

数据论文

太阳全日面视向磁图活动区数据集

潘劲琿¹ 刘佳佳^{1,2} 刘睿¹

1(中国科学技术大学地球和空间科学学院 合肥 230026)

2(深空探测全国重点实验室 合肥 230026)

摘要 太阳活动区是太阳剧烈爆发活动例如太阳耀斑和日冕物质抛射等的重要源区, 对于太阳活动区的识别有助于预报太阳剧烈活动对地球空间环境的影响. 本数据集使用 2010–2019 年来自太阳动力学天文台 (SDO) 搭载的日震与磁像仪 (HMI) 拍摄的间隔 720 s 的太阳全日面视向磁图数据, 结合 Solar Monitor (<https://solarmonitor.org/>) 网站给出的美国国家海洋和大气管理局 (NOAA) 活动区编号, 并基于传统图像处理算法的识别方法, 结合人工标注进行了磁图中活动区的标注. 本数据集包含 6975 张每天间隔 12 h 的太阳全日面视向磁图, 给出了每张磁图上的活动区位置信息, 共 26531 个活动区标注. 本数据集旨在为基于深度学习识别活动区的算法提供一个标准训练集, 从而为空间天气预报提供数据支持.

关键词 太阳活动区, 太阳全日面视向磁图, 图像识别, 太阳磁场

中图分类号 P353

数据集基本信息

数据集名称	太阳全日面视向磁图活动区数据集
数据通信作者	刘睿 (rliu@ustc.edu.cn)
数据作者	潘劲琿 (图像数据采集、图像预处理与文章撰写), 刘佳佳 (数据处理与文章修改), 刘睿 (文章修改、组织与指导)
数据时间范围	2010–2019年
数据量	11.1 GByte
数据格式	*.fits, *.txt
数据服务系统网址	DOI: https://doi.org/10.12176/01.01.00113-V01 CSTR: https://cstr.cn/CSTR:14804.11.01.01.00113-V01
基金项目	国家重点研发计划项目 (2022YFF0711402)
数据集组成	数据集包括了6975张*.fits格式的太阳全日面视向磁图以及与每张磁图对应的*.txt格式的活动区标注信息. 磁图数据下载自JSOC数据网站 (https://jsoc.stanford.edu/)

2025-06-02 收到原稿, 2025-09-26 收到修定稿

©The Author(s) 2025. This is an open access article under the CC-BY 4.0 License (<https://creativecommons.org/licenses/by/4.0/>)

Dataset of Solar Active Regions in the Solar Full-disk Magnetograms

PAN Jinhui¹ LIU Jiajia^{1,2} LIU Rui¹

¹(School of Earth and Space Sciences, University of Science and Technology of China, Hefei 230026)

²(National Key Laboratory of Deep Space Exploration, Hefei 230026)

Abstract Solar active regions are key source regions of intense solar phenomena such as solar flares and Coronal Mass Ejections (CMEs). Accurate identification of these regions can help forecast the impact of solar activities on the geospace environment. This dataset utilizes solar full-disk line-of-sight magnetograms observed by the Helioseismic and Magnetic Imager onboard the Solar Dynamic Observatory (2010–2019), combined with NOAA AR numbers provided by the Solar Monitor website (<https://solarmonitor.org/>). The active regions are annotated using an image processing-based recognition method along with manual labeling in accordance with the NOAA catalog. The dataset consists of 6975 solar full-disk magnetograms, taken every 12 hours, with a total number of 26531 annotations consisting of the active-region information for each magnetogram. This dataset serves as a benchmark resource for developing deep learning solar active region detection models, and aims to enhance predictive capabilities for severe space weather phenomena through physics-informed training samples.

Key words Solar active regions, Solar full-disk magnetogram, Image recognition, Solar magnetic field

Dataset profile

Title	Dataset of Solar Active Regions in the Solar Full-disk Magnetograms
Corresponding Author	LIU Rui (rliu@ustc.edu.cn)
Data Authors	PAN Jinhui (Image data collection, image preprocessing and article writing), LIU Jiajia (Data processing and article modification), LIU Rui (Article modification, organization and guidance)
Time Range	2010–2019
Size	11.1 GByte
Formats	*.fits, *.txt
Data Service System URL	DOI: https://doi.org/10.12176/01.01.00113-V01 CSTR: https://cstr.cn/CSTR:14804.11.01.01.00113-V01
Funding	National Key R&D Program of China (2022YFF0711402)
Dataset Composition	The dataset includes 6975 fits-format solar full-disk magnetograms, along with corresponding txt-format active region annotation files for each magnetogram. The magnetogram data is downloaded from the JSOC data website (https://jsoc.stanford.edu/)

0 引言

太阳活动区是太阳大气中众多太阳活动现象的重要源区. 在这些区域内, 经常出现一些不同时间、空间尺度而相互联系的太阳活动现象, 例如太阳黑子、日珥、喷流、耀斑和日冕物质抛射等^[1,2]. 这些剧烈的太阳活动会对日地空间环境造成严重影响^[3]. 对太阳活动区的精准识别可以为预报这些太阳剧烈活动现象提供有效的样本支持, 减小剧烈空间天气活动带来的灾害.

近年来, 太阳和太阳风层探测器 (SOHO) 与太阳动力学天文台 (SDO) 的成功发射, 实现了对太阳连续、多通道、高分辨率的观测, 显著推动了太阳物理学的研究. 然而海量的太阳图像数据也亟需更客观、自动化的太阳特征提取方法, 以实现高效的处理. 目前已经有很多关于太阳活动区自动识别的研究. 早期的识别方法主要基于传统图像处理技术, 例如, 阈值分割、形态学开闭运算和区域生长算法^[4,5]. 随着深度学习和计算机视觉的发展, 利用深度学习开发的太阳特征检测识别算法也越来越成熟, 例如基于 R-CNN 算法^[6] 的对于冕洞、太阳黑子和日珥的识别^[7]; 基于 YOLOv3 算法^[8] 对于太阳活动区的识别^[9]; 基于 U-net 算法对于日冕喷流的识别^[10]. 这些有监督的深度学习算法离不开带有标签的数据集的支持.

数据集使用 JSOC 提供的 hmi.M_720 s 太阳全日面视向磁图数据, 基于 SolarMonitor 网站给出的美国国家海洋和大气管理局 (NOAA) 活动区编号, 使用传统图像处理方法, 结合人工辅助的方式完成数据集标注, 旨在为基于深度学习识别活动区的算法提供一个标准训练集. Pan 等^[11] 基于 YOLOv8 深度学习算法开发了 SARD 算法, 该算法能够实现 SDO/HMI 太阳全日面磁图中的太阳活动区的自动检测. 相比传统的基于图像处理的太阳活动区检测方法, SARD 在检测速度和准确率方面均有提升.

1 数据采集与处理方法

研究使用的图像数据源自太阳动力学天文台 (SDO^[12]) 搭载的日震与磁像仪 (HMI^[13]). HMI 自 2010 年 5 月 1 日起持续运行, 这里采用其全日面

720 s 视向磁图 (hmi.M_720 s) 数据构建数据集.

考虑到人工标注海量数据耗时费力, 本研究结合传统的图像处理技术^[4,6], 开发了自动化标注流程 (见图 1), 具体步骤如下.

第 1 步 阈值分割, 通过 OTSU 阈值法^[13] 自动确定最佳阈值, 确保算法在不同图像条件下的自适应性. 通过利用太阳全日面视向磁图的固有的对比度差异, 实现活动区的初步分离, 如图 1(b) 所示.

第 2 步 应用形态学开运算优化上一步得到的区域. 使用 11×11 的核结构消除细小噪声, 如图 1(c) 所示.

第 3 步 使用区域生长算法以前一步获取的像素为基础, 延展出活动区的完整形态. 选择磁场强度大于 50 GByte 的像素点作为种子点, 根据强度与纹理等相似性迭代扩展, 确保形成精确的连续区域, 并且设置了 0.05 的强度阈值, 用来恢复超过阈值且与种子点相连通的像素, 如图 1(d) 所示.

第 4 步 采用形态学闭运算融合相邻像素以填充检测区域内部的空隙, 使用了 5×5 大小的核结构, 用以消除活动区区域内的空隙. 最后使用标注框标注联通的区域, 获得如图 1(f) 所示的最终活动区标注图像, 如图 1(e) 所示.

需要注意的是, 该自动检测结果与美国国家海洋和大气管理局 (NOAA) 活动区目录给出的活动区位置信息仍存在差异. 例如 2013 年 10 月 12 日太阳全日面视向磁图预处理结果 (见图 1f) 与 Solar Monitor 平台提供的活动区标注 (见图 1g) 略有不同. 本方法将 AR 11861 识别为两个活动区, 并在日面中心附近检测到额外区域. 因此, 通过人工调整标注 (见图 1h) 使其与 NOAA 活动区信息保持一致.

2 数据样本描述

数据集由来自 HMI 的太阳全日面 720 s 视向磁图 (hmi.M_720 s) 数据和每张磁图上对应的活动区标注构成. 活动区标注以 txt 格式独立存储 (见表 1), 每个文件对应一张磁图, 其内容包含经归一化处理的的活动区域边界坐标——即各活动区在磁图的像素坐标系中左上角与右下角的纵向及横向的像素位置, 以 4 个数值表示. 每个独立活动区的标注信息独占一行, 形成规范化的存储结构 (格式示意: $[x_min, y_min,$

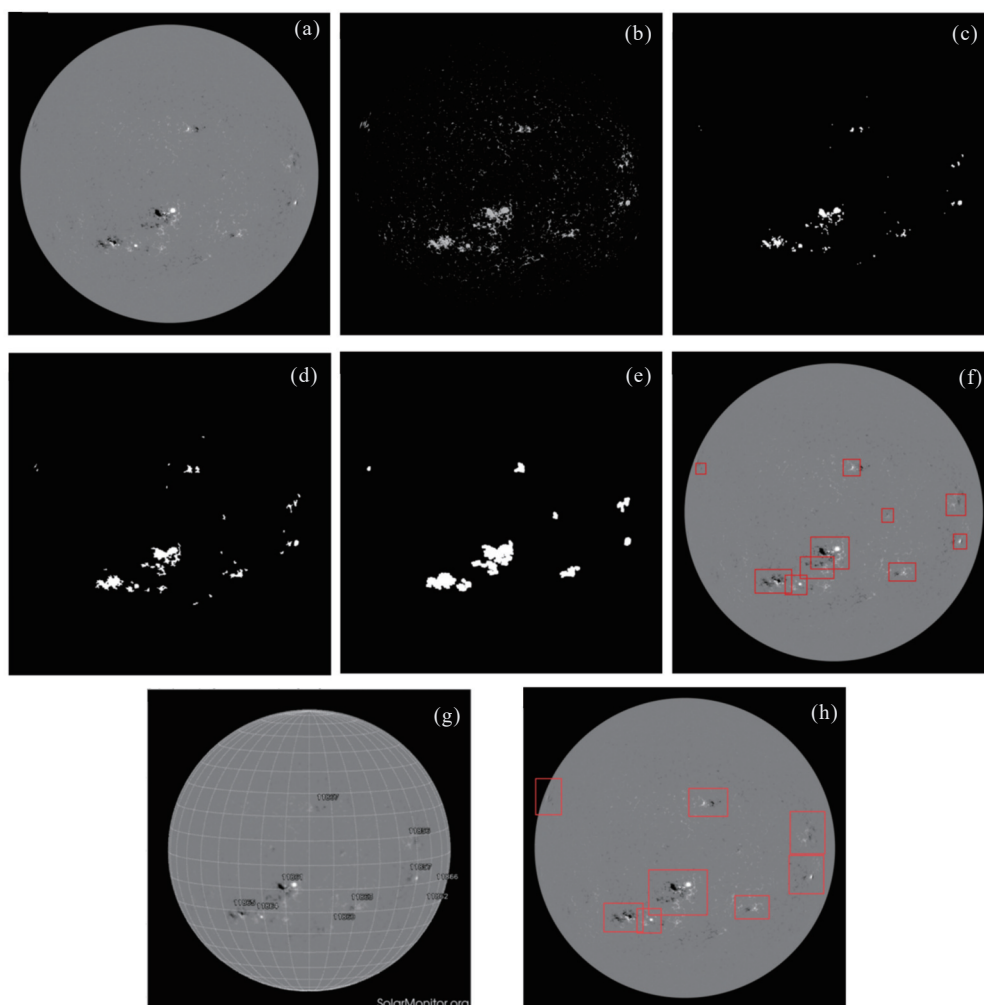


图 1 数据集制备过程中的图像预处理流程. (a) 2013 年 10 月 12 日 00:00:00 UTC 时获取的原始视向磁图, (b) 经 OTSU 阈值化处理后的结果, (c) 形态学开运算处理效果, (d) 区域生长算法的处理结果, (e) 形态学闭运算处理结果, (f) 显示最终检测到的活动区, (g) Solar Monitor 获取的 2013 年 10 月 11 日 23:31:38 UTC 时的 NOAA 活动区 (AR) 标注, (h) 经过人工调整后的活动区标注

Fig. 1 Flowchart of image preprocessing during dataset preparation. (a) Original Line-Of-Sight (LOS) magnetogram acquired at 00:00:00 UTC on 12 October 2013, (b) result of OTSU thresholding, (c) processing effect of morphological opening operation, (d) processing result of region growing algorithm, (e) processing result of morphological closing operation, (f) final detected active regions, (g) NOAA Active-Region (AR) labels from Solar Monitor at 23:31:38 UTC on 11 October 2013, (h) manually adjusted activity area labels

$x_max, y_max]$), 便于目标检测任务的批量读取与解析.

3 数据价值

太阳活动区是太阳剧烈爆发现象的主要源区, 对其准确识别可以为恶劣空间天气的预报提供有效的样本支持. 本数据集包含 2010—2019 年每天间隔

12 h 的太阳全日面视向磁图及其对应的活动区标注信息, 可以为太阳全日面视向磁图上活动区的智能识别模型的构建工作提供支持, 具备时空连续性强、标注一致性高的特点. 进一步, 可为基于监督学习的太阳活动区检测模型训练、验证以及太阳爆发事件与磁结构关联性研究提供标准化数据支撑, 同时为太阳活动周演化规律分析和空间天气预报模型优化建立基础数据平台.

表 1 太阳活动区数据集信息
Table 1 Information of the solar active region dataset

数据集信息	获取方式	文件格式
太阳全日面视向磁图	下载自JSOC数据网站	fits
活动区标注	自动检测结合人工校正	txt

4 数据质量控制与评估

在数据集中, 使用了 SDO/HMI 发布的标准 hmi.M_720 s 视向磁图数据. 该级别数据存在约 10 GByte 的像元随机噪声作为主要测量误差来源^[14], 建议用户在后续分析中对 $|B| < 10$ GByte 的弱磁场区域实施阈值屏蔽处理, 以提高信噪比.

在时间完整性方面, 数据集共检测到 136 个缺失帧, 主要集中分布于 2010 年 4 月卫星初始在轨调试阶段. 所有缺帧的起止时间与持续长度已汇总于附录表中, 便于用户在进行时序分析或数据插值时作相应处理.

在数据集的标注过程中, 太阳活动区标注环节采用自动化识别方法进行初标注, 经与 NOAA 官方活动区目录比对, 发现存在差异. 遵循惯例, 定义准确检出的活动区的样本数量为 TP (定义符号 N_{TP}), 错误检出活动区的样本数量为 FP (定义符号 N_{FP}), 没有检测到的活动区样本数量为 FN (定义符号 N_{FN}), 则准确率和召回率定义为^[15]

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}, R = \frac{N_{TP}}{N_{TP} + N_{FN}}.$$

以 NOAA 官方活动区目录为基础进行对比, 活动区自动检测标注算法的准确率为 74.1%, 召回率为 75.1%. Zhang 等^[4]使用同样的方法进行综合磁图的活动区检测, 检测算法的准确率为 77.8%, 召回率为 73.8%. 为了与 NOAA 活动区目录保持一致, 通过 Solar Monitor 网站给出的 NOAA 活动区编号进行校正, 确保最终标注与 NOAA 提供的活动区目录一致, 有效保障了数据集的可靠性.

作为参考和对比, 目前广泛用于太阳活动区磁场研究与空间天气预报评估的 SHARP (Space-weather HMI Active Region Patches) 数据基于 HMI 的全日面矢量磁场观测, 通过自动检测并跟踪连贯磁结构, 形成 HARP (HMI Active Region Patches) 裁剪区域, 并且对于每个裁剪区域提供直接裁剪版数据和

圆柱等面积投影 (CEA) 版数据, 并在经过投影后的数据上每 12 min 计算一组活动区参数. 由于 HARP 是按磁结构连贯性而非 NOAA 目录进行定义的, 因此 SHARP 数据中的裁剪区域常常对应多个 NOAA 活动区. 相比而言, 本数据集按照 NOAA 活动区编号进行标注与校正, 便于围绕按 NOAA 定义的活动区开展统计与建模. 在需要物理量的计算结果时, 建议参照 SHARP 数据, 并根据其参数与关键字获取对应的物理参量.

5 数据使用方法与建议

数据集由太阳全日面视向磁图 (fits 格式) 与活动区标注文件 (txt 格式) 构成, 专为基于深度学习识别太阳活动区的算法研发设计. 磁图数据源自 JSOC 数据网站的标准化采集, 时间跨度为 2010—2019 年, 每日 00:00 UT 和 12:00 UT 各提供一张全日面磁图, 总计 6975 张. 标注文件通过自动化检测结合人工校正生成, 采用规范化存储结构, 每个文件包含对应磁图中所有活动区的标注位置信息, 每行数据以统一格式记录活动区的边界框坐标 ($[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$), 便于程序化读取与解析.

在深度学习应用中, 建议将磁图统一缩放至 1024 pixel $\times$ 1024 pixel 分辨率, 以减轻计算资源需求. 若需增强模型泛化能力, 可将标注边界框与 SDO/AIA 1700 Å 波段的图像进行空间配准, 构建多模态训练样本. 对于时序分析任务, 可利用相邻 12 h 间隔的磁图序列研究活动区演化特征. 若需使用原始磁场数据, 可通过 JSOC 接口检索 hmi.M_720 s 序列原始 FITS 文件.

致谢 本数据论文得到了国家科技基础条件平台——国家空间科学数据中心 (<https://www.nssdc.ac.cn>) 的大力支持和帮助.

参考文献

- [1] VAN DRIEL-GESZTELYI L, GREEN L M. Evolution of active regions[J]. *Living Reviews in Solar Physics*, 2015, 12(1): 1
- [2] LIU R, KLIEM B, TÖRÖK T, *et al.* Slow rise and partial eruption of a double-decker filament. I. Observations and interpretation[J]. *The Astrophysical Journal*, 2012, 756(1): 59

- [3] DANG T, LI X L, LUO B X, *et al.* Unveiling the space weather during the Starlink satellites destruction event on 4 February 2022[J]. *Space Weather*, 2022, **20**(8): e2022SW-003152
- [4] ZHANG J, WANG Y M, LIU Y. Statistical properties of solar active regions obtained from an automatic detection system and the computational biases[J]. *The Astrophysical Journal*, 2010, **723**(2): 1006-1018
- [5] WANG R H, JIANG J, LUO Y K. Toward a live homogeneous database of solar active regions based on SOHO/MDI and SDO/HMI synoptic magnetograms. I. Automatic detection and calibration[J]. *The Astrophysical Journal Supplement Series*, 2023, **268**(2): 55
- [6] GIRSHICK R, DONAHUE J, DARRELL T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580-587
- [7] BAEK J H, KIM S, CHOI S, *et al.* Solar event detection using deep-learning-based object detection methods[J]. *Solar Physics*, 2021, **296**(11): 1-15
- [8] REDMON J, DIVVALA S, GIRSHICK R, *et al.* You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 779-788
- [9] GONG L, YANG Y F, FENG S, *et al.* Solar active regions detection and tracking based on deep learning[J]. *Solar Physics*, 2024, **299**(8): 1-22
- [10] LIU J J, JI C Y, WANG Y M, *et al.* Improving the automated coronal jet identification with U-NET[J]. *The Astrophysical Journal*, 2024, **972**(2): 187
- [11] Pan J, Liu J, Fang S, *et al.* SARD: A YOLOv8-based system for solar active region detection with SDO/HMI magnetograms[J]. *Solar Physics*, 2025, **300**(8): 111
- [12] PESNELL W D, THOMPSON B J, CHAMBERLIN P C, *et al.* The Solar Dynamics Observatory (SDO)[J]. *Solar Physics*, 2012, **275**(1): 3-15
- [13] OTSU N. A threshold selection method from gray-level histograms[J]. *Automatica*, 1975, **11**: 23-27
- [14] COUVIDAT S, SCHOU J, HOEKSEMA J T, *et al.* Observables processing for the Helioseismic and Magnetic imager instrument on the Solar dynamics observatory[J]. *Solar Physics*, 2016, **291**(7): 1887-1938
- [15] BOBRA M G, SUN X, HOEKSEMA J T, *et al.* The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: SHARPs-space-weather HMI active region patches[J]. *Solar Physics*, 2014, **289**(9): 3549-3578

作者简介



潘劲琨 男, 2001年6月18日出生于甘肃省兰州市, 现为中国科学技术大学地球和空间科学学院博士研究生。
E-mail: panjh@mail.ustc.edu.cn



刘睿 (通信作者) 男, 中国科学技术大学博士生导师、讲席教授. 主要从事太阳大气中各种爆发和瞬变现象的研究, 在国际期刊发表论文 130 余篇。
E-mail: rliu@ustc.edu.cn