

团体标准

T/CES XXX□XXXX

电力视觉检测算法评价方法

Evaluation method of power vision detection algorithm

(征求意见稿)

XXXX-XX-XX 发布 XXXX-XX-XX 实施

中国电工技术学会发布

目次

前言.....	II
1 范围.....	1
2 规范性引用文件.....	1
3 术语和定义.....	1
4 符号、代号和缩略语.....	2
5 评价指标与计算方法.....	2
5.1 性能指标.....	2
5.2 效率性指标.....	3
6 评价数据集的质量要求.....	3
6.1 可见光图片的质量要求.....	3
6.2 红外线图片的质量要求.....	3
6.3 紫外线图片的质量要求.....	3
7 评价方法.....	错误!未定义书签。
7.1 评价流程.....	4
7.2 评价方法.....	4
8 模型功能性等级判定.....	7
8.1 分类模型等级判定.....	8
8.2 检测模型等级判定.....	8
8.3 分割模型等级判定.....	9
附录 A（规范性）评价指标计算方法.....	10

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》给出的规则起草。

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这些专利的责任。

本文件由中国电工技术学会提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本文件起草单位：国网河北省电力有限公司电力科学研究院、中国电力科学研究院有限公司、河北工业大学、智洋创新科技股份有限公司。

本文件主要起草人：刘良帅、曾四鸣、赵建利、陈泽、霍振飞、赵建斌、赵劲康、姬艳鹏、杜晓东、冯海燕、景皓、董娜、王春璞、王晓辉、李道兴、郑碧煌、陈海永、周颖、林涛、刘坤、张东岳、龚奕霖、刘博洋、张磊、朱言庆、申朕。

本文件为首次发布。

电力视觉检测算法评价方法

1 范围

本文件规定了电力视觉检测算法的评价指标与计算方法、评价数据集的质量要求、评价方法、模型功能性等级判定。

本文件适用于对电力设备运行时含有大目标、中目标或小目标的可见光场景目标图片、红外光场景目标图片或紫外光场景目标图片进行目标分类、检测或分割的人工智能模型进行评价。

2 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，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.28 信息技术 词汇 第 28 部分：人工智能 基本概念与专家系统

AIOSS—01—2018 人工智能深度学习算法评估规范

ITU-T F.748.12: Deep learning software framework evaluation methodology

ITU-T L Suppl. 48: Data centre energy saving: Application of artificial intelligence technology in improving energy efficiency of telecommunication room and data centre infrastructure

ITU-T L Suppl.53: Guidelines on the implementation of environmental efficiency criteria for artificial intelligence and other emerging technologies.

3 术语和定义

下列术语和定义适用于本文件。

3.1 人工智能 artificial intelligence

利用数字计算机或者由数字计算机控制的机器，模拟、延伸和扩展人类的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术 and 应用系统。

3.2 人工智能模型 artificial intelligence model

通过学习海量样本数据中的内在规律和表现层次，实现包括分类、聚类、目标检测、分割等任务。

3.3 视觉检测 visual inspection

针对可见光、红外光、紫外光三种光线场景下的图像数据，使用分类、检测、分割等深度学习模型进行训练后，模型能够准确预测出图像中待检测的目标并确定其位置。

3.4 分类模型 classification model

判断一个新的观察样本所属的类别的模型。

3.5 检测模型 detection model

确定图像上目标类别和位置属性信息的模型。

3.6 分割模型 segmentation model

将图像上的目标以像素为单位标记出来的模型。

3.7 训练数据集 training set

模型构建过程中使用的数据集合。

3.8 测试数据集 testing set

评估模型构建质量的数据集合。

3.9 召回率 recall

模型正确判定的正类样本占有所有正类样本的比例。

3.10 信噪比 signal-to-noise ratio

信号均值与背景标准偏差的比值。

3.11 色差值 chromatic aberration value

表示样品颜色空间和标准品颜色空间之间的欧氏距离。

3.12 照度均匀度 illumination uniformity

最小照度值与平均照度值的比值。

3.13 热灵敏度 thermal sensitivity

红外探测器能探测到的最小温差。

3.14 电力业务场景 power business scenarios

包括电力系统发、输、配、用等全环节传统业务场景，及新兴业务场景，如储能、电碳交易等。

3.15 大目标 big object

在图像中的尺寸大于 96×96 像素的物体。

3.16 中目标 middle object

在图像中的尺寸小于 96×96 像素，并且大于 32×32 像素的物体。

3.17 小目标 small object

在图像中的尺寸小于 32×32 像素的物体。

4 符号、代号和缩略语

下列符号、代号和缩略语适用于本文件。

TP: 真正例 true positive

FP: 假正例 false positives

TN: 真负例 true negative

FN: 假负例 false negative

a_{ij} : 分割模型中标注为第 i 类，且被预测为第 j 类的样本

5 评价指标与计算方法

本文件规定了用于模型评价的指标与计算方法。

5.1 性能指标

5.1.1 精度 accuracy

评价要素内容如下：

- a) 用于目标分类模型，定义为分类正确的样本数占总样本数的比例；
- b) 表征模型分类正确水平，精度越高，模型分类越正确；
- c) 计算方法见附录 A.1.1。

5.1.2 准确率 precision

评价要素内容如下：

- a) 用于目标分类模型，定义为真正例数量占预测结果为正例的数量的比例；
- b) 表征模型分类正样本查准水平，准确率越高，模型正样本分类越准确；
- c) 计算方法见附录 A.1.2。

5.1.3 召回率 recall

评价要素内容如下：

- a) 用于目标分类模型，定义为真正例数量占样本正例数量的比例；
- b) 表征模型分类正样本查全水平，召回率越高，模型分类正样本识别越完整；
- c) 计算方法见附录 A.1.3。

5.1.4 场景准确率 scene accuracy

评价要素内容如下：

- a) 用于目标检测模型，表示模型检测结果中包含检测目标的图片数量占总图片数量的比例；
- b) 表征模型检测包含检测目标的图片水平，场景准确率越高，模型检测图片的性能越好；
- c) 计算方法见附录 A.2.1。

5.1.5 平均准确率 average precision

评价要素内容如下：

- a) 用于目标检测模型，表示以准确率为纵轴、召回率为横轴绘制的曲线下面积；
- b) 表征模型查准和查全的综合水平，平均精度越高，模型查准和查全的综合性能越好；
- c) 计算方法见附录 A.2.2。

5.1.6 平均准确率均值 mean average precision

评价要素内容如下：

- a) 用于目标检测模型，表示模型检测结果中各类标签平均精度的平均值；
- b) 表征模型查准和查全的综合水平，平均精度均值越高，模型在查准和查全的综合性能越好；
- c) 计算方法见附录 A.2.3。

5.1.7 平均交并比 mean intersection over union

评价要素内容如下：

- a) 用于目标分割模型，表示图像分割结果中各预测区域和标准区域交并比的平均值；
- b) 表征图像分割的准确程度，平均交并比越高，模型对图像分割越准确；
- c) 计算方法见附录 A.3.1。

5.2 效率性指标

5.2.1 推理速度 inference speed

评价要素内容如下：

- a) 用于所有模型，评价模型对单张图片的处理所需要的时间；
- b) 表征模型的运算速度，所需时间越少，模型对图像推理速度越快；

5.3 稳定性评价

评价要素内容如下：

a) 用于所有模型，对测试数据集加入噪声、改变亮度、平移等操作后，模型对该数据集改变前后的测试指标变化程度；

b) 表征模型的泛化能力和稳定性，两次指标相差越小，模型的泛化能力和稳定性越好；

5.4 可信性评价

5.4.1 K 折交叉验证 K-fold cross validation

评价要素内容如下：

a) 用于所有模型，评价模型结果的可信性；

b) 表征模型采用交叉验证进行训练和验证时输出的 K 个精度值，这些值的方差越小，模型的可信性越高；

6 评价数据集的质量要求

本标准规定了用于模型评价的数据集的质量要求。

6.1 可见光图片的质量要求

a) 采集的原始图像分辨率不小于 1280 像素×720 像素，主要目标大小不小于 1024 像素，横向色差小于等于 0.08；

b) 色差值小于等于 6.0；

c) 信噪比大于等于 50dB；

d) 照度均匀度大于等于 90%；

e) 可见光图片类别覆盖全体实际应用场景；

f) 每个实际应用场景中的测试图片数不少于该场景图片的 20%

6.2 红外线图片的质量要求

a) 采集的原始图像分辨率不小于 640 像素×480 像素；

b) 热灵敏度小于等于 50mk；

c) 红外线波长大于 7.5 微米，小于 14 微米；

d) 可见光图片类别覆盖全体实际应用场景；

e) 每个实际应用场景中的测试图片数不少于该场景图片的 20%

6.3 紫外线图片的质量要求

a) 采集的原始图像分辨率不小于 320 像素×240 像素；

b) 紫外线波长大于 185 纳米，小于 260 纳米；

c) 可见光图片类别覆盖全体实际应用场景；

d) 每个实际应用场景中的测试图片数不少于该场景图片的 20%

7 评价方法

本文件规定了模型评价的方法。

7.1 评价流程

评价流程应包含模型基础信息完整性评价、确定模型任务和场景、选择评价指标、选择测试数据集、样本循环测试、评价结果汇总等 7 个关键步骤，详见图 1。

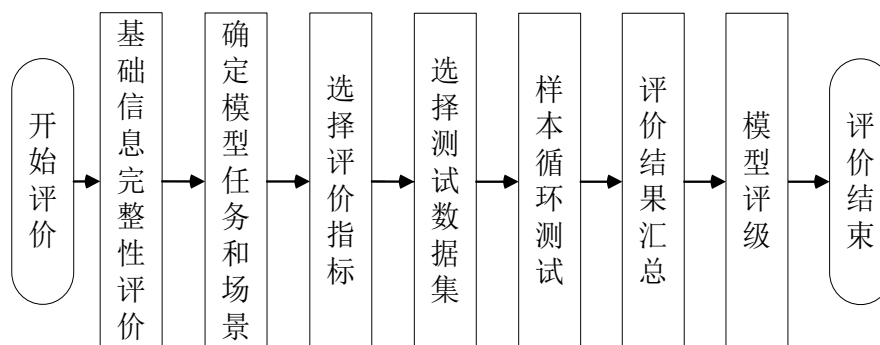


图 1 模型评价流程

7.2 评价方法

7.2.1 模型基础信息完整性评价

用于测试的模型应进行完整性评价，模型基础信息具体应包含如下内容：

- a) 模型负责人及所属单位；
- b) 模型名称；
- c) 模型光线场景，主要包括可见光场景、红外线场景、紫外线场景；
- d) 模型应用场景，描述测试目标的电力业务场景和适用目标大小；
- e) 模型任务类型，主要包括分类、目标检测和语义分割；
- f) 模型编写语言，主要包括 C/C++ 语言、Java 语言、Python 语言；
- g) 模型配置文件，描述所需库的名称、版本、依赖关系；
- h) 模型参考文献，描述模型结构主要参考的论文名称；
- i) K 折交叉验证结果，描述 K=10 时，模型在采用交叉验证的方法进行训练的结果；
- j) 模型全部代码；
- k) 模型运行指令，描述如何操作模型进行训练和测试。

7.2.2 确定模型任务及场景

确定模型任务，包括图像分类、目标检测、图像分割；

确定评价数据集的电力业务场景及光线场景，包括可见光场景、红外线场景、紫外线场景；

确定模型适用目标大小，包括大目标、中目标、小目标。

7.2.3 选择评价指标

7.2.3.1 针对可见光样本的模型性能评价指标

- a) 按照附录 A.1 选择分类模型指标；
- b) 按照附录 A.2 选择检测模型指标；
- c) 按照附录 A.3 选择分割模型指标；
- d) 评价模型对评价数据集的处理并输出运算结果所需要的时间。

7.2.3.2 针对红外样本的模型性能评价指标

- a) 按照附录 A.1 选择分类模型指标；
- b) 按照附录 A.2 选择检测模型指标；
- c) 按照附录 A.3 选择分割模型指标；
- d) 评价模型对评价数据集的处理并输出运算结果所需要的时间。

7.2.3.3 针对紫外样本的模型性能评价指标

- a) 按照附录 A.1 选择分类模型指标；
- b) 按照附录 A.2 选择检测模型指标；

- c) 按照附录 A.3 选择分割模型指标；
- d) 评价模型对评价数据集的处理并输出运算结果所需要的时间。

7.2.4 选择测试数据集

测试数据集选择要求如下：

- a) 按照 7.2.2 规定的场景下的目标选择数据集；
- b) 测试数据集应与训练数据集具有互斥性；
- c) 测试数据集样本格式参照《人工智能样本基本要求和标注规范》；
- d) 测试数据集、训练数据集的比例宜为 2：8；
- e) 测试数据集样本包含的各类别的样本数量宜相同。

7.2.5 样本循环测试

将测试数据集分成 10 等份，对模型进行多次测试，记录每次得到的指标，去除异常数值后取平均值，避免误差的干扰。

分类和检测模型功能性测试流程图如图2所示，首先将TP、TN、FP、FN初始化为0，输入图片，按行读取JSON结果，遍历所有的检出值。如果检出值数量为0，且标注数量为0，则TN+1，继续读取下一张图片；如果标注数量不为0，则FN+a（a等于标注的数量）。如果检出值数量大于0，则遍历所有检出值，如果检出值为0（即只检测到背景时），则FN+b（b等于剩余标注值数量）；如果标注为0，则FP+c（c等于剩余检出值数量）。标注值和检出值都不为0时，比对标注值的交并比，若交并比最大值大于等于0.5且标签正确，则TP+1，标签错误则FP+1，然后读取下一个检出值，继续上述过程；若交并比最大值小于0.5且标签正确，则FP+1，读取下一个检出值，继续上述过程。依此法检测并判断完所有的图片，最后，根据公式计算相应指标。

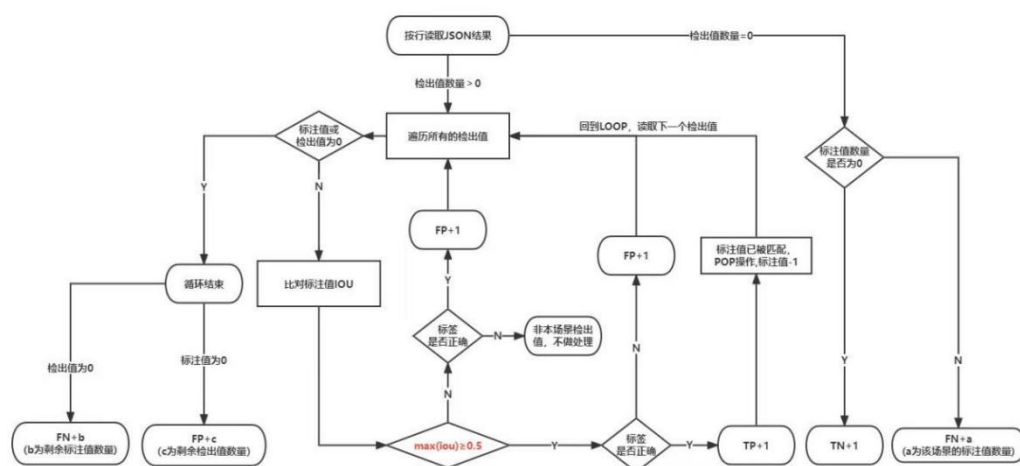


图 2 分类和检测测试流程图

分割模型功能性测试流程图如图3所示。首先将a_{ii}、a_{ij}、a_{jj}、a_{ji}初始化为0，输入图片，遍历图片上的所有像素。将分割值为第i类且标注值为第i类的像素个数记为a_{ii}，分割值为第i类且标注值为第j类的像素个数记为a_{ji}，分割值为第j类且标注值为第i类的像素个数记为a_{ij}，分割值为第j类且标注值为第j类的像素个数记为a_{jj}。更新参数，依此法检测并判断完所有的图片，并按照附录A.3.1所给公式计算平均交并比。

分类模型场景测试流程图如图 4 所示，首先输入图片，按行读取 JSON 结果，遍历所有的检出值。如果有输出值，则场景预测正确数量+1，并进行下一张图片的判断。直到所有图片测试结束，统计正确预测的数量，结束。

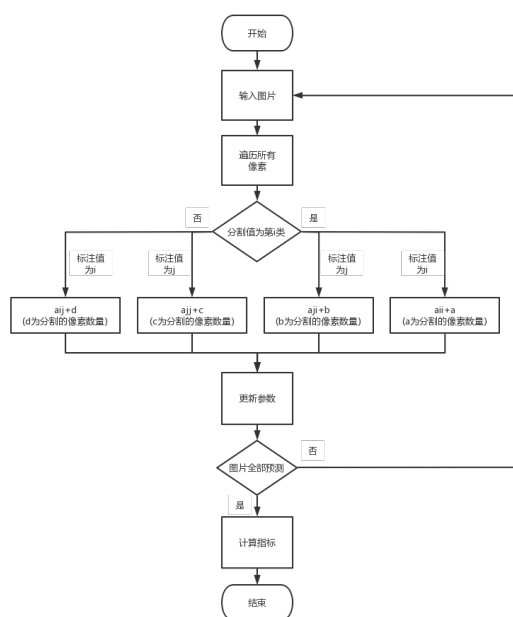


图3 分割测试流程图

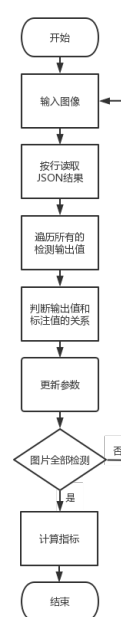


图4 分类场景测试流程图

7.2.6 评价结果汇总

被评价模型如涉及分类、检测、分割相关功能，宜优先选用本标准中模型功能性指标进行评价。评价结果包括以下内容：

模型分类性能测试。按照7.2.5给出的参数计算方法得到标签的正确预测个数TP、标签漏检FN或误检FP个数、场景预测正确个数，然后根据附录A给出的公式，计算精度、准确率、召回率和场景准确率。

模型目标检测性能测试，需要将测试结果可视化，评估人员对检测类别标签、预测框的置信度和预测框框选区域做出评价。其次按照附录A给出的公式，代入TP、FN、FP计算标签的平均准确率、平均准确率均值。

模型像素分割性能测试，对于分割模型的检测需根据附录A提供的指标进行评估，还需要将分割结果可视化，评估人员对分割区域及类别信息做出评价。

模型效率性能测试，对于所有模型，评估人员需随机挑选单张图片进行推理速度测试，并对推理时间做出评价。

模型稳定性能测试，对于所有模型，评估人员需将数据集随机地加入高斯或椒盐噪声、提高或降低亮度、进行平移，得到一份用于测试模型泛化能力和稳定性的数据集，再次执行样本循环测试，对前后两次测试得到的指标进行比较并做出评价。

模型可信性测试，对于所有模型，评估人员需记录10次样本循环测试时每次测试得到的指标，对每个指标的方差做出评价。

8 模型功能性等级判定

本标准适用于电力视觉检测算法的应用效果评价，对算法的功能性进行等级判定，实际应用中应考虑算法效率、稳定性、可信性等其他因素。

等级判定总体要求如下：

a) 模型等级对应的模型性能为：A代表模型优秀，可以直接使用；B代表模型良好，稍加修改后可以投入使用；C代表模型一般，需要提升部分性能后，再根据情况而定能否投入使用；D代表模型较差，需在大量修改后，重新对模型进行测试，再根据测试结果判定模型能否投入使用；E代表模型不合格，需要建立全新的模型。

b) 评价模型等级时，所有指标需同时满足等级要求，例如：可见光场景下的分类模型的测试结果为，场景准确率88%、精度87%、准确率86%、召回率85%，则模型等级为B；

c) 针对分类和检测模型，若指标中存在一个或两个指标低于判定阈值，则模型等级需根据情况而定，例如：红外线场景下的检测模型测试结果为，平均准确率83%、平均准确率均值79%，则模型等级为A-；可见光场景下的分类模型测试结果为，场景准确率86%、精度83%、准确率82%、召回率82%，则模型等级为B+；紫外线场景下的分类模型测试结果为，场景准确率82%、精度78%、准确率77%，召回率64%，则模型等级为B-。

d) 针对不同的电力业务场景，模型能否投入使用由该场景下的最高等级决定，例如：可见光场景下的输电通道烟雾数据集，现有的检测模型能达到的最高等级为C，则判定待测试模型时，等级达到D就能符合测试标准。

e) 针对中目标和小目标的测试数据集，模型等级的指标判定阈值分别比大目标数据集的阈值降低5%、10%

8.1 分类模型等级判定

功能性评价等级可参考以下规则：

表 1 针对大目标可见光图片的分类模型评价价值计算

指标判定	模型等级
场景准确率≥90%，精度≥90%，准确率≥90%，召回率≥90%	A
场景准确率≥85%，精度≥85%，准确率≥85%，召回率≥85%	B
场景准确率≥80%，精度≥80%，准确率≥80%，召回率≥80%	C
场景准确率≥70%，精度≥70%，准确率≥70%，召回率≥70%	D
场景准确率≥60%，精度≥60%，准确率≥60%，召回率≥60%	E

表 2 针对大目标红外线图片的分类模型评价价值计算

指标判定	模型等级
场景准确率≥80%，精度≥80%，准确率≥80%，召回率≥80%	A
场景准确率≥75%，精度≥75%，准确率≥75%，召回率≥75%	B
场景准确率≥70%，精度≥70%，准确率≥70%，召回率≥70%	C
场景准确率≥60%，精度≥60%，准确率≥60%，召回率≥60%	D
场景准确率≥50%，精度≥50%，准确率≥50%，召回率≥50%	E

表 3 针对大目标紫外线图片的分类模型评价价值计算

指标判定	模型等级
场景准确率≥80%，精度≥80%，准确率≥80%，召回率≥80%	A
场景准确率≥75%，精度≥75%，准确率≥75%，召回率≥75%	B
场景准确率≥70%，精度≥70%，准确率≥70%，召回率≥70%	C
场景准确率≥60%，精度≥60%，准确率≥60%，召回率≥60%	D
场景准确率≥50%，精度≥50%，准确率≥50%，召回率≥50%	E

8.2 检测模型等级判定

功能性评价等级可参考以下规则：

表 4 针对大目标可见光图片的检测模型评价价值计算

指标判定	模型等级
平均准确率≥90%，平均准确率均值≥90%	A
平均准确率≥85%，平均准确率均值≥85%	B
平均准确率≥80%，平均准确率均值≥80%	C
平均准确率≥70%，平均准确率均值≥70%	D
平均准确率≥60%，平均准确率均值≥60%	E

表 5 针对大目标红外线图片的检测模型评价价值计算

指标判定	模型等级
平均准确率 $\geq 80\%$ ，平均准确率均值 $\geq 80\%$	A
平均准确率 $\geq 75\%$ ，平均准确率均值 $\geq 75\%$	B
平均准确率 $\geq 70\%$ ，平均准确率均值 $\geq 70\%$	C
平均准确率 $\geq 60\%$ ，平均准确率均值 $\geq 60\%$	D
平均准确率 $\geq 50\%$ ，平均准确率均值 $\geq 50\%$	E

表 6 针对大目标紫外线图片的检测模型评价价值计算

指标判定	模型等级
平均准确率 $\geq 80\%$ ，平均准确率均值 $\geq 80\%$	A
平均准确率 $\geq 75\%$ ，平均准确率均值 $\geq 75\%$	B
平均准确率 $\geq 70\%$ ，平均准确率均值 $\geq 70\%$	C
平均准确率 $\geq 60\%$ ，平均准确率均值 $\geq 60\%$	D
平均准确率 $\geq 50\%$ ，平均准确率均值 $\geq 50\%$	E

8.3 分割模型等级判定

功能性评价等级可参考以下规则：

表 7 针对大目标可见光图片的分割模型评价价值计算

指标判定	模型等级
平均交并比 $\geq 80\%$	A
平均交并比 $\geq 75\%$	B
平均交并比 $\geq 70\%$	C
平均交并比 $\geq 60\%$	D
平均交并比 $\geq 50\%$	E

表 8 针对大目标红外线图片的分割模型评价价值计算

指标判定	模型等级
平均交并比 $\geq 80\%$	A
平均交并比 $\geq 75\%$	B
平均交并比 $\geq 70\%$	C
平均交并比 $\geq 60\%$	D
平均交并比 $\geq 50\%$	E

表 9 针对大目标紫外线图片的分割模型评价价值计算

指标判定	模型等级
平均交并比 $\geq 80\%$	A
平均交并比 $\geq 75\%$	B
平均交并比 $\geq 70\%$	C
平均交并比 $\geq 60\%$	D
平均交并比 $\geq 50\%$	E

附录 A 评价指标计算方法 (规范性)

A.1 分类模型性能指标

A.1.1 精度

精度的计算方式见公式 (A.1)：

$$A = (c_{TP} + c_{TN}) / (c_{TP} + c_{TN} + c_{FP} + c_{FN}) \quad (A.1)$$

式中：

A ——精度；

c_{TP} ——模型分类正确的正样本数量；

c_{FP} ——模型分类成正类的负样本数量；

c_{TN} ——模型分类正确的负样本数量；

c_{FN} ——模型分类成负类的正样本数量。

A.1.2 准确率

准确率的计算方式见公式 (A.2)：

$$P = c_{TP} / (c_{TP} + c_{FP}) \quad (A.2)$$

式中：

P ——准确率；

c_{TP} ——模型分类正确的正样本数量；

c_{FP} ——模型分类成正类的负样本数量。

A.1.3 召回率

召回率的计算方式见公式 (A.3)：

$$R = c_{TP} / (c_{TP} + c_{FN}) \quad (A.3)$$

式中：

R ——召回率；

c_{TP} ——模型分类正确的正样本数量；

c_{FN} ——模型分类成负类的正样本数量。

A.2 检测模型性能指标

A.2.1 场景准确率

场景准确率的计算方式见公式 (A.4)：

$$P_c = n / N \quad (A.4)$$

式中：

P_c ——场景准确率；

n ——检测出目标的图片数量；

N ——总图片数量。

A.2.2 平均准确率

平均准确率的计算方式见公式（A.5）：

$$V_{AP} = \int_0^1 p(r) dr \quad (\text{A.5})$$

式中：

V_{AP} ——平均准确率；

$p(r)$ ——以准确率为纵轴、召回率为横轴绘制的曲线函数。

A.2.3 平均准确率均值

平均准确率均值的计算方式见公式（A.6）：

$$V_{MAP} = (1/M) \sum_{i=1}^M V_{AP_i} \quad (\text{A.6})$$

式中：

V_{MAP} ——平均准确率均值；

V_{AP_i} ——第*i*类的平均准确率；

M ——类别总数。

A.3 分割模型性能指标

A.3.1 平均交并比

平均交并比的计算方式见公式（A.7）：

$$V_{MIOU} = \frac{1}{M+1} \sum_{i=0}^M \frac{a_{ii}}{\sum_{j=0}^M a_{ij} + \sum_{j=0}^M a_{ji} - a_{ii}} \quad (\text{A.7})$$

式中：

V_{MIOU} ——平均交并比；

M ——类别总数；

a_{ij} ——属于第*i*类，且被预测为第*j*类的样本数量。