质区别、收敛过程、训练时间、以及通用性分析。具体而言，本质剖析部分将对 TFN 与同类模型的差异进行公式推导和可视化，以解释 TFN 在诊断性能上的优越性原因。训练过程和训练时间的分析部分将讨论 TFN 在训练过程中的收敛速度优势以及 TFN 的训练成本。通用性分析则以基准模型作为实验变量，用以验证将时频卷积层推广到其他 CNN 模型的可行性。

2.5.1　时频卷积网络与现有信号融入网络的本质剖析

同类方法中的 SincNet[87]、WKN[89] 和 W-CNN[151]，其本质是通过特定的核函数对传统卷积核进行初始化或控制，但这些模型仅考虑实数卷积核，等价于一系列带通 FIR 滤波器，其输出为滤波后的子信号。相反，时频卷积层使用实虚部机制来模拟复数卷积核，并通过后续的复数求模运算，使得其输出为时频分布，即信号在时频域中的能量分布。这与传统时频变换方法相一致，但时频卷积层的参数是可训练的，可以根据数据集特性进行自适应调整。

现通过公式推导和可视化两种方式来展示 TFN 模型和对比模型之间的区别。在公式推导部分，TFN 模型使用的“复数核”和对比模型使用的“实数核”在输出具有显著区别。以 STFT 这一基本的时频变换方法为例。给定输入信号 $x(t)$，“复数核”的计算过程可以表示为

$$X(\tau, f)=\frac{1}{2\pi}\int x(t)w(t-\tau)\mathrm{e}^{-\mathrm{i}2\pi ft}\,\mathrm{d}t, \tag{2-13}$$

式中 $X(\tau,f)$ 代表“复数核”过程的输出，τ 和 f 分别代表时间和频率。$w(t)$ 是 STFT 的窗口函数，其长度表示为 T。“实数核”的计算过程可以表示为

$$\widehat{X(\tau, f)}=\frac{1}{2\pi}\int x(t)w(t-\tau)\cos 2\pi ft\,\mathrm{d}t. \tag{2-14}$$

根据 Fourier 级数，上式可以展开为

$$x(t)w(t-\tau)=\frac{a_0(\tau)}{2}+\sum_{n=1}^{\infty}a_n(\tau)\sin\left[n2\pi f_0t+\phi_n(\tau)\right], \tag{2-15}$$

式中 $f_0=1/T$ 代表时间 T 下的基准频率，$a_n(\tau)$ 和 $\phi_n(\tau)$ 分别代表输入信号 $x(t)$ 在时间 τ 和频率 nf_0 的幅值和相位。

由此，将式 (2-15) 代入式 (2-13) 并考虑正弦函数的正交性，可以得到“复数核”计算过程的输出：

$$\begin{aligned}|X(\tau,f)| &= \left|\frac{1}{2\pi}\int x(t)w(t-\tau)\mathrm{e}^{-\mathrm{i}2\pi ft}\,\mathrm{d}t\right| \\ &= \left|\frac{1}{2\pi}\int\left\{\frac{a_0(\tau)}{2}+\sum_{n=1}^{\infty}a_n(\tau)\sin\left[nf_0t+\phi_n(\tau)\right]\right\}\mathrm{e}^{-\mathrm{i}2\pi ft}\,\mathrm{d}t\right| \\ &= \left|\frac{T}{2}\cdot\left[a_f(\tau)\sin\left(\phi_f(\tau)\right)+\mathrm{i}\cdot a_f(\tau)\cos\left(\phi_f(\tau)\right)\right]\right| \\ &= \frac{T}{2}a_f(\tau),\end{aligned} \tag{2-16}$$

式中 $a_f(\tau)$ 是输入信号 $x(t)$ 在时间 τ 和频率 f 的幅值，即时频分布。

同样地将式 (2-15) 代入式 (2-14)，可以得到“实数核”过程的输出：

$$\begin{aligned}\widehat{X(\tau,f)} &= \frac{1}{2\pi}\int x(t)w(t-\tau)\cos(2\pi ft)dt \\ &= \frac{1}{2\pi}\int\left\{\frac{a_0(\tau)}{2}+\sum_{n=1}^{\infty}a_n(\tau)\sin\left[n2\pi f_0t+\phi_n(\tau)\right]\right\}\cos(2\pi ft)dt \\ &= \frac{T}{2}a_f(\tau)\cdot\sin\phi_f(\tau),\end{aligned} \tag{2-17}$$

式中包含两部分，前一部分 $a_f(\tau)$ 是时频分布的能量，后一部分是与时间 τ 和频率 f 相关的相位信息。总结而言，TFN 采用的“复数核”的输出是不包含相位信息的时频分布；而对比模型采用的“实数核”输出的是滤波后的子信号，同时包含能量信息和相位信息。

在公式推导部分之后，现通过可视化分析进一步展示 TFN 模型和 SincNet、WKN-Morlet、WKN-Laplace、W-CNN 这些对比模型之间的区别。所有模型的预处理层通道数目均设置为 64，实验设置与 CWRU 数据集上的可解释性分析相同。不同模型处理过程的示意图如图 2-19 所示，图的左边展示了一个模拟输入信号和通过传统 STFT 处理获得的时频分布。图的中间展示了各个模型训练后的预处理层第 16 个卷积核，以展示模型在卷积核形状上的差异。图的右边则展示了该模拟输入信号经过这些模型预处理层后的输出，以展示模型在输出上的差异。

从图中各个模型的卷积核可以看出，CNN 预处理层采用了传统卷积层，其卷积核是随机初始化的。SincNet、WKN-Morlet、WKN-Laplace、W-CNN 这类对比模型均使用特定函数将卷积核约束至特定频率，从而实现 FIR 带通滤波器的功能。而 TFN-STTF、TFN-Chirplet 和 TFN-Morlet 这类 TFN 模型则使用复数卷积核并对其进行参数化，在提取幅值信息和相位信息后，通过复数求模运算来模拟时频变换。

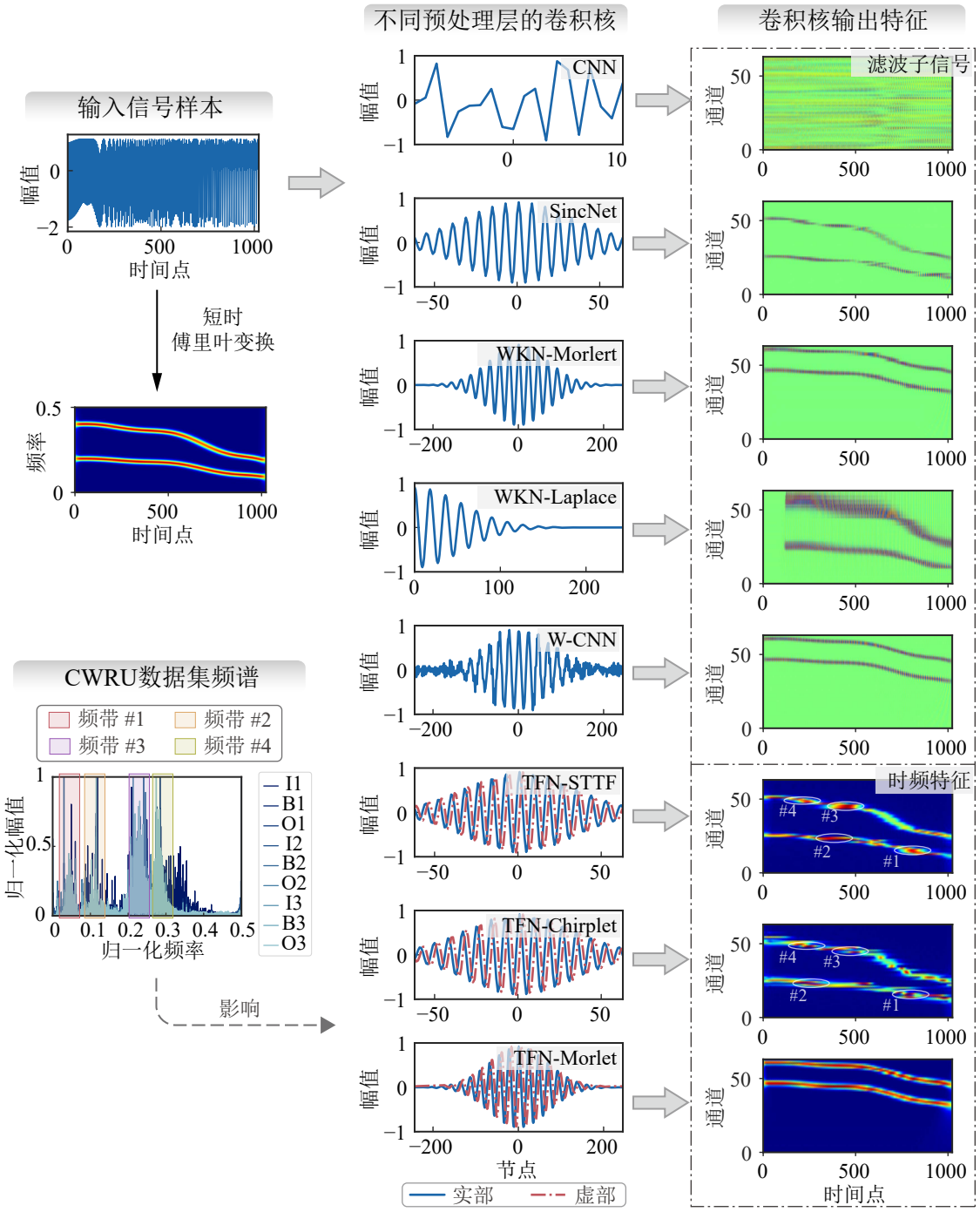


图 2-19　不同模型在卷积核和处理过程方面的对比图

Fig. 2-19　Comparative diagram of different models in terms of convolution kernels and processing

从图中各个模型预处理层的输出可以看出，CNN 的输出难以观察出可理解的信息。SincNet、WKN-Morlet、WKN-Laplace、W-CNN 这类对比模型通过将各类核函数融入预处理层，使得其输出与输入模拟信号的时频分布有着明确的对应关系，表明这些模型在一定程度上可以提取时频信息。但这些模型仅考虑实数卷积核，如式 (2-17) 所示，其输出同时包含输入信号的能量信息和相位信息。这使得这些对比的输出较为模糊，受到相位信息的干扰，仍然与输入信号的时频分布有明显差异。相反，TFN 模型使用实虚部机制来模拟复数卷积核，使得 TFN 模型的输出与输入信号的时频分布完全对应，只是在某些特定频带上有一些“失真”。可以发现，这些“失真”是由卷积核函数的训练过程引起的。图 2-9 的可解释性实验表明，TFN 模型在训练过程中会改变其幅频响应以更好地捕捉数据集的故障特征，而图 2-19 中 TFN-STTF 和 TFN-Chirplet 所展现出的“失真”正好对应于 CWRU 数据集频谱中的信息频带。这也与之前在 CWRU 数据集的可解释性分析中讨论的现象一致，进一步验证了 TFN 模型的解释逻辑。

总结而言，对比模型采用实数卷积核，其输出是一系列滤波后的子信号，尽管能与输入信号的时频分布有着一定程度的对应，但却受到相位信息的干扰，表现出明显的差异。相反地，TFN 模型采用复数卷积核来模拟时频变换，其输出是输入信号的自适应时频分布。不同于普通的时频变换，TFN 模型的预处理层能够通过网络训练过程对核函数参数进行优化，使得预处理层更加关注训练数据集的信息频带，自适应地提取与故障相关的特征，这也是 TFN 模型在诊断性能方面更具优势的原因。

2.5.2　时频卷积网络的收敛速度和训练时间分析

为了全面分析 TFN 的性能，现对 TFN 模型的收敛速度和训练时间进行测试。和之前的实验一样，所采用的模型仍是表 2-3 所示的九种模型。预处理层的通道数均设置为 64，训练周期数目设置为 80 以获得完整的训练记录，其余实验设置与 CWRU 数据集上的诊断实验相同。

不同模型在 CWRU 数据集的训练过程如图 2-20 所示，其中 CNN-Backbone 表现最差，收敛速度最慢。CNN-Random、W-CNN 和 WKN-Laplace 的收敛速度略好于 CNN-Backbone。SincNet 和 WKN-Morlet 属于第二梯队，它们的性能略优于前述五种模型。借助时频卷积层的时频特征提取能力，三种 TFN 模型比所有其他模型具有更快的收敛速度，其中 TFN-Morlet 的性能最佳，远远领先其他模型。收敛速度与图 2-8 所示的诊断性能具有良好对应关系。

故障诊断实验表明，预处理层的通道数越多，模型的故障诊断精度越高，但预处

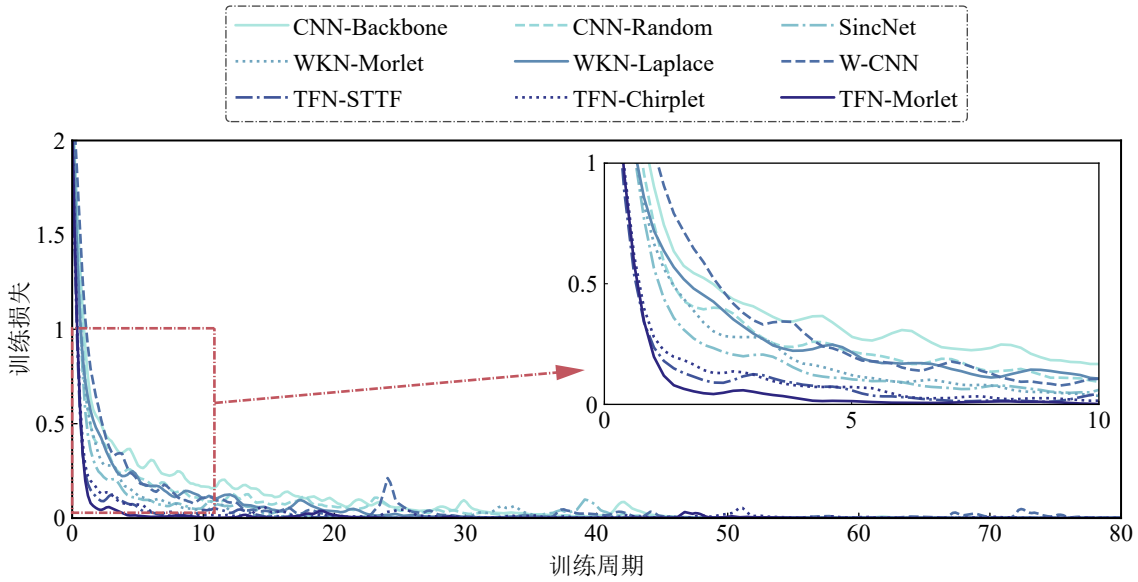


图 2-20　不同模型在 CWRU 数据集上的训练过程

Fig. 2-20　The training process of different models on the CWRU dataset

理层的通道数也会显著增加模型训练时间。此外，卷积核调控可以显著提高 CNN 模型的诊断能力，但诊断能力的提高是以训练时间为代价的，而目前的文献中很少提及这一点[87,89,151]。为了量化卷积核调控和预处理层通道数等变量的计算代价，现记录 CWRU 数据集的诊断实验中不同通道数下各个模型的训练时间，如图 2-21 所示。

(1) CNN-Backbone 和 CNN-Rondom 的每次训练时间接近 22 秒。W-CNN 仅对卷积核进行初始化而非调控，因此 W-CNN 的每次训练时间接近 30 秒，与 Backbone 模型接近。

(2) SincNet、WKN-Morlet 和 WKN-Laplace 则对实数卷积核进行调控，它们的训练时间相比 CNN-Backbone 显著增加，且 128 个通道的训练时间接近 400 秒，几乎是 Backbone 模型的 18 倍。此外，如表 2-1 所示，WKN-Morlet 和 WKN-Laplace 具有更长的核长度，因此它们的训练时间多于 SincNet。

(3) TFN-STTF、TFN-Chirplet 和 TFN-Morlet 对复数卷积核进行参数化，且 128 个通道的训练时间接近 500 秒，几乎是 Backbone 模型的 23 倍。在这三个模型中，TFN-Chirplet 的训练时间高于 TFN-STTF，因为 Chirplet 核函数具有额外的控制参数（即线性调频因子 α）需要训练，而 TFN-Morlet 的训练时间最高，由于 Morlet 核函数的卷积核长度更长。

(4) 随着通道数量的增加，这些参数化卷积核的模型的训练时间显著增加，因此需

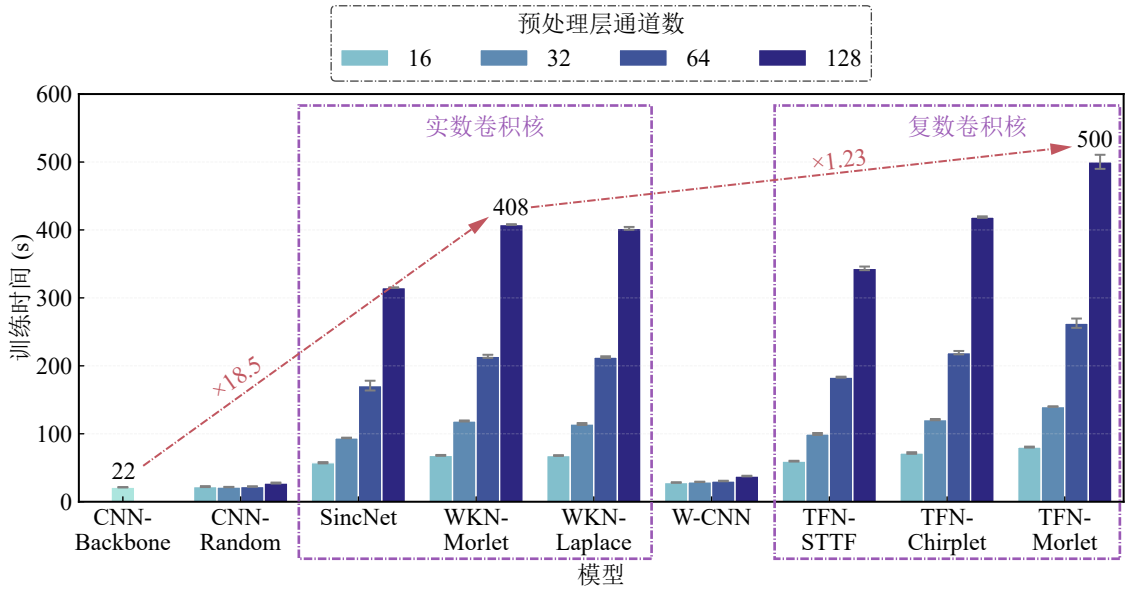


图 2-21　不同预处理层通道数下各类模型在 CWRU 数据集上的训练时间

Fig. 2-21　The training time of different models with different preprocessing layer channels on the CWRU dataset

要仔细考虑此类模型的通道数量，以在诊断准确率和训练时间之间取得平衡。

总结而言，调控卷积核非常耗时，与 Backbone 模型相比，调控卷积核的模型需要大量的训练时间。然而，相比调控实数卷积核的同类方法[87,89]，调控复数卷积核的 TFN 模型并未显著增加训练时间，这使得 TFN 模型相比对比模型同样具有竞争力，并未表现出明显劣势。

2.5.3　时频卷积层的通用性分析

为了验证时频卷积层的通用性，现以基准网络为实验变量，选择三种具有不同深度的典型 CNN 作为基准网络并进行准确率测试。使用 CWRU 数据集在 3HP 负载条件下的振动信号作为输入样本，并将时频卷积层的通道数全部设置为 128。其他实验设置与之前在 CWRU 数据集上进行的诊断实验一致，得到的故障诊断准确率如表 2-5 所示。

从结果可以看出，针对不同的基准网络，融入时频变换的时频卷积层均能够显著提高其诊断准确率。其中，Morlet 时频卷积层始终获得最佳的诊断性能，其次是 STTF 时频卷积层。Chirplet 时频卷积层略逊于 Morlet 时频卷积层和 STTF 时频卷积层，但仍然优于基准网络。该实验表明，时频卷积层是一种通用方法，可以应用于具有不同深度的 CNN，以提高其诊断性能。此外，实验结果同时也强调了基准网络的

表 2-5　不同基准网络的 TFN 在 CWRU 数据集上的故障诊断结果
Table 2-5　Fault diagnosis accuracies of TFNs with different backbone networks on the CWRU dataset

基准网络	预处理层	准确率（%）	方差（%）
LeNet	无	90.02	0.177
	STTF 时频卷积层	95.71	0.355
	Chirplet 时频卷积层	94.25	0.339
	Morlet 时频卷积层	**98.96**	0.346
AlexNet	无	97.32	0.476
	STTF 时频卷积层	98.27	0.413
	Chirplet 时频卷积层	97.84	0.339
	Morlet 时频卷积层	**99.89**	0.129
ResNet	无	97.74	0.243
	STTF 时频卷积层	99.58	0.183
	Chirplet 时频卷积层	98.79	0.464
	Morlet 时频卷积层	**99.96**	0.058

重要性，尽管 Morlet 时频卷积层能够使 LeNet 这一基准网络的准确率从 90.02% 提升至 98.96%，但仍然不如 Morlet 时频卷积层和 ResNet 组合后得到的 98.96% 这一准确率。因此，建议研究人员采用足够深度的基准 CNN 来构建 TFN 模型，以保证模型的诊断性能。

2.6　本章小结

针对旋转机械智能诊断中主动解释效果、诊断性能及可拓展性的多方共赢问题，本章聚焦智能诊断模型输入层，提出一种融入传统时频变换的新型时频卷积层及其完整应用流程，在保证诊断性能的前提下，有效揭示模型决策依据, 并通过三个实测数据集进行实验验证。本章的主要内容可总结如下：

(1) 建立了时频变换与卷积层相融合的时频变换核函数。两者本质均为输入信号与窗函数或卷积核的内积运算，其差异主要体现为窗函数和卷积核的区别。时频变换窗函数被精心设计为具有固定参数的复数形式 FIR 滤波器，而传统卷积核则采用随机初始化并通过训练优化。基于两者在内积运算的等价性，可以为卷积核赋予复数形式并引入核函数约束来使卷积层与时频变换相等价，进而从三种经典时频变换方法中提取对应核函数（STTF、Chirplet 和 Morlet），为时频卷

积层的设计奠定了坚实的理论基础。

(2) 构建了融入时频变换的可解释时频卷积层及相应解释分析方法。时频卷积层通过实虚部机制与核函数约束，不仅能够有效模拟传统时频变换，还可参与网络训练过程，实现核函数参数的自适应优化。将时频卷积层作为输入层与现有模型结合并构建时频卷积网络，实现信号处理先验知识对神经网络的有效赋能。时频卷积层训练后的综合幅频响应能够准确揭示数据集信息频带，为神经网络故障诊断决策提供可靠的频域解释。

(3) 三类实测数据集的实验结果表明，时频卷积网络在诊断准确率、收敛速度和少样本学习能力方面均显著优于传统 CNN 和同类网络，尤其在复杂工况和小样本条件下优势更为突出。时频卷积层的综合幅频响应与数据集信息频带高度吻合，证实了时频卷积层能够通过自适应学习聚焦故障相关频率成分，从而从频域视角揭示模型诊断的决策依据。三种核函数的对比研究表明，STTF 核函数和 Morlet 核函数在诊断准确率方面表现卓越，而 Chirplet 核函数则在可解释性分析方面更具优势，为不同应用场景提供了差异化选择。此外，时频卷积层展现出优异的通用性和灵活性，能够与不同深度的 CNN 架构相结合，实现诊断性能的全面提升。

第三章　基于原型匹配的智能诊断模型决策层主动解释

3.1　引言

时频卷积网络的主动解释主要局限于智能诊断模型的输入层，而未能深入阐释模型的末端决策逻辑。在实际应用场景中，用户不仅需要了解模型对输入信号的关注点分布，还需要理解模型的决策推理过程，即模型如何基于从振动信号中提取的高维特征映射到特定的故障类别判断。

针对上述问题，本章将研究重点转向决策层的主动解释，旨在揭示智能诊断模型分类逻辑的内在机制。为实现此目标，本章将可解释的原型匹配概念融入智能诊断模型的决策层，并通过与自编码器结构的深度结合，构建原型匹配网络 (Prototype-Matching Network, PMN)，从而实现智能诊断网络决策层的透明化和可解释性。原型匹配作为人类固有的分类逻辑范式，能够显式构建各类别的原型表征，并基于输入样本与各类别原型之间的相似度量进行分类判断。所提出的原型匹配网络具有三个维度的可解释性：(1) 在决策逻辑方面，原型匹配网络基于直观的原型匹配原理，通过计算样本与类别原型间的相似性实现故障类别的精准诊断；(2) 在类别原型方面，原型匹配网络能够明确提取各类别的特征原型，并借助解码器将其映射回样本域，从而增强信号中的故障特征；(3) 在相似性来源方面，原型匹配网络通过引入额外的归因解释，以原型匹配层的输出距离为起点，阐明输入信号与匹配原型之间的相似性在各频率成分上的来源。

本章首先系统介绍原型匹配和自编码器的理论基础，继而详细阐述基于原型匹配的旋转机械智能诊断决策层主动解释方法，包括原型匹配网络的架构设计、损失函数构建、解释性体现以及完整的应用流程。最后，通过传统故障诊断场景和领域泛化故障诊断场景的实证研究，全面验证原型匹配网络在诊断性能和可解释性方面的显著优势。

3.2　原型匹配逻辑和神经网络自编码器及其在智能诊断中的应用

3.2.1　基于距离分类的可解释原型匹配逻辑

原型匹配是一种显式构建类别原型、并基于样本与原型的相似性进行分类的直观逻辑。如图 3-1 所示，原型匹配是人类与生俱来的分类逻辑，人类会构建各个类别

的原型概念。当对特定物体进行判断时，人类会将该物体和各个原型概念进行相似度匹配，从而实现物体正确分类。并且，人类所构建的原型，具有该类别最典型的特征，比如“苹果”原型概念会具有“红颜色”、“圆形状”、“甜/酸口味”等典型的苹果类别特征。相反地，神经网络（人工智能）只是一系列非线性映射的黑箱组合，并不天然地具备原型匹配的能力。尽管经过优化和学习后，神经网络能够准确判定输入振动信号的故障类别，但其分类的逻辑却不为人所理解，也并难以准确描绘各故障类别的典型特征。因此，如何将原型匹配逻辑引入神经网络，以提高智能诊断模型的决策层可解释性，是本章的后续研究重点。

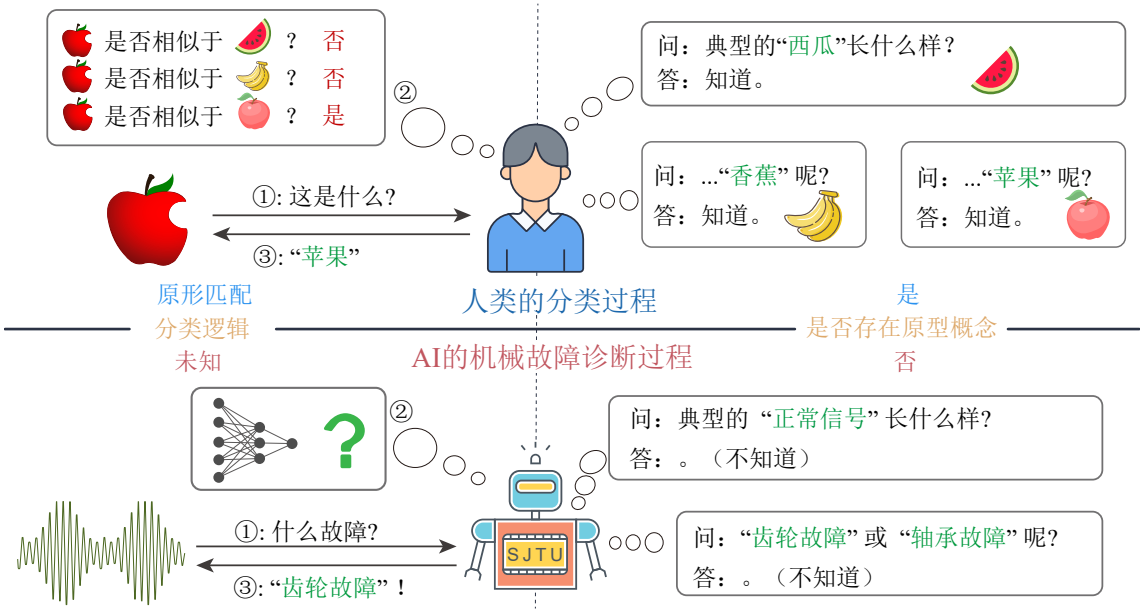


图 3-1　原型匹配分类逻辑示意图

Fig. 3-1　The illustration of prototype-matching logic

原型匹配的计算过程可大致分为三部分，原型向量的构建、距离的计算以及分类的决策。首先，原型向量可通过样本特征均值或网络学习等方法构建。样本特征均值的方式构建样本原型的过程可表示为

$$\boldsymbol{p}^k = \frac{\sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{x}_i \cdot \mathbb{I}(y_i = k)}{\sum_{i=1}^{n} \mathbb{I}(\boldsymbol{y}_i = k)} = \frac{1}{n_k} \cdot \sum_{i=1}^{n_k} \boldsymbol{w}_i \boldsymbol{x}_i^k, \tag{3-1}$$

式中 $\boldsymbol{p}^k$ 代表类别 k 的样本原型，n 代表数据集样本数量，$(x)_i$ 和 $\boldsymbol{w}_i$ 分别代表第 i 个输入样本和其对应的样本权重。$\mathbb{I}(\cdot)$ 代表指示函数，当括号内的表达式成立则为 1，不成立则为 0。n_k 代表第 k 类样本的数量，$\boldsymbol{x}_i^k$ 代表第 k 类样本的第 i 个样本。

距离的计算则是将输入样本与各个原型向量进行距离计算，常用的距离度量包

括 L_2 距离（欧式距离）、L_1 距离和余弦距离。各距离度量的计算可表示为

$$\begin{aligned} d_{L_2}(\boldsymbol{x},\boldsymbol{p}) &= \|\boldsymbol{x}-\boldsymbol{p}\|_2 = \sqrt{\sum_{i=1}^{d}(\boldsymbol{x}_i-\boldsymbol{p}_i)^2}, \\ d_{L_1}(\boldsymbol{x},\boldsymbol{p}) &= \|\boldsymbol{x}-\boldsymbol{p}\|_1 = \sum_{i=1}^{d}|\boldsymbol{x}_i-\boldsymbol{p}_i|, \\ d_{L_{\cos}}(\boldsymbol{x},\boldsymbol{p}) &= \frac{\boldsymbol{x}\cdot\boldsymbol{p}}{\|\boldsymbol{x}\|_2\cdot\|\boldsymbol{p}\|_2} = \frac{\sum_{i=1}^{d}\boldsymbol{x}_i\cdot\boldsymbol{p}_i}{\sqrt{\sum_{i=1}^{d}\boldsymbol{x}_i^2}\cdot\sqrt{\sum_{i=1}^{d}\boldsymbol{p}_i^2}}, \end{aligned} \tag{3-2}$$

式中 $d_{L_2}(\cdot,\cdot)$、$d_{L_1}(\cdot,\cdot)$、$d_{L_{\cos}}(\cdot,\cdot)$ 分别代表 L_2 距离、L_1 距离和余弦距离，$\boldsymbol{x}$ 代表输入样本，$\boldsymbol{p}$ 代表原型向量，d 代表样本特征维度。

最后，分类的决策是将输入样本分配给与其距离最小的原型向量所对应的类别，常用的方式即是将距离结果的负值作为 Softmax 函数的输入，进而得到各类别的概率分布，并实现分类决策。Softmax 函数的计算公式为

$$p(y=k|\boldsymbol{x}) = \frac{\exp\left(-d(\boldsymbol{x},\boldsymbol{p}^k)\right)}{\sum_{i=1}^{K}\exp\left(-d(\boldsymbol{x},\boldsymbol{p}^i)\right)}, \tag{3-3}$$

式中 $p(y=k|\boldsymbol{x})$ 代表输入样本 $\boldsymbol{x}$ 属于类别 k 的概率，$d(\boldsymbol{x},\boldsymbol{p}^k)$ 代表输入样本 $\boldsymbol{x}$ 与类别 k 的原型向量 $\boldsymbol{p}^k$ 的距离。原型匹配的优势在于其直观性和可解释性，使得模型的决策逻辑变得透明，易于人类理解。

原型匹配的实际应用如图 3-2 所示，可大致可以分为三个发展阶段。在初始阶段，原型匹配主要应用于原始样本或人工提取的特征，代表方法包括 K-means 或多元高斯分布。这些方法以距离相似性为基准，显式地估计各个类别的原型，如 K-means 的距离中心和多元高斯分布的高斯元中心。方法在简单任务下效果尚佳，具有可解释性，但缺乏足够的非线性映射能力，却难以满足复杂任务的需求。在第二阶段，借助神经网络在特征提取方面的强大映射能力，原型匹配被应用于神经网络的高维特征之上，通过嵌入特征的均值以构建各类别原型，使其能够处理更复杂的任务[155]。但其中原型的构建受到样本数据的约束，且获得的原型仅为特征级别，使得解释效果较为受限。在当前阶段，学者们进一步地将随机初始化的可学习向量视为类别原型，通过神经网络的学习能力对原型向量进行优化，并使用解码后的原型来解释图像的类别原型[49] 和语义原型[50]。然而，上述方法主要集中在计算机视觉领域，而视觉图像和振动信号在解释形式和可理解性上存在显著差异，无法直接应用于智能诊断领域。

近年来，原型匹配被证明在小样本学习中有效[155]，并被引入故障诊断领域以解决样本有限的问题[156,157]。凭借其出色的性能，原型匹配被拓展至更广泛的复杂问题，

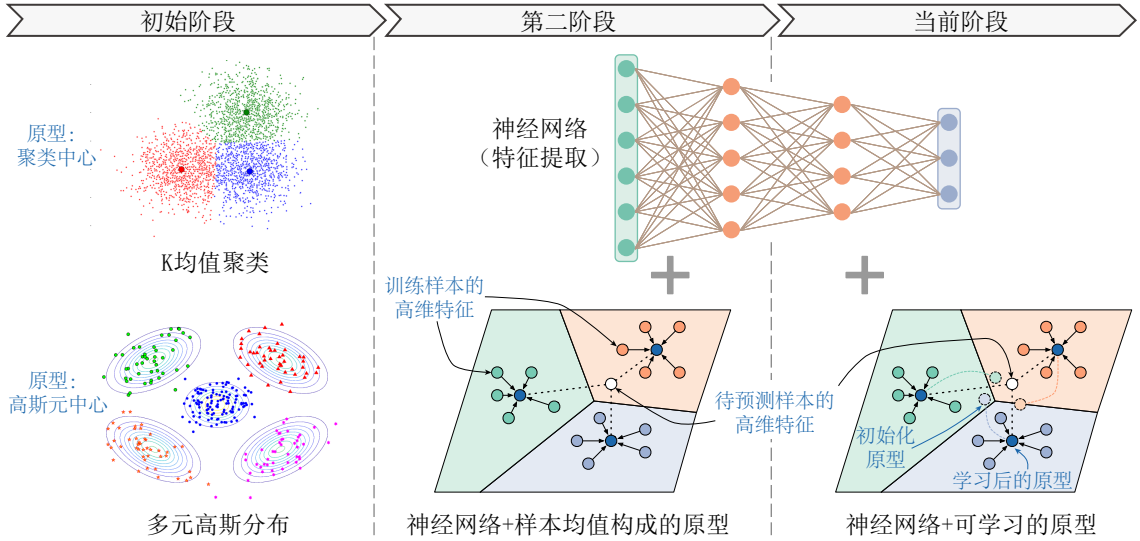


图 3-2　原型匹配应用的三个发展阶段

Fig. 3-2　The three stages of prototype-matching application

包括半监督学习[158-160]、迁移学习[161-168]、联邦学习[159,169] 等方面。原型匹配应用在智能诊断领域的研究工作如表 3-1 所示。在小样本学习方面，Li 等[156] 结合多尺度特征提取和原型聚类用于行星齿轮箱故障诊断。在迁移学习方面，Zhang 等[170] 提出了一种原型对比学习模块和原型校准策略，用于多源域适应，能够以更少的源标签成功实现优异的域对齐结果。在联邦学习方面，Wang 等[169] 将联邦对比学习与原型匹配相结合，以减轻客户端和服务器端的数据差异，实现联邦学习场景下的故障诊断。尽管在许多方面进行了大量关于原型匹配的研究，但其在解释性方面的巨大潜力仍尚待发掘。

表 3-1　原型匹配在故障诊断领域的应用文献

Table 3-1　The literature review of prototype matching in fault diagnosis

原型匹配在故障诊断中的应用目的	研究工作
小样本学习	[156,157]
小样本学习 + 半监督学习	[158-160]
小样本学习 + 封闭集迁移学习	[161-165,170]
小样本学习 + 开集迁移学习	[166-168]
小样本学习 + 联邦学习	[159,169]
小样本学习 + 原型漂移抑制	[158,160,164]
其他方面（元学习，抗噪声）	[157,163]
可解释性	原型匹配网络

3.2.2　用于特征提取和信息降维的神经网络自编码器

自编码器（AutoEncoder, AE）[171] 是一种无监督学习的神经网络模型，广泛应用于数据降维、特征提取和生成建模等领域。其核心思想是通过学习输入样本的隐含特征（编码过程），并利用这些特征对输入样本进行重构（解码过程）。自编码器在深度学习中具有重要的地位，尤其在处理高维数据时表现出色。

如图 3-3 所示，自编码器主要包括编码器和解码器两部分。编码器 f 负责将输入样本 $\boldsymbol{x}$ 通过非线性映射 f 压缩为潜在空间表征 $f(\boldsymbol{x})$，而解码器 g 则将潜在空间表征 $f(\boldsymbol{x})$ 重构回原始输入样本 $g\circ f(\boldsymbol{x})$。编码器和解码器通常是由多层神经网络构成，其参数通过损失函数的反向传播来优化。自编码器的目标是最小化原始输入样本 $\boldsymbol{x}$ 与重构样本 $g\circ f(\boldsymbol{x})$ 之间的差异，从而保证潜在空间表征 $f(\boldsymbol{x})$ 尽可能地包含输入样本 $\boldsymbol{x}$ 的完整信息。均方误差（Mean Squared Error, MSE）是自编码器训练中常用的损失函数，可表示为

$$\mathcal{L}(\boldsymbol{x}, g\circ f(\boldsymbol{x})) = \|\boldsymbol{x} - g\circ f(\boldsymbol{x})\|_2^2, \tag{3-4}$$

式中 $\mathcal{L}(\boldsymbol{x}, g\circ f(\boldsymbol{x}))$ 代表输入样本 $\boldsymbol{x}$ 与重构样本 $g\circ f(\boldsymbol{x})$ 之间的均方误差，$\|\cdot\|_2^2$ 代表 L_2 范数的平方，算子 $\circ$ 代表多个函数的串联，即 $g\circ f(\boldsymbol{x}) = g\big(f(\boldsymbol{x})\big)$。通过最小化损失函数，自编码器能够学习输入样本的有效表示，从而实现数据的降维和特征提取。

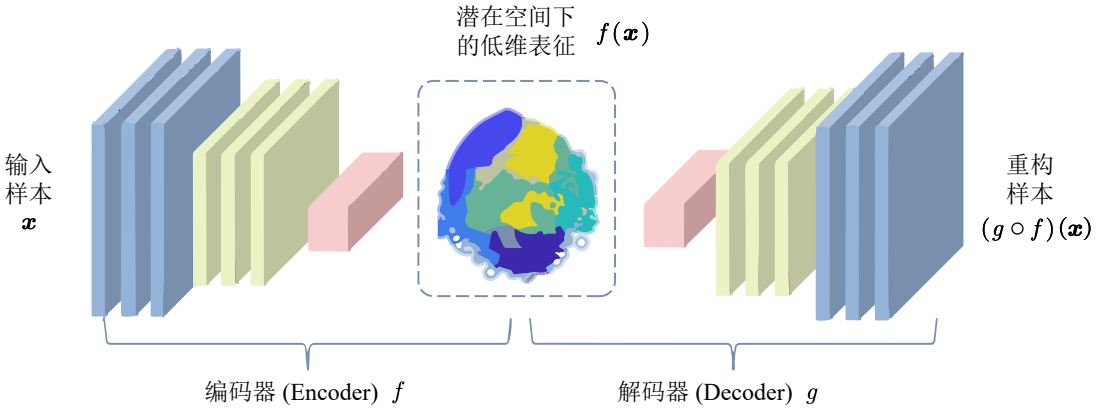


图 3-3　自编码器结构示意图

Fig. 3-3　The illustration of autoencoder achitecture

需要说明的是，潜在空间的维度通常远远小于样本空间，同时自编码器还尽可能地保证潜在空间表征 $f(\boldsymbol{x})$ 能够有效重构出原始样本 $\boldsymbol{x}$，这使得自编码器能够有效地提取输入样本的重要隐含特征。与传统的降维方法（如主成分分析）相比，自编码器能够捕捉数据中更复杂的非线性关系，因此在特征提取和降维任务中表现更为优越。

自编码器在机械设备故障诊断中具有重要的应用价值[172]。机械设备在运行过程中会产生大量高维且复杂的传感器数据，这些数据中往往包含噪声和冗余信息。自编码器凭借其强大的降维和特征提取能力，能够有效地将高维传感器数据压缩为低维潜在空间表征，从而去除噪声和冗余信息，保留与故障相关的关键特征。这些低维特征不仅提高了故障诊断模型的训练效率，还显著提升了故障诊断的准确性和可靠性，使得自编码器在智能故障诊断领域具有广泛应用，并展现出巨大的潜力。

3.3　基于原型匹配网络的决策层主动解释

3.3.1　原型匹配网络的结构设计

将训练数据集记为 $D^{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$，其中包括振动信号频谱构成的输入样本 $\boldsymbol{x} \in \mathbb{R}^p$，以及对应的故障类别标签 $y \in \{1, 2, \cdots, K\}$。所提出的原型匹配网络如图 3-4 所示，由三个部分组成：编码器 $f : \mathbb{R}^p \to \mathbb{R}^q$，解码器 $g : \mathbb{R}^q \to \mathbb{R}^p$，以及分类器 $h : \mathbb{R}^q \to \mathbb{R}^K$。与传统自编码器类似，编码器 f 将输入样本 $\boldsymbol{x}$ 压缩为低维编码特征 $f(\boldsymbol{x}) \in \mathbb{R}^q$，解码器 g 则将其恢复为重构样本 $(g \circ f)(\boldsymbol{x})$。通过编码和解码过程，编码特征 $f(\boldsymbol{x})$ 可以在低维潜在空间中获取输入样本 $\boldsymbol{x}$ 的关键信息。在此基础上，分类器 h 基于编码特征 $f(\boldsymbol{x})$ 进行进一步预测故障类别，最终获得故障诊断的分类结果。

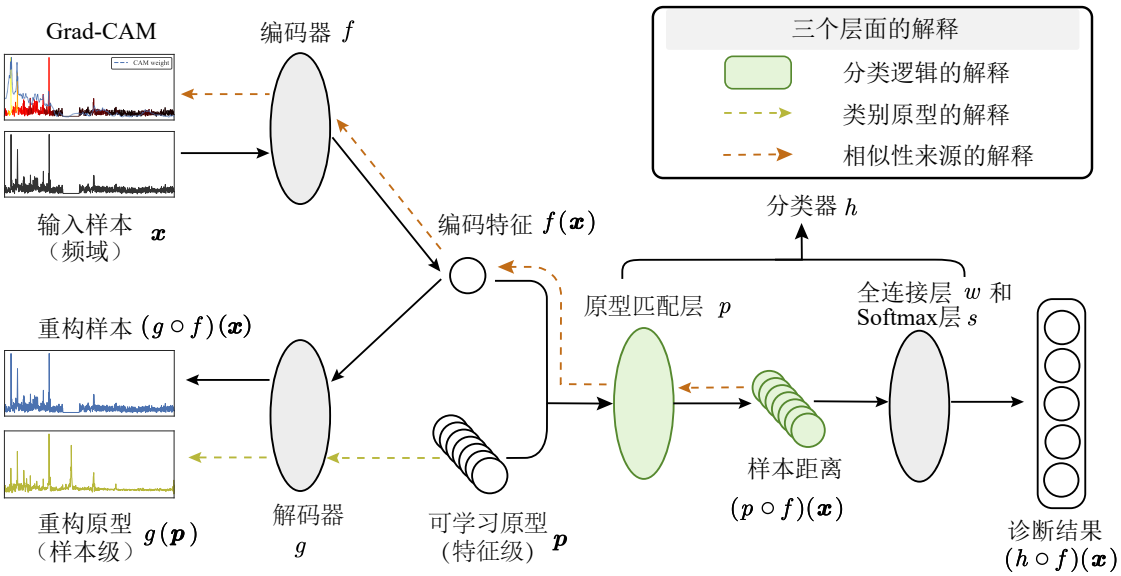


图 3-4　原型匹配网络的结构示意图

Fig. 3-4　The architecture of prototype-matching network

上述自编码器结构在故障诊断任务中并不少见[19,173]，而所提原型匹配网络的独

特之处在于分类器。该分类器包括一个原型匹配层（Prototype-Matching Layer, PML）、一个全连接层和一个 Softmax 层。具体而言，原型匹配层在编码器潜在空间中初始化 m 个向量 $\boldsymbol{p} \in \mathbb{R}^q$ 作为可学习的特征级原型，并计算输入样本的编码特征 $\boldsymbol{z} = f(\boldsymbol{x}_i)$ 与每个特征级原型向量之间的距离：

$$p(\boldsymbol{z}) = [d(\boldsymbol{z}, \boldsymbol{p}_1), d(\boldsymbol{z}, \boldsymbol{p}_2), \cdots, d(\boldsymbol{z}, \boldsymbol{p}_m)]^T \in \mathbb{R}^m, \tag{3-5}$$

式中 $d(\cdot,\cdot)$ 代表距离度量。根据文献[155] 和实际测试，距离度量采用平方欧氏距离 $d_{L_2}(x,y)^2 = \|x-y\|_2^2$ 能够比余弦距离 d_{cos} 和 L_1 距离 d_{L_1} 有着更好的表现。

在获得距离结果 $p(\boldsymbol{z})$ 后，全连接层将其映射到 K 维的故障类别空间，获得故障诊断的未规范化概率（logits）：

$$\boldsymbol{v} = \boldsymbol{W} p(\boldsymbol{z}) \in \mathbb{R}^K, \ \boldsymbol{W} \in \mathbb{R}^{K \times m}, \tag{3-6}$$

式中 $\boldsymbol{W}$ 代表全连接层的权重矩阵。为了使每个原型与特定的故障类别相绑定，并考虑距离与类别概率之间的负相关性，需要使 $\boldsymbol{W}$ 的每一列近似于一个负独热（one-hot）向量。因此，将权重矩阵初始化为 $\boldsymbol{W}_{i,j} = -\mathbb{I}\left(\mathrm{mod}(j,K) = i\right)$，当 $m = K$ 时，初始化的 W 等于负单位矩阵 $-I$。

Softmax 层将未规范化概率 $\boldsymbol{v}$ 归一化为 K 个故障类别的具体概率。第 k 个故障类别的概率 $s(\boldsymbol{v})_k$ 可以表示为

$$s(\boldsymbol{v})_k = \frac{\exp\left(\boldsymbol{v}_k\right)}{\sum_{k'=1}^{K} \exp\left(\boldsymbol{v}_{k'}\right)}. \tag{3-7}$$

所提出的 PMN 模型是自编码器和基于原型匹配层的分类器的有机融合，自编码器的降维能力为原型匹配提供了可行性，而原型匹配的聚合特性增强了自编码器的表征学习能力。本质上，图 3-4 所示的分类器是在低维的编码特征空间中基于距离进行分类，通过寻找与输入样本最近的原型作为预测输出，与混合密度估计（Mixture Density Estimation，MDE）算法相类似。原型数量 m 相当于 MDE 中的密度成分数量，当 $m = K$ 时，每个故障类别有一个原型，而当 $m > K$ 时，每个故障类别有多个原型。原型数量 m 的设置是原型匹配网络的一个重要超参数，由于自编码器具有足够的非线性映射能力，将原型数量设置为故障类别数据 $m = K$ 即可获得最优表现，这将在后续章节进行讨论与验证。

3.3.2　原型匹配网络的损失函数设计

原型匹配网络的训练目标包括准确性和可解释性两部分，准确性方面通过传统的分类损失 $\mathcal{L}_{\mathrm{cla}}$ 和自编码器的重构损失 $\mathcal{L}_{\mathrm{recon}}$ 来实现，解释性方面则通过 $\mathcal{R}_1$、$\mathcal{R}_2$ 和

$\mathcal{R}_3$ 这三个原型匹配距离损失项来实现，来鼓励每个样本找到一个足够接近的原型。

分类损失 $\mathcal{L}_{\text{cla}}$ 使用标准的交叉熵来对错误分类进行惩罚，可表示为

$$\mathcal{L}_{\text{cla}}(h \circ f, D) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{I}(y_i = k)\cdot \log\left[(h \circ f)(\boldsymbol{x}_i)\right]_k. \tag{3-8}$$

重构损失 L_{recon} 使用 MSE 来促使重构样本与输入样本保持一致，可表示为

$$\mathcal{L}_{\text{recon}}(g \circ f, D) = \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i - (g \circ f)(\boldsymbol{x}_i)\right\|_2^2. \tag{3-9}$$

在可解释性方面，原型匹配距离损失 $\mathcal{R}_1$ 和 $\mathcal{R}_2$ 以最小化编码特征 $f(\boldsymbol{x})$ 和原型向量 $\boldsymbol{p}$ 在各自视角下的互相最小距离为目标，其计算公式分别为

$$\mathcal{R}_1(\boldsymbol{p}, f, D) = \frac{1}{n}\sum_{i}^{n}\min_{j\in[1,m]} d\left(f(\boldsymbol{x}_i), \boldsymbol{p}_j\right), \tag{3-10}$$

$$\mathcal{R}_2(\boldsymbol{p}, f, D) = \frac{1}{m}\sum_{j}^{m}\min_{i\in[1,n]} d\left(f(\boldsymbol{x}_i), \boldsymbol{p}_j\right), \tag{3-11}$$

式中 $\mathcal{R}_1$ 从编码特征角度出发，促使编码特征接近任意一个原型，保证编码特征香原型向量靠拢。$\mathcal{R}_2$ 则从原型向量的角度出发，鼓励原型向量与特征空间中的至少一个样本紧密对齐，便于解码器处理后的原型重构。需要注意的是，$\mathcal{R}_2$ 需要在整个数据集上进行全局最小化，而这对于大规模数据集而言是不现实的。因此，在原型匹配网络的训练中，将式 (3-11) 中的全局最小化简化至局部的训练批次最小化，以保证模型训练的可行性。

$\mathcal{R}_1$ 和 $\mathcal{R}_2$ 保证了编码特征和原型向量之间的相互靠拢，但为了确保各原型向量的区分性，还需要引入 $\mathcal{R}_3$ 以最大化原型之间的互相最小距离：

$$\mathcal{R}_3(\boldsymbol{p}) = -\frac{1}{m}\sum_{i}^{m}\min_{j\in[1,m]} d\left(\boldsymbol{p}_i, \boldsymbol{p}_j\right). \tag{3-12}$$

将上述五个损失项相结合，原型匹配网络的完整损失函数 $\mathcal{L}$ 可以表示为

$$\begin{aligned}\mathcal{L}(g \circ h \circ f, \boldsymbol{p}, D) = \mathcal{L}_{\text{cla}}(h \circ f, D) + \lambda\mathcal{L}_{\text{recon}}(g \circ f, D)\\ + \lambda_1\mathcal{R}_1(\boldsymbol{p}, f, D) + \lambda_2\mathcal{R}_2(\boldsymbol{p}, f, D) + \lambda_3\mathcal{R}_3(\boldsymbol{p}),\end{aligned} \tag{3-13}$$

式中 λ、λ_1、λ_2、λ_3 是平衡不同损失比例的超参数，同原型数目 m 一样，会在后续的分析部分加以讨论。

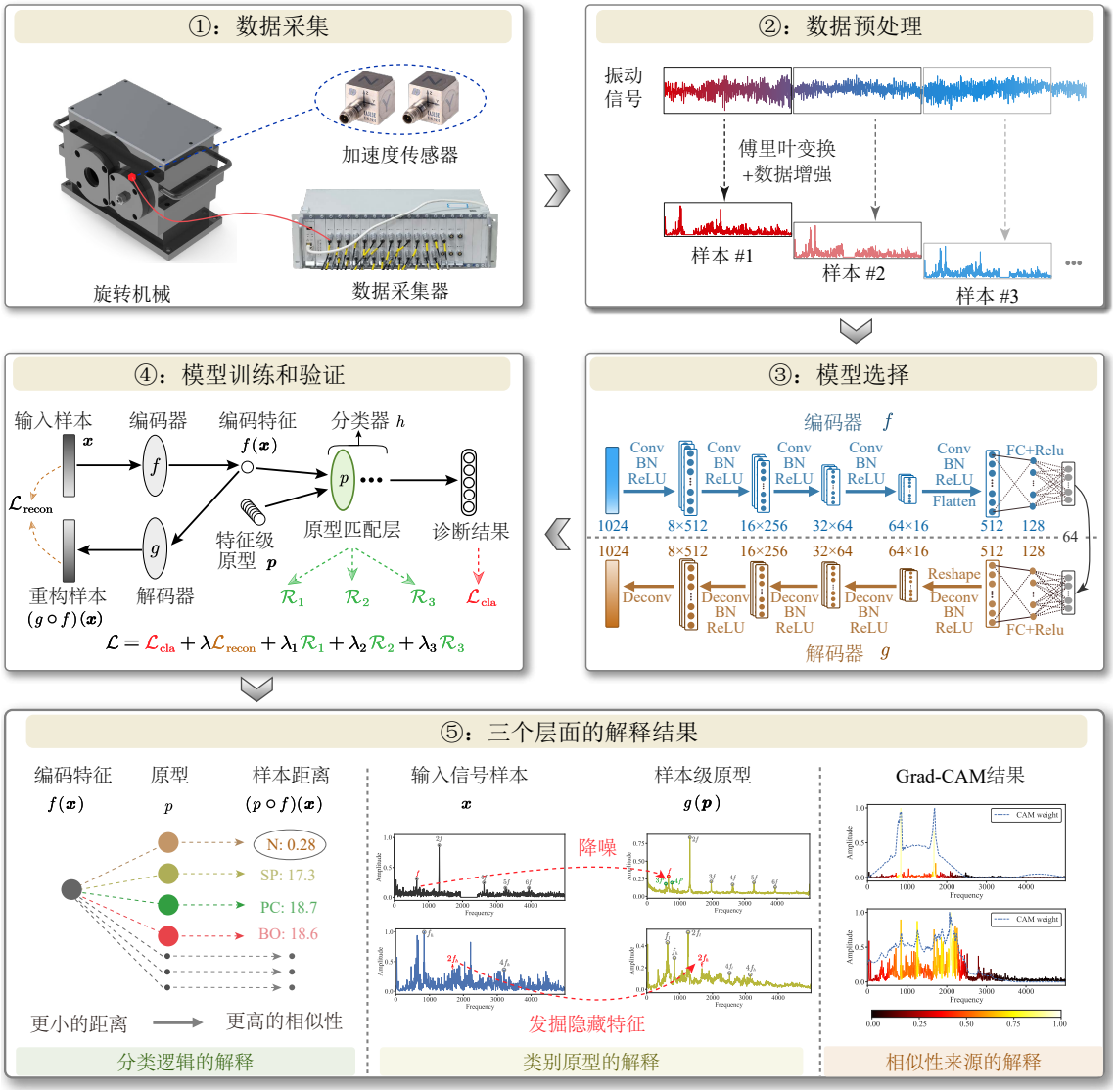


图 3-5　原型匹配网络应用于智能机械故障诊断的全过程

Fig. 3-5　The entire process of applying Prototype-Matching Network to intelligent mechanical fault diagnosis

3.3.3　原型匹配网络的三类解释层面及其故障诊断应用流程

原型匹配网络三个可解释性层面如图 3-4 所示。首先是分类逻辑可解释，通过原型匹配概念的融入，原型匹配层具有明确和清晰的分类逻辑。它将编码特征 $f(\boldsymbol{x})$ 与每个特征级原型 $\boldsymbol{p}$ 进行相似性比较，并选择最相似原型的故障类别作为预测结果。这种分类遵循人类固有的原型匹配逻辑，使得黑箱的神经网络在决策方面更为透明，成为部分可理解的灰箱模型。

其次是类别原型可解释，一方面，原型匹配层能够显式地构建特征级原型向量 $\boldsymbol{p}$

表 3-2　实验中所采用的原型匹配网络架构

Table 3-2　The architecture of prototype-matching network used in the experiment

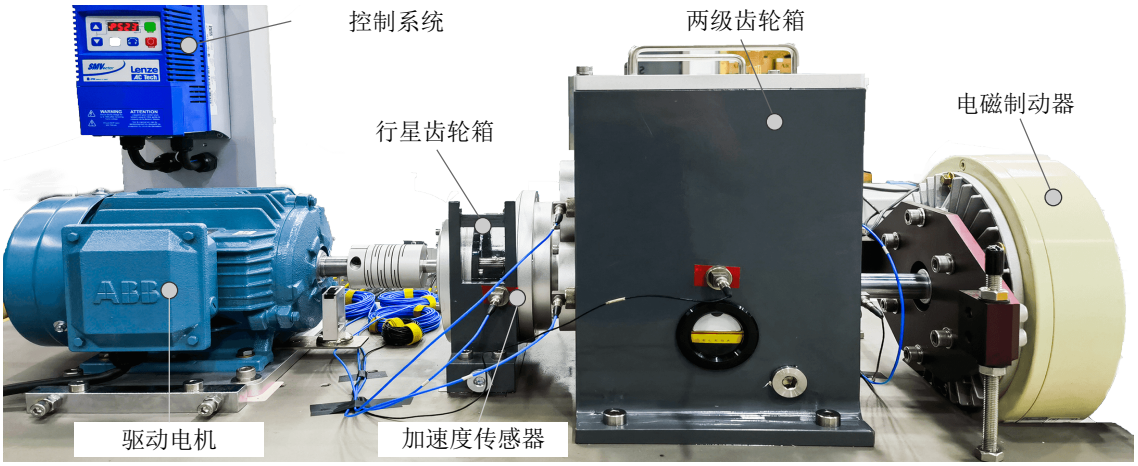
网络部分	序号	网络层参数	输出尺寸
编码器	-	Input	1×1024
	1	Conv(9@2@4)[a] -BN-ReLU	8×512
	2	Conv(9@2@4)-BN-ReLU	16×256
	3	Conv(11@4@5)-BN-ReLU	32×64
	4	Conv(11@4@5)-BN-ReLU	64×16
	5	Conv(11@4@5)-BN-ReLU-Flatten	128×4
	6	FC(128)-ReLU-FC(64)	64
解码器	1	FC(128)-ReLU-FC(512)-Reshape	128×4
	2	Deconv(10@4@3)-BN-ReLU	64×16
	3	Deconv(10@4@3)-BN-ReLU	32×64
	4	Deconv(8@2@3)-BN-ReLU	16×256
	5	Deconv(8@2@3)-BN-ReLU	8×512
	6	Deconv(8@2@3)	1024
分类器	1	PML(m[b])-FC(K[c])	K

[a] *(x@y@z)*: x、y 和 z 分别表示卷积核大小、步长和样本边缘填充长度。
[b] m: 作为原型匹配层超参数的原型数量。
[c] K: 由数据集确定的故障类别数量。

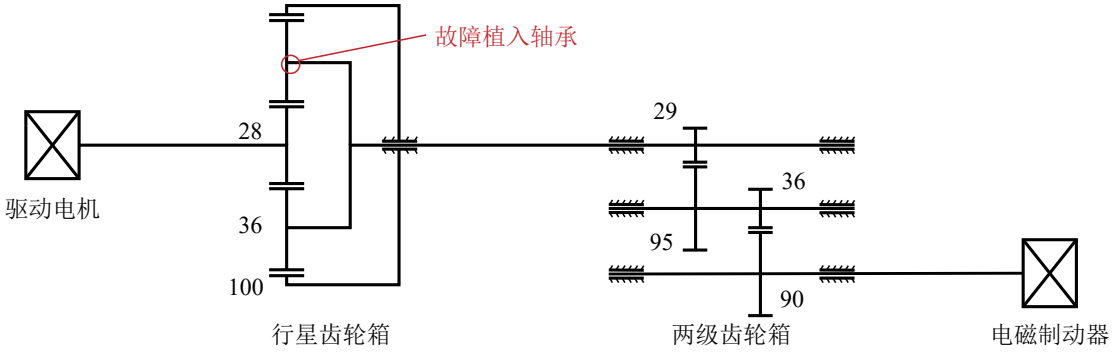
并通过训练过程进行优化；另一方面，借助于自编码器和 PM 层的有机组合，所学习的特征级原型向量 $\boldsymbol{p}$ 也能通过解码器 g 重构至样本域。重构出的样本级原型 $g(\boldsymbol{p})$，从模型视角描绘了典型故障信号，从而有效强化了故障机理特征。

最后是相似性来源可解释，通过引入额外的归因方法，对原型匹配层的距离结果进行贡献度溯源，从而获得输入样本各部分对高相似结果的贡献度，揭示出导致输入信号样本与匹配原型之间高相似性的关键故障相关频率。额外归因方法可由任务场景自由选择，实验中选择 Grad-CAM[71] 开展分析。

原型匹配网络应用于智能机械故障诊断的全过程如图 3-5 所示。首先，通过安装在旋转机械上的加速度传感器来采集振动信号，然后通过滑动窗截取、Fourier 变换、和数据增强来获得频域样本，并随机划分出训练集和测试集。其次，选择合适的自编码器网络来构建原型匹配网络，本章在实验全程采用如表 3-2 所示的、基于 CNN 的简单自编码器。随后，借助训练数据集 $\mathcal{D}^{\text{train}}$ 按照式 (3-13) 对原型匹配网络进行训练，并使用测试数据集 $\mathcal{D}^{\text{test}}$ 评估原型匹配网络的诊断性能。最后，在训练好的原型匹配



(a) 行星齿轮箱和两级齿轮箱复合试验台



(b) 试验台传动系统

图 3-6　复合齿轮箱数据集的试验台和传动系统示意图

Fig. 3-6　The experimental rig and transmission system of the comprehensive gearbox dataset

网络上进行可解释性分析，包括解释故障诊断分类逻辑、获取类别原型以描绘典型故障信号、以及从模型视角解释导致高相似性的关键故障频率来源。

3.4　决策层主动解释方法的故障诊断性能和解释效果实验验证

3.4.1　基于复合齿轮箱的传统故障诊断任务

复合齿轮箱数据集的实验装置如图 3-6 所示，包括行星齿轮箱和两级齿轮箱这两个传动系统，以及驱动电机、电磁制动器、加速度传感器、控制系统等组件。加速度传感器安装在行星齿轮箱的壳体，以 12 kHz 的采样频率采集振动信号，电动机的转速设定为 1800 rpm。

该数据集包括行星齿轮箱的六种齿轮故障和四种轴承故障。其中，齿轮故障为太阳轮点蚀（Sun Gear Pitting, SP）、太阳轮裂纹（Sun Gear Crack, SC）、太阳轮部分齿磨

(a) SP：太阳轮点蚀

(b) SC：太阳轮裂纹

(c) Sw：太阳轮部分齿磨损

(d) SW：太阳轮全齿磨损

(e) PC：行星轮裂纹

(f) PW：行星轮全齿磨损

(g) BI：轴承内圈故障

(h) BO：轴承外圈故障

图 3-7　复合齿轮箱数据集所考虑的故障部件

Fig. 3-7　The fault components considered in the comprehensive gearbox dataset

损（Sun Gear Partial Wear, Sw）、太阳轮全齿磨损（Sun Gear Full Wear, SW）、行星轮裂纹（Planetary Gear Crack, PC）和行星轮全齿磨损（Planetary Gear Full Wear, PW）。轴承故障包括轴承内圈故障（Bearing Inner Race Fault, BI）、轴承外圈故障（Bearing Outer Race Fault, BO）、轴承保持架故障（Bearing Cage Fault, BC）和轴承滚动体故障（Bearing Rolling Ball Fault, BB）。复合齿轮箱数据集的部分故障件如图 3-7 所示，轴承保持架故障和轴承滚动体故障由于故障部位难以拍摄，因此未展示在图中。考虑健康状态（Health, H），复合齿轮箱数据集可以看作是一个 11 类别的分类任务。

在数据准备过程中，使用滑动窗口对原始振动信号进行无重叠地截断，并随后转换为频域以生成输入样本。每个类别包含 190 个样本，每个样本的长度为 1024。随后，这些样本的 70% 被随机分配为训练集，其余样本则作为测试集。训练集和测试

集均通过 0-1 标准化进行预处理，其公式可表示为

$$\boldsymbol{x}_i = \frac{\boldsymbol{x}_i - \min(\boldsymbol{x}_i)}{\max(\boldsymbol{x}_i) - \min(\boldsymbol{x}_i)}. \tag{3-14}$$

为模拟实际工业环境中遇到的数据差异，本实验对训练集和测试集均进行数据增强。参照文献[113]，共选择随机加噪、随机缩放和随机掩码三种数据增强方法。其中，随机加噪可表示为

$$\boldsymbol{x}_i = \boldsymbol{x}_i + v \cdot 10 \cdot \mathcal{N}\left(0, \mathrm{std}(\boldsymbol{x}_i)\right), \tag{3-15}$$

随机缩放可表示为

$$\boldsymbol{x}_i = \mathcal{N}\left(1, v\right) \cdot \boldsymbol{x}_i, \tag{3-16}$$

随机掩码可表示为

$$\boldsymbol{x}_i = \mathrm{Mask}(\boldsymbol{x}_i, d), \tag{3-17}$$

式中 v 和 p 代表数据增强的超参数，$\mathcal{N}(a,b)$ 代表以 a 为均值、b 为方差的正态分布随机函数，$\mathrm{Mask}(\cdot)$ 函数代表将 $\boldsymbol{x}_i$ 的一段长度为 d 的片段设置为 0 的操作。所有三种增强的随机概率均设为 0.5。

在模型超参数方面，将原型匹配层的原型数量设置为与数据集的故障类别相等 $m = 11$，并通过网格搜索寻优方式将损失系数设置为 $(\lambda, \lambda_1, \lambda_2, \lambda_3) = (1, 0.25, 0.25, 0.01)$。训练周期设置为 50，批次大小为 128，优化器为 Adam，学习率为 0.001，衰减系数为每个周期 0.99。

除了传统的诊断准确性这一评估指标外，本章还引入了一个新的无量纲指标 R_{rps} 用以评估模型的表征学习能力。表征评估指标 R_{rps} 由类间距离 D_{inter} 和类内距离 D_{intra} 构建：

$$R_{\mathrm{rps}} = \frac{D_{\mathrm{intra}}}{D_{\mathrm{inter}}}. \tag{3-18}$$

将属于第 k 类的学习特征的均值向量表示为 $\overline{\boldsymbol{z}}_k$，类内距离 D_{intra} 通过计算每个样本到其对应类别质心的距离来衡量每个类别内的聚集性，其数学表达式为

$$D_{\mathrm{intra}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \left\|\boldsymbol{z}_i - \overline{\boldsymbol{z}}_k\right\|_2 \mathbb{I}(y_i = k). \tag{3-19}$$

类间距离 D_{inter} 则通过计算每个类别质心之间的距离来衡量不同类别之间的离散性，可表示为

$$D_{\mathrm{inter}} = \frac{1}{K(K-1)}\sum_{i=1}^{K}\sum_{j=1}^{K} \left\|\overline{\boldsymbol{z}}_i - \overline{\boldsymbol{z}}_j\right\|_2. \tag{3-20}$$

表 3-3　复合齿轮箱数据集下不同噪声参数的各模型故障诊断准确率结果（%）

Table 3-3　The fault diagnosis accuracy of different models on the comprehensive gearbox dataset under different noise parameters (%)

模型	实验噪声参数 (v-d)				
	0-0[a]	0.1-100	0.2-100	0.2-200	平均
CNN[174]	98.65	96.23	89.01	87.45	92.84
BiLSTM[174]	99.26	98.11	92.66	90.36	95.10
ResNet18[174]	99.63	97.68	92.74	92.83	95.72
Transformer[113]	**100.0**	98.72	94.23	91.33	96.07
DCAE[174]	**100.0**	98.89	**95.63**	94.63	97.29
ProtypicalNet[155]	99.50	98.31	94.66	93.45	96.48
CNN-PML[168]	99.52	98.76	94.57	93.60	96.61
原型匹配网络	99.98	**99.31**	95.41	**94.81**	**97.37**

[a] v-d：表示噪声增强的强度和掩码长度分别为 v 和 d。

R_{rps} 值越小越好，表明所提取的表征具有较小的类内距离和较大的类间距离，即该模型具有更强的表征学习能力。

在对比方法方面，本章选择七种典型的先进模型与原型匹配网络开展对比实验，这些模型包括 CNN[174]、BiLSTM[174]、ResNet18[174]、Transformer[113] 和 DCAE[174]，以及 PrototypicalNet[155] 和 CNN-PML[168]（CNN 基础上将分类层替换为原型匹配层）。上述七种模型包括基于原型匹配的方法（即 PrototypicalNet 和 CNN-PML），以及当前深度学习中使用的三种主流方法：卷积神经网络（CNN、ResNet、DCAE）、循环神经网络（BiLSTM）和注意力机制网络（Transformer）。此外，DCAE 也是原型匹配层的骨干自编码器，但使用多层感知机（Multi-Layer Perceptron, MLP）而非本章的原型匹配层作为分类器。

上述七种模型和原型匹配网络在复合齿轮箱数据集不同噪声参数下的诊断准确率如表 3-3 所示。总体而言，所提出的原型匹配网络在诊断性能和表征学习能力方面相较于其他七种模型表现出优异竞争力。具体来说，在无噪声条件下，所有八种模型几乎都能达到近乎 100% 的准确率。然而，随着噪声强度的增加，卷积神经网络（CNN、ResNet）、循环神经网络（BiLSTM）和注意力机制网络（Transformer）的准确率显著下降，在 0.2-200 噪声增强下仅能达到约 91%。显式融入原型匹配逻辑的 PrototypicalNet 和 CNN-PML 表现较好，但其准确率在 0.2-200 噪声增强下仍下降至约 93.5%。更进一步地，借助自编码器结构，DCAE 和原型匹配网络通过编码-解码过

程保留了提取过程的信息完整程度，更为鲁棒。随着噪声增加，它们的诊断准确率下降较少，在 0.2-200 噪声增强下约为 94.6%。此外，将分类器的 MLP 替换为原型匹配层的原型匹配网络，相较于基准的 DCAE 表现出略微提高的诊断准确率，同样的精度提高现象也发生在 CNN-PML 和 CNN 的准确率差异上，这有效证明了原型匹配层在提高模型诊断能力的作用。

复合齿轮箱数据集下不同噪声参数的各模型表征评估指标值 R_{rps} 如表 3-4 所示。借助自编码器强大的特征提取和将为能力，DCAE 和原型匹配网络的表征评估指标显著优于其他六种模型。但所提出的原型匹配网络在所有噪声增强条件下表现出色，其表征评估指标 R_{rps} 的平均值达到 0.270，是所有模型的最低值。这表明原型匹配网络通过善于特征提取的自编码器结果和显式的原型匹配分类逻辑，能够提取比其他七种模型更具类别区分度的故障特征，从而有效促进后续的故障分类诊断。

表 3-4　复合齿轮箱数据集下不同噪声参数的各模型表征评估指标值（R_{rps}）

Table 3-4　The representation metric values of different models on the comprehensive gearbox dataset under different noise parameters (R_{rps})

模型	实验噪声参数 (v-d)				
	0-0	**0.1-100**	**0.2-100**	**0.2-200**	**平均**
CNN[174]	0.792	0.912	1.201	1.323	1.057
BiLSTM[174]	0.631	0.863	1.256	1.365	1.028
ResNet18[174]	0.361	0.457	0.594	0.645	0.514
Transformer[113]	0.618	0.788	0.987	1.012	0.851
DCAE[174]	0.271	0.329	0.416	0.474	0.372
ProtypicalNet[155]	0.465	0.631	0.895	0.981	0.743
CNN-PML[168]	0.511	0.577	0.752	0.793	0.658
原型匹配网络	**0.195**	**0.225**	**0.316**	**0.345**	**0.270**

尽管表征评估指标 R_{rps} 能够直接反映模型表征学习能力，但将模型提取出的表征进行 t-SNE 可视化分析则更为直观。将噪声强度设置为 0.2-200，复合齿轮箱数据集下各模型表征的 t-SNE（t-distributed Stochastic Neighbor Embedding）可视化结果如图 3-8 所示，CNN、BiLSTM、ResNet18 和 Transformer 的表示过于模糊，难以区分其类别，而 DCAE 表现显著更好，与其在表 3-4 中的较低 R_{rps} 值相呼应。ProtypicalNet 和 CNN-PML 都具有较好的表征学习结果，各类别样本围绕在类别原型周围，显示出较高的可区分度。尽管前三种模型表现出色，但原型匹配网络的 t-SNE 结果最为优秀。借助自编码器结构和显式的原型匹配逻辑，原型匹配网络获得的样本表征紧密聚

集在原型向量周围，并且各原型向量保持足够距离，使得学习到的特征更具可分性。

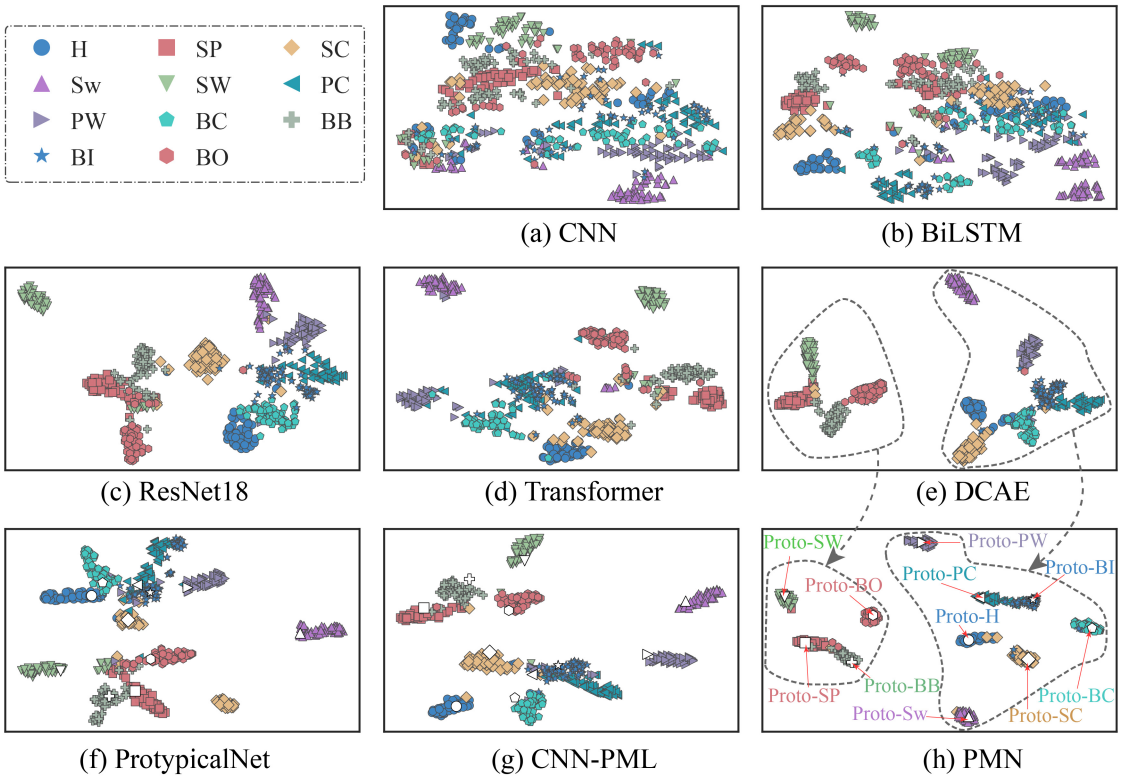


图 3-8　复合齿轮箱数据集噪声强度为 0.2-200 下各模型表征的 t-SNE 可视化结果

Fig. 3-8　The t-SNE visualization results of representation of different models on the comprehensive gearbox dataset under noise intensity of 0.2-200

诊断性能分析结果表明，通过引入原型匹配逻辑并结合自编码器结构，原型匹配网络在复合齿轮箱数据集上取得了最佳的诊断性能和表征学习能力。原型匹配网络在 0.2-200 噪声增强下的平均诊断准确率达到 97.37%，表征评估指标 R_{rps} 为 0.270，t-SNE 可视化结果表现最优。这表明原型匹配网络在智能机械故障诊断任务具有显著优势。

在上述的诊断性能分析之外，还需对原型匹配网络的可解释性进行深入分析。实验中，驱动电机的转速为 800 rpm，结合图 3-6(b) 所示的传动系统原理图，可以计算出故障所在的行星齿轮箱的啮合频率 f 和后续健康的两级齿轮箱的啮合频率 f' 分别为 f=656.25 Hz 和 f'=190 Hz。复合齿轮箱数据集噪声强度为 0.1-100 下原型匹配网络的解释结果如图 3-9 所示。其中，第一列是不同故障类别的输入样本频谱，用以展现数据信息。第二列是模型训练后并经解码器重构出的对应类别原型，用以从模型角度描绘典型故障信号。第三列是输入样本与不同类别原型的距离结果，用以解释基于相

似性的显式分类逻辑。最后一列是 Grad-CAM 结果，用以揭示导致输入信号样本与匹配原型之间高相似性的关键故障相关频率。

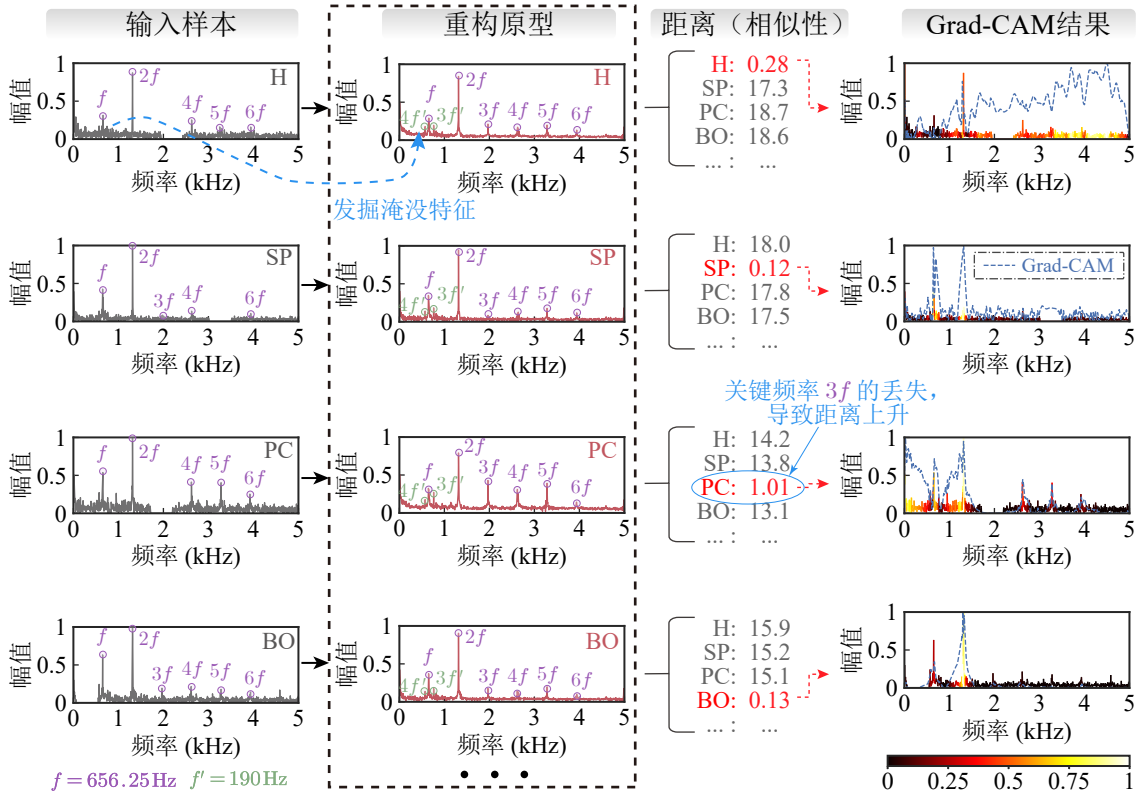


图 3-9　复合齿轮箱数据集噪声强度为 0.1-100 下原型匹配网络的三类解释结果

Fig. 3-9　The three types of interpretability results of Prototype-Matching Network on the comprehensive gearbox dataset under noise intensity of 0.1-100

类别原型可解释方面，重构原型是故障类别样本的典型概况，相比输入样本具有更清晰的故障机理特征。如图 3-9 所示，输入样本被数据增强的噪声所模糊，并难以辨认后续两级齿轮箱的啮合频率 f'，但重构的原型对噪声足够鲁棒，不仅有效降低噪声干扰，还将被淹没的关键特征 $3f'$ 和 $4f'$ 发掘出来。这种去噪和发掘淹没特征的能力，有助于提高模型的故障诊断准确性，并有助于进一步加深对故障机理的认识。

分类逻辑可解释方面，原型匹配网络在特征层面将输入样本的编码特征与每个原型向量进行匹配，并选择最相似原型（距离最小）的故障类别作为预测结果。此外，随机的 Mask(·) 数据增强操作可能会模糊关键频率，从而增大故障分类的难度。例如，第 3 行 PC 故障的 $3f$ 关键频率被掩盖掉，导致最近距离从其他样本的 0.2 左右上升到 1.01，该现象符合常理。

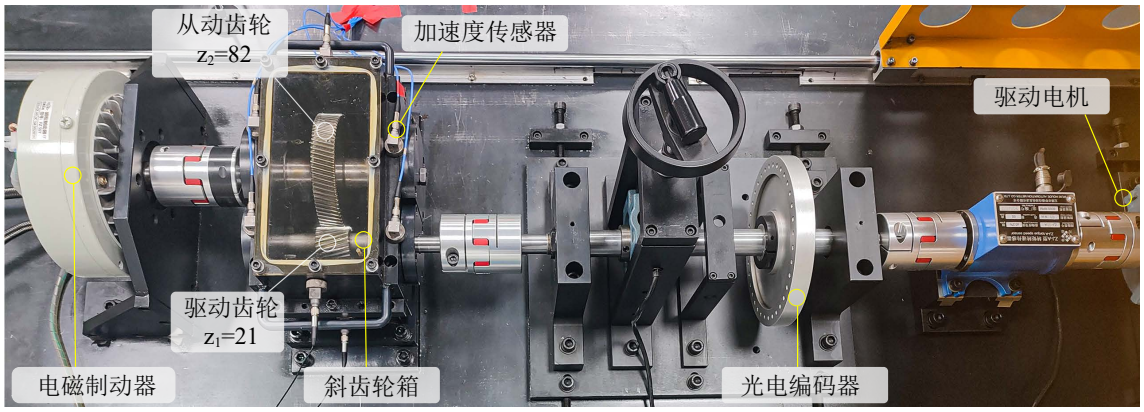
相似性来源可解释方面，Grad-CAM 结果揭示了导致输入信号样本与匹配原型之

间高相似性的关键故障相关频率。其中，行星齿轮箱的啮合频率及其谐波成分，特别是 f 和 $2f$，通常能够获得较高贡献。这表明行星齿轮箱的啮合频率是行星齿轮箱故障诊断中最关键的故障相关频率，这与行星齿轮箱故障的先验知识相一致。

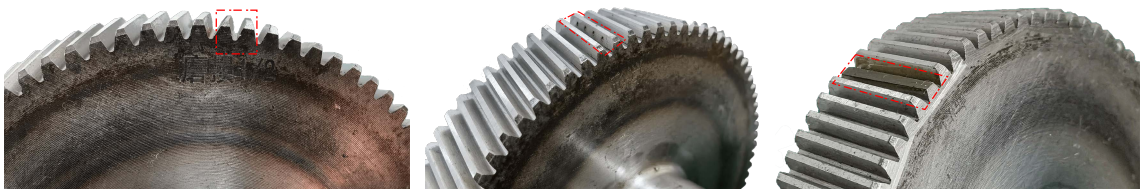
复合齿轮箱传统故障诊断任务下的可解释性分析，表明原型匹配网络在故障诊断任务中具有较强的可解释性。原型匹配网络通过重构原型、分类逻辑和相似性来源三个方面，有效地揭示分类逻辑和强化信号故障成分，有助于用户更好地理解模型的决策过程，提高故障诊断的可信度。

3.4.2　基于斜齿轮箱的领域泛化故障诊断任务

斜齿轮数据集的试验台和故障类型如图 3-10 所示，包括斜齿轮传动系统、驱动电机、电磁制动器等部件。其中，驱动齿轮和从动齿轮的齿数分别为 21 和 82。加速度传感器安装在斜齿轮箱的轴承端盖上，其采样频率设置为 10 kHz。数据集共考虑四种工况：D_{L0}、D_{L1}、D_{H0} 和 D_{H1}，其中下标 L 和 H 分别表示 1800 rpm 和 2400 rpm，下标 0 和 1 则分别表示空载和加载工况。数据集的故障类别包括：健康（Health, H）、从动齿轮表面磨损故障（Wear, W）、从动齿轮表面点蚀故障（Pitting, P）和从动齿轮断齿故障（Crack, C）。



(a) 斜齿轮箱试验台



(b) W：从动齿轮表面磨损　(c) P：从动齿轮表面点蚀　(d) C：从动齿轮断齿

图 3-10　斜齿轮数据集的试验台和故障部件

Fig. 3-10　The experimental rig and fault components of the helical gearbox dataset

不同负载、不同转速工况下收集的振动信号，具有不同的数据分布。但这种数据分布差异是由工况不同导致的，其潜在的故障本质仍然不变。一个具有良好领域泛化能力的故障诊断模型，能够降低甚至避免工况差异的干扰，捕捉故障类别的本质特征并实现高精度故障诊断。为了测试原型匹配网络的领域泛化能力，现将斜齿轮数据集的各个工况视为独立的域分布，从而制定了如表 3-5 所示的六个跨领域子任务。同样地，该实验中原型匹配网络的原型数目设置与故障类别数目相同 $m=4$，其他实验设置与3.4.1 小节的复合齿轮箱数据集实验保持一致。

表 3-5　斜齿轮数据集领域泛化场景的子任务设置

Table 3-5　The subtask setting of domain generalization experiment on the helical gearbox dataset

子任务	源域	目标域
T_1	D_{L1} & D_{H0} & D_{H1}	D_{L0}
T_2	D_{L0} & D_{H0} & D_{H1}	D_{L1}
T_3	D_{L0} & D_{L1} & D_{H1}	D_{H0}
T_4	D_{L0} & D_{L1} & D_{H0}	D_{H1}
T_5	D_{L0} & D_{H0}	D_{L1} & D_{H1}
T_6	D_{L1} & D_{H1}	D_{L0} & D_{H0}

斜齿轮数据集下不同领域泛化子任务的各模型故障诊断准确率结果如表 3-6 所示。相比于 3.4.1 小节的传统故障诊断任务，领域泛化故障诊断任务具有更高的诊断难度，没有任何单一方法能够在所有子任务中获得压倒性优势。CNN 在 T_3 表现最佳，BiLSTM 在 T_1 和 T_6 表现最佳，而 DCAE 在 T_5 中表现最佳。但所提出的原型匹配网络具有最优的整体表现，其平均准确率达到 85.39%，是所有方法中的最高值。

斜齿轮数据集下不同领域泛化子任务的各模型表征评估指标值 R_{rps} 如表 3-7 所示，实验结果和上一小节传统故障诊断任务的结果大体上保持一致。由于自编码器的特征提取能力和原型匹配逻辑的显式约束，原型匹配网络在表征学习能力方面优于所有其他模型。它的表征评估指标 R_{rps} 在所有子任务中都显示出压倒性优势，其平均值达到最低的 0.253。上述实验表明，原型匹配网络在领域泛化故障诊断任务中仍具有显著的诊断性能优势，平均诊断精度超过其他模型，且具有压倒性的表征学习优势。

斜齿轮箱数据集领域泛化子任务 T_4 下各模型表征的 t-SNE 可视化结果如图 3-11 所示。一个良好的跨域泛化模型，应该不受领域差异的影响，将工况不同而故障类型相同的样本进行聚类。然而，如图 3-11 (a)-(g) 所示的其他七种模型，受到领域差异的

表 3-6　斜齿轮数据集下不同领域泛化子任务的各模型故障诊断准确率结果（%）

Table 3-6　The fault diagnosis accuracy of different models on the helical gearbox dataset under different domain generalization subtasks (%)

模型	T_1	T_2	T_3	T_4	T_5	T_6	平均
CNN[174]	37.70	94.93	**92.68**	86.27	56.33	51.93	69.97
BiLSTM[174]	**86.62**	97.44	74.73	75.53	69.59	**70.97**	79.15
ResNet18[174]	57.52	80.63	65.23	74.70	66.31	51.53	65.99
Transformer[113]	70.43	80.46	83.42	74.92	83.67	57.02	74.99
DCAE[174]	80.85	99.83	75.05	76.73	**96.02**	61.51	81.66
ProtypicalNet[155]	53.98	54.26	79.78	68.56	53.91	60.60	61.85
CNN-PML[168]	48.91	52.77	78.78	65.17	50.14	59.81	59.26
原型匹配网络	75.35	**99.92**	75.00	**99.45**	95.45	67.19	**85.39**

表 3-7　斜齿轮数据集下不同领域泛化子任务的各模型表征评估指标值（R_{rps}）

Table 3-7　The representation metric values of different models on the helical gearbox dataset under different domain generalization subtasks (R_{rps})

模型	T_1	T_2	T_3	T_4	T_5	T_6	平均
CNN[174]	1.384	0.975	1.148	0.954	0.964	1.106	1.089
BiLSTM[174]	0.941	0.640	0.881	0.820	0.781	0.744	0.801
ResNet18[174]	0.602	0.361	0.544	0.698	0.376	0.538	0.520
Transformer[113]	0.840	0.649	0.480	0.548	0.721	0.513	0.625
DCAE[174]	0.538	0.319	0.518	0.239	0.215	0.571	0.400
ProtypicalNet[155]	0.619	0.469	0.573	0.556	0.514	0.560	0.548
CNN-PML[168]	0.745	0.632	0.566	0.659	0.508	0.590	0.617
原型匹配网络	**0.281**	**0.119**	**0.396**	**0.131**	**0.123**	**0.470**	**0.253**

严重影响，故障类型相同的样本由于工况差异不能很好地聚集，进而难以正确对故障进行分类。相反地，所提出的原型匹配网络通过显式原型匹配，在训练过程中积极鼓励相同类别的样本围绕原型聚集，在压缩源域分布的同时有效降低了工况差异的影响，进而也降低目标域和源域的表征距离。因此，如图 3-11 (h) 所示的原型匹配网络有效克服了领域差异，将来自不同领域的相同故障样本紧密聚集在一起，并使目标域的样本也更接近源域相应故障的样本表征簇。

在可解释性分析部分，首先需要确定斜齿轮数据集的特征频率。如图 3-10 所示，驱动齿轮的齿数为 $z_1 = 21$，则低速（1800 rpm）和高速（2400 rpm）工况下的齿轮啮

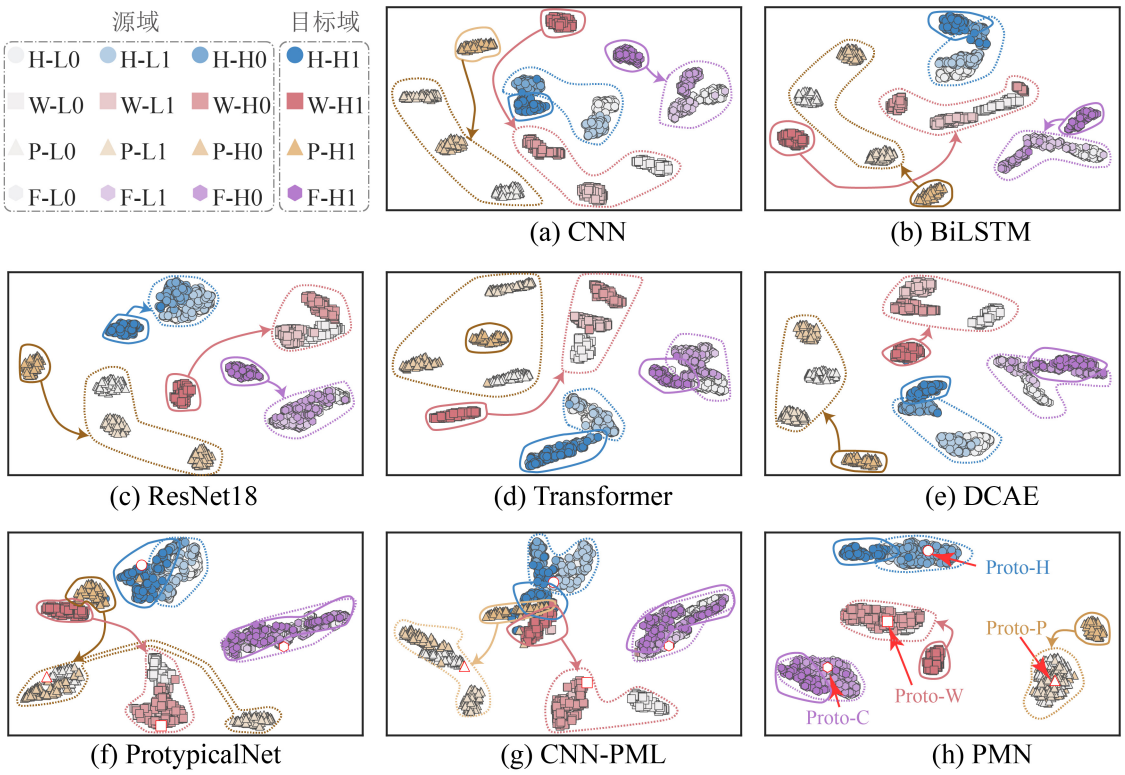


图 3-11　斜齿轮箱数据集领域泛化子任务 T_4 下各模型表征的 t-SNE 可视化结果

Fig. 3-11　The t-SNE visualization results of representation of different models on the bevel gearbox dataset under domain generalization subtask T_4

合频率可分别确定为 f_l= 630 Hz 和 f_h= 840 Hz。

斜齿轮箱数据集领域泛化子任务 T_4 下原型匹配网络的可解释性分析如图 3-12 所示，前两列分别是用于训练的源域样本和用于测试的目标域样本，总的解释结果基本与前一实验的结论大体一致。首先，通过引入显式的原型匹配概念，原型匹配层根据输入样本和类别原型的相似性，来实现故障预测。其次，相比于源域样本，重构原型的故障成分更为明显，具有降噪和发掘淹没特征的能力。最后，Grad-CAM 能够计算输入样本中每个频率对高相似性的贡献，揭示了啮合频率及其谐波（尤其是 f_h 和 $2f_h$）在故障识别中的关键作用。

此外，由于子任务 T_4 涉及三个源域（D_{L0}、D_{L1}、D_{H0}），由此重构的原型包含了 f_l 和 f_h 及其谐波分量。这表明从模型的角度来看，多源域场景下各类别典型故障信号是各个源域特征的现象组合。

作为神经网络视角下的典型故障信号解释，重构原型具有在高噪声背景下提取微弱关键故障频率的能力。具体而言，频率 $2f_h$ 在目标域 D_{H1} 故障分类中发挥关键

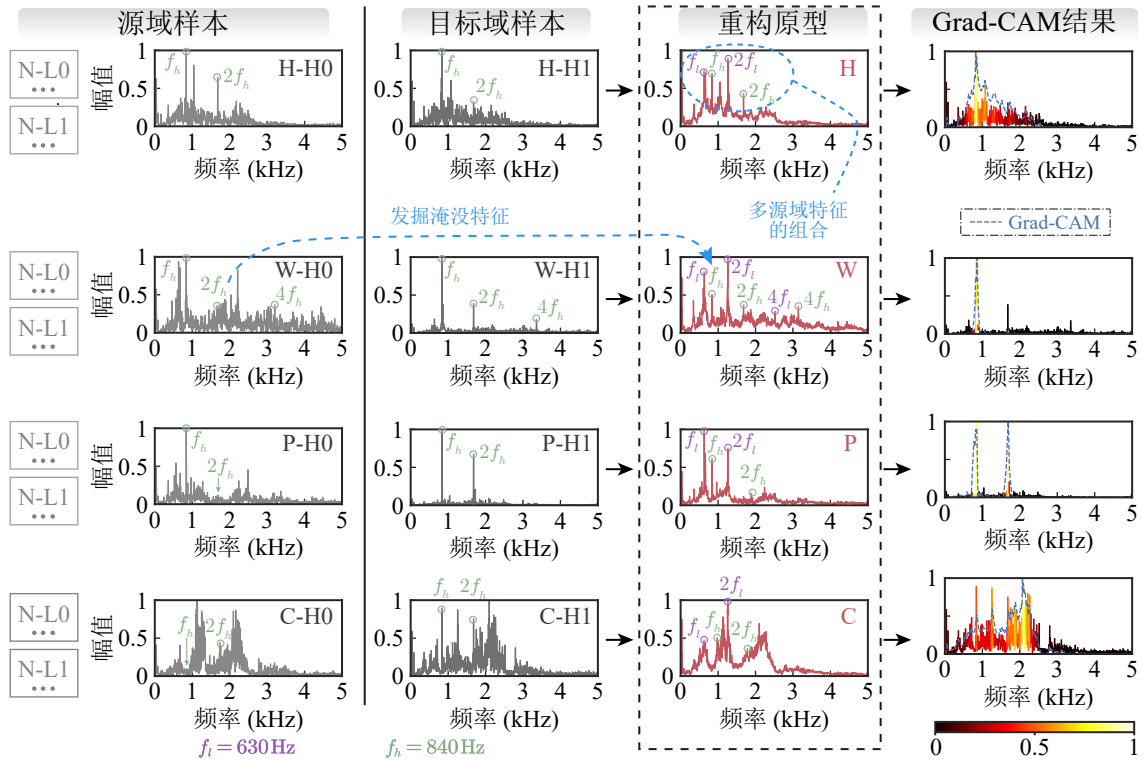


图 3-12　斜齿轮箱数据集领域泛化子任务 T_4 下原型匹配网络的三类解释结果

Fig. 3-12　The three types of interpretability results of Prototype-Matching Network on the bevel gearbox dataset under domain generalization subtask T_4

作用，但它们却在源域中被淹没（如图 3-12 的 W-H0 样本）或极其微弱（如图 3-12 的 P-H0 样本），而在基于源域训练出的原型匹配网络，却巧妙捕捉到这个关键频率 $2f_h$，使其在重构原型得到显著加强。原型匹配网络的可解释性不仅能够从模型角度描绘了典型故障信号，也为提取微弱故障成分提供了可行途径。

3.5　决策层主动解释方法的影响参数分析

原型匹配网络的性能优劣受到多个参数的影响，包括原型数目、原型匹配层距离度量和训练损失系数等。本节将分别对这些参数进行分析，以揭示原型匹配网络的性能优劣，更好地评估其在智能机械故障诊断任务中的适用性。

3.5.1　原型匹配层中距离度量和损失函数对诊断性能的影响

原型匹配层会根据式 (3-5) 计算各个样本编码特征与每个原型之间的距离，其中可以应用各种距离度量。此外，式 (3-13) 中的各项损失系数也需深入分析。因此，现开展实验以讨论不同距离度量和损失系数对模型诊断准确率的影响。实验设置与斜

齿轮箱数据集的领域泛化实验一致，包含如表 3-5 所示的六个领域泛化子任务。

值得注意的是，当使用平方 L_2 距离时，原型匹配层可等价于线性模型。原型匹配层中使用的平方欧氏距离度量 L_2 可表示为

$$d_{L_2}(f(\boldsymbol{x}), \boldsymbol{p}^k) = \| f(\boldsymbol{x}) - \boldsymbol{p}^k \|_2^2, \tag{3-21}$$

其展开形式为

$$\| f(\boldsymbol{x}) - \boldsymbol{p}^k \|_2^2 = f(\boldsymbol{x})^T f(\boldsymbol{x}) - 2\boldsymbol{p}_k^T f(\boldsymbol{x}) + \boldsymbol{p}_k^T \boldsymbol{p}^k. \tag{3-22}$$

第一项 $f(\boldsymbol{x})^T f(\boldsymbol{x})$ 相对于样本 $\boldsymbol{x}$ 是常数，不影响 Softmax 概率。剩余项可以表示为

$$-2\boldsymbol{p}_k^T f(\boldsymbol{x}) + \boldsymbol{p}_k^T \boldsymbol{p}^k = \mathbf{w}_k^T f(\boldsymbol{x}) + b_k, \tag{3-23}$$

式中 $\mathbf{w}_k = -2\boldsymbol{p}^k$，$b_k = \boldsymbol{p}_k^T \boldsymbol{p}^k$，上式是一个典型的线性模型。

斜齿轮箱数据集不同领域泛化子任务下不同距离度量和损失系数的原型匹配网络故障诊断准确率如图 3-13 所示，图中的“BaseAE”代表一个简化的、仅使用一个全连接层作为分类器的 DCAE 模型作为对照组。其余九个模型代表了使用三种不同距离度量（L_2、$L_{\cos}$ 和 L_1）和三组不同损失系数的原型匹配网络。其中 $\emptyset$ 表示没有额外损失，Λ_1 表示 $(\lambda_1, \lambda_2, \lambda_3) = (0.05, 0.05, 0.002)$，$\Lambda_2$ 表示 $(\lambda_1, \lambda_2, \lambda_3) = (0.25, 0.25, 0.01)$。

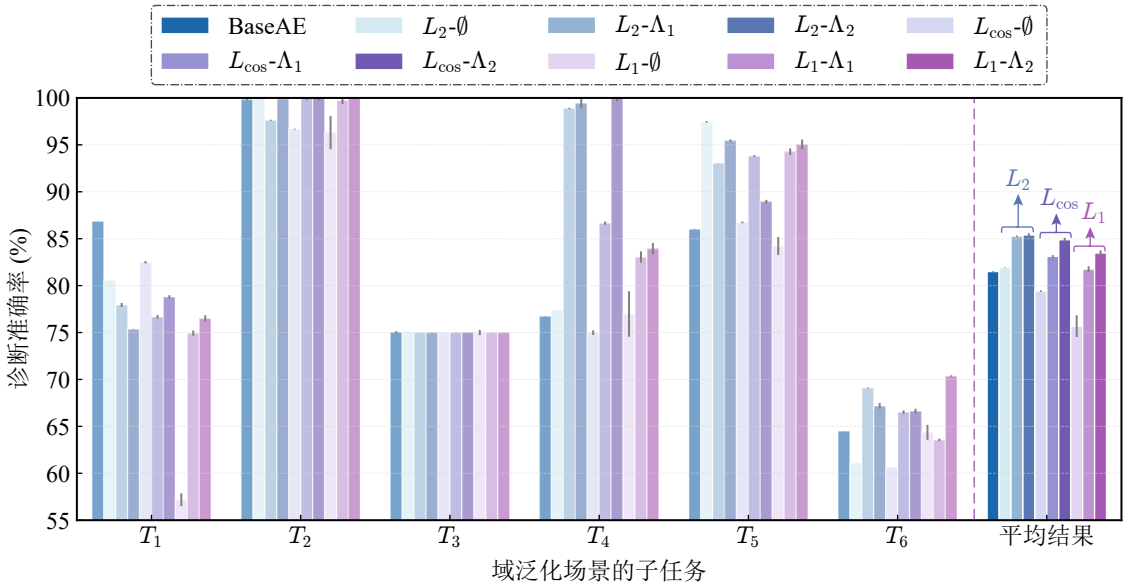


图 3-13　斜齿轮箱数据集不同领域泛化子任务下不同距离度量和损失系数的原型匹配网络故障诊断准确率

Fig. 3-13　The fault diagnosis accuracy of PMN with different distance metrics and loss coefficients on the helical gearbox dataset under different domain generalization subtasks

首先，就平均诊断准确率而言，距离度量 L_2 在相同损失系数下优于距离度量 L_1 和 $L_{\cos}$，由此式 (3-5) 中的距离度量建议采用 L_2。其次，当损失系数为 0 时，L_2 情况下的原型匹配网络和 BaseAE 具有相近的诊断准确率，这有效验证了具有 L_2 距离度量的原型匹配层与线性模型的等价性。此外，随着损失系数的增加，所有原型匹配网络的准确率都随之升高，这表明所提出的原型匹配距离损失（即 $\mathcal{R}_1$、$\mathcal{R}_2$ 和 $\mathcal{R}_3$）通过约束样本编码表征和类别原型的距离，能够有效提高模型诊断性能。

综上所述，具有 L_2 距离度量的原型匹配层等价于线性模型，但原型匹配层通过引入原型匹配距离损失来主动促进编码特征与其对应原型的相互聚集，从而超越了基础的线性模型，使得原型匹配网络获得具有竞争力的诊断性能和压倒性的表征学习优势。

3.5.2　原型匹配层中原型数量对诊断性能的影响

原型匹配层的本质是在低维的编码空间中基于距离进行分类，原型数量 m 相当于混合密度估计算法中的密度成分数量，是由具体任务决定的关键超参数。一个自然的问题便是原型数量 m 的最优设置，即需要对比每个类别对应单个原型向量亦或多个原型向量情况下的故障诊断表现。因此，本章设计了多领域故障诊断任务来探索原型数量 m 对原型匹配网络诊断性能的影响。

多领域故障诊断任务是指训练集和测试集均来自包含多个领域的相同分布，通过组合不同的领域能够控制训练数据的领域数量，进而全面探索原型数量 m 的更优参数设置。现基于斜齿轮箱数据集构建如表 3-8 所示多领领域故障诊断任务，其中包括七个子任务和四种不同的领域数量。

表 3-8　斜齿轮数据集多域故障诊断场景的子任务设置

Table 3-8　The task setting of multi-domain diagnostic experiment on the helical gearbox dataset

子任务	包含的具体域	域数量
Q_1	L_0	1
Q_2	H_0	1
Q_3	L_0 & H_0	2
Q_4	L_1 & H_1	2
Q_5	L_0 & H_0 & L_1	3
Q_6	L_0 & H_0 & H_1	3
Q_7	L_0 & L_1 & H_0 & H_1	4

实验中将原型匹配网络的原型数量 m 设置为故障类别数目的倍数（4-8-12-16），

并通过故障诊断的准确率进行对比。这些模型记为 PMN-x，其中 x 表示原型数量。不同原型数量的原型匹配网络故障诊断准确率结果如图 3-14 所示。

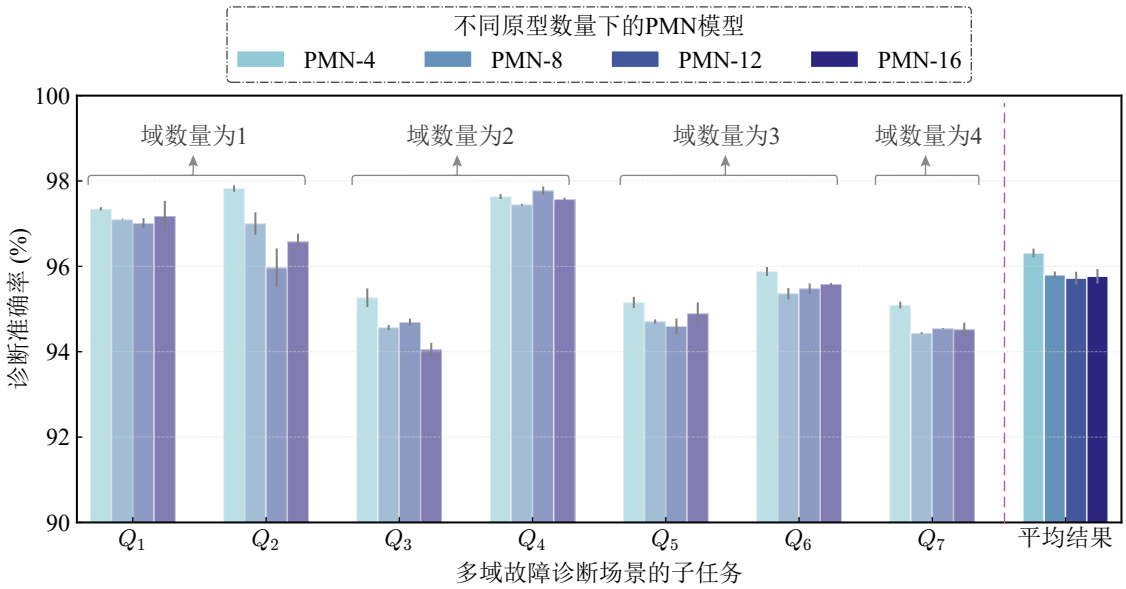


图 3-14　斜齿轮数据集多域故障诊断场景下不同原型数量的原型匹配网络故障诊断准确率

Fig. 3-14　The fault diagnosis accuracy of PMN with different prototype numbers on the helical gearbox dataset under multi-domain diagnostic scenario

由图可知，PMN-8、PMN-12 和 PMN-16 表现出相似的诊断准确率，但均略微落后于 PMN-4。PMN-4 不仅在单域子任务 Q_1 中表现出明显优势，在其他子任务中也展现出显著优势，如域数量为 2 的子任务 Q_3 和域数量为 4 的子任务 Q_7 。不同于应用于样本层面的混合密度估计，原型匹配网络是自编码器和原型匹配的结合，其中的自编码器能够在编码过程中将来自不同域的故障样本非线性映射到相近的低维编码表征。因此，各故障类别仅对应单个原型的 PMN-4（即 $m = K$）时便能实现最优的分类准确率，无需像混合密度估计那样根据数据集特征调整密度成分数量。总结而言，借助于自编码器强大的非线性映射能力，将原型数量 m 设置为与故障类别数量 K 相同，便足以使原型匹配网络获得最佳诊断性能。

3.6　本章小结

针对旋转机械智能诊断中主动解释效果、诊断性能及可拓展性的多方共赢问题，本章聚焦智能诊断模型决策层，将原型匹配概念融入决策层，并与自编码器模型有机结合以构建原型匹配网络，在提高模型诊断能力的同时，解释模型的分类逻辑以及模

型视角下的典型故障样本。通过对三个实测数据集的系统实验验证，本章的主要内容可总结如下：

(1) 提出了基于原型匹配概念的智能诊断模型决策层主动解释方法。原型匹配层通过显式构建可学习的各类别原型向量，并基于输入样本与各原型间的距离相似度进行故障分类。所设计的原型匹配网络将原型匹配层与自编码器有机融合，其中自编码器的降维能力为原型匹配提供基础，而原型匹配又反向促进自编码器表征学习能力的提升。在传统的分类损失、重构损失基础上，本章引入三项原型匹配距离损失，通过协同作用促进原型匹配网络更加优化的收敛。

(2) 梳理了原型匹配网络所具有的三方面解释结果。所提原型匹配网络具备三个维度的可解释性：在分类逻辑方面，原型匹配层通过显式的相似性距离实现故障分类，使决策逻辑直观清晰；在类别原型方面，学习得到的原型向量能够通过解码器重构至样本域，有效刻画模型视角下的典型故障样本特征；在相似性来源方面，通过额外的归因方法揭示导致原型匹配高相似性的关键频率贡献成分。

(3) 传统故障诊断和领域泛化的两类任务表明，所提原型匹配网络在诊断性能和表征学习能力上均优于现有方法，展现了优异的诊断能力。在可解释性方面，重构出的原型样本呈现出更为明显的故障特征成分，能够有效抑制噪声干扰并发掘被淹没的微弱故障频率特征，为故障机理认知提供了新的视角和途径。在原型匹配层的参数影响研究中，三类原型匹配距离损失能够通过约束样本和原型距离有效提升模型诊断性能，L_2 距离度量的效果优于其他距离度量方式，且受益于自编码器的强大非线性映射能力，将原型数量设置为与故障类别数相等（$m = K$）即可获得最佳诊断表现。

第四章　结合域变换的智能诊断模型被动解释形式优化

4.1　引言

前两章工作分别从时频变换和原型匹配两方面出发，主动构建可解释的特殊网络模块，从而为黑箱智能诊断网络的输入层和决策层赋予可解释性。但这两种主动解释方法均要求在训练前对模型架构进行针对性修改，虽然能够获得独特的解释能力，但同时也限制了网络结构。

与具有诸多约束的主动解释方法不同，被动解释方法无需参与模型构建与训练过程，仅针对已完成预训练的模型进行事后场景解释，从而在不牺牲性能的前提下保证模型的灵活性和可扩展性。然而，大多数被动解释方法的解释形式由模型的输入域决定，所获得的时域归因结果往往不够直观，难以有效传达机械故障机理信息。尽管可以通过数据前处理或归因结果后处理方式对时域解释进行优化，但这种方式要么破坏模型的端到端特性，要么因时域分析的本质局限而难以获得满意的解释效果。

传统故障诊断方法通常借助成熟的信号处理域转换技术，将时域信号转换至故障特征更为凸显的其他域（如频域、时频域等），从而更准确地判定故障类别。受此启发，部分学者将现有被动解释方法与域变换相结合，对被动解释的形式进行优化。例如，Gwak 等[133] 通过在频域中扰动样本来识别端到端模型的关键频率和决策边界；Herwig 等[135] 将 SHAP 扩展至频域和时频域，有效揭示了时域样本中不同谱频率 f_c 的贡献；进一步地，Decker 等[134] 将 SHAP 应用于包络域，从而将贡献度归属到更能体现轴承故障的调制频率 f_m 上。然而，这些方法仅从谱频率 f 或循环频率 α 的单一角度分析故障，可能导致片面甚至误导性的解释。事实上，不同故障可能具有共同的谱频率 f_c 或调制频率 f_m，使得这些相近的故障成分难以被上述方法有效区分。相比之下，循环域能够同时从谱频率 f 或循环频率 α 两个维度揭示故障成分，具有更强大的故障区分能力。

基于上述分析，本章将传统的 SHAP 解释扩展到更为清晰的循环域，建立面向旋转机械故障诊断模型的 CS-SHAP （Cyclic Spectral - SHapley Additive exPlanations）被动解释方法，从而有效优化智能诊断的被动解释形式。具体而言，CS-SHAP 在不改变端到端架构的情况下，将模型诊断结果归因到不同载波频率 f_c 和调制频率 f_m 的信号成分之上。借助循环域变换，CS-SHAP 的归因解释能够有效区分紧邻故障成分，更符合旋转机械故障机制，从而提供更清晰和准确的解释结果。此外，与传统 SHAP

一致，CS-SHAP 作为与模型无关的被动解释方法，可对任意智能诊断模型进行事后场景分析，具有更广泛的应用前景。

本章首先介绍 SHAP 和循环谱相关算法作为 CS-SHAP 的理论基础。考虑到循环谱相关仅适用于随机信号，本章随后针对确定性信号的循环域变换 $\mathcal{D}$ 及其对应的逆变换 $\mathcal{D}^{-1}$ 进行理论推导。然后，详细阐述 CS-SHAP 算法的具体实现以及将其应用于旋转机械故障诊断模型事后解释的完整流程。最后，在开源数据集和实测数据集上深入开展实验，以验证 CS-SHAP 在故障诊断领域的优异解释效果，并对相关参数的影响进行系统分析。本章的方法代码已开源在 *https://github.com/ChenQian0618/CS-SHAP*。

4.2　SHAP 被动解释方法和循环谱相关分析

4.2.1　面向机器学习模型被动解释的 SHAP

SHAP（SHapley Additive exPlanations）[175] 是一种受博弈论启发的方法，旨在解释机器学习模型的预测结果。它通过衡量样本中每个特征对预测结果的贡献来实现归因。在介绍 SHAP 之前需要首先介绍 Shapley 值[78]，这是一种用来确定团体内公平有效的资源分配策略的数学方法，其应用场景包括股东分配利润、合作者分配成本和功劳等。

在博弈论中，将玩家记为 p，玩家构成的集合记为 $\boldsymbol{S}$，集合的价值函数记为 $v:\boldsymbol{S}\to\mathbb{R}$，则用来衡量玩家 p 加入集合 $\boldsymbol{S}$ 作用的边际贡献 $\Delta v(p,\boldsymbol{S})$ 可以记为

$$\Delta v(p,\boldsymbol{S})=v(\boldsymbol{S}\cup\{p\})-v(\boldsymbol{S}). \tag{4-1}$$

进而，Shapley 值 $\psi_{v,\boldsymbol{U}}(p_i)$，用以衡量玩家 p_i 在全集 $\boldsymbol{U}=\{p_1,p_2,\cdots,p_n\}$ 的所有可能子集 $\boldsymbol{S}$ 中的期望边际贡献，可以表示为

$$\psi_{v,\boldsymbol{U}}(p_i)=\sum_{\boldsymbol{S}\subseteq\boldsymbol{U}\setminus\{p_i\}}\frac{s!(n-s-1)!}{n!}\cdot\Delta v(p_i,\boldsymbol{S}), \tag{4-2}$$

式中 s 表示子集 $\boldsymbol{S}$ 的元素数量，权重 $s!(n-s-1)!\,/\,n!$ 表示玩家 p_i 加入子集 $\boldsymbol{S}$ 的概率。

虽然 Shapley 值能够有效衡量每个集合成员的贡献，但不能直接应用于机器学习模型。这是因为式 (4-1) 中价值函数 v 的输入是玩家构成的集合，其输入维度可随集合成员的数量而改变，但机器学习模型的输入维度是固定的。

为了解决这个问题，SHAP 将模型的所有特征维度视为全集 $\boldsymbol{U}=\{1,2,\cdots,d\}$，其中 d 是特征维度的数量。然后将子集内的特征固定为常量，而将子集外的特征视

为随机变量。从而通过计算这些随机变量分布的期望输出来构建价值函数 v。具体而言，对于给定的子集 $\boldsymbol{S}$，子集内的特征设置为输入样本 $\tilde{\boldsymbol{x}}$ 中的固定值，而子集外的特征从数据分布 $\boldsymbol{X}$ 中重新采样。组合出的随机变量 $\tilde{\boldsymbol{x}}^{\boldsymbol{S}}$ 可表示为

$$\tilde{\boldsymbol{x}}_i^{\boldsymbol{S}} = \begin{cases} \tilde{\boldsymbol{x}}_i\,(\text{常量}), & \text{如果 } i \in \boldsymbol{S} \\ \boldsymbol{X}_i\,(\text{变量}), & \text{如果 } i \notin \boldsymbol{S}. \end{cases} \tag{4-3}$$

将待分析的模型记为 $\mathcal{M}:\mathbb{R}^d \to \mathbb{R}$，SHAP 中的价值函数 $v_{\mathcal{M},\boldsymbol{X},\tilde{\boldsymbol{x}}}(S)$ 可以表示为

$$\begin{aligned} v_{\mathcal{M},\boldsymbol{X},\tilde{\boldsymbol{x}}}(S) &= \mathbb{E}_{\boldsymbol{X}}[\mathcal{M}(\tilde{\boldsymbol{x}}^{\boldsymbol{S}})] - \mathbb{E}_{\boldsymbol{X}}[\mathcal{M}(\boldsymbol{X})] \\ &= \int \mathcal{M}(\tilde{\boldsymbol{x}}^{\boldsymbol{S}})\,\mathrm{d}\mathbb{P}_{\boldsymbol{X}} - \int \mathcal{M}(\boldsymbol{X})\,\mathrm{d}\mathbb{P}_{\boldsymbol{X}}. \end{aligned} \tag{4-4}$$

上式中的价值函数 $v_{\mathcal{M},\boldsymbol{X},\tilde{\boldsymbol{x}}}(S)$ 可以理解为将子集 $\boldsymbol{S}$ 内的特征从数据分布 X 中采样的随机变量约束为样本 $\tilde{\boldsymbol{x}}$ 的固定值所引起的模型输出差异。然后，将 $v_{\mathcal{M},\boldsymbol{X},\tilde{\boldsymbol{x}}}(S)$ 代入式 (4-2)，SHAP 的解释结果 $\psi_{\mathcal{M},\boldsymbol{X}}(\tilde{\boldsymbol{x}})$ 可以计算为

$$\psi_{\mathcal{M},\boldsymbol{X}}(\tilde{\boldsymbol{x}})_i = \sum_{\boldsymbol{S}\subseteq \boldsymbol{U}\setminus\{i\}} \frac{s!(d-s-1)!}{d!} \left(\mathbb{E}[\mathcal{M}(\tilde{\boldsymbol{x}}_{\boldsymbol{S}\cup\{i\}})] - \mathbb{E}[\mathcal{M}(\tilde{\boldsymbol{x}}_{\boldsymbol{S}})]\right). \tag{4-5}$$

需要注意的是，SHAP 的计算需要枚举全集 $\boldsymbol{U}$ 的特征子集 $\boldsymbol{S}$ 并估计数据分布上的期望 $\mathbb{E}[\mathcal{M}(\boldsymbol{x})]$，这使得实现过程具有较大的计算复杂度。为了解决这个问题，Lundberg 等[175] 开发了一个开源的 Python 库[176]，大大简化 SHAP 的实际应用难度。

4.2.2　面向随机信号的循环谱相关分析

域转换分析是信号处理中常用的技术，用于将信号从时域转换到其他域，以揭示信号的隐含特征，常见的域包括频域、时频域、包络域等。为了更直观地理解不同域的差别，现通过图 4-1 展示故障信号在不同域下故障特征。如图 4-1(a) 所示，当旋转机械发生故障时，故障部位会在每个旋转周期中参与啮合，进而产生周期性的时域脉冲响应，形成如图 4-1(b) 所示的时域信号。其中，故障诊断包含两种关键信息：反映机械结构固有特性的载波频率 f_c 和反映冲击激励周期属性的调制频率 f_m。

然而，图 4-1(b) 所示的时域信号通常仅能揭示冲击时刻、响应波形等故障特征，且极易受到噪声干扰。因此，借助域转换技术能够为故障分析提供更清晰的故障表征。具体而言，图 4-1(c) 所示的频域能将故障信息体现为频谱边频带，能够有效揭示载波频率 f_c 并间接地反映调制频率 f_m。图 4-1(d) 所示的包络域，将故障特征表现为更明显的调制频率 f_m 及其谐波，使故障依据更准确地定位到特定的调制频率 f_m。图 4-1(e) 所示的时频域，则将故障特征体现为周期性脉冲，同时从载波频率 f_c 和冲

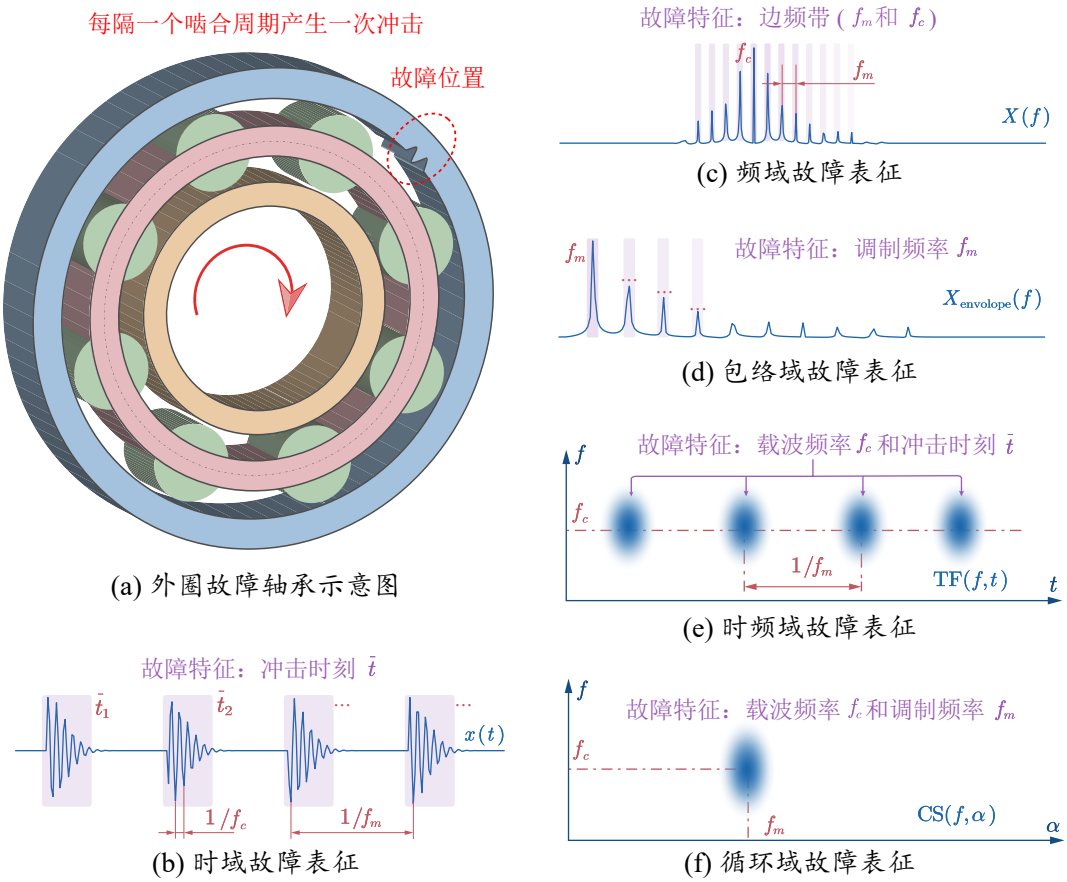


图 4-1　外圈轴承故障的示意图及对应信号在不同域下的故障特征

Fig. 4-1　The schematic diagram of the relationship between rotating machinery fault and vibration signal, and the characteristics and attribution results of vibration signal in different domains

击时刻两方面对故障进行刻画，提供故障诊断依据。相比之下，如图 4-1(f) 所示的循环域能够同时从载波频率 f_c 和调制频率 f_m 表现故障成分，更符合图 4-1(a) 中从故障冲击到信号激发的机理过程，具有更强大的故障区分能力。

时域到循环域的转换，源于循环谱相关（Cyclic-Spectral Correlation, CSC）这一信号处理技术，它专门用于分析循环平稳信号[177]。与具有恒定统计特性的平稳信号不同，N 阶循环平稳信号的 N 阶统计量具有随时间周期性变化的特性。

以二阶循环平稳信号为例，其自相关函数具有周期性，如被周期信号调制的白噪声。这类信号通常表示系统对周期性激励的响应。在旋转机械中，转子在旋转过程中会周期性地与故障部件接触，从而激发周期性的系统响应。由此产生的振动信号是典型的二阶循环平稳信号。这种循环平稳信号同时包含了两类关键信息：周期性激励频率 f_c 和系统响应频率 f_m。尽管传统的频谱分析对这种信号效果不佳，但循环谱相关

能够有效揭示其隐含特征。

对于循环平稳信号 $x(t)$，其二维自相关函数（$R_x(\tau,t)$ 可以表示为

$$R_x(\tau,t)=\mathbb{E}[x(t-\tau/2)x^*(t+\tau/2)], \tag{4-6}$$

式中 $*$ 代表共轭运算符，τ 代表时延，$\mathbb{E}$ 代表统计期望。它代表随机信号 $x(t)$ 在特定时间 t 下的时域特性。ACF $R_x(\tau,t)$ 具有 t 和 τ 两个维度，t 轴代表全局时间轴，而 τ 轴代表局部时域特征。

通过对 ACF 的时间轴 t 施加 Fourier 变换，可以得到循环自相关函数（Cyclic Autocorrelation Function, CAF）$R_x(\tau,\alpha)$：

$$\begin{aligned} R_x(\tau,\alpha)&=\int R_x(\tau,t)\mathrm{e}^{-\mathrm{i}2\pi\alpha t}\,\mathrm{d}t \\ &\approx \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2}x(t-\tau/2)x^*(t+\tau/2)\mathrm{e}^{-\mathrm{i}2\pi\alpha t}\,\mathrm{d}t, \end{aligned} \tag{4-7}$$

式中 α 代表循环频率（Cyclic frequency）。$R_x(\tau,\alpha)$ 揭示了随机信号 $x(t)$ 在不同循环频率 α 下的时域特性，有效反映了激励的频率信息。

进一步地，对循环自相关函数 $R_x(\tau,\alpha)$ 的时延轴 τ 进行 Fourier 变换，可以得到循环谱相关：

$$\begin{aligned} S_x(f,\alpha)&=\int R_x(\tau,\alpha)\mathrm{e}^{-\mathrm{i}2\pi f\tau}\,\mathrm{d}\tau \\ &=\int\int R_x(t,\tau)\mathrm{e}^{-\mathrm{i}2\pi(f\tau+\alpha t)}\,\mathrm{d}t\,\mathrm{d}\tau, \end{aligned} \tag{4-8}$$

式中 f 代表谱频率（Spectral frequency）。从本质上讲，循环谱相关可以视为二维自相关函数 $R_x(\tau,t)$ 在时间 t 和时延 τ 两个轴上的二维 Fourier 变换，从而获得的关于谱频率 f 和循环频率 α 的二维函数。

与传统频谱分析不同，循环谱相关通过引入循环频率的额外维度，从谱频率和循环频率两个维度对故障信号进行刻画。具体而言，谱频率 f 对应于故障信号中的载波成分 f_c，而循环频率 α 对应于故障信号中的调制成分 f_m。这种方法自上世纪以来被广泛应用于随机信号分析[178,179]，近年来也逐渐被引入故障诊断中的数据预处理环节[20]。循环谱相关的这种双维度分析特性，具有强大的故障成分刻画能力，也为后文提出的 CS-SHAP 算法奠定了重要的理论基础。

4.3　将传统 SHAP 拓展至循环域以优化解释形式的 CS-SHAP

4.3.1　面向确定性信号的循环域变换

针对端到端故障诊断模型，现有的 SHAP 方法可以直接计算振动信号时域各成分的归因贡献，或扩展到频域、时频域和包络域。然而，这些方法往往无法全面揭示故障成分。在 4.2.2 小节中介绍的循环谱相关分析能够同时揭示振动信号载波和调制频率，能更清晰的刻画故障成分，这促使我们将 SHAP 拓展到循环域来优化被动解释形式。

然而，循环谱相关分析是为随机信号设计的，不能直接应用于故障诊断模型中使用的确定性信号。将循环谱相关分析应用于确定性信号的关键，在于估计式 (4-6) 所示的二维自相关函数 $R_x(\tau,t)$。对于二阶平稳信号，随着时延 τ 的增加，$R_x(\tau,t)$ 会衰减至零。也就是说，二维自相关函数本质上是信号 $x(t)$ 在特定局部时间 t 的自相关函数。因此，可以对确定性信号 $x(t)$ 应用窗函数 $h(t)$，并获得自相关函数来近似于二维自相关函数 $R'_x(\tau,t)$。加窗信号 $\hat{x}(t',t)$ 可以表示为

$$\hat{x}(t',t)=x(t'-t)w(t'), \tag{4-9}$$

式中 t' 代表时间轴，$w(t)$ 代表窗函数，t 代表窗函数的中心时刻。然后，二维自相关函数 $R'_x(\tau,t)$ 可以由加窗信号 $\hat{x}(t',t)$ 沿 t' 轴的自相关函数 $\mathcal{R}(\cdot)$ 进行近似：

$$\begin{aligned} R'_x(\tau,t)&=\mathcal{R}_{t'\to\tau}(\hat{x}(t',t))\\ &=\int \hat{x}(t'-\tau/2,t)\cdot\hat{x}^*(t'+\tau/2,t)\,\mathrm{d}t'. \end{aligned} \tag{4-10}$$

将近似得到的二维自相关函数 $R'_x(\tau,t)$ 代入式 (4-8)，则确定性信号 $x(t)$ 的循环谱表征 $\mathrm{CS}_x(f,\alpha)$ 可以表示为

$$\begin{aligned} \mathrm{CS}_x(f,\alpha)&=\int\int R_x(\tau,t)\mathrm{e}^{-\mathrm{i}2\pi(f\tau+\alpha t)}\,\mathrm{d}\tau\,\mathrm{d}t\\ &=\int\left[\int\mathcal{R}_{t'\to\tau}\Big(\hat{x}(t',t)\Big)\mathrm{e}^{-\mathrm{i}2\pi f\tau}\mathrm{d}\tau\right]\cdot\mathrm{e}^{-\mathrm{i}2\pi\alpha t}\,\mathrm{d}t\\ &=\int\mathcal{F}_{\tau\to f}\left(\mathcal{R}_{t'\to\tau}\Big(\hat{x}(t',t)\Big)\right)\cdot\mathrm{e}^{-\mathrm{i}2\pi\alpha t}\,\mathrm{d}t, \end{aligned} \tag{4-11}$$

式中 $\mathcal{F}(\cdot)$ 表示 Fourier 变换。由信号处理知识可知，对自相关函数 $\mathcal{R}(\cdot)$ 进行 Fourier 变换 $\mathcal{F}(\cdot)$ 可以得到功率谱密度，而功率谱密度和 Fourier 变换存在如下关系：

$$\mathcal{F}_{\tau\to f}\Big(\mathcal{R}_{t'\to\tau}\Big(x(t')\Big)\Big)=\Big|\mathcal{F}_{t\to f}\Big(x(t')\Big)\Big|^2. \tag{4-12}$$

基于式 (4-11) 和 (4-12)，可以推导出

$$\begin{aligned}\mathrm{CS}_x(f,\alpha) &= \int \left|\mathcal{F}_{t'\to f}\left(\hat{x}(t',t)\right)\right|^2 \cdot \mathrm{e}^{-\mathrm{i}2\pi\alpha t}\,\mathrm{d}t \\ &= \mathcal{F}_{t\to\alpha}\left[\left|\int x(t'-t)w(t')\mathrm{e}^{-\mathrm{i}2\pi f t'}\mathrm{d}t'\right|^2\right] \\ &= \mathcal{F}_{t\to\alpha}\left[\left|\mathrm{STFT}_x(f,t)\right|^2\right],\end{aligned} \tag{4-13}$$

式中 $\mathrm{STFT}_x(f,t)$ 代表信号 $x(t)$ 的短时傅里叶变换结果。因此，我们可以对确定性信号 $x(t)$ 的短时傅里叶变换能量谱 $|\mathrm{STFT}_x(f,t)|^2$ 的时间轴 t 进行 Fourier 变换 $\mathcal{F}(\cdot)$，从而获得信号在循环域的表征 $\mathrm{CS}_x(f,\alpha)$。该做法在文献[104] 中被隐式地提及，但本章是首次推导并严格证明了时频变换和时间轴 Fourier 变换的结合，与随机信号中 CSC 的等价性。

上述工作理论地推导了获得确定性信号在循环域表征的正向过程，称之为循环域变换 $\mathcal{D}$。它不仅能有效地提取二阶循环平稳故障信号的载波频率和调制频率，也能揭示普通正弦故障信号的特征频率。具体而言，对于普通正弦信号 $x_1(t)$：

$$x_1(t) = A_1\sin(2\pi f_1 t + \phi_1), \tag{4-14}$$

式中 A_1、f_1、ϕ_1 分别代表输入信号 $x_1(t)$ 的幅值、故障频率和相位。那么，$x_1(t)$ 的短时 Fourier 变换的能量谱为

$$|\mathrm{STFT}_x(f,t)|^2 = \begin{cases} K, & \text{if } f=\pm f_1 \\ 0, & \text{else} \end{cases}, \quad \text{其中, } K = \frac{A_1}{2}\cdot\left|\int h(t')\,\mathrm{d}t'\right|^2. \tag{4-15}$$

进而，正弦信号 $x_1(t)$ 的循环域表征可得

$$\mathrm{CS}_{x_1}(f,\alpha) = \begin{cases} KT, & \text{if } f=\pm f_1, \alpha=0 \\ 0, & \text{else} \end{cases}, \tag{4-16}$$

式中输入信号 T 代表 $x_1(t)$ 的信号长度。综上，普通正弦信号可以视为循环频率 $\alpha=0$ 的二阶循环平稳信号，其循环域表征将出现在 $f=\pm f_1, \alpha=0$ 处，表明循环域变换 $\mathcal{D}$ 仍适用于常见的正弦信号。

为了将此方法与 SHAP 集成，还必须建立对应的循环域逆变换 $\mathcal{D}^{-1}$。图 4-2 展示了循环域变换 $\mathcal{D}$ 及其逆变换 $\mathcal{D}^{-1}$ 的计算过程。需要注意的是，整个计算过程需要保留 STFT 结果的相位信息 $\phi_x(f,t)$，以确保循环域逆变换中的准确重构，从而避免式 (4-13) 中求模运算导致的信息丢失。

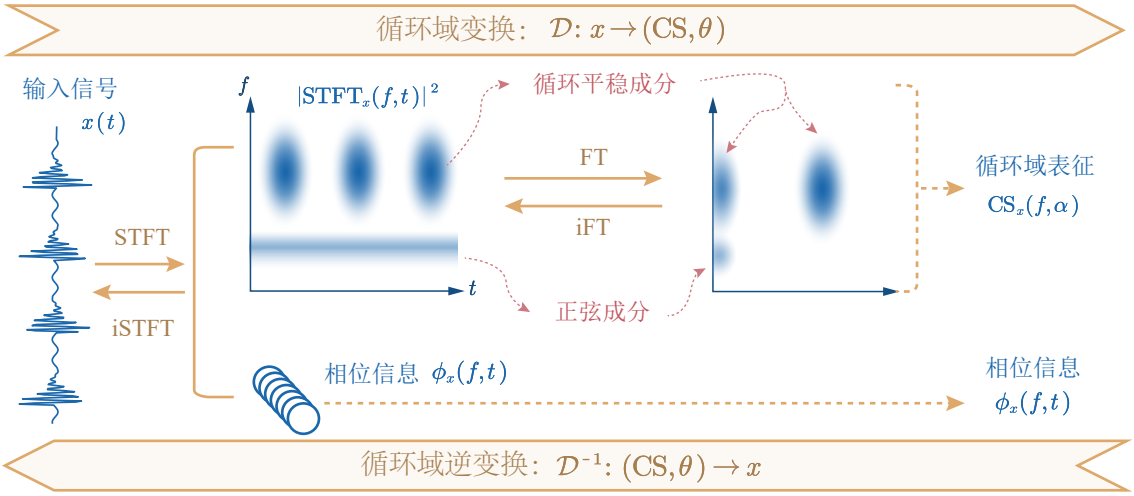


图 4-2　循环域变换及其逆变换的示意图

Fig. 4-2　The processes of the CS transform and inverse CS transform

对于输入信号 $x(t)$，循环域变换 $\mathcal{D}: x \to (\mathrm{CS}, \phi)$ 可以表示为

$$\begin{aligned} \mathrm{STFT}_x(f,t) &= \int x(t'-t)h(t')\mathrm{e}^{-\mathrm{i}2\pi f t'}\mathrm{d}t', \\ \phi_x(f,t) &= \mathrm{Ang}(\mathrm{STFT}_x(f,t)), \\ \mathrm{CS}_x(f,\alpha) &= \int |\mathrm{STFT}_x(f,t)|^2 \cdot \mathrm{e}^{-\mathrm{i}2\pi\alpha t}\,\mathrm{d}t, \end{aligned} \tag{4-17}$$

式中 $\mathrm{Ang}(\cdot)$ 代表用于提取复数相位的函数。

给定循环域表征 $\mathrm{CS}_x(f,\alpha)$ 及其相位信息 $\phi_x(f,t)$，循环域逆变换 $\mathcal{D}^{-1}: (\mathrm{CS}, \phi) \to x$ 可以表示为

$$\begin{aligned} \mathrm{STFT}_x(f,t) &= \sqrt{\int \mathrm{CS}_x(f,\alpha) \cdot \mathrm{e}^{\mathrm{i}2\pi\alpha t}\,\mathrm{d}t} \cdot (\cos\phi + \mathrm{i}\sin\phi), \\ x(t) &= \mathrm{iSTFT}\left(\mathrm{STFT}_x(f,t)\right), \end{aligned} \tag{4-18}$$

式中 $\mathrm{iSTFT}(\cdot)$ 代表短时 Fourier 逆变换。

借助循环域变换 $\mathcal{D}$ 和循环域逆变换 $\mathcal{D}^{-1}$，能够将信号在时域和循环域直接相互转换，为后续 CS-SHAP 的实现提供基础。

4.3.2　面向旋转机械智能诊断模型的 CS-SHAP 被动解释及其应用流程

通过式 (4-17) 和式 (4-18) 中所示的循环域变换 $\mathcal{D}$ 和循环域逆变换 $\mathcal{D}^{-1}$，CS-SHAP 可以将传统的时域 SHAP 解释扩展到循环域。传统时域 SHAP 归因和 CS-SHAP 归因的计算过程如图 4-3 所示。

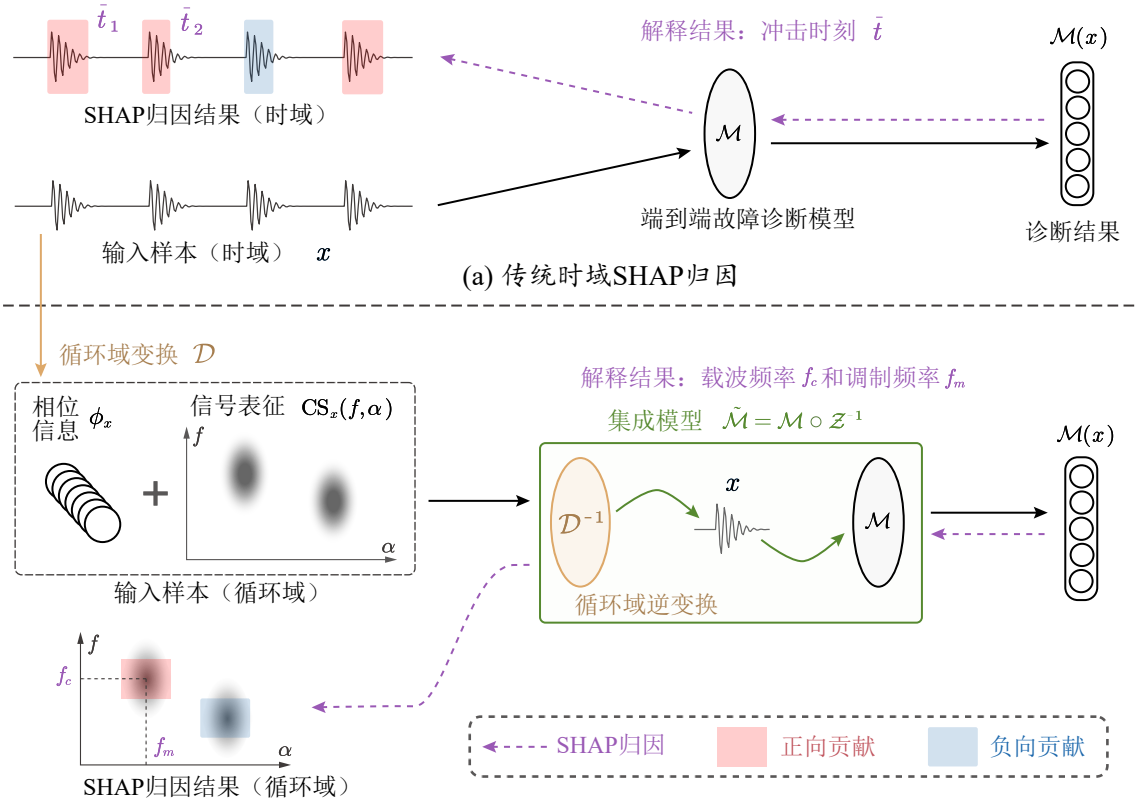


图 4-3　传统时域 SHAP 归因和 CS-SHAP 归因的计算过程

Fig. 4-3　The illustration of SHAP attribution methods in time-domain and CS-domain

传统时域 SHAP 使用式 (4-5) 直接计算时域样本 x 的不同部分对预测结果 $\mathcal{M}(\boldsymbol{x})$ 的贡献度，其归因结果通常是为各个故障部分的冲击时刻。它在故障特征显著、噪声强度低的信号上具有一定解释效果，但在实际的高噪声场景中面临挑战，因为故障特征往往被噪声掩盖。

相比于传统的时域 SHAP，CS-SHAP 有两方面改动：(1) 样本预处理：循环域变换 $\mathcal{D}$ 将时域样本 x 预处理为包括循环域表征 CS_x 和相位信息 ϕ_x 的循环域样本。(2) 模型集成：通过循环域逆变换 $\mathcal{D}^{-1}$ 与端到端模型 $\mathcal{M}$ 集成，从而在不改变端到端型架构的前提下，实现将循环域样本作为模型输入。

所提出的 CS-SHAP 将归因解释从时域拓展至循环域，从而具有如下三方面优势：(1) 可靠的解释性标签：时域难以辨认高噪声信号的冲击时刻，而循环域更能凸显信号中的故障信息，从而获得更实测准确的标签用以评估解释效果。(2) 更清晰的解释效果：时域只能揭示故障成分的冲击时刻，而循环域能同时揭示载波频率 f_c 和调制频率 f_m，这是故障定位和推理的重要依据。(3) 鲁棒性：循环域比时域受噪声干扰的

影响更小，更适合高噪声场景。

将 CS-SHAP 应用于智能故障诊断模型可解释性分析的完整流程如图 4-4 所示，包括五个关键步骤：模型准备、模型集成、样本预处理、SHAP 归因、以及可视化和可解释性分析。

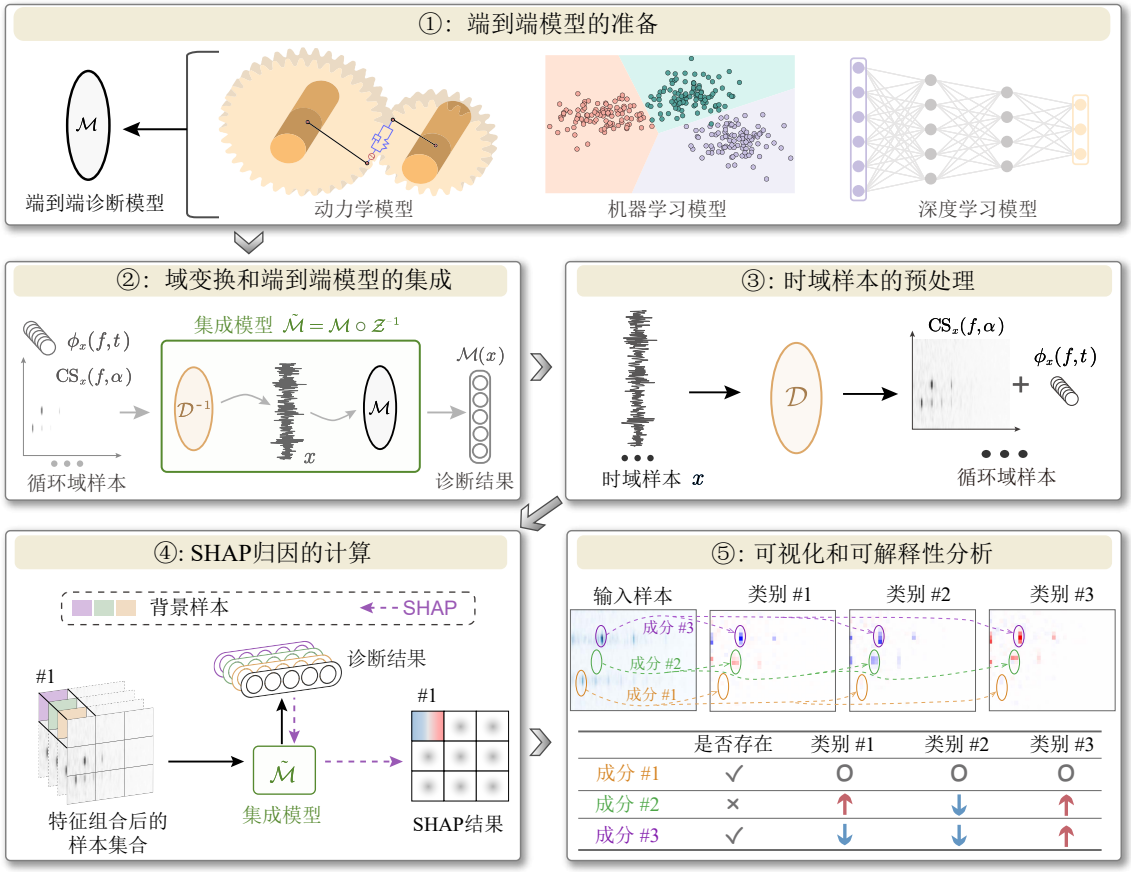


图 4-4　将 CS-SHAP 应用于智能故障诊断模型可解释性分析的完整流程

Fig. 4-4　The complete process of applying CS-SHAP to the explainability analysis of intelligent fault diagnosis models

首先，用户可以根据任务场景准备特定的、不受限的端到端模型 $\mathcal{M}$，该模型可由动力学、机器学习或深度学习等各种方法构建。其次，将准备好的模型 $\mathcal{M}$ 与式 (4-17) 所示的循环域变换 $\mathcal{D}$ 集成，获得以循环域样本为输入的集成模型 $\tilde{\mathcal{M}}$。之后，使用式 (4-18) 所示的循环域逆变换 $\mathcal{D}^{-1}$ 将现有时域样本 x 转换为循环域样本（CS_x 和 ϕ_x）。接着，根据式 (4-5) 进行 SHAP 分析，不断迭代各部分组合并计算相对于数据分布的期望，从而获得循环谱表征 CS_x 中各部分对预测结果 $\mathcal{M}(x)$ 的贡献度 $\psi_{\tilde{\mathcal{M}},\boldsymbol{x}}(\tilde{\boldsymbol{x}})$。最后，对 CS-SHAP 结果 ψ 进行可视化和可解释性分析，其中，红色代表样本的此区域成分对预测类别具有正向贡献，蓝色则代表负向贡献。

需要提及的是，CS-SHAP 结果中的贡献受两种因素影响：一种是信号成分与类别的相关性，另一种是信号成分的存在性。具体而言，存在（或不存在）与当前类别相关的故障成分会导致正（负）贡献，存在（或不存在）与其他类别相关的故障成分则会导致负（正）贡献。如图 4-4 的可视化部分所示，信号成分 #1 与所有三个类别无关，无论其是否存在，对三类类别的贡献均为零。信号成分 #2 对应于类别 #2，其不存在，对类别 #2 具有负贡献，而对类别 #1 和 #3 具有正贡献。信号成分 #3 对应于类别 #3，其存在，对类别 #3 具有正贡献，而对类别 #1 和 #2 具有负贡献。

4.4　CS-SHAP 被动解释效果的实验验证

本章将通过三个不同的数据集来对 CS-SHAP 的被动解释效果进行实验验证，包含一个仿真数据集、图 2-7 所示的 CWRU 轴承开源数据集和图 3-10 所示的斜齿轮数据集。其中，仿真数据集的故障逻辑完全可知，具有实测的解释性标签，能够有效评估解释效果；CWRU 数据集作为广泛使用的开源基准，以确保 CS-SHAP 的实测性和可复现性；斜齿轮数据集则来自实测工业环境的，以突显 CS-SHAP 在实际场景中的应用价值。

对于预测模型 $\mathcal{M}$，本章选择如表 4-1 所示的端到端卷积神经网络进行分析。同 SHAP 一致，CS-SHAP 也属于事后解释类型，这意味着 CS-SHAP 并不限定于特定模型。本章在实验部分统一使用该模型，以确保对 CS-SHAP 可解释性性能的评估不受模型影响。

表 4-1　CS-SHAP 验证实验中所采用的端到端模型架构

Table 4-1　The architecture of the end-to-end model used in CS-SHAP validation experiments

序号	网络层参数	输出尺寸
-	Input	1×2000
1	Conv(8@7)[a] -BN-ReLU-MaxPool(2)	8×997
2∼7:→ i	[Conv(2^{i+2}@3)-BN-ReLU-MaxPool(2)] *6	512×13
9	Conv(1024@3)-BN-ReLU-AdapMaxPool(1)	1024×1
10	Flatten-FC(256)-ReLU-FC(64)-ReLU-FC(K)	K[b]

[a] Conv(x @y): 表示一个输出通道为 x 、卷积核大小为 y 的卷积层。此外, 步长为 1, 填充为 0。其中，输入通道数由前一层的输出大小决定。

[b] K: 表示数据集中的类别数量。

至于对比方法，本章选择了主流的归因方法用于对比解释效果，包括梯度类别激活映射（grad-CAM）、普通的的时域 SHAP (Time-SHAP)、拓展至频域的 SHAP (Freq-

SHAP)、拓展至包络谱域的 SHAP (Env-SHAP) 和拓展至时频域的 SHAP (TF-SHAP)。

4.4.1　故障逻辑已知的仿真数据集

首先介绍各类别信号的构造方式并梳理故障逻辑，然后基于仿真数据集对模型进行训练，最后用不同的归因方法获得被动解释结果并加以分析。在各类别信号的构造上，定义周期脉冲分量 x_p：

$$x_p(f_m, f_c, t) = \sum_{k\in\mathbb{N}} \mathrm{e}^{-\beta(t-k/f_m)} \sin\Big(2\pi f_c(t-\frac{k}{f_m})+\phi\Big), \tag{4-19}$$

式中 f_m 代表故障脉冲的激励频率，f_c 代表故障的响应频率，$\beta = 0.04$ 代表阻尼系数，$\phi \sim \mathcal{U}(0, 2\pi)$ 代表初始相位，其中 $\mathcal{U}(a, b)$ 代表下界为 a、上界为 b 的均匀分布。仿真信号可以表示为

$$x = \sum_{i=1}^{2} A \cdot x_p^i(f_m^i, f_c^i, t) + n(t), \tag{4-20}$$

式中 $A \sim \mathcal{U}(0.8, 1)$ 代表幅值系数，$n(t)$ 代表信噪比为 0 的高斯白噪声。不同类别的故障样本具有不同的 f_m 和 f_i。

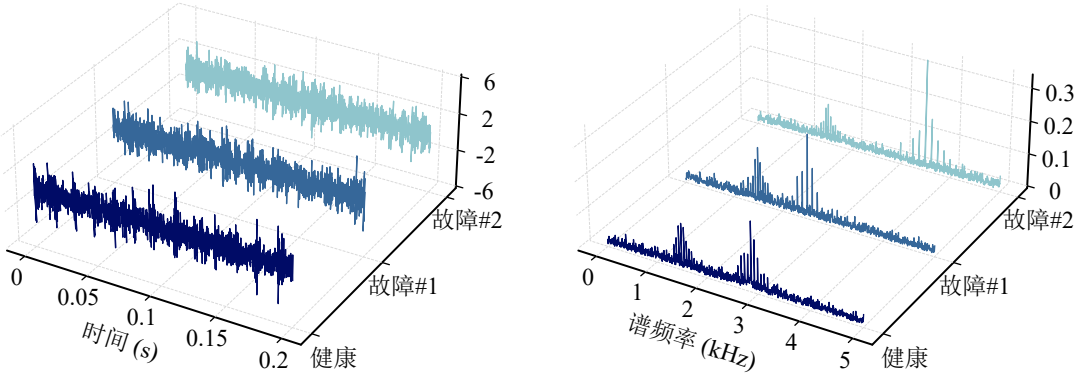
在该数据集中，采样频率设置为 10 kHz，定义了三种故障类别：健康状态、故障 #1 和故障 #2。各信号成分的参数设置及其与故障类别的关系如表 4-2 所示，每个故障类别各包含两个周期脉冲分量。具体而言，分量 C_0 在所有三个类别中均存在，对任何类别都没有贡献。相反，其他分量（即 C_H、C_1、C_2）仅存在于单个类别中，对其对应类别具有正贡献，而对其他类别具有负贡献。该仿真数据集的故障逻辑完全已知，这有助于进行可解释性评估，而其他数据集由于缺乏可靠解释标签是难以做到的。为了便于理解，图 4-5 展示了三种故障类别的时域和频域表征。

表 4-2　仿真数据集中各信号成分的参数设置及其与故障类别的关系

Table 4-2　The parameter settings of each signal component in the simulation dataset and their relationships with fault classes

信号成分	f_c (kHz)	f_m (Hz)	健康	故障 #1	故障 #2
C_0	1.5	50	✓	✓	✓
C_H	$\mathcal{U}(1, 4)$	$\mathcal{U}(20, 200)$	✓		
C_1	2.5	100		✓	
C_2	3.5	125			✓

在训练方面，仿真数据集实验可以被视为一个三分类任务。每个类别包含 5000 个样本，每个样本长度为 2000，通过均值-标准差归一化处理。其中 70% 的样本随机



(a) 时域表征　　(b) 频域表征

图 4-5　仿真数据集中各故障类别的时域和频域表征

Fig. 4-5　The time-domain and frequency-domain representations of different fault classes in the simulation dataset

选择用于训练，剩余 30% 用于测试。训练参数包括 20 个训练周期，批次大小为 64，采用 Adam 优化器，学习率为 0.001 且每个周期衰减 0.99。最终，训练好的卷积网络模型在测试集上达到了 99.98% 的准确率。

为了测试算法的解释能力，本章将 CS-SHAP 和对比方法应用于训练好的卷积网络模型，三种类别的样本表征和归因解释结果分别如图 4-6、4-7 和 4-8 所示。首先对不同域的样本表征效果进行对比。对于时域表征，由于噪声干扰，难以直接辨认故障的冲击时刻，导致缺乏故障的解释标签。在频域中，周期脉冲分量表现为边带，能够清晰揭示载波频率 f_c 和调制频率 f_m。然而，实际应用中的噪声经常掩盖调制频率 f_m，从而将解释标签限制在载波频率 f_c。在包络谱域中，周期脉冲分量表现为调制频率 f_m 及其谐波，能够提供 f_m 的解释标签。在时频域中，周期脉冲分量表现为载波频率 f_c 及冲击时刻，但它同样易受噪声影响，仅能够提供载波频率 f_c 的解释标签。在循环域中，周期脉冲分量表现为载波频率 f_c 和调制频率 f_m 及其谐波，且对噪声具有鲁棒性不易受噪声影响，为载波频率 f_c 和调制频率 f_m 提供全面的解释标签。由此可见，循环域能够全方面地揭示故障成分，而其他域仅揭示部分信息或易受噪声干扰。

健康样本由共同分量 C_0 和随机分量 C_H 组成，其不同归因方法的解释结果如图 4-6(b)-(f) 所示。其中，图 4-6(b) 中 grad-CAM 结果比较粗糙，并不理想，尽管图 4-6(c) 中的 Time-SHAP 更为精细，其中部分冲击时刻具有高贡献，但仍无法获得有效信息。图 4-6(d) 所示的 Env-SHAP 表明，结合各成分的载波频率 f_c 可知，C_0 对任何类别都没有贡献，而健康类的预测来自 C_1 和 C_2 的不存在。同理，图 4-6(e) 所示的 Env-SHAP 从对应于 f_m 的循环频率角度表明，C_0 的存在无影响，而 C_H 的存在和 C_1、

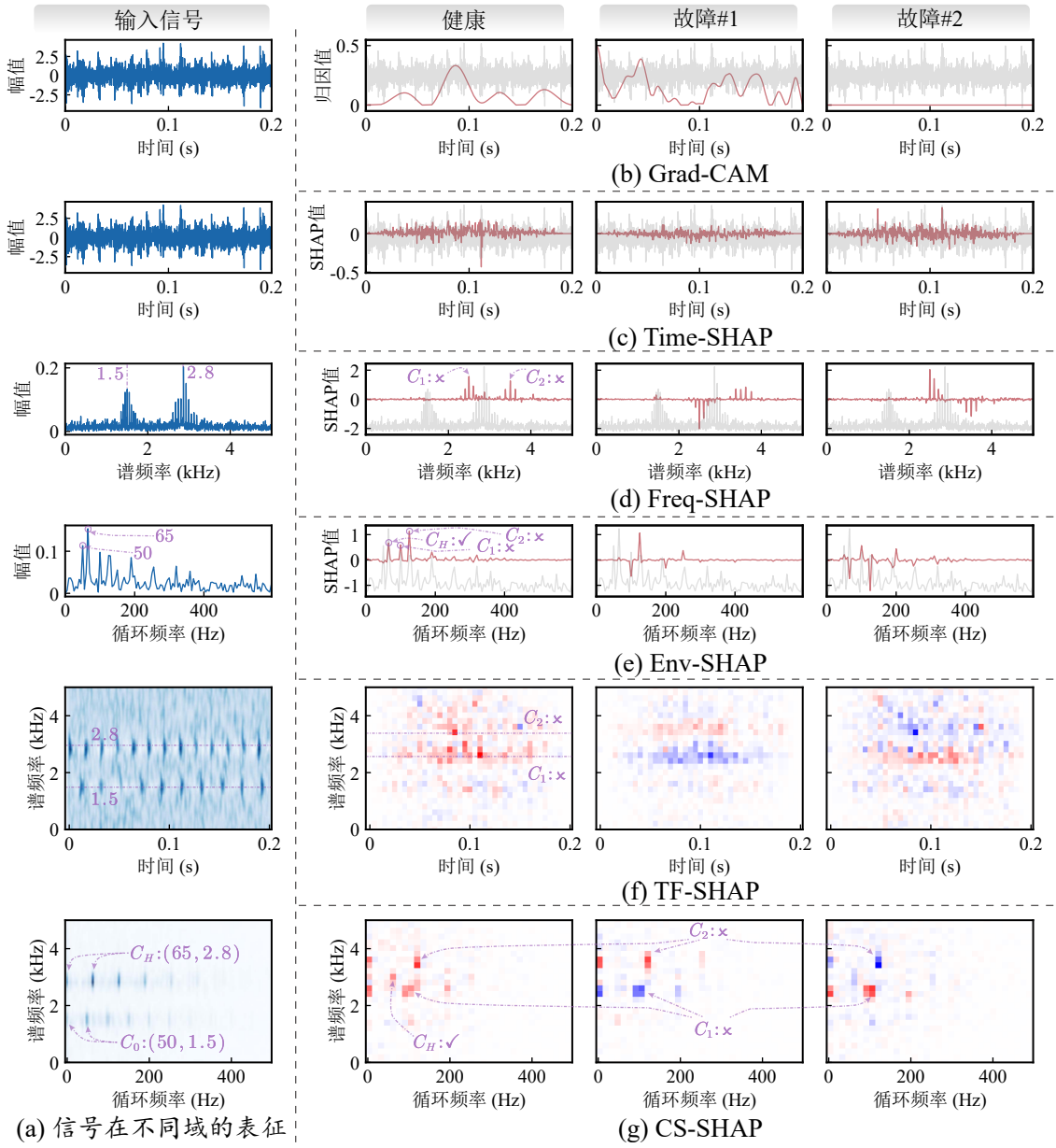


图 4-6　仿真数据集下健康样本的各域表征及不同归因方法的结果

Fig. 4-6　The domain representations of a health sample from the simulation dataset and its attribution results using various methods

C_2 的不存在对健康类别具有正贡献。图 4-6(f) 所示的 TF-SHAP 与 Freq-SHAP 的结果保持一致，但由于时间轴 t 的不确定性，其结果更为粗糙。从各成分的载波频率 f_c 可知，C_1 和 C_2 的不存在对健康类有正贡献。图 4-6(g) 所示的 CS-SHAP 则从对应于 f_c 的谱频率和对应于 f_m 的循环频率两个维度表明，C_0 没有贡献，C_H 的存在对健康类略有正贡献，对其他类别有负贡献；而 C_1 和 C_2 的不存在，对健康类别均为正贡献，

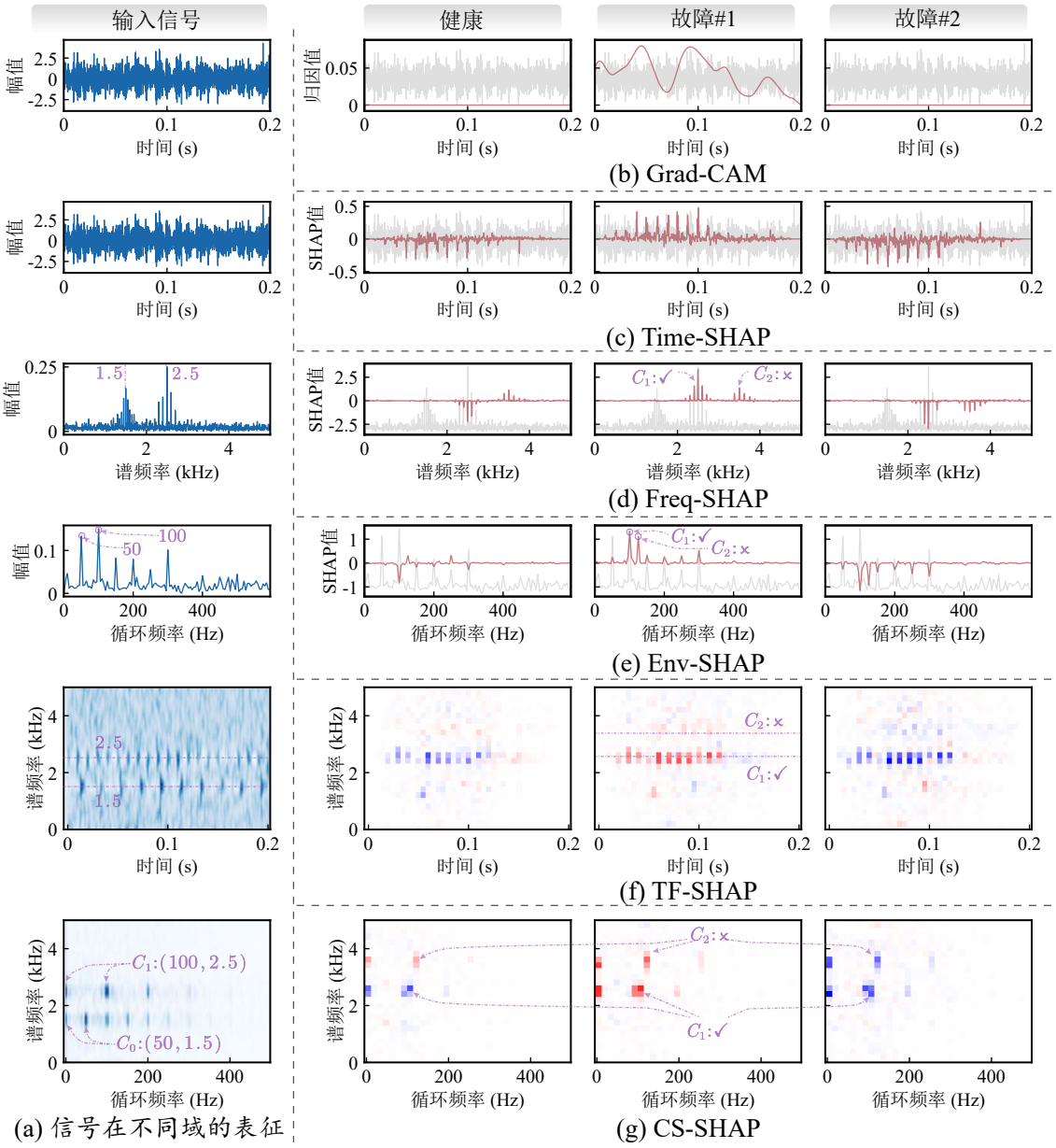


图 4-7　仿真数据集下故障 #1 样本的各域表征及不同归因方法的结果

Fig. 4-7　The domain representations of a Fault #1 sample from the simulation dataset and its attribution results using various methods

而对他们的对应类别则为负贡献。CS-SHAP 不仅能够全面揭示载波频率 f_c 和调制频率 f_m，而且其解释结果和表 4-2 所预设的故障逻辑完全一致，表明了 CS-SHAP 的显著解释优势。

图 4-7 和图 4-8 的其他两类样本归因解释结果和上述分析过程一致。为简洁起见，本章仅对其中的关键部分进行分析。本质上，故障 #1 和故障 #2 样本均包含共同分量

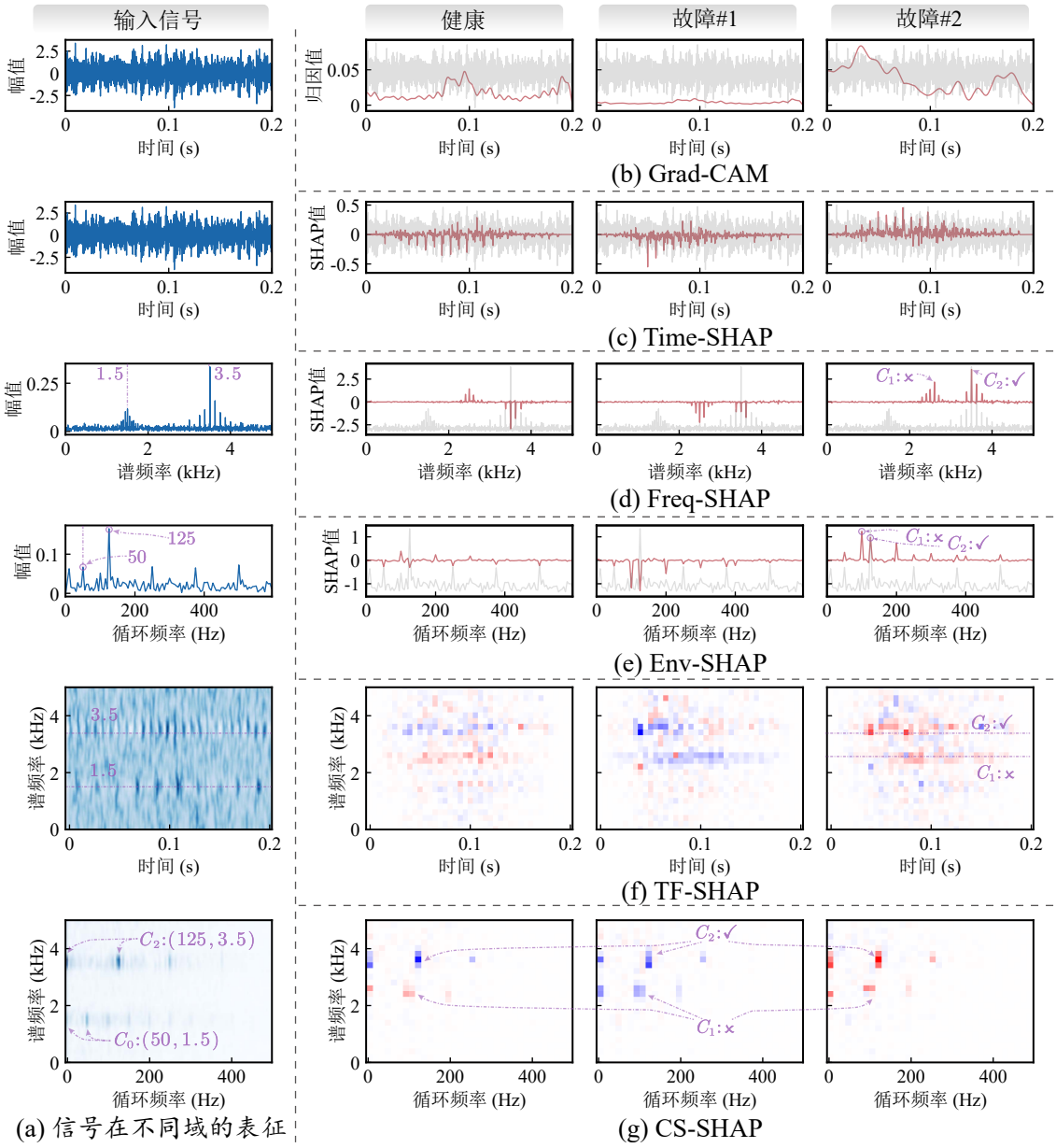


图 4-8　仿真数据集下故障 #2 样本的各域表征及不同归因方法的结果

Fig. 4-8　The domain representations of a Fault #2 sample from the simulation dataset and its attribution results using various methods

C_0 和独特分量 C_1 或 C_2。总的来说，由于 C_0 在所有类别中都存在，属于共有成分，因此其存在对所有类别都没有贡献。由于 C_1 对应故障 #1，因此其存在对故障 #1 是正贡献，而对其他类别是负贡献，反之亦然。C_2 的效果类似于 C_1，只是其正贡献与故障 #2 相关。图 4-7 和 4-8 有效验证了上述逻辑，证明了全体 SHAP 类方法在事后解释方面的逻辑正确性，但 CS-SHAP 在解释效果方面更具优势，其解释结果相比其

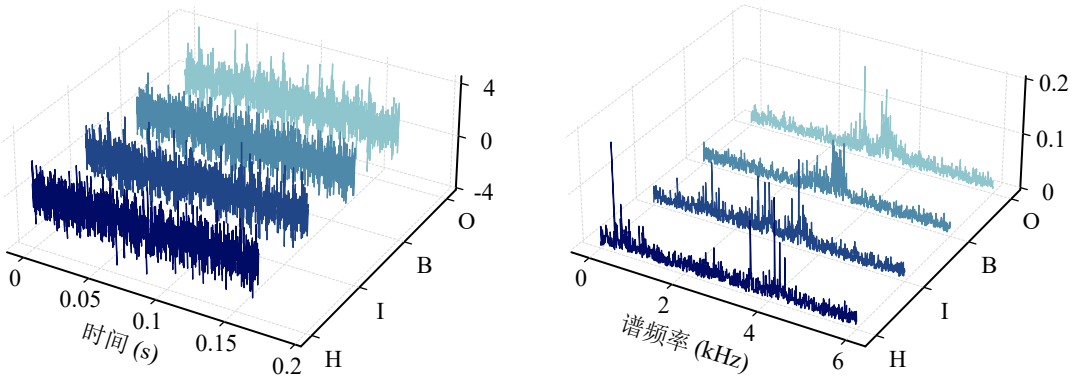
他 SHAP 方法更为全面和清晰。

至于不同域的解释效果，由于冲击时刻的可变性，Grad-CAM 和 Time-SHAP 的效果都比较差，表现不佳。其他域中的 SHAP 方法表现更好。具体而言，Freq-SHAP 和 TF-SHAP 解释了载波频率 f_c 的贡献，前者更精细，后者粒度更粗。Env-SHAP 解释了调制频率 f_m 的贡献度。然而，上述方法的解释性结果都是片面的，只有 CS-SHAP 能够全面地解释不同载波频率 f_c 和调制频率 f_m 的贡献，提供比其他方法更完整和清晰的解释。

4.4.2　可复现的 CWRU 轴承开源数据集

CWRU 轴承开源数据集是故障诊断领域广泛使用的基准数据集，本章在该数据集上开展解释性分析以验证 CS-SHAP 的实测性和可复现性。CWRU 轴承开源数据集如图 2-7 所示，本试验选择了在 1800 rpm 转速、1 HP 负载下的工况，相应的轴承特征频率如表 4-3 所示。采样频率选择 12 kHz，包括四种故障类别: 健康（Health, H）、内圈故障（Inner race fault, I）、滚动体故障（Rolling ball fault, B）和外圈故障（Outer race fault, O），故障尺寸均为 0.007 英寸。每种类别包含 119 个长度为 2000 点的样本。为了便于理解，图 4-9 展示了四种类别在时域和频域的表现，其中健康类别在转频的滚珠数目倍频 360 Hz (nf_r) 处有显著幅值。实验设置和训练参数与 4.4.1 小节的仿真数据集实验保持一致。最终，端到端卷积网络模型在测试集上达到 100% 的准确率。

不同于故障逻辑完全已知的仿真数据集，CWRU 数据集需要首先确定各类别的故障特征。为方便描述，本章将各信号成分表示为 $P:(a,b)$ 并在图中进行标注，其



(a) 时域表征　　(b) 频域表征

图 4-9　CWRU 轴承数据集中各故障类别的时域和频域表征

Fig. 4-9　The time-domain and frequency-domain representations of different fault classes in the CWRU bearing dataset

表 4-3　CWRU 轴承数据集的特征频率

The characteristic frequencies of CWRU bearing dataset

滚动体数目 (n)	旋转频率 (f_r / Hz)	滚珠内圈通过频率 (f_{BPFI} / Hz)	滚珠外圈通过频率 (f_{BPFO} / Hz)
12	30	162.45	107.55

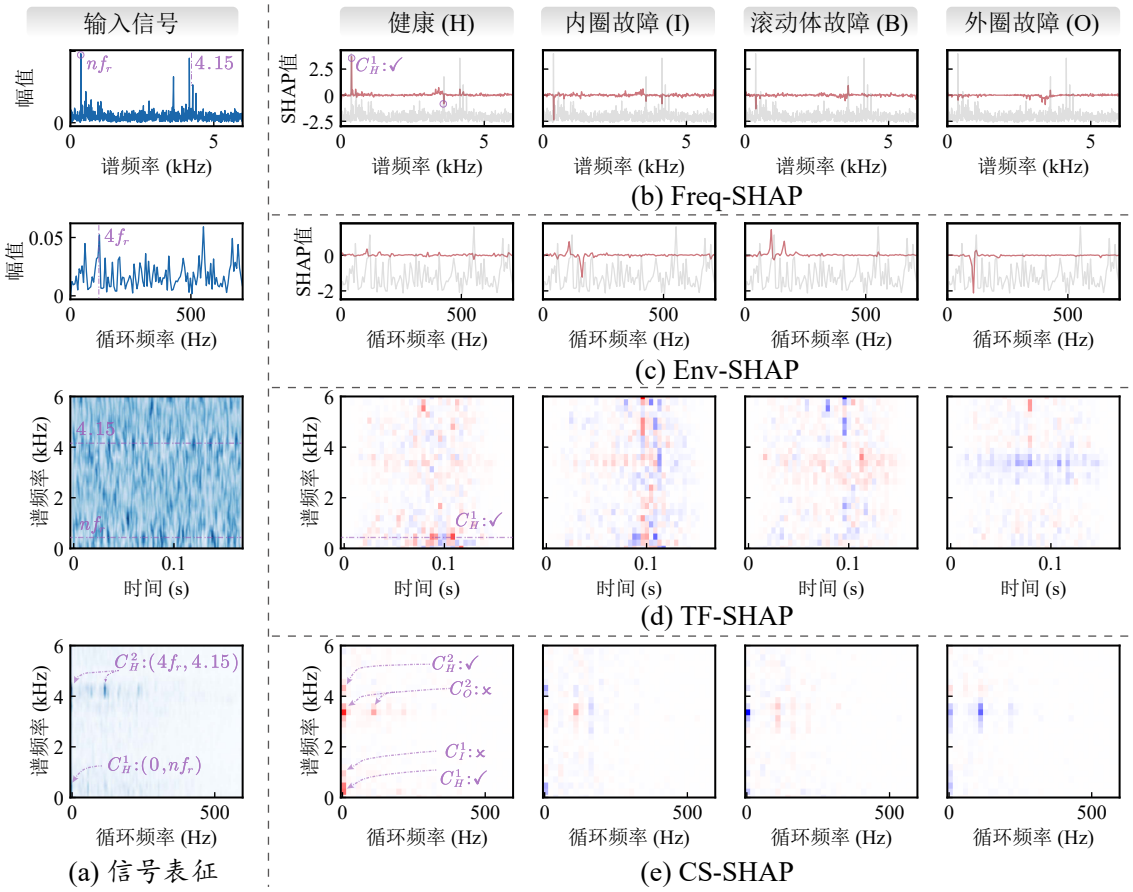


图 4-10　CWRU 数据集下健康样本的各域表征及不同归因方法的结果

Fig. 4-10　The domain representations of a health sample from the CWRU dataset and its attribution results using various methods

中 a (Hz) 和 b (kHz) 分别表示载波频率 f_c 和调制频率 f_m。由各类别的信号表征可知，图 4-10(a) 所示的健康样本包含一个恒定频率分量 $C_H^1:(0,nf_r)$ 和一个调制分量 $C_H^2:(4f_r,4.15)$。图 4-11(a) 所示的内圈故障样本包含 $C_I^1:(0,1.46)$、$C_I^2:(f_{\mathrm{BPFI}},2.74)$ 和 $C_I^3:(f_{\mathrm{BPFI}},3.54)$。图 4-12(a) 所示的滚动体故障样本表现为宽带调制分量 $C_B^1:(0-200,3.3)$。图 4-13(a) 所示的外圈故障样本包含 $C_O^1:(f_{\mathrm{BPFO}},2.87)$ 和 $C_O^2:(f_{\mathrm{BPFO}},3.4)$。

为简化分析，Grad-CAM 和 Time-SHAP 由于表现不佳而被排除在后续分析中，其

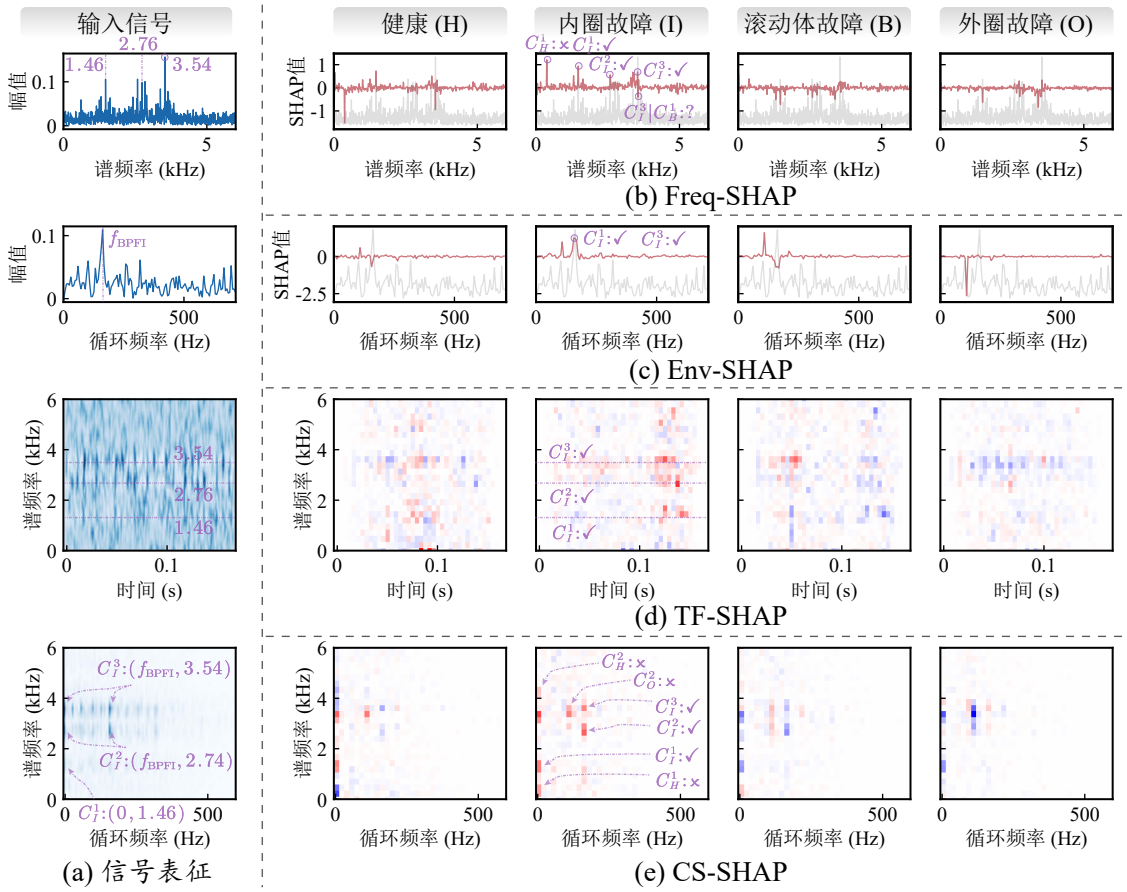


图 4-11　CWRU 数据集下内圈故障样本的各域表征及不同归因方法的结果

Fig. 4-11　The domain representations of a inner race fault sample from the CWRU dataset and its attribution results using various methods

他归因方法对 CWRU 数据集不同故障类别样本的解释结果分别如图 4-10、图 4-11、图 4-12 和图 4-13(b)-(e) 所示。

首先分析健康样本，图 4-10(b) 中的 Freq-SHAP 表明，模型将该样本分类为健康类别 y_H 的贡献主要来自于 C_H^1 的存在。图 4-10(c) 中的 Env-SHAP 则表明，各个调制频率对健康类别 y_H 的预测均无显著影响，这与不存在显著循环频率的健康样本包络谱相对应。图 4-10(d) 中的 TF-SHAP 同样将健康类别 y_H 主要归因于 C_H^1 的存在，与 Freq-SHAP 的解释结果一致。

不同于上述三种模型，图 4-10(e) 中的 CS-SHAP 更多地考虑了其他类别特征不存在所带来的贡献，其归因结果表明，健康类别 y_H 的正贡献不仅来自于健康类别特征 C_H^1 和 C_H^2 的存在，也来自于其他类别特征 C_I^1 和 C_O^2 的不存在，只是 C_H^1 和 C_O^2 的正贡献更为显著。综上，CS-SHAP 通过捕捉其他类别分量不存在导致的正贡献（即

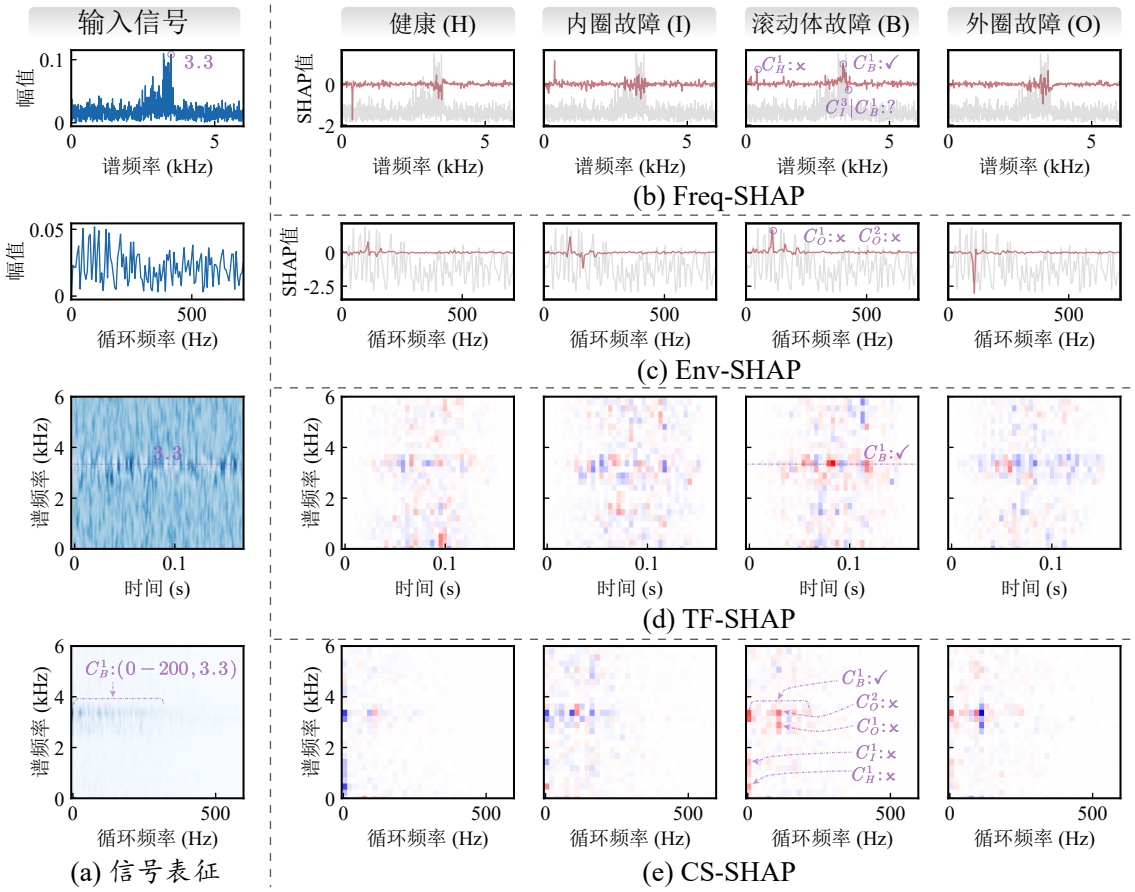


图 4-12　CWRU 数据集下滚动体故障样本的各域表征及不同归因方法的结果

Fig. 4-12　The domain representations of a rolling ball falt sample from the CWRU dataset and its attribution results using various methods

C_I^1 和 C_O^2）展现出更高的精确性，这一现象很少体现于其他 SHAP 方法的结果上。

图 4-11、图 4-12 和图 4-13 的解释效果分析流程与上述图 4-10 一致。借助图中对故障成分和贡献来源的详细标注，读者不难理解各归因方法的效果差异。

值得注意的是，不同信号成分可能具有相似甚至同样的载波频率 f_c 或调制频率 f_m，这给其他 SHAP 方法在区分相近信号成分上带来巨大挑战，进而导致他们的解释效果不够准确。在载波频率方面，内圈故障信号成分 C_I^3 和滚动体故障成分 C_B^1 有相似的、位于 3.3 kHz 附近的谱频率 f_c。图 4-11(b) 所示的内圈故障样本归因结果表明，C_I^3 和 C_B^1 的相似性导致 Freq-SHAP 难以判定 3.3 kHz 频率附件的信号成分对内圈故障类别 y_I 的贡献，使得该区域错误地同时存在由正贡献和负贡献。这种错误现象在图 4-12(b) 中也有体现，表明这并非是偶然现象，而是 Env-SHAP 方法的固有局限。相比之下，图 4-11 和4-12(e) 的 CS-SHAP 归因解释结果表明，CS-SHAP 通过能

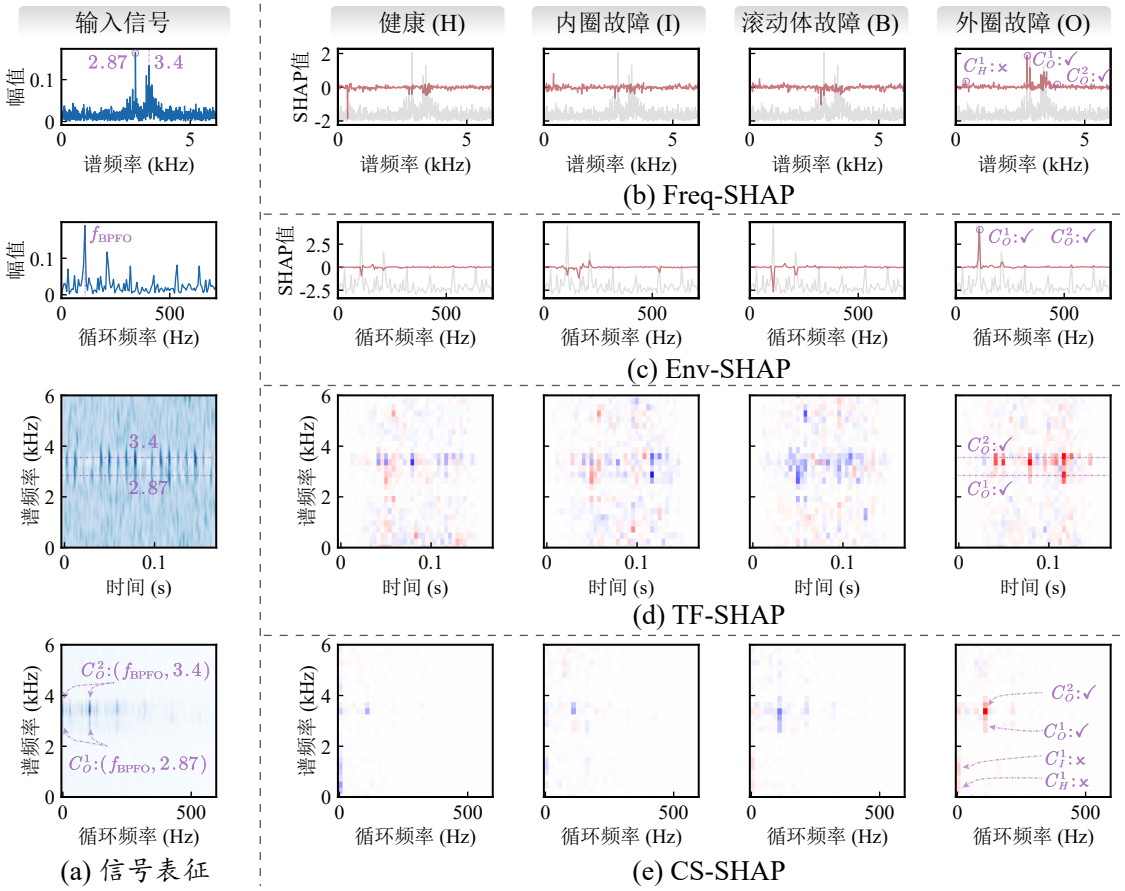


图 4-13　CWRU 数据集下外圈故障样本的各域表征及不同归因方法的结果

Fig. 4-13　The domain representations of a outer race fault sample from the CWRU dataset and its attribution results using various methods

通过载波频率 f_c 和调制频率 f_m 两个维度有效区分 C_I^3 和 C_B^1 这两种不同类别的近似故障成分，从而避免了上述由于故障成分混淆而引发的错误归因现象，保证了归因解释结果的正确性。

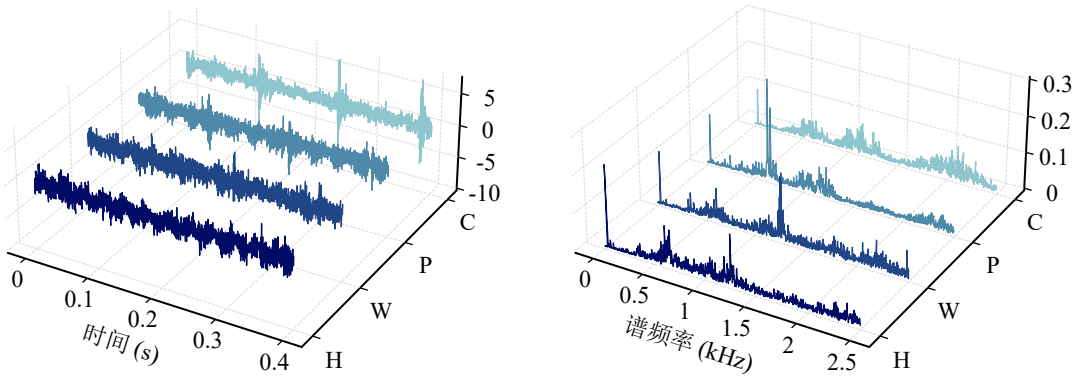
在调制频率方面，内圈故障信号成分 C_O^1 和 C_O^2 具有相同的调制频率 f_{BPIO}。如图 4-13(c) 所示的外圈故障样本归因结果表明，尽管 Env-SHAP 识别出 f_{BPIO} 对外圈故障类别 y_O 的预测具有显著贡献，但无法分离这种贡献的来源中 C_O^1 还是 C_O^2 的占比。相反地，图 4-13(e) 中的 CS-SHAP 能够从另外的载波频率 f_c 角度成功分离 C_O^1 和 C_O^2 这两种相近分量，表明 C_O^2 的贡献大于 C_O^1。这一结果与图 4-13(f) 中的 TF-SHAP 结果和图 4-13(a) 中的循环域表征相符合。

综上，CWRU 数据集进一步验证了 CS-SHAP 相对于其他 SHAP 方法的解释优势。一方面，CS-SHAP 同时考虑故障分量的存在和其他故障分量的不存在所具有的

影响，提供了更清晰准确的归因解释结果。另一方面，CS-SHAP 能够从载波频率和调制频率更有效地区分信号成分，不仅能够获得更清晰的贡献度解释，更能避免潜在的归因错误。

4.4.3　实验室场景下的斜齿轮数据集

斜齿轮数据集的试验台和故障类型如图 3-10 所示，用于验证 CS-SHAP 在实践中的有效性。电机转速设置为 1800 rpm，对应的特征频率如表 4-4 所示。斜齿轮数据集的采样频率设置为 5 kHz 以更好展现故障信息，并考虑四种故障类别：健康（Health, H）、从动齿轮表面磨损故障（Wear, W）、从动齿轮表面点蚀故障（Pitting, P）和从动齿轮断齿故障（Crack, C）。每个类别各包含 76 个长度为 2000 的样本，各故障类别的时域和频域表征如图 4-14 所示。实验和训练设置与 4.4.1 小节的仿真数据集保持一致，端到端卷积网络模型在测试集上达到 100% 的准确率。



(a) 时域表征　　(b) 频域表征

图 4-14　斜齿轮数据集中各故障类别的时域和频域表征

Fig. 4-14　The time-domain and frequency-domain representations of different fault classes in the helical gearbox dataset

表 4-4　斜齿轮数据集的特征频率

The characteristic frequencies of helical gearbox dataset

驱动轮齿数	从动轮齿数	驱动轮转动频率 (f_1 / Hz)	从动轮转动频率 (f_2 / Hz)	啮合频率 (f_{mesh} / Hz)
21	82	30	7.683	630

与 CWRU 数据集类似，在开展分析之前需要首先需要确定每个类别的故障特征，作为可解释性分析的评估标签。具体而言，图 4-15(a) 所示的健康样本包含 C_H^1：

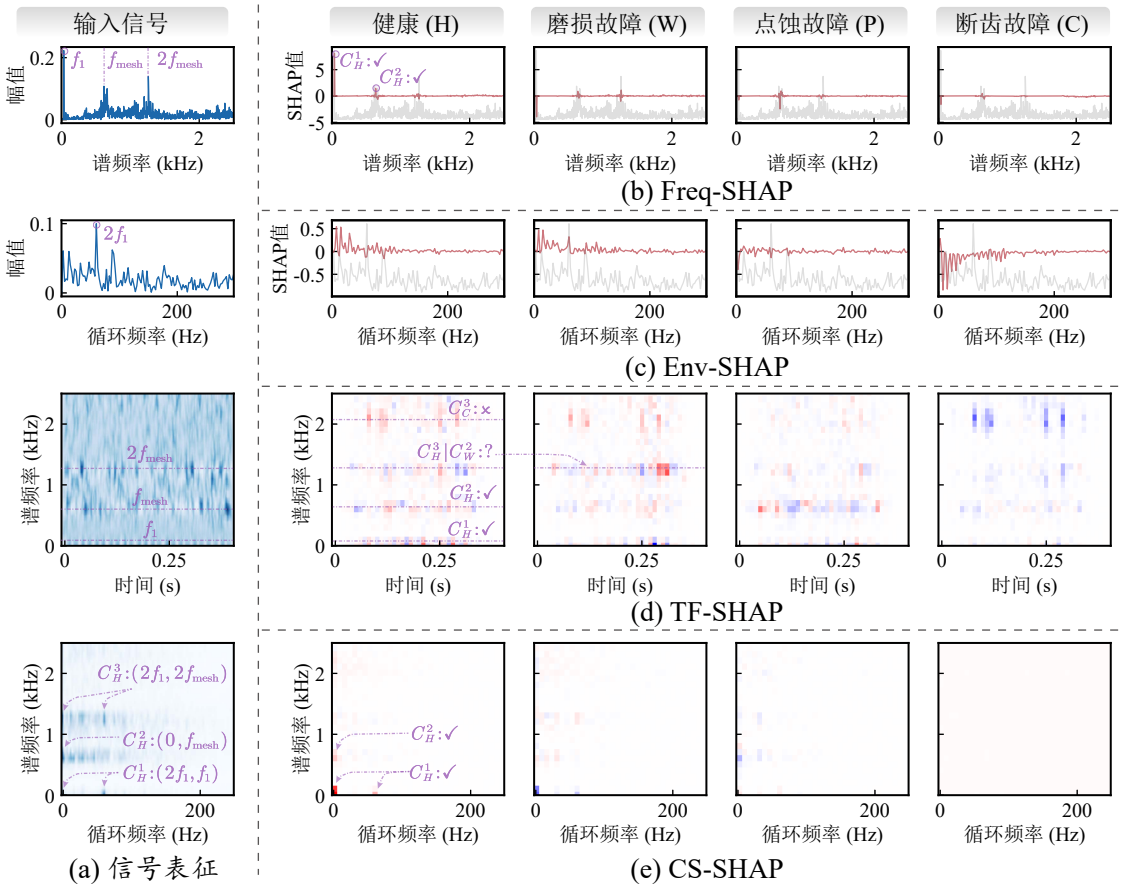


图 4-15　斜齿轮数据集下健康样本的各域表征及不同归因方法的结果

Fig. 4-15　The domain representations of a health sample from the helical gearbox dataset and its attribution results using various methods

$(2f_1, f_1)$、$C_H^2:(0, f_{\text{mesh}})$ 和 $C_H^3:(2f_1, 2f_{\text{mesh}})$ 三类信号成分。图 4-16(a) 所示的磨损故障样本包含 $C_W^1:(0, f_{\text{mesh}})$、$C_W^2:(f_2, 2f_{\text{mesh}})$ 和 $C_W^3:(f_2, 2.2)$ 三类信号成分。图 4-17(a) 所示的点蚀故障样本表现为 $C_P^1:(f_1, f_{\text{mesh}})$、$C_P^2:(f_2, f_{\text{mesh}})$ 和 $C_P^3:(f_2, 2.3)$ 三类信号成分。图 4-18(a) 所示的齿裂故障样本包含 $C_C^1:(f_2, f_{\text{mesh}})$、$C_C^2:(f_2, 1.1)$ 和 $C_C^3:(f_2, 2.1)$ 三类信号成分。

与 CWRU 数据集相比，斜齿轮数据集的难度更高，许多信号成分都具有共同的载波频率 f_c 或共同的调制频率 f_m。就载波频率 f_c 而言，磨损故障信号成分 C_W^1 和点蚀故障信号成分 C_P^1 具有共同的载波频率 $f_c = f_{\text{mesh}}$。如图 4-16(b) 所示，Freq-SHAP 认为磨损故障样本中谱线 f_{mesh} 的存在对磨损类别 y_W 和点蚀类别 y_P 的同时具有正负贡献，产生矛盾的解释结果。相比之下，图 4-16(e) 中的 CS-SHAP，有效地区分了相似的 C_W^1 和 C_P^1，载波频率 f_{mesh} 区域对磨损类别 y_W 均为正贡献，而对点蚀类别 y_P

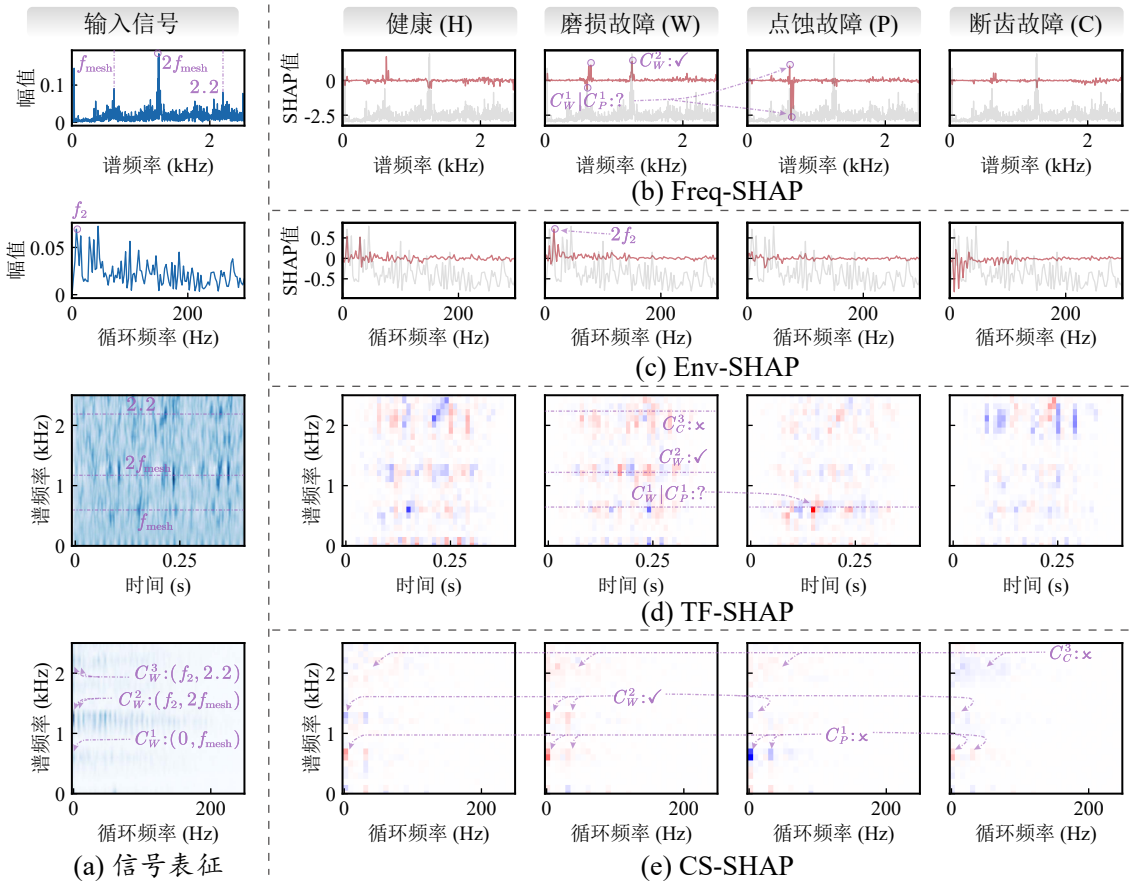


图 4-16　斜齿轮数据集下磨损故障样本的各域表征及不同归因方法的结果

Fig. 4-16　The domain representations of a wear fault sample from the helical gearbox dataset and its attribution results using various methods

均为负贡献。

在调制频率 f_m 方面，断齿故障信号成分 C_C^2 和 C_C^3 具有共同的调制频率 $f_m = f_2$，两者均在图 4-18(a) 中具有显著能量幅值。图 4-18(c) 中的 Env-SHAP 表明调制频率 f_2 对断齿类别 y_C 有显著正贡献，但却无法区分这些贡献来自 C_C^2 还是 C_C^3，还需载波频率 f_c 来进一步澄清。图 4-18(b) 中的 Freq-SHAP 则表明 C_C^2 和 C_C^3 的贡献极低，这种不合理的解释凸显了 Freq-SHAP 的局限性。相反地，图 4-18(d) 中的 TF-SHAP 表明 C_C^3 的贡献大于 C_C^2。图 4-18(e) 中的 CS-SHAP 同样表明 C_C^3 的贡献更为显著，与图 4-18(d) 中的结果大致一致。

综上，CS-SHAP 通过载波频率 f_c 和调制频率 f_m 提供了全面的解释。一方面，它能够有效地分离相近成分，使得解释结果更为清晰；另一方面，它从两个维度对样本进行归因，更贴合信号的故障本质，从而获得更正确的解释结果，避免如图 4-18(b)

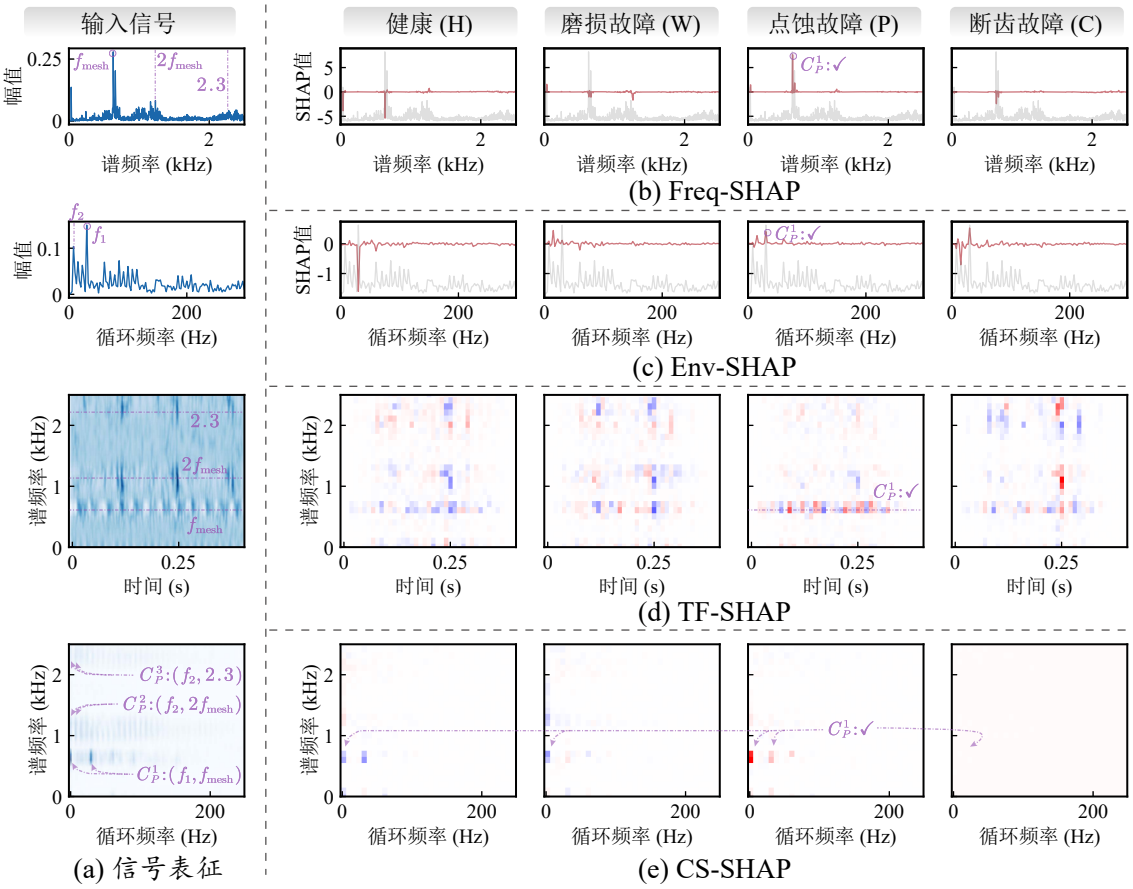


图 4-17　斜齿轮数据集下点蚀故障样本的各域表征及不同归因方法的结果

Fig. 4-17　The domain representations of a pitting fault sample from the helical gearbox dataset and its attribution results using various methods

种所有部分都几乎无贡献的问题。

4.5　CS-SHAP 被动解释效果的影响参数分析

根据式 (4-5) 所示，SHAP 解释主要取决于两个因素：模型 $\mathcal{M}$ 和数据 $\boldsymbol{x}$。由此，本节分别将模型 $\mathcal{M}$ 和数据 $\boldsymbol{x}$ 视为实验变量，分析不同模型和噪声强度下 CS-SHAP 的归因结果。所有实验均在 CWRU 数据集上进行，采用与 4.4.1 小节相同的实验设置。

4.5.1　不同模型下 CS-SHAP 的通用性

为了验证 CS-SHAP 的通用性，实验选择了三种具有代表性的模型 $\mathcal{M}$：MLP、Transformer 和 ResNet。三类模型在 CWRU 数据集的类别层级诊断准确率如图 4-19(a)

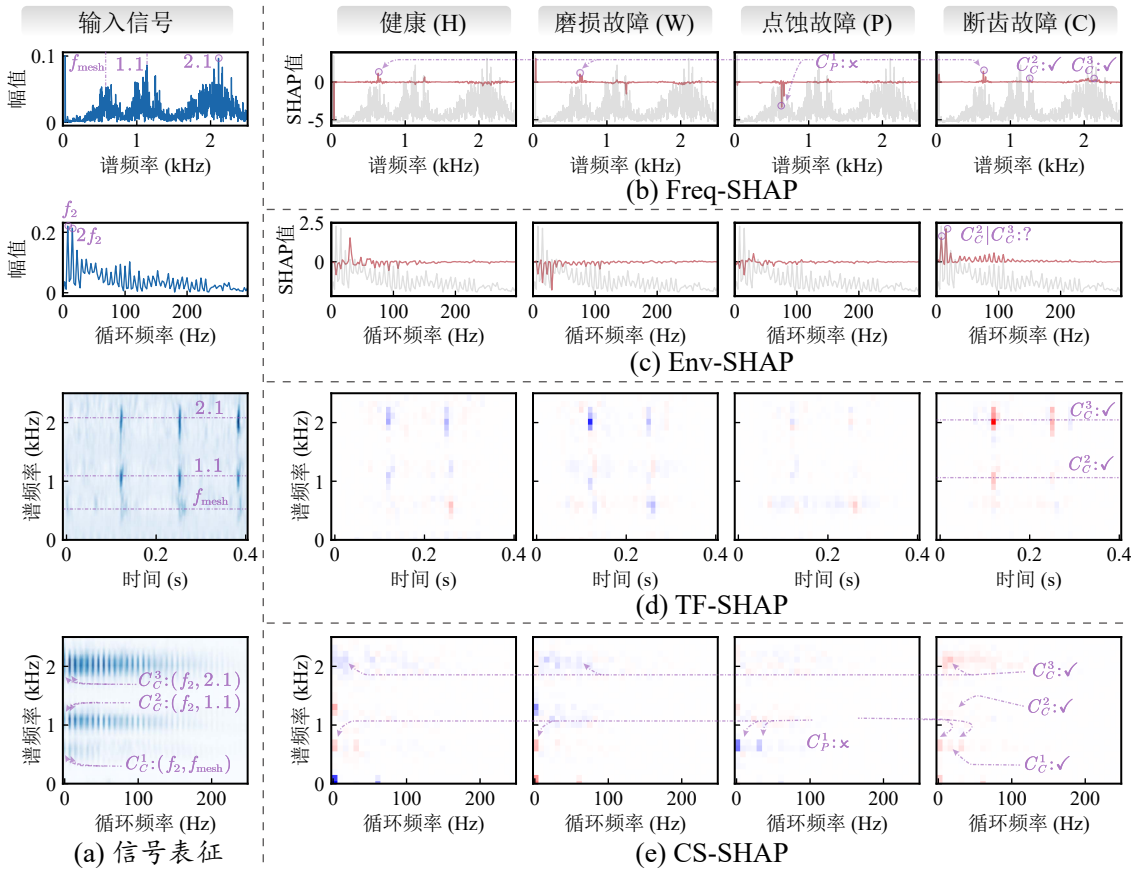


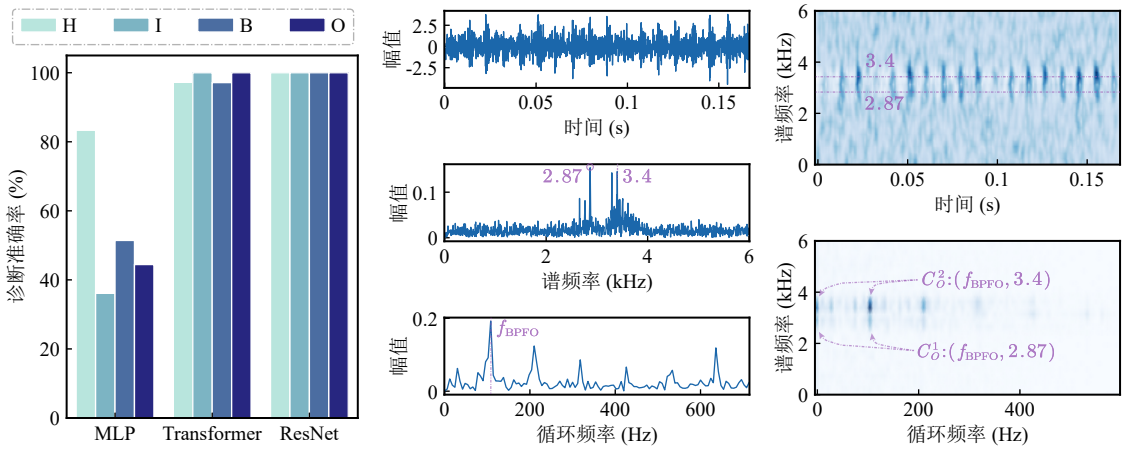
图 4-18　斜齿轮数据集下断裂故障样本的各域表征及不同归因方法的结果

Fig. 4-18　The domain representations of a tooth crack fault sample from the helical gearbox dataset and its attribution results using various methods

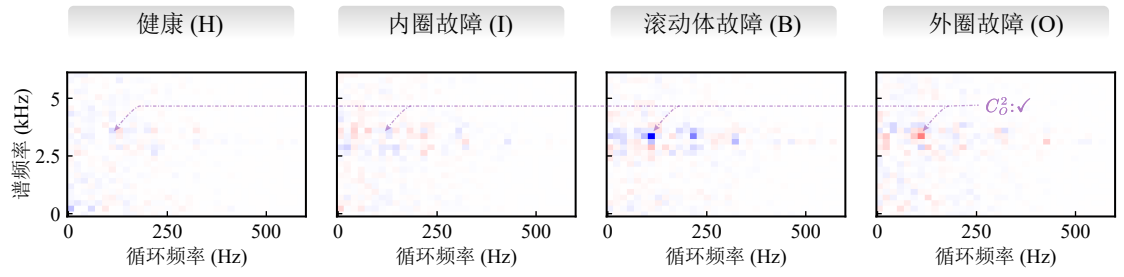
所示。MLP 的平均诊断准确率约为 50%, 其中健康类的准确率提高至 80%，这可能是由于健康类别具有显著的转频特征 nf_r。Transformer 和 ResNet 则表现出优异的性能，诊断准确率均接近 100%。在可解释性方面，三个模型的输入均选择统一的外圈故障样本，其多域表征如图 4-19(b) 所示，包含两个显著信号分量 C_O^1 和 C_O^2。

三种模型在外圈故障样本的 CS-SHAP 归因结果如图 4-19(c)-(e) 所示。MLP 较弱的诊断能力显著影响了 CS-SHAP 结果，其归因结果表明 C_O^1 的贡献几乎为零，C_O^2 虽然存在但却不够显著。图 4-19(d) 的 Transformer 由于诊断性能的提高，其 C_O^2 的贡献显著增，但 C_O^1 的贡献仍然很微弱。而且空白地方均存在些许幅值，表明 transformer 可能由于注意力机制导致结果的稳定性不佳。图 4-19(e) 所示的 ResNet 取得了最佳性能，其 CS-SHAP 结果种 C_O^1 和 C_O^2 均显示出明显的贡献，且有 $C_O^1 < C_O^2$。这与图 4-19(b) 中 C_O^1 和 C_O^2 的能量相符合。

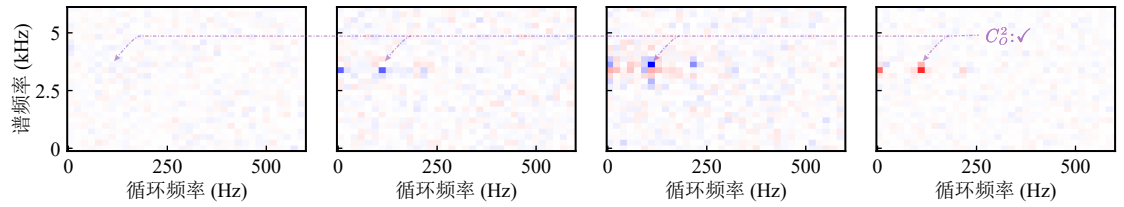
综上，CS-SHAP 的解释结果确实受到模型 $\mathcal{M}$ 差异的影响。一方面，较弱的模型



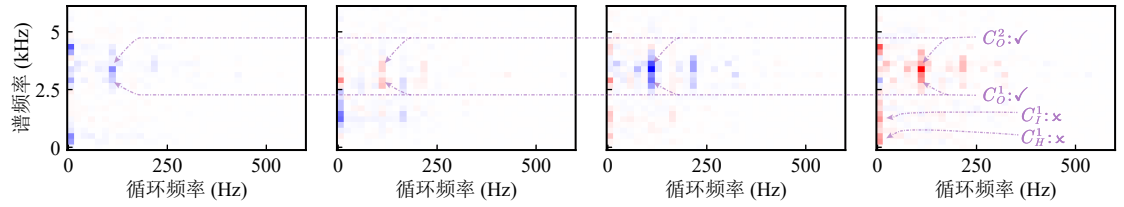
(a) 类别层级诊断准确率　　(b) 外圈故障样本的各域表征



(c) MLP



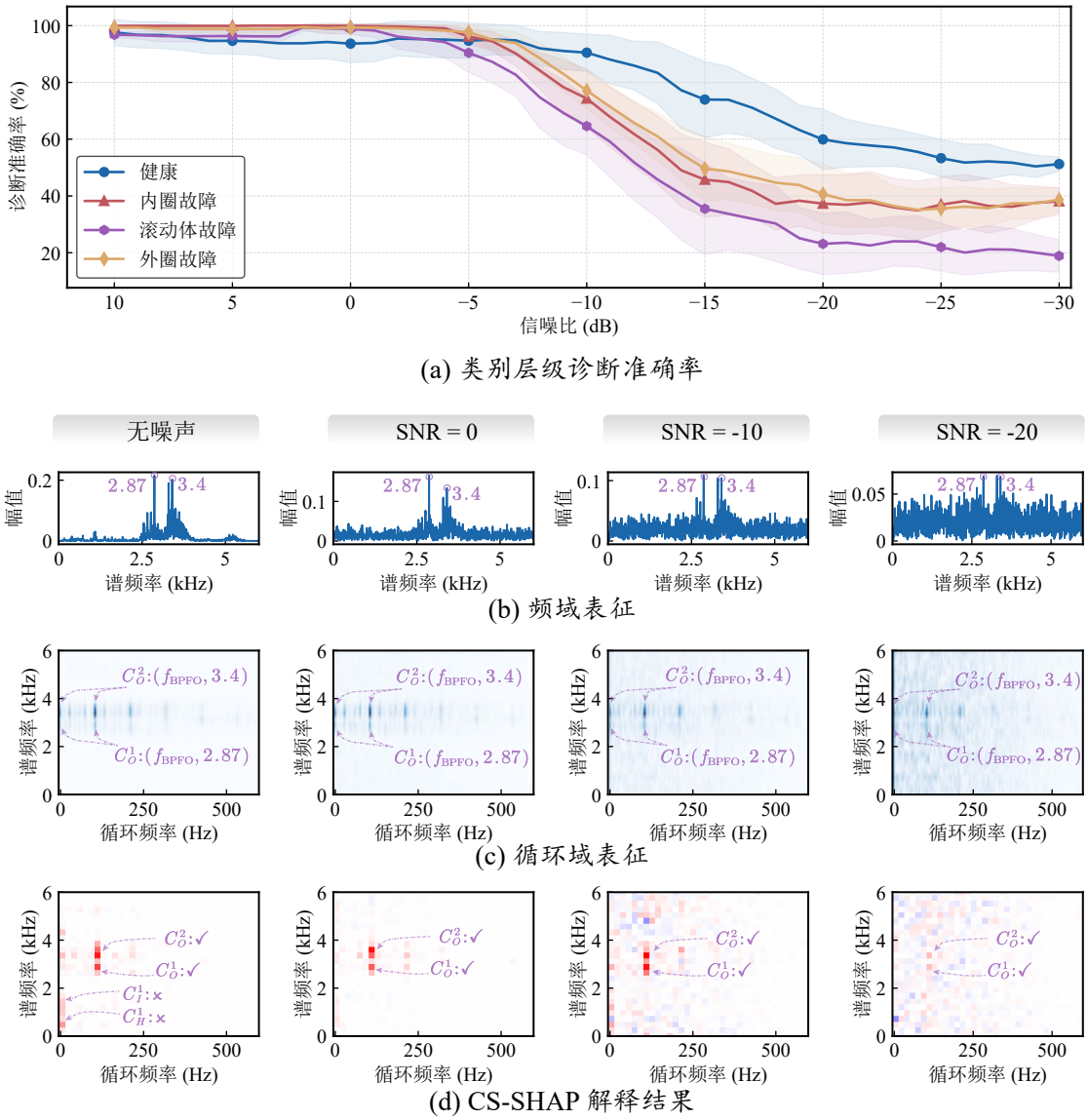
(d) Transformer



(e) ResNet

图 4-19　三类模型在 CWRU 数据集的类别层级诊断准确率、外圈故障样本各域表征和各模型在对应样本的 CS-SHAP 归因结果

Fig. 4-19　The class-wise test accuracies, domain representations of out race fault sample, and CS-SHAP results of three models under the CWRU dataset



(a) 类别层级诊断准确率

(b) 频域表征

(c) 循环域表征

(d) CS-SHAP 解释结果

图 4-20　CWRU 数据集不同 SNR 噪声下的类别层级诊断准确率及外圈故障的各域表征和 CS-SHAP 解释结果

Fig. 4-20　Class-wise diagnostic accuracy under different SNR noises in the CWRU dataset, along with cross-domain representations of outer race faults and CS-SHAP interpretation results

诊断能力会导致其预测结果缺少差异，进而使 CS-SHAP 难以有效捕捉各信号成分的贡献度。另一方面，模型类型也可能存在影响，例如，Transformer 具有特定的注意力机制，在擅长捕捉嵌入向量关系的同时，也可能导致归因结果的不稳定。尽管如此，CS-SHAP 仍然在不同模型 $\mathcal{M}$ 取得了总体上正确的解释结果，表明 CS-SHAP 的适用范围非常广泛，有望成为面向智能诊断模型的通用解释算法。

4.5.2　不同噪声强度下 CS-SHAP 的稳定性

基于表 4-1 中的卷积网络模型，CWRU 数据集在不同噪声强度下的测试准确率如图 4-20(a) 所示。随着噪声强度的增加，模型诊断准确率逐渐下降。在可解释性方面，实验统一地选择外圈故障样本，其频谱、循环域表征和 CS-SHAP 结果如图 4-20(b)-(e) 所示。

首先，随着噪声强度增加，频域表征的外圈故障信号分量 C_O^1 和 C_O^2 逐渐被噪声淹没，SNR = -20 噪声下便难以辨认。但循环域表征更为清晰，在 SNR = -20 噪声下仍能够辨认出 C_O^1 和 C_O^2，验证循环域在微弱特征提取方面的优势。此外，噪声强度的增加也使得模型难以做出准确的诊断预测，进而导致 CS-SHAP 得解释结果逐渐变得模糊。尽管如此，CS-SHAP 仍然能够持续揭示信号分量 C_O^1 和 C_O^2 对外圈故障类别 y_O 的正向贡献。即使在 SNR = -20 的情况下也能够模糊地识别。

综上，借助循环域表征强大的特征提取能力，CS-SHAP 表现出对噪声的极强的鲁棒性。即使在高噪声场景下，CS-SHAP 仍然能够识别出关键信号分量的贡献度。

4.6　本章小结

针对旋转机械智能诊断模型被动解释形式直观性不足的问题，本章系统地推导了面向确定性信号的循环域变换 $\mathcal{D}$ 及其逆变换 $\mathcal{D}^{-1}$，通过时域样本预处理以及逆变换与端到端模型的有机集成，构建了将解释形式拓展至循环域的 CS-SHAP 方法，以获得更为清晰准确的被动解释归因结果。基于仿真数据集和两个实测数据集的对比验证，本章的主要内容可总结如下：

(1) 建立了面向确定性信号的循环域变换方法。以面向随机信号的循环谱分析为基础，提出确定性信号二维自相关函数的近似估计，并系统推导了循环域变换 $\mathcal{D}$ 及其逆变换 $\mathcal{D}^{-1}$ 的计算过程。进而可知，通过对确定性信号 STFT 能量谱沿时间轴进行 Fourier 变换，即可获得信号在循环域的完整表征，为 CS-SHAP 的实现奠定了坚实的技术基础。

(2) 提出了将解释形式从传统时域拓展至循环域的 CS-SHAP 方法。基于域变换 $\mathcal{D}$ 及其逆变换 $\mathcal{D}^{-1}$，通过样本预处理和模型集成，将传统时域 SHAP 方法扩展至循环域。所提方法从谱频率和循环频率两个维度来量化各信号分量对模型决策的贡献，与旋转机械故障机理更为契合，可以有效提高解释结果的清晰度和准确性。

(3) 仿真数据集和实测数据集的实验结果表明，所提 CS-SHAP 方法能够精确量化

各信号分量对模型诊断决策的贡献程度。其载波频率和调制频率的双维度分析特性使归因解释结果更为清晰直观，且与故障已知的仿真数据集中信号成分与故障类别的对应关系完全一致。在存在多个谱特性相近的信号分量情况下，CS-SHAP 仍能借助双维度分析能力有效区分各成分的贡献，避免了因分量特征混淆而导致的错误归因，从而保证了解释结果的准确性和可靠性。此外，CS-SHAP 对不同学习能力的诊断模型均展现出良好的通用解释效果，并在高噪声环境下依然保持了显著的鲁棒性，体现了该方法在实际应用场景中的适用性。

第五章　针对振动信号高耗时计算的诊断模型被动解释效率优化

5.1　引言

CS-SHAP 在不修改端到端架构的前提下，将归因解释从时域扩展至故障区分能力更强的循环域，显著提升了解释结果的准确性与清晰度。然而，相较于传统特征归因场景，旋转机械智能诊断涉及的振动信号具有明显的高维特性，导致 CS-SHAP 计算呈现更为突出的耗时问题。其原因主要有二点：首先，CS-SHAP 的计算过程需对所有特征维度执行完整的子集枚举，使得计算复杂度呈指数级增长；其次，域变换方法虽能有效揭示故障特征，却往往使数据维度进一步膨胀，进而导致 SHAP 计算成本急剧上升。这种计算效率瓶颈严重限制了其在实时监测等场景的应用可行性，凸显了被动解释效率优化的迫切需求。

针对 SHAP 解释效率优化这一核心目标，本章从降低数据维度和优化计算复杂度两个关键方面入手，提出了组合块归因策略和 SHEP（SHapley Estimated exPlanation）算法。组合块归因策略通过将相邻特征组合成单个块，以更粗解释粒度为代价来降低特征维度；而 SHEP 算法则对复杂的子集枚举过程进行有效简化，通过两个具有代表性的过程来近似高耗时的 SHAP，成功地将计算复杂度从指数级降低至线性级。这两种方法的协同作用显著提升了 SHAP 的计算效率，为实时监测场景下的被动解释提供了可行性保障。

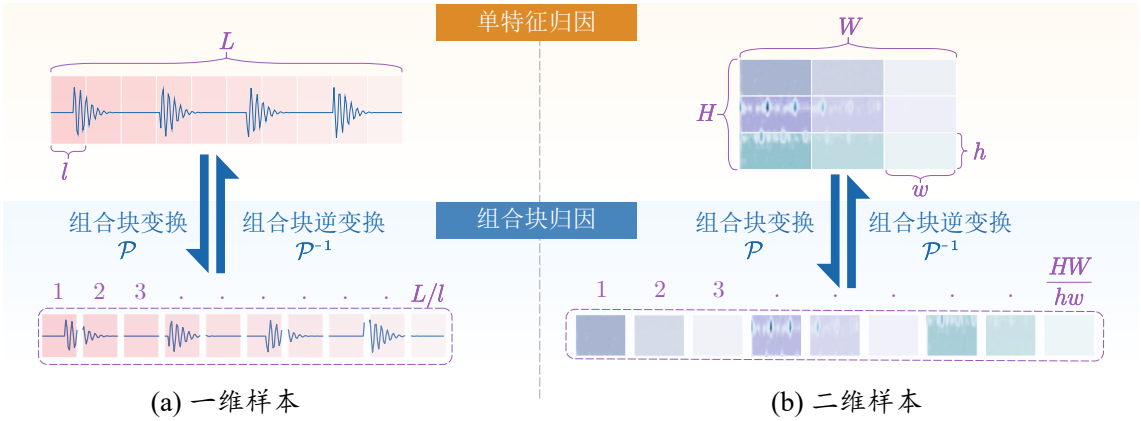
本章首先系统阐述组合块归因策略和 SHEP 算法的理论基础，随后利用仿真数据集进行效果对比、效率评估和可解释性验证，最后通过两个实测数据集验证 SHEP 算法在实际应用场景中解释结果的可靠性与有效性。本章的方法代码已开源在 *https://github.com/ChenQian0618/SHEP*。

5.2　用以降低特征维度的组合块归因策略

SHAP 的计算过程如式 (4-5) 所示，其计算成本主要包括两部分：估计数据分布的期望 $\mathbb{E}[\mathcal{M}(\boldsymbol{x})]$ 和枚举全集 $\boldsymbol{U}$ 的特征子集 $\boldsymbol{S}$。其中数据分布期望的计算成本相对较小，因为它与数据集样本数量 n 呈线性关系，即 $\mathcal{O}(n)$。此外，通过选择较小的代表性数据集可以降低这一部分的耗时。相反，子集枚举的计算成本则非常高，随着样本

维度 d 的增加呈指数级增长，即 $\mathcal{O}(2^d)$。尽管将枚举过程退化为置换方式的简化方法可以将复杂度降低到 $\mathcal{O}(2k_p d)$，其中 k_p 表示置换次数，但计算负担仍然显著[175,180]。

考虑到 SHAP 中子集枚举的复杂度为 $\mathcal{O}(2^d)$，降低样本的特征维度 d 是一种降低计算成本的显然方案。SHAP 计算中，原始样本中的每个一维点或二维像素都被视为一个独立的维度。为降低样本的特征维度，本节采用组合块归因策略，将多个相邻的一维点或二维像素绑定为组合块，以计算它们的联合贡献。这样的绑定操作称之为组合块变换 $\mathcal{P}$，其操作过程如图 5-1 所示。



(a) 一维样本　　(b) 二维样本

图 5-1　组合块变换的示意图

Fig. 5-1　Illustration of patch-wise transform

将组合块的尺寸记为 k，代表一维样本的长度或者二维样本的长和宽。那么，组合块变换 $\mathcal{P}$ 会按设定的 k 将相邻的一维点或二维像素等间隔地划分为尺寸为 k 的组合块，从而将原始样本 $\boldsymbol{x}$ 转换为组合块形式的样本 $\boldsymbol{p}$：

$$\mathcal{P}_k(\boldsymbol{x}) = \boldsymbol{p}. \tag{5-1}$$

$\boldsymbol{p}$ 中的每个组合块包含多个一维点或二维像素，并且都将作为一个单独的维度参与到 SHAP 计算中。此外，组合块变换具有完全可逆性，逆组合块变换 $\mathcal{P}^{-1}$ 可以通过拼接操作，将这些块还原至原始样本：

$$\mathcal{P}^{-1}(\boldsymbol{p}) = \boldsymbol{x}. \tag{5-2}$$

组合块归因策略通过降低特征维度 d 来提高 SHAP 的计算效率，但这种提升是以归因结果的粒度为代价的。原始样本的特征维度为 d，计算复杂度变为 $\mathcal{O}(2^d)$，可以获得每个一维点或二维像素的贡献，其解释粒度为 1。通过尺寸为 k 的组合块变换 $\mathcal{P}$，样本维度变为 d/k，计算复杂度也同步降低至 $\mathcal{O}(2^{d/k})$，但解释粒度会升高至 k，

即分辨率变为原来的 k 倍。然而，解释粒度的升高并不影响组合块归因策略的实际价值。一方面，解释粒度线性劣势是远小于计算成本指数劣势，仍具有竞争力；另一方面，可以通过调整组合块尺寸来对解释粒度和计算成本进行平衡，提供了可选方案。由此，在具体应用中可以根据实际情况灵活调整组合块的大小，在计算效率和解释精度之间找到最佳平衡点。

5.3　用以降低归因计算复杂度的 SHEP 算法

SHAP 计算需要对输入样本的所有特征维度进行完整的子集枚举，由此引发对完整枚举必要性的思考。换言之，仅考虑少数典型子集是否也能获得令人满意的归因解释结果。由此，本文提出了 SHEP 来对 SHAP 作近似，以降低归因计算复杂度。SHEP 由式 (4-5) 所示子集枚举过程中的两个关键子集组成，分别称之为 SHEP-Remove 和 SHEP-Add。

SHEP-Remove 的概念与主流的基于扰动的归因方法一致，其计算过程如图 5-2(a) 所示。具体而言，HEP-Remove 从当前样本 $\tilde{\boldsymbol{x}}$ 中移除指定特征（或组合块）$\tilde{\boldsymbol{x}}_i$，并通过模型输出的变化来衡量该特征的贡献。与常见的掩码、模糊或缩放等技术不同，移去特征的操作遵循了如式 (4-3) 的 SHAP 框架。移除特征 $\tilde{\boldsymbol{x}}_i$ 意味着将样本 $\tilde{\boldsymbol{x}}$ 中的该特征退化为数据分布 $\boldsymbol{X}_i$，从而获得分析样本 $\tilde{\boldsymbol{x}}^{\boldsymbol{U}\setminus\{i\}}$，其中 U 表示完整的特征维度集合。由此，SHEP-Remove 归因的计算可表示为

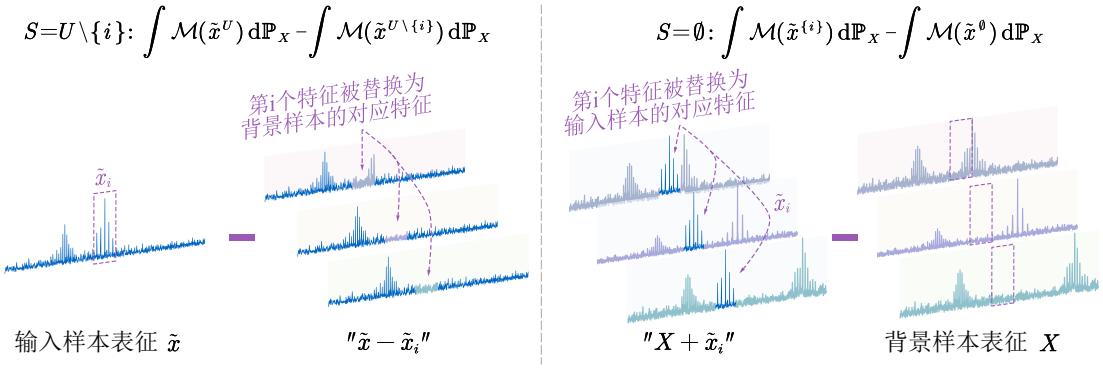
$$\begin{aligned}\psi^{\mathrm{Rm}}_{\mathcal{M},\boldsymbol{X}}(\tilde{\boldsymbol{x}})_i &= \mathcal{M}(\tilde{\boldsymbol{x}}) - \mathbb{E}\left[\mathcal{M}\left(\tilde{\boldsymbol{x}}^{\boldsymbol{U}\setminus\{i\}}\right)\right] \\ &= \mathbb{E}\left[\mathcal{M}(\tilde{\boldsymbol{x}}^{\boldsymbol{U}})\right] - \mathbb{E}\left[\mathcal{M}\left(\tilde{\boldsymbol{x}}^{\boldsymbol{U}\setminus\{i\}}\right)\right].\end{aligned} \tag{5-3}$$

显然，SHEP-Remove 是式 (4-5) 所示 SHAP 计算中 $\boldsymbol{S} = \boldsymbol{U} \setminus \{i\}$ 的特例。

不同于基于输入样本视角的 SHEP-Remove，SHEP-Add 则是从数据分布视角出发。如图 5-2(b) 所示，SHEP-Add 通过将输入样本的特征 $\tilde{x}_i$ 添加到从数据分布中采样出的背景样本 $\boldsymbol{X}$ 上，从而获得分析样本 $\tilde{\boldsymbol{x}}^{\{i\}}$，并同样地通过模型输出的变化来衡量该特征的贡献度 $\psi^{\mathrm{Add}}_{\mathcal{M},\boldsymbol{X}}(\tilde{\boldsymbol{x}})_i$。SHEP-Add 归因的计算可表示为

$$\begin{aligned}\psi^{\mathrm{Add}}_{\mathcal{M},\boldsymbol{X}}(\tilde{\boldsymbol{x}})_i &= \mathbb{E}\left[\mathcal{M}\left(\tilde{\boldsymbol{x}}^{\{i\}}\right) - \mathcal{M}(\boldsymbol{X})\right] \\ &= \mathbb{E}\left[\mathcal{M}(\tilde{\boldsymbol{x}}^{\{i\}})\right] - \mathbb{E}\left[\mathcal{M}\left(\tilde{\boldsymbol{x}}^{\emptyset}\right)\right].\end{aligned} \tag{5-4}$$

不难看出，SHEP-Add 同样是式 (4-5) 所示 SHAP 计算中的特例，其中集合 $\boldsymbol{S}$ 为空集 $\boldsymbol{S} = \emptyset$。



(a) SHEP-Remove　　(b) SHEP-Add

图 5-2　SHEP-Remove 和 SHEP-Add 的计算过程

Fig. 5-2　The calculation process of SHEP-Remove and SHEP-Add

最终，SHEP 的输出可以表示为这两种子集情况的组合：

$$\psi^{\mathrm{SHEP}}_{\mathcal{M},\boldsymbol{X}}(\tilde{\boldsymbol{x}})_i = \frac{\psi^{\mathrm{Rm}}_{\mathcal{M},\boldsymbol{X}}(\tilde{\boldsymbol{x}})_i + \psi^{\mathrm{Add}}_{\mathcal{M},\boldsymbol{X}}(\tilde{\boldsymbol{x}})_i}{2}. \tag{5-5}$$

SHAP 的子集枚举过程存在大量可选子集，而本章选择 SHEP-Remove 和 SHEP-Add 的原因在于归因解释的基本逻辑。样本对预测类别的正向贡献不仅来自于该类别特有特征的存在，还来自于其他类别相关特征的缺失。一个有效的归因方法必须同时考虑这两个因素。事实上，由于神经网络的性能优势，其预测结果通常是稀疏的，即主导类别的预测概率极高而其他类别的预测概率极低。式 (5-3) 和 (5-4) 所示的分析样本由于特征重组，原本的预测结果稀疏性被破坏，导致主导类别预测概率的显著下降和其他类别的微弱上升，而非预测结果简单叠加。

由此，以当前样本为视角的 SHEP-Remove 更善于捕捉输入样本中特征存在的影响，这与传统的基于扰动的归因方法理念一致。另一方面，以背景样本为视角的 SHEP-Add 则更有效地捕捉背景样本中特征缺失的影响。而 SHEP 通过结合这两种方法的优势，能够提供更为全面的特征贡献评估。所提出的 SHEP，很好地结合 SHEP-Remove 和 SHEP-Add 的优势，能够更全面地衡量特征贡献度。此外，本章将在后续实验部分进行更深入的分析，来验证这一逻辑的正确性。

5.4　组合块归因和 SHEP 相结合的高效率智能诊断被动解释流程

在应用上，可以将本章的组合块归因和 SHEP 算法融入上一章图 4-4 所示的解释框架，从而建立一个解释效果清晰、计算效率高的智能诊断模型被动解释方案。其中，域变换方法 $\mathcal{D}$ 将解释载体从时域扩展到其他更为清晰的域，有效地揭示信号各成分

的贡献度。同时，组合块变换 $\mathcal{P}$ 和 SHEP 则分别用于降低特征维度、算法复杂度，从而提高归因计算速度，为实时监测场景下的模型解释提供基础。

有必要说明的是，域变换方法 $\mathcal{D}$ 不局限于上一章所述的循环域，应用流程中仍可考虑基础的频域、包络域和时频域来保证分析方法的多样性。四类域变换方法获取信号表征和残余信息的计算公式如表 5-1 所示，各种域变换从不同角度刻画故障信号特征，可由任务特定加以选择。其中，频域关注常见的谱频率，包络域关注调制频率，时频域关注谱频率和时间维度，循环域则同时识别载波频率和调制频率。

表 5-1　四类域变换方法获取信号表征和残余信息的计算公式

Table 5-1　The calculation of representations and remains across four domain transforms

域	信号表征[a] (z)	残余信息[b] (r)
频域	$X(f)=\int x(t)\mathrm{e}^{-\mathrm{i}2\pi ft}\,\mathrm{d}t,$ $z=\lvert X(f)\rvert^2$	$r=\mathrm{Ang}[X(f)]$
包络域	$z(t)=x(t)+\mathrm{i}\hat{x}(t),\ \hat{x}(t)=\frac{1}{\pi}\int\frac{x(\tau)}{t-\tau}\,\mathrm{d}t,$ $Z(f)=\int\left(\lvert z(t)\rvert-\overline{\lvert z(t)\rvert}\right)\mathrm{e}^{-\mathrm{i}2\pi ft}\,\mathrm{d}t,$ $z=\lvert Z(f)\rvert^2$	$r_1=\mathrm{Ang}(z(t)),$ $r_2=\overline{\lvert z(t)\rvert},$ $r_3=\mathrm{Ang}[Z(f)]$
时频域	$Z(f,t)=\int x(\tau-t)w(\tau)\mathrm{e}^{-\mathrm{i}2\pi f\tau}\mathrm{d}\tau,$ $z=\lvert Z(f,t)\rvert^2$	$r=\mathrm{Ang}[Z(f,t)]$
循环域	$Z(f,t)=\int x(\tau-t)w(\tau)\mathrm{e}^{-\mathrm{i}2\pi f\tau}\mathrm{d}\tau,$ $Z(f,\alpha)=\int\lvert Z(f,t)\rvert^2\mathrm{e}^{-\mathrm{i}2\pi\alpha t}\mathrm{d}t,$ $z=\lvert Z(f,\alpha)\rvert^2$	$r_1=\mathrm{Ang}[Z(f,t)],$ $r_2=\mathrm{Ang}[Z(f,\alpha)]$

[a] $w(t)$ 表示 STFT 中使用的窗函数。
[b] $\mathrm{Ang}(\cdot)$ 表示提取复数相位的操作。

将组合块变换和 SHEP 应用于旋转机械智能诊断模型被动解释的流程图如图 5-3 所示，包括样本预处理、诊断模型集成、SHEP 归因、以及可视化和可解释性分析四部分。在样本预处理阶段，依次使用域变换 $\mathcal{D}$ 和组合块变换 $\mathcal{P}$ 对输入样本 $\boldsymbol{x}$ 进行预处理，获得组合块形式的目标域表征 $\boldsymbol{p}$，可表示为

$$\boldsymbol{p}=\mathcal{P}\circ\mathcal{D}(\boldsymbol{x})=\mathcal{P}\left[\mathcal{D}(\boldsymbol{x})\right]. \tag{5-6}$$

与此同时，模型集成阶段则将域逆变换 $\mathcal{D}^{-1}$、组合块逆变换 $\mathcal{D}^{-1}$ 与待分析的端到端智能诊断模型 $\mathcal{M}$ 进行依次组合，构建集成模型 $\tilde{\mathcal{M}}$：

$$\tilde{\mathcal{M}}=\mathcal{M}\circ\mathcal{D}^{-1}\circ\mathcal{P}^{-1}. \tag{5-7}$$

在 SHEP 归因阶段，首先用公式 (5-3)-(5-5) 计算出组合块形式的目标域归因结果 $\psi_{\tilde{\mathcal{M}}}(\boldsymbol{p})$，然后通过组合块逆变换 $\mathcal{D}^{-1}$ 将其恢复至目标域的归因结果 $\psi_{\tilde{\mathcal{M}}}(\boldsymbol{z})$，最后

基于归因结果 $\psi_{\tilde{\mathcal{M}}}(z)$ 进行可视化并开展可解释性分析，可解释性分析的具体流程与上一章一致。

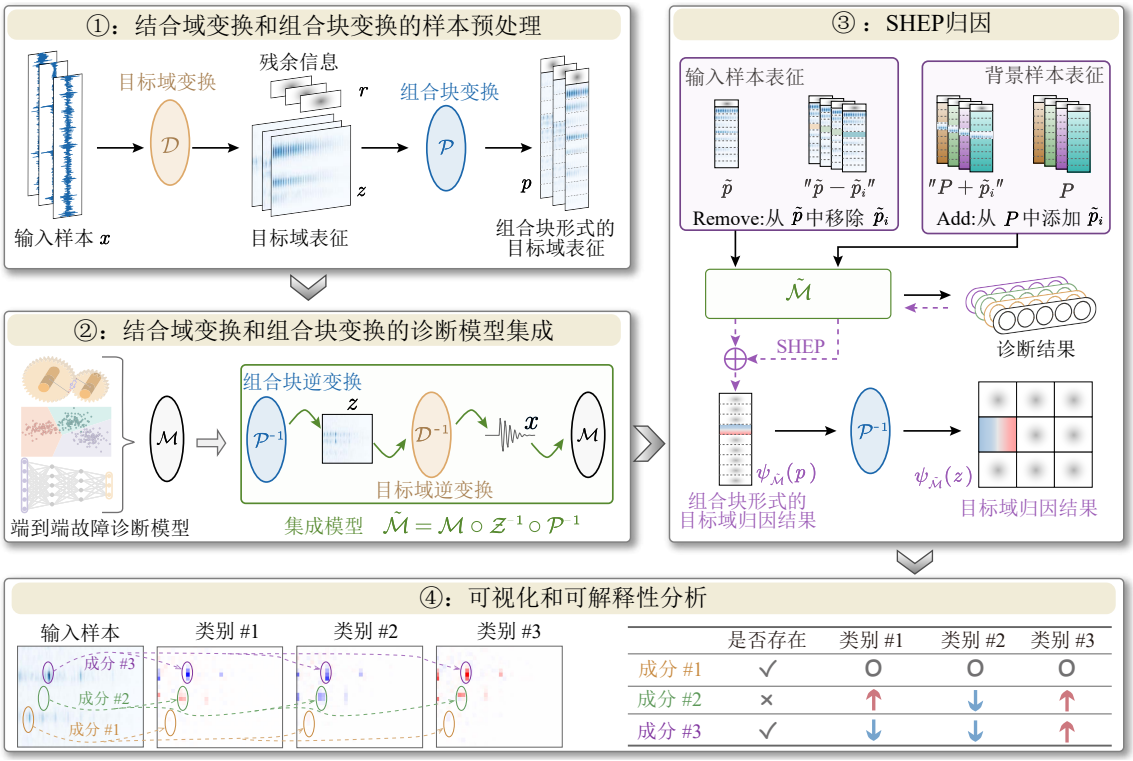


图 5-3　将组合块变换和 SHEP 应用于旋转机械智能诊断模型被动解释的流程图

Fig. 5-3　The flowchart of applying patch transform and SHEP to passive explanation of intelligent diagnosis model for rotating machinery

5.5　仿真场景下 SHEP 被动解释的全面验证和分析

故障逻辑完全可知的仿真数据集具有真实的解释性标签，能够更为有效地评估解释效果。因此，本节将基于仿真数据集对组合块归因策略和 SHEP 开展一系列分析。具体而言，首先讨论组合块归因测略的影响，然后验证 SHEP-Remove 和 SHEP-Add 的理论假设，最后从解释效果和计算效率两个方面与其他方法进行对比。实验中选择上一章所使用的端到端卷积神经网络作为预测模型 $\mathcal{M}$ 进行分析，其具体架构见表 4-1。

5.5.1　故障逻辑已知的仿真数据集及实验参数介绍

仿真数据集的设置和 4.4.1 小节保持一致，其采样频率为 10 kHz，包含三种故障类型：健康（H）、故障 #1（F1）和故障 #2（F2）。表 4-2 详细列出了它们与各信号

成分的关系及其对应参数，每种故障类型由两个周期脉冲成分构成。为了便于理解，图 5-4 给出了这三种故障类型在多个域中的表征。值得注意的是，信号成分 C_0 在所有三个类别中均存在，因此对任何类别都不具有贡献。与之相反，其他信号成分（即 C_H、C_1、C_2）均专属于单个类别，对其对应类别产生正向贡献的同时，而对其他类别产生负向贡献。

实验的训练参数和 4.4.1 小节保持一致，训练后的卷积网络模型在测试集的诊断准确率为 99.98%。

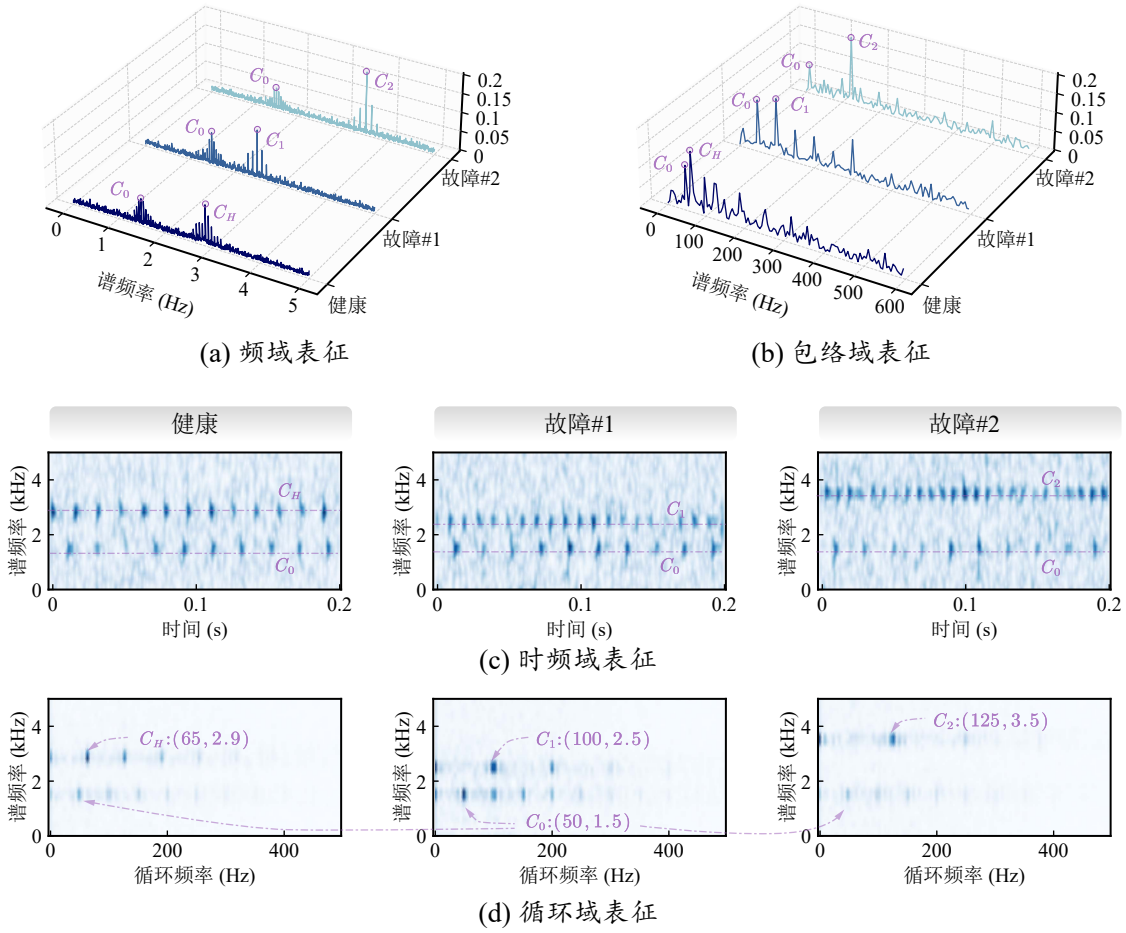


图 5-4　仿真数据集下各类样本在不同域的表征

Fig. 5-4　The representation of each class sample in different domains of the simulation dataset

5.5.2　组合块归因策略的解释效果分析

与传统 SHAP 归因相比，组合块归因策略通过降低特征维度显著降低了计算复杂度，以解释结果的粒度为代价，能够通过降低特征维度来显著降低计算复杂度。现

对组合块尺寸对解释粒度的影响进行分析，而组合块尺寸与计算效率之间的关系将在后续部分进行统一讨论。

为方便分析，现按照五档级别，定义了如表 5-2 所示的不同域下组合块尺寸。随着组合块级别的提高，块的尺寸逐渐增加，而特征维度则相应减少。需要注意的是，一维域（如频域和包络域）和二维域（如时频域和循环谱域）在组合块的设计上存在差异。

表 5-2　不同域下不同组合块级别的尺寸设置及对应特征维度

Table 5-2　The setting of patch size in different domains and corresponding feature dimensions

组合块级别	频域	包络谱域	时频域	循环域
无	1&1001[a]	4&120	1&26×205	2&26×103
#1	(3)-335[b]	(1)-124	(1,5)-1067[c]	(1,3)-912
#2	(6)-168	(2)-60	(2,5)-534	(2,3)-457
#3	(12)-84	(4)-34	(2,10)-274	(2,6)-236
#4	(24)-43	(8)-19	(2,20)-148	(4,6)-128
#5	(48)-22	(16)-12	(4,20)-78	(4,12)-65

[a] *p*&*q*: 表示残余信息和信号表征的维度分别为 *p* 和 *q*。
[b] (*l*)-*x*: 表示一维输入情况下，组合块的大小为 *l*，组合出的特征维度为 *x*。
[c] (*h*,*w*)-*x*: 表示二维输入情况下，组合块的高度和宽度分别为 *h* 和 *w*，组合出的特征维度为 *x*。

为了全面评估组合块尺寸对归因结果的影响，现以故障 #2 类别的样本作为输入，分析其在不同域和不同组合块尺寸下的 SHEP 结果，如图 5-5 所示。从结果来看，不同域的归因有着不同的侧重点，包括谱频率、循环频率和冲击时间。域的差异在上一章已详细讨论，其选择取决于任务的需求，并非本章的关注重点。

在组合块尺寸的影响方面，随着组合块级别的提高，所有域中的解释粒度均显著变粗，但都获得了正确的归因结果。SHEP 方法准确地将模型对故障 #2 类别 y_2 的预测结果归因于信号成分 C_2 的存在（C_2:✓）和信号成分 C_1 的缺失（C_1:✗），同时信号成分 C_0 的存在（C_0:✓）并未产生任何贡献。这种归因结果与仿真数据集的预设故障逻辑完全一致，证实了组合块归因的正确性。

然而，解释粒度的上升也导致不同成分贡献度难以区分。例如，组合块级别为 #1 甚至 #3 时，C_2:✓ 和 C_1:✗ 的贡献仍然可以被有效区分，而在组合块级别为 #5 时，这些信号成分的贡献变得难以区分。这种区分能力的降低源于较大的组合块将原本独立的特征维度合并在一起，导致归因结果的分辨率下降。

这种影响在二维域中表现得更为明显。随着组合块尺寸的增加，时频域和循环谱

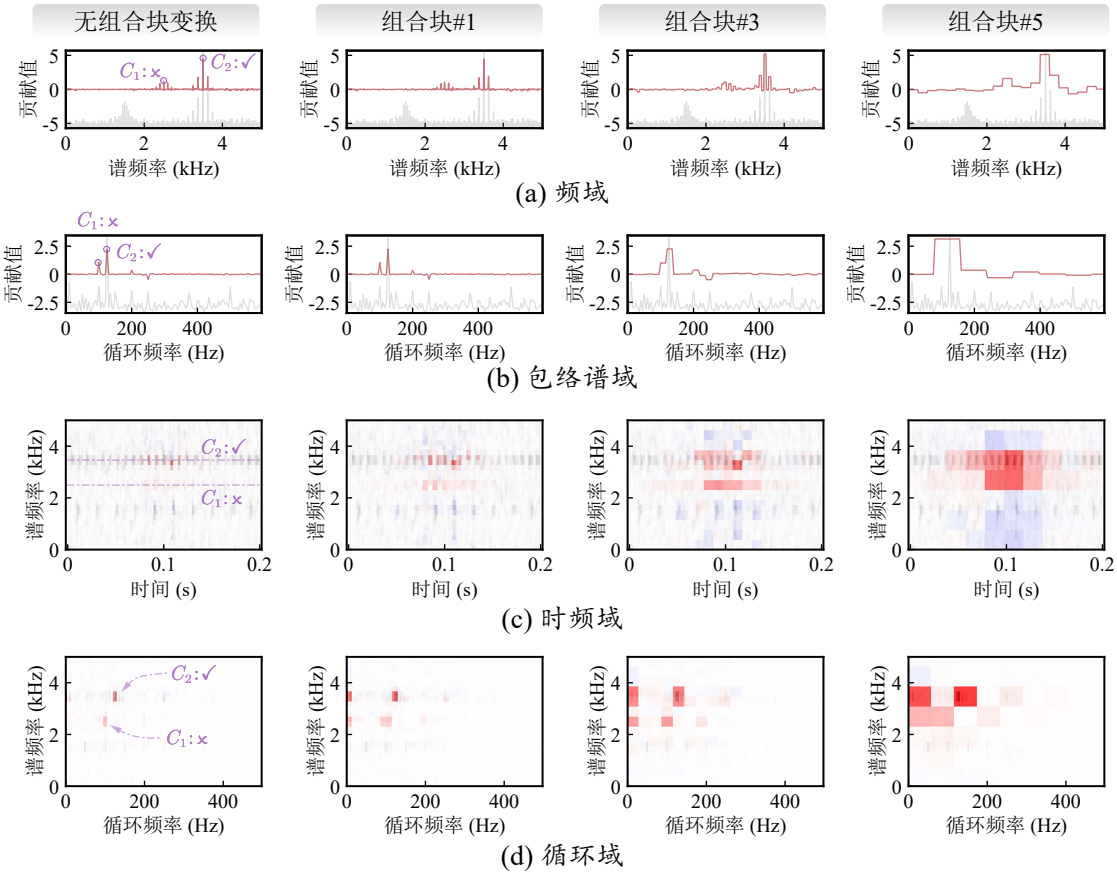


图 5-5　不同域变换、不同组合块尺寸下以故障 #2 样本为输入、故障 #2 类别为输出的 SHEP 归因结果

Fig. 5-5　The SHEP attribution results with different domain transforms and different patch sizes, with the input as the Fault #2 sample and the output as the Fault #2 class

域中的局部特征逐渐被平均化，使得细节信息逐渐丢失。这一现象提示本章在实际应用中需要谨慎选择组合块尺寸，在计算效率和解释精度之间寻求平衡。过大的组合块虽然能显著降低计算复杂度，但可能会严重影响解释结果的质量，不利于深入理解故障机理。

综上，随着组合块尺寸的增加，解释结果的粒度也随之增大。它不影响解释结果的正确性，但会导致不同成分的贡献度难以区分。基于这一分析，实际应用中避免使用过大的组合块尺寸，以免导致解释结果的显著恶化。在具体实践中，组合块尺寸的选择应当综合考虑计算资源约束、实时性要求和解释粒度需求等多个因素，以实现最优的性能平衡。

5.5.3　SHEP-Add 和 SHEP-Remove 的解释效果分析

5.3 节介绍了 SHEP-Remove 和 SHEP-Add 的理论假设，现通过实验对其进行验证。以故障 #2 类别的样本作为输入，SHEP-Add、SHEP-Remove、SHEP 和 SHAP 四种归因方法对不同预测类别下的归因结果如图 5-6 所示。为了理解背后的原理，图中将背景样本的类别作为 y 轴变量，分析了不同类别背景样本 $\boldsymbol{X}$ 的归因贡献，其中的“均值” 代表所有背景样本结果的平均值，这也是最终各方法输出的归因结果。

图 5-6(a) 所示的 SHEP-Remove 从输入样本 $\tilde{\boldsymbol{x}}$ 的视角进行归因。输入样本 $\tilde{\boldsymbol{x}}$ 的类别为为 F2，它包含各类共有的信号成分 C_0 独有的信号成分 C_2，并对预测类别 y_2 具有稀疏性特征。从 y 轴的背景样本类别来看，同类（F2）背景样本的贡献为零，其他类别（H 和 F1）的背景样本才具有贡献，这与同类样本特征交换不影响预测结果的理论预期相一致。从 x 轴的谱频率 f 而言，对 C_2 对应频率的特征交换破坏了输入样本 $\tilde{\boldsymbol{x}}$ 对 F2 预测类别的稀疏性，导致 y_2 显著降低的同时，其他类别的预测概率不具区分性地上升。相比之下，其他频率（包括 C_0 和 C_1 对应频率）的特征交换几乎不影响预测结果，从而不具有贡献。综上，SHEP-Remove 擅长捕捉当前样本 $\tilde{\boldsymbol{x}}$ 中存在特征（如 C_2）的影响。

与之相反，图 5-6(b) 所示的 SHEP-Add 则从背景样本 $\tilde{\boldsymbol{X}}$ 的视角进行归因。不同类别的背景样本表现出截然不同的行为特征。对于健康类别的背景样本，其独有信号成分 C_H 的贡献由于随机性可忽略。对于 F1 类别的背景样本，将其独有信号成分 C_1 对应频率的特征替换为输入样本 $\tilde{\boldsymbol{x}}$ 的特征，同样破坏了其对 F1 预测类别 y_1 的预测性，因此 F1 类别背景样本的 C_1 对应频率会表现出出明显贡献。F2 类别的背景样本由于与输入样本 $\tilde{\boldsymbol{x}}$ 属于同一类别，显然也不具有贡献。综上，SHEP-Add 擅长捕捉不存在于当前样本 $\tilde{\boldsymbol{X}}$ 但存在于背景样本 $\tilde{\boldsymbol{X}}$ 中的特征（如 C_1）。

SHEP 对不同预测类别下的归因结果图 5-6(c) 所示。它通过结合 SHEP-Remove 和 SHEP-Add 的优势，有效地捕捉了当前样本 $\tilde{\boldsymbol{x}}$ 中特征存在和背景样本 $\tilde{\boldsymbol{X}}$ 中特征缺失的双重影响。其归因结果与如图 5-6(d) 所示的原始 SHAP 结果高度一致。这种一致性不仅体现在平均结果上，还体现在不同背景样本类别的贡献上。这充分证明了 SHEP 不仅具有出色的归因性能，同时也是 SHAP 的有效近似。此外，这一结果也验证了 5.3 节提出的理论假设，即通过组合基于当前样本和背景样本的两种视角，可以获得更全面和准确的特征贡献评估。

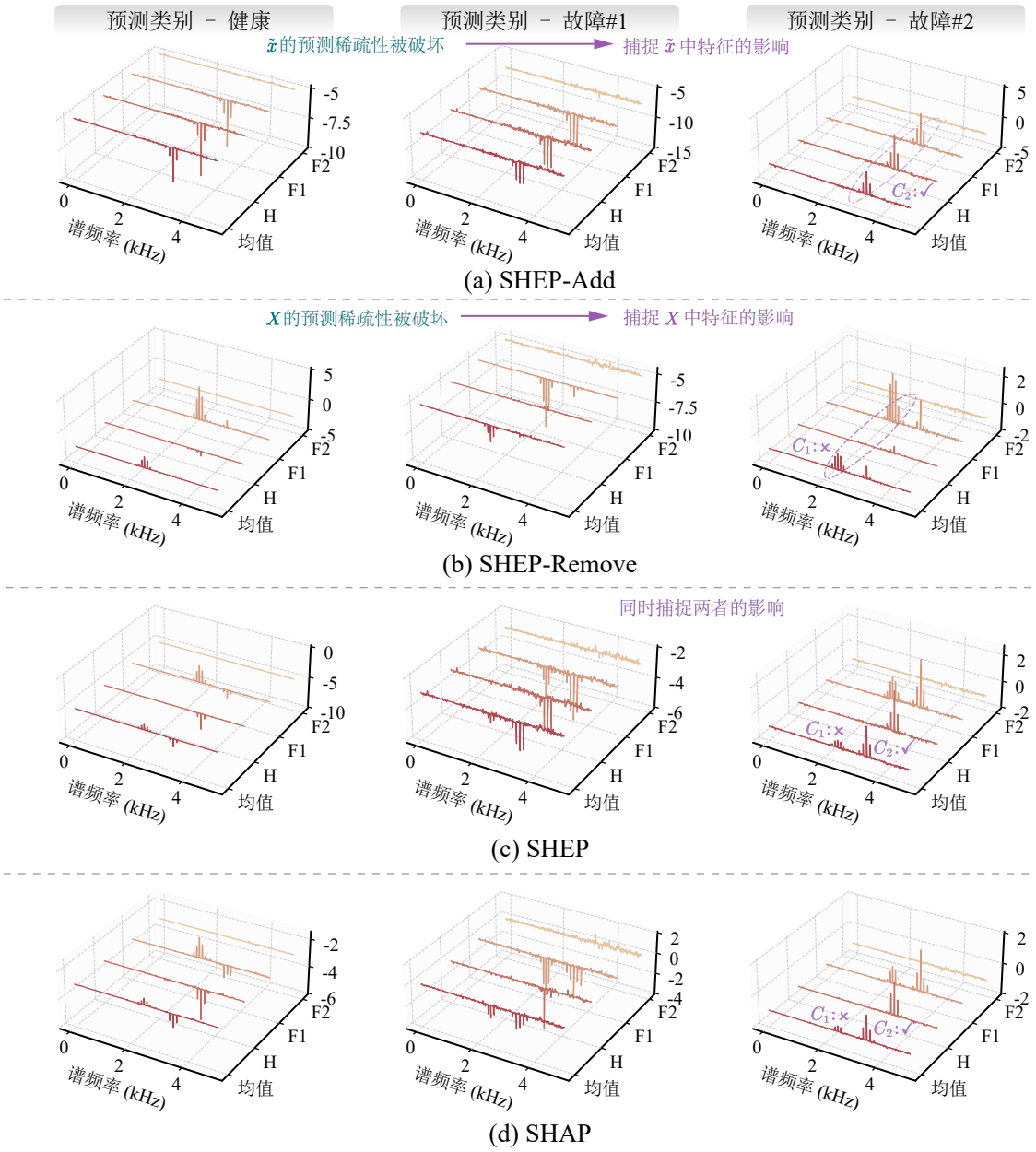


图 5-6　输入样本 $\tilde{x}$ 为 F2 时四种归因方法对不同预测类别下的归因结果

Fig. 5-6　The attribution results of four attribution methods for different predicted classes with Fault #2 as the input sample $\tilde{x}$

5.5.4　SHEP 和同类方法的解释效果对比

常见的归因解释方法可分为三类：基于模型简化的方法（如 LIME[76]）、基于梯度的方法（如 Grad-CAM[71]、SmoothGrad[181]）以及基于扰动的方法（即 Mask[81]和 Scale[133]）。然而，模型简化类方法难以处理具有异构特征的组合块样本，而基于梯度

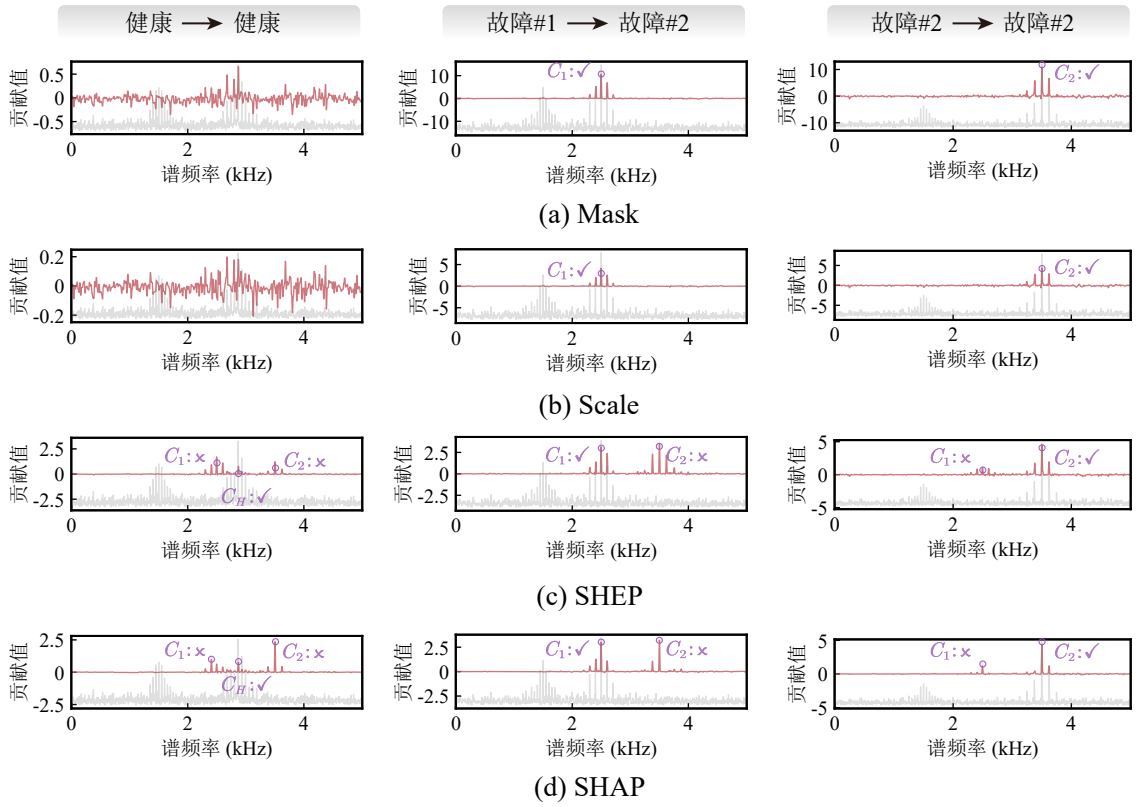


图 5-7　组合块为 #1 时不同归因方法对仿真数据集不同类别样本的对应预测类别频域归因结果

Fig. 5-7　Attribution results of different attribution methods in the frequency domain for different class samples towards thier corresponding predicted class in the simulation dataset under patch size level #1

的方法则难以适用于可能破坏梯度传播的域变换。为了确保公平性，实验中选择与域变换和组合块变换相兼容的扰动类归因解释方法作为对比，同时也将原始 SHAP（使用置换模式，置换次数设置为 $k_p = 5$）的结果作为评价基准。

将组合块级别设置为 #1，不同归因方法对仿真数据集中不同类别样本的对应预测类别频域归因和循环域归因，分别如图 5-7 和图 5-8 所示，图中已对各信号成分进行标注以方便理解。从图 5-7 和图 5-8(a)-(b) 中可以看出，Mask 和 Scale 作为基于扰动的方法，和图 5-6 所示的 SHAP-Remove 具有类似的行为特征。它们仅仅关注于当前输入样本中信号成分的存在所带来的贡献，而忽视了其他类别信号成分的缺失所带来的贡献。这种单一视角的归因方式难以全面反映特征对预测结果的影响机制。

相比之下，图 5-7 和图 5-8(c)-(d) 所示的 SHAP 和 SHEP，能够有效地捕捉这两种类型的贡献，从而提供更加全面和准确的归因结果。独属于健康类别的信号成分 C_0，具有如表 4-2 所示的随机参数，Mask 和 Scale 方法难以准确捕捉其贡献。然而，SHEP

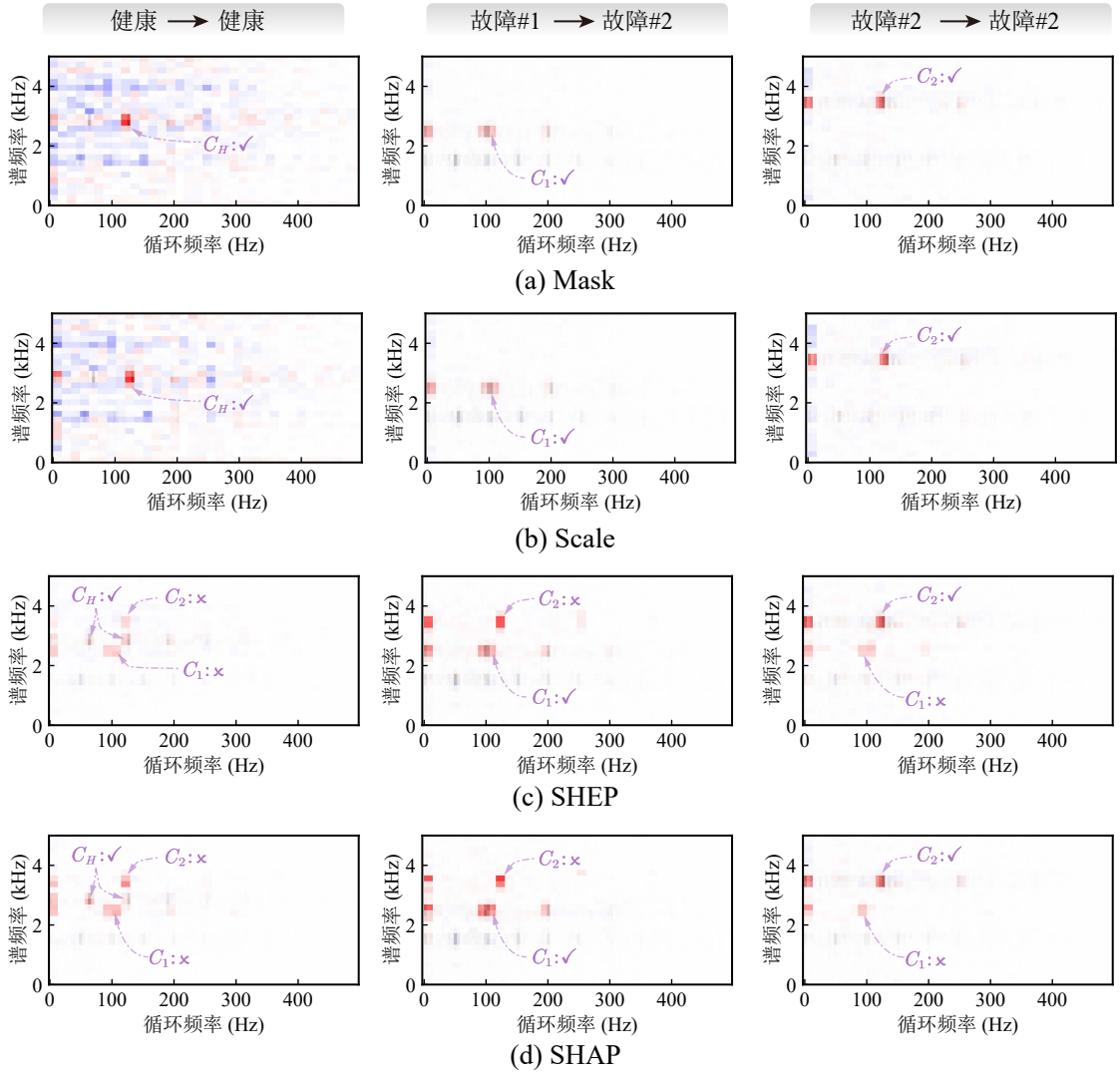


图 5-8　组合块为 #1 时不同归因方法对仿真数据集不同类别样本的对应预测类别循环域归因结果

Fig. 5-8　Attribution results of different attribution methods in the cyclic-stationary domain for different class samples towards thier corresponding predicted class in the simulation dataset under patch size level #1

和 SHAP 仍然能够相对准确地反映这种微弱的贡献，这充分展示了它们更强大的归因能力。这一点对于实际应用具有重要意义，因为在实际故障诊断中，信号往往包含不确定性，需要归因方法具有较强的微弱特征解释能力。

从深层次来看，这种差异主要源于不同方法在归因机制设计上的差异。基于扰动的方法通过直接观察特征扰动对模型输出的影响来评估特征重要性，这种简单直接的机制虽然计算效率较高，但难以捕捉特征之间的复杂交互关系。而 SHAP 和 SHEP

则通过系统性地评估特征组合的影响来归因，这种基于博弈论的方法能够更全面地考虑特征的协同效应，从而提供更准确的归因结果。

尽管图 5-7 和图 5-8 中的可视化结果非常直观，但这还不足以全面评估各归因方法解释结果的优劣。为实现定量评估，现以 SHAP 结果为基准，将各方法与 SHAP 结果的余弦相似性作为可解释性的定量化评价指标。余弦相似度 $d_{\cos}$ 的计算公式可表示为

$$d_{\cos}(\boldsymbol{p},\boldsymbol{q}) = \frac{\boldsymbol{p}\cdot\boldsymbol{q}}{\|\boldsymbol{p}\|_2 \times \|\boldsymbol{q}\|_2} = \frac{\sum_i \boldsymbol{p}_i \boldsymbol{q}_i}{\sqrt{\sum_i \boldsymbol{p}_i^2}\sqrt{\sum_j \boldsymbol{q}_j^2}}. \tag{5-8}$$

该公式不仅适用于一维向量，也可以将二维图像进行展平操作，并开展相似度计算。这种度量方法的优势在于它能够有效捕捉归因结果的相对数值相似性，而不受其幅值差异的影响。

将组合块级别设置为 #1，不同归因方法在不同域的仿真数据集各样本类别对不同预测类别的归因结果余弦相似度如图 5-9 所示。从方法层面来看，Mask 和 Scale 方法具有相近的表现，它们的余弦相似度显著低于 SHEP。这是由于他们仅仅能捕捉当前样本中特征存在的影响，而无法评估其他类别特征缺失所带来的贡献。相比之下，SHEP 通过结合 SHEP-Remove 和 SHEP-ADD 两种典型子集，能够全面考虑两方面的影响，因此与 SHAP 的结果具有更高的一致性。

从样本类别来看，Mask 和 Scale 方法在处理健康类别样本时表现极差，这主要源于信号成分 C_0 的随机性特征，正如在图 5-7 和图 5-8(a)-(b) 中所讨论的。这种随机性使得基于简单扰动的归因方法难以准确评估其贡献，从而导致与 SHAP 结果的显著偏差。

从分析域来看，所有方法在二维域（即时频域和循环谱域）中的余弦相似度大体上均低于一维域（即频域和包络域）。二维域的特征维度更高，其归因结果更为复杂。

图 5-9 仅仅展示了组合块级别 #1 下的余弦相似性结果，为了进一步展示更全面的统计结果，现对类别矩阵进行压缩并引入块大小维度，获得的余弦相似度统计结果如图 5-10 所示。从组合块尺寸来看，随着组合块尺寸的增加，SHEP 的余弦相似度显著提高，这表明在较低的解释粒度下，SHEP 的归因结果与 SHAP 更为接近。这种随组合块尺寸增加而相似度提升的现象，主要源于低粒度下特征空间的简化使得 SHEP 的近似更加准确。相比之下，Mask 和 Scale 方法的平均相似度随块大小变化不明显，甚至出现方差增大的趋势。这种差异进一步证实了基于扰动的方法难以准确捕捉特征间的复杂交互关系。

从分析域的角度来看，一方面，一维域（频域和包络域）的相似度普遍高于二维

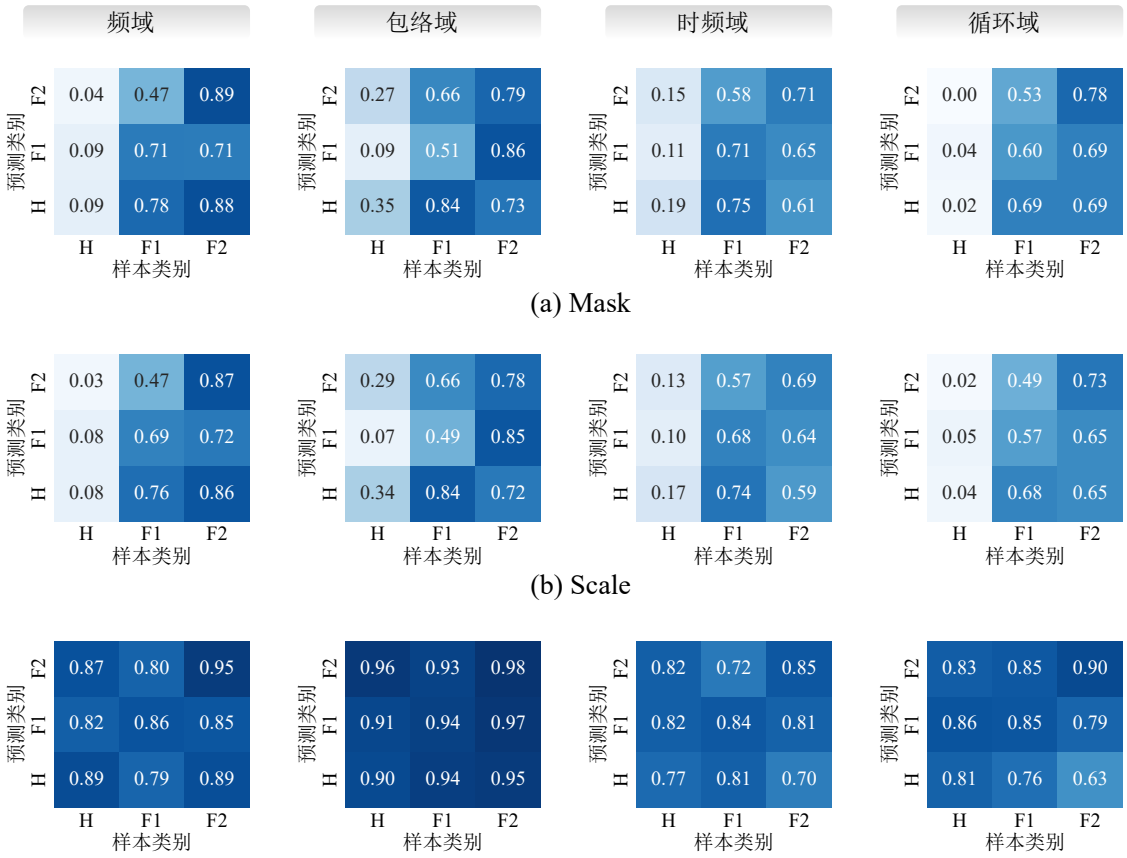


图 5-9　组合块为 #1 时不同归因方法在不同域的仿真数据集各样本类别对不同预测类别的归因结果余弦相似度

Fig. 5-9　The cosine similarity between the attribution results of different methods and domains for each sample class in the simulation dataset and SHAP when the patch size is #1

域（时频域和循环谱域）。另一方面，包络域和循环谱域的相似度方差显著大于频域和时频域。这种方差增大主要归因于循环频率 α 中复杂的谐波成分，这些谐波成分使得不同方法在处理这些域时的表现差异更为明显。

总结而言，Mask 和 Scale 作为基于扰动的方法，仅仅关注当前类别相关特征存在的影响，这种单一视角的归因机制导致其结果与 SHAP 存在显著偏差。相比之下，SHEP 通过同时考虑特征的存在和缺失两个方面，能够更全面地评估特征贡献。这种优势使得 SHEP 在不同块大小、不同分析域和不同样本类别下都表现出与 SHAP 的良好一致性。

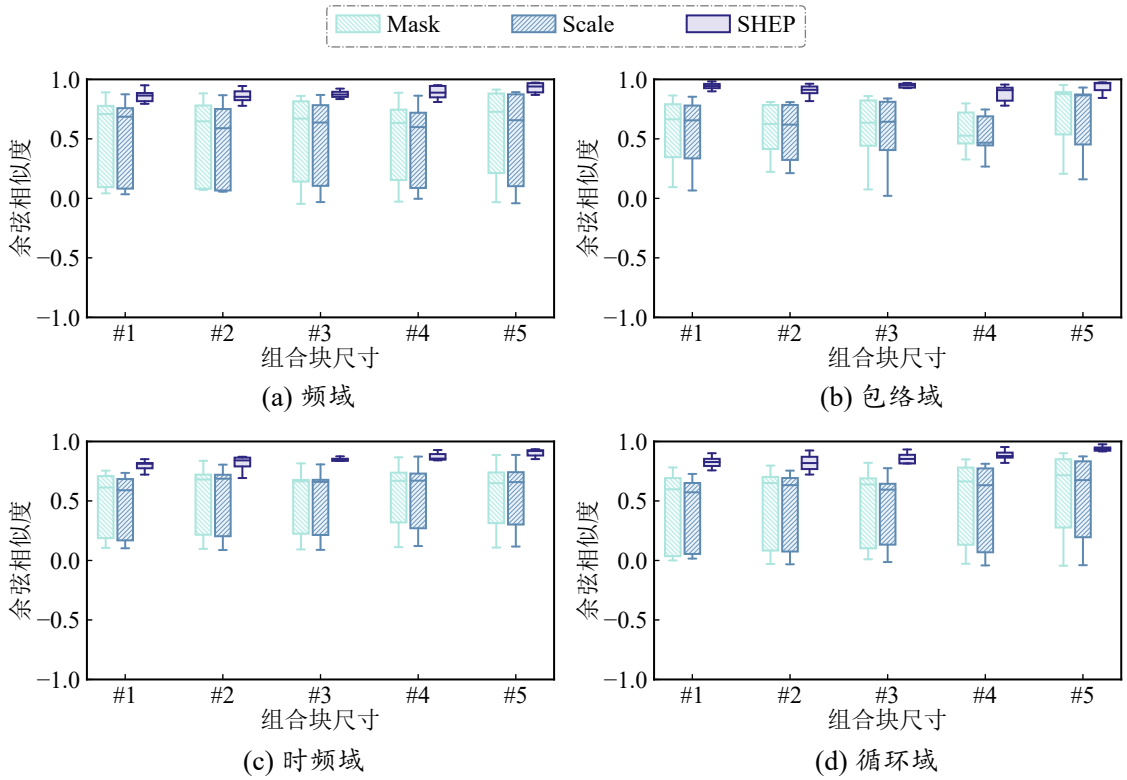


图 5-10　仿真数据集中不同归因方法在不同域的的余弦相似度统计结果

Fig. 5-10　The statistic result of cosine similarity under different attribution methods and domains in the simulation dataset

5.5.5　SHEP 和同类方法的解释效率对比

组合块归因和 SHEP 分别从降低特征维度和降低计算复杂度两方面来提高归因解释的计算效率。不同归因方法的理论计算复杂度可总结为如表 5-3 所示，其中，组合块归因在效率上的提高体现在特征维度 d 的降低。由于原始的 SHAP 包含完整的子集枚举，具有指数级复杂度，实验中采用排列模式的 SHAP 作为其近似实现。

表 5-3　不同归因方法的计算复杂度对比

Table 5-3　The computational complexity comparison of different attribution methods

Mask 归因	**Scale 归因**	**SHAP (枚举模式)**	**SHAP (排列模式)**	**SHEP**
$\mathcal{O}(d)$[a]	$\mathcal{O}(k_s \cdot d)$	$\mathcal{O}(2^d \cdot n)$	$\mathcal{O}(k_p \cdot d \cdot n)$	$\mathcal{O}(d \cdot n)$

[a] 注：d 表示特征维度，k_s 表示 Scale 操作的次数，n 表示背景样本数量，k_p 表示排列次数。在实验中，$k_s = 3$，$k_p = 5$，且 $n = 5 \times c$，其中 c 为类别数。

理论上，归因解释的计算时间主要取决于两个因素：一是由模型复杂度 $\mathcal{O}(\mathcal{M})$

和硬件设备的推理速度，二是解释算法所需的模型调用次数。模型复杂度和推理速度并非本章的研究内容，实际的优化对象是模型调用次数。仿真数据集下不同归因方法在不同域的单次归因计算耗时如图 5-11 和表 5-4，其结果与表 5-3 中的理论复杂度高度一致。

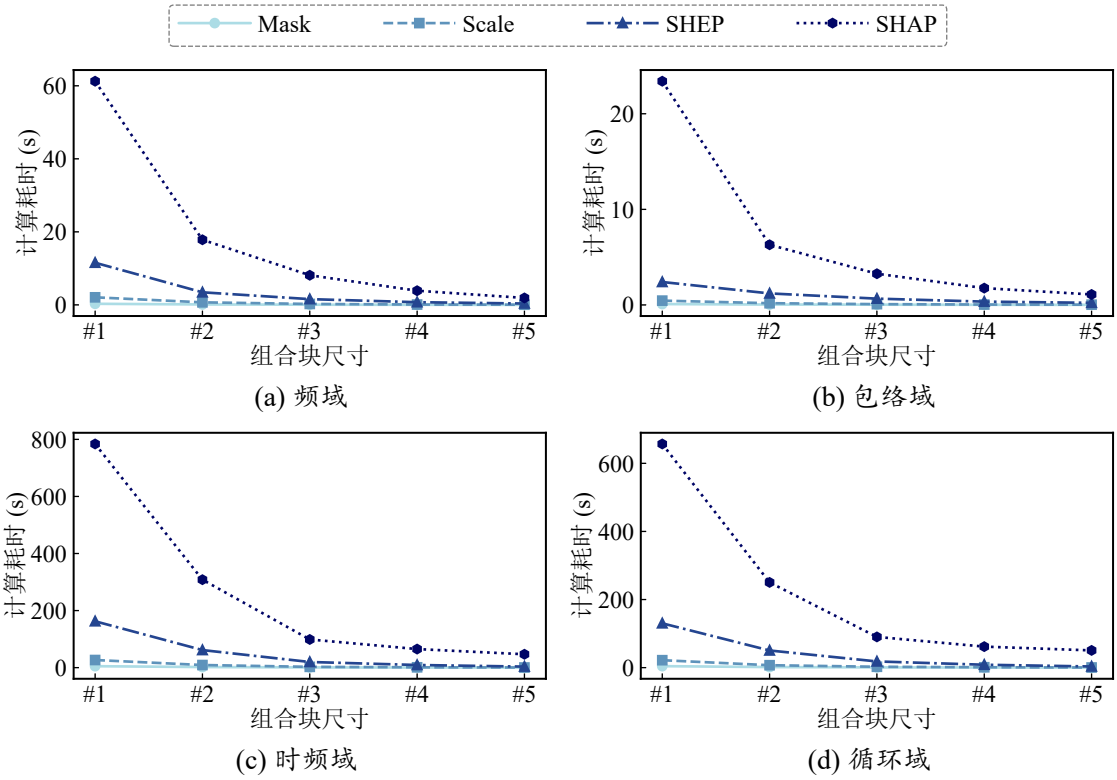


图 5-11　仿真数据集下不同归因方法在不同域的单次归因计算耗时对比

Fig. 5-11　The computational times for each sample across different attribution methods, patch sizes, and domains in the simulation dataset

从方法层面来看，Mask 归因具有最短的计算时间，其次是 Scale 归因，两者相差 k_s 倍 (即 Scale 操作次数)，但整体的复杂度均为 $\mathcal{O}(d)$。SHEP 和排列模式的 SHAP 由于引入 $\mathcal{O}(n)$ 的数据集期望，其复杂度为 $\mathcal{O}(dn)$，其计算时间显著增加。然而，SHEP 通过将置换次数从 k_p 降低到常数 1，大大缩短了其计算时间。这种计算效率的提升使得 SHEP 在实际应用中更具优势，尤其是在需要实时解释的场景中。

从域和块大小的角度来看，它们主要通过影响特征维度 d 来改变计算时间。二维域由于具有更高的维度 d 和更耗时的域变换 $\mathcal{D}$，其计算时间显著高于一维域。此外，块大小的选择也会显著影响计算效率。较大的块尺寸能够有效降低特征维度，从而减少计算时间。然而，正如前文讨论的，这种效率提升是以解释粒度为代价的。

表 5-4　仿真数据集下不同归因方法在不同域的单次归因计算耗时对比

Table 5-4　The computational times for each sample across different attribution methods, patch sizes, and domains in the simulation dataset

域变换	归因方法	#1[a] (s)	#2 (s)	#3 (s)	#4 (s)	#5 (s)
频域	Mask	0.30	0.12	0.06	0.03	0.02
	Scale	2.08	0.70	0.26	0.10	0.05
	SHEP	11.55	3.45	1.58	0.74	0.38
	SHAP	61.26	18.21	17.71	17.81	18.17
包络谱域	Mask	0.09	0.04	0.03	0.02	0.01
	Scale	0.45	0.18	0.09	0.04	0.03
	SHEP	2.40	1.21	0.66	0.35	0.23
	SHAP	23.40	18.27	19.72	19.14	18.53
时频域	Mask	5.04	2.07	0.72	0.37	0.16
	Scale	26.71	9.15	2.79	1.20	0.48
	SHEP	162.48	61.57	19.79	9.27	3.64
	SHAP	783.90	308.25	98.72	65.07	46.91
循环域	Mask	4.37	1.69	0.69	0.35	0.16
	Scale	21.87	7.14	2.60	1.14	0.44
	SHEP	130.17	50.39	18.24	8.53	3.41
	SHAP	656.88	250.31	90.25	61.73	50.62

[a]：组合块尺寸级别

总的来说，所提出的 SHEP 方法虽然计算时间长于 Mask 和 Scale 等基础方法，但相较于原始 SHAP 仍显著降低了计算耗时。这种计算复杂度的降低主要得益于 SHEP 将置换次数从 k_p 降低到 1，从而将计算复杂度从 $\mathcal{O}(k_p \cdot d \cdot n)$ 降低到 $\mathcal{O}(d \cdot n)$。此外，域变换类型和组合块设置都会通过影响特征维度 d 来影响计算时间，二维域（如时频域和循环谱域）具有更高的特征维度，其计算时间显著高于一维域（如频域和包络域），而增大组合块尺寸则能有效降低特征维度，从而减少计算时间。因此，在实际应用中，需要根据具体场景（如实时性要求、硬件性能限制）和分析目标（如解释粒度需求、可视化偏好）来权衡选择合适的域变换类型和组合块尺寸，以实现计算效率和解释结果之间的最优平衡。

5.6　实测数据集下 SHEP 被动解释的实验验证

在相同参数设置下，改变数据集并不会影响各种归因方法的计算效率。因此，本节将主要关注可解释性方面的表现。为了验证 SHEP 在实测场景下的可解释性能，本节将其应用于 CWRU 轴承数据集和斜齿轮数据集，包括数据集介绍、SHEP 可视化

和解释结果相似性分析三部分。

5.6.1　可复现的 CWRU 轴承开源数据集

CWRU 轴承数据集是广受认可的开源数据集，基于此数据集开展实验验证可以确保 SHEP 的可复现性，其试验台如图 2-7 所示。实验中选取了负载为 1 马力、转速为 1800 rpm、采样频率为 12 kHz 的工况，并考虑四种故障状态：健康状态（Health, H）、内圈故障（Inner race fault, I）、滚动体故障（Rolling ball fault, B）和外圈故障（Outer race fault, O）。表 4-3 列出了相应的特征频率。为了更好地理解数据特征，图 5-12 展示了这四种类别样本在不同域中的表示。其中，健康状态样本在 360 Hz（$n_r f_z$）处表现出显著特征。实验设置和训练过程与 5.5 节的仿真数据集保持一致，最终端到端 CNN 模型在测试集上达到了 100% 的识别准确率。

与具有完全已知故障逻辑（即真实标签）的仿真数据集不同，CWRU 数据集首

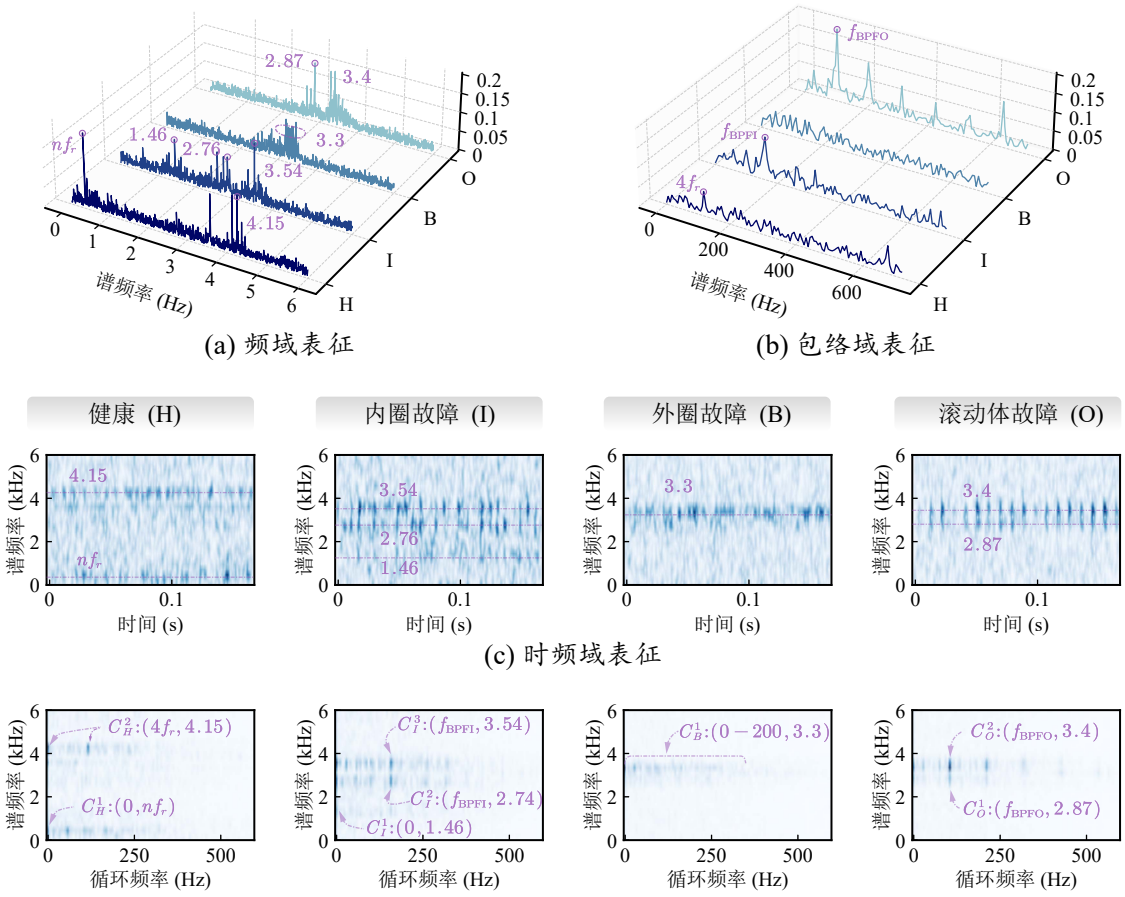


图 5-12　CWRU 数据集下各类样本在不同域的表征

Fig. 5-12　The representation of each class sample in different domains of the CWRU dataset

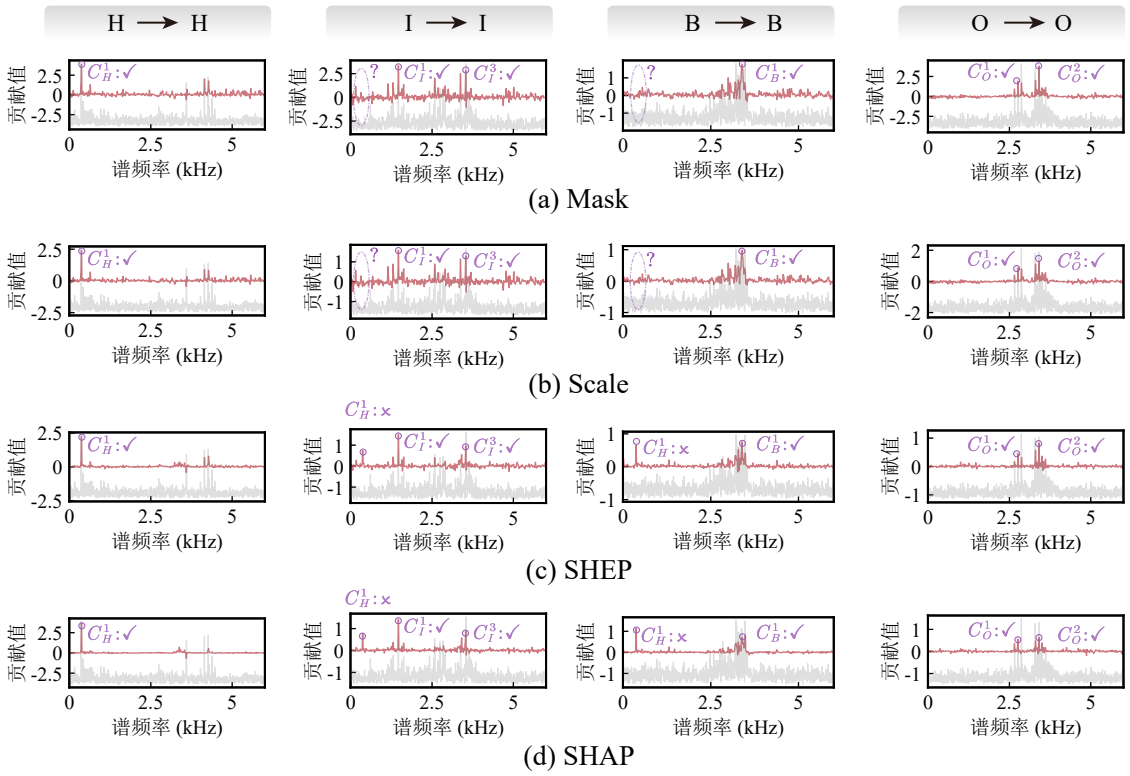


图 5-13　组合块为 #1 时不同归因方法对 CWRU 数据集中不同类别样本的对应预测类别频域归因结果

Fig. 5-13　Attribution results of different attribution methods in the frequency domain for different class samples towards thier corresponding predicted class in the CWRU dataset under patch size level #1

先需要识别每个类别的故障特征。健康类别（H）样本包含一个主要的恒定频率分量 $C_H^1:(0,nf_r)$ 和一个调制分量 $C_H^2:(4f_r,4.15)$；内圈故障类别（I）样本包含 $C_I^1:(0,1.46)$、$C_I^2:(f_{\mathrm{BPFI}},2.74)$ 和 $C_I^3:(f_{\mathrm{BPFI}},3.54)$；滚动体故障类别（B）样本包含 $C_B^1:(0-200,3.3)$；外圈故障类别（O）样本包含 $P_O^1:(f_{\mathrm{BPFO}},2.87)$ 和 $P_O^2:(f_{\mathrm{BPFO}},3.4)$。这些特征频率的组合构成了每个类别的独特故障特征，为后续的可解释性分析提供了重要的参考基准。值得注意的是，与仿真数据集相比，CWRU 数据集的故障特征更为复杂，这为验证所提出方法的实际应用效果提供了更具挑战性的测试场景。

将组合块级别设置为 #1，不同归因方法对 CWRU 数据集中不同类别样本的对应预测类别频域归因结果如图 5-13 所示。以健康类别为例，信号成分 C_H^1 是健康类别的关键特征，健康类别样本中信号成分 C_H^1 的存在（即 C_H^1:✓）能够被 Mask 和 Scale 方法准确识别，进而对健康类别的预测产生显著正向贡献。然而，这两种方法无法捕捉到该信号成分在其他类别样本中缺失所带来的贡献（即 C_H^1:✗）。相比之下，SHEP

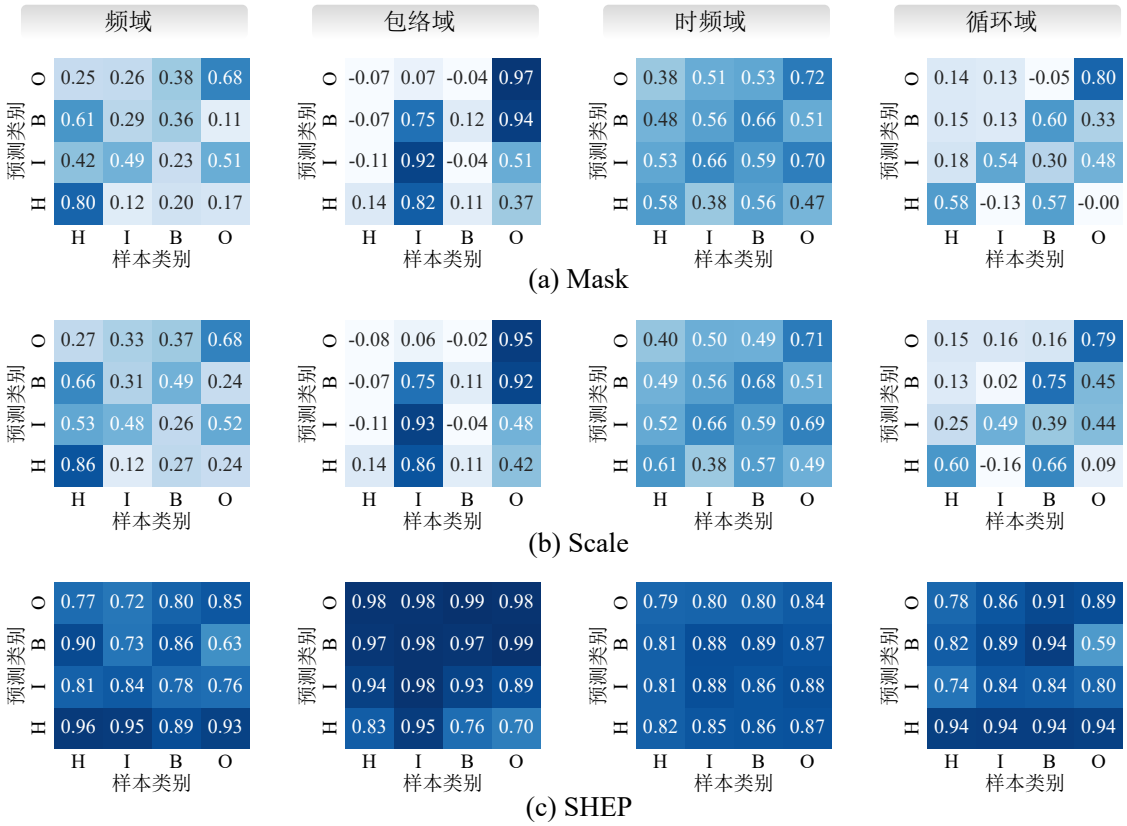


图 5-14　组合块为 #1 时不同归因方法在不同域的 CWRU 数据集各样本类别对不同预测类别的归因结果余弦相似度

Fig. 5-14　The cosine similarity between the attribution results of different methods and domains for each sample class in the CWRU dataset and SHAP when the patch size is #1

和 SHAP 不仅准确识别了 C_H^1 存在的贡献，还有效地捕捉到了其在内圈故障和滚动体故障类别样本中缺失所具有的正向贡献（即 C_H^1:✕），展现出更优越的解释能力。这种双重视角的归因机制使得 SHEP 和 SHAP 能够更全面地解释模型的预测逻辑。

将组合块级别设置为 #1，不同归因方法在不同域的 CWRU 数据集各样本类别对不同预测类别的归因结果余弦相似度如图 5-14 所示。从整体来看，SHEP 表现出极高的余弦相似度，大多数情况下超过 0.8，这显著优于 Mask 和 Scale 方法。这种高相似度不仅证实了 SHEP 是 SHAP 的有效近似，还表明其能够保持与 SHAP 相同的归因逻辑。值得注意的是，在包络域中，Mask 和 Scale 方法在处理健康类别（H）和滚动体故障类别（B）样本时表现较差。这种性能下降主要源于这两个类别缺乏显著的循环特征，如内圈故障类别（I）中的 f_{BPFI} 或外圈故障类别（O）中的 f_{BPFO}，这使得基于简单扰动的归因方法难以准确衡量各成分贡献度。然而，即使在这种困难场景下，SHEP 仍然保持了与 SHAP 的高度一致性，充分展示了其归因机制的鲁棒性和通

用性。这种出色的表现进一步证实了 SHEP 在处理复杂故障特征方面的优势。

CWRU 数据集中不同归因方法在不同域的的余弦相似度统计结果如图 5-15 所示，其变化趋势与仿真数据集基本一致。从组合块尺寸来看，随着组合块尺寸的增加，解释的粒度变得更粗，这通常导致余弦相似性随之上升。这一现象的主要原因在于，较大的组合块会将相邻的特征组合在一起，从而简化了特征空间，使得不同归因方法的结果更容易趋同。此外，Mask 和 Scale 方法的相似度方差随组合块尺寸增加而显著增大，这种趋势在包含循环频率 α 的包络域和循环谱域中尤为明显。

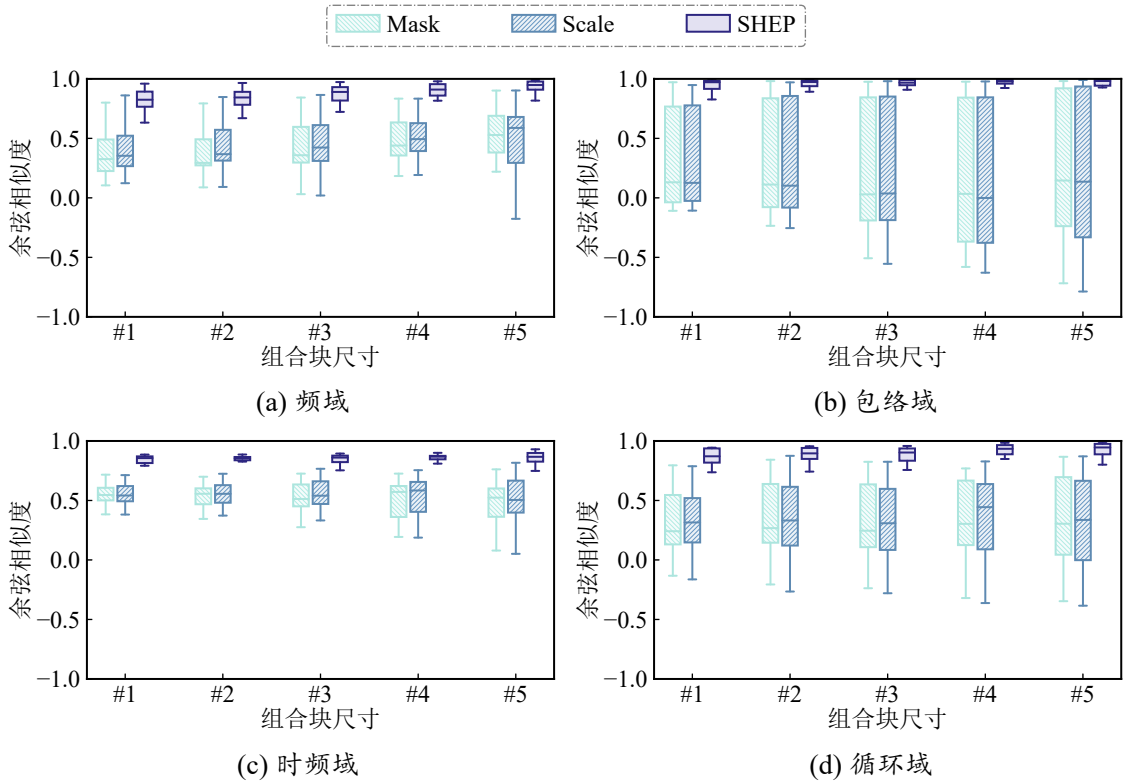


图 5-15　CWRU 数据集中不同归因方法在不同域的的余弦相似度统计结果

Fig. 5-15　The statistic result of cosine similarity under different attribution methods and domains in the CWRU dataset

从域的角度分析，一维域（频域和包络域）的相似度普遍高于二维域（时频域和循环谱域），这与特征维度和空间复杂性直接相关。同时，包络域和循环谱域由于涉及调制频率 f_c 的提取，其相似度方差明显大于频域和时频域，这表明在处理复杂调制信号时，不同归因方法的表现差异更为显著。

总的来说，CWRU 轴承开源数据集的实验充分证明，SHEP 在不同域和不同组合块尺寸下均能保持与 SHAP 的一致性，并且在归因效果上显著优于 Mask 和 Scale 方

法。这种优势不仅体现在与原始 SHAP 的相似性上，还体现在对复杂故障特征的准确捕捉能力上，这对于实际故障诊断应用具有重要意义。

5.6.2　实验室场景下的斜齿轮数据集

为了验证 SHEP 方法在实际工程应用中的有效性，现选用实测场景的斜齿轮箱试验数据进行验证，斜齿轮数据集的试验台和故障类型如图 3-10 所示。其中，电机的转速设置为 1800 rpm，相应的特征频率如表 4-4 所示。数据集的采样频率设置为 5 kHz，包含四种工况类别：健康（Health, H）、从动齿轮表面磨损故障（Wear, W）、从动齿轮表面点蚀故障（Pitting, P）和从动齿轮断齿故障（Crack, C）。每个类别包含 76 个样本，每个样本的长度为 2000 个采样点。为确保实验的一致性和可对比性，本实验的实验设置和训练过程与 5.5 节的仿真数据集保持一致。经过训练后，端到端 CNN

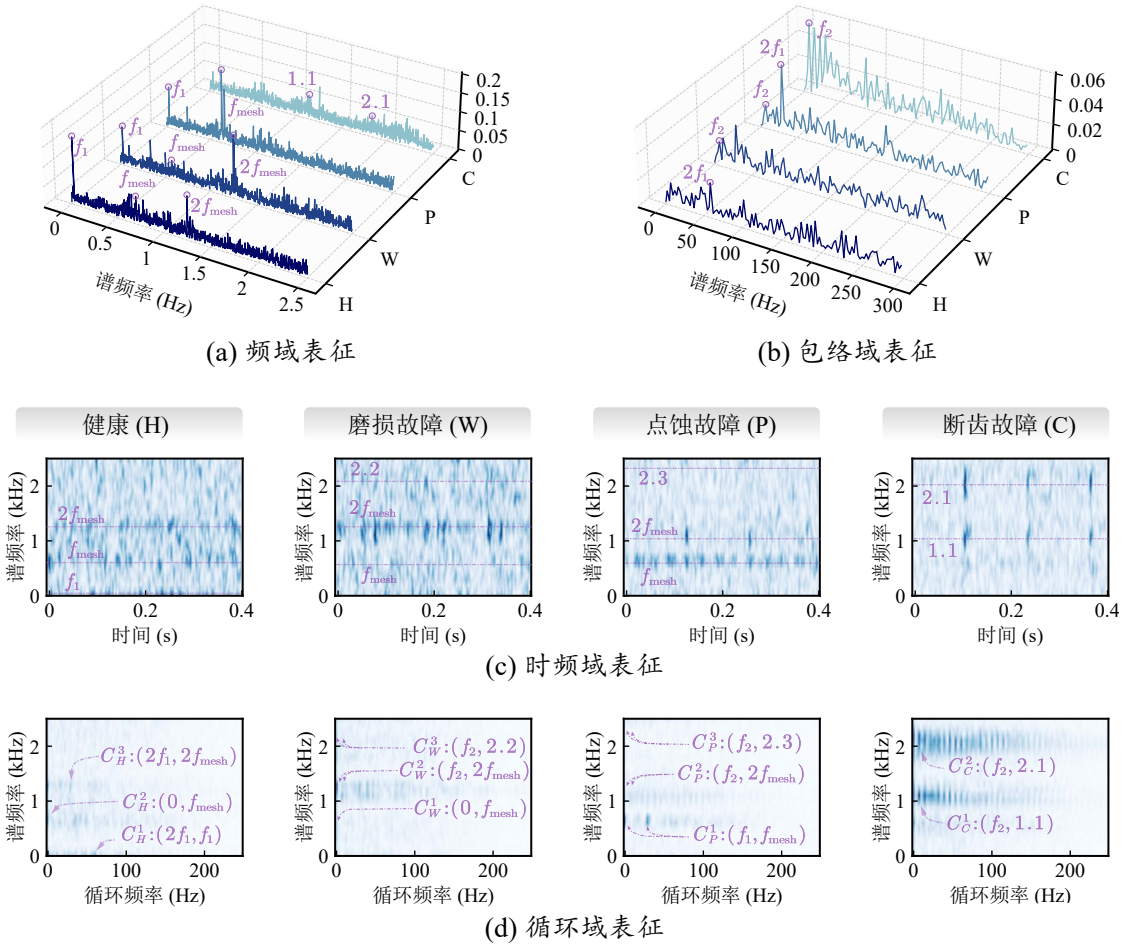


图 5-16　斜齿轮数据集下各类样本在不同域的表征

Fig. 5-16　The representation of each class sample in different domains of the helical gearbox dataset

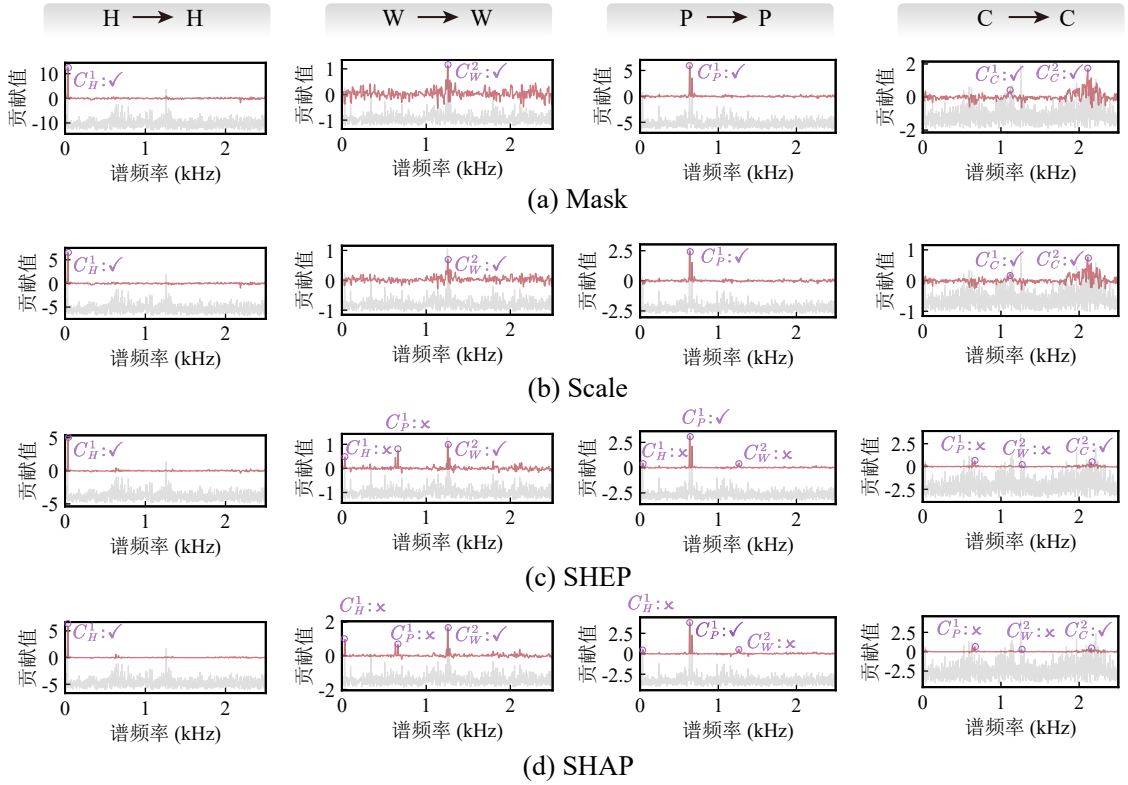


图 5-17　组合块为 #1 时不同归因方法对斜齿轮数据集中不同类别样本的对应预测类别频域归因结果

Fig. 5-17　Attribution results of different attribution methods in the frequency domain for different class samples towards thier corresponding predicted class in the helical gearbox dataset under patch size level #1

模型在测试集上取得了 100% 的识别准确率，这为后续的可解释性分析提供了可靠的基础。

与 CWRU 数据集类似，在进行可解释性分析之前，首先需要明确每个类别的故障特征。斜齿轮数据集下各类样本在不同域的表征如图 5-16 所示，其中，健康类别 (H) 样本包含信号成分 $C_H^1:(2f_1, f1)$、信号成分 $C_H^2:(0, f_{\text{mesh}})$ 以及信号成分 $C_H^3:(2f_1, 2f_{\text{mesh}})$。磨损故障类别 (W) 样本包含信号成分 $C_W^1:(f_2, 2f_{\text{mesh}})$。点蚀故障类别 (P) 样本包含 $C_P^1:(f_1, f_{\text{mesh}})$ 和 $C_P^2:(f_2, f_{\text{mesh}})$ 两个信号成分。断齿故障类别 (C) 样本包含 $C_C^1:(f_2, 1.1)$ 和 $C_C^2:(f_2, 2.1)$ 两个信号成分。这些故障特征的识别对于理解和验证后续归因结果的合理性至关重要。相比于 CWRU 数据集，斜齿轮箱的故障特征呈现出更复杂的调制关系，这为验证 SHEP 方法的实际应用效果提供了更具挑战性的测试场景。

将组合块级别设置为 #1，不同归因方法对斜齿轮数据集中不同类别样本的对应

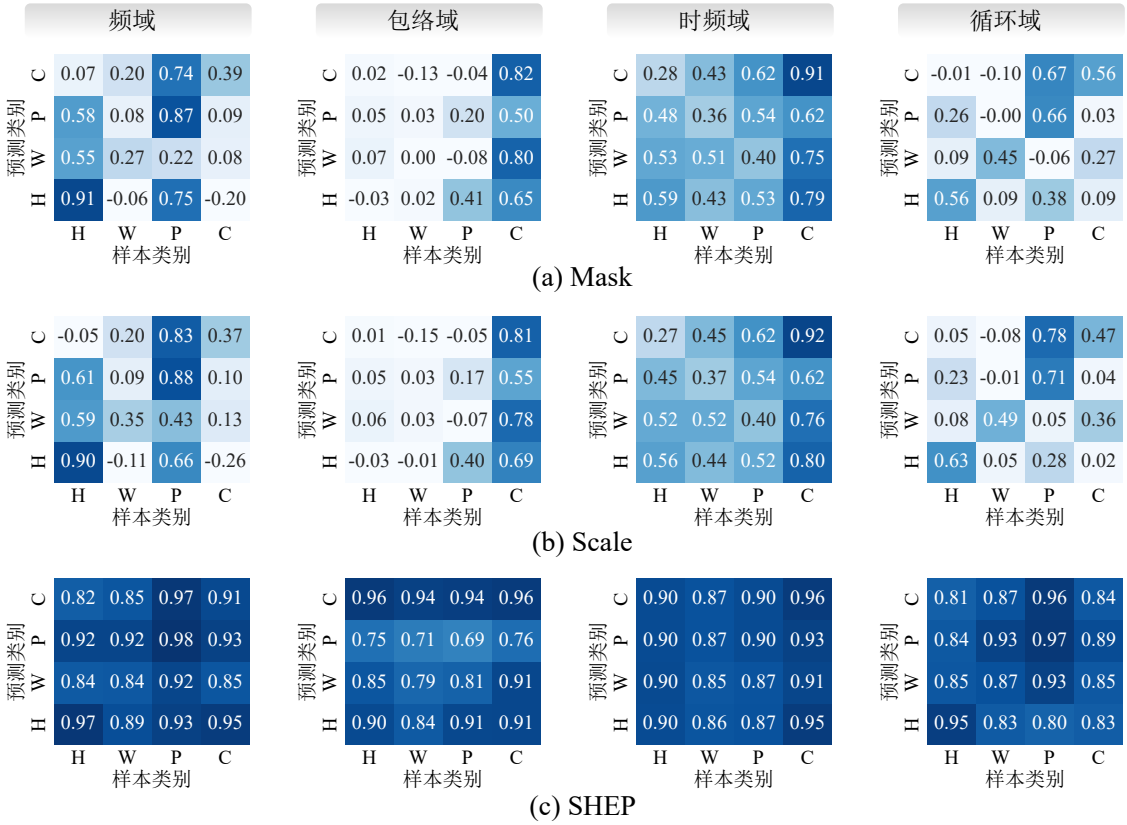


图 5-18　组合块为 #1 时不同归因方法在不同域的斜齿轮数据集各样本类别对不同预测类别的归因结果余弦相似度

Fig. 5-18　The cosine similarity between the attribution results of different methods and domains for each sample class in the helical gearbox dataset and SHAP when the patch size is #1

预测类别频域归因结果如图 5-17 所示。与前文的观察结果一致，Mask 和 Scale 方法仅能捕捉当前类别中存在的特征组成部分所带来的贡献，如对健康（H）样本中信号成分 C_H^1:✓ 的存在对健康类别预测的贡献，又或磨损故障（W）样本中 C_W^1 : ✓ 对磨损故障类别预测的贡献。然而，这些方法无法识别其他类别相关特征的缺失所带来的贡献，如磨损故障（W）和点蚀故障（P）样本中 C_H^1:✗ 的贡献，以及磨损故障（W）和断齿故障（C）类别中 C_P^1:✗ 的贡献。这种单一视角的归因机制显著限制了这些方法的解释能力，使其难以全面反映特征与预测结果之间的复杂关系。

将组合块级别设置为 #1，不同归因方法在不同域的斜齿轮数据集各样本类别对不同预测类别的归因结果余弦相似度如图 5-18 所示。实验结果表明，SHEP 方法在所有分析场景下均显著优于 Mask 和 Scale 方法，其与 SHAP 的相似度大多数情况下超过 0.8。这种高度一致性不仅证实了 SHEP 是 SHAP 的有效近似，还表明其能够准确保持 SHAP 的归因逻辑。相比之下，Mask 和 Scale 方法在频域中对磨损故障（W）和

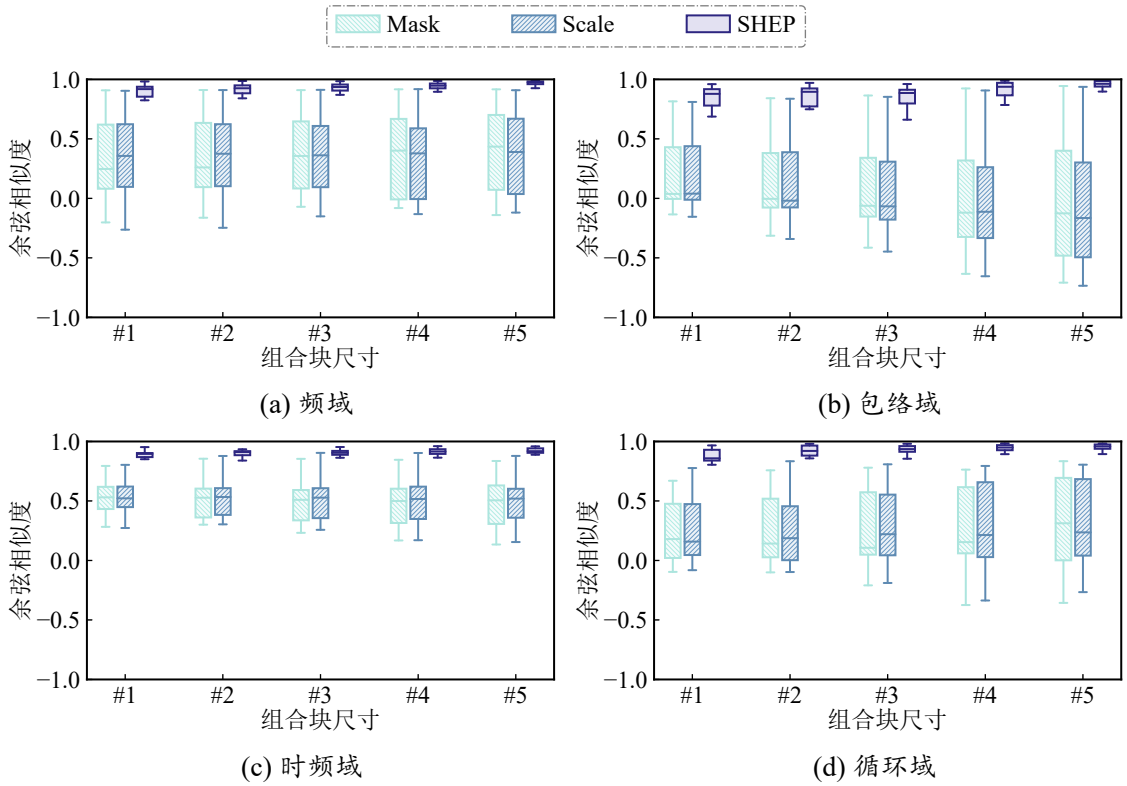


图 5-19　斜齿轮数据集中不同归因方法在不同域的的余弦相似度统计结果

Fig. 5-19　The statistic result of cosine similarity under different attribution methods and domains in the helical gearbox dataset

断齿故障（C）类别的处理，以及在包络域中对健康状态（H）、磨损故障（W）和点蚀故障（P）类别的处理上表现出明显过低的余弦相似性。

深入分析这种性能差异的原因，在频域中的相似度降低主要源于这些方法无法捕捉关键信号成分的缺失所具有的影响。如图 5-17 所示，它们既不能识别 C_P^1:✕ 对磨损故障（W）类别的影响，也不能捕捉 C_W^1:✕ 对断齿故障类别（C）的影响。而在包络域中的性能下降则与信号特征的本质有关。如图 5-16(b) 所示，健康状态（H）、磨损故障（W）和点蚀故障（P）类别缺乏显著的调制频率特征，这与断齿故障（C）类别中以输出轴频率 f_2 形成的明显调制频率成分形成鲜明对比。

值得注意的是，即使在这些具有挑战性的场景下，SHEP 仍然保持了与 SHAP 的高度一致性。这种优异的表现不仅证实了 SHEP 归因机制的鲁棒性和通用性，还凸显了其在处理复杂故障特征时的显著优势。这对于实际工程应用具有重要意义，因为实际故障往往表现出更为复杂和多样的特征模式。

斜齿轮数据集中不同归因方法在不同域的的余弦相似度统计结果如图 5-19 所示。

从组合块尺寸的角度来看，随着块大小的增加，解释的粒度变得更粗，这导致 SHEP 与 SHAP 的相似度显著提升。不同于之前两个数据集，Mask 和 SHEP 方法的相似度方差不仅在包含循环频率的包络域和循环谱域中表现出高度波动，在频域中也呈现出显著的变化。这种现象主要归因于斜齿轮数据集更高的诊断难度，如图 5-16 所示，多个信号成分具有相同的载波频率（例如，磨损故障的 C_W^1 和点蚀故障的 C_P^2）或调制频率（例如，断齿故障的 C_H^3 和 C_P^2）。过于紧邻的信号成分增加了准确归因的难度，使得不同归因方法在处理这些复杂特征时表现出更大的差异。

总结而言，斜齿轮数据集的分类难度较高，导致 Mask 和 Scale 方法的相似度性能明显下降，主要原因在于这些方法仅考虑特征存在带来的贡献，而忽视了特征缺失的影响。相比之下，SHEP 通过同时考虑特征的存在和缺失这两方面的影响，即使在这种更具挑战性的诊断场景下，仍然保持了与 SHAP 的高度一致性。

5.7　本章小结

由于振动信号的高维特性以及域变换所带来的维度升高，采用完整子集枚举的 SHAP 被动归因方法在智能故障诊断领域面临严峻的计算成本问题。为提升计算效率，本章提出了以增加解释粒度为代价来降低特征维度的组合块归因策略，以及通过缩减子集枚举过程来降低算法复杂度的 SHEP 归因方法。基于一个故障已知的仿真数据集和两个实测数据集的实验验证，本章主要内容总结如下：

(1) 提出了用以降低特征维度的组合块归因策略。对特征维度在 SHAP 计算耗时中的影响进行理论性分析，并建立了将多个相邻特征绑定的组合块变换 $\mathcal{P}$ 及其逆变换 $\mathcal{P}^{-1}$。由此，组合块归因策略将所得组合块视为整体以计算联合贡献，以更粗解释粒度为代价，在不改变输入数据的前提下降低归因计算中的特征维度。

(2) 提出了用以降低计算复杂度的 SHEP 归因方法。通过两个具有代表性的实例（SHEP-Remove 和 SHEP-Add）来近似原始 SHAP 计算过程，将复杂度由指数级降低至线性级。其中，SHEP-Remove 从输入样本的角度出发，聚焦当前特征存在所带来的贡献，SHEP-Add 则从背景样本的角度出发，捕捉其他类别特征缺失的贡献。这两者互补的归因视角相结合，使 SHEP 能够提供全面且与 SHAP 近似的归因解释结果。

(3) 在故障逻辑已知的仿真数据集和实测数据集实验证明，组合块归因策略虽然会使结果的解释粒度更粗，但并不影响解释的正确性，为实际应用中平衡解释粒度与计算效率提供了调控手段。SHEP 方法在多种数据集、不同域、各类组合

块尺寸和不同类别样本下均表现出与 SHAP 的高度一致性，余弦相似度的平均值超过 0.85，显示了其作为 SHAP 近似算法的可行性。在复杂故障特征场景中，SHEP 仍能保持与 SHAP 相近的优异解释能力。在计算效率方面，各方法实际计算时间和理论复杂度均与预期相符，SHEP 相较于 SHAP 展现出明显的效率提升，其计算耗时随特征维度和背景样本数量呈线性 $\mathcal{O}(d \cdot n)$ 增长。

第六章　面向旋转机械智能诊断模型的综合解释框架及应用验证

6.1　引言

第二章和第三章提出的时频卷积网络和原型匹配网络在增强模型诊断能力的同时，分别聚焦智能诊断模型的输入层和决策层实现主动解释。第四章和第五章则将SHAP归因与振动信号特性相结合，分别从解释效果和解释效率两方面对智能诊断场景中的被动解释方法进行了优化。尽管上述方法均展现出良好的智能诊断可解释性，但各自的解释结果侧重点不同，应用条件也存在差异。其中，主动解释方法通过对模型针对性修改获得独特的解释结果，但同时受到模型结构的约束；被动解释方法仅对已有模型进行事后分析，保持了模型设计的自由度，但解释形式相对单一。

基于此，本章综合前述研究成果，构建了一种面向旋转机械智能诊断模型的综合解释框架，旨在指导用户根据实际任务需求选择最适合的解释方法，从而提升智能诊断可解释性研究的系统性和实用性。针对能够参与模型设计阶段的解释场景，本框架将时频卷积层和原型匹配层分别整合进模型的输入层和决策层，形成联合解释网络，进而实现对模型关注频带、分类逻辑和典型故障原型的全面解释；针对无法参与模型设计的解释场景，本框架采用清晰、准确且高效的SHEP归因方法对现有智能诊断模型进行事后分析，揭示模型决策的内在依据。

本章首先阐述综合解释框架的设计思路，以及将时频卷积层和原型匹配层相结合的联合解释网络。考虑到主动解释方法可能对模型诊断性能产生影响，本章依托实测重载行星齿轮传动系统构建了强噪声和少样本两种挑战性诊断任务，对联合解释网络的诊断性能进行了全面验证。最后，从输入层主动解释、决策层主动解释和被动归因解释三个方面，系统评估了综合解释框架各模型的解释效果。

6.2　面向旋转机械智能诊断模型的综合解释框架

本文共提出三种智能诊断可解释方法：时频变换层实现的输入层主动解释、原型匹配层实现的决策层主动解释和与域转换相结合的SHEP归因被动解释。这三种解释方法的特点、实现方式和解释结果如表6-1所示。

表 6-1　三类解释方法的综合对比

Table 6-1　Comparison of three types of interpretation methods

解释方法	特点	实现方式	解释结果
输入层主动解释	定制预处理层、很强兼容性	将时频卷积层作为模型预处理层，并开展幅频响应分析	关键频率：解释模型对不同频带的关注程度
决策层主动解释	定制分类层、需要自编码器、较强兼容性	将原型匹配层作为模型分类层，并引入额外训练损失	决策逻辑：强化各故障类别关键频谱特征
SHEP被动解释	任意端到端模型、完全兼容	指定转换域和组合块尺寸，对现有模型进行 SHEP 归因分析	决策依据：特定域、特定粒度下各特征对决策的贡献度

具体而言，输入层主动解释方法将具有物理意义的时频变换卷积层作为模型预处理层，并开展幅频响应分析，从而揭示模型对不同频带的关注程度。它仅需为现有模型添加预处理层，由此具有很强的兼容性。决策层主动解释方法通过将模型分类层替换为原型匹配层并引入额外距离约束损失项，进而解释模型的决策逻辑、并通过解码器对所学习原型进行重构，强化模型对各故障类别的关键频谱特征。它需要特定模型分类层、用以重构原型的自编码器结构以及获得频域样本的数据预处理，但仍可选择不同模型作为编码器，具有较强的兼容性。SHEP 归因方法根据指定的转换域和组合块尺寸，对现有模型进行 SHEP 归因分析，获取特定域、特定粒度下输入样本各特征对决策的贡献度，即模型的决策依据。作为一种通用的被动解释方法，SHEP 归因不需要对模型进行修改，具有完全兼容性。

为了同时获得输入层和决策层的解释结果，本章将时频卷积网络和原型匹配网络进行组合，从而构建输入层和决策层均可解释的联合解释网络。时频卷积网络、原型匹配网络和联合解释网络的整体架构如图 6-1 所示。

时频卷积网络以时域信号 $\boldsymbol{x}$ 为输入，通过时频卷积层 f_{TF} 进行预处理并通过特征提取器 f_{FE} 获得高层次特征 $\boldsymbol{z}_{\mathrm{TF}}$，最后通过分类器 f_{Cla} 获得预测结果 $\boldsymbol{c}$ 并借助分类损失 $\mathcal{L}_{\mathrm{Cla}}$ 进行模型训练。原型匹配网络以 Fourier 变换后的频域样本 $\mathcal{F}(\boldsymbol{x})$ 为输入，通过编码器 f_{Enc} 获得编码特征 $\boldsymbol{z}_{\mathrm{PM}}$ 后，一方面借助定制的原型匹配层 p 获取预测结果 $\boldsymbol{c}$，另一方面也通过解码器重构频域样本 $f_{\mathrm{Dec}}(\boldsymbol{z})$，最终通过式 (3-13) 所示的分类损失 $\mathcal{L}_{\mathrm{Cla}}$、重构损失 $\mathcal{L}_{\mathrm{Recon}}$ 和原型匹配距离损失（$\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$）进行模型训练。

联合解释网络则在原型匹配网络的基础上，将时频卷积网络从时域样本 $\boldsymbol{x}$ 中所提取的特征 $\boldsymbol{z}_{\mathrm{TF}}$ 和编码器从频域样本 $\mathcal{F}(\boldsymbol{x})$ 中提取的特征 $\boldsymbol{z}_{\mathrm{PM}}$ 进行拼接，并通过简

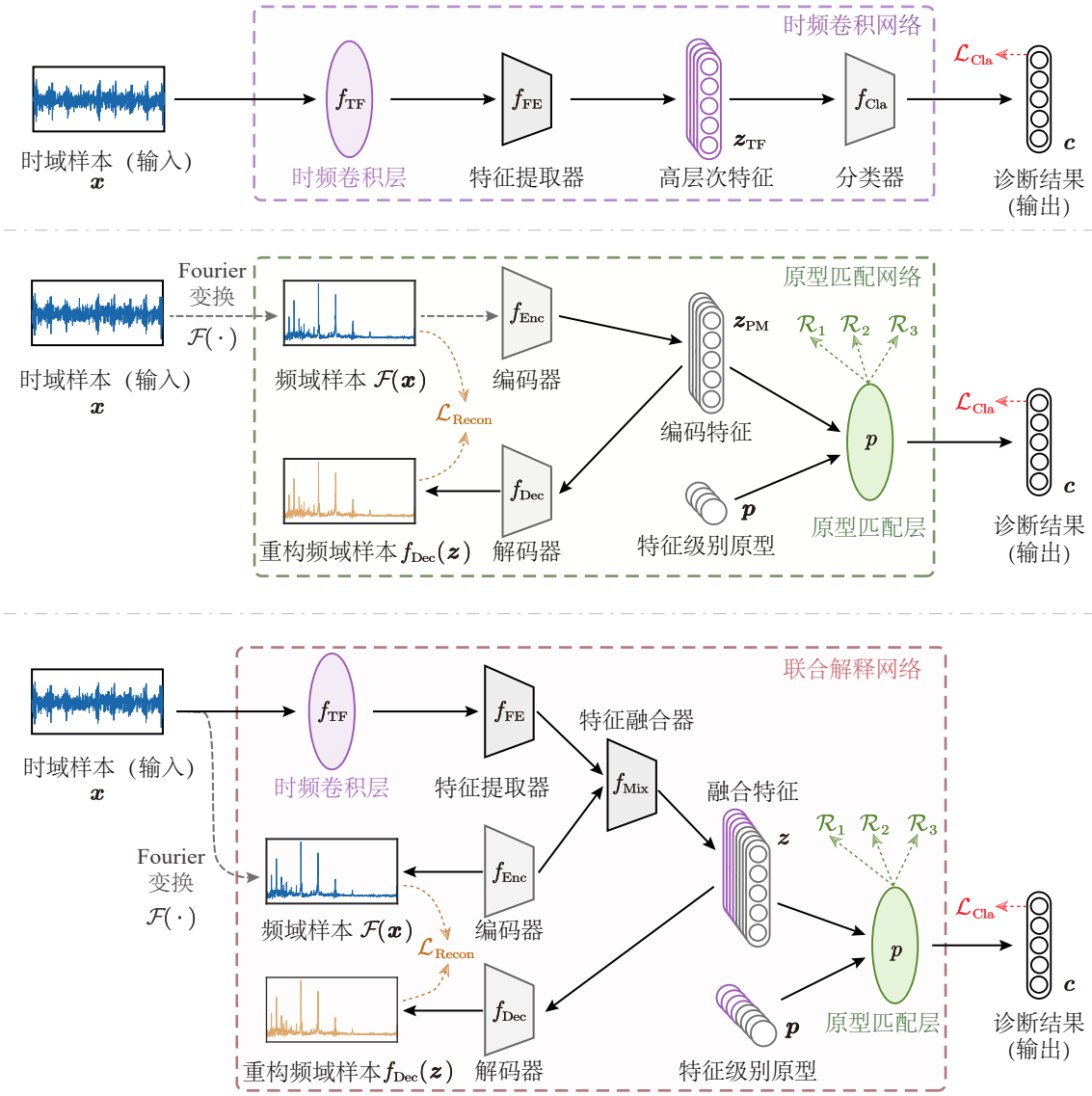


图 6-1　时频卷积网络、原型匹配网络和联合解释网络的整体架构

Fig. 6-1　The overall architecture of the TFN, PMN, and Joint Interpretation Network

单的特征融合器 f_{Mix} 构建融合特征 $\boldsymbol{z}$:

$$\boldsymbol{z} = f_{\mathrm{Mix}}(\boldsymbol{z}_{\mathrm{TF}} \oplus \boldsymbol{z}_{\mathrm{PM}}), \tag{6-1}$$

式中 $\oplus$ 代表向量拼接操作，实验中 f_{Mix} 采用简单的单个全连接层。除上述特征融合外，联合解释网络的其余部分与原型匹配网络相同，即联合解释网络的分类器、解码器和损失函数均与原型匹配网络保持一致。一方面，联合解释网络的输入层为时频卷积层，可通过对训练后的时频卷积层进行频响分析，解释模型对不同频带的关注程度；另一方面，联合解释网络的决策层为原型匹配层，可以在解释模型分类逻辑的同

时，利用解码器对特征级别原型进行重构，解释模型视角下各故障类别的关键频谱特征。

总的来看，图 6-1 所示的三类主动解释模型都具有较强的兼容性，现有先进网络均可作为特征提取器和编码器与之结合。在实际应用中，用户可根据任务需求选择合适的先进网络作为基准网络，从而发挥神经网络规模化、灵活性的优势，在保证模型诊断性能的前提下进行可解释性分析。

综合解释框架的流程图如图 6-2 所示，首先根据能否参与模型设计这一条件划分出两种解释场景。在能够参与模型设计的解释场景下，用户再根据关键频率、决策逻辑这两种解释需求，来针对性地选择输入层主动解释、决策层主动解释或联合主动解释并构建对应的解释网络。在主动解释完成后，也可选择通过 SHEP 这一被动解释方法对已训练模型进行事后分析，以获取模型决策依据相关的解释结果。在无法参与模型设计的解释场景下，主动解释方法由于需要模型修改便不在适用，但仍可直接采用 SHEP 归因方法对给定训练模型进行解释，以获取模型的决策依据。上述两种解释场景和四种解释方法的组合，构成了本文所提出的综合解释框架。

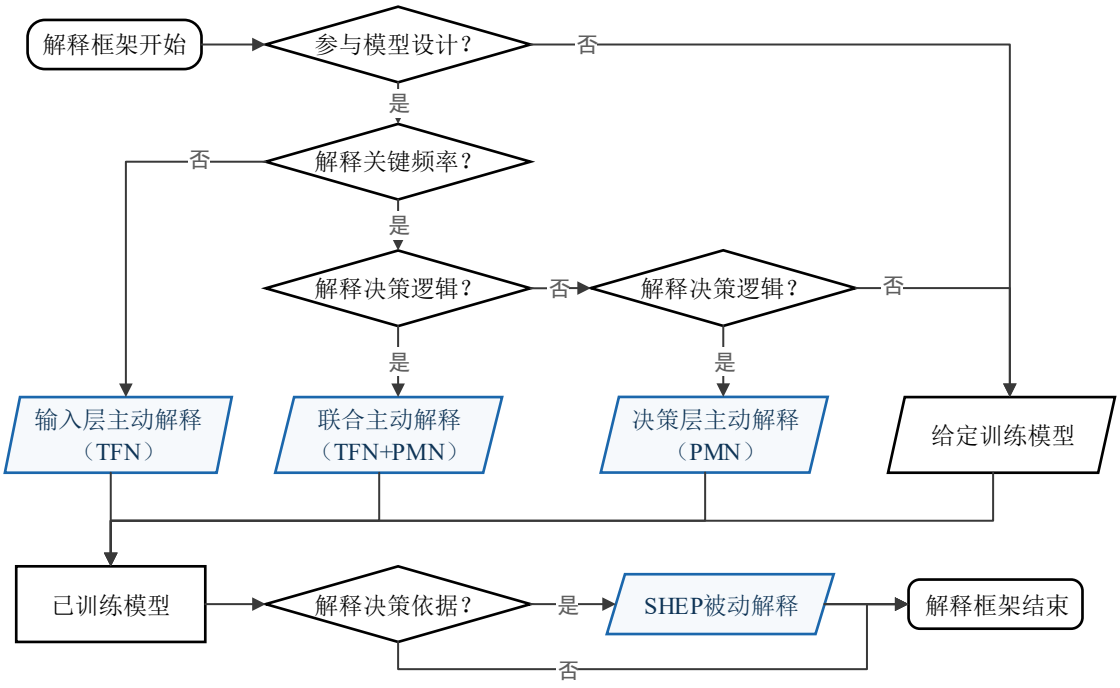


图 6-2　综合解释框架的流程图

Fig. 6-2　Flowchart of the comprehensive interpretation framework

6.3　高速重载行星齿轮传动系统数据集

输入层主动解释、决策层主动解释和 SHEP 被动解释三类方法都在各自数据集进行了有效验证，但一方面数据集大多来自实验室小型试验台，实测应用场景下的效果尚待证实；另一方面，解释方法和验证数据集各自独立，缺少综合解释框架下的统一验证。由此，依托课题组两机重大专项，搭建 300 KW 高速重载行星齿轮传动系统试验平台，用以全面测试所提综合解释框架在故障诊断性能和解释效果方面的可行性。

300 KW 高速重载行星齿轮传动系统试验平台如图 6-3 所示，实验台由两台交流异步电机（TT AMP 280-4C, 450 KW）、SYT1300 行星齿轮传动系统、电机控制系统、油液润滑系统、计算机和数据采集系统（东华 DH8303）组成。加速度传感器安装在行星齿轮传动系统的外齿圈上，用于采集行星齿轮箱不同工况下的稳态振动信号。行星齿轮传动系统的关键参数和特征频率如表 6-2 所示。

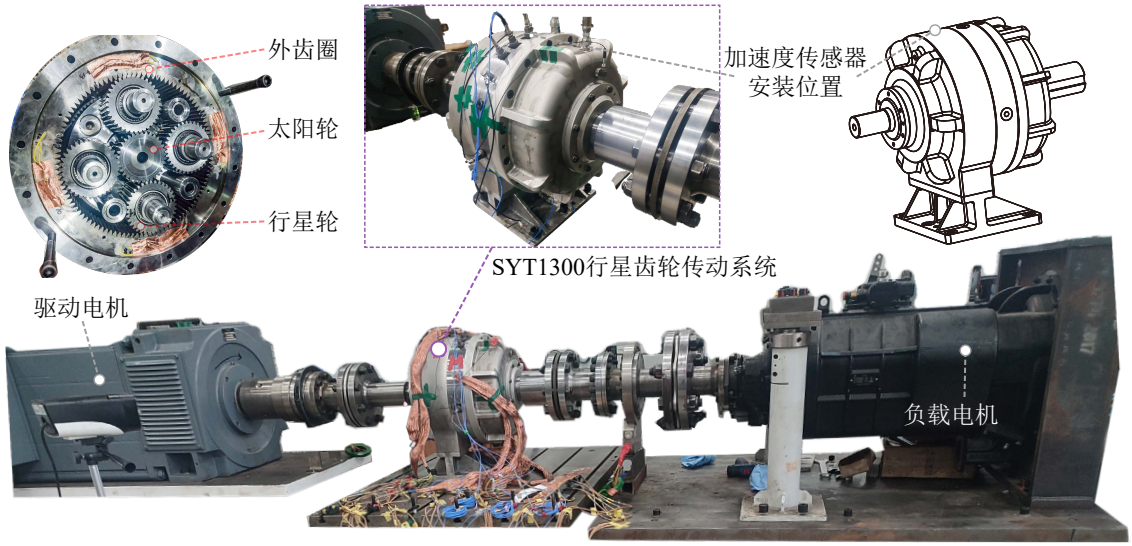


图 6-3　行星齿轮传动系统试验平台

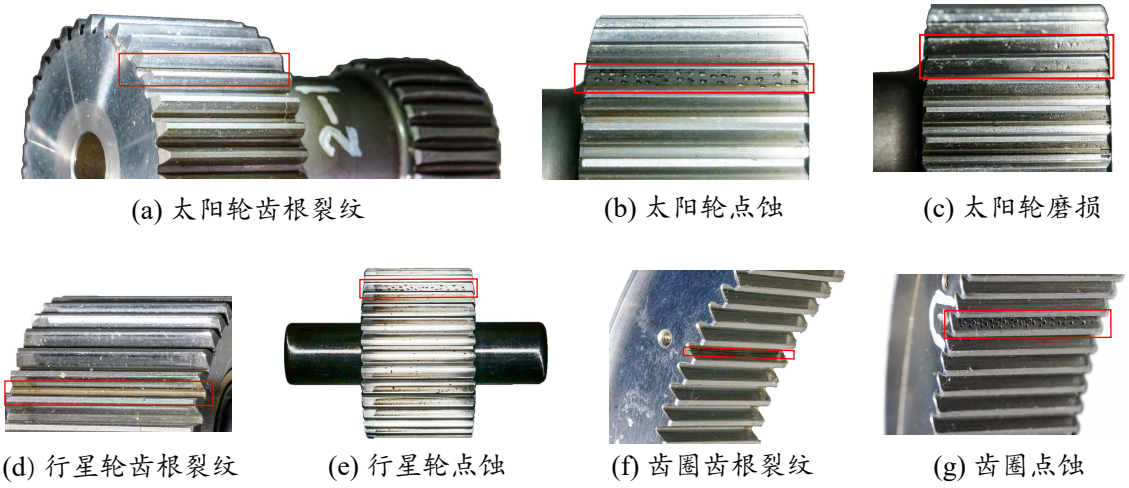
Fig. 6-3　Experimental platform of the planetary gear transmission system

行星传动系统的各故障部件如图 6-4 所示，包括太阳轮的齿根裂纹（SC）、齿面点蚀（SP）和齿面磨损（SW）故障，行星轮的齿根裂纹（PC）和齿面点蚀故障（PP）、以及齿圈的齿根裂纹（RC）和齿面点蚀（RP）故障，共计 7 种故障状态。各类故障均采用人工植入方式，其中齿轮裂纹的深度设置为 1 mm，齿面点蚀和齿面磨损则分别为 0.1 mm 和 0.02 mm。实验中，将输入轴转速 f_s 设置为 2400 RPM，采样频率设置为 20 kHz，采集包括健康（H）状态的八种工况下的两分钟稳态振动信号，用以构建后续分析的数据集。

表 6-2　行星传动系统试验台的关键参数和特征频率

Table 6-2　Key parameters and characteristic frequencies of the planetary gear transmission system test rig

试验台参数	值	试验台参数	值
太阳轮齿数 (Z_s)	34	行星轮齿数 (Z_p)	36
外齿圈齿数 (Z_r)	106	行星轮个数 (k)	4
太阳轮转频 (f_s / Hz)	$n/60$	保持架转频 (f_p / Hz)	$\approx 4.118 f_s$
啮合频率 (f_m / Hz)	$\approx 25.743 f_s$	太阳轮特征频率 (f_{ps} / Hz)	$\approx 3.029 f_s$
行星轮特征频率 (f_{pp} / Hz)	$\approx 0.715 f_s$	外齿圈特征频率 (f_{pr} / Hz)	$\approx 0.971 f_s$



(a) 太阳轮齿根裂纹　(b) 太阳轮点蚀　(c) 太阳轮磨损

(d) 行星轮齿根裂纹　(e) 行星轮点蚀　(f) 齿圈齿根裂纹　(g) 齿圈点蚀

图 6-4　行星齿轮传动系统的故障部件

Fig. 6-4　Fault components of the planetary gear transmission system

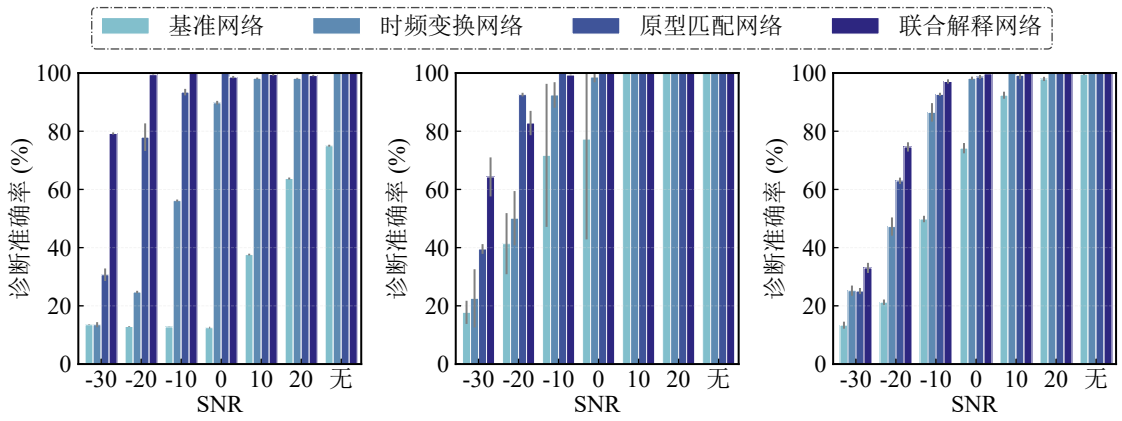
在训练方面，行星传动系统数据集实验可以被视为一个八分类任务。每个类别包含 800 个样本，每个样本长度为 5000，通过均值-标准差归一化处理，训练参数和 2.4.1 小节保持一致。

6.4　综合解释框架中主动解释模型的诊断性能实验验证

主动解释方法需要定制模型特定结构，一定程度上会影响模型的拓展性和灵活性，甚至损害模型的诊断性能。为保证如图 6-1 所示的三类主动解释模型在故障诊断方面的表现，现将主流的、具有不同映射能力的 MLP、CNN 和 BiLSTM 分别作为基准网络，并通过行星齿轮传动系统数据集构建高噪声和少样本两类故障诊断场景，对三类主动解释模型开展诊断性能实验验证。

6.4.1　噪声场景下的旋转机械故障诊断性能实验

原始故障诊断任务的难度较低，难以有效区分各模型的诊断性能，由此通过数据加噪方式来提高诊断任务难度。分别以 MLP、CNN 和 BiLSTM 三类不同学习能力的网络为基准，来构建对应的时频卷积网络、原型匹配网络和联合解释网络。所得到的四类模型在行星齿轮传动系统数据集不同信噪比下的诊断准确率如图 6-5 所示。



(a) 基准网络为 MLP　(b) 基准网络为 CNN　(c) 基准网络为 BiLSTM

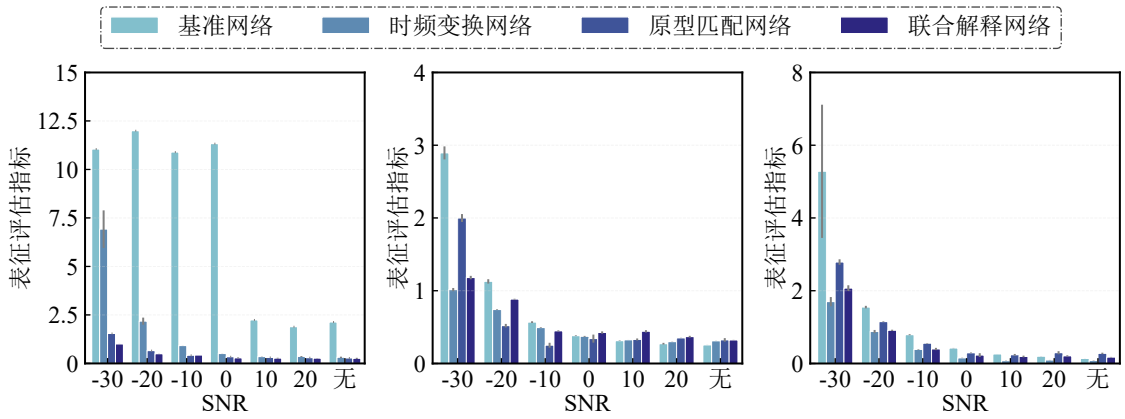
图 6-5　基准网络和对应主动解释模型在行星齿轮传动系统数据集不同信噪比下的诊断准确率

Fig. 6-5　The diagnostic accuracy of four models on the planetary gear transmission system dataset at different signal-to-noise ratios

从基准网络来看，MLP 的学习能力远低于 CNN 和 BiLSTM，其在 SNR 低于 0 的高噪声情况准确率低于 20%。随着 SNR 的升高，MLP 的准确率随之提升，但最终的准确率仍低于 85%。CNN 和 BiLSTM 的表现相近，在无噪声情况下均能达到 100%，但 CNN 在高噪声场景下的准确率比 BiLSTM 略高。再从三类解释网络来看，时频卷积网络和原型匹配网络分别将时频变换和原型匹配逻辑融入网络特定结构，有效提高基准模型的诊断精度，其诊断精度优势在三类基准网络上得到充分验证。作为两者的组合，联合解释网络的诊断准确率则是更为优异，在同种基准网络、同噪声强度情况下均领先其他模型。

除诊断准确率这一指标外，还可通过式 (3-18) 所示的指标 R_{rps} 来评估模型的表征学习能力。基准网络和对应主动解释模型在行星齿轮传动系统数据集不同信噪比下的表征评估指标 R_{rps} 如图 6-6 所示。对于学习能力较差的 MLP，三种主动解释方法能够显著的提高基准网络的表征学习能力，将表征评估指标 R_{rps} 从 SNR = -30 高噪声场景下的 12.5 降低至 2 附近。对于有充足学习能力的 CNN 和 BiLSTM，三种主动解释方法对基准网络的表征学习能力仍有一定程度的提高，且在高噪声场景下更

为明显。



(a) 基准网络为 MLP　(b) 基准网络为 CNN　(c) 基准网络为 BiLSTM

图 6-6　基准网络和对应主动解释模型在行星齿轮传动系统数据集不同信噪比下的表征评估指标 R_{rps}

Fig. 6-6　The representation metric R_{rps} of four models on the planetary gear transmission system dataset at different signal-to-noise ratios

6.4.2　少样本场景下的旋转机械故障诊断性能实验

同样的思路，可通过降低训练样本数量来构建难度更高的少样本场景故障诊断任务，进而有效区分模型诊断性能。将行星齿轮传动系统数据集的噪声设置为 SNR=0，基准网络和对应主动解释模型在不同训练样本数目下的诊断准确率如图 6-7 所示。三类基准网络的诊断能力各有不同，在各类训练样本数目为 5 的极端少样本场景下，MLP 的准确率低于 20%，CNN 和 BiLSTM 接近 40%。而三类主动解释模型均能有效提高对应基准网络的诊断准确率，其中联合解释网络的提升效果最为明显，其在不同基准网络下的准确率分别为 45%、65%、75%。随着训练样本数目的增加，除 MLP 外基准网络的准确率都随之增加，但三类主动解释模型，特别是联合解释网络，仍保持显著的诊断精度优势。

在表征学习能力方面，基准网络和对应主动解释模型在行星齿轮传动系统数据集不同训练样本数目下的表征指标 R_{rps} 如图 6-8 所示。MLP 由于性能限制，其表征评估指标 R_{rps} 始终较低，三种主动解释方法对 MLP 的表征能力提升极为显著。对于具有更强学习能力的 CNN 和 BiLSTM，三种主动解释方法仍对表征能力有一定程度的提高。

总结而言，通过构建高噪声和少样本两种高难度故障诊断场景，表明综合解释框架下的三类主动解释模型在不同基准网络下均能有效提高模型的诊断性能（诊断准

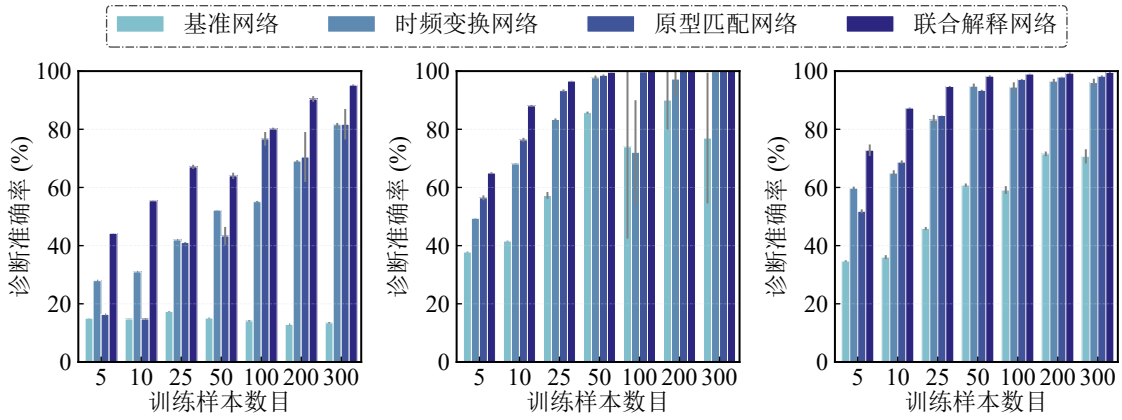


图 6-7　基准网络和对应主动解释模型在行星齿轮传动系统数据集不同训练样本数目下的诊断准确率

Fig. 6-7　The diagnostic accuracy of four models on the planetary gear transmission system dataset at different numbers of training samples

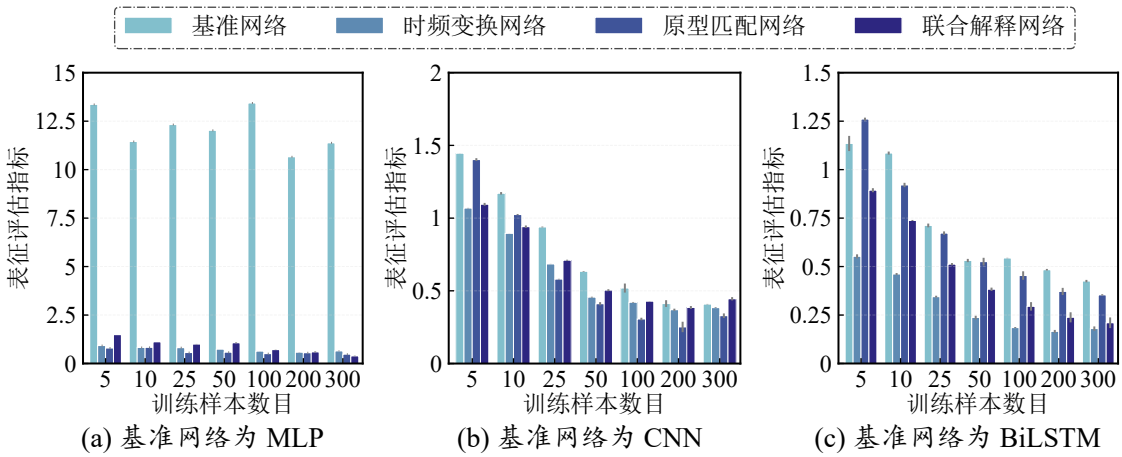


图 6-8　基准网络和对应主动解释模型在行星齿轮传动系统数据集不同训练样本数目下的表征指标 R_{rps}

Fig. 6-8　The representation metric R_{rps} of four models on the planetary gear transmission system dataset at different numbers of training samples

确率和表征评估指标 R_{rps}），且联合解释网络在诊断准确率和表征学习能力上均具有优势。这为综合解释框架的实际应用提供了有力支撑，用户可根据解释需求选择合适的主动解释方法，而无须担心构建的主动解释模块对模型诊断性能的潜在损害。

6.5　综合解释框架的全面解释效果实验验证

在验证主动解释方法的诊断性能优势之后，本节将分别对综合解释框架下的输入层主动解释、决策层主动解释和被动归因解释进行实验验证。由于 8 个故障类别对应的解释结果较为复杂，而实际应用中我们更关注故障的发生位置。因此，可解释性实验仅考虑健康（H）、太阳轮齿根裂纹（SC）、行星轮齿根裂纹（PC）和外齿圈齿根裂纹（RC）四种故障状态，以获得更清晰的解释效果。

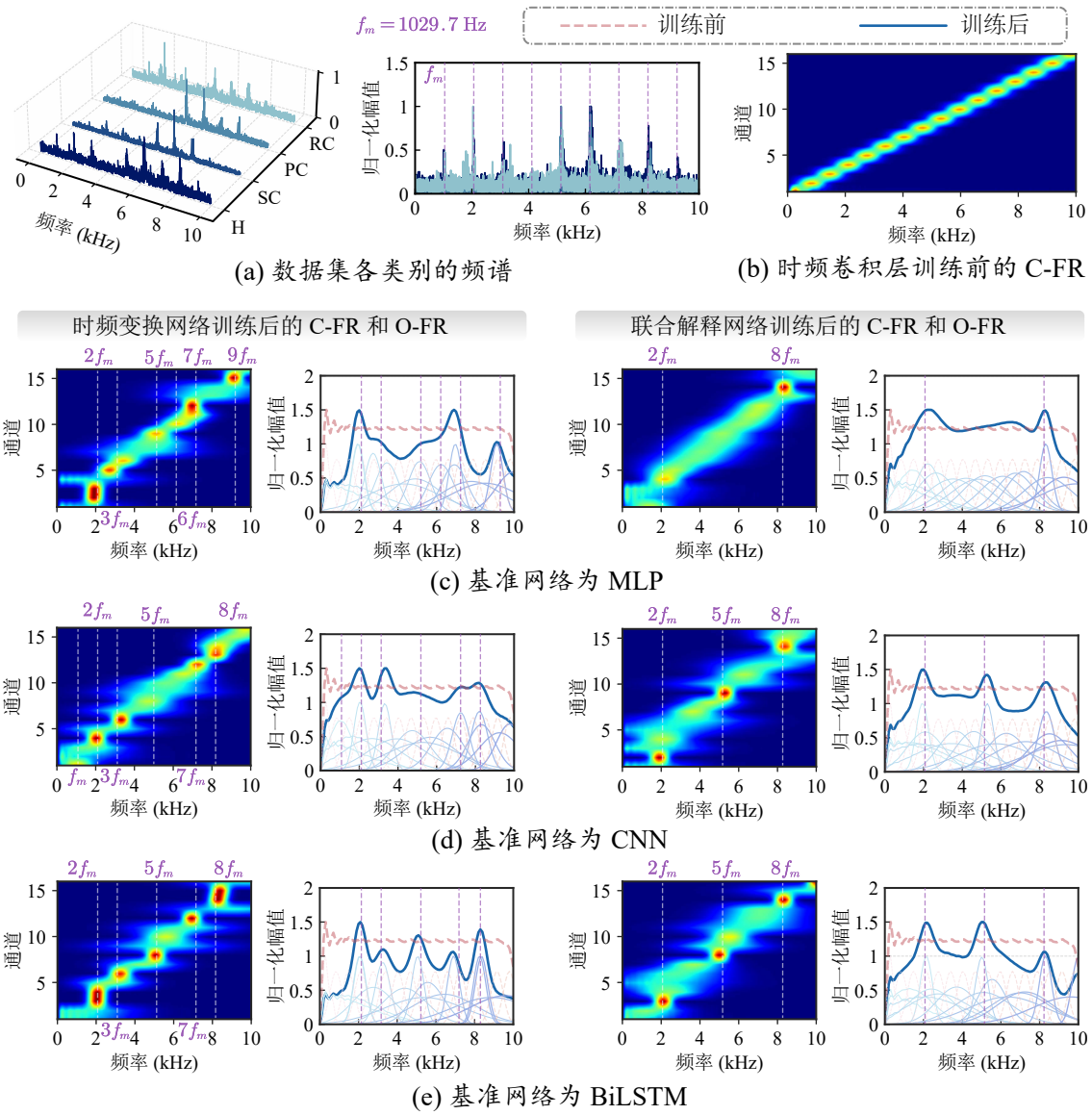
6.5.1　融入时频变换的输入层主动解释

时频卷积网络和联合解释网络均采用可解释的时频变换卷积层作为模型的预处理层，由此可通过 2.3.2 小节所示的幅频响应分析方法，解释模型对不同频带的关注程度，即输入层主动解释。

以 SNR = 0 的行星齿轮传动系统数据集为输入，其频谱以及不同基准网络下 Chirplet 时频卷积层训练前后的 C-FR 和 O-FR 如图 6-9 所示。从频谱图中可看出，故障类别的信息主要体现在齿轮啮频 $f_m = 1029.7\,\mathrm{Hz}$ 及其倍频上。一方面，一个良好的诊断模型应该对这些频率重点关注，另一方面，各频率的重要性也并非等同，部分频率可能承载着更多的故障频率调制信息则更为关键，而部分频率可能不具有区分性则应该忽视。智能诊断领域可解释性研究缺乏真实标签是一个普遍问题，但大体的评估准则可总结为：(1) 模型关注频率应该为齿轮啮频 f_m 及其倍频；(2) 不同模型的解释结果应尽可能保持统一，保证内在解释逻辑的一致性；(3) 输出层主动解释的结论需与后续的 SHEP 解释相互印证。满足上述评估准则的解释结果便可充分证实所提综合解释框架的有效性。

由图 6-9 可知，时频卷积层参照信号处理时频变换方法进行初始化，其训练前频响的各个通道中心频率呈均匀分布，对数据集不具有针对性。经训练后，各类模型的 C-FR 和 O-FR 普遍收敛至齿轮啮频 f_m 及其倍频上，以 $2f_m$、$5f_m$ 和 $8f_m$ 最为突出，呈现出强烈的数据集针对性。

从主动解释方法来看，时频卷积网络以时频卷积层作为模型的唯一入口，需要尽可能的关注所有承载类别信息的频带以提高模型诊断精度，由此时频卷积网络训练后的 C-FR 和 O-FR 会收敛至较多的频带，包括 $2f_m$、$3f_m$、$5f_m$、$7f_m$ 和 $8f_m$。而联合解释网络则在时频卷积网络的基础上，通过特征融合器将时频卷积层的特征与编码器获取的特征进行融合，时频卷积层和编码器输入频谱共同作为模型的数据入口。借助额外的编码器作为补充，此时的时频卷积层仅需关注更为关键的信息频带，使得



(a) 数据集各类别的频谱　(b) 时频卷积层训练前的 C-FR

(c) 基准网络为 MLP

(d) 基准网络为 CNN

(e) 基准网络为 BiLSTM

图 6-9　行星齿轮传动系统数据集频谱、以及时频卷积网络和联合解释网络中时频卷积层训练前后的 C-FR 和 O-FR

Fig. 6-9　The spectrum of the planetary gear transmission system dataset, and the C-FR and O-FR of the TFconv layer in the TFN and Joint Interpretation Network after training

联合解释网络训练后的 C-FR 和 O-FR 更为集中，如 $2f_m$、$5f_m$ 和 $8f_m$。

从基准网络差异来看，不同基准下的时频卷积网络在训练后的频响上确实存在些许差异，MLP 的部分频响关注到独特的 $6f_m$ 和 $9f_m$，而 CNN 的部分频响则收敛到 f_m 上，但三类基准模型的频响均是正确地收敛至齿轮啮频和其倍频的。而联合解释网络由于频响更为集中，不同基准下的频响结果更为一致，MLP 收敛至 $2f_m$ 和 $8f_m$，而 CNN 和 BiLSTM 则均收敛至 $2f_m$、$5f_m$ 和 $8f_m$。

总结而言，不同类型的输入层主动解释方法都能收敛至至齿轮啮频和其倍频，表明解释结果是合理的。其中，时频卷积网络需要提取尽可能多的类别信息，使得其训练后频响的收敛频带更为丰富，也更易受基准网络差异的影响；而联合解释网络则通过编码器的补充，使得时频卷积层能够更为集中地关注到关键频带，其训练后的解释结果更为清晰，在不同基准网络下有更好的解释结果一致性。输入层主动解释获得的结果主要集中在 $2f_m$、$5f_m$ 和 $8f_m$ 上，该解释结论可在后续的 SHEP 被动解释分析中，进行进一步地验证。

6.5.2　基于原型匹配的决策层主动解释

原型匹配网络和联合解释网络均采用可解释的原型匹配层作为模型的决策层，由此可通过 3.3.3 小节所示的分析方法，解释模型基于相似性的分类逻辑和各类别的典型频域原型，即决策层主动解释。

以 SNR = 0 的行星齿轮传动系统数据集为输入，以 CNN 网络为基准，输入样本的频谱、以及原型匹配网络和联合解释网络的重构原型和输出距离的 C-FR 和 O-FR 如图 6-10 所示。从频谱图中可看出，由于额外噪声的引入，输入样本的频谱变得较

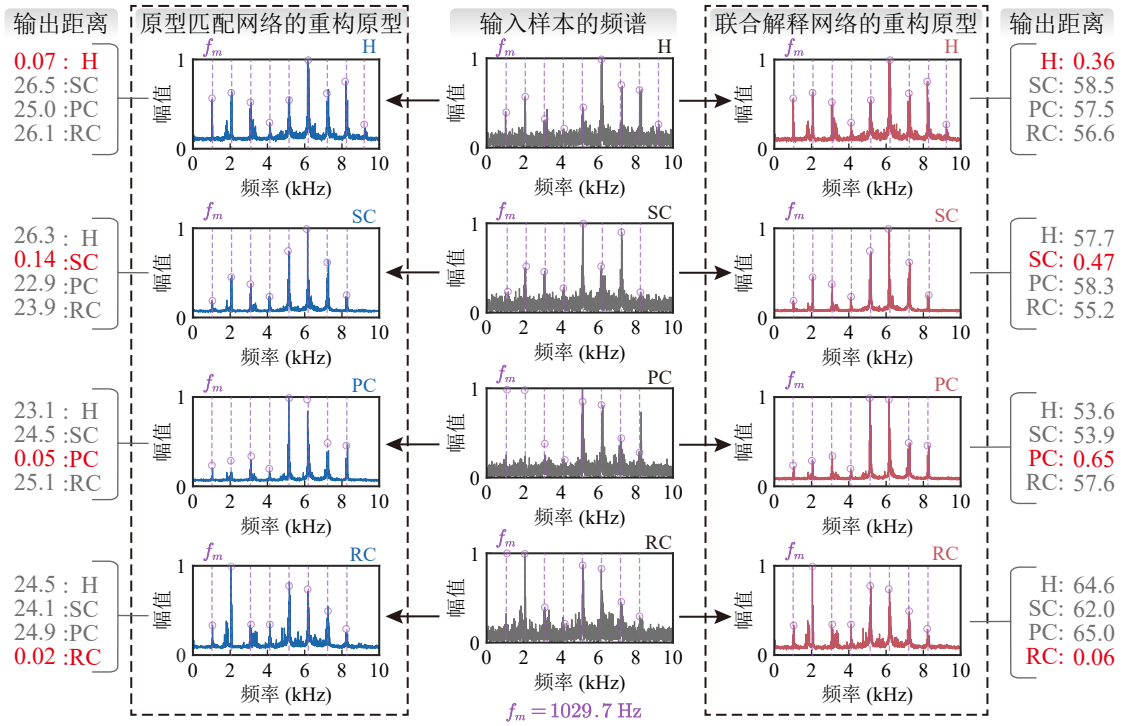


图 6-10　输入样本的频谱、以及原型匹配网络和联合解释网络的重构原型和输出距离

Fig. 6-10　The spectrum of input samples, and the reconstructed prototypes and output distances of the PMN and Joint Interpretation Network

为模糊，部分关键啮频的倍频（如健康类别的 $9f_m$）被噪声所掩盖。一方面，这增加了模型诊断的难度，需要模型能够从噪声中捕捉故障特征，提取除具有足够区分性的高维特征用以原型匹配；另一方面，所学到的原型是同类别样本在原型空间的几何中心，其重构出的样本级原型也有望克服噪声的干扰，恢复原本的频谱表征。

从重构原型来看，原型匹配网络和联合解释网络的重构原型均能较好地恢复输入样本的频谱特征，被噪声淹没的关键特征（如健康类别的 $9f_m$）也随着噪声去除而凸显，且两者网络的重构原型近乎一致，表明决策层主动解释具有良好的一致性。

从分类逻辑来看，原型匹配网络和联合解释网络的输出距离均能有效区分不同故障类别，样本与对应类别原型的距离显著小于其他类别的距离，进而实现正确的故障诊断预测。尽管联合解释网络的最近距离的幅值略高于原型匹配网络，但其它类别的幅值也有响应提升，实际上的表征学习能力是相近且优秀的。

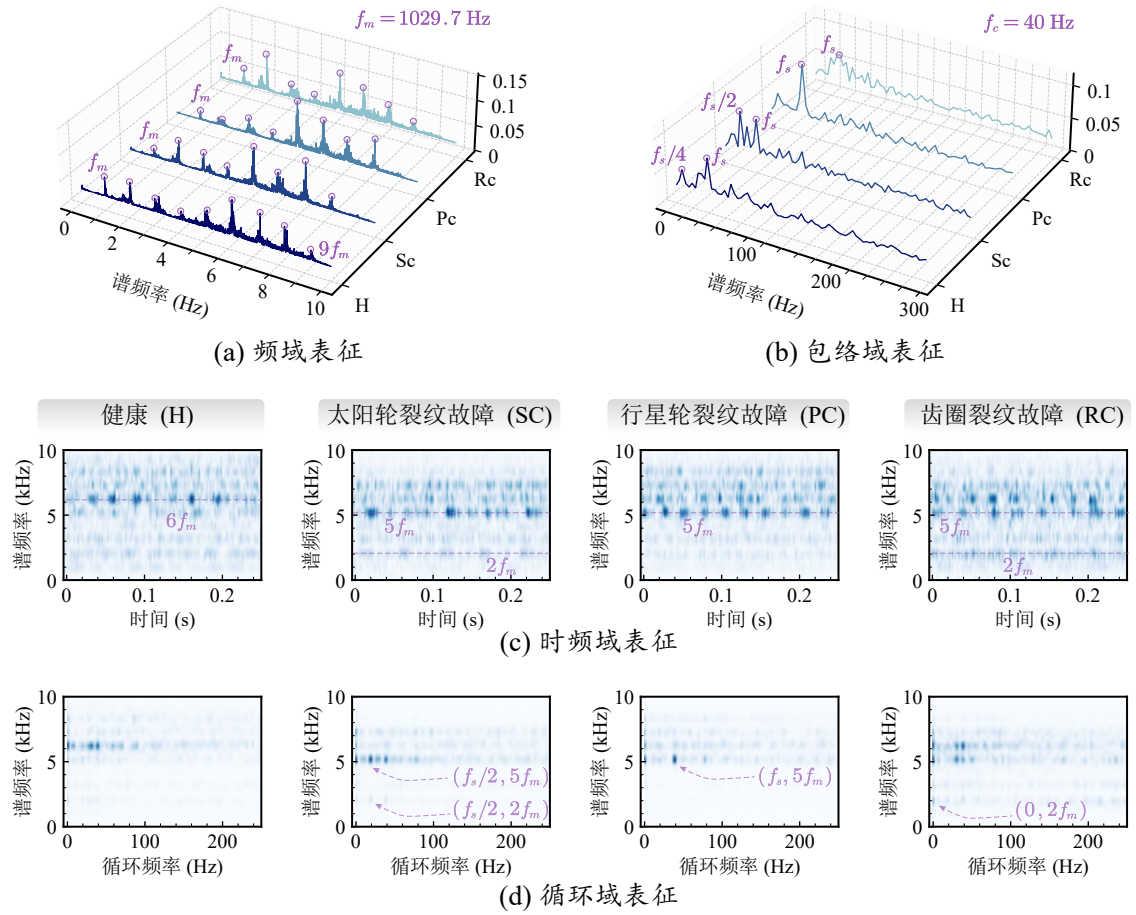
原型匹配网络和联合解释网络的解释结果表明，综合解释框架的决策层主动解释方法均能够克服噪声干扰，通过重构原型强化故障特征信息，对各类别典型频谱进行清晰刻画。另一方面，通过引入显式的原型匹配逻辑，决策层主动解释方法也能有效的区分不同故障类别，具有优秀的表征学习能力。

6.5.3　结合域变换的 SHEP 归因被动解释

SHEP 归因被动解释无须引入额外的模型约束，具有完全的兼容性，既可适用于无法参与模型设计的解释场景，也可为综合解释框架中的主动解释模型提供补充和交叉验证。

在进行 SHEP 解释前，还需对数据集的特征进行分析，以构建更准确的可解释性评估标签。无噪声行星齿轮传动系统数据集下各类样本在不同域的表征如图 6-11 所示。从频域表征来看，各类别样本的幅值主要体现在啮频 f_m 及其倍频。值得注意的是，H 类别具有独特的高幅值 f_m 和 $9f_m$ 频率，而 PC 类别具有独特的低幅值 $2f_m$ 频率。从包络域表征来看，各类别均具有显著的太阳轮转频 f_s 成分，但 H 类别和 SC 类别还分别具有额外的 $f_s/4$ 频率和 $f_s/2$ 频率。从时频域表征来看，各类别的谱频率以啮频 f_m 及其倍频为主，而冲击的周期则以太阳轮转频 f_s 及其分频为主，但较为杂乱，难以清晰辨认。循环域的结果进一步验证了时频域表征的结论，但各类别的信号成分更为直观。其中，信号的谱频率非常清晰，但循环频率则具有多种成分。

以无噪声行星齿轮传动系统数据集作为训练样本，对以 CNN 为基准的联合解释网络开展 SHEP 归因解释分析，所获得的各类样本对所属故障类别的不同域 SHEP 解释结果如图 6-12 所示。



(a) 频域表征　　(b) 包络域表征

(c) 时频域表征

(d) 循环域表征

图 6-11　行星齿轮传动系统数据集下各类样本在不同域的表征

Fig. 6-11　The representation of samples of each class in different domains under the planetary gear transmission system dataset

从频域解释结果（Freq-SHEP）可以发现，决策贡献度与频率幅值并非成正相关，高幅值的频率并非一定具有高贡献度。众多的啮频 f_m 及其倍频都具有显著的幅值，但仅有部分频率具有足够的贡献度。其中，PC 和 RC 类别的贡献度来源非常单一，仅仅为 $2f_m$ 频率；SC 类别的贡献度来源则包括 $2f_m$、$5f_m$ 和 $8f_m$；H 类别的贡献度来源则非常丰富，还包括额外的 f_m 和 $9f_m$。事实上，决策贡献度应与频率的区分性成正相关。H 类别中的高贡献频率 f_m 和 $9f_m$ 与图 6-11(a) 相吻合，其中 H 类别中 f_m 和 $9f_m$ 频率的幅值远远高于其他类别，具有明显的区分性。此外，PC 类别的 $2f_m$ 频率显著低于其他类别，这种区分性也支撑了图 6-12(a) 中 $2f_m$ 频率在 PC 类别上的高贡献。

从包络域解释结果（Env-SHEP）来看，各类样本中具有显著幅值的循环频率 f_s 均未具有明显贡献。相反地，H 类别中独特的 $f_s/4$ 和 SC 类别中独特的 $f_s/2$ 循环频

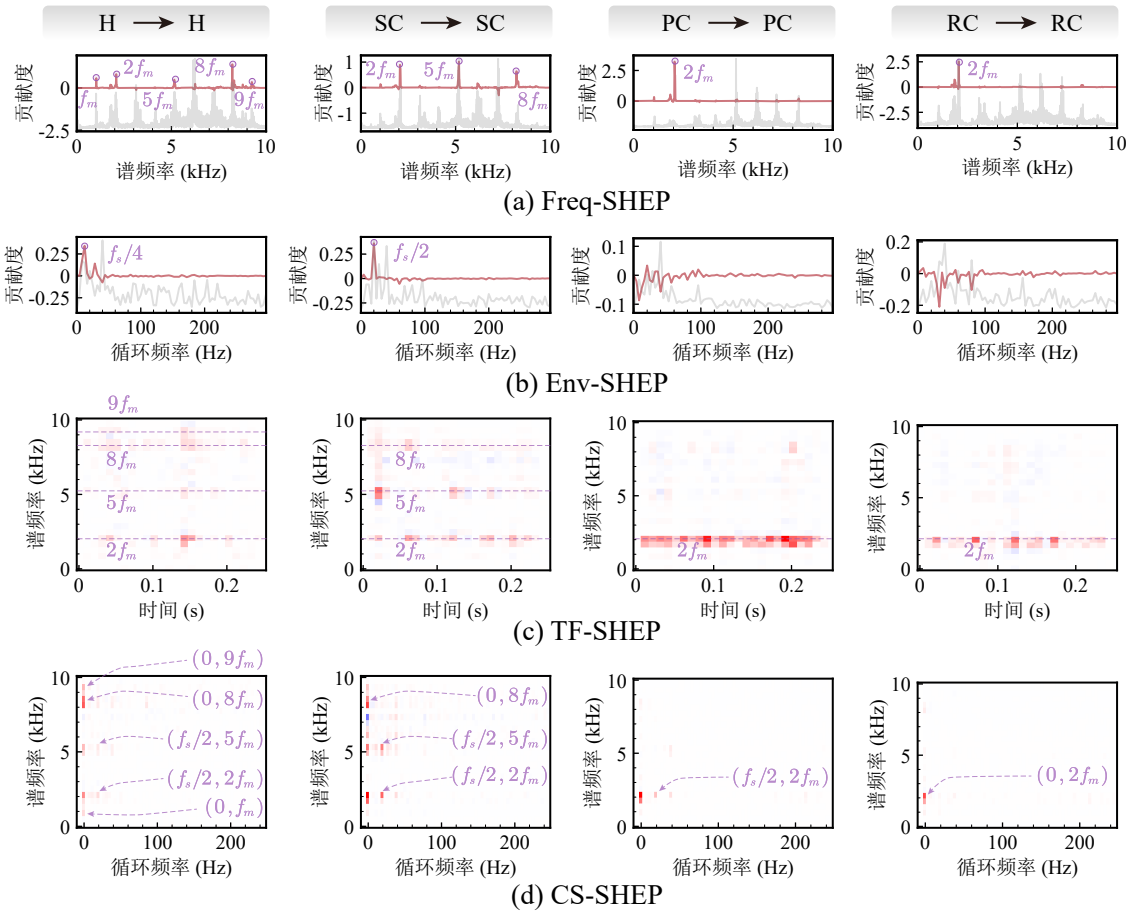


图 6-12　行星齿轮传动系统数据集下各类样本对所属故障类别的不同域 SHEP 归因解释结果

Fig. 6-12　The SHEP attribution results of samples of each class to their fault categories in different domains under the planetary gear transmission system dataset

率能为对应类别提供足够区分度，因而在对应类别的 Env-SHEP 解释结果中具有显著的贡献。这一现象再次验证，决策贡献度与频率幅值并非成正相关，而应与频率的区分性成正相关。

时频域解释结果（TF-SHEP）和 Freq-SHEP 大体保持一致，其中 PC 和 RC 类别的贡献度来源仍然为 $2f_m$ 频率，SC 类别的贡献度来源则额外地增加了 $5f_m$ 和 $8f_m$ 频率，而 H 类别的贡献度来源则进一步增加了 $9f_m$ 频率。此外，TF-SHEP 在具有明显调制现象的情况还能由贡献度的周期时间来解释信号成分的调制频率，例如 H 类别的 $2f_m$ 成分具有明显的调制现象，其贡献度的调制频率为 $f_s/2 = 20$Hz，周期时间约为 0.05 s。

循环域解释结果（CS-SHEP）能够从谱频率和循环频率两方面对信号成分进行刻画，从而获得更全面和准确的解释结果。其中，H 类别的贡献度可大体来源于

$(0, f_m)$、$(f_s/2, 2f_m)$、$(f_s/2, 5f_m)$、$(0, 8f_m)$ 和 $(0, 9f_m)$ 信号成分，SC 类别的贡献度来源于 $(f_s/2, 2f_m)$、$(f_s/2, 5f_m)$ 和 $(0, 8f_m)$ 信号成分，PC 类别和 RC 类别的贡献度来源则分别仅为 $(f_s/2, 2f_m)$ 和 $(0, 2f_m)$ 信号成分。这一结果与 Freq-SHEP 的解释结果保持一致，但更为直观和全面。

此外，上述的 SHEP 解释结果还可与 6.5.1 小节的输入层主动解释结果进行相互印证。如图 6-9所示，联合解释网络的频响在训练后会稳定地收敛至 $2f_m$、$5f_m$ 和 $8f_m$ 频率，而这三个频率也正是图 6-12 所示 SHEP 解释结果中各故障类别的主要贡献度来源。两种不同方式的解释结果成功地相互印证，侧面地验证了综合解释框架的有效性，能够为用户提供更为全面和准确的解释结果。

总结而言，结合域变换的 SHEP 归因被动解释能够准确、全面地揭示输入样本中不同成分对智能诊断模型不同决策的贡献度。SHEP 解释结果表明，决策贡献度的大小并非由输入信号的频率幅值决定，而应与各成分的区分性成正相关。最终，行星齿轮传动系统数据集中 $2f_m$、$5f_m$ 和 $8f_m$ 频率成分对模型的决策贡献度最大，这一结论与输入层主动解释的结果相互印证，验证了综合解释框架的有效性。此外，不同域的 SHEP 解释结果对信号成分的分析角度（时间、谱频率、循环频率）不同，但都能够很好地对模型决策依据进行解释，用户可根据实际解释任务需求对转换域进行择优选取。

6.6　本章小结

针对当前智能诊断可解释性研究中的零散性与表面性问题，本章将前四章独立的高兼容主动解释、高性能被动解释进行整合，开发了面向旋转机械智能诊断模型的综合解释框架，以帮助用户根据任务需求选择合适解释方法，从而提高本研究的系统性和实用性。其中，输入层主动解释和决策层主动解释相结合以形成联合解释网络，通过分情况讨论，建立了各解释方法的综合应用流程。最后依托高速重载行星齿轮传统系统，对框架中主动解释方法进行诊断性能测试，并对全体解释方法的解释能力进行验证。本章的研究内容可总结为：

(1) 建立了输入层和决策层共同解释的联合解释网络。将现有基准网络作为编码器，与输入层主动解释的时频变换层共同提取特征并进行特征融合，并使用原型匹配层基于所得融合特征进行故障分类，从而构建联合解释网络。所得联合解释网络同时具备输入层和决策层的主动解释能力，在揭示模型关注频带的同时，还能显式地学习各类故障原型，强化故障关键特征。

(2) 提出了包含主动解释和被动解释的综合解释框架。以能否参与模型设计为界限，并以关键频率、决策逻辑和决策依据为解释需求，对现有三类主动解释方法和SHEP归因被动解释方法进行整合，形成了可指导用户进行方法选择的综合解释框架。

(3) 开展了三类主动解释方法的诊断性能实验。考虑主动解释对模型结构的约束，依托实测行星齿轮传动系统构建高噪声和少样本两类故障诊断场景，验证了综合解释框架下的三类主动解释模型在不同基准网络下均能有效提高模型的诊断性能，且联合解释网络在诊断准确率和表征学习能力上均具有优势。这为综合解释框架的实际应用提供了有力支撑，使用户可根据解释需求选择合适的主动解释方法，而无须担心构建的主动解释模块对模型诊断性能的潜在损害。

(4) 验证了综合解释框架的全面解释效果。依托实测行星齿轮试验台，开展输入层主动解释、决策层主动解释和SHEP归因被动解释三方面分析。输入层主动解释的实验表明，时频卷积网络和联合解释网络均能有效关注到齿轮啮频及倍频，但联合解释网络的解释结果更为稀疏，且在不同基准网络下的解释结果更为一致。决策层主动解释的实验表明，原型匹配网络和联合解释网络均能有效克服噪声干扰，通过重构原型强化故障特征。SHEP归因被动解释的实验表明，结合域变换的SHEP方法能够获得准确、全面的归因解释结果，与输入层主动解释的结果成功印证，且各成分贡献度大小与其区分性成正相关。

第七章 总结与展望

7.1 全文工作总结

深度学习智能诊断是当前大容量、低密度、多样性和时效性的大数据背景下实现旋转机械运维保障的有力途径。但深度学习作为典型的黑箱模型，其诊断依据、诊断逻辑和适用范围均不明晰，且在可信性、性能优化、缺陷分析等方面缺少科学指导，缺乏可解释性已成为了深度学习从学术研究走向工业应用的巨大阻碍。为此，本文瞄准我国重大装备高可靠性和安全性的迫切需求，以提高旋转机械智能诊断模型可解释性为目标，针对现有主动解释中拓展性差、缺少明确解释结果，以及被动解释中解释形式差、计算高耗时的问题，分别建立约束程度低且解释效果明确的高兼容性主动解释网络，以及解释形式清晰且计算效率高效的被动解释方法。最终形成面向旋转机械智能诊断模型的综合解释框架，为智能诊断的解释需求提供系统性和实用性方案。本文的主要研究结论如下：

(1) 针对主动解释效果与模型诊断性能、可拓展性的多方共赢问题，提出基于时频变换的诊断模型输入层主动解释方法。首先阐明时频变换与神经网络卷积层在内积运算上的等价性，然后通过实虚部机制和核函数约束，建立等效于可学习时频变换、具有三类核函数的时频卷积层，并提出以训练后频响揭示模型关注频带的解释分析方法。多个实测数据集的实验结果充分表明，由于引入先验信号处理知识，所提输入层主动解释方法，能够有效提高基准模型在诊断精度、少样本学习能力和收敛速度等方面的性能，同时所学习到的频响能够与数据集关键频带准确对应，有效验证输入层主动解释方法的有效性。

(2) 在提高输入层可解释性之后聚焦智能诊断模型的决策层，提出基于原型匹配的诊断模型决策层主动解释方法。首先基于原型匹配概念，建立能够显式构建原型向量、并基于样本与各原型相似程度进行故障分类的原型匹配层。然后将原型匹配层作为决策层与自编码器相结合，并通过分类损失、重构损失和原型匹配距离损失对网络进行联合训练。最后，梳理原型匹配网络在分类逻辑、类别原型、相似性来源三方面的解释能力，以形成直观、实用的解释结果输出。传统故障诊断和领域泛化两类任务表明，所提原型匹配网络与当前先进方法相比，在诊断准确率方面略有领先，且在表征能力上具有显著优势。此外，重构的原型样本能够有效抑制噪声干扰并发掘微弱故障特征，强化信号中的机理信息，有

效验证决策层主动解释方法的解释能力。

(3) 针对被动解释形式的直观性不足问题，提出结合域变换的诊断模型被动解释形式优化方法。首先基于传统的循环谱分析，提出确定信号的二维自相关函数的近似估计，进而建立时域信号到循环域表征的域变换及其逆变换方法。然后基于域变换开展样本预处理和模型集成，实现将解释形式从时域拓展至循环域的CS-SHAP 被动解释归因方法。仿真数据集和两个实测数据集的实验结果表明，所提 CS-SHAP 借助循环域的双维度分析特点，不仅使归因解释结果更为清晰、直观，而且能够有效区分多个相近故障成分的不同贡献，避免成分混淆，保证解释的准确性和可靠性。此外，CS-SHAP 在不同模型和噪声强度下，都获得较为不错的解释效果，充分验证所提被动解释方法的优势。

(4) 针对被动解释的高耗时计算问题，提出结合组合块归因和子集枚举缩减的诊断模型被动解释效率优化方法。首先对 SHAP 计算过程进行理论性分析，建立绑定多个相邻特征以求解联合贡献的组合块归因策略，从而以解释粒度为代价降低归因计算的特征维度。然后对耗时的子集枚举过程进行简化，通过 SHEP-Remove 和 SHEP-Add 两个代表性实例形成 SHEP 方法，从而将计算复杂度从原始 SHAP 的指数级降低到线性级。仿真数据集和两个实测数据集的实验结果表明，SHEP-Remove 和 SHEP-Add 分别擅长捕捉特征的存在和不存在的贡献，结合两者的 SHEP 方法在多数据集、不同域和不同组合块尺寸下都与 SHAP 保持高度一致，充分验证适用 SHEP 来近似 SHAP 的可行性。在计算效率上，SHEP 相比 SHAP 具有显著的计算加速，其计算耗时与特征维度和背景样本数量 $\mathcal{O}(d\cdot n)$ 呈线性增长。

(5) 针对当前可解释性研究的零散性、表面性问题，建立面向旋转机械智能诊断模型的综合解释框架。首先对所提两类主动解释方法进行整合，建立联合解释网络，实现对模型输入层和决策层的同时主动解释。然后对各类主动解释方法及 SHEP 被动解释方法进行整合，以模型设计为界限、以解释结果为需求，形成指导用户根据任务场景选择适合方法的综合解释框架。最后依托行星齿轮传动系统，对综合解释框架进行诊断性能和解释效果的全面验证。在高噪声和少样本实验中，三类主动解释网络均能有效提高不同基准模型的诊断准确率和表征学习能力，使得用户可专心于解释效果而无须担心模型约束带来的潜在损害。在解释效果方面，输入层主动解释能够有效且一致地关注到齿轮啮频及倍频，决策层主动解释能够克服噪声干扰通过重构原型以强化故障特征，SHEP 归因被

动解释则能准确、全面地评估各信号成分对决策的贡献，且与输入层主动解释的结果成功呼应，有效验证所提综合解释框架的实用性和有效性。

7.2 本文创新点

本文创新点在于：

(1) 提出了基于时频变换的诊断模型输入层主动解释方法。阐明了时频变换和卷积层的等价性，以此设计了融入时频变换的时频卷积层以及相应的核函数，建立了基于幅频响应的解释分析方法，最终作为输入层与基准模型相结合，在提高诊断性能的同时实现对模型关注频带的解释。

(2) 提出了基于原型匹配的诊断模型决策层主动解释方法。设计了基于原型匹配概念的原型匹配层以及促进原型收敛的原型匹配损失项，梳理了原型匹配网络在分类逻辑、类别原型、相似性来源三方面的解释能力，最终通过与自编码器模型相结合，在提高诊断性能的同时从模型视角刻画各类别的原型样本。

(3) 建立了结合域变换、组合块归因的诊断模型被动解释优化方法。一方面推导了循环域变换及其逆变换，提出了将解释结果拓展至循环域的 CS-SHAP 归因方法，不仅获得清晰直观的归因解释结果，而且能够有效区分近邻成分；另一方面设计了降低特征维度的组合块归因策略并对 SHAP 的子集枚举过程进行简化，建立更低计算复杂度的 SHEP 归因方法，最终在保持与原始 SHAP 解释一致性的同时，显著降低计算耗时。

7.3 研究展望

本文以提高旋转机械智能诊断模型的可解释性为目标，一方面通过时频变换和原型匹配分别对模型输入层和决策层进行深入设计，建立约束程度低、解释效果明确的高兼容主动解释方法；另一方面通过域变换、组合块归因和子集枚举缩减分别对模型被动解释形式和计算效率进行优化，建立解释形式清晰、计算成本低的高性能被动解释方法，为解释诊断模型的决策逻辑、决策依据和提高模型可信性提供可行方案和技术源泉。但本工作仍存在一定局限，在如下方面尚需进一步探索：

(1) 建立包含解释标签、解释载体和评价指标的规范化可解释性评价体系。本文以旋转机械的特征频率及其转频作为解释标签，结合任务特点以训练频响、重构原型和归因贡献度为载体，实现了对旋转机械诊断模型的定性解释。但决策依据标签的合理性、解释载体的接受性和认可性、以及解释能力的量化评价指标

都有待商榷与完善，将各类独特的解释方法纳入统一赛道是当前可解释性领域亟需解决的关键问题。

(2) 开展迁移、无监督等复杂场景下的可解释性研究。本文聚焦于模型可解释性，主要在数据同分布的传统有监督故障诊断场景进行分析，对于迁移、无监督等复杂任务则尚未纳入研究范围。然而，可解释性研究的实际应用离不开这些复杂场景，后续将致力于在这些场景下设计更具有普适性和适应性的解释方法。

(3) 优化基于梯度传播的诊断模型被动解释方法。本文以基于扰动的 SHAP 为基础，在将解释结果拓展到循环域、频域等的同时，使模型交互次数降低为线性级。但梯度传播类解释方法仅需单次交互，拥有难以媲美的解释效率，是大规模或实时场景下被动解释的首选方案。因此，针对梯度传播类被动解释，优化其解释形式不直观、结果不稳定等问题，对实现诊断模型实时解释具有重要意义。

参考文献

[1] 铁道部安全监察司. 2004、2005 年铁路行车事故案例（合订本）[M]. 中国铁道出版社, 2006.

[2] 刘亮. 挪威直升机坠毁事故原因系技术故障[EB/OL]. (2016-05-04). https://news.cctv.com/2016/05/04/ARTIsxsLfORweHitAc6XcHyn160504.shtml.

[3] 陈光. PW4077 风扇叶片断裂引发的重大故障简析[J]. 航空动力, 2021(05): 36-38.

[4] 马娟. 国务院关于印发《中国制造 2025》的通知[EB/OL]. (2015-05-08). https://www.gov.cn/zhengce/content/2015-05/19/content_9784.htm.

[5] 国家自然科学基金会. 机械工程学科发展战略报告（2021-2035）[M]. 科学出版社, 2021.

[6] 雷亚国, 贾峰, 孔德同, 等. 大数据下机械智能故障诊断的机遇与挑战[J]. 机械工程学报, 2018, 54(5): 94-104.

[7] 王潘, 刘魁. 大数据技术在航空发动机中的应用[J]. 航空动力, 2018(01): 48-51.

[8] 李心萍. 用大数据发掘大价值（中国制造 2025 调研行）[EB/OL]. (2016-11-30). http://politics.people.com.cn/n1/2016/1130/c1001-28911045.html.

[9] 俞陶然. 图灵奖得主姚期智院士：人工智能存在三大技术瓶颈[EB/OL]. (2020-10-23). https://www.shkjdw.gov.cn/c/2020-10-23/525023.shtml.

[10] 上海人工智能实验室. 龚克：下一阶段 AI 基础研究应主攻 “可解释性”[EB/OL]. [2025-03-20]. http://www.shlab.org.cn/news/5443078.

[11] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge [J]. Nature, 2017, 550(7676): 354-359.

[12] Lei Y, Yang B, Jiang X, et al. Applications of machine learning to machine fault diagnosis: A review and roadmap[J]. Mechanical Systems and Signal Processing, 2020, 138: 106587.

[13] Zhao Z, Wu J, Li T, et al. Challenges and opportunities of AI-enabled monitoring, diagnosis & prognosis: A review[J]. Chinese Journal of Mechanical Engineering, 2021, 34: 56.

[14] Jia F, Lei Y, Lin J, et al. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data[J]. Mechanical Systems and Signal Processing, 2016, 72–73: 303-315.

[15] Jia F, Lei Y, Guo L, et al. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines[J]. Neurocomputing, 2018, 272: 619-628.

[16] Lu C, Wang Z Y, Qin W L, et al. Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification[J]. Signal Processing, 2017, 130: 377-388.

[17] Shao H, Jiang H, Lin Y, et al. A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders[J]. Mechanical Systems and Signal Processing, 2018, 102: 278-297.

[18] 李巍华, 单外平, 曾雪琼. 基于深度信念网络的轴承故障分类识别[J]. 振动工程学报, 2016, 29(2): 340-347.

[19] Wu X, Zhang Y, Cheng C, et al. A hybrid classification autoencoder for semi-supervised fault diagnosis in rotating machinery[J]. Mechanical Systems and Signal Processing, 2021, 149: 107327.

[20] Chen Z, Mauricio A, Li W, et al. A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks[J]. Mechanical Systems and Signal Processing, 2020, 140: 106683.

[21] Jing L, Zhao M, Li P, et al. A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox[J]. Measurement, 2017, 111: 1-10.

[22] Li X, Zhang W, Ding Q, et al. Intelligent rotating machinery fault diagnosis based on deep learning using data augmentation[J]. Journal of Intelligent Manufacturing, 2018, 31(2): 433-452.

[23] Yuan M, Wu Y, Lin L. Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network[C]//2016 IEEE International Conference on Aircraft Utility Systems (AUS). Beijing, China: IEEE, 2016: 135-140.

[24] Zhao R, Wang D, Yan R, et al. Machine health monitoring using local feature-based gated recurrent unit networks[J]. IEEE Transactions on Industrial Electronics, 2018, 65(2): 1539-1548.

[25] Shi J, Peng D, Peng Z, et al. Planetary gearbox fault diagnosis using bidirectional-convolutional LSTM networks[J]. Mechanical Systems and Signal Processing, 2022, 162: 107996.

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017: 6000-6010.

[27] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: vol. 1. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019: 4171-4186.

[28] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//9th International Conference on Learning Representations (ICLR). Austria: OpenReview.net, 2021.

[29] Hou Y, Wang J, Chen Z, et al. Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved Transformer[J]. Engineering Applications of Artificial Intelligence, 2023, 124: 106507.

[30] Ding Y, Jia M, Miao Q, et al. A novel time–frequency transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings[J]. Mechanical Systems and Signal Processing, 2022, 168: 108616.

[31] Xiao Y, Shao H, Wang J, et al. Bayesian variational transformer: A generalizable model for rotating machinery fault diagnosis[J]. Mechanical Systems and Signal Processing, 2024, 207: 110936.

[32] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[33] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[34] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montréal, Canada: MIT Press, 2014: 3104-3112.

[35] Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine[J]. Nature Biotechnology, 2018, 36(9): 829-838.

[36] Xiong H Y, Alipanahi B, Lee L J, et al. The human splicing code reveals new insights into the genetic determinants of disease[J]. Science, 2015, 347(6218): 1254806.

[37] Parks D, Prochaska J X, Dong S, et al. Deep learning of quasar spectra to discover and characterize damped Lyα systems[J]. Monthly Notices of the Royal Astronomical Society, 2018, 476(1): 1151-1168.

[38] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[EB/OL]. arXiv: 1312.6199(2014-02-19). https://arxiv.org/abs/1312.6199.

[39] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 427-436.

[40] 联合国教科文组织. 人工智能伦理问题建议书[EB/OL]. (2021-11-23). https://unesdoc.unesco.org/ark:/48223/pf0000381137_chi.

[41] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning[EB/OL]. arXiv: 1702.08608(2017-03-02). https://arxiv.org/abs/1702.08608.

[42] Zhang Y, Tino P, Leonardis A, et al. A survey on neural network interpretability[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2021, 5(5): 726-742.

[43] Wu M, Hughes M, Parbhoo S, et al. Beyond sparsity: Tree regularization of deep models for interpretability[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 32. New Orleans, Louisiana, USA: AAAI, 2018.

[44] Wu M, Parbhoo S, Hughes M, et al. Regional tree regularization for interpretability in deep neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 34. New York, NY, USA: AAAI, 2020: 6413-6421.

[45] Zhang Q, Wu Y N, Zhu S C. Interpretable convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 8827-8836.

[46] Plumb G, Al-Shedivat M, Cabrera Á A, et al. Regularizing black-box models for improved in-

terpretability[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates Inc., 2020: 10526-10536.

[47] Weinberger E, Janizek J, Lee S I. Learning deep attribution priors based on prior knowledge[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates, Inc., 2020: 14034-14045.

[48] Wojtas M, Chen K. Feature importance ranking for deep learning[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates, Inc., 2020: 5105-5114.

[49] Li O, Liu H, Chen C, et al. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 32. New Orleans, Louisiana, USA: AAAI, 2018.

[50] Chen C, Li O, Tao C, et al. This looks like that: Deep learning for interpretable image recognition[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019.

[51] Fu L. Rule learning by searching on adapted nets[C]//Proceedings of the Ninth National Conference on Artificial Intelligence. Anaheim, California, USA: AAAI, 1991: 590-595.

[52] Towell G G, Shavlik J W. Extracting refined rules from knowledge-based neural networks[J]. Machine learning, 1993, 13: 71-101.

[53] Setiono R, Liu H. Understanding neural networks via rule extraction[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995: 480-485.

[54] Odajima K, Hayashi Y, Tianxia G, et al. Greedy rule generation from discrete data and its use in neural network rule extraction[J]. Neural Networks, 2008, 21(7): 1020-1028.

[55] Nayak R. Generating rules with predicates, terms and variables from the pruned neural networks [J]. Neural Networks, 2009, 22(4): 405-414.

[56] Castro J, Mantas C, Benitez J. Interpretation of artificial neural networks by means of fuzzy rules [J]. IEEE Transactions on Neural Networks, 2002, 13(1): 101-116.

[57] Dhurandhar A, Chen P Y, Luss R, et al. Explanations based on the missing: Towards contrastive explanations with pertinent negatives[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates, Inc., 2018: 590-601.

[58] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR[J]. Harvard Journal of Law & Technology, 2017, 31: 841.

[59] Wang Y, Su H, Zhang B, et al. Interpret neural networks by identifying critical data routing paths [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 8906-8914.

[60] Erhan D, Bengio Y, Courville A, et al. Visualizing higher-layer features of a deep network[J]. University of Montreal, 2009, 1341(3): 1.

[61] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, 2015: 5188-5196.

[62] Yosinski J, Clune J, Nguyen A, et al. Understanding neural networks through deep visualization [EB/OL]. arXiv: 1506.06579(2015-06-22). https://arxiv.org/abs/1506.06579.

[63] Minematsu T, Shimada A, Taniguchi R i. Analytics of deep neural network in change detection[C]//2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Lecce, Italy: IEEE, 2017: 1-6.

[64] Bau D, Zhou B, Khosla A, et al. Network dissection: Quantifying interpretability of deep visual representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 6541-6549.

[65] Fong R, Vedaldi A. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 8730-8738.

[66] Dalvi F, Durrani N, Sajjad H, et al. What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 33. Honolulu, Hawaii, USA: AAAI, 2019: 6309-6317.

[67] Baehrens D, Schroeter T, Harmeling S, et al. How to explain individual classification decisions [J]. Journal of Machine Learning Research, 2010, 11(61): 1803-1831.

[68] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[EB/OL]. arXiv: 1312.6034(2014-04-19). https://arxiv.org/abs/1312.6034.

[69] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net [EB/OL]. arXiv: 1412.6806(2015-06-22). https://arxiv.org/abs/1412.6806.

[70] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 2921-2929.

[71] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 618-626.

[72] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017: 3145-3153.

[73] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLOS ONE, 2015, 10(7): e0130140.

[74] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017: 3319-3328.

[75] Fong R C, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 3429-3437.

[76] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, California, USA: ACM, 2016: 1135-1144.

[77] Plumb G, Molitor D, Talwalkar A S. Model agnostic supervised local explanations[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates, Inc., 2018: 2520-2529.

[78] Shapley L S. A value for n-person games[J]. Contribution to the Theory of Games, 1953.

[79] Chen J, Song L, Wainwright M, et al. Learning to explain: An information-theoretic perspective on model interpretation[C]//Proceedings of the 35th International Conference on Machine Learning: vol. 80. Stockholm, Sweden: PMLR, 2018: 883-892.

[80] Ivanovs M, Kadikis R, Ozols K. Perturbation-based methods for explaining deep neural networks: A survey[J]. Pattern Recognition Letters, 2021, 150: 228-234.

[81] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//Computer vision – ECCV 2014. Cham, Switzerland: Springer, 2014: 818-833.

[82] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017: 1885-1894.

[83] Yeh C K, Kim J, Yen I E H, et al. Representer point selection for explaining deep neural networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates, Inc., 2018: 9311-9321.

[84] 严如强, 商佐港, 王志颖, 等. 可解释人工智能在工业智能诊断中的挑战和机遇：先验赋能[J]. 机械工程学报, 2024, 60(12): 1-20.

[85] 严如强, 周峥, 杨远贵, 等. 可解释人工智能在工业智能诊断中的挑战和机遇：归因解释[J]. 机械工程学报, 2024, 60(12): 21-40.

[86] Abid F B, Sallem M, Braham A. Robust interpretable deep learning for intelligent fault diagnosis of induction motors[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(6): 3506-3515.

[87] Ravanelli M, Bengio Y. Interpretable convolutional filters with sincnet[EB/OL]. arXiv: 1811.09725(2019-08-09). https://arxiv.org/abs/1811.09725.

[88] Zhong J, Zheng Y, Ruan C, et al. M-IPISincNet: An explainable multi-source physics-informed

neural network based on improved SincNet for rolling bearings fault diagnosis[J]. Information Fusion, 2025, 115: 102761.

[89] Li T, Zhao Z, Sun C, et al. WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(4): 2302-2312.

[90] He C, Shi H, Liu X, et al. Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis[J]. Knowledge-Based Systems, 2024, 288: 111499.

[91] He C, Shi H, Li R, et al. Interpretable modulated differentiable STFT and physics-informed balanced spectrum metric for freight train wheelset bearing cross-machine transfer fault diagnosis under speed fluctuations[J]. Advanced Engineering Informatics, 2024, 62(A): 102568.

[92] He C, Shi H, Si J, et al. Physics-informed interpretable wavelet weight initialization and balanced dynamic adaptive threshold for intelligent fault diagnosis of rolling bearings[J]. Journal of Manufacturing Systems, 2023, 70: 579-592.

[93] He C, Shi H, Li J. IDSN: A one-stage interpretable and differentiable STFT domain adaptation network for traction motor of high-speed trains cross-machine diagnosis[J]. Mechanical Systems and Signal Processing, 2023, 205: 110846.

[94] Yuan J, Cao S, Ren G, et al. LW-Net: An interpretable network with smart lifting wavelet kernel for mechanical feature extraction and fault diagnosis[J]. Neural Computing and Applications, 2022, 34(18): 15661-15672.

[95] Han Y, Lv S, Huang Q, et al. AMCW-DFFNSA: An interpretable deep feature fusion network for noise-robust machinery fault diagnosis[J]. Knowledge-Based Systems, 2024, 301: 112361.

[96] Li T, Sun C, Fink O, et al. Filter-informed spectral graph wavelet networks for multiscale feature extraction and intelligent fault diagnosis[J]. IEEE Transactions on Cybernetics, 2024, 54(1): 506-518.

[97] Li T, Sun C, Li S, et al. Explainable graph wavelet denoising network for intelligent fault diagnosis [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(6): 8535-8548.

[98] Zhu P, Deng L, Tang B, et al. Digital twin-enabled entropy regularized wavelet attention domain adaptation network for gearboxes fault diagnosis without fault data[J]. Advanced Engineering Informatics, 2025, 64: 103055.

[99] Michau G, Frusque G, Fink O. Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series[J]. Proceedings of the National Academy of Sciences, 2022, 119(8): e2106598119.

[100] Wang H, Liu Z, Peng D, et al. Interpretable convolutional neural network with multilayer wavelet for noise-robust machinery fault diagnosis[J]. Mechanical Systems and Signal Processing, 2023, 195: 110314.

[101] Wang H, Li Y F, Men T, et al. Physically interpretable wavelet-guided networks with dynamic frequency decomposition for machine intelligence fault prediction[J]. IEEE Transactions on Sys-

tems, Man, and Cybernetics: Systems, 2024, 54(8): 4863-4875.

[102] Li S, Li T, Sun C, et al. WPConvNet: An interpretable wavelet packet kernel-constrained convolutional network for noise-robust fault diagnosis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(10): 14974-14988.

[103] Liu C, Qin C, Shi X, et al. TScatNet: An interpretable cross-domain intelligent diagnosis model with antinoise and few-shot learning capability[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-10.

[104] Liu C, Ma X, Han T, et al. NTScatNet: An interpretable convolutional neural network for domain generalization diagnosis across different transmission paths[J]. Measurement, 2022, 204: 112041.

[105] Li Q, Li H, Hu W, et al. Transparent operator network: a fully interpretable network incorporating learnable wavelet operator for intelligent fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2024, 20(6): 8628-8638.

[106] Wang D, Chen Y, Shen C, et al. Fully interpretable neural network for locating resonance frequency bands for machine condition monitoring[J]. Mechanical Systems and Signal Processing, 2022, 168: 108673.

[107] Shang Z, Zhao Z, Yan R. Denoising fault-aware wavelet network: A signal processing informed neural network for fault diagnosis[J]. Chinese Journal of Mechanical Engineering, 2023, 36: 9.

[108] An B, Wang S, Zhao Z, et al. Interpretable neural network via algorithm unrolling for mechanical fault diagnosis[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-11.

[109] Zhao Z, Li T, An B, et al. Model-driven deep unrolling: Towards interpretable deep learning against noise attacks for intelligent fault diagnosis[J]. ISA Transactions, 2022, 129(B): 644-662.

[110] An B, Wang S, Qin F, et al. Adversarial algorithm unrolling network for interpretable mechanical anomaly detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(5): 6007-6020.

[111] Wu Q, Ding X, Zhao L, et al. An interpretable multiplication-convolution sparse network for equipment intelligent diagnosis in antialiasing and regularization constraint[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-12.

[112] Liu H, Xu Q, Han X, et al. Attention on the key modes: Machinery fault diagnosis transformers through variational mode decomposition[J]. Knowledge-Based Systems, 2024, 289: 111479.

[113] Li Y, Zhou Z, Sun C, et al. Variational attention-based interpretable transformer network for rotary machine fault diagnosis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(5): 6180-6193.

[114] Kim M S, Yun J P, Park P. An explainable convolutional neural network for fault diagnosis in linear motion guide[J]. IEEE Transactions on Industrial Informatics, 2021, 17(6): 4036-4045.

[115] Guo C, Zhao Z, Ren J, et al. Causal explaining guided domain generalization for rotating machinery intelligent fault diagnosis[J]. Expert Systems with Applications, 2024, 243: 122806.

[116] Wu H, Huang A, Sutherland J W. Layer-wise relevance propagation for interpreting LSTM-RNN decisions in predictive maintenance[J]. The International Journal of Advanced Manufacturing Technology, 2021, 118(3-4): 963-978.

[117] Grezmak J, Zhang J, Wang P, et al. Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis[J]. IEEE Sensors Journal, 2020, 20(6): 3172-3181.

[118] Chen H Y, Lee C H. Vibration signals analysis by explainable artificial intelligence (XAI) approach: Application on bearing faults diagnosis[J]. IEEE Access, 2020, 8: 134246-134256.

[119] Liu C, Meerten Y, Declercq K, et al. Vibration-based gear continuous generating grinding fault classification and interpretation with deep convolutional neural network[J]. Journal of Manufacturing Processes, 2022, 79: 688-704.

[120] Kim I, Wook Kim S, Kim J, et al. Single domain generalizable and physically interpretable bearing fault diagnosis for unseen working conditions[J]. Expert Systems with Applications, 2024, 241: 122455.

[121] Sun H, Cao X, Wang C, et al. An interpretable anti-noise network for rolling bearing fault diagnosis based on FSWT[J]. Measurement, 2022, 190: 110698.

[122] Brito L C, Susto G A, Brito J N, et al. Fault diagnosis using eXplainable AI: A transfer learning-based approach for rotating machinery exploiting augmented synthetic data[J]. Expert Systems with Applications, 2023, 232: 120860.

[123] Mey O, Neufeld D. Explainable AI algorithms for vibration data-based fault detection: Use case-adadpted methods and critical evaluation[J]. Sensors, 2022, 22(23): 9037.

[124] Yu S, Wang M, Pang S, et al. Intelligent fault diagnosis and visual interpretability of rotating machinery based on residual neural network[J]. Measurement, 2022, 196: 111228.

[125] Chen Z, Qin W, He G, et al. Explainable deep ensemble model for bearing fault diagnosis under variable conditions[J]. IEEE Sensors Journal, 2023, 23(15): 17737-17750.

[126] Chen B, Liu T, He C, et al. Fault diagnosis for limited annotation signals and strong noise based on interpretable attention mechanism[J]. IEEE Sensors Journal, 2022, 22(12): 11865-11880.

[127] Miettinen J, Haikonen S, Koene I, et al. Comparing torsional and lateral vibration data for deep learning-based drive train gear diagnosis[J]. Mechanical Systems and Signal Processing, 2023, 203: 110710.

[128] Li J, Wang Y, Zi Y, et al. Whitening-net: A generalized network to diagnose the faults among different machines and conditions[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(10): 5845-5858.

[129] Li S, Li T, Sun C, et al. Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis[J]. Journal of Manufacturing Systems, 2023, 69: 20-30.

[130] Brito L C, Susto G A, Brito J N, et al. An explainable artificial intelligence approach for unsu-

pervised fault detection and diagnosis in rotating machinery[J]. Mechanical Systems and Signal Processing, 2022, 163: 108105.

[131] Lee J, Park S, Kim S, et al. LiteFDNet: A lightweight network for current sensor-based bearing fault diagnosis[J]. IEEE Access, 2024, 12: 100493-100505.

[132] Jang K, Pilario K E S, Lee N, et al. Explainable artificial intelligence for fault diagnosis of industrial processes[J]. IEEE Transactions on Industrial Informatics, 2025, 21(1): 4-11.

[133] Gwak M, Kim M S, Yun J P, et al. Robust and explainable fault diagnosis with power-perturbation-based decision boundary analysis of deep learning models[J]. IEEE Transactions on Industrial Informatics, 2023, 19(5): 6982-6992.

[134] Decker T, Lebacher M, Tresp V. Does your model think like an engineer? Explainable AI for bearing fault detection with deep learning[C]//ICASSP 2023 - 2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[135] Herwig N, Borghesani P. Explaining deep neural networks processing raw diagnostic signals[J]. Mechanical Systems and Signal Processing, 2023, 200: 110584.

[136] Liu S, Huang J, Ma J, et al. SRMANet: Toward an interpretable neural network with multi-attention mechanism for gearbox fault diagnosis[J]. Applied Sciences, 2022, 12(16): 8388.

[137] Xiang L, Bing H, Li X, et al. A frequency channel-attention based vision Transformer method for bearing fault identification across different working conditions[J]. Expert Systems with Applications, 2025, 262: 125686.

[138] Li X, Zhang W, Ding Q. Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism[J]. Signal Processing, 2019, 161: 136-154.

[139] Wang H, Liu Z, Peng D, et al. Understanding and learning discriminant features based on multi-attention 1DCNN for wheelset bearing fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2020, 16(9): 5735-5745.

[140] Tang J, Zheng G, Wei C, et al. Signal-transformer: A robust and interpretable method for rotating machinery intelligent fault diagnosis under variable operating conditions[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-11.

[141] Cui L, Tian X, Wei Q, et al. A self-attention based contrastive learning method for bearing fault diagnosis[J]. Expert Systems with Applications, 2024, 238(A): 121645.

[142] Wang G, Liu D, Cui L. Auto-embedding transformer for interpretable few-shot fault diagnosis of rolling bearings[J]. IEEE Transactions on Reliability, 2024, 73(2): 1270-1279.

[143] Decker T, Lebacher M, Tresp V. Explaining deep neural networks for bearing fault detection with vibration concepts[C]//2023 IEEE 21st International Conference on Industrial Informatics (INDIN). Lemgo, Germany: IEEE, 2023: 1-6.

[144] Yang H, Li X, Zhang W. Interpretability of deep convolutional neural networks on rolling bearing

fault diagnosis[J]. Measurement Science and Technology, 2022, 33(5): 055005.

[145] Borghesani P, Herwig N, Antoni J, et al. A Fourier-based explanation of 1D-CNNs for machine condition monitoring applications[J]. Mechanical Systems and Signal Processing, 2023, 205: 110865.

[146] Pang P, Tang J, Luo J, et al. An explainable and lightweight improved 1-D CNN model for vibration signals of rotating machinery[J]. IEEE Sensors Journal, 2024, 24(5): 6976-6997.

[147] Yang Y, Zhang W, Peng Z, et al. Multicomponent signal analysis based on polynomial chirplet transform[J]. IEEE Transactions on Industrial Electronics, 2013, 60(9): 3948-3956.

[148] Tu G, Dong X, Chen S, et al. Iterative nonlinear chirp mode decomposition: A Hilbert-Huang transform-like method in capturing intra-wave modulations of nonlinear responses[J]. Journal of Sound and Vibration, 2020, 485: 115571.

[149] Cohen M X. A better way to define and describe Morlet wavelets for time-frequency analysis[J]. NeuroImage, 2019, 199: 81-86.

[150] Oppenheim A V, Willsky A S, Nawab S H, et al. Signals & systems[M]. Pearson Educación, 1997.

[151] Ganguly B, Chaudhuri S, Biswas S, et al. Wavelet kernel-based convolutional neural network for localization of partial discharge sources within a power apparatus[J]. IEEE Transactions on Industrial Informatics, 2021, 17(3): 1831-1841.

[152] Andrearczyk V, Whelan P F. Using filter banks in convolutional neural networks for texture classification[J]. Pattern Recognition Letters, 2016, 84: 63-69.

[153] Smith W A, Randall R B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study[J]. Mechanical Systems and Signal Processing, 2015, 64–65: 100-131.

[154] Case School of Engineering. Apparatus & Procedures[EB/OL]. [2025-03-20]. https://engineering.case.edu/bearingdatacenter/apparatus-and-procedures.

[155] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates, Inc., 2017: 4080-4090.

[156] Li B, Tang B, Deng L, et al. Multiscale dynamic fusion prototypical cluster network for fault diagnosis of planetary gearbox under few labeled samples[J]. Computers in Industry, 2020, 123: 103331.

[157] Jiang C, Chen H, Xu Q, et al. Few-shot fault diagnosis of rotating machinery with two-branch prototypical networks[J]. Journal of Intelligent Manufacturing, 2022, 34(4): 1667-1681.

[158] Zhang X, Su Z, Hu X, et al. Semisupervised momentum prototype network for gearbox fault diagnosis under limited labeled samples[J]. IEEE Transactions on Industrial Informatics, 2022, 18(9): 6203-6213.

[159] Zhou F, Xu W, Wang C, et al. A semi-supervised federated learning fault diagnosis method based on adaptive class prototype points for data suffered by high missing rate[J]. Journal of Intelligent & Robotic Systems, 2023, 109(4): 93.

[160] Su Z, Zhang X, Wang G, et al. The semisupervised weighted centroid prototype network for fault diagnosis of wind turbine gearbox[J]. IEEE/ASME Transactions on Mechatronics, 2024, 29(2): 1567-1578.

[161] Wang H, Bai X, Tan J, et al. Deep prototypical networks based domain adaptation for fault diagnosis[J]. Journal of Intelligent Manufacturing, 2020, 33(4): 973-983.

[162] Yang J, Wang C, Wei C. A novel Brownian correlation metric prototypical network for rotating machinery fault diagnosis with few and zero shot learners[J]. Advanced Engineering Informatics, 2022, 54: 101815.

[163] Tang T, Wang J, Yang T, et al. An improved prototypical network with L2 prototype correction for few-shot cross-domain fault diagnosis[J]. Measurement, 2023, 217: 113065.

[164] Chen X, Yang R, Xue Y, et al. A novel momentum prototypical neural network to cross-domain fault diagnosis for rotating machinery subject to cold-start[J]. Neurocomputing, 2023, 555: 126656.

[165] Zhang X, Huang W, Wang R, et al. Dual prototypical contrastive network: a novel self-supervised method for cross-domain few-shot fault diagnosis[J]. Journal of Intelligent Manufacturing, 2023, 36(1): 475-490.

[166] Sun H, Yang B, Lin S. An open set diagnosis method for rolling bearing faults based on prototype and reconstructed integrated network[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-10.

[167] Long J, Chen Y, Huang H, et al. Multidomain variance-learnable prototypical network for few-shot diagnosis of novel faults[J]. Journal of Intelligent Manufacturing, 2023, 35(4): 1455-1467.

[168] Mei J, Zhu M, Liu S, et al. Cross-domain open-set fault diagnosis using prototype learning and extreme value theory[J]. Applied Acoustics, 2024, 216: 109749.

[169] Wang R, Huang W, Zhang X, et al. Federated contrastive prototype learning: An efficient collaborative fault diagnosis method with data privacy[J]. Knowledge-Based Systems, 2023, 281: 111093.

[170] Zhang X, Huang W, Ding C, et al. Cross-supervised multisource prototypical network: A novel domain adaptation method for multi-source few-shot fault diagnosis[J]. Advanced Engineering Informatics, 2024, 61: 102538.

[171] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008: 1096-1103.

[172] Zhao R, Yan R, Chen Z, et al. Deep learning and its applications to machine health monitoring [J]. Mechanical Systems and Signal Processing, 2019, 115: 213-237.

[173] Shao H, Xia M, Wan J, et al. Modified stacked autoencoder using adaptive morlet wavelet for intelligent fault diagnosis of rotating machinery[J]. IEEE/ASME Transactions on Mechatronics, 2022, 27(1): 24-33.

[174] Zhao Z, Li T, Wu J, et al. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study[J]. ISA Transactions, 2020, 107: 224-255.

[175] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017: 4768-4777.

[176] Slundberg, Connortann, Ryserrao, et al. Shap[EB/OL]. (2018-06-29) [2025-03-20]. https://github.com/shap/shap.

[177] Randall R B. Vibration-based condition monitoring: Industrial, automotive and aerospace applications[M]. John Wiley & Sons, 2021.

[178] Gardner W, Spooner C. The cumulant theory of cyclostationary time-series. I. Foundation[J]. IEEE Transactions on Signal Processing, 1994, 42(12): 3387-3408.

[179] Spooner C, Gardner W. The cumulant theory of cyclostationary time-series. II. Development and applications[J]. IEEE Transactions on Signal Processing, 1994, 42(12): 3409-3429.

[180] Ancona M, Oztireli C, Gross M. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation[C]//Proceedings of the 36th International Conference on Machine Learning: vol. 97. Long Beach, California, USA: PMLR, 2019: 272-281.

[181] Smilkov D, Thorat N, Kim B, et al. SmoothGrad: Removing noise by adding noise[EB/OL]. arXiv: 1706.03825(2017-06-12). https://arxiv.org/abs/1706.03825.

致 谢

岁月骎过，山陵浸远。五载直博光阴如白驹过隙，初入交大时的青涩懵懂却依稀如昨日。这段旅程中，有拨穗加冕的憧憬，也有路不知何方的彷徨；有灵光乍现的雀跃，也有辗转难眠的焦虑；有一气呵成的酣畅，也有屡投不中的苦涩。漫漫求学路，幸得良师益友相伴左右，在此谨向他们致以最诚挚的谢忱。

承蒙导师董兴建副教授悉心栽培。董师深谙因材施教之道，既给予学生充分的学术自由，又能在关键时刻指点迷津。基于我的研究志趣，董老师在动力学研究之外，为我提供了智能诊断这一课题，让我更好地发挥自身长处。董老师治学严谨，见解独到，每周组会上鞭辟入里的点拨，总能为我们的研究指明方向，开放的学术氛围也让热烈的讨论变为常态，成为夜里久久不愿熄灭的明灯。董老师始终强调科研要立足实际需求，从管道载荷辨识到座椅强度预测，多个工程项目让我得以将理论付诸实践。“纸上得来终觉浅，绝知此事要躬行”，这份知行合一的治学理念令我受益终身。在论文撰写与答辩准备期间，董老师更是倾注大量心血，从整体架构到字句推敲无不精益求精。师恩如山，谨此致以最深切的感激，并祝愿恩师阖家安康，万事胜意。

承蒙导师程长明副教授殷切指导。程老师既是我来到交大的引路人，也是我学术生涯的启明星。自入学伊始，程老师便给予无微不至的关怀，助我顺利适应研究生生涯。在学术上，程老师严谨治学，循循善诱，总能给我的研究带来新的启发。他特别强调团队合作的重要性，鼓励我与大课题组的同门多交流、多合作，这让我在科研道路上结识了许多志同道合的伙伴。程老师的谆谆教诲，我将永远铭记于心。

求学期间，承蒙诸位师长厚爱。感谢彭志科老师在项目支持与平台搭建方面的贡献，感谢王冬老师、何清波老师、熊玉勇老师的悉心指导，让我深刻体会到团队协作的力量。同时感谢瞿叶高老师、龙新华老师，以及振动所实验室严莉老师、塔娜老师、章振华老师等在科研实验与项目开展中的支持。谨祝各位师长工作顺遂，家和人安。

在交大的点滴成长，离不开师兄师姐的提携。涂国伟师兄聪慧洒脱，开导我走出博士入学时的迷茫；皇甫一樊师兄和陈康康师兄手把手带我做实验，为我树立榜样的力量；赵保璇师兄率先尝试智能诊断新方向，在提供参考的同时，也在求职过程对我帮助匪浅。感谢四位师兄，以及李松旭师兄、吴高阳师兄、位莎师姐、于小洛师兄、李崇师兄、孔金震师姐、周鹏师兄、魏莎师姐、代鹤师兄、吴新亚师兄、李占伟师兄，在读博期间给予我科研以及生活上的帮助与指导。

感谢 Dong Group 的成员冉启平、宋华庆、张津毓、侯宁、曹意翔、吴非晗、张孟珂、赵锐源和鲍宇风，大家一起集思广益、攻坚克难，分享科研的乐趣与挑战。感谢课题组的赵尚宇、李鑫宇、胡奎、毕志昊、张济旭、姚锦涛、刘昭宇，作为 403 工作室的饭搭子，在工作之余一同结伴畅聊。也感谢课题组的林苗苗、李天奇、黄浩、李问渠、魏存驹、李伟涛、任泽生、王胤博、罗顺安、杨纪楠、侯炳昌、陈一锴、刘洁、严彤彤、付奕楚等伙伴给予我的陪伴和鼓励。愿诸位前程似锦，各展宏图。

校园时光中，结识了许多志同道合的朋友。感谢我的室友刘昇，作为同一届的直博生，共同鼓励，为彼此的科研和生活提供支持。感谢我的健身搭子暴汗哥、夏教练、巨肩哥、羊教练，在健身房的枯燥锻炼中，分享撸铁的技巧和欢乐。也感谢我的骑车搭子曹意翔、鲍宇风和李怀瑾，一起踏足申城的每一个角落。另外感谢 SJTUG 团队在交大 LaTeX 模板上的辛勤付出，为毕业论文撰写提供便利。

感谢女友长久以来的理解与支持，让我在科研的道路上走得更加坚定。也感谢孙浩然、杨沅松和董亚北等挚友的支持与鼓励，大家在一起的时光总是充满欢声笑语，分享生活中的点滴快乐。

感谢父亲和母亲的含辛茹苦，用坚实的臂膀托起我求学的梦想。不善言辞、不够成熟的我，曾经让你们担心和失望，但你们始终给予我无条件的支持与信任。也感谢胞弟的温暖陪伴，手足情深，六年相隔未减情谊，愿我们永远守望相助。谨祝家人身体康泰，平安喜乐。

行文至此，落笔为终。五载春秋的求索，终化为涓涓墨痕，且将这段珍贵时光铭记于心，整装再赴新程。

陈　钱

二〇二五年四月

上海交通大学

攻读博士学位期间的科研成果

期刊论文

[1] **Chen Q**, Dong X J, Tu G W, Wang D, Cheng C M, Zhao B X, Peng Z K. TFN: An interpretable neural network with time-frequency transform embedded for intelligent fault diagnosis[J]. Mechanical Systems and Signal Processing, 2024, 207: 110952. (SCI, 中科院一区TOP, IF=7.9, ESI高被引, 对应第二章)

[2] **Chen Q**, Dong X J, Peng Z K. Interpreting what typical fault signals look like via prototype-matching[J]. Advanced Engineering Informatics, 2024, 62: 102849. (SCI, 中科院一区TOP, IF=8.0, 对应第三章)

[3] **陈钱**, 陈康康, 董兴建, 皇甫一樊, 彭志科, 孟光. 一种面向机械设备故障诊断的可解释卷积神经网络[J]. 机械工程学报, 2024, 60(12): 65-76. (EI)

[4] **陈钱**, 董兴建, 陈康康, 刘文博, 袁顺, 吴培桂, 倪洪斌. 基于迭代式局部加权线性回归的汽车座椅滑轨剥离强度预测[J]. 机械工程学报, 2025. (EI, 已录用)

[5] **Chen Q**, Dong X J, Hu K, Chen K K, Peng Z K, Meng G. CS-SHAP: Extending SHAP to cyclic-spectral domain for better interpretability of intelligent fault diagnosis[EB/OL]. arXiv: 2502.06424(2025-02-10). https://arxiv.org/abs/2502.06424. (预印本, 对应第四章)

[6] **Chen Q**, Dong X J, Gao S, Luo J, Peng Z K, Meng G. SHapley Estimated exPlanation (SHEP): A fast post-hoc attribution method for interpreting intelligent fault diagnosis[EB/OL]. arXiv: 2504.03773(2025-04-03). https://arxiv.org/abs/2504.03773. (预印本, 对应第五章)

[7] Dong X J, **Chen Q**, Liu W B, Wang D, Peng Z K, Meng G. A systematic framework of constructing surrogate model for slider track peeling strength prediction[J]. Science China Technological Sciences, 2024, 67(10): 3261-3274. (SCI, 中科院一区TOP, IF=4.4)

[8] Hu K, **Chen Q**, Yao J T, He Q B, Peng Z K. An interpretable deep feature aggregation framework for machinery incremental fault diagnosis[J]. Advanced Engineering Informatics, 2025, 65: 103189. (SCI, 中科院一区TOP, IF=8.0)

[9] Chen K K, Dong X J, Gao P L, **Chen Q**, Peng Z K, Meng G. Physics-informed neural networks for topological metamaterial design and mechanical applications[J]. International Journal of Mechanical Sciences, 2025, 301: 110489. (SCI, 中科院一区TOP, IF=7.1)

会议论文

[1] **陈钱**, 董兴建, 彭志科. 通过原型匹配从神经网络视角解释典型故障信号[C] // 2024 年全国设备监测诊断与维护学术会议. 大理, 2024.

[2] 皇甫一樊, **陈钱**, 董兴建, 彭志科. 考虑轴系变形的齿面损伤动态演化模型[C] // 2022 年全国设备监测诊断与维护学术会议. 太原, 2022.

发明专利

[1] 董兴建, **陈钱**, 陈康康, 袁顺, 刘文博, 吴培桂. 一种基于数据驱动的汽车座椅滑轨剥离强度预测方法:202310769536.9[P]. 2023-06-27. (发明专利)

[2] 董兴建, 皇甫一樊, 于小洛, 陈康康, **陈钱**. 一种基于原位测量的传递路径分析与齿轮故障溯源方法:202210804850.1[P]. 2022-07-08[2023-06-02]. (发明专利，已授权)

[3] 董兴建, 刘山尖, 皇甫一樊, 陈康康, **陈钱**, 曹意翔. 一种齿轮传动系统局部非线性快速分析方法:202411274641.6[P]. 2024-09-11. (发明专利)

[4] 董兴建, 吴峰崎, 皇甫一樊, 陈康康, **陈钱**, 曹意翔. 一种行星齿轮系统时变传递路径分析与故障溯源方法:202411273288.X[P]. 2024-09-11. (发明专利)

[5] 董兴建, 皇甫一樊, 于小洛, 陈康康, **陈钱**. 一种考虑齿圈柔性的行星齿轮动力学分析方法:202310120066.3[P]. 2023-02-14. (发明专利)

攻读学位期间所获荣誉与奖励

[1] 2025 年度上海交通大学优秀毕业生

[2] 2024 年度 SMC 高田奖学金（一等）

[3] 2024 年度机械与动力工程学院优秀学生党员