



文献 CSTR:

32001.14.11-6035.nasdc.2021.0058.zh

文献 DOI:

10.11922/11-6035.nasdc.2021.0058.zh

数据 DOI:

10.12205/E0290.20211018.26.cs.1909

文献分类: 农学

收稿日期: 2021-11-04

开放同评: 2022-03-18

录用日期: 2022-04-11

发表日期: 2022-07-20

1. 宁夏大学信息工程学院, 银川 750021
2. 中国农业科学院农业信息研究所, 北京 100081
3. 中国农业科学院国家南繁研究院, 海南三亚 572024
4. 国家农业科学数据中心, 北京 100081

摘要:枸杞作为宁夏重要的经济作物之一,被多种虫害寄生且产量极易受到影响,因此提高枸杞虫害防治能力对于稳固当地经济具有十分重要的意义。在大数据时代背景下,图像、文本等多模态数据爆发式增长,传统的单模态检索已不能满足人们多元化信息检索的需求。跨模态检索针对图文等多模态数据进行互检,更符合人们全面灵活的信息检索要求,构建可供跨模态检索技术使用的相关数据集对于农作物虫害防治具有非常重要的实际意义。为此我们构建了枸杞虫害图文跨模态检索数据集,包含枸杞虫害图像采集、文本撰写、数据增广、分类、图文对应等多方面内容。本数据集有 17 类常见枸杞病虫害图像文本数据共 492 MB,其中虫害图像数据及与其对应的文本描述文件各为 9496 份,数据充足可供图文跨模态检索训练样本使用。本数据集将为枸杞虫害跨模态检索提供宝贵的基础数据资源,同时可作为农业领域大数据环境下机器学习的标准数据集,对促进跨模态在农业领域的发展和枸杞虫害防治研究,提高枸杞产量具有重要的实际应用价值。

关键词: 枸杞虫害; 跨模态检索; 训练样本; 标准图文库

数据库(集)基本信息简介

数据库(集)名称	2020年宁夏枸杞虫害图文跨模态检索数据集
数据作者	刘立波
数据通信作者	刘立波(liulib@163.com)、王晓丽(wangxiaoli@caas.cn)
数据时间范围	2020年
地理区域	中国宁夏境内
数据量	492 MB
数据格式	*.jpg, *.txt
数据服务系统网址	http://dx.doi.org/10.12205/E0290.20211018.26.cs.1909
基金项目	国家自然科学基金(61862050); 农业基础性长期性科学数据治理与挖掘平台构建,中国农业科学院院级基本科研业务费(Y2022LM20)
数据库(集)组成	本数据集共三部分数据组成:其一为增广的图像,包含17个数据文件,为尺蠖、大青叶蝉、枸杞负泥虫、枸杞黑盲蝽、黑圆角蝉等17种常见枸杞虫害高质量图像,平均每类约558张。每个JPG文件以流水号命名,共计9496张。其二为增强文本数据,共包含17个TXT文件,均以流水号命名,且流水号与对应图像数据一致,文本数据共计9496条。其三为未经增广处理的原始图像和文本数据,共包含两个数据文件,分别为1900张图像原始数据与1900条英文文本的原始数据。

* 论文通信作者

刘立波: liulib@163.com

王晓丽: wangxiaoli@caas.cn

引言

枸杞原产地位于宁夏，且在我国有着二千余年的药用史，具有调节自身免疫、滋肾、润肺、补肝之作用，在国内外享有很高的声誉。由于宁夏区域气候条件变化以及栽培技术不断引进，近年来枸杞生长环境得到改善，种植面积逐渐扩大，已成为宁夏乃至整个西北地区重要的经济作物之一。枸杞属于多种虫类的寄主而且抗虫害能力差，极易遭受虫害侵扰。虫害对枸杞产量及品质影响巨大，严重影响了生产效益^[1-3]。因此，快速准确检索枸杞虫害多方面信息并及时给予精准防治，对避免虫害进一步扩散、提高枸杞产量与品质、推进枸杞产业发展从而带动区域经济发展至关重要。

传统的农作物虫害检索主要是使用人力将虫害与农作物虫害图像图谱进行对比或者使用单模态检索技术获取相关信息。但是，随着大数据时代的到来，来自文本、图像和视频等不同模式的数据正在以前所未有的速度增长，多种模态数据之间互相关联、互相补充，农业领域中的作物虫害信息检索方式也更加多元化，传统模式已无法满足人们全面灵活的检索要求。将跨模态检索技术引入农业领域，从模态不同但是语义相互关联的数据中获取有价值的信息，对满足人们对农作物虫害多样化检索需求具有重要意义。

本文将跨模态检索技术引入枸杞虫害防治领域，实现不同模态之间的信息互检，使得检索方式多元化，从而获得更多模态量化信息，对枸杞虫害及时防治提供帮助。这不仅有助于减少非必要人力、物力的资源浪费，而且有利于展开精准虫害防治，提高枸杞产业链带来的收益。但目前跨模态检索尚未在农业领域应用，且我国还没有建立起可供深度学习使用的农作物病虫害跨模态研究标准数据集^[4]。因此，为了支撑跨模态检索在农业领域内的研究，本文构建了枸杞虫害图文跨模态检索数据集，涵盖了枸杞虫害图像采集、文本撰写、数据增强等多方面内容。本数据集目前共包含 17 类常见枸杞病虫害图像文本数据共 492 MB，数据充足可供图文跨模态检索研究使用。

1 数据采集和处理方法

1.1 数据来源

本数据集以 17 种宁夏地区常见的枸杞虫害为研究对象，具体为：尺蠖、大青叶蝉、负泥虫、黑盲蝽、黑圆角蝉、红斑芫菁、红长蝽、红缘天牛、黄斑大蚊、蓟马、毛跳甲、实蝇、印度谷螟、蛀果蛾、小地老虎、蚜虫、木虱。主要采用实地调研拍照、网络爬虫、书籍扫描 3 种采集方法构建图像数据子集，占比依次为 20%、50%、30%。

(1) 实地调研拍照

团队使用 Canon EOS 1D X Mark III 数码单反相机，在宁夏回族自治区中卫市中宁县舟塔乡万亩枸杞种植基地进行实地调研，并在专家指导下于自然露天环境下对虫害拍照，且保证了自然露天条件下虫害受光均匀且位于画面中央主要位置。

(2) 网络爬虫

在 python 环境下利用 Scrapy 在维基百科 (http://en.volupedia.org/wiki/Main_Page)，百度百科 (<https://baike.baidu.com/>)，药用植物病虫害数据库 (<https://www.pests.com.cn/>)，植保驿站 (<http://www.51agritech.com/>) 4 个公开网站进行图片爬取，并删除低分辨率图像。其中这 4 个网站爬虫最终得到数据占比分别约为 30%、50%、10%、10%。

(3) 书籍扫描

利用 Huawei p40 pro 自带文档扫描功能，扫描并保存现存农业书籍与图谱^[5-6]中的相关枸杞虫害图片。

在构建数据集的文本数据时，团队通过翻阅相关书籍、网络检索以及专家描述对每类枸杞虫害的原始图像分别撰写包含学名简介、来源分布、生活习性、防治方法等文本描述。

1.2 数据处理

针对自建枸杞虫害跨模态图文检索数据集学习样本过少,在复杂网络中容易发生过拟合的问题,本数据集在数据预处理过程中采用数据增广技术对原始数据进行扩充。数据增广后可以使数据集更具多样性,从而减少在复杂网络训练中产生过拟合的可能性,提高模型泛化能力。

对图像数据进行增广处理主要通过通过对原始图像进行垂直翻转、随机调整亮度、随机裁剪、随机旋转得到 4 类增强图像数据共计 7596 张,部分图像扩增前后对比如图 1 所示。



图 1 枸杞虫害图像数据增广操作

Figure 1 Data augmentation of a wolfberry pest image

采用文本分类任务的简单数据增强(Easy Data Augmentation for Text Classification Task, EDA)方法对原始数据集中的每一个文本描述分别进行随机插入、同义词替换、随机删除以及随机交换得到 4 类增强文本数据共计 7596 条,部分文本增强结果如图 2 所示。

```
Original data:
0 This measuring worm lies on a tender green bud with four prominent ridges.
It eats into the buds of plants and does so much damage to crops that it is
sprayed with chemicals to kill them.
Synonyms Replace:
0 This measuring worm lying in a tender green bud with four prominent ridges.
It eats into the buds of plants and does so much harm to crops that it is
sprayed with chemicals to eliminate them.
Randomly Insert:
0 This measuring worm lying in a tender bright green bud with four prominent ridges.
It eats into the buds of plants and does so much harm to crops.It affects the rate
at which it grows.That it is sprayed with chemicals to eliminate them.
Randomly Swap:
0 This measuring worm lying in a tender bright green bud with four prominent ridges.
It affects the rate at which it grows.It eats into the buds of plants and does so
much harm to crops.That it is sprayed with chemicals to eliminate them.
Randomly Delete:
0 This measuring worm lies on a tender green bud with four prominent ridges.
It eats into the buds of plants and does harm to crops.It affects the rate at
which it grows.That sprayed with chemicals.
```

图 2 枸杞虫害文本数据增强操作

Figure 2 Text data augmentation of wolfberry pests

将预处理后的枸杞虫害数据,按照虫害类别进行区分,首先给 17 类枸杞虫害分配类别标签,标签 0 为尺蠖,1 为大青叶蝉,以此类推标签 16 为蛀果蛾;最后以 Wikipedia 数据集^[7]结构为基准,构建枸杞虫害图像-文本对列表形成以虫害类别为判别标准的多个图像数据子集。部分图文对应如图 3 所示。

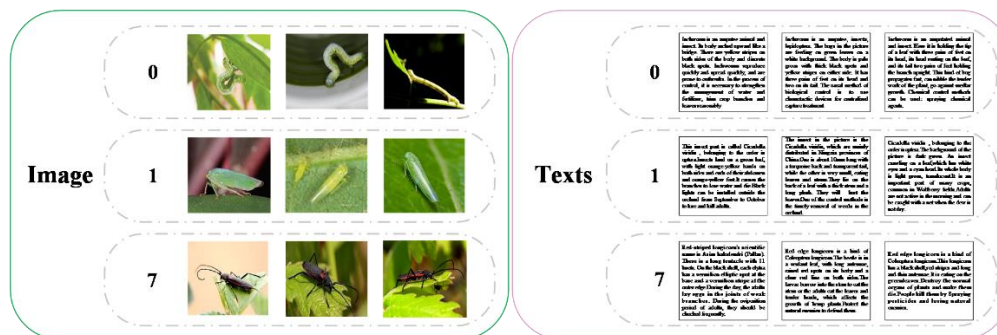


图 3 自建枸杞虫害数据集部分类别图像及对应文本示例

Figure 3 The image categories and text samples from the dataset

2 数据样本描述

2.1 图像数据描述

为确保数据的一致性在采集时统一收录高质量 JPG 格式图像，图像数据包含增广与原始数据，其中原始数据在路径为“gouqi\raw_data\raw_img”文件中，共包含 1900 张 JPG 文件。增广后的图像按类别存放在“gouqi\img_reinforced\split_img”文件夹中，其中有 990 张尺蠖图像、1005 张大青叶蝉图像、540 张枸杞负泥虫图像、600 张黑圆角蝉图像、860 张枸杞红长蝽等 17 类常见枸杞虫害图像共计 9496 张，平均每类约为 558 张。文件夹中每张图像命名方式为“虫害名_流水号.jpg”。部分图像数据如图 4 所示。



图 4 部分图片数据样本
Figure 4 Image data samples

2.2 文本数据描述

本数据集文本数据主要来源为相关书籍翻阅、专家撰写等，在得到文本数据后建立以虫害类别为判别标准的多个文本数据子集。文本数据包含扩增与原始数据，其中包含与图像一一对应的 1900 条中文原始数据与 1900 条英文原始数据，分别在路径为“raw_data\raw_text\ChineseText”，“raw_data\raw_text\EnglishText”的文件中。在对原始英文文本进行随机同义词替换、随机同义词插入、随机单词交换以及随机单词删除操作获得增强数据共 7596 条与原始数据共同存在路径为“gouqi\text_reinforced\split_text”文件中，共计 9496 条文本数据。为与图像数据形成数据对，文本数据命名方式为“虫害名_流水号.txt”。表 1 对数据集的详细信息进行列举，包括每类虫害图文数据对、虫害对应文件夹名称及图文数据量。

表 1 数据集详细信息列举

Table 1 A list of dataset details

虫害名 文件夹名 数据量	图文数据对示例	虫害名 文件夹名 数据量	图文数据对示例
尺蠖 chihuo, 图:990,文:990	 chihuo_27.jpg estimating worms belongs to artemisia forest diet leaf class hypodermis observation is likely the first on dark green leaves and looked like dead branches its body was extremely with a winking wave and true ability lost at the end of its appetite is large facilitates moving usually in the summer and autumn because the growth and product quality of lycium leafhopper through the use of insecticide can have effective control chihuo_27.txt	蓟马 jima 图:830,文:830	 jima_441.jpg lilya pest is called thrips belonging tree leaf family thrips and belongs to the family thrips and there are five fresh thrips in the picture their body color is yellow brown without wings and small black tip on the tail is not apparent it is brown yellow like the with black in the upper left leafhopper pest is fast moving and has a variety of serious field and quality strengths the thrips which fertilizes and water to promote cross growth of plants and reduce harm jima_441.txt
大青叶蝉 daqingyechan 图:1005,文:1005	 daqingyechan_27.jpg harm the insect is cradella pull up status viridis a member of the family thymus leafhopper family diet leaf the insect landed on the green leaves in body was yellow green and the head was yellow in the head and there were small black spot on the left and right side of the wings light blue above the antennae circular and there is a pair of black spots between the two eyes cicada harm the stimulant leaves and stems of many plants making them necrotic or wilted you can use light to trap and kill and use light traps to trap and kill the blackened adults daqingyechan_27.txt	毛跳甲 maotiaojia 图:510,文:510	 maotiaojia_264.jpg this pestifer is called genus lycium thymus and belongs to the family lygidae in the picture the pest is an round and the color is red plus black where black curved red the background is a leaf that has been eaten where the eaten area is exposed in red when the thymus wolfberry germinates in april the growing worms eat raw buds and deminish the growth point clean the postural area remove use stable and other ever-wetting hidden areas and bold down the source of spring worms maotiaojia_264.txt

3 数据质量控制和评估

为确保数据质量，本数据集在图片采集时严格把控，力争图片来源可靠。在图片采集中以人工实地拍摄照片与利用仪器扫描相关书籍图片两种方式所得到的枸杞虫害图片完全能够保障分类准确。在利用网络爬虫技术获取枸杞虫害图像数据过程中，获取图片后与专家进行人工筛选，剔除错误分类的文件，确保了本数据集在建立过程中虫害图像数据来源的质量和可靠性。

3.2 数据质量评估

中国科学数据, 2022, 7(3) | 5

认为最为优秀的有标签公共数据集的 Wikipedia 数据集进行对比，Wikipedia 数据集实验结果见表 3。

表 2 自建数据集实验结果

Table 2 Experiment results from the self-created Wikipedia

方法	图像检索文本	文本检索图像	平均值
CCA	0.274	0.252	0.263
Corr-AE	0.431	0.420	0.426
MCSM	0.466	0.414	0.447

表 3 Wikipedia 数据集实验结果

Table 3 Experiment results from the Wikipedia

方法	图像检索文本	文本检索图像	平均值
CCA	0.267	0.222	0.245
Corr-AE	0.402	0.395	0.399
MCSM	0.516	0.458	0.487

其中表 2 实验数据来源于本实验室实验结果，表 3 实验数据来源于论文《基于关联约束的对抗跨模态检索方法》^[11]。

通过对比可知在某些方法上在本数据集上检索性能优于 Wikipedia 数据集，在另外方法上差距较小，可见本数据集在与 Wikipedia 数据集对比之下也具有一定的优势，因此在定量评估下也展现出了在本数据集在建立时充分保障了数据的质量。

4 数据价值

本数据集与现有农业虫害数据集相比不仅包含了虫害图像，还有与图像一一对应的文本描述，其不仅能为宁夏枸杞的虫害防治研究领域提供基础数据资源，同时也可以弥补国内枸杞虫害数据集的空缺。随着本数据集的发展，今后还将进一步建设并形成全国性的枸杞虫害跨模态检索研究数据资源，以便为相关研究人员提供统一数据，方便不同算法在同一标准下进行比较。本数据集不仅能够促进枸杞虫害跨模态检索的发展，并且能进一步对枸杞虫害的及时防治提供帮助。此外，“第十三个五年规划纲要”指出要推动信息技术与农业农村全面深度融合，确保农业农村信息化发展取得明显进展。本数据集可为宁夏枸杞虫害防治系统提供研究基础，进一步推动宁夏地区农业生产过程信息化。

5 数据使用方法和建议

本数据集包含大量图像与文本描述，可用于图像分类、跨模态检索等研究。在使用时根据研究任务仅选取图像或使用全部数据，根据需要进行训练、测试数据划分。其中原始及增广后的图像足以应对小规模图像研究，单独的文本数据不能够用于自然语言处理研究中，主要原因在于增广后的文本存在部分单词的缺失。增广后的图文数据主要在跨模态研究中被同时使用。需要注意的是，本数据集图像来源不一导致尺寸大小不同，因此基于深度学习使用本数据集时建议对图像进行裁剪或填充。本数据集建立之初主要应用于无监督方法，若使用有监督方法研究人员需要自主生成标签文件。

致 谢

感谢张炳炎等编著的枸杞虫害图谱，百科、药用植物病虫害数据库，植保驿站等公开数据网站，它们为本数据集的建立带来莫大便利。

数据作者分工职责

刘立波（1974—），女，宁夏人，博士，教授，中国计算机学会会员，研究方向为智能信息处理。主要承担工作：本数据集的规划、建立。

王晓丽（1982—），女，河北人，博士，助理研究员，主要从事科学数据管理相关研究。主要承担工作：论文撰写指导。

陈磊（1997—），男，四川人，硕士，研究方向为跨模态检索。主要承担工作：图像处理、论文撰写。

参考文献

- [1] 许盼盼. 枸杞抗盐种质资源筛选与抗盐基因的克隆鉴定[D]. 西北农林科技大学, 2018. [XUE P. Screening of lycium salt-resistant germplasm Resources and identification of salt-resistant genes[D]. Northwest A&F University, 2018.]
- [2] 陈君, 程惠珍, 张建文, 等. 宁夏枸杞害虫及天敌种类的发生规律调查[J]. 中药材, 2003 (06): 391-394. [CHEN J, CHENG H Z, ZHANG J, et al. Investigation on Occurrence of Lycium barbarum Pests and Its Natural Enemies at Ningxia[J]. Journal of Cultivation and Breeding, 2003(06): 391-394.]
- [3] 任月萍, 胡忠庆. 宁夏枸杞主要病虫害化学防治研究进展[J]. 宁夏农学院学报, 2004(03): 88-91. [REN Y, HU Z. The Progress of Research on Chemical Control for Main Diseases and Insect Pests of Chinese Lycium barbarum in Ningxia[J]. Journal of Ningxia Agricultural College, 2004 (03): 88-91.]
- [4] 陈雷, 袁媛. 大田作物病害识别研究图像数据集[J/OL]. 中国科学数据, 2019, 4(4). (2019-06-11). DOI: 10.11922/csdata.2019.0008.zh. [CHEN L, YUAN Y. An image dataset for field crop disease identification[J/OL]. Journal of Chinese Scientific Data, 2019, 4(4). (2019-06-11). DOI: 10.11922/csdata.2019.0008.zh.]
- [5] 张炳炎, 胡冠芳. 枸杞病虫害及其防治彩色图谱[M]. 甘肃: 甘肃文化出版社, 2011. [ZHANG B Y, HU G F. Color Atlas of Lycium Barbarum Diseases, Insect Pests and Their Control[M]. GanSu: Gansu Culture Press, 2011.]
- [6] 张炳炎. 枸杞病虫害及防治原色图册[M]. 北京: 金盾出版社, 2011. [ZHANG B Y. Lycium Barbarum Pests and Control Primary Color Atlas[M]. BeiJing: JINDUN CHUBANSHE, 2011.]
- [7] PEREIRA J C, COVIELLO E, DOYLE G, et al. On the role of correlation and abstraction in cross-modal multimedia retrieval[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(3): 521-535.
- [8] RASIWASIA N, COSTA Pereira, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]//Proceedings of the 18th ACM international conference on Multimedia. 2010: 251-260.
- [9] FENG F, WANG X, LI R. Cross-modal retrieval with correspondence autoencoder[C]//Proceedings of the 22nd ACM international conference on Multimedia. 2014: 7-16.
- [10] PENG Y X, QI J W, YUAN Y X. Modality-specific cross-modal similarity measurement with recurrent attention network[J]. IEEE Transactions on Image Processing, 2018, 27(11): 5585-5599.
- [11] 郭倩, 钱宇华, 梁新彦. 基于关联约束的对抗跨模态检索方法[J]. 模式识别与人工智能, 2021, 34(01): 68-76. [GUO Q, QIAN Y, LIANG X. Adversarial Cross-Modal Retrieval Based on Association Constraint[J]. Pattern Recognition and Artificial Intelligence, 2021, 34(01): 68-76.]

论文引用格式

陈磊, 刘立波, 王晓丽. 2020 年宁夏枸杞虫害图文跨模态检索数据集[J/OL]. 中国科学数据, 2022, 7(3). (2022-07-20). DOI:10.11922/11-6035.nasdc.2021.0058.zh.

数据引用格式

刘立波. 2020 年宁夏枸杞虫害图文跨模态检索数据集[DS/OL]. 国家农业科学数据中心, 2021. (2021-06-28). DOI: 10.12205/E0290.20211018.26.cs.1909.

A dataset of image-text cross-modal retrieval of *Lycium barbarum* pests in Ningxia in 2020

CHEN Lei¹, LIU Libo^{1*}, WANG Xiaoli^{2,3,4*}

1. School of Information Engineering, Ningxia University, Yinchuan 750021, P.R. China

2. Agricultural Information Institute of CAAS, Beijing 100081, P.R. China

3. National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, Sanya 572024, P.R. China

4. National Agriculture Science Data Center, Beijing 100081, P.R. China

*Email: liulib@163.com (LIU Libo); wangxiaoli@caas.cn (WANG Xiaoli)

Abstract: *Lycium barbarum*, one of the major cash crops in Ningxia, is susceptible to pests. As a result, it is of great significance to improving the pest control capability of *Lycium barbarum* for the stability of the local economy. As a result of the explosive growth of multimodal data (e. g. images and texts) in the new era of big data, traditional single-mode retrieval cannot meet the needs of diverse information retrieval. Cross-modal retrieval for multimodal data is better to cater to the need for comprehensive and flexible information retrieval. Building relevant datasets for cross-modal retrieval technology is critical for crop pest control. As a result, we created a retrieval dataset of *Lycium barbarum* pests, including image acquisition, text writing, data augmentation, classification, image-text correspondence, and so on. This dataset contains 492 MB of image and text data of 17 common diseases and insect pests of *Lycium barbarum*, with 9,496 entries pest images and 9,496 corresponding text description files. The sufficient data in the dataset can be used for the training of cross-modal retrieval samples. This dataset offers valuable basic data resources for the cross-modal retrieval of *Lycium barbarum* pests and can be used as a standard for machine learning under the background of big data in agriculture. It has significant practical application value for promoting cross-modal development in agriculture, as well as the research on the prevention and control of *Lycium barbarum* pests, as well as improving *Lycium barbarum* yield.

Keywords: *Lycium barbarum* pests; cross-modal retrieval; training samples; standard graphic library

Dataset Profile

Title	A dataset of image-text cross-modal retrieval of <i>Lycium barbarum</i> pests in Ningxia in 2020
Data corresponding author	LIU Libo (liulib@163.com), WANG Xiaoli (wangxiaoli@caas.cn)
Data authors	LIU Libo
Time range	2020
Geographical scope	Ningxia, China
Data volume	492 MB
Data format	*.jpg / *.txt
Data service system	< http://dx.doi.org/10.12205/E0290.20211018.26.cs.1909 >
Sources of funding	National Natural Science Fund (61862050); Construction of Basic Long-term Scientific Data Governance and Mining Platform, Basic scientific research fees of CAAS (Y2022LM20).
Dataset composition	This dataset is composed of three subsets in total. The first is augmented image data, which contains 17 data files, including <i>Lycium barbarum</i> species such as Measuring worm, <i>Cicadella viridis</i> , <i>Crioceridae</i> , <i>Lygus</i> sp, <i>Gargara genistae</i> Fabricius, etc. named in the form of stream number JPG format file; with an average of about 558 images per category, there are 417MB files of 9,496 high-quality JPG images. The second is the augmented text data, a total of 17 data files, with each data file named in the form of stream number TXT file, and the stream number is consistent with the corresponding image data. There are about 4MB files of 9,496 entries of text data. The third is the original data, including two data files of 1,900 original image data and 1,900 original English texts, with a data size of 72 MB.