## 智慧(生产)园区视频监控方案

### 一 背景介绍

#### 1.1 定义



完整意义的智慧园区是个复杂的综合概念，如上图所示，是指融合新一代信息与通信技术，具备迅捷信息采集、高速信息传输、高度集中计算、智能事务处理和无所不在的服务提供能力，实现园区内及时、互动、整合的信息感知、传递和处理，以提高园区产业集聚能力、企业经济竞争力、园区可持续发展为目标的先进园区发展理念。

在当前报告中，我们仅聚焦其中安防监控这一个模块展开。智慧园区视频系统是指基于现代信息与通信技术，利用摄像头(常规+红外)视频采集设备，实现对园区内各个场景和设备的实时监控、预警和管理的系统。具有以下优势:

1. 实现了对园区的全面监控，提升安全感，威慑不法分子，有效防范和打击各类安全风险；
2. 减少人力投入，提高监控效率，避免不必要的人力、物力资源浪费；
3. 提供了可追溯的数据，为安全管理及责任鉴定提供依据和参考。

#### 1.2 建设内容

传统安防主要是以硬件设备为主，依靠人力进行监控和巡查，存在很多管理盲区，而且无法实现数据的深度挖掘和分析园区，建设还是处于基础的IT信息化阶段，只是以弱电工程还有基础管理的软件为主，业务并没有实现数字化。而智慧园区是基于数字化技术兴起的新兴产业，智慧安防则通过物联网和人工智能等技术，将各个安防系统进行集成，实现对园区的全面感知和智能化管理。通过数据分析和挖掘，智慧安防能够预测园区可能面临的安全风险，并提前采取相应的防护措施。

建设方案的核心, 是通过视频采集设备的全方位部署，系统并实时获取园区各个角落的图像信息，并将其传输至中央监控室，为管理者提供全局、直观的视觉感知。不论是白天还是夜晚，系统都能够提供清晰、稳定的画面，确保整个园区的实时监控，当突发异常状况，监控室能第一时间了解详情，给出应急预案。

智慧安防在具体实施过程中，以平台化的思维，借助物联网、互联网以及人工智能等先进的技术去打造一个系统性工程，一般分为三大部分, 第一部分是智能化的基础设施, 第二部分是智慧化的信息平台, 第三部分是一体化的运营服务。

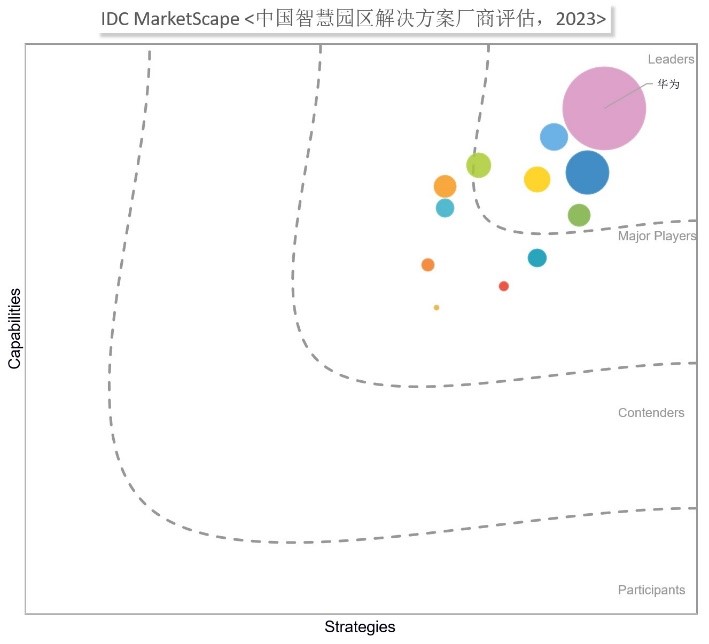
基础设施是指采集、传输、存储信息的设备，例如数据的设备, 交换的设备, 及各类控制设备，这些设备与信息管理系统连接起来, 同时也是这些设备和外界通讯, 网络相连, 它还包括建筑物, 外部网络和电信线路的连接点, 及应用系统, 设备之间的所有的线缆。如何系统规划、安装、管理这些硬件不是当前报告重点，此处略过。

智慧化的信息平台是指对采集的数据进行分析，然后分析结果连同视频一起展示在显示屏幕上，信息包括各处人员流动，重要设备性能监控，停车场车位数量，门岗进入申请等，我们提供的解决部分是对视频数据的深层解析，可划分到这一部分。

运营服务由人参与，监控信息平台显示的各种数据，汇总统计各项参数，为园区良好运行提供数据支撑，发现异常后及时采取相应措施。这个部分也不在当前报告中体现。

### 二 业界动态

目前深耕该领域的企业很多，搜索到的部分厂家见附录一。这些厂家大多有自己硬件产品，他们的园区方案通常部署在自家硬件上。同时，交付的是一个园区区完整的智慧生态，投入人力、物力、时间都十分巨大。在2023年12月国际信息咨询公司IDC MarketScape发布《中国智慧园区解决方案2023年厂商评估》，对12家典型的智慧园区解决方案厂商进行重点研究后给出影响力排名：



以下给出部分厂商方案架构图，所有资料都从官网获得。从结构图可以看出，硬件设备（感知层）几乎都一致，所有创新点与技术壁垒都在软件生态的构建上。

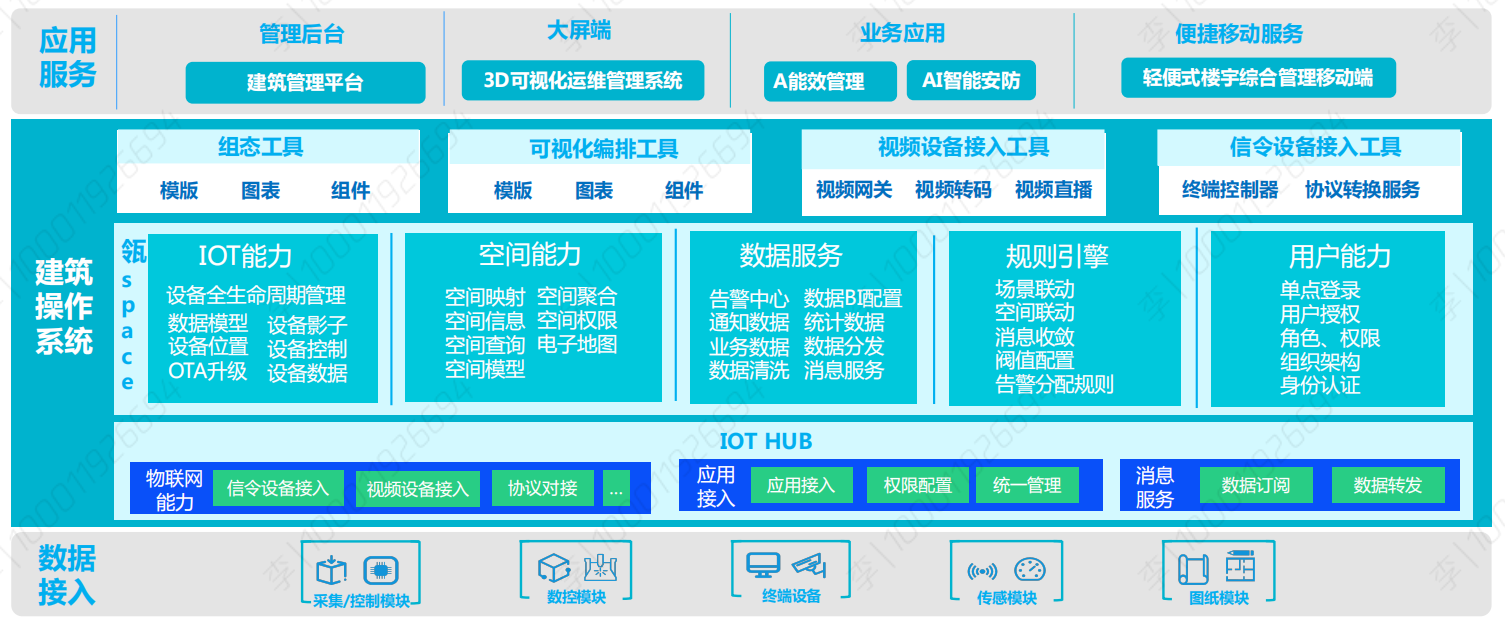
#### 2.1 华为园区智慧园区



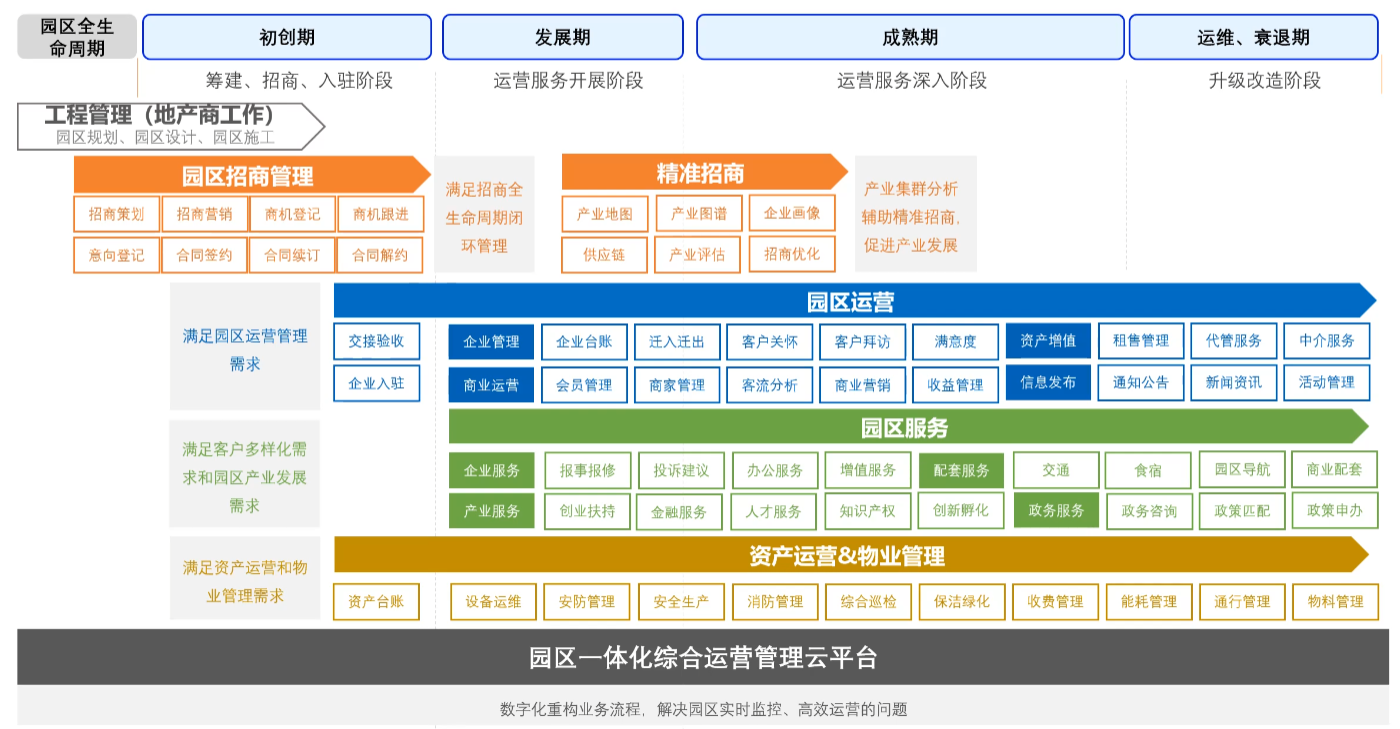
#### 2.2 旷视



#### 2.3 腾讯卫瓴BIM智慧园区



#### 2.4 世茂物联



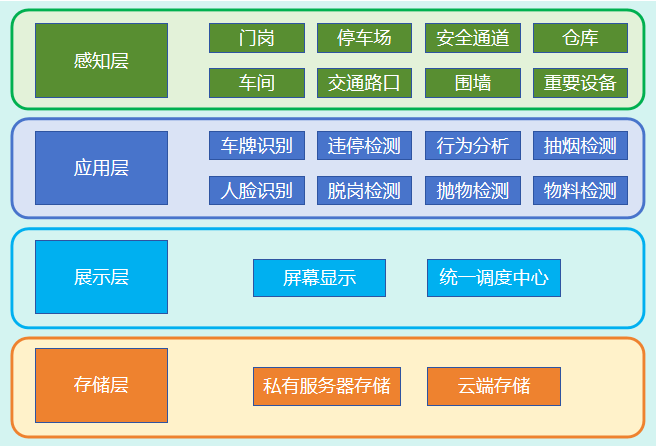
### 三 解决方案

我们给出的解决方案，属于视频分析范畴，因此主要针对视频监控场景，反应在智慧园区规划中，属于智慧安防模块。

#### 3.1 设计流程

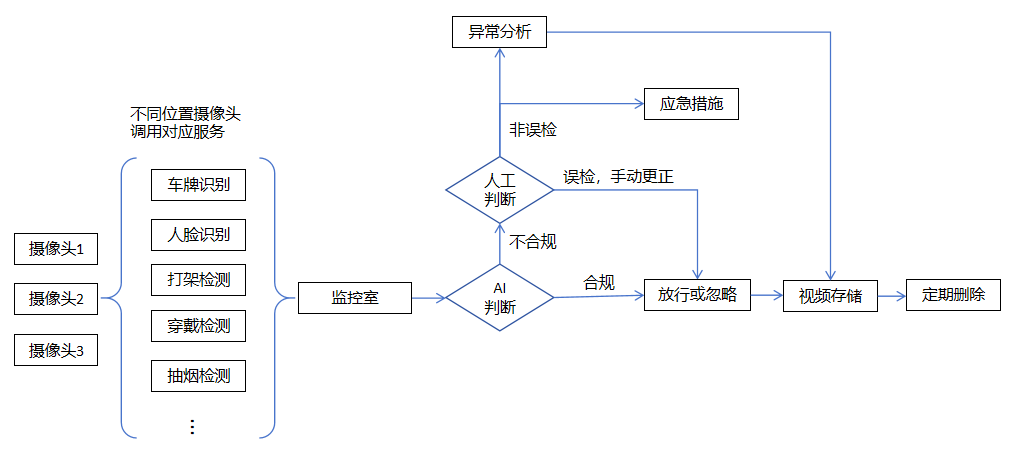
在监控场景中，所有视频来自摄像头，已知所有摄像头视频都会接入总监控室中，在接入总控前，会先经过数据服务器，调用部署的视频类业务模型，如场景检测，人脸识别，穿着检测等，在原图上绘制检测结果，并将分析结果与处理后视频输出到控制室屏幕上，再对触发的特定事件提供声音报警功能。

总体流程可分为以下4个层级：



1. 感知层：在各个重要设施和出入口安装摄像头，数据实时接入服务器后台。
2. 应用层：应用部署在服务器上，根据部署摄像头的位置，对捕获的视频调用指定应用，处理后的视频传入下一级展示层。
3. 展示层：通常位于监控室，将监控画面实时展示在显示器上，应用层检测到的异常也会在这里显示出来，再由这里的调度中心统一处理。
4. 存储层：所有录入视频可选择性保留。有私有，云端两种存储方式，考虑到数据安全性，建议存储在本地私有服务器上。

#### 3.2 使用流程



在3.1中我们划分了流程的层级结构，在这里给出使用规范，

1. 从摄像头捕获的视频，实时传入服务器中，根据摄像头分配的功能，调用指定服务。
2. 给出ai判断结果，实时传输给监控室。
3. 人员通过监控器观察，结果是否误报，是误报，手动忽略；不是误报，人工处理。

#### 3.3 感知层部署方案

园区应用场景分为园区出入口、重要设备、围栏、建筑出入口、大厅、食堂、电梯、走廊、办公区、收银、步梯、停车场、园区开阔地带等。根据位置划分，可粗略分为室内、室外两种。

##### 3.3.1 室外安装

该场景中环境亮度变化较大，白天存在逆光环境，夜间环境较暗，需要全天候看清进出人员的脸部特征；推荐使用高清摄像头并兼具红外摄像功能，便于夜晚也能看清目标。对于监控人员聚集与行为识别的摄像头，若对被观察人ID不感兴趣，可以适当降低摄像头像素，节约成本。对于围栏处的视频采集设备，需要太阳能备用电池和无线网络传输功能，保证电源线和网络被破坏时也能工作。室外摄像头建议不低于500万像素（3K画质），同时长期处于室外恶劣环境中，因此产品必须防水防尘，建议安装方式为壁装和吊顶装，并定时清理镜头灰尘。

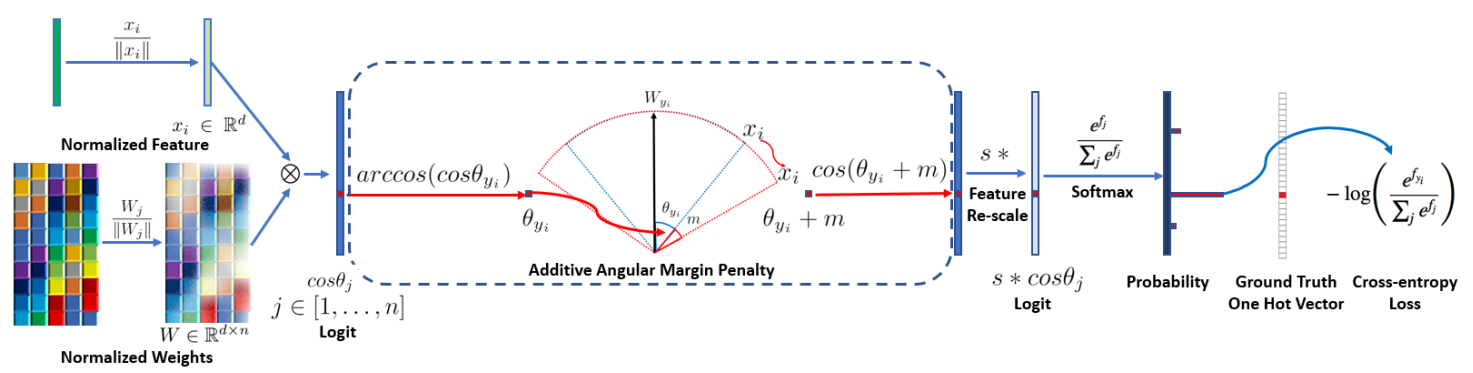
##### 3.3.2 室内安装

一般室内灯光充足且稳定，除监控重要资产摄像头需配备红外功能外，其他都不需要，该场景视频采集设备都推荐广域摄像头，扩大监控面积。安装位置需要确保能以合适角度排到经过人员面部特征，便于身份ID识别，与室外不需鉴定ID的行为分析不同，室内的行为识别还需要定位人员身份，如室内抽烟影响他人，对重要设备的破坏行为等。室内摄像头像素最小不低于200万像素。

#### 3.4 基础技术分析

在应用层涉及的大量功能应用，从技术角度深度解析，可归纳为以下算法:人脸识别、目标检测、OCR识别、行为分析这三大类，他们的下游任务都可在实现过程中通过具体的限定条件完成。

##### 3.4.1 人脸识别



该模块用于身份核验，适用于需要鉴定的身份的所用摄像头模块，通常发生在门岗，重要设备、车间作业、打卡位置，需要使用高清摄像头（建议800w像素），安装在能拍到人脸的合适位置。可以预见该场景下的挑战和问题：如光照条件、姿态变化、遮挡等因素可能影响识别性能。

① 人脸识别是一种基于人脸图像或视频的生物特征识别技术，通过分析和比对人脸的特征来进行身份验证、识别等任务。主要以下步骤：

② 人脸检测：识别的第一步是检测图像或视频中的人脸位置。常见的方法包括yolo\_face, resnet, retinaface和scrfd等技术，人脸检测是人脸视频的前置技术，可作为单独一个模块使用，这样与人脸识别合并作为一个模块使用。

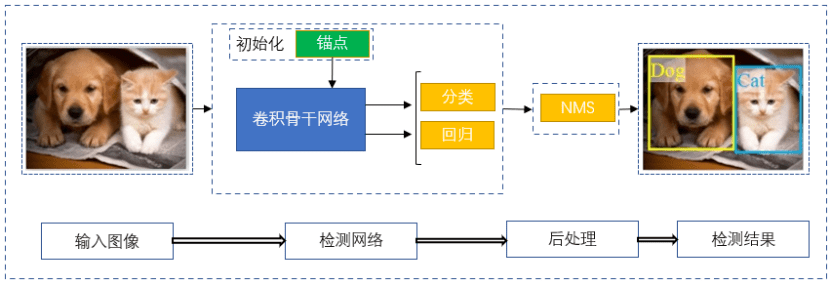
③ 人脸特征提取：基于深度学习的方法训练模型，如成熟的技术arcface。在推理阶段，将人脸图像转换为具有辨别能力的高维特征向量。

④ 特征匹配和比对：在人脸特征提取后，需要将提取的特征与已知的人脸特征进行比对。在此之前，要先将经过园区认可的人员人脸数据存入底库，从摄像头捕获的人脸，经过同样模型提取特征后，与底库数据比较，匹配相似度。

在以上步骤，特征提取模型是人脸识别的核心算法，将人脸图片转换为有区分度的高维向量，本质上是一个度量学习问题，其关键是学习到有判别力的大裕度特征。与分类问题相比，度量学习的目标是大的类间（class-inter）方差和小的类间（class-intra） 方差，因此模型设计的核心是损失函数。

人脸识别算法的损失函数可以分为**基于欧式距离的损失**、**基于角/余弦裕度的损失**和 **softmax 损失**及其变种等三种类型。基于欧式距离的损失将图像嵌入到欧式空间，使得类内方差减小，类间方差增大；基于角裕度(angular margin)的损失使学到的特征具有潜在的大的角距离。

##### 3.4.2 目标检测



目标检测是计算机视觉领域中的重要任务，旨在从图像或视频中准确地定位和识别出不同类别的目标物体，园区内许多任务都依赖该算法来实现，如安全帽检测，口罩检测，统一工服检测，脱岗检测。违停检测等，虽然检测内容不同，不过它们技术思路一致。都可归为目标检测范畴。

目标检测算法是当前AI落地技术中最成熟的算法之一，经过学术界/工业界多年技术推进，目前已有众多优秀算法投入工业应用，而从实时性考虑，单阶段算法最符合我们的需求。

单阶段检测器是一类直接从图像中输出目标框和类别的方法，与Faster-RCNN两阶段系列算法不同，它们无需生成候选区域。常见的单阶段检测器包括YOLO（You Only Look Once）和SSD（Single Shot MultiBox Detector）。这些方法通过在不同尺度上预测目标框和类别，可实现实时目标检测。算法组成有以下三部分：

骨干网络：骨干网络是目标检测中的核心组件，用于提取图像特征。常用的骨干网络包括CspDarkNet、ResNet、Inception、MobileNet等，根据硬件不同选择合适的网络，这些网络通过层叠卷积层和池化层，逐渐提取出图像的高级特征，用于目标检测任务。

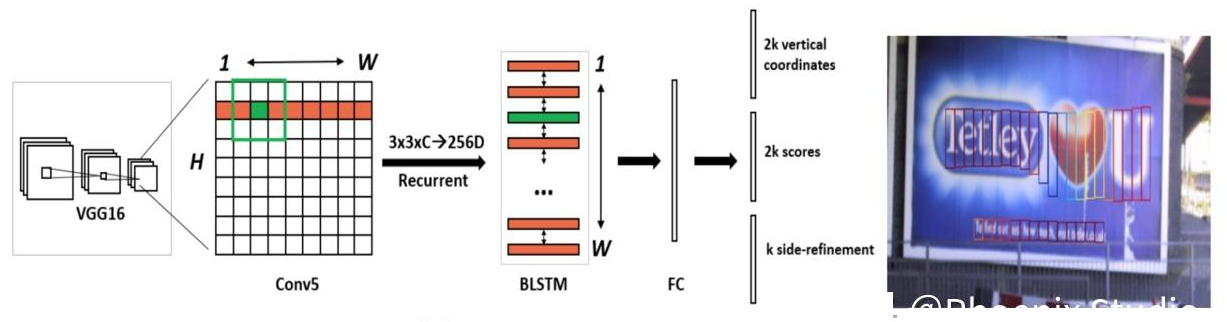
特征金字塔网络：特征金字塔网络用于解决目标在不同尺度上的变化问题。由于目标可能出现在不同大小的区域中，特征金字塔网络可以在不同层次上提取多尺度的特征，并结合这些特征进行目标检测。当骨干网络输出特征与特征金字塔网络结合，就能检测出图片上不同尺度的同类目标。

后处理模块：对输出特征解码后能获得很多目标框，这时要进行NMS（非极大值抑制）过滤多余的重叠框，这个过程中，NMS参数设置（iou与threshold）对结果输出由很大影响，合适的参数选择能提高检测精度。

目标检测领域仍有部分问题待解决，如小目标，复杂背景与遮挡，少样本等问题，需要根据具体任务选择合适算法。

最新transformer模块在该领域学术圈取得巨大成功，如DETR、transformerable DETR、在多个检测领域性能指标超越CNN网络，RT-DETR也在实时性、准召率上超越YOLOv8，这些优秀算法也是工业场景应用的有力备选方案。

##### 3.4.3 OCR识别



通用识别模块，OCR（Optical Character Recognition，光学字符识别）是指对图片中的文字进行查找、提取、识别的一种技术，通过检测暗、亮的模式确定其形状，然后用字符识别方法将形状翻译成计算机文字的过程。在园区中可用于出入车辆车牌号、员工衣服编号、不恰当的标语、身份证等文字识别。算法过程由以下两个步骤构成：

①文字检测：解决的问题是哪里有文字，文字的范围有多少

②文字识别：对定位好的文字区域进行识别，主要解决的问题是每个文字是什么，将图像中的文字区域进转化为字符信息。

文字检测的任务是定位出输入图像中的文字区域。近年来学术界关于文本检测的研究非常丰富，一类方法将文本检测视为目标检测中的一个特定场景，基于通用目标检测算法进行改进适配，如TextBoxes基于一阶段目标检测器SSD算法，调整目标框使之适合极端长宽比的文本行，CTPN则是基于Faster RCNN架构改进而来。但是文本检测与目标检测在目标信息以及任务本身上仍存在一些区别，如文本一般长宽比较大，往往呈“条状”，文本行之间可能比较密集，弯曲文本等，因此又衍生了很多专用于文本检测的算法，如EAST、PSENet、DBNet等。目前较为流行的文本检测算法可以大致分为基于回归和基于分割的两大类文本检测算法，也有一些算法将二者相结合。

文本识别的任务是识别出图像中的文字内容，一般输入来自于文本检测得到的文本框截取出的图像文字区域。可以根据待识别文本形状分为规则文本识别和不规则文本识别两大类。规则文本主要指印刷字体、扫描文本等，文本大致处在水平线位置；不规则文本往往不在水平位置，存在弯曲、遮挡、模糊等问题。不规则文本场景具有很大的挑战性，也是目前文本识别领域的主要研究方向。规则文本识别的算法根据解码方式的不同可以大致分为基于CTC和Sequence2Sequence两种，将网络学习到的序列特征 转化为 最终的识别结果。不规则文本的识别算法相比更为丰富，如STAR-Net等方法通过加入TPS等矫正模块，将不规则文本矫正为规则的矩形后再进行识别；RARE等基于Attention的方法增强了对序列之间各部分相关性的关注.此外，随着近年来Transfomer的快速发展和在各类任务中的有效性验证，也出现了一批基于Transformer的文本识别算法，这类方法利用transformer结构解决CNN在长依赖建模上的局限性问题，也取得了不错的效果。

##### 3.4.4 行为分析

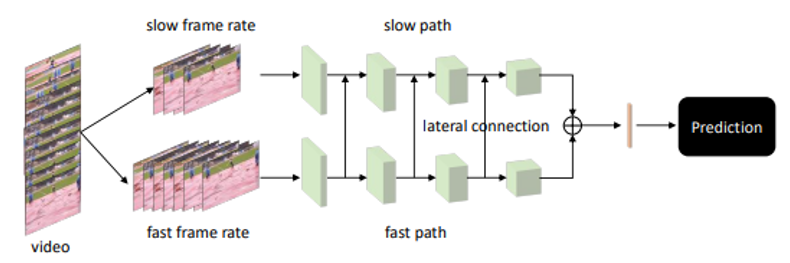
行为识别旨在从视频中识别和分类不同的行为或动作。这可以是简单的动作，如走路、跑步，也可以是复杂的行为，如打斗、交谈等。行为识别可以通过传统的机器学习方法或深度学习方法来实现，行为本身可以大致分为3个子任务，具体如下：

1、动作识别：动作识别不仅要分析视频中每帧图像的内容，还需要从视频帧之间的时序信息中挖掘线索。动作识别是视频理解的核心领域，虽然动作识别主要是识别视频中人的动作，但是该领域发展出来的算法大多数不特定针对人，也可以用于其他视频分类场景。

2、时序动作检测：时序动作识别可以认为在动作识别的基础上增加了时间维度上的信息，动作识别可以看作是一个纯分类问题，而在时序动作定位领域，视频通常没有被剪辑，视频时长较长，动作通常只发生在视频中的一小段时间内，视频可能包含多个动作，也可能不包含动作，即为背景类。时序动作定位不仅要预测视频中包含了什么动作，还要预测动作的起始和终止时刻。

3、时空动作检测：时空动作检测在时序动作的基础上又增加了空间维度，不仅仅需要定位事件发生的时段，还需要定位感兴趣的人或物体所在的空间位置。

行为分析算法在近年来快速发展，这里我们主要介绍目前效果较好的基于深度学习的主流算法。通过使用卷积神经网络（CNN）或循环神经网络（RNN）等深度学习模型，可以对视频数据进行端到端的学习和预测。这些模型可以从原始视频数据中提取特征，并进行行为分类、检测和跟踪等任务。该领域已积累大量研究算法，如VideoMAE、MViT V2、TCANet、LFB、STGCN++等，这些算法从不同思路给出了解决思路。



以SlowFast算法为例简要描述下视频行为的分析流程，结构图如上所示，slowflat的思想在于外观和动作的变化速度是不同的，外观的变化通常很慢，而动作变化速度很快，使用低速分支基于低帧率图像捕捉外观特征，使用高速分支基于高帧率图像捕捉运动特征，并进行分阶段融合，这种方式对计算量及识别性能具有一个非常好的平衡。

#### 3.5 应用层功能设计



##### 3.5.1 功能项目

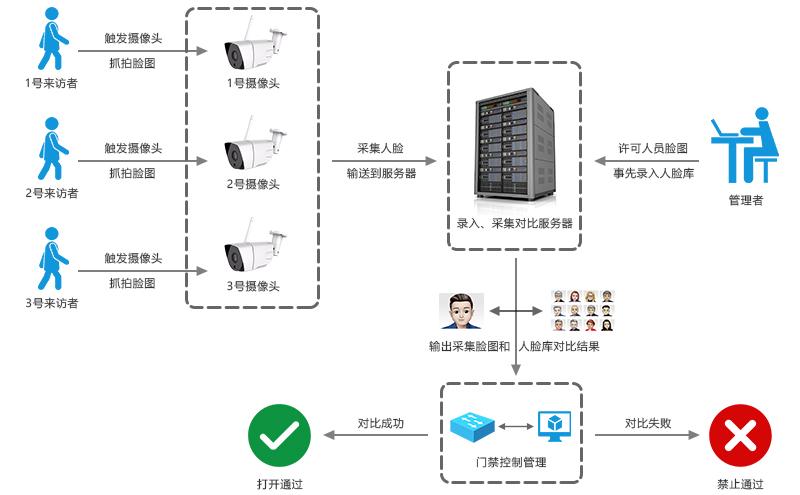
针对工业生产环境, 应用对接入的视频流应具备以下功能的处理能力：

1. 警戒功能: 检查到黑名单或非录入系统的人员，在终端监控室显示器视频画面上标记。
2. 打卡功能: 在园区，车间，餐厅入口, 人脸识别员工ID，记录最早、最晚出现的时间。
3. 行为分析: 在围墙上的监控设备，需要对传回终端的视频出现的人员做出行为分析， 如向外抛物，偷盗，翻越栏杆，抽烟，打架，聚集，破坏设备等行为给出预警。
4. 安全检测: 对指定区域物品排放进行监控，发现未按规定摆放的物品向监控室发出警告，避免安全通道堵塞，生产原料随意堆放等情况发生。
5. 脱岗检测: 需要人脸识别辅助，记录人员离开工作岗位时间。
6. 着装检测: 检测工人是否按规定着装，如安全帽，口罩，工装等.
7. 车牌识别：对进入园区的车牌进行识别，确认是否为园区已登记车辆；记录在园区停留时间。
8. 违停检测：检测车牌停放位置是否合格，避免不能及时发现而堵塞道路。
9. 场景检测：监控设备覆盖的场景，对不能正常工作的设备进行预警，如路灯不亮、栏杆倒塌、烟雾（火灾）等。
10. 缺陷检测：对产线产品进行外观检测，确认是否合格。
11. 目标追踪：视频监控阶段, 追踪视频中出现的指定目标。
12. 视频检索：视频回放阶段, 根据文字描述检出指定人与物品出现的时间段。

##### 3.5.2 功能实现

在3.5.1节描述的各种功能，都可通过3.4节阐述的基础技术组合来实现。在此给出各功能算法实现逻辑，更具体的细节在项目实施过程中再做调整。

1. 警戒功能



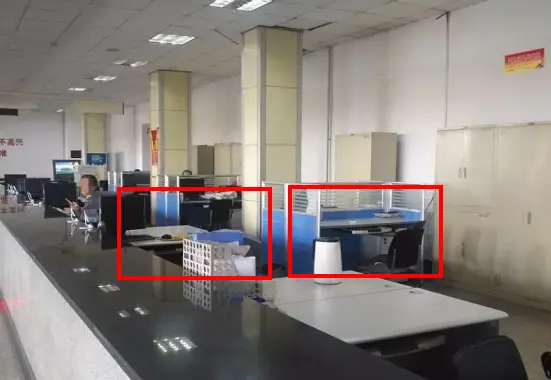
将黑名单（例如已被辞退员工）与园区成员人脸照片录入数据库，对门口、围栏、重要设备等位置部署的摄像头采集的视频流，每一帧图像进行人脸检测，即找出图像中的人脸位置和边界框。然后，对每个检测到的人脸应用已训练好的人脸识别模型，将得到的特征向量与数据库中已知人脸特征进行比对，以确定其身份，根据人脸识别的结果，可以根据需要采取相应的行动：

①检测到陌生人，向终端控制室发出提示。

②检测到黑名单中的人脸，可以触发报警或采取进一步的安全措施，记录当前目标出现时间、位置等信息，以便后续分析和审查，同时二次调用行为分析与目标跟踪算法，时刻显示在监控屏幕上。

③检测到正常员工，不做后续处理。

1. 打卡功能、脱岗检测



与作息时间相关的功能都可归为当前类别，拥有该功能的视频采集设备必须具有时间自校准能力。因该功能与员工考勤关联，因此，为提高人脸辨识率，摄像头必须高清，环境采光良好且部署位置恰当，安装位置最好能捕捉人脸正面。在算法内使用可配置的时间节点值，该功能需要人脸识别算法辅助，在上下班时间段内，记录人员出现的最早、最晚在指定位置的时间。在工作时间内，使用目标检测技术观察指定位置是否有员工，若未检测到，开始计时，直到该工位员工回来（因为涉及隐私问题，此处不使用人脸识别算法，可能出现人员顶替现象）。记录离开工位次数与暂离时间，最后向相关统计全部写入Excel表格，以备考勤回溯。

1. 着装检测



适用于目标检测算法，实现步骤如下：

① 数据采集：收集各子任务的图像数据，这些数据可以通过摄像机或其他设备在园区内进行采集。

② 模型训练：使用标注好的数据集，将提取的特征和对应的标签输入到深度学习模型中进行训练。训练过程中，模型将调整权重和参数，以最小化损失函数，使得模型准召率达到用户标准。

③ 检测：将摄像头采集的视频实时输入到模型中，通过设置合适阈值，可以判断员工着装是否合规。

④ 报警或记录：对着装不合规人员，进一步调用人脸检测+人脸识别算法核验身份，记录违规事件的时间、地点和工人信息，以便后续分析和处理。

4. 安全检测、违停检测、场景检测

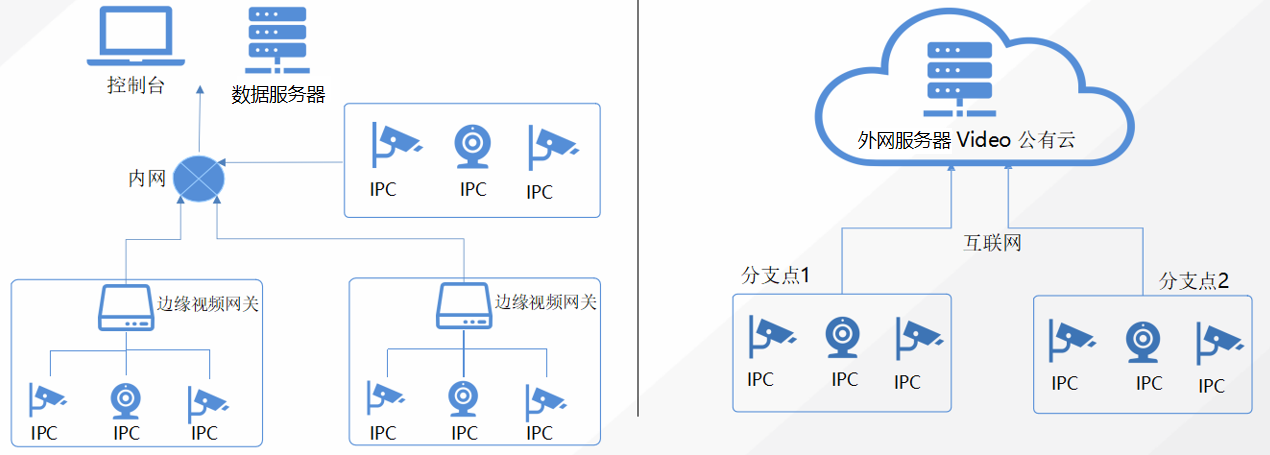


这些类型算法有共同特点：不需要实时监控，每隔一段时间观察指定位置是否被遮挡或被其他物体占据即可，没有实时性压力。由于指定区域需要人工标定，因此需要部署固定角度监控设备，避免画面转动导致标定区域移位。

算法实现步骤同着装检测等，只是不再需要人脸相关技术辅助。在违停检测中，当发现违规时，需要OCR技术提取车牌号，然后将车牌号和事发位置信息返回到主控制。

#### 3.6 组件分析

##### 3.6.1 数据传输



从摄像头获取的视频数据流，需要实时传输到服务端，通过光缆、以太网光电转换器、超五类以太网双绞线等进行信息的传输、交换和控制，能够有效地进行通信和共享数据，实现各系统前端设备和后台中心交换机的通信。视频传输协议有网络传输协议（如RTSP、RTP、RTMP）和实时传输协议（WebRTC）供选择，推流方式的选择具体需求和系统架构。在接入服务器后，首先要对数据流解码，然后调用图像处理算法，实际经验表明，解码操作比较耗时，会影响算法那整体流畅度，大尺寸的高分辨率视频帧表现尤为明显，可以考虑GPU解码。

##### 3.6.2 算法调用

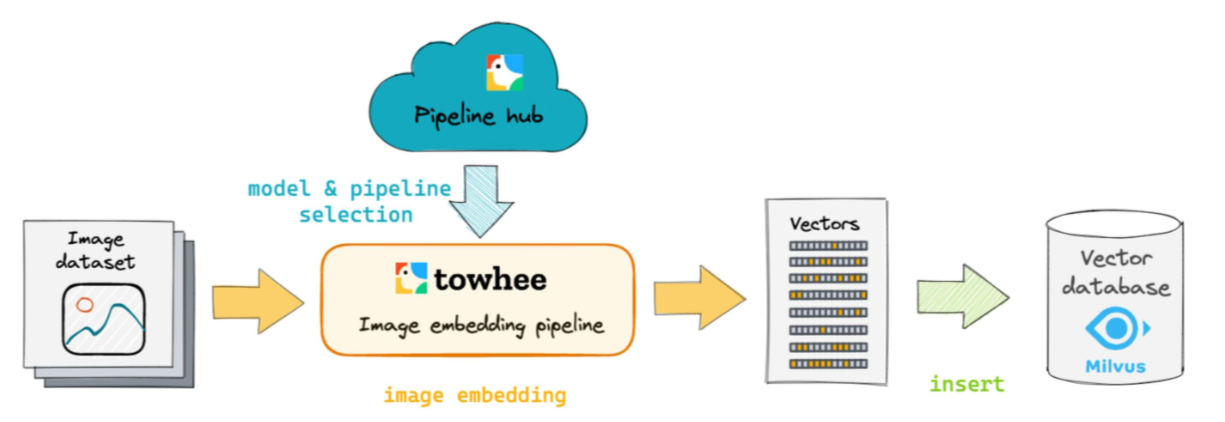


根据以上分析可知，园区会部署大量摄像头，AI算法有时会嵌入摄像机内部，通过搭载高性能计算芯片，将摄像机从一个单一化的视频采集前端摄像机进化到一个智慧数据采集、分析平台，这种方式非常高效，弊端也很明显，即需要事先将算法参数固化在芯片上，后续迭代升级困难。我们建议将算法部署在专用服务器上，多个摄像头部署的算法功能有时会重叠，此时需要算法有高效的并发能力，通常以多实例形式呈现，由Nginx调度协调，保证返回结果的实时性。若算法不能完成每一帧的实时处理，应丢弃相邻的部分帧，使用间隔采样方式，保证处理图片帧的实时性。

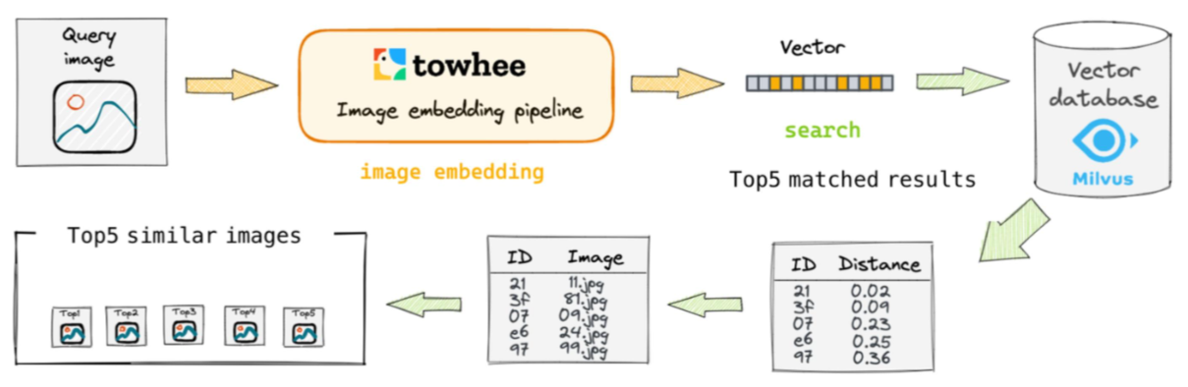
##### 3.6.3 特征存储与访问

使用milvus工具构建人脸信息存储库，通过人脸识别模型录入人脸数据后，获得256/512维特征，不断加入特征库。需要识别人脸时，提取人脸特征与特征库所有人脸比对，返回最接近的低库图片和相似度。关于该向量库，我们曾详细研究过不同参数配置对查询性能影响，可依据部署时具体数据量给出最佳参数设置。存储与访问的两个流程如下：

**① 向量存储**：输入人脸证件照 --> 转化成 向量 --> 写入向量数据库



**② 向量访问**：待查询人脸图片 --> 转化成向量 --> 查询向量数据库 --> 返回最符合结果 --> 展示



#### 3.7 我们的优势

经过调研发现，多数TOP厂家都把精力放在系统的构建上，为整个园区搭建完善的管理系统，监控可收集的各种数据细节，虽然做到了全面、实时，但仍然需要人员肉眼监控所有监控画面，一旦不能及时发现视频中问题，实时性就失去意义。同时他们提供的多是通用系统，很难处理园区独有问题。而我们的设计方案, 可以从源头接入视频，在发送到监控室前，就对视频进行解析，一旦发现不符合预定义行为，直接输出警告信息，对实时性监控起到很大的辅助作用；同时，我们在视频领域由深厚积淀，能为园区提供定制化算法，灵活处理各种技术问题。

### 四 附录

##### 4.1 相关企业排名

