XXXX

基于人工智能的无线通信数据标注技术研究进展

张瀚文[1)]，白崧言[1)]，李凡[2)]，高琳[1)]，罗亚梅[3)]

（1. 1）成都信息工程大学 人工智能学院 ， 成都 中国 610225；

2. 2）成都市武侯区数字健康与医学智能产业研究院 医学智能研究所， 成都 中国 610041；

3. 3）西南医科大学 医学信息与工程学院， 泸州 中国 646000）

摘　要：随着无线通信系统向 5G-A 与 6G 演进，网络环境的复杂性和动态性不断增强，数据驱动方法在无线系统中的应用日益广泛。然而，无线通信数据具有连续流、强时变和场景依赖等特点，高质量标注数据的获取与长期维护成本高昂，已成为制约电信智能化发展的关键瓶颈。围绕无线通信中的数据标注问题，本文系统梳理了无线数据的主要模态、标签形态及真值来源，并分析了标注噪声与数据分布漂移的形成机制。进一步结合频谱监测、干扰识别、调制识别和 Wi-Fi 感知等典型场景，综述主动学习、弱监督、半监督、自监督、噪声鲁棒学习、生成式模型、大模型等技术路线，并比较其在降低标注成本与提升标注质量方面的优势与局限。最后，面向未来 6G 与 AI-RAN 场景，探讨了基础模型、持续学习与隐私约束下标注优化等潜在研究方向，为无线通信智能化标注体系的构建提供参考。

关键词：无线通信；数据标注；人工智能；弱监督学习；主动学习；自监督学习

中图分类号：

文献标志码：

doi: 10.11959/j.issn.1000−0801.

Recent Advances in AI-Based Data Annotation for Wireless Communications

Zhang Hanwen[1)], Bai Songyan[1)], Li Fan[2)], Gao Lin[1)], Luo Yamei[3)]

1. 1)(College of Artificial Intelligence (CUIT Shuangliu Industrial College) Chengdu University of Information Technology Chengdu, Sichuan Province, China 610225

2. 2) Academy of Digital Medicine and Medical Intelligence Chengdu, Sichuan Province, China 610041

3. 3) School of Medical Informatics and Engineering, Southwest Medical University, Luzhou 646000, Sichuan Province, China 646000

Abstract: With the evolution of wireless communication systems toward 5G-Advanced and 6G, network environments are becoming increasingly complex and dynamic, making data-driven approaches and artificial intelligence essential for intelligent wireless systems. However, wireless communication data exhibit continuous streams, strong temporal variations, and high scenario dependency, resulting in high costs for acquiring and maintaining high-quality labeled data, which has become a critical bottleneck for practical deployment. Focusing on the problem of data annotation in wireless communications, this paper systematically reviews the main data modalities, label forms, and ground-

收稿日期：XXXX−XX−XX；修回日期：XXXX−XX−XX

truth acquisition mechanisms, and analyzes the sources of annotation noise and data distribution drift. We further survey several representative learning paradigms, including active learning, weak supervision, semi-supervised learning, self-supervised learning, noise-robust learning, generative models, and large-model-based approaches, and discuss their advantages and limitations in reducing annotation cost and improving label quality across typical applications such as spectrum monitoring, interference identification, modulation recognition, and Wi-Fi sensing. Finally, future research directions are outlined toward 6G and AI-RAN, including foundation models, continual learning, and privacy-aware annotation, aiming to provide insights for building low-cost and reliable annotation pipelines for intelligent wireless systems.

Key words: Wireless communications, data annotation, artificial intelligence, weak supervision learning, active learning, self-supervised learning

0 引言

无线通信系统正从传统模型驱动范式向数据驱动范式转型。随着通信场景复杂化和频谱环境动态变化，传统方法在应对非理想因素时逐渐显现局限性[1, 2]。深度学习凭借其非线性表征能力，可从海量数据中自动提取特征并提升系统性能[3]。

数据驱动方法依赖高质量标注数据，但无线通信中的标注面临显著挑战：一是标注依赖专业射频知识与复杂设备，成本高昂；二是无线环境动态变化导致标签易失效；三是实际场景中普遍存在弱标签与未标注数据问题。

与视觉和文本数据相比，无线数据具有连续流、强时变和场景依赖等特性，其真值获取依赖协议解析、外部测量或仿真等多种方式，标注流程复杂且难以规模化。同时，噪声、同步误差及域漂移等因素进一步影响标签质量[4]。

面对上述挑战，一个核心问题是：能否将人工智能技术深度融入数据标注流程，构建数据采集、智能标注、模型训练、部署应用的闭环优化体系。该范式通过引入人工智能驱动的标注策略，降低对大规模人工标注的依赖，并提升标注结果的质量与一致性，从而在减少标注成本与人工复核开销的同时，提高数据的去噪能力、抗分布漂移能力。本文系统综述人工智能驱动的无线通信数据标注技术，包括主动学习、弱监督、半监督、自监督以及噪声鲁棒学习等方法，并结合频谱监测、调制识别和 Wi-Fi 感知等场景分析其应用效果与工程价值[5, 6]。

针对现有研究在电信数据标注方面缺乏系统性分析的问题，本文围绕面向无线数据标注的学习驱动方法开展综述研究。首先构建无线数据标注的问题框架，梳理无线通信数据模态、标签形态及真值来源，并分析标注成本与精度之间的权衡关系；随后结合频谱监测与干扰识别、调制识别与信号分类以及 CSI 感知与人体动作识别等典型任务，分析实际场景中标注链路的关键挑战；在此基础上，系统总结主动学习、弱监督、半监督、自监督及噪声标签鲁棒学习等主要技术路线；进一步构建任务、数据集、标签来源的综合评测框架[7, 8]，引入面向工程应用的评价指标；最后，讨论开放问题与未来研究方向，并对全文进行总结。

1 无线数据与标签

随着无线通信系统智能化发展，数据驱动方法广泛应用于网络优化、信号识别与无线感知，高质量标注数据成为关键基础。然而，相较于视觉与文本数据，无线数据在结构形态、标签获取及标注流程上差异显著，如连续流生成、标签依赖协议解析或外部测量、标注质量受环境影响等。因此，理解其模态结构、标签来源与标注流

程是分析智能标注技术的前提。在此基础上，本章从数据模态与切片粒度、标签形态与真值来源及标注流程工具链三个方面进行梳理。

1.1　数据模态与切片粒度

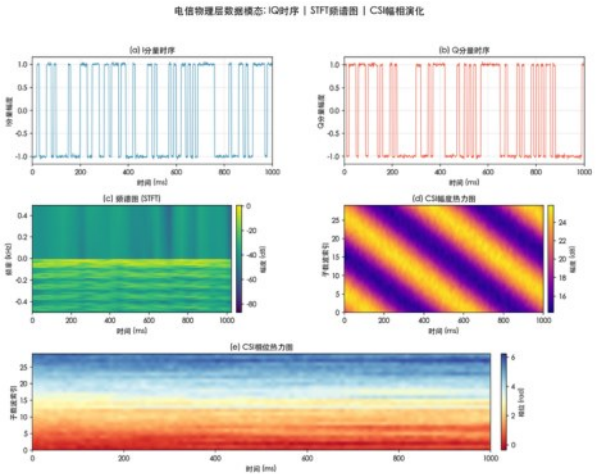

图1　无线物理层数据模态

从观测形式看，无线数据可以粗略划分为物理层与网络层两大类。物理层侧，典型数据包括复基带IQ时序、时频谱图（如STFT/瀑布图）、CSI及其幅相随时间与子载波的演化等；网络侧则包括KPI指标序列、告警与工单日志、信令与计费记录等。不同模态的共同特征是连续性与强上下文依赖：原始观测往往以长时间连续流产生，而可用于学习的样本通常需要通过切片与对齐得到。数据可视化如图1所示。

1.1.1　基础信号数据：IQ样本与频谱图

在物理层数据中，复基带IQ样本是最基础表示形式，其复数序列可完整表征调制结构、脉冲成形及信道效应，广泛应用于调制识别、信号分类与频谱感知。研究表明，IQ样本既可直接作为深度学习输入，也可转换为星座图或时频图等形式[9]。其优势在于保留细粒度特征，支持从物理量到语义的映射；但其不含语义标签，需依赖协议解析或外部测量建立标注，增加了标注复杂度[10]。频谱图通过短时傅里叶变换将IQ样本映射为时频能量分布，具有良好的可解释性，便于识别带宽、中心频率及出现时间等特征，已成为频谱监测与干扰识别中的常用格式。然而，其信息压缩会导致细节丢失，不同信号可能呈现相似特征，增加分类歧义。

在数据集构建方面，调制识别等任务通常在受控或仿真环境下生成固定长度样本，而实际空口采集中，样本边界、同步及干扰叠加会影响切片质量与标签可靠性。在频谱监测场景中，连续频谱流需通过时间窗、频率分辨率及阈值策略生成训练片段，易引入弱标注偏差[11]。

1.1.2　信道状态信息：CSI与信道响应

信道状态信息（CSI）刻画无线信道对信号传播的影响，是波束成形、资源调度及无线感知的重要输入。在Wi-Fi与蜂窝系统中，CSI通常表示为天线—子载波维度的复数矩阵。相较IQ样本，CSI可在已知导频条件下通过信道估计直接获取，具有一定“自标注”特性。然而，CSI对环境变化高度敏感，人体移动、设备位置及场景变化均会引起分布漂移，增加标注与模型维护难度[12]。

1.1.3　高层语义数据：关键性能指标、日志与事件

在网络运维层，高层语义数据主要包括关键性能指标（KPI）与系统日志。KPI以吞吐量、时延、丢包率等时间序列形式呈现，具有明确数值语义；日志则记录设备与网络事件，包含大量非结构化信息。此类数据标注成本较低，KPI为客观测量，日志异常亦可自动标记，但其挑战在于语义鸿沟，即由指标或告警映射至具体故障根因依赖领域知识，标注粒度与一致性难以保障。

1.1.4　切片粒度与标注成本

切片粒度直接影响模型输入统计特性、标签定义及标注成本。窗口过短易破坏协议结构或动作连续性，导致语义模糊；过长则增加复核成本并加剧类不平衡与长尾问题。在CSI/Wi-Fi感知中，样本需与视频或穿戴传感器对齐，切片边界

及跨模态同步误差会放大标签噪声。因此，采集、切片与对齐应作为标注质量工程的重要组成，切片策略、同步信息及采集元数据的规范记录是评测对齐与实验复现的基础[13]。

1.2 标签形态与真值来源

无线数据标签呈现多样化与层级化特征，包括离散类别标签（如调制方式、信号体制、干扰类型、设备身份）、片段或事件级标签，以及连续值真值（如到达角、位置、速度、信道参数、波束索引等）。其中，连续真值通常依赖外部测量系统、定位融合算法或仿真器获取[14]。图 2 给出了主要真值来源的整体关系框架。

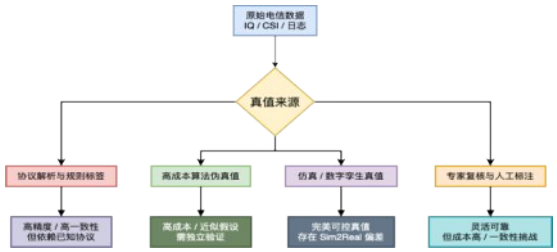

图2 真值来源关系框架图

1.2.1 协议解析与规则标签

协议解析是最直接的真值获取方式。在已知协议条件下，通过解析帧头、控制字段或同步序列，可获取调制方式、编码速率及传输格式等信息，实现 IQ 样本的自动标注。O’Shea 等人指出，基于 GNU Radio 的软件无线电平台可构建从信号采集到协议解析的自动化链路，为大规模数据集生成提供了技术基础。该方法具有高精度与高一致性，但适用范围受限，仅适用于已知协议信号，对未知协议或非标准设备难以处理；且在低信噪比、强干扰或非标准实现条件下，易解析失败并产生漏标[10]。

1.2.2 高成本算法伪真值

在协议解析难以覆盖的场景中，可采用高成本算法生成伪真值。典型方法包括：基于 MUSIC、ESPRIT 的高分辨率谱估计用于参数提取，专家规则匹配用于复杂信号识别，以及基于物理模型的信道仿真用于生成可控 CSI 真值。该类方法成本显著高于人工标注，且精度受限于算法假设与近似。Alkhateeb 等人构建的 DeepMIMO 数据集即采用电磁仿真生成毫米波大规模 MIMO 通道数据。伪真值的质量验证是关键问题，需依赖独立高精度参照（如实测数据或更高成本算法）评估其误差水平[14]。

1.2.3 可控生成机制下的真值构建

仿真环境提供完全可控的真值生成范式，信号参数、设备配置及环境条件均由研究者设定，标签与数据具有确定对应关系。O’Shea 等人指出，仿真数据的核心价值在于提供完美标签，使算法比较可在可控条件下进行。然而，仿真与真实环境存在差距，信道、硬件及非线性建模偏差会导致模型在实际场景中的性能下降。尽管仿真与射线追踪等机制可生成一致真值并覆盖极端工况，广泛用于波束与信道任务，但仍需应对 Sim2Real 偏差与域外泛化问题[15]。

1.2.4 专家复核与人工标注

在自动化方法失效或需真值验证时，人工标注仍是不可替代手段，其优势在于灵活性与领域知识融合，能够处理复杂场景并给出符合语义的判断。然而，其面临成本高、效率低及一致性不足等问题，且射频领域专业人才稀缺、标注周期长，不同专家间亦存在差异。Electrosense 项目通过众包方式动员无线电爱好者参与频谱标注，在一定程度上缓解了人才瓶颈，但其质量控制与一致性保障仍待解决[11, 16]。

1.3 标注噪声与域漂移机制

1.3.1 标签噪声来源

无线数据标注中的噪声来源复杂。首先，切片边界误差源于连续信号中起止时刻划分的主观性，易导致标签不一致；其次，多模态场景下的同步误差（如 CSI 与视频对齐）会引入时间偏移；再次，协议解析在低信噪比、信号叠加或非

标准实现下易失效，产生错误或漏标。此外，多规则并行标注可能引发规则冲突，而标注员在边界与复杂场景中的主观分歧亦难以避免，众包场景尤为突出[11, 12]。

1.3.2　域漂移因素

域漂移是无线数据的重要挑战，使历史标注在新条件下可能失效。其主要表现为信噪比漂移，由电磁环境、干扰及传播条件变化引起，使同一信号特征随SNR显著变化。信道漂移源于不同场景（室内外、城市与郊区）的多径差异，影响模型跨场景泛化。设备漂移来自射频前端、振荡器及采样策略差异，使同类信号在不同设备上呈现不同特征。此外，在CSI感知中，场景与人群变化同样会引起信号模式偏移[13]。

1.4　标注流程工具链

1.4.1　数据采集与切片

高质量标注以规范化数据采集为前提。在无线通信中，采集涉及射频前端配置、采样参数设置及存储管理等环节，其中采样率、中心频率与增益等参数直接影响频率覆盖与信噪比。Hilburn等人指出，采集配置应作为元数据完整记录，以支持数据溯源与质量评估。数据切片将连续时域信号划分为独立样本单元，其策略（如固定时长、基于事件或信号检测）决定标注粒度与样本多样性[8]。

1.4.2　标注管理与版本控制

标注流程的工程化管理是保障质量的关键。有效系统应支持任务分发与回收、多标注者进度与质量统计、冲突自动检测与人工仲裁，以及标注历史的完整记录与回溯。其中，版本控制机制尤为重要，当标注标准变更、新增数据或修正错误时，应可追溯各版本变更内容，保障数据集可复现性。借鉴软件工程实践，数据集应采用语义化版本管理，对新增样本、标签修正及数据划分调整等变更进行版本发布并记录日志[17]。

1.4.3　质量控制与一致性检验

标注质量控制贯穿全过程。事前控制包括标注规范制定、人员培训及界面优化；事中控制涵盖进度与质量指标监控、基于置信度阈值的复核触发，以及通过交叉验证评估标注一致性；事后控制包括抽样复核与错误率统计、基于模型的异常检测，以及错误根因分析与流程改进。Gebru等人提出的数据集文档化框架从创建动机、数据组成、采集过程及维护计划等方面提供系统指南，为标注质量的可追溯评估提供支撑[7]。

1.5　元数据标准与可复现发布

1.5.1　信号元数据格式

信号元数据格式（SigMF）是由GNU Radio社区推动的开放标准，用于统一射频数据的元数据描述。其核心设计为元数据与原始采样数据分离存储，采用JSON格式记录采集参数、信号特征、标注信息及版权等属性。Hilburn等人在介绍SigMF时指出，该格式实现了不同采集系统与处理工具之间的数据互操作，支撑数据集共享与复用。同时，SigMF支持扩展命名空间，允许用户自定义字段以适应特定应用需求[8]。

1.5.2　数据集可复现发布的最佳实践

可复现性是科研与工程落地的基础。对于无线标注数据集，其发布应遵循以下原则：数据集需包含完整的原始数据、元数据与标注文件，以支持基于公开信息的重建；创建过程应充分文档化，涵盖采集设备、时空条件、标注人员及流程规范；应采用开放许可协议明确使用条件并保障知识产权，同时提供基准评测结果或基线代码，以支持方法的公平比较与验证。DeepMIMO数据集即为典型实践，作者公开了完整信道数据以及生成脚本与使用文档，支持在相同条件下复现实验结果[14]。

2　典型无线通信任务中的数据标注链路与

关键挑战

本章选取频谱监测与干扰识别、调制识别与信号分类以及Wi-Fi感知与人体动作识别三类典型任务，作为贯穿全文的案例，分析标注链路中的核心问题。这些场景覆盖从被动监测到主动感知、从物理层到应用层的不同层级，面临差异化标注挑战。基于统一分析框架（标签来源、噪声来源、漂移来源及AI介入机会），本章总结无线数据标注的共性规律，并为后续技术路线提供问题导向的依据。

2.1 频谱监测与干扰识别

2.1.1 任务背景与标注需求

频谱监测是无线电管理、频谱共享与干扰排查的基础，其任务包括信号检测、体制识别、干扰分类及频谱地图构建。随着设备规模增长与场景复杂化，自动化监测需求日益迫切。其标注需求具有典型特征：信号类型多样，涵盖广播、电视、雷达、卫星通信及未授权干扰等，分类体系复杂且动态演化；同时，信号呈现随机性与瞬时性，需在宽频段与长时域内完成事件捕获与标注[11, 18]。

2.1.2 标签来源与成本分析

Electrosense项目是频谱监测众包标注的典型实践[11, 16]。该项目通过部署分布式低功耗频谱传感器网络采集IQ样本并上传云端，由志愿者完成标注与验证。众包模式可覆盖广域频谱并获取多样化数据，但也带来质量挑战，包括志愿者专业水平差异导致的标注分歧、标注质量依赖参与度而影响可持续性，以及新型信号体制持续演进带来的维护负担。

另一类标签来源为基于规则与信号数据库的自动标注。系统通过内置调制方式、带宽及中心频率等特征模板进行模式匹配，实现已知信号识别。该方法成本低、效率高，但仅适用于已知类型，未知或异常信号多被标记为“其他”，形成弱标签；同时，模板精度受设备差异与信道影响限制，易产生系统

2.1.3 噪声来源与标注偏差

频谱数据的标注噪声主要体现在多个方面。首先，信号边界模糊，实际频谱中信号常与其他信号叠加或呈现动态功率变化，起止时间与频率边界存在主观性；其次，体制分类存在歧义，不同系统或设备在时频特性上相近，分类依赖领域知识；再次，时变信号具有标签时效性，特征随时间、环境及设备状态变化而漂移，使历史标注失效；此外，众包标注中的二值或粗粒度标签属于弱标签，难以刻画细粒度属性，易引入信息损失与训练偏差[19, 20]。

2.1.4 漂移因素与环境变化

频谱数据的分布漂移是影响模型泛化的关键因素。其一，信噪比漂移由电磁环境、干扰及传播条件变化引起，使同一信号在不同SNR下特征显著差异；其二，设备漂移源于传感器硬件差异，如射频前端响应、时钟稳定性及增益控制策略不同，导致系统性偏差；其三，场景漂移体现在地理环境与频谱政策差异，不同区域的信号构成与干扰水平存在变化；此外，技术演进带来时间维度漂移，新型体制（如5G、Wi-Fi 6/7）持续改变频谱生态，要求标注系统具备动态更新能力[21]。

2.1.5 人工智能应用的可行性

针对上述挑战，AI可在多环节提供支持。在数据筛选阶段，主动学习可优先选择高信息量样本以提升标注效率；在弱标签融合阶段，可通过加权聚合或概率模型整合众包标注以降低个体偏差；在伪标签生成阶段，利用模型预测为未标注数据提供初步标签并辅助专家复核；在持续更新阶段，在线或增量学习可适应分布变化，缓解模型性能随时间退化[22]。

2.2　调制识别与信号分类

2.2.1　任务背景与标注需求

调制识别是无线信号分析的核心任务，旨在依据接收信号的时频特征判别调制方式，广泛应用于频谱感知、信号情报及电子对抗等领域。其标注需求具有细粒度特征，需区分 QPSK、16QAM、64QAM 等具体类型，且在特定条件下不同调制方式可能呈现相似特征，增加分类难度。相较频谱监测，该任务数据规模更大且标注精度要求更高，是深度学习在无线通信中的早期典型应用之一[3, 9]。

2.2.2　标签来源与真值困境

调制识别的标签获取面临根本性困难。O’Shea 等人指出，在实测场景中难以获知发射端真实调制方式，尤其在信号经信道传输与干扰后，仅由接收信号反推调制方式具有病态性。这一问题催生了两类数据集构建路径[3]。

一是基于仿真的数据生成。RadioML 系列通过 GNU Radio 生成可控信号并赋予完备标签，具有标签准确、可控性强及成本低等优势，但受限于仿真与真实信道差异，模型性能难以直接迁移。二是基于开源数据的实测采集。DeepSig 等发布的 RadioML 数据集尝试弥合仿真与真实差距，但其标签仍依赖仿真辅助或专家标注，数据规模与多样性受限。

2.2.3　噪声来源与标注偏差

调制识别中的标注噪声来源具有特殊性。首先，仿真与真实差异会引入偏差，模型可能学习仿真器特有特征而非可泛化表示。Boegner 等人指出，该问题在深度学习中尤为突出，神经网络易过拟合仿真分布及噪声[23]。其次，信道条件不匹配会导致标签偏差，当训练与测试在 SNR 或多径特性上存在差异时，预测误差可能源于标签与输入关系变化。再次，标注粒度不一致，如粗粒度与细粒度标签混用，会导致训练目标模糊。

2.2.4　漂移因素与泛化挑战

调制识别面临多源漂移因素。SNR 漂移最为直接，模型在特定信噪比下训练后，在更高或更低 SNR 条件下性能显著下降。O’Shea 等人指出，深度学习模型在低 SNR 区域退化尤为明显，与人工识别能力变化趋势一致。信道漂移源于不同场景（城市、郊区、室内外）的多径差异，影响跨场景泛化；设备漂移则由发射与接收硬件差异引起，不同射频前端、振荡器及采样策略会使同类调制信号呈现不同特征[3]。

2.2.5　人工智能应用的可行性

针对调制识别的标注挑战，AI 可在多个方面提供支持。在数据增强方面，可基于仿真生成可控扰动样本，覆盖更广 SNR 与信道条件以提升鲁棒性；在域适应方面，迁移学习可缓解仿真与真实之间的差距；在标注效率方面，主动学习通过选择高不确定样本提升信息利用率；在标签去噪方面，Confident Learning 等方法可识别错误标签及分布外样本；在持续维护方面，增量学习可适应新调制方式与环境变化[24, 25]。

2.3　CSI 感知与人体动作识别

2.3.1　任务背景与标注需求

基于 Wi-Fi 等商用设备的无线感知近年来受到关注，其中 CSI 感知是核心路径。其通过利用信号与人体或物体交互产生的多径变化，实现动作识别、手势识别、跌倒检测及呼吸监测等应用。不同于频谱监测与调制识别，CSI 感知的标注对象为信号与环境交互结果，即人体动作或状态，因此面临特定挑战：标签需刻画时间序列动态过程，粒度从离散类别到连续轨迹不等[12]。

2.3.2　标签来源与真值困境

CSI 感知的标签来源主要包括三类。其一，基于视频同步的标注，通过摄像头与 Wi-Fi 设备协同采集，并利用视觉方法提取动作标签，Wi-dAR3.0 等数据集广泛采用该方式。其优势在于标签精度高、时间对齐准确，但存在隐私风险，仅

适用于受控环境，难以规模化部署。其二，基于实验设计的标注，在预设场景下由受试者按脚本执行动作并同步采集数据，具有较强可控性，但数据多样性有限，且与真实行为存在偏差。其三，基于众包的自然采集，在不干预用户行为的情况下收集数据，后期通过人工或算法推断标签，更贴近实际场景，但噪声大且质量不稳定[26, 27]。

2.3.3 噪声来源与标注偏差

CSI感知的标注噪声具有典型特征。首先，时间对齐误差来源于连续CSI序列与离散标签的不匹配，即使视频提供时间戳，帧率差异仍需插值或重采样，引入误差。其次，动作边界具有主观性，不同标注者对起止时刻判断存在偏差，SenseFi作者指出该问题构成系统性误差来源。再次，执行变异性显著，包括不同个体间及同一个体不同时刻的差异，增加建模难度。最后，环境依赖性强，家具布局、人员密度及设备位置变化均会改变CSI模式[12]。

2.3.4 漂移因素与跨域泛化

CSI感知的分布漂移尤为显著。环境漂移最为关键，WidAR3.0研究表明，即使同一室内环境中家具位置或人员变化亦会显著影响模型性能。设备漂移源于硬件差异，不同路由器或网卡在天线配置、采样策略及信号处理上的差异会导致CSI分布变化。人员漂移体现为个体差异，不同身高、体型及动作习惯对应不同信号模式。此外，时间漂移包括环境渐变与设备老化等因素。跨域泛化能力不足已成为限制CSI感知规模化应用的关键瓶颈[13]。

2.3.5 人工智能应用的可行性

针对CSI感知的挑战，AI可提供针对性支持。在标注层面，弱监督学习可利用视频提供的粗粒度标签降低标注精度要求，自监督预训练可从无标注CSI序列中学习通用表征[28]。在去噪层面，时序模型可识别边界误差与执行异常。在跨域泛化层面，域适应与域泛化方法可学习对环境、设备及人员变化不敏感的表示。在持续更新层面，迁移学习可实现新环境与新用户的快速适配，降低重标注成本[29]。

3 人工智能驱动的数据标注方法

针对无线数据标注中的关键问题，如人工成本高、标签噪声及分布漂移等，本章基于前述案例分析，系统总结提升标注效率与质量的智能方法。表1对比了相关方法在标签来源、目标、优劣及应用场景等方面的差异，表2进一步总结了具体应用及其优势。

3.1 基于主动学习的标注方法

3.1.1 主动学习的基本原理

主动学习是一种人机协同的学习范式，其核心在于从大量未标注数据中主动选择高信息量样

表 1 面向无线通信数据标注的人工智能方法体系对比

方法类别	标注来源	主要目标	优势	局限性	应用场景
主动学习	少量人工精确标签	最大化单位标注信息量	显著降低人工标注数量	依赖人工专家；采样策略设计复杂	频谱监测异常信号标注、调制识别边界样本标注
弱监督学习	规则标签、协议解析、众包标签	低成本生成大规模标签	标注成本低；易于规模化部署	标签噪声大；依赖领域规则质量	频谱授权识别、规则驱动信号分类
半监督学习	少量标注 + 大量无标注数据	利用数据分布结构	能充分利用未标注数据；实现简单	对分布假设敏感；伪标签错误易放大	调制识别、频谱分类
自监督学习	无人工标签	学习通用特征表示	不依赖人工标注；适合大规模数据	代理任务设计难；下游适配仍需标注	CSI感知预训练、频谱表征学习
噪声鲁棒学习	含噪标签	抑制错误标签影响	可在高噪声标注条件下训练模型	对真实噪声分布敏感；参数调节复杂	众包频谱标注、仿真数据去噪

表 2　面向无线通信数据标注的人工智能方法具体应用

方法类别	典型任务	采用的技术	优势	参考文献
主动学习	毫米波波束选择	不确定性采样	在保持性能的同时，将所需标注样本量减少约50%	[4]
主动学习	毫米波吞吐量预测	不确定性采样	仅增加40个标注样本，RMSE从389下降到365，明显优于随机采样	[30]
弱监督学习	调制识别	Snorkel框架	程序化生成大规模弱标签，减少人工逐样本标注	[31]
半监督学习	频谱分类	伪标签自训练	利用少量标注数据结合大量无标注数据训练，降低标注需求	[32-34]
自监督学习	CSI感知	对比学习、掩码预测	学习环境无关表征，减少特定环境标注需求	[28]
自监督学习	频谱表征学习	通用预训练	可从无标注数据学习通用频谱模式，减少下游任务标注成本	[35]
生成式方法	调制识别	GAN半监督学习	在合成RF数据集上，相比传统深度学习模型 分类准确率提升约 0.1%–12%；在中高SNR条件下明显优于 CNN / SVM 等方法	[36]
生成式方法	信道估计	GAN数据增强	通过生成数据减少约70%的导频需求	[37]
生成式方法	CSI感知	RF-Diffusion	作为数据增强可使CSI感知任务准确率 提升4%–11%	[38]
噪声鲁棒学习	众包标注	Confident Learning	自动识别错误标签，减少人工复核成本	[25]
噪声鲁棒学习	仿真去噪	DivideMix	高噪声环境下可以区分干净、噪声样本	[39]
基础模型	信道建模	WiFo预训练	支持 zero-shot / few-shot 下游任务，显著减少人工标注数据需求	[40]
基础模型	调制识别	大模型伪标签	利用通信基础模型自动预测通信特征生成伪标签，减少人工标注数据需求	[41]
大语言模型	标注规则生成	LLM 程序化标注	利用 LLM 理解通信协议文档和网络规范，自动生成弱监督标注规则并融合多源监督信号	[42]
大语言模型	规则知识抽取	文档规则自动提取	从网络配置文档和技术规范中自动提取规则知识，用于自动化标注函数构建	[43]

本进行标注，以较少数据获得更高性能，在标注成本高或依赖专家知识的场景中具有重要价[22]。其典型流程为：基于已有标注训练模型，利用不确定性或多样性准则选择待标注样本，提交专家标注并加入训练集，循环迭代优化。该方法在高成本标注场景中优势显著，被视为提升无线数据标注效率的关键路径之一。

3.1.2　关键采样策略

主动学习的效果依赖于样本选择策略。常见方法包括不确定性采样、多样性采样以及基于表示或梯度信息的策略，其目标是在有限标注预算下优先选择高信息量样本以提升效率。不确定性采样通过选择预测置信度最低的样本，获取靠近决策边界的数据，从而减少标注需求[22]。在深度学习场景中，研究者进一步提出了针对高维特征空间的主动学习策略。Sener 和 Savarese 提出的 Core-Set 方法将问题转化为覆盖问题，通过选取代表性子集降低标注成本[24]。Ash 等人提出的 BADGE 方法结合梯度信息，选择兼具不确定性与多样性的样本。

在无线通信场景中，上述策略需结合数据特性进行调整。由于信号具有连续时间结构并受信道影响，样本选择除考虑分类不确定性外，还需兼顾片段相关性及数据分布变化[44]。

3.1.3　无线数据标注中的主动学习应用

在无线通信中，主动学习被视为降低标注成本的重要方法。Soltani 等人指出，数据采集相对容易，而标签获取依赖复杂信号处理或专家参

与，成本较高，主动学习通过选择高信息量样本可在保持性能的同时减少标注需求[4]。在应用方面，Soltani等人将其用于毫米波波束选择，通过筛选关键样本降低标注规模，实验表明标注量可减少约50%。进一步地，Alhussein等人将主动学习嵌入6G数据闭环，提出“按需标注”，在采集阶段依据模型增益决定是否触发人工标注，该机制契合频谱监测中信号随机性强、标注窗口短的特点[30]。

需要注意，无线信号特征空间不同于视觉数据，其连续性、物理约束及信道扰动要求主动学习策略进行针对性设计[45]。已有研究表明，直接迁移视觉领域的不确定性或多样性采样难以取得最优效果，需结合调制结构、SNR分布及频谱占用特性设计采样准则。总体而言，主动学习可显著降低标注需求，尤其适用于高成本或专家依赖场景，但现有方法仍依赖不确定性度量或启发式策略，易产生选择偏差。结合通信物理特性优化样本选择机制，是未来的重要方向。

3.2 弱监督与程序化标注

3.2.1 弱监督学习的基本原理

弱监督学习利用弱于精确标签的监督信号进行训练，这类标签通常具有噪声大、粒度粗或不完整等特点，但获取成本显著较低。在无线场景中，弱监督来源广泛，包括基于规则的自动标注、协议解析生成的标签以及带有置信度差异的众包标注。其核心价值在于利用低成本弱信号构建初始模型，并通过后续精炼逐步提升标注质量与模型性能[28]。

3.2.2 Data Programming与Snorkel框架

弱监督学习以低成本、低精度或不完整标签替代人工精确标注。Ratner等人提出的Data Programming框架通过构建多个标注函数自动生成弱标签，并利用生成模型估计其质量，在减少人工标注的同时构建大规模训练数据。其核心在于将标注函数视为启发式规则单元，每个函数从数据映射至标签，可能存在噪声[31]。Snorkel实现了该范式，通过学习标注函数的准确性及相关性生成高质量合成标签。在无线场景中，标注函数可对应各类规则，如基于中心频率或带宽阈值进行判别，多函数组合可扩展覆盖范围并实现相互纠错[46]。

3.2.3 工业级弱监督实践

Bach等人系统介绍了Snorkel DryBell在谷歌等工业场景中的部署，验证了弱监督在大规模数据处理中的可扩展性，并强调其与现有IT基础设施的集成能力。标注函数可调用外部服务（如知识库与模型推理），实现跨系统与多源数据的弱标签融合。在无线领域，该思路可扩展至融合网络配置、频谱授权及设备指纹等信息，通过程序化方式生成统一标签。相关实践还表明，需关注标注函数的管理与版本控制、弱标签质量的可视化监控，以及与oracle标注的协同优化[47]。

3.2.4 无线数据标注中的弱监督应用

弱监督学习为无线场景中规则化、低成本标注提供了系统化建模框架。从应用角度看，其优势在于能够将领域规则转化为可扩展的标注机制。在频谱监测中，基于地理位置与频段授权的规则可自动生成授权或未授权标签，在大规模筛选阶段显著降低人工复核需求；在调制识别中，可利用带宽、符号速率及中心频率等物理特征构造启发式规则，生成调制类型弱标签并形成候选样本池；在CSI感知中，视频或多传感器融合得到的粗粒度标签可作为弱监督信号，缓解时间边界不确定性带来的标注噪声。

从方法局限看，弱监督性能高度依赖标注函数设计与组合策略，其本质是在覆盖度与准确性之间权衡。过宽的规则易引入噪声，过严则限制数据利用效率。因此，该方法通常依赖领域知识，并需结合少量高质量标注对规则进行持续校准。

总体而言，弱监督通过利用不完整或间接监

督信号，实现低成本数据标注与模型训练，在大规模无线数据处理中具有显著优势。但其效果受限于标签噪声与不确定性，如何提升标签融合质量并增强模型鲁棒性，是其进一步发展的关键问题。

3.3　基于半监督学习的标注方法

3.3.1　半监督学习的基本原理

半监督学习利用少量标注数据与大量无标注数据联合训练，其核心假设是数据的流形或聚类结构有助于提升泛化能力。在无线场景中，无标注数据易于获取，持续采集的信号流可直接作为无标注样本。该方法通过挖掘无标注数据中的分布信息，在有限标注条件下提升模型性能。其中，自训练是典型范式，即利用已训练模型对无标注数据进行预测，并将高置信度结果作为伪标签迭代加入训练。从机制上看，半监督方法主要分为基于伪标签的自训练与基于一致性正则化的方法，后者通过约束模型在输入扰动下预测一致性来利用无标注数据。

3.3.2　Mean Teacher与一致性正则化

Tarvainen和Valpola提出的Mean Teacher是一致性正则化的代表方法，通过对student模型参数进行指数滑动平均构建teacher模型，从而提升伪标签稳定性[32]。其核心思想是在输入扰动下保持预测一致，teacher模型由于融合历史信息，较student更稳定，能够生成更可靠的伪标签。在无线信号分类中，该方法与信号受噪声与信道扰动的物理特性相契合，可用于提升模型对环境变化的鲁棒性[32]。

3.3.3　MixMatch与FixMatch

MixMatch将半监督学习与数据增强技术相结合，通过混合标注与无标注样本、猜测无标注样本的标签并应用锐化处理等策略，在多个基准数据集上取得了显著的性能提升[34]。FixMatch则进一步简化了半监督学习流程，仅使用弱增强的预测作为伪标签，强制模型在强增强下保持一致。FixMatch的简洁设计使其在实践中易于部署，同时保持了与复杂方法相当甚至更好的性能[33]。这些先进的半监督方法为无线数据标注提供了有力工具：在标注稀缺的情况下，模型可以从大量无标注信号中学习通用特征，仅需少量标注即可达到较高性能。

3.3.4　无线数据标注中的半监督应用

在无线通信中，半监督学习通过利用大量未标注数据缓解标注不足问题。在频谱监测中，长期监测数据大多未标注，可与少量专家标注联合使用；在调制识别中，RadioML等仿真数据可作为无标注数据，与真实标注样本混合训练；在CSI感知中，不同用户与环境下的CSI序列可用于学习环境无关表征。这表明半监督方法在数据丰富但标注稀缺的场景中具有较高适用性。

从方法前提看，半监督依赖流形或聚类假设，当数据分布不满足该假设时可能失效甚至产生负作用。无线信号的多模态性与信道时变性进一步增加了方法设计难度。

总体而言，半监督学习通过伪标签、自训练及一致性约束等机制挖掘未标注数据信息，有助于提升模型泛化能力。但在实际场景中，未标注数据分布复杂且含噪，伪标签错误传播仍是主要瓶颈。因此，构建稳定可靠的伪标签生成机制，是其进一步发展的关键方向。

3.4　基于自监督学习的标注方法

3.4.1　自监督学习的基本原理

自监督学习（Self-Supervised Learning）通过设计代理任务从无标注数据中学习有用的表征，无需人工标注即可获得通用的特征表示。近年来，自监督学习在自然语言处理和计算机视觉领域取得了突破性进展，大规模自监督预训练已成为构建高性能模型的标准范式。在无线通信领域，自监督学习同样展现出巨大潜力：无线信号具有丰富的时频结构、周期性和上下文相关性，这些特性可以用于设计有效的代理任务[35]。

3.4.2 对比学习与掩码预测

对比学习是自监督学习的主流范式之一，通过拉近同一样本的不同视图、拉远不同样本来学习表征。在无线信号处理中，同一信号的不同增强视图（如时移、频移、添加噪声）可以构成正样本对，不同信号的视图构成负样本对。掩码预测是另一种有效的代理任务，通过随机遮蔽信号的部分内容并预测被遮蔽部分，模型可以学习信号的上下文结构和完整性信息。Ferrand等人将自监督学习思想应用于无线信道分析，提出了基于自监督的无线信道charting方法，利用信道数据之间的几何关系进行预训练[29]。

3.4.3 无线数据标注中的自监督应用

在无线通信中，自监督学习通过挖掘无标注数据的结构信息提升表示能力。在频谱监测中，可从长期频谱数据中学习通用模式表征以提升后续标注与分类效率；在调制识别中，可捕捉信号的时频结构特征，为下游任务提供有效初始化；在CSI感知中，可学习对环境变化鲁棒的感知表示，从而降低对特定场景标注数据的依赖。这表明自监督方法在数据充足但标签稀缺的场景中具有基础性作用[48]。

从方法关键看，自监督性能依赖代理任务设计，其需与下游任务相关且难度适中，否则难以学习有效表征。总体而言，自监督通过预训练任务实现无标注条件下的表示学习，在特征提取与任务迁移中具有显著潜力。但现有方法多依赖经验设计，不同任务间适配性存在差异，如何结合通信信号的物理结构设计更具针对性的自监督任务，是未来的重要方向。

3.5 基于生成式人工智能的数据生成与辅助标注方法

3.5.1 生成式学习的基本原理

生成式人工智能（Generative Artificial Intelligence）通过学习数据的潜在分布结构生成新的样本数据，为解决标注数据不足问题提供了新的技术路径。与主动学习、弱监督或半监督学习主要关注如何更高效利用已有数据不同，生成式方法通过构建概率模型直接模拟数据生成过程，从而生成与真实数据统计特性一致的合成样本。在数据标注场景中，这类方法可以通过数据生成与数据增强的方式扩展训练集规模，从而缓解人工标注成本高昂的问题。近年来，随着深度生成模型的发展，生成式方法在计算机视觉、语音处理和自然语言处理等领域取得了显著进展，也逐渐被引入无线通信数据处理任务中。

3.5.2 生成对抗网络与扩散模型

生成对抗网络（GAN）是早期广泛应用的数据生成模型，通过生成器与判别器的对抗训练学习真实数据分布，从而生成高相似度样本。在无线通信中，GAN已用于射频信号生成、调制数据增强及信道特征模拟。通过学习IQ样本或频谱图分布，可生成合成数据扩展数据集；条件GAN进一步支持在给定标签或信道参数条件下按需生成带标签样本，适用于标注稀缺场景[49]。

扩散模型近年来成为重要生成范式，通过逐步加噪与反向去噪生成高质量样本，在训练稳定性与多样性方面优于GAN。在无线数据处理中，扩散模型可生成具有复杂信道特性的信号样本，如不同多径与传播条件下的数据；条件扩散模型还可依据调制类型、信噪比及频谱特征生成对应样本，为监督学习提供更丰富的数据支持[50]。

3.5.3 无线数据标注中的生成式数据应用

在无线通信数据标注中，生成式模型主要用于扩展训练数据规模，以缓解高质量标注不足。Li等人在认知无线电调制识别中引入GAN进行半监督学习，通过生成伪样本结合少量标注数据完成训练，在减少人工标注需求的同时保持较高分类精度，表明生成模型可有效降低数据集构建对人工标注的依赖[36]。

在信道与频谱数据方面，Balevi和Andrews利用GAN学习信道分布，并将其作为先验辅助

宽带信道估计，从而减少对真实测量数据的依[37]。Roy 等人提出 Generative Adversarial Radio Spectrum Networks，通过生成频谱样本扩展训练数据。这些研究表明，生成模型可在信号与频谱层面补充数据，支持数据集扩展[51]。

随着模型发展，扩散模型亦被引入无线数据生成。Chi 等人提出 RF-Diffusion，通过扩散机制生成具有时频结构的信号样本，并在 Wi-Fi 感知与信道估计中验证其有效性。相较 GAN，扩散模型在稳定性与多样性方面具有优势，被视为重要发展方向[38]。

总体而言，生成式方法通过构造具有真实统计特性的合成样本扩展数据规模，降低对人工标注的依赖。但其在物理一致性、跨场景泛化及样本真实性评估方面仍存在不足，如何在满足信号物理约束下构建高可信生成模型，是后续研究的关键方向。

3.6　噪声标签检测与鲁棒学习

3.6.1　标签噪声问题的普遍性

在实际标注流程中，错误标签几乎不可避免。人工标注易受疲劳与主观差异影响，程序化标注受规则覆盖与边界处理限制，众包标注质量亦存在波动。研究表明，深度神经网络可拟合任意噪声标签，即使随机标签也可达到零训练误差，使标签噪声问题尤为突出。因此，开发噪声检测与鲁棒学习方法，对提升标注质量具有重要意义[25]。

3.6.2　Confident Learning 理论框架

Northcutt 等人提出的 Confident Learning（CL）为噪声标签处理提供了系统框架，其核心在于识别模型高置信但可能错误的样本，即预测与标注不一致的高置信数据。CL 通过估计噪声矩阵刻画标签在类别间的转移概率，并据此修正损失函数或重估标签以降低噪声影响。cleanlab 工具库实现了该方法，为噪声检测与处理提供了实用接口[25]。

3.6.3　鲁棒训练策略

除了事后检测与清洗，另一类方法是设计对噪声标签鲁棒的训练策略。Co-Teaching 采用双网络协同训练的策略，每个网络选择其认为干净的样本供对方训练，相互纠错。这样的设计利用了两个网络的不同学习能力，提高了在噪声环境下的泛化能力。在电信应用中，鲁棒训练策略可以应对标注质量波动的场景：例如，当标注人员发生变化或标注标准发生调整时，模型能够自动适应新的噪声模式。鲁棒学习的挑战在于理论保证与实践效果之间的 gap。许多方法在合成噪声下表现优异，但在真实噪声分布下可能失效[52]。

3.6.4　无线数据标注中的噪声标签处理

在无线通信中，噪声标签处理方法可针对不同任务发挥作用。在频谱监测中，众包标注存在较高错误率，Confident Learning 可识别可疑样本并指导复核资源分配；在调制识别中，仿真数据可能因参数设置不当产生标注偏差，噪声检测可用于修正；在 CSI 感知中，动作边界模糊导致标签不准确，鲁棒学习可降低其对训练的影响。此外，DivideMix 等方法结合半监督与噪声处理，对干净与噪声样本进行区分，在高噪声条件下表现出良好性能[39]。

总体来看，噪声标签检测与鲁棒学习为提升标注质量提供了关键支撑，通过错误识别、样本筛选与鲁棒优化降低噪声影响。但在实际无线场景中，噪声类型复杂且分布未知，现有方法多基于理想或合成噪声验证，实际适用性仍有限。因此，面向真实环境构建鲁棒且可解释的噪声处理机制，是未来的重要方向。

3.7　基于基础模型与大模型的智能标注方法

3.7.1　基础模型与大语言模型

础模型指在大规模数据上预训练并可迁移至多任务的通用模型。Bommasani 等人指出，其通过自监督学习获得通用表示，并通过微调或提示适配下游任务。随着算力与数据规模提升，模型

参数持续增长，形成大模型范式[53]。

在该体系中，大语言模型是典型代表。Brown等人提出的GPT-3通过大规模预训练实现少样本学习，并可通过提示完成多任务预测。由此可见，基础模型是总体范式，大语言模型是其重要类型[54]。

在数据标注中，基础模型具备重要价值。其可利用预训练知识对未标注数据进行理解与预测，生成伪标签或辅助人工标注，从而形成“模型生成+人工复核”的新型标注范式。

3.7.2 大模型辅助数据标注的关键技术

大模型在数据标注中的应用通常依赖三类关键技术：伪标签生成、提示学习以及人机协同标注。

首先，伪标签方法（Pseudo-Labeling）利用模型对未标注数据进行预测，并将高置信度预测结果作为训练标签。Lee提出的伪标签方法表明，在少量标注数据的条件下，通过模型生成伪标签能够有效扩展训练数据规模[55]。在基础模型框架下，预训练模型能够提供更准确的初始预测，从而提高伪标签质量。

其次，提示学习（Prompt Learning）使得大语言模型能够通过自然语言提示直接完成数据标注任务。例如，通过设计合适的提示，大语言模型可以对文本、日志或结构化数据进行分类或标注，从而替代部分人工标注工作。这种方式在知识抽取、文本分类和事件识别等任务中已得到广泛应用。

此外，人机协同标注（Human-in-the-Loop）成为大模型时代的重要标注模式。在这种模式下，大模型首先对数据进行自动标注，然后由人工进行快速审核和修正。这种方式能够显著减少人工标注工作量，并提高标注效率。

3.7.3 无线数据标注中的大模型应用

随着基础模型的发展，其在无线通信中的应用逐渐显现。Buffelli等人指出，通信系统正由任务专用模型向基础模型范式转型，通过统一预训练与微调机制，可降低新任务对标注数据的依赖并提升跨任务迁移能力。这一趋势表明，数据标注问题有望从数据驱动转向模型驱动[41]。

在具体应用中，Liu等人提出的WiFo通过大规模多场景信道数据预训练，学习通用传播特征，实现跨频段与环境迁移，并在少量标注条件下保持较好性能，验证了基础模型在降低标注需求方面的潜力。同时，生成式基础模型可学习信号分布并生成不同信道与噪声条件下的样本，用于数据增强与标注扩展，提供新的数据构建路径[40]。此外，一些研究还指出，生成式基础模型能够学习无线信号数据的统计分布，并生成具有不同信道条件和噪声水平的合成样本，从而用于数据增强和标注数据扩充，为构建大规模无线通信数据集提供新的技术路径[56]。

除信道建模任务外，基础模型同样可以辅助构建无线通信数据集。例如，在自动调制识别任务中，大量IQ信号样本需要标注其对应的调制类型（如BPSK、QPSK、16QAM等）。传统方法通常依赖仿真生成或人工标注信号数据集，以获得准确的调制类别标签[9]。通过利用基础模型在未标注信号数据上的预训练能力，可以首先对新采集的信号样本进行自动分类预测，并生成初始伪标签，再通过人工抽样验证与修正，从而加速调制识别数据集的构建过程。

与此同时，大语言模型的发展也为通信数据标注流程提供了新的工具。Boateng等人总结了大语言模型在通信网络管理中的应用，指出LLM能够理解协议文档、网络规范和运维规则等文本信息。在数据标注场景中，这种能力可以用于自动生成弱监督标注规则，并辅助构建程序化标注函数，从而实现多源监督信号的融合与冲突分析[42]。Siino等人的研究进一步展示了LLM从网络配置文档中自动提取规则知识的能力，为自动化标注规则生成提供了新的思路[43]。

综合来看，基础模型与大语言模型正在共同推动无线通信数据标注范式的转型：从人工主导的数据生产走向模型生成与人机协同”的混合模式。其关键价值不仅在于降低标注成本，更在于改变数据获取与利用的基本逻辑。然而，该方向仍处于早期阶段，在信号语义建模、多模态融合以及模型可信性等方面仍存在明显不足。因此，面向通信信号特性的专用基础模型构建，将是推动标注智能化落地的关键路径。

3.8　技术路线对比

3.8.1　各方法族的定位与适用场景

综合前文分析，面向无线通信数据标注的AI方法可归纳为七类：主动学习、弱监督、半监督、自监督、生成式方法、噪声鲁棒学习以及基础模型辅助标注。这一划分反映了不同方法在标签来源、数据规模假设及系统目标上的本质差异，也决定了其适用场景的边界。

从适用条件看，主动学习面向高质量标注可获得但成本受限的场景，通过选择高信息量样本提升标注效率；弱监督依赖规则系统、协议解析或外部知识源，适合低成本构建大规模数据，但受限于规则质量并需处理噪声与冲突；半监督利用少量标注与大量无标注数据的分布结构，在标注稀缺条件下提升性能；自监督则进一步放宽对标签的依赖，通过代理任务学习通用表征，在无线信号时频特征建模中具有基础性作用。

在数据扩展与质量提升方面，生成式方法通过学习数据分布生成合成样本，可模拟复杂信道并缓解数据不足问题；噪声鲁棒学习则针对标签不可靠场景，通过噪声检测与鲁棒优化降低错误标签影响。近年来，基础模型提供了新的统一范式，通过大规模预训练实现跨任务迁移，并结合大语言模型实现规则生成与标注审核，显著提升标注自动化水平。

总体而言，这些方法并非替代关系，而是围绕数据稀缺、标签噪声、分布漂移这三类核心问题形成互补体系。实际系统中往往需要多种方法协同设计，以在标注成本、数据规模与模型性能之间取得平衡。

3.8.2　组合策略与系统设计

在实际系统中，单一技术难以同时应对标注成本、标签噪声与分布漂移等问题，因此多方法协同成为更可行的路径。综合来看，可形成分阶段的组合策略：首先利用弱监督或规则系统生成初始伪标签以快速构建数据基础；随后通过主动学习筛选高信息量样本并引入专家复核以提升关键标签质量；在模型训练阶段结合半监督与自监督方法，充分利用无标注数据增强表示能力；在数据质量控制阶段引入噪声鲁棒学习进行错误检测与清洗；同时利用生成式模型扩展样本并模拟多样信道条件；最后通过基础模型或大模型实现跨任务迁移与自动化标注辅助。

上述流程本质上构建了一种由弱标注、主动校正、鲁棒训练、数据扩展、持续迁移组成的闭环机制。该框架表明，无线数据标注问题更适合通过系统级协同优化解决，而非依赖单一算法，从而在控制标注成本的同时兼顾模型性能与系统稳定性。

4　无线标注评测体系

4.1　无线领域核心数据集

本文汇总了无线通信与智能感知领域中具有代表性的公开数据集，并从主要任务、数据类型、标签来源及标注特征等维度进行对比分析，如表3所示。该表旨在为后续标注方法评测与工程实践中的数据集选择提供参考依据。

4.1.1　RadioML 系列数据集

RadioML是调制识别领域最具影响力的数据集系列，由DeepSig维护发布。该数据集基于GNU Radio生成，包含2FSK至256QAM等多种调制方式的IQ样本，并覆盖不同信噪比条件。其发展经历多版本迭代，早期以简单仿真为主，后

期逐步引入更真实的信道衰落、载波频偏及采样率偏差，以缩小仿真与真实差距。RadioML提供统一评测基准，支持算法在相同条件下比较，但其局限在于仿真与真实信号仍存在差异，模型性能难以完全反映实测表现[9, 10, 57]。

4.1.2 DeepMIMO数据集

DeepMIMO是Alkhateeb等人构建的大规模MIMO信道数据集，面向毫米波与大规模MIMO应用。该数据集基于射线追踪仿真生成，可刻画路径损耗、阴影衰落及多径效应等传播特性，并支持场景几何、频率、天线配置与带宽等参数的自定义。由于为仿真数据，信道矩阵真值已知，无需额外标注。其局限在于仿真与真实信道存在差异，且参数设置依赖专业知识，可能引入不现实场景[14]。

4.1.3 SenseFi数据集

SenseFi是杨等人于2023年发布的Wi-Fi感知基准数据集，系统整理了基于深度学习的Wi-Fi人体感知研究。该数据集不仅包含CSI数据，还集成CSI Tool、Nexmon等处理工具，并提供标准化预处理流程与评测基准。其覆盖人体存在检测、动作识别、手势识别及呼吸监测等任务，包含多用户与多环境标注数据。SenseFi的优势在于综合性与可复现性，提供完整采集协议、标注规范及评测指标，支持公平比较。然而，其数据规模有限且主要集中于室内场景，跨场景泛化评估仍需进一步扩展[12]。

4.1.4 Widar3.0数据集

Widar3.0是清华大学团队发布的Wi-Fi手势识别数据集，重点研究跨域泛化问题。该数据集采集同一空间内不同环境配置（如家具布局与人员密度）下的CSI数据，用于评估模型在环境变化中的性能退化。其采用视频同步标注与动作捕捉获取精确时间边界作为真值。Widar3.0提供了重要的跨域评测基准，表明即使在同一房间内，环境变化亦会显著影响模型性能[13]。

4.1.5 Electrosense数据集

Electrosense项目提供了分布式频谱监测网络采集的大规模数据，并构建了众包标注的质量控制与聚合机制。其标签来源于志愿者社区，多人标注通过聚合生成最终标签并附带置信度信息。该项目验证了众包标注在频谱监测中的可行性，同时为研究标注可靠性与聚合方法提供了数据基础。

在此基础上，Electrosense+引入低成本IoT接收节点，扩展至对信号内容与调制体制的众包解码与识别。其主要挑战在于标注质量的波动，不同区域与时段存在差异，使用时需进行质量评估[11, 16]。

4.1.6 WiMANS数据集

WiMANS是帝国理工团队发布的Wi-Fi多用户行为感知基准数据集，其核心在于首次系统支持并评测多用户场景下的Wi-Fi感知问题。该数据集采集0–5名用户在同一空间内执行相同或不同活动的CSI数据，突破了以往仅覆盖单用户的局限。在教室、会议室及空房间等环境中，基于2.4 GHz与5 GHz双频Wi-Fi共采集超过9.4小时数据，并同步记录高分辨率视频，支持多模态分析。通过预设多用户协同脚本实现精确标注，提供用户身份、空间位置及活动类别等细粒度标签，支持身份识别、定位与行为识别等任务。同时，WiMANS评估了主流Wi-Fi与视频模型在多用户条件下的性能，揭示用户数量增加带来的性能退化及特征解耦挑战，为多用户场景研究提供基准[27]。

4.2 评测指标体系

4.2.1 任务性能指标

任务性能指标是评估模型有效性的基础。对于分类任务，标准指标包括准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1分数和混淆矩阵；对于回归任务，常用指标包括均方误差（MSE）、平均绝对误差（MAE）和决定系

数（R^2）。在电信场景中，还需要考虑任务特定的指标：例如，调制识别的性能应当以SNR函数的形式呈现，展示模型在不同信噪比条件下的表现；频谱检测的性能需要考虑检测概率与虚警率的权衡，以ROC曲线或PR曲线呈现；CSI感知的性能需要区分不同动作类别的识别效果。

4.2.2　标注节省率

标注节省率（Annotation Saving Rate）是评估主动学习、弱监督等技术路线价值的关键指标。其定义为：达到相同任务性能所需的标注数量，相比于随机采样或基线方法的减少比例。计算公式为：节省率 = (基准标注量 - 优化标注量) / 基准标注量 × 100%。例如，如果主动学习方法仅需标注20%的样本即可达到基线90%的性能，则标注节省率为80%。这一指标直接反映了技术路线在标注成本上的收益，是工程决策的重要依据。[58]

4.2.3　标签去噪收益

标签去噪收益（Label Denoising Gain）评估噪声标签检测与鲁棒学习方法的价值。其定义为：经过去噪处理后模型性能相比处理前的提升程度。去噪收益可以从多个角度量化：检测准确率（正确识别噪声标签的比例）、清洗有效率（被识别为噪声的样本中实际错误标签的比例）、性能恢复率（去噪后模型性能与干净标签模型性能的差距缩小程度）。Sculley等人指出，数据质量问题的影响往往远超模型选择，系统的去噪机制对于维持模型性能至关重要[59]。

4.2.4　跨域泛化能力

跨域泛化能力衡量模型在不同环境条件下保持性能的能力。在电信场景中，该问题尤为关键，模型需在新的SNR、设备及场景下仍具有效性。评测通常基于异于训练域的测试域，常用指标包括跨域准确率、域间性能方差及适配所需新增标注量。Breck等人提出的ML测试评分卡将泛化能力视为生产就绪的重要维度，强调模型在未见分布下的稳定性[60]。

4.2.5　漂移维护成本

漂移维护成本（Drift Maintenance Cost）评估模型在长期运行中应对数据分布漂移的持续投入。在电信系统中，信号环境随时间演变，模型性能可能逐渐退化，需要定期更新。维护成本的度量包括：性能监控所需的人工/计算资源、性能下降到阈值以下后的检测延迟、模型更新所需的标注数量与时间。理想的标注优化方案应当降低而非增加维护成本。例如，主动学习策略可以选择性地标注最能帮助模型适应漂移的样本，以最小成本维持模型性能。

5　未来研究方向

随着基础模型、大语言模型与持续学习等人工智能技术的快速发展，无线通信数据标注正由

表3　核心数据集

数据集名称	主要任务	数据类型	标签来源	典型标注特征	核心挑战	代表性文献
RadioML	调制识别	IQ	仿真生成	完美调制标签；多 SNR 覆盖	仿真-真实差距	[10]
DeepMIMO	信道建模 / 波束预测	信道矩阵	射线追踪仿真	可控真值；参数可复现	Sim2Real 偏差	[14]
Electrosense	频谱监测	IQ / 频谱	众包标注 + 规则	多人标注；置信度融合	标签噪声大	[11]
Electrosense+	频谱解码	IQ	众包 + 解码规则	解码级弱标签	覆盖不均衡	[16]
Widar3.0	手势识别	CSI	视频同步	精确时间边界	跨域泛化难	[13]
SenseFi	Wi-Fi 感知	CSI	实验设计 + 人工	多任务统一评测	数据规模有限	[13]
WiMANS	多人活动感知	CSI	人工标注	多用户场景	干扰复杂	[27]

任务驱动的辅助环节逐步演化为支撑智能网络运行的关键基础能力。未来标注方法不仅需要关注标注效率与成本控制，还需兼顾模型长期演进、系统可信性与隐私保护等约束。在此背景下，表4对面向6G与AI-RAN的无线通信数据标注研究方向进行了系统归纳。

5.1 漂移驱动的持续学习与终身标注

电信系统运行环境持续变化，数据分布漂移普遍存在。A. Tziouvaras对6G网络中的概念漂移进行了系统研究，指出用户行为变化、设备更新和网络演进均可能导致模型性能退化[61]。在此背景下，持续学习将模型部署视为动态过程，需要通过持续标注更新来维持性能。

概念漂移检测的关键在于区分真实分布变化与噪声波动。M. A. Mohsin提出了面向无线信道预测的持续学习方法，通过监测预测误差变化检测漂移，并触发增量学习。该工作表明，漂移检测应与标注成本联合考虑，而非对所有变化均触发再标注[62]。

仅依赖漂移检测难以支撑模型的长期自适应运行，仍需将漂移感知、样本选择、标注获取与模型更新形成闭环。主动学习与持续学习的融合为此提供了有效途径。在数据流场景下，系统需在模型性能与标注成本之间进行权衡，通过不确定性采样等策略选择高价值样本进行标注[24, 30, 66]。

5.2 可信标注与可解释性AI-RAN

AI-RAN将人工智能深度嵌入无线接入网，对系统透明性和可审计性提出了更高要求。S. Guo指出，可解释性是AI在电信系统中大规模部署的前提条件，可信AI-RAN需要从模型可解释性延伸至数据和标注过程的可信性[63]。

在信道定位等任务中，标注的可解释性直接影响模型决策的可追溯性。可解释标注体系需支持标签生成依据的记录、标注主体的可追溯性以及标注规则的可审计性，并对标注结果提供不确定性描述[29]。在网络规模较小的场景下，集中式标注仍具可行性；而在大规模动态网络中，标注需与持续学习机制相结合，以支撑在线智能决策。

5.3 隐私约束下的标注优化

感知通信一体化（ISAC）场景下，标注数据可能涉及用户位置和行为等敏感信息。Y. Qu指出，隐私风险不仅存在于数据存储与传输阶段，也可能通过模型推理过程被间接泄露[64]。

K. Roy提出的FedWiLoc框架展示了在WiFi室内定位中结合联邦学习进行协作训练的可行性。联邦学习通过在本地完成标注与模型更新，

表4 无线通信数据标注未来研究方向

研究方向	核心思想	关键技术要素	对数据标注的价值	主要挑战	参考文献
多模态联合标注与对齐	融合IQ、频谱、CSI、视频和日志等多模态信息进行联合建模	多模态对齐、跨模态表示学习	提升标注精度，缓解单一模态歧义	模态同步误差，数据采集复杂	[12, 13, 27]
漂移感知的持续学习标注	将概念漂移检测、再标注与模型更新形成闭环	概念漂移检测、持续学习、增量训练	降低历史标签失效成本，支持长期运行	漂移判定困难，系统复杂度高	[61, 62]
主动学习与持续学习融合	在数据流场景下联合优化标注时机与样本选择	流式主动学习、不确定性采样	以最小标注代价维持模型性能	采样策略设计复杂，实时性要求高	[22, 24, 30]
可解释与可审计标注体系	构建标签生成过程可追溯、可解释的标注链路	可解释AI、标注溯源、置信度建模	提升标注可信度，支撑AI-RAN决策审计	工程实现复杂，缺乏统一标准	[59, 63]
隐私约束下的标注优化	在隐私法规约束下进行标注与模型训练	联邦学习、差分隐私、自监督学习	支持隐私敏感场景下的大规模标注	隐私噪声影响标注质量	[64, 65]
联邦与去中心化标注	标注与模型更新在本地完成，避免数据集中	联邦标注、参数聚合	降低隐私风险，提升用户参与意愿	通信开销大，标注一致性难控	[11, 65]

避免原始数据集中存储，并可结合差分隐私机制进一步降低隐私泄露风险[65]。此外，自监督学习可与联邦学习结合，在保护隐私的同时利用无标注数据进行表征学习[35]。

差分隐私为隐私约束下的标注提供了严格的数学框架，但隐私噪声会影响标注质量和模型性能。在众包标注等场景中，隐私保护强度、标注效用与法规合规要求之间的权衡仍是重要挑战。未来研究需在现有机制基础上构建隐私感知的标注优化框架，实现隐私保护与标注收益的系统性平衡[11, 16]。

6　总结

本文面向 5G-A 与 6G，无线系统的规模化与智能化对高质量标注数据提出更高要求，而信号连续性、强时变性及场景依赖使传统人工标注难以持续适用。本文从数据模态、标签形态与真值来源出发，系统分析典型场景中的标注流程与核心挑战，并归纳主动学习、弱监督、半监督、自监督及噪声鲁棒学习等主要技术路线，同时讨论基础模型与大语言模型在自动标注与伪标签生成中的作用。

综合来看，人工智能方法能够在降低标注成本的同时提升标签质量，是支撑无线通信数据驱动智能化的关键技术。进一步地，数据标注正由离线准备环节向支撑网络持续优化的基础能力演化。面向 6G 与 AI-RAN，基于大模型的自动标注、人机协同与持续学习机制将成为重要发展方向。

参考文献：

[1] O’SHEA T, HOYDIS J. An introduction to deep learning for the physical layer [J]. IEEE Transactions on Cognitive Communications and Networking, 2017, 3(4): 563-75.

[2] 谷志群, 张佳玮, 纪越峰, et al. 数据与模型协同驱动的智能光网络架构与关键技术 [J]. 电信科学, 2024, 38(7): 18-30.

[3] O’SHEA T J, ROY T, CLANCY T C. Over-the-air deep learning based radio signal classification [J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(1): 168-79.

[4] SOLTANI N, ZHANG J, SALEHI B, et al. Learning from the best: Active learning for wireless communications [J]. IEEE Wireless Communications, 2024, 31(4): 177-83.

[5] 章坚武, 王路鑫, 孙玲芬, et al. 人工智能在 5G 系统中的应用综述 [J]. 电信科学, 2024, 37(5): 14-31.

[6] 李攀攀, 谢正霞, 乐光学, et al. 基于深度学习的无线通信接收方法研究进展与趋势 [J]. 电信科学, 2024, 38(2): 1-17.

[7] GEBRU T, MORGENSTERN J, VECCHIONE B, et al. Datasheets for datasets [J]. Communications of the ACM, 2021, 64 (12): 86-92.

[8] HILBURN B, WEST N, O'SHEA T, et al. SigMF: The signal metadata format; proceedings of the Proceedings of the GNU Radio Conference, F, 2018 [C].

[9] O’SHEA T J, CORGAN J, CLANCY T C. Convolutional radio modulation recognition networks; proceedings of the International conference on engineering applications of neural networks, F, 2016 [C]. Springer.

[10] O'SHEA T J, WEST N. Radio machine learning dataset generation with gnu radio; proceedings of the Proceedings of the GNU radio conference, F, 2016 [C].

[11] VAN DEN BERGH B, GIUSTINIANO D, CORDOBéS H, et al. Electrosense: Crowdsourcing spectrum monitoring; proceedings of the 2017 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), F, 2017 [C]. IEEE.

[12] YANG J, CHEN X, ZOU H, et al. SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing [J]. Patterns, 2023, 4(3).

[13] ZHANG Y, ZHENG Y, QIAN K, et al. Widar3. 0: Zero-effort cross-domain gesture recognition with Wi-Fi [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44 (11): 8671-88.

[14] ALKHATEEB A. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications [J]. arXiv preprint arXiv:190206435, 2019.

[15] RAPPAPORT T S, MACCARTNEY G R, SAMIMI M K, et al. Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design [J]. IEEE transactions on Communications, 2015, 63(9): 3029-56.

[16] CALVO-PALOMINO R, CORDOBéS H, ENGEL M, et al. Electrosense+: Crowdsourcing radio spectrum decoding using IoT receivers [J]. Computer Networks, 2020, 174: 107231.

[17] WELINDER P, BRANSON S, PERONA P, et al. The multidimensional wisdom of crowds [J]. Advances in neural informa-

tion processing systems, 2010, 23.

[18] 王金强, 孙闽红, 唐向宏, et al. 小样本下雷达复合干扰半监督迁移学习识别方法 [J]. 电信科学, 2024, 39(10): 15-28.

[19] HAO C, WAN X, FENG D, et al. Satellite-based radio spectrum monitoring: Architecture, applications, and challenges [J]. IEEE Network, 2021, 35(4): 20-7.

[20] PFAMMATTER D, GIUSTINIANO D, LENDERS V. A software-defined sensor architecture for large-scale wideband spectrum monitoring; proceedings of the Proceedings of the 14th International Conference on Information Processing in Sensor Networks, F, 2015 [C].

[21] LIU W, CHWALISZ M, FORTUNA C, et al. Heterogeneous spectrum sensing: challenges and methodologies [J]. EURASIP Journal on Wireless Communications and Networking, 2015, 2015(1): 70.

[22] SETTLES B. Active learning literature survey [J]. 2009.

[23] BOEGNER L, GULATI M, VANHOY G, et al. Large scale radio frequency signal classification [J]. arXiv preprint arXiv: 220709918, 2022.

[24] SENER O, SAVARESE S. Active learning for convolutional neural networks: A core-set approach [J]. arXiv preprint arXiv: 170800489, 2017.

[25] NORTHCUTT C, JIANG L, CHUANG I. Confident learning: Estimating uncertainty in dataset labels [J]. Journal of Artificial Intelligence Research, 2021, 70: 1373-411.

[26] ZHU G, HU Y, GAO W, et al. CSI-Bench: A Large-Scale In-the-Wild Dataset for Multi-task WiFi Sensing [J]. arXiv preprint arXiv:250521866, 2025.

[27] HUANG S, LI K, YOU D, et al. Wimans: A benchmark dataset for wifi-based multi-user activity sensing; proceedings of the European Conference on Computer Vision, F, 2024 [C]. Springer.

[28] ZHOU Z-H. A brief introduction to weakly supervised learning [J]. National science review, 2018, 5(1): 44-53.

[29] FERRAND P, GUILLAUD M, STUDER C, et al. Wireless channel charting: Theory, practice, and applications [J]. IEEE Communications Magazine, 2023, 61(6): 124-30.

[30] ALHUSSEIN O, ZHANG N, MUHAIDAT S, et al. Active ML for 6G: Towards Efficient Data Generation, Acquisition, and Annotation [J]. arXiv preprint arXiv:240603630, 2024.

[31] RATNER A J, DE SA C M, WU S, et al. Data programming: Creating large training sets, quickly [J]. Advances in neural information processing systems, 2016, 29.

[32] TARVAINEN A, VALPOLA H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results [J]. Advances in neural information processing systems, 2017, 30.

[33] SOHN K, BERTHELOT D, CARLINI N, et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence [J]. Advances in neural information processing systems, 2020, 33: 596-608.

[34] BERTHELOT D, CARLINI N, GOODFELLOW I, et al. Mixmatch: A holistic approach to semi-supervised learning [J]. Advances in neural information processing systems, 2019, 32.

[35] ERICSSON L, GOUK H, HOSPEDALES T M. How well do self-supervised models transfer?; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2021 [C].

[36] LI M, LI O, LIU G, et al. Generative adversarial networks-based semi-supervised automatic modulation recognition for cognitive radio networks [J]. Sensors, 2018, 18(11): 3913.

[37] BALEVI E, ANDREWS J G. Wideband channel estimation with a generative adversarial network [J]. IEEE Transactions on Wireless Communications, 2021, 20(5): 3049-60.

[38] CHI G, YANG Z, WU C, et al. RF-diffusion: Radio signal generation via time-frequency diffusion; proceedings of the Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, F, 2024 [C].

[39] FOOLADGAR F, TO M N N, MOUSAVI P, et al. Manifold DivideMix: A semi-supervised contrastive learning framework for severe label noise; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2024 [C].

[40] LIU B, GAO S, LIU X, et al. WiFo: Wireless foundation model for channel prediction [J]. Science China Information Sciences, 2025, 68(6): 162302.

[41] BUFFELLI D, DAS S, LIN Y-W, et al. Towards a Foundation Model for Communication Systems [J]. arXiv preprint arXiv: 250514603, 2025.

[42] BOATENG G O, SAMI H, ALAGHA A, et al. A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions [J]. IEEE Communications Surveys & Tutorials, 2025.

[43] SIINO M, GIULIANO F, TINNIRELLO I. Integrating Large Language Models into Network Testbeds: A Novel Approach for Automated Experimentation and Optimization [J]. Available at SSRN 5187143.

[44] ASH J T, ZHANG C, KRISHNAMURTHY A, et al. Deep batch active learning by diverse, uncertain gradient lower bounds [J]. arXiv preprint arXiv:190603671, 2019.

[45] SHI Y, DAVASLIOGLU K, SAGDUYU Y E, et al. Deep learn-

ing for RF signal classification in unknown and dynamic spectrum environments; proceedings of the 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), F, 2019 [C]. IEEE.

[46] RATNER A, BACH S H, EHRENBERG H, et al. Snorkel: rapid training data creation with weak supervision [J]. The VLDB Journal, 2020, 29(2): 709-30.

[47] BACH S H, RODRIGUEZ D, LIU Y, et al. Snorkel drybell: A case study in deploying weak supervision at industrial scale; proceedings of the Proceedings of the 2019 International Conference on Management of Data, F, 2019 [C].

[48] STUDER C, MEDJKOUH S, GONULTAş E, et al. Channel charting: Locating users within the radio environment using channel state information [J]. IEEE Access, 2018, 6: 47682-98.

[49] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-44.

[50] YANG L, ZHANG Z, SONG Y, et al. Diffusion models: A comprehensive survey of methods and applications [J]. ACM computing surveys, 2023, 56(4): 1-39.

[51] ROY T, O'SHEA T, WEST N. Generative adversarial radio spectrum networks; proceedings of the Proceedings of the ACM Workshop on Wireless Security and Machine Learning, F, 2019 [C].

[52] CHEN Y, SHEN X, HU S X, et al. Boosting co-teaching with compression regularization for label noise; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2021 [C].

[53] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models [J]. arXiv preprint arXiv:210807258, 2021.

[54] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33: 1877-901.

[55] LEE D-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks; proceedings of the Workshop on challenges in representation learning, ICML, F, 2013 [C]. Atlanta.

[56] NGUYEN D N, HOANG D T, DOBRE O A, et al. Generative AI for Communications Systems: Fundamentals, Applications, and Prospects [M]. John Wiley & Sons, 2026.

[57] KAYA O, KARABULUT M A, SHAH A S, et al. Modulation Classifier Based on Deep Learning for Beyond 5G Communications; proceedings of the 2024 47th International Conference on Telecommunications and Signal Processing (TSP), F, 2024 [C]. IEEE.

[58] ELEZI I, YU Z, ANANDKUMAR A, et al. Not all labels are equal: Rationalizing the labeling costs for training object detection; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2022 [C].

[59] SCULLEY D, HOLT G, GOLOVIN D, et al. Hidden technical debt in machine learning systems [J]. Advances in neural information processing systems, 2015, 28.

[60] BRECK E, CAI S, NIELSEN E, et al. The ML test score: A rubric for ML production readiness and technical debt reduction; proceedings of the 2017 IEEE international conference on big data (big data), F, 2017 [C]. IEEE.

[61] TZIOUVARAS A, FORTUNA C, FLOROS G, et al. Towards Reliable AI in 6G: Detecting Concept Drift in Wireless Network [J]. arXiv preprint arXiv:250800042, 2025.

[62] MOHSIN M A, UMER M, BILAL A, et al. Continual Learning for Wireless Channel Prediction [J]. arXiv preprint arXiv: 250622471, 2025.

[63] GUO S, ZHONG Y, FENG Z, et al. Towards Transparent 6G AI-RAN: A Survey on Explainable Deep Reinforcement Learning for Intelligent Network Slicing [J]. Journal of Information and Intelligence, 2025.

[64] QU Y, ZHANG Y, WANG Y, et al. Secure and privacy-preserving issues in integrated sensing and communication-enabled wireless networks: a survey [J]. EURASIP Journal on Advances in Signal Processing, 2025.

[65] ROY K, HASAN T F, WU C, et al. FedWiLoc: Federated Learning for Privacy-Preserving WiFi Indoor Localization [J]. arXiv preprint arXiv:251218207, 2025.

[66] YASUNAGA M, LIANG P. Graph-based, Self-Supervised Program Repair from Diagnostic Feedback [Z]//HAL D, III, AARTI S. Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research; PMLR. 2020: 10799--808