ICS 35.240
CCS L70

SJ

中华人民共和国电子行业标准

SJ/T XXXXX—XXXX

人工智能 工业视觉模型训练推理平台技术规范

Artificial Intelligence- Technical specification of the industrial visual model training and inference platform

报批稿

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中华人民共和国工业和信息化部　　发布

目　次

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利，本文件的发布机构不承担识别专利的责任。

本文件由工业和信息化部科技司提出。

本文件由中国电子技术标准化研究院归口。

本文件起草单位：中国电子技术标准化研究院、深圳思谋信息科技有限公司、北京思谋智能科技有限公司、赛西(深圳)电子信息产品标准化工程中心有限公司、上海思谋信息科技有限公司、清华大学、中国电信股份有限公司上海研究院 、浙江大华技术股份有限公司、中国联合网络通信有限公司、杭州汇萃智能科技有限公司、成都数之联科技股份有限公司、海关总署研究中心、中移互联网有限公司、神华新街能源有限责任公司、杭州海康威视数字技术股份有限公司、海信视像科技股份有限公司、广东中科凯泽信息科技有限公司、卡奥斯工业智能研究院（青岛）有限公司、中国电子技术标准化研究院华东分院、中科慧远视觉技术（洛阳）有限公司、浪潮通信信息系统有限公司、西北工业大学、北京百度网讯科技有限公司、武汉精测电子集团股份有限公司 、上海燧原科技股份有限公司、浪潮云信息技术股份公司、上海计算机软件技术开发中心、中国移动通信集团有限公司、京东方科技集团股份有限公司、中移（苏州）软件技术有限公司、深圳市优必选科技股份有限公司、浪潮电子信息产业股份有限公司、昆仑数智科技有限责任公司。

本文件主要起草人：刘枢、李睿宇、马珊珊、吕江波、张驰、刘凯、沈小勇、李悦、季向阳、陈鹏光、史敏锐、王泽琨、沈芷月、聂简荻、方贵明、孔维生、陶蒙华、朱知法、张璐、傅彦、崔爱香、张雅杰、赵优、黄首盛、徐诚率、王旭东、赵剑、任文奇、钟巧勇、何超、王烨东、涂小芳、吴军、王宇、严小格、张正涛、张蔚、梁秉豪、肖红梅、张艳宁、王冀、呼啸、郭世泽、梅敬青、王思善、郑佳佳、陈敏刚、曹汐、姜幸群、闫伟、刘彬、王小宏、袁立杰。

人工智能 工业视觉模型训练推理平台技术规范

1 范围

本文件确立了面向二维图像的工业AI视觉模型训练推理平台的架构，规定了功能要求、性能要求，描述了对应的检查方法。

本文件适用于工业视觉研发与应用相关机构开展面向2D图像的工业AI视觉模型训练推理平台的规划、设计、构建和应用。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 41864-2022 信息技术 计算机视觉 术语

GB/T 41867-2022 信息技术 人工智能 术语

3 术语和定义

GB/T 41864-2022 和 GB/T 41867-2022 界定的以及下列术语和定义适用于本文件。

3.1

工业视觉模型训练推理平台 industrial visual model training and inference platform

专为工业领域设计的，用于训练和优化AI视觉模型，并进行实时推理分析，以提高生产效率、质量检测和智能化水平为目标，包含数据管理、模型训练、在线推理等核心模块的智能化系统。

3.2

方案 solution

为实现特定智能化目标，由一个或多个模型及其配套参数、处理流程组成的运作架构或实施体系。

3.3

模型推理 model inference

使用训练好的机器学习模型对新数据进行分析和预测的过程，包括输入数据、推理计算和输出结果，以实现决策支持和问题解决。

4 缩略语

下列缩略语适用于本文件。

ROI: 感兴趣区域（Region of Interest）

OCR: 光学字符识别（Optical Character Recognition）

SDK：软件开发工具包（Software Development Kit）
RGB：红、绿、蓝图像格式（Red Green Blue）
RGB-D：红、绿、蓝-深度图像格式（Red, Green, Blue - Depth）
YUV：亮度-色度（Luminance-Chrominance）
CUDA：统一计算设备架构（Compute Unified Device Architecture）
cuDNN：CUDA深度神经网络（CUDA Deep Neural Network）
Mono 8/10/12：单通道像素深度8/10/12（Monochrome 8/10/12）
IoU：交集面积除以并集面积的比值（Intersection over Union）
Bayer8：基于Bayer阵列的单通道8位彩色图像格式

5 平台架构

工业视觉模型训练推理平台架构见图1。

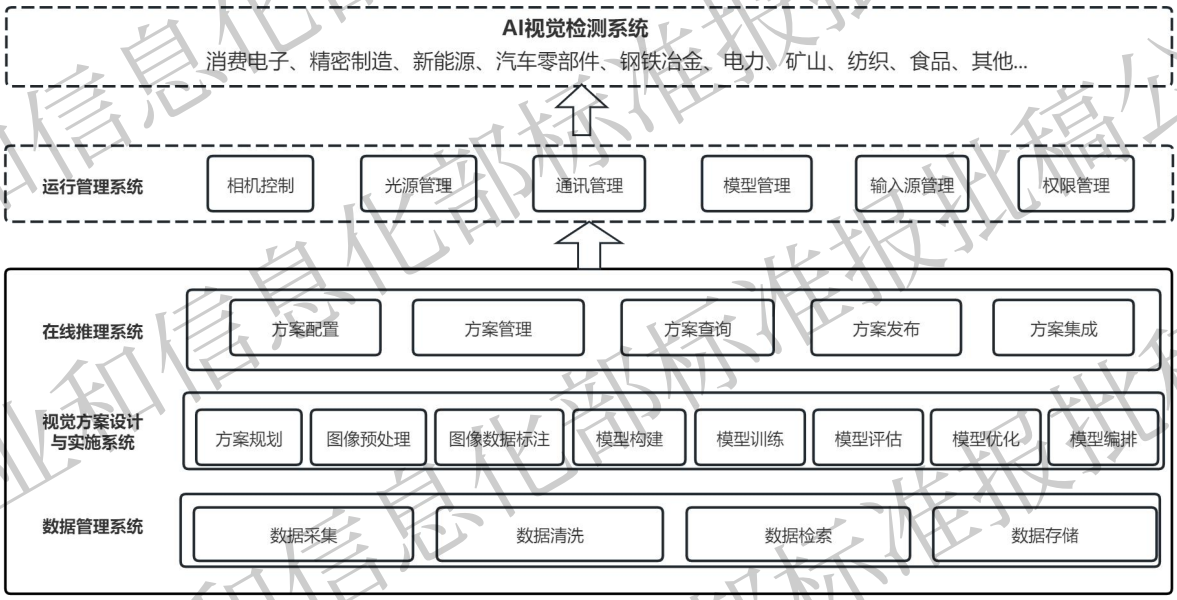

图 1　工业视觉模型训练推理平台架构

工业视觉模型训练推理平台架构见图1。平台基于分层架构设计，支持分布式部署，采用模块化管理，主要包含数据管理系统、视觉方案设计与实施系统、在线推理系统，通过运行管理系统实现场景应用；

a) 数据管理系统主要涉及机器学习算法基础数据库搭建，其功能模块包含数据采集、数据清洗、数据检索、数据存储等；
b) 视觉方案设计与实施系统主要涉及实现视觉检测业务需要具备的算法和模型训练功能，包含方案规划、图像预处理、图像数据标注、模型构建、模型训练、模型评估、模型优化、模型编排等；
c) 在线推理系统主要涉及面向不同工业业务场景的机器视觉检测需求所需的解决方案管理，包含方案配置、方案管理、方案查询、方案发布、方案集成等；
d) 运行管理系统对接工业视觉模型训练推理平台，负责相机控制、光源管理、通讯管理、模型管理、输入源管理、权限管理等，在消费电子、汽车制造等行业场景中完成技术落地。

6 平台功能要求

6.1 数据管理系统

6.1.1 数据采集

图像采集功能要求如下。

a) 图像源：
   1) 具有实时相机采集和本地图像文件输入功能；
   2) 具有网络摄像机、工业线阵相机、工业面阵相机、深度相机等一种或多种的不同类型相机的同步连接及其数据流的实时处理功能，同时兼容 GigE Vision、USB3 Vision 相机协议；
   3) 图像成像像素具备 30 万、130 万、200 万、500 万、1200 万、2500 万等一种或多种分辨率；
   4) 图像成像像素格式宜具备 Bayer8、RGB、YUV、RGB-D、Mono 8/10/12 等一种或多种格式；
   5) 图像成像文件格式宜具备 jpeg、tiff、jpg、bmp、png 等一种或多种格式；
   6) 图像帧率具有 20FPS 以上功能，匹配业务应用需求。
b) 图像窗口：
   1) 具有一个或多个图像窗口显示功能；
   2) 具有多个图像窗口内设置不同图像源、光源和图像处理方案并能灵活适应不同检测场景功能。

6.1.2 数据清洗

数据清洗功能要求如下：

a) 数据去重：去除重复数据；
b) 异常值处理：检测并处理数据集中因成像噪声、设备故障等导致的异常值，确保数据质量；
c) 数据标准化：将数据格式标准化为一致的格式，以便于处理和分析。

6.1.3 数据检索

数据检索功能要求如下：

a)数据准确性：检索结果具备相关性和准确性，以匹配用户的实际需求，避免误导和错误数据；
b)检索效率：快速响应时间，具有实时或近实时的应用场景功能，确保用户能够迅速获取所需数据；
c)可扩展性：具有灵活的扩展、保持性能和效率的功能，以适应数据量的增长和用户需求的变化。

6.1.4 数据存储

数据存储应提供数据、模型及用户信息的存储管理功能，包括但不限于基于权限的实体创建、删除、访问、编辑、导出指定的数据。

6.2 视觉方案设计与实施系统

6.2.1 方案规划

方案应根据已采集数据和项目需求，选择对应的算法，以对检测项进行智能识别。

6.2.2 图像预处理

图像预处理功能要求如下：

a) 图像尺寸调整：将图像调整为模型所需的输入尺寸；
b) 数据增强：通过对图像进行旋转、翻转、缩放、平移等操作来增加数据样本的多样性，以提高模型的泛化能力；
c) 图像数据校正：对图像进行几何校正、光学校正、色彩校正等；
d) 归一化：将图像的像素值缩放到一个固定的范围，通常是[0, 1]或[-1, 1]，以便更好地适应模型的输入要求；
e) 噪声去除：通过滤波等技术去除图像中的噪声，以提高图像质量和模型的稳定性；
f) 图像增强：通过调整图像的对比度、亮度、饱和度等参数来增强图像的特征，以提高模型的性能；
g) 图像标准化：将图像转换为特定颜色空间或表示形式的过程。例如，将彩色图像转换为灰度图像、将图像转换为特定的色彩空间（如 RGB、HSV 等）；
h) 外部自定义图像处理算法导入：宜具备自定义图像处理算法，便于平台灵活支持各类实际项目需求。

6.2.3 图像数据标注

图像数据标注功能要求如下。

a) 标注工具：
   1) 有笔刷、多边形、矩形、特征标签与带文本的矩形和多边形工具用于进行数据的标注；
   2) 标注工具产生标注实例会附带特征标签，并通过特征列表对特征标签集合进行管理，对特征标签进行新增、编辑、删除和合并；
   3) 有放大缩小功能、抓手功能、隐藏/显示标注/推理结果功能以及撤销与回复功能辅助标注；
   4) 应具备标注结果即时预览与置信度可视化，标注结果导出格式等功能，并兼容 COCO、VOC、YOLO 等图像数据集或算法模型。
b) 智能标注：

   智能标注宜具备自动识别和精确标记多类别对象、执行像素级图像分割、关键点检测，并提供强大的数据集管理及人工审核机制。
c) 标注管理：
   1) 对标注进行存储，编辑，按条件的筛选查看；导出及导入标注使得标注复用。对不同算法模型的训练应用不同的数据集进行训练；
   2) 通过编辑 ROI 区域，让算法方案只关注图片局部信息；串联模块下，ROI 区域作为前置模块输出结果的调整。

6.2.4 模型构建

模型构建功能要求如下：

a) 灵活且可扩展：平台宜具备各种图像数据源和数据类型，包括 bmp、png、jpg、jpeg、tif、tiff 等，同时具备灵活的数据处理和分析能力，具有不同的模型构建需求功能；
b) 模块化设计：平台的架构应采用模块化的设计，以便于扩展和维护。每个算法模块具有独立的功能和责任，并且模块之间的交互应该清晰和简洁；
c) 高效计算能力：平台宜具备高效的计算能力，包括高性能的硬件资源和优化的软件算法，以具有大规模数据的处理和复杂的模型构建功能；

d) 算法模型工具箱：平台应提供全面的模型工具箱，包括二分类算法、多分类算法、目标检测算法、语义分割算法、OCR 算法、无监督算法等。同时具有自定义模型构建功能，以匹配特定的业务需求；
e) 易用性：平台宜具备友好的用户界面和简洁的操作流程，以便用户方便地进行模型构建、调试和评估；
f) 可视化分析：平台提供可视化分析功能，包括数据探索、特征选择、模型评估等过程的可视化展示，以帮助用户更好地理解和解释模型结果；
g) 安全性：平台宜具备完善的安全性措施，包括数据加密、访问控制、身份鉴别、安全审计和漏洞管理等，以保护用户的数据和模型安全；
h) 集成性：平台与其他系统或工具集成，包括数据仓库、报表系统、可视化工具等，以实现更全面的模型构建和管理；
i) 技术选型：根据任务目标、数据特性、算法优缺点及计算资源等因素选择合适的模型算法、架构和硬件配置,确保模型性能最优化,匹配应用需求。

6.2.5 模型训练

模型训练功能要求如下：

a) 模型训练：使用标注数据通过优化算法调整人工智能模型的参数和策略，以使其准确地执行如分类、预测或识别等特定任务；
b) 模型推理：利用训练好的模型对指定的数据进行分析、处理和预测；
c) 模型信息查看：查看历史版本信息、训练与推理在数据集上的表现等；
d) 运行参数配置：配置模型的运行参数，如阈值、检测算法的 IOU 等，实现在不重新训练模型的前提下快速调节模型效果。

6.2.6 模型评估

模型评估功能要求如下：

a) 评估能力：平台选择已有模型和数据进行模型评估，评估要素包括安全性、稳定性、准确性、鲁棒性、泛化能力等方面；
b) 评估方法：平台针对算法模型进行不同维度的效果评估，如图像级别、实例级别、像素级别、字符级别等。

6.2.7 模型优化

模型优化功能要求如下：

a) 性能优化：提高模型的准确性、减少误差，并在不同的评估要素（如精确率、召回率、F1 分数）上优化结果；
b) 计算效率优化：减少模型所需的计算资源，提升推理速度，降低延迟，适应快速响应的应用场景；
c) 资源使用优化：优化模型的内存占用和能耗，使其在资源有限的设备上运行，如移动设备或嵌入式系统；
d) 模型简化：通过模型剪枝、量化、知识蒸馏等技术减少模型大小，便于部署；
e) 泛化能力：通过数据增强、正则化、交叉验证等策略提高模型的泛化能力，防止过拟合；
f) 自动调参：使用超参数优化算法，如网格搜索、随机搜索或贝叶斯优化，自动寻找最优的模型参数；

g) 部署能力：平台具备将模型部署至低功耗设备的能力；具备将模型部署至特殊加速框架下运行的能力。

6.2.8 模型编排

模型编排功能要求如下：

a) 管理多模型间的数据流转与交互：确保多个模型可以协同工作，正确地交换数据和信息，包括模型间的输入输出管理以及处理依赖关系；
b) 动态资源管理：根据模型的需求和系统的可用资源自动分配和优化计算资源，包括负载均衡、模型的部署和扩展；
c) 版本控制和更新：对模型进行版本管理，具有模型的平滑更新和回滚，以及实验跟踪和管理的功能；
d) 容错和异常处理：确保在某个模型发生故障或表现下降时，整个系统持续运行，通过重启、替换或绕过问题模型来保证服务的稳定性。

6.3 在线推理系统

6.3.1 方案配置

方案配置功能要求如下：

a) 具有将图像采集和图像处理相关参数保存为方案配置文件的功能；
b) 具有方案二次开发功能，包括算法前后处理、运行逻辑、结果输出等方案，帮助 AI 算法适用于实际生产需求；
c) 具有运行界面下快速选择运行方案的功能，适用于产线更换物料情况下的方案快速切换。

6.3.2 方案管理

方案管理功能要求如下：

a) 方案操作：对方案的追加、删除、更新、导出、查询等操作；
b) 方案编排：对方案进行串联或并联的编排；
c) 方案备份：定期将方案数据备份到安全的地方，以防止数据丢失。

6.3.3 方案查询

方案查询功能要求如下：

a) 支持按方案名称、创建时间、适用场景等关键字段进行模糊或精确查询；
b) 支持查询结果的排序、筛选及分页展示，便于用户快速定位目标方案；
c) 提供方案详情预览功能，包括参数配置、关联算法模型及历史运行记录等核心信息。

6.3.4 方案发布

方案发布功能要求如下：

a) 导出模型：导出模型的使用场景主要为已经构建好调用方案的用户，这些用户仅需要在各自方案中对算法模型进行更新，无需重新编译或修改调用代码；
b) 二次开发：提供的接口包含方案信息获取，方案运行结果获取，运行参数设置，结果可视化。并提供环境安装指引，示例程序，调用示例和丰富的调用文档。开发人员可快速稳定应用目标检测、图片分类、目标分割、OCR 等功能，充分覆盖多类工业视觉检测场景，匹配针对场景的定制化开发需求，让使用者更加聚焦于业务本身。

6.3.5 方案集成

方案集成功能要求如下：

a) 支持配置：SDK 具有自定义方案的集成功能；

b) 二次开发：SDK 具有 C、C#、C++、Python 中至少一种语言的二次开发功能。

7 平台性能要求

7.1 准确率

准确率不低于85%。

计算方法见公式（1）。

$$P_a = \frac{M_1+N_1}{M_1+N_1+M_2+N_2} \times 100\% \cdots\cdots (1)$$

式中：

P_a——准确率；

M_1——被正确判断为正例的样本数；

M_2——被错误判断为正例的样本数；

N_1——被正确判断为负例的样本数；

N_2——被错误判断为负例的样本数。

7.2 精确率

精确率应不低于80%。

计算方法见公式（2）。

$$P_p = \frac{M_1}{M_1+M_2} \times 100\% \cdots\cdots (2)$$

式中：

P_p——精确率；

M_1——被正确判断为正例的样本数；

M_2——被错误判断为正例的样本数。

7.3 召回率

召回率应不低于80%。

计算方法如公式（3）。

$$P_r = \frac{M_1}{M_1+N_2} \times 100\% \cdots\cdots (3)$$

式中：

P_r——召回率；

M_1——被正确判断为正例的样本数；

N_2——被错误判断为负例的样本数。

7.4 F1 分数

F1分数不应低于0.8。

计算方法如公式（4）。

$$P_f = 2 \times \frac{P_p \times P_r}{P_p + P_r} \quad \cdots\cdots (4)$$

式中：

P_f——F1分数；

P_p——精确率；

P_r——召回率。

7.5 数据传输性能

数据管理系统数据上传和下载速度均应不低于10MB/s。

7.6 数据管理系统的单张图片处理

数据管理系统的单张图片处理支持单张图片文件1G的存储、查看、编辑。

7.7 数据标注响应时间

数据标注响应时间不应超过1s、数据标注工具延时不应超过100ms。

7.8 方案管理系统

方案管理系统支持至少2个方案同时训练、推理处理，1个方案下应支持最少创建2个算法模块。

7.9 图像处理速度

处理帧率应不低于20FPS。

8 检查方法

8.1 检查条件

检查环境要求如下：

a) 操作系统：64 位桌面级操作系统，建议支持图形加速与多线程任务调度；

b) CPU：性能不低于主流高性能桌面级处理器（如 8 核心/16 线程、主频 3GHz 及以上）；

c) 内存：16GB 及以上；

d) USB：USB3.0 接口；

e) GPU：支持深度学习计算、计算能力不低于 6.0 的主流显卡，需支持张量计算与加速库；

f) 其他关键环境配置：

1）支持 CUDA 11.4 及以上功能的计算平台；

2）支持 cuDNN 8.2.4 及以上的深度学习加速库；

3）显卡驱动版本需支持上述计算平台与加速库的正常运行。

8.2 数据管理系统功能检查

8.2.1 数据采集

图像采集检查方法如下。

a) 图像源：

1) 检查图像源是否支持相机和本地图像源；

2） 检查图像源是否支持一个或多个相机连接；
3） 检查图像源是否支持选择相机触发方式；
4） 检查图像传输是否支持 USB 3.0、GIGE、CameraLink、CoaxPress 等一种或多种传输协议；
5） 检查图像成像像素是否支持 30 万、130 万、200 万、500 万、1200 万、2500 万等一种或多种分辨率；
6） 检查图像成像编码格式是否支持 Bayer8、RGB、YUV、RGB-D 等一种或多种格式；
7） 检查图像成像文件格式是否支持 jpeg、bmp、png 等一种或多种格式；
8） 检查图像成像是否支持 8 位、24 位；
9） 检查图像帧率是否匹配业务应用需求，支持 20FPS 以上。

b） 图像窗口：
1） 检查图像窗口是否支持一个或多个图像窗口显示；
2） 检查图像窗口是否支持多个图像窗口内设置不同图像源、光源和图像处理方案，能灵活适应不同检测场景。

8.2.2 数据清洗

数据清洗检查方法如下：
a） 数据去重：检查数据集中的重复数据，如根据 md5 去重；
b） 异常值处理：检查是否对不支持的格式进行异常处理，如禁止采集不支持的图像；
c） 数据标准化：检查数据格式标准化为一致的格式，如统一存储格式为 jpg。

8.2.3 数据检索

数据管理检查方法如下：
a） 选择一组已知结果的查询条件，检索相应数据，并核对返回结果的准确性，对检索结果进行人工审核，确保返回的数据符合预期；
b） 响应时间测量：使用性能检查工具（如 JMeter、LoadRunner）对系统进行压力检查，记录不同负载下的响应时间；并发检查：模拟多个用户同时进行数据检索，测量系统在高并发情况下的响应时间和处理能力；
c） 负载检查：逐步增加数据量和用户数，观察系统性能变化，评估系统的承载能力。

8.2.4 数据存储

验证系统是否能够根据用户权限正确执行数据的创建、删除、访问、编辑和导出操作，确保所有功能符合预期的存储管理要求。

8.3 视觉方案设计与实施系统功能检查

8.3.1 方案规划

方案规划检查方法如下：
a） 检查是否支持新建方案；
b） 检查是否支持删除方案；
c） 检查是否支持修改方案；
d） 检查方案中是否支持模块拖拽构建。

8.3.2 图像预处理

逐项验证图像预处理功能，确保图像尺寸调整、数据增强、校正、归一化、噪声去除、增强、标准化及自定义算法导入均符合功能要求，并有效提高模型性能和稳定性。

8.3.3 图像数据标注

数据标注检查方法如下：

a) 打开平台标注工具，选择预定目录下的数据集，点击打开选中某张待标注图像，在特征编辑区域点击右键，选择创建类别；
b) 点击创建的类别框，对类别进行编辑；
c) 选中合适的标注工具；按照标注工具说明进行标注，完成后进行保存，导出标注文件，打开导出的标注文件，检查信息是否正确。

8.3.4 模型构建

模型构建检查方法如下：

a) 灵活且可扩展：上传不同格式（BMP、PNG、JPG、JPEG、TIF、TIFF）的图像数据，验证平台能否正确解析并处理；检查自定义数据处理流程的扩展能力，如添加新的数据增强方法或特征提取方式；
b) 模块化设计：检查各算法模块是否可独立部署、更新或替换；模拟模块间通信，验证接口是否清晰、低耦合；检查新增模块时是否无需修改核心架构；
c) 高效计算能力：使用大规模数据集（如百万级图像）检查训练和推理时间，对比基准性能；监控 CPU/GPU 利用率，验证资源优化效果；压力检查下检查系统稳定性；
d) 算法模型工具箱：逐一调用二分类、多分类、目标检测、语义分割、OCR、无监督等算法，验证功能完整性；检查自定义模型接口，如加载用户定义的 PyTorch/TensorFlow 模型；
e) 易用性：招募目标用户进行可用性检查，记录完成模型构建、调试、评估的关键操作步骤数和耗时；收集用户反馈优化 UI/UX 设计；
f) 可视化分析：检查数据分布、特征重要性、混淆矩阵、ROC 曲线等可视化组件是否清晰可用；验证交互式操作（如参数调整实时更新结果）是否流畅；
g) 安全性：渗透检查验证数据加密（传输/存储）、角色权限控制（RBAC）的有效性；扫描常见漏洞（如 SQL 注入）；模拟异常访问检查审计日志完整性；
h) 集成性：检查与数据仓库（如 Hive）、报表系统（如 Tableau）、可视化工具（如 Grafana）的 API/插件集成，验证数据流和功能兼容性；
i) 技术选型：针对不同任务（如小样本分类、高分辨率检测）对比平台推荐的算法、硬件配置与实际性能（如准确率、延迟），验证选型合理性。

8.3.5 模型训练

模型训练检查方法如下：

a) 模型训练：使用不同标注数据集训练模型，验证损失下降及评估要素（如准确率）是否符合预期；
b) 模型推理：加载训练好的模型对检查集进行预测，统计结果精度（如分类准确率、检测 mAP）并检查推理耗时；
c) 模型信息查看：查看历史训练记录、评估要素及可视化图表（如 Loss 曲线），验证信息完整且可追溯；
d) 运行参数配置：动态调整参数（如阈值、IOU）并立即推理，验证新参数是否生效且结果符合预期。

8.3.6 模型评估

模型评估检查方法如下：

采用交叉验证法：将数据集 D 划分为 k 个大小相似的互斥子集，每个子集都尽可能保持数据分布的一致性。然后每次用 k-1 个子集的并集作为训练集，剩下的作为检查集。最终返回 k 个检查结果的均值，验证其在安全性、稳定性、准确性、鲁棒性和泛化能力等各项要素上的表现是否符合预期要求。

8.3.7 模型优化

模型优化检查方法如下：

a) 性能优化：使用验证集对比优化前后的模型，验证精确率、召回率和 F1 分数等关键要素是否显著提升；
b) 计算效率优化：在相同硬件环境下检查优化前后的模型，对比推理速度（FPS）和延迟时间，验证计算资源消耗是否降低；
c) 资源使用优化：在资源受限设备（如移动端）上运行优化后的模型，监测内存占用和能耗，验证是否匹配部署要求；
d) 模型简化：对剪枝/量化后的模型进行精度检查，验证模型大小减少的同时，性能下降是否在可接受范围内；
e) 泛化能力:在独立检查集上评估采用数据增强/正则化后的模型,验证其过拟合程度是否降低，泛化性能是否提升；
f) 自动调参：运行超参数优化算法（如贝叶斯优化），对比调参前后的模型性能，验证是否找到更优参数组合；
g) 部署能力：将优化后的模型部署至目标设备（如嵌入式硬件或 TensorRT 框架），验证推理速度和资源占用是否符合预期。

8.3.8 模型编排

模型编排检查方法如下：

a) 管理多模型数据流转：构建多模型协同工作流，验证数据输入/输出格式匹配性及依赖关系处理正确性，确保信息交互无误；
b) 动态资源管理：模拟高负载场景，验证系统能否自动调整计算资源分配（如 GPU/CPU 动态调度）并保持负载均衡；
c) 版本控制和更新：执行模型版本切换、回滚及 A/B 检查，验证历史版本可追溯且更新过程不影响服务可用性；
d) 容错和异常处理：主动触发模型故障（如进程终止），监测系统是否自动启用备用模型或恢复机制，保障服务不中断。

8.4 在线推理系统功能检查

8.4.1 方案配置

方案配置检查方法如下：

a) 检查方案配置应支持将图像采集和图像处理相关参数保存为方案配置文件；
b) 检查方案配置应支持方案二次开发，包括算法前后处理、运行逻辑、结果输出方案等，帮助 AI 算法适用于实际生产需求；

c) 检查方案配置应支持运行界面下快速选择运行的方案，适用于产线更换物料情况下的方案快速切换。

8.4.2 方案管理

方案管理检查方法如下：

a) 方案操作：检查方案应支持追加、删除、更新、查询操作；
b) 方案编排：检查方案应支持进行串联或并联的编排；
c) 方案备份：定期触发备份流程，并验证备份文件是否完整、可读且可成功恢复至检查环境。

8.4.3 方案查询

方案查询检查方法如下：

a) 模糊/精确查询检查：输入完整或部分方案名称、创建时间、适用场景等关键字段，验证系统能否正确返回匹配结果；
b) 排序与筛选检查：对查询结果按名称、时间等字段排序，并检查筛选功能，确保展示符合预期；
c) 分页与详情预览检查：检查分页功能是否正常，并验证方案详情（参数、关联模型等）能否正确显示。

8.4.4 方案发布

方案发布检查方法如下：

a) 导出模型：检查支持可导出模型；
b) 二次开发：检查方案发布支持提供的接口包含方案信息获取，方案运行结果获取，运行参数设置，结果可视化；
c) 检查支持提供环境安装指引，示例程序，调用示例和丰富的调用文档。

8.4.5 方案集成

检查 SDK 支持自定义方案的集成；

8.5 性能参数检查

8.5.1 准确率

使用标准检查数据集，运行系统后计算分类正确的样本数占总样本数的比例，验证该比例是否不低于 85%。

8.5.2 精确率

在标准检查数据集上，计算系统预测为正类的样本中实际为正类的比例，验证该比例是否不低于 80%。

8.5.3 召回率

在标准检查数据集上，计算实际为正类的样本中被系统正确预测为正类的比例，验证该比例是否不低于 80%。

8.5.4 F1 分数

在标准测试数据集上，根据精确率和召回率计算 F1 分数，验证该分数是否不低于 0.8。

8.5.5 数据传输性能速度检查

选择一个大文件（如1GB的文件），然后使用数据管理系统进行上传和下载测试，记录测试时间；或使用专用测试工具。

8.5.6 数据管理系统的单张图片处理检查

使用单张为1G的文件检查其存储、查看、编辑功能。

8.5.7 数据标注保存及数据标注响应时间检查

数据标注保存及数据标注响应时间检查方法如下：

a) 创建一个标注任务，包括一定数量的数据样本；
b) 使用数据标注工具进行标注，并记录每个样本的标注时间；
c) 在标注完成后，计算平均保存延时，即从标注完成到数据保存完成所花费的时间；
d) 重复上述步骤多次，并计算平均值以获取更准确的结果。

8.5.8 方案管理系统方案及算法模块并行检查

同时训练和推理处理2个方案、在一个方案下创建至少2个算法模块进行检查验证。

8.5.9 图像处理后质量检查方法

以信噪比为例，检查方法如下：

a) 选择一组具有已知信噪比的图像作为参考图像；
b) 使用图像处理算法对这些图像进行处理，并记录处理后的图像；
c) 计算处理后图像与参考图像之间的信噪比，以评估图像处理算法对噪声的影响；
d) 常用的信噪比计算方法包括峰值信噪比（PSNR）和结构相似性指数（SSIM）等。

8.5.10 图像处理速度检查

图像处理速度检查方法如下：

a) 选择一组输入图像或视频，代表实际应用场景；
b) 使用图像处理算法对每帧图像进行处理，并记录处理时间每帧；
c) 根据处理时间计算每秒处理的帧数(FPS)，图像处理帧数每秒；
d) 进行N组检查，并计算平均值。