中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：21721974

上海大学

专业硕士学位论文

SHANGHAI UNIVERSITY
PROFESSIONAL MASTER'S DISSERTATION

题目	基于迁移学习的轴承故障诊断技术研究

作　　者　于福超
学科专业　电子信息
导　　师　修贤超
完成日期　二〇二四年四月

上海大学工学专业硕士学位论文

基于迁移学习的轴承故障诊断技术研究

作　　者: 于福超

导　　师: 修贤超

学科专业: 电子信息

机电工程与自动化学院

上海大学

2024 年 4 月

A Dissertation Submitted to Shanghai University for the
Degree of Professional Master in Engineering

Research on Bearing Fault Diagnosis Technology Based on Transfer Learning

Candidate: Fuchao Yu
Supervisor: Xianchao Xiu
Major: Electronic Information

School of Mechatronic Engineering and Automation
Shanghai University
April, 2024

摘　要

轴承故障诊断作为一种关键技术，能从数据中及时发现安全隐患，直接影响了工业的安全生产效率和成本效益。随着制造业的迅速发展，我国对工业安全的需求逐渐增加，《中国制造 2025》中明确指出要开发自主可控的高端工业系统，建立完善的工业智能运维安全体系。然而不同工业系统之间存在的巨大差异、庞大数据群中包含的冗余信息、异质数据所造成的分布漂移等问题，对故障诊断技术提出了巨大的挑战。因此，如何设计一种安全性高、适应性广、迁移性好的轴承故障诊断方法已成为亟待解决的技术难题。本论文聚焦于轴承故障诊断中不同工况下缺少标签的问题，利用迁移学习方法实现无监督跨工况诊断，主要工作如下：

（1）针对故障数据中包含大量不同种类的噪声以及子空间变换矩阵自由度较低的问题，提出了基于迁移子空间学习的故障诊断方法，建立了两种具有低秩稀疏结构的鲁棒迁移学习模型。首先，通过引入一个额外的矩阵来建模高斯噪声的方式，实现了模型的鲁棒性，建立了一种基于高斯噪声改进的迁移学习方法。进一步，通过自适应地从样本中学习标签矩阵，增强了回归算法的自由度，建立了一种基于松弛回归矩阵改进的迁移学习方法。在算法方面，利用交替方向乘子法开发了有效的优化策略来求解所提出的模型。大量的数值实验验证了所提出的方法在跨域轴承故障诊断任务中的优越性和有效性。

（2）针对现有的深度迁移模型鉴别性较低的问题，提出了深度判别联合概率适应网络的智能故障诊断方法，建立了一种鉴别性强、灵活度高的深度领域自适应网络。通过使用多核的方法来优化希尔伯特空间的核选择，提高了分布差异度量的精确度，增强了网络的适应性。进一步，将判别联合概率最大均值差异度量准则嵌入到神经网络中，提高了神经网络的可迁移性和可鉴别性。最后在多个轴承故障诊断数据集以及不同的神经网络上进行了实验，结果表明该方法不仅能有效地抑制领域偏移，而且具有更好的聚类的效果，能准确地识别出不同工况下的故障种类。

关键词：轴承故障诊断；迁移学习；稀疏表示；特征空间变换；深度领域自适应

ABSTRACT

Bearing fault diagnosis, as a key technology, can detect potential safety hazards from data promptly, which directly affects the safety productivity and cost-effectiveness of the industry. With the rapid development of the manufacturing industry, the country's demand for industrial safety is gradually increasing. "Made in China 2025" clearly states that it is necessary to develop autonomous industrial systems and establish perfect industrial safety systems. However, the great differences between different industrial systems, the redundant information contained in the huge data cluster, and the distribution drift caused by heterogeneous data all pose great challenges to fault diagnosis. Therefore, designing a bearing fault diagnosis method with high security, wide adaptability, and good transferability has become an urgent technical challenge. In this paper, we focus on the the problem of fault diagnosis under unknown operating conditions and utilize the transfer learning methods to realize cross-operating condition fault diagnosis, and the main work is summarized as follows:

(1) To overcome the problem that the fault data contains different kinds of noise and the subspace transform matrix has low flexibility, we proposed fault diagnosis methods based on transfer subspace learning and established two robust transfer learning models with low-rank sparse and relaxation regression. First, by introducing an additional matrix to model Gaussian noise for robustness, we established a transfer learning method based on Gaussian noise improvement. In addition, by adaptively learning the label matrix from the samples for flexibility, we established another transfer learning method based on relaxation regression matrix improvement. Finally, we developed an efficient algorithms using the alternating direction method of multipliers to solve the models. Extensive experiments verified the superiority and effectiveness of proposed methods.

(2) To overcome the problem of low discriminatory nature of existing deep transfer models, we proposed a deep discriminative joint probability adaptive network. The method uses a multi-kernel approach to optimize kernel selection in reproducing kernel hilbert space, which improves the accuracy of the distributional discrepancy metric and the adaptability

of networks. Further, we embedded the discriminative joint probability maximum mean difference metric into neural network to improve the transferability and discriminability. We conducted experiments on several bearing fault diagnosis datasets as well as different neural networks, the results show that our method effectively suppresses domain bias, has better clustering results, and accurately identifies the types of faults under different operating conditions compared to other deep domain adaptation methods.

Keywords: Bearing fault diagnosis; Transfer learning; Sparse representation; Feature space transform; Deep domain adaptation

目 录

摘 要 I

ABSTRACT II

第一章 绪论 1

1.1 研究背景及意义 1

1.2 国内外研究现状 3

1.2.1 智能故障诊断技术 4

1.2.2 现阶段面临的挑战 7

1.3 本文主要研究内容和结构组织 9

第二章 迁移学习预备知识 11

2.1 迁移学习简介 11

2.1.1 理论基础 11

2.1.2 发展概况 12

2.2 基于模型的深度迁移学习 15

2.2.1 预训练微调 16

2.2.2 自训练 17

2.2.3 Transformer 结构 18

2.3 基于差异的深度迁移学习 19

2.3.1 暹罗网络架构 19

2.3.2 特征差异补偿法 22

2.4 基于对抗的深度迁移学习 22

2.4.1 特征提取法 23

2.4.2 特征变换法 25

2.5 本章小结 27

第三章 基于迁移子空间学习的轴承故障诊断方法 28

3.1 引言 28

3.2 迁移子空间学习 30

3.3 基于高斯噪声改进的子空间学习 32
3.3.1 数学模型 32
3.3.2 优化算法 33
3.4 基于松弛回归矩阵改进的子空间学习 36
3.4.1 数学模型 36
3.4.2 优化算法 37
3.5 数值实验 40
3.5.1 实施细节 40
3.5.2 实验数据 41
3.5.3 实验分析 46
3.6 本章小结 53
第四章 基于深度域适应网络的轴承故障诊断方法 54
4.1 引言 54
4.2 差异度量相关工作 56
4.2.1 最大均值差异 56
4.2.2 联合最大均值差异 59
4.2.3 判别联合概率最大均值差异 60
4.3 模型建立 63
4.3.1 深度联合概率适应网络框架 64
4.3.2 基于 AlexNet 的深度联合概率适应网络 68
4.3.3 基于 ResNet50 的深度联合概率适应网络 69
4.4 数值实验 71
4.4.1 实施细节 71
4.4.2 实验数据 72
4.4.3 实验分析 73
4.5 本章小结 81
第五章 总结与展望 82
5.1 总结 82
5.2 展望 83

插图索引 84

表格索引 86

参考文献 87

攻读专业硕士学位期间取得的研究成果 97

致　谢 98

附录 A　本文使用的英文缩写 99

第一章　绪论

1.1　研究背景及意义

随着信息科学技术的迅速发展，新一轮的工业科技革命也逐渐兴起，高端制造业成为各国努力发展和积极布局的重要领域。预测和健康管理系统 (Prognostics and Health Management，PHM) 能有效评估工业系统中结构和部件的演变衰老过程，为工业安全建设提供可靠的保障，其大致流程如图1.1所示。故障诊断技术作为 PHM 中的重要一环，不仅对社会工业发展有着重要的经济意义，而且与我国的战略需求深度绑定。国务院印发的《中国制造 2025》一文中明确提出，要集中优势力量发展高端装备，推动实现航空航天、轨道交通、航海船舶、数控机床、新能源汽车、生物医学以及各类先进制造业等战略领域的重点突破[1]。

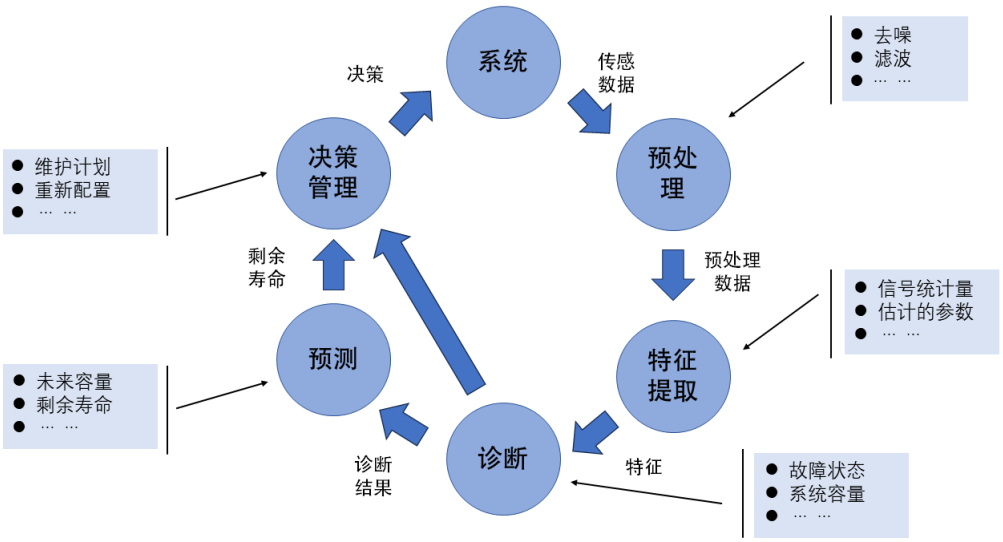


图 1.1　PHM 系统流程

Figure 1.1　PHM system flow

旋转机械的轴承部件能够有效减少机械设备运行时的摩擦阻力，维持设备的正常运转，是上述各个领域中高端机械运转的核心部件，也是故障频发部件[2]。滚动轴承的发展水平一定程度上能反映出国家机械工业的发展水平，其良好的运行状态能有效减少工业事故的发生。据相关数据统计，超过 30% 的机械设备故障都是由于轴承发生故障造成的[3]。在设备运行时，滚动轴承长期处于高温、高转速、高负载

的运行状况，容易造成轴承受腐蚀磨损、高速摩擦过热、冲击负荷较大脱落等情况的发生，而复杂多变的运行环境加剧了这一情况的发生。轴承的损坏将直接影响设备的运行状态，导致设备运行效率下降造成直接的经济损失，更严重的会导致设备失控引发重大安全事故，造成人员伤亡的情况发生。2008 年，惠州抽水蓄能电站发电机组其中一台发电机在进行负荷调试时，由于振动过大导致轴承脱轴，机组轴系完全损毁；2010 年，日本海南电厂发电机组一台 600MW 超临界火力发电机组在进行超速试验时，由于轴承失效导致机组发生强烈振动并最终损毁，造成了高达 50 亿日元的经济损失；2015 年，加拿大一化工厂的一台离心机在运行过程中，由于轴承过热导致润滑失效，最终引发离心机解体爆炸，事故造成化工厂停产；2021 年，美国一煤矿由于输送带轴承故障导致整个输送系统瘫痪，严重影响了煤矿的正常生产，事故原因是轴承润滑不足，长时间摩擦导致过热损坏。这些事故不断警醒着人们工业安全生产与设备安全运行的重要性。

图 1.2　工业安全事故

Figure 1.2　Industrial safety accidents

图1.2展示了由于设备轴承故障引发的重大安全事故，包括化工厂监测机械轴承转子损坏引发的爆炸、飞机起落架轴承故障引发的飞机失事以及运载火箭因轴承配件故障导致的发射失败。这类事故一旦发生，就一定会造成大量的经济损失，且事后运维的成本远远高于故障诊断的成本。如果能做到准确地定位到轴承的故障点，无

误地判别出轴承的故障类型，并由工作人员及时对损坏轴承进行更换和维护，那么就能大大减少轴承故障的发生，延长设备的使用寿命，有效地避免由于轴承损坏带来的直接经济损失和伤亡事故。

针对上述问题，本文聚焦基于迁移学习的轴承故障诊断技术，力求开发实现一种对跨工况问题具有良好鲁棒性的故障检测方法，应对工况改变且新工况缺少标签的场景，能够及时发掘轴承的故障类型，减少经济损失和安全事故，提高旋转机械运行的可靠性。

1.2 国内外研究现状

故障诊断技术涉及了传感器技术、信号处理、计算机视觉、数据分析、机器学习等各大关键领域[4-6]，引起了学者们的广泛关注。目前对于故障诊断技术的研究大致可分为三类：基于模型的方法、基于信号处理的方法以及基于机器学习的方法，其中后两类方法也称为数据驱动的方法，如图1.3所示。下面对这三类方法的国内外研究现状以及目前存在的挑战进行介绍。

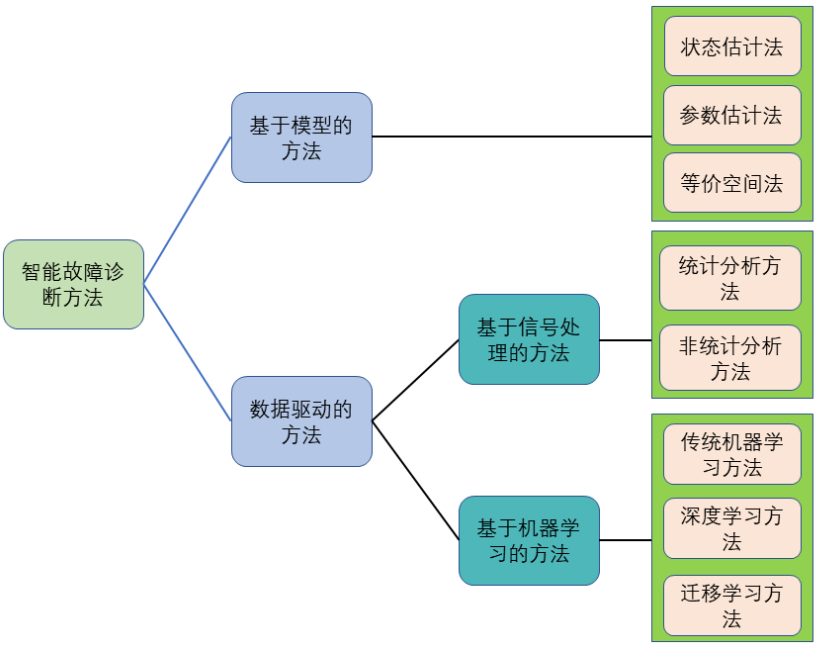


图 1.3 故障诊断技术分类

Figure 1.3 Classification of intelligent fault diagnosis

1.2.1 智能故障诊断技术

(1) 基于模型的方法

基于模型的故障诊断技术最早是在 1971 年由 Beard 等人[7]提出的，该方法利用精确的数学模型对可观测的输入输出量进行建模来反映系统实际行为与期望行为之间的不一致进行故障诊断。基于模型的故障诊断方法在工业系统应用中取得了一系列的研究成果，Isermann 等人[8]全面总结了基于模型的故障诊断方法近 30 年的研究，并将其分为参数估计法、状态估计法以及等价估计法。

参数估计法是根据参数估计值与正常值之间的偏差来判断故障情况。在已知控制模型的基础上，参数的值很容易由闭环控制系统的输入与输出得到，因此学者们对模型参数与物理参数的对应关系进行了大量的研究。Patton 等人[9]第一次提出了参数特征结构赋值法，将观察增益参数表示为特征值和特征向量的函数。这类方法在燃气轮机、航天器和风力轮机系统中得到了广泛的应用。此外，对于参数选择的多目标优化问题可以使用线性矩阵不等式进行重新表述，这类方法在 Takagi–Sugeno 模糊非线性系统、开关系统以及时间延迟系统等各种动态系统中都有广泛的应用[10]。

状态估计法是通过某种类型的观测器来估计系统的状态，对输出误差进行适当的加权来构造残差从而检测系统故障。由于非线性系统的状态观测器难以设计，因此这类方法大多应用于线性系统。例如，利用卡尔曼滤波器[11]以及高增益自适应观测器[12]来进行线性系统的状态估计。对于非线性系统来说状态观测器的设计一直是一项具有挑战性的工作，目前得到广泛使用的有滑模观测器[13]、基于 Hamilton-Jacobi-Bellman 的非线性观测器[14]。此外，Ka 等人[15]利用微分几何法扩展线性系统中不可观测子空间概念，将原系统的非线性坐标进行变换实现非线性观测器的设计。

等价空间法是通过比对系统输入输出的测量值来检验模型的一致性从而检测系统故障。等价空间法具有有限脉冲响应的结构并且其推导出的残差与初始状态完全解耦[16]，在某种特定情况下可以等价地转换为状态估计法[17]。因此 Zhong 等人[18]将等价空间方法扩展到确定性系统上，此后 Jiang 等人[19]进一步研究了将等价空间方法扩展到具有适当修改的随机系统上。需要注意的是，等价空间法不同于参数估计法和状态估计法，该方法只适用于线性系统。

然而，这类方法过于依赖观测对象精准的数学模型，而实际问题中被诊断对象的数学模型往往难以建立，因此下面的两类方法逐渐引起了人们的关注。

(2) 基于信号处理的方法

基于信号处理的故障诊断技术不再受限于某一特定的数学模型，而是根据从传感器中收集的信号来反映轴承健康状态。该类方法通常包括以下四个步骤：数据采集、特征提取、特征选择和特征分类。在基于信号处理的方法中，为了获得高精度的识别结果，往往会将原始的特征信号进行处理。例如，从旋转机械收集的原始轴承故障信号是时域信号，通过使用适当的工具将轴承故障信号转换到相应的领域，使其可以在频域和时频域中研究轴承故障信号。基于信号处理的故障诊断技术大致可分为两类：统计分析方法[20]和非统计方法[21]。

统计分析方法是利用统计学工具对信号进行分析，该方法对快速故障检测非常有效。由于故障信号形成的特征矩阵结构复杂，特征相关性和冗余度较高，因此统计学领域的相关方法为解决这类问题提供了很好的方案。Wen 等人[22]采用主成分分析的方法 (Principal Component Analysis，PCA) 降低数据维度，减少数据之间的冗余，以此提高故障诊断的准确性。Cong 等人[23]从矩阵的奇异值出发，提出了矩阵奇异值分解的方法构建了滚动轴承智能检测策略。此外，Yu 等人[24]针对条件属性节点冗余和通用图推理能力差导致计算量大、诊断精度低的问题，提出了一种利用流程图和非朴素贝叶斯推理的滚动轴承故障严重程度识别方法。这类方法能准确地提供过程信息，有效地解释复杂系统的行为，并能处理大量相关数据因而在工业故障诊断中得到广泛应用。

非统计方法是利用常见的信号处理方法对原始信号处理后进行诊断。由于工业现场环境比较复杂,实际所测得的振动信号几乎都是非平稳多分量信号，因此通常利用谱分析法和相关函数小波变换 (Wavelet Transform，WT) 等模型对信号进行分析，判断其是否发生故障。傅里叶变换 (Fourier Transformation，FT) 与 WT 是两类常用的时频域变换方法，大量学者基于这两类方法进行了研究。Zhu 等人[25]提出了基于频域窗函数的短时 FT 方法，通过频域窗函数实现二维时频平面中时间的精准定位，减少了振动信号中无关成分的干扰。相较于 FT，WT 具有可调节的窗口，能够提供更好的时间和频率位置。根据 Li 等人[26]在综述中进行的调查，WT 已被广泛用于提取快速、非平稳变化的特征以及诊断具有弱复合故障的应用中。这类非统计方法能够准确提取故障信号中的有用特征，在故障诊断中得到了广泛的应用。

(3) 基于机器学习的方法

随着大数据时代的到来，待处理的故障信号数量级呈指数倍增加，基于机器学习的故障诊断技术近年来引起了人们的大量关注。机器学习是一门多领域交叉学科，专门研究计算机怎样模拟或实验人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自己的性能。故障诊断技术发展至今，已经提出了较多的方法，从开始的基于解析模型方法到现在的基于机器学习的方法，在不需要太多的先验知识以及系统精确解析模型的情况下完成系统的故障诊断，机器学习拥有很广泛的应用空间。根据目前的研究，现有的基于机器学习的方法大致可以分为三类：传统的机器学习方法、深度学习方法和迁移学习方法[27]。我们在图1.4中给出了详细的分类，其中绿色部分为本文的研究方向。

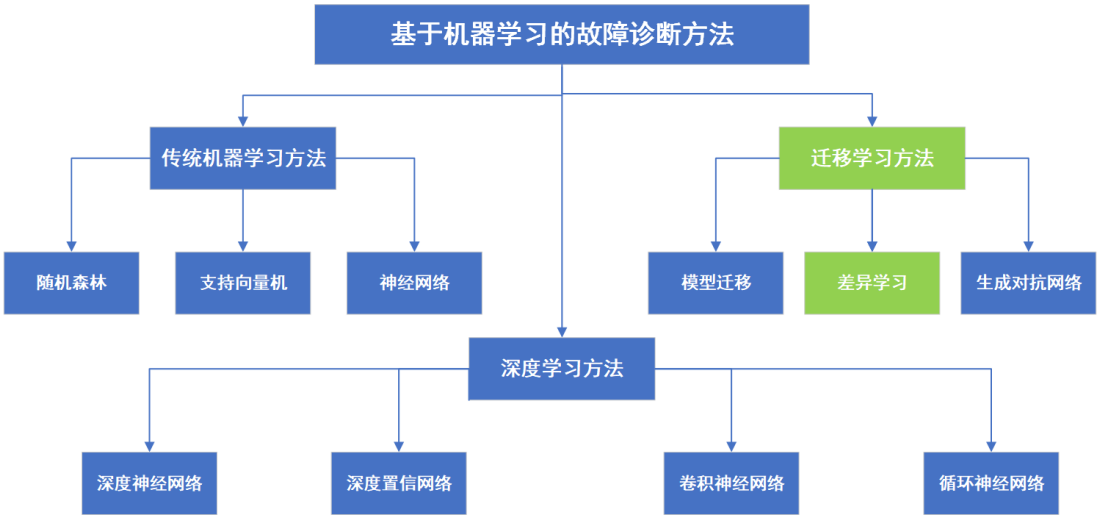


图 1.4　基于机器学习的故障诊断技术分类

Figure 1.4　Classification of machine learning based fault diagnosis

传统的机器学习方法与基于信号的方法类似，都需要对采集后的数据进行特征提取，不同的是经过特征提取后可以直接通过机器学习的方法对故障进行分类。故障的分类是利用机器学习模型在特征和故障之间建立映射关系实现的，目前常用的几个模型包括：决策树 (Decision Tree，DT)[28]、支持向量机 (Support Vector Machines，SVM)[29]、人工神经网络 (Artificial Neural Networks，ANN)[30]以及隐式马尔科夫模型 (Hidden Markov Model，HMM)[31]等方法。该类方法较为依赖特征选择和提取模型的准确性，信号中不相关的特征会对分类精度造成较大影响。

深度学习方法直接建立故障信号与故障类别直接的映射，可以在没有先验知识

的情况下自动地从输入数据中学习特征并输出诊断类别。这类方法不再依赖手动的特征选择和提取，将特征提取的工作交给卷积层来实现，并通过池化层和全连接层剔除杂余数据，减少了特征提取模型不准确带来的数据冗余的影响，提供了一种端到端的故障诊断方法。这类方法随着深度学习的研究而兴起，在近十年得到了广泛的关注和研究。具有代表性的基于深度学习的故障诊断技术包括：卷积神经网络 (Convolutional Neural Networks，CNN)[32]、循环神经网络 (Recursive Neural Network，RNN)[33]以及深度信念网络 (Deep Belief Network，DBN)[34]等。但是这类方法的前提是训练数据和测试数据服从相同的特征分布，且必须包含大量标记的数据来对网络进行训练，然而实际场景中工作条件会因任务不同有较大差距，这为深度学习方法提出了挑战。

基于迁移学习的方法利用训练的数据和任务对诊断模型进行预训练，然后将学习到的知识重复用于实际的故障诊断任务。这类方法的动机在于模仿人类可以基于历史经验与知识对新问题进行适应与解决的能力。如图1.5所示，不同于传统机器学习方法针对不同的域进行独立建模与决策，迁移学习针对源域中存在的历史知识进行挖掘与迁移，从而增强在目标域上的诊断性能。这类方法能够在训练过程中减少训练数据和目标数据之间的差异，将模型或参数复用减少重新训练的成本。在这方面有许多学者对迁移学习在故障诊断中的研究进行了全面的回顾。例如，Li 等人[35]回顾了近年来迁移学习在故障诊断中的应用，并将现有的研究分为迁移成分分析、联合分布适应、深度适应网络和对抗性领域适应。Yan 等人[36]对基于知识的机械故障诊断进行了综述，并根据具体问题将所有方法分为四类：不同工作条件之间迁移、不同位置之间的迁移、机器之间的迁移和故障类型之间的迁移。Chen 等人[37]重点研究了轴承故障诊断上的无监督深度迁移学习，搭建了一个统一的深度学习框架来衡量各类深度迁移学习方法在轴承和齿轮箱数据集上的有效性。

1.2.2 现阶段面临的挑战

虽然故障诊断方法经过了多年的发展，但是目前仍处于起步阶段。这是因为工业应用大多方法 (包括机器学习、深度学习、多元统计等方法) 的有效性存在几个基本前提：历史数据同源且充足，训练与测试数据满足独立同分布条件，先验机理完备且故障知识包含在预设假设中，模型的表达能力较强。当前工业故障诊断不仅存在建模精度低和通用性差的难题，而且在一些场景下工业故障样本的获取是困难的。针对轴承机械故障诊断和健康管理技术的研究，目前存在的主要问题和难点在于以

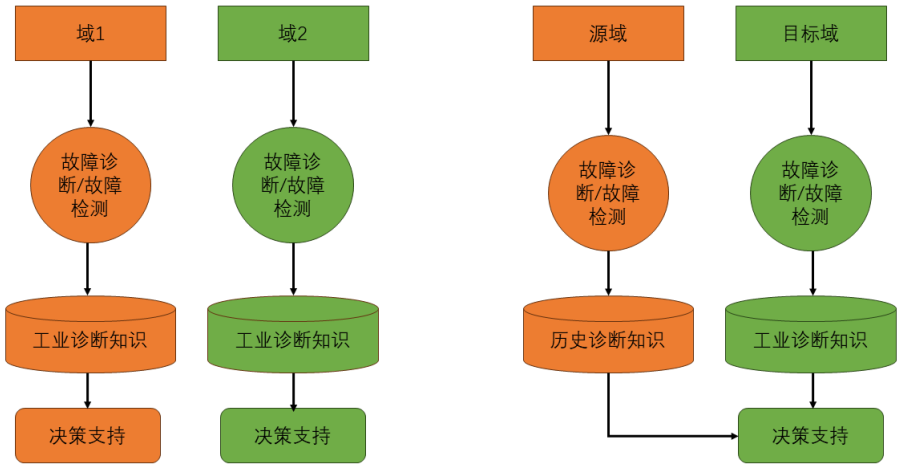


图 1.5 传统机器学习与迁移学习的工业监控流程对比

Figure 1.5 Traditional and transfer learning for industrial monitoring processes

下几个方面。

（1）数据体量不够。智能运维与健康管理的基础是对设备结构和数据进行分析，这里的数据包括各种传感器采集的状态监测信息。很多工业设备在进行预测性维护时都会面临一个共性的问题，自身的传感器数量不够或传感器精度较低，这就造成了数据体量不足，无法形成有效的长期积累。此外，采集的数据可能来源不同，不一定能够成为有效数据，反映设备的故障特征。

（2）标签数据不完整。除了数据的体量，还要考虑数据标签的完整性。以高铁轴承故障诊断为例，轴承是关系到高铁安全运行的一个关键部件。在对轴承进行故障建模时，其中重点和难点之一在于进站出站时，准确地诊断轴承故障。然而设备的故障样本往往很少，因为一旦出现故障，企业一般不会允许其持续运转。因此，出于高可靠性的考虑，往往在设备刚出现故障，或者出现故障前对其进行维护或者替换，造成故障标签数据的遗失。

（3）数据质量难以保证。数据质量是工业故障诊断能够完整分析的重要基础。在进行故障诊断时可通过分析设备的振动信号发现设备是否存在异常，如果错误地选择采样频率，则可能由于信号分辨率过低造成数据无法使用。此外，数据采样过程中，由于传感器或工作环境的干扰，数据中难免会包含各种噪声。这些噪声种类复杂、数据量较大，会降低数据质量，进而影响对数据故障类型的分析。如何剔除数据中噪声的干扰，保证数据质量也是故障诊断方法面临的重要挑战。

1.3 本文主要研究内容和结构组织

本文针对轴承故障诊断应用中目标域缺少标签或没有标签场景下的跨工况诊断问题，提出了基于迁移学习的故障诊断方法。本文共由五个章节组成，各章研究内容总体框架展示于图1.6中。

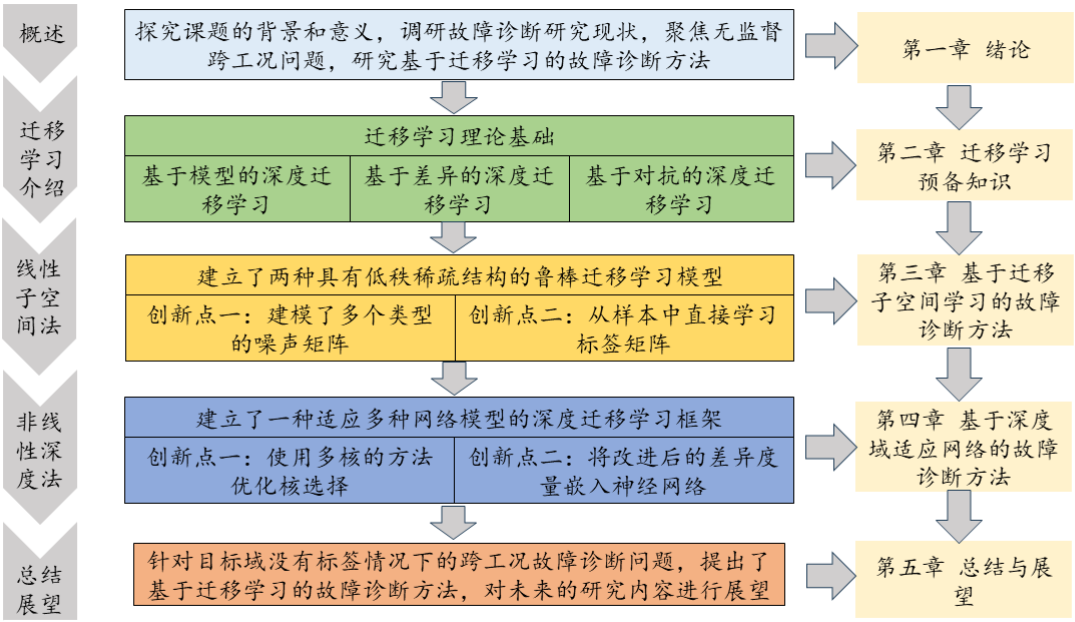


图 1.6　文章主要内容与章节安排

Figure 1.6　Main content of the article and chapter organization

本文的研究内容和结构组织安排如下：

第一章为绪论，阐述了本文的研究背景、研究目的和意义以及目前国内外故障诊断技术的研究现状。

第二章充分地介绍了迁移学习的理论基础，给出了迁移学习的概念，回顾了迁移学习技术的发展概况，并对深度迁移学习技术进行系统性地总结和概括。根据实现迁移的方法不同，将深度迁移学习技术总结为三大类：基于模型的深度迁移学习、基于差异的深度迁移学习以及基于对抗的深度迁移学习。对每类方法的代表性研究进行了详尽的介绍并给出了相应的数学模型，进一步根据模型、函数和操作对象的不同将每个类别中的方法划分为不同的子类。

第三章对迁移子空间学习在故障诊断中的应用进行了研究。首先针对采集的故

障数据中存在大量噪声的问题，在迁移子空间学习模型的基础上引入一个额外的矩阵建模了高斯噪声，有效减少了数据中高斯噪声对模型精度的影响。其次，针对子空间变换矩阵自由度较低的问题，通过直接从样本中学习标签矩阵的方式改进了回归算法，增强了模型的鲁棒性。在西储大学故障诊断的数据集以及江南大学轴承数据集上进行了大量的数值实验，验证了所提出方法的有效性。

第四章对深度迁移学习在故障诊断中的应用进行了研究。针对现有的深度迁移模型鉴别性较低的问题，利用判别联合概率最大均值差异度量代替了最大均值差异度量，提出了基于深度联合概率网络的智能故障诊断方法。这类方法通过最小化类内距最大化类间距提高了模型的可鉴别性，为分布差异度量提供了准确的理论基础。此外，在域适应的过程中引入多核的方法，优化了再生核希尔伯特空间的核选择。针对不同的神经网络上进行了实验，实验结果表明所提出模型的优越性。

第五章是总结与展望部分。总结了本文主要的的研究成果，并在此工作的基础上，对不足之处做了分析并对未来的研究方向进行了展望。

第二章　迁移学习预备知识

本章给出了迁移学习的预备知识和相关概念，回顾了迁移学习的发展概况。重点介绍了深度迁移学习近十年的发展概况，并根据模型、函数和操作对象的不同将其总结为基于模型的深度迁移学习、基于差异的深度迁移学习以及基于对抗的深度迁移学习。在此基础上，对其中的代表性工作进行了详细的介绍，给出了它们的神经网络架构和数学模型，阐述了不同方法的核心思想，总结了各类方法的优缺点，为具体应用提供了依据。

2.1　迁移学习简介

2.1.1　理论基础

根据维基百科的定义，迁移学习是运用已有的知识对不同但相关领域问题进行求解的一种机器学习方法[38]，它能够有效地利用目标域和源域中数据分布的相似性进行机器学习。迁移学习与我们的生活息息相关，例如从骑自行车到骑摩托车，从下五子棋到下围棋。这种看起来相似的事情我们都可以将“经验”重新利用，把“知识”进行迁移来减少学习成本。

在后续的章节中，我们用大写粗体字母表示矩阵，小写粗体字母表示向量，用小写字母表示标量。对于矩阵 $\mathbf{X} \in \mathbb{R}^{m \times n}$，$\mathbf{x}_i$ 为其第 i 行，x_{ij} 则表示其第 i 行第 j 列的元素。$\mathbf{X}$ 的 ℓ_1 范数和 Frobenius 函数分别写为 $\|\mathbf{X}\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n}|x_{ij}|$ 和 $\|\mathbf{X}\|_{\mathrm{F}}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}^2$。此外，令 $\sigma_i(\mathbf{X})$ 为 $\mathbf{X}$ 的第 i 个奇异值，可以得到 $\mathbf{X}$ 的核范数为 $\|\mathbf{X}\|_* = \sum_i \sigma_i(\mathbf{X})$。对于两个大小相同的矩阵 $\mathbf{X}, \mathbf{Y}$，$\langle \mathbf{X}, \mathbf{Y} \rangle$ 表示为内积。$\mathbf{X}^\top$ 代表矩阵 $\mathbf{X}$ 的转置，$\circ$ 表示矩阵进行 Hadamad 乘积运算。将 $P(\mathbf{Y}|\mathbf{X})$ 定义为实例集合 $\mathbf{X}$ 的条件概率分布，其中 $\mathbf{Y} = \{\mathbf{y}|\mathbf{y}_i \in \mathcal{Y}, i = 1, 2, ..., n\}$ 是 $\mathbf{X}$ 对应的标签集合。

下面将通过数学语言具象化地描述迁移学习。域可以表示为 $\mathcal{D} = \{\mathcal{X}, P(\mathbf{X})\}$，由特征空间 $\mathcal{X}$ 和边缘概率分布 $P(\mathbf{X})$ 组成，其中 $\mathbf{X} = \{\mathbf{x} \mid \mathbf{x}_i \in \mathcal{X}, i = 1, 2, ..., n\}$ 是样本矩阵 $\mathbf{x}_i$ 的实例集合。对于特定域 $\mathcal{D}$，$\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$ 为机器学习任务，$\mathcal{Y}$ 表示标签空间，$f(\cdot)$ 是一个隐式的决策函数能够从样本中进行学习。此时给定一个源域 $\mathcal{D}_{\mathrm{S}}$ 和源域学习任务 $\mathcal{T}_{\mathrm{S}}$，一个目标域 $\mathcal{D}_{\mathrm{T}}$ 和目标域学习任务 $\mathcal{T}_{\mathrm{T}}$，注意 $\mathcal{D}_{\mathrm{S}} \neq \mathcal{D}_{\mathrm{T}}$ 并且 $\mathcal{T}_{\mathrm{S}} \neq \mathcal{T}_{\mathrm{T}}$。

此外，在实际应用中，通常源域的样本是被完全标记的，而目标域样本缺少标签或没有标签。迁移学习旨在利用 $\mathcal{D}_S$ 和 $\mathcal{T}_S$ 中的知识提高决策函数 $f(\cdot)$ 在 $\mathcal{T}_T$ 中的性能。

图2.1直观地展示了迁移学习的学习过程，可以看出机器学习模型利用源域的数据进行训练来完成分类任务，在这个过程中利用一些方法减少源域和目标域数据特征的差异，或者直接将模型中训练好的参数 (知识) 迁移到新模型中来减少目标域数据的训练成本来完成新的分类任务。

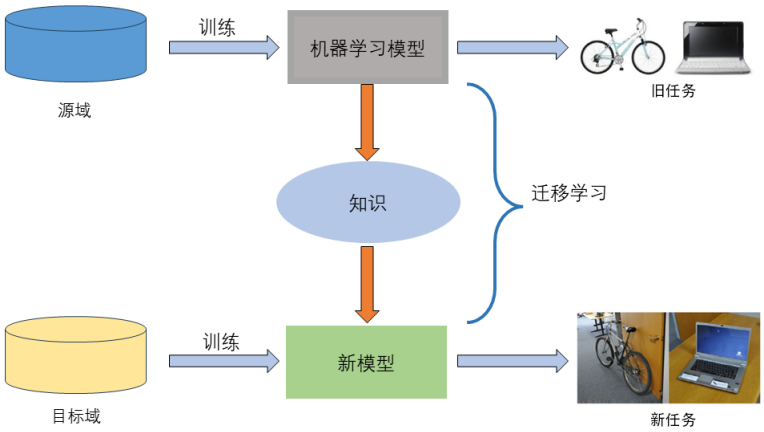


图 2.1 迁移学习的学习过程

Figure 2.1 Process of transfer learning

2.1.2 发展概况

自 1995 年以来，迁移学习的研究以不同的名称引起了越来越多的关注：终身学习、知识迁移、归纳迁移、多任务学习、基于知识的归纳偏差以及元学习等[39]。2009 年，Pan 等人[40]对迁移学习进行了全面的回顾和总结，并最终根据方法的不同将迁移学习分为四大类：基于实例的迁移学习、基于模型的迁移学习、基于特征的迁移学习以及基于实例的迁移学习，下面对四种方法进行简单介绍。

基于实例的方法通过为不同的实例分配不同的权重来完成迁移。这种方法首先会度量有标签的训练样本与无标签的测试样本之间的相似度，若训练样本和测试样本相似度大，则增加其权重。Yao 等人[41]将 Boosting 学习算法扩展到迁移学习中，通过不断的迭代改变不同样本被采样的权重。这种方法利用 Boosting 技术去除源域中与目标域数据最不像的样本数据，可以建立自动调整权重的机制来增加重要的源域样本数据权重。此外，Huang 等人[42]提出了一种核均值匹配的方法，能够利用再生

核希尔伯特空间匹配源域和目标域实例之间的均值来完成迁移。

基于模型的方法通过构建具有共享参数的模型来完成迁移。这种方法利用源域中训练的模型对目标域数据进行正则化与微调，大致分为两类：基于共享模型组件的知识迁移和基于 SVM 正则化的知识迁移。第一类方法通过对源域模型或者超参数的重新利用来学习目标域模型[43]，第二类方法则是与 SVM 相结合，通过正则化项约束模型共享的超参数来防止过拟合。例如，Aytar 等人[44]利用其他领域训练的图像检测器作为正则化项，提出了形变自适应 SVM 模型，使用尽量少的样本完成目标类别的学习。

基于特征的方法通过变换不同域的特征减少源域和目标域的差异来完成迁移。一类代表性的方法是统计特征变换法[45-47]，它是通过统计学的方式最小化源域和目标域的分布差异。例如迁移成分分析 (Transfer Component Analysis，TCA)[45]方法通过最大均值差异度量准则，减少特征空间的分布差异从而实现跨域特征的良好表示。在此基础上，Long 等人[46]通过联合边缘分布与条件分布，提出了联合分布自适应 (Joint Distribution Adaptation，JDA) 方法来减小域差异。之后，Long 等人[47]和 Wang 等人[48]分别提出迁移联合匹配 (Transfer Joint Matching，TJM) 和平衡分布自适应 (Balance Distribution Adaptation，BDA) 方法。另一种值得注意的方法是几何特征变换方法。它是通过将源域数据和目标域数据进行特征变换，隐式地对齐二者特征空间，这其中的代表方法有测地线流形核 (Geodesic Flow Kernel，GFK) 与子空间对齐法 (Subspace Alignment，SA)。GFK 是在各类流形空间中将源域和目标域进行特征变换，并利用测地线距离不断缩小二者特征分布的差异，最后利用机器学习的方法实现分类[49]。SA 通过特征空间变换来找到一个合适子空间，在这个子空间中源域和目标域的特征分布差异减少从而实现迁移[50]。

基于关系的方法通过构建源域和目标域的逻辑映射关系来完成迁移。假设源域和目标域之间的逻辑关系具有共同的模式，那么在源域中学习到的逻辑关系或规则可以转移到目标域。基于关系的迁移学习研究中的一个重要工具是马尔可夫逻辑网络 (Markov Logic Network，MLN)。MLN 是一种逻辑概率的混合模型，为表示结构关系提供了理想工具。代表工作有 Mihalkova 等人[51]提出的用于迁移学习的一阶 MLN，它的思想是首先从源域中发现逻辑公式作为关系规则，基于此规则将具有来自目标域的逻辑公式作为为候选，然后对这些候选进行筛选、修订和重新加权完成对目标域建模。我们在表2.1中对上述四种迁移学习方法进行了总结。

表 2.1 传统迁移学习方法的总结

Table 2.1 Summary of traditional transfer learning

方法	简介
基于实例的方法	为不同样本分配不同的权重实现源域和目标域的样本迁移[41-42]
基于模型的方法	将源域模型的参数共享给目标域模型并进行微调实现迁移[43-44]
基于特征的方法	特征变换使源域和目标域的特征分布相同或相似实现迁移[45,50]
基于关系的方法	通过现有的逻辑关系网络构建源域和目标域关系实现迁移[51]

上述提到的传统迁移学习方法需要准确的特征提取器对源域和目标域的特征进行提取，然后利用迁移学习方法来减少源域和目标域特征的差异，这种方法依赖特征提取器的设计。随着深度学习的发展，学者们开始将深度学习技术与迁移学习相结合，从而共享深度学习和迁移学习的优势。在 2012 年，CNN 被提出后，大量的深度迁移学习框架被构建。图2.2按时间顺序给出了这些方法发展的时间线。

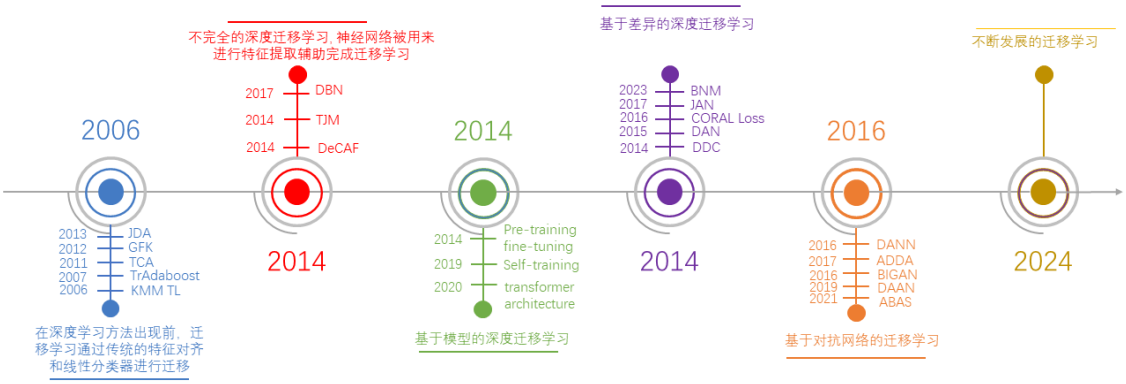


图 2.2 迁移学习发展的时间线

Figure 2.2 Timeline of transfer learning development

深度迁移学习方法的核心思想是将迁移学习方法嵌入到神经网络模型中，在网络的训练过程中减少网络对源域和目标域差异的敏感性，从而实现端到端的深度迁移学习模式。该方法有效消除了传统迁移学习对特征提取器的依赖，使迁移学习算法进入大数据、大模型时代。根据使用场景不同，深度迁移学习分为以下三类：基于模型的深度迁移学习，基于差异的深度迁移学习，基于生成对抗网络的深度迁移学习，我们在图2.3中给出了更为详细的分类。

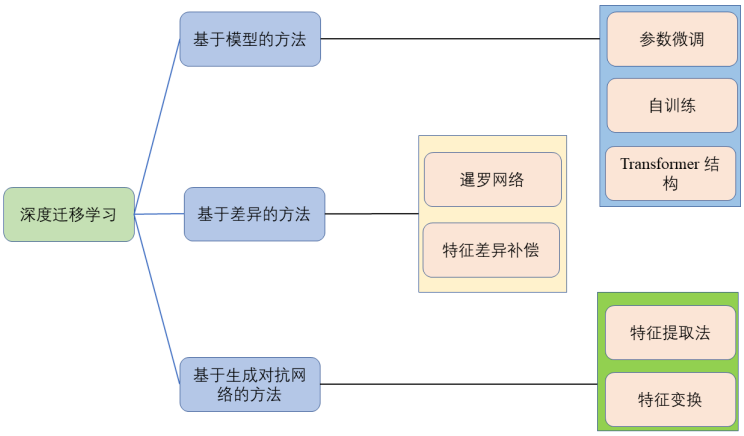


图 2.3　深度迁移学习方法分类

Figure 2.3　Classification of deep transfer learning

2.2　基于模型的深度迁移学习

基于模型的深度迁移学习方法可以总结为三类。一是预训练微调的方法，这类方法将源域网络的参数固定并根据目标域数据进一步训练微调，从而在目标域中获得良好的性能。第二种是自训练方法，这种方法克服了上一种方法在数据增强和注释增加情况下的局限性。第三种是基于 Transformer 的体系结构，它引入了图像识别领域中的注意机制。我们在表2.2中简要总结了基于模型的深度学习方法。

表 2.2　基于模型的深度迁移学习方法总结

Table 2.2　Summary of model based deep transfer learning

基于模型的方法	代表性工作	简介
预训练微调	Yosinski 等人[52] Chopra 等人[53] Rozantsev 等人[54]	共用在源域中预训练的神经网络不同层的参数
自训练	He 等人[55] Xie 等人[56] Zoph 等人[57]	使用预测的伪标签和噪声来提高模型的性能
Transformer 结构	Chen 等人[58] Bao 等人[59] Yang 等人[60]	共享和微调 Transformer 的参数

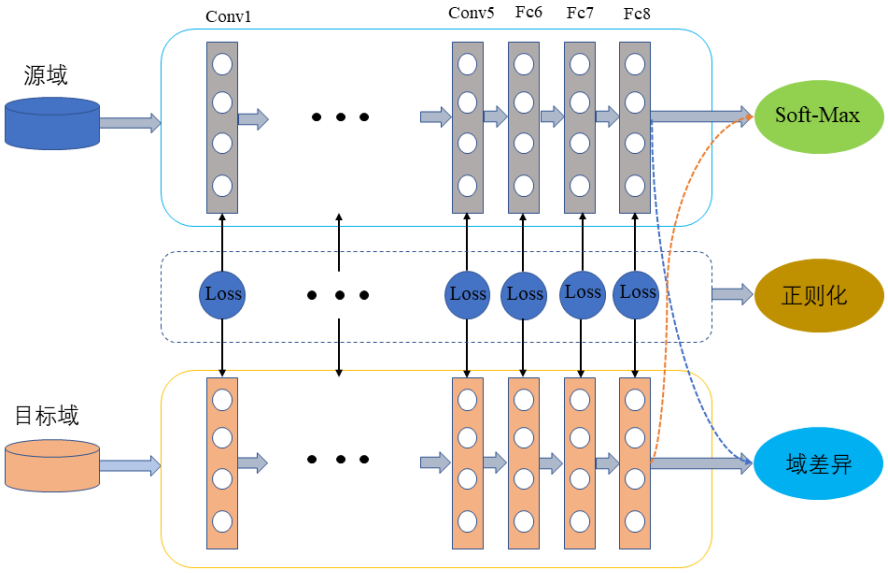


图 2.4 基于双流网络的参数微调网络

Figure 2.4 Parameter fine-tuning network architecture based on two-stream networks

2.2.1 预训练微调

预训练微调法是基于模型的深度迁移学习的最早尝试之一，可以追溯到由 Chopra 等人[53]提出的插值路径方法，该方法通过在域间的插值实现域适应的深度学习。随后，Yosinski 等人[52]第一次系统性地研究了在神经网络中进行迁移的可行性，他们对来自两个不同域的数据集用 AlexNet 网络进行训练，通过固定不同层的参数来探究卷积层在模型迁移过程中的作用。这项研究表明：卷积网络的前三层有利于传输通用特征，固定前三层的参数并微调后续层的参数可以克服数据的可变性，提高网络性能，减少训练时间和成本。尽管这种方法得到了广泛的应用，但是对于固定和微调层的确定并没有明确的依据。2018 年，Rozantsev 等人[54]提出了一种提出了一种选择性地共享和限制不同层参数的深度域自适应方法。该方法在双流结构中引入了最大均值差异 (Maximum Mean Difference，MMD) 损失函数，利用 MMD 衡量神经网络相同层的损失以确定参数共享层，对其余层进行权重正则化，其中心思想如图2.4所示。这里 Conv1-Conv5 表示卷积层，Fc6-Fc8 表示网络的全连接层。该方法可以对神经网络所有层的输出进行度量，根据 MMD 损失的不同来确定哪些层的参数共享，哪些层的参数随机初始化，这为预训练微调方法中参数共享层的确定提供了理论依据。

2.2.2 自训练

虽然参数微调的方法在一些应用中取得了较大的成功，但对于目标域缺少标签或没有标签的弱监督域适应问题有一定的局限性。He 等人[55]在跨数据集执行目标检测和语义分割任务时首次发现了微调模型的局限性。研究发现，ImageNet 数据集上的预训练模型在 COCO 数据集上的表现不如随机初始化参数方法，预训练的模型可以在早期阶段加速收敛，但不能提高最终任务的准确性。Xie 等人[56]针对弱监督领域适应问题提出了一种自训练模型，该方法使用源领域的少量标签来实现跨数据集的迁移学习，下面将对这种方法的训练过程进行介绍。给定一个网络 θ (θ 表示网络参数)，对于有标记的图像数据 $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ 和未标记的图像数据 $\{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_m\}$，进行如下的训练：

(1) 利用已标记的数据集训练一个教师网络 θ 来使其交叉熵损失 $L(\cdot)$ 最小，即

$$\frac{1}{n}\sum_{i=1}^{n} L\left(\mathbf{y}_i, f\left(\mathbf{x}_i, \theta\right)\right), \tag{2.1}$$

其中 $f(\cdot)$ 是能够进行标签预测的决策函数。

(2) 利用教师网络 θ 对未标记图像进行预测，并将预测结果 $\tilde{\mathbf{y}}_i$ 作为伪标签，即

$$\tilde{\mathbf{y}}_i = f\left(\tilde{\mathbf{x}}_i, \theta\right), \forall i = 1, 2, \ldots, m. \tag{2.2}$$

(3) 在有标记的数据和有伪标签的数据上训练一个学生网络 η 来使其交叉熵损失 $L(\cdot)$ 最小，即

$$\frac{1}{n}\sum_{i=1}^{n} L\left(\mathbf{y}^i, f_{\text{noised}}\left(\mathbf{x}^i, \eta\right)\right) + \frac{1}{m}\sum_{i=1}^{m} L\left(\tilde{\mathbf{y}}^i, f_{\text{noised}}\left(\tilde{\mathbf{x}}^i, \eta\right)\right), \tag{2.3}$$

其中 $f_{\text{noised}}(\cdot)$ 是一个新的决策函数。训练过程中对数据进行 Dropout、随机深度以及数据增强等噪声化处理，以此增强学生网络的性能。

(4) 随后学生网络 η 成为一个新的教师网络 θ_*，重复 (2)。

自训练方法通过在无标记的数据集上进行训练，来获得通用的数据表示。自训练的关键在于在学生网络的训练过程中加入噪声，来增强决策函数在有标签和无标签数据中的平滑性，从而获得比教师网络更高的性能。在多次迭代过程中，该网络性能会不断提高。随后，Zoph 等人[57]将自训练方法与预训练方法进行了对比实验，并得出了三点结论。(i) 更强的数据增强和更多的标记数据进一步降低了预训练的价

值。(ii) 与预训练不同，在进行数据增强和数据标记的情况下，自训练总是有助于训练精度的提高。(iii) 即使在预训练方法起作用的情况下，结合自训练的策略能够进一步提高模型的性能。

2.2.3 Transformer 结构

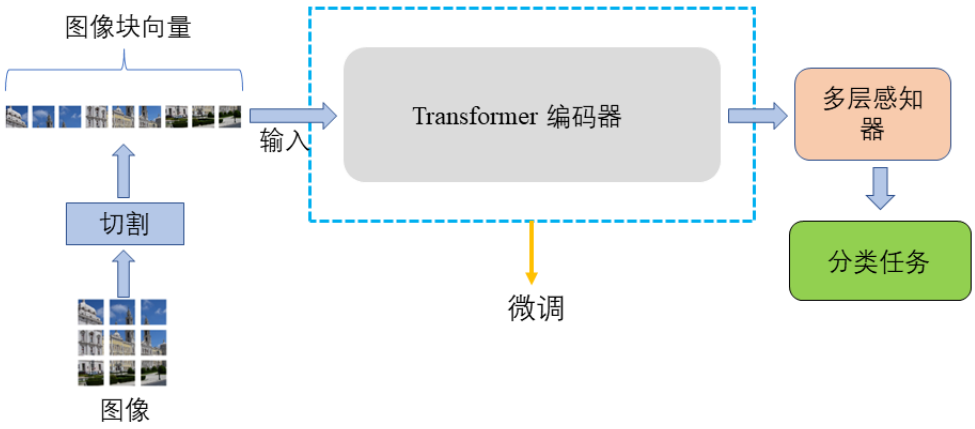


图 2.5 IPT 方法架构

Figure 2.5 Architecture of the IPT

Transformer 注意力机制在自然语言处理、目标检测、视频流处理等领域取得了巨大的成功，这种机制为迁移学习提供了一种不同于 CNN 的预训练微调框架。Chen 等人[58]首先实现了 Transformer 为核心的预训练-微调方法，他们提出的图像处理转换器，将 Transformer 应用于计算机视觉的底层任务。图2.5展示了基于 Transformer 的微调架构，对于不同的数据集，Transformer 模块是共享的，根据任务要求只需要更换新的头和尾结构即可完成迁移。这种方法将预训练微调思想引入了 Transformer，将在大模型中训练好的 Transformer 进行迁移，大大降低了训练成本和网络搭建的难度。Bao 等人[59]在此基础上提出一种自监督视觉表示模型，该方法借鉴自然语言处理中的 Bert 方法。在数据预训练阶段，每张图像都有原始图像和随机屏蔽部分图像块的损坏图像两个视图，预训练的目标是利用 Transformer 根据损坏图像恢复原始图像。在预训练的编码器上附加的任务层中，直接微调下游任务的模型参数，能够让 Transformer 不借助标记自动获取到语义区域知识，大大提高了微调能力。此外，Yang 等人[60]利用视觉 Transformer 的注意力机制和顺序图像的优势进行知识迁移，提出了可迁移的视觉转换器。

基于模型的深度迁移学习具有部署简单、训练成本低以及可操作性强的优点，但是这种方法依赖目标域样本的标签信息，在模型的可解释性以及网络迁移层的规范性的问题上还有待进一步研究。

2.3 基于差异的深度迁移学习

基于模型的方法往往要求数据包含大量的有标注的样本，并且源域和目标域数据的特征分布相似，因此不能很好地解决目标域没有标签的无监督迁移学习问题以及目标域和源域相差较大的跨领域学习问题。基于差异的深度迁移学习进一步探索了神经网络的架构以及源域和目标域的特征分布差异，有效地解决了预训练方法存在的缺陷。根据方法不同，将基于差异的深度迁移学习分为暹罗网络结构法和特征差异补偿法两类。其本质都是通过各种手段最小化源域和目标域数据集之间的特征分布差异来实现迁移学习，我们在表2.3中简要总结了基于差异的深度迁移学习方法。

表 2.3 基于差异的深度迁移学习方法总结

Table 2.3 Summary of difference based deep transfer learning

基于差异的方法	代表性工作	简介
暹罗网络架构	Tzeng 等人[61] Long 等人[62] Zhu 等人[63] Cui 等人[64]	在神经网络中引入自适应层来减少域的差异
特征差异补偿法	Nam 等人[65] Yoon 等人[66] Yu 等人[67]	直接处理网络提取的特征来对齐源域和目标域

2.3.1 暹罗网络架构

暹罗网络架构法是目前基于差异的深度迁移学习的重要框架，通过引入自适应层来减少两个域的差异。这种深度架构最大的好处是采用端到端的思想，减少了错误在冗杂步骤中的积累。目前广泛使用的神经网络，如 AlexNet、VGG、GoogleNet 以及 ResNet 等，都能被用来搭建双流的暹罗网络。模型共享的暹罗网络架构设计思路是：采用相同的神经网络架构对源域数据和目标域数据同时进行训练，在该网络中加入自适应层来减少两个域之间的差异。这样就能使得训练的网络拥有较差的域

判别能力，因此能够在两个数据集同时生效。其中不同方法的区别除了采用的网络不同以外，主要在于自适应层的设计和自适应层损失函数的不同。MMD 损失是暹罗架构中最常见的损失函数，可以形式化地写为如下的形式

$$\mathrm{MMD}\left(\mathbf{X}_{\mathrm{S}}, \mathbf{X}_{\mathrm{T}}\right)=\left\|\frac{1}{n_{\mathrm{S}}} \sum_{i=1}^{n_{\mathrm{S}}} \Phi\left(\mathbf{x}_{\mathrm{S}}\right)_i-\frac{1}{n_{\mathrm{T}}} \sum_{j=1}^{n_{\mathrm{T}}} \Phi\left(\mathbf{x}_{\mathrm{T}}\right)_j\right\|_{\mathcal{H}}^2, \tag{2.4}$$

其中，$\left(\mathbf{x}_{\mathrm{S}}\right)_i \in \mathbf{X}_{\mathrm{S}}$ 和 $\left(\mathbf{x}_{\mathrm{T}}\right)_j \in \mathbf{X}_{\mathrm{T}}$ 分别是源域和目标域的样本。n_{S} 和 n_{T} 分别表示源域和目标域样本的数量，$\Phi(\cdot)$ 是从原始数据空间到可再生核希尔伯特空间 (Reproducing Kernel Hilbert Space，RKHS) 的映射函数，而 $\mathcal{H}$ 则代表了 RKHS 中的距离度量。

Tzeng 等人[61]将 MMD 与深度学习相结合，开始了迁移学习暹罗架构的研究，提出了深度域混淆 (Deep Domain Confusion，DDC) 的方法，图2.6的灰色部分展示了 DDC 方法的架构。通常 CNN 会利用损失函数输出的分类损失在进行反向传播中优化网络参数，而 DDC 方法则利用分类损失 $L_{\mathrm{C}}(\cdot)$ 和域损失 $L_{\mathrm{MMD}}(\cdot)$ 来共同对网络参数进行优化。DDC 方法在输出层 Fc8 的前一层引入了一个适应层 Fc_Adapt，并通过适应层的输出计算出域损失 $L_{\mathrm{MMD}}(\cdot)$。文中利用源域及目标域特征之间的 MMD 距离作为域损失函数，通过最小化 MMD 距离来减小源域与目标域之间的差异。DDC 方法最终的目标损失函数如下

$$L_{\mathrm{DDC}}=L_{\mathrm{C}}+\lambda L_{\mathrm{MMD}}, \tag{2.5}$$

其中，参数 λ 是一个权衡系数，用来确定域损失对优化的影响程度。通过最小化分类损失与最小化 MMD 损失来共同优化该损失函数，并利用参数的梯度公式在反向传播中更新网络参数。此后，Tzeng 等人[68]在该研究的基础上进一步改进了 DDC 方法。具体来说，在式(2.5)中加入了由目标域网络输出的软标签损失 $L_{\mathrm{soft}}(\cdot)$。该方法同时优化了这三种损失实现了跨域的无监督迁移学习，其具体框架如图2.6所示。

Long 等人[62]对暹罗网络架构进行了进一步研究，提出了深度域适应网络 (Deep Adaptation Network，DAN) 方法。DAN 通过在分类器的前三层同时加入自适应层来进行特征约束，三层的适应层可以同时匹配源域与目标域之间的低阶矩和高阶矩。之后的研究开始从 MMD 损失的改进入手，针对 DAN 无法对齐不同类别子域的问题，Zhu 等人[63]提出了局部最大均值差异 (Local Maximum Mean Difference，LMMD) 损失，并将该损失用于子域适应的深度神经网络架构。它通过捕获每个类别的细粒度信息来

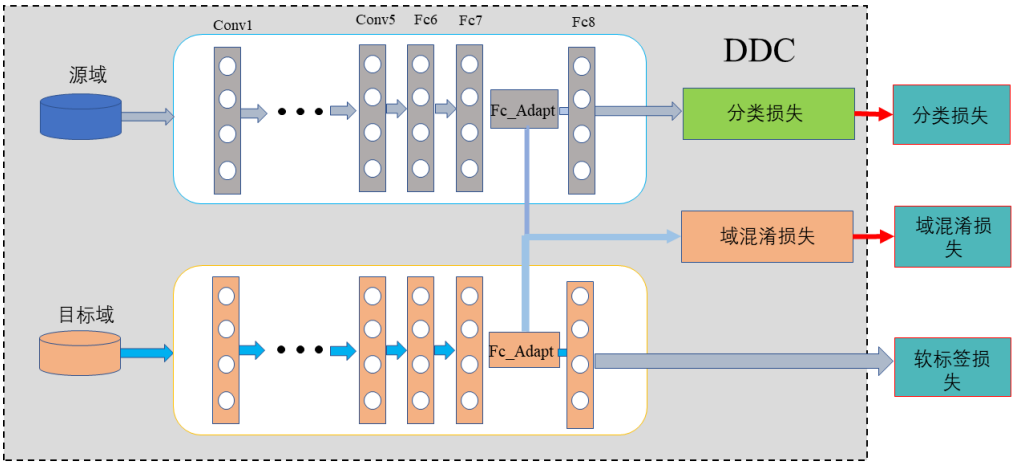


图 2.6 基于差异的深度迁移学习网络

Figure 2.6 Difference based deep transfer learning network

扩展 DAN 的能力，能对齐源域和目标域中具体某一类样本之间的差异。深度子域适应网络 (Deep Subdomain Adaptation Network，DSAN) 可以准确地对齐源域和目标域同一类别内的相关子域分布，在大部分跨域分类任务中都有较好的表现。图2.7展示了 DSAN 的方法框架，卷积神经网络部分可以替换为目前任何主流的神经网络。通过全连接层每一层输出的源域数据特征矩阵 $P_{\mathrm{S}}^{n-i},\ldots,P_{\mathrm{S}}^{n-1}$ 和目标域数据特征矩阵 $P_{\mathrm{T}}^{n-i},\ldots,P_{\mathrm{T}}^{n-1}$，计算局部最大均值差异，最后将该损失与分类损失结合共同优化网络参数。

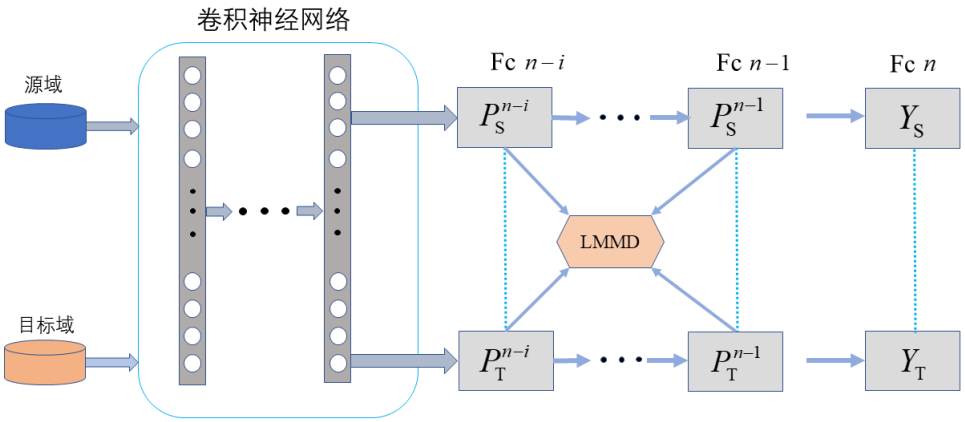


图 2.7 DSAN 网络架构

Figure 2.7 Architecture of the DSAN network

其他形式的度量准则也被广泛应用于暹罗网络架构中。例如，Sun 等人[69]将相关对齐 (Correlation Alignment, CORAL) 损失嵌入到网络中提出了 DeepCORAL 方法，通过优化源域分类损失和 CORAL 损失，对齐域分布的二阶统计量，实现了目标域未标记时的图像分类问题。Cui 等人[64]提出了一种快速批核范数最大最小化 (Batch Nuclear-norm Maximization，BNM) 的方法，通过对全连接层的批输出矩阵进行操作减少源域和目标域的差异。该研究通过理论证明标签预测的鉴别性和多样性与输出矩阵的 F 范数和秩有关，并且可以通过核范数对二者进行约束。具体来说，BNM 对目标域输出矩阵进行核范数最大化，以提高目标预测能力；对源域批输出矩阵进行核范数最小化，以提高源域的适用性。

2.3.2 特征差异补偿法

暹罗网络架构不可避免的需要在网络训练过程中引入各类损失函数对齐源域和目标域。虽然这类方法部署难度低、操作简单，但是各类损失与距离度量函数在应用时存在鲁棒性差、无法很好地适应训练数据与测试数据分布差异的问题。下面介绍的方法直接对图像特征进行操作，试图通过各种方式弥补特征之间的差异。

深度学习在图像分类任务中无法有效完成跨数据集检测的主要原因是 CNN 对图像纹理特征较为敏感，因此不同数据集图像纹理特征差异较大是阻碍迁移学习有效进行的主要原因之一。Nam 等人[65]提出了风格无关网络，实现将风格编码与图像内容分离以减少域偏差。该网络的特征提取器不仅提取图像的内容，还提取图像的样式。在内容偏置网络中，通过自适应实例归一化对样式进行随机初始化，使该网络专注于图像内容。在偏向风格的网络中，情况正好相反。对于风格特征的利用，Yoon 等人[66]提出的知识蒸馏方法，通过在标记样本和未标记目标之间转移中间风格来生成助手特征。最后极小化助手特征与未标记目标之间的输出差异来逐渐弥合两个领域，从而完成域泛化。此外，Yu 等人[67]通过与元学习思想结合，以数据驱动的方式学习分布匹配来减少归纳偏差。

2.4 基于对抗的深度迁移学习

随着生成式对抗网络 (Generative Adversarial Network，GAN) 在图像处理领域取得了的巨大成功，研究人员试图将 GAN 与迁移学习相结合。我们首先对生成对抗网络中的概念进行简单介绍，GAN 由生成器网络 $G(\cdot)$ 和鉴别器网络 $D(\cdot)$ 两个子网络

组成，目标函数如下

$$\min_{G}\max_{D} \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{x}}}\left(\log\left(D\left(\mathbf{x}\right)\right)+\mathbb{E}_{\mathbf{z}\sim P_{\mathbf{z}}}\log\left(1-D\left(G\left(\mathbf{z}\right)\right)\right)\right), \tag{2.6}$$

其中，$\mathbf{x},\mathbf{z}\in\mathbb{R}^{d}$ 表示从真实数据分布 $P_{\mathbf{x}}$ 和先验分布 $P_{\mathbf{z}}$(通常是均匀分布或高斯分布)采集的样本，$G\left(\cdot\right)$ 是从 $P_{\mathbf{x}}$ 到 $P_{\mathbf{z}}$ 的映射，$D\left(\cdot\right)$ 的作用是来区分真实数据和生成器生成的数据并输出一个概率，$\mathbb{E}\left(\cdot\right)$ 表示不同分布的期望分布。$G\left(\cdot\right)$ 和 $D\left(\cdot\right)$ 相互对抗来完成训练，当模型收敛时生成器就可以生成真实的样本。

基于 GAN 的迁移学习可以在源域样本和目标域样本之间进行“翻译”，同时保留原始的标签信息。这种方法通过解决最大最小的博弈问题来减少源域和目标域的差异，方法简单高效引起了学者们的广泛关注，表2.4对该方法进行了简单的总结。

表 2.4　基于差异的深度迁移学习方法总结

Table 2.4　Summary of difference based deep transfer learning

基于差异的方法	代表性工作	简介
特征提取法	Ganin 等人[70] Yu 等人[71] Long 等人[72]	用对抗训练的方式从源域和目标域提取域不变特征
特征变换法	Zhu 等人[73] Tzeng 等人[74] Shrivastava 等人[75]	通过对抗训练转换域间特征来减少领域偏差

2.4.1　特征提取法

特征提取方法是通过对抗训练提取源域和目标域的域不变特征来进行迁移学习。Ganin 等人[70]第一次在迁移学习中加入了对抗训练的机制，提出了领域对抗神经网络 (Domain Adversarial Neural Network，DANN)。DANN 利用了生成对抗网络的特点，通过特征提取器和域鉴别器相互竞争来学习域不变特征。训练稳定后，从源域和目标域中提取的特征变得越来越相似，此时一个来自源域的分类器可以在目标域中完成分类任务。图2.8展示了该方法的框架，DANN 由三个自网络组成：在源域和目标域中共享的特征提取器 $G_{\mathrm{f}}\left(\theta_{\mathrm{f}}\right)$，能够对源域的数据进行分类的标签预测器 $G_{\mathrm{y}}\left(\theta_{\mathrm{y}}\right)$，以及一个能够对源域和目标域特征进行分类的域分类器 $G_{\mathrm{d}}\left(\theta_{\mathrm{d}}\right)$。$F_S$ 和 F_T 是来自源域和目标域的特征，是 $G_{\mathrm{y}}\left(\theta_{\mathrm{y}}\right)$ 和 $G_{\mathrm{d}}\left(\theta_{\mathrm{d}}\right)$ 的输入。L_{y} 和 L_{d} 表示源域的分类损失和域分

类损失，DANN 将 L_{y} 和 L_{d} 这两个损失通过梯度反转层传递到特征提取器，来进行反向传播优化。训练的最终目标是使域分类器 $G_{\mathrm{d}}\left(\theta_{\mathrm{d}}\right)$ 无法区分由特征提取器传递来的特征是属于源域还是目标域。接下来我们将对训练过程进行介绍。

(1) 最小化分类损失和特征提取器损失来优化 $G_f\left(\theta_f\right)$ 和 $G_y\left(\theta_y\right)$ 的参数 θ_{f} 和 θ_{y}，即

$$\left(\hat{\theta}_{\mathrm{f}}, \hat{\theta}_{\mathrm{y}}\right)=\arg \min _{\theta_{\mathrm{f}}, \theta_{\mathrm{y}}} \mathbb{E}\left(\theta_{\mathrm{f}}, \theta_{\mathrm{y}}, \theta_{\mathrm{d}}\right). \tag{2.7}$$

(2) 最大化 $G_{\mathrm{d}}\left(\cdot\right)$ 的损失来优化参数 θ_{d}，即

$$\left(\hat{\theta}_{\mathrm{d}}\right)=\arg \max _{\theta_{\mathrm{d}}} \mathbb{E}\left(\theta_{\mathrm{f}}, \theta_{\mathrm{y}}, \theta_{\mathrm{d}}\right). \tag{2.8}$$

(3) DANN 通过添加一个梯度反转层来合并上述两个训练过程，最终得到的整体优化目标如下

$$\mathbb{E}\left(\theta_{\mathrm{f}}, \theta_{\mathrm{y}}, \hat{\theta}_{\mathrm{d}}\right)=\sum_{\mathbf{x} \in \mathcal{D}_S} L_{\mathrm{y}}\left(G_{\mathrm{y}}\left(G_{\mathrm{f}}\left(\mathbf{x}\right)\right), \mathbf{y}\right)-\lambda \sum_{\mathbf{x} \in \mathcal{D}_S \cup \mathcal{D}_T} L_{\mathrm{d}}\left(G_{\mathrm{d}}\left(G_{\mathrm{f}}\left(\mathbf{x}\right)\right), j\right), \tag{2.9}$$

其中 j 代表域标签。当数据来自源域时，$j=0$，否则为 $j=1$，$\mathbf{x} \in \mathbb{R}^d$ 和 $\mathbf{y}$ 表示输入的图像样本和样本对应的标签独热码矩阵。λ 是一个权衡参数，能够平衡两个损失的权重。

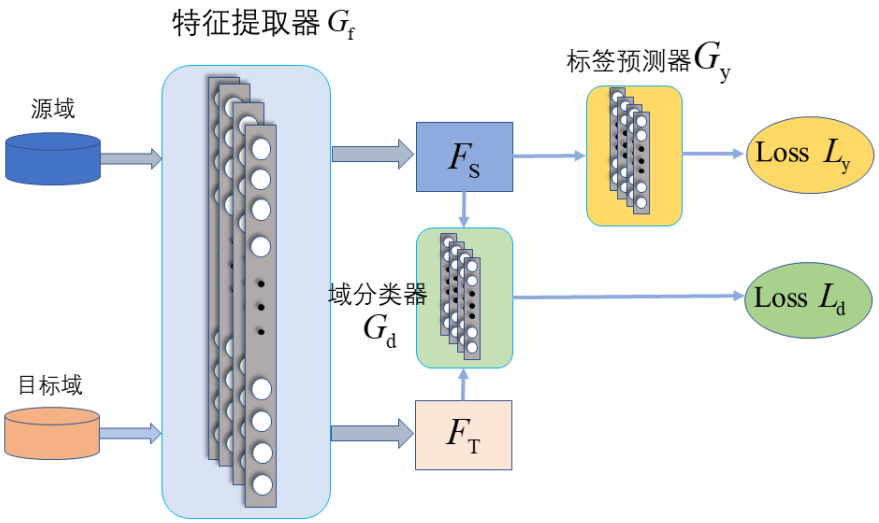


图 2.8 DANN 网络架构

Figure 2.8 Architecture of the DANN network

DANN 考虑了数据特征的整体分布，却忽略了类别之间的相关性。Long 等人[72]

提出了条件对抗域适应网络，这种方法对特征和类别同时进行自适应以得到深度特征之间的关系。此外，该方法使用多线性映射来优化 GAN，在一定程度上改善了负迁移。受到上述工作的启发，Yu 等人[71]进一步优化了 DANN，提出了动态对抗性自适应网络 (Dynamic Adversarial Adaptation Network，DAAN)。这种方法在域分类损失的设计中引入了自适应因子，动态定量地评估边缘分布和条件分布两种决策对学习的贡献度。图2.9展示了 DAAN 方法的框架，深度特征提取器 (蓝色) 完成源域和目标域的特征提取，标签预测器 (橙色) 利用提取的特征进行训练输出分类损失。与 DANN 方法不同的是域分类损失部分，该损失由全局域分类损失 (紫色) 和多个子域分类损失 (绿色) 两部分构成。子域分类损失对输入样本的每一类进行计算输出域损失，通过动态测量整体特征域分类器和多个子域分类器的权值来更新动态测量因子 ω。作为权重参数，ω 能够衡量全局域鉴别器和子域鉴别器的比重。最后将分类损失和两个域损失作为整体的损失函数，在网络的反向传播中完成更新。

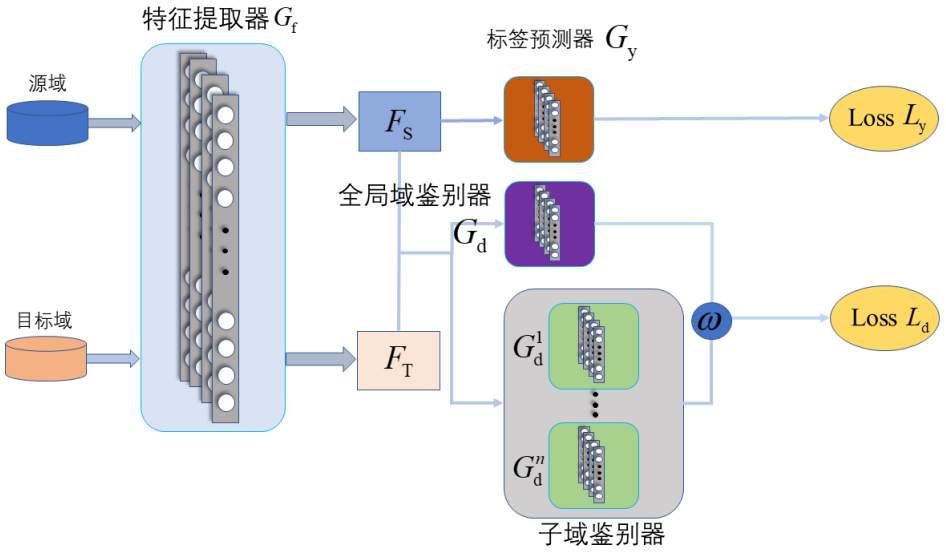


图 2.9 DAAN 网络架构

Figure 2.9 Architecture of the DAAN network

2.4.2 特征变换法

特征变换是对抗迁移学习的一种重要方法，它通过对抗性训练对特征进行变换或对齐来减少域偏差。域映射是特征变换的一种代表性方法，下面对这个方法的几个研究进行介绍。Tzeng 等人[74]结合判别建模、非共享权值和 GAN 损失提出了一种

对抗鉴别域适应方法。该方法首先在源域中使用标签学习判别器表示，然后通过域对抗性损失学习非对称映射，将目标数据映射到相同空间单独编码。这种方法的核心是将目标域数据映射到源域，在这个过程中生成器对模拟器生成的合成图像进行优化，并且在判别器网络优化后加入正则化损失到模拟器标注中，进行交替更新完成训练。还有部分研究是将源域数据映射到目标域，Shrivastava 等人[75]提出了模拟对抗网络，这种方法能够将源域样本转换为目标域，以学习在两个域上可用的识别分类器。Zhu 等人[73]提出了周期一致的对抗网络，该网络通过测量源域映射到目标域，再映射到源域的数据与原始数据之间的差异来进行训练。

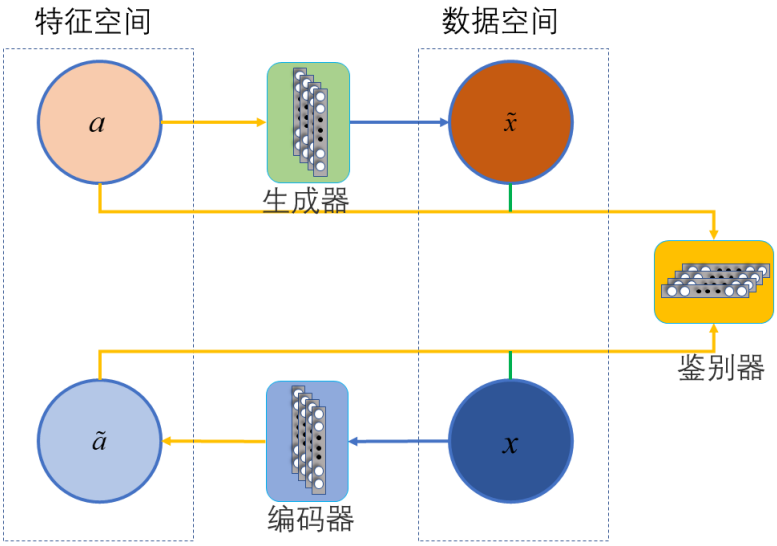


图 2.10　逆向学习推理方法框架

Figure 2.10　Framework of the reverse learning reasoning

基于对抗性训练的特征对齐是特征变换的另一种代表性方法。Kurmi 等人[76]运用 Dropout 正则化的方法进行特征对齐。该方法以蒙特卡洛 Dropout 鉴别器获得的分布估计，代替单个判别器获得的点估计。该分布鉴别器能够逐渐增加基于样本的分布方差，使用相应的反向梯度对齐源域和目标域特征。Saito 等人[77]从减少决策边界入手来对齐源域和目标域的分布，具体做法是最大化两个分类器输出之间的差异，以检测远离源域支持的目标样本，特征生成器学习在源域附近生成的目标特征以最小化差异。在特征空间变换方面，Hoffman 等人[78]提出了循环一致的对抗性领域适应方法。该方法结合对抗训练和特征空间变换的思想，通过交叉熵损失保持对抗训练前后图像的语义特征，优化循环一致性损失与对抗损失完成域适应。Dumoulin 等

人[79]从双向映射的角度出发，提出了逆向学习推理的方法。图2.10展示了该方法的思想框架，其中 **x** 表示数据空间中真实数据的样本, **a** 表示特征空间中真实数据的样本, $\tilde{\mathbf{x}}$ 表示生成器 (绿色) 输出的样本, $\tilde{\mathbf{a}}$ 表示编码器 (蓝色) 输出的样本。生成器完成特征空间到数据空间的映射，编码器完成数据空间到特征空间的反向映射，鉴别器用于区分生成器和编码器中的数据是来自特征空间还是数据空间，鉴别器完成区分，三者之间完成对抗性训练。

2.5 本章小结

本章介绍了迁移学习预备知识，回顾了深度迁移学习近十年的发展概况。首先，简单概括了传统的迁移学习方法。然后，重点总结了深度迁移学习技术的发展，根据模型、函数和操作对象的不同将其分为：基于模型的深度迁移学习、基于差异的深度迁移学习以及基于生成对抗的深度迁移学习。此外，对于其中的一些代表方法进行了详细介绍，并给出了其目标函数和训练过程，总结了它们的优点和不足之处。

第三章　基于迁移子空间学习的轴承故障诊断方法

本章首先介绍了迁移子空间学习方法，给出了迁移子空间学习目标函数的一般形式，在此基础上改进并提出了两种基于迁移子空间学习的故障诊断方法。针对一般的迁移子空间学习方对高斯噪声鲁棒性较低的问题，引入了一个矩阵建模高斯噪声，有效减少了数据中高斯噪声对模型精度的影响，提出了基于高斯噪声改进的子空间学习方法记为 TSL-1。在此基础上，针对其算法自由度较低、空间变换矩阵不够灵活的问题，使用松弛回归矩阵代替原本的标签矩阵，有效增强了回归算法的自由度，提出了基于松弛回归矩阵改进的子空间学习方法记为 TSL-2。大量的数值实验和分析表明，我们提出的方法在分类和抗噪声方面取得了优势。与目前最好的方法 RLCSL 相比，TSL-2 在 CWRU 数据集的 12 个迁移任务上平均精度提升了 1.77%，TSL-1 和 TSL-2 在 JNU 数据集的 6 个迁移任务上的平均精度分别提升了 0.17% 和 2.21%。实验代码可以在https://github.com/FuchaoYu进行下载。

3.1　引言

在传统的迁移学习领域，特征空间变换法是解决数据分布不一致问题的有效方法。其思想是改变原始数据的表示，在另一个空间中将原始数据特征矩阵用等价的矩阵表示。比较常见的几类方法有在第二章介绍过的：TCA、JDA、TJM 以及 BDA 等方法，但是这类方法有一定的缺陷。首先，由于数据分布的不同很难捕获它的内在结构，如全局结构和局部结构。其次，没有对特征矩阵中的噪声进行有效处理，不利于模型的鲁棒性。最后，只关注如何改变数据的表示，却忽视了将分类器设计和改变数据表示的方法整合为一项任务，而这可以减少中间步骤，避免误差累积从而更好地解决问题。

在过去的几十年里，基于低秩表示的子空间学习方法被广泛应用于机器学习与模式识别等领域。与传统的假设有特定噪声的子空间恢复方法相比，基于低秩表示的方法可以有效地处理幅度较大的噪声。Xu 等人[50]在变换矩阵中引入了低秩稀疏表示 (Low-rank and Sparse Representation，LRSR) 来寻求最优子空间，其中低秩矩阵可以捕捉全局结构，而稀疏矩阵可以捕捉局部结构。此外，数据的 LRSR 可以在一定

程度上减轻噪声的影响，保留子空间中的标签信息。随后，Lu 等人[80]通过流形学习将局部几何结构信息从源域转移到目标域，Liu 等人[81]将核函数嵌入 LRSR，以增强模型的鲁棒性。最近，Zhan 等人[82]将一个联合学习框架集成到 LRSR 方法中，提出了鲁棒潜在公共子空间学习方法 (Robust Latent Common Subspace Learning，RLCSL)。

在故障诊断领域，基于低秩表示的子空间学习方法已经有了一定的研究和应用[83-85]。2023 年，Li 等人[86]对稀疏正则化在故障诊断中的研究进行了全面的回顾。该综述根据正则化项和优化特点将稀疏正则化模型分为凸稀疏正则化和非凸稀疏正则化两类，并介绍了解决稀疏正则化模型的优化算法。然而，这类关于低秩稀疏方法的研究并未考虑数据在不同条件下的偏移，只是在同一种工况下展开故障诊断研究。基于迁移子空间学习的方法在故障诊断领域尚未得到广泛的关注，这其中存在许多挑战。首先，收集到的数据来自不同的工作环境，它们是混合的、无序的。数据的这些特点不利于模型的迁移，这将会导致负迁移。其次，数据中的噪声会降低模型的鲁棒性，大多数方法只考虑了某种特定类型噪声的影响。最后，虽然一些标签拖曳技术为变换矩阵提供了一些自由度，但这些方法都是从固定的二进制标签矩阵出发，这会导致子空间变换矩阵的灵活度降低。

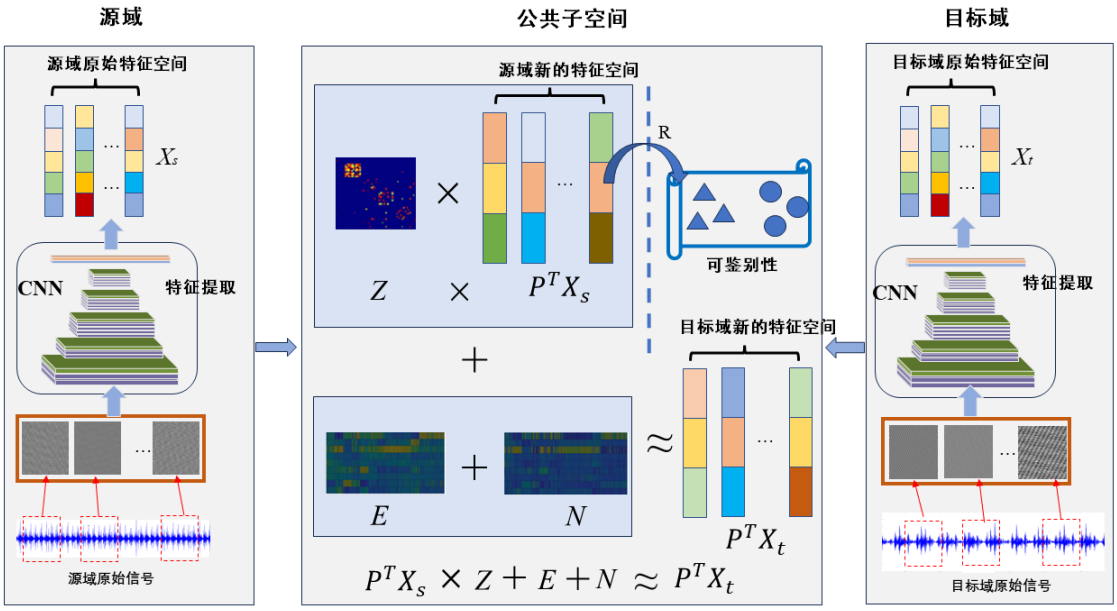


图 3.1 所提方法的框架

Figure 3.1 Framework of the proposed methods

图3.1直观地展示我们的方法，$\mathbf{X}_S$ 和 $\mathbf{X}_T$ 是由两种不同工况下产生的故障信号并

经过 CNN 特征提取后生成的特征矩阵。我们的方法尝试找到一个变换矩阵 **P**，从而把数据变换到一个能够减少数据分布差异的公共子空间中。同时在子空间中对矩阵 **Z** 施加低秩和稀疏约束, 用矩阵 **E** 和 **N** 来对噪声建模, 以及引入松弛回归矩阵 **R** 来提高模型鉴别性等操作来改进原始的迁移子空间学习方法。

综上，本章工作的创新点如下：

(1) 将一维信号图像化并进行提取特征，减少了数据中冗余信息的影响。

(2) 引入了矩阵建模高斯噪声，进而增强模型的鲁棒性。

(3) 从样本中直接学习标签矩阵，提高子空间变换矩阵的自由度。

3.2 迁移子空间学习

迁移子空间学习的目标是找到一个变换矩阵，通过这个变换矩阵可以将源域数据 $\mathbf{X}_S \in \mathbb{R}^{m\times n_S}$ 和目标域数据 $\mathbf{X}_T \in \mathbb{R}^{m\times n_T}$ 变换到一个公共的流形子空间中，其中 m 为原始空间的数据维数，n_S 与 n_T 分别为源域和目标域数据的样本数。在这个子空间中，源域数据和目标域的数据分布相似，因此目标域中的每一个样本可以由源域中的样本线性表示。进而可以形式化地描述为

$$\mathbf{P}^\top \mathbf{X}_T = \mathbf{P}^\top \mathbf{X}_S \mathbf{Z}, \tag{3.1}$$

其中，$\mathbf{P} \in \mathbb{R}^{m\times d}$ 为变换矩阵，d 为公共子空间的维数，而 $\mathbf{Z} \in \mathbb{R}^{n_S\times n_T}$ 则表示线性变换矩阵。此外，为了使线性变换矩阵 **Z** 能够捕捉数据的全局结构，式(3.1)可以写为

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{P}} \ \operatorname{rank}(\mathbf{Z}) \\ &\text{s.t.} \ \ \mathbf{P}^\top \mathbf{X}_T = \mathbf{P}^\top \mathbf{X}_S \mathbf{Z}. \end{aligned} \tag{3.2}$$

这里的 rank(**Z**) 表示矩阵 **Z** 的秩。由于上式中求解 **Z** 的秩最小化问题是非凸且 NP 难的，因此式(3.2)可以被进一步表示为如下凸优化形式

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{P}} \ \|\mathbf{Z}\|_* \\ &\text{s.t.} \ \ \mathbf{P}^\top \mathbf{X}_T = \mathbf{P}^\top \mathbf{X}_S \mathbf{Z}. \end{aligned} \tag{3.3}$$

为了能够在变换过程中保持数据的局部结构，可以对 $\mathbf{Z}$ 施加稀疏性约束，从而得到

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{P}} \ \|\mathbf{Z}\|_* + \alpha\|\mathbf{Z}\|_1 \\ &\text{s.t.} \ \ \mathbf{P}^\top\mathbf{X}_T = \mathbf{P}^\top\mathbf{X}_S\mathbf{Z}, \end{aligned} \tag{3.4}$$

其中 $\alpha > 0$ 是用于权衡矩阵 $\mathbf{Z}$ 低秩性和稀疏性权重的参数。为了减少噪声的影响，引入了矩阵 $\mathbf{E} \in \mathbb{R}^{m\times n_t}$ 来对噪声进行建模，式(3.4)可以改写为

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{E},\mathbf{P},\mathbf{M}} \ \|\mathbf{Z}\|_* + \alpha\|\mathbf{Z}\|_1 + \beta\|\mathbf{E}\|_1 \\ &\text{s.t.} \quad P^\top\mathbf{X}_T = P^\top\mathbf{X}_S\mathbf{Z} + \mathbf{E}, \end{aligned} \tag{3.5}$$

其中，$\beta > 0$，其作用与 α 作用相同也是一个权衡参数。

最后，迁移子空间学习方法引入了判别子空间学习函数，保证在数据变换的过程中源域数据的标签特征不被破坏。这个函数本质上是一个最小二乘回归算法，可以写为

$$\|\mathbf{P}^\top\mathbf{X}_S - (\mathbf{Y} + \mathbf{B}\circ\mathbf{M})\|_{\mathrm{F}}^2, \tag{3.6}$$

其中 $\mathbf{Y}$ 是源域数据的标签矩阵，$\mathbf{B}$ 是一个二进制矩阵，定义为

$$b_{ij} = \begin{cases} +1, & y_{ij} = 1, \\ -1, & y_{ij} = 0, \end{cases} \tag{3.7}$$

其中矩阵 $\mathbf{M}$ 是一个需要学习的非负标签松弛矩阵。LRSR 的最终目标函数如下

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{E},\mathbf{P},\mathbf{M}} \ \frac{1}{2}\|\mathbf{P}^\top\mathbf{X}_S - (\mathbf{Y} + \mathbf{B}\circ\mathbf{M})\|_{\mathrm{F}}^2 + \|\mathbf{Z}\|_* + \alpha\|\mathbf{Z}\|_1 + \beta\|\mathbf{E}\|_1 \\ &\text{s.t.} \quad \mathbf{P}^\top\mathbf{X}_T = \mathbf{P}^\top\mathbf{X}_S\mathbf{Z} + \mathbf{E}, \ \mathbf{M} \geq 0, \end{aligned} \tag{3.8}$$

其中，$\mathbf{M} \geq 0$ 表示 $\mathbf{M}$ 中每个元素非负。这种方法可以在一定程度上缓解一般噪声的影响，并在子空间中保留标签信息。在此基础上，Liu 等人[81]将核函数嵌入到上述模型，以增强模型的鲁棒性。

LRSR 方法存在两个不足：一是迁移子空间学习对噪声的鲁棒性较低，在建模时只考虑了非高斯噪声的影响，无法有效处理高斯噪声；二是判别子空间学习函数 $\phi(\cdot)$ 中标签矩阵 $\mathbf{Y}$ 的自由度较低，限制了变换矩阵 $\mathbf{P}$ 的灵活度。下面我们针对这两个问

题对 LRSR 进行了改进，分别在 3.3 节和 3.4 节提出了基于高斯噪声改进的子空间学习方法和基于松弛回归矩阵改进的子空间学习方法。

3.3 基于高斯噪声改进的子空间学习

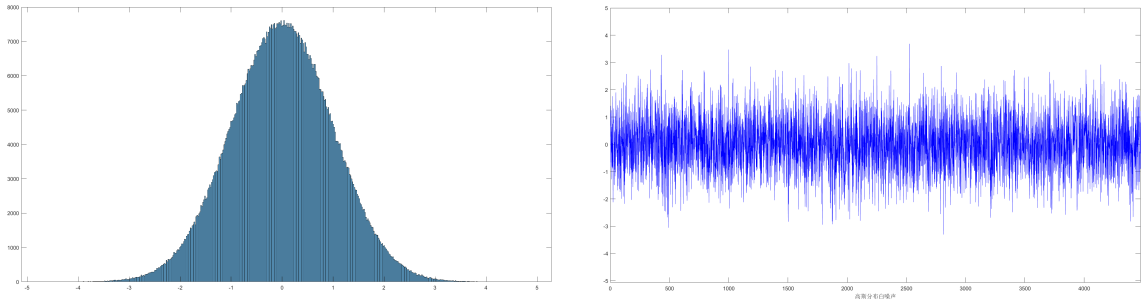
式(3.8)通过引入矩阵 **E** 对非高斯噪声进行建模，并添加稀疏约束来过滤噪声，这个过程可以实现对白噪声的有效过滤。但是数据中还存在高斯噪声，这会在迁移过程中模糊数据的某些特征影响模型的精度，因此本节通过对高斯噪声进行建模来提高模型对噪声的鲁棒性。针对提出的基于高斯噪声改进的子空间学习方法，设计了一种有效的交替方向乘子算法 (Alternating Direction Method of Multipliers，ADMM)。

3.3.1 数学模型

高斯噪声一般指噪声点的概率密度函数服从正态分布的一类噪声。它区别于椒盐噪声的随机分布，高斯噪声的噪声点分布与噪声强度之间服从高斯分布，即在某个噪声强度下噪声点分布最多，远离这个强度噪声点就越少。高斯噪声的概率密度函数表达式为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\nu)^2}{2\sigma^2}\right). \tag{3.9}$$

这是一个服从均值为 ν，方差为 σ^2 的高斯噪声。图3.2展示了在 1000 个采样点下均值为 0，标准差为 0.5 的高斯噪声的频率分布直方图。



(a) (b)

图 3.2 高斯噪声图

Figure 3.2 Figures of gaussian noise

我们通过在式(3.8)中嵌入矩阵 **N** 并对其进行 F-范数正则化来建模高斯噪声，所

得的的模型可以同时过滤数据中的高斯噪声和随机噪声，模型的鲁棒性进一步提升。TSL-1 最终的数学模型如下

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{E},\mathbf{N},\mathbf{Y},\mathbf{P}}\ & \frac{1}{2}\|\mathbf{P}^\top\mathbf{X}_S-(\mathbf{Y}+\mathbf{B}\circ\mathbf{M})\|_{\mathrm{F}}^2+\|\mathbf{Z}\|_*+\alpha\|\mathbf{Z}\|_1+\beta\|\mathbf{E}\|_1+\gamma\|\mathbf{N}\|_{\mathrm{F}}^2\\ \text{s.t.}\quad & \mathbf{P}^\top\mathbf{X}_T=\mathbf{P}^\top\mathbf{X}_S\mathbf{Z}+\mathbf{E}+\mathbf{N},\ \mathbf{M}\geq 0,\end{aligned}\tag{3.10}$$

其中，$\|\mathbf{N}\|_{\mathrm{F}}^2$ 表示 $\mathbf{N}$ 的 F 范数。

3.3.2 优化算法

本节将对上节中提出的数学模型进行求解。为了便于求解，我们引入 $\mathbf{Z}_1,\mathbf{Z}_2$，将式(3.10)改写为

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{E},\mathbf{N},\mathbf{M},\mathbf{P}}\ & \frac{1}{2}\|\mathbf{P}^{k\top}\mathbf{X}_S-(\mathbf{Y}+\mathbf{B}\circ\mathbf{M})\|_{\mathrm{F}}^2+\|\mathbf{Z}_1\|_*+\alpha\|\mathbf{Z}_2\|_1+\beta\|\mathbf{E}\|_1+\gamma\|\mathbf{N}\|_{\mathrm{F}}^2\\ \text{s.t.}\quad & \mathbf{P}^\top\mathbf{X}_T=\mathbf{P}^\top\mathbf{X}_S\mathbf{Z}+\mathbf{E}+\mathbf{N},\ \mathbf{Z}_1=\mathbf{Z},\ \mathbf{Z}_2=\mathbf{Z}.\end{aligned}\tag{3.11}$$

ADMM 是求解等式约束优化问题的一种有效算法，它能够将原始的优化问题拆解成几个相对好解决的子优化问题进行迭代求解。式(3.11)的增广拉格朗日函数为

$$\begin{aligned}&\mathcal{L}_\mu(\mathbf{Z},\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{E},\mathbf{N},\mathbf{R},\mathbf{P},\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\mathbf{\Lambda}_3)\\ &=\frac{1}{2}\|\mathbf{P}^\top\mathbf{X}_S-(\mathbf{Y}+\mathbf{B}\circ\mathbf{M})\|_{\mathrm{F}}^2+\|\mathbf{Z}_1\|_*+\alpha\|\mathbf{Z}_2\|_1+\beta\|\mathbf{E}\|_1\\ &\quad+\gamma\|\mathbf{N}\|_{\mathrm{F}}^2+\frac{\mu}{2}\|\mathbf{P}^\top\mathbf{X}_T-\mathbf{P}^\top\mathbf{X}_S\mathbf{Z}-\mathbf{E}-\mathbf{N}\|_{\mathrm{F}}^2\\ &\quad+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{Z}_1\|_{\mathrm{F}}^2+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{Z}_2\|_{\mathrm{F}}^2\\ &\quad+\langle\mathbf{\Lambda_1},\mathbf{P}^\top\mathbf{X}_t-\mathbf{P}^\top\mathbf{X}_S\mathbf{Z}-\mathbf{E}-\mathbf{N}\rangle\\ &\quad+\langle\mathbf{\Lambda_2},\mathbf{Z}-\mathbf{Z}_1\rangle+\langle\mathbf{\Lambda}_3,\mathbf{Z}-\mathbf{Z}_2\rangle,\end{aligned}\tag{3.12}$$

其中，$\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\mathbf{\Lambda}_3$ 是拉格朗日乘子，$\mu\geq 0$ 是一个惩罚参数。接下来将逐步更新增广拉格朗日函数中待求解的变量，直至收敛。

(1) 更新 $\mathbf{Z}^{k+1}$：将其他变量固定，$\mathbf{Z}^{k+1}$ 可以通过最小化下式进行更新

$$\begin{aligned}\min_{\mathbf{Z}}\ & \frac{\mu}{2}\|\mathbf{P}^{k\top}\mathbf{X}_T-\mathbf{P}^{k\top}\mathbf{X}_S\mathbf{Z}-\mathbf{E}^k-\mathbf{N}^k+\mathbf{\Lambda}_1^k/\mu\|_{\mathrm{F}}^2\\ &+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{Z}_1^k+\mathbf{\Lambda}_2^k/\mu\|_{\mathrm{F}}^2+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{Z}_2^k+\mathbf{\Lambda_3}^k/\mu\|_{\mathrm{F}}^2.\end{aligned}\tag{3.13}$$

令 $\mathbf{T}_1^k=\mathbf{Z}_1^k-\mathbf{\Lambda_2}^k/\mu+\mathbf{Z}_2^k-\mathbf{\Lambda_3}^k/\mu$，$\mathbf{T}_2^k=\mathbf{P}^{k\top}\mathbf{X}_T-\mathbf{E}^k-\mathbf{N}^k-\mathbf{\Lambda}_1^k/\mu$。通过化简

可以得到以下显示解

$$\mathbf{Z}^{k+1} = \left(\mu \mathbf{X}_S^\top \mathbf{P}^k \mathbf{P}^{k\top} \mathbf{X}_S + 2\mu \mathbf{I}\right)^{-1} (\mathbf{T}_1^k - \mathbf{X}_S^\top \mathbf{P}^k \mathbf{T}_2^k). \tag{3.14}$$

(2) 更新 $\mathbf{Z}_1^{k+1}$：将其他变量固定，$\mathbf{Z}_1^{k+1}$ 的求解可以化简为

$$\min_{\mathbf{Z}_1} \|\mathbf{Z}_1\|_* + \frac{\mu}{2}\|\mathbf{Z}^{k+1} - \mathbf{Z}_1 + \mathbf{\Lambda}_2^k/\mu\|_{\mathrm{F}}^2. \tag{3.15}$$

根据文献[87]中提供的求解核范数的思路，$\mathbf{Z}_1^{k+1}$ 有以下显示解

$$\mathbf{Z}_1^{k+1} = \mathcal{D}_{1/\mu}(\mathbf{Z}^{k+1} + \mathbf{\Lambda}_2^k/\mu), \mathbf{Z}_1^{k+1} = \mathcal{D}_{1/\mu}(\mathbf{Z}^{k+1} + \mathbf{\Lambda}_2{}^k/\mu). \tag{3.16}$$

这里 $\mathcal{D}_\tau(\mathbf{T}) = \mathbf{U}\mathcal{S}_\tau(\boldsymbol{\Sigma})\mathbf{V}^\top$ 为奇异值算子，其中 $\mathcal{S}_\tau(\Sigma_{ii}) = \text{sign}(\Sigma_{ii})\text{max}(0, |\Sigma_{ii} - \tau|)$ 是软阈值算子。

(3) 更新 $\mathbf{Z}_2^{k+1}$：按照上面的求解思路，$\mathbf{Z}_2^{k+1}$ 可以表示为

$$\min_{\mathbf{Z}_2} \alpha\|\mathbf{Z}_2\|_1 + \frac{\mu}{2}\|\mathbf{Z}^{k+1} - \mathbf{Z}_2 + \mathbf{\Lambda}_3^k/\mu\|_{\mathrm{F}}^2, \tag{3.17}$$

进而可以得到 $\mathbf{Z}_2^{k+1}$ 的显示解为

$$\mathbf{Z}_2^{k+1} = \mathcal{S}_{\alpha/\mu}(\mathbf{Z}^{k+1} + \mathbf{\Lambda}_3^k/\mu). \tag{3.18}$$

(4) 更新 $\mathbf{E}^{k+1}$：同理通过化简，可以得到非高斯噪声矩阵 $\mathbf{E}^{k+1}$ 的解

$$\mathbf{E}^{k+1} = \mathcal{S}_{\beta/\mu}(\mathbf{P}^{k\top}\mathbf{X}_T - \mathbf{P}^{k\top}\mathbf{X}_S\mathbf{Z}^{k+1} - \mathbf{N}^k + \mathbf{\Lambda}_1^k/\mu). \tag{3.19}$$

(5) 更新 $\mathbf{N}^{k+1}$：对于高斯噪声矩阵 $\mathbf{N}^{k+1}$，其显示解为

$$\mathbf{N}^{k+1} = \frac{\mu}{2\gamma + \mu}(\mathbf{P}^{k\top}\mathbf{X}_T - \mathbf{P}^{k\top}\mathbf{X}_S\mathbf{Z}^{k+1} - \mathbf{E}^{k+1} + \mathbf{\Lambda}_1^k/\mu). \tag{3.20}$$

(6) 更新 $\mathbf{M}^{k+1}$：容易得到 $\mathbf{M}^{k+1}$ 的求解表达式为

$$\min_{\mathbf{V}\geq 0} \frac{1}{2}\|\mathbf{P}^{k\top}\mathbf{X}_s - (\mathbf{Y} + \mathbf{B} \circ \mathbf{V})\|_{\mathrm{F}}^2. \tag{3.21}$$

令 $\mathbf{U}^k = \mathbf{P}^{k\top}\mathbf{X}_s - \mathbf{Y}$，上述问题可简化为

$$\min_{m_{ij}\geq 0} \frac{1}{2}(u_{ij}^k - b_{ij}m_{ij})^2. \tag{3.22}$$

Algorithm 1 式(3.11)的优化算法

输入： 数据 $\mathbf{X}_S, \mathbf{X}_T, \mathbf{Y}$，参数 α, β, γ

初始化： $(\mathbf{Z}^0, \mathbf{Z}_1^0, \mathbf{Z}_2^0, \mathbf{E}^0, \mathbf{N}^0, \mathbf{M}^0, \mathbf{P}^0, \mathbf{\Lambda}_1^0, \mathbf{\Lambda}_2^0, \mathbf{\Lambda}_3^0), \mu = 0.1, \rho = 1.01, \mu_{\max} = 10^7, \epsilon = 10^{-7}$

迭代： $k = 1, \ldots, 200$

1: 通过式(3.14)更新 $\mathbf{Z}^{k+1}$

2: 通过式(3.16)更新 $\mathbf{Z}_1^{k+1}$

3: 通过式(3.18)更新 $\mathbf{Z}_2^{k+1}$

4: 通过式(3.19)更新 $\mathbf{E}^{k+1}$

5: 通过式(3.20)更新 $\mathbf{N}^{k+1}$

6: 通过式(3.23)更新 $\mathbf{M}^{k+1}$

7: 通过式(3.25)更新 $\mathbf{P}^{k+1}$

8: 通过式(3.26)更新 $\mathbf{\Lambda}_1^{k+1}, \mathbf{\Lambda}_2^{k+1}, \mathbf{\Lambda}_3^{k+1}$

输出： $(\mathbf{Z}, \mathbf{P}, \mathbf{E}, \mathbf{N})$

此时，最优解为 $m_{ij} = \max(u_{ij}^k b_{ij}, 0)$。因此，式(3.21)的显示解可以写为如下的形式

$$\mathbf{M}^{k+1} = \max(\mathbf{U}^k \circ \mathbf{B}, 0). \tag{3.23}$$

(7) 更新 $\mathbf{P}^{k+1}$：将其他变量固定，$\mathbf{P}^{k+1}$ 的迭代公式为

$$\begin{aligned} \min_{\mathbf{P}} \ & \frac{1}{2}\|\mathbf{P}^\top \mathbf{X}_S - \mathbf{Y} - \mathbf{B} \circ \mathbf{M}^{k+1}\|_{\mathrm{F}}^2 \\ & + \frac{\mu}{2}\|\mathbf{P}^\top \mathbf{X}_T - \mathbf{P}^\top \mathbf{X}_S \mathbf{Z}^{k+1} - \mathbf{E}^{k+1} - \mathbf{N}^{k+1} + \mathbf{\Lambda}_1^k/\mu\|_{\mathrm{F}}^2. \end{aligned} \tag{3.24}$$

令 $\mathbf{T}_3^{k+1} = \mathbf{X}_S(\mathbf{Y} + \mathbf{B} \circ \mathbf{M}^{k+1})$, $\mathbf{T}_4^{k+1} = \mathbf{X}_T - \mathbf{X}_S \mathbf{Z}^{k+1}$, $\mathbf{T}_5^{k+1} = \mathbf{E}^{k+1} + \mathbf{N}^{k+1} - \mathbf{\Lambda}_1^k/\mu$，通过化简得到 $\mathbf{P}^{k+1}$ 的最终求解表达式为

$$\mathbf{P}^{k+1} = (\mathbf{X}_S \mathbf{X}_S^\top + \mu \mathbf{T}_4^{k+1} \mathbf{T}_4^{(k+1)\top} + \lambda \mathbf{I})^{-1} (\mathbf{T}_3^{(k+1)\top} + \mu \mathbf{T}_4^{k+1} \mathbf{T}_5^{(k+1)\top}). \tag{3.25}$$

(8) 更新 $\mathbf{\Lambda}_1^{k+1}, \mathbf{\Lambda}_2^{k+1}, \mathbf{\Lambda}_3^{k+1}$：当其它所有的原始变量都更新完毕后，可以得到拉格

朗日乘子的迭代公式为

$$\begin{cases} \mathbf{\Lambda}_1^{k+1} = \mathbf{\Lambda}_1^k + \mu(\mathbf{P}^{(k+1)\top}\mathbf{X}_t - \mathbf{P}^{(k+1)\top}\mathbf{X}_s\mathbf{Z}^{k+1} - \mathbf{E}^{k+1} - \mathbf{N}^{k+1}), \\ \mathbf{\Lambda}_2^{k+1} = \mathbf{\Lambda}_2^k + \mu(\mathbf{Z}^{k+1} - \mathbf{Z}_1^{k+1}), \\ \mathbf{\Lambda}_3^{k+1} = \mathbf{\Lambda}_3^k + \mu(\mathbf{Z}^{k+1} - \mathbf{Z}_2^{k+1}). \end{cases} \tag{3.26}$$

最后我们在算法1中给出了问题(3.11)的完整求解过程。

3.4 基于松弛回归矩阵改进的子空间学习

事实上，TSL-1 通过引入了判别子空间学习函数来保证子空间变换前后源域数据标签特征的稳定，同时也为变换矩阵 **P** 提高了自由度。但是该方法是在严格的二进制标签矩阵 **Y** 的基础上迭代的，提供的自由度有限，也就是说矩阵 **P** 容易被独热码标签矩阵 **Y** 限制。本节将通过引入松弛回归矩阵 **R** 来放松子空间变换矩阵 **P** 的限制，以提高模型的灵活性和判别能力。

3.4.1 数学模型

注意到，式(3.6)的作用是为了保证源域数据标签特征在变换后不被丢失。但是，该式中的 **Y** 是独热码形式的标签矩阵，对矩阵 **P** 的限制过于严格。我们期望通过迭代能够找到一个灵活的 **P**，一方面它可以尽可能地扩大不同类之间的边界，另一方面能够最小化两个域分布之间的差异。

受 Jin 等人[88]的启发，我们引入了一个松弛回归矩阵 **R**，它从数据中直接学习标签特征来放松二进制标签矩阵 **Y** 的限制，为 **P** 提供更高的自由度。下面将通过一个简单的例子来具体说明。假设有三个训练样本 $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ 分别来自第三、第一和第二类，此时可以得到其标签矩阵的独热码形式为

$$\mathbf{Y} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \tag{3.27}$$

这里列表示标签，行表示样本。

当任意两个来自不同类别的样本投影到其标签空间时，它们之间的欧式距离为 $\sqrt{2}$，此时样本距离固定矩阵 **P** 的灵活度受限。我们的目标是对于不同类的距离尽可

能扩大，而不是固定为一个常数。为此引入一个松弛回归矩阵来代替 $\mathbf{Y}$，在经过更新后该矩阵可以简单写为

$$\mathbf{R}=\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{32} \\ r_{31} & r_{32} & r_{33} \end{bmatrix},\ r_{ij}=\begin{cases} >0,\ ij=iy_i, \\ \leq 0,\ ij\neq iy_i, \end{cases} \tag{3.28}$$
$$r_{iy_i}-\max_{j\neq y_i} r_{ij}\geq 1,\ i=1,...,n,$$

其中 iy_i 可以理解为真实标签所在的位置。此时，第一个样本和第二个样本之间的距离可以写为

$$\sqrt{(r_{21}-r_{11})^2+(r_{22}-r_{12})^2+(r_{32}-r_{13})^2}\geq\sqrt{2}. \tag{3.29}$$

相对于式(3.6)，这里的 $\mathbf{R}$ 不再受限于二进制标签矩阵 $\mathbf{Y}$ 的 0 和 1，它将从源域数据变换矩阵 $\mathbf{P}^\top\mathbf{X}_S$ 中直接学习标签特征，因此具有更高的自由度。

于是，我们在 TSL-1 的基础上进一步改进，引入了松弛回归矩阵 $\mathbf{R}\in\mathbb{R}^{n_s\times c}$，最终得到了 TSL-2 的数学模型为

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E},\mathbf{N},\mathbf{R},\mathbf{P}}\ & \frac{1}{2}\|\mathbf{P}^\top\mathbf{X}_S-\mathbf{R}\|_{\mathrm{F}}^2+\|\mathbf{Z}\|_*+\alpha\|\mathbf{Z}\|_1+\beta\|\mathbf{E}\|_1+\gamma\|\mathbf{N}\|_{\mathrm{F}}^2 \\ \text{s.t.}\quad & \mathbf{P}^\top\mathbf{X}_T=\mathbf{P}^\top\mathbf{X}_S\mathbf{Z}+\mathbf{E}+\mathbf{N},\ r_{it}-\max_{j\neq t} r_{ij}\geq 1,\ i=1,\ldots,n_S, \end{aligned} \tag{3.30}$$

其中 t 为样本 $\mathbf{x}_i$ 的真实索引此外，为了使算法能够学习到正确的松弛回归矩阵 $\mathbf{R}$，公共子空间的维数 d 等于分类的种类数 c。

3.4.2 优化算法

本节将对式(3.30)中提出的数学模型进行求解。与 TSL-1 求解的区别在于 $\mathbf{R}$ 与 $\mathbf{P}$ 的不同，两个问题中相同的求解步骤我们将不再赘述。式(3.30)的增广拉格朗日函数

可以写为

$$\begin{aligned}
&\mathcal{L}_{\mu}(\mathbf{Z},\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{E},\mathbf{N},\mathbf{R},\mathbf{P},\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\mathbf{\Lambda}_3)\\
&=\frac{1}{2}\|\mathbf{P}^{\top}\mathbf{X}_S-\mathbf{R}\|_{\mathrm{F}}^2+\|\mathbf{Z}_1\|_*+\alpha\|\mathbf{Z}_2\|_1+\beta\|\mathbf{E}\|_1\\
&\quad+\gamma\|\mathbf{N}\|_{\mathrm{F}}^2+\frac{\mu}{2}\|\mathbf{P}^{\top}\mathbf{X}_T-\mathbf{P}^{\top}\mathbf{X}_S\mathbf{Z}-\mathbf{E}-\mathbf{N}\|_{\mathrm{F}}^2\\
&\quad+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{Z}_1\|_{\mathrm{F}}^2+\frac{\mu}{2}\|\mathbf{Z}-\mathbf{Z}_2\|_{\mathrm{F}}^2\\
&\quad+\langle\mathbf{\Lambda}_1,\mathbf{P}^{\top}\mathbf{X}_t-\mathbf{P}^{\top}\mathbf{X}_S\mathbf{Z}-\mathbf{E}-\mathbf{N}\rangle\\
&\quad+\langle\mathbf{\Lambda}_2,\mathbf{Z}-\mathbf{Z}_1\rangle+\langle\mathbf{\Lambda}_3,\mathbf{Z}-\mathbf{Z}_2\rangle,
\end{aligned}\tag{3.31}$$

其中 $\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\mathbf{\Lambda}_3$ 是拉格朗日乘子，$\mu\geq 0$ 是惩罚参数。接下来我们将对 $\mathbf{R}$ 与 $\mathbf{P}$ 的求解进行介绍。

(1) 更新 $\mathbf{R}^{k+1}$：通过固定其他变量并进行化简，可以得到 $\mathbf{R}^{k+1}$ 的表达式为

$$\begin{aligned}
&\min_{\mathbf{R}}\ \frac{1}{2}\|\mathbf{P}^{k\top}\mathbf{X}_S-\mathbf{R}\|_{\mathrm{F}}^2\\
&\text{s.t. }\ r_{iy_i}-\max_{j\neq y_i}r_{ij}\geq 1,\ i=1,...,n.
\end{aligned}\tag{3.32}$$

令 $\mathbf{Q}=\mathbf{P}^{k\top}\mathbf{X}_S$，上式可以写为

$$\begin{aligned}
&\min_{\mathbf{r}_i}\ \frac{1}{2}\|\mathbf{q}_i-\mathbf{r}_i\|_2^2\\
&\text{s.t. }\ r_{iy_i}-\max_{j\neq y_i}r_{ij}\geq 1,
\end{aligned}\tag{3.33}$$

其中 $\mathbf{q}_i$ 和 $\mathbf{r}_i$ 是矩阵 $\mathbf{Q}$ 和 $\mathbf{R}$ 的第 i 行。令 $r_{iy_i}=q_{iy_i}+\eta$，其中 η 是一个可以被优化的参数。上述目标可以被改写为 $\sum_{j=1}^{c}(q_{ij}-r_{ij})^2$，其中 c 是样本的种类，进而得到式(3.33)的另一种形式

$$\begin{aligned}
&\min_{r_{ij}}\ \frac{1}{2}(q_{ij}-r_{ij})_2^2\\
&\text{s.t. }\ q_{iy_i}+\eta-r_{ij}\geq 1,\ \forall j\neq y_i.
\end{aligned}\tag{3.34}$$

引入新的变量 d，将 $\mathbf{Q}$ 中的第 ij 个元素表示为 $d_j=q_{ij}+1-q_{iy_i},\ \forall j\neq y_i$，其中 $d_j\leq 0$ 和 $d_j>0$ 分别表示违反和满足第 i 个样本第 y_i 类的边际约束。r_{ij} 可以通过下面的式子进行计算

$$r_{ij}=\begin{cases}q_{ij}+\eta, & j=y_i,\\ q_{ij}+\min\left(\eta-d_j,0\right), & \text{其他}.\end{cases}\tag{3.35}$$

通过上式可以看到 r_{ij} 可以用 η 进行代替，于是 $\mathbf{r}_i$ 的求解可以通过参数 η 的求解来完成。因此式(3.33)可以改写为

$$\min_{\eta} f(\eta) = \eta^2 + \sum_{j \neq y_i} \left(\min\left(\eta - d_j, 0\right)\right)^2, \tag{3.36}$$

经过化简可以得到 η 的解为

$$\eta = \frac{\sum_{j \neq y_i} d_j \mathrm{g}\left(f'\left(d_j\right) > 0\right)}{1 + \sum_{j \neq y_i} \mathrm{g}\left(f'\left(d_j\right) > 0\right)}, \tag{3.37}$$

其中 $\mathrm{g}(\cdot)$ 表示指示符算子，即如果 $f'\left(d_j\right) > 0$，则 $\mathrm{g}(\cdot) = 1$，反之 $\mathrm{g}(\cdot) = 0$。算法2给出了求解 $\mathbf{R}$ 的详细步骤。

Algorithm 2 式(3.33)的求解

输入： $\mathbf{p}_i, \mathbf{y}_i$

初始化： $d_j = q_{ij} + 1 - q_{iy_i}, \forall i = 1, 2, 3, ..., n, \eta = 0, iter = 0$

循环 $j = 1$ to $c(j \neq y_i)$

如果 $f'\left(d_j\right) > 0$

$\eta = \eta + d_j, iter = iter + 1$

$\eta = \eta / (iter + 1)$

通过式(3.35)更新 $\mathbf{r}_i = [r_{i1}, r_{i2}, ..., r_{ij}]$

输出： $\mathbf{R}$

(2) 更新 $\mathbf{P}^{k+1}$：将其他变量固定，通过化简得到下式

$$\begin{aligned} \min_{\mathbf{P}} \ & \frac{1}{2}\|\mathbf{P}^\top \mathbf{X}_S - \mathbf{R}^{k+1}\|_{\mathrm{F}}^2 \\ & + \frac{\mu}{2}\|\mathbf{P}^\top \mathbf{X}_T - \mathbf{P}^\top \mathbf{X}_S \mathbf{Z}^{k+1} - \mathbf{E}^{k+1} - \mathbf{N}^{k+1} + \mathbf{\Lambda}_1^k / \mu\|_{\mathrm{F}}^2. \end{aligned} \tag{3.38}$$

令 $\mathbf{T}_7^{k+1} = \mathbf{X}_S \mathbf{R}^{k+1}, \mathbf{T}_8^{k+1} = \mathbf{X}_T - \mathbf{X}_S \mathbf{Z}^{k+1}$，$\mathbf{T}_9^{k+1} = \mathbf{E}^{k+1} + \mathbf{N}^{k+1} - \mathbf{\Lambda}_1^k / \mu$，得到

$$\mathbf{P}^{k+1} = (\mathbf{X}_S \mathbf{X}_S^\top + \mu \mathbf{T}_8^{k+1} \mathbf{T}_8^{(k+1)\top} + \lambda \mathbf{I})^{-1} (\mathbf{T}_7^{(k+1)\top} + \mu \mathbf{T}_8^{k+1} \mathbf{T}_9^{(k+1)\top}), \tag{3.39}$$

综上，在算法3中给出了式(3.30)的完整求解过程。

Algorithm 3 式(3.30)的优化算法

输入： 数据 $\mathbf{X}_S, \mathbf{X}_T, \mathbf{Y}$，参数 α, β, γ

初始化： $(\mathbf{Z}^0, \mathbf{Z}_1^0, \mathbf{Z}_2^0, \mathbf{E}^0, \mathbf{N}^0, \mathbf{R}^0, \mathbf{P}^0, \mathbf{\Lambda}_1^0, \mathbf{\Lambda}_2^0, \mathbf{\Lambda}_3^0)$, $\mu = 0.1$, $\rho = 1.01$, $\mu_{\max} = 10^7$, $\epsilon = 10^{-7}$

迭代： $k = 1, \ldots, 200$

1: 通过式(3.14)更新 $\mathbf{Z}^{k+1}$

2: 通过式(3.16)更新 $\mathbf{Z}_1^{k+1}$

3: 通过式(3.18)更新 $\mathbf{Z}_2^{k+1}$

4: 通过式(3.19)更新 $\mathbf{E}^{k+1}$

5: 通过式(3.20)更新 $\mathbf{N}^{k+1}$

6: 通过算法2更新 $\mathbf{R}^{k+1}$

7: 通过式(3.39)更新 $\mathbf{P}^{k+1}$

8: 通过式(3.26)更新 $\mathbf{\Lambda}_1^{k+1}, \mathbf{\Lambda}_2^{k+1}, \mathbf{\Lambda}_3^{k+1}$

输出： $(\mathbf{Z}, \mathbf{P}, \mathbf{E}, \mathbf{N})$

3.5 数值实验

本节通过跨域故障诊断的实验来评估所提出的 TSL-1 与 TSL-2。我们首先给出了实验的实施细节，然后对使用的两个轴承故障诊断数据集进行介绍，并给出了数据预处理手段和特征提取的方法，最后讨论了算法的噪声鲁棒性、参数敏感性、模型稳定性和收敛性。

3.5.1 实施细节

(1) 对比的方法

在实验部分，本文提出的两种方法将会与各类具有代表性的迁移学习方法进行比较，包括 GFK[49]，TCA[45]，JDA[46]，TJM[47]，BDA[48]，LRSR[50] 以及 RLCSL[82]。最后，我们还选择了 k-最近邻 (K Nearest Neighbor，KNN) 作为基线分类器完成迁移后的分类任务。

(2) 参数设置

在本实验中，将模型参数 α, β, γ 在 $\{10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3\}$ 的范围内进行搜索。对于增广拉格朗日函数式(3.12)中的参数 μ，选择 $\mu = 0.1$ 为初始值，然

后通过式 $\mu = \min(\rho\mu, \mu_{\max})$ 进行更新，其中 $\rho = 1.01$，$\mu_{\max} = 10^7$。

(3) 停机准则

对于我们提出的方法，设置其收敛条件为

$$\begin{cases} \delta_1 = \|\mathbf{P}^\top \mathbf{X}_T - \mathbf{P}^\top \mathbf{X}_S \mathbf{Z} - \mathbf{E} - \mathbf{N}\|_\infty < \epsilon, \\ \delta_2 = \|\mathbf{Z} - \mathbf{Z}_1\|_\infty < \epsilon, \\ \delta_3 = \|\mathbf{Z} - \mathbf{Z}_2\|_\infty < \epsilon, \end{cases} \tag{3.40}$$

其中，$\epsilon = 10^{-7}$。此外，最大迭代次数设置为 200。

(4) 评价指标

为了衡量分类精度，使用以下指标进行评价

$$\text{Acc} = \frac{\text{number of samples } (y_{pre} = y_T)}{\text{total of samples } (y_T)} \times 100\%, \tag{3.41}$$

其中，y_{pre} 是通过将转换后的目标样本输入 KNN 后得到的。KNN 通过转换后的源域样本进行训练，y_T 表示目标样本的真实标签。

3.5.2 实验数据

在本次实验中，我们使用西储大学 (Case Western Reserve University，CWRU) 和江南大学 (Jiang Nan University，JNU) 的轴承故障诊断数据集，下面将介绍两个轴承故障数据集、数据预处理的方法以及迁移学习任务划分的具体方式。

(1) CWRU 数据集

CWRU 数据集[①]是轴承故障诊断任务中常用的数据集之一，该数据集是将加速度传感器置于轴承的不同位置采集振动信号并存储。图3.3展示了该数据集采集数据的实验平台，电机在 4 种不同负载下 (1730，1750，1772，1797r/min) 运行，4 种负载分别对应 4 种不同工况。每种工况下都存在共包含 3 种故障类型内圈故障 (IF)、外圈故障 (OF) 和滚珠圈故障 (RF) 以及健康状态 (NB) 下采集的信号。此外，在 IF、RF 和 OF 区域存在单点故障，故障直径分别为 7mil、14mil 和 21mil，因此总共可以分为 10 类 (1 种健康状态和 9 种故障状态)，表3.1对 10 种类型进行了详细的说明。我们使用 12kHz 采样频率下采集的驱动端的故障数据作为实验数据。每种电机负载对

① https://engineering.case.edu/bearingdatacenter/download-data-file

应一个数据集，针对 CWRU 共有 4 个数据集 (S0，S1，S2 和 S3)。任务 S0→S2 表示源域 S0 向目标域 S2 的迁移，我们总共得到了 12 个迁移学习任务。

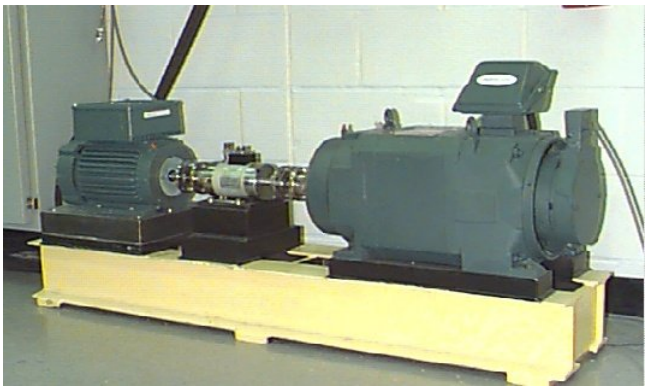


图 3.3　CWRU 数据集的实验平台

Figure 3.3　Experimental platform for the CWRU dataset

表 3.1　CWRU 数据集的 10 种故障类型

Table 3.1　10 fault types for the CWRU dataset

标签	1	2	3	4	5
故障位置	NB	IF	RF	OF	IF
故障尺寸 (mils)	0	7	7	7	14
标签	6	7	8	9	10
故障位置	RF	OF	IF	RF	OF
故障尺寸 (mils)	14	14	21	21	21

(2) JNU 数据集

JNU 数据集[①]是来自江南大学的一个轴承故障基准数据集，该数据集广泛应用于故障诊断任务中。数据的获取方式是将加速度传感器置于轴承的不同位置，采集振动信号并存储。图3.4展示了该数据集采集数据的实验平台。不同于 CWRU 数据集，JNU 数据集只包含 4 种故障类型内圈故障 (IF)、外圈故障 (OF) 和滚珠圈故障 (RF) 以及健康状态 (NB)。JNU 数据集使用 50kHz 采样频率采集的故障数据，分别包含了 600rpm、800rpm 和 1000rpm 转速下采集的故障数据。同样地，每种转速对应 1 个数据集，总共包含 3 个数据集 (T0，T1，T2) 和 6 个迁移学习任务。

① https://github.com/ClarkGableWang/JNU-Bearing-Dataset

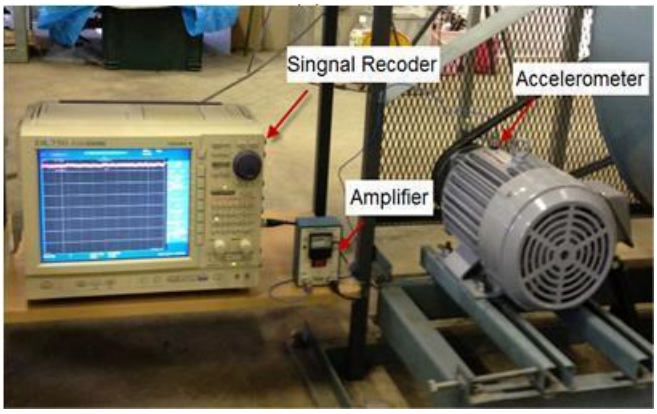


图 3.4　JNU 数据集的实验平台

Figure 3.4　Experimental platform for the JNU dataset

(3) 数据预处理

与实验室实验数据相比，工业数据具有数量庞大、低价值密度、多源异构和监测数据连续流四个关键特征。在数据驱动的故障诊断范式中，训练数据的数量主要影响最终诊断模型的性能。大量的干净数据对于提高诊断模型的性能至关重要，因此数据预处理是至关重要的一步。轴承故障数据集包含大量复杂的振动信号，如果直接将原始的振动信号作为诊断信号，则数据量太大无法执行矩阵运算。如果在时频域内分析振动信号，则可能导致数据中部分有用信息被消除。在信号处理领域，一种二维图片化的方法被广泛使用，这种方法能将成千上万的一维振动信号转化为二维图像，提取出一维数据中隐藏的重要特征和信息，有助于特征分析和深度学习模型的训练等任务。相对于一维信号，二维图像包含更多有价值的信息，能够提供数据更直观的视角，对数据有更好的解释性。

受 Wen 等人[89]的启发，我们将上述提到的 CWRU 和 JNU 数据集中的振动信号转换为像素为 128*128 灰度像素图像，以便于迁移子空间学习方法的应用。此时在 CWRU 数据集上转换后的 4 个工况对应的图像数据集为 (SP0，SP1，SP2，SP3)，JNU 数据集上转换后的 3 个工况对应的图像数据集为 (JP0，JP1，JP2)。图3.5展示了 CWRU 在 SP0 中 10 种不同故障类型的图像，图3.6展示了 JNU 在 JP0 中 4 种不同故障类型的图像。观察到转换后的不同故障类型的图像完全不同，这为我们提供了一种直观的方法进行分类。

迁移子空间学习方法要求输入的样本是矩阵，因此为了获得转换后二维图像的特征矩阵，我们利用卷积神经网络 VGG19 对图像特征进行提取。VGG 网络是牛津大学 Visual Geometry Group 研究小组提出的一个卷积网络模型。它与一般卷积神经网

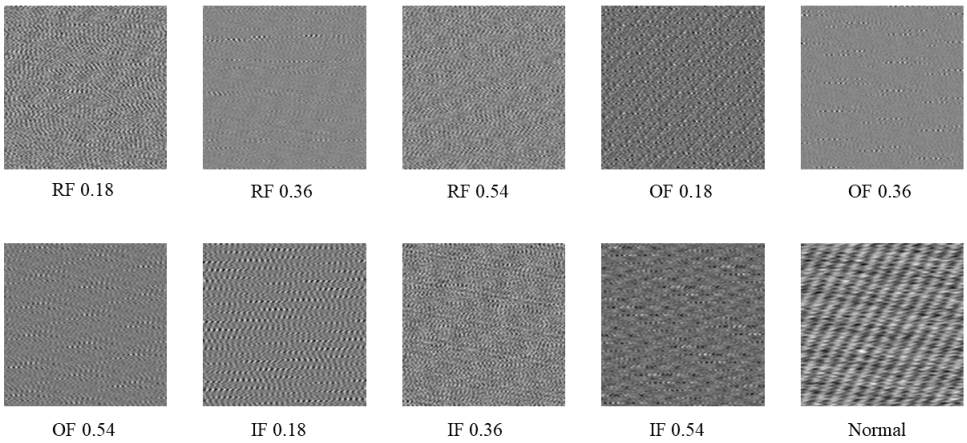


图 3.5　CWRU 数据集的 10 种故障类型示例

Figure 3.5　Examples of 10 fault types for the CWRU dataset

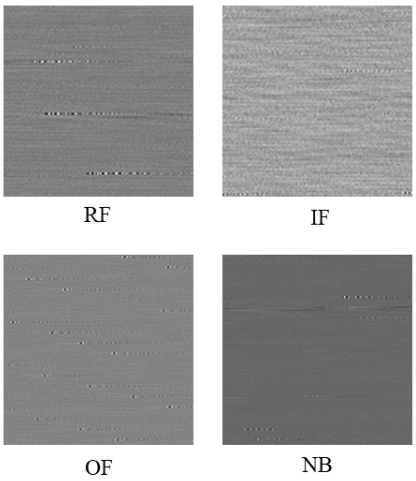


图 3.6　JNU 数据集的 4 种故障类型示例

Figure 3.6　Examples of 4 fault types for the JNU dataset

络的区别是，在 VGG 中使用了 3 个 3*3 卷积核来代替 7*7 卷积核，使用了 2 个 3*3 卷积核来代替 5*5 卷积核，即使用连续的几个较小的卷积核代替较大卷积核。这样就能使得在保证具有相同感知野的条件下，不仅提升了网络的深度，而且具有更少的网络参数减少了训练成本。此外，这种多层的、小尺寸的卷积层的设计能够提升卷积层特征提取的能力。图3.7展示了本节使用的 VGG 网络的架构，同时在表3.2中给出了网络各层的参数。VGG 提取特征将每张图像转换为有 4096 个特征点的一维向量，此时矩阵的列为图像的特征维度，矩阵的行为图像的数量。针对 CWRU 数据集的 4 种工况，获得了可用于迁移子空间学习的 4 个样本矩阵：SF0、SF1、SF2 和

SF3。对于 JNU 数据集的 4 种工况，同样获得了 3 个样本矩阵：TF0、TF1 和 TF2。我们将样本的特征矩阵以 mat 的格式存储，方便后续的算法迭代和矩阵运算。

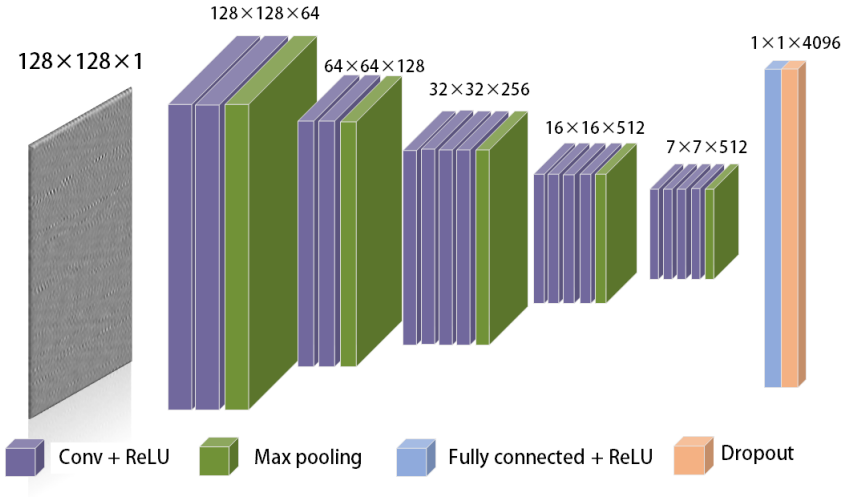


图 3.7　所使用的 VGG19 网络架构

Figure 3.7　Architecture of the VGG19 network

表 3.2　VGG19 网络的参数

Table 3.2　Parameters of the VGG19 network

网络层	参数
Conv1/Conv2	out_channels=64, kernel_size=5
MaxPool1	kernel_size=3, stride=2, dilation=1
Conv3/Conv4	out_channels=128, kernel_size=3
MaxPool2	kernel_size=2, stride=2, dilation=1
Conv5/Conv6/Conv7/Conv8	out_channels=256, kernel_size=3
MaxPool3	kernel_size=2, stride=2, dilation=1
Conv9/Conv10/Conv11/Conv12	out_channels=512, kernel_size=3
MaxPool4	kernel_size=2, stride=2, dilation=1
Conv13/Conv14/Conv15/Conv16	out_channels=512, kernel_size=3
MaxPool5	kernel_size=2, stride=2, dilation=1
Fully connected	out_feature=4096
Dropout	p=0.5

(4) 数据集划分

在实际应用中，通常利用实验室已知工况下的故障数据来预测未知工况下没有标签数据的故障类别。因此，我们将不同工况下的数据划分为含有样本标签的源域和不包含标签信息的目标域，来进行无监督迁移学习模拟实际工况。对于 CWRU 数据集，根据负载速度获得了 4 个不同工作条件下的数据集，即 SF0、SF1、SF2 和 SF3。对于 JNU 数据集，获得了 3 个不同工作条件下的数据集，即 TF0、TF1 和 TF2。每个数据集都可以被看作是源域或目标域。在优化过程中，我们的模型训练源域的数据矩阵对齐目标域的矩阵。

3.5.3 实验分析

(1) 实验结果

表3.3展示了所有比较的方法在 CWRU 数据集上的分类精度，其中加粗的数值表示最优结果，加下划线的数值为次优的结果。结果表明，TSL-2 方法对于无监督跨域故障诊断任务的性能优于其他 7 种迁移学习方法。TSL-1 在 12 个任务上的平均准确率为 62.30%，与 BDA 方法相比效果提升了 1.57%，这表明引入的高斯噪声矩阵能够减少数据中噪声的干扰。TSL-2 在 12 个任务上的平均准确率为 65.30%，与 RLCSL 相比性能提升了 1.77%。此外，与 TSL-1 相比，TSL-2 在大多数任务中都取得了最高的分类准确度。这表明引入的可学习的标签学习矩阵能够维持数据结构的稳定，提供更高的自由度。上述结果验证了所提出的 TSL-1 和 TSL-2 可以为跨工况故障诊断任务构建更有效、更稳健的表示。同时，我们也注意到在某些任务中迁移子空间学习方法表现不佳，这可能是由于源域样本和目标域的特征矩阵无法通过线性变换完成重建导致的。

表3.4显示了所有比较方法在 JNU 数据集上的分类准确率。可以观察到，所提出的方法在大多数任务中都取得了高的分类准确率。从所有任务的平均值来看，TSL-1 和 TSL-2 方法对于跨域无监督故障诊断的效果远远优于其他方法，这说明了去除高斯噪声和为变换矩阵提供自由度的重要性。具体来说，与 LRSR 和 RLCSL 相比，TSL-2 方法的平均分类精度分别提高了 6.47% 和 2.21%。与其他迁移学习方法相比，TSL-1 和 TSL-2 方法有更好的鲁棒性和更强的适应性。

表 3.3 不同方法在 CWRU 数据集上的准确率 (%)

Table 3.3 Accuracy of different methods on the CWRU dataset

迁移任务	方法									
	KNN	GFK	TCA	JDA	TJM	BDA	RLCSL	LRSR	TSL-1	TSL-2
SF0 → SF1	66.61	69.20	55.70	71.79	69.64	70.27	68.95	64.73	70.18	**73.21**
SF0 → SF2	67.61	68.83	64.60	64.78	74.01	72.13	70.69	69.21	74.34	**74.95**
SF0 → SF3	43.77	48.99	35.9	50.82	**53.11**	52.29	47.45	43.96	43.77	51.92
SF1 → SF0	70.64	72.83	69.60	72.65	73.21	72.83	72.45	71.02	69.97	**74.26**
SF1 → SF2	61.68	60.17	55.60	53.58	48.87	57.16	64.33	61.02	65.07	**67.33**
SF1 → SF3	51.74	48.63	40.70	44.30	48.62	48.90	59.94	55.77	53.30	**61.72**
SF2 → SF0	72.93	72.93	71.7	73.12	76.59	74.83	76.84	74.17	74.27	**78.74**
SF2 → SF1	64.46	**68.04**	60.90	64.55	66.16	64.64	66.95	65.09	64.55	67.14
SF2 → SF3	56.50	58.06	46.80	54.30	53.20	56.78	61.85	62.73	**63.10**	62.36
SF3 → SF0	47.38	51.38	42.10	**61.18**	53.19	54.62	47.41	44.43	52.24	49.29
SF3 → SF1	47.59	49.55	37.90	55.54	51.78	52.32	57.64	55.36	59.46	**63.75**
SF3 → SF2	50.09	50.66	39.60	52.17	52.82	51.98	55.82	56.03	57.34	**58.95**
Average	58.42	59.94	51.76	59.90	60.10	60.73	62.53	60.29	62.30	**65.30**

表 3.4 不同方法在 JNU 数据集上的准确率 (%)

Table 3.4 Accuracy of different methods on the JNU dataset

迁移任务	方法									
	KNN	GFK	TCA	JDA	TJM	BDA	RLCSL	LRSR	TSL-1	TSL-2
TF0 → TF1	72.41	81.61	70.54	71.88	76.87	73.04	**83.69**	75.98	82.50	83.32
TF0 → TF2	74.63	71.54	72.79	66.54	66.54	71.76	87.34	74.85	86.54	**93.38**
TF1 → TF0	69.00	**80.33**	70.08	75.05	74.00	76.33	61.94	63.33	65.92	65.25
TF1 → TF2	88.24	72.87	65.00	68.97	69.34	80.59	84.96	77.87	83.34	**92.28**
TF2 → TF0	71.17	75.00	69.00	74.67	**76.33**	75.25	73.41	73.35	72.58	73.50
TF2 → TF1	83.75	78.93	67.95	71.34	73.39	71.79	85.46	85.98	**86.96**	85.34
Average	76.53	75.73	69.23	71.44	72.75	74.79	79.47	75.21	79.64	**81.68**

(2) 噪声鲁棒性分析

本小节讨论了在不同数据集中加入不同程度的高斯噪声和非高斯噪声对不同方法准确率的影响。表3.5展示了在 CWRU 数据集中加入 0%, 30%, 60% 和 90% 不同程度的高斯噪声后，不同方法在 SF1→SF0 迁移任务中的准确率。根据结果可以观察到，所提出的 TSL-2 方法在所有噪声水平下表现均为最好，其次是 TSL-1。此外，随着噪声水平的增加，所有迁移子空间学习的性能都在下降，但是提出的 TSL-1 和 TSL-2 方法性能相对稳定，表现出对高斯噪声的鲁棒性，这说明了在模型中引入矩阵建模高斯噪声能够有效抑制数据中高斯噪声的影响，提高模型的抗干扰能力。

表 3.5　在 CWRU 数据集上加入不同程度的噪声对迁移任务的影响 (%)

Table 3.5　Effects of adding different levels of noises on the CWRU dataset

方法	0%	30%	60%	90%
KNN	70.64	65.34	63.16	60.45
GFK	72.83	67.45	65.32	61.59
TCA	69.60	63.89	61.78	58.06
JDA	72.65	<u>70.85</u>	67.51	64.74
TJM	<u>73.21</u>	69.95	67.76	65.33
BDA	72.83	69.04	67.24	64.86
RLCSL	72.45	69.11	65.67	62.31
LRSR	71.02	68.64	65.36	62.81
TSL-1	69.97	68.52	<u>67.99</u>	<u>65.46</u>
TSL-2	**74.26**	**73.26**	**72.87**	**71.13**

此外，我们还比较了在 JNU 数据集中添加高斯噪声和非高斯噪声后对 JNU 数据集中迁移任务 TF0→TF1 的影响。图3.8展示了添加噪声后下不同方法准确率的直方图。结果表明，RLCSL 方法和 LRSR 方法在添加非高斯噪声的情况下性能表现优于添加高斯噪声的情况，这说明这两种方法不能有效抑制高斯噪声的影响。相比之下，我们提出的 TSL-1 和 TSL-2 方法，在添加噪声后性能没有明显下降，因此提出的方法对高斯噪声和非高斯噪声都有很好的鲁棒性。而其他方法在添加噪声后准确率大幅下降，无法有效抑制数据中各种噪声的影响，该实验说明在模型中建模不同类型的噪声能够有效抑制数据中噪声的影响。

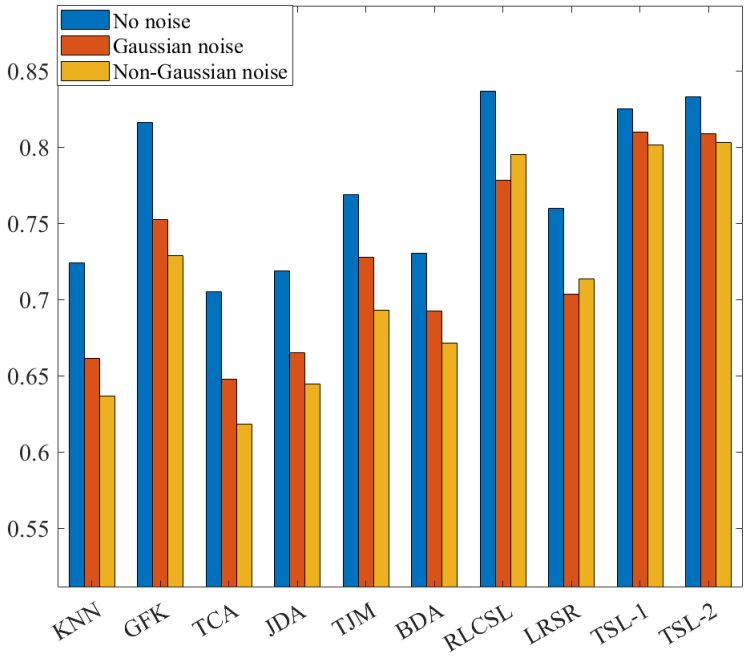


图 3.8　在 JNU 数据集中添加高斯噪声和非高斯噪声的影响

Figure 3.8　Accuracy of different methods on the JNU dataset with different noises

(3) 参数敏感性分析

对于提出的 TSL-2 方法，我们验证了该模型对参数变化的敏感性。通过在实验中分别固定参数 α, β 和 γ，其余两个参数在设定的范围内变化观察其性能。图3.9显示了 TSL-2 模型在不同参数值下的性能，其中图3.9(a)，图3.9(b)，图3.9(c)是在 CWRU 数据集上 SF0→SF1 任务下的结果三维直方图，图3.9(d)，图3.9(e)，图3.9(f)是在 CWRU 数据集上 SF2→SF0 任务下的结果三维直方图，图3.9(g)，图3.9(h)，图3.9(i)是在 JNU 数据集上 TF1→TF2 任务下的结果三维直方图。

图中的结果显示，3 个参数的变化都会对精度结果产生一定的影响，所有的任务都对参数变化敏感但是不同任务的敏感程度不同。JNU 数据集上的任务对参数变化的敏感度较小，我们猜测这是因为 JNU 数据集的数据有一定的稀疏性并且噪声含量少，因此对于不同性质的系数变化不敏感。此外，同一任务的不同的参数也会对分类精度造成影响，这说明每个跨域故障诊断任务具有不同的特征。例如，SF0→SF1 对 α 和 β 的变化敏感程度较高，这说明该任务对数据的稀疏性和其中包含的非高斯噪声都很敏感。SF2→SF0 对 γ 的变化敏感程度较高，这说明该任务对数据中包含的高斯噪声非常敏感。

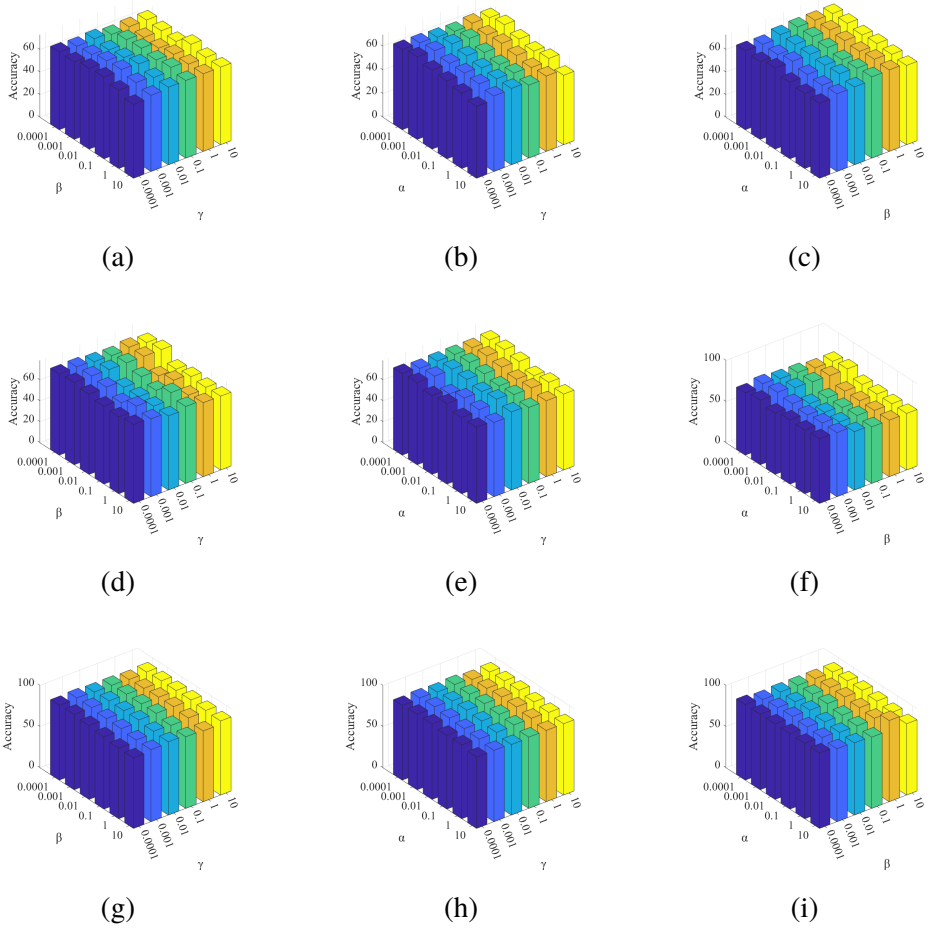


图 3.9 TSL-2 在不同参数值下的性能

Figure 3.9 Performance of TSL-2 at different parameter values

(4) 模型稳定性分析

箱形图包含最大值、最小值、中值、上四分位数和下四分位数和异常值，它们可以描述一组数据的离散度，并提供了一种观察稳定性的有效方法。图3.10展示了不同数据集不同迁移任务不同方法的箱形图，(a) CWRU 数据集 SF0→SF2 迁移任务, (b) CWRU 数据集 SF2→SF0 迁移任务, (c) JNU 数据集 TF0→TF1 迁移任务, (d) JNU 数据集 TF1→TF2 迁移任务。在图中，方框的宽度表示数据的分散度，分散度越大说明该模型对于不同任务准确率变化较大，模型的鲁棒性较低。红色的 + 表示数据中的异常值，箱线图中的红线代表数据的平均值。

通过观察我们可以得出结论，与其他方法相比 TSL-1 和 TSL-2 在大多数任务中获得了更好的精度和更小的数据离散度。这表明我们提出的方法具有更好的稳定性和更集中的分类结果。在某些任务中，KNN(不进行迁移) 的准确率优于其他方法，这

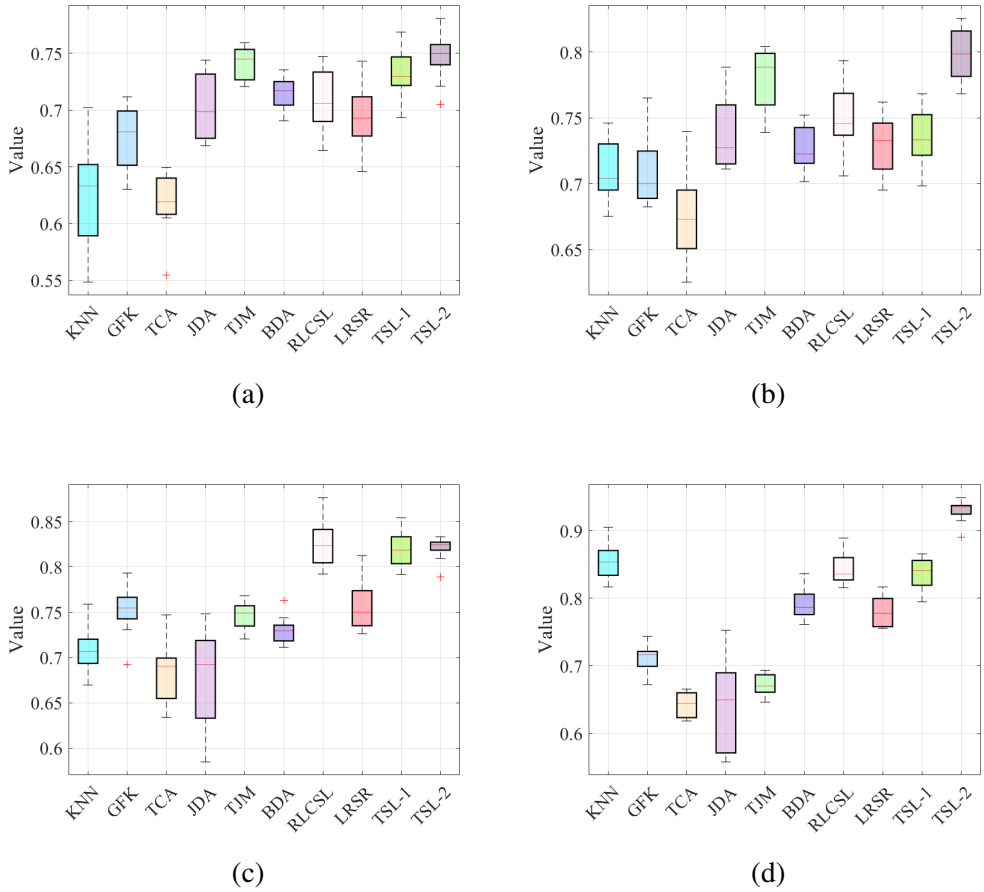


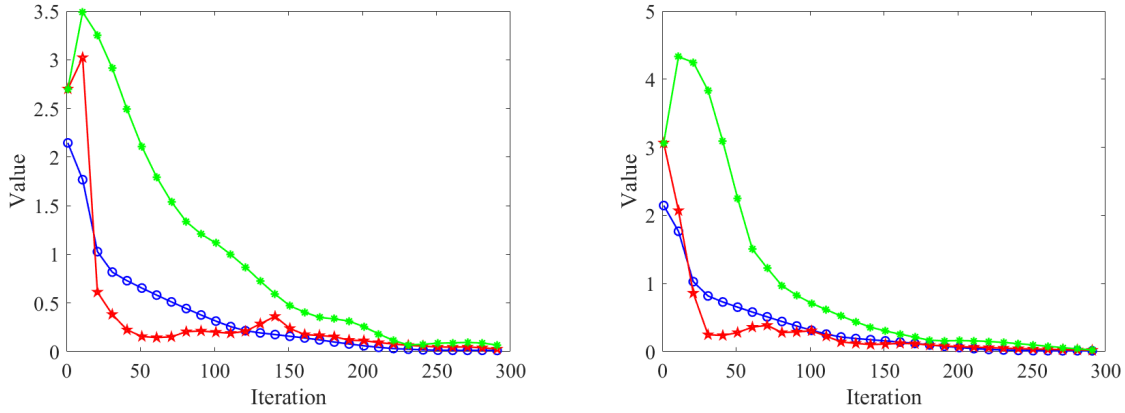
图 3.10　不同方法的模型稳定性分析

Figure 3.10　Model stability analysis for different methods

说明在子空间中从源域学习到的知识对目标域有副作用不利于迁移。与其他方法相比，TSL-2 的平均精度最好。并且在发生负迁移的情况下，TSL-2 不仅能有效抑制负迁移的发生，而且具有良好的稳定性。

(4) 收敛性验证

本小节验证了所提出的 TSL-2 模型的收敛性。图3.11分别绘制了在 JNU 数据集上 TF0→TF1 和 TF1→TF2 两个任务算法更新过程中 TSL-2 模型收敛条件的值随迭代次数的变化而变化。其中，红色的折线表示收敛条件式(3.40)中相对误差 δ_1 的值，蓝色折线表示收敛条件式(3.40)中相对误差 δ_2 的值，绿色折线表示收敛条件式(3.40)中相对误差 δ_3 的值。通过观察可以得到，随着迭代次数的增加这些相对误差的值迅速减小，在经过大约 150 次迭代后，所有相对误差的值均趋近于零，这表明模型的收敛性较强。



(a) (b)

图 3.11　收敛条件的值随着迭代次数的变化

Figure 3.11　Convergence condition versus the number of iterations on the JNU dataset

3.6　本章小结

在本章中，我们提出了用于轴承故障诊断的迁移子空间学习方法。针对轴承故障数据的特点，在模型中引入噪声矩阵来建模高斯噪声。此外，还引入了可学习的标签矩阵来对迁移子空间学习方法进行改进，增强了模型的灵活性和辨别性。这些改进不仅提高了模型的抗噪声能力，而且为子空间变换矩阵提供了更高的自由度。在数据预处理阶段，将一维振动信号转换为二维灰度像素图像并通过 VGG 网络提取特征，这种方式增加了迁移子空间方法在故障诊断任务中的可用性。与 BDA、LRSR 以及 RLCSL 等最具有代表性的迁移学习方法相比，所提出的 TSL-1 和 TSL-2 在大多数跨工况轴承故障诊断任务中表现出更理想的性能。

第四章　基于深度域适应网络的轴承故障诊断方法

本章针对迁移子空间学习线性变换的局限性以及其故障识别精度较低的问题，改进了差异度量函数，在此基础上优化了深度域适应网络框架。首先，使用多核的方法改进了判别联合概率最大均值差异 (Discriminative Joint Probability Maximum Mean Discrepancy，DJP-MMD)，优化 RKHS 的核选择提出了多核 DJP-MMD 来提高分布差异度量的精确度。其次，将多核 DJP-MMD 度量准则嵌入到神经网络框架中，有效提高了深度域适应网络的可迁移性和可鉴别性，提出了基于深度联合概率适应网络 (Deep Joint Probability Adaptation Network，DJPAN) 的智能故障诊断方法。最后，将振动信号进行图像化构建了故障诊断的图像数据集，增加了深度域适应网络在智能故障诊断领域的可用性，完成了目标域没有标签情况下的跨工况故障诊断任务。该方法的 Python 代码可以在https://github.com/FuchaoYu/DJPAN进行下载。

4.1　引言

传统的迁移学习方法通过学习不变的特征表示或估计实例的重要性来连接源域和目标域，这个过程通常涉及到特征提取、领域适应以及分类三个步骤。随着深度学习的发展，研究者们发现神经网络可以学习更多可迁移的特征来完成域适应[52]。这种端到端的深度迁移学习方法减少了误差在中间过程的积累，构建了图像与类别的关系映射。基于深度迁移学习的故障诊断技术避免了耗时且不可靠的人工分析，吸引了越来越多的关注。其中一种直接的深度迁移学习方法是利用目标域内的标记数据对模型进行微调，但是新收集的数据或不同工作条件下的数据通常是没有标记的，因此这种微调的深度迁移学习方法在实际应用中无法大规模使用。

近年来，在神经网络中嵌入域适应层来构建深度域适应网络的方法在轴承故障诊断中引起了广泛关注。目前的研究主要是集中在无监督域适应下的智能故障诊断问题，即新收集的目标域中的测试数据是未标记的，并且假设源域与目标域具有相同的标签空间。Zhao 等人[90]和 Guo 等人[91] 应用 MMD 测量源域轴承数据集和目标域轴承数据集的分布差异，不过他们并未考虑内部协变量偏移的影响，仅在全连接层中对源域和目标域数据分布进行了对齐。Yang 等人[92]提出了一种多层域适应轴承

故障诊断模型，通过在网络各层中应用 MMD 来最小化源域和目标域数据的分布差异。上述方法的目标是减少源域和目标域数据的边缘分布，没有考虑两个域之间的条件分布。然而，以往的研究表明，同时最小化边缘分布和条件分布对构建稳健的迁移学习模型至关重要。基于这一观点，Wu 等人[93]将 JDA 应用于轴承故障诊断。数据集上的测试结果表明，最小化联合分布可以在有少量故障数据的情况下实现有效的故障诊断。Cao 等人[94]发现，忽略类权重差异会导致基于 MMD 的方法性能下降，因此他们在 MMD 中引入了类概率惩罚项，以解决类别不平衡问题。Li 等人[95]提出了一种强化集合深度迁移学习的方法，利用各种内核 MMD 学习不同来源的故障特征。上述各类基于深度迁移学习的故障诊断方法旨在学习两个域的不变特征，忽略了模型在目标域上的泛化，因此，Zhang 等人[96]提出使用 MMD 损失最小化全域分布差异，并使用监督对比学习损失来实现类对齐。深度域适应网络由于其优越的性能在各领域都引起的学者们广泛的关注和讨论，但是也存在着一定的问题。具体来说，上述各类方法只是将不同领域之间的差异最小化，对齐源域分布和目标域的分布，却忽略了不同类之间的差异。因此，所有来自源域和目标域的数据都会被混淆，数据的可鉴别性会进一步降低。如果能够在整体对齐源域分布和目标域的分布的基础上，增加不同类之间的距离，那模型的迁移性和可鉴别性都能得到保证。

改进后的网络通过增加不同类之间的距离，扩展了深度域适应网络的特征表示能力。大多数前馈网络模型都可以使用该方法，并且可以使用标准的反向传播方法进行高效训练。我们在图4.1中直观展示了判别联合概率域适应与全局域适应的区别，左边的例子展示了全局域适应在迁移的过程中全局对齐源域和目标域的样本，没有考虑将样本中不同的类加以区分。右边的例子展示了判别联合概率域适应在迁移过程中不仅能对齐源域和目标域的样本，而且能够增加样本中不同类之间的距离，使来自不同域的同一类更一致，不同的类更分离，增强网络的分类性能。

综上，本章的主要工作如下：

(1) 利用多个核进行线性组合构造新的核函数来优化 DJP-MMD 中 RKHS 的选择，提高了分布差异度量的精确度。

(2) 将改进后的 DJP-MMD 嵌入神经网络的自适应层，增加了深度域适应网络的迁移性和鉴别性。

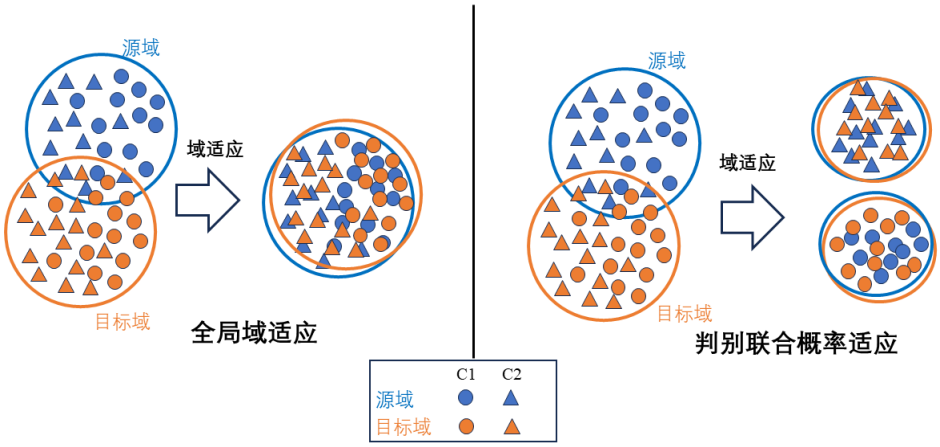


图 4.1 判别联合概率域适应与全局域适应的区别

Figure 4.1 Characterization of discriminative joint probability domain adaptation

4.2 差异度量相关工作

在第二章中介绍过深度迁移学习的发展近况，暹罗网络架构由于其网络移植简单、部署性强的特点近年来得到了广泛的关注。这种方法将前馈网络模型的分类损失和迁移损失相结合，并在反向传播中优化来进行高效训练，从而实现深度域适应网络的设计。不难看出，深度域适应网络的关键就是迁移损失函数的设计，迁移损失函数的好坏直接关系到网络迁移性能的好坏。研究者们设计了大量的差异度量函数来完成深度域适应网络的设计，本节将对部分差异度量函数进行介绍。

4.2.1 最大均值差异

MMD 是迁移学习中广泛使用的一种损失函数，主要用于判断两个分布是否相同。对于来自两个分布的样本，有很多标准 (如 Kullback–Leibler 散度等) 都可以对它们之间的距离进行估计，但是这些标准大多是参数化的或者需要中间密度的估计量来完成距离度量。Gretton 等人[97]通过在 RKHS 中嵌入分布，引入 MMD 度量来比较基于 RKHS 距离的分布，设计出了一种非参数距离估计法。接下来我们将对这一过程进行介绍。

任意函数空间的 MMD： 基于 MMD 的统计检验方式原理如下，利用来自两个分布的数据集 $\mathbf{A} = \{\mathbf{a}_i\}$ 和 $\mathbf{B} = \{\mathbf{b}_j\}$，通过在样本空间中寻找一个能够将变量映射到高维空间映射函数 f，通过计算得到不同分布的随机变量在映射后的均值。将两个均值相减便能得出两个分布在 f 上的平均差异，即期望的差值。我们的目标是要寻找

这样一个 f，使得这个均值差异能够达到最大值即平均差异的上确界，这个最大值即为 MMD。使用 $\mathcal{F}$ 表示在样本空间上的连续函数域，MMD 的一般形式如下

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathbf{a} \sim p} \left[f\left(\mathbf{a}\right) \right] - \mathbb{E}_{\mathbf{b} \sim q} \left[f\left(\mathbf{b}\right) \right] \right), \tag{4.1}$$

其中，**a** 和 **b** 分别是从分布 p 和 q 中满足独立同分布采样得到的两个数据样本，假设 **A** 和 **B** 数据集中的样本数分别是 n 和 m，可以得到基于 **A** 和 **B** 的 MMD 经验估计

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} f\left(\mathbf{a}_i\right) - \frac{1}{m} \sum_{j=1}^{m} f\left(\mathbf{b}_j\right) \right), \tag{4.2}$$

该值可以用于判断两个分布之间的相似程度。当且仅当 **A** 和 **B** 的数据分布相同时 MMD 的值为 0，这就要求函数域 $\mathcal{F}$ 必须足够丰富。另一方面，检验统计量必须具有足够的连续性，这样 MMD 的经验估计才能随着观测集规模增大而收敛到它的期望，要求函数域 $\mathcal{F}$ 必须是有限的。而当 $\mathcal{F}$ 是 RKHS 中单位球内的一个任意向量时，满足上述两个约束。

RKHS 中的 MMD：在 RKHS 中，MMD 中的函数域定义为 RKHS 中的单位球中的一个向量 (即 $\|f\| < 1$)。将 $\phi(\mathbf{a})$ 定义为 **a** 在 RKHS 上的映射，式(4.2)中的 $f\left(\mathbf{a}\right)$ 表示 RKHS 中的向量 f 与该空间中的向量 $\phi(\mathbf{a})$ 的内积

$$f(\mathbf{a}) = \langle f, \phi(\mathbf{a}) \rangle_{\mathcal{H}}, \tag{4.3}$$

即变量经过核函数映射到 RKHS 中的向量。在 RKHS 上，每个 f 对应一个特征映射，对于 $\mathbb{E}_{\mathbf{a} \sim p} \left[f\left(\mathbf{a}\right) \right]$，根据内积的性质有

$$\mathbb{E}_{\mathbf{a} \sim p} \left[f\left(\mathbf{a}\right) \right] = \mathbb{E}_{\mathbf{a} \sim p} \left[\langle f, \phi(\mathbf{a}) \rangle_{\mathcal{H}} \right] = \langle f, \mathbb{E}_{\mathbf{a} \sim p} \left[\phi(\mathbf{a}) \right] \rangle_{\mathcal{H}} = \langle f, \mu_p \rangle_{\mathcal{H}}, \tag{4.4}$$

其中，$\mu_p = \mathbb{E}_{\mathbf{a} \sim p} \left[\phi(\mathbf{a}) \right]$。由上述定义，利用 RKHS 的性质，可以将 MMD 距离进行如下推导

$$\begin{aligned} \mathrm{MMD}[\mathcal{F}, p, q] &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}_p \left[f\left(\mathbf{a}\right) \right] - \mathbb{E}_q \left[f\left(\mathbf{b}\right) \right] \right) \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}_p \left[\langle f, \phi(\mathbf{a}) \rangle_{\mathcal{H}} \right] - \mathbb{E}_q \left[\langle f, \phi(\mathbf{b}) \rangle_{\mathcal{H}} \right] \right) \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \left[\langle f, \mu_p - \mu_q \rangle_{\mathcal{H}} \right] \\ &= \|\mu_p - \mu_q\|_{\mathcal{H}}. \end{aligned} \tag{4.5}$$

在第二个等号中，利用了 RKHS 的再生性，先将向量通过函数 $\phi(\mathbf{a})$ 映射到希尔伯特空间，然后与该空间中的单位球内给定的向量 f 作内积，完成映射到高维的变换。第三个等号我们可以由式(4.4)得到。第四个等号则是因为内积的性质，即两向量的内积小于等于两向量模的乘法，并且有限制条件 $\|f\| < 1$。

将 μ_p 和 μ_q 的值由均值替代，通过引入核技巧隐式地表示映射函数来简化映射函数的求解过程。式(4.5)可以进一步化简，并用核函数进行表达

$$\begin{aligned}
\mathrm{MMD}^2[\mathcal{F},p,q] &= \left\| \frac{1}{n}\sum_{i=1}^{n}\phi\left(\mathbf{a}_i\right) - \frac{1}{m}\sum_{j=1}^{m}\phi\left(\mathbf{b}_j\right) \right\|_{\mathcal{H}}^2 \\
&= \left\| \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{n}\phi\left(\mathbf{a}_i\right)\phi\left(\mathbf{a}_i'\right) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\phi\left(\mathbf{a}_i\right)\phi\left(\mathbf{b}_j\right) + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{j'=1}^{m}\phi\left(\mathbf{b}_j\right)\phi\left(\mathbf{b}_j'\right) \right\|_{\mathcal{H}} \\
&= \left\| \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{n}k\left(\mathbf{a}_i,\mathbf{a}_i'\right) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k\left(\mathbf{a}_i,\mathbf{b}_j\right) + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{j'=1}^{m}k\left(\mathbf{b}_j,\mathbf{b}_j'\right) \right\|_{\mathcal{H}},
\end{aligned} \tag{4.6}$$

其中，$k(\cdot,\cdot)$ 表示两个向量在 RKHS 中的内积。为简化计算，将上式化为矩阵形式

$$\begin{aligned}
\mathrm{MMD}^2[\mathcal{F},p,q] &= \mathrm{tr}\left(\begin{bmatrix} \phi(\mathbf{A})^{\mathrm{T}} \\ \phi(\mathbf{B})^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \phi(\mathbf{A}) & \phi(\mathbf{B}) \end{bmatrix} \begin{bmatrix} \frac{1}{n^2}\mathbf{1}\mathbf{1}^{\mathrm{T}} & \frac{-1}{nm}\mathbf{1}\mathbf{1}^{\mathrm{T}} \\ \frac{-1}{nm}\mathbf{1}\mathbf{1}^{\mathrm{T}} & \frac{1}{m^2}\mathbf{1}\mathbf{1}^{\mathrm{T}} \end{bmatrix} \right) \\
&= \mathrm{tr}\left(\begin{bmatrix} \langle\phi(\mathbf{A}),\phi(\mathbf{A})\rangle & \langle\phi(\mathbf{A}),\phi(\mathbf{B})\rangle \\ \langle\phi(\mathbf{B}),\phi(\mathbf{A})\rangle & \langle\phi(\mathbf{B}),\phi(\mathbf{B})\rangle \end{bmatrix} \mathbf{M} \right) \\
&= \mathrm{tr}\left(\begin{bmatrix} K_{\mathbf{A},\mathbf{A}} & K_{\mathbf{A},\mathbf{B}} \\ K_{\mathbf{B},\mathbf{A}} & K_{\mathbf{B},\mathbf{B}} \end{bmatrix} \mathbf{M} \right) = \mathrm{tr}\left(\mathbf{K}\mathbf{M}\right),
\end{aligned} \tag{4.7}$$

其中，

$$\left(M\right)_{ij} = \begin{cases} \frac{1}{n^2}, & \mathbf{a}_i \in \mathbf{A}, \\ \frac{1}{m^2}, & \mathbf{b}_j \in \mathbf{B}, \\ \frac{-1}{nm}, & \text{其他}. \end{cases} \tag{4.8}$$

此时，$\mathbf{a}_i$，$\mathbf{b}_j$ 两个向量在高维特征空间的内积，就可以简化为它们在原始样本空间中的内积通过核函数 $k(\cdot,\cdot)$ 的计算结果。进一步的对于 $\mathbf{A}$ 和 $\mathbf{B}$ 在特征空间的内积 $\langle\phi(\mathbf{A}),\phi(\mathbf{B})\rangle$ 等于在原始样本空间中通过核函数 $K_{\mathbf{A},\mathbf{B}}$ 的计算结果。通过 RKHS 的核函数限制了函数域的范围，选定了一个核函数就对应了一系列的函数矩，而两个分布在 RKHS 中均值的距离就是这一系列函数矩的差值的上界，即所求的 MMD 距离。

4.2.2 联合最大均值差异

将来自两个分布的样本 **A** 和 **B** 分别看作是来自源域的数据和目标域的实例集合，MMD 就是通过在 RHSK 中减少源域和目标域的边缘分布概率来完成域适应，然而减少边缘分布的差异并不能保证条件分布的差异也被减小。$P(\mathbf{Y}|\mathbf{X})$ 样本的条件分布概率对领域自适应同样重要，如果能够同时减少两个域之间的边缘分布和条件分布，那么域适应的鲁棒性将大大增强。但是条件分布概率要求目标域样本的标签已知，否则无法对 $P(\mathbf{Y}|\mathbf{X})$ 进行建模完成条件分布的匹配。针对目标域没有标签的无监督迁移学习，如何获取目标域的标签成为解决适配条件分布概率的关键。

Long 等人[46]提出利用在有标记的源域上训练的基准分类器 (如 SVM，KNN) 预测目标域数据，产生伪标签来建模条件分布概率，这种同时适配两个域之间的边缘分布和条件分布方法被称为 JDA。由于后验分布概率 $P(\mathbf{Y}|\mathbf{X})$ 的计算非常复杂，他们根据充分统计量的思想，采用类条件分布 $P(\mathbf{X}|\mathbf{Y})$ 代替 $P(\mathbf{Y}|\mathbf{X})$，完成条件概率适配。此时对于源域数据矩阵 $\mathbf{X}_S = \{\mathbf{x}_i\}$ 和目标域的数据矩阵 $\mathbf{X}_T = \{\mathbf{x}_j\}$，样本数量分别为 n 和 m，总共包含 C 个类别。此时，我们有了目标域的 “标签”，可以对齐类条件分布 $P_S(\mathbf{x}_S|\mathbf{y}_s = c)$ 和 $P_T(\mathbf{x}_T|\mathbf{y}_T = \hat{c})$，标签空间 $\mathcal{Y}$ 中的每个类 $\hat{c} \in \left\{1, \ldots, \hat{C}\right\}$。此时，我们套用式(4.7)，可以得出相同的类与类之间的 MMD 距离为

$$\begin{aligned}
&\sum_{c=1}^{C} \left\| \frac{1}{n_c} \sum_{\mathbf{x}_i \in \mathcal{D}_S^{(c)}} \phi(\mathbf{x}_i) - \frac{1}{m_c} \sum_{\mathbf{x}_j \in \mathcal{D}_T^{(c)}} \phi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2 \\
&= \operatorname{tr}\left(\begin{bmatrix} K_{\mathbf{X}_S,\mathbf{X}_S} & K_{\mathbf{X}_S,\mathbf{X}_T} \\ K_{\mathbf{X}_T,\mathbf{X}_S} & K_{\mathbf{X}_T,\mathbf{X}_T} \end{bmatrix} \mathbf{M} \right) \\
&= \operatorname{tr}\left(\mathbf{K}_x \mathbf{M}_c\right),
\end{aligned} \tag{4.9}$$

其中，n_c，m_c 分别表示源域和目标域中来自第 c 类的样本个数。$\mathcal{D}_S^{(c)} = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathcal{D}_S \wedge y(\mathbf{x}_i) = c\}$ 是源域中属于第 c 类的一组示例集，$y(\mathbf{x}_i)$ 是 $\mathbf{x}_i$ 的真实标签。对应的，$\mathcal{D}_T^{(c)} = \{\mathbf{x}_j \mid \mathbf{x}_j \in \mathcal{D}_T \wedge \hat{y}(\mathbf{x}_j) = \hat{c}\}$ 是目标域域中属于第 c 类的一组示例集，$\hat{y}(\mathbf{x}_j)$ 是对 $\mathbf{x}_j$ 预测的伪标签。

此时将边缘概率分布

$$\left\| \frac{1}{n} \sum_{\mathbf{x}_i \in \mathcal{D}_S} \phi(\mathbf{x}_i) - \frac{1}{m} \sum_{\mathbf{x}_j \in \mathcal{D}_T} \phi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2 = \operatorname{tr}\left(\mathbf{K}_x \mathbf{M}_0\right). \tag{4.10}$$

与式(4.9)相结合，可以得到联合最大均值差异 (Joint Maximum Mean Difference，JMMD)

$$\sum_{c=0}^{C} \mathrm{tr}\left(\mathbf{K}_x \mathbf{M}_c\right), \tag{4.11}$$

其中，$\mathbf{M}_c$ 为

$$(\mathbf{M}_c)_{ij} = \begin{cases} \frac{1}{n_c^2}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_S^{(c)}, \\ \frac{1}{m_c^2}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_T^{(c)}, \\ -\frac{1}{m_c n_c}, & \begin{cases} \mathbf{x}_i \in \mathcal{D}_S^{(c)}, \mathbf{x}_j \in \mathcal{D}_t^{(c)}, \\ \mathbf{x}_i \in \mathcal{D}_T^{(c)}, \mathbf{x}_j \in \mathcal{D}_s^{(c)}, \end{cases} \\ 0, & \text{其他.} \end{cases} \tag{4.12}$$

JMMD 在 MMD 的基础上，同时适配源域和目标域之间的边缘分布和条件分布，对跨域分类问题具有更好的鲁棒性。

4.2.3 判别联合概率最大均值差异

JMMD 方法是利用类条件分布概率近似后验分布概率，这种方法虽然降低了计算复杂度，但也一定程度上降低了域适应的精度。此外，MMD 和 JMMD 方法只考虑了通过对齐不同的分布来增加域之间的可迁移性，而忽略了不同类之间的可鉴别性。Zhang 等人[98]，在充分考虑可迁移性和可鉴别性的基础上，对最大均值差异度量方法进行改进，采用贝叶斯定律使用严格的后验分布对条件分布概率进行计算，提出了 DJP-MMD。

联合概率差异：此处，沿用上面提到的源域和目标域的概念，以及两个域中关于类的定义。对于源域和目标域的数据矩阵 $\mathbf{X}_S$、$\mathbf{X}_T$ 和对应标签的独热码矩阵 $\mathbf{Y}_S$、$\mathbf{Y}_T$，令 $P(\mathbf{X}|\mathbf{Y})$ 为条件概率，$P(\mathbf{Y})$ 是先验概率，根据贝叶斯定理，可以得到联合概

率差异如下

$$
\begin{aligned}
d(\mathcal{D}_S, \mathcal{D}_T) &= d\left(P(\mathbf{X}_S, \mathbf{Y}_S), P(\mathbf{X}_T, \mathbf{Y}_T)\right) \\
&= d\left(P(\mathbf{X}_S|\mathbf{Y}_S)P(\mathbf{Y}_S), P(\mathbf{X}_T|\mathbf{Y}_T)P(\mathbf{Y}_T)\right) \\
&= \sum_{c=\hat{c}}^{C}\sum_{\hat{c}=1}^{C} d\left(P(\mathbf{X}_S|\mathbf{Y}_S^c)P(\mathbf{Y}_S^c), P(\mathbf{X}_T|\mathbf{Y}_T^{\hat{c}})P(\mathbf{Y}_T^{\hat{c}})\right) \\
&+ \sum_{c\neq\hat{c}}^{C}\sum_{\hat{c}=1}^{C} d\left(P(\mathbf{X}_S|\mathbf{Y}_S^c)P(\mathbf{Y}_S^c), P(\mathbf{X}_T|\mathbf{Y}_T^{\hat{c}})P(\mathbf{Y}_T^{\hat{c}})\right) \\
&= \sum_{c=1}^{C} d\left(P(\mathbf{X}_S|\mathbf{Y}_S^c)P(\mathbf{Y}_S^c), P(\mathbf{X}_T|\mathbf{Y}_T^{\hat{c}})P(\mathbf{Y}_T^{\hat{c}})\right) \\
&+ \sum_{c\neq\hat{c}}^{C}\sum_{\hat{c}=1}^{C} d\left(P(\mathbf{X}_S|\mathbf{Y}_S^c)P(\mathbf{Y}_S^c), P(\mathbf{X}_T|\mathbf{Y}_T^{\hat{c}})P(\mathbf{Y}_T^{\hat{c}})\right) \\
&= \mathcal{M}_B + \mathcal{M}_D,
\end{aligned}
\tag{4.13}
$$

其中，$\mathcal{M}_B$ 和 $\mathcal{M}_D$ 分别表示源域和目标域在同一类别和不同类别上的联合概率差异，$\hat{c}$ 表示目标域预测的种类。

不同于 JMMD，联合概率差异是基于类条件概率和类先验概率的乘积来计算域差异，是直接从数据中计算出来而不需要进行充分统计量的近似。此时，如果直接最小化式(4.13)可以提高源域和目标域之间的可迁移性，但是无法增加可鉴别性。因此，将鉴别联合概率分布差异定义为

$$
d(\mathcal{D}_S, \mathcal{D}_T) = \mathcal{M}_B - \mu\mathcal{M}_D, \tag{4.14}
$$

其中，$\mu > 0$ 是一个平衡参数。$\mathcal{M}_B$ 能够测量同一类别在不同域之间的可迁移性，$\mathcal{M}_D$ 则能够对不同类别在不同域之间的鉴别性进行测量。这个式子可以最小化类内联合概率来增加可迁移性，最大化类间联合概率来增加可鉴别性。

$\mathcal{M}_B$ 类内联合概率差异计算：根据式(4.13)有

$$
\begin{aligned}
\mathcal{M}_B &= \sum_{c=1}^{C} d\left(P(\mathbf{X}_S|\mathbf{Y}_S^c)P(\mathbf{Y}_S^c)P(\mathbf{X}_T|\mathbf{Y}_T^{\hat{c}})P(\mathbf{Y}_T^{\hat{c}})\right) \\
&= \sum_{c=1}^{C} \left\|\mathbb{E}[f(\mathbf{x}_S)|\mathbf{y}_S^c]P(\mathbf{y}_S^c) - \mathbb{E}[f(\mathbf{x}_T)|\mathbf{y}_T^{\hat{c}})]P(\mathbf{y}_T^{\hat{c}})\right\|^2,
\end{aligned}
\tag{4.15}
$$

其中，$\mathbb{E}[f(\mathbf{x}_S)|\mathbf{y}_S^c]$ 为

$$\mathbb{E}[f(\mathbf{x}_S)|\mathbf{y}_S^c] = \frac{1}{n^c}\sum_{i=1}^{n^c} A^\top \mathbf{x}_i^c, \quad P(\mathbf{y}_S^c) = \frac{n^c}{n}, \tag{4.16}$$

其中，n^c 表示第 C 类中样本的数量。此时，可以得到

$$\mathbb{E}[f(\mathbf{x}_S)|\mathbf{y}_S^c]P(\mathbf{y}_S^c) = \frac{1}{n}\sum_{i=1}^{n^c} \phi(\mathbf{x}_i^c), \tag{4.17}$$

其中，$\phi(\mathbf{x}_i^c)$ 表示样本 $\mathbf{x}_i$ 在 RKHS 中的映射函数。同样的，针对目标域数据

$$\mathbb{E}[f(\mathbf{x}_T)|\hat{\mathbf{y}}_T^{\hat{c}}]P(\hat{\mathbf{y}}_T^{\hat{c}}) = \frac{1}{m}\sum_{j=1}^{m^c} \phi(\mathbf{x}_j^c), \tag{4.18}$$

将式(4.17)和式(4.18)在式(4.15)进行替换，可以得到 $\mathcal{M}_B$

$$\mathcal{M}_B = \sum_{c=1}^{C} \left\| \frac{1}{n}\sum_{i=1}^{n^c} \phi(\mathbf{x}_i^c) - \frac{1}{m}\sum_{j=1}^{m^{\hat{c}}} \phi(\mathbf{x}_j^{\hat{c}}) \right\|_{\mathcal{H}}^2. \tag{4.19}$$

已知源域的独热码标签矩阵为 $\mathbf{Y}_S = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$，目标域的独热码伪标签矩阵为 $\hat{\mathbf{Y}}_T = [\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_m]$，式(4.19)可以进一步化简

$$\mathcal{M}_B = \|\phi(X_S)N_S - \phi(X_T)N_T\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\mathbf{K}_x \mathbf{R}_{\min}\right), \tag{4.20}$$

其中，

$$\mathbf{R}_{\min} = \begin{bmatrix} \mathbf{N}_S\mathbf{N}_S^\top & -\mathbf{N}_S\mathbf{N}_T^\top \\ -\mathbf{N}_T\mathbf{N}_S^\top & \mathbf{N}_T\mathbf{N}_T^\top \end{bmatrix}, \quad \mathbf{N}_S = \frac{\mathbf{Y}_S}{n}, \quad \mathbf{N}_T = \frac{\hat{\mathbf{Y}}_T}{m}. \tag{4.21}$$

$\mathcal{M}_D$ 类间联合概率差异计算：根据式(4.13)有

$$\begin{aligned} \mathcal{M}_D &= \sum_{c\neq\hat{c}}\sum_{\hat{c}=1}^{C} d(P(\mathbf{X}_S|\mathbf{Y}_S^c)P(\mathbf{Y}_S^c), P(\mathbf{X}_T|\mathbf{Y}_T^{\hat{c}})P(\mathbf{Y}_T^{\hat{c}})) \\ &= \sum_{c\neq\hat{c}}\sum_{\hat{c}=1}^{C} \left\|\mathbb{E}[f(\mathbf{x}_S)|\mathbf{y}_S^c]P(\mathbf{y}_S^c) - \mathbb{E}[f(\mathbf{x}_T)|\mathbf{y}_T^{\hat{c}}]P(\mathbf{y}_T^{\hat{c}})\right\|^2. \end{aligned} \tag{4.22}$$

将式(4.17)和式(4.18)带入上式，得到

$$\mathcal{M}_D = \sum_{c\neq\hat{c}}\sum_{\hat{c}=1}^{C} \left\| \frac{1}{n^c}\sum_{i=1}^{n^c} \phi(\mathbf{x}_i^c) - \frac{1}{m^{\hat{c}}}\sum_{j=1}^{m^{\hat{c}}} \phi(\mathbf{x}_j^{\hat{c}}) \right\|_{\mathcal{H}}^2. \tag{4.23}$$

定义两个矩阵 $\mathbf{F}_S \in \mathbb{R}^{n\times(C(C-1))}$ 和 $\hat{\mathbf{F}}_T \in \mathbb{R}^{m\times(C(C-1))}$

$$\begin{aligned}\mathbf{F}_s &= [\mathbf{Y}_S(:,1)*(C-1),...,\mathbf{Y}_S(:,C)*(C-1)],\\ \hat{\mathbf{F}}_T &= [\hat{\mathbf{Y}}_T(:,1:C)_{\hat{c}\neq 1},...,\hat{\mathbf{Y}}_T(:,1:C)_{\hat{c}\neq C}],\end{aligned} \tag{4.24}$$

其中，$\mathbf{Y}_S(:,c)$ 表示源域 $\mathbf{Y}_S$ 的第 c 列，$\hat{\mathbf{Y}}_T(:,1:C)_{\hat{c}\neq 1}$ 由目标域的 $\hat{\mathbf{Y}}_T$ 的第 1 列到第 C 列（不包括第一列）组成。$\hat{\mathbf{F}}_T$ 由伪标签组成，能够在迭代中进行更新，以此来减少伪标签对迁移结果的影响。

此时，式(4.23)可以进一步写为

$$\mathcal{M}_D = \|\phi(\mathbf{X}_S)\mathbf{M}_S - \phi(\mathbf{X}_T)\mathbf{M}_T\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\mathbf{K}_x\mathbf{R}_{\max}\right), \tag{4.25}$$

其中，

$$\mathbf{R}_{\max} = \begin{bmatrix}\mathbf{M}_s\mathbf{M}_s^\top & -\mathbf{M}_s\mathbf{M}_t^\top\\ -\mathbf{M}_t\mathbf{M}_s^\top & \mathbf{M}_t\mathbf{M}_t^\top\end{bmatrix},\quad \mathbf{M}_s = \frac{\mathbf{F}_s}{n},\quad \mathbf{M}_t = \frac{\hat{\mathbf{F}}_t}{m}. \tag{4.26}$$

此时，将式(4.20)和式(4.25)代入式(4.14)中得到 DJP-MMD 的最终形式为

$$d(\mathcal{D}_S,\mathcal{D}_T) = \mathcal{M}_B - \mu\mathcal{M}_D = \mathrm{tr}\left(\mathbf{K}_x\left(\mathbf{R}_{\min} - \mu\mathbf{R}_{\max}\right)\right). \tag{4.27}$$

DJP-MMD 方法通过最小化类内的联合概率 MMD，最大化类间的联合概率 MMD 来提高源域和目标域的可迁移性和可鉴别性，这种方法弥补了 JMMD 的不足，进一步提高了域适应的鲁棒性。

4.3 模型建立

本节将通过多核的方法对 DJP-MMD 进行改进，增强 RKHS 变换的核选择来提高域适应的效果，提出多核 DJP-MMD(Multi Kernel DJP-MMD，MKDJP-MMD) 方法。我们把该方法拓展到神经网络上，提出深度联合概率适应网络。最后，分别在两个卷积神经网络上实现了基于 MKDJP-MMD 方法的自适应层设计，接下来对提出的方法进行详细介绍。

4.3.1 深度联合概率适应网络框架

(1) 基于多核方法改进的 DJP-MMD

在最大均值差异的计算中，利用核函数将原始数据送入高维的 RKHS，在该核函数对应的高维 RKHS 中计算两个数据分布的差异。由于映射函数 $\phi(\cdot)$ 未知，因此在这个过程中利用核技巧的方法，隐式地表示映射函数 $\phi(\cdot)$，直接获得了数据的高维差异。RKHS 的核再生性体现在两个方面，首先是 $\langle f, \phi(\mathbf{X})\rangle$ 表示了 $\mathbf{X}$ 在希尔伯特空间的映射点 $\phi(\mathbf{X})$ 与其他映射点 f 的相似度，其次是 $\langle K(\mathbf{X}_S), K(\mathbf{X}_T)\rangle$ 可以由另一个核表示，反映了源域和目标域数据在 RKHS 中的相似度。

核函数的选择对于 MMD 的计算至关重要，确定了一个核函数就确定了一个 RKHS，而这个核函数通常是固定的由高斯核或线性核表示。对于不同的数据，不同的核函数有效性不同，但是在以往的方法中是根据经验来确定核函数的。Gretton 等人[99]针对核函数的选择问题，提出了多核最大均值差异 (Multiple Kernel Maximum Mean Discrepancy，MK-MMD) 的方法。我们在本节中借鉴该方法的思想，用多个核的线性组合构造一个总的核来对 DJP-MMD 的核选择进行改进。

将 $\phi(\cdot)$ 函数看作是一个单射函数，那么与其相关联的核函数 $k(\mathbf{x}_S, \mathbf{x}_T) = \langle \phi(\mathbf{x}_S), \phi(\mathbf{x}_T)\rangle$ 是一个特征核。我们的目标是从一个特定的内核族 $\mathcal{K}$ 中选择一个可用于假设测试的内核，定义如下。令 $\{k_u\}_{u=1}^m$ 是 m 个半正定内核的凸组合，此时有

$$\mathcal{K} = \left\{ k \mid k = \sum_{u=1}^{m} \beta_u k_u, \sum_{u=1}^{m} \beta_u = 1, \beta_u \geq 0, \forall u \in \{1, \ldots, m\} \right\}, \tag{4.28}$$

其中，对系数 $\{\beta_u\}$ 施加约束来保证得到的多核 k 是特有的。每一个 $k \in \mathcal{K}$ 都与唯一一个 RKHS 中的 $\mathcal{H}$ 相关联，此时 DJP-MMD 改写为

$$\begin{aligned} d(\mathcal{D}_S, \mathcal{D}_T) &= \mathcal{M}_B - \mu \mathcal{M}_D \\ &= \text{tr}\left(\sum_{u=1}^{m} \beta_u \begin{bmatrix} K_{\mathbf{X}_S, \mathbf{X}_S} & K_{\mathbf{X}_S, \mathbf{X}_T} \\ K_{\mathbf{X}_T, \mathbf{X}_S} & K_{\mathbf{X}_T, \mathbf{X}_T} \end{bmatrix} (\mathbf{R}_{\text{min}} - \mu \mathbf{R}_{\text{max}}) \right) \\ &= \text{tr}\left(\sum_{u=1}^{m} \beta_u \mathbf{K}_x (\mathbf{R}_{\text{min}} - \mu \mathbf{R}_{\text{max}}) \right). \end{aligned} \tag{4.29}$$

在这个过程中假设最优的核可以由多个不同的核线性组合得到，由于多个不同的核中每个核都是特有的，并且其系数 $\beta_u \geq 0$，可以得知产生的最优核也是特有的。

我们使用多核的方法加强了 DJP-MMD 方法，提出了 MKDJP-MMD 方法。这种方法避免了单核选择的局限性，通过多个核的线性组合选择最优的核函数，这使得将数据变换到 RKHS 中时，核函数的选择合理、准确。

(2) 深度联合概率适应网络 DJP-MMD

本小节在卷积神经网络的架构基础上，设计了基于 MKDJP-MMD 的深度域适应网络。由于应对的任务是目标域没有故障标记信息的无监督迁移学习问题，因此无法利用基于模型方法将源域训练的神经网络适配目标域数据进行参数微调。我们提出的 DJPAN 同时利用源域的标记数据和目标域的未标记数据，能够在网络的不同层对两个域的数据进行适配。此外针对源域和目标域中来自不同类的数据，引入的域适应损失能够增加不同种类数据的差异，使得模型对不同的种类更为敏感，更容易地完成分类任务。

图4.2中给出了深度联合概率适应网络的架构图，该方法能够在不同的卷积神经网络上进行应用。对于不同网络而言，网络的基本架构不会变化，DJPAN 在网络的全连接层进行域适应，并将域适应损失作为目标函数的一部分，在反向传播中优化网络参数。

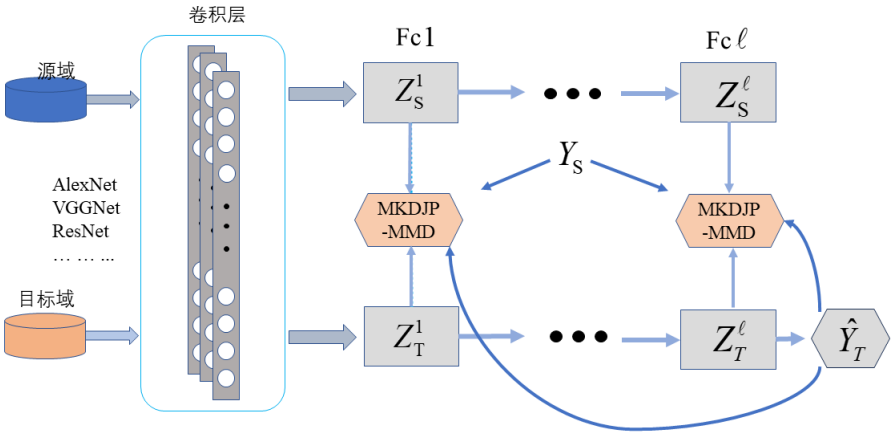


图 4.2　深度联合概率适应网络

Figure 4.2　Deep joint probabilistic adaptation network

根据第二章对深度迁移学习的介绍，可以知道卷积神经网络的卷积层能够完成对图片通用特征的提取。对于迁移学习来说，源域和目标域数据的区别在于高级特征的不同，因此我们对全连接层每一层输出的层激活输出矩阵 $\mathbf{Z}$ 进行适配，$\mathbf{Z}_S^\ell = \left\{\mathbf{z}_{S,i}^\ell, \ldots, \mathbf{z}_{S,i}^\ell\right\}_{i=1}^n$ 表示源域数据经过卷积层后在第 ℓ 个全连接层的层激活矩阵，对应

地 $\mathbf{Z}_T^\ell = \left\{\mathbf{z}_{T,j}^\ell, \ldots, \mathbf{z}_{T,j}^\ell\right\}_{j=1}^m$ 表示目标域数据经过卷积层后在第 ℓ 个全连接层的层激活矩阵。此外，在源域数据标签已知的情况下，可以非常容易地利用源域数据训练出一个神经网络。将该网络应用于目标域数据进行标签预测，可以得到目标域数据的伪标签 $\hat{\mathbf{Y}}_T$，此时引入 MKDJP-MMD 来减少在第 ℓ 全连接层中输出的两个域的激活矩阵之间的分布差异。

接下来，我们将对网络的优化过程进行介绍。神经网络的每个不同层 ℓ 都能学习一个非线性映射 $\mathbf{h}_i^\ell = f^\ell\left(\mathbf{W}^\ell \mathbf{h}_i^{\ell-1} + \mathbf{b}^\ell\right)$，其中 $\mathbf{h}_i^\ell$ 是关于数据 $\mathbf{x}_i$ 第 ℓ 层的隐藏表示，$\mathbf{W}^\ell$ 和 $\mathbf{b}^\ell$ 是第 ℓ 层神经元的权重和偏置，f^ℓ 作为输出层的激活函数，根据需要定义为修正线性单元 (Rectified Linear Unit，ReLU)$f^\ell(\mathbf{x}) = \max(\mathbf{0}, \mathbf{x})$ 或 Softmax$f^\ell(\mathbf{x}) = e^{\mathbf{x}} / \sum_{j=1}^{|\mathbf{x}|} e^{x_j}$。将卷积神经网络参数统一定义为 $\Theta = \left\{\mathbf{W}^\ell, \mathbf{b}^\ell\right\}_{\ell=1}^l$，此时在源域数据集上，可以将神经网络的经验风险损失写为如下的形式

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} J\left(\theta\left(\mathbf{x}_{S,i}\right), y_{S,i}\right), \tag{4.30}$$

其中，J 表示交叉熵损失函数，$\theta\left(\mathbf{x}_{S,i}\right)$ 能够输出网络对数据 $\mathbf{x}_{S,i}$ 预测为 $y_{S,i}$ 的概率。

我们此前介绍过，卷积网络的卷积层能够提取通用特征，因此源域和目标域数据在卷积层的差异不大。后续的全连接层能够提取高级特征，是影响领域适应的关键。对两个域在全连接层的层激活 $\mathbf{Z}_S$ 和 $\mathbf{Z}_T$，使用 MKDJP-MMD 进行差异度量

$$\begin{aligned}
&\widehat{d}_{\mathcal{L}}\left(\mathcal{D}_S^\ell, \mathcal{D}_T^\ell\right) = \mathcal{M}_B - \mu \mathcal{M}_D \\
&= \sum_{c=1}^{C} \left[\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k^\ell\left(\mathbf{z}_{S,i}^\ell, \mathbf{z}_{S,j}^\ell\right) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k^\ell\left(\mathbf{z}_{T,i}^\ell, \mathbf{z}_{T,j}^\ell\right) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k^\ell\left(\mathbf{z}_{S,i}^\ell, \mathbf{z}_{T,j}^\ell\right) \right] \\
&- \mu \sum_{c \neq \hat{c}} \sum_{\hat{c}=1}^{C} \left[\frac{1}{(n^c)^2} \sum_{i=1}^{n^c} \sum_{j=1}^{n^c} k^\ell\left(\{\mathbf{z}_{S,i}^\ell\}^c, \{\mathbf{z}_{S,j}^\ell\}^c\right) + \frac{1}{(m^{\hat{c}})^2} \sum_{i=1}^{m^{\hat{c}}} \sum_{j=1}^{m^{\hat{c}}} k^\ell\left(\{\mathbf{z}_{T,i}^\ell\}^c, \{\mathbf{z}_{T,j}^\ell\}^c\right) \right. \\
&\left. - \frac{2}{n^c m^{\hat{c}}} \sum_{i=1}^{n^c} \sum_{j=1}^{m^{\hat{c}}} k^\ell\left(\{\mathbf{z}_{S,i}^\ell\}^c, \{\mathbf{z}_{T,j}^\ell\}^{\hat{c}}\right) \right].
\end{aligned} \tag{4.31}$$

将域适应损失整合到 CNN 的分类损失中，得到 DJPAN 网络的整体目标函数

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} J\left(\theta\left(\mathbf{x}_{S,i}\right), y_{S,i}\right) + \lambda \sum_{\ell \in \mathcal{L}} \widehat{d}_{\mathcal{L}}\left(\mathcal{D}_S^\ell, \mathcal{D}_T^\ell\right), \tag{4.32}$$

其中，$\lambda > 0$ 是平衡参数，用于衡量迁移损失在目标函数中的权重，ℓ 表示层索引。$\widehat{d}_{\mathcal{L}}\left(\mathcal{D}_S^\ell, \mathcal{D}_T^\ell\right)$ 用于在全连接层评估源域和目标域之间的 MKDJP-MMD。

接下来将说明域适应损失的引入不会增加网络训练负担，并且深度域适应网络在大型数据集上也能够应用。在网络的反向传播中，利用式(4.32)的损失来对网络参数 Θ 进行学习优化。通过使用核技巧，MKDJP-MMD 可以写为核函数的期望

$$\mathbb{E}_{\mathbf{x}_S\mathbf{x}'_S}k(\mathbf{x}_S,\mathbf{x}'_S)+\mathbb{E}_{\mathbf{x}_T\mathbf{x}'_T}k(\mathbf{x}_T,\mathbf{x}'_T)-2\mathbb{E}_{\mathbf{x}_S\mathbf{x}_T}k(\mathbf{x}_S,\mathbf{x}_T), \tag{4.33}$$

其中，$\mathbf{x}_S \overset{iid}{\sim} p$，$\mathbf{x}_T \overset{iid}{\sim} q$，$k\in\mathcal{K}$，很明显这种计算会产生 $o(n^2)$ 的复杂度，会增加卷积网络参数在随机梯度下降 (Stochastic Gradient Descent，SGD) 时的求解难度，不利于在大规模数据集上的应用。在提出的方法中，采用 MKDJP-MMD 的无偏估计，来减少计算的复杂度。设一个四元组矩阵 $\mathbf{z}_i=(\mathbf{x}_{S,2i-1},\mathbf{x}_{S,2i},\mathbf{x}_{T,2i-1},\mathbf{x}_{T,2i})$，通过

$$\begin{aligned} g_k(\mathbf{z}_i)=&k(\mathbf{x}_{S,2i-1},\mathbf{x}_{S,2i})+k(\mathbf{x}_{T,2i-1},\mathbf{x}_{T,2i})\\ &-k(\mathbf{x}_{S,2i-1},\mathbf{x}_{T,2i})-k(\mathbf{x}_{S,2i-1},\mathbf{x}_{T,2i}), \end{aligned} \tag{4.34}$$

在四元组矩阵上计算多核函数 k，此时 MKDJP-MMD 为

$$\widehat{d}_{\mathcal{L}}(p,q)=\frac{2}{n}\sum_{i=1}^{n/2}g_k(\mathbf{z}_i). \tag{4.35}$$

这种方式直接计算自变量的期望，计算复杂度降低为为 $o(n)$。

使用小批量的 SGD 对卷积神经网络进行训练时，只需要计算目标函数(4.32)在每个数据点的梯度。我们在式(4.35)中，将域适应损失写为 $g_k(\mathbf{z}_i)$ 求和的形式，此时只需要计算关于第 ℓ 层的隐藏表示四元组 $\mathbf{z}_i^\ell=\left(\mathbf{h}_{S,2i-1}^\ell,\mathbf{h}_{S,2i}^\ell,\mathbf{h}_{T,2i-1}^\ell,\mathbf{h}_{T,2i}^\ell\right)$ 上的梯度 $\partial g_k(\mathbf{z}_i^\ell)/\partial\Theta^\ell$。网络参数的更新过程中，需要与域适应损失的梯度计算方式一致，将相应的卷积网络梯度写为 $\partial J(\mathbf{z}_i)/\partial\Theta^\ell$，其中 $J(\mathbf{z}_i)=\sum_{i'}J\left(\theta\left(\mathbf{x}'_{S,i}\right),y'_{S,i}\right)$。此时，目标函数(4.32)相对于第 ℓ 层参数 Θ^ℓ 的梯度可以写为

$$\nabla_{\Theta^\ell}=\frac{\partial J(\mathbf{z}_i)}{\partial\Theta^\ell}+\lambda\frac{\partial g_k\left(\mathbf{z}_i^\ell\right)}{\partial\Theta^\ell}. \tag{4.36}$$

针对梯度 $\partial g_k\left(\mathbf{z}_i^\ell\right)/\partial\Theta^\ell$ 的计算，假设给定核函数 k 为 m 个高斯核的线性组合 $\{k_u(\mathbf{x}_i,\mathbf{x}_j)=e^{-\|\mathbf{x}_i-\mathbf{x}_j\|^2/\gamma_u}\}$，那么梯度根据链式法则进行求导计算有

$$\begin{aligned} \frac{\partial k(\mathbf{h}_{S,2i-1}^\ell,\mathbf{h}_{T,2i}^\ell)}{\partial\mathbf{W}^\ell}=&-\sum_{u=1}^{m}\frac{2\beta_u}{\gamma_u}k_u\left(\mathbf{h}_{S,2i-1}^\ell,\mathbf{h}_{T,2i}^\ell\right)\times(\mathbf{h}_{S,2i-1}^\ell-\mathbf{h}_{T,2i}^\ell)\\ &\times\left(\mathbb{I}\left[\mathbf{h}_{S,2i-1}^{(\ell-1)}\right]-\mathbb{I}\left[\mathbf{h}_{T,2i}^{(\ell-1)}\right]\right)^{\mathsf{T}}, \end{aligned} \tag{4.37}$$

这里，式(4.37)的最后一行计算卷积神经网络第 ℓ 层激活函数的梯度。当激活函数为 ReLU 即 $f^{\ell}(\mathbf{x}) = \max(\mathbf{0}, \mathbf{x})$ 时，$\mathbb{I}$ 被定义为指示器函数

$$\mathbb{I}\left[\mathbf{h}_{j,i}^{(\ell-1)}\right] = \begin{cases} \mathbf{h}_{j,i}^{(\ell-1)}, & \mathbf{W}_j^{\ell}\mathbf{h}_i^{\ell-1} + \mathbf{b}_j^{\ell} \geqslant 0, \\ \mathbf{0}, & \text{其他}. \end{cases} \tag{4.38}$$

这种小批次的域适应损失 SGD 计算方法，可以在几乎任何卷积神经网络上实现，具有很高的可移植性。

4.3.2 基于 AlexNet 的深度联合概率适应网络

在本小节中，我们在 AlexNet 卷积神经网络的架构基础上，设计了基于 MKDJP-MMD 的深度域适应网络。针对 AlexNet 卷积网络的 8 层结构，Conv1-5 表示 AlexNet 的前 5 个卷积层，Fc1-3 表示该网络的 3 个全连接层，我们对 3 个全连接层的输出进行适配并输出域适应损失，图4.3展示了基于 AlexNet 的深度联合概率适应网络。

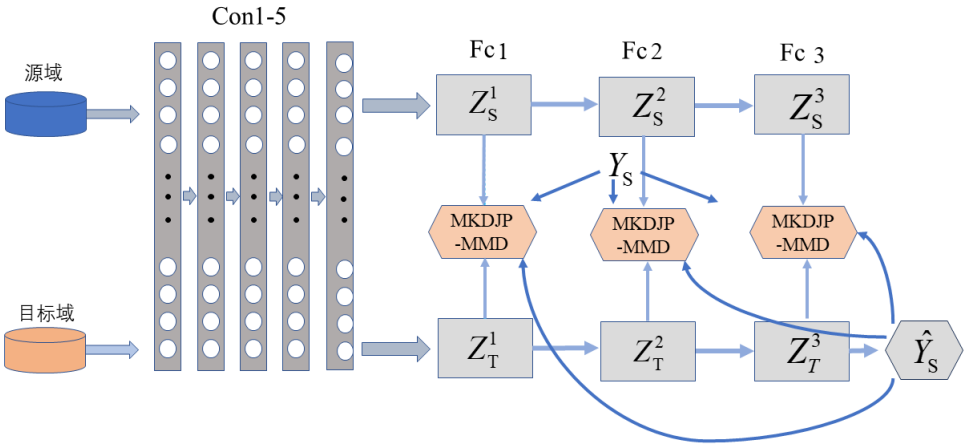


图 4.3 基于 AlexNet 的深度联合概率适应网络

Figure 4.3 Deep joint probabilistic adaptation network on AlexNet

AlexNet 网络结构简单，学习的参数较少，该网络采取了一系列措施来增强网络性能，例如引入 ReLU 这一非线性激活函数来解决梯度消失的问题，采用的 Dropout 处理来减少过拟合。表4.1中给出了我们实验所使用的 AlexNet 网络的参数，C 表示不同数据分类的种类数。在该网络模型中，通过整合源域的分类损失和域损失可以得到与式(4.32)相同的目标函数。针对该目标函数，通过式(4.36)完成其梯度计算，最终通过 SGD 完成该网络参数的优化。

表 4.1　AlexNet 网络的参数

Table 4.1　Parameters of the AlexNet network

网络层	参数
Conv1	out_channels=64, kernel_size=11
MaxPool1	kernel_size=3, stride=2, dilation=1
Conv2	out_channels=192, kernel_size=5
MaxPool2	kernel_size=3, stride=2, dilation=1
Conv3	out_channels=384, kernel_size=3
Conv4	out_channels=256, kernel_size=3
Conv5	out_channels=256, kernel_size=3
MaxPool3/Dropout	kernel_size=3, stride=2, dilation=1/p=0.5
Fully connected/Dropout	out_feature=4096/p=0.5
Fully connected	out_feature=256
Fully connected	out_feature=C

4.3.3　基于 ResNet50 的深度联合概率适应网络

在这个小节中，我们在 ResNet50 残差卷积神经网络架构的基础上，设计了基于 MKDJP-MMD 的深度域适应网络。深度残差网络引入了残差网络的设计，针对了超深度卷积网络的训练，解决了网络性能饱和退化以及梯度消失等问题。ResNet50 网络的卷积部分简单来说可以分为 1 个卷积层以及 4 个卷积阶段，全连接层为 2 层，我们的深度域适应工作在全连接层完成。图4.4中展示了我们所使用的 ResNet50 网络以及基于 ResNet50 的深度联合概率适应网络的架构。

表4.2中展示了我们搭建的网络模型的参数，所有 BatchNorm 层的设计是相同 (eps=1e-5, momentum=0.1)，输入尺寸与卷积层的输出有关，负责将卷积层的输出进行归一化处理。在每个 STAGE 中都包含了不同数量的 Bottleneck (BTNK)，这些 BTNK 的设计可以总结为两类，表4.2中给出了 BTNK1 和 BTNK2 的参数，同时图4.4展示了两种 BTNK 的结构。残差的设计是在每个 BTNK 中实现的，中心思想是不让网络直接拟合原先的映射，而是拟合残差映射。将不同层的数据经过 ReLU 进行非线性激活，将两个在不同卷积层的映射在堆叠的非线性层中进行拟合。这种方法在不引入额外的参数增加训练成本的情况下，大幅降低了训练更深层次神经网络的难度，也使准确率得到显著提升。同样地，整合 ResNet50 网络源域的分类损失和域损失可以得到与式(4.32)相同的目标函数，网络参数的优化通过 SGD 来完成。

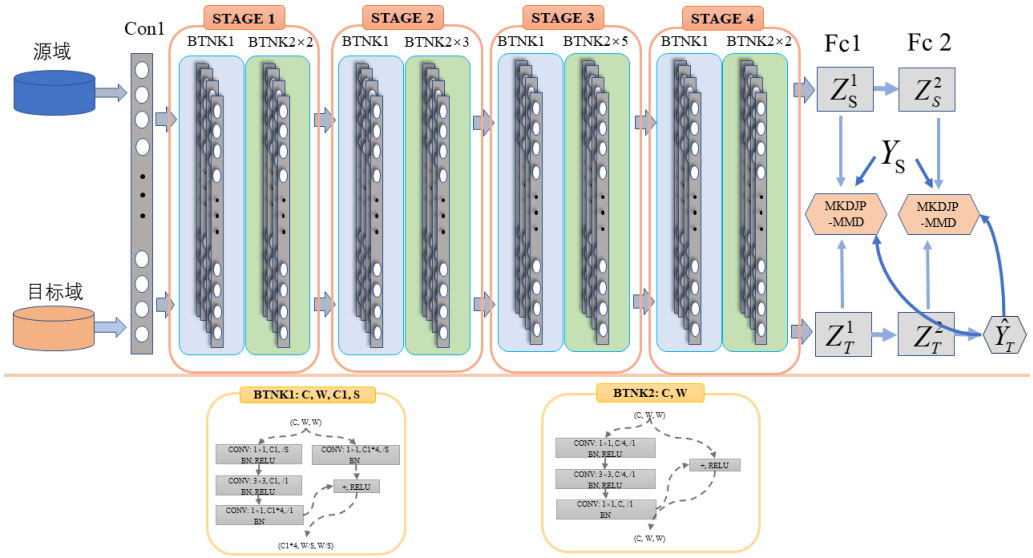


图 4.4　基于 ResNet50 的深度联合概率适应网络

Figure 4.4　Deep joint probabilistic adaptation network on ResNet50

表 4.2　ResNet50 网络的参数

Table 4.2　Parameters of the ResNet50 network

网络层	参数
Conv1/BatchNorm	out_channels=64, kernel_size=7,stride=2, padding=3,
MaxPool1	kernel_size=3, stride=2, padding=1, dilation=1
STAGE1	**BTNK1:**
Conv1/BatchNorm	out_channels=64, kernel_size=1, stride=1,
Conv2/BatchNorm	out_channels=64, kernel_size=3, stride=2, padding=1
Conv3/BatchNorm	out_channels=256, kernel_size=1, stride=1,
Conv4/BatchNorm	out_channels=256, kernel_size=1, stride=1,
STAGE1	**BTNK2:**×2
Conv1/BatchNorm	out_channels=64, kernel_size=1, stride=1,
Conv2/BatchNorm	out_channels=64, kernel_size=3, stride=2, padding=1
Conv3/BatchNorm	out_channels=256, kernel_size=1, stride=1,
STAGE2	**BTNK1,BTNK2 ×3**
STAGE3	**BTNK1,BTNK2 ×5**
STAGE4	**BTNK1,BTNK2 ×2**
Fully connected	out_feature=256
Fully connected	out_feature=C

我们提出的这种深度域适应网络有以下特点：(1) 该方法对卷积网络进行多层域适应。在深度域适应的过程中，根据所使用的骨干网络的不同，对不同的全连接层进行域适应，从本质上弥合了数据的边缘分布和条件分布的领域差异。(2) 该方法采用多核的方法进行 RKHS 的核选择。内核的选择对差异度量的能力至关重要，不同的内核可以在不同的 RKHS 中嵌入概率分布，因此通过不同核的线性组合来选择一个最优的核可以强调不同阶的充分统计量。(3) 该方法增加了神经网络的鉴别性。使用 MKDJP-MMD 作为域适应损失函数，该度量一方面能够减少相同类别之间的差异，另一方面能够增加不同类别之间的距离，在反向传播过程中这种特性能够增加神经网络的可迁移性和可鉴别性。

4.4 数值实验

本节通过跨域故障诊断的实验来评估所提出深度联合概率适应网络，我们分别使用了 AlexNet 卷积网络和 ResNet50 卷积网络作为骨干网络验证方法的有效性。首先给出了实验的实施细节，然后介绍了本章节使用的另一个轴承故障诊断数据集。最后进行了大量的数值实验来表明我们方法的有效性。本实验所有的方法都在 PyTorch 的框架下进行构建，硬件环境为搭配 RTX3090，8GB 显存 GPU，以及锐龙 16 核 4-GHz CPU 的设备。

4.4.1 实施细节

(1) 对比的方法

在实验部分，我们提出的 DJPAN 方法将会与具有代表性的深度迁移学习的代表性方法进行比较，DDC[61]，DeepCoral[69]，DAN[62]，DANN[70]，DSAN[63]以及批核范数最大 BNM[64]。为了保证实验的公平性，所有的方法在相同的深度神经网络架构上进行比较。

(2) 参数设置

对于所有方法的网络训练，我们使用小批量 SGD 和学习速率退火策略来完成网络优化，其中 SGD 的动量为 0.9。由于通过网格搜索来选择学习率 η_p 的计算成本较高，我们选择在 SGD 过程中通过 $\eta_p = \eta_0/(1+\alpha p)^{\varepsilon}$ 来调整学习速率。其中，p 为从 0 到 1 线性变化的训练进度，$\eta_0 = 0.01$，$\alpha = 10$，$\varepsilon = 0.75$，这种方式能够促进源域

的收敛并且降低训练误差。此外，为了抑制训练初期的噪声激活，我们没有选择将迁移损失的权衡因子 λ 和类间差异的权衡因子 μ 进行固定，而是逐步将其从 0.01 变为 1 进行选择。这种渐进策略显著稳定了参数敏感性，简化了 DJPAN 的模型选择。

4.4.2 实验数据

在本次实验中，我们使用了 CWRU 和帕德伯恩大学 (Paderborn University，PU) 的轴承故障数据集。关于 CWRU 数据集在上一章节中已经进行了详细的介绍，此处我们将不再赘述。下面将介绍 PU 轴承故障数据集的详细信息以及在该数据集中迁移学习任务的划分方式。

(1) PU 数据集

PU 数据集[①]是由帕德伯恩大学的团队在轴承加速平台上得到的轴承故障数据集，图4.5中给出了该数据集的实验平台。试验台的组成模块从右到左依次为：飞轮和负载电机、滚动轴承测试模块、扭矩测量轴以及电动机，将不同损伤类型的滚动轴承安装在滚动轴承测试模块中，在 64KHZ 的振动信号采样频率下通过加速度传感器来获取实验数据。

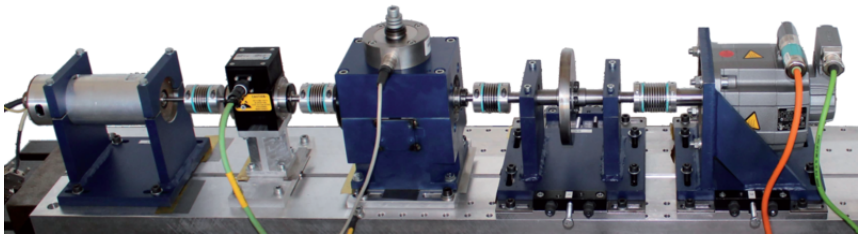


图 4.5　PU 数据集的实验平台

Figure 4.5　Experimental platform for the PU dataset

这些损伤类型包括人工诱发的轴承损伤和真实的轴承损伤，人工损伤是通过对轴承进行电火花、钻孔以及电动雕刻形成的，真实损伤则是通过加速寿命试验台得到的。通过改变驱动电机的转速、试验轴承的径向力和传动系统的负载力矩，PU 数据集总共包含四种不同的工况，每种工况对应一个迁移数据集，针对 PU 总共包含四个数据集 (PU0，PU1，PU2，PU3)，我们在表4.3中进行了总结。本次实验中，我们采用加速寿命实验下损伤的 13 个轴承研究不同工作条件下的迁移学习，轴承编号分别为 KA04，KA15，KA16，KA22，KA30，KB23，KB24，KB27，KI14，KI16，KI17，

① http://groups.uni-paderborn.de/kat/BearingDataCenter/

KI18，KI21，每个编号对应一种故障类型，具体信息可以参考文献[100]。

表 4.3　PU 数据集迁移学习任务分类

Table 4.3　Classification of transfer learning tasks for the PU dataset

迁移任务	PU0	PU1	PU2	PU3
负载转矩 (Nm)	0.7	0.7	0.1	0.7
径向力 (N)	1000	1000	1000	400
转速 (Rpm)	1500	900	1500	1500

(2) 数据集预处理及划分

为了保存更多的故障信息和更直观地解释不同故障类型的区别，我们在本章节采用与第三章相同的数据预处理方式，将一维的振动信号二维图像化，生成轴承故障信号的图像数据集。图4.6展示了在相同工况下，PU 数据集的 13 种不同故障类型的图像。同样地，我们将不同工况下的数据划分为含有标签信息的源域和不包含标签信息的目标域，来进行无监督迁移学习模拟实际工况。对于 PU 数据集，根据负载情况的不同获得了 4 个不同工作条件下的图像数据集，即 PU0、PU1、PU2 和 PU3。针对这 4 个迁移数据集，任务 PU0→PU2 表示源域 PU0 向目标域 PU2 的迁移，总共有 12 个迁移学习任务。

4.4.3　实验分析

(1) 实验结果

表4.4展示了基于 AlexNet 卷积网络的无监督深度域适应方法在 CWRU 数据集上的分类精度结果。结果表明这类深度域适应的方法与第三章的线性方法相比，性能提升较大，这说明将迁移学习的域适应损失嵌入卷积网络的方法能有效提升跨工况故障诊断的结果。与其他的深度迁移学习方法相比，我们提出的 DJPAN 模型在大多数跨工况故障诊断任务中都取得了最优的分类结果。与对抗式学习方法 DANN 相比，DJPAN 的平均分类精度提高了 1.42%，与同类型中最好的的深度域适应方法 DSAN 相比，DJPAN 的平均分类精度提升了 6.80%。这表明在 AlexNet 的架构下对于 CWRU 数据集，增加不同种类数据的类间距，减小相同种类数据的类内距确实能够增加网

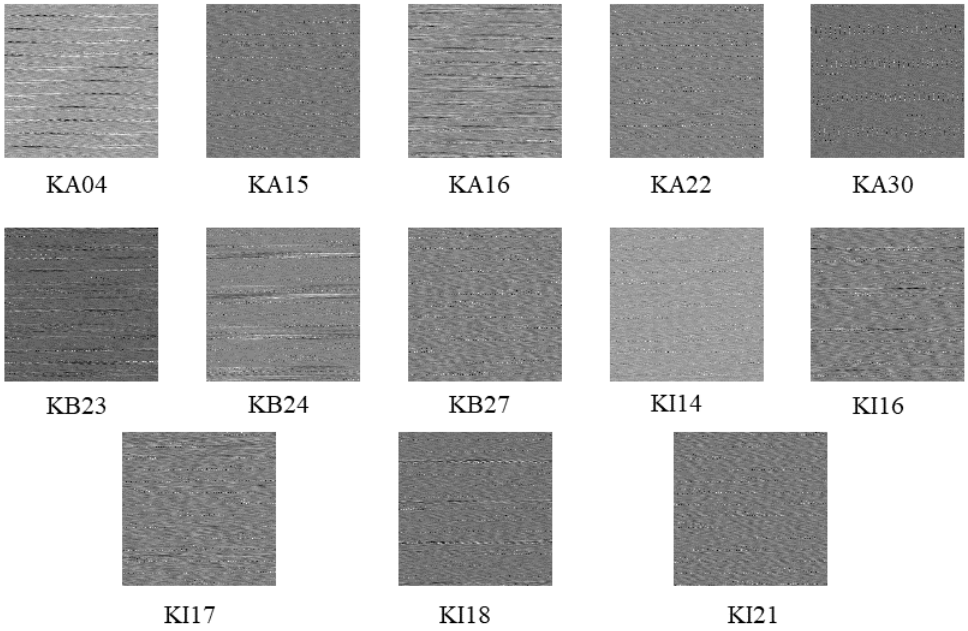


图 4.6　PU 数据集的 13 故障类型示例

Figure 4.6　Example of 13 fault types for the PU dataset

络的迁移性和鉴别性，有助于提高网络完成不同领域的分类任务。此外，所有使用多核度量的方法，如 DAN，DSAN，DJPAN 性能都优于使用单核方法的 DDC。具体地，DJPAN 的准确率相较于 DDC 更是提升了 17%，这说明了多核度量方式能够优化核函数的选择，提高差异度量的精度。

表4.5展示了基于 ResNet50 卷积网络的无监督深度域适应方法在 CWRU 数据集上的分类精度。我们可以观察到，在网络中增加域适应损失的方法同样适用于更为复杂的网络。在 ResNet50 卷积网络的架构下分类精度进一步提升，大大优于基于 AlexNet 的方法。这说明更深的网络架构能够学习到更多的可迁移的深度特征，因此增加网络复杂度，提高网络深度的可以进一步提升深度迁移学习方法的性能。此外对于 ResNet50 卷积网络架构，DJPAN 准确率比次好的深度迁移学习方法 BNM 增加了 1.67%，相较于 DSAN 增加了 2.21%。我们可以看到不同方法之间分类精度的差距减少了，这说明更深的网络架构能够缩小不同域适应方法的差距。因此，我们认为在训练过程中更多的卷积网络训练参数在一定程度上能够减少故障诊断中不同工况的差异。此外，所有增加了迁移损失的深度方法的性能都优于直接使用 ResNet50，这表明即使是非常深的网络也只能减少而不能完全消除域差异。

表4.6展示了基于 ResNet50 卷积网络的无监督深度域适应方法在 PU 数据集上的

表 4.4 AlexNet 网络的不同方法在 CWRU 数据集上的准确率 (%)

Table 4.4 Accuracy of different methods of AlexNet network on the CWRU dataset

迁移	方法							
任务	AlexNet	DDC	DeepCoral	DAN	DANN	DSAN	BNM	DJPAN
SP0 → SP1	74.82	75.18	73.57	82.50	88.84	74.91	88.21	**90.98**
SP0 → SP2	71.09	72.13	77.12	78.06	81.45	82.11	85.78	**89.74**
SP0 → SP3	76.01	76.28	72.99	74.08	81.32	**84.25**	71.25	78.93
SP1 → SP0	44.42	47.38	59.68	78.74	**88.27**	75.69	73.98	81.98
SP1 → SP2	64.41	68.64	68.08	85.40	**99.44**	92.09	82.77	96.23
SP1 → SP3	67.22	77.47	74.48	83.42	86.45	84.98	**87.73**	86.27
SP2 → SP0	58.91	61.20	65.59	67.21	69.30	64.92	69.49	**81.22**
SP2 → SP1	70.89	75.36	78.30	92.41	**96.25**	85.71	91.96	95.45
SP2 → SP3	68.32	75.27	85.35	90.11	81.32	87.82	79.58	**90.29**
SP3 → SP0	61.98	64.92	65.97	71.12	73.21	72.16	69.59	**79.03**
SP3 → SP1	57.86	62.59	62.41	68.30	72.14	61.88	**74.73**	70.89
SP3 → SP1	52.26	54.71	54.33	75.33	**80.13**	67.14	76.46	74.20
Average	64.02	67.59	69.82	78.89	83.18	77.81	79.29	**84.60**

表 4.5 ResNet50 网络的不同方法在 CWRU 数据集上的准确率 (%)

Table 4.5 Accuracy of different methods of ResNet50 network on the CWRU dataset

迁移任务	方法							
	ResNet50	DDC	DeepCoral	DAN	DANN	DSAN	BNM	DJPAN
SP0 → SP1	89.55	90.54	90.36	92.41	93.93	95.83	**97.68**	95.63
SP0 → SP2	76.84	79.94	90.11	96.99	97.27	99.81	**99.91**	97.63
SP0 → SP3	87.91	90.38	91.67	92.86	91.85	92.12	**98.72**	98.08
SP1 → SP0	67.30	70.07	74.64	82.17	91.42	97.14	93.90	**98.47**
SP1 → SP2	74.77	79.38	79.47	92.47	99.44	**99.81**	97.83	99.72
SP1 → SP3	69.96	84.07	80.31	90.38	95.70	98.08	95.60	**99.45**
SP2 → SP0	67.49	78.17	78.65	91.90	91.80	86.46	88.37	**94.95**
SP2 → SP1	83.04	88.57	88.04	90.45	88.04	**96.34**	92.77	94.64
SP2 → SP3	82.14	94.69	95.05	97.99	**99.35**	98.72	99.08	98.63
SP3 → SP0	66.25	73.98	74.07	86.60	80.17	86.56	88.66	**90.85**
SP3 → SP1	62.50	68.57	73.21	86.07	88.39	87.77	89.55	**93.57**
SP3 → SP2	67.61	72.98	79.47	92.37	95.01	95.10	98.11	**98.31**
Average	74.61	80.95	82.92	91.05	92.69	94.48	95.01	**96.69**

表 4.6 ResNet50 网络的不同方法在 PU 数据集上的准确率 (%)

Table 4.6 Accuracy of different methods of ResNet50 network on the PU dataset

迁移任务	方法							
	ResNet50	DDC	DeepCoral	DAN	DANN	DSAN	BNM	DJPAN
PU0 → PU1	28.24	33.76	34.07	33.93	37.25	36.15	36.10	**37.47**
PU0 → PU2	36.97	41.62	41.51	38.56	40.82	**53.59**	39.26	42.59
PU0 → PU3	23.69	26.70	29.66	31.28	30.97	31.02	31.09	**31.54**
PU1 → PU0	13.92	21.37	19.11	21.90	23.73	20.53	19.98	**24.11**
PU1 → PU2	56.87	65.82	66.54	65.46	64.56	67.21	**71.46**	70.90
PU1 → PU3	28.85	28.39	29.32	29.81	37.56	38.73	**39.77**	35.79
PU2 → PU0	15.22	29.42	30.84	27.43	32.64	32.55	21.95	**33.05**
PU2 → PU1	59.12	58.19	60.38	63.21	67.01	56.92	65.14	**71.81**
PU2 → PU3	31.63	35.93	34.98	38.14	33.12	36.20	42.76	**43.97**
PU3 → PU0	13.29	13.65	13.61	14.93	13.22	15.63	**18.17**	17.48
PU3 → PU1	33.85	38.93	36.76	45.14	38.71	**60.40**	55.63	43.24
PU3 → PU2	39.79	40.77	39.51	41.26	44.92	**52.62**	50.44	46.44
Average	31.79	36.21	36.36	37.58	38.71	**41.79**	40.98	41.53

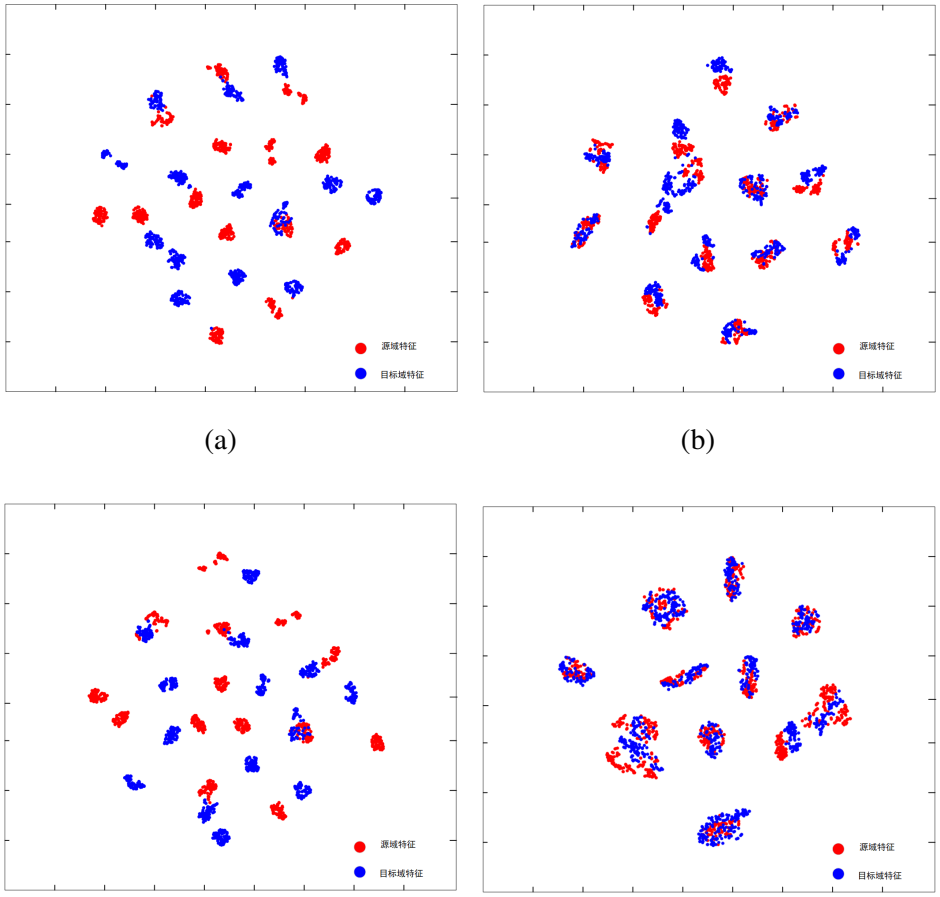
分类精度。PU 数据集相对于 CWRU 数据集数据量更大，分类任务更多。我们想在该数据集上探究在数据量增加，领域差异加大情况下 DJPAN 性能的有效性。实验结果表明，在大部分跨域故障诊断任务中，DJPAN 方法都取得了最优的故障分类精度，但是整体的平均精度略低于深度子域适应 DSAN 方法 0.51%。这是因为 DSAN 方法是对每一类分别进行对齐，所以在类别增加的情况下 DSAN 具有一定的优势。但是，DSAN 在 PU2→PU1 任务中出现了负迁移，性能低于直接使用 ResNet50 进行分类。我们认为，这是因为对于该任务不同类别之间的差异较小，DSAN 的子域 (子类别) 对齐策略无法很好地应对这一现象。而 DJPAN 在网络训练的过程中增加了不同类之间的距离，同时减小相同类之间的差异，增加了网络对的鉴别性和迁移性，这使得分类精度进一步提升。与其他方法相比，DJPAN 性能提升明显，相比于 BNM 准确率提高了 1.07%，相比于 DANN 准确率提高了 3.34%，这表明即使在数据量增加，分类任务增加的情况下，DJPAN 的策略仍然有效。

(2) 特征可视化分析

图4.7(a)-4.7(d)分别展示了 DAN，DSAN，BNM 以及 DJPAN 在 AlexNet 网络架构下的 SF0→SF2 迁移任务网络层激活的可视化结果。我们将源域特征与目标域特征在同一张图片下进行展示，以便说明领域自适应工作的有效性，红色点代表源域特征，蓝色点代表目标域特征。对于 DAN 方法，图4.7(a)中部分类别的源域特征和目标域特征重合，这说明 DAN 方法对域适应有一定的促进效果，但是对于不同的类别聚类效果较差，这导致分类精度较低。对于 DSAN 方法，图4.7(b)中大部分的源域特征和目标域特征重合，这表明该方法具有较好的域适应效果，但是不同种类的特征也出现重合，也导致了分类精度的不理想。对于 BNM 方法，图4.7(c)显示能够减少部分类别之间的差异，不同种类之间的距离也较远，但是聚类的效果并不理想。对于 DJPAN，图4.7(d)的结果显示该方法不仅减少了源域和目标域的差异，具有良好的迁移性，而且不同种类之间的距离较远，聚类效果理想，这说明 DJPAN 方法中减少类内距增加类间距的策略有助于完成迁移分类任务。

(3) 跨工况差异分析

在迁移学习研究中，$\mathcal{A}$-distance 是一种分布差异的度量标准，通常被定义为 $d_{\mathcal{A}} = 2(1-2\epsilon)$，其中 ϵ 是区分源域和目标域分类器 (如核 SVM 等) 的泛化误差。如果两个域之间的差异越小，那么 ϵ 的就越大，这在一定程度上能反映两个域之间的分布差

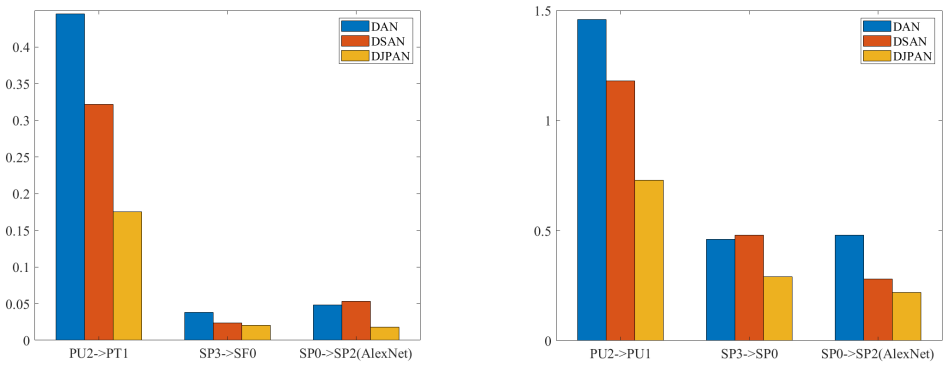


(a) (b)

(c) (d)

图 4.7 不同方法 t-SNE 特征可视化的结果

Figure 4.7 Results of t-SNE feature visualization by different methods



(a) MMD (b) $\mathcal{A}$-distance

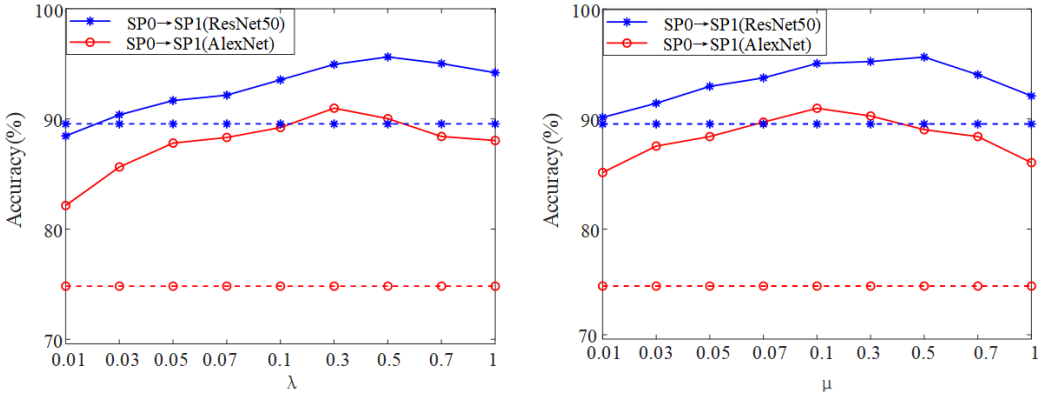
图 4.8 差异分析

Figure 4.8 Analysis of distribution discrepancy

异。我们将 $d_{\mathcal{A}}$ 和 d_{MMD} 作为本节中分布差异的度量标准，分析了不同工况下的数据进行域适应后的 d_{MMD} 和 $d_{\mathcal{A}}$ 的值。图4.8(a)展示了 DAN，DSAN 以及 DJPAN 在不同跨域分类任务中进行域适应后，网络输出的层激活的 d_{MMD} 的值。我们可以发现在经过 DJPAN 后网络输出层激活的 MMD 要比其他两种方法小的多，这表明 DJPAN 可以更有效地缩小不同工况之间的差异。图4.8(b)展示了 DAN，DSAN 以及 DJPAN 的 $d_{\mathcal{A}}$ 的值，同样可以验证上述结论。此外，由于 PU 数据集中不同工况下的数据集差异较大，因此它们的 d_{MMD} 和 $d_{\mathcal{A}}$ 较高，这也解释了 PU 数据集中跨工况故障诊断任务效果不理想的原因。

(4) 参数敏感性分析

本小节分析了迁移损失的权衡因子 λ 和类间差异的权衡因子 μ 对 DJPAN 模型的灵敏度。图4.9(a)和图4.9(b)分别展示了 λ 和 μ 对不同网络框架下的 DJPAN 迁移精度的影响，其中虚线表示不同网络不进行域适应情况下的跨域分类精度。我们将参数范围设置为 $\{0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7, 1\}$，在实验过程中固定一个调整另一个参数直到出现最优结果。结果显示，DJPAN 的精度随着参数的变化先增后减，这表明了在迁移损失和网络分类损失之间进行适当的权衡能够提升网络跨域的性能。另一方面，类间距离和类内距离之间也需要根据迁移任务的不同进行权重分配，以此提升 MKDJP-MMD 度量的准确度。此外，结果显示 AlexNet 网络对于参数的变化更敏感，这说明域适应损失对简单的网络影响更大。



(a) MMD　　(b) $\mathcal{A}$-distance

图 4.9　参数敏感性分析

Figure 4.9　Analysis of parameter sensitivity

(5) 网络收敛性分析

本小节将七种不同的方法在同一网络设置下进行训练，并对其收敛性进行了验证，图4.10展示了基于 ResNet50 网络的七种方法在任务 SP3→SP1 上目标域的测试误差随着训练轮数增加而变化的折线图。结果表明 DJPAN 在训练轮数达到 15 轮左右模型收敛，相较于其他方法具有更快的收敛速度。此外，DJPAN 方法训练的第一轮目标域的误差在 0.7 左右，低于所有比较的方法，这表明我们的方法在网络训练初期就能减小源域和目标域的差异，验证了 MKDJP-MMD 的有效性。同时，不同方法在模型稳定时的误差与分类精度呈现正相关的规律，这与我们的实验结果相符。

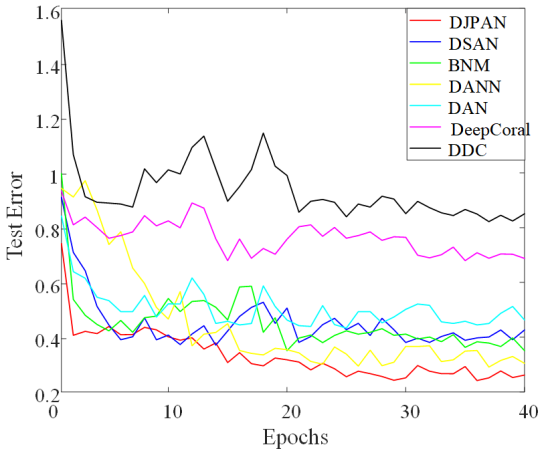


图 4.10　不同方法在相同训练轮数下的误差

Figure 4.10　Test errors of different methods for the same epochs

4.5　本章小结

在本章中，我们提出了一种新的用于轴承故障诊断的深度迁移学习方法。首先采用多核的策略通过多个核的线性组合来选择最优的核，大大提高了分布差异度量的精确度。然后将这种 MKDJP-MMD 的度量方法嵌入到神经网络的非线性运算中，充分发挥深度学习端到端的优势，搭建了两个基于深度联合概率的域适应网络。该网络通过增加类间距的方式，扩展了深度自适应网络的特征表示能力，能够自行构建图像到故障类型的映射关系。最后进行了大量的数值实验，与 DAN、DSAN 和 BNM 等最具有代表性的深度迁移学习方法相比，所提出的 DJPAN 在大多数跨工况轴承故障诊断任务中表现出更理想的性能。

第五章　总结与展望

5.1　总结

随着人工智能和数据分析技术的持续进步，智能故障诊断技术日益成为学术界的研究焦点。在实际工业场景中，由于设备的运行环境、设备负载以及设备类型的不同，待检测的故障信号通常与实验环境下采集的信号有一定的差异，具体表现为标签不足、特征分布不一致以及数据量不足等问题，这为故障诊断技术带来了挑战。本文针对目标域没有标签场景下的跨工况轴承故障诊断问题展开了一系列研究，提出了基于迁移子空间学习的故障诊断方法和基于深度域适应网络的故障诊断方法，主要工作有以下几点：

（1）全面回顾了深度迁移学习技术近十年的发展概况，对各类深度迁移学习方法进行系统地总结和概括。根据模型、函数和操作对象的不同，对其进行详细的分类，在此基础上分析了各类方法的优缺点。

（2）针对采集的轴承振动信号中存在大量噪声的问题，在迁移子空间学习模型的基础上，同时对高斯噪声和非高斯噪声进行矩阵建模，有效减少了数据中各类不同的噪声对故障诊断结果的影响。此外，通过直接从待诊断的数据样本中学习标签矩阵的方式改进了回归方法，增强了模型的鲁棒性。同时，利用交替方向乘子法开发了有效的优化算法来求解所提出的模型。总的来说，我们提出的基于迁移子空间学习的故障诊断方法，建立了低秩稀疏的、松弛回归的鲁棒迁移学习模型，能够有效的完成高噪声情况下的无监督跨域工况故障诊断任务。

（3）针对现有的深度迁移网络鉴别性较低的问题，在深度不同的网络模型中使用 MKDJP-MMD 度量代替了目前常用的 MMD 度量，在增加网络迁移性的同时，通过增加不同种类之间特征的距离来增强网络的鉴别性。具体来说，我们对 DJP-MMD 度量方式进行了优化，在 RKHS 的变换过程中，通过多个核的线性组合来选择最优的核函数，增加了度量的准确性。将优化后的度量准则嵌入神经网络的损失函数，通过增加类间距的方式，进一步减少了源域和目标域的差异。总的来说，我们提出的这种深度域适应网络，利用深度学习的方法，为解决目标域没有标签的跨域故障诊断故障问题提供了一种端到端的学习方式。

5.2 展望

针对轴承故障诊断应用中，训练数据和待检测数据存在差异、数据标签不足的问题，从线性子空间方法到非线性神经网络方法，开展了一系列的研究，提出了两种不同的跨工况故障诊断方法。随着研究的深入，我们发现仍有许多问题需要深入探讨，主要有以下几个方面：

（1）跨工况故障诊断的负迁移问题：迁移学习的有效性建立在以下两个假设的基础上：一是源域和目标域中的数据分布具有相关性。二是两个域中的学习任务是相似的。如果目标域的先验知识是从一个不相关的源域中迁移过来的，那故障诊断的有效性将大打折扣，甚至引发负迁移。例如，各类方法在 PU 数据集诊断效果普遍较差，甚至出现负迁移的情况。因此，如何结合不同方法的优点，将 DJPAN 增加类间距的特点与 DSAN 的子域对齐的特点相结合，成为下一步工作的重点。

（2）故障诊断数据隐私保护问题：在迁移模型的训练过程中，不可避免的需要对已知工况的数据和待检测工况的数据进行学习。然而设备的故障数据在某些方面能够反映设备的信息，这些信息可能会被利用，泄露国家机密，对工业生产安全造成隐患。联邦学习能有效保护设备隐私和数据安全，可以让参与方在不共享数据的基础上联合建模。因此未来将联邦学习与跨域故障诊断技术相结合，来保护设备隐私，确保敏感信息不被泄露，是一个值得关注的研究。

（3）不同信号之间的知识迁移问题：近年来，传感器技术发展迅速，不同类型的传感器被广泛应用于轴承故障检测。目前对于跨工况故障诊断的研究，大多是使用相同信号类型进行知识迁移。但是某些情况下，如果源域数据是振动信号，目标域中只有电流或温度信号可用。此时由于信号类型的不同，故障数据的分布、物理意义和特征嵌入都是不同的。因此，如何在不同类型的信号之间进行迁移，成为了一个具有研究价值的课题。

（4）跨工况迁移的可解释性问题：知识迁移的可解释研究一直以来都是迁移学习的重点研究内容，尤其是在深度迁移学习技术迅速发展的阶段。不同工况下采集的故障数据可迁移的部分有多少，迁移模型的训练过程是否透明可知，这些都是迁移可解释研究的关注内容。在实际工业环境中，操作人员对模型所依据的判断逻辑和基本原则的深入理解是增强模型信任度的前提。因此，增加知识迁移的可解释性研究，对于推动迁移学习在工业故障诊断应用的长期发展极为关键。

插图索引

图 1.1 PHM 系统流程 1
图 1.2 工业安全事故 2
图 1.3 故障诊断技术分类 3
图 1.4 基于机器学习的故障诊断技术分类 6
图 1.5 传统机器学习与迁移学习的工业监控流程对比 8
图 1.6 文章主要内容与章节安排 9
图 2.1 迁移学习的学习过程 12
图 2.2 迁移学习发展的时间线 14
图 2.3 深度迁移学习方法分类 15
图 2.4 基于双流网络的参数微调网络 16
图 2.5 IPT 方法架构 18
图 2.6 基于差异的深度迁移学习网络 21
图 2.7 DSAN 网络架构 21
图 2.8 DANN 网络架构 24
图 2.9 DAAN 网络架构 25
图 2.10 逆向学习推理方法框架 26
图 3.1 所提方法的框架 29
图 3.2 高斯噪声图 32
图 3.3 CWRU 数据集的实验平台 42
图 3.4 JNU 数据集的实验平台 43

图 3.5 CWRU 数据集的 10 种故障类型示例 44
图 3.6 JNU 数据集的 4 种故障类型示例 44
图 3.7 所使用的 VGG19 网络架构 45
图 3.8 在 JNU 数据集中添加高斯噪声和非高斯噪声的影响..... 50
图 3.9 TSL-2 在不同参数值下的性能..... 51
图 3.10 不同方法的模型稳定性分析 52
图 3.11 收敛条件的值随着迭代次数的变化..... 53
图 4.1 判别联合概率域适应与全局域适应的区别..... 56
图 4.2 深度联合概率适应网络..... 65
图 4.3 基于 AlexNet 的深度联合概率适应网络..... 68
图 4.4 基于 ResNet50 的深度联合概率适应网络 70
图 4.5 PU 数据集的实验平台 72
图 4.6 PU 数据集的 13 故障类型示例 74
图 4.7 不同方法 t-SNE 特征可视化的结果..... 79
图 4.8 差异分析 79
图 4.9 参数敏感性分析 80
图 4.10 不同方法在相同训练轮数下的误差..... 81

表格索引

表 2.1 传统迁移学习方法的总结 14
表 2.2 基于模型的深度迁移学习方法总结 15
表 2.3 基于差异的深度迁移学习方法总结 19
表 2.4 基于差异的深度迁移学习方法总结 23
表 3.1 CWRU 数据集的 10 种故障类型 42
表 3.2 VGG19 网络的参数 45
表 3.3 不同方法在 CWRU 数据集上的准确率 (%) 47
表 3.4 不同方法在 JNU 数据集上的准确率 (%) 48
表 3.5 在 CWRU 数据集上加入不同程度的噪声对迁移任务的影响 (%) 49
表 4.1 AlexNet 网络的参数 69
表 4.2 ResNet50 网络的参数 70
表 4.3 PU 数据集迁移学习任务分类 73
表 4.4 AlexNet 网络的不同方法在 CWRU 数据集上的准确率 (%) 75
表 4.5 ResNet50 网络的不同方法在 CWRU 数据集上的准确率 (%) 76
表 4.6 ResNet50 网络的不同方法在 PU 数据集上的准确率 (%) 77

参考文献

[1] 周济. 智能制造——“中国制造 2025”的主攻方向[J]. 中国机械工程, 2015, 26 (17): 2273-2284.

[2] 李军宁, 罗文广, 陈武阁. 面向振动信号的滚动轴承故障诊断算法综述[J]. 西安工业大学学报, 2022, 42(02): 105-122.

[3] JARDINE A K, LIN D, BANJEVIC D. A review on machinery diagnostics and prognostics implementing condition-based maintenance[J]. Mechanical Systems and Signal Processing, 2006, 20(7): 1483-1510.

[4] 张少敏, 毛冬, 王保义. 大数据处理技术在风电机组齿轮箱故障诊断与预警中的应用[J]. 电力系统自动化, 2016, 40(14): 129-134.

[5] 王政, 赵娟. 基于信号处理的滚动轴承故障诊断[J]. 电工技术, 2023(02): 150-152+157.

[6] 王守鹏, 赵冬梅. 基于免疫克隆约束多目标优化方法的电网故障诊断[J]. 电网技术, 2017, 41(12): 4061-4068.

[7] BEARD R V. Failure accomodation in linear systems through self-reorganization.[D]. Massachusetts Institute of Technology, 1971.

[8] ISERMANN R. Model-based fault-detection and diagnosis-status and applications [J]. Annual Reviews in control, 2005, 29(1): 71-85.

[9] PATTON R, WILLCOX S, WINTER J. Parameter-insensitive technique for aircraft sensor fault analysis[J]. Journal of Guidance, Control, and Dynamics, 1987, 10(4): 359-367.

[10] QIN C, LIN W J. Adaptive event-triggered fault-tolerant control for Markov jump nonlinear systems with time-varying delays and multiple faults[J]. Communications in Nonlinear Science and Numerical Simulation, 2024, 128: 107655.

[11] WANG Z, SHEN Y, ZHANG X. Attitude sensor fault diagnosis based on kalman filter of discrete-time descriptor system[J]. Journal of Systems Engineering and Electronics, 2012, 23(6): 914-920.

[12] XU A, ZHANG Q. Residual generation for fault diagnosis in linear time-varying systems[J]. IEEE Transactions on Automatic Control, 2004, 49(5): 767-772.

[13] VAN GORP J, DEFOORT M, DJEMAI M, et al. Fault detection based on higher-order sliding mode observer for a class of switched linear systems[J]. IET Control Theory & Applications, 2015, 9(15): 2249-2256.

[14] ADHYARU D M. State observer design for nonlinear systems using neural network [J]. Applied Soft Computing, 2012, 12(8): 2530-2537.

[15] KAZEMI H, YAZDIZADEH A. Optimal state estimation and fault diagnosis for a class of nonlinear systems[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(2): 517-526.

[16] LI Y, KARIMI H R, ZHONG M, et al. Fault detection for linear discrete time-varying systems with multiplicative noise: The finite-horizon case[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2018, 65(10): 3492-3505.

[17] DING S X. Model-based fault diagnosis techniques: design schemes, algorithms, and tools[M]. Springer Science & Business Media, 2008.

[18] ZHONG M, XUE T, DING S X. A survey on model-based fault diagnosis for linear discrete time-varying systems[J]. Neurocomputing, 2018, 306: 51-60.

[19] CHO S, JIANG J. Optimal fault classification using Fisher discriminant analysis in the parity space for applications to NPPs[J]. IEEE Transactions on Nuclear Science, 2018, 65(3): 856-865.

[20] ABBASI A R, MAHMOUDI M R, AVAZZADEH Z. Diagnosis and clustering of power transformer winding fault types by cross-correlation and clustering analysis of FRA results[J]. IET Generation, Transmission & Distribution, 2018, 12(19): 4301-4309.

[21] ABAD M R A A, MOOSAVIAN A, KHAZAEE M. Wavelet transform and least square support vector machine for mechanical fault detection of an alternator using vibration signal[J]. Journal of Low Frequency Noise, Vibration and Active Control, 2016, 35(1): 52-63.

[22] WEN X, XU Z. Wind turbine fault diagnosis based on ReliefF-PCA and DNN[J]. Expert Systems with Applications, 2021, 178: 115016.

[23] CONG F, ZHONG W, TONG S, et al. Research of singular value decomposition based on slip matrix for rolling bearing fault diagnosis[J]. Journal of Sound and Vibration, 2015, 344: 447-463.

[24] YU J, XU Y, YU G, et al. Fault severity identification of roller bearings using flow graph and non-naive Bayesian inference[J]. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2019, 233 (14): 5161-5171.

[25] 朱亚军, 胡建钦, 李武. 基于频域窗函数的短时傅里叶变换及其在机械冲击特征提取中的应用[J]. 机床与液压, 2021, 49(18): 177-182.

[26] LI S, XIN Y, LI X, et al. A review on the signal processing methods of rotating machinery fault diagnosis[C]//2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, 2019: 1559-1565.

[27] 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.

[28] AMARNATH M, SUGUMARAN V, KUMAR H. Exploiting sound signals for fault diagnosis of bearings using decision tree[J]. Measurement, 2013, 46(3): 1250-1256.

[29] KONAR P, CHATTOPADHYAY P. Bearing fault detection of induction motor using wavelet and support vector machines (SVM)[J]. Applied Soft Computing, 2011, 11 (6): 4203-4211.

[30] YU Y, JUNSHENG C, et al. A roller bearing fault diagnosis method based on EMD energy entropy and ANN[J]. Journal of Sound and Vibration, 2006, 294(1-2): 269-277.

[31] YUWONO M, QIN Y, ZHOU J, et al. Automatic bearing fault diagnosis using particle swarm clustering and Hidden Markov Model[J]. Engineering Applications of Artificial Intelligence, 2016, 47: 88-100.

[32] LIANG C M, LI Y W, LIU Y H, et al. Segmentation and weight prediction of grape ear based on SFNet-ResNet18[J]. Systems Science & Control Engineering, 2022, 10 (1): 722-732.

[33] CHOI D J, HAN J H, PARK S U, et al. Comparative study of CNN and RNN for motor fault diagnosis using deep learning[C]//2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA). IEEE, 2020: 693-696.

[34] ZHANG C, HE Y, YUAN L, et al. Analog circuit incipient fault diagnosis method using DBN based features extraction[J]. IEEE ACCESS, 2018, 6: 23053-23064.

[35] LI C, ZHANG S, QIN Y, et al. A systematic review of deep transfer learning for machinery fault diagnosis[J]. Neurocomputing, 2020, 407: 121-135.

[36] YAN R, SHEN F, SUN C, et al. Knowledge transfer for rotary machine fault diagnosis [J]. IEEE Sensors Journal, 2019, 20(15): 8374-8393.

[37] CHEN X, YANG R, XUE Y, et al. Deep transfer learning for bearing fault diagnosis: A systematic review since 2016[J]. IEEE Transactions on Instrumentation and Measurement, 2023.

[38] 庄福振, 罗平, 何清. 迁移学习研究进展[J]. 软件学报, 2015, 26: 26-39.

[39] ZHUANG F, QI Z, DUAN K, et al. A comprehensive survey on transfer learning[J]. Proceedings of the IEEE, 2020, 109(1): 43-76.

[40] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359.

[41] YAO Y, DORETTO G. Boosting for transfer learning with multiple sources[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 1855-1862.

[42] HUANG J, GRETTON A, BORGWARDT K, et al. Correcting sample selection bias by unlabeled data[J]. Advances in Neural Information Processing Systems, 2006, 19.

[43] FEI-FEI L, FERGUS R, PERONA P. One-shot learning of object categories[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4): 594-611.

[44] AYTAR Y, ZISSERMAN A. Tabula rasa: Model transfer for object category detection [C]//2011 International Conference on Computer Vision. IEEE, 2011: 2252-2259.

[45] PAN S J, TSANG I W, KWOK J T, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks, 2010, 22(2): 199-210.

[46] LONG M, WANG J, DING G, et al. Transfer feature learning with joint distribution adaptation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2013: 2200-2207.

[47] LONG M, WANG J, DING G, et al. Transfer joint matching for unsupervised domain adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1410-1417.

[48] WANG J, CHEN Y, FENG W, et al. Transfer learning with dynamic distribution adaptation[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11 (1): 1-25.

[49] GONG B, SHI Y, SHA F, et al. Geodesic flow kernel for unsupervised domain adaptation[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2066-2073.

[50] XU Y, FANG X, WU J, et al. Discriminative transfer subspace learning via low-rank and sparse representation[J]. IEEE Transactions on Image Processing, 2015, 25(2): 850-863.

[51] MIHALKOVA L, HUYNH T, MOONEY R J. Mapping and revising markov logic networks for transfer learning[C]//AAAI: Vol. 7. 2007: 608-614.

[52] YOSINSKI J, CLUNE J, BENGIO Y, et al. How transferable are features in deep neural networks?[J]. Advances in Neural Information Processing Systems, 2014, 27.

[53] CHOPRA S, BALAKRISHNAN S, GOPALAN R. Dlid: Deep learning for domain adaptation by interpolating between domains[C]//ICML Workshop on Challenges in Representation Learning: Vol. 2. Citeseer, 2013.

[54] ROZANTSEV A, SALZMANN M, FUA P. Beyond sharing weights for deep domain adaptation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(4): 801-814.

[55] HE K, GIRSHICK R, DOLLÁR P. Rethinking imagenet pre-training[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4918-4927.

[56] XIE Q, LUONG M T, HOVY E, et al. Self-training with noisy student improves imagenet classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10687-10698.

[57] ZOPH B, GHIASI G, LIN T Y, et al. Rethinking pre-training and self-training[J]. Advances in Neural Information Processing Systems, 2020, 33: 3833-3845.

[58] CHEN H, WANG Y, GUO T, et al. Pre-trained image processing transformer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12299-12310.

[59] BAO H, DONG L, PIAO S, et al. Beit: Bert pre-training of image transformers[A]. arXiv:2106.08254, 2021.

[60] YANG J, LIU J, XU N, et al. Tvt: Transferable vision transformer for unsupervised domain adaptation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 520-530.

[61] TZENG E, HOFFMAN J, ZHANG N, et al. Deep domain confusion: Maximizing for domain invariance[A]. arXiv:1412.3474, 2014.

[62] LONG M, CAO Y, WANG J, et al. Learning transferable features with deep adaptation networks[C]//International Conference on Machine Learning. PMLR, 2015: 97-105.

[63] ZHU Y, ZHUANG F, WANG J, et al. Deep subdomain adaptation network for image classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(4): 1713-1722.

[64] CUI S, WANG S, ZHUO J, et al. Fast batch nuclear-norm maximization and minimization for robust domain adaptation[A]. arXiv:2107.06154, 2022.

[65] NAM H, LEE H, PARK J, et al. Reducing domain gap by reducing style bias[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8690-8699.

[66] YOON J, KANG D, CHO M. Semi-supervised domain adaptation via sample-to-sample self-distillation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 1978-1987.

[67] YU C, WANG J, LIU C, et al. Learning to match distributions for domain adaptation [A]. arXiv:2007.10791, 2020.

[68] TZENG E, HOFFMAN J, DARRELL T, et al. Simultaneous deep transfer across domains and tasks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4068-4076.

[69] SUN B, SAENKO K. Deep coral: Correlation alignment for deep domain adaptation [C]//Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. Springer, 2016: 443-450.

[70] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. Journal of Machine Learning Research, 2016, 17(59): 1-35.

[71] YU C, WANG J, CHEN Y, et al. Transfer learning with dynamic adversarial adaptation network[C]//2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019: 778-786.

[72] LONG M, CAO Z, WANG J, et al. Conditional adversarial domain adaptation[J]. Advances in Neural Information Processing Systems, 2018, 31.

[73] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//roceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2223-2232.

[74] TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7167-7176.

[75] SHRIVASTAVA A, PFISTER T, TUZEL O, et al. Learning from simulated and unsupervised images through adversarial training[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2107-2116.

[76] KURMI V K, SUBRAMANIAN V K, NAMBOODIRI V P. Exploring dropout discriminator for domain adaptation[J]. Neurocomputing, 2021, 457: 168-181.

[77] SAITO K, WATANABE K, USHIKU Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3723-3732.

[78] HOFFMAN J, TZENG E, PARK T, et al. Cycada: Cycle-consistent adversarial domain adaptation[C]//International Conference on Machine Learning. Pmlr, 2018: 1989-1998.

[79] DUMOULIN V, BELGHAZI I, POOLE B, et al. Adversarially learned inference[A]. arXiv:1606.00704, 2016.

[80] LU Y, WANG W, YUAN C, et al. Manifold transfer learning via discriminant regression analysis[J]. IEEE Transactions on Multimedia, 2021, 23: 2056-2070.

[81] LIU Z, SHI K, NIU D, et al. Dynamic classifier approximation for unsupervised domain adaptation[J]. Signal Processing, 2023, 206: 108915.

[82] ZHAN S, SUN W, KANG P. Robust latent common subspace learning for transferable feature representation[J]. Electronics, 2022, 11(5): 810.

[83] WANG L, MA S, HAN Q. Enhanced sparse low-rank representation via nonconvex regularization for rotating machinery early fault feature extraction[J]. IEEE/ASME Transactions on Mechatronics, 2022, 27(5): 3570-3578.

[84] ZHANG Q, LV Y, YUAN R, et al. A local transient feature extraction method via periodic low rank dynamic mode decomposition for bearing incipient fault diagnosis [J]. Measurement, 2022, 203: 111973.

[85] MA J, HUANG W, LIAO Y, et al. Sparse low-rank matrix estimation with nonconvex enhancement for fault diagnosis of rolling bearings[J]. IEEE Transactions on Instrumentation and Measurement, 2023.

[86] LI Q. A comprehensive survey of sparse regularization: Fundamental, State-of-the-art Methodologies and Applications on Fault Diagnosis[J]. Expert Systems with Applications, 2023: 120517.

[87] CAI J F, CANDÈS E J, SHEN Z. A singular value thresholding algorithm for matrix completion[J]. SIAM Journal on Optimization, 2010, 20(4): 1956-1982.

[88] JIN J, QIN Z, YU D, et al. Relaxed least square regression with $\ell_{2,1}$-norm for pattern classification[J]. International Journal of Wavelets, Multiresolution and Information Processing, 2023.

[89] WEN L, LI X, GAO L, et al. A new convolutional neural network-based data-driven fault diagnosis method[J]. IEEE Transactions on Industrial Electronics, 2018, 65(7): 5990-5998.

[90] ZHAO K, JIANG H, WU Z, et al. A novel transfer learning fault diagnosis method based on manifold embedded distribution alignment with a little labeled data[J]. Journal of Intelligent Manufacturing, 2022, 33: 151-165.

[91] GUO L, LEI Y, XING S, et al. Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data[J]. IEEE Transactions on Industrial Electronics, 2018, 66(9): 7316-7325.

[92] YANG B, LEI Y, JIA F, et al. An intelligent fault diagnosis approach based on transfer learning from laboratory bearings to locomotive bearings[J]. Mechanical Systems and Signal Processing, 2019, 122: 692-706.

[93] WU Z, JIANG H, ZHAO K, et al. An adaptive deep transfer learning method for bearing fault diagnosis[J]. Measurement, 2020, 151: 107227.

[94] CAO X, WANG Y, CHEN B, et al. Domain-adaptive intelligence for fault diagnosis based on deep transfer learning from scientific test rigs to industrial applications[J]. Neural Computing and Applications, 2021, 33: 4483-4499.

[95] LI X, JIANG H, XIE M, et al. A reinforcement ensemble deep transfer learning network for rolling bearing fault diagnosis with Multi-source domains[J]. Advanced Engineering Informatics, 2022, 51: 101480.

[96] ZHANG Y, REN Z, ZHOU S, et al. Supervised contrastive learning-based domain adaptation network for intelligent unsupervised fault diagnosis of rolling bearing[J]. IEEE/ASME Transactions on Mechatronics, 2022, 27(6): 5371-5380.

[97] SMOLA A, GRETTON A, SONG L, et al. A Hilbert space embedding for distributions[C]//International Conference on Algorithmic Learning Theory. Springer, 2007: 13-31.

[98] ZHANG W, WU D. Discriminative joint probability maximum mean discrepancy (DJP-MMD) for domain adaptation[C]//2020 International Joint Conference on Neural Networks (IJCNN). 2020: 1-8.

[99] GRETTON A, SEJDINOVIC D, STRATHMANN H, et al. Optimal kernel choice for large-scale two-sample tests[J]. Advances in Neural Information Processing Systems, 2012, 25.

[100] LESSMEIER C, KIMOTHO J K, ZIMMER D, et al. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification[C]//PHM Society European Conference: Vol. 3. 2016.

攻读专业硕士学位期间取得的研究成果

一、学术论文

[1] **YU F**, XIU X, LI X, et al. Robust transfer subspace learning based on low-rank and sparse representation for bearing fault diagnosis[J]. Measurement Science and Technology, 2024, 35(6): 066204.(第一作者，已发表，SCI，JCR Q1)

[2] **YU F**, XIU X, LI Y. A survey on deep transfer learning and beyond[J]. Mathematics, 2022, 10(19): 3619. (第一作者，已发表，SCI，JCR Q1)

[3] **YU F**, XIU X. Robust transfer subspace learning: a novel data-driven scheme for fault diagnosis[C]//2023 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS). IEEE, 2023: 1-5. (第一作者，已发表，EI)

二、知识产权

[1] 修贤超, 于福超, 李云辉, 等. 基于鲁棒转移子空间学习的故障诊断方法及系统 [P]. 上海市: CN202310850712.1, 2023-10-10.

[2] 修贤超, 于福超, 李云辉, 等. 小样本下基于深度适应网络的轴承故障诊断方法和系统 [P]. 上海市: CN202310927514.0, 2023-10-27.

[3] 轴承振动信号可视化及特征提取软件 V1.0, 授权号: 2023SR1659938.

三、科研项目

[1] 国家自然科学基金面上项目, 12371306, 大规模黎曼流形稀疏优化算法及应用, 2024.01-2027.12.

[2] 国家自然科学基金面上项目, 62173003, 基于关键性能指标的过程监测与性能恢复理论与方法研究, 2022.01-2025.12.

[3] 上海市教育委员会科研创新计划项目 (重大), 2023ZKZD47, 跨域开放作业环境下智慧农场无人系统群智协同关键技术研究, 2023.01-2027.12.

四、获得荣誉

[1] 上海大学优秀毕业生, 2024.02.

[2] 研究生国家奖学金, 2023.12.

[3] 上海大学优秀学生, 2023.09.

致　谢

现在是西历 2024 年，农历甲辰年三月十五日，不出意外的话，我的学生生涯会随着正文最后一个句号的终笔就此结束。从道义南大街 37 号到上大路 99 号，在我深深向下扎根寻知求果时，这个阶段给予了我前进所需要的所有养料，这一路何其精彩，这一程何其有幸。我一直认为，任何阶段经历的任何事情，无论好坏都是一场修行。很喜欢加缪的一段话，只要我还一直读书，我就能够一直理解自己的痛苦，一直与自己的无知、狭隘、偏见、阴暗见招拆招。很多人说和自己握手言和，我不要做这样的人，我要拿石头打磨我这块石头。一直读书，一直痛苦，一直爱着从痛苦荒芜里生出的喜悦。乘兴而来，尽兴而归。

昨日种种历历在目，明日种种譬如新生。一直以来我自认为很幸运，占尽人间怙恩后，全数归还流落身。在这个真正的与校园时光告别之际，我要感谢陪我走过三年桎梏的所有人，无论是挚友、老师、同学还是有过几面之缘的同路人，与你们的经历丰满了我三年的读研时光，你们带给了我太多的欢乐与感动。三年的时间里，我竟然如此幸运与这么多人相识结为挚友，我们结伴同行至此，互相扶持至此，感谢你们的包容和鼓励，与你们的经历将成为我记忆深处最宝贵的一部分埋藏心底，历久弥香，永不褪色。这里，我要特别感谢我的导师修贤超老师在我这张科研白纸上的画作，感谢您三年的教导之恩。您严谨务实的学术态度使我受益良多，您的用心雕琢使我这块朽木略微成器，我相信您的言传身教会对我的未来产生长远的影响。最后，是我的父母，我不知该从何感谢你们，从起初的蹒跚学步到如今的弱冠之年，你们虽然有时是那么固执，但是仍默默支持我至今，感谢你们无理由的支持与信任。

前路漫漫亦灿灿，这里的终点将是我新生活的起点。最后，我想引用当年明月的一句话送给未来的自己，“无论怎样成功的方式只有一个，那就是按照自己喜欢的方式度过一生。”希望此去一生坦荡，一生纯善。

于福超

机自大楼 345

2023 年 4 月 23 日

附录 A　本文使用的英文缩写[①]

预测和健康管理系统	Prognostics and Health Management，PHM
主成分分析	Principal Component Analysis，PCA
傅里叶变换	Fourier Transformation，FT
小波变换	Wavelet Transform，WT
决策树	Decision Tree，DT
支持向量机	Support Vector Machines，SVM
人工神经网络	Artificial Neural Networks，ANN
隐式马尔可夫模型	Hidden Markov Model，HMM
卷积神经网络	Convolutional Neural Networks，CNN
循环神经网络	Recursive Neural Network，RNN
深度信念网络	Deep Belief Network，DBN
迁移成分分析	Transfer Component Analysis，TCA
联合分布自适应	Joint Distribution Adaptation，JDA
迁移联合匹配	Transfer Joint Matching，TJM
平衡分布自适应	Balance Distribution Adaptation，BDA
测地线流形核	Geodesic Flow Kernel，GFK
子空间对齐法	Subspace Alignment，SA
马尔可夫逻辑网络	Markov Logic Network，MLN
最大均值差异	Maximum Mean Difference，MMD
再生核希尔伯特空间	Reproducing Kernel Hilbert Space，RKHS
深度域混淆	Deep Domain Confusion，DDC
深度适应网络	Deep Adaptation Network，DAN
局部最大均值差异	Local Maximum Mean Difference，LMMD
深度子域适应网络	Deep Subdomain Adaptation Network，DSAN
相关对齐	Correlation Alignment，CORAL
快速批核范数最大最小化	Batch Nuclear-norm Maximization，BNM

① 按文中出现的先后顺序排列

生成对抗网络	Generative Adversarial Network，GAN
领域对抗神经网络	Domain Adversarial Neural Network，DANN
动态对抗自适应网络	Dynamic Adversarial Adaptation Network，DAAN
低秩稀疏表示	Low-rank and Sparse Representation，LRSR
鲁棒潜在公共子空间学习	Robust Latent Common Subspace Learning，RLCSL
交替方向乘子算法	Alternating Direction Method of Multipliers，ADMM
k-最邻近	K Nearest Neighbor，KNN
西储大学	Case Western Reserve University，CWRU
江南大学	Jiang Nan University，JNU
判别联合概率 MMD	Discriminative Joint Probability MMD，DJP-MMD
深度联合概率适应网络	Deep Joint Probability Adaptation Network，DJPAN
联合 MMD	Joint MMD，JMMD
多核 MMD	Multiple Kernel MMD，MK-MMD
多核 DJP-MMD	Multi Kernel DJP-MMD，MKDJP-MMD
修正线性单元	Rectified Linear Unit，ReLU
随机梯度下降	Stochastic Gradient Descent，SGD
帕德伯恩大学	Paderborn University，PU