ICS

T/GXDSL

团　　体　　标　　准

T/GXDSL —2026

人工智能图像识别算法性能测试规程

Specification for Performance Testing of Artificial Intelligence Image Recognition Algorithms

（工作组讨论稿）

（本草案完成时间：2026 - 6 - 12 ）

2026 - - 发布　　2026 - - 实施

广西电子商务企业联合会　发布

目　　次

前　言

本文件依据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。

本文件由广西产学研科学研究院提出。

本文件由广西电子商务企业联合会归口。

本文件起草单位：

本文件主要起草人：

本文件为首次发布。

人工智能图像识别算法性能测试规程

1 引言

人工智能图像识别技术是我国智能制造、智慧城市、安防交通等重点领域数字化、智能化升级的核心支撑。当前行业普遍存在测试数据集不统一、评价指标口径混乱、测试流程不规范、结果公信力不足等问题，严重制约了视觉人工智能技术创新、产业标准化建设及安全可控落地应用。为落实国家人工智能标准化发展战略，健全行业合规评价体系，本文件立足产业发展、监管合规、技术创新需求，构建分级分类、可溯源、可复现的标准化测试与评价体系。本规范用于统一行业测试标准、破除技术评价壁垒，为算法研发迭代、国产化选型、合规准入及第三方权威认定提供技术依据，助力我国视觉人工智能产业规范、高质量、安全可控发展。

2 范围

规定了人工智能图像识别算法性能测试的术语定义、测试分级分类、测试样本量化要求、核心评价指标体系、软硬件测试环境规范、标准化测试流程及分级通过准则。适用于图像分类、目标检测、图像分割三类主流计算机视觉任务人工智能算法的性能测试、量化评价、横向比对与合规判定。通用场景图像识别算法、嵌入式轻量化图像识别算法、国产化适配图像识别算法的性能测试可参照本文件执行。覆盖算法研发验证、产品市场化选型、第三方合规认定全场景，可作为行业机构、企事业单位、第三方检测机构开展图像识别算法质量评测、标准落地、合规核查的通用技术规范。

3 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件；凡是不注日期的引用文件，其最新有效版本（包含所有修改单、修订版）适用于本文件。

GB/T 41867-2022 信息技术人工智能术语

GB/T 41864-2022 信息技术计算机视觉术语

GB/T 25000.51-2016 系统与软件工程系统与软件质量要求和评价（SQuaRE）第 51 部分：就绪可用软件产品（RUSP）的质量要求和测试细则

T/BICA 067-2026 人工智能系统性能检测与评估技术要求

T/CAMETA 001078-2025 视觉检测算法性能测评规范

4 术语和定义

GB/T 41867-2022、GB/T 41864-2022 界定的术语和定义适用于本文件。除上述标准外，以下术语和定义适用于本文件。

4.1 图像识别算法

基于人工智能深度学习技术，对输入图像数据进行自动化特征提取、特征分析与模式匹配，精准输出图像类别属性、目标物体位置坐标、像素级区域分割结果的计算模型及程序实现，是支撑计算机视觉产业化应用的核心算法载体。

4.2 混淆矩阵

用于量化可视化图像识别算法预测结果与数据集真实标注的匹配关系，统计并记录真正例、假正例、真负例、假负例四类样本数量分布的矩阵结构，是算法精度、误差分析的核心基础工具。

4.3 交并比

算法预测的目标区域与图像真实标注目标区域的交集面积与并集面积的比值，是量化目标检测定位精度、图像分割像素匹配精度的核心技术参数，广泛应用于视觉算法精度评价。

4.4 测试样本量

在指定置信水平、允许误差范围内，为保障测试结果具备统计显著性、真实性与代表性，所需的最小独立测试样本数量，是规避测试偏差、保障评测公信力的基础前提。

5 缩略语

本文件适用缩略语如下：

AP：平均精度（Average Precision）

F1：F1 分数（F1-Score）

FN：假负例（False Negative）

FP：假正例（False Positive）

mAP：平均精度均值（mean Average Precision）

TN：真负例（True Negative）

TP：真正例（True Positive）

6 测试分级与任务分类

6.1 测试分级

立足我国人工智能产业研发迭代、市场应用、合规监管三级发展体系，结合测试用途、应用场景、公信力要求，将图像识别算法性能测试划分为三个等级，分级标准适配产业不同阶段发展需求：

6.1.1 一级——研发验证测试：适配算法创新研发、技术迭代优化场景，服务国内人工智能企业技术攻关与模型升级，测试样本量不少于 5000 幅图像，整体测试周期不超过 72 小时，重点验证算法基础性能与技术短板；

6.1.2 二级——产品选型测试：适配行业市场化应用、项目采购部署、国产化替代选型场景，服务各行业数字化转型落地，测试样本量不少于 20000 幅图像，测试周期不超过 120 小时，重点完成算法横向比对、商用性能核验；

6.1.3 三级——第三方质量认定测试：适配国家级、行业级合规认证、市场准入、质量评级场景，服务人工智能行业监管与标准化治理，测试样本量不少于 50000 幅图像，测试周期不超过 240 小时，且必须由具备 CNAS 国家级资质的第三方权威检测机构执行，测试结果具备行业公信力与合规效力。

6.2 任务分类与指标对应体系

结合三类核心视觉任务的产业应用特点，构建“核心指标定性能、辅助指标补短板”的分层评价体系，全面覆盖算法精度、稳定性、适应性能力，适配国家各行业应用标准要求，各任务类型对应的核心评价指标与辅助评价指标具体规定如下：

6.2.1 图像分类任务：核心评价指标为 Top-1 准确率、Top-5 准确率；辅助评价指标为精确率、召回率、F1 分数。

6.2.2 目标检测任务：核心评价指标为 mAP@[0.5:0.95]、mAP@0.5；辅助评价指标为 AP50、AP75、召回率。

6.2.3 图像分割任务：核心评价指标为 mIOU、Dice 系数；辅助评价指标为边界 F1 分数、平均绝对误差。

7 测试样本量规范

7.1 最小样本量计算依据

为保障全国范围内算法测试结果统一可比、具备统计科学性，规避小样本测试导致的结果偏差，统一采用固定公式计算最小测试样本量，满足 95%置信水平、±1%误差的国家级评测统计要求，计算公式如下：$$n = Z^2 \times p \times (1-p) / e^2$$ ……（1）式中：n —— 所需最小独立测试样本量；Z —— 置信水平对应 Z 值，统一取 1.96（对应行业通用 95%置信水平）；p —— 预估算法准确率，取 0.5（方差最大化最严苛测试场景）；e —— 可接受误差范围，取 0.01（即±1%量化误差）。经标准化计算：$$n = 1.96^2 \times 0.5 \times 0.5 / 0.01^2 = 9604$$，行业统一取整为 10000 幅，作为各类算法基础最小测试样本基准。

7.2 各任务专项样本量要求

结合国家各行业应用场景的复杂性与多样性，针对三类视觉任务制定差异化样本量标准，保障测试覆盖全面、贴合产业实际：

7.2.1 图像分类任务：单类别测试样本不少于 1000 幅，测试集类别分布均衡可控，整体分布偏差不超过±15%，规避类别不均衡导致的评测失真；

7.2.2 目标检测任务：单类别有效目标图像不少于 800 幅，单类别目标实例数量不少于 2000 个，充分覆盖不同尺寸、不同场景目标，保障检测性能评测全面性；

7.2.3 图像分割任务：单类别目标图像不少于 500 幅，像素级有效标注占比不低于总像素的 5%，满足高精度像素分割性能评测要求。

8 测试数据集标准化规范

8.1 数据集构成原则

为兼顾国际通用性与国内本土化场景适配性，适配我国各行业复杂应用环境，统一采用“基准通用样本+国内场景样本+边缘干扰样本”6:2:2 的标准化数据集构成比例，构建适配中国产业场景的标准化测试数据集体系：

8.1.1 公开基准数据集样本（60%）：采用 ImageNet、COCO、Cityscapes 等国际通用权威基准数据集，保障算法评测的国际可比性与通用性；

8.1.2 国内场景增补样本（20%）：聚焦我国智慧城市、智能制造、交通安防、民生监测等本土核心应用场景，实地采集或仿真合成样本，全面覆盖 100 lx～10000 lx 光照变化、0°～360°全角度旋转、

目标尺寸占比 0.5%～50%尺度变化等国内复杂场景特征；

8.1.3 干扰与边缘场景样本（20%）：针对国内户外复杂工况、设备拍摄误差、环境干扰等实际问题，纳入 3×3 至 15×15 梯度运动模糊、15dB～30dB 信噪比噪声、10%～70%梯度遮挡等极端场景样本，验证算法的环境鲁棒性与场景适配能力。

8.2 标注质量国家级规范

标注质量是测试结果公信力的核心基础，为统一全国评测数据精度，制定严格的标注质量管控标准目标检测任务标注框与真实目标边缘 IOU 不低于 0.95，保障定位标注高精度；图像分割任务像素级标注准确率不低于 99%，满足高精度分割评测需求；所有标注数据必须执行双人独立标注、交叉校验机制，整体标注一致率不低于 98%，杜绝人工标注误差，确保测试数据真实可信、可溯源、可复用。

9 核心测试指标体系

立足国家人工智能产业高质量发展要求，兼顾算法精度、场景鲁棒性、运行效率、国产化适配性，构建覆盖精度、稳定性、效率、场景适配性的全方位量化指标体系，分级设定国家级性能准入阈值，引领行业技术升级。

9.1 图像分类指标

9.1.1 Top-1 准确率：算法最高置信度预测类别与真实类别一致的样本占比，三级权威测试阈值不低于 95.0%，二级商用测试阈值不低于 92.0%，保障商用及合规场景基础精度；

9.1.2 Top-5 准确率：算法输出前五高置信度类别中包含真实类别的样本占比，各级测试统一阈值不低于 99.0%，适配复杂多类别识别场景；

9.1.3 F1 分数：精确率与召回率的调和均值，综合反映算法识别精准度与完整性，计算公式见式（2），各级商用及合规测试阈值不低于 0.93。$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$ ……（2）

9.2 目标检测指标

9.2.1 mAP@[0.5:0.95]：IOU 阈值 0.50 至 0.95、步长 0.05 共 10 个梯度阈值下的平均精度均值，全面量化算法全域检测精度，三级权威测试阈值不低于 72.5%；

9.2.2 mAP@0.5：IOU 阈值 0.50 条件下的平均精度，适配行业通用检测评价标准，测试阈值不低于 88.0%；

9.2.3 小目标检测能力：针对国内安防、交通等小目标高频应用场景，尺寸小于 32 像素×32 像素

的小目标 AP 值，不低于算法整体 mAP 的 70%，补齐复杂场景检测短板。

9.3 图像分割指标

9.3.1 mIOU：所有类别预测分割区域与真实分割区域的交并比均值，是像素级分割精度核心指标，计算公式见式（3），三级权威测试阈值不低于 82.0%。$$mIOU = (1/C) \times \sum (TP_i / (TP_i + FP_i + FN_i))$$ ……（3）式中：C 为总类别数，i 为类别索引；

9.3.2 Dice 系数：量化预测分割与真实区域的相似度，反映算法整体分割匹配度，测试阈值不低于 0.85；

9.3.3 类别不均衡适应性：针对工业缺陷、小众目标等稀缺场景，像素占比低于 5%的小样本类别，IOU 值不低于全类别均值的 50%，提升算法行业落地适配性。

9.4 运行效率指标

适配我国人工智能轻量化部署、嵌入式落地、规模化商用的产业需求，统一软硬件环境下的效率阈值，兼顾性能与落地实用性：

9.4.1 推理延迟：NVIDIA T4 同级 GPU 环境下单图平均推理耗时≤50ms，Intel Xeon Gold 6248 同级 CPU 环境下单图平均推理耗时≤500ms，满足实时监测场景需求；

9.4.2 吞吐量：GPU 环境下图像处理帧率≥20 帧/秒，适配规模化、高并发商用场景；

9.4.3 显存占用峰值：不超过测试硬件显存容量的 85%，保障硬件资源高效利用、适配国产化硬件部署。

10 标准化测试环境配置

为实现全国范围内测试结果统一、可复现、可比对，规避环境差异导致的评测偏差，统一规定软硬件标准化配置与环境管控要求，构建国家级统一测试环境基准。

10.1 硬件配置标准

10.1.1 通用控制硬件：CPU 主频≥2.5GHz，运行内存≥32GB，硬盘剩余存储空间≥500GB，保障测试程序稳定运行；

10.1.2 加速运算硬件：GPU 显存≥8GB（加速测试场景），基准性能测试可采用无 GPU 环境，兼顾通用与轻量化测试需求；

10.1.3 图像采集硬件：分辨率≥1920×1080，24 位真彩色，统一图像输入质量基准。

10.2 软件配置标准

统一通用、稳定、国产化适配的软件生态，保障测试环境标准化：

10.2.1 操作系统：Ubuntu 20.04 LTS、Windows 11 专业版（64 位）；

10.2.2 深度学习框架：PyTorch 1.12.0 及以上、TensorFlow 2.10.0 及以上；

10.2.3 编程语言：Python 3.8 及以上；

10.2.4 核心依赖库：OpenCV 4.5.0 及以上、NumPy 1.21.0 及以上。

10.3 环境一致性管控要求

测试全流程严禁变更软硬件配置、参数设置，杜绝环境变量干扰测试结果；正式测试前需完成不少于 100 幅样本预热，待性能指标波动幅度≤±2%后方可启动正式测试，保障运行状态稳定；全程留存软硬件环境快照、配置日志，实现测试环境全溯源、结果全复现，满足国家级检测合规要求。

11 标准化测试流程

构建“准备-执行-核验-归档”全流程标准化测试体系，明确各环节管控要求，实现全国测试流程统一、操作规范、质量可控。

11.1 测试准备阶段

11.1.1 需求梳理与方案编制：明确被测算法任务类型、应用场景、测试等级，编制专项测试方案，经技术负责人审核、归档后实施，确保测试工作合规有序；

11.1.2 数据集构建与质检：严格依据本规范样本量、构成、标注质量要求完成数据集搭建，落实双人校验机制，确保标注一致性≥98%；

11.1.3 环境搭建与验证：按标准化要求部署软硬件环境，运行验证脚本，确认所有依赖项、版本参数匹配，环境状态达标。

11.2 测试执行阶段

11.2.1 功能正确性验证：覆盖多类图像格式、尺寸适配、异常输入等场景，执行不少于 100 项功能用例，通过率 100%，保障算法基础可用性；

11.2.2 性能指标测试：按任务类型完成全部核心、辅助、效率指标测试，每项指标独立测试 3 次及以上，取算术平均值为最终结果，测试相对偏差≤±5%；

11.2.3 统计显著性检验：基于 95%置信水平完成统计检验，指标置信区间宽度≤中心值±3%，保障结果具备统计科学性。

11.3 缺陷复核与数据归档

11.3.1 将算法预测结果与标准标注偏差超阈值的样本标记为失效样本，建立缺陷台账；

11.3.2 每类别随机抽取不少于 50 幅失效样本开展人工复核，精准区分标注误差与算法性能缺陷；

11.3.2 所有失效样本、算法输出结果、测试日志全部归档留存，保存期限不少于 12 个月，满足行业监管与溯源要求。

12 分级通过准则

结合研发、商用、合规三级应用场景，制定差异化、梯度化的通过标准，适配产业不同发展阶段的质量要求，引领行业质量分级升级。

12.1 一级研发验证测试通过条件

同时满足以下全部条件，判定测试通过：所有核心指标实测值不低于本规范对应目标值的 90%；测试用例完整执行率 100%，无测试遗漏；无 P0 级阻塞性缺陷，保障算法迭代研发基础可用性。

12.2 二级产品选型测试通过条件

同时满足以下全部条件，判定测试通过：核心指标基本满足规范目标值，允许单项指标最大负偏差不超过 5%；功能正确性测试通过率≥99.5%，保障商用场景稳定运行；无 P0 级阻塞性缺陷、P1 级严重缺陷，杜绝商用落地重大风险。

12.3 三级第三方权威认定测试通过条件

同时满足以下全部条件，判定测试通过，具备行业合规公信力：所有精度、效率、稳定性指标全部达标，无指标偏差；测试全流程经 CNAS 资质第三方机构核验、确认合规；测试原始数据、脚本、日志、台账完整归档，全流程可追溯、可复现；统计显著性检验合格，置信区间宽度≤中心值±2%，结果精准可靠。

13 测试报告规范

为统一全国人工智能图像识别算法评测报告标准，保障报告规范性、权威性、可用性，测试报告必须包含以下标准化模块，可作为技术验证、产品认证、项目招投标、合规监管的官方依据：

13.1 测试基础标识

含唯一报告编号、版本号、编制日期、编制人、审核人、审批人完整签名，落实质量责任；

13.2 被测算法信息

算法名称、版本、模型参数量、输入输出规格、研发单位、国产化适配情况；

13.3 标准化测试环境

完整软硬件配置清单、系统版本、框架及依赖库版本、环境校验记录；

13.4 测试数据集信息

总样本量、分类别样本量、数据集构成比例、标注质量核验报告、场景覆盖说明；

13.5 核心测试结果

所有指标实测值、规范目标值、偏差率、达标判定结果，数据清晰可查；

13.6 失效样本分析

失效样本数量、类别分布、典型失效场景、缺陷根源分析、问题汇总统计；

13.7 测试结论与产业建议

明确测试等级与达标判定，结合国家产业应用需求，给出算法优化、场景适配、国产化升级、落地应用的针对性建议。