董自飞

纳什维尔, 美国 | 13920559013 (微信号) | zifei.dong@vanderbilt.edu | 领英 | 谷歌学术

教育背景

范德堡大学 (Vanderbilt University) **2024.08 – 2026.05 (预计)**
数据科学硕士, *GPA: 3.8/4.0* 纳什维尔, *TN*

北卡罗来纳大学教堂山分校 (UNC-Chapel Hill) **2021.05 – 2024.05**
计算机科学、统计与分析双学士, *GPA: 3.625/4.0,* 优秀毕业生 教堂山, *NC*

凯斯西储大学 (Case Western Reserve University) **2020.08 – 2021.05**
计算机科学 *(*转学至 *UNC), GPA: 4.0/4.0,* 校长荣誉名单 克利夫兰, *OH*

自我介绍

- 精通 CV 复杂数据构造与增强策略（3D 合成、多阶段、弱监督），有效突破模型在复杂与极端场景下的泛化瓶颈。
- 具备扎实的 CNN/Diffusion/LLM 建模基础与跨界迁移能力，独立主导过从技术选型、端到端训练调优到方案落地的全流程。
- 擅长海量异构数据的端到端处理与训练/推理链路优化，在医疗影像与量化分析项目中，持续推动模型结果向业务决策的稳定转化。

专业技能

方向与任务：计算机视觉（CV）、大语言模型（LLM）、医学影像分析、生成式建模、信息抽取、文本分类、数据增强与模型鲁棒性优化
模型训练与推理：PyTorch、TensorFlow、QLoRA、指令微调、CoT 训练、RLHF 风格对齐、SGLang、Unsloth
数据与工程：Pandas、NumPy、Spark、Hadoop、Multiprocessing、Linux
编程语言：Python、R、SQL、Bash（熟练）；Java、C++、MATLAB（具备开发经验）

代表性学术成果

AnyCXR（胸部 X 光任意位姿解剖分割）：面向复杂采集位姿的胸片分割任务，提出基于 **3D 合成数据增强**与 **不完美标注学习**的训练框架，提升模型在跨视角场景下的鲁棒性与泛化能力。
Medical Image Analysis（医学影像顶刊，在审，arXiv）
Dong Z, Wu W, Hao J, Chen T, Weng Z, Zhou B | **角色：项目负责人、算法实现与主笔**

深度学习项目经历

医学影像生成式去噪研究，AIMP Lab，西北大学 **2025 年 12 月 – 至今**
负责人，远程 埃文斯顿，伊利诺伊州

- 开发基于 **VAE** 的 2.5D 医学图像去噪框架，在编码器中引入 **时空注意力机制（STA）**，增强跨帧信息建模能力，缓解运动伪影对图像结构恢复的影响。
- 设计融合 **Dirichlet 连续松弛**与分组 **FSQ** 的潜空间建模方法，改善离散表示与连续图像特征之间的衔接，减少量化误差，并取得 **31.23** 的重建 PSNR。
- 构建潜空间与像素空间联合优化的双阶段 **CFM** 去噪流程，结合 **OT** 进行流场约束与细化，提升去噪结果的结构保真度与纹理细节恢复能力。

联博基金–范德堡联合项目：领域自适应大语言模型训练 **2026 年 1 月 – 至今**
研究协作者，兼职 纳什维尔，田纳西州

- 参与面向长文本理解、结构化信息抽取与细粒度文本分类的大语言模型项目，主要负责基于 Qwen3-4B 的模型训练、调参与实验迭代。
- 围绕真实长文档场景训练并适配大语言模型，完成主题信息抽取与分类任务建模，提升模型在指令跟随、结构化输出与任务泛化方面的表现。
- 基于 **Unsloth**、**QLoRA** 与 **SGLang** 搭建参数高效微调与可扩展推理流程，支持多组提示词、适配器与评测设置下的快速实验迭代。
- 进一步探索 **Chain-of-Thought (CoT)** 监督训练与 **RLHF** 风格对齐方法，提升模型推理能力、输出稳定性与下游任务可用性。

AnyCXR 项目，AIMP Lab，西北大学 **2024 年 4 月 – 至今**

首席研究员，远程 埃文斯顿，伊利诺伊州

- 独立主导 AnyCXR 项目的算法设计与系统实现，面向高噪声、复杂采集位姿下的胸部 X 光影像，构建任意角度解剖结构分割框架；相关成果作为第一作者论文投稿至医学影像顶级期刊。
- 搭建端到端 3D 多阶段合成数据生成与增强流水线，清洗并处理超过 **8TB**、来源于 **5 万余例** CT 的异构数据，有效缓解真实长尾样本不足问题，并提升模型的跨视角泛化能力。
- 设计结合增强 U-Net 与条件联合标注正则化的弱监督学习方案，以处理部分标注数据场景；实现 **54** 类精细解剖结构分割，PA 视图平均 Dice 达 **94%**。
- 进一步构建“分割—分类”一体化下游应用流程，开发基于分割引导的 DenseNet 自动筛查模型，用于心脏肥大、脊柱侧弯等任务，并带来 **2%** 的绝对分类性能提升。

宁波金戈量锐资产管理有限公司 **2025 年 6 月 – 2025 年 8 月**

量化研究实习生 上海，中国

- 负责复杂高噪声时序数据的端到端处理与建模流程，面向 10 年规模的高维日频数据构建清洗、对齐、特征生成与训练流水线，并基于领域先验设计深度学习驱动的预测特征。
- 主导数据加载与模型训练的计算优化，使用 **Numba** 与 **NumPy** 重构关键矩阵运算模块，实现训练速度约 **10 倍**提升，显著缩短大规模实验迭代周期。
- 开发并集成广义特征神经网络（Generalized Signature Neural Networks）用于高噪声时序预测任务，在全量数据集上取得 **1.2%** 的 R-squared，验证了深度学习方法在复杂时序场景下的建模有效性。

工程与实习经历

宁波金戈量锐资产管理有限公司 **2023 年 5 月 – 2023 年 8 月**

量化研究实习生 上海，中国

- 面向高噪声、多因子时序数据开展特征筛选与稳定性分析，使用 Python 线性模型进行特征选择，提取具备长期稳定性的低频信号。
- 结合时间序列建模与统计分析方法，评估特征在不同时间窗口下的鲁棒性与可迁移性，支持后续建模方案的选择与迭代。
- 构建基于 **Decision Tree** 与 **ARIMA** 的 Stacking 模型，用于风险估计与回撤预测，并完成回测验证，形成可落地的模型评估流程。

北卡罗来纳州大流行恢复办公室 (NCPRO) **2022 年 6 月 – 2022 年 8 月**

软件工程师实习生 罗利，北卡罗来纳州

- 使用 **pandas** 设计并自动化数据处理流水线，集成 OneDrive 支出数据，支持 **200+** 企业预算审核与资格分析，使审计团队每周工作量减少约 **1 天**。
- 基于 **Python**（**shutil, os**）开发文档管理与索引系统，完成 **1,000+** 份遗留文档的整理、检索与归档，显著提升数据访问效率。

Ultimate Consequences 研究小组，范德堡大学 **2025 年 3 月 – 至今**

数据架构师与工程师，兼职 纳什维尔，田纳西州

- 构建并部署端到端数据基础设施，涵盖标准化数据库架构、自动化 R 数据采集管道（5 个动态数据源）及基于 **LLM** 的数据增强与验证流程，支持 **639 名个体**生命史与死亡率数据的稳定管理与分析。
- 重构 R 遗留代码并搭建可复用数据处理框架，提升跨研究人员协作效率与分析流程稳定性，使团队整体工作效率提升 **50%**。

董伟实验室，天津大学 **2023 年 8 月 – 2024 年 4 月**

实验室研究助理，兼职 天津，中国

- 处理并清洗多模态研究数据（问卷、EEG、EDA），使用 **Python (pandas, scipy)** 构建数据整理流程，并在 **R** 中完成基础统计分析，支持 VR 混合学习研究。
- 开发访谈文本的情感与主题分析流程（NLTK、TextBlob），并协助基于 **SteamVR SDK** 的实验数据采集系统，实现行为数据与生理信号的同步记录。

刘新桥研究小组，天津大学 **2022 年 8 月 – 2023 年 5 月**

实验室研究助理，兼职 天津，中国

- 清洗并标准化 **2,812 名**学生的纵向量表数据，使用 **Python (pandas, NumPy)** 搭建数据处理流程，并在 **R (lavaan)** 中完成统计建模分析。
- 使用 **Matplotlib** 与 **Plotly** 进行变量关系与模型路径可视化，并优化数据处理与结果展示流程，支持模型拟合与研究结论解释。