

华为云AI视频

技术白皮书



目录 / CONTENTS

01 引言 / 01

02 AI 与视频的碰撞 / 05

2.1 发展趋势	05
2.2 典型场景	07

03 AI 视频介绍 / 09

3.1 参考架构	09
3.2 行业视频管理服务（IVM）	11
3.3 视频智能分析服务（VIAS）	13
3.4 盘古CV大模型	15
3.5 盘古视频解译大模型	21

04

AI 视频应用案例 / 23

4.1 华为门店	23
4.2 物流	25
4.3 铁路	26
4.4 矿山	28
4.5 电力	30

05

展望未来，从感知到生成 / 31



01 引言

用摄像机拍摄，记录并播放视频可以回溯到 19 世纪，大家公认的第一部电影是法国影片《工厂的大门》，1895 年由路易斯·卢米埃尔摄影。表现当时法国里昂卢米埃尔工厂放工时的情景，片长仅一分多钟。从这以后，摄影技术持续发展，从模拟技术到数字技术，从电影摄像机到家用摄像机，直到电脑，手机等便携终端内置摄像头，摄像已经成为当今社会人们记录信息并传播交流的最重要工具和手段，也是大众百姓所需要和掌握的一项基本生活技能。

摄像技术应用也从电影，广播电视发展到生活中的方方面面，包括城市治理、安全防护、工业质检等等。每个城市，每个企业都有大量的摄像机，不断在记录发生的一切。海量的视频数据，在方便大众的生活的同时，也带来了很多管理上的困扰。数据如何有效存储，如何能够感知并记录关键事件，如何能够将屏幕面前的工作人员解放出来或者减轻他们工作的强度，已经成为视频使用者最关心的问题。与此同时，AI 技术虽然起步较晚，但随着其快速的发展，已经在诸多方面与视频技术产生了深度的融合。



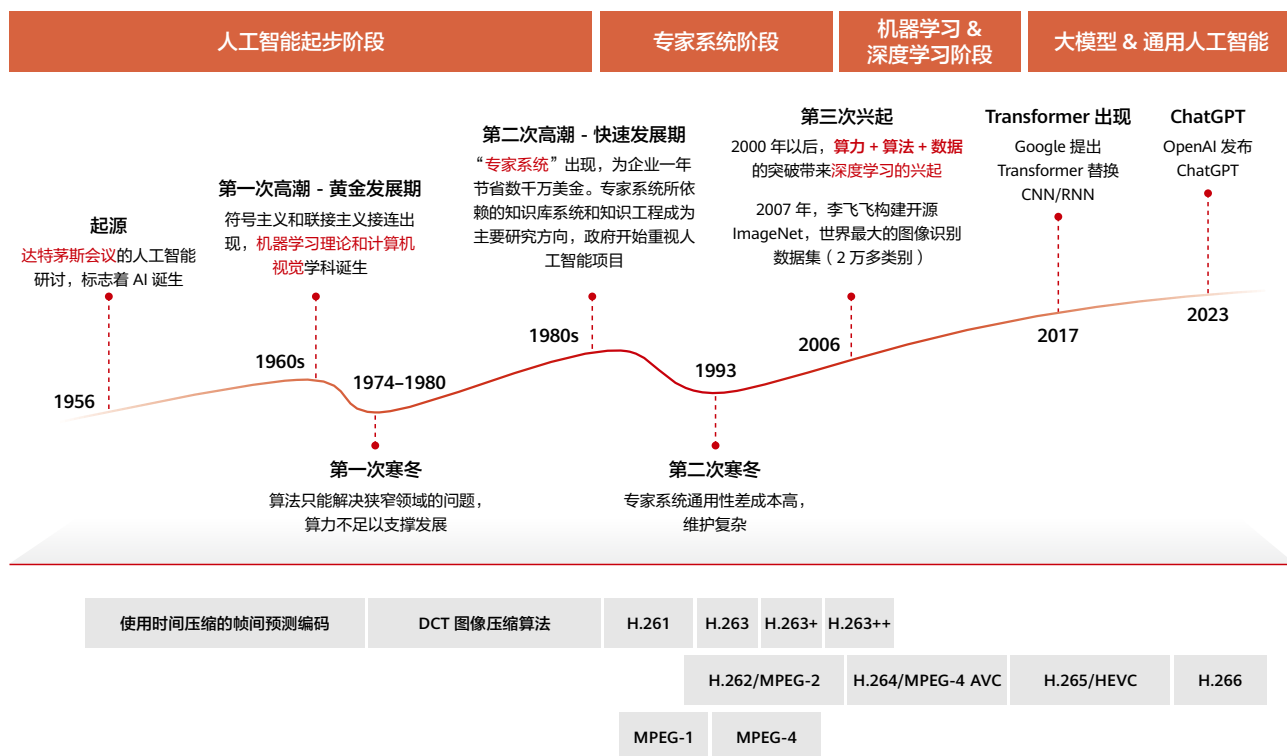
1950 年，艾伦·图灵（Alan Turing）在论文《计算机与智能（Computing Machinery and Intelligence）》中提出了著名的“图灵测试”，给出了判定机器是否有“智能”的试验方法。1956 年夏，麦卡锡、明斯基等科学家在美国达特茅斯学院开会研讨“如何用机器模拟人的智能”，首次提出“人工智能（Artificial Intelligence，简称 AI）”这一概念，标志着人工智能学科的诞生。

人工智能从诞生之初，就希望让机器理解这个世界，将人类从繁琐，重复性的事务中脱离出来。从符号主义、专家系统、神经网络、深度学习、强化学习到预训练大模型，一个又一个技术的突破，让人类看到了通用人工智能的希望。自然而然，用人工智能识别视频数据，甚至生成视频数据成为了技术路线的必然选择。华为公司在视频技术和人工智能技术上都有深厚的积累和丰富的实践，华为云 AI 视频产品正是将 AI 和视频技术相结合的优秀典范。此篇《华为云 AI 视频白皮书》，是我们团队研究和实践经验的总结，希望能够更好地促进产业的发展，让摄像机“看得懂、会说话”。



02 AI与视频的碰撞

没有孤立的技术，融合才能发展。视频技术和 AI 技术虽然起步时间不同，但在发展中却交相辉映，在最需要彼此的时候相遇。如下图所示，人工智能和视频编解码，都经历了漫长的摸索期，并先后于 21 世纪取得突破式发展。深度学习将人工智能带入千行百业的生产流程，H.264 编码技术将视频业务带入移动互联网，成为每个消费者日常的必需品。两条平行发展的技术踩着同样的步点，在视频数据爆发增长，预训练大模型横空出世的时代相遇。AI 和视频的结合是趋势和必然，给所有人，所有行业无限的想象空间和发展潜力。两个超万亿产值的行业相乘，将带来无法估量的产业价值。



图表 1 人工智能 VS 视频编解码发展历史

2.1 发展趋势



趋势 1：视频流云上集中管理

大量的摄像机安装完成后，面临的问题是如何集中式管理。摄像机分散在不同的地方，管理者需要在一个平台上，跨区域、大范围集中管理，通过完善的分权分域能力保护隐私安全。同时，各摄像机采集的视频需要集中存储，因为本地化分散存储会造成信息碎片化，无法形成多个视频流之间的联动，且本地存储易丢失、管理成本高。统一的云上存储，则可以有效解决以上问题。咨询报告指出，2023 年到 2027 年，视频流上云和云存储的年复合增长率超过 27%。在云化的趋势下，视频流云上管理、存储越来越成为业界趋势，企业的主流选择。

趋势 2：用预训练的大模型生成场景模型

AI 技术中，处理视频的相关技术一般被称为计算机视觉（Computing Vision）。计算机视觉是一种利用计算机和数学算法来模拟和自动化人类视觉的过程。它涉及到从数字图像或视频中提取信息，如对象识别、场景理解、运动跟踪、三维重建等。计算机视觉技术在许多领域都有应用，如自动驾驶、医学影像分析、机器人视觉等。

计算机视频分析视频流或者图片时采用计算机视觉模型。计算机视觉模型是指使用深度学习技术训练的神经网络模型，用于解决计算机视觉领域的各种问题。这些模型通常由数百万或更多个参数组成，可以对图像、视频等视觉数据进行高级别的理解和分析，例如图像分类、目标检测、语义分割、人脸识别等任务。



随着大数据和 AI 算力的发展，模型参数越来越大，大模型应运而生。大模型指网络规模巨大的深度学习模型，具体表现为模型的参数量规模较大，其规模通常在百亿以上级别。研究发现，模型的性能（指精度）通常与模型的参数规模息息相关。模型参数规模越大，模型的学习能力越强，最终的精度也将更高，泛化性也越强。

用大模型可以有效提升场景模型的准确率和泛化性，加上预训练的海量数据，用少量样本，甚至零样本就可以生成场景模型，解决视频算法长尾的问题。

趋势 3：用视频解译大模型理解视频内容

视频场景模型可以用确定的规则对视频流进行分析，识别关键事件，辅助人工进行判别并给出决策建议。但现实世界纷繁复杂，花鸟鱼虫，春夏秋冬，都在表达着自己的个性和不同，规则是无法穷尽的，判别式算法不断遇到新的需求和挑战。如何能够用泛化性强的模型理解视频，并通过自然语言的方式进行交互和报告，真正让人类从繁琐、重复性的事务中脱离出来是行业内普遍的需求。

视频解译大模型融合了视觉大模型、多模态大模型、自然语言大模型多种模型，可以实现对视频、图片、声音、文本多种模态组合分析，感知视频流发生的各种事件，实现让摄像机开口说话，实现真正的智能分析、智能交互、智能决策。

2.2 典型场景



城市日常管理

在城市治理场景中，往往建设有庞大复杂的城市事件类别体系，包含了繁多细碎的事项类别，如垃圾暴露、道路破损、围栏破损等等，一个城市一般有几百种事件类别。同时，不同城市可能还有不同的标准，可能某城市关注某一些特定事件类别，另一个城市又关注另一些特定事件类别。因此，城市政务场景面临着众多碎片化 AI 需求场景。城市事件的类别数量众多，同时绝大多数的城市事件又难以采集到大量数据来训练 AI 模型，这种问题我们称为“碎片化长尾需求场景”。

“碎片化长尾需求”一直是 AI 开发面临的难题，传统的 AI 开发模式需要对每种目标类别单独采集数据、训练模型，依赖专家经验进行算法参数调优，最后才能上线应用，每种算法的开发周期耗费几周至几个月，低下的效率难以满足当前高速的城市建设发展。华为 AI 视频方案，基于 AI 开发工作流，将数据标注、模型训练、部署上线等繁杂的流程固化为一个流水线的步骤，无需编写代码，任何人只要有准备数据，都可以通过流水线交互步骤快速地完成一个 AI 应用的开发和上线。每个 AI 算法的开发周期缩短至几天便可完成。同时基于预训练 CV 大模型的能力，依托于海量的大规模数据预训练，即便只有少量样本，也可以达到良好的模型泛化性和鲁棒性，解决碎片化长尾需求的问题。

由此可见，AI 视频方案中的预训练 CV 大模型 + AI 开发工作流，可以更好地契合城市治理的痛点需求，解决碎片化长尾需求场景的问题，更好地将 AI 落地到智慧城市的建设发展中。



城市应急处置

在城市建设过程中，除了事先设定好的事件类别，还经常有突发性的临时需求。比如突发暴雨，很多地方会临时地希望检测各地是否有积水内涝的情况，以便及时预防与救援；或者某地突发交通事故，相关部门也想快速地排查周边受影响交通拥堵的路段，以及时安排人力疏通车流。这些突发性的临时需求，可能根据天气、地点、时间等不同因素千变万化，这在城市政务的场景十分常见，也对于保障城市正常运转有着非常重要的作用。

然而，这种临时性的需求对于传统 AI 开发来说是灾难性的。传统的 AI 开发需要对每种待识别的事件采集数据、训练模型，而训练出来的模型也仅能解决这一特定的任务。当一个临时性的 AI 需求来临时，既往训练出来的模型肯定是无法适应这个新的任务的，那又要基于这个新的需求采集对应数据、训练模型，这一流程走下来即便有 AI 开发工作流支撑，少说也要几天的时间开发上线。但是临时性的需求往往是紧急的，比如对于积水内涝的场景，时间就是生命，业务往往要求算法立刻就能发挥作用、识别事件。因此传统的 AI 开发模式面对这种紧急的临时性需求就显得捉襟见肘了。华为 AI 视频方案，基于业界最新的多模态大模型技术，构建了开放式的目标检测和分割模型。该算法模型基于海量数据预训练的大模型，具备通用的特征提取能力，同时内嵌预言大模型，可以理解用户输入文本的语义信息。因此，该模型可以结合用户输入的任意文本信息，实现对应物体的检测，即便这个物体之前没有出现在模型的训练集里。这种特性非常符合城市治理里突发性的临时需求场景。比如面对积水内涝的场景，就不需要再针对积水事件重新训练一个模型，而是简单地输入一个类似“请问画面中是否有积水内涝？”的语句，算法通过图片和文本的语义理解，就可以识别出来画面中是否有积水内涝的事件了。这样一来，算法就不再局限于仅能识别特定范围的一些事件，应用的广度被无限地拉大，也能更好地满足城市政务场景中灵活变化的业务需求。

更详细的应用场景请参考章节“4 AI 视频应用案例”



03

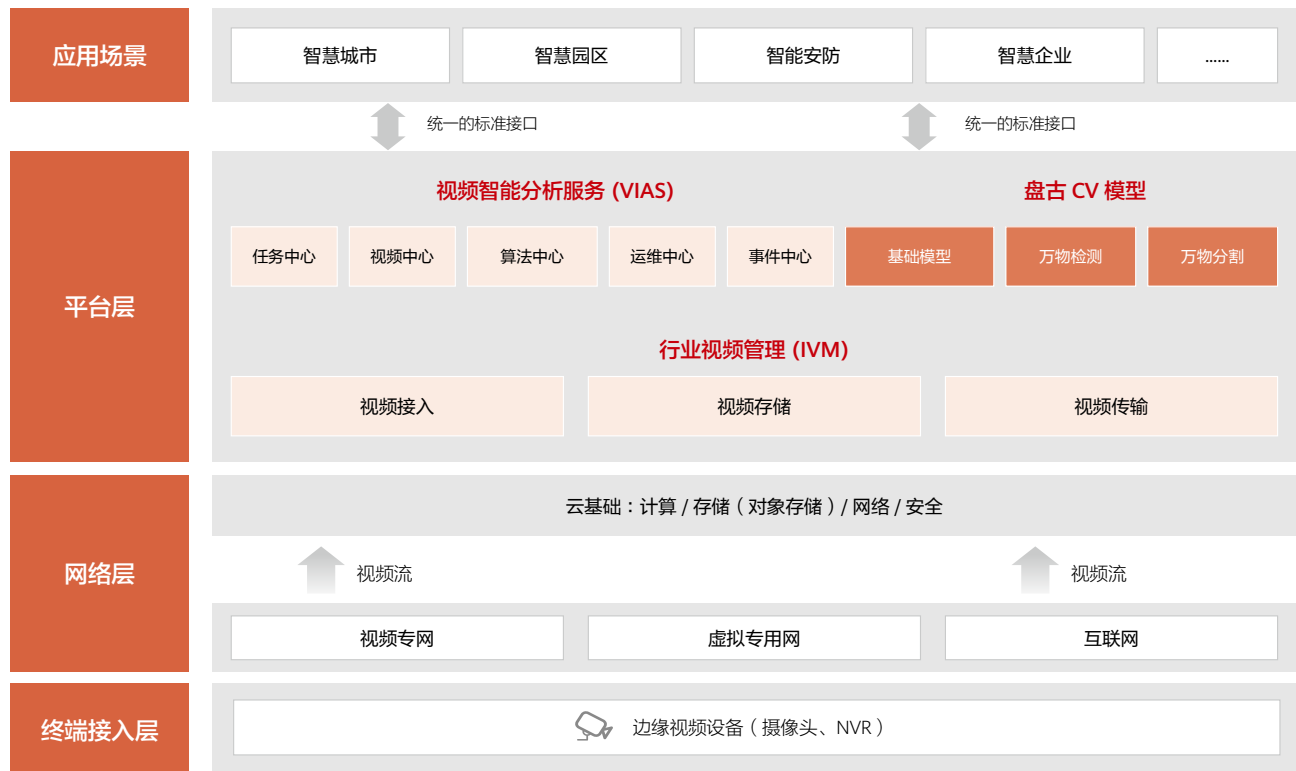
AI 视频介绍

3.1 参考架构

AI 视频服务依托联接、云、AI、计算等新一代 ICT 技术与知识创新融合，将感知、认知、决策、行动实现深度协同。其充分利用云计算能力，系统功能可靠、稳定、完整。平台设计秉承模块化、框架化、集群化、服务化的设计理念，提供电信级系统可靠性、可扩展性和可维护性，满足不同场景对接整合、兼容应用和可持续发展的需要。

AI 视频服务是面向行业视频场景的组合方案，参考华为公司架构，包括行业视频管理（IVM），智能视频分析服务（VIAS）和盘古 CV 大模型三个产品。可以提供一站式服务，将视频流从摄像机接入上来，进行调阅管理、存储管理，同时对视频流中的内容进行分析，识别关键事件，捕捉异常场景，并上报给上层应用系统进行处理，实现用人工智能的方式，用摄像机感知万物、掌控全局。

如下图所示，典型的组网分为接入层、网络层、平台层和应用层。AI 视频服务属于平台层网元，和接入层、网络层、应用层设备互联互动，相互配合，形成整体解决方案。



图表 2 AI 视频系统架构图

接入层主要设备为摄像机和 NVR，负责采集视频数据，并通过标准协议注册到平台层，被平台进行管理；NVR 可以实现对多个摄像机汇聚管理，并在本地存储视频。

网络层主要为网络设备，负责通过 IP 协议，将摄像机，NVR 等边缘设备接入到平台层，同时负责保障视频流传输的质量，包括但不限于丢包，抖动，时延等等。

平台层的 AI 视频服务，采用云化架构，支持公有云，混合云等多种模式。AI 视频属于 SaaS 服务，依赖云平台

虚拟机，OBS 存储，网络传输等能力，完成对摄像机设备信息，以及产生的视频流管理。同时基于人工智能技术，AI 视频提供对视频流的分析能力，理解视频流内容、关键事件，并将相应的结果推送给上层应用系统。

应用层负责面向行业，提供图形化页面和管理功能。不同行业有不同的应用系统，例如连锁门店客流管理系统、智慧安防系统、智慧园区管理系统等等。华为云 AI 视频服务和应用层是松耦合关系，采用消息接口对接。

3.2 行业视频管理服务（IVM）

3.2.1 业务需求

部署摄像机（SDC/IPC）等端侧设备后，首先需要集中管理功能，包括设备信息注册、远程配置、分权分域等功能。同时还要能够实现基于互联网任意时间、任意地点查看摄像机状态、视频流内容。同时，为了追溯、回溯，还要能够实现视频流存储，将视频流录制下来，长时间安全保存。

3.2.2 方案建议

行业视频管理服务（Industry Video Management Service）依托于华为云基础设施与音视频领域技术优势，为摄像机（SDC/IPC）、网络视频录像机（NVR）、智能视频存储（IVS1800）等华为及第三方设备，提供云端视频接入、视频传输及视频存储能力，适用于安全防范、生产管理、智慧运营等场景。行业视频管理服务可以帮助企业快速完成视频设备上云和智能化，助力企业数字化转型。

主要包括设备接入，视频调阅，录像管理等功能。

1. 设备接入

行业视频管理服务（IVM）支持国际标准协议接入摄像机，

也支持中国标准 GB/T28181 协议，同时还支持私有协议接入，通过私有协议或者 SDK，实现视频流解码显示能力。

2. 视频调阅

行业视频管理服务具备为公众及其他业务系统提供媒体流播放能力。媒体转码主要满足互联网 web/H5 技术和视频系统媒体流之间的转换适配需求，通过将码流转换为 RTMP、HTTP-FLV、HLS 等 PC 端可直接播放的视频流，为业务集成和开放提供快速的技术方案，同时提供基于视频技术的富媒体应用技术。

支持远程查看前端摄像机的实时视频，根据现场情况进行事件预判，实现视频实时浏览播放，实时播放时可显示视频相关信息，便于视频流的状态查询和故障诊断。

支持多布局能力，支持多个视频点位同时进行实况预览；单击摄像机开启视频按照从左到右、从上到下的顺序选择播放窗口；如果当前所有窗格已经用完，可手动增加另一个多窗格布局；系统客户端支持同时播放多个前端设备的实时视频。

支持显示当前实况摄像头的视频信息，其中视频信息包括：当前码率、平均码率、编码格式、分辨率等；支持声音控制，例如静音、取消静音；支持以拖动摄像头的



方式进行播放和停止；支持单画面停止播放，支持全部画面停止。

3. 录像管理

行业视频管理服务提供大容量的云端存储，通过互联网实时将前端数据传出至云端，依托于华为云 OBS 服务，为客户提供可靠的数据数据备份，帮助客户实现更长周期、更大容量、更高安全的云上数据管理。

用户可以在客户端上回放录像，也可以将系统录像文件下载到本地，支持使用通用播放器进行回放。用户可进

行事后录像的检索，通过录像可查看之前发生的事件现场视频，实现事后取证功能。同时支持查询平台录像、前端录像；支持自定义时间范围进行录像查询；支持录像查询结果以进度条方式展现，进度条可以前后拖动，支持精度缩放等功能。

3.2.3 小结

行业视频管理服务（IVM）基于华为公有云，提供摄像机设备管理、接入、调阅、存储等服务。主要功能服务方式如下，供项目参考。

产品组合	商品	量纲	应用场景
行业视频管理服务	视频接入	路 / 年	公有云必选
	调阅带宽	Mbps/ 年	公有云必选
	视图云存储	GB/ 年	云存储、云备份、告警录像

图表 3 行业视频管理 (IVM) 方案建议

3.3 视频智能分析服务（VIAS）

3.3.1 业务需求

完成摄像机和行业视频管理平台建设后，实现了视频流集中管理、集中存储。如果仅仅依靠人工监看的方式，必然消耗大量人力，识别准确率依赖人员技能。如何实现视频流的自动分析、准确识别关事件主动上报成为普遍的业务需求。基于人工智能的视频分析服务，要能够为上层的行业应用提供 AI 能力，包括但不限于：

- » 丰富的视频分析算法，满足复杂场景分析需求；
- » 建设视频统一分析平台，集中管理，充分盘活视频资源；
- » 算法统一管理，算法和算力解耦，多厂家算法共享算法仓，算力统一调度。

3.3.2 方案建议

视频智能分析服务（VIAS）是集成视频 AI 分析、事件感知等能力的一体化平台，实现智慧园区、城市治理、安全生产等场景的事件感知、分析和决策能力，助力业务闭环。视频智能分析服务提供丰富的“开箱即用”的算法模型，包括城市治理、公共安全、连锁门店、智慧物流、智慧园区等等，帮助千行百业快速使用成熟的人工智能技术，提效降本。



主要包括分析服务、算法中心、视频中心、任务中心、事件中心等功能。

1. 分析服务

视频分析服务是承载视频 AI 算法的弹性计算引擎，提供视频数据接入、分析及告警输出的能力，可通过 API 支撑业务开发应用，同时能够帮助 AI 开发人员提升视频 AI 集成效率，助力其核心业务价值开发。

视频分析能力主要基于如下技术构建：

1) 物体检测技术

物体检测是视觉感知的第一步，也是计算机视觉的一个重要分支。物体检测的目标，就是用框去标出物体的位置，并给出物体的类别。在当前视频分析服务构建的能力中，人或者车的检测是第一步，也是最关键的一步。人与车目标检测的准确率也会直接影响后续算法的效果，但由于目标环境的多样性复杂性，对于物体的检测，通常会受到不同环境的干扰。所以为了提高算法的准确率，通常会针对实际的应用场景进行定制化的训练，以此排除复杂的环境带来的干扰。

2) 图像分类技术

一张图像中是否包含某种物体，对图像进行特征描述是物体分类的主要研究内容。一般说来，物体分类算法通过手工特征或者特征学习方法对整个图像进行全局描述，然后使用分类器判断是否存在某类物体。图像分类的研究，通常衍生出来对特定目标物体进行检测的能力，比如识别大货车、公交车等特定的目标。

3) 物体定位技术

如果说图像识别解决的是 what，那么物体定位解决的则是 where 的问题。利用计算视觉技术找到图像中某一目标物体在图像中的位置，即定位。对物体的定位，通常能衍生出很广的应用场景。比如在安防领域，判断目标

物体的位置，可以进行入侵检测、徘徊检测以及过线计数等等算法。

基于如上技术，视频分析服务可提供面向智慧园区、水利、交通、应急管理场景的视频 AI 分析能力，不但能保证自研 AI 算法的接入，还能保证第三方算法和行业共享算法的对接，最终实现 AI 能力的稳步提升。

华为视频分析算法，基于 100+ 项目实践经验持续积累、优化，已沉淀形成多种类型的算法能力。

2. 算法中心

算法中心提供多厂商、多框架、多功能的统一管理能力，支持用户将导入的算法镜像进行统一管理，支持算法版本的全生命周期管理，为后续算法部署提供基础管理能力。用户可在该模块查看已上线的算法能力，同时为三方开发者提供账号体系，开发者可在该模块发布新算法、更新算法版本。算法中心可跳转算法商城，算法商城展示了可上线的算法能力清单，可根据用户业务需求上线。

3. 视频中心

视频中心提供视频源数据接入管理能力，是算法的前置输入模块，通过该模块的配置，任务中心即可选择输入源，

实现视频算法的整体功能性配置，构建基于视频数据的智能分析应用。视频中心支持视频源管理，视频质量巡检，摄像机分组管理等功能。

4. 任务中心

任务中心提供算法作业配置、算法作业管理能力，是算法的核心配置模块，通过该模块的配置，算法即可具备分析功能。任务中心支持作业配置、作业管理、批量配置、公共模板、定时任务等功能。

5. 事件中心

事件中心提供事件统一管理，是算法的分析结果输出模块，委办单位可通过该模块查看视频分析的事件结果，同时支持将事件分析结果上报到现网业务系统，及时发现事件并形成工单分派，提升网格处置效率。事件中心支持事件管理，事件重复聚合，事件审核，事件订阅，运营报告生成等功能。

3.3.3 小结

视频智能分析服务（VIAS）基于华为公有云，提供视频算法分析服务、算法管理、算力管理、任务管理、事件管理等。主要功能服务方式如下，供项目参考。



图表 4 视频智能分析服务方案建议

3.4 盘古 CV 大模型

3.4.1 业务需求

随着工业生产越来越强调智能化，大量传统行业开始积累领域数据，并寻求人工智能算法以解决生产和研发过程中遇到的重复而冗杂的问题。这意味着，人工智能算法在落地的过程中，将会面对大量不同场景、不同需求的用户。这对算法的通用性提出了很高的要求。然而我们注意到，当前业界大部分人工智能开发者，正在沿用传统的“小作坊模式”，即针对每个场景，独立地完成模型选择、数据处理、模型优化、模型迭代等一系列开发环节。由于无法积累通用知识，同时不同领域的调试方法有所不同，这样的开发模式往往比较低效。特别地，当前人工智能领域存在大量专业水平不高的开发者，他们往往不能掌握规范的开发模式和高效的调优技巧，从而使得模型的精度、性能、可扩展性等指标都不能达到令人满意的水平。我们将上述问题，称为人工智能算法落地的碎片化困境。

因此如何能够类似流水线的方式，用少量样本，快速生成场景化模型，成为行业的迫切需求。

3.4.2 方案建议

华为盘古 CV 大模型瞄准人工智能在工业场景应用中的困境，创造性提出用经过海量数据预训练的视觉大模型作为训练 workflow，用类似工业流水线的方式快速生成场景化模型。盘古 CV 大模型收集大量图像数据，以及图像和文本对比数据，利用无监督或者自监督学习方法将数据中蕴含的知识提取出来，存储在具有大量参数的神经网络模型中。遇到特定任务时，只要调用一个通用的流程，就能够将这些知识释放出来，并且与行业经验结合，解决实际问题。



图表 5 盘古 CV 大模型 workflow 原理



图表 6 盘古 CV 大模型应用场景和优势

对于常见的视觉处理任务，盘古 CV 大模型通过自动化模型抽取、参数自动化调优等模块实现场景模型的训练和推理。盘古 CV 大模型包括物体检测、姿态估计、视频分类、图像分类、异常检测、目标跟踪、语义分割、实例分割等多条预训练工作流，可以全面覆盖场景模型训练需求，并在矿山、钢铁、铁路、交通等多个行业进行验证和实践，成为行业首选。

由于盘古 CV 大模型配套完善的工程套件，可以基于图形化界面，零代码前提下，实现数据标注、模型开发、推理部署，实现 AI 落地零门槛。购买盘古 CV 大模型的企业，实现人工智能转型，构建“内生的，持续发展”的 AI 能力。

小样本，结合数据检索及数据增广技术，相对传统训练方式，数据需求减少 80% 以上；

高精度，受益于更好的语义对齐效果，在小样本学习上表现优异，显著超越对比方法；

高效率，利用行业模型高效表征及数据筛选能力，数据处理效率提升 5 倍以上；

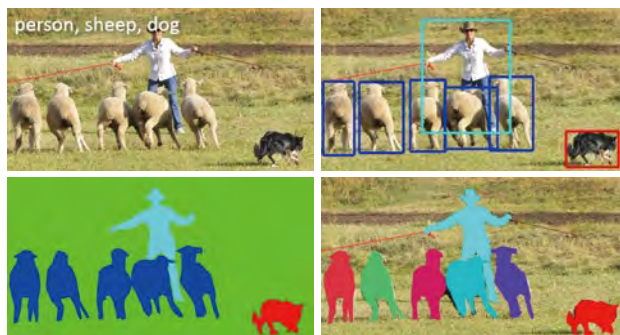
由于盘古 CV 大模型配套完善的工程套件，可以基于图形化界面，零代码前提下，实现数据标注、模型开发、推理部署、实现 AI 落地零门槛。购买盘古 CV 大模型的企业，实现人工智能转型，构建“内生的持续发展”的 AI 能力。

技术原理

计算机视觉的主要目标，是设计出能够识别视觉信号，并且对其进行各种处理和分析的程序。换句话说，计算机视觉是研究计算机如何去“看”的学科。其中，较为典型的任务包括图像分类、物体检测、物体分割、物体追踪、姿态估计等。下图展示了图像分类中最著名的 ImageNet 数据集（超过 2 万个物体类别）和 MS-COCO 数据集（包括检测、分割等多种任务）。



The ImageNet dataset
~15M images, ~21K categories, ~1.5TB



The MS-COCO dataset
detection, segmentation, pose estimation, etc.

图表 7 测试数据集

在计算机中，视觉信号一般以“密集采样强度”的方式存储：不同方向入射的光线在每个信道（如红绿蓝）上的强度被记录下来，用于呈现图像的基本内容。图像中的每个基本单元被称为像素——很显然，这些像素并不能代表基本的语义信息，因而图像的基本存储形态和人类能够理解的语义之间，存在很大的差距。在学界，这种差距被称为“语义鸿沟”，这也是几乎所有计算机视觉研究所需要处理的核心问题。

进一步探究图像的存储形态，我们会发现图像信号的若干特点：



内容较复杂

图像信号的基本单位是像素，但是单个像素往往不能表达语义。图像识别的任务，就是构建特定函数，使得像素级输入能够产生语义级输出。这种函数往往非常复杂，很难通过手工方式定义。



信息密度低

图像信号能够忠实地反映事物的客观表征；然而其中相当部分的数据被用于表达图像中的低频区域（如天空）或者无明确语义的高频（如随机噪声）区域。这就导致了图像信号的有效信息密度较低，特别是相比于文本信号而言。



域丰富多变

图像信号受到域的影响较大，而且这种影响通常具有全局性质，难以和语义区分开来。例如，同样的语义内容，在强度不同的光照下，就会体现出截然不同的表征。同时，相同的物体能够以不同的大小、视角、姿态出现，从而在像素上产生巨大差异，为视觉识别算法带来困难。



鉴于上述特点，基于深度神经网络的预训练大模型就成为了计算机视觉落地的最佳方案之一。预训练过程能够一定程度上完成视觉信号的压缩，神经网络能够抽取层次化的视觉特征，而预训练结合微调的范式则能够应对丰富多变的域。

数据收集

图像是一种复杂的非结构化数据，包含丰富的语义信息。现如今，还没有任何一种方法能够对图像数据的数学规律进行准确的描述，因而人们只能通过收集大量的数据，来近似现实中图像数据的分布。2009 年出现的 ImageNet 数据集是计算机视觉领域的重要里程碑，它使

得训练、评估大规模图像处理方法成为可能。随着计算机视觉技术的进步和更多应用的出现，ImageNet 数据集的局限性逐渐显现出来，包括规模、复杂性等。

为了解决这一问题，我们必须收集更大规模、更加复杂的图像数据，而这也是业界的一致趋势。

通过多种渠道收集图像数据，包括但不限于公共数据集下载、自有数据集扩充、各搜索引擎关键字爬取、以图搜图、视频图像抽帧等。从这些原始数据中，我们筛除了低分辨率、低曝、过曝、简单背景等低质量图像数据，再通过已有预训练视觉模型进行重复图像的判断和去除，最终保留超过 10 亿张高质量图像数据，占据约 40TB 空间。



10 亿 +
图像数据



~40 TB
存储空间

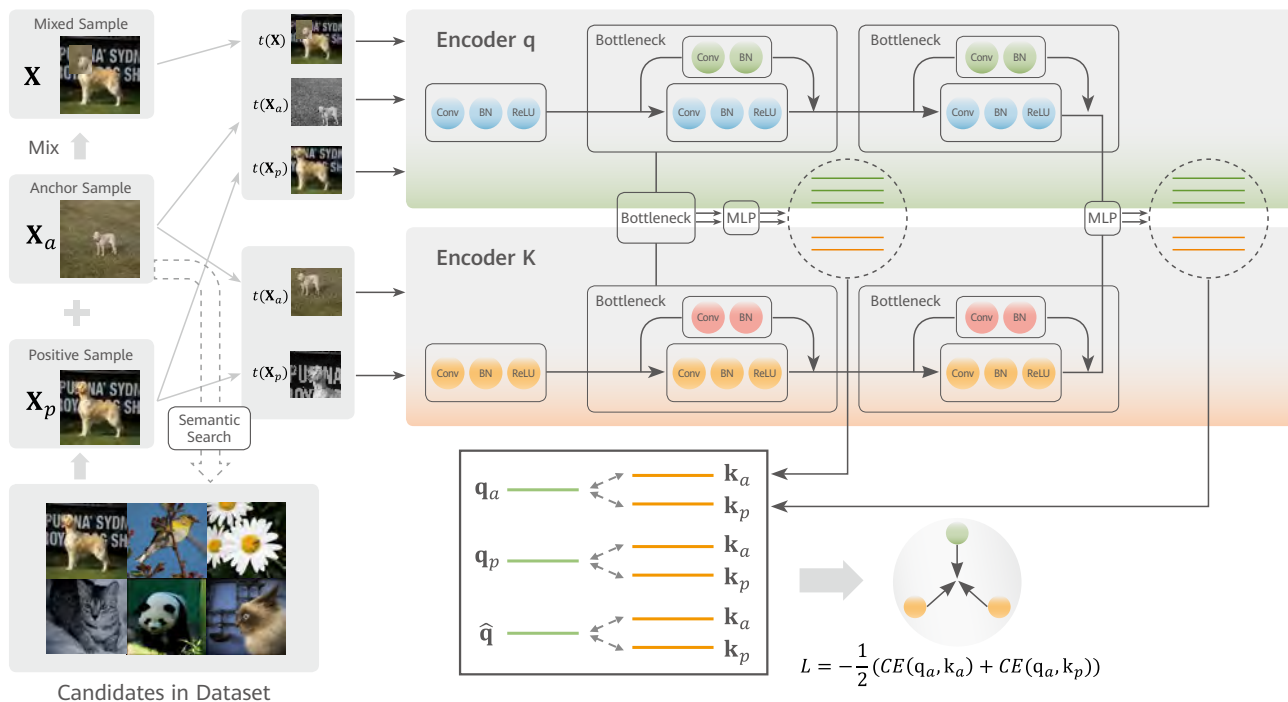


覆盖
自动驾驶，电力，
铁路，遥感等

预训练方法

我们使用的神经网络模型覆盖了计算机视觉领域最为常见的卷积网络和 transformer 架构，两者既可以分开使用，也可以按需结合以达到更好的效果。利用自动机器学习算法，能够支持并调用不同大小的神经网络，其中最大的计算模型具有接近 30 亿参数，最小的模型只有数十万参数，其大小相差超过 1000 倍，为适配不同的视觉任务提供了可能性。

我们收集的训练数据大部分来自互联网，不仅具有相当程度的噪声，而且不包含准确的语义标签。为了充分利用这些数据，我们设计了自监督学习方法，即通过某种或者某几种代理任务，教会模型如何理解视觉数据，在没有语义标签的情况下也能拟合复杂的数据分布。与此同时，我们在对比学习的基础上优化了相关代理算法，业界首创在对比度自监督学习中引入等级化语义相似度，即挑选那些距离相应聚类中心更近的最近邻作为正样本，并且在拉近语义相似样本的时候引入了混合样本增强，以减少样本选取过程中的噪声影响。在此基础上，我们拓展自监督学习算法中正样本的数目，使得正样本集合能够更加高效地被聚集，同时避免受到大量负样本优化的影响。我们采用的预训练算法（发表于 TPAMI 上）的简略示意图如下所示：



（注：基于等级化语义聚集的对比度自监督学习）

效果展示

盘古视觉大模型在 ImageNet 数据集的线性分类评估上，首次达到了与全监督相比拟的结果。

同时，受益于更好的语义对齐效果，我们的方法在小样本学习上表现优异：使用 ImageNet 上 1% 和 10% 的标签训练，我们的方法达到了 66.7% 和 75.1% 的分类精度，均显著超

越对比方法。以此方法为基础，我们设计了具有 10 亿参数量基础模型，并在超过 10 亿张无标注图像组成的数据集上进行预训练。所得到的模型，在 ImageNet 上达到了 88.7% 的分类精度，而 1% 标签的半监督学习精度也达到 83.0%。同时，盘古大模型在超过 20 项下游任务上进行了测试，展现出良好的泛化能力，如下表所示。

	数据集	业界最佳模型	盘古预训练模型
1	Aircraft (飞行器)	90.43	89.32
2	CUB-200-2011 (鸟类)	86.90	91.80
3	DTD (纹理)	80.05	85.00
4	EuroSAT (卫星图块)	98.85	98.98
5	Flowers102 (花)	97.07	99.69
6	Food101 (食物)	92.21	94.58
7	Pets (动物)	95.29	95.91
8	SUN397 (场景)	71.51	78.92
9	StanfordCars (车)	92.48	94.09
10	StanfordDogs (狗)	87.41	91.28
11	Average	89.22	91.96

图表 8 盘古预训练模型分类性能比较列表

	数据集	业界最佳模型	盘古预训练模型
1	VOC (自然场景)	72.2	76.6
2	Comic (风格变换)	35.6	38.0
3	Clipart (风格变换)	57.5	61.0
4	Watercolor (风格变换)	34.4	36.9
5	DeepLesion (医疗)	36.7	38.1
6	Dota 2.0 (遥感)	21.2	21.0
7	Kitti (自动驾驶)	29.6	32.9
8	Wider Face (人脸)	35.3	36.3
9	LISA (红绿灯)	43.5	42.7
10	Kitchen (厨房场景)	53.6	55.0
	average	41.96	43.85

图表 9 盘古预训练模型检测性能比较列表



3.5 盘古视频解译大模型

3.5.1 业务需求

在特定场景分析基础上，开放式场景分析和识别需求越来越强烈，尤其是针对应急事件的处理。包括但不限于以下需求：

智能视频检索，通过自然语言对摄像机，或者视频存储进行开放式检索，如检索发生在特定时间，地点的特殊事件；检索多个线索关联的场景等等；

视觉标签库，通过对视觉数据进行标签化处理，可以对所有视频流的标签进行精细化管理，提升全域摄像机标

签数据的准确性和实用性，同时还可以动态刷新，确保数据的实时性和有效性；

关键帧定位，借助视频向量化能力，可以实现对关心事件检索时，可以定位到摄像头关键帧，并对关键帧前后视频直接查看，提升问题定位的效率；

智能视频摘要，借助大语言模型的能力，汇总摄像机关键标签，摘要文本数据，生成一句话摘要或分析报告，将摄像机所拍摄的关键内容报告给管理者，实现让摄像机说话。



3.5.2 方案建议

盘古视频解译大模型，是在视频智能分析服务（VIAS）和盘古 CV 大模型基础上，融合多模态大模型能力，进一步延伸人工智能在视频领域的应用。此方案依托大模型的万物理解能力，实现视频检索、视频标签、以及视频摘要能力。将摄像机拍摄的画面描述出来，实现让摄像机开口说话。

如上图所示，此方案主要包括四个主要部分。中间核心为“盘古大模型重构 AI 视频服务交互”，CV 大模型 + 多模态大模型双轮驱动，开放场景视觉分析，快速覆盖

数千个场景，并兼容专家模型支持专属场景准确识别。盘古大模型基础上，用“Agent 驱动视觉感知”，作为视觉感知能力入口，通过可编排可组合可插拔特性，实现大语言模型对视觉感知能力的驱动。

最后向上，可以覆盖海量“场景应用”从视觉 + 文本协调应用出发，牵引视觉感知能力“可看”向“可交互”转变；向下重构“摄像头 & 标签资源”，构建分层分类视觉标签体系、动态标注，并实现视频存储资源的精细治理，释放视频数据资源价值。



图表 10 视频解译大模型架构图



04

AI 视频应用案例

4.1 华为门店

秉承“自己的降落伞自己先跳”的原则，华为率先将行业视频管理服务（IVM）应用于华为门店管理。华为终端 BG 有超过 1 万家门店，每个门店都有若干摄像机，对门店进行管理。为了管理高效，华为终端 BG 需要一套集中式管理系统，实现对超过 10 万路摄像机统一管理，统一调阅和统一存储。

客户需求 and 痛点



- 摄像机统一管理
- 视频数据安全
- 全国所有门店统一监管
- 外墙广告
- 清洁墙壁



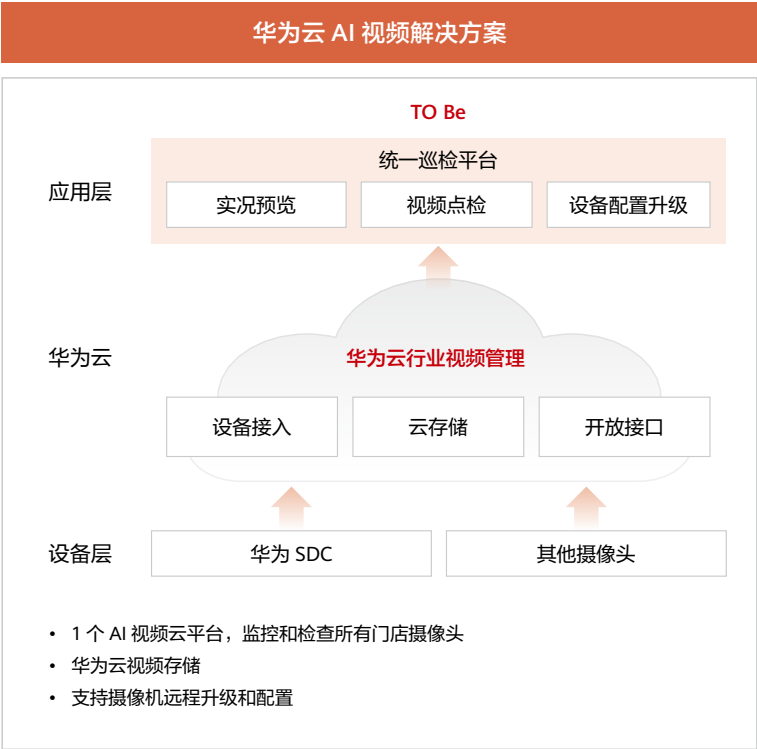
- 员工着装和行为
- 店内广告
- 商品和样品放置
- 店内地板清洁
- 店内橱窗品牌推广

图表 11 华为门店关键需求



行业视频管理服务（IVM）完美满足以上需求，提供云化管理平台，实现视频接入、视频调阅和视频存储，一个管理中心可以管理分布在全国的所有华为门店。

在华为门店项目中，IVM 实现接入多厂家摄像机，充分保护前期摄像机建设成本，门店不需要任何物理改造，通过软件适配实现统一管理。IVM利用云存储可靠性高，永不丢失等能力，确保门店关键事件被记录、可回溯；同时支持视频水印，视频加密，端到端可回溯等能力，确保视频传输和存储的安全，并不被盗取。同时 IVM 还提供完善的分权分域管理机制，隔离多级管理者，确保顾客隐私和数据安全。行业视频管理服务已经成为华为门店管理环节中不可或缺的组成部分，融入到华为终端销售的管理体系中，为华为终端业务增长保驾护航。



图表 12 IVM 华为门店解决方案

4.2 物流

伴随电子商务的蓬勃发展，物流是近些年发展快速的行业，无论是营业额还是覆盖地区的数量都在快速增长。物流行业属于劳动力密集型行业，有大量分支机构、仓库，一般都采用摄像机方式进行远程管理，确保安全，有序传输，既保证效率，又保证客户端满意度。因此物流行业普遍存在以下需求：

- » 上千个分支机构，超万路摄像机的集中管理
- » 摄像机产生的视频数据，需要采用高安全的手段进行存储
- » 物品传递过程中，要最大程度避免暴力分拣、错误配送等问题，亟需人工智能的方式进行监管，改善服务质量

华为云提供行业视频管理服务（IVM）和视频智能分析服务（VIAS），满足以上需求。IVM 基于华为公有云提供摄像机管理、视频流传输和存储功能。确保物流公司管理者在任何地方，都可以远程查看视频画面。VIAS 提供 AI 分析算法，包括暴力分拣、吸烟检测等，实时识别不符合工作规范要求的行为并上报，对工作质量进行监督，极大提升了物流行业的工作规范性和服务质量。



图表 13 物流行业视频接入分析系统架构图

4.3 铁路

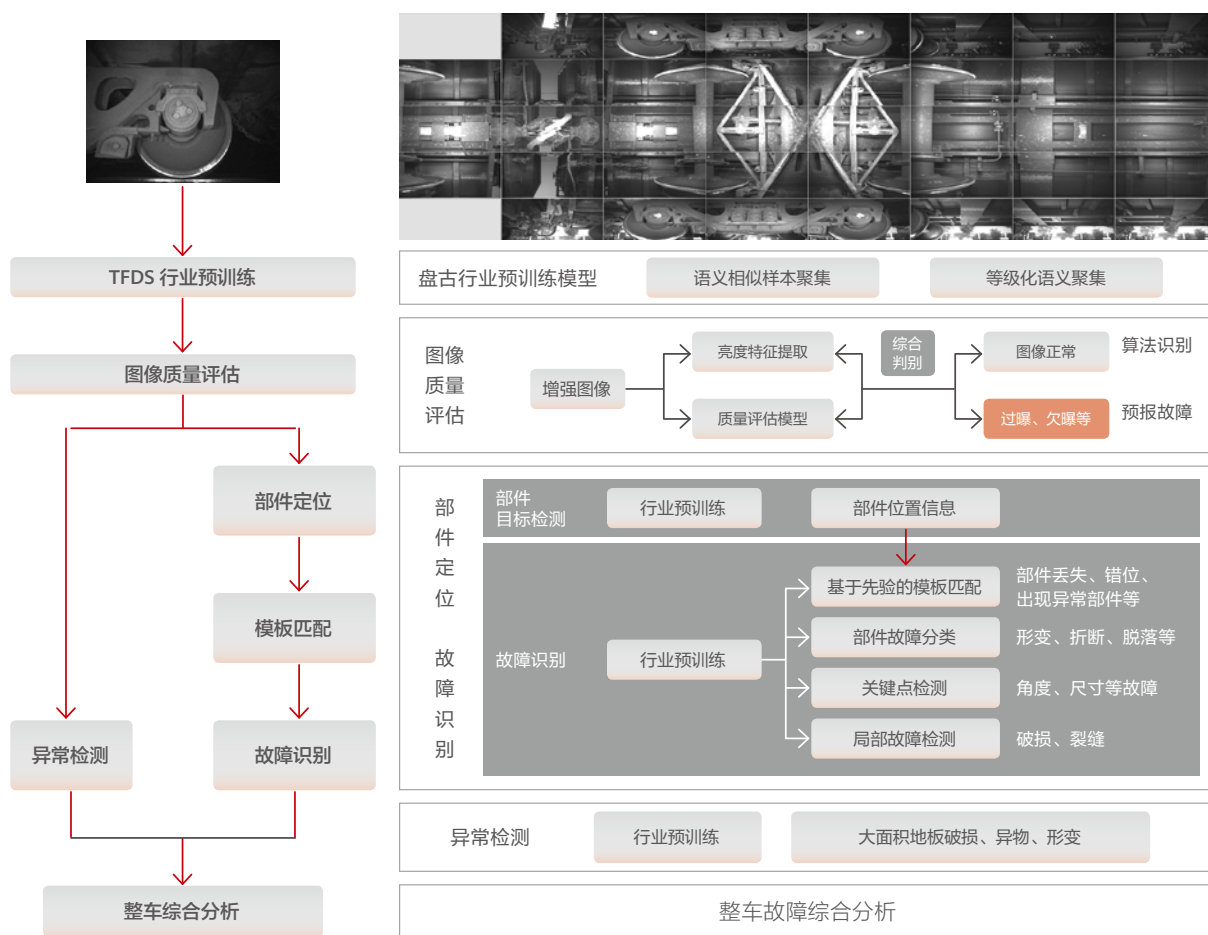
中国漫长的铁路线上运行着近百万辆铁路货车，安全运行的重要性不言而喻。当前，铁路系统广泛采用 TFDS（Trouble of moving Freight car Detection System，货车运行故障动态图像检测系统）来检测列车安全，简言之，就是利用部署在铁轨旁的高速相机拍摄通过 TFDS 探测站的列车部件图像，再由列检员对这些图像逐一分析，识别车辆故障隐患并预警处置。

受制于技术发展，TFDS 过去大多采用人工方式进行故障识别。以郑州北车辆段 5T 检测车间为例，日均检车 4 万余辆，识别图片 280 万余张。列检员每天需要检查大量极其相似的图片，并且需要在 5 秒左右的时间及时发现细微的差别，找出列车存在的故障。人工方式识别劳动强度大、人力成本高，高强度的重复劳动也极易产生疲劳，造成误判。

2021 年，国铁集团货车事业部把 TFDS 故障图像智能识别项目作为国铁集团第一批科研计划“揭榜挂帅”课题，指定郑州局集团公司郑州北车辆段 5T 检测车间作为该项目的试点单位，与华为公司、慧铁科技公司共同研究、联手推进。

在项目中，盘古大模型充当了 TFDS 系统的“AI 训练师”，它能够基于海量无标注数据进行预训练，还可以“边用边学”；具备小样本学习和样本生成能力，能够生成大量训练样本提升模型质量。比如摇枕心盘脱出的故障，全国范围内只找到一张故障样本，借助小样本学习，目前盘古大模型已经能正确识别这个故障。

下图展示了盘古视觉大模型在为 TFDS 定制的方案。依托于盘古行业预训练大模型，定制化地开发了整体解



图表 14 基于盘古行业预训练模型的铁路 TFDS 开发方案

决方案，包括车型筛选、工位分类、配件筛选、图像质量评估、已与车型先验的模板匹配、多车级联分析等模块，其中盘古大模型核心解决方案包含以下组成部分：



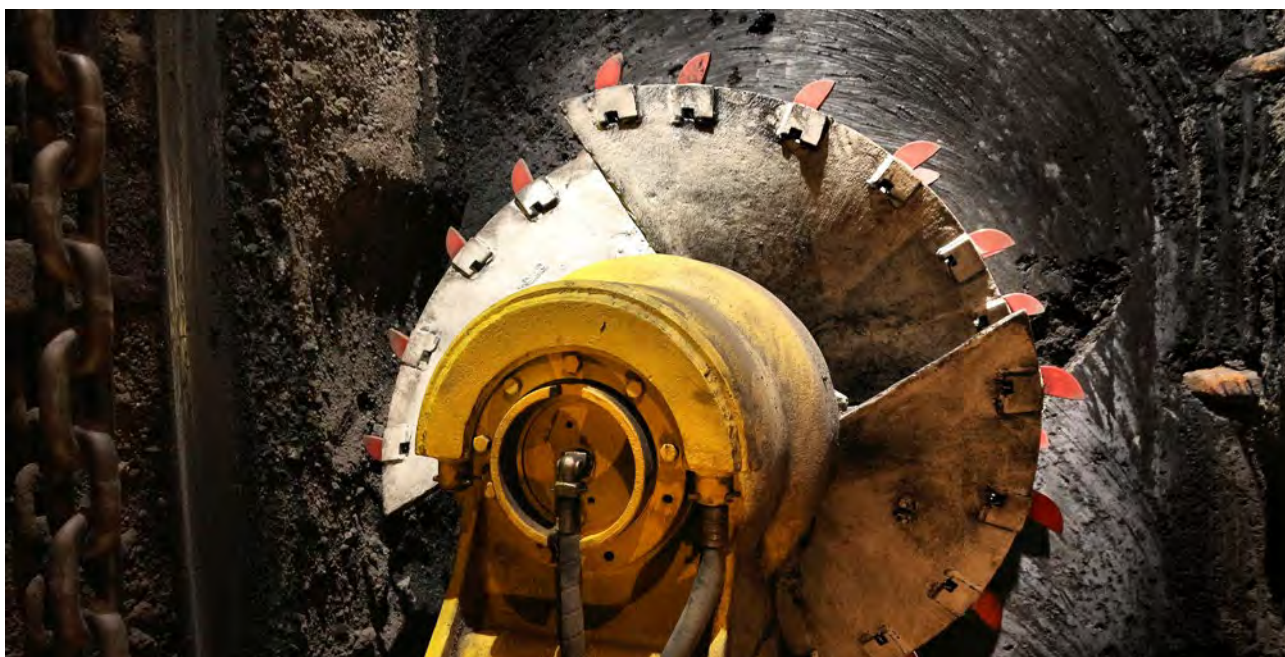
基于盘古视觉大模型的整体解决方案，在 5T 检测车间集中作业分析的 14 条线路进行了验证。由 5T 检测车间动态检车员预报并经组长确认为提报故障的数据样本（故障图片）共计 32007 张。在测试环境下，这些故障图片与大量正常图片混合，送入盘古大模型进行判断。如下表所示，实测结果表明，当前盘古大模型的识别精度已经超过人类检测员水平。



2022 年 12 月，华为云盘古大模型加持的 TFDS 系统正式投入使用，可根据大量的数据样本，自动总结部件特征、自动寻找故障规律，并在实际试用中持续改善分析效果，实现从整体到局部、再到故障细节特征的逐步精细识别。

得益于一双又快又准的“盘古眼”，这套 TFDS 系统能精准识别 67 种货车 430 多种故障类别，重大异常故障 100% 识别，综合故障识别率达 99.8%；原来人工需要识别 4000 张图片，现在仅需要复检 170 多张图片，工人劳动强度下降 95.75%，极大提升了检测效率和故障识别率，提高了列车安全性。

4.4 矿山



山东能源集团（简称山能集团）以矿业、高端化工、电力、新能源新材料、高端装备制造、现代物流贸易为主导产业。其中，煤炭产量位居全国煤炭行业第三位，矿井智能化生产水平居行业前列，9处矿井成为首批国家级智能化示范矿井。

山能集团依托盘古大模型建设了集团人工智能训练中心，把盘古矿山大模型全面应用到采、掘、机、运、通、洗选等9大业务系统，具备视觉大模型、预测大模型、自然语言大模型等三大能力，探索和发掘煤矿生产领域全场景的人工智能应用，通过技术创新实现“人工智能大规模下矿”，让员工远离井下作业环境，实现“高效、安全、可持续性”的生产运营管理。

1. 支持智能生产模式创新

1) 样本训练效率高

通过云边协同方案，打通集团中心云和矿山边缘云数据，低代码小样本训练，模型自动优化，边用边学，能以更少的数据达到其他模型相同乃至更高的精度；通过云边协同方案，在其AI标杆兴隆庄煤矿的一处训练，全集团

共享，未来可复制到集团其他70+矿井。

2) 海量吞吐信息处理

利用无监督训练策略对海量信息（图片数量10亿+，视频信号>100TB）进行归纳抽取训练得到的模型，具备强大视觉表征识别能力。

3) 模型移植能力

大模型相比小模型有良好的泛化性能，在相似场景上训练的模型可迁移到未进行训练的新场景上，并且可以快速地在新的矿井进行部署和上线应用，无需从零开始大量重复训练。

4) 数据筛选效率高

大模型具有在全新场景实现缺陷样本高效筛选的能力，相对传统小模型训练方式，可以节省85%的标注人力。

5) 模型识别精度高

基于“非正常即异常”识别原则，快速训练生产、安监、决策的L2场景化模型，在同等少量样本训练的情况下，大模型精度高出小模型10%。

2. 提升生产质量效益

如在洗选煤和配煤场景中，相关生产工艺数据输入因素关系复杂，无法完全凭人工经验来确定。大模型通过厂矿实际数据进行建模，协助解决相关参数准确预测和控制的问题，平衡生产质量与成本，提高生产效率和效益。

在洗选煤参数优化场景中，通过预测大模型构建自主预测分选密度模型和产品灰分预测模型，进行旋流器 / 全流程控制参数优化，根据系统观测到的灰分比，快速自动调整悬浮液密度以及入口压力等工作参数，实现稳定精煤灰分、提升精煤回收率 0.1% ~ 0.2%，每年多产出 8000 吨精煤。这个能力推广到全国，可让每个煤矿每年平均多产出 2000 吨精煤。

在焦化配煤优化场景中，利用图网络技术训练配煤优化模型，可帮助配煤师提升输出配比效率，预计人工耗时可从 1-2 天缩短到分钟级。



3. 降低安全生产风险

通过盘古矿山大模型和 AI 应用的视觉识别能力，原恶劣作业环境下每天巡检改为每周巡检一次，节省人力的同时，也改善了巡检人员的作业环境。

在兴隆庄一期项目中，对于危险区域人员入侵识别等场景，大模型识别率达 90% 以上。基于大模型算法，系统识别精度比传统小模型提高 10%。在实际应用中，系统可通过告警提醒，避免潜在危险发展为安全事故，并进一步规范井下人员的作业行为，提升安全意识。

钻孔深度是防冲卸压工程的关键参数之一是防冲工程管理人员人工核验的重点。基于盘古矿山大模型，实现了防冲卸压施工孔深度智能监管。基于专用摄像仪对施工过程动态监管，现场视频可实现实时上传、智能核验，在孔深不足时及时进行声光数字化告警。系统还可设置施工计划管理、识别结果查询、施工深度核验、施工数量统计等功能，便于工程核验和监管，提高监管时效性与准确性，降低人工核验工作量 80%。



4.5 电力



电力公司负责电网规划建设、运行管理、电力销售和供电服务工作。

通过引入华为云针对电力行业开发的盘古预训练大模型 AI 推理服务，实现了山区高压输电线的无人机智能巡检。一个大模型替代原有 20 多个小模型，并且模型精度提升 18.4%。原来需要人工登塔才能完成的杆塔巡检，现在通过操作无人机就可以完成。原本人工需要 16 天才能完成的杆塔巡检缩减至 2 天、效率提升 8 倍、线路故障率降低 60%。

“以前我去巡山，一出门就是半个月，现在有了无人机这个‘千里眼’，又快又准，2 天收工。”某电力巡检工作人员如是说。



05

展望未来，从感知到生成

视频智能分析服务（VIAS）开箱即用的算法，可以实现智慧园区、城市治理、安全生产等场景的事件感知、分析和决策能力。盘古 CV 大模型提供预训练工作流，可以用类似工业流水线的方式快速生成场景化模型，助力企业实现人工智能转型，构建“内生的，持续发展”的 AI 能力。盘古视频解译大模型更进一步，实现视频、图像、文本、语音之间的自由转换，提供多模态理解能力，让摄像机开口说话，可以为多个行业带来变革。

例和城市治理领域，将摄像机拍摄的视频流转换成文本，可以实现异常事件（例如火灾，极端天气等）的主动上报；自动驾驶领域，将车载摄像头捕捉的道路图像转换成文本，以帮助自动驾驶系统理解周围环境；客户服务和智能助手领域，将视频通话中的对话、动作、表情转换成文本，以便智能客服代理理解用户需求并提供支持；广告和内容推荐领域，将在线视频的内容转换成文本标签，以便更好地匹配广告或推荐相关内容，同时分析用户观看的视频流，将其转换成文本，以便为用户提供个性化的推荐；教育和培训领域，将教学视频转换成文本，以便学生搜索和理解课程内容。

善于利用工具让人类从众多生灵中脱颖而出，成为世界的主宰。摄像机是人类眼睛的延伸，功能甚至比眼睛还要强大。红外摄像机、紫外摄像机、高速摄像机、偏振摄像机等等，可以捕捉到人类眼睛无法捕捉的画面。海量摄

像机产生的海量视频数据，记录了世界的点点滴滴，也带来了数据爆炸的困扰。海量的视频数据存储成本高昂，导致很多视频流数据被忽略，很多摄像机拍摄画面没有被利用，形同虚设。人工智能技术和视频技术的深度结合，相信可以解决这个问题。用更高效的方式记录和存储，用有效的方式感知和判断，忽略无效画面，不放过任何有效信息。城市角落的安全时间，工业生产的细微缺陷，四季更替的极端天气，甚至浩瀚星空中的天外来客，都应该触发人工智能的识别，感知和分析，让城市更安全，工业更高效，生活更美好，实现科技服务生活。

著名的物理学大师费曼说过“凡我不能创造的，我都不能理解”。意味着理解和生成是相伴而生的。如果能够充分理解视频，那么生成视频所需的训练数据问题就迎刃而解，可控、高质量视频生成也自然水到渠成。所有人都对虚拟世界充分幻想、影视、游戏、短视频等娱乐领域，会因为视频生成释放万亿的产值。工业生产中，也大量需要模拟和仿真、异常场景构建等等。生活中，也需要用增强现实的技术，提升沟通的效率，彻底解决“可意会不可言传”。

技术发展会起起伏伏，但没有人会忽视人工智能和视频这两个和人类生活息息相关的领域。人工智能一定能和视频技术碰撞出绚烂的火花，进一步释放人类的想象力，让生活变得更加美好。



华为技术有限公司



深圳龙岗区坂田华为基地

电话：+86 755 28780808

邮编：518129

www.huawei.com

商标声明

 HUAWEI, HUAWEI,  是华为技术有限公司商标或者注册商标，在本手册中以及本手册描述的产品中，出现的其它商标，产品名称，服务名称以及公司名称，由其各自的所有人拥有。

免责声明

本文档可能含有预测信息，包括但不限于有关未来的财务、运营、产品系列、新技术等信息。由于实践中存在很多不确定因素，可能导致实际结果与预测信息有很大的差别。因此，本文档信息仅供参考，不构成任何要约或承诺，华为不对您在本文档基础上做出的任何行为承担责任。华为可能不经通知修改上述信息，恕不另行通知。

版权所有 © 华为技术有限公司 2024。保留一切权利。

非经华为技术有限公司书面同意，任何单位和个人不得擅自摘抄、复制本手册内容的部分或全部，并不得以任何形式传播。