

测光数据和光谱数据的匹配

——LAMOST 同源星表

孔 啸

(中国科学院国家天文台 100101)

一、引言：天文学的数据宝藏

天文学是一门以观测为基础的科学,其基本研究范式是先通过各种观测设备获取大量的观测数据,再根据这些数据建立理论模型,并预测出各种可观测的现象,最后再通过后续观测来验证模型的准确性。因此,天文学研究对数据的质量、数量及准确性具有极高的要求。

在光学波段,观测研究一般分为两类:测光和拍摄光谱。测光观测,利用测光望远镜拍摄天区图像。这些图像经过专门的数据处理软件分析后,能够得到天体的位置、亮度、颜色、形状等基本信息^[1,2]。这一过程类似于进行人口普查,测光观测能

够快速获得大量天体的初步信息,但这些信息相对粗略。图 1 展示了一幅斯隆数字巡天项目(Sloan Digital Sky Survey, SDSS)^[3,4]拍摄的测光图像(伪彩图, RUN-CAMCOL-FIELD= 001140-1-0057)。这一张测光图像的大小约为 0.037 平方度,曝光时间约为 1 分钟。全地球的面积是 41252.96 平方度,SDSS 拥有 6 组 CCD 相机,理论上只需约 3000 多个小时就可以把天球所有的区域都拍摄一遍。但在实际观测中,由于 SDSS 是坐落在北半球的地基望远镜,无法拍摄南半天球的天空;而且考虑到不同的拍摄区域有重叠部分,有些区域需要进行多次重复曝光,有的区域又由于种种原因不在拍摄计划中,所以实际的拍摄时长取决于通过多种因素所制定的观测计划。



图 1 测光图像

光谱观测是通过光谱望远镜获取天体的光谱信息,详细分析天体的化学成分、运动速度、温度、表面重力、金属丰度等物理参数^[5]。图2展示了一幅郭守敬望远镜(The Large Sky Area Multi-object Spectroscopic Telescope, LAMOST)^[6,7]拍摄的光谱图像(obsid=116603144)。图2上方的图片是拍摄到的原始二维光谱图像,每一行亮条代表一颗星的光谱,光谱中的暗条纹代表吸收线。图2下方的图片是对其中一条二维光谱进行定标、减天光、抽谱之后得到的一维光谱数据,横向代表波长方向,纵向代表强度^[8,9]。

与测光观测相比,光谱观测的数据精度更高、信息更全面,但观测的效率和速度相对较低,难以快速完成大面积全天域的观测。如果对图1中测光图像上所有的星体进行光谱拍摄,假设图像中包含了100个天体(实际上会更多,每张测光图像都包含了上千个天体),每个天体需要进行总共约1小时的光谱观测,则需要至少100个小时才能完成光谱数据的获取。

测光和光谱观测数据相互配合,能够让天文学家更加深入地理解宇宙的本质,比如银河系的结构与演化、暗物质和暗能量的分布特征、恒星的生命周期及特殊天体的性质等。因此,高质量的测光和光谱数据是现代天文学研究的基础资源。

二、LAMOST、Gaia 和 Pan-STARRS 项目介绍

LAMOST^[10-12]是中国自主设计并建造的世界领先的大型光谱巡天望远镜,位于河北省承德市的兴隆观测基地(图3),具有口径大、视场广、观测效率高的特点。它一次曝光可以同时获取4000个天体的光谱数据,极大地提高了观测效率。截至目前,LAMOST已累计获得了超过千万个天体的光谱数据,成为全球最大规模的恒星光谱数据库。

然而,LAMOST望远镜虽然具备卓越的光谱观测能力,却缺乏专门配套的测光观测设备,因此无法独立进行大规模的测光观测来选源。为了弥补

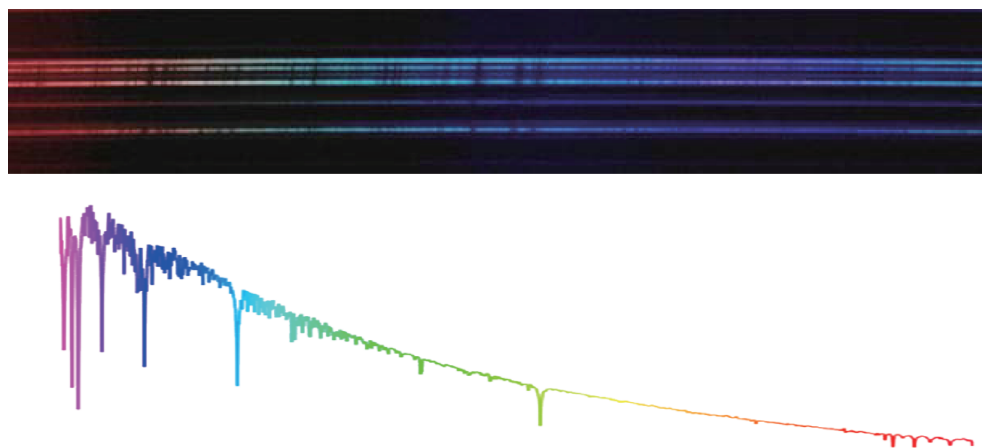


图2 天体的光谱(上图:二维光谱;下图:一维光谱)



图3 LAMOST望远镜

这一缺陷, LAMOST 的数据团队需要借助其他测光巡天项目的数据, 通过测光数据和光谱数据的精确匹配形成同源星表, 来作为 LAMOST 数据产品的补充。

Gaia^[13,14]是欧洲航天局(ESA)发射的一颗天基测光卫星, 具有高精度测量位置、视差和自行等能力。Gaia 卫星覆盖全天, 能够精确地测量数十亿颗恒星的位置和亮度, 提供 BP(蓝端)和 RP(红端)波段测光数据。BP 和 RP 波段的波长范围分别约为 330 ~ 660 nm 和 620 ~ 1050 nm, 覆盖了全部的可见光波段和部分近紫外、近红外波段。这些高精度测光数据为 LAMOST 提供了极为宝贵的测光参考。图 4 描绘了 Gaia DR3 的测光数据在天球上的覆盖区域。由于 Gaia 是天基卫星, 所以可以拍摄全部天球上的星体, 进而图 4 的全部天球区域都被覆盖。图 4 中天球的南北极区域颜色较深, 是由于区域折叠导致, 并非代表拍摄密度较高。

泛星计划(Panoramic Survey Telescope And Rapid Response System, Pan-STARRS)^[15-17]由美国夏威夷大学主导, 使用位于夏威夷哈雷阿卡拉山的地基望远镜系统进行巡天测光观测。Pan-STARRS 具有覆盖面积广、数据量大、观测周期长的特点, 能够提供包括 grizy 五个波段在内的详细测光数据, 并形成丰富的测光星表数据库。grizy 波段的波长覆盖范围

为 400 ~ 1100 nm, 覆盖了全部的可见光波段和部分近红外波段。图 5 展示了 Pan-STARRS1 DR2 所发布的测光数据在天球的覆盖情况。由于 Pan-STARRS 的望远镜坐落在北半球, 所以只能拍摄赤纬 -30° 以北的区域。此项目对拍摄区域的所有天区面积进行了无死角的扫描式巡天拍摄, 获得了丰富的测光数据样本。

通过借助 Gaia 和 Pan-STARRS 的测光数据, LAMOST 望远镜能够准确、高效地追溯观测目标, 从而显著提升了数据的科学质量和观测的整体效率。

三、如何制作 LAMOST 同源星表和同源测光星等^[18]

1. 同源星表

为了实现 LAMOST 数据与 Gaia 和 Pan-STARRS 数据的精确匹配, 进而制作出科学价值高、可信度强的同源星表, 我们设计了一套精细的交叉匹配流程(图 6)。

首先, 对于每个 LAMOST 拍摄到的天体光谱, 我们以 LAMOST 观测的空间定位精度为标准(大约 3 角秒^[19]), 进行与 Gaia 测光星表的初步匹配。如果在 3 角秒范围内, 仅找到一个唯一的 Gaia 测光源(图 6 中的“1-1”), 则使用这个 Gaia 测光源的数据作为

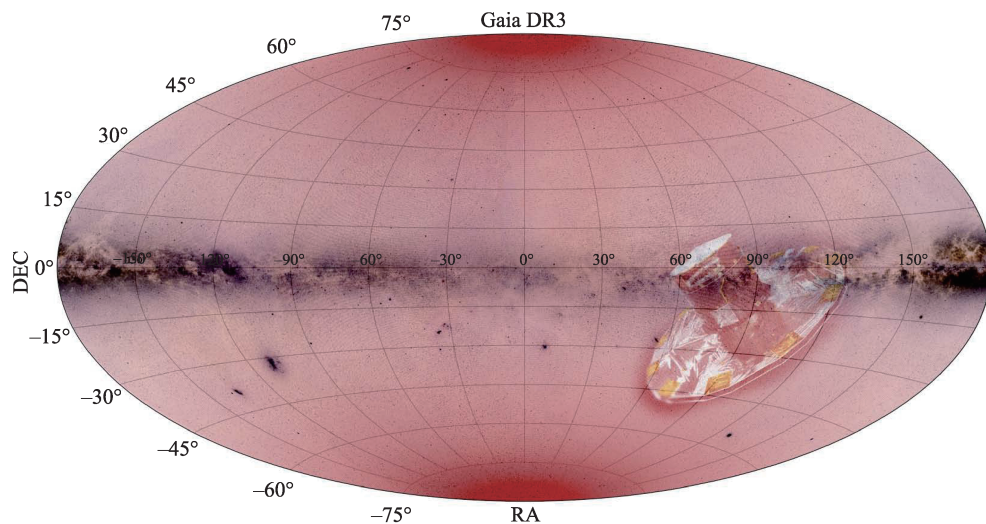


图 4 Gaia DR3 测光数据的天球覆盖图

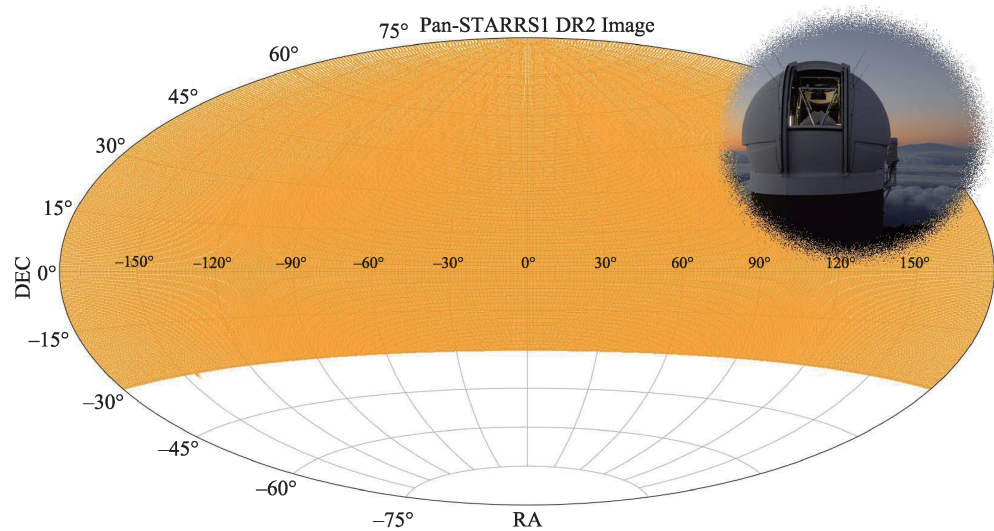


图5 Pan-STARRS1 DR2测光数据的天球覆盖图

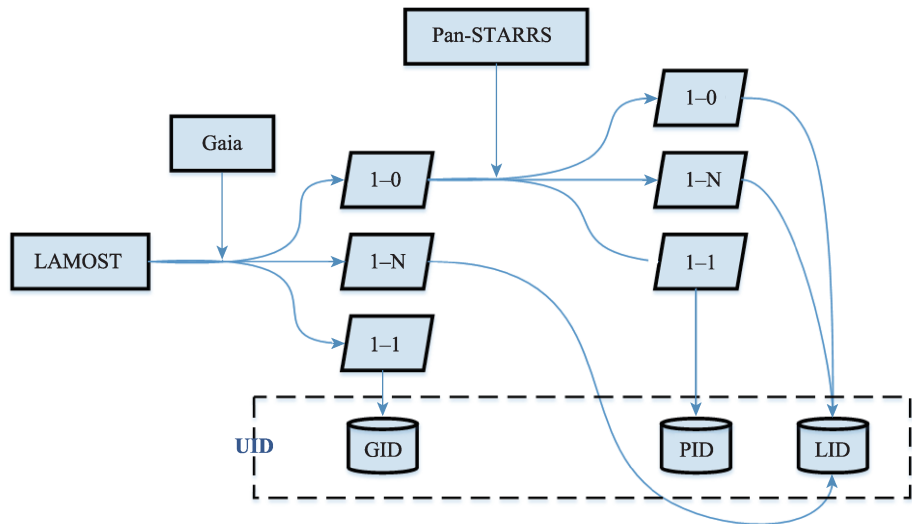


图6 同源星表的制作流程

LAMOST数据的对应测光数据(GID)。

如果在 Gaia 星表中未找到对应的测光数据,则继续使用 Pan-STARRS 测光星表进行匹配。同样,在 Pan-STARRS 数据中找到唯一的测光源后,即采用该测光数据作为 LAMOST 的对应测光数据(PID)。

若以上两个测光星表匹配失败,即未能找到唯一匹配测光源,LAMOST 数据将使用首次观测该目标时记录的初始测光数据作为默认测光数据(LID),以保证 LAMOST 数据能够始终具备基础测光信息。在 LAMOST 的发布数据中,GID、PID 和

LID 统称为 UID,作为以目标天体为单位的唯一 ID 标识。

在实际进行同源星表制作的过程中,由于 LAMOST 观测的空间分辨能力有限,通常设定 3 角秒作为匹配半径。但在个别区域,星场密集度较高,可能出现一条 LAMOST 光谱数据在交叉半径范围内同时对应多个测光源的情况。这种情况在匹配过程中被称为“一对多”的问题(图 6 中的“1-N”)。在 LAMOST 数据的同源计算中,这种情况出现的概率小于 5%。图 7 展示了两种“一对多”的测光图

像和光谱。图7中a有两个测光目标天体距离非常近,所以对应的光谱呈现出了双星的特征;图7中b的两个天体距离也小于3角秒,但是一颗星的亮度远大于另外一颗星,所以对应的光谱基本都呈现出亮星的特征(蓝色的早型星)。

一旦出现这种情况,如果仅依靠空间位置的信息,将很难准确判断哪一个测光源是真正与LAMOST光谱对应的目标。这时需要额外引入其他更精准的判据,如测光星等、色指数、光谱特征的吻合程度,甚至历史观测记录等辅助信息,进行进一步的甄别与确认。然而在实际制作LAMOST标准同源星表时,为了确保数据产品的客观性与可重复性,我们并未加入额外的主观判据。相反,我们采取了一种更为保守的策略:如果匹配半径内存在多于一个的测光源(即一对多情形),并且无其他额外判据予以区分和确认,则默认不采用这些测光源的数据。

此种情况使用LAMOST首次观测该目标时所采用的原始测光输入数据,作为其测光星等的备选方案,以尽可能避免人为因素对最终数据产品产生的潜在偏差。尽管这种处理方式客观且可复现,但也不可避免地带来了某种程度上的信息损失,尤其在密集星区(如银河系盘面附近区域),这种现象更加明显。后续的研究工作中,我们也在尝试引入更加精细的统计学和机器学习方法,结合天体的运动学特征、测光光谱联合信息等多维度数据特征,以减少一对多情形所带来的数据损失,更精确地确定LAMOST光谱所对应的准确测光信息。

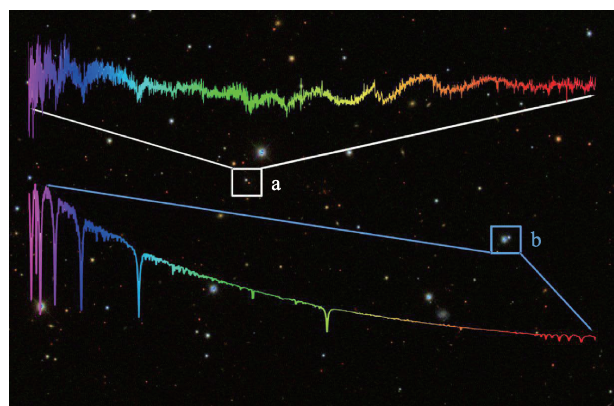


图7 光谱与测光一对多的情况

2. 同源星等

同源星等,顾名思义,就是通过精确匹配LAMOST光谱数据与测光数据,得到的与LAMOST观测目标完全对应的测光星等数据。在同源星等的制作过程中,考虑到测光数据本身的客观性与广泛适用性,我们采用了不同于同源星表的匹配策略。

我们在确定同源测光星等时采取的是距离优先的原则。对于LAMOST低分辨率光谱(LRS)数据,其输入星等的分布范围与Pan-STARRS的星等分布范围较为一致;波长覆盖范围亦与Pan-STARRS五色滤光片的波长覆盖范围重叠。以LAMOST观测目标的位置为中心,在Pan-STARRS测光星表中以3角秒为半径进行匹配,若此范围内存在多个测光目标,则直接选择与LAMOST目标距离最近的Pan-STARRS测光目标的grizy波段星等作为同源星等。

对于LAMOST中分辨率光谱(MRS)数据^[20],我们同样以3角秒的匹配半径,但采用Gaia测光星表,由于两者的星等分布范围和波长覆盖范围都较为一致。选择Gaia的BP波段(蓝端)和RP波段(红端)星等作为LAMOST MRS光谱数据对应的同源星等,同样遵循距离优先原则,即选取3角秒范围内最近的Gaia测光目标的数据。

上述处理方式在客观性上具备明显优势,即完全依靠空间位置进行匹配,没有额外引入其他可能主观或人为选择的判据。然而,在实际观测中,星场中可能存在这样一种特殊情况:在匹配半径内距离LAMOST目标更远的天体,其本身亮度可能显著高于LAMOST实际观测的天体。在此情形下,这个较亮天体发出的光线可能进入LAMOST光纤内,从而对LAMOST观测目标的光谱产生一定程度的污染,导致测得的光谱数据与实际测光数据存在一定的一致性。

从测光数据本身来看,由于测光信息的获取本身是通过图像分析得出的客观结果,测光星等本身并不会受到LAMOST光谱观测过程中可能存在的污染的影响。换句话说,测光星等仍然可以客观地代表该空间位置附近目标的实际观测结果。也正

是基于这个客观性和稳定性,我们仍然坚持在测光星等制作中采用距离最近的原则,而非采用其他人为主观的挑选方式。

四、科学前景展望

LAMOST 同源星表和同源星等的建立,不仅显著增强了 LAMOST 数据的科学应用价值,更为未来天文学研究提供了丰富的可能性。

1. 多模态数据融合的深入探索

近年来,以 AstroCLIP^[21]为代表的多模态学习模型在天文学领域快速发展,LAMOST 同源星表正好提供了丰富的数据基础。这些模型可以实现 LAMOST 低分辨(LRS)和中分辨(MRS)光谱数据与 Gaia、Pan-STARRS、SDSS 测光/光谱,以及各自星表数据的同源映射,建立统一的高维数据空间,使同一天体在不同波段、不同模态下的数据能够有效对齐(图 8)。

在此基础上,我们已经成功开发出 SpecCLIP 光谱模型,实现了对恒星大气参数(如有效温度、金属丰度、表面重力)和运动学参数(如径向速度)等的高精度预测。这类先进算法与模型的进一步发展,将深刻推动多模态数据融合在天文学研究中的应用与进步。

2. 特殊天体的挖掘与研究

借助 SpecCLIP 光谱模型,LAMOST 同源星表和星等的深入分析使我们能够更容易、更有效地挖掘出特殊和稀有的天体。例如,通过光谱与测光数据的综合分析,我们能够更精确地识别出极端贫金属星,这类恒星对研究宇宙早期化学丰度、恒星形成历史具有重大意义。此外,LAMOST 同源星表也为超高速星、蓝离散星等特殊天体的发现提供了高效、精准的数据基础。这些特殊天体的发现与分析,不仅对恒星演化理论、银河系动力学研究提供重要的实测证据,也极大丰富了天文学的研究内容和深度。

3. 为大规模天文巡天项目提供数据支撑

未来十年,天文学将迎来大规模巡天观测的黄金时代。中国空间站巡天望远镜(CSST)、Vera Rubin 天文台的大型综合巡天望远镜(LSST)等国际旗舰项目的陆续投入运行,必将把天文观测推向前所未有的高度。在这一历史性机遇面前,LAMOST 同源星表项目积累的宝贵经验——包括创新的多模态数据融合框架、高效的数据处理流水线和成熟的质量控制体系——将为新兴巡天项目提供重要的技术范式。尤其在海量异构数据的实时处理、跨波段数据的关联分析等关键领域,LAMOST 的成功实践

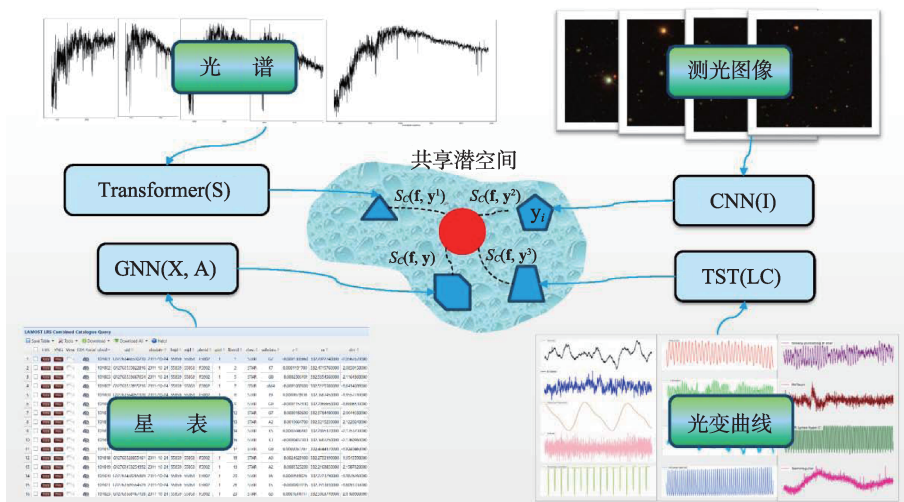


图 8 天文多模态数据的高维空间对齐

不仅具有重要的示范意义,更将持续引领下一代天文大数据处理方法的发展方向。

五、附录

1. LAMOST 数据的获取方式

LAMOST 的数据产品定期向全球发布,数据公开透明,用户可免费获取。LAMOST 官方发布网站(<https://www.lamost.org/lmusers/>)提供了所有版本发布数据的下载与查询服务。每个数据发布(Data Release)版本都会包括国内版(需要注册账号审核)和国际公开版(无需审核)的数据集,用户可以根据需求进行访问和下载。同时,LAMOST 网站还提供丰富的使用手册、操作说明、视频教程,帮助用户快速上手。此外,LAMOST 团队也提供了微信端的用户帮助群,数据分析、处理人员都在群里,可为用户提供各方面的技术支持和学术交流。

2. 如何进行同源数据交叉匹配

LAMOST 同源星表的建立,需要将 LAMOST 光谱与 Gaia、Pan-STARRS 测光数据进行精确的位置匹配。研究人员可以通过 Python 语言编写脚本,利用 Astropy 包中的 SkyCoord 等模块进行快速精准的位置匹配计算。此外,天文数据处理软件 TOPCAT (<http://www.starlink.ac.uk/topcat/>)也被广泛应用,它具有便捷的图形化界面和高效的交叉匹配算法,用户无需编程即可实现高效、精准的数据匹配与分析。

参考文献

- [1] E. Bertin and S. Arnouts. SExtractor: Software for source extraction. *Astron. Astrophys. Suppl. Ser.* 117, 393-404 (1996).
- [2] 钟靖, 侯金良, 沈世银. 银河系光学和红外巡天进展. *天文学进展*. 第 31 卷, 第 4 期, 2013 年 11 月.
- [3] Kevork N. Abazajian, Jennifer K. Adelman-McCarthy, et al. THE SEVENTH DATA RELEASE OF THE SLOAN DIGITAL SKY SURVEY. *The Astrophysical Journal Supplement Series*, 182:543-558, 2009 June.
- [4] Donald G. York, J. Adelman, et al. The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, Volume 120, Number 3.
- [5] 宁长春, 索郎桑姆. 天文望远镜 400 年发展中对于光学知识的应用和拓展. *大学物理*. 第 32 卷, 第 10 期, 2013 年 10 月.
- [6] A-Li Luo, Hao-Tong Zhang, et al. Data release of the LAMOST pilot survey. *Research in Astron. Astrophys.* 2012 Vol. 12 No. 9, 1243-1246.
- [7] A-Li Luo, Yong-Heng Zhao, et al. The first data release (DR1) of the LAMOST regular survey. *Research in Astronomy and Astrophysics*, Volume 15, Number 8.
- [8] Bai Zhong-Rui, Zhang Hao-Tong, et al. Sky subtraction for LAMOST. *Research in Astronomy and Astrophysics*, Volume 17, Issue 9, article id. 091 (2017).
- [9] Bai Zhongrui, Zhang Haotong, et al. Cosmic Ray Removal in Fiber Spectroscopic Image. *Publications of the Astronomical Society of the Pacific*, Volume 129, Issue 972, pp. 024004 (2017).
- [10] Cui Xiangqun. LAMOST project and its progress. *Progress in Astronomy*, Vol. 19, No. 2, p. 123 - 128 (2001).
- [11] Cui Xiangqun. Progress and prospect of LAMOST project. *Ground-based and Airborne Telescopes*. Edited by Stepp, Larry M.. *Proceedings of the SPIE*, Volume 6267, id. 626703 (2006).
- [12] Cui Xiangqun. LAMOST Project and its First Flight. *American Astronomical Society, AAS Meeting #213*, id.226.01; *Bulletin of the American Astronomical Society*, Vol. 41, p.473. January 2009.
- [13] Gaia Collaboration, Prusti T., et al. The Gaia mission. *Astronomy & Astrophysics*, Volume 595, id.A1, 36 pp.. November 2016.
- [14] Gaia Collaboration, Vallenari, A., et al. Gaia Data Release 3: Summary of the contents and survey properties. *Astron. Astrophys.* Volume 674, June 2023.
- [15] K. C. Chambers, E. A. Magnier, et al. The Pan-STARRS1 Surveys. *arXiv:1612.05560*.
- [16] Eugene A. Magnier, K. C. Chambers, et al. The Pan-STARRS Data Processing System. *arXiv:1612.05240*.
- [17] H. A. Flewelling, E. A. Magnier, et al. The Pan-STARRS1 Database and Data Products. *arXiv:1612.05243*.
- [18] 孔啸, 李双, 等. 同源星表构建方法及装置. 专利号: ZL 2024 1 1026396.7.
- [19] Jian-Jun Chen, Zhong-Rui Bai, et al. Measure fiber position errors from spectra data for LAMOST. *SPIE*. 91491N (2014).
- [20] Bai Zhong-Rui, Zhang Hao-Tong, et al. The first data release of LAMOST low-resolution single-epoch spectra. *Research in Astronomy and Astrophysics*, Volume 21, Issue 10, id.249, 13 pp. November 2021.
- [21] Liam Parker, Francois Lanusse, et al. AstroCLIP: A Cross-Modal Foundation Model for Galaxies. *arXiv:2310.03024*.