附件3

项目实施报告

1. 项目承担单位概况

四川云从天府人工智能科技有限公司是提供人机协同智能操作系统和解决方案的独角兽企业。公司在双层异构深度神经网络、图像（人脸）识别、自然语言处理、时空复合并行计算技术、时空复合并行计算技术等领域，打通了从感知到认知到决策的全链技术。在人脸识别算法模块上也从2011年就从中国科学院开始研究并取得许多成果。与五大行、三大运营商、公安部、民航总局、中科院、证通股份等成立联合实验室，利用特殊行业非结构化数据资源，训练现有算法模型，不断优化算法模型使之更契合行业发展的深层需求，力争实现相关行业的大数据智能化转变。在行业应用方面，云从科技业务涵盖智慧金融、智慧治理、智慧交通、智慧商业等领域，通过行业领先的人工智能、认知计算与大数据技术形成的整合解决方案，已服务400余家银行14.7万网点、30个省级行政区公安。四川云从2020年中标天府新区数字城市建设项目，该项目基于国产化自主可控的系统，围绕“四梁四柱”：天府大脑“四梁”：数据治理、人工智能平台、数字孪生平台、物联网平台，让城市更聪明；建好城市运行“四柱”：“人”、“城”、“境”、“业”的智慧城市应用，让公园城市智慧场景落地。

二、项目概况

（一）项目研究内容

四川云从天府人工智能科技有限公司作为本项目参与单位，负责面向多模态边缘云计算平台建设，集成智能云计算协同管理，重点解决设备接入管理子模块、算法容器引擎、感知数据解析、算法基础管理、级联管理等，形成人工智能算法应用平台协同工作架构，为项目提供整个管理与工作机制。集成云边端一体化应用，重点搭建云边端的架构，支持纵向级联和横向扩展，支持国产化芯片和关键部件，支持软硬件一体化终端的研究与接入管理，达到硬件设备接入及服务的水准，作为智慧城市的硬件基础，为平台AIoT场景落地提供终端载体

1. 项目总体任务和预期目标

1. 项目总体任务

面向多模态边缘云计算平台，集成智能云计算协同管理，重点解决设备接入管理子模块、算法容器引擎、感知数据解析、系统基础管理、开发者中心、级联管理等，形成人工智能算法应用平台协同工作架构，为任务提供整个管理与工作机制。集聚AI产业，加速行业龙头企业发展。助推人工智能产业能级的整体提升，成为新兴的支柱性产业。赋能传统领域应用，助推产业结构升级改造。

2. 预期目标

（1）主要技术、成果和知识产权等指标

1）研制面向不少于10种场景智能终端的边缘计算软件平台，集成视听觉不少于20种算法模型，多模态融合和音视频协同的响应时延小于200ms，可同时支持100路和100种多媒体终端接入。

2）研制统一计算软件架构，支持Tensorflow、Pytorch、Caffe、MXNet等主流深度学习框架，兼容X86、GPU、ARM、RISC-V、ASIC等异构芯片。

3）支持不少于2层的纵向级联和不少于100个计算节点的横向扩展；实现边缘计算的容器化应用管理，支持实时动态的服务加载与资源回收；集成各类安全软件，实现内容加密、隐私保护、代码可溯源。

（2）经济、社会和环境效益

任务围绕新型基础设施、智慧城市和区域数据中心建设需要，通过研究面向多模态的人工智能云边端一体化、深度学习专用算法等关键技术，构建人工智能算法应用平台协同工作架构，集成人工智能数据能力仓和人工智能训练等系统，实现数据资源动态调度、可扩展、可级联的边缘和云边计算，拓展应用场景，在教育、政务、城市管理、医疗等领域开展应用示范，将形成直接经济效益3000万元。

任务实施将有助于形成一批人工智能与经济社会发展深度融合的典型模式，促进人工智能在社会民生领域的广泛应用，打造一批具有重大引领带动作用的人工智能创新高地。

二、项目实施效果及合同指标完成情况

（一）项目总体完成情况及取得的成果、实施的效果综述

根据任务书的要求，项目执行情况如下：

1. 主要技术、成果和知识产权等指标

（1）研制面向不少于10种场景智能终端的边缘计算软件平台，集成视听觉不少于20种算法模型，多模态融合和音视频协同的响应时延小于200ms，可同时支持100路和100种多媒体终端接入。

完成情况：研发完成谛格AI定义小站软硬件一体的边缘计算产品，该产品已经在加油站、环保屋、学校食堂、工厂安全生产、出入口疫情防控、银行网点、监控中心、保安室、园区管理、停车场等12种场景得到应用；该产品集成了包括离岗报警、玩手机报警、打电话报警、烟火报警、抽烟报警、跌倒报警、垃圾遗留报警、人员聚集报警、区域入侵报警、未带安全帽报警、睡岗报警、未带口罩报警、未带厨师帽报警、人数统计、非机动车报警、未穿厨师服报警、车位占用、人流密度、堆物检测、未戴手套、未穿反光衣报警等35种算法模型；多模态融合和视频协同的响应时延测试结果小于200ms；系统接入128路多媒体终端，系统接入138种设备。

1. 研制统一计算软件架构，支持Tensorflow、Pytorch、Caffe、MXNet等主流深度学习框架，兼容X86、GPU、ARM、RISC-V、ASIC等异构芯片。

完成情况：系统支持4种主流深度学习框架，包括Tensorflow、Pytorch、Caffe、MXNet；系统支持的芯片包括：X86（Intel（R） Core（TM）i5-11500）、GPU（登临加速卡）、ARM（海思3559A芯片）。

1. 支持不少于2层的纵向级联和不少于100个计算节点的横向扩展；实现边缘计算的容器化应用管理，支持实时动态的服务加载与资源回收；集成各类安全软件，实现内容加密、隐私保护、代码可溯源。

完成情况：系统支持3层架构纵向级联，同时支持132个计算节点的横向扩展；系统使用Docker启动运行，支持实时动态的服务加载；停止Docker后，CPU的占用率从75%降到7.3%，实现资源的快速回收；通过加密、授权校验、解密实现数据可用不可见，不流通数据本身，只流通数据的计算结果。系统包含50种边缘端基础AI应用，同时系统支持自动生成的SKU应用套件2974种。

1. 知识产权

超额指标完成情况：申请或获得发明专利共计8项，其中授权5项；登记软件著作权4件；团体标准2项。

专利：

[1] 陆先，姜思君，候朝能. 自适应视频文件加速解析方法、装置、设备及存储介质，专利号：ZL202110485426.0，授权日期：2023年1月

[2] 王洋娟. 带数聚治理平台图形用户界面的显示屏幕面板，专利号：ZL202230867680.2，授权日期：2023年4月

[3] 黄娜. 带数聚治理平台图形用户界面的显示屏幕面板，专利号：ZL202230875531.0，授权日期：2023年4月

[4] 吴学纯. 车辆拥堵检测方法、装置及计算机存储介质，专利号：ZL202210428109.X，授权日期：2023年5月

[5] 杜凯，门雷. 存储系统的数据存储方法、装置、控制装置和介质，专利号：ZL202111484720.6，授权日期：2021年12月

[6] 赵希桐，祖广盛，王仲勋，兰天翼，肖万勇. 视频标注方法、装置、电子设备及介质，申请号：CN202310706332.0，申请日期：2023年6月

[7] 段丹丹,晏超,王仲勋,兰天翼,肖万勇. 模型训练方法、韵律预测方法、控制装置及存储介质，申请号：CN202310713141.7，申请日期：2023年6月

[8] 王彦添,张泽伦,赵德芳,王仲勋,张正伟. 图像超分辨重建方法、设备及计算机可读存储介质，申请号：CN202310869533.2，申请日期：2023年7月

软件著作权：

[1] 四川云从天府人工智能科技有限公司. 四川天府新区智慧门户APP软件，版本号: V1.0，登记号: 2022SR0459456，软件开发完成日期: 2021年11月30日，登记日期: 2022年04月12日，权利取得方式: 原始取得，权利范围: 全部权利

[2] 四川云从天府人工智能科技有限公司. 区域核酸检测预决策调度系统，版本号: V1.0，登记号: 2022SR0994783，软件开发完成日期: 2022年05月10日，登记日期: 2022年08月03日，权利取得方式: 原始取得，权利范围: 全部权利

[3] 四川云从天府人工智能科技有限公司. 数字孪生智慧水生态治理系统，版本号: V1.0，登记号: 2022SR0994810，软件开发完成日期: 2022年05月10日，登记日期: 2022年08月03日，权利取得方式: 原始取得，权利范围: 全部权利

[4] 四川云从天府人工智能科技有限公司. 四数字孪生智慧城市治理平台，版本号: V1.0，登记号: 2022SR1309172，软件开发完成日期: 2022年05月10日，登记日期: 2022年08月26日，权利取得方式: 原始取得，权利范围: 全部权利

标准：

[1] 团体标准，《人工智能 深度合成图像系统技术规范》，T/CESA 1197-2022.

[2] 团体标准，《人工智能 智能字符识别技术规范》，T/CESA 1199-2022.

1. 经济、社会和环境效益

任务围绕新型基础设施、智慧城市和区域数据中心建设需要，通过研究面向多模态的人工智能云边端一体化、深度学习专用算法等关键技术，构建人工智能算法应用平台协同工作架构，集成人工智能数据能力仓和人工智能训练等系统，实现数据资源动态调度、可扩展、可级联的边缘和云边计算，拓展应用场景，在教育、政务、城市管理、医疗等领域开展应用示范，将形成直接经济效益3000万元。

任务实施将有助于形成一批人工智能与经济社会发展深度融合的典型模式，促进人工智能在社会民生领域的广泛应用，打造一批具有重大引领带动作用的人工智能创新高地。

完成情况：任务围绕新型基础设施、智慧城市和区域数据中心建设需要，通过研究面向多模态的人工智能云边端一体化、深度学习专用算法等关键技术，构建人工智能算法应用平台协同工作架构，集成人工智能数据能力仓和人工智能训练等系统，与四川天府新区、重庆市梁平区以及多家公司开展合作，形成经济效益1.34亿元。

|  |  |  |  |
| --- | --- | --- | --- |
| 合同名称 | 合同甲方 | 合同乙方 | 合同金额（元） |
| 四川天府新区成都直管区数字城市建设2020-2021年行动计划项目服务合同 | 四川天府新区行政审批局 | 四川云从天府人工智能科技有限公司 | 538,285,467.00（其中1.2亿归属四川云从） |
| 梁平区双桂湖智慧体验公园管理服务平台升级项目软件开发服务合同 | 重庆市梁平区大数据应用发展管理局 | 四川云从天府人工智能科技有限公司 | 1,319,137.00 |
| 数字天府智慧交通建设服务项目采购合同 | 成都雷科特毫米波技术有限公司 | 四川云从天府人工智能科技有限公司 | 6,260,000.00 |
| 2022年天府新区感知源项目（标段一）合同 | 成都杰迈德电子科技有限公司 | 四川云从天府人工智能科技有限公司 | 990,000.00 |
| 天府新区公安分局感知源销售合同 | 四川警桦科技有限公司 | 四川云从天府人工智能科技有限公司 | 1,080,000.00 |
| 独角兽岛启动区公园城市展示厅数字孪生智慧水生态治理系统软件购销合同 | 力方数字科技集团有限公司 | 四川云从天府人工智能科技有限公司 | 888,514.32 |
| 2022年拉萨经开区公安局视频大数据项目销售合同 | 成都甜心商旅科技有限公司 | 四川云从天府人工智能科技有限公司 | 680,000.00 |
| 蒲江县公安局指挥中心改造设备采购合同 | 蒲江县公安局 | 四川云从天府人工智能科技有限公司 | 2,535,646.00 |
| 天府新区公安分局感知源销售合同 | 四川警桦科技有限公司 | 四川云从天府人工智能科技有限公司 | 1,134,000.00 |
| 合 计（万元） | | | 13,488.73 |

（二）项目技术指标完成情况

1. 项目技术方案

多模态边缘云计算平台是集数据通、设备通、应用通为一体，提供AI标准场景应用方案的通用开放平台，支持设备管理、应用管理、引擎管理、权限配置，支持IO配置、语音设置、用户管理、设备共享、SDK服务、预览布局和授权管理等高级设置，支持录像回放和实时预览，通过云边端协同方式为传统行业提供智能化升级服务。



图1 多模态边缘云计算平台搭建

平台基于广泛的智能设备接入能力和行业领先的AI能力，支持云从、海康、大华、宇视IPC、NVR、抓拍机、DCS/DCR等海量设备，以一体机、解决方案和云服务的方式，为用户提供出入口、区域布控、门店、疫情防控和空间视觉五大应用集，支撑机场业务、智慧交通、安防、智慧工地等场景解决方案的落地。提供访客门禁、区域布控、轨迹抓拍、车辆抓拍、人数统计、行为报警、工地监管等核心功能。

平台集成视听觉20多种算法模型，多模态融合和音视频协同的响应时延小于200ms，可同时支持100路和100种多媒体终端接入；支持3层的纵向级联和100个计算节点的横向扩展；实现边缘计算的容器化应用管理；实现内容加密、隐私保护、代码可溯源等安全功能。

2. 关键技术及解决途径（项目解决的关键共性技术、达到的技术性能指标及其技术水平，实际实现的技术指标与合同约定指标对比情况）；

项目研发的deepnet算法框架，通过异构计算的运行时库大量运行在各种终端、边缘和云端的硬件平台和设备上。云从科技的异构计算运行时库兼容性强，广泛支持不同类型的CPU、GPU、NPU等通用的硬件架构。以CPU硬件平台进行划分，可以广泛运行在ARM CPU、Intel CPU、AMD CPU、华为的鲲鹏CPU、飞腾CPU、海光CPU、龙芯CPU等一系列商用的CPU硬件平台上。以GPU／NPU等硬件平台进行划分，可以广泛运行在ARM Mali GPU、Qualcomm Adreno GPU、NVIDIA GPU、AMD GPU、曙光的DCU、华为的昇腾NPU、寒武纪的思元NPU等一系列商用的AI加速硬件平台上。



图2 deepnet算法框架

异构计算的运行时库包括多个主要的模块，一个是模型转换模块，一个是异构推理计算模块，一个是模型压缩量化模块，一个是云端多卡推理服务。

①模型转换

模型转换模块负责将不同标准类型的神经网络模型，譬如Caffe、Tensorflow、Pytorch／ONNX模型转换成云从的模型表达格式，然后进行算子替换、算子融合、内存布局调整等图优化操作，最大化地提升模型对内存的复用，在模型推理阶段提升推理的速度与性能，如下图所示：



图3 模型转换

②型压缩量化

模型压缩量化模块负责对训练好的神经网络模型进行压缩，以在尽量保持模型精度损失最小的情况下进一步提升模型推理的速度和性能。压缩的策略包括低比特的量化（譬如INT8或者INT4的量化）、知识蒸馏（用一个Teacher的大模型训练一个Student的小模型）、稀疏化剪枝（譬如结构化剪枝、非结构化剪枝）。

③异构推理计算

异构推理计算模块负责神经网络模型的在线推理和计算，需要对模型需要的所有算子在不同的硬件平台上进行高效地实现，包括上述的CPU、GPU、NPU等硬件平台。其中最重要的工作是对模型的卷积运算在不同的硬件平台上进行优化，以获取最高性能的推理计算。卷积运算的优化策略包括直接卷积计算、Im2Col＋Pack＋GEMM计算、Winograd计算、Strassen计算、FFT计算等。需要根据硬件平台的型号、支持的指令集、L1/L2/L3 Cache大小、Cache类型、寄存器大小和容量、计算核的数量、卷积核的大小等硬件特性和卷积算子特性，优化上述卷积计算的算法在不同硬件平台上的性能以获取最大化的模型推理速度，如下图所示：

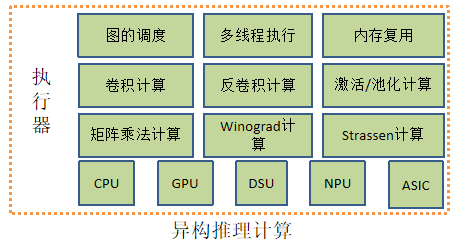


图4 异构推理计算

异构推理计算模块支持多种编程语言，包括SSE、SSE2、AVX2、AVX512、NEON汇编编程语言，CUDA、OpenCL编程语言，DSL、TIK编程语言，BANG编程语言，BM Lang编程语言。其中，SSE、SSE2、AVX2、AVX512、NEON汇编编程语言用在X86 CPU、ARM CPU硬件平台上的算子优化和加速。CUDA、OpenCL编程语言用在NVIDIA GPU、AMD GPU、曙光DCU硬件平台上的算子优化和加速。DSL、TIK编程语言用在华为昇腾硬件平台上的算子优化和加速。BANG编程语言用在寒武纪硬件平台上的算子优化和加速。BM Lang编程语言用在比特大陆硬件平台上的算子优化和加速。

3. 试验方法及结果

研制了四种边缘计算智能终端产品

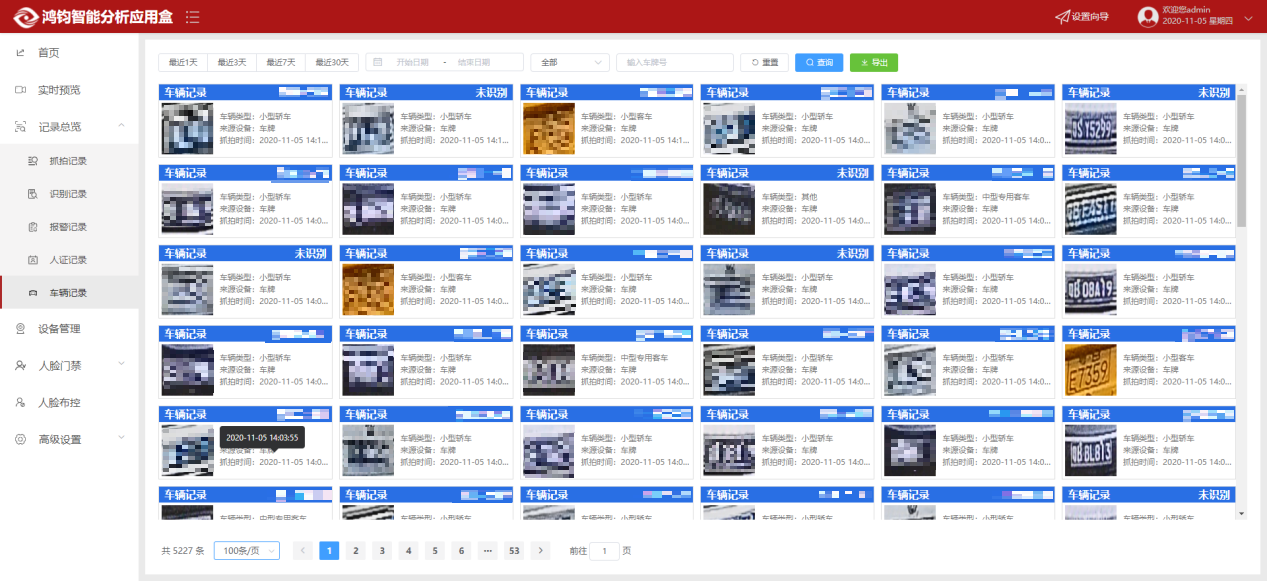
（1）鸿钧智能分析盒



图5 鸿钧智能分析盒

本产品搭载云从自研的人脸抓拍、人脸识别、以及九种人脸属性算法，支持接入云从或第三方网络摄像机、人脸抓拍机的视图信息进行解析识别，可管理云从人脸识别终端、认证合一设备、人脸门控控制终端等智能设备，可实现边缘侧人员管理、数据采集、设备管理闭环。此边缘设备适用于社区、园区、学校、医院、企业、政府单位等场景，实现人员车辆布控报警、智能门禁改造升级。





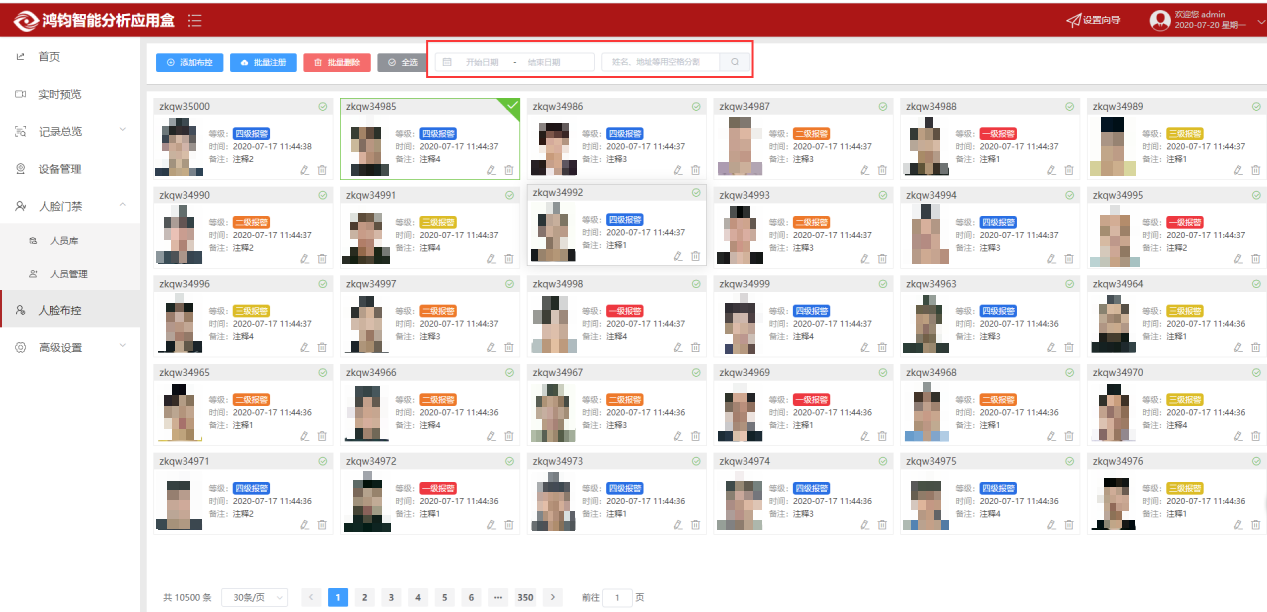


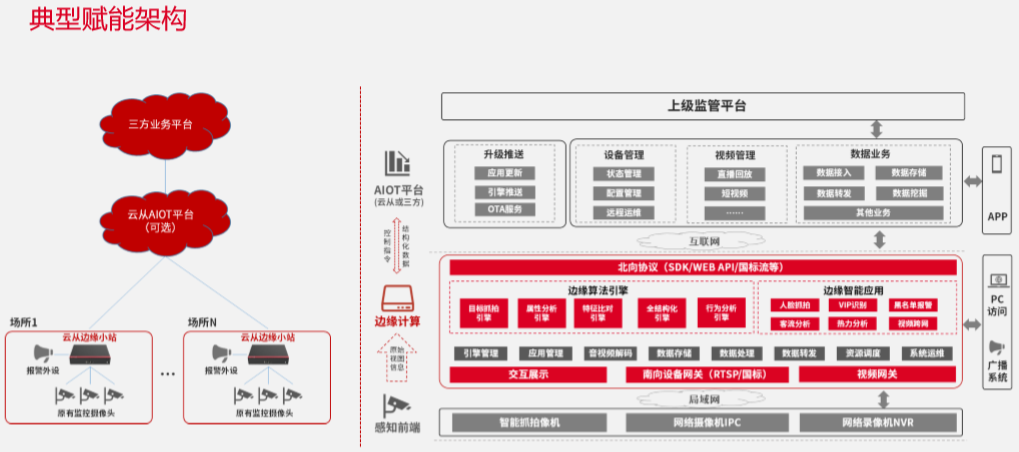
图6 应用界面

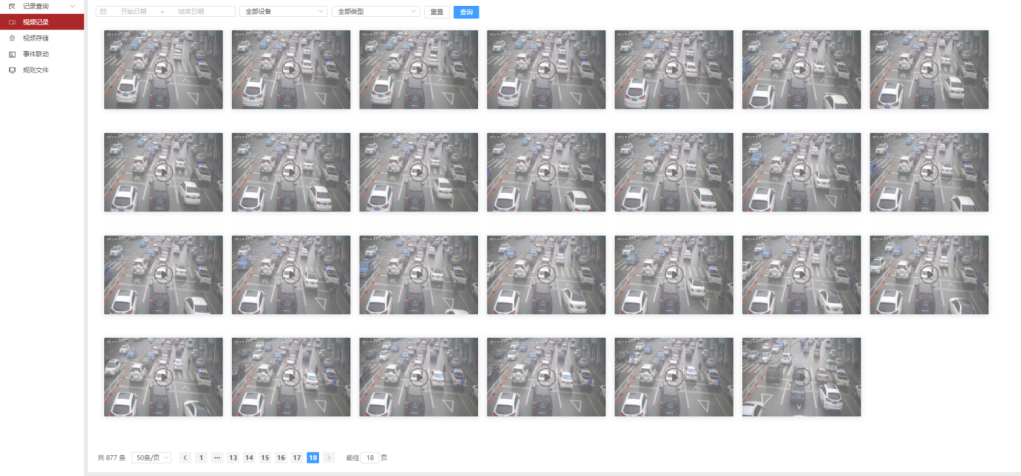
（2）谛格AI定义小站



图7 谛格AI定义小站

本产品搭载云从自研的安全监测相关算法，实现烟火监测、跌倒监测、安全帽监测、人员聚集、人员在离岗检测、周界入侵报警。该产品支持接入云从或第三方网络摄像机，实现视频的相关算法解析。此边缘设备产品适用于工地、生产车间、值班室、仓库等场景，实现相关作业、管理人员、场景的安全监测。





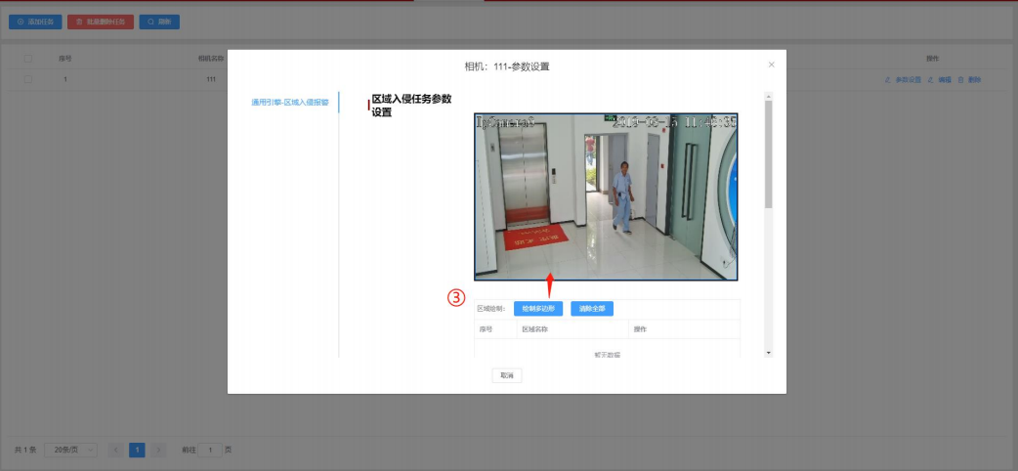


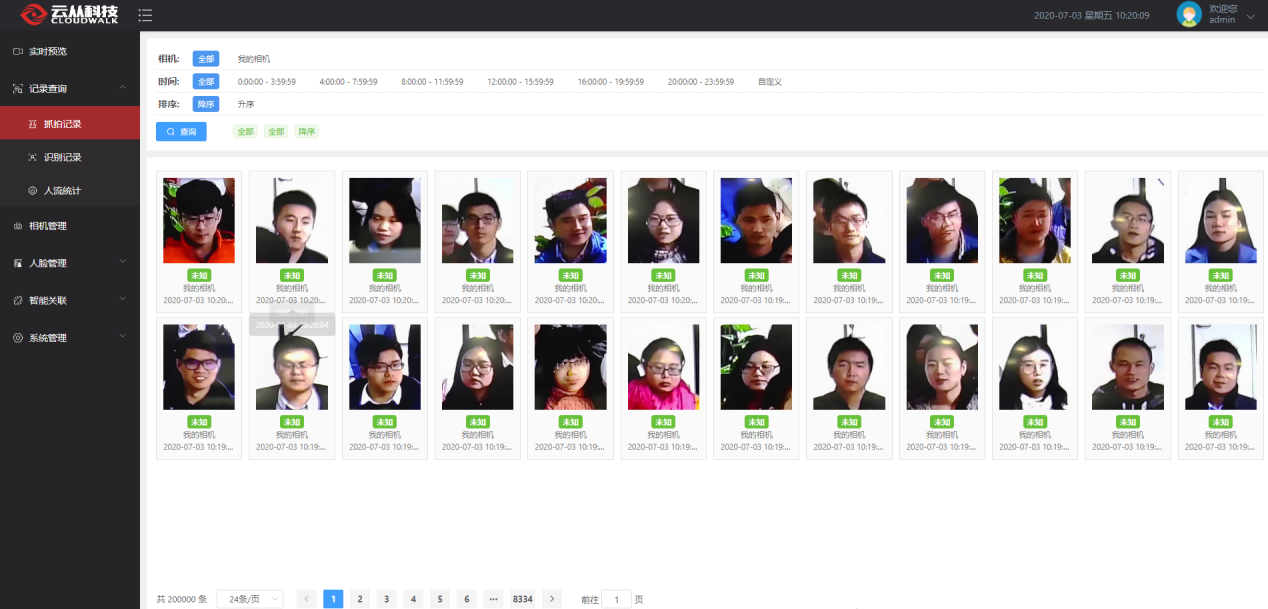
图8 应用界面

（3）扬帆AI智能盒



图9 扬帆AI智能盒

本产品搭载云从科技自有先进智能算法，针对边缘场景智能应用设计研发的嵌入式一体机设备。扬帆安全工地分析版本，通过接入8路网络摄像机RTSP视频流采集图像，自动识别现场施工人员是否佩戴安全帽等特征，判断人员是否按照要求进行安全施工，进行语音提醒和警告并对报警信息进行记录留存，为安全施工监督提供有力保障。适用于建筑工地、电力作业、路桥施工等对于安全监管有要求场景。



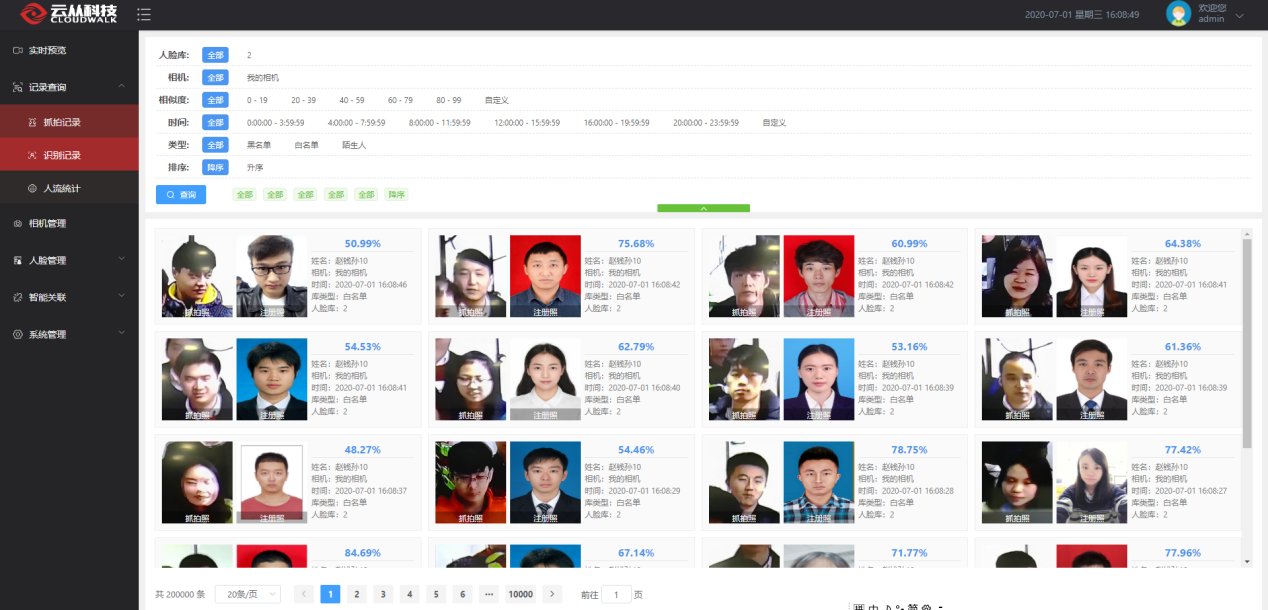




图10 应用界面

（4）新一代多模态大模型



图11 新一代多模态大模型

本产品基于大模型技术，预置大模型训练推理一体化平台和模型服务应用。采用全国产化芯片，支持对不同参数量的从容基础大模型以及开源基础大模型进行微调优化和推理部署验证。让大模型通过私域数据和知识经验学习，便捷快速地构建大模型应用生产一站式平台，并迅速将其应用到用户专属的业务场景中。针对客户进行量身定制了对话开发、任务编排、插件执行、知识接入和提示工程等五种定制优化模式，以及办公、代码、客服、运维、营销和采购等十多种即开即用的丰富场景包，以满足用户的个性化需求，帮助客户低成本且快速形成自用的大模型。



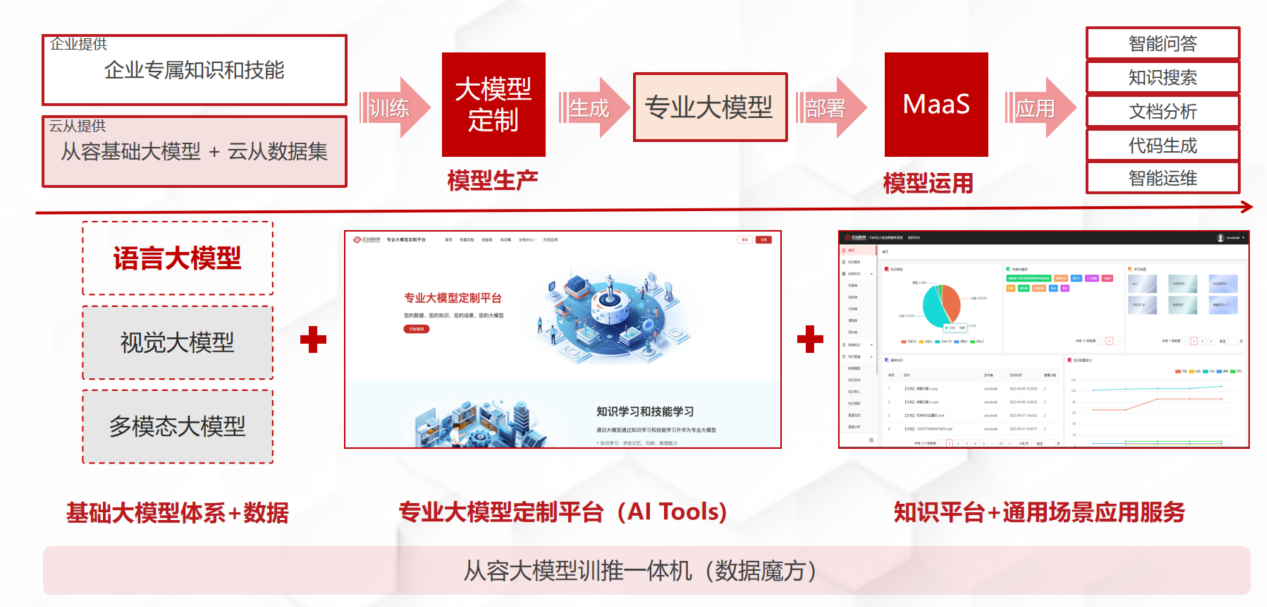
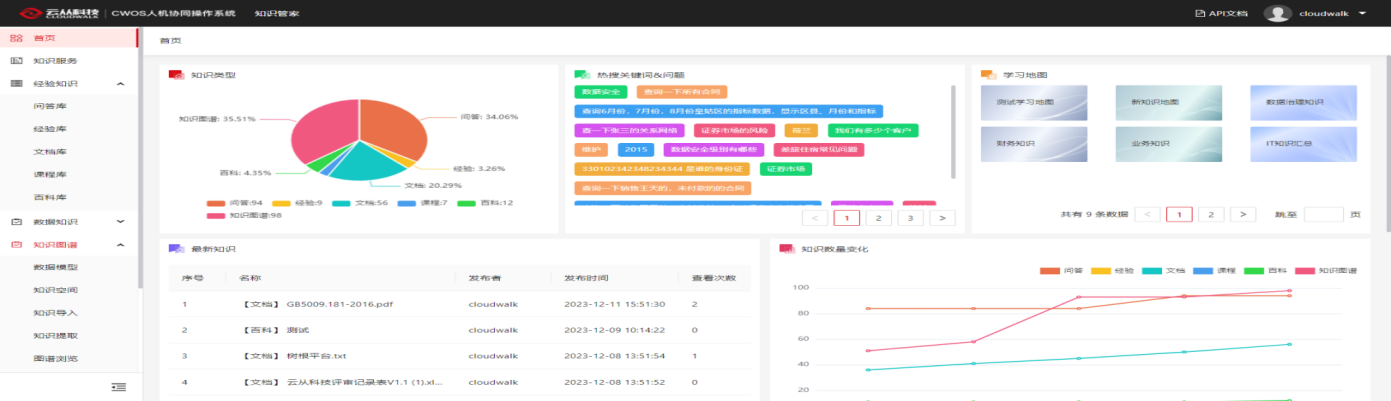


图12 应用界面

（三）项目实现的经济效益和社会效益情况

该技术与产品服务于各类智慧城市应用场景，包括天府数字城市项目，天府新区疫情防控系统，南沙智慧图书馆，大亚湾核电站视频分析平台，天津港港口大模型等项目，相关场景累计营收超过3000万。

该技术提高云边端技术和硬件的水平，答复降低终端运算负载和数据传输压力，降低能耗和硬性成本，达到了降本增效的目的，同时也降低了整体系统的碳排放总量，实现绿色节能，环保减碳。

大模型的推广使用，能大幅促进技术进步优化，科技创新，降本增效，解决企业数字化转型实际应用价值的最后一公里问题。

三、项目存在的问题及解决方案

大模型的自然语言和多模态处理领域取得了显著进展，但仍然存在一些关键问题

1. 资源消耗大：训练和运行大模型需要庞大的计算资源和能源，成本高昂且对环境不友好。
2. 泛化与常识理解局限：尽管模型参数量大，但在处理未见过的场景或应用人类基本常识时，其泛化能力和理解深度仍有待提高。
3. 数据偏见与公平性问题：模型可能从训练数据中继承并放大社会偏见，导致结果不公平，不符合伦理要求。
4. 可解释性低：大模型的决策过程复杂难解，缺乏透明度和可解释性，用户难以理解模型为何做出某种预测或决策。
5. 安全性与鲁棒性问题：大模型容易受到对抗样本攻击，即通过细微改动输入就能误导模型产生错误输出，显示出模型在安全性和鲁棒性上的缺陷。

四、企业未来发展思路

针对通用大模型现存的问题，未来的研究和发展可以从以下几个方向着手：

1. 模型优化与效率提升：研发更高效的训练方法和技术，比如知识蒸馏、模型剪枝、量化等，降低模型的计算和存储需求，同时保持或提高模型性能。
2. 增强模型泛化与常识推理能力：研究如何将更多形式的知识（如图谱、规则）融入模型，以及如何设计能够更好地模拟人类常识推理机制的新颖模型结构。
3. 解决偏见与公平性问题：开发有效的算法和工具监测、减轻甚至消除模型中的偏见，确保AI系统的公平性和包容性，例如引入公平性约束进行模型优化。
4. 提高模型可解释性：研究可解释AI技术，让模型不仅能给出准确的结果，还能提供清晰的决策依据和过程说明，增加用户对模型的信任度。

5、增强模型的安全性和鲁棒性：开发针对对抗样本的防御策略，提高模型在面对恶意攻击时的稳定性和准确性，同时研究模型的内在脆弱性，预防潜在的安全风险。

6. 丰富和改善训练数据：收集更加全面、多元且质量更高的训练数据，并探索如何在没有充足标注数据的情况下进行有效学习，例如自监督学习和半监督学习等方法。