

2020 年青藏高原湖泊语义分割数据集

王兆滨^{1,2*}, 高雄¹, 康建芳^{2,3}, 艾鸣浩^{2,3}

ISSN 2096-2223

CN 11-6035/N



文献 DOI:

10.11922/11-6035.ncdc.2021.0005.zh

数据 DOI:

10.12072/ncdc.NIEER.db0112.2021

文献分类: 地球科学

收稿日期: 2021-06-15

开放同评: 2021-11-11

录用日期: 2021-11-19

发表日期: 2021-12-13

1. 兰州大学信息科学与工程学院, 兰州 730000
2. 中国科学院西北生态环境资源研究院, 兰州 730000
3. 国家冰川冻土沙漠科学数据中心, 兰州 730000

摘要: 青藏高原是我国湖泊分布最密集的两个地区之一。随着全球气候变暖, 冰川径流量增加, 青藏高原湖泊存在着明显的扩张。湖泊水体的自动提取是进行灾害监测的基础性工作, 也对水资源和生态环境保护具有重要的意义。本研究以青藏高原(空间范围大致介于 26°00'N–39°47'N, 73°19'E–104°47'E 之间)为研究区域, 以 14 级(空间分辨率为 17 m)谷歌地球遥感影像为基础构建小型的青藏高原湖泊语义分割数据集。本数据集共包含随机选取的 6774 张大小为 256×256 的 RGB 图像, 以及结合深度学习数据集标注软件 labelme 和目视解译方法获得的真值图像(红色为湖泊水体, 黑色为背景)。本数据集一方面可以用作青藏高原湖泊水体的提取, 为该区域工程建设提供基础的数据支撑; 另一方面可以用来预训练语义分割模型的权重, 提高其他区域水体提取的精度。

关键词: 谷歌地球; 青藏高原; 深度学习; 语义分割

数据库(集)基本信息简介

数据库(集)名称	青藏高原语义分割数据集
数据作者	王兆滨、高雄、康建芳、艾鸣浩
数据通信作者	王兆滨 (wangzhib@lzu.edu.cn)
数据时间范围	2020年
地理区域	26°00'N–39°47'N, 73°19'E–104°47'E
空间分辨率	17 m
数据量	545.8 MB
数据格式	*.png
数据服务系统网址	http://dx.doi.org/10.12072/ncdc.NIEER.db0112.2021
基金项目	寒旱区环境演变研究“科技领域云”的建设与应用 (XXH13506-103) 国家冰川冻土沙漠科学数据中心 (No.E01Z7902) 中国科学院冰川冻土沙漠科学数据中心能力提升 (No.Y9298302)
数据库(集)组成	本数据集由两部分构成: (1) 6774张大小为256×256的RGB谷歌遥感图像。(2) 与谷歌遥感图像相对应的6774张真值图像(红色为湖泊水体, 黑色为背景)。

* 论文通信作者

王兆滨: wangzhib@lzu.edu.cn

引言

湖泊水体提取是遥感图像分析的一个重要研究领域，并且青藏高原上分布着约 1612 个湖泊，面积超过 1 km² 的湖泊约占 1200 个，占中国湖泊数量与面积的一半^[1]。一方面，湖泊对全球温度变化非常敏感，且在碳循环中扮演着重要的角色，而青藏高原上湖泊密集，为研究者们提供了重要的机会^[2]。另一方面，湖泊与人类的生产和生活息息相关，比如农业灌溉、交通运输和提供人类赖以生存的饮用水源等。因此，研究基于遥感图像的湖泊水体的自动提取具有重要意义。

语义分割是计算机视觉领域的一项关键技术，其是对图像中的所有像素进行分类，并将所有的图像分割成若干个具有语义类别的区域来表示现实生活的实际物品。与自然图像语义分割数据集相比，遥感图像的拍摄角度是不同的，图像中目标通常是向上的，且在图像中以任意的方向出现^[3]。

对于遥感图像的语义分割任务已经得到了广泛的研究，大部分的研究方法都是将基于自然图像的语义分割方法迁移到遥感图像处理领域。由于深度学习对数据集的规模存在一定的要求，若数据集过小，则模型不能对目标进行充分的学习。但遥感图像的获取存在一定的限制，其语义分割数据集往往都不会太大。研究者们通常使用在大规模的自然图像数据集上预训练而获得的预训练权重，然后在遥感图像数据集上精细调整权重，用于遥感领域的语义分割^[4]。

虽然此方法可以提高模型的精度，但是由于自然图像与遥感图像有较大的差别，主要体现在同一类别尺度变化较大，且遥感图像中具有复杂的背景。因此，使用自然图像数据集获得的语义分割模型也就不适用于遥感图像，使用基于自然图像数据集的预训练方法对模型准确率的提升还是具有较大局限性。

随着深度学习技术的快速发展，可以使用基于深度学习方法来对遥感图像中湖泊进行分割^[5-6]。虽然，基于阈值法和常规机器学习算法对湖泊水体提取的研究仍具有一定的优势，但是深度学习方法对特征的学习能力相比其他方法还是具有较大的优势。本数据集，可用于青藏高原区域湖泊的灾害分析和评估研究。一方面可以基于语义分割方法对湖泊进行提取，也可以对后面的研究进行数据支持。另一方面，可以使用本数据集来对语义分割模型进行预训练，解决了使用自然数据集进行预训练的局限性，从而提高研究模型的准确率。

1 数据采集和处理方法

1.1 数据来源

本研究建立了以青藏高原湖泊水体为研究对象且由 RGB 图像组成的可见光光谱遥感图像数据集，其中的图像全部来自于谷歌地球。在遥感图像中，大小和方向的差异是产生目标物体偏差的主要原因，为了消除偏差和增加多样性，我们可以使用数据增广来扩充数据集，使得模型可以充分学习目标物体的语义特征。

1.2 数据采集方法

1.2.1 训练样本制作

与自然图像相比，遥感图像的尺寸往往很大。在本数据集中，图像的原始大小一般为 1457×971，

个别图像的大小会有细微的差别,但由于 GPU 显存和其他软硬件的限制,不能将原始图像直接输入到模型,需要对原始图像进行缩放,并将原始图像分割成不重叠的块。本数据集中只有湖泊与非湖泊两种类别,在青藏高原上,湖泊只占一小部分,而非湖泊则占有较大的部分,因此会产生数据集正负样本不平衡问题,进而模型不能充分学习到正样本的特征,产生欠拟合问题。因此,使用数据平衡策略来平衡正负样本。如果正样本在图像中的占比非常小,将丢弃图像。反之,我们则保留该图像。最后从所有的子图像中随机选取了共 6774 张大小为 256×256 的包含湖泊的图像。

1.2.2 数据标注

在得到了裁剪后的子图像后,需要将所得到的图像进行数据标注。本研究使用的数据标注软件为 labelme,其是麻省理工大学人工智能实验室采用 Python 语言编写研发的图形界面的图像标注软件。使用 labelme 软件打开裁剪后的子图像并对图像中的湖泊水体沿着湖泊边缘进行标注,对所有的标注结果分配类别为 lake,最终得到标注后的 json 文件。数据标注形式及类别如图 1 所示。

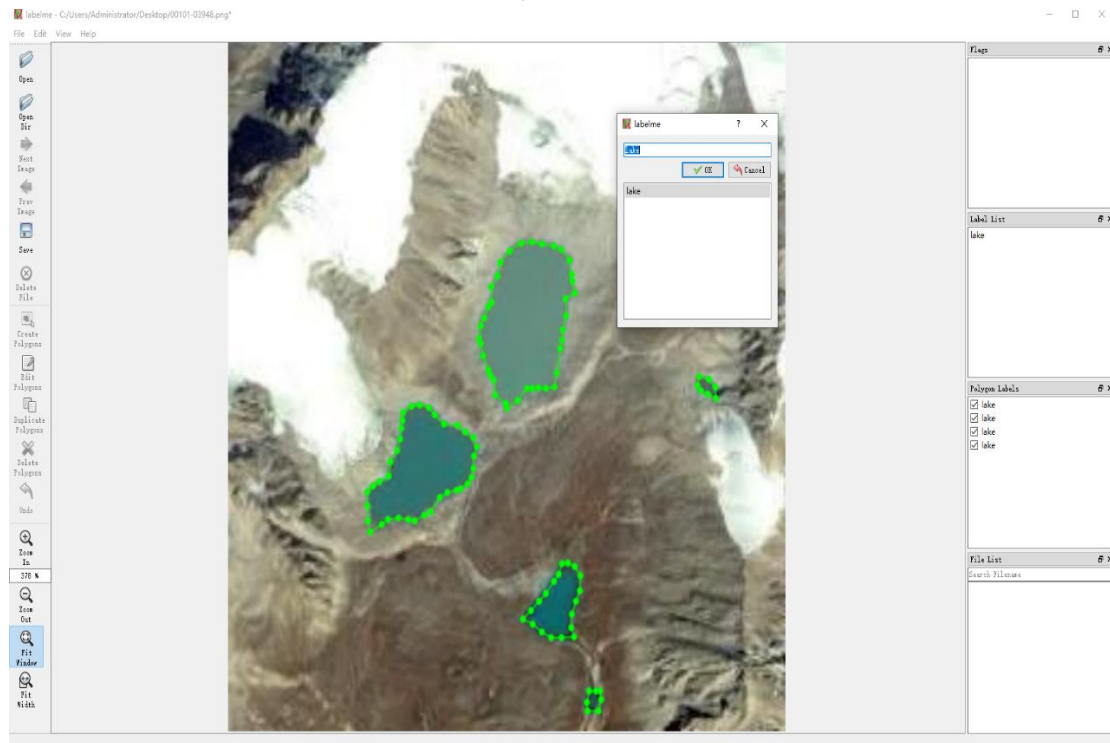


图 1 labelme 对图像进行数据标注

1.2.3 文件格式转换

在完成数据标注后,得到的 json 文件并不能直接用于语义分割模型的训练。因此,需要对得到的 json 文件进行批量转换。转换后的文件如图 2 所示,其中共包含 5 个文件,需提取 img.png 和 label.png 作为训练图像和真值图像。对于图片中所有相同的类别,赋予其相同的颜色。在本数据集中只有两个类别,一个是湖泊,另一个是背景。

1.2.4 数据增广

数据增广是深度学习任务中用于在训练样本较小的情况下生成新的样本以增加训练集样本数量

的一种常用技术。在本研究每个训练时代之前，对所有训练样本使用最常见的翻转（上-下、左-右）和在（ 0° ， 90° ）中以任意角度旋转方法来提高模型的泛化能力。数据增广的结果如图 3 所示，其中 A 为原始图像；B 是将原始图像左右翻转；C 是将原始图像上下翻转；D-F 是将原始图像以（ 0° ， 90° ）中的任意角度旋转；a-f 为 A-F 对应的真值图像。



图 2 json 文件转换后的图像文件

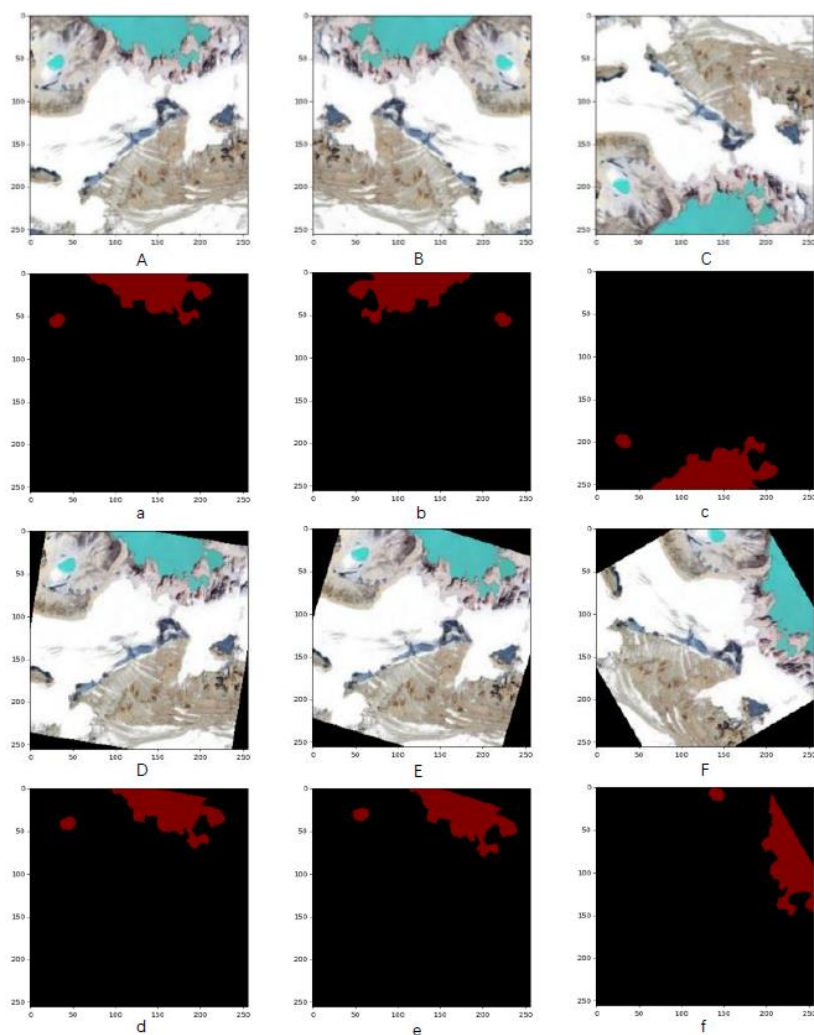


图 3 数据增广的示例

1.3 数据集的划分

一般来说，深度学习模型学习的本质是学习数据分布，为了保证模型所学习到的特征可以对数

据进行正确的预测，需要保证训练集和测试集具有近似的分布。随机选择 6164 张 RGB 图像作为训练集，剩下的 610 张 RGB 图像用作测试集。

1.4 模型训练和应用细节

语义分割模型训练的过程一般分为训练阶段和测试阶段。在训练阶段，每次可将批量的 RGB 图像和通过数据标注以及格式转换后的真值图像作为训练样本输入到模型中。在前向传播过程中，经过一系列的卷积（反卷积）层、标准化层、池化层和激活函数层来进行特征的提取，最后通过 softmax 函数来得到最后的标签映射。将得到的标签映射和真值图像输入到损失函数中来计算损失值。最后，在反向传播过程中通过梯度下降法来更新模型的参数。在测试阶段，使用在训练集上获得最佳评价指标的模型参数来预测测试集中的图像，并将得到的预测标签映射与真值图像进行比较，通过相关的评价指标确定模型性能的好坏。语义分割模型训练的流程如图 4 所示。

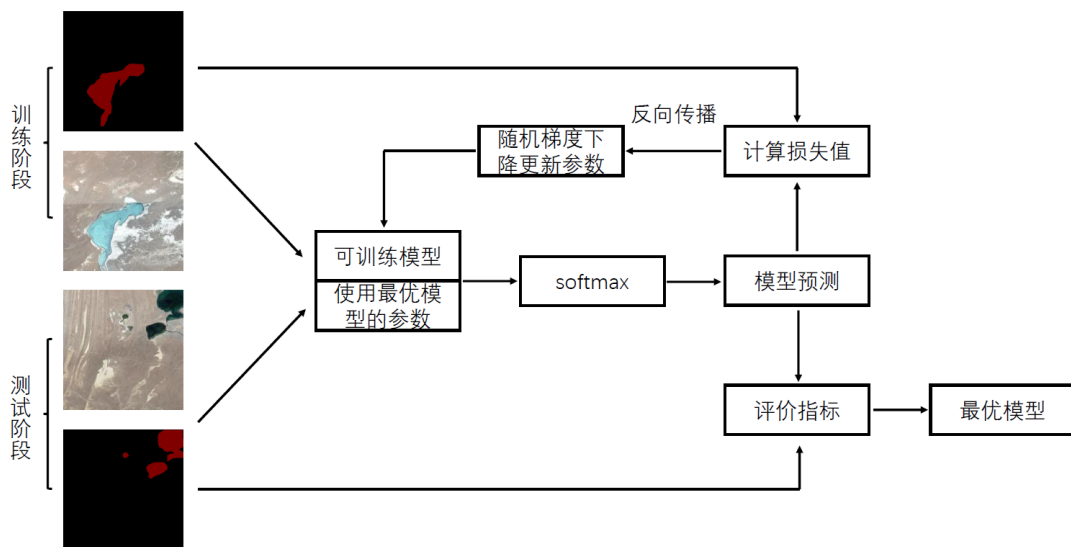


图 4 语义分割模型流程图

本模型算法是基于深度学习开源框架 pytorch^[7]，且模型的训练参数使用 Kaiming^[8]初始化来得到。交叉熵损失函数^[9]被用来计算预测标签和真值图像的损失值，为了最小化交叉熵损失函数，使用 Adam^[10]优化器来更新模型的参数，将权重衰减系数设置为 0.001，并使用学习率衰减策略来防止模型的过拟合。使用学习率衰减策略的主要原因是模型训练的后期，较大的学习率会越过最优解从而导致模型不能收敛到最小值。最终模型经过 70 个时代后收敛到最小值。

2 数据样本描述

本数据集共分为 train_img 和 train_label 两部分，其中 train_img 文件夹内共包含 6774 张大小为 256×256 的 RGB 谷歌地球遥感影像，其空间分辨率为 17 m，train_label 文件夹内为 6774 张原始影像相对应的二分类标签图像，其中红色为湖泊水体，黑色为背景。

图 5 中展示了青藏高原中 RGB 图像及其对应的真值图像。

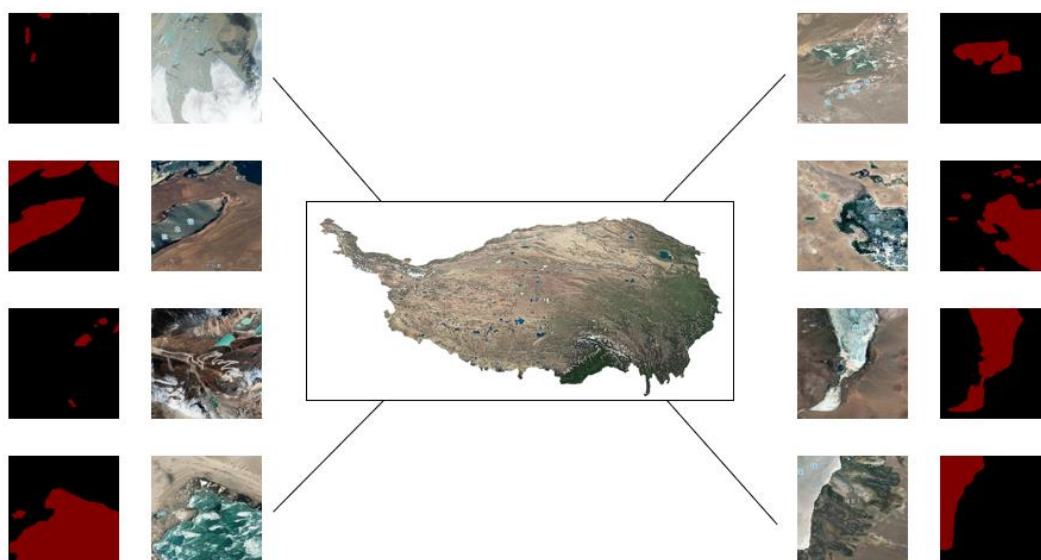


图 5 湖泊水体提取研究区

3 数据质量控制和评估

3.1 评价指标

语义分割是深度学习中热门的研究领域之一，不同于图像分类方法，但归根结底也是一种图像像素分类方法。所以，其分类结果由 True Positive(TP), False Positive(FP), True Negative(TN), False Negative(FN), 即真阳性, 假阳性, 真阴性, 假阴性 4 种情况来构成混淆矩阵。常用的衡量算法精度的评价指标如下:

(1) Overall Accuracy (OA, 总体准确率): 这是一个评价算法精度最直接的评价指标, 即正确分类的像素数与总像素数的比值。

$$OA = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

(2) Recall (RC, 召回率): 召回率是正确预测为湖泊的像素数与湖泊所有像素数的比率。

$$RC = \frac{TP}{TP+FN} \times 100\% \quad (2)$$

(3) Mean Intersection Over Union (MIoU, 平均交并比): MIoU 是正确分类的像素数和真值图像像素数之比与对应类别中检测到的像素数之比的平均值。

$$MIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{FN+TP+FN} \times 100\% \quad (3)$$

(4) True Water Rate (TWR, 真水率): TWR 是正确分类的湖泊像素数与标记的湖泊像素数的比值。

$$TWR = \frac{TP}{TP+FP} \times 100\% \quad (4)$$

(5) False Water Rate (FWR, 误水率): FWR 为错误分类湖泊像素数与标记的湖泊像素数之比。

$$FWR = \frac{FP}{TP+FP} \times 100\% \quad (5)$$

3.2 湖泊水体提取结果评价

通过使用 MSLWENet^[5]、DeepLab V3+^[11]、PSPNet^[12]、MWEN^[13]和 Unet^[14]对本数据集进行验证，其结果表明，由本数据集训练出的语义分割模型还是比较理想的，能够有效的解决湖泊类内方差大、类间方差小以及小湖泊误分类等问题。其中 MSLWENet 在验证集上 OA 达到了 98.53%，并且在类内方差大地区的 OA 为 98.31%、类间方差小地区的 OA 为 98.78%以及对小湖泊识别的 OA 达到了 98.70%。除此之外，MSLWENet 的 RC、MIoU、TWR 和 FWR 评价指标分别为 97.67%、96.09%、97.47%、2.53%，都明显优于其他 4 个对比模型，具体信息如表 1 所示。DeepLab V3+和 PSPNet 是自然图像分割领域的经典模型，由于使用了具有强特征提取能力的 ResNet^[15]作为编码器，所以在湖泊水体提取任务也获得了不错的性能，但是在细节上与 MSLWENet 还是具有一定的差异。MWEN 是专用于遥感图像水体提取的神经网络，Unet 是医学图像分割领域的经典模型，但是由于这两个模型使用了浅层编码器，所以将其应用到青藏高原区域水体提取任务时，不能很好地处理噪声干扰和提取足够的特征用于像素分类。其中 DeepLab V3+、PSPNet、MWEN 和 Unet 在验证集上的 OA 分别为 98.10%、98.08%、97.75%和 96.12%，在类内方差大地区的 OA 分别为 97.07%、97.70%、96.22%和 92.97%，在类间方差小地区的 OA 为 95.87%、96.85%、95.30%和 84.89%，以及对小湖泊识别的 OA 为 98.57%、98.26%、98.30%和 96.81%。

图 6 中第 1–2 行为小湖泊提取可视化结果，第 3–4 行为类间方差小区域提取的可视化结果，第 5–6 行为类内方差大区域提取的可视化结果。其中 a(1)–a(5)为测试集中的原始图像；(b(1)–b(5))是原始图像所对应的真值图像；c(1)–c(5)、d(1)–d(5)、e(1)–e(5)、f(1)–f(5)、g(1)–g(5)分别是 MSLWENet、DeepLab V3+、PSPNet、MWEN 和 Unet 的提取结果。通过对图 6 分析可以得到，非湖泊类别中存在的冰川积雪、山体阴影与湖泊具有相似的光谱特征，同为湖泊但是在纹理光谱等特征却存在着较大的差异以及对小湖泊水体提取的准确率是衡量算法性能的重要方面。

表 1 不同神经网络模型在验证集上的定量比较

模型	MSLWENet	DeepLab V3+	PSPNet	MWEN	Unet
OA	98.53%	98.10%	98.08%	97.75%	96.12%
RC	97.67%	97.01%	96.29%	96.94%	96.06%
MIoU	96.09%	95.61%	95.55%	94.84%	91.36%
TWR	97.47%	97.01%	97.11%	96.94%	96.06%
FWR	2.53%	2.99%	2.89%	3.06%	3.94%

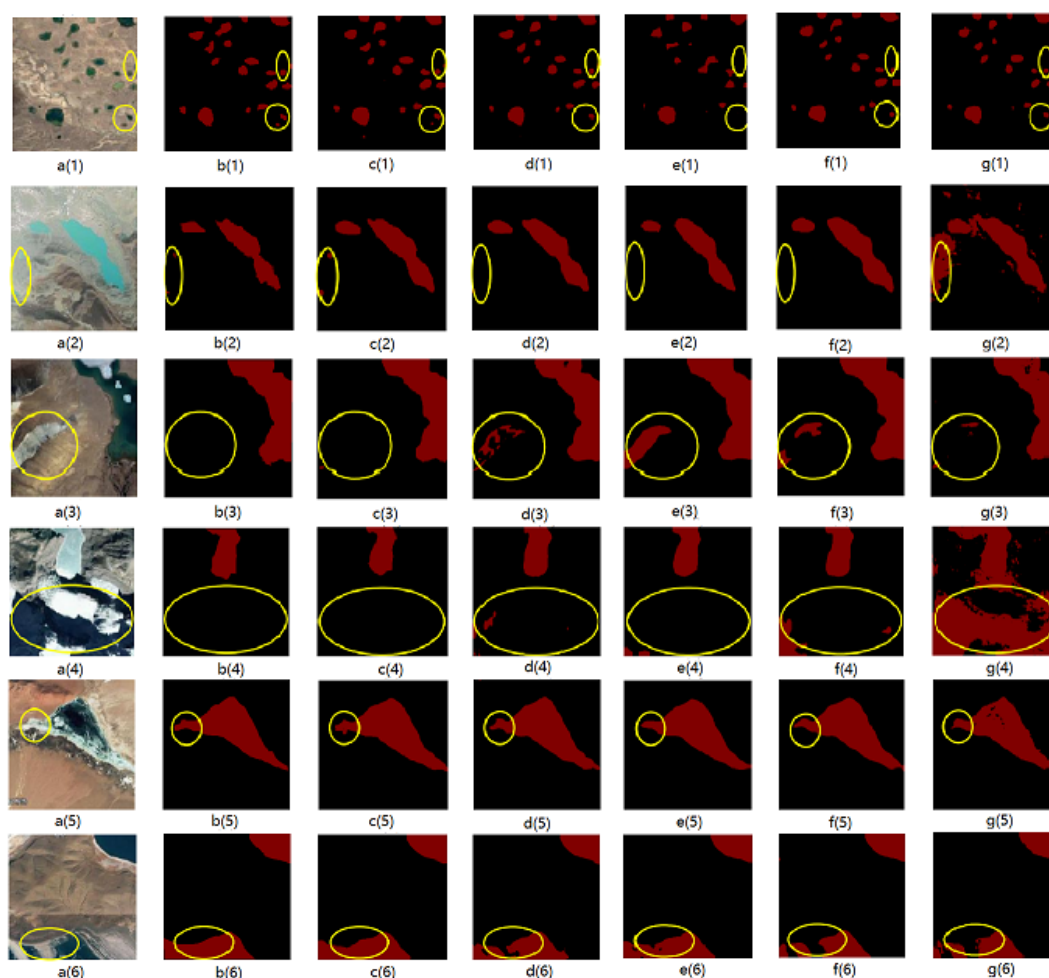


图 6 不同模型提取湖泊水体的性能比较

4 数据价值

基于阈值法和传统机器学习等方法的湖泊水体提取，虽然取得了不错的进展，但仍存在一定的局限性。比如，阈值法及其相关算法，存在最佳阈值的选择问题；而传统的机器学习算法，则涉及到特征工程中特征的组合和选择问题。最后算法的性能好坏主要就是取决于特征工程所获得的特征能否正确表达样本的真实属性。

深度学习语义分割这一计算机视觉领域中的重点研究领域，近些年来逐渐被应用到遥感图像湖泊水体提取中，但是这也存在着一个弊端，就是数据集的制作问题。本文所制作的数据集一方面可以用来作为湖泊语义分割数据集，来对青藏高原上的湖泊进行水体提取的研究。另一方面也可以用本数据集来做预训练获得预训练权重，从而提高湖泊水体相关研究者们在自己数据集上的精度。这对青藏高原的生态稳定以及环境的可持续发展具有重要的意义。

5 数据使用方法和建议

本数据集所保存的图像文件，可以通过 Python 中的 PIL 和 Opencv 库中的相关模块或包来进行

读取，并输入到深度学习模型中。

数据作者分工职责

王兆滨（1980—），男，山东菏泽人，博士研究生，副教授。研究方向：遥感影像信息处理。主要承担的工作：组织实施数据集的制作。

高雄（1995—），男，山东烟台人，硕士生。研究方向：遥感影像信息处理。主要承担的工作：具体数据处理与数据集的整理。

康建芳（1981—），女，甘肃天水人，硕士生。研究方向：科学数据中心运行管理及科学数据的管理研究。主要承担的工作：数据集的管理。

艾鸣浩（1986—），男，山东济宁人，博士研究生，研究方向：地理信息系统。主要承担工作：部分数据处理方法的实现。

参考文献

- [1] 张国庆. 青藏高原湖泊变化遥感监测及其对气候变化的响应研究进展[J]. 地理科学进展, 2018, 37: 214 – 223.
- [2] TRANVIK L J, DOWNING J A, COTNER J B, et al. Lakes and reservoirs as regulators of carbon cycling and climate[J]. Limnology and Oceanography, 2009, 54(6part2): 2298–2314. DOI:10.4319/lo.2009.54.6_part_2.2298.
- [3] XIA G S, BAI X, DING J, et al. DOTA: a large-scale dataset for object detection in aerial images[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 3974–3983. DOI:10.1109/CVPR.2018.00418.
- [4] CUI B G, CHEN X, LU Y. Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection[J]. IEEE Access, 2020, 8: 116744–116755. DOI:10.1109/ACCESS.2020.3003914.
- [5] WANG Z B, GAO X, ZHANG Y N, et al. MSLWENet: a novel deep learning network for lake water body extraction of google remote sensing images[J]. Remote Sensing, 2020, 12(24): 4140. DOI:10.3390/rs12244140.
- [6] WANG Z B, GAO X, ZHANG Y N. HA-net: a lake water body extraction network based on hybrid-scale attention and transfer learning[J]. Remote Sensing, 2021, 13(20): 4121. DOI:10.3390/rs13204121.
- [7] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high-performance deep learning library[EB/OL]. 2019: arXiv: 1912.01703[cs.LG]. <https://arxiv.org/abs/1912.01703>
- [8] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. IEEE, 2015: 1026–1034. DOI:10.1109/ICCV.2015.123.
- [9] DE BOER P T, KROESE D P, MANNOR S, et al. A tutorial on the cross-entropy method[J]. Annals of Operations Research, 2005, 134(1): 19–67. DOI:10.1007/s10479-005-5724-z.

- [10] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. 2014: arXiv: 1412.6980[cs.LG]. <https://arxiv.org/abs/1412.6980>
- [11] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Computer Vision - ECCV 2018. DOI:10.1007/978-3-030-01234-2_49.
- [12] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 6230–6239. DOI:10.1109/CVPR.2017.660.
- [13] GUO H X, HE G J, JIANG W, et al. A multi-scale water extraction convolutional neural network (MWEN) method for GaoFen-1 remote sensing images[J]. ISPRS International Journal of Geo-Information, 2020, 9(4): 189. DOI:10.3390/ijgi9040189.
- [14] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. DOI:10.1007/978-3-319-24574-4_28.
- [15] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 770–778. DOI:10.1109/CVPR.2016.90.

论文引用格式

王兆滨, 高雄, 康建芳, 等. 2020 年青藏高原湖泊语义分割数据集[J/OL]. 中国科学数据, 2021, 6(4). (2021–12–13). DOI: 10.11922/11-6035.ncdc.2021.0005.zh.

数据引用格式

王兆滨, 高雄, 张耀南. 青藏高原湖泊语义分割数据集[DB/OL]. 国家冰川冻土沙漠科学数据中心 (www.ncdc.ac.cn), 2021. DOI:10.12072/ncdc.NIEER.db0112.2021.

A dataset of semantic segmentation for the Qinghai-Tibet Plateau in 2020

WANG Zhaobin^{1,2*}, GAO Xiong¹, KANG Jianfang^{2,3} AI Minghao^{2,3}

1. School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, P.R. China
2. Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, P.R. China
3. National Cryosphere Desert Data Center, Lanzhou 730000, P.R. China

*Email: wangzhaobin@lzu.edu.cn

Abstract: The Qinghai-Tibet Plateau is one of China's top 2 regions with the most densely distributed lakes. Due to effect of global warming and increasing glacial runoff, there is a significant expansion the lakes on the plateau has significantly expanded. The automatic extraction of lake water bodies is fundamental for disaster monitoring, and for the protection of water resources and ecological environment. In this research, we studied the Qinghai-Tibet Plateau (spatial range roughly between 26°00' N–39°47' N latitude and 73°19' E–104°47' E longitude), and used the 14-level (17 m spatial resolution) Google Earth remote sensing images as the basis for constructing a small-scale dataset of lake semantic segmentation for the Qinghai-Tibet Plateau. This dataset consists of 6,774 randomly selected RGB images (256×256 in size), as well as ground truth images (lake water bodies in red, background in black) obtained by combining the deep learning dataset annotation software—labelme and visual interpretation methods. On the one hand, this dataset can be used for the extraction of lake water bodies on the Qinghai-Tibet Plateau, supporting the engineering construction in this region. On the other, it can be used to pre-train the weight of the semantic segmentation model and improve the accuracy of water extraction in other regions.

Keywords: Google Earth; the Qinghai-Tibetan Plateau; deep learning; semantic segmentation

Dataset Profile

Title	A dataset of semantic segmentation for the Qinghai-Tibet Plateau in 2020
Data corresponding author	WANG Zhaobin (wangzhib@lzu.edu.cn)
Data authors	WANG Zhaobin, GAO Xiong, KANG Jianfang, AI Minghao
Time range	2020
Geographical scope	26°00'N–39°47' N, 73°19'E–104°47'E
Spatial resolution	17 m
Data volume	545.8 MB
Data format	*.png
Data service system	< http://dx.doi.org/10.12072/ncdc.NIEER.db0112.2021 >
基金项目	Information Program of the Chinese Academy of Sciences“Construction and Application of ‘Technology Cloud’ in Cold and Arid Regions Environment Evolution” (XXH13506) National Cryosphere Desert Data Center (No.E01Z7902) Capacity improvement of Cryosphere Desert Data Center of Chinese Academy of Sciences (No.Y9298302)
Dataset composition	This dataset consists of two parts: 6,774 RGB Google remote sensing images (256×256 in size); 6,774 ground truth images corresponding to the Google remote sensing images (lake water bodies in red, background in black).