ICS 35.240
CCS L70

团　体　标　准

T/CESA XXXX—202X

人工智能 计算机视觉系统可信赖技术规范

Artificial intelligence - Specification for trustworthiness for computer vision systems

征求意见稿

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

已授权的专利证明材料为专利证书复印件或扉页，已公开但尚未授权的专利申请证明材料为专利公开通知书复印件或扉页，未公开的专利申请的证明材料为专利申请号和申请日期。

202X-XX- XX 发布　　202X-XX- XX 实施

中国电子工业标准化技术协会　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。

本文件由中国电子技术标准化研究院提出。

本文件由中国电子技术标准化研究院、中国电子工业标准化技术协会归口。

本文件起草单位：。

本文件主要起草人：。

人工智能 计算机视觉系统可信赖技术规范

1 范围

本文件规定了基于人工智能的计算机视觉系统的可信赖技术要求和测试方法。

本文件适用于基于人工智能的计算机视觉系统及其硬件、软件和模型等部件的可信赖设计及测试，为计算机视觉相关研制机构、测试机构、服务提供机构及用户提供参考。

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 35273－2020 信息安全技术 个人信息安全规范

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 42018－2022 信息技术 人工智能 平台计算资源规范

T/CESA 1169－2021 信息技术 人工智能 服务器系统性能测试规范

3 术语和定义

GB/T 35273－2020，GB/T 41864－2022及GB/T 41867－2022界定的以及下列术语和定义适用于本文件。

3.1

计算机视觉系统 computer vision system

具备获取、处理及解释图形、图像数据能力的计算机系统

注1：计算机视觉系统，在特定频域内，完成目标识别、目标跟踪、目标测量等任务。

注2：计算机视觉系统一般可作为其它作业系统的子系统。

注3：计算机视觉系统的功能藉由人工智能技术实现。

[参考：ISO/IEC 2382－2015,2123787]

3.2

故障 failure

<计算机视觉系统>整体或其部分无法提供或无法在限定条件下提供用户要求功能的状态。

[来源：ISO/IEC 25040－2011，4.26，有修改]

3.3

异常 exception

<计算机视觉系统>运行过程中的一类条件或现象，引起与业务逻辑过程不同的另一个处理序列。后者实现对此条件或现象的识别、忽略或处理过程。

[来源：ISO/IEC 2382－2015，2121997，有修改]

3.4

视觉功能模块 vision function module

计算机视觉系统中，单一视觉功能（图像、视频处理或分析）的实现。

注：人工智能是实现视觉功能的技术途径之一。

3.5

工作流 workflow

<计算机视觉>系统中为完成用户特定业务逻辑的必要活动的序列。

注：计算机视觉系统的工作流一般由一个或多个视觉模块组合实现。

[来源：ISO 20186.2－2019,3.28，有修改]

4 缩略语

DIIM	多样性输入迭代法 （Diverse Input Iterative Method）
FGSM	快速梯度符号法 （Fast Gradient Sign Method）
IT	信息技术 （Information Technology）
JSMA	雅可比矩阵的显着性图攻击 （Jacobian-based Saliency Map Attack）
MIM	动量迭代法 （Momentum Iterative Method）
PGD	投影梯度下降 （Projected Gradient Descent）
RISE	基于随机输入采样的解释 （Randomized Input Sampling for Explanation）
ROC	接收者操作特征曲線 （Receiver Operating Characteristic curve）
RSSLLCM	随机最不可能类方法 （Random Single Step Least-Likely Class Method）

5 概述

5.1 计算机视觉可信赖要素

对基于人工智能的计算机视觉系统的可信赖，考察系统对图1中各可信赖要素相关要求的满足程度。

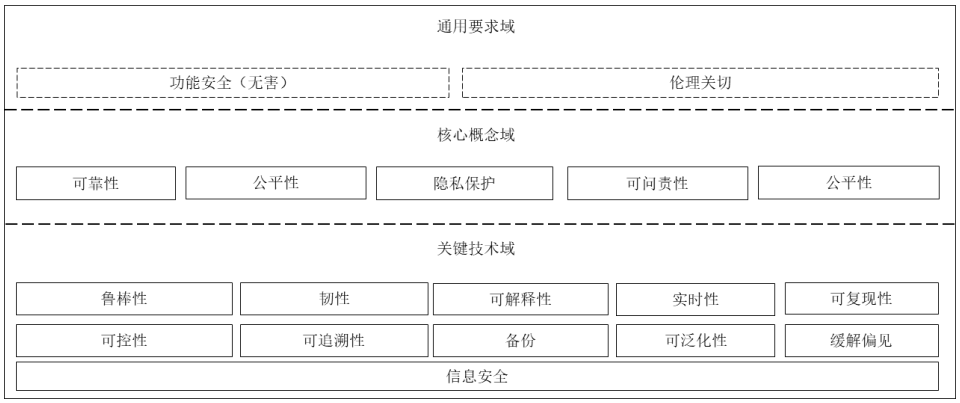

图 1 人工智能可信赖框架

5.2 技术要求的适用原则

符合本文件的计算机视觉系统，并不需满足全部可信赖技术要求。不包含某些视觉功能（如“水印”）的系统，则不必满足这些功能对应的可信赖要求。

6 可信赖概念要求

6.1 可靠性

计算机视觉系统，符合以下可靠性要求：

a) 应能对附录 B 中规定的图像、视频输入给出反馈；

b) 附录 B 中规定的图像、视频输入，应不致使系统进入异常或无法持续服务的状态；

c) 对不在附录 B 中规定的图像、视频输入，其中间或最终处理、分析结果应可用；

d) 训练过程应能保存断点，在训练因故障停止后，从断点继续训练，宜能自动重启训练。

6.2 公平性

计算机视觉系统，符合以下公平性要求：

a) 图像、视频处理、分析结果不应对个体、社群造成歧视性伤害；

注：对视觉业务需求本身要求的情况除外（如某医疗视觉业务本身要求从步态判别残障及恢复状态，则不属于歧视性伤害）。

b) 在未得到法律授权时，图像、视频处理、分析过程对同类数据应采用相同的自动决策逻辑；

c) 对特殊人群（如残障人群），宜实现特别的信息处理或传递方式，使该群体能使用系统的功能。

示例：某视觉系统识别输入图像输出对图像的解释音频，但对聋哑人群提供解释音频的文本以便阅读。

6.3 隐私保护

6.3.1 隐私数据的判定

计算机视觉系统涉及的隐私数据，包含GB/T 35273－2020中3.1和3.2规定的类型，按GB/T 35273－2020中附录A和附录B提出的方法判定。

6.3.2 隐私保护要求

计算机视觉系统隐私数据保护应满足以下要求，包含但不限于：

a) 明确隐私数据保护在数据生命周期管理内的工作和要求，形成方案体系，内容应符合法律、行政法规、部门规章和其他规范性文件的强制性要求（包括 GB/T 35273—2020 要求）。隐私数据保护的方案体系，内容应至少覆盖：

   1) 涉及隐私数据存储时，应履行告知义务、获取同意隐私数据存储请求及相应安全管理措施；

   2) 业务目标完成或图像、视频采集、处理分析计算设施更换、报废时，删除用户隐私数据并实施防恢复措施；

   3) 可再分配资源（视频、图像云空间）再分配之前，删除原有用户隐私数据并实施防恢复措施；

   4) 隐私数据留存方案应对用户透明，明确界定需要留存的用户隐私数据范围、留存时间、目的及用户告知机制。

b) 控制用户隐私数据可见范围，宜使用以下方法，包含但不限于：

   1) 主动隐私提醒；

2） 隐私符号；
3） 透明日志。

c） 控制数据关联性，宜使用解关联方法，包含但不限于：
1） K 匿名，L 多样性，T 接近性；
2） 数据屏蔽（如数据掩码、数据截断等）；
3） 随机化（如使用图像、视频加噪、置换等）；
4） 同态加密；
5） 安全多方计算。

d） 增强基本安全性，宜使用安全加密方法，包含但不限于：
1） 可搜索加密；
2） 基于身份认证的访问控制；
3） 安全加密传输。

e） 检查数据完整性，宜使用完整性检测技术，包含但不限于：
1） 数字签名；
2） 远程证明。

f） 确保对数据的可干预性，宜使用数据存取干预技术，包含但不限于：
1） 图像、视频数据溯源；
2） 细粒度访问控制；
3） 隐私风险度量；
4） 用户（图像、视频数据主体）同意机制及工具。

g） 提高用户身份标识数据安全性，宜实施数据脱敏，包含但不限于：
1） 将居民身份证、银行卡、刑事证据、用户面部、行为动作、病灶、伤残部位的图像、视频数据与用户身份标识数据分离存储；
2） 使用动态脱敏技术，确保数据从数据库中取出时已脱敏；
3） 针对用户身份标识数据设计脱敏算法和混淆冗余数据，抗重识别，满足以下要求；
- 在不慎泄露部分图像、视频数据及部分用户身份识别数据时，用户身份不被重识别；
- 在不慎泄露部分身份标识数据时，用户身份不被重识别；
- 在不慎泄露部分敏感属性数据时，用户身份不被重识别。

h） 宜采用差分隐私技术，保护训练数据中的个人信息；

i） 宜采用联邦学习，使训练数据不被泄露。

6.4 可问责性

计算机视觉系统，符合以下关于可问责性的要求：

a） 宜建立系统技术、管理方面的责任模型，明确各子系统提供者的角色和责任，至少包含：
1） 数据处理者和控制者；
2） 计算机视觉计算设施提供者，包含计算设备提供者、软件工具提供者、集成服务提供者和云计算设施提供者；
3） 视觉算法或模型提供者；
4） 应用提供者；
5） 部署、运维、运营服务提供者；
6） 用户；

注：对不同的应用，因工作逻辑的不同可导致角色及责任的差别。

b） 应建立系统设计、研发过程管理机制，其中包含计算机视觉系统相关业务被问责主体的责任及其代表；
c） 在研发过程中，应具备问责依据，即对法律、协议或授权遵从的说明；
d） 在使用过程中，应定义系统对其利益相关方行使的权能，提供供方免责声明；
e） 宜对计算机视觉系统服务过程中产生的问题进行分析，按服务重要性和技术、管理完备性，分析问题等级（越重要的服务，其技术、管理完备性越差时，问题等级越高）；
f） 计算机视觉系统或服务提供方，宜建立责任认定及应对机制，至少规定以下内容：
1）责任认定的起点（如合同违约、服务质量下降、资产损失、社会影响等）；
2）法律或合同规定的任责或免责的事项（如不可抗力）；
3）明确责任认定的发起者及方式；
4）责任认定的第三方监督方式；
g） 宜建立计算机视觉算法周期性审计制度，规定以下内容：
1）审计频次（如年审等）；
2）审计方式（如全局、局部）；
3）稽查方法（如顺查、逆查、抽查等）和使用原则；
4）记录方法及其生效和保存条件；
5）报告，其中应明确调查者与被调查者，包含范围段、意见段、签章及日期等要素；
6）审计结果的利用方法，复核方法。

6.5 透明性

计算机视觉系统，符合以下关于透明性的要求：
a） 应分析视觉处理、子系统对外部的影响，包含但不限于：
1）工作流上下游子系统；
2）外部关联系统；
b） 应披露视觉处理、子系统对外部的影响，包含但不限于a）中提出的系统能完全、部分或无法获知的影响其业务的数据及行为；
c） 应定义人工智能子系统的潜在受益、受害群体，并保证这些群体能获知完整信息；
d） 应对数据集的利益相关方，就数据使用情况（含使用方法、展示方式、周期等）建立通报机制；
e） 对涉及重识别的数据拥有者，应提供关于其数据使用、展示、维护过程等的信息；
f） 在系统部署前，应将系统行为告知利益相关方，包含但不限于：决策的效果、目的、依据和局限性；
g） 应明确告知最终用户或其他主体其正在与机器交互；
h） 应对用户在使用过程中受不公平对待的情形，提出维护权益的途径；
i） 应提供对系统所使用或生成数据的来源及变化的描述，宜提供检查手段。

7 可信赖实践要求

7.1 鲁棒性

7.1.1 图像、视频处理

计算机视觉系统图像、视频处理鲁棒性要求，包含但不限于：
a) 如系统提供编解码功能，则该功能满足以下要求：
1） 对合法的图像输入，应能输出指定格式、分辨率、位分辨率的图像数据；

2) 对合法的视频输入，应能输出指定编码格式、分辨率、位分辨率、帧速率、码率的视频数据；

b) 如系统提供去噪功能，则该功能应能识别并去除不同水平的和类型（如高斯、泊松、斑点等）的噪声；

c) 如系统提供边缘检测功能，则该功能满足以下要求：
1) 检测算法（核或算子）的输出质量不因输入图像的方向变化而改变；
2) 能识别带噪边缘，区别噪声与边缘；

d) 如系统提供图像、视频填充功能，则该功能满足以下要求：
1) 应能对含锯齿状边缘、自交多边形边缘、内或外多边形边缘的输入输出准确填充结果；
2) 具备图像完整性和逻辑性检查能力，填充结果不应出现边界判别错误、边界丢失；

e) 如系统提供拼接功能，则该功能应满足以下要求：
1) 拼接图像结果符合尺度不变性、旋转不变性；
2) 拼接图像结果的质量不因图像噪声、采集角度差异、光照差异等因素下降；
3) 拼接过程应能平滑拼接处的光照、色泽差异，不引起重影、形变；

f) 如系统提供图像、视频分割功能，对边界识别不应受光照、噪点等因素影响；

g) 系统提供水印处理功能时：
1) 人眼不可见水印（盲水印），满足以下要求：
- 应保证人眼不可见；
- 宜不因图像域变换、扰动等操作而复见；
- 应能被授权操作检出，水印类别可识别（可识别性）；
- 应能被授权操作去除，所得图像与原图像一致（可还原性）；
- 宜不因图像亮度变化、对比度变化、噪声等操作而去除或无法检出；

2) 人眼可见水印的处理，满足以下要求：
- 应保证人眼可见；
- 宜不因图像域变换、扰动等操作而消失。

7.1.2 图像、视频分析

计算机视觉系统图像、视频分析鲁棒性要求，包含但不限于：

a) 如系统提供颜色识别功能，则该功能满足以下要求：
1) 应能在不同的噪声环境下，准确标识颜色；
2) 识别过程、结果宜具备颜色恒常性；
3) 宜能在不同环境遮挡物、覆盖物、天候影响下，准确标识颜色。

b) 如系统提供物体识别功能，则该功能宜满足以下要求：
1) 不受表面其它图形、噪声、（部分）遮挡等因素影响，输出错误分类结果；
2) 不受物体光照、角度、对抗样本等因素影响，输出错误分类结果。

c) 如系统提供人脸识别功能，则识别结果准确率不应受图像灰度值的影响，不宜受角度及脸部涂鸦遮挡等因素的影响。

d) 如系统提供物体检测功能，则检测过程不应受图像灰度值、噪音等因素影响，而输出错误的物体边界；

e) 如系统提供字符识别功能，则该功能满足以下要求：
1) 应能识别不同颜色的字符；
2) 应能识别混色字符；
3) 宜能识别复杂背景中的字符；

4) 宜能识别扭曲（弯曲、非直线走向、不规则形状）的字符；
5) 应能识别大小混写字符；
6) 宜能识别失真（如移动中拍摄）字符；
7) 宜能识别非均匀光照（明暗变化、曝光过度）图像中的字符；
8) 宜能识别形态退化（灰度/清晰度退化或形态结构规划）图像中的字符。

7.2 韧性

7.2.1 概述

计算机视觉系统的韧性，包含：

a) 故障防御：系统的视觉功能具备防止外部恶意侵害的能力；
b) 故障承受：系统视觉功能失能或部分失能时，系统不降低服务质量或停止服务；
c) 系统恢复：系统能够自动检出，定位图像、视频处理、分析模块及工作流的故障，并重置故障模块，恢复其正常功能。

7.2.2 故障防御要求

计算机视觉系统，应能够防御如下攻击，包含但不限于：

a) 闪避攻击：通过修改图像、视频输入，使机器学习模型无法正确判断、决策，导致系统错误输出的行为。包含但不限于：
   1) 对抗样本攻击：对输入图像、视频帧实施微小扰动，从而使机器学习模型决策失效；
   2) 物理实体攻击：在现实物理世界中放置加入扰动之后的对抗样本图像、视频帧等，以使计算机视觉系统对物理世界认知出现偏差，导致错误决策；

注：对不涉及物理世界视觉业务的计算机视觉系统，2）不适用。

   3) 模型窃取攻击：在未知机器学习模型结构及参数的情况下，通过多次查询得出相同或近似的训练数据集，并以此训练功能类似的模型来生成对抗样本，使系统在未察觉数据、模型参数泄露的情况下遭受攻击，导致错误决策；
b) 过载攻击：产生并高频发送大量图像、视频帧，使处理、分析模块所依赖的计算软硬件发生过载、缓冲溢出、拥塞丢失输入、过热烧毁等情况，造成计算机视觉系统故障。

7.2.3 故障承受要求

在部分组件故障的情况下，计算机视觉系统应能提供视觉功能服务，宜保持故障发生前的服务水平（如预测准确率，吞吐率等）。

7.2.4 故障恢复要求

计算机视觉系统，宜具备如下故障恢复能力：

a) 能直接察觉决策错误，并确定是否存在故障；
b) 能定位故障，或配备故障定位的辅助措施；
c) 能重置故障组件，使之在业务逻辑可接受的时间范围内恢复正常工作。

7.3 可解释性

计算机视觉系统，符合以下关于可解释性的要求：

a) 应能提供事前（固有）解释（内在、设计时的解释），在投入应用前，向利益相关方总结、说明系统行为特性；

b） 应能提供事后解释，在投入使用后，向利益相关方解释特定输入与输出的逻辑关系，此逻辑关系应能被利益相关方理解；

c） 应能结合训练数据及训练过程，对模型中某区域的权值及变化给出解释，说明训练数据及其使用是如何影响模型特定区域的权值的；

d） 应能结合训练数据及训练过程，对系统表现出的行为给出解释，说明训练数据及其使用是如何影响模型行为的；

e） 宜能对基于生产环境数据的训练和推理提供a）～ d）规定的解释；

f） 宜通过可视化等方式增强视觉系统可解释性，在模型训练、推理过程中，提供对特定现象的解释。

7.4 实时性

计算机视觉系统，符合以下实时性要求：

a) 如系统提供编解码功能，则该功能应在设计要求延时范围内完成编、解码。编解码参数，按计算机视觉系统的形态，参考 GB/T 42018－2022 中，6.1.2.1 i）~ 6.1.2.1 k），6.1.2.2 e），6.1.2.2 g）~ 6.1.2.2 j）和 6.1.3 i）~ 6.1.3 j）的规定；

b) 如系统提供去噪功能，则该功能应满足以下要求：
   1) 对不同噪声功率的图像，在规定延时内输出符合设计要求；
   2) 视频去噪不降低帧率；

c) 如系统提供边缘检测功能，则该功能满足以下要求：
   1) 应在设计要求延时范围内完成检测边缘；
   2) 图像分辨率变化不宜显著增加边缘检测操作延时（延时差距不大于 10%）；

d) 如系统提供拼接功能，则该功能满足以下要求：
   1) 对给定张数、分辨率的输入图像，应在设计要求延时范围内输出拼接结果；
   2) 光照差异不宜显著影响拼接操作延时（延时差距不大于 10%）；

e) 如系统提供分割功能，则该功能满足以下要求：
   1) 应在设计要求延时范围内识别像素区域语义界线；
   2) 光照变化不宜显著影响分割操作延时（延时差距不大于 10%）。

f) 如系统提供颜色识别功能，则该功能满足以下要求：
   1) 应在设计要求延时范围内标识像素颜色；
   2) 图像中渐变阴影覆盖及其复杂度的变化不宜显著增加操作延时（延时差距不大于 10%）；

g) 如系统提供物体识别功能，则该功能应在设计要求延时范围内归类物体；

h) 如系统提供物体检测功能，则该功能应在设计要求延时范围内检测物体，输出包围框或等效指示数据。

7.5 可复现性

计算机视觉系统，应符合以下关于可复现性的要求：

i） 相同服务条件下，对相同的图像、视频输入，计算机视觉系统总输出一致的决策结果；

注 1：服务条件，包含作业工具（人工智能计算设施软硬件配置，启停状态，版本，功能）,操作账号,业务处理组件（计算机视觉系统业务逻辑软、硬件实现）,业务处理方法（模型初始化、调参、优化过程、分布式运行策略等）,运行环境条件（处理器、磁盘、内存、数据通道利用率，系统中其他作业运行状态等）。

注 2：a）中的“一致”指对于给定的测试集，不一致结果不超过总样本数量的 5%。

j） 相同服务条件下，对相同的图像、视频输入，处理、分析过程耗时和能耗一致。

注：b）中的“一致”指对于给定的测试集中的任一输入样本，耗时、能耗与该测试集上均值的差距不大于 20%。

7.6 可控性

计算机视觉系统，符合以下可控性要求：

a) 计算机视觉系统模块、工作流存在一定程度的自主能力，控制过程宜包含系统状态、过程观测、控制指令调用、系统响应、系统状态迁移及观测和状态满足度度量等过程；
b) 视觉处理模块的运行状态及其迁移控制，应满足业务权限（角色）设计要求；
c) 应具备权限控制策略，授权控制命令的执行；
d) 应具备权限控制策略，约束状态及迁移过程对于业务逻辑规定的用户群的可观测性；
e) 控制命令宜与业务逻辑指令分离传输和执行；
f) 应提供能 “终止”服务的操作；
g) 宜实现人工接管机制。

7.7 可追溯性

7.7.1 异常可追溯

计算机视觉系统应能追溯以下异常信息：

a) 逻辑异常，即与业务逻辑相关的异常，由业务程序实现定义。对逻辑异常，能至少追溯异常的描述，促使、判决系统进入异常状态的组件（或人）和异常发生时间；
b) 软件异常，即与业务逻辑无关的异常，由 IT 系统或程序本身的错误（如空指针，IO 错误等）触发。对软件异常，能至少追溯异常组件的名称，或进程标识，异常发生的缓冲区标识（如适用）和异常类型；
c) 硬件异常，即硬中断（如 stm32 硬件错误），指明硬件到达了某种暂时不可恢复的状态。对硬件异常，能至少追溯硬中断标识，物理缓冲区标识（如适用）和物理缓冲区、寄存器、扇区信息（如适用）。

7.7.2 组成可追溯

计算机视觉系统，满足以下组成可追溯要求：

a) 记录以下图像、视频处理、分析过程所使用的硬件信息及使用过程，包含但不限于：
   1) 硬件标识、名称、品牌、用途、型号、生产日期、产地、生产商、使用手册获取方法；
   2) 采购、运输、库存、交接信息；
   3) 拆封、组装、调试、安装等过程涉及的仪器、人员、时间、地点信息；
   4) 在 3）中涉及的图形驱动，工具软件等，应按 7.7.2 b）中的要求记录；
   5) 在调试、运行配置脚本时发生异常时，应按 7.7.1 的规定记录；
b) 记录以下图像、视频分析、处理软件的获取、调试和使用过程，包含但不限于：
   1) 软件名称、版本号、用途、运行环境要求信息；
   2) 配置参数、实施人员信息；
   3) 实现图像处理、分析功能的每个处理模块的名称，版本号，参数配置信息；
   4) 视觉机器学习模型或算法的部署过程调用及参数，宜包含完整性校验方法和结果；
   5) 如计算机视觉系统采用在线训练视觉模型，应记录每次训练过程的信息，包含操作组件（人）标识，训练数据标识，训练时段，新模型校验结果，部署过程（按 4）的要求记录）。
   6) 如存在调试过程，应记录调试程序版本和存放路径；
   7) 在调试过程中出现异常，按 7.7.1 的要求记录。

7.7.3 运行可追溯

计算机视觉系统，满足以下运行可追溯要求：

a) 应支持单一功能模块、工作流功能调用记录；
b) 功能调用的记录，应包含调用行为的时间戳、操作组件（人），参数值（对于图片、视频等二进制参数，需记录存取地址）和返回值；
c) 图像、视频处理模块运行的记录，应包含模块名称、版本号、输入、输出图像、视频存取地址，运行起止时间、操作人（组件）；
d) 涉及决策的模块的运行记录应包含模块名称、版本号、输入图像（或图像序列）存取地址、输出值等信息，运行起止时间戳，操作人员（组件）；
e) 对从系统外部施加的视频盗取等非授权操作，宜能追溯如下信息：
   1) 录屏、盗摄片段信息，包含视频文件标识和被录屏、盗摄片段的起止时间；
   2) 录屏、盗摄片段的流出渠道信息，包含操作者账号和片段流向目标信息；
   3) 对从网络流出的片段，流向目标应包含接收方地址；
f) 运行过程中，有异常发生时，按照 7.7.1 的要求记录异常信息。

7.7.4 维护可追溯

计算机视觉系统，满足以下维护可追溯要求：

a) 计算机视觉系统维护信息应至少包含操作人、操作过程信息（如存在自动维护过程，则应记录控制组件信息及调用过程）和维护过程的起止时间；
b) 当硬件发生变更时（新加入或移除），应按 7.7.2 a）的规定记录硬件信息。
a) 当软件发生变更时（新安装或卸载），应按 7.7.2.b）的规定记录软件信息。对新安装的软件，记录安装结果（成功或失败），对卸载的软件，记录卸载结果（成功或失败），残余数据、组件地址；
b) 数据集变更的记录，满足以下要求：
   1) 如新加入数据，供图像处理、分析机器学习模型在线学习时，应记录数据标识及存取地址；
   2) 如淘汰旧数据，应具备介质来存储淘汰数据，记录数据标识及存取地址；
   3) 如整体更换新的图像、视频数据集，应记录数据集完整性校验值及其存储过程；
c) 人工智能算法或模型变更的记录，满足以下要求：
   1) 如在线更新图像处理、分析模型，应按 7.7.2 b）5）的规定，记录信息；
   2) 应配备存储媒体，保存所移除模型或算法，并对存储媒体的访问实施控制；
d) 模块或系统启停的记录，应包含启停操作的原因，宜包含相关业务的影响范围及通告复本；
e) 模块或系统配置参数变更的记录，应满足以下要求：
   1) 配置文件被替换时，记录原配置文件及新配置文件的存取地址，完整性校验值；
   2) 配置文件被修改时，记录新的文件完整性校验值；
f) 计算机视觉系统销毁时，应记录移除软硬件的操作和去向。

7.8 备份

计算机视觉系统，宜制备备份系统，符合以下要求：

a) 备份系统实现与视觉子系统相同的业务逻辑，满足业务要求；
b) 备份系统采用与主系统不同的技术路线实现（如对基于深度学习的子系统的备份，可由非人工智能技术实现业务逻辑），在无不同技术路线备份时，可采用多模型互备；
c) 系统失能情况下，却无备份系统时，已定义人工操作流程。

7.9 可泛化性

计算机视觉系统，应符合以下关于可泛化性的要求：

a） 在研发过程中，使用至少 1 种方法增强可泛化性，包含 Dropout、数据增强（如缩放、裁剪等）、噪声（如白噪声）注入、提前停止、批量归一化或梯度下降干预（如梯度截断）等；

b） 在与测试集光照,角度,目标尺寸等因素相近但场景不同的生产环境中,系统预测准确率不低于模型在测试集上准确率的 50%。

7.10 缓解偏见

计算机视觉系统，宜符合以下关于缓解偏见行为的要求：

a) 在研发阶段，识别可能涉及的偏见类型，包含但不限于：性别，人种，民族，特殊群体（如残障）等；

b) 在研发阶段，识别可能导致偏见的原因，包含但不限于：

1) 样本分布（如在特定特征明显的图像集合上训练模型，使模型在此特征不明显的情况下获得较低预测准确率）；

2) 错误排除（如未将非显著特征作为训练输入）；

3) 不一致测量（仅使用 1 种照相或摄像设备采集图像，使模型在其他设备所采集的图像上获得较低的预测准确率）；

4) 不一致标记（采用不一致的标注尺度，标注类似的图像样本）；

5) 主观态度（标注带有一致的主观因素；

c) 在运行阶段，提供偏见结果复议机制，至少提供问题反馈渠道，包含但不限于电话、网页等。

7.11 信息安全

7.11.1 数据保护要求

计算机视觉系统训练及测试数据保护，符合以下要求，包含但不限于：

a) 不被非授权篡改：

1）应能实施数据完整性校验，隔离存储校验值；

2）应能以日志的方式记录训练数据修改操作，且不提供日志修改、删除操作，确保日志独立存储；

3）应能对训练数据的访问操作设置权限；

4）应能实施访问源过滤机制，宜能对非授权的访问源给予训练数据假象；

5）应能实施图像、视频训练数据加密，独立保存解密密钥；

6）宜能实施图像、视频训练数据同态加密，并在密文上训练；

b) 不被非授权拷贝：

1）应实施权限控制，能针对训练数据目录，最小化读权限授予对象（操作者）集合；

2）在多参与方协同训练时，应实现方案（如联邦学习），保护各方数据不被非授权获取；

3）应符合a）中3）~ 5）的要求。

c) 不被非授权检视：

1）应符合b）中1）和2）的要求；

2）应符合a）中3）~ 6）的要求。

7.11.2 模型保护要求

计算机视觉系统机器学习模型保护，符合以下要求，包含但不限于：

a) 模型不被非授权修改和拷贝：

1）应能实施模型完整性校验，隔离存储校验值；
2）应能以日志记录模型修改操作，不提供日志修改、删除操作，确保日志独立存储；
3）应能对模型的访问操作设置权限；
4）应能实施访问源过滤机制，宜对来自非授权访问给予模型假象；
5）在模型编译后，宜能转移编译前模型到独立存储空间，该空间与网络物理断开；
6）应能实施模型加密，独立保存解密密钥。

b） 模型不被窃取：
1）应能够实施访问行为控制，熔断高频大量推理任务；
2）宜能实施查询控制，识别恶意查询模式（如特定频率、类型等），降低模型被仿造的风险；

c） 宜能实施基于模型结构的防御策略，在训练中调整模型结构，降低模型输出结果对样本偏差的敏感性；

d） 宜能实施信息混淆防御，修改模型输出、参数更新等交互数据，保证模型有效性的前提下，降低信息泄露的风险。

8 计算机视觉系统可信赖测试

8.1 通则

计算机视觉系统可信赖测试的实施：

a） 由运行测试系统完成或由测试者人为评价：
1） 由运行测试系统获得的测试结果，宜由测试者（抽样）复查；
2） 由测试者人为评价获得的结果，应由第三方（或其他测试者）复查或评审；

b） 人为评价，应基于及必要证据的检查，证据包含但不限于：
3） 用户反馈或服务记录；
4） 系统组成、说明书、技术文件、应用案例、自声明或其他任何具有证明效力的文件；

c） 应按 5.2 的要求及系统的功能，选择实施测试；

d） 测试配置见附录 A。

8.2 可靠性

计算机视觉系统的可靠性测试，鲁棒性测试，参照表1的对应关系，按8.2.2规定的操作实施，测试的配置见附录A.1。系统不具备测试集时，使用8.7.2和8.7.3提出的数据集。具备时，则使用系统提供的测试集。

表 1 计算机视觉系统可靠性测试方法与技术要求的对应关系

测试类项	技术要求	测试方法
非法输入	6.1 a）	8.2.2 a），8.2.2 b）
	6.1 b）	8.2.2 a），8.2.2 b），8.2.2 c）
合法输入	6.1 c）	8.2.2 d）
视觉模型训练恢复	6.1 d）	8.2.2 e），8.2.2 f），8.2.2 g），8.2.2 h）

按以下规定，实施计算机视觉系统可靠性测试：

a) 对测试集中的任 1 样本，按附录 B，实施格式或编码格式转换、缩放、位分辨率调整等操作，生成不合法样本；
b) 将不合法样本输入计算机视觉系统或模块，观察输出或其他信息；
c) 交替输入合法及不合法样本，观察计算机视觉系统或模块的行为或反馈；
d) 对合法图像、视频输入，至少应能从视觉系统获得处理结果。按业务逻辑要求，人为评价结果的可用性；
e) 启动计算机视觉系统训练过程，观察断点保存情况，至少包含训练中间结果模型；
f) 在计算机视觉系统训练过程中，注入故障（见 8.8.3 e）~ 8.8.3 g））， 获取系统报错、警告或停止信息；
g) 设置系统训练从保存的断点重启，观察中间模型的准确率变化情况（重启之后，训练过程获得的中间模型在验证集上的准确率，不显著低于断点模型在相同验证集上的准确率（偏差 < 5%））；
h) 在系统因故障停止后，观察系统自动按断点恢复训练的情况，按 g）的规定观察准确率变化情况。

8.3 公平性

计算机视觉系统或其提供方，向测试方提供关于6.3的规定的材料，证明技术、管理过程的准备情况。

8.4 隐私数据保护

计算机视觉系统或其提供方，向测试方提供关于6.3的规定的材料，证明技术、管理过程的准备情况，对6.3.2的要求，提供实现、运行情况或规划；

8.5 可问责性

计算机视觉系统或其提供方，向测试方提供关于6.4的规定的证明材料，证明技术、管理过程的准备情况。

8.6 透明性

计算机视觉系统或其提供方，向测试方提供关于6.5的规定的证明材料，证明技术、管理过程的准备情况。

8.7 鲁棒性

鲁棒性的测试，按表2的对应关系，执行8.7.2和8.7.3规定的操作实施，测试的配置见附录A.6。如计算机视觉系统没有专用测试集，可使用8.7.2和8.7.3提出的数据集。具备专用测试集时，则使用专用测试集：

表 2 计算机视觉系统鲁棒性测试方法与技术要求的对应关系

测试类项	技术要求	测试方法
编解码	7.1.1 a）1） 7.1.1 a）2）	8.7.2 a）
去噪	7.1.1 b）	8.7.2 b），8.7.2 c），8.7.2 d）
边缘检测	7.1.1 c）1）	8.7.2 e）

	7.1.1 c） 2）	8.7.2 f）
填充	7.1.1 d） 1）	8.7.2 g）
	7.1.1 d） 2）	8.7.2 h）
拼接	7.1.1 e）1）	8.7.2 i） 1），8.7.2 i） 2）
	7.1.1 e） 2）	8.7.2 i） 3），8.7.2 i） 4）
	7.1.1 e） 3）	8.7.2 i） 5）
分割	7.1.1 f）	8.7.2 j）
水印	7.1.1 g） 1）	8.7.2 k）~ q）
	7.1.1 g） 2）	8.7.2 k）~ m）
颜色识别	7.1.2 a）1）	8.7.3 a），8.7.3 b），8.7.3 e）
	7.1.2 a）2）	8.7.3 a），8.7.3 c），8.7.3 e）
	7.1.2 a 3）	8.7.3 a），8.7.3 b），8.7.3 d），8.7.3 e）
物体识别	7.1.2 b） 1）	8.7.3 f），8.7.3 g），8.7.3 l）；8.7.3 f），8.7.3 h），8.7.3 l）
	7.1.2 b） 2）	8.7.3 f），8.7.3 c），8.7.3 l）；8.7.3 f），8.7.3 j），8.7.3 l）；8.7.3 f），8.7.3 k），8.7.3 l）
人脸识别	7.1.2 c）	8.7.3 m），8.7.3 i），8.7.3 n)；8.7.3 m），8.7.3 j），8.7.3 n)；8.7.3 m），8.7.3 d），8.7.3 n)
物体检测	7.1.2 d）	8.7.3 o），8.7.3 i），8.7.3 p），8.7.3 o），8.7.3 h），8.7.3 p）
字符识别	7.1.2 e)1）~7.1.2 e）5）	8.7.3 q），8.7.3 s）
	7.1.2 e）6）~ 7.1.2 e）8）	8.7.3 r），8.7.3 s）

按以下方法实施图像、视频处理的鲁棒性测试：

a) 向系统输入合法数据，系统应能按要求编解码，并输出正确结果；

b) 采用 Mnist 或 cifar10 数据集，各随机抽样 100 例，分别加入高斯、泊松、斑点噪声。对于每种噪声，设置 2 种强度参数配置，获得 18 个带噪数据集。噪声点的像素值为全 1（白色），参考参数如下：

1) 高斯噪声：（ $\mu = 0.0, \sigma^2 = 0.12$），（ $\mu = 0.5, \sigma^2 = 0.25$）；

2) 泊松噪声：（ $\lambda = 1, k = 5$），（ $\lambda = 4, k = 10$）；

3) 斑点噪声（负指数分布）：（ $\lambda = 0.1$），（ $\lambda = 0.3$）；

c) 运行计算机视觉系统噪声识别功能，应能对 60%的图像，正确输出噪声类型；

d) 运行计算机视觉系统噪声去除功能，对比输出图像与未加噪图像的差异（在 60%的测试样本上，接近全 0）；

e) 对特定测试集（如 BSDS500），对任一输入样本图像做旋转操作，每隔 60 度，45 度形成 1 个新样本（原样本和旋转后的样本共 13 个），输入边缘检测模块或系统，检查边缘检出情况；
f) 对 e）中描述的测试集中任一输入样本，施加 b）中提出的任意噪声，输入边缘检测模块或系统，检查边缘检出情况；
g) 选取锯齿状边缘，自交多边形边缘，内外多边形边缘的图像（如从 Mnist 数据集衍生），输入视觉系统或模块，选取与色块不同的颜色（如对于 Mnist 数据集，选取红色）实施填充（如对 Mnist 样本，填充原字符黑色所占的区域），检查填充结果（扫描填充后的图像，是否含有原字符色值像素）；
h) 随机从 cifar10 或 Imagenet 中选取特定数量（如 100）的样本，以其中前景物体为目标，使用视觉系统或模块实施填充操作（填充色可随机选取原前景中不存在的颜色值），检查填充结果（扫描填充后的图像，是否含有原图前景的任何像素上的色值）；
i) 随机从 cifar10 或 Imagenet 中选取特定数量（如 100）的样本作为集合，对其中任 1 图像，实施以下操作：
   4) 切分为两幅图像（A，B），切分线为直线，A、B 分别获得原图像的不少于 40%像素，并补充背景。对图像 A 实施等比缩放（尺度为 50%，150%），获得 A1 和 A2。使用视觉系统或模块，分别将 A1，A2，A 与 B 拼接，检查拼接结果；
   5) 对 1）中 A1，A2 和 A 实施旋转操作（角度在 0 度-90 度之间，随机选取）获得 A1’，A2’ 和 A’，使用视觉系统或模块，分别将 A1’，A2’ 和 A’与 B 拼接，检查拼接结果；
   6) 对 1）中 A1，A2 和 A 实施，施加 b）中的任何 1 种噪声，再与 B 拼接，检查拼接结果；
   7) 对 1）中 A1，A2 和 A 实施，施加亮度调整（如通过调整 c 通道），再与 B 拼接，检查拼接结果；
   8) 对 1）～4）中的结果，人工检查，是否引起肉眼能分辨的光照、色泽差异或重影、形变；
j) 随机从 Pascal VOC 2012 中选取特定数量（如 100）的样本作为集合，对其中任 1 图像，实施亮度调整（增加及减小），或添加 c）中的噪声，分别形成新的图像，对原图和新图像实施分割操作，人工检查新图像与原图像分割后的边界的一致性和正确性；
k) 使用 LVM 和 Pascal VOC 2012，选取特定数量（如 100）的样本作为集合，对 Pascal VOC 2012 中选中的图像，施加 LVM 中规定的（LVM 包含位置）水印（可见或不可见）；
l) 对 k）中的结果图像，人工检查是否肉眼可分辨；
m) 对 k）中的结果图像，添加 b）中任 1 噪声，人工检查水印是否肉眼可分辨；
n) 对 k）中的原始图像，在图内不同区域，添加相同水印，记录位置，人工检查水印是否肉眼可分辨；
o) 对 k）中的结果图像，利用视觉系统检出水印所在位置（或包围框），对比标注值；
p) 对 k）中的结果图像，利用视觉系统去除水印，将结果图像与 1）中对应原图逐像素对比；
q) 对 1）中的结果图像，实施 8.7.3 b），8.7.3 c），8.7.3 i）规定的操作，再按 o），p）的规定，检查检出和去除效果。

按以下方法实施图像、视频分析的鲁棒性测试：

a) 随机选取 100 种颜色，记录颜色值，生成纯色图像；
b) 施加 8.7.2 b）中任 1 种噪声；
c) 改变图像亮度（如增加 50%和减小 50%）；
d) 随机选取 Mnist 中的样本作为前景，将已有图像作为背景，合成新图像；
e) 识别颜色值，并与记录的颜色值对比；
f) 从 cifar10 中随机选取一定数量样本（如 100）；
g) 对每个样本，将样本作为背景，从 Mnist 中随机选取样本作为前景，合成新图像；