ICS 35 030
CCS

DB11

北　京　市　地　方　标　准

DB 11/ XXXXX—2025

信息安全 无人平台智能感知安全分析技术规范

Information security—Technical specifications for security analysis of intelligent perception for unmanned platforms

（征求意见稿）

XXXX - XX - XX 发布　　XXXX - XX - XX 实施

北京市市场监督管理局　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。

本文件由北京市公安局提出并归口。

本文件由北京市公安局组织实施。

本文件起草单位：北京市公安局、北京市公安局人工智能安全研究中心、中国标准化协会、清华大学、北京理工大学、北京瑞莱智慧科技有限公司、北京百度网讯科技有限公司。

本文件主要起草人：

信息安全 无人平台智能感知安全分析技术规范

1 范围

本文件规定了无人平台智能感知系统的安全测试操作步骤、安全分析指标和结果分析。

本文件适用于无人平台智能感知系统的基础感知能力、非理想条件下感知能力、对抗性攻击下感知能力安全分析等。

本文件适用对象包括但不限于无人平台智能感知系统研发单位、感知软硬件供应商、测试验证机构以及标准化与监管部门。在系统设计验证、安全性评估、测试验证及标准符合性评定等工作中均可参考使用。

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 5271.31—2006 信息技术 词汇 第31部分：人工智能 机器学习

3 术语和定义

GB/T 5271.31—2006界定的以及下列术语和定义适用于本文件。

3.1

无人平台 unmanned platforms

搭载多路传感器、控制器、执行器等装置，具备复杂环境感知、智能决策、协同控制等功能，可实现自主运作的系统设备。

3.2

对抗样本 adversarial examples

在数据中通过故意添加细微扰动所形成的输入样本，该样本可导致模型给出错误的输出。

4 无人平台智能感知系统构成

4.1 无人平台系统

无人平台的典型架构通常由多个子系统组成，涵盖感知、定位、规划、决策和控制等关键模块。这些模块协同工作，通过感知周围环境、识别障碍物、计算行动路径并控制无人平台实现整体功能。常见如自动驾驶汽车、无人机、机器人等。

4.2 智能感知系统

无人平台智能感知系统通常由传感器和智能感知算法构成，用于实时采集无人平台周边环境信息，包括行人、车辆、道路、建筑、交通标志及障碍物等。

传感器负责多源异构数据的实时采集，包括但不限于摄像头、激光雷达、毫米波雷达等，详见A.2。

智能感知算法对传感器获取的数据进行融合、处理与解释。算法输出结构化的感知信息，如目标类型、位置、速度、置信度等，支撑无人平台后续的决策规划与控制执行。其核心功能包括但不限于目标

检测、语义分割、路径规划与避障、环境场景分类与事件检测等，详见 A.3。

系统通常还包括感知中间件与数据接口模块，用于协调不同设备间的数据交互、时序同步与数据格式转换，保障系统的实时性与稳定性。

5 安全测试

5.1 设计测试方案

根据无人平台的具体类型及安全需求等设定需测试目标及测试方式。测试可采用仿真测试、封闭场地实验、实体平台测试等方式开展。

5.2 准备测试环境

5.2.1 搭建测试环境

搭建与无人平台实际操作环境相似的测试环境，应包含必要的硬件及软件。

5.2.2 选择传感器与智能感知算法

应准备与实际应用环境一致的传感器和智能感知算法，确保其技术规格与实际运行条件相符。应确保所选传感器已完成标定与验证，所选算法已完成训练与验证，传感器参考A.2，智能感知算法参考A.3。

5.3 构建测试数据集

收集或使用传感器构建测试数据集，并对数据集进行标注，整理数据集格式，确保其能被系统正确读取和处理。测试数据集可包含但不限于基准数据集、特定场景数据集、对抗样本数据集。

5.3.1 基准数据集

基准数据集用于测试系统完成在典型场景下的基础感知任务能力。构建基准数据集应基于无人平台在典型运行场景，不应添加扰动或补丁。数据集场景应不少于10种，每个场景数据应不少于100个。

5.3.2 特定场景数据集

特定场景数据集用于测试系统在非理想条件下的感知识别安全性。

构建本数据集应基于基准数据集环境，添加包括但不限于边界样本、极端情况样本、多源干扰、高噪声、严重遮挡等情况。干扰方法参考但不限于B.2。

每种传感数据的采集的数据中干扰类型应不少于2种，每类样本数据应不少于100个。

5.3.3 对抗样本数据集

对抗样本数据集用于测试系统在遭受对抗攻击时的安全性。

构建本数据集应基于基准数据集环境，使用攻击算法生成可通过传感器（物理空间）或识别算法接口（数字空间）输入的对抗样本数据。

攻击算法应不少于5种，每类样本数据应不少于100个。通过传感器采集的方式构建本数据集时，每类对抗样本数量应不少于10个。攻击算法参考但不限于B.3。

5.3.4 其他数据集

存在特定安全需要时，可构建其他数据集用于测试系统的其他安全性能。

构建本数据集应基于基准数据集环境，每类样本数据应不少于100个。

5.4 逐项测试

将按照 5.3 构建的数据集逐项输入到智能感知系统中，执行 3 轮或以上测试。

5.5 过程记录

记录测试过程所使用的设备的型号、系统参数、数据集组成信息及示例，每次测试的测试轮次、输入数据标签、输出结果、处理时间及异常情况。

6 结果分析

6.1 确定分析指标

根据具体应用需求，制定安全分析的指标体系，涉及4大种5小类，分别为精确率、召回率/漏检率、误检率、攻击成功率，见表1。

表 1 分析指标表

关键维度	指标	定义	安全影响
检测准确性	精确率	检测结果中正确目标的占比（TP / (TP + FP)）	低精确率意味着存在较多误检，可能导致系统错误决策。
真实目标的检测情况	召回率	真实目标中被正确检测的占比（TP / (TP + FN)）	低召回率表示存在漏检，真实目标未被识别，可能导致系统未能规避风险或及时响应，带来严重的安全隐患。
	漏检率（1−召回率）	真实目标的漏检占比（FN / (TP + FN)）	高漏检率表示系统未能识别出真实威胁或障碍物。
虚假目标的检出情况	误检率	虚假目标的误判占比（FP / (FP + TN)）	误检可能使系统对无威胁目标产生误判，浪费资源或引发误操作
抵御攻击能力	攻击成功率	攻击成功占比（成功干扰样本数 / 总样本数）	高攻击成功率说明系统在面对对抗样本时易被误导，可能威胁系统的安全性。

精确率衡量被系统判定为目标的样本中，真实为目标的比例。计算公式见式（1）。

$$Precision = \frac{TP}{TP+FP} * 100\% \quad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

式中：

Precision——检测结果中正确目标的数量占比；

TP——目标被正确判断的数量，单位为个；

FP——非目标被错误判断为目标的数量，单位为个。

召回率则是衡量所有真实目标中，被系统正确检测到的比例。计算公式见式（2）。

$$Recall = \frac{TP}{TP+FN} * 100\% \quad \cdots\cdots\cdots\cdots\cdots\cdots (2)$$

式中：

Recall——真实目标中被正确检测的数量占比；

FN——目标被错误判断为非目标的数量，单位为个。

误检率表示在所有真实为非目标的样本中，被错误检测为目标的比例。计算公式见式（3）。

$$FPR = \frac{FP}{FP+TN} * 100\% \quad \cdots\cdots\cdots\cdots\cdots\cdots (3)$$

式中：

FPR——虚假目标的误判数量占比。

漏检率表示在所有真实为目标的样本中，被漏检（未检测到）的比例。计算公式见式（4）。

$$FNR = \frac{FN}{TP+FN} * 100\% \quad \cdots\cdots\cdots\cdots\cdots\cdots (4)$$

式中：

FNR——真实目标的漏检数量占比。

攻击成功率表示数据成功干扰系统的比例。计算公式见式（5）。

$$ASR = \frac{N_{attack\_success}}{N_{attack\_total}} * 100\% \quad \cdots\cdots\cdots\cdots\cdots\cdots (5)$$

式中：

ASR——样本攻击成功数量占比；

$N_{attack\_success}$——成功使系统出错的样本数量，单位为个；

N_{attack_total}——测试时使用的样本总数量，单位为个。

被分析的智能感知算法执行目标检测任务时，宜将每一个对象类别视为一个单独的目标进行计算；执行语义分割任务时，宜将对象区域每一个像素视为一个单独目标进行计算；执行路径规划与避障、环境场景分类、事件检测等任务时，宜将每一个任务视为一个单独的目标进行计算。

6.2 撰写分析报告

根据具体应用需求，开展综合分析并撰写分析报告，报告模版参见附录C。

附 录 A
（资料性）
无人平台智能感知系统说明

A.1 概述

本附录列举了无人平台使用的常见传感器和智能感知算法。

A.2 传感器

A.2.1 摄像头

具有视频摄像/传播和静态图像捕捉等基本功能，借由镜头采集图像后，由摄像头内的感光组件电路及控制组件对图像进行处理并转换成电脑所能识别的数字信号，然后借由并行端口或USB连接输入到电脑后由软件再进行图像还原。摄像头包含：

a) 前视/前远视相机：单目/双目摄像头，通常安装在无人平台前部，主要功能为前进时的环境感知等；
b) 后视相机：采用广角摄像头，安装在无人平台尾部，主要功能为后退时的环境感知等；
c) 周视/环视：周视相机或环视鱼眼相机采用多个鱼眼镜头安装于无人平台左右侧下方以及前后侧下方四个位置，其视角较大，可以达到180°以上，图像采集后，经过图像拼接，输出无人平台周围的全景图。由于其感知范围不大，因此主要被用于无人平台周围5 m~10 m内障碍物检测等。

A.2.2 激光雷达

激光雷达是激光探测及测距系统的简称，由发射系统、接收系统、信息处理等组成，利用可见光和近红外光发射一个信号,经目标反射后被接收系统收集,通过测量反射光的运行时间而确定目标的距离。通常被安装在无人平台顶部用于探测无人平台在移动过程中的路况及障碍物。激光雷达的主要参数包括：探测距离、测距精度、线束、垂直/水平视场角（FOV）、角分辨率以及出点数等。

A.3.3 毫米波雷达

毫米波雷达的工作频段主要分布在30GHz-300GHz，实际应用中会根据国家法规和用途选择特定频段。通过对目标物体发送电磁波并接收回波来检测目标物体的距离、速度和角度。毫米波雷达通常被安装于无人平台最前端和最后端，被广泛应用于避障等。

A.2.4 麦克风

具有音频采集和声波信号转换等基本功能，通过振膜或其它敏感元件接受声波后，由麦克风内的换能组件电路及控制组件对声波信号进行处理并转换成电子设备所能识别的模拟或数字信号,然后借由音频接口（如3.5mm接口、USB或无线连接）输入到电子设备后，由软件进行信号处理、分析和还原，实现声音的录制、传输或进一步应用。麦克风主要包含：

a) 定向麦克风：采用单指向性或超指向性设计，通常安装在无人平台前部，主要功能为采集特定方向的声音信号，用于目标声源定位、语音识别、事件检测等；
b) 全向麦克风：采用全指向性设计，安装在无人平台顶部或中部，主要功能为均匀采集周围环境中的声音信号，适用于环境噪声监测或小型场景的环境声采集；

c) 麦克风阵列：由多个麦克风按特定几何结构排列组成，安装于无人平台的关键位置，通过多通道信号处理实现声源定位、噪声抑制和语音增强等功能。由于其高精度和抗干扰能力，主要被用于复杂环境下的目标检测、语音交互和场景分析等。

A.3 智能感知算法

A.3.1 目标检测

目标检测任务通常运用图像、点云等数据处理技术，并借助深度学习等人工智能技术在智能交通、自主飞行、自主行动等场景中检测识别出行人、车辆、建筑、标识、障碍物等环境信息。

A.3.2 语义分割

语义分割任务旨在为图像、点云等数据的每个像素或空间位置分配对应的语义标签，其结果是将给定图像、点云划分为若干个有意义或感兴趣的区域，便于无人平台后续对环境进行理解，例如可行驶区域检测等。

A.3.3 路径规划与避障

路径规划与避障任务是帮助无人平台在复杂环境中找到从起点到目标点的最优或可行路径，同时避开障碍物。

A.3.4 环境场景分类

环境场景分类任务是通过分析传感器采集的数据，识别和分类无人平台当前所处的环境类型，支持导航、决策和任务执行。

A.3.5 事件检测

事件检测任务是通过分析传感器采集的数据，识别和分类特定的事件，支持无人平台的决策、控制、预警等。

附 录 B
（资料性）
无人平台智能感知干扰方法及攻击算法示例

B.1 概述

本附录列举了无人平台进行智能感知安全分析涉及的常见干扰方法和常见攻击算法。

B.2 常见干扰方法简介

B.2.1 对激光雷达的干扰方式

常见干扰方式如下：

a) 将激光雷达暴露在具备相同波长的强光源下，造成雷达无法从光源方向感知物体；
b) 通过添加噪声信号使激光雷达收集虚假数据来降低传感器数据的质量，从而影响无人平台的行为决策；
c) 通过激光脉冲在不同位置创建一个点来干扰无人平台系统。

B.2.2 对摄像头的干扰方式

常见干扰方式如下：

a) 通过向摄像头发射强光而使的相机拍摄的图像过度曝光而无法被感知系统失效；
b) 将干扰光束投射在物体上，使感知系统将物体识别错误。

B.2.3 对麦克风的干扰方式

常见干扰方式如下：

a) 通过向麦克风发射高强度噪声或干扰音源，使得麦克风采集的音频信号被干扰或淹没，导致感知系统失效；
b) 使用定向声源，将特定频率或模式的声音投射到麦克风附近，使感知系统对声源定位或音频事件识别错误。

B.3 常见攻击算法简介

B.3.1 FGSM（Fast Gradient Sign Method）

通过一阶梯度信息，快速生成对抗样本。扰动通过模型损失的梯度符号乘以微小常数得到。计算速度快，适用于初步测试模型的鲁棒性。但是攻击效果有限，尤其是对高防御性的模型。

B.3.2 BIM（Basic Iterative Method）

在FGSM的基础上迭代多次，每步生成微小扰动，逐步累加。相比FGSM生成更有效的对抗样本。但是计算成本较高。

B.3.3 PGD（Projected Gradient Descent）

在BIM的基础上增加投影步骤，使生成的样本始终在可控范围内。PGD被认为是白盒攻击中的“最强”基准方法。能生成鲁棒性较强的对抗样本。对计算资源需求较大。

B.3.4 CW Attack（Carlini & Wagner Attack）

利用优化算法生成对抗样本，通过调整目标函数的优化来确保样本能够成功迷惑模型。攻击效果强且控制扰动量。优化过程复杂，计算开销大。

B.3.5 DeepFool

逐步寻找最小扰动，使样本跨越分类决策边界，达到干扰分类器的目的。生成的对抗样本扰动小。计算复杂度较高，不适合实时应用。

B.3.6 JSMA（Jacobian-based Saliency Map Attack）

利用模型的Jacobian矩阵，通过对重要特征施加扰动来生成对抗样本。可以更精准地控制特定区域的扰动。仅适用于白盒场景，且计算开销较高。

B.3.7 EAD（Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples）

结合CW攻击和Elastic Net正则化，使用优化方法生成稀疏扰动的对抗样本。攻击效果强，且可控制扰动的稀疏性。计算资源需求大。

B.3.8 STA（Spatial Transformation Attack）

通过对输入样本进行空间变换（如旋转、缩放）来生成对抗样本。对抗样本的变化具有物理可实现性。效果受限，尤其是对防御性强的模型。

B.3.9 UPGD（Unrestricted Projected Gradient Descent）

在PGD的基础上允许不受限的输入扰动，增强攻击灵活性。适合生成广义对抗样本。不可控性增加，应用场景受限。

B.3.10 NES（Natural Evolution Strategies）

使用进化算法来估计梯度，并使用这些估计值生成对抗样本。可用于模型不可访问的场景。攻击效率相对较低。

B.3.11 SPSA（Simultaneous Perturbation Stochastic Approximation）

利用随机采样估计梯度，通过有限次数的查询来实现黑盒攻击。适用于查询次数有限的场景。在复杂模型上攻击效果不佳。

B.3.12 GenAttack

基于遗传算法，通过模拟自然选择的方式生成对抗样本。灵活性高，适合无梯度信息的黑盒场景。计算代价高。

B.3.13 Transfer Attack

通过一个可访问的代理模型生成对抗样本，期望样本在目标模型上具有同样的攻击效果。可以实现跨模型攻击。攻击效果依赖于模型间的相似性。

B.3.14 Adversarial Patch

在输入图像的某一位置贴上一个对抗补丁，利用模型对局部特征的敏感性生成对抗样本。适用于实际物理世界的攻击。攻击效果容易受到模型防御策略的影响。

B.3.15 AdvGAN

利用生成对抗网络（GAN）生成对抗样本，通过对抗训练来生成更鲁棒的对抗样本。生成的对抗样本多样性强。训练成本高。

B.3.16 UAP（Universal Adversarial Perturbations）

生成对多个输入样本都有效的通用扰动，通过白盒或黑盒方法实现。可以针对多样本一次性生成对抗样本。对某些输入的攻击效果有限。

B.3.17 TFS (Towards Feature Space Adversarial Attack)

通过扰动特定的中间层特征，诱导模型对输入的误分类。可以精确控制对抗扰动的位置。在黑盒场景下不适用。

B.3.18 声源干扰攻击（Spoofing Attack）

通过伪造声源特征，干扰声源定位或声纹识别系统，或者使用声学反射器伪造声源位置，误导无人平台的声源定位系统。

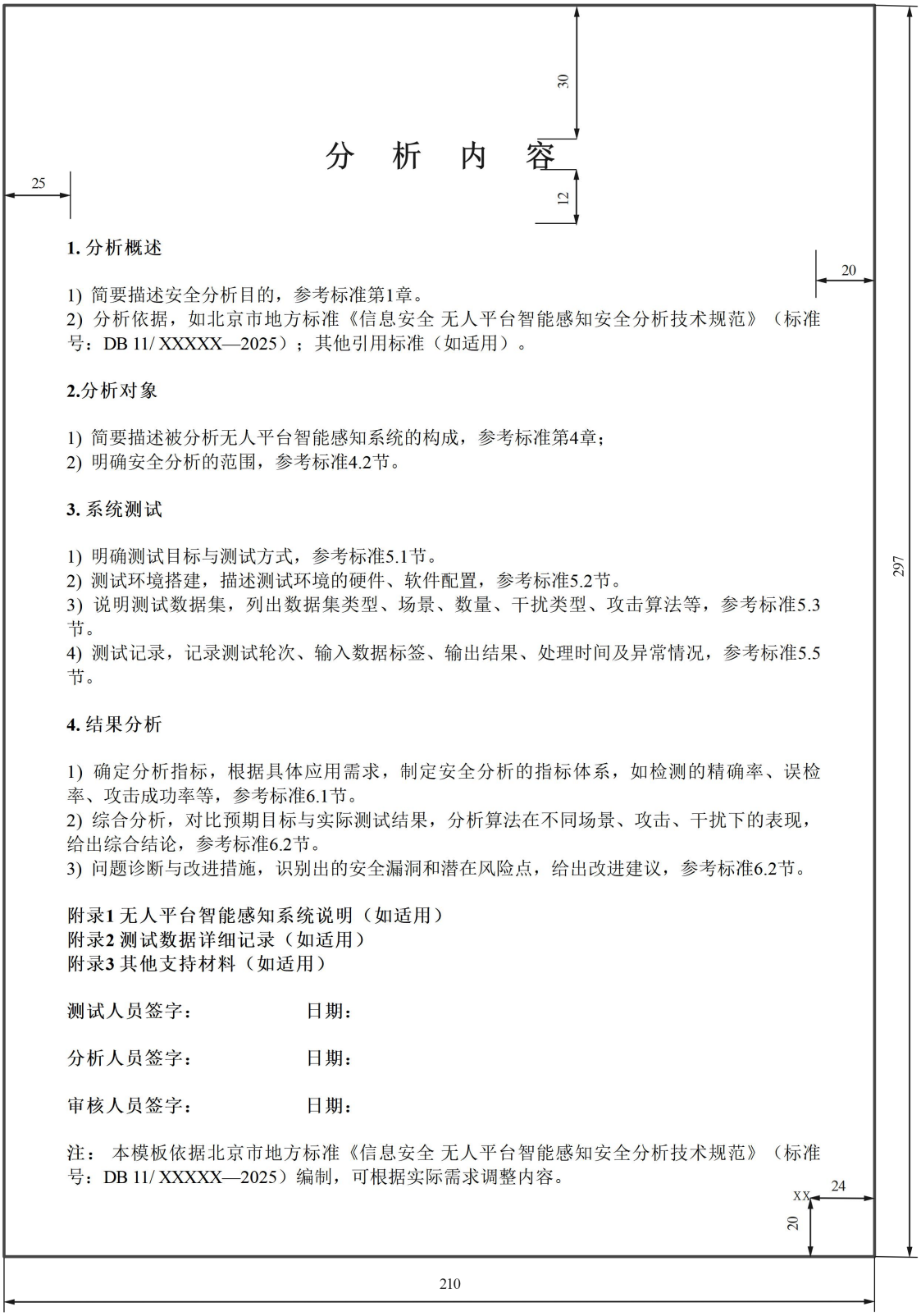
附 录 C
（资料性）
无人平台智能感知安全分析报告模板

图 C.1~图 C.3规定了报告文件的页面格式。

单位为毫米

图 C.1 封面页格式

单位为毫米



1. 分析概述

1) 简要描述安全分析目的，参考标准第1章。
2) 分析依据，如北京市地方标准《信息安全 无人平台智能感知安全分析技术规范》（标准号：DB 11/ XXXXX—2025）；其他引用标准（如适用）。

2.分析对象

1) 简要描述被分析无人平台智能感知系统的构成，参考标准第4章；
2) 明确安全分析的范围，参考标准4.2节。

3. 系统测试

1) 明确测试目标与测试方式，参考标准5.1节。
2) 测试环境搭建，描述测试环境的硬件、软件配置，参考标准5.2节。
3) 说明测试数据集，列出数据集类型、场景、数量、干扰类型、攻击算法等，参考标准5.3节。
4) 测试记录，记录测试轮次、输入数据标签、输出结果、处理时间及异常情况，参考标准5.5节。

4. 结果分析

1) 确定分析指标，根据具体应用需求，制定安全分析的指标体系，如检测的精确率、误检率、攻击成功率等，参考标准6.1节。
2) 综合分析，对比预期目标与实际测试结果，分析算法在不同场景、攻击、干扰下的表现，给出综合结论，参考标准6.2节。
3) 问题诊断与改进措施，识别出的安全漏洞和潜在风险点，给出改进建议，参考标准6.2节。

附录1 无人平台智能感知系统说明（如适用）
附录2 测试数据详细记录（如适用）
附录3 其他支持材料（如适用）

测试人员签字： 日期：

分析人员签字： 日期：

审核人员签字： 日期：

注： 本模板依据北京市地方标准《信息安全 无人平台智能感知安全分析技术规范》（标准号：DB 11/ XXXXX—2025）编制，可根据实际需求调整内容。

图 C.2 内容页格式

单位为毫米

图 C.3 附录页格式