



文献 CSTR:

32001.14.11-6035.noda.2025.0209.zh



文献 DOI:

10.11922/11-6035.noda.2025.0209.zh

数据 DOI:

10.57760/sciencedb.j00001.01659

文献分类: 地球科学

收稿日期: 2025-11-11

修回日期: 2026-01-04

录用日期: 2026-01-21

发表日期: 2026-06-24

* 论文通信作者

李国庆: ligq@aircas.ac.cn

基于 WorldView 数据的 2021–2022 年加拿大亚北极地区白鲸分布数据集

郑祎杰^{1,2}, 吴冰玥³, 王娅萱⁴, 李国庆^{1*}

1. 中国科学院空天信息创新研究院, 北京 100094
2. 中国科学院大学, 电子电气与通信工程学院, 北京 101408
3. 中国科学院地理科学与资源研究所, 北京 100101
4. 中国科学院大学, 未来技术学院, 北京 101408

摘要: 本研究基于 2021–2022 年获取的 WorldView-2/3 卫星影像, 构建了加拿大亚北极地区的坎伯兰湾西北部、坎伯兰湾东南部、丘吉尔河口的白鲸个体分布数据集。数据集包含三类目标标注: 白鲸、疑似白鲸与竖琴海豹。通过结合目视标注与深度学习预测的方法构建。首先使用坎伯兰湾西北部、丘吉尔河口地区影像进行目视解译, 构建目标检测训练数据集。随后用训练好的 YOLOv8 模型在坎伯兰湾东南部进行大范围预测, 并对预测结果进行人工审核与修正。最终数据集包含 721 个白鲸样本、477 个疑似白鲸样本与 551 个海豹样本的地理位置信息。本数据集可用于白鲸和竖琴海豹分布与丰度估计、目标检测算法训练与评测等研究, 对极地生态保护与旅游资源合理开发具有重要价值。

关键词: 目标检测; 白鲸; WorldView; 亚北极地区

数据库(集)基本信息简介

数据库(集)名称	基于 WorldView 数据的 2021–2022 年加拿大亚北极地区白鲸分布数据集
数据通信作者	李国庆 (ligq@aircas.ac.cn)
数据作者	郑祎杰、吴冰玥、王娅萱、李国庆
数据时间范围	2021–2022 年
地理区域	加拿大北部的亚北极地区, 包括坎伯兰湾西北部 (66°00'57"–66°42'03"N, 68°02'26"–66°39'03"W)、丘吉尔河口 (58°47'38"–58°49'19"N, 94°12'30"–94°09'05"W)、坎伯兰湾东南部 (63°50'35"–64°26'35"N, 64°48'25"–64°29'16"W)。
数据量	95 KB
数据格式	*.shp
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01659
基金项目	第四届 IEEE 地球科学与遥感学会学生挑战赛; 河南省重大科技专项 (E1D7180200)。

数据库（集）组成

本数据集包含4个目录，分别存储了不同地区和时间的白鲸分布数据文件：(1) northwestern_Cumberland_Sound_2021.shp是坎伯兰湾西北部2021年的白鲸分布数据，包含196只白鲸，215只疑似白鲸，526只竖琴海豹；(2) northwestern_Cumberland_Sound_2022.shp是坎伯兰湾西北部2022年白鲸分布数据，包含103只白鲸，79只疑似白鲸，16只竖琴海豹；(3) Churchill_River_2022.shp是丘吉尔河口2022年白鲸分布数据，包含378只白鲸和113只疑似白鲸；(4) southeastern_Cumberland_Sound_2021.shp是坎伯兰湾东南2021年白鲸分布数据，包含44只白鲸，70只疑似白鲸，9只竖琴海豹。每个文件中，label字段的值为0、1、2，分别代表白鲸、疑似白鲸、竖琴海豹。

引言

全球气候变化、海洋污染及日益频繁的人类活动正持续对海洋生态环境构成威胁，使包括白鲸（*Delphinapterus leucas*）在内的多种海洋哺乳动物的生存面临严峻挑战^[1]。作为北极与亚北极地区生态系统中的关键物种，白鲸种群数量的变动不仅反映其自身的生存状态，也在一定程度上指示了区域生态系统的健康状况。因此，准确、高效地监测白鲸的分布与丰度，对于生态保护、渔业管理、科学研究和相关政策制定均具有重要意义^[2]。

传统的白鲸调查方法，如船舶观测与无人机监测，其覆盖范围有限，且存在干扰野生动物自然行为的潜在风险。近年来，超高分辨率卫星影像技术的发展，为大范围、非侵入式的海洋哺乳动物种群监测提供了新的有效途径^[3-4]。

目前已有的白鲸分布数据集较少，且主要依赖专家标注，成本较高。Charry 等人采集了 2019 年坎伯兰湾西北部的 WorldView-3 影像，由两名专家标注者进行目视标注，共识别出 292 条白鲸和 86 条疑似白鲸^[5]。Cortney 等人采集了东高北极 - 巴芬湾地区 3271 km² 范围的 WorldView-2、WorldView-3、GeoEye-1 影像，由众包标注员对影像中白鲸进行人工标注，但其标注精度（准确率 71%，召回率 65%）难以满足科研需求^[6]。

近年来，以 YOLO^[7] 为代表的目标检测算法的进步，为自动化检测鲸类提供了新的技术思路。该类模型能够输出检测框、类别及置信度，已经在基于卫星影像的灰鲸（*Eschrichtius robustus*）检测中得到初步应用^[8]。

当前研究仍面临两大挑战：其一，基于深度学习的鲸类检测方法大多依赖于人工标注的边界框，成本较高^[9]；其二，鲸类作为海洋生物，在游动时于卫星影像中未必清晰可见。已有研究通常仅选用影像中形态明确的鲸鱼个体作为训练样本，而忽略了外观模糊的“疑似鲸鱼”样本，以及体型相近的其他海洋哺乳动物^[10]，限制了模型在真实复杂场景下的适用性。

为应对上述问题，本研究选取亚北极地区坎伯兰湾西北部等关键白鲸栖息地为研究区，采用专家标注与深度学习相结合的方法，对其中的中小型海洋哺乳动物，包括白鲸和竖琴海豹进行系统标注，构建了一套开放获取的白鲸空间分布数据集。与现有研究相比，本研究验证了“模型预测+人工修正”的半自动化流程可以在保证样本标注质量的同时提升标注效率。本数据集覆盖了河口、峡湾和近海等多种典型生境，其多物种标注体系为训练能有效区分相似海洋哺乳动物的目标检测模型提供了独特支持，有助于训练和验证遥感目标检测算法区分相似物种的能力，直接支持栖息地利用分析^[6]、种群聚集行为^[10]等生态学研究。

1 数据采集和处理方法

1.1 数据源与采集方法

本文的研究区域覆盖了加拿大北部亚北极地区白鲸关键的夏季栖息地，包括坎伯兰湾西北部（northwestern Clearwater Fiord）、丘吉尔河口（Churchill River Estuary）和坎伯兰湾东南部（southeastern Cumberland Sound）三个区域（图 1）。数据采集时间为 2021 年至 2022 年夏季（7 月至 9 月初），此期间海冰消融、水域开阔，且白鲸常聚集于河口与峡湾地带，有利于卫星观测。

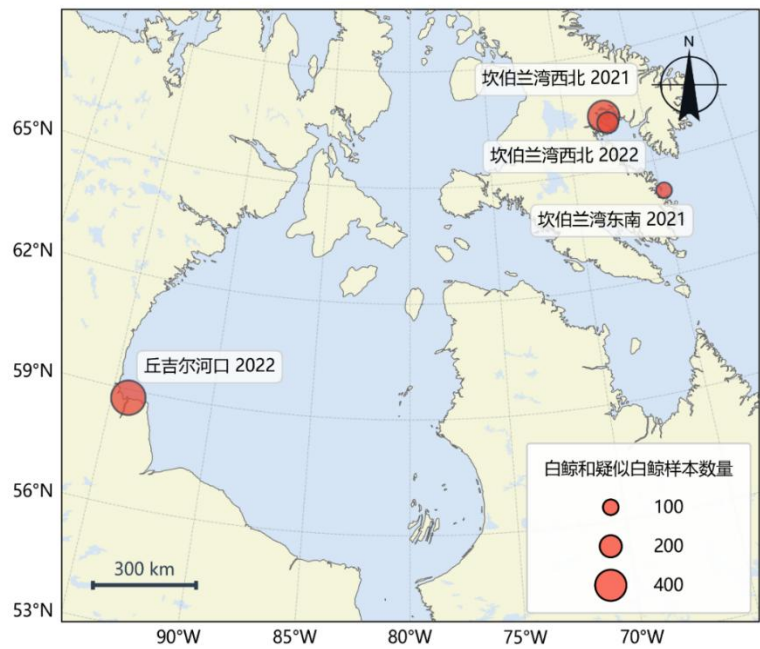


图 1 研究区域及样本分布概况

Figure 1 Study areas and sample distribution overview

根据加拿大渔业与海洋部利用人工布设的白鲸模型所开展的观测结果，在水质条件和鲸鱼活动深度一致的情况下，采用全色波段与全色锐化后的多光谱波段组合进行目视解译，其对白鲸的识别效果并无显著差异^[1]。因此，本研究选用 WorldView-3 与 WorldView-2 的 Map-Ready 级别（Level 3D）全色影像产品。该产品已完成辐射校正、传感器几何校正及三维正射校正，可直接用于目视解译。WorldView-3 与 WorldView-2 的空间分辨率分别为 0.3 米与 0.46 米，其超高空间分辨率是识别白鲸与竖琴海豹（*Pagophilus groenlandicus*）等中小型海洋哺乳动物的关键。所获取产品中，每景影像已经被统一裁切为 16384×16384 像素的瓦片。每个研究区域的样本来源、影像获取日期、样本数量等信息在表 1 中汇总。

表 1 各研究区域的数据来源与样本统计

Table 1 Data sources and sample statistics of the each study areas

研究区域	坎伯兰湾西北	坎伯兰湾西北	丘吉尔河口	坎伯兰湾东南
样本来源	人工标注	人工标注	人工标注	模型预测+人工修正
日期	2021-09-07	2022-08-01	2022-07-31	2021-08-29

研究区域	坎伯兰湾西北	坎伯兰湾西北	丘吉尔河口	坎伯兰湾东南
样本来源	人工标注	人工标注	人工标注	模型预测+人工修正
拍摄卫星	WorldView-3	WorldView-2	WorldView-3	WorldView-3
覆盖面积 (km ²)	779	1281	146	836
鲸鱼数量	196	103	378	41
疑似鲸鱼数量	215	79	113	62
海豹数量	526	16	0	9

1.2 数据处理方法和流程

本研究的总体路线图如图 2 所示，包含以下数据处理步骤：

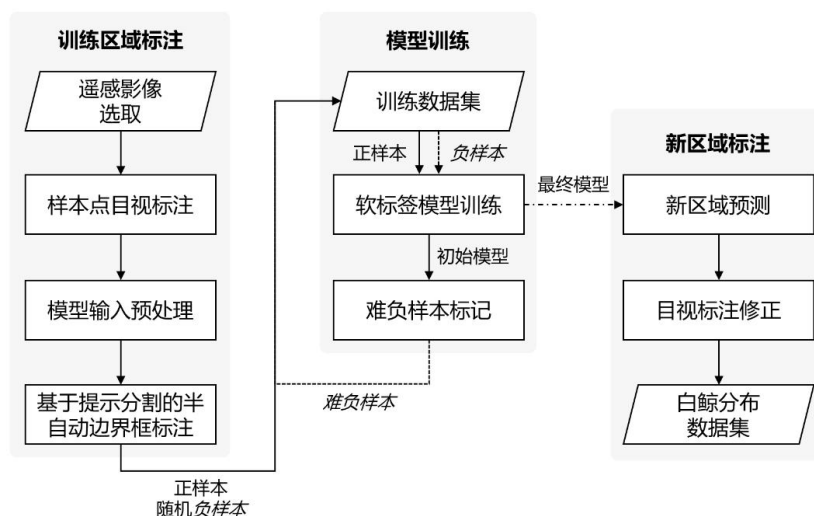


图 2 数据集构建流程

Figure 2 Workflow for constructing the dataset

(1) 样本标注：针对丘吉尔河口与坎伯兰湾西北部区域的影像，首先进行人工点标注，标识白鲸与海豹个体；随后将裁切后的小图像块，基于点标注，使用 Segment Anything Model (SAM)^[12]进行半自动化的实例分割与边界框标注。

(2) 训练数据集构建和模型训练：采用两阶段分步构建包含正样本图像块（至少一个白鲸、疑似白鲸或竖琴海豹目标的图像块）和负样本图像块（不包含任何感兴趣目标的图像块）的训练数据集。首先，基于正样本训练初始模型；随后，从随机采样的负样本图像块（随机负样本）中，筛选出被初始模型错误识别为目标的负样本图像块（难负样本图像块），将其与部分随机负样本图像块共同加入训练集，最终训练出误检率更低的检测模型。

(3) 大规模推理与应用：将训练好的最终模型应用于坎伯兰湾东南部区域影像，进行大规模预测，并对模型检测结果进行人工核查与修正。

1.2.1 点标注

点标注是本数据集构建的基础。两位研究者独立完成所有样本的标注；对标注结果不一致的样本，由第三位研究者介入讨论，直至达成一致意见。所有标注均在 ArcMap 软件中完成，采用系统化的目视解译方法：首先将研究区的超高分辨率卫星影像划分为 250 m×250 m 的规则网格，然后逐

网格进行细致排查，以确保标注的完整性与一致性。

在研究区域内，除白鲸外，还广泛分布着另一类中小型海洋哺乳动物——竖琴海豹。成年白鲸体长约 4.2 m，而竖琴海豹体长约 2 m。由于白鲸在水中活动，在清澈的海水中，卫星图像上的可见深度大约是 2 米^[11]。当个体潜入更深处时，将难以在卫星影像中被识别。根据先前研究^[5]的分类标准，将识别出的对象根据形态特征和空间上下文信息分为三类：

(1) 白鲸：呈现清晰的鲸类体型特征，如纺锤形身体，或可辨识的头部、尾鳍等结构。

(2) 疑似白鲸：目标特征模糊，难以明确判定为鲸类，但依据其亮度、尺寸及周边存在确定白鲸等空间情境，推测可能为白鲸。

(3) 竖琴海豹：体型通常较白鲸更小，呈更圆润的斑点状，通常以大规模群体的形式活动。

标注时，在每个识别目标的身体中心位置标记一点。对于彼此临近的个体，确保相邻鲸鱼的标注点的垂直平分线可以将两条鲸鱼分开。图 3 展示了三类目标在 WorldView 影像中的典型形态特征，可作为目视解译的参考标志。最终，共获得 1626 个点标注（表 1）。

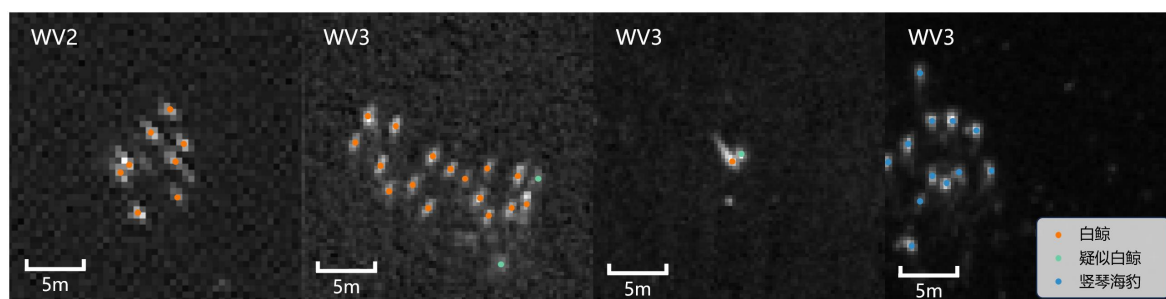


图 3 不同卫星数据和研究区域下的目视标注示例

Figure 3 Visual annotation examples from different satellite imagery and the study areas

1.2.2 模型输入预处理

为适配深度学习模型输入要求并提升处理效率，对原始影像进行了以下预处理操作：

位深转换：将原始 16 位的全色波段影像线性拉伸并转换为 8 位深度，在保留目标对比度的同时兼容主流模型框架。

图像裁剪：将大幅影像裁切为小的图像块。为保证不同分辨率影像中的同类目标在缩放至模型输入尺寸时具有相近的像素大小，根据平台分辨率差异，将 WorldView-3 影像裁剪为 320×320 像素，WorldView-2 影像裁剪为 192×192 像素，使每个图像块覆盖的实际地理范围基本一致。最终，仅保留包含至少一个感兴趣目标的图像块，共得到 538 个正样本图像块。

鉴于鲸鱼分布极为稀疏（含有鲸鱼的图像块数量不足所有图像块数量的 1%），为增强模型鲁棒性，从每个有海洋覆盖的 WorldView-3 瓦片中随机裁剪 50 个不包含任何感兴趣目标的图像块，作为负样本图像用于训练与测试。

1.2.3 基于提示分割的边界框标注

为减少标注的时间成本，设计了半自动化流程对坎伯兰湾西北部、丘吉尔河口地区的样本进行边界框标注。采用提示分割模型 SAM 来自动生成高质量的实例分割掩膜与边界框。SAM 在十亿规模自然图像数据集上训练，在遥感图像上也表现出了强大的零样本泛化能力^[13]。

针对每一个标记采样点，将该点作为 SAM 模型的点提示输入；同时以该标记点为中心，构建适

配目标物种专属尺寸的方形缓冲区域，作为 SAM 的矩形框提示。其中白鲸对应方形缓冲区边长设为 4 m，竖琴海豹对应方形缓冲区边长设为 2 m。采用点提示与物种专属矩形框提示相结合的复合提示策略，可更精准地引导 SAM 完成单一个体目标的定位与分割。

在鲸群密集区域，SAM 初始生成的分割掩码可能存在重叠。为此，开发了一种基于欧氏距离的重叠像素分配算法：对每一个重叠像素，计算其到所有邻近标注点的距离，并将其分配给距离最近的标注点所对应的目标。此过程确保了每个像素仅归属于一个实例，从而生成彼此分离的实例级分割掩码。

最后，将每个实例分割掩膜转换为其最小外接矩形（即边界框）。为保障标注质量，对 SAM 生成的所有边界框进行了人工检查与修正，其中 19% 的白鲸边界框、21% 的疑似白鲸边界框及 4% 的海豹边界框经过了修正。

1.2.4 目标检测训练集构建和模型训练

为实现鲸鱼的自动化检测，本研究选用在速度与精度间取得良好平衡的 YOLOv8-s 模型进行训练。由于研究区域内鲸鱼目标十分稀疏，若仅使用正样本图像块训练，模型在大规模应用时会产生大量误检。为此，设计了两阶段的目标检测训练集构建与模型训练策略。

第一阶段，将 538 个正样本图像块按 4:1 的比例随机划分为初始训练集和初始测试集。训练采用迁移学习技术，使用在 MS-COCO 数据集^[4]上的预训练权重进行微调，并应用了包括随机翻转、马赛克增强、HSV 色彩变换等在内的数据增强技术。模型训练 100 个轮次后，得到初始模型。

第二阶段，从无目标区域随机采样得到随机负样本图像块中，选取 25% 加入初始训练集，25% 加入测试集。对剩余 50% 的随机负样本，使用初始模型进行推理，将其中被错误检测为目标的图像块定义为“难负样本”，加入初始训练集。最终训练集包含 430 个正样本图像块、882 个随机负样本图像块及 227 个难负样本图像块；测试集包含 108 个正样本图像块与 882 个随机负样本图像块。使用相同超参数在最终训练集上重新训练模型，并在测试集上评估性能。

1.2.5 新区域预测和人工修正

为扩展数据集并减少人工标注负担，使用训练好的最终模型对未标注的坎伯兰湾东南部区域影像进行推理。该区域共包含 88763 个图像块，模型预测得到 467 个候选目标，分布在 387 个图像块中。随后，对所有候选图像块进行人工复核，修正误检、补充漏检。为进一步确保数据完整性，还对其余 88376 个图像块进行了系统目视解译。最终，该区域共确认 44 只白鲸、70 只疑似白鲸和 9 只竖琴海豹。

2 数据样本描述

本数据集基于 WorldView-2 与 WorldView-3 卫星影像构建，涵盖 4 个时空分布区域。数据集以压缩包形式提供，内含 5 个文件夹，分别以研究区域名称及影像获取年份命名。每个文件夹包含一个 Shapefile 文件，该文件含有 label 与 geometry 两个字段。其中，label 字段取值为 0、1、2，分别对应白鲸、疑似白鲸、竖琴海豹；geometry 字段则存储了目标的地理位置点坐标。

图 4 至图 7 分别提供了各区域在相应年份的白鲸的空间分布格局，可用于后续的模式分析、栖息地偏好研究以及种群聚集区识别。为了展示研究区域及其周边的地理环境，使用开放获取的

ESRI World Imagery 作为底图。

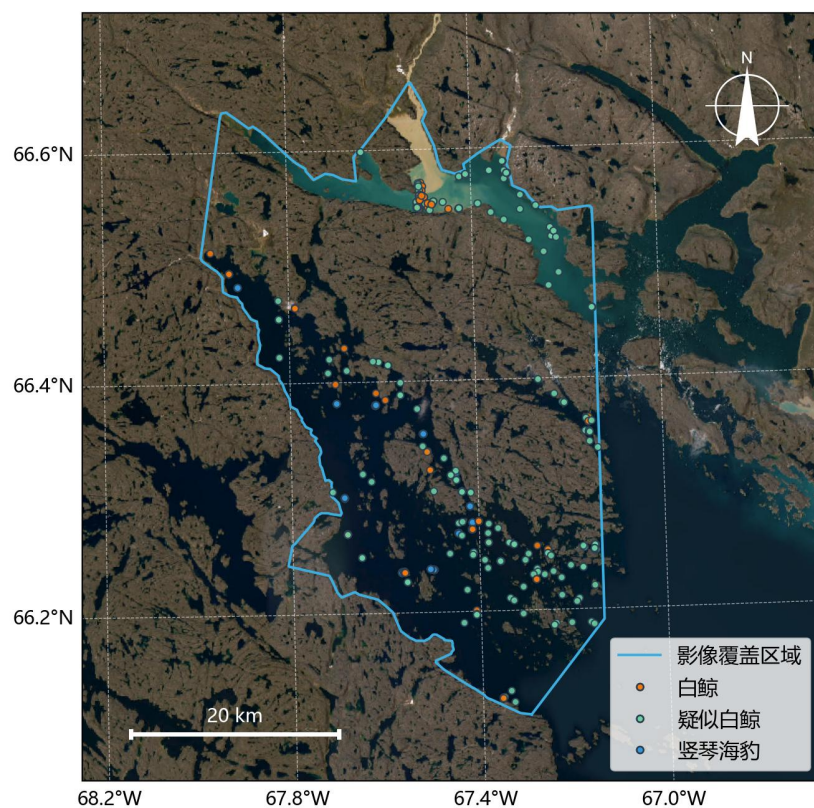


图 4 2021 年坎伯兰湾西北部白鲸分布情况

Figure 4 Beluga distribution in northwestern Cumberland Sound in 2021

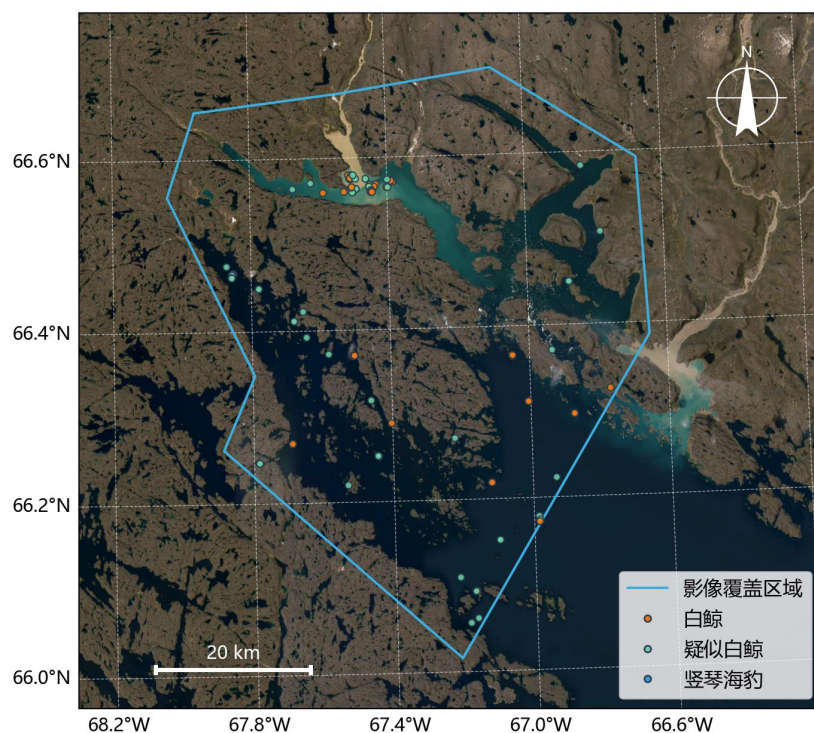


图 5 2022 年坎伯兰湾西北部白鲸分布情况

Figure 5 Beluga distribution in northwestern Cumberland Sound in 2022

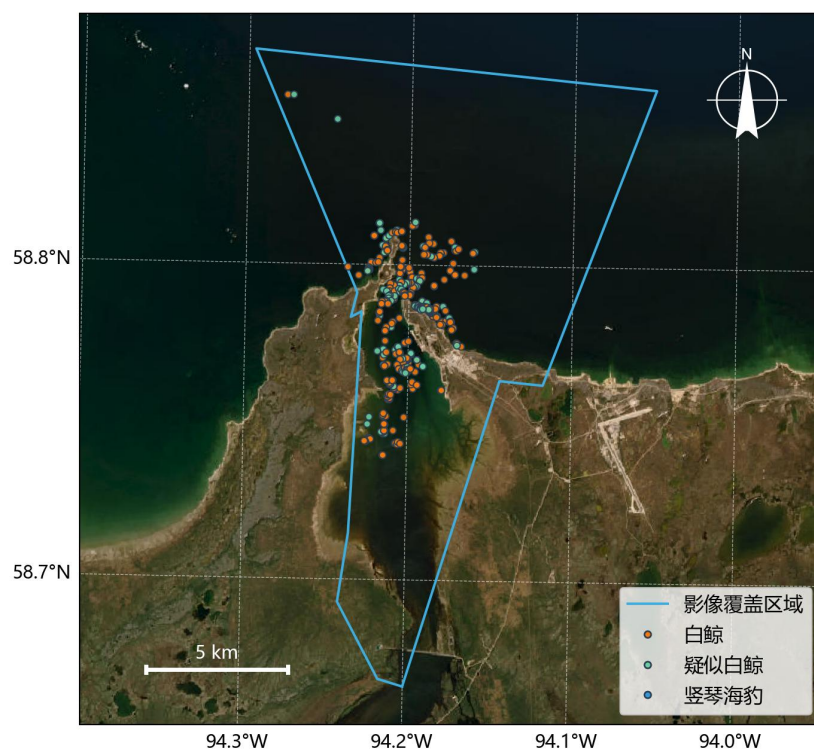


图 6 2022 年丘吉尔河口白鲸分布情况

Figure 6 Beluga distribution in the Churchill River estuary in 2022

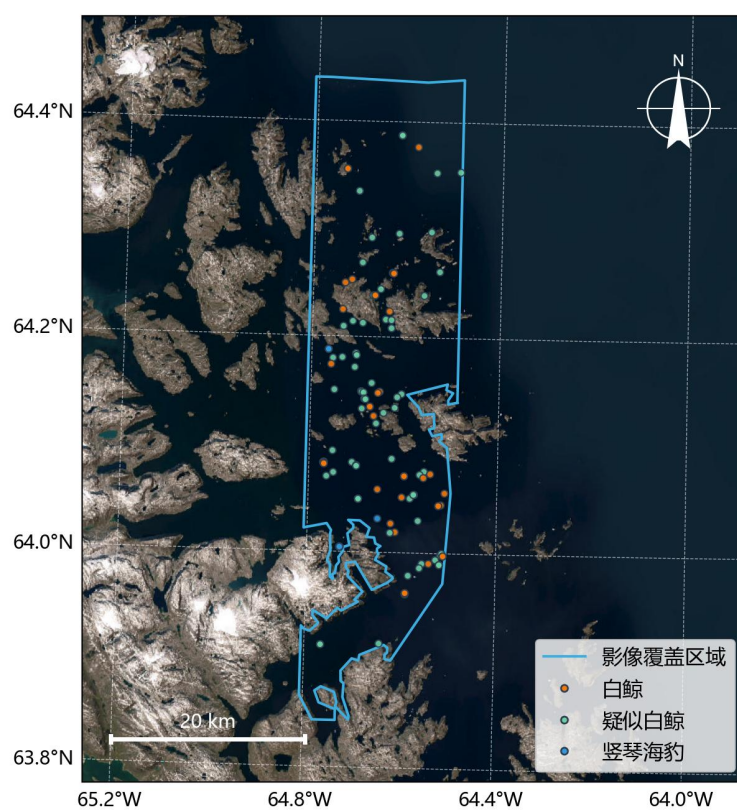


图 7 2021 年坎伯兰湾东南部白鲸分布情况

Figure 7 Beluga distribution in southeastern Cumberland Sound in 2021

3 数据质量控制和评估

为确保数据的准确性与可靠性，在数据生产过程中实施了多层次的质量控制措施。

在数据选择时，优先选取云覆盖量少、成像质量高的影像，以最大限度减少因云层遮挡或图像质量导致的鲸鱼漏检与误判。

在标注过程中，采用双人独立进行标注的方式，标注者间一致性达到 90%，和加拿大渔业部专家在 2019 年坎伯兰湾西北部影像中的标注一致性（89.7%）^[5]相当。对于存在分歧的样本，引入第三名标注者参与集体讨论，直至确定最终类别。对影像中特征模糊的目标，充分利用相邻条带的重叠区域或不同时相的影像进行对比分析，以排除礁石等固定干扰物。

对于白鲸检测模型，使用准确率、召回率和 F_1 分数作为评价指标。在目标检测中，准确率（Precision）指模型预测为感兴趣目标中实际为标注目标的比例；召回率（Recall）指所有标注目标中被模型正确预测的比例； F_1 分数是准确率与召回率的调和平均数，综合反映模型性能。在划分的含有负样本图像块的测试集上，模型的表现如表 2 所示。其中白鲸（含疑似）是指将白鲸和疑似白鲸都划分至同一个类别。结果表明，置信度阈值为 0.36 时，白鲸、疑似白鲸、竖琴海豹的平均 F_1 分数最高。其中疑似白鲸的 F_1 分数最低，这与其自身定义中所固有的不确定性相符。但是白鲸（含疑似）的 F_1 分数达到 69.3%，竖琴海豹的 F_1 分数达到 73.7%，表明模型区分鲸鱼和海豹的能力较强。尽管“疑似白鲸”类别本身具有一定主观性，但其在整体白鲸（含疑似）识别中的贡献仍有助于提升模型在复杂场景下的适用性。

表 2 模型在测试集上的目标检测性能评估结果

Table 2 Evaluation results of the model's object detection performance on the test dataset

评价指标	白鲸	疑似白鲸	白鲸（含疑似）	竖琴海豹
准确率（%）	70.3	34.1	71.2	73.7
召回率（%）	58.4	40.3	67.5	73.7
F_1 分数（%）	63.8	36.9	69.3	73.7

训练白鲸检测模型的目的是在训练集以外的时间和区域进行预测，得到候选感兴趣目标，随后通过仅对包含候选目标的图像块进行人工复核，得到较为高质量的解译结果。因此，模型需要有较高的召回率以保证检测到尽可能多的感兴趣目标。为了进一步测试所训练模型在实际使用中的能力，在坎伯兰湾东南部，以更低的 0.25 为置信度阈值进行预测，并分两步进行修正。第一步，在 387 张含有候选目标的图像块中进行复核（删除多检的候选目标、纠正分类错误的候选目标、补充漏检的目标），得到 41 个白鲸样本，62 个疑似白鲸样本，9 个竖琴海豹样本。第二步，在剩余 88376 个图像块中进行目视解译，补充了 3 个白鲸样本和 8 个疑似白鲸样本。模型预测的结果和两步修正后的样本数量如表 3 所示。模型预测的总体准确率达到 19.1%，召回率达到 72.3%。模型在新区域准确率较低主要源于域适应（domain adaptation）挑战：新区域的影像条件（如光照、海水浊度）与训练集存在分布差异，导致模型在新环境中产生较多误检。此外，低置信度阈值（0.25）的设定虽然提高了召回率，但也引入了更多背景噪声。仅对含有候选目标的图像块进行人工修正后，召回率达到 91.1%。这验证了“模型预测+候选图像块人工修正”策略的有效性，既大幅减少了人工标注工作量，又保证了数据质量。

表 3 模型在坎伯兰湾东南区域的预测与人工修正结果对比

Table 3 Comparison of model predictions and manually corrected results in southeastern Cumberland Sound

样本类别	白鲸 (个)	疑似白鲸 (个)	竖琴海豹 (个)	样本总数 (个)
模型预测正确	31	49	9	89
模型多检	101	243	22	366
模型错分	6	6	0	12
候选图像块修正后	41	62	9	112
剩余图像块标注后	44	70	9	123

4 数据使用方法和建议

本数据集以 Shapefile 格式提供, 建议用户借助 ArcGIS、QGIS 等地理信息系统软件, 或 GeoPandas、Shapely 等 Python 库进行数据读取、可视化及空间分析操作。在科学研究方面, 可基于本数据集开展核密度估计、种群丰度估计、聚集行为分析等工作, 从而深入揭示白鲸的空间分布规律及其与海洋环境因子之间的关联; 在生态保护与管理实践中, 本数据集也为相关管理部门识别关键栖息地、探究物种行为习性、划定科学保护范围以及制定可持续生态旅游策略提供了重要的科学数据支撑。

5 AIGC 使用声明

在撰写本手稿期间, 作者仅使用 DeeSeek-V3.2 进行语法检查、语言润色和语句流畅度优化。所有生成的内容均由作者修改和确认, 作者对论文的整体内容、数据真实性及学术诚信负全部责任。

作者分工职责

郑伟杰 (2001—), 男, 浙江省衢州市人, 硕士研究生, 研究方向为遥感智能解译。主要承担工作: 文献调研, 数据处理, 数据质量控制, 论文撰写。

吴冰玥 (2000—), 女, 广东省中山市人, 硕士研究生, 研究方向为地理空间数据可视化和古气候重建。主要承担工作: 数据处理, 数据可视化, 数据质量控制, 论文修改。

王娅萱 (2001—), 女, 云南省楚雄市人, 硕士研究生, 研究方向为极低温制冷材料和海洋野生动物保护。主要承担工作: 文献调研, 数据处理, 论文修改。

李国庆 (1968—), 男, 河南省洛阳市人, 博士, 研究员, 研究方向为高性能计算、地学大数据管理和共享。主要承担工作: 研究方法指导, 论文修改。

参考文献

- [1] WATT C A, MARCOUX M, FERGUSON S H, et al. Population dynamics of the threatened Cumberland Sound beluga (*Delphinapterus leucas*) population[J]. Arctic Science, 2021, 7(2): 545 – 566. DOI:10.1139/as-2019-0030.
- [2] BELANGER A M, SHERBO B A H, ROTH J D, et al. Use of satellite imagery to estimate distribution and abundance of Cumberland Sound beluga whales reveals frequent use of a glacial river estuary[J].

- Frontiers in Marine Science, 2024, 10: 1305536. DOI:10.3389/fmars.2023.1305536.
- [3] 曾志伟, 李可, 郭浩, 等. 动物遥感与遥测的研究进展与挑战[J]. 遥感学报, 2025, 29(10): 2859-2890. DOI:10.11834/jrs.20254437. [ZENG Z W, LI K, GUO H, et al. Research advances and challenges in animal remote sensing and telemetry. National Remote Sensing Bulletin, 29(10): 2859-2890 DOI:10.11834/jrs.20254437.]
- [4] ATTARD M R G, PHILLIPS R A, BOWLER E, et al. Review of satellite remote sensing and unoccupied aircraft systems for counting wildlife on land[J]. Remote Sensing, 2024, 16(4): 627. DOI:10.3390/rs16040627.
- [5] CHARRY B, TISSIER E, IACOLAZZA J, et al. Mapping Arctic cetaceans from space: a case study for Beluga and narwhal[J]. PLoS One, 2021, 16(8): e0254380. DOI:10.1371/journal.pone.0254380.
- [6] WATT C A, CARLYLE C G, HORNBLY C A, et al. Eastern High Arctic–Baffin Bay beluga whale (*Delphinapterus leucas*) estuary abundance and use from space[J]. Frontiers in Mammal Science, 2023, 2: 1208276. DOI:10.3389/fmamm.2023.1208276.
- [7] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 779 – 788. DOI:10.1109/CVPR.2016.91.
- [8] GREEN K M, VIRDEE M K, CUBAYNES H C, et al. Gray whale detection in satellite imagery using deep learning[J]. Remote Sensing in Ecology and Conservation, 2023, 9(6): 829 – 840. DOI:10.1002/rse2.352.
- [9] HÖSCHLE C, CUBAYNES H C, CLARKE P J, et al. The potential of satellite imagery for surveying whales[J]. Sensors, 2021, 21(3): 963. DOI:10.3390/s21030963.
- [10] RIVAS M L, GUIRADO E, ORTEGA Z. Relation between beluga whale aggregations and sea temperature on climate change forecasts[J]. Frontiers in Marine Science, 2024, 11: 1359429. DOI:10.3389/fmars.2024.1359429.
- [11] STEWART J B, HUDSON J M, SHERBO B A H, et al. Estimating Beluga whale abundance from space: using drones to ground-validate VHR satellite imagery[J]. Remote Sensing in Ecology and Conservation, 2024, 10(6): 673 – 685. DOI:10.1002/rse2.396.
- [12] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France. IEEE, 2024: 3992 – 4003. DOI:10.1109/ICCV51070.2023.00371.
- [13] WANG D, ZHANG J, DU B, et al. SAMRS: scaling-up remote sensing segmentation dataset with segment anything model[EB/OL]. 2023: arXiv: 2305.02034. <https://arxiv.org/abs/2305.02034>.
- [14] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[M]//Computer Vision – ECCV 2014. Cham: Springer International Publishing, 2014: 740 – 755. DOI:10.1007/978-3-319-10602-1_48.

论文引用格式

郑伟杰, 吴冰玥, 王娅萱, 等. 基于 WorldView 数据的 2021 – 2022 年加拿大亚北极地区白鲸分布数据

集[J/OL]. 中国科学数据, 2026, 11(2). (2026-06-24). DOI: 10.11922/11-6035.noda.2025.0209.zh.

数据引用格式

郑祎杰, 吴冰玥, 王娅萱, 等. 基于 WorldView 数据的 2021 – 2022 年加拿大亚北极地区白鲸分布数据集 [DS/OL]. V2. Science Data Bank, 2025. (2026-01-05). DOI: 10.57760/sciencedb.j00001.01659.

A dataset of beluga distribution in the Canadian sub-Arctic derived from WorldView imagery (2021–2022)

ZHENG Yijie^{1,2}, WU Bingyue³, WANG Yaxuan⁴, LI Guoqing^{1*}

1. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, P.R. China

2. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, P.R. China

3. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, P.R. China

4. School of Future Technology, University of Chinese Academy of Sciences, Beijing 101408, P.R. China

*Email: ligq@aircas.ac.cn

Abstract: This dataset was derived from WorldView-2/3 satellite imagery acquired in 2021–2022 over northwestern and southeastern Cumberland Sound and the Churchill River estuary in the Canadian sub-Arctic. It contains three annotated classes: beluga, suspected beluga, and harp seal. The dataset was developed through a workflow combining manual annotation, deep learning inference, and manual verification. Systematic manual interpretation was first conducted in northwestern Cumberland Sound and the Churchill River estuary to build training samples for object detection. A trained YOLOv8 model was then applied to southeastern Cumberland Sound for large-scale inference, followed by manual review and correction. The final dataset includes the geolocations of 721 belugas, 477 suspected belugas, and 551 harp seals. This resource supports studies on the distribution and abundance estimation of belugas and harp seals, as well as training and evaluation of object detection algorithms, and provides valuable data support for polar ecological conservation and sustainable development of tourism resources.

Keywords: object detection; beluga whale; WorldView; sub-Arctic region

Dataset Profile

Title	A dataset of beluga distribution in the Canadian sub-Arctic derived from WorldView imagery (2021–2022)
Data corresponding author	LI Guoqing (ligq@aircas.ac.cn)
Data authors	ZHENG Yijie, WU Bingyue, WANG Yaxuan, LI Guoqing
Time range	2021–2022
Geographical scope	Sub-Arctic regions of northern Canada, including northwestern Cumberland Sound

	(66°00'57"–66°42'03"N, 68°02'26"W–66°39'03"W), the Churchill River Estuary (58°47'38"–58°49'19"N, 94°12'30"W–94°09'05"W), and southeastern Cumberland Sound (63°50'35"–64°26'35"N, 64°48'25"–64°29'16"W).
Data volume	95 KB
Data format	*.shp
Data service system	https://doi.org/10.57760/sciencedb.j00001.01659
Sources of funding	4th IEEE GRSS Student Grand Challenge; Development of an Integrated Storage—Computing Environment for Remote Sensing Big Data Analysis (Grant No. E1D7180200).
Dataset composition	The dataset archive contains four directories, each storing beluga distribution data files from different regions and periods. (1) northwestern_Cumberland_Sound_2021.shp: Beluga distribution data for northwestern Cumberland Sound in 2021, including 196 belugas, 215 suspected belugas, and 526 harp seals. (2) northwestern_Cumberland_Sound_2022.shp: Beluga whale distribution data for northwestern Cumberland Sound in 2022, including 103 belugas, 79 suspected belugas, and 16 harp seals. (3) Churchill_River_2022.shp: Beluga whale distribution data for the Churchill River estuary in 2022, including 378 belugas and 113 suspected belugas. (4) southeastern_Cumberland_Sound_2021.shp: Beluga whale distribution data for southeastern Cumberland Sound in 2021, including 44 belugas, 70 suspected belugas, and 9 harp seals. In each file, the field “label” takes the values of 0, 1, and 2, representing beluga, suspected beluga, and harp seal, respectively.