

数 字 山 东 工 程 标 准

SZSD 0016—2024

公共视频监控 AI 服务安全指南

Public video surveillance—Guide to AI Service Security

2024 - 04 - 15 发布

2024 - 06 - 01 实施

目 次

前言..... II

1 范围..... 3

2 规范性引用文件..... 3

3 术语和定义..... 3

4 缩略语..... 3

5 安全规则..... 3

6 平台安全..... 3

 6.1 概述..... 3

 6.2 平台自身安全功能..... 4

 6.3 AI 核心资产保护..... 4

7 模型安全..... 5

8 数据安全..... 5

 8.1 数据来源..... 5

 8.2 数据内容..... 6

 8.3 数据标注..... 6

参考文献..... 7

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。

本文件由山东省大数据局提出、归口并组织实施。

本文件起草单位：山东省大数据局、泰山信息科技有限公司、山东新一代标准化研究院有限公司、北京格灵深瞳信息技术股份有限公司、山东省计算中心（国家超级计算济南中心）、浙江大华技术股份有限公司、济南海康威视数字技术有限公司、北京市商汤科技开发有限公司、百度（中国）有限公司。

本文件主要起草人：魏淑斌、冯政、孙健勇、赵硕、魏海峰、扈玮、杨昌平、刘国振、梅岭、刘婉婷、逢锦山。

公共视频监控 AI 服务安全指南

1 范围

本文件规定了公共视频监控AI服务在安全方面的基本要求，包括安全规则、平台安全、模型安全、数据安全等内容。

本文件适用于面向公共视频监控领域提供人工智能服务的提供者，提高AI服务安全水平。

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 25069—2022 信息安全技术 术语

GB/T 35273 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

SZSD 0018—2024 公共视频监控 视图数据应用安全规范

3 术语和定义

GB/T 25069—2022、GB/T 41867—2022界定的以及下列术语和定义适用于本文件。

3.1

提供者 provider

以交互界面、可编程接口等形式面向各级公共视频监控平台，提供人工智能服务的组织或个人。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

CPU：中央处理器（Central Processing Unit）

5 安全规则

安全规则主要包括应用合法合规、功能可靠可控、数据安全可信、决策公平公正、行为可以解释、事件可以追溯。

6 平台安全

6.1 概述

AI 服务平台安全框架分为平台自身安全功能和 AI 核心资产保护两大安全组件，每一类安全组件包含多个具体安全功能，AI 服务平台安全框架见图 1。

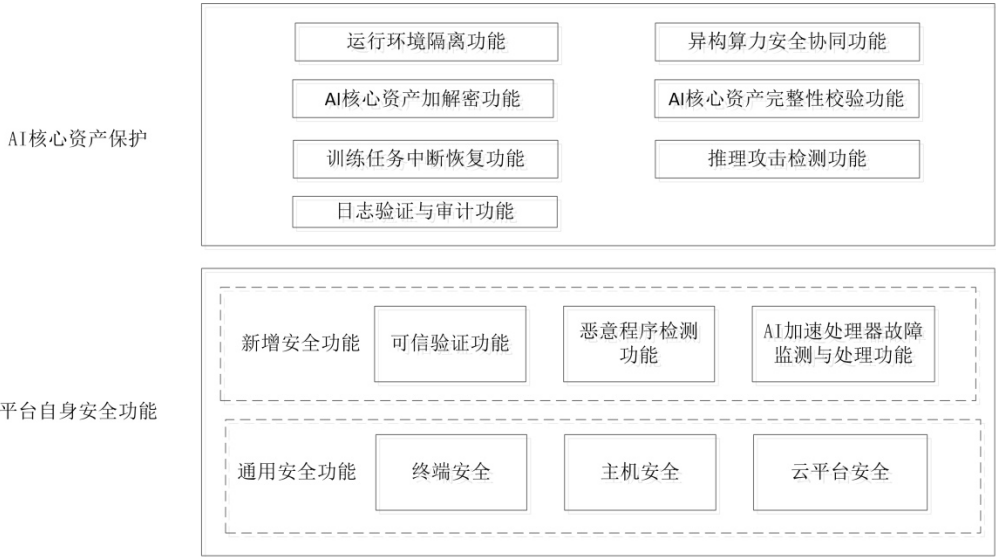


图 1 AI 服务平台安全框架

6.2 平台自身安全功能

AI 服务平台自身安全功能为平台使用方提供安全的运算环境，平台自身安全功能分为两类，其中：

- a) 通用安全功能：包括终端安全、主机安全和云平台安全，符合 SZSD 0018—2024 的第 8 章要求，削减通用场景下平台软硬件面临的安全风险；
- b) 新增安全功能：根据 AI 服务平台中新增的安全威胁及 AI 应用场景中的差异化需求，分为可信验证功能、恶意程序检测功能和 AI 加速处理器故障监测与恢复功能，各安全功能的作用如下：
 - 1) 可信验证功能：对 AI 服务平台进行完整性度量，包括 CPU 侧、AI 加速处理器侧的固件、软件等，并支撑平台使用方对度量值进行验证，以检测 AI 服务平台是否遭受篡改；
 - 2) 恶意程序检测功能：监测 AI 服务平台中为 AI 训练、推理任务运行开启的虚拟机、容器及其他用户进程运行状态，检测已知或未知的恶意程序；
 - 3) AI 加速处理器故障监测与处理功能：监测 AI 服务平台中的 AI 加速处理器资源状态，及时发现故障并进行恢复处理，保障 AI 训练/推理任务持续运行。

6.3 AI 核心资产保护

AI 核心资产保护保障 AI 核心资产传输、存储、训练、推理等各环节安全，各安全功能的作用包括但不限于：

- a) 运行环境隔离功能：为 AI 训练、推理等计算任务提供安全隔离的 CPU 资源、AI 加速运算资源、存储资源等；
- b) 异构算力安全协同功能：保障 AI 训练、推理等计算任务均在 CPU 与 AI 加速处理器协同构筑的安全隔离环境中完成，防范 AI 核心资产在 AI 训练、推理等计算任务执行过程中遭窃取、篡改；
- c) AI 核心资产加解密功能：对 AI 核心资产在不同计算节点之间流转过程中进行加解密保护，保障 AI 核心资产传输、存储过程中的保密性；

- d) AI 核心资产完整性校验功能：在 AI 核心资产传输、存储、训练、推理各环节中，对其进行完整性校验，支撑用户鉴别 AI 核心资产传输、存储过程中的完整性；
- e) 训练任务中断恢复功能：监测并识别训练资源故障，及时保存训练任务上下文状态及 AI 模型参数等信息，并在训练资源恢复后，快速恢复训练任务，防范 AI 模型参数等训练过程数据丢失；
- f) 推理攻击检测功能：计算推理请求为对抗样本、模型萃取、属性推断等攻击行为的置信度，并为 AI 应用提供响应策略；
- g) 日志验证与审计功能：为平台使用方提供集中化的日志收集与完整性校验服务，支撑其定位安全问题及审计追溯，该功能同样适用于平台自身操作产生的日志。

7 模型安全

对提供者的模型安全要求，包括但不限于：

- a) 提供者如使用基础模型进行研发，不使用未经主管部门备案的基础模型；
- b) 模型生成内容安全要求：
 - 1) 在训练过程中，将产生内容安全性作为评价生成结果优劣的主要考虑指标之一；
 - 2) 在每次训练中，对使用者输入信息进行安全性检测，引导模型生成积极正向内容；
 - 3) 对提供服务过程中以及定期检测时发现的安全问题，通过针对性的指令微调、强化学习等方式优化模型。
- c) 模型概要信息要求：
 - 1) 以交互界面提供服务的，在网站首页等显著位置向社会公开服务适用的人群、场合、用途、第三方基础模型使用情况等信息；
 - 2) 以交互界面提供服务的，在网站首页、服务协议等便于查看的位置向使用者公开服务的局限性、所使用的模型架构、训练框架等有助于使用者了解服务机制的概要信息；
 - 3) 以可编程接口形式提供服务的，在说明文档中公开序号 1 和 2 中的信息。
- d) 模型评估要求：
 - 1) 产生内容准确性要求：内容准确响应使用者输入意图，所包含的数据及表述符合科学常识或主流认知、不含错误内容；
 - 2) 产生内容可靠性要求：服务按照使用者指令给出的回复，格式框架合理、有效内容多，能够有效帮助使用者使用 AI 服务。

8 数据安全

8.1 数据来源

对提供者的数据来源安全要求，包括但不限于：

- a) 数据来源管理要求：
 - 1) 建立数据来源黑名单，不使用黑名单来源的数据进行训练；
 - 2) 对各来源数据进行安全评估，对各来源数据进行安全评估，单一来源数据内容中含不良信息的，将该来源加入黑名单。
- b) 不同来源数据搭配要求：
 - 1) 提高多样性，对每一种语言（如中文、英文等），以及每一种数据类型（如图片、视频等），都有多个数据来源；
 - 2) 合理搭配境内外来源数据。

- c) 数据来源可追溯要求：
 - 1) 使用开源数据时，具有该数据来源的开源授权协议或相关授权文件；
 - 2) 使用自采数据时，具有采集记录，不采集他人已明确声明不可采集的数据；
 - 3) 使用商业数据时，有具备法律效力的交易合同、合作协议等，交易方或合作方不能提供数据合法性证明材料时，不使用该数据；
 - 4) 将使用者输入信息当作数据时，具有使用者授权记录。

8.2 数据内容

对提供者的数据内容安全要求，包括但不限于：

- a) 数据内容过滤要求：采取分类模型、人工抽检等方式，充分过滤全部数据中不良信息；
- b) 知识产权要求，包括：
 - 1) 设置数据以及产生内容的知识产权负责人，并建立知识产权管理策略；
 - 2) 数据用于训练前，知识产权相关负责人等对数据中的知识产权侵权情况进行识别，提供者不使用有侵权问题的数据进行训练；
 - 3) 在用户服务协议中，向使用者告知生成内容使用时的知识产权相关风险，并与使用者约定关于知识产权问题识别的责任与义务。
- c) 个人信息要求，按 GB/T 35273 的相关规定，对个人信息进行保护，包括：
 - 1) 使用包含个人信息的数据时，获得对应个人信息主体的授权同意，或满足其他合法使用该个人信息的条件；
 - 2) 使用包含敏感个人信息的数据时，获得对应个人信息主体的单独授权同意，或满足其他合法使用该敏感个人信息的条件；
 - 3) 使用包含人脸等生物特征信息的数据时，获得对应个人信息主体的书面授权同意，或满足其他合法使用该生物特征信息的条件。

8.3 数据标注

对提供者的数据标注安全要求，包括但不限于：

- a) 标注人员要求：
 - 1) 自行对标注人员进行考核，给予合格者标注资质，并有定期重新培训考核以及必要时暂停或取消标注资质的机制；
 - 2) 将标注人员职能至少划分为数据标注、数据审核等，在同一标注任务下，同一标注人员不承担多项职能；
 - 3) 为标注人员执行每项标注任务预留充足、合理的标注时间。
- b) 标注规则要求：
 - 1) 标注规则至少包括标注目标、数据格式、标注方法、质量指标等内容；
 - 2) 对功能性标注以及安全性标注分别制定标注规则，标注规则至少覆盖数据标注以及数据审核等环节；
 - 3) 功能性标注规则能指导标注人员按照特定领域特点生产具备真实性、准确性、客观性、多样性的标注数据；
 - 4) 安全性标注规则能指导标注人员围绕数据及产生内容的主要安全风险进行标注。
- c) 标注内容准确性要求：
 - 1) 对安全性标注，每一条标注数据至少经由一名审核人员审核通过；
 - 2) 对功能性标注，对每一批标注数据进行人工抽检，发现内容不准确的，需重新标注。发现内容中包含不良信息的，该批次标注数据作废。

参 考 文 献

- [1] GB/T 20272—2019 信息安全技术操作系统安全技术要求
 - [2] GB/T 22239—2019 信息安全技术 网络安全等级保护基本要求
 - [3] GB/T 28181—2022 公共安全视频监控联网系统信息传输、交换、控制技术要求
 - [4] GB/T 37939—2019 信息安全技术 网络存储安全技术要求
 - [5] GB/T 39680—2020 信息安全技术 服务器安全技术要求和测评准则
 - [6] GB/T 39786—2021 信息安全技术信息系统密码应用基本要求
 - [7] GB/T 42018—2022 信息技术 人工智能平台计算资源规范
-