赛题九：材料科学图像曲线识别与智能解析

一、赛题背景

材料科技文献中的曲线是承载材料属性的核心载体之一。例如应力-应变曲线包含材料的强度、延伸率和加工硬化率等力学性能信息，循环伏安曲线反映电化学反应动力学过程，差示扫描量热曲线表征热诱导的相变特性等，谱曲线包含材料成分或结构信息等。据统计，超过 70%的材料科技文献中包含着曲线图像。然而，目前科研实践中仍普遍依赖如 GetData 等人工数字化软件进行曲线数据的提取，存在效率低、主观误差大（可达 1%–3%）等问题，无法满足曲线图像数值自动化和高精度提取和分析的需求。本赛题旨在借助人工智能，突破曲线图像数值化和智能解析的技术瓶颈，推动材料科学曲线图像数据获取流程的自动化和智能化发展。

二、赛题应用场景

以新型耐蚀材料研发为例，在分析海量历史文献时，需从数千份研究报告中提取循环伏安曲线数据，并基于曲线的特征峰谷值解析氧化还原电位等关键参数。传统人工标注方式不仅存在观误差问题，更难以应对海量数据快速增长对处理效率的要求。常规图表解析工具在应对分段坐标、曲线遮挡等复杂场景时，往往存在识别精度不足的问题，导致实际应用中数据提取成功率显著下降。开发能够精准解析曲线图像的新型智能算法已成为跨行业数据提取的共性技术需求。本赛题要求开发智能解析算法，实现多类型曲线的自动坐标识别、数据点提取及批量处理，输出结构化数据。本赛题旨在借助人工智能技术突破曲线智能解析的技术瓶颈，推动材料科学数据获取流程向自动化、智能化变革迈进。

三、赛题任务

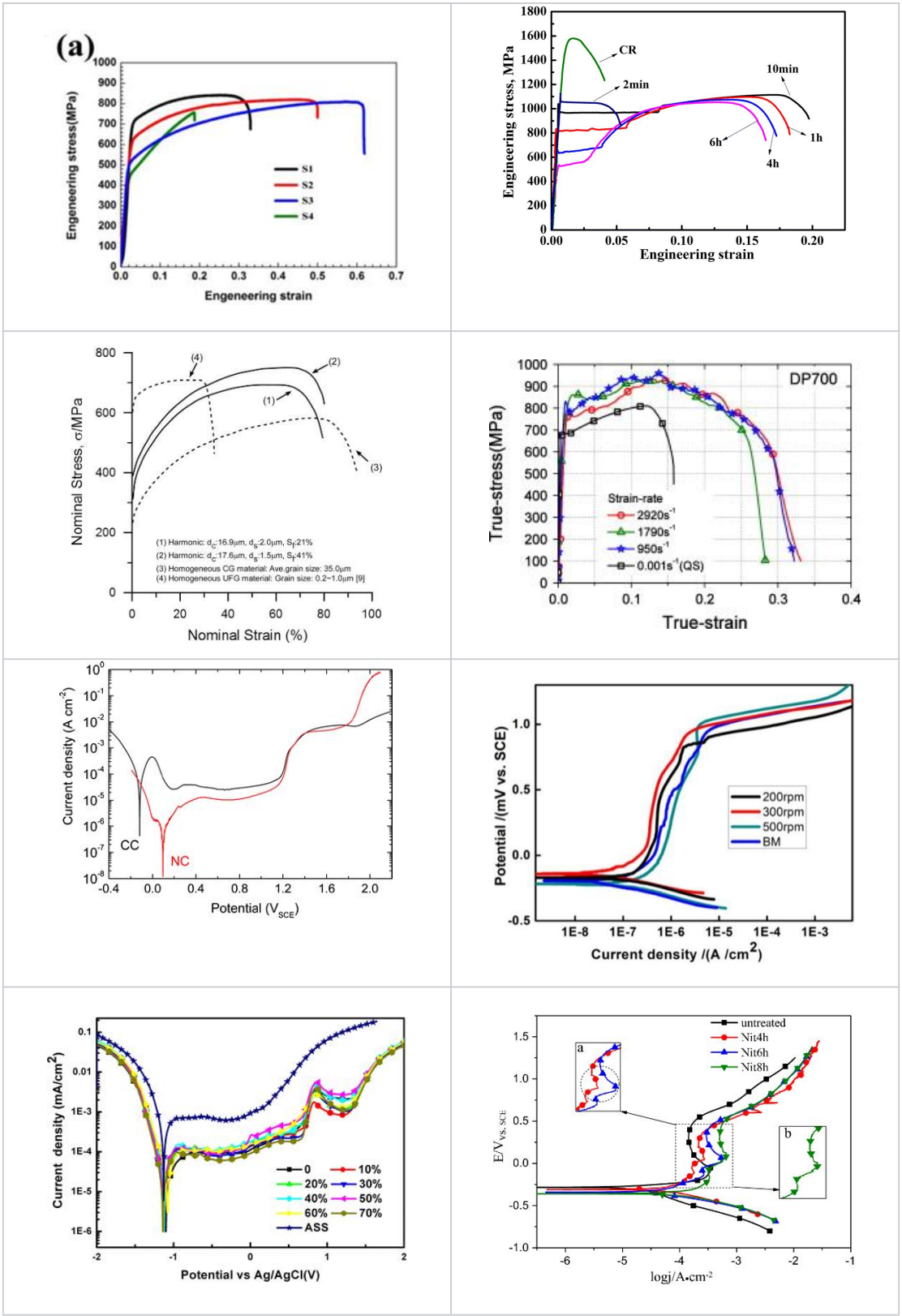
参赛团队需构建针对曲线图片场景解析能力的算法系统，着重解决材料曲线图像中存在的坐标轴分段，非线性刻度、多重曲线交叉/粘连、图例遮挡、扫描噪点等实际干扰问题。算法需实现像素级曲线提取、坐标语义解析与图例匹配的多模态特征融合，并可选将材料学物理约束条件知识（如应力-应变曲线的单调性、循环曲线的周期性等）融入其中，以提升识别的可靠性。此外，算法还需克服设备差异性带来的挑战，并能够兼容电化学极化曲线、差示扫描量热曲线等多种材料表征图像格式，满足工业级跨域泛化的应用需求。

四、数据集及数据说明

数据集采用需参赛团队自行采集或人工生成仿真数据方式获取。参赛团队可采用公开文献数据，从文献的 pdf 文件中，提取各类曲线图片（带图例）用于训练。

数据格式案例如下：

1. 输入图片（通常包含坐标系、曲线、图例等关键要素）：



2. 标准提取结果：

输出结果主要包括以下三类结构化信息：（1）每条曲线对应的数值序列；（2）图像元数据的提取结果，包括坐标轴物理量及其单位、曲线与图例的对应关系。

3. 输出结果示例：

最终输出的结果形成一个 Excel 文件，下方结果为图片（a）对应的抽取结果，从左到右的数据列，依次需要填入的是图片名称（figure_name）、子图标号（figure_index）、图题（figure_title）、x 轴标题（x-lable）、y 轴标题（y-label）、曲线对应图例（sample）、曲线采样点（point_coordinates）、note（图片出现的其它文字说明）。

figure_name (id)	figure_index	figure_title	x-label	y-label	sample	point_coordinates	note
stress_strain_curve_test.png	(a)		Engeneering strain	Engenerring stress(MPa)	S1	0.00003, 12.22707 0.00277, 27.94760	
stress_strain_curve_test.png	(a)		Engeneering strain	Engenerring stress(MPa)	S2	0.00003, 12.22707 -0.00127, 35.80786	
stress_strain_curve_test.png	(a)		Engeneering strain	Engenerring stress(MPa)	S3	0.00001, 4.36681 0.00276, 25.32751	
stress_strain_curve_test.png	(a)		Engeneering strain	Engenerring stress(MPa)	S4	-0.00133, 6.98690 0.00279, 35.80786	

（1） figure_name：直接采用原始图片文件名，保持原名称无需修改；

（2） figure_index：若为多子图形式，标注对应子图编号（如（a）、（b）等），单个曲线图情况为空；

（3） figure_title：提取图片顶部的主标题内容，若无则为空；

（4） 按顺序连续标注数据点，特别注意曲线拐点处，确保每条曲线采点总数量不少于 128；

（5） 严格遵循 X 轴升序方向采集（从左到右），若遇到震荡曲线时，按主要趋势采样；

（6） 部分图片包含对数坐标轴，需要与传统线性坐标轴区分开，最终采样点的值统一转换到线性轴表示。

待解决关键问题：

关键问题	重要程度
坐标轴分段	★★
坐标轴非线性	★★★
多重坐标轴	★
曲线遮挡	★★★
多曲线重叠	★★★
点线图提取关键点	★★★
曲线图按步长取值	★★★
曲线图例对应	★★★
提取值包含极值点	★★
采样点与图例对齐	★★★
部分图片分辨率较低	★★

五、算法设计要求

（一） 鼓励参赛者采用基于深度学习的算法框架，构建具备端到端处理能力的视觉目标检测模型。可以尝试融合 YOLO、Faster R-CNN 等主流检测网络的优势结构，同时引入 Transformer、DETR、ViT 等具备全局建模能力的跨模态特征融合方法，以提升对图像中坐标轴、刻度、文本及曲线的多任务解析能力。

（二） 鼓励参赛者在现有基础上提出具有创新性的融合策略，如注意力机制增强的多尺度特征提取、自监督预训练辅助定位、或领域知识驱动的先验结构引导，旨在进一步提升检测精度、泛化能力与鲁棒性。明确禁止使用任何封装式商业软件 API（如 WebPlotDigitizer、PlotDigitize 等）直接提取图像数据，所有算法模块须由参赛队自主设计或基于开源框架进行合理扩展。

（三） 算法设计应兼顾实用性与可扩展性，具备良好的模块化结构，能够灵活部署于不同计算资源配置的设备（如 GPU 服务器、边缘计算终端）上，支持批量处理与多线程加速。在面对高分辨率图像或千量级图像数据时，系统应保持稳定的运行性能与可控的内存占用，并具备故障容错机制及高可靠性的数据输出能力。

六、性能指标要求

本赛题以钢铁类材料的典型曲线图像为基准测试对象，构建标准化的性能评估体系，初赛和复赛各包含 50 张标注图像（应力-应变、极化曲线、谱图等）。性能评价主要包括三方面：使用动态时间规整距离（DTW）度量曲线形态相似度，以 F1-score 评估识别的语义准确性并测算 GPU 环境下的运行效率。参赛队需提交完整的训练代码与模型参数，支持可复现性验证。赛题最终目标在于实现超越专业人工数字化精度的智能系统，通过引入物理约束增强的小样本学习方法，显著提升材料数据处理效率，为高通量材料研发与数据库构建提供核心支撑能力。系统性能需达到：曲线数值提取误差小于 3%、图例匹配准确率高于 97%。

七、功能要求

本系统需支持多格式输入图像（如 PNG、JPG 等），输入图像通常包含坐标系、曲线、图例等关键要素，输出结果包括以下三类结构化信息：（1）每条曲线对应的数值序列（CSV 格式）；（2）图像元数据的提取结果，包括坐标轴物理量及其单位、曲线与图例的对应关系。

八、开发环境

（一） 软件环境

1. 参赛者开发过程的软件系统不限（Windows、Linux、Mac 等），最终代码应当保证在 Linux 系统下测试通过。开发工具（IDE）不限。

2. 开发语言应当以 Python 为主语言。若采用其他语言，参赛者应当合理封装，并以 python 作为胶水语言实现相关接口。相关支持库和框架等，注明版本要求。

3. 深度学习建议以 pytorch 为框架，以提高版本兼容性。

（二） 硬件环境

硬件开发环境不限。最终代码应当至少在标准机器（32 核心，显存 24G，Nvidia 4090）硬件资源下实现数据抽取。标准机器下的图片处理速度作为模型推理效率评分。

九、成绩评价

线下最终成绩由客观评分（85%）与主观评分（15%）组成。客观评分依据赛方提供的人工标注测试集结果计算，包括：曲线提取精度（DTW 度量，占比 70%）、识别准确率（20%）和模型推理效率（GPU 内存占用与运行时间加权评分，占比 10%）。主观评分则由材料科学专家从算法创新性（50%）、工程落地价值（50%）两个维度进行综合评审。总成绩以 100 分为满分，其他方案按照比例进行归一化调整。（国二、国三成绩不涉及 30%主观评分部分）

十、解题思路

通过自适应滤波、颜色空间分析、动态阈值分割和透视变换，对曲线图像进行降噪、校正和增强处理，确保图像清晰、结构完整。随后采用双阶段坐标解析系统，先通过霍夫变换与连通域分析定位坐标轴与刻度，再利用改进 OCR 识别数值，结合语义分割和趋势预测技术提取曲线数据，并自动转换为物理量，兼容线性、对数等多种坐标系。模型选用神经网络并结合迁移学习，加入风格迁移与物理约束提升小样本表现。部署阶段开发并行处理流水线与误差修正机制，支持多格式图像的高效解析。重点提升坐标识别鲁棒性与遮挡数据的重构能力，融合材料专业知识并建立多维度评估体系，确保提取结果工程上的实用性。

十一、参考资源

[1] Cliche, M., Rosenberg, D., Madeka, D., & Yee, C. （2017）. Scatteract: Automated extraction of data from scatter plots. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10 （pp. 135-150）. Springer International Publishing.

[2] Siegel, N., Horvitz, Z., Levin, R., Divvala, S., & Farhadi, A. （2016）. Figureseer: Parsing result-figures in research papers. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14 （pp. 664-680）. Springer International Publishing.

[3] Luo, J., Li, Z., Wang, J., & Lin, C. Y. （2021）. Chartocr: Data extraction from charts images via a deep hybrid framework. In Proceedings of the IEEE/CVF winter conference on applications of computer vision （pp. 1917-1925）.

[4] Drevon, D., Fursa, S. R., & Malcolm, A. L. （2017）. Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. Behavior modification, 41（2）,

323-339.

十二、提交要求

1. 可运行的 python 源代码及结果文件。参赛方需要在赛事方提供的测试集上进行预测，代码应当包含算法的主体结构，并可直接运行成功，确保结果的可复现性。代码最终需要生成名为"extractor.xlsx"文件，用以统一评分。extractor.xlsx 数据结果格式如附件所示。

2. 代码的报告文档 word 文件: "算法说明.docx"，介绍算法的主要原理、创新点说明、代码逻辑（训练、外推等）、参考文献等，不少于 3000 字。

3. 附属文件夹"support"。代码运行所需参数文件、权重文件等支撑材料。

4. 针对曲线图像数据集，将抽取结果保存为 "extractor.xlsx" 文件，并压缩为 ZIP 格式后提交。

十三、联系方式

赛项交流 QQ 群：982830277

邮　　箱：shaohan.tian@hotmail.com

报名官网：www.aicomp.cn