

基于深度学习的自动化小行星搜索

孙畅，邹桐，周佳颖

(南京外国语学校，江苏省南京市玄武区北京东路 30 号)

摘要

为了以后能摆大烂而创造了一个模板，为了展现转行效果而开始啊对对
对对对对对对对对对对对对对

关键词：摆大烂、啊对对对

Abstract

Attention ! If you input "different", the computer will output "different",
but if you input "dif{}ferent", the computer will output "different"

目录

1	引言	3
2	数据集构建	4
2.1	数据选取	4
2.2	图像预处理	6
2.3	训练目标	7
3	深度学习模型构建	8
3.1	研究背景	8
3.2	模型构建方法	9
3.3	训练模型	10

1 引言

小行星 (Figure 1)，一般指太阳系内的一类天体。该类天体类似行星环绕太阳运动，但体积和质量比行星小得多。小行星一般被认为是由太阳系形成时期的微行星演变而来，是目前发现数量最多的太阳系天体。

了解小行星的位置和轨道参数细节，对于人类在地球上的生存安全有极为重要的意义。1994 年，苏梅克-列维九号彗星在分裂成 21 颗碎片后撞击木星，其中最大的一颗碎片直径达 35 公里，产生了明显的撞击坑。据估计，这次撞击相当于 10 亿颗原子弹同时爆炸的当量，对木星大气的影响直到三个月后才基本恢复。如果这样的撞击发生在地球上，将会对地球大气产生极为显著的影响。自此之后，IAU 建立起完备的小行星观测、报备、核验系统，确认了相当数量的小行星，大大增加了我们对于地球所处的宇宙环境的了解。



Figure 1: 宇宙中的小行星示意图

此外，包括小行星在内的太阳系小天体已成为人类了解太阳系的起源和发展的重要观测对象，对于人类揭开恒星系演化过程具有重要的意义。

截至目前，尽管已有相当多的太阳系内的小行星被发现、编号，仍然有大量的太阳系内小天体未被发现。世界各地的先进天文观测望远镜每天都在产生体量极大的观测数据，但是对于观测数据的处理、筛选、分析，仍然处于相对较低的水平。尽管在观测图像中存在相当数量的小行星踪迹，但由于拍摄出的照片体量相当大，且分析手段仍存在优化空间，所以仍可能有大量已经被拍摄到的小行星被我们忽略。

传统的小行星检测算法往往从小行星本身的物理特性出发，结合本观测站的观测条

件，精心设计出特定的检测算法。然而，这类算法在不同的观测设备、观测方法和观测条件下迁移较为困难，且几乎所有的大型观测项目都不提供开源的小行星检测算法，在实际使用中存在较多限制。

一些大型观测项目逐渐意识到，在现有条件下，小行星观测数据往往存在图像分析能力不够、精确度不高的问题。于是，他们发起了一系列公众科学项目（如 IASC 和 Hubble Asteroid Hunter），将望远镜产生的数据分发给参与项目的志愿者，借助社会力量来补充现有的数据分析手段。在这些公众科学项目中，参与项目的志愿者往往需要通过肉眼进行小行星检测，总体效率很低，且准确度无法保证。

针对以上问题，本文提出了使用深度学习方法来进行小行星搜寻工作。我们拟采取对经典的图像分类卷积神经网络进行微调的方法，实现高效、自动化小行星搜索工作。

2 数据集构建

2.1 数据选取

从观测的角度来说，地面观测通常受到昼夜更替、晴夜数量、夜天光情况和大气视宁度等因素的影响，对小行星的巡天观测条件比较苛刻。所以，选取拍摄质量更高的空间望远镜数据或将成为更好的选择。哈勃空间望远镜（Figure 2）作为部署在大气层外的先进光学望远镜，具有视场大、观测能力强、干扰较少、不受地表天气因素约束等优势，是进行小行星搜寻的良好器材。



Figure 2: 哈勃空间望远镜

为能够捕获到星际中微弱的电磁波信息，哈勃空间望远镜采用的是对指定区域的长时间曝光（平均每张照片曝光 35 分钟）的观测方法。在长时间曝光下，距离更远的恒星、星系、星云往往不会发生明显的位置变化，所以在图像中能够看到一个清晰的像。由于小行星是太阳系内天体，离望远镜距离要近很多，在一次曝光中会产生明显的位移，所以表现在照片上是一条线状轨迹。我们展示了几张哈勃望远镜的观测数据（Figure 3）。

哈勃团队为提高小行星搜索检测的准确度，发起了“Hubble Asteroid Hunter”公众科学项目。该项目借助公众力量和机器学习方法，发现了 1701 条小行星轨迹，其中 670 条被确认为已知的小行星。在项目结束后，哈勃团队公开了 1701 条数据的观测号、曝光时间、图像中小行星轨迹起始和终止位置等信息，为深度学习提供了较高质量的训练数据来源。该数据集是目前我们已知的唯一公开的标记数据集。

我们展示了哈勃团队提供的表格中的局部行列（Table 1）。表格中提供的小行星轨迹是使用天球坐标系上的 RA(经度) 和 Dec(纬度) 来描述的，单位是度，这是天文常用的坐标系统，但是不适合我们进行进一步的深度学习训练。在后续处理中，我们将会把天球坐标系中的坐标转换成图像上的像素点坐标。

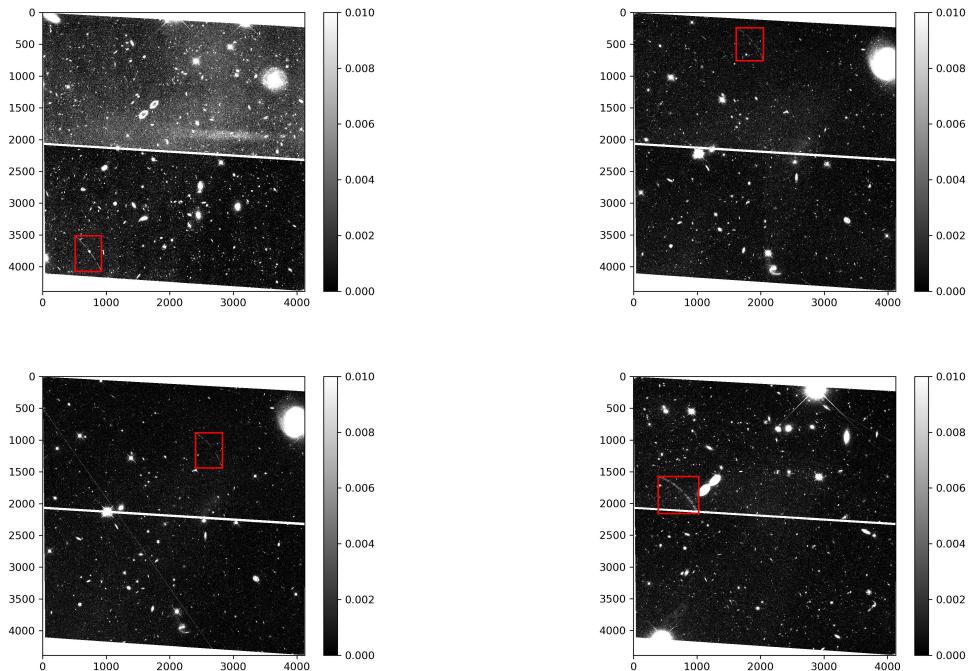


Figure 3: 哈勃观测图像与小行星轨迹样例

Table 1: Hubble Asteroid Hunter 数据集样例

哈勃望远镜观测号	轨迹曝光时间 (秒)	轨迹起始位置 RA	轨迹起始位置 Dec	轨迹终止位置 RA	轨迹终止位置 Dec
ib1901010	2520.0	146.7329823	10.0969372	146.72596	10.0971402
ib2r03020	1340.0	57.8625451	28.3094616	57.866537	28.3105017
ib4801010	1772.0	135.3464886	18.2331854	135.3422318	18.2361665
ib4803010	3686.0	135.0326681	22.5615199	135.0202912	22.549794
ib4803020	2799.0	135.0172415	22.5440068	135.0042467	22.5320243

2.2 图像预处理

尽管该数据集提供了含有小行星照片的详细参数信息，但没有直接提供可以用于训练的哈勃望远镜观测照片。所以，我们设计了一套处理流程，实现由 Python 自动化完成数据批量下载、剪裁、导出并可视化等工作。以下是图像预处理的流程 (Figure 4):

1. 读取哈勃团队提供的表格，获得指定观测号、小行星位置等关键信息；
2. 使用 Python 中的 astroquery 第三方库，调用相关 API，下载 Mikulski Archive for Space Telescopes (MAST) 数据库中指定观测号的数据¹；
3. 提取天文标准数据格式 FITS 文件中的图像数据，处理成 NumPy Array 的格式；
4. 对小行星轨迹标记的位置进行坐标转化。将原先天球坐标系统中的坐标转换成图像上的像素坐标，方便后续处理；
5. 根据现有的小行星位置信息，引入 x, y 方向上的随机偏移，获取 1000*1000 像素点尺寸的包含小行星 (positive 类别) 数据集。在这种构造方式下，可以在确保图像上保留全部或部分的小行星轨迹的同时，实现图像上的小行星轨迹随机出现，模拟真实应用场景下的图像。
6. 将所有图像按照 70/15/15 的比例划分训练集 (train)/验证集 (val)/检测集 (test)

哈勃团队只提供了含有小行星的图像观测号，而没有提供不含小行星的观测号。为保证不包含小行星 (negative 类别) 数据集的纯净性，我们从这些确保经过人工筛查的图像中，在远离小行星标记位置随机选取 1000*1000 像素点大小的图像，构成 negative 类别数据集。最终，我们数据集的大小为 positive 类别 1696 张，negative 类别 1586 张。

¹<https://astroquery.readthedocs.io/en/latest/mast/mast.html>

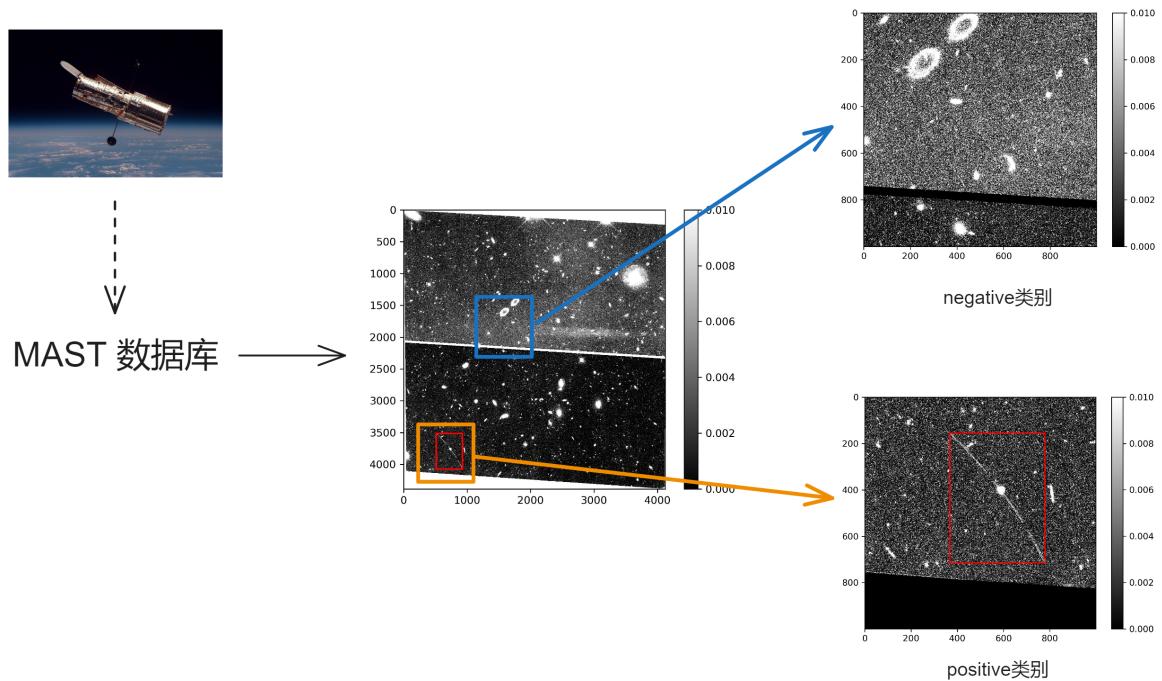


Figure 4: 预处理流程示意图

2.3 训练目标

“小行星搜索”看起来是一个目标检测问题，但实际应当被定义为一个图像分类问题。根据文献给出的数据统计，在所有哈勃拍摄的照片中，含有一条或一条以上小行星轨迹的图像仅占总数的 1.4%，只有相当少量的照片中含有小行星轨迹目标，所以有必要优先进行分类模型的训练。

在实际操作中，我们研究了哈勃团队提供的小行星轨迹标记，发现由于小行星轨迹存在弯曲的情况，根据标记数据并不能够良好的框选出小行星的轨迹区域，训练目标检测模型的效果不佳，所以我们仅专注于提高分类模型的表现。

考虑到实际应用场景下，尽管望远镜产生了大量的图片，含有小行星的照片只占 1.4%，在已经完成高精度的分类工作后将会大大减少天文工作者人工判别的时间，这是本应用场景下的痛点问题，而进一步确定小行星轨迹位置本身不是整个任务的核心。所以，我们将“小行星搜索”转化成“图像分类问题”是符合实际应用场景需求、紧抓主要矛盾的。

综上，我们的训练目标是：将没有小行星轨迹的图像 (negative 类别) 和有小行星轨迹的图像 (positive 类别) 区分开。这是一个典型的二元分类问题。

3 深度学习模型构建

3.1 研究背景

近年来，随着深度学习，尤其是卷积神经网络的快速发展，越来越多的天文领域问题开始使用深度学习的方式进行解决。在图像分类领域，一系列令人振奋的成果涌现，大大提高了自动化图像分类技术的实现精度，拓宽了卷积神经网络的应用范畴。

自 2012 年起，出现了 AlexNet, VGGNet, GoogLeNet, ResNet 等一系列现代卷积神经网络，实现了在图像分类问题上跨越式的进步。AlexNet 在大型数据集（如 ImageNet 数据集）上表现出了相比传统方法更优的性能，引领了机器视觉领域的一场革命性的进步。Simonyan et al. 提出的 VGGNet 使用多个连续的 VGG 块相连，使用简单重复的网络架构实现了更深层次的卷积神经网络，进一步提升了深度学习模型在图像分类领域的表现。GoogLeNet 将若干精心设计的 Inception 块串联起来，每个 Inception 使用多条并行路径，最终实现了较好的性能。残差神经网络 ResNet，创造性地在 VGG 块的基础上，使用残差进行训练，在 ImageNet 数据集上展示了其出色的性能。

而在小行星发现问题上，曾经的做法往往是根据具体的观测条件下精心设计某种基于小行星物理特性的算法。这些算法本身具有高度的特化性，往往难以在不同的观测方法、观测条件中迁移。深度学习方法在不同的应用场景下都具有良好的适应性，是在小行星搜索问题上更具普遍性的解决方案。此外，随着时间的推移、观测到小行星的照片不断增加，深度学习模型可以在更大的数据集上实现更高的分类精度。

2022 年，由 Kruk et al.（哈勃团队成员）首次提出了结合强化学习和人工判别的方法进行自动化搜索。然而，该团队选取的模型是使用 Google AutoML 平台生成的目标检测模型，根据论文中提供的数据，在小行星图像分类问题上，仅实现了 73.6% 的精准度 (precision)，58.2% 的召回率 (recall)，和 65.0% 的 F1 值 (f1-score)。作者没有提供开源的训练模型。尽管我们无从了解模型训练的具体细节，但显然这样的结果还有较大的优化空间。Cowan et al. 在 2023 年初将深度学习算法用于 MOA 微引力透镜巡天数据库上，实现了较好的小行星搜索结果。遗憾的是，由于作者使用的标记数据集并不公开，我们无法了解作者进行数据预处理的方法，也无法在同样的数据库上复现作者的成果或进行二次开发。

本文提出使用迁移学习的方式，对预训练模型进行微调训练，以适应我们选取的哈勃望远镜图像分类任务。我们从公开数据库中构建数据集，通过微调训练实现优异的分类精度，希望能为今后的小行星搜索工作提供可以直接迁移借鉴的成果，提供一些启发。

3.2 模型构建方法

考虑到我们的数据集体量相对较小，直接从零训练成本高且容易过拟合，我们提出使用预训练模型进行迁移学习的方法来减少训练时间并提高训练成果精度。我们使用 Pytorch 深度学习框架完成模型构建代码，使用 torchvision 第三方库中的预训练参数进行迁移训练。这些参数是经典的图像分类模型在 ImageNet 数据集上训练的成果。

为使得模型能够更好地进行深度学习训练，我们首先将所有图像 (NumPy Array 格式) 归一化到 0 至 1 之间；然后将 NumPy Array 格式的数据转换成 PIL 图像，并将图片尺寸设置为 224*224，符合预训练模型在 ImageNet 数据集上训练时的尺寸；接着，将训练集图片进行随机水平翻转；最后，将 PIL 格式的图片转换成 tensor 格式并完成 DataLoader 构建。

不同于 ImageNet 数据集中的 RGB 三通道图像，我们从哈勃获取的图像是单通道的灰度图，所以在进行迁移训练的时候需要重新设计第一个卷积层。在此过程中，我们选取的方法是将三通道修改为单通道，微调的初始权重取原先 RGB 三个通道的平均值。在这种情况下，我们成功将原本的预训练模型适配到现有的数据集。此外，模型本身实在 ImageNet 数据集上训练的，所以输出数量是 1000 类。而我们现在的分类问题是简单的二元分类，还需要将原先的输出层的输出数量设置为 2 并进行参数随机初始化。

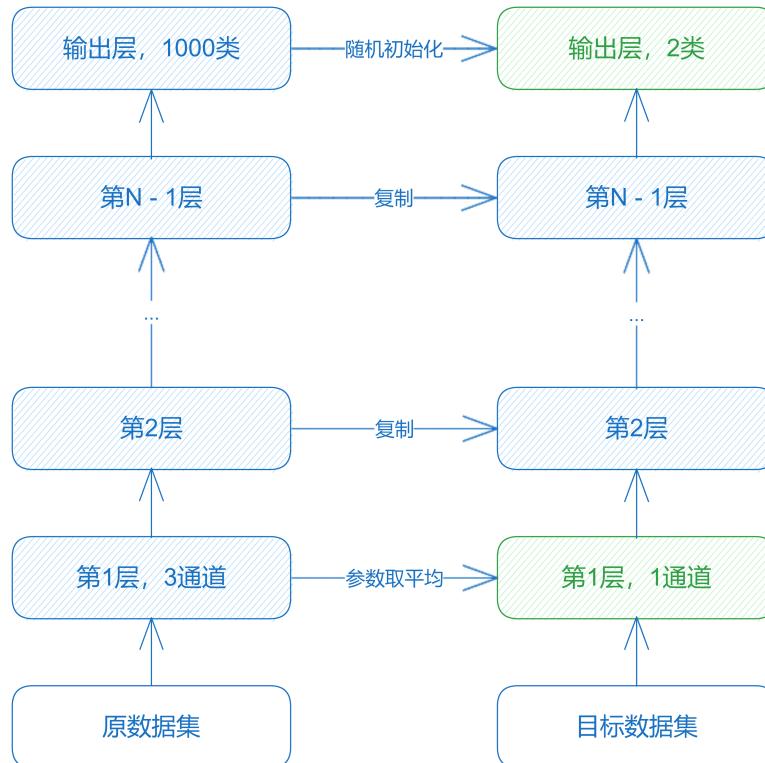


Figure 5: 微调模型构建示意图

3.3 训练模型

参考文献

- [1] Ronald A. Remillard and Jeffrey E. McClintock. X-Ray Properties of Black-Hole Binaries. *Annual Review of Astronomy and Astrophysics*, 44(1):49–92, 2006. _eprint: <https://doi.org/10.1146/annurev.astro.44.051905.092532>.