

编号: JT000001



君同智能
www.juntong.ai

空相-图像模型测评专家 用户使用手册

浙江君同智能科技有限责任公司

目录

1	概览.....	1
1.1	产品简介	1
1.1.1	产品技术架构	1
1.1.2	产品优势	1
1.2	基础知识	2
1.2.1	对抗攻击	2
1.2.2	样本检测	2
1.2.3	算法评估	2
2	使用说明.....	2
2.1	创建任务	2
2.1.1	任务信息	2
2.1.2	测评配置	3
2.1.3	测评策略	3
2.1.4	测评过程	5
2.1.5	测评结果	5
2.2	任务列表	6

1 概览

1.1 产品简介

您好！欢迎使用「空相-图像模型测评专家」。该产品可用于对市面上的图像模型提供测评服务。客户上传需要测评的图像模型，该产品可对上传的模型进行多种攻击，测评其有效性并生成测评报告。

1.1.1 产品技术架构

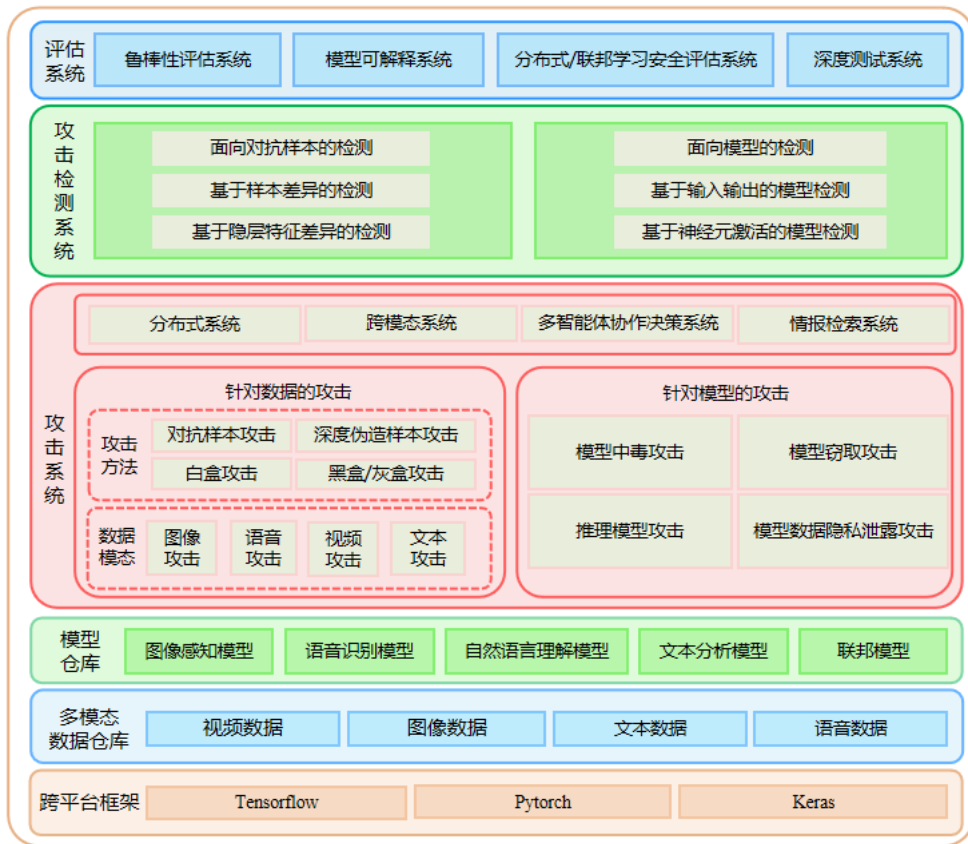


图 1-1-1 平台技术架构

1.1.2 产品优势

- 强大的科研能力，领先的核心算法：
 - 自主原创国际领先的攻防技术；深受国际学界业界认可的技术积累；
 - AI 安全领域的多项世界大赛冠军
- 低/零编码：
 - 全面 API 化，一键式定制产品模块；零/低成本的安全部署；客户使用门槛及学习成本最小化

- 用户友好的量度评估，可视可量的报告数据：
通俗易懂的安全性可视化表达；交互式的安全评分及报表；简明直观的安全风险演示还原；模型安全性一目了然
- 白盒服务，黑盒实施：
通俗易懂的安全性可视化表达；让客户黑盒一不获取用户模型及敏感信息，严格保护被测评方知识产权

1.2 基础知识

1.2.1 对抗攻击

对抗攻击即恶意攻击者通过设计有针对性的数值型向量让机器学习模型做出错误判断；本平台融合了 100 多种攻击方法，全面筛选 AI 模型漏洞，为用户提供一键式攻击入口，展现 AI 算法对抗攻击的直观威胁力。

1.2.2 样本检测

样本检测即在攻击者实施针对性的攻击；生成对抗样本后，由本平台内置的 100 多种对抗样本检测算法，对样本进行检测，判断样本的恶意性，杜绝恶意样本干扰 AI 模型。

1.2.3 算法评估

算法评估即对模型使用的算法进行多个方面的评估；本平台内置多种 AI 算法评估手段，包括鲁棒性评估、可解释性评估、公平性评估、隐私性评估等，对算法进行全方面、全角度的深度评测。

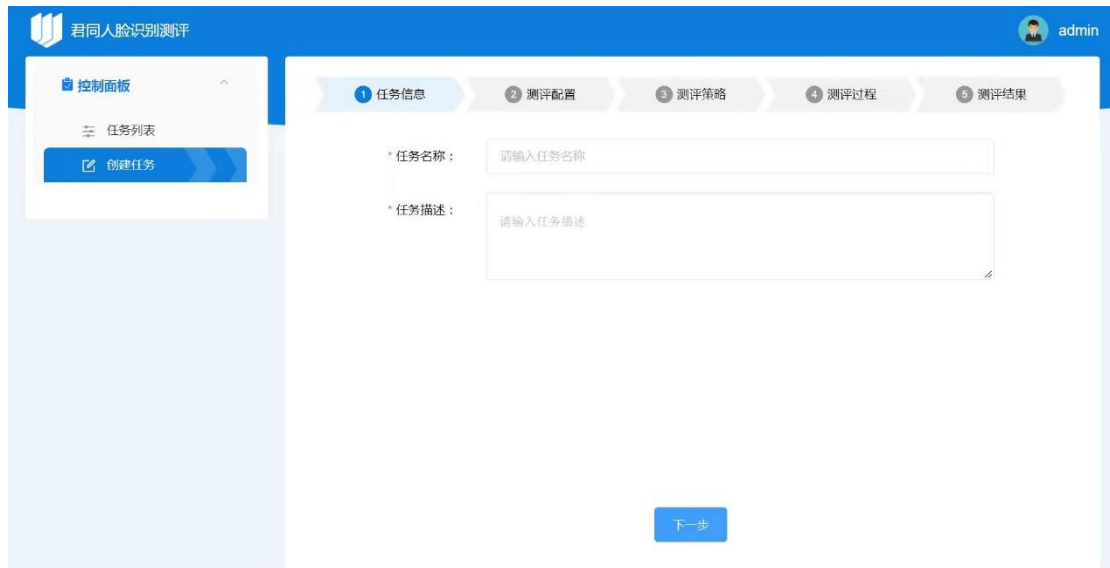
2 使用说明

该产品的主要界面包括任务列表模块和创建任务模块。

2.1 创建任务

2.1.1 任务信息

点击创建任务，首先进行任务信息的设置，包括任务名称和任务描述，点击[下一步]完成任务信息设置。



2.1.2 测评配置

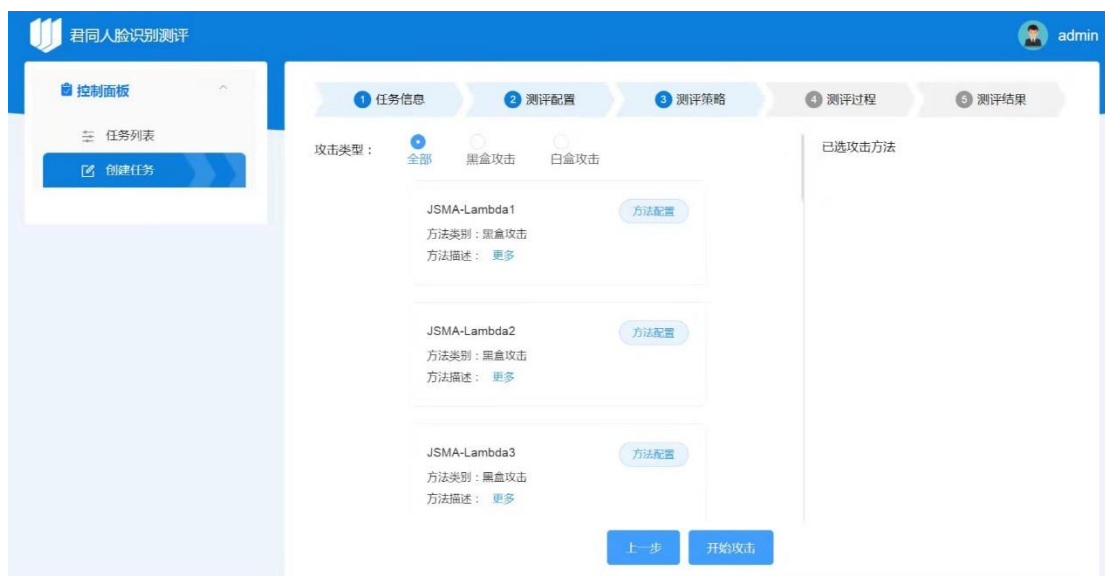
在测评配置页面中，完成场景名称、平台框架、数据集选择的填写，点击[点击上传]按钮完成深度模型代码、深度模型参数的上传，点击[下一步]完成测评配置。



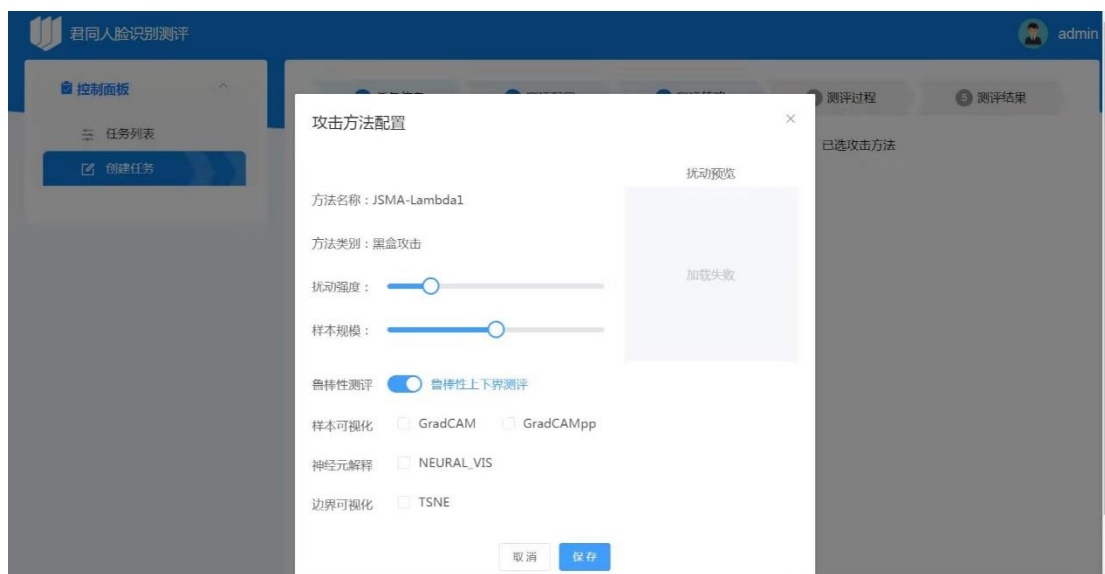
2.1.3 测评策略

在测评策略页面，选择需要的攻击方法，点击[方法配置]进行攻击方法的设置。

空相-图像模型测评专家用户使用手册



在攻击方法配置弹出界面中进行扰动强度设置、样本规模设置、选择的测评方案，分别有鲁棒性测评、样本可视化、神经元解释以及边界可视化，点击[保存]：

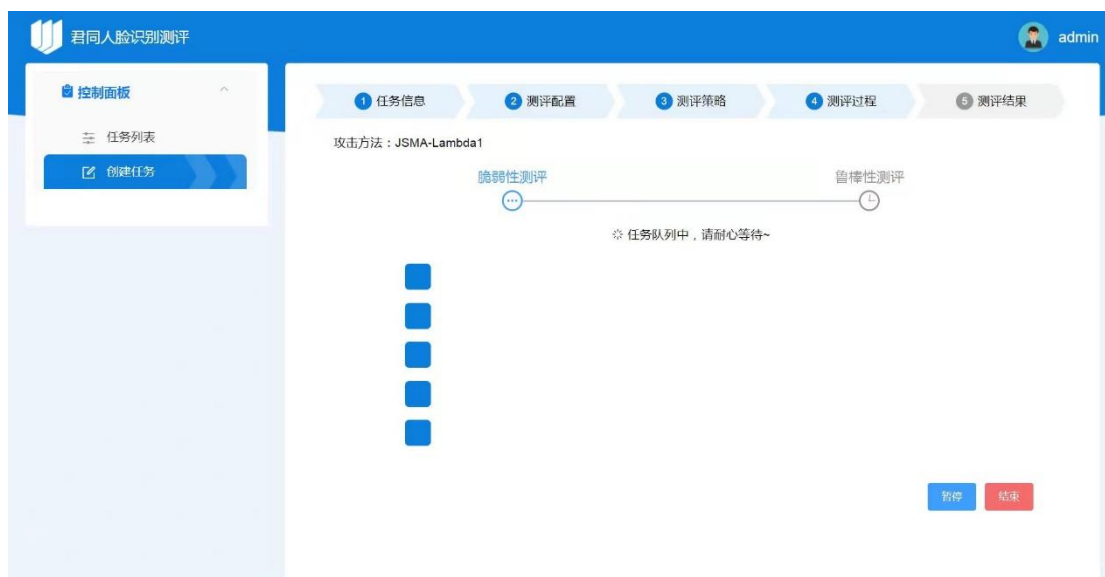


测评策略右侧为已选攻击方法预览，可进行多选。点击[编辑]可对攻击方法配置进行修改，点击[移除]可以删除当前已选攻击方法，点击[开始攻击]进行测评。



2.1.4 测评过程

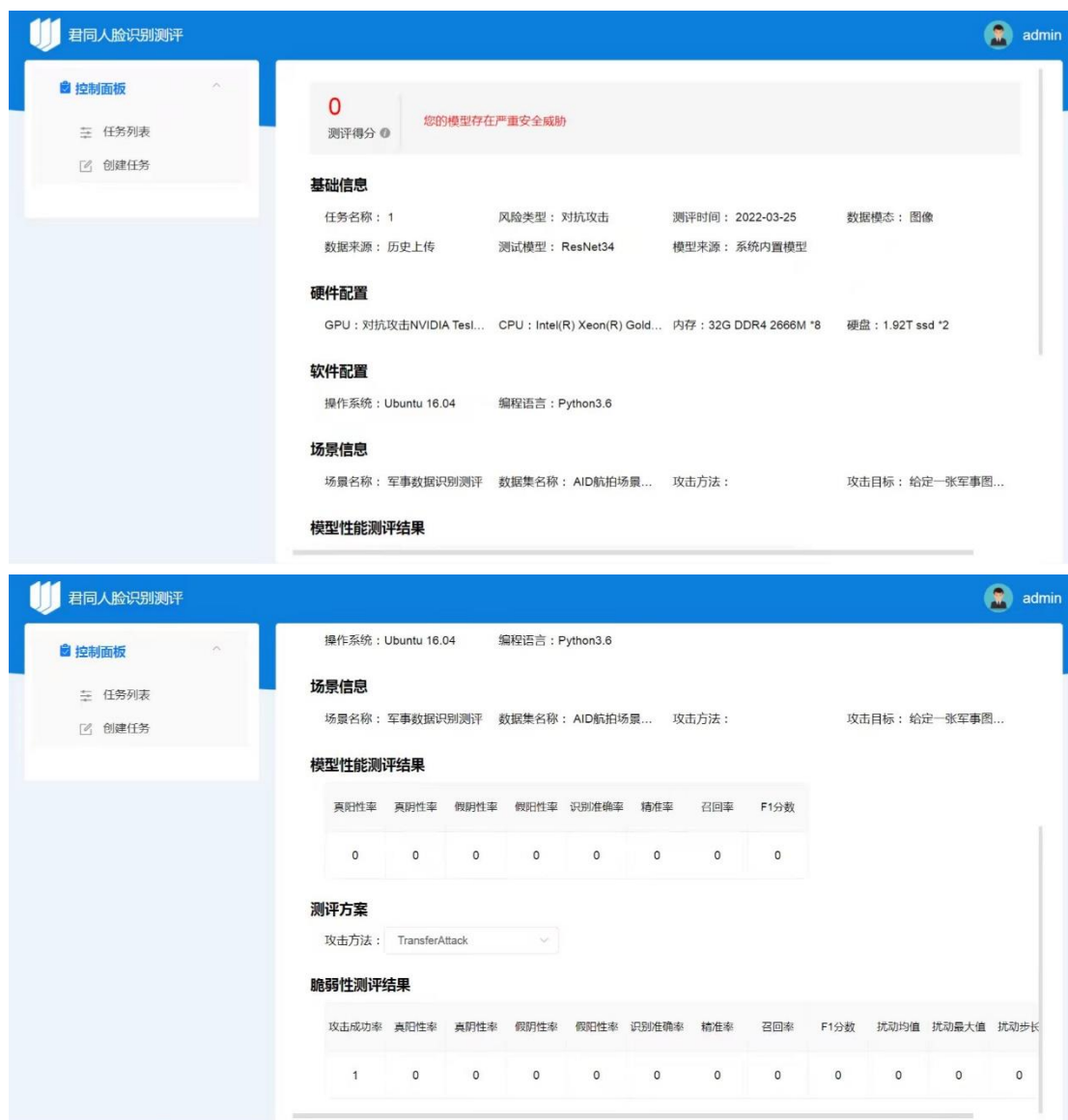
在脆弱性测评过程界面，会显示原本样本，扰动范数可视化和生成的对抗样本及其置信度。点击[暂停]可暂停测评过程，点击[结束]可结束该测评环节；在鲁棒性测评界面，会显示原始样本和对抗样本的图像及其鲁棒边界，点击[暂停]可暂停测评过程，点击[结束]可结束该测评环节。



2.1.5 测评结果

在测评结果界面中，包含测评得分、基础信息、硬件配置、软件配置、模型性能测评结果、测评方案、脆弱性测评结果等信息。点击测评方案下拉菜单可选择不同攻击方法的测评结果：

空相-图像模型测评专家用户使用手册



2.2 任务列表

在任务列表模块中，用户可以查看任务名称、任务 id、创建日期、任务模态、场景名称、模型类型等信息，并且可以查看任务报告。

空相-图像模型测评专家用户使用手册

君同人脸识别测评

控制面板

任务列表

创建任务

任务名称	任务id	创建日期	任务模式	场景名称	模型类型	操作
1	5	2022-03-25 14:43:10	图像	军事数据识别测评	系统内置模型	查看报告
1	4	2022-03-25 14:28:05	图像	军事数据识别测评	用户上传模型	查看报告
1	3	2022-03-25 14:19:29	图像	动物识别-API场景	商用API接口模型	查看报告
测试1	2	2022-03-25 13:23:10	图像	军事数据识别测评	用户上传模型	查看报告
测试	1	2022-03-25 13:19:29	图像	军事数据识别测评	系统内置模型	查看报告

1