



文献 CSTR:
32001.14.11-6035.ncdc.2025.0108.zh



文献 DOI:
10.11922/11-6035.ncdc.2025.0108.zh
数据 DOI:

10.12072/ncdc.dpsl.db6865.2025
文献分类: 地球科学

收稿日期: 2025-06-04
修回日期: 2025-10-22
录用日期: 2025-12-04
发表日期: 2026-05-31

* 论文通信作者
张耀南: yaonan@lzb.ac.cn

面向 AI-Ready 的黄土高原坝地标准化语义分割数据集

刘景琦^{1,2,3}, 黄波^{1,2,3}, 张耀南^{1,2*}, 敏玉芳^{1,2}, 钱安良^{1,2,3}

1. 中国科学院西北生态环境资源研究院, 兰州 730000
2. 国家冰川冻土沙漠科学数据中心, 兰州 730000
3. 中国科学院大学, 北京 100049

摘要: 淤地坝作为黄土高原地区关键的水土保持工程, 兼具控制土壤侵蚀与保障粮食安全的核心功能, 但其智能化管理长期受限于数据获取效率低、模型泛化能力不足及标准化数据集缺失等技术瓶颈。本研究依托吉林一号 0.75 m 高分辨率遥感影像, 以黄土高原典型流域韭园沟为采样区, 构建了一套面向 AI-Ready 的黄土高原坝地标准化语义分割数据集。本数据集通过系统化样本制备流程, 涵盖格网划分、样本筛选、像素级语义标注及图像增强等关键环节, 形成了包含 2920 个样本的高精度数据集, 有效满足数据空间表征的完整性和算法泛化能力需求。数据质量评估实验显示, 基于该数据集训练的 DenseUnet 等模型在验证集上表现优异, 平均交并比 (mIoU) 超过 80%, 总体精度 (OA) 达 89% 以上, 且与公开数据集相比, 其坝地提取结果的空间匹配性和可靠性显著提升。本数据集实现了复杂地貌环境下坝地与背景的精细化语义区分, 弥补了当前坝地智能识别领域标准化数据集不足的问题, 不仅为淤地坝高精度时空制图、溃坝风险动态评估及水土保持效应量化研究提供了关键数据支撑, 还为人工智能驱动的水土保持工程优化开辟了新的技术路径, 对推进黄河流域生态保护与高质量发展具有重要的实践价值。

关键词: AI-Ready; 语义分割; 深度学习; 淤地坝; 黄土高原

数据库(集)基本信息简介

数据库(集)名称	面向 AI-Ready 的黄土高原坝地标准化语义分割数据集
数据作者	刘景琦, 黄波, 张耀南, 敏玉芳
数据通信作者	刘景琦 (liujingqi@nieer.ac.cn)
数据量	662 MB
数据格式	*.tif, *.png
数据服务系统网址	http://dx.doi.org/10.12072/ncdc.dpsl.db6865.2025 https://www.scidb.cn/doi/10.12072/ncdc.dpsl.db6865.2025
基金项目	国家冰川冻土沙漠科学数据中心 (E01Z7902)
数据库(集)组成	数据集由4个数据文件夹和1个python脚本文件组成, 具体包括: (1) origin 文件夹, 包含未经图像增强的原始的584组TIF-PNG样本对; (2) train, val, test文件夹, 分别存放图像增强后划分的训练集(1752组), 验证集(584组)和测试集(584组), 所有子集均采用TIF格式存储图像、PNG格式存储标签, 严格保持图像-标签映射关系; (3) dataAug.py用于自定义图像增强。

引言

淤地坝是黄土高原地区重要的水土保持工程，自 16 世纪自然形成至 20 世纪人工大规模修建以来，在控制土壤侵蚀、调节水沙关系、提升耕地生产力方面具有不可替代的生态与经济价值^[1]。其形成的坝地虽然仅占黄土高原耕地面积的 9%，却贡献了 20.5% 的粮食产量，且土壤有机质含量显著高于普通耕地^[2-3]。

作为国家三大水土保持工程之一^[4]，淤地坝在服役数十年后，其时空分布、溃坝风险及功能演变缺乏系统性探索^[5]。传统研究多聚焦于工程结构、拦沙机理及经济收益^[6]。2000 年后，虽逐步引入 3S 技术^[7]，但研究仍以案例分析和局部实证为主，缺乏宏观层面的统一认知。近年来，伴随着遥感影像获取与深度学习技术的飞速发展，地物智能识别、提取与分割研究取得了显著进展。然而，针对淤地坝这一特定对象的智能化识别技术研究仍相对薄弱。将高分辨率遥感影像与深度学习方法进行融合，被认为是突破这一瓶颈、实现淤地坝高效精准识别的重要路径，不仅能大幅度提升其空间分布与形态特征信息的提取效率，也有望为水土保持决策与管理提供更为科学、精确的数据支撑^[8-9]。

然而，要实现这一目标，淤地坝的智能化识别技术在实际应用中仍面临三大关键瓶颈：一是数据获取难，淤地坝空间分布零散、占地面积小，普通光学影像难以直接解译，需依赖高分辨率遥感影像或无人机影像^[10-11]，但此类数据的获取成本与效率难以平衡^[12]；二是模型适配性较差，U-Net、DSU-Net 等现有方法^[13]仅适用于特征明显的坝地，对于背景复杂区域的泛化能力较弱^[14]；三是高质量数据集稀缺，开源的黄土高原坝地数据集质量参差不齐，例如基于 Google 影像提取的坝地数据^[15]与我国高分辨率遥感影像目视解译结果对比时，存在明显的系统性空间偏移以及坝地几何形态不完整等问题，难以满足实际应用对数据精度与完整性的要求，亟需挖掘我国海量高分辨率遥感数据资源，构建更精准的坝地数据集。

针对上述问题，本研究依托吉林一号遥感影像，构建了一套面向 AI-Ready 的高精度坝地语义分割数据集。该数据集通过端到端的标准化处理流程，降低了模型训练前的预处理成本。具体涵盖格网裁剪、样本筛选、语义标注、标签转换、数据增强以及数据集划分等规范化环节。在此基础上，经实地核查与多模型验证，数据质量得到了充分保障。依托标准化处理与高质量标注，该数据集可直接输入主流语义分割模型开展训练、评估与部署，体现了“AI-Ready”数据的高效性、可靠性与即用性。此外，本研究弥补了当前坝地智能识别领域标准化数据集不足的问题，为后续相关研究与应用提供了关键数据支撑。具体而言，本数据集在三个方向具有重要应用价值：一是推动坝地语义分割模型技术创新，为开发轻量化、高鲁棒性的坝地专用算法提供数据基础；二是用于构建淤地坝高精度时空分布图，支撑大尺度动态监测与溃坝风险评估；三是助力生态效应量化研究，通过分割结果分析坝地水土保持效率与碳汇潜力。

综上所述，淤地坝作为黄土高原生态治理的核心载体，其智能化管理亟需数据支撑与技术创新的双重突破。本研究构建的数据集经系统性标注与处理，不仅有效弥补了当前研究中的数据缺口，也为人工智能驱动的水土保持工程优化提供了全新路径和技术支点。

1 数据采集和处理方法

1.1 数据采集方法

韭园沟流域（N37°33′–37°38′、E110°16′–110°26′，图 1）作为黄土高原丘陵沟壑区的典型代表，其地

理特征与治理背景具有显著的科学研究价值。选择该流域作为数据采集区，一方面源于其作为黄土高原第一副区地貌特征的典型性^[16]，另一方面则得益于其所在地区坝地系统规模大、类型全（涵盖大中小型坝）且空间分布密集，能够为提取坝地信息提供丰富的样本基础。

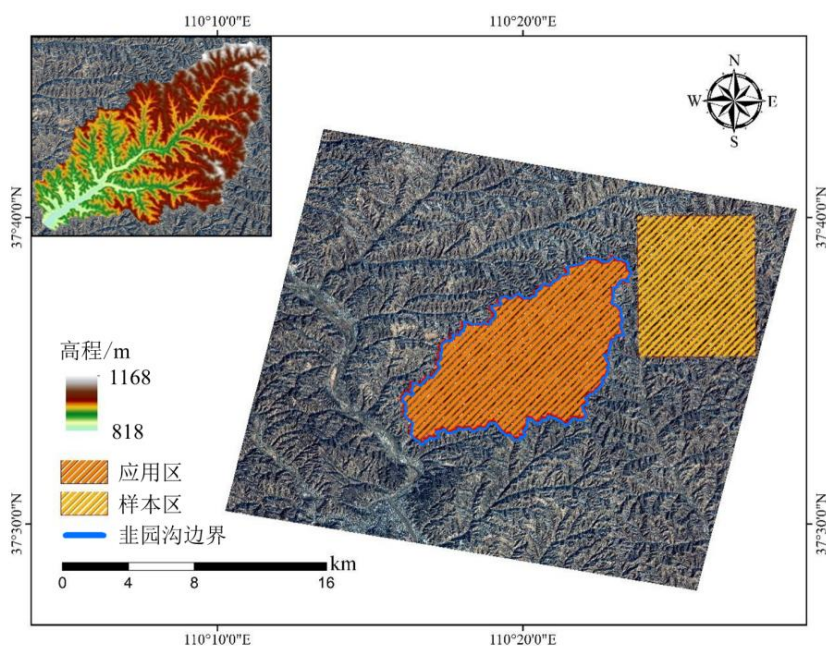


图1 研究区概况图

Figure 1 Overview of the study area

基于上述区域的典型特征，本研究采用吉林一号宽幅 01A 卫星 PMS02 传感器获取的 0.75 m 高分辨率影像作为核心数据源（原始影像数据访问地址：<https://www.jl1mall.com/store/>）。经过多时相检索与质量评估，选定 2021 年 11 月 11 日无云覆盖影像（云量=0%），该数据采用 WGS1984 坐标系 UTM6° 带投影，覆盖面积 548.30 km²，太阳高度角 29.7493°，传感器侧摆角 -2°。该影像的亚米级空间精度可清晰解析淤地坝边界几何形态、坝地结构特征及周边土地利用细节，其强纹理表征能力尤其适应沟壑地貌的语义分割建模需求，为构建标准化坝地样本库提供数据保障，样本制备以流域内特定矩形区域作为样本采集区，保留完整韭园沟流域作为模型应用区。

1.2 数据处理

在语义分割任务中，数据预处理对模型的驱动和训练稳定性起着至关重要的作用。这一过程并非简单的数据清洗，而是通过规范化、增强化、适配化等操作，为模型提供高质量、多样化的输入，从根本上影响模型的学习效率、收敛速度与泛化能力。本研究采用的系统化数据处理流程，在涵盖常规语义标注的基础上，进一步纳入数据筛选、数据增强等关键步骤，旨在优化数据质量，最大限度地减少潜在的数据偏差，从而为后续的模型训练与应用提供可靠的数据基础。数据预处理的完整流程如图 2 所示，具体流程分为六个阶段：（1）格网划分阶段，将样本区原始影像规则裁剪为 256×256 像素网格单元，共生成 3689 个初始样本；（2）样本筛选阶段，通过目视解译剔除不含坝地的无效样本，保留 584 个有效像素网格单元；（3）语义标注阶段，基于 Labelme 开源平台完成全样本矢量化标注，构建以“gully”为统一地物类别的语义分割数据集；（4）标签转换阶段，通过 GDAL 库将 JSON 矢量标注文件批量栅格化为二值图像，确保标签与原始影像空间坐标严格匹配；（5）数据增强阶段，对原始样本实施镜像翻转、旋转及亮度调整等几何-辐射组合变换，将样本量扩展至 2920 个；（6）数据集划分阶段：将增

强后的数据集按 6 : 2 : 2 的比例划分为训练集、验证集和测试集三个独立子集。该标准化流程兼顾样本空间表征完整性与算法泛化需求, 其产出的多尺度增强数据集可支撑主流卷积神经网络进行特征学习, 最终数据分布如表 1 所示。

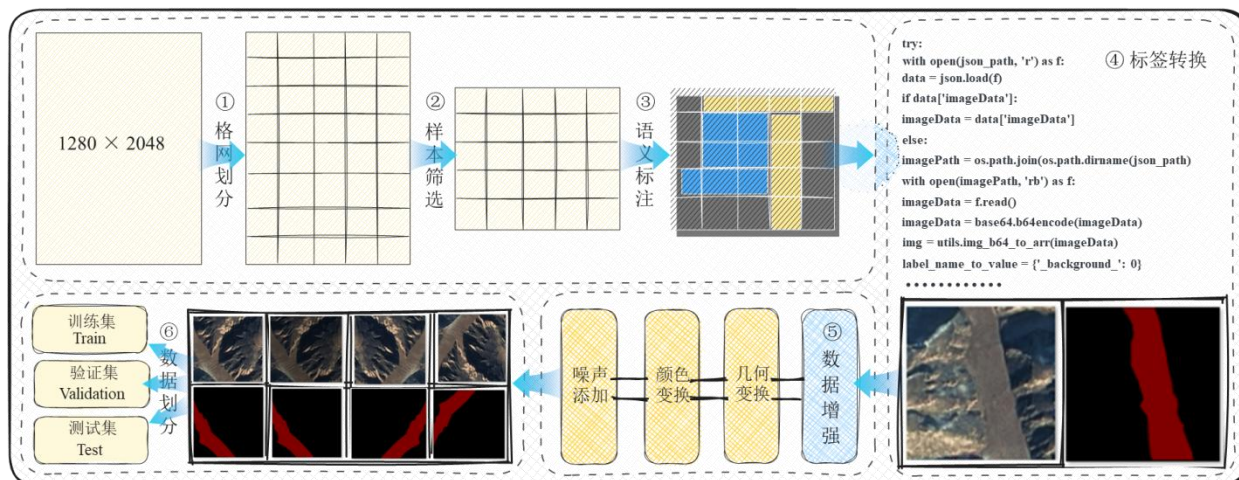


图 2 数据预处理流程

Figure 2 Data preprocessing workflow

表 1 数据分布

Table 1 Dataset distribution

数据	样本图像	增强图像	训练集	验证集	测试集
数据数量	584	2920	1752	584	584
数据大小 (MB)	110	551	328	109	109

1.2.1 格网划分

由于原始遥感图像通常具有较大的尺寸规格, 若直接输入深度学习网络进行处理, 会因数据量过大导致显存溢出。因此, 将其分割为尺寸较小的块单元是不可或缺的预处理环节。本研究采用 256×256 像素的固定尺寸滑动窗口策略, 以无重叠的划分方式对采样区域进行规则化裁剪, 生成相互独立的图像块单元。该操作既能确保空间分辨率的一致性, 又能大幅度提升后续样本处理与模型训练的标准化程度。选择 256×256 像素作为图像块尺寸, 本质上是因为 2 的幂次尺寸 (如 $2^8=256$ 、 $2^9=512$) 在内存分配时能被硬件高效处理, 这类尺寸在内存分配时可充分适配 GPU 的存储机制, 减少计算资源的浪费。本步骤共提取出 3689 个初始样本单元, 这些样本单元完整覆盖了采样区域的全部空间范围, 为后续的影像分析与模型训练工作奠定了坚实的数据基础。

1.2.2 样本筛选

在初始样本中, 存在大量不包含目标地物 (坝地) 的无效样本。为确保训练数据的针对性与有效性, 采用人工目视解译方法对样本进行初筛。以坝地的可识别性与空间显著性为标准, 剔除未包含任何坝地信息的图像块, 仅保留 584 个具有明确坝地特征的有效样本单元。该步骤有助于降低样本噪声, 提高语义分割模型的学习效率。

1.2.3 语义标注

在有效样本基础上，利用 Labelme 开源图像标注平台对每一个图像块中的坝地区域进行精确矢量化标注。所有标注均采用“gully”作为唯一的语义类别，确保标签一致性和语义统一性。标注过程严格遵循人工精细描绘原则，确保坝地边界的完整性与空间精度，为后续构建高质量语义分割数据集奠定基础，语义标注过程如图 3 所示。

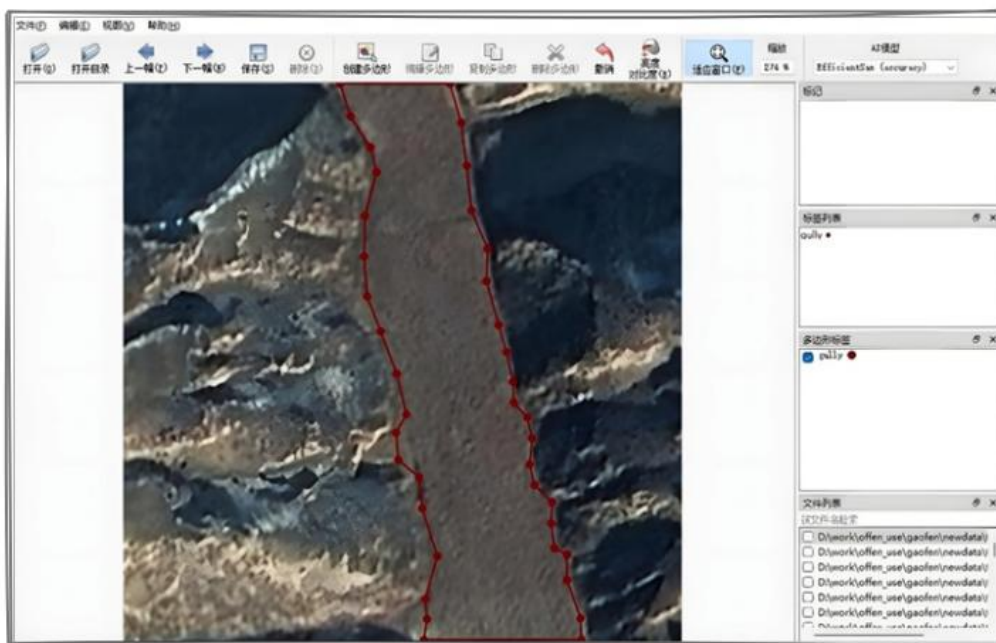


图 3 样本数据标注过程

Figure 3 Sample data annotation process

1.2.4 标签转换

完成矢量标注后，需将 JSON 格式的矢量数据转换为可用于模型训练的栅格标签图像。本阶段采用 GDAL 地理数据处理库对矢量数据进行批量栅格化处理，生成与原始影像分辨率和地理配准严格一致的二值标签图像，如图 4 所示。转化后图像中，坝地区域以像素值 1 表示，其余区域为 0，形成标准的语义分割标签。该过程确保标签数据与输入影像在空间坐标系上的一一对应关系，避免配准误差。转换后的数据标签示例见表 2。

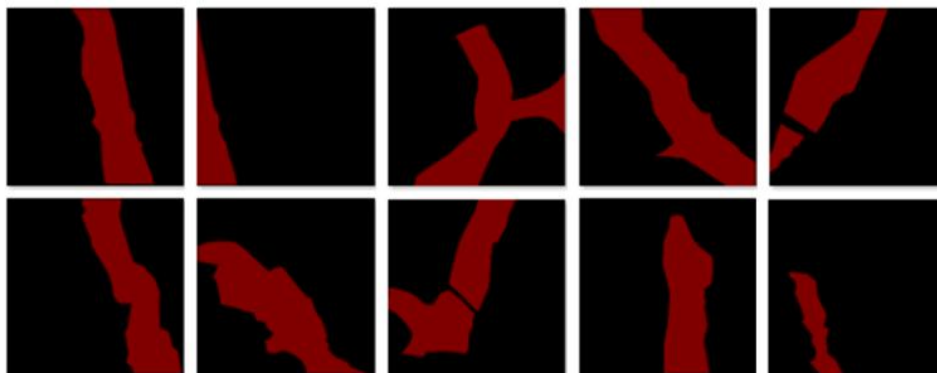


图 4 一组标签转换结果

Figure 4 A set of label conversion results

表 2 数据标签示例表

Table 2 Example of data annotations

序号	标签	像素值	像素点 x	像素点 y
1	gully	1	94.8350	1.2323
2	gully	1	105.9461	17.3939
3	background	0	113.5821	32.1376

1.2.5 数据增强

为提升模型泛化能力并扩充样本数量，采用组合式图像增强技术对原始样本进行处理，具体参数详见表 3。其中，翻转操作通过水平或垂直镜像变换，改变图像空间布局；旋转操作以图像中心为轴心，在 -90° 至 90° 范围内随机调整角度，进一步丰富图像空间朝向变化；亮度调整则将图像亮度在原始亮度的 0.5 至 1.5 倍区间内线性调节，模拟真实场景中的光照差异。这三类操作分别从几何变换（翻转、旋转）与辐射变化（亮度调整）两个维度，使样本在空间结构和光谱特征上产生多样化差异，有效避免模型对特定样本样式的过拟合现象。最终，训练样本数量成功扩展至原始规模的 5 倍，形成包含 2920 组样本-标签对的图像数据集，为模型训练提供了更充足的数据支撑。数据增强示例如图 5 所示。

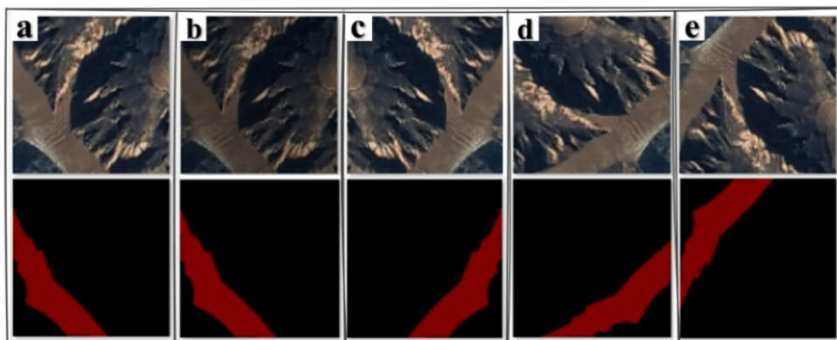
图 5 一组数据增强结果。(a) 原始图像；(b) 亮度调整；(c) 水平翻转；(d) 90° 旋转；(e) 垂直翻转

Figure 5 A set of data augmentation results. (a) Original image; (b) Brightness adjustment; (c) Horizontal flip; (d) 90° rotation; (e) Vertical flip

表 3 数据增强参数表

Table 3 Data augmentation parameters

序号	操作	描述	参数值
1	翻转	水平/垂直的镜像变换	水平/垂直翻转
2	旋转	随机改变图像空间朝向	旋转角度: $(-90^{\circ}, 90^{\circ})$ 旋转中心: $(\text{width}/2, \text{height}/2)$
3	亮度调整	模拟不同光照条件, 提升模型光照鲁棒性	亮度范围: $(0.5, 1.5)$

1.2.6 数据划分

为保证模型训练的科学性与评估的客观性，将完整增强后的数据集按照 6 : 2 : 2 的比例划分为训练集、验证集和测试集三部分。其中，训练集用于模型参数的优化学习，验证集用于超参数调优与模型早停判据的设定，测试集则独立用于模型最终性能评估。最终的数据统计如表 4 所示。

表 4 数据统计表

Table 4 Data statistics

数据	原始图像	增强图像	训练集	验证集	测试集
数据数量	584	2920	1752	584	584
数据大小 (MB)	110	551	328	109	109

2 数据样本描述

本数据集采用层级式目录结构进行组织，确保数据管理的规范性，包含 origin、train、val、test 四个子集。原始数据（origin）由 584 组经过严格筛选的样本-标签对构成，其中高分辨率 TIF 格式图像统一存储于 tif_sample 子目录，采用四位补零连续编号规则，编号为 0000.tif 至 0583.tif。对应的 PNG 格式语义分割标签存储于 label 子目录，编号为 0000.png 至 0583.png，与图像样本精确匹配。

train、val、test 三个子集是原始数据基于 1.2.5 方法进行图像增强后，按 6:2:2 比例进行划分的训练集、验证集、测试集。训练集（train）包含 1752 组样本对，图像以 TIF 格式存储于 image 子目录，编号为 0000.tif 至 1751.tif。对应标签以 PNG 格式存储于 label 子目录，编号与图像一致（0000.png 至 1751.png）。验证集（val）与测试集（test）各包含 584 组样本对，图像与标签的存储格式、目录结构及命名规则与训练集一致。数据集完整层级结构及文件映射关系详见图 6。

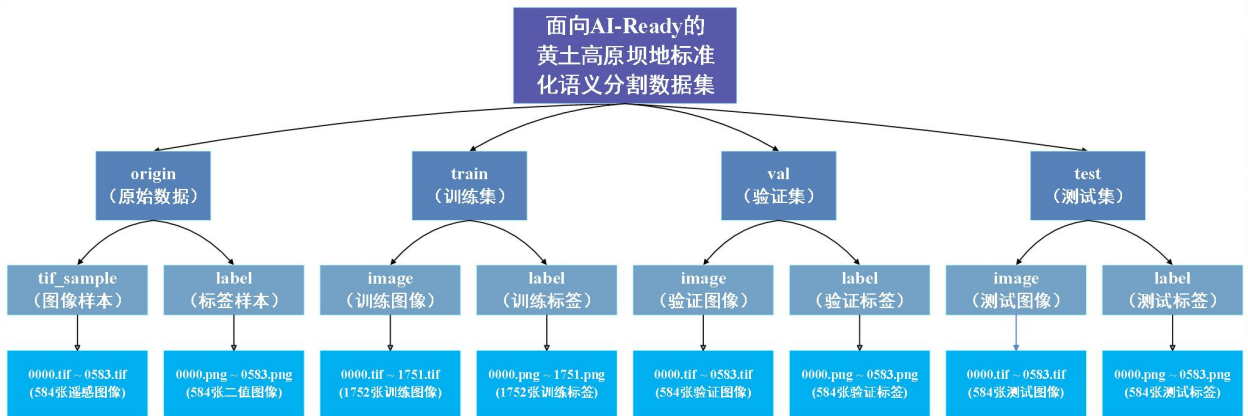


图 6 文件组成与结构关系

Figure 6 Composition and structure of the dataset files

3 数据质量控制和评估

3.1 实地考察验证

为确保坝地语义分割数据的准确性与可靠性，研究团队首先基于遥感影像对样本区进行了系统性的目视解译，提取坝地边界信息。随后，于 2024 年 7 月初开展了野外实地调查，选取 20 余处典型淤地坝进行现场核查，并使用无人机获取坝地影像与位置信息（图 7）。通过将实地调查结果与解译成果进行对比验证，对典型样本中的空间偏差进行了修正，进一步确认了解译结果的准确性与代表性。在此基础上，研究团队将经过实地验证的目视解译数据作为原始样本，采用 256×256 像素网格方式裁剪，最终生成了共计 3689 个样本单元，并依据标准化流程构建了坝地语义分割数据集。由于淤地坝通常是坝地

的起始和核心区域，其识别准确与否直接关系到坝地区域的划定与语义分割的质量。因此，通过野外实地验证强化了解译结果的真实性，从源头上保障了坝地语义标注数据的准确性。

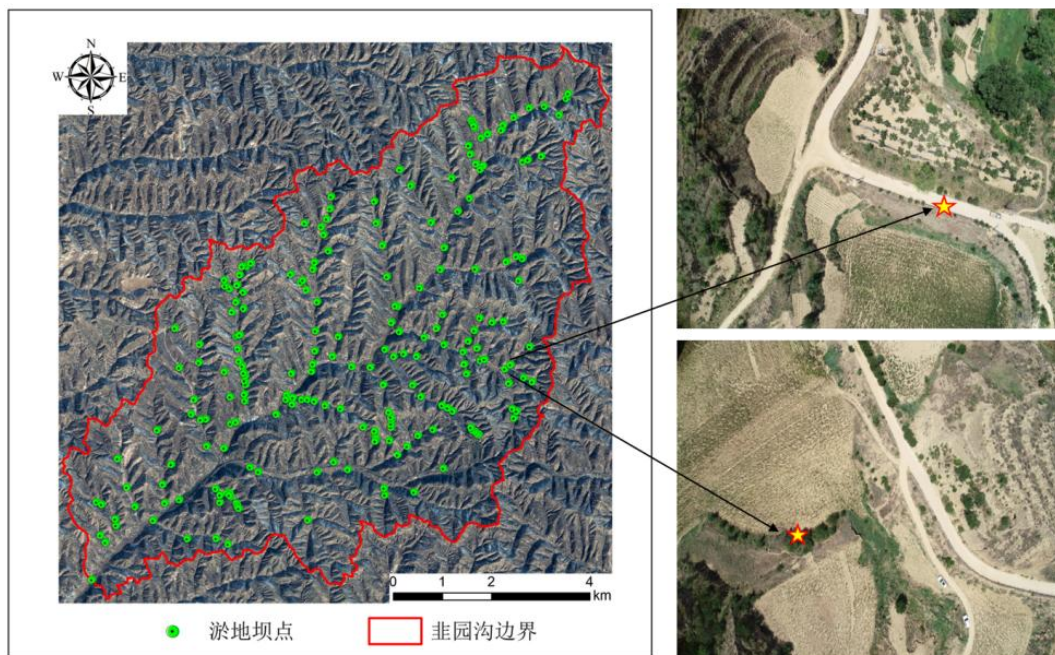


图 7 基于吉林一号数据目视解译淤地坝的空间分布（五角星代表淤地坝实际位置）

Figure 7 Visual interpretation of the spatial distribution of check dams based on Jilin-1 data (the red five-pointed star indicates the actual locations of check dams)

3.2 深度学习模型应用评估

本研究构建了系统性实验框架，旨在验证坝地语义分割数据集的有效性。通过设计控制变量实验组，将不同骨干网络与超参数组合（包括学习率、批处理规模、权重衰减系数及优化器类型）作为独立变量，观察相同训练条件下不同模型在验证集上的收敛特性和泛化表现。实验采用交叉熵损失函数结合余弦衰减学习率策略构建基准训练流程，通过标准化训练范式确保实验结果能够直接映射出数据集的质量特征，为数据集的有效性验证提供科学依据。表 5 结果显示，平均交并比（mIoU）、精确率（P）、召回率（R）、F1 分数（F1）及总体精度（OA）等指标整体呈现一致性，多数模型的 mIoU 超过 80，且 OA 超过 89，表明数据集类别划分清晰、标注质量高、样本分布均衡且多样性充足，对不同模型架构适配性强。实验结果从算法层面进一步证实了该坝地语义分割数据集的有效性与可靠性，也为相关领域的研究与工程应用提供了重要的参考依据。

表 5 不同骨干网络模型在测试集上指标

Table 5 Performance metrics of different backbone network models on the test set

Backbone	mIoU	P	R	F1	OA
Vgg 16 ^[17]	80.56	88.74	89.74	89.24	89.24
GoogLeNet ^[18]	70.5	85.94	79.69	82.7	82.68
Resnet50 ^[19]	83.52	90.54	91.51	91.02	91.02
Densenet121 ^[20]	85.23	91.22	92.85	92.03	92.03
Inceptionv4 ^[21]	83.97	90.62	91.96	91.29	91.28

Backbone	mIoU	P	R	F1	OA
InceptionResnet2 ^[21]	84.2	89.93	92.96	91.42	91.42
Resnext ^[22]	84.34	90.85	92.17	91.51	91.51
Convnext ^[23]	81.41	89.61	89.89	89.75	89.75
Resvit50 ^[24]	83.47	90.09	91.91	90.99	90.99

3.3 坝地提取结果与公开坝地数据对比验证

目前，公开获取的坝地矢量数据^[15]在与吉林一号卫星遥感影像叠加时存在明显的偏移和残缺等现象。本研究依托吉林一号遥感影像构建高质量坝地数据集，并以此为基础训练的最优模型 DenseUnet 在韭园沟流域进行坝地提取实验。结果显示，相比当前公开数据集，模型提取结果在空间分布上与实际地物更加吻合，且在精度和可靠性方面表现更优。如图 8 所示，A 图与 B 图中的公开坝地数据几何形态不完整，坝地轮廓断裂，而本研究提取结果能够准确反映典型坝地的空间形态，C 图中基于公开数据集的提取结果则出现了明显的空间偏移。该对比充分证明，本研究构建的数据集具备更高精度，大幅度提升了模型的特征提取效果，体现了数据集的高质量与应用价值。

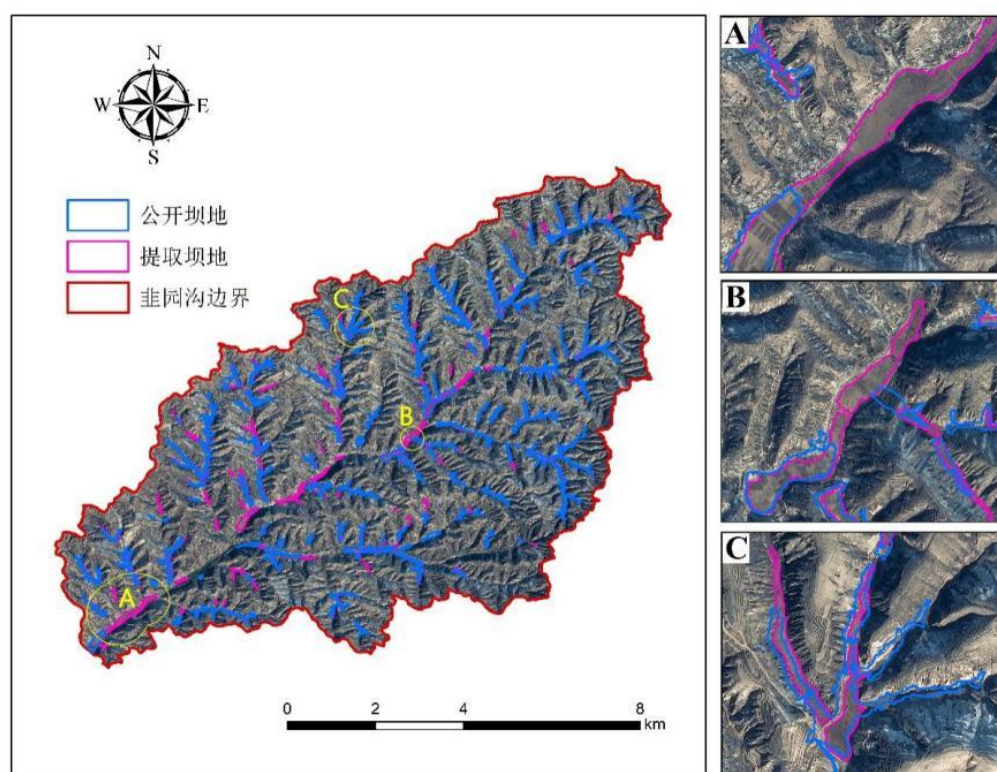


图 8 比较不同的坝地数据

Figure 8 Comparison of different check dam land data

4 数据价值

本研究构建的面向 AI-Ready 的黄土高原坝地标准化语义分割数据集，具有多方面的数据价值：

(1) 高精度坝地数据集。当前国际公开的黄土高原坝地数据集质量参差不齐，尤其是基于 Google

影像提取的数据,与我国高分辨率遥感数据存在明显空间偏移,难以满足实际应用需求。本研究基于吉林一号遥感影像,构建了高精度、标准化的坝地语义分割数据集,精准实现坝地与背景的像素级区分,在复杂地貌条件下依然具备良好适应性,有效提升遥感影像中坝地智能识别的能力。

(2) 标准化数据资源。该数据集弥补了当前坝地智能识别领域标准化数据集不足的问题,为坝地智能识别领域 提供了标准化数据集支撑。其高质量标注和规范化构建大幅度提升了研究的可复现性,也为未来构建跨区域、跨时相的坝地监测数据提供了技术基准和数据参考。

(3) 赋能多尺度监测与风险生态协同评估。一方面,基于该数据集训练的坝地语义分割模型,可实现淤地坝高精度的时空分布图构建,结合溃坝易发性影响因子,开展溃坝风险评估,支撑大尺度坝地分布的动态监测与更新管理;另一方面,通过分析坝地分割结果的空间特征,可定量评价水土保持效能,基于流域源-汇系统理论框架,系统测算土壤碳汇潜力,为黄河流域生态保护工程提供多维度数据支撑。

AIGC 使用声明

在撰写本手稿期间,作者仅使用 Qwen3.5 进行部分语法检查、语言润色和语句流畅度优化。所有生成的内容均由作者进行修改和确认,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

刘景琦(1997—),男,甘肃平凉人,博士,主要研究方向为数据工程防灾减灾。主要承担工作:数据审核和论文撰写。

黄波(1997—),男,四川资阳人,硕士,主要研究方向为数据工程防灾减灾。主要承担工作:数据采集,标准化处理和模型验证。

张耀南(1966—)男,甘肃天水人,研究员,博士生导师,主要研究方向为地学数据工程及数据工程防灾减灾、基于高性能计算环境的地学模型模拟、遥感图像处理及多源数据融合。主要承担工作:论文修改。

敏玉芳(1983—),女,甘肃省甘南州人,高级工程师,硕士生导师,主要研究方向为地学大数据管理与分析、基于人工智能的遥感数据信息提取。主要承担工作:数据质量控制。

钱安良(2001—),男,河南三门峡人,硕士,主要研究方向为数据工程防灾减灾。主要承担工作:数据标注。

参考文献

- [1] JIN Z, CUI B L, SONG Y, et al. How many check dams do we need to build on the Loess Plateau?[J]. Environmental Science & Technology, 2012, 46(16): 8527–8528. DOI:10.1021/es302835r.
- [2] XU X Z, ZHANG H W, OUYANG Z. Development of check-dam systems in gullies on the Loess Plateau, China[J]. Environmental Science & Policy, 2004, 7(2): 79–86. DOI:10.1016/j.envsci.2003.12.002.
- [3] XU X Z, ZHANG H W, WANG G Q, et al. A laboratory study on the relative stability of the check-dam system in the Loess Plateau, China[J]. Land Degradation & Development, 2006, 17(6): 629–644. DOI:10.1002/ldr.740.
- [4] LI X G, WEI X, WEI N. Correlating check dam sedimentation and rainstorm characteristics on the Loess

- Plateau, China[J]. *Geomorphology*, 2016, 265: 84–97. DOI:10.1016/j.geomorph.2016.04.017.
- [5] 刘晓燕, 高云飞, 马三保, 等. 黄土高原淤地坝的减沙作用及其时效性[J]. *水利学报*, 2018, 49(2): 145–155. DOI:10.13243/j.cnki.slxb.20170925. [LIU X Y, GAO Y F, MA S B, et al. Sediment reduction of warping dams and its timeliness in the Loess Plateau[J]. *Journal of Hydraulic Engineering*, 2018, 49(2): 145–155. DOI:10.13243/j.cnki.slxb.20170925.]
- [6] WEI X, LI X G, WEI N. Fractal features of soil particle size distribution in layered sediments behind two check dams: Implications for the Loess Plateau, China[J]. *Geomorphology*, 2016, 266: 133–145. DOI:10.1016/j.geomorph.2016.05.003.
- [7] 党维勤. 3S 技术在淤地坝建设工作中的应用[C]. “全国水土流失与江河泥沙灾害及其防治对策”学术研讨会会议文摘. 武汉, 2003: 101. [DANG W Q. Application of 3S technology in check dam construction[C]. *Proceedings of the National Academic Symposium on Soil Erosion and River Sediment Disaster and Its Prevention and Countermeasures*. Wuhan, 2003: 101.]
- [8] SUN L Q, GUO H L, WANG H M, et al. Deep learning for check dam area extraction with optical images and digital elevation model: a case study in the hilly and Gully Regions of the Loess Plateau, China[J]. *Earth Surface Processes and Landforms*, 2023, 48(14): 2645–2660. DOI:10.1002/esp.5652.
- [9] 张大旗. 淤地坝在国产高分辨率卫星影像中的解译特征研究[J]. *内蒙古水利*, 2025(3): 105–107. [ZHANG D Q. Study on interpretation characteristics of warping dam in domestic high-resolution satellite images[J]. *Inner Mongolia Water Resources*, 2025(3): 105–107.]
- [10] DAI W, QIAN W, LIU A L, et al. Monitoring and modeling sediment transport in space in small loess catchments using UAV-SfM photogrammetry[J]. *Catena*, 2022, 214: 106244. DOI:10.1016/j.catena.2022.106244.
- [11] 李云飞. 融合多源数据和地形特征的黄土高原流域淤地坝识别研究 [D]. 西安: 长安大学, 2023. [LI Y F. Recognition of check dams in Loess Plateau watersheds by integrating multi-source data and topographic features [D]. Xi'an: Chang'an University, 2023.]
- [12] 马煜栋, 杨帅, 韩静, 等. 陕西榆林地区无定河流域淤地坝遥感解译[J]. *干旱区地理*, 2022, 45(3): 786–791. DOI:10.12118/j.issn.1000-6060.2021.425. [MA Y D, YANG S, HAN J, et al. Remote sensing interpretation of check dams in Wuding River Basin in Yulin Area of Shaanxi Province[J]. *Arid Land Geography*, 2022, 45(3): 786–791. DOI:10.12118/j.issn.1000-6060.2021.425.]
- [13] 侯景伟, 朱默研, 侯博. 基于 DSU-Net 的黄土高原淤地坝智能识别[J]. *湖南科技学院学报*, 2023, 44(5): 23–27. DOI:10.16336/j.cnki.cn43-1459/z.2023.05.020. [HOU J W, ZHU M Y, HOU B. Intelligent identification of warping dams on Loess Plateau based on DSU-Net[J]. *Journal of Hunan University of Science and Engineering*, 2023, 44(5): 23–27. DOI:10.16336/j.cnki.cn43-1459/z.2023.05.020.]
- [14] 孙立全. 黄土丘陵沟壑区淤地坝遥感监测及泥沙调控模拟研究[D]. 陕西杨陵: 西北农林科技大学, 2024. [SUN L Q. Remote sensing monitoring and sediment regulation simulation of check dams in the hilly and gully region of the Loess Plateau [D]. Yangling, Shaanxi: Northwest A&F University, 2024.]
- [15] ZENG Y, JING T G, XU B D, et al. Vectorized dataset of silted land formed by check dams on the Chinese Loess Plateau[EB/OL]. (2024-04-06) [2024-05-28]. <https://www.nature.com/articles/s41597-024-03198-z>.
- [16] 徐乃民. 韭园沟流域的治理与试验 [J]. *中国水土保持*, 1984 (11): 40-42. DOI:10.14123/j.cnki.swcc.1984.11.018. [XU N M. Treatment and experiment of Jiuyuanguou watershed [J].

- Soil and Water Conservation in China, 1984 (11): 40-42. DOI:10.14123/j.cnki.swcc.1984.11.018.]
- [17] MAGGIORI E, TARABALKA Y, CHARPIAT G, et al. Convolutional neural networks for large-scale remote-sensing image classification[J]. IEEE Transactions on geoscience and remote sensing, 55(2), 645-657. DOI: 10.1109/TGRS.2016.2612821
- [18] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. IEEE, 2015: 1-9. DOI:10.1109/CVPR.2015.7298594.
- [19] 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.
- [20] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 2261-2269. DOI:10.1109/CVPR.2017.243.
- [21] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 2818-2826.. DOI:10.1109/cvpr.2016.308.
- [22] XIE S N, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. IEEE, 2017: 5987-5995. DOI:10.1109/CVPR.2017.634.
- [23] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA. IEEE, 2022: 11966-11976. DOI:10.1109/CVPR52688.2022.01167.
- [24] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]. In Proceedings of the 9th International Conference on Learning Representations, Vienna, Austria, May 04 2021; pp. no.

论文引用格式

刘景琦, 黄波, 张耀南, 等. 面向 AI-Ready 的黄土高原坝地标准化语义分割数据集[J/OL]. 中国科学数据, 2026, 11(2). (2026-05-31). DOI: 10.11922/11-6035.ncdc.2025.0108.zh.

数据引用格式

刘景琦, 黄波, 张耀南, 等. 面向 AI-Ready 的黄土高原坝地标准化语义分割数据集[DS/OL]. 国家冰川冻土沙漠科学数据中心(<http://www.ncdc.ac.cn>), 2025. (2025-06-04). DOI: 10.12072/ncdc.dpsl.db6865.2025.

A dataset of AI-Ready standardized semantic segmentation for silted land formed by check dam in the Loess Plateau

LIU Jingqi^{1,2,3}, HUANG Bo^{1,2,3}, ZHANG Yaonan^{1,2*}, MIN Yufang^{1,2}, QIAN Anliang^{1,2,3}

1. Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, P.R. China
2. National Cryosphere Desert Data Center, Lanzhou 730000, P.R. China
3. University of Chinese Academy of Sciences, Beijing 100049, P.R. China

*Email: liujingqi@nieer.ac.cn

Abstract: As a critical soil and water conservation project in the Loess Plateau region, check dams play a central role in controlling soil erosion and ensuring regional food security. However, their intelligent management has long been constrained by technical bottlenecks such as inefficient data acquisition, insufficient model generalization capabilities, and lack of standardized datasets. This study utilized 0.75-meter high-resolution Jilin-1 remote sensing imagery and selected the Jiuyuangou watershed, a typical watershed in the Loess Plateau. Through a systematic sample preparation workflow, including grid division, sample screening, pixel-level semantic annotation, and image enhancement, the dataset comprises 2,920 high-precision samples that comprehensively ensure the completeness of spatial representation and satisfy the requirements for algorithm generalization. Data quality evaluation experiments demonstrated that models trained on this dataset (e.g., DenseUnet) achieved excellent performance on the validation set, with a mean Intersection over Union (mIoU) exceeding 80% and Overall Accuracy (OA) reaching over 89%. Compared with public datasets, the spatial consistency and reliability of silted land formed by check dam extraction results were significantly improved. This dataset enables fine-grained semantic differentiation between check dam sedimentation land structures and surrounding backgrounds under complex geomorphological conditions, addressing the lack of standardized datasets for intelligent identification of silted land formed by check dams. It not only provides crucial data support for high-precision spatiotemporal mapping of silted land formed by check dams, dynamic assessment of dam-break risks, and quantitative analysis of soil conservation effects, but also opens new technical pathways for AI-driven optimization of soil and water conservation engineering. These advancements hold substantial practical value for promoting ecological protection and high-quality development in the Yellow River Basin.

Keywords: AI-Ready; semantic segmentation; deep learning; check dam; the Loess Plateau

Dataset Profile

Title	A dataset of AI-Ready standardized semantic segmentation for silted land formed by check dam in the Loess Plateau
Data authors	LIU Jingqi, HUANG Bo, ZHANG Yaonan, MIN Yufang
Data corresponding author	LIU Jingqi (liujingqi@nieer.ac.cn)
Data volume	662 MB
Data format	*. tif, *. png

Data service system	http://dx.doi.org/10.12072/ncdc.dpsl.db6865.2025 https://www.scidb.cn/doi/10.12072/ncdc.dpsl.db6865.2025
Sources of funding	The National Cryosphere Desert Data Center (E01Z7902)
Dataset composition	The dataset consists of four data folders and one Python script, specifically including: (1) the “origin” folder, which contains 584 original TIF-PNG sample pairs without image enhancement; (2) the “train”, “val”, and “test” folders, which store the training set (1,752 sample pairs), validation set (584 sample pairs), and test set (584 sample pairs) divided after image augmentation, respectively. All subsets store images in TIF format and labels in PNG format, while strictly preserving the correspondence between images and their associated labels; (3) the file dataAug.py is used for user-defined image augmentation.