0254-6124/2023/43(2)-0219-12 *Chin. J. Space Sci.* 空间科学学报

ZHANG Lingshu, ZOU Ziming, BAI Xi. Clustering of Ultraviolet Auroral Oval Images Based on Deep Representation Learning (in Chinese). *Chinese Journal of Space Science*, 2023, **43**(2): 219–230. DOI:10.11728/cjss2023.02.220127012

基于深度表征学习的紫外极光卵图像聚类*

张龄舒[1,2,3] 邹自明[1,3] 白 曦[1,3]

1(中国科学院国家空间科学中心 北京 100190)

2(中国科学院大学 北京 100049)

3(国家空间科学数据中心 北京 100190)

摘 要 极光受太阳风驱动的地磁亚暴等大尺度动力学影响，其形态及演化因不同的太阳风-磁层-电离层耦合作用可能表现不同。目前，极光卵及其形态的归类大多依据极光演化理论作主观定性分析，没有明确的分类标准，故难以借助统计分析方法和有监督分类模型开展客观定量研究。建立了基于深度表征学习的紫外极光卵图像聚类模型（MoCo-GMM），并利用空间环境参数设计了评估模型物理合理性的方法，在大规模 POLAR 卫星紫外极光卵图像数据上进行了实验，聚类结果不仅具有良好的簇内凝聚性和簇间分散性，且具备一定的物理可解释性，有效实现了基于图像的极光卵及其形态的客观归类。

关键词 极光卵及形态归类，紫外极光卵图像聚类，MoCo-GMM 模型

中图分类号 P353

Clustering of Ultraviolet Auroral Oval Images Based on Deep Representation Learning

ZHANG Lingshu[1,2,3] ZOU Ziming[1,3] BAI Xi[1,3]

1(*National Space Science Center*, *Chinese Academy of Sciences*, *Beijing* 100190)

2(*University of Chinese Academy of Sciences*, *Beijing* 100049)

3(*National Space Science Data Center*, *Beijing* 100190)

Abstract Aurora is affected by large-scale dynamics such as geomagnetic substorm driven by solar wind, due to varies solar wind-magnetosphere-ionosphere coupling effects, its morphology and evolution can be different. Currently, categorization of aurora oval and its morphology is mostly based on auroral evolution theory to do subjective qualitative analysis and has no clear classification standard, which makes it challenging to conduct objective quantitative research with statistical analysis method and supervised classification models. Ultraviolet (UV) auroral oval image clustering model (MoCo-GMM) was

* 中国科学院网信专项资助（CAS-WX2021PY-0101）

2022-01-27 收到原稿，2022-11-07 收到修定稿

E-mail：zhanglingshu19@mails.ucas.edu.cn. 通信作者 邹自明，E-mail：mzou@nssc.ac.cn

established based on deep representation learning, also a method was designed to evaluate physical rationality of the model by using space environment parameters. Additionally, experiments on large-scale POLAR UV auroral oval image data were carried out. Clustering results of MoCo-GMM obtained not only delightful intra-cluster cohesion and inter-cluster separation, but a certain degree of physical interpretability, which means we effectively realized objective categorization of aurora oval and its morphology based on images.

Key words Categorization of auroral oval and its morphology, Ultraviolet auroral oval image clustering, MoCo-GMM model

0 引言

极光主要分布在以地磁极为中心的椭圆环带状区域,称作极光卵。极光卵形态及其变化受太阳风驱动的地磁亚暴等大尺度动力学影响,与太阳风–磁层–电离层耦合作用关系密切。例如,极光卵的极向和赤道向边界受太阳风和磁层顶相互作用控制;极光卵的强度变化对应不同的磁层现象,其突然增强和极向扩张可能伴随着亚暴[1];极光卵的位置、大小和强度随太阳风动压脉冲的变化而显著变化,行星际磁场(Interplanetary Magnetic Field, IMF)保持稳定且向南时会产生最强烈的观测现象[2]。因此,研究极光卵形态归类,对极光的统计分析、太阳风和地磁场的耦合关系以及空间天气动力学过程的研究具有重要意义。

极光卵形态归类使用天基观测的极光数据,属于全域极光形态归类,与地基观测的局部极光形态归类不同[3–5],尚无明确的分类标准,难以利用有监督分类模型开展定量研究,大多依据极光形态演化理论和稀疏观测事例作定性分析。Akasofu[6]定性描述了极光亚暴期间极光卵的主要形态演化,但刻画的特征与实际观测偏差较大。Henderson[7]分析了极光亚暴演化周期的边界活动、Omega带、极光流的变化特点,但统计个别特征无法确立归类标准。Grodent等[8]通过观察分析,以简化的极光卵形态描述了哈勃望远镜观测的木星北极成分完整的极光卵,并将其定性地分为6类,但因类间区分带有一定的主观因素,且极光卵形态复杂多变,在某些情况下可能同时符合多类判断标准。已有研究表明,因太阳风–磁层–电离层耦合的动力学过程复杂,尺度跨幅大,极光卵及其形态的定性归类难免受物理认知所局限,存在观测者偏差,可能忽略一些具有统计意义的特征,误判观测结果,以致难以归纳极光的演化规律并解析与之相关的物理过程。

因缺乏极光卵及其形态的真实类别标签,其归类无法借助成熟的分类模型,但可尝试利用图像聚类技术开展研究。Nichols等[9]采用基于特征的图像聚类思想,以主成分分析法(Principle Component Analysis, PCA)和基于密度的空间聚类(Density Based Spatial Clustering of Applications with Noise, DBSCAN)对哈勃望远镜观测的29张100 s平均木星极光卵图像进行了聚类,其结果大多能与文献[8]建立的分类标准相对应,说明机器学习聚类技术可在无监督条件下学习图像中的极光卵形态特征并对其客观定量归类。

与传统PCA算法相比,深度学习模型能挖掘更大规模图像数据的有效特征,并获得紫外极光卵图像蕴含的与极光卵形态背后潜在物理过程相关的抽象信息。因此,本文利用大规模POLAR卫星紫外极光卵图像数据,结合无监督深度表征学习模型和特征聚类算法,建立了紫外极光卵图像聚类模型MoCo-GMM,实现了极光卵形态的客观归类,并设计了基于聚类质量评估指标和空间环境参数的聚类评估方法,验证了该模型对紫外极光卵图像聚类的有效性。

1 方法

1.1 MoCo-GMM模型

紫外极光卵图像样本属于高维数据,包含噪点等无效信息,像素点仅表示极光卵亮度,若直接拉伸为向量进行聚类,难以有效利用图像中的极光卵带宽、大小等其他重要的形态特征信息。本文采用基于特征的图像聚类思想对紫外极光卵图像数据进行聚类,建立的MoCo-GMM聚类模型主要包括两个模块:紫外极光卵图像特征提取和紫外极光卵图像特征聚类。因图像特征与图像一一对应,将特征标签赋予图

像即可获得紫外极光卵图像聚类结果, MoCo-GMM聚类模型的聚类流程如图 1 所示。

紫外极光卵图像特征提取模块选用对比式深度表征学习模型 MoCo(1.2), 其无监督学习的图像特征不仅能够有效表征极光卵形态, 还包含图像中可能与潜在、未知的物理过程相关的抽象信息, 为将来归纳分析不同类别极光卵形态相关的物理机制奠定基础。紫外极光卵图像特征聚类模块选用实验效果最优的高斯混合聚类模型。

1.2 MoCo 图像特征提取

MoCo[10] 能够无监督学习图像的视觉表征, 构建了一个查询编码器(query encoder)、一个动态队列、以及一个键编码器(key encoder)。训练完成的查询编码器可将紫外极光卵图像编码为可区分的表征, 从而提取其有效特征, 核心原理如图 2 所示[9]。

首先, 当前小批次(mini-batch)的紫外极光卵图像经两套随机增强变换得到查询 x^{query} 及其正键 x_+^{key}。随即经查询编码器和键编码器分别编码, 得到查询表征 q 及其正键表征 k_+。而负键表征则来自动态队列$\{k_-^0, k_-^1,\cdots\}$, 其中保存了先前小批次图像随机增强变换得到的键$\{x_-^{\text{key}0}, x_-^{\text{key}1},\cdots\}$经键编码器编码后的结果。最后, 计算查询表征与键表征的对比损失(contrastive loss)[10], 并反向更新查询编码器, 键编码器则伴随查询编码器动量更新[10]。

对比损失函数为

$$L_q = -\log\frac{\exp\left(q\cdot\dfrac{k_+}{\tau}\right)}{\sum_{i=0}^{K}\exp\left(q\cdot\dfrac{k_i}{\tau}\right)}. \tag{1}$$

对比损失函数是一个无监督目标函数, 是 MoCo 能够无监督学习无标注紫外极光卵图像表征的关键。其中 L_q 表示 q 的对比损失, τ 为温度超参, 当 q 与 k_+相似而与动态队列的负键表征不相似时, 对比损失较小。

此外, 动量更新使不同小批次的键编码器差别很小, 是动态队列中的键表征保持一致性的关键。动量更新表达式为

$$\theta_k' = m\theta_k + (1-m)\,\theta_q. \tag{2}$$

其中, 动量 $m\in[0,1)$, θ_q 和 θ_k 分别表示查询编码器和键编码器的网络参数, 仅 θ_q 反向更新, 而 θ_k 伴随 θ_q 动量更新, 演化更为平稳。较大动量、缓慢行进的键编码器对 MoCo 的学习效果是有益的。

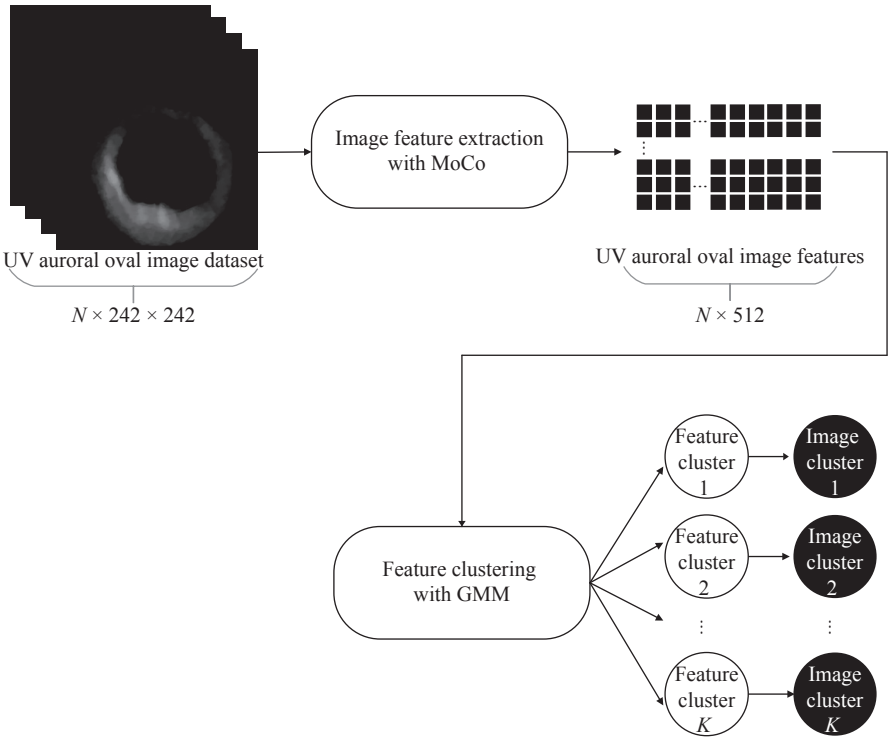

图 1 紫外极光卵图像聚类流程

Fig. 1 Pipeline of UV auroral oval image clustering

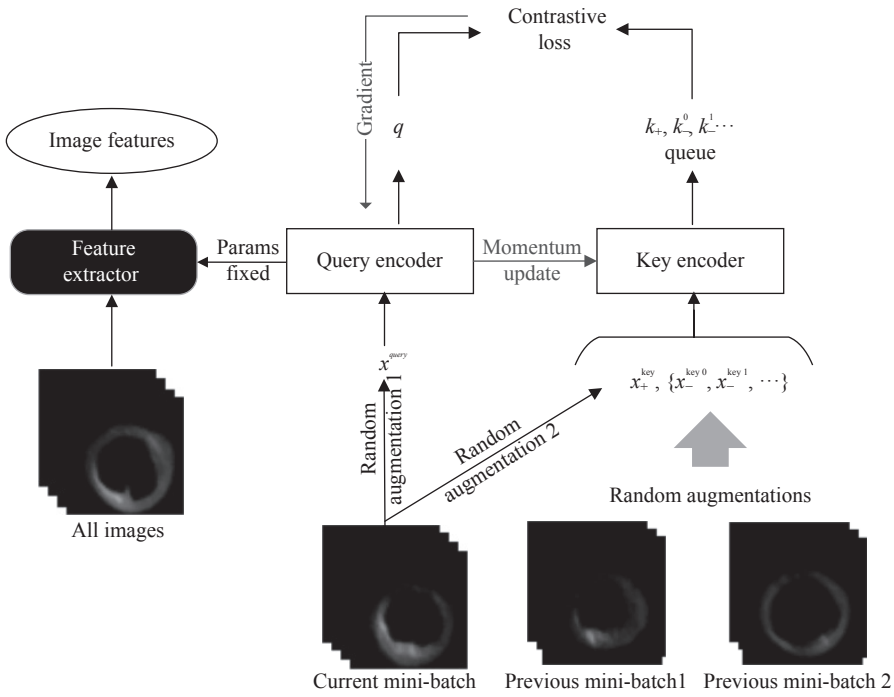

图 2 MoCo 对比学习紫外极光卵图像表征

Fig. 2 Contrastive learning of UV auroral oval image representations with MoCo

MoCo 编码器通常使用 ResNet[11], 其骨干网络的具体结构及参数需根据实验环境及数据集进行调优适配。提取图像特征的维度由查询编码器最后一层卷积层的输出通道数决定。紫外极光卵图像的增强变换主要包括随机尺寸裁剪、色彩失真(随机亮度/对比度/饱和度/色调变换)及随机水平翻转等, 旨在训练模型区分复杂变换的图像。改进的 MoCo[12] 将编码器后的一层全连接层(Full Connected layer, FC)扩展为多层感知机(Multilayer Perceptron, MLP), 提高了非线性学习的能力, 并增加了模糊增强变换, 进一步提升了图像样本的复杂度。本文使用改进的 MoCo 学习紫外极光卵图像表征。

1.3 GMM 图像特征聚类

由于缺少极光卵形态特征分布的先验知识, 图像特征聚类假设数据为多个符合高斯分布的簇叠加, 并采用高斯混合模型(Gaussian Mixture Model, GMM)[13] 进行拟合聚类。因无法直接得到 GMM 参数, 故通常以观察到的一系列样本点, 在确定聚类簇数 K 后, 求解最佳高斯子模型, 即求每个高斯子模型的均值、方差和权重的最大似然。由于 GMM 的最大似然估计目标函数是非凸的, 难以展开求偏导, 故使用期望最大化(Estimation Maximization, EM)算法[13] 进行迭代。其中, 簇数 K 会影响聚类结果, 可根据肘部法和间隔统计量法[14] 进行确定。

1.4 聚类结果合理性评估方法

紫外极光卵图像特征是极光卵形态的有效表示, 依据极光卵形态与物理背景条件关联的客观事实, 可合理推断图像特征聚类结果应符合一定的物理规律。若已知的对极光卵形态影响较强的空间环境参数与聚类结果关联较强或反之亦然, 则表明紫外极光卵图像聚类结果具有一定的合理性。本文利用空间环境参数设计了对应的分析方法, 以判断聚类结果是否符合科学认知。

因参数记录–图像按时刻一一对应, 故可对比空间环境参数聚类结果与紫外极光卵图像聚类结果, 从而分析不同参数与图像聚类结果的关联强弱。首先, 依照贪心原则确定两个聚类的对应簇, 将重合样本最多的两个簇优先确定为对应关系, 并按重合由多至少得到簇对。其后, 计算所有簇对重合数之和占总样本数的比值, 作为图像聚类结果与参数的关联度, 有

$$D = \frac{\sum_{i=0}^{K-1} N_{\mathrm{rep}_i}(I_j, P_k)}{N_{\mathrm{total}}};$$
$$0 \leqslant j,\ k < K. \tag{3}$$

其中, K 表示簇数, I_j 和 P_k 分别表示相互对应的特征

聚类簇和空间参数聚类簇, $N_{\mathrm{rep}}(\cdot)$ 表示簇对中的重合数, N_{total} 表示总样本数。

2 数据集

2.1 原始数据

POLAR 卫星携带的紫外成像仪(Ultraviolet Imager, UVI)能够获取北半球极光卵的形态信息, 其观测图像基本不会受云和下垫面的影响, 相较于 IMAGE 卫星光学成像仪的观测图像, 能更纯粹地反映极光卵形态。为减少日晖等干扰源对紫外极光卵图像的影响, 本文选取 POLAR UVI LBHL 波段(约 170 nm)1996 年 12 月的共 25056 幅紫外极光卵图像(冬季北半球极区处于永夜), 每幅图像的像素为 228×200, 两帧图像的时间间隔为 0.5 min 左右。

为评估 MoCo-GMM 聚类结果是否具备物理可解释性, 使用由 OMNI 数据库提供的空间环境参数数据进行聚类合理性评估。参考对极光卵形态参数与空间环境参数进行回归建模的相关研究所使用的空间环境参数[15,16], 选用 IMF 三分量 B_x, B_y 和 B_z, 太阳风参数 v_{p}(太阳风速度)和 N_{p}(太阳风密度), 以及地磁指数 AE, 共 6 项空间环境参数。空间环境参数数据的一条记录包括: 年、天、时、分、B_x、B_y、B_z、v_{p}、N_{p} 和 AE, 每条记录间隔 1 min, 共 44640 条记录。由于 OMNI 数据库中的 IMF 和太阳风参数在磁层顶的前端给出, IMF 和太阳风参数等效于地球弓激波顶点处的特征参数, 这些空间环境变化与电离层中的极光形态变化的时延估计为穿过磁鞘层的 5 min 和 Alfvén 过境的 2 min 之和[15,16], 而 AE 是描述极光带电急流强度的指数, 表示磁层亚暴强度, 取瞬时值即可。本文 2.3 节构建的实验数据集对时延进行了修正。

2.2 数据预处理

原始的紫外观测极光卵图像数据包含无极光卵图像, 且受观测条件限制, 数据普遍存在有噪声、对比度低的问题, 可能影响最终的聚类效果, 故而需在聚类分析前对其进行预处理。

为避免低对比度导致的样本误筛, 以及对比度增强导致的噪声放大, 原始观测数据按图像去噪、对比度增强、数据筛选的顺序先进行常规预处理。首先, 利用滑动窗口去除图像亮斑[17], 并对全图进行双边滤波, 以较好地保留极光卵的边缘信息; 随后, 利用限制对比度自适应直方图均衡化法得到明显、自然的对比度增强图像; 进行数据筛选处理, 去除不包含极光卵的图像。

为减少人工损耗, 本文利用 Vgg-Net16 分类网络[18] 迭代式筛选数据, 将其分为有/无极光卵两类图像。迭代前, 选取标记有/无极光卵图像各 2528 幅, 并以其中 1/5 作为测试集, 其余用于第 0 轮的训练和验证。第 1 轮新增未标注图像 1000 幅, 此后第 2～5 轮新增数为上一轮的两倍, 新增图像的类别标签由上一轮训练后的网络分类得到, 并人工纠正个别误分标签, 进而按 4:1 划分数据进行本轮训练和验证, 随着迭代次数增加, 误分标签将减少。第 6 轮利用除测试集外的所有数据进行训练, 由测试集进行测试。网络的最终平均分类准确率为 97%, 准确率及损失值变化曲线如图 3 所示, 从第 2 轮开始分类网络趋于收敛, 后续迭代每轮训练不超过 5 期。与完全标注的基础分类网络训练相比, 迭代法只需标注不到 1/4 的数据, 最终筛选出 18986 幅有极光卵图像。

为使紫外极光卵图像聚类更准确地实现极光卵

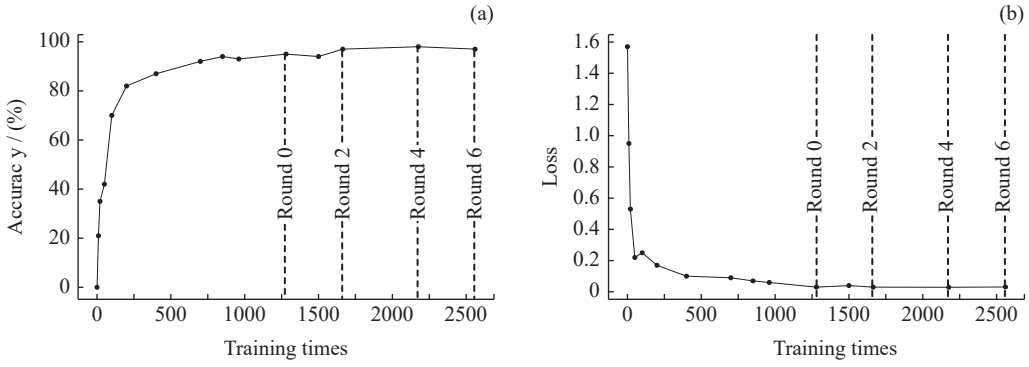

图 3 VggNet-16 迭代式分类的准确率和损失值变化曲线

Fig. 3 Accuracy and loss curves of iterative classification with VggNet-16

形态归类，本文在以上常规预处理基础上增加了极光卵形态提取处理，尽量去除每张图像中包含光学干扰的背景区域，而保留极光卵区域。通过引入 Graph-Cut[19] 算法，进一步改善了 U-Net[20] 模型对极光卵边缘模糊、有日晖干扰等图像的形态提取效果，如图 4 所示。此外，为使客观相似的极光卵不因卫星拍摄角度变化而在图像中形态各异，为图像聚类结果符合客观物理规律打下数据基础，仅含极光卵形态的图像被定位到地磁坐标系下，图像像素大小为 242×242。紫外极光卵图像数据预处理管线及处理效果如图 5 所示。

2.3 实验数据集构建

由于空间环境参数数据与紫外极光卵图像数据的观测时间分辨率不同，且包含非 LBHL 波段图像对应的参数记录，故在修正时延后需进一步处理，以构建图像和参数记录一一对应的数据集。首先，舍弃非 LBHL 波段图像对应的参数记录。而后处理空间环境参数数据的缺省值，对单侧紧邻正常值的记录进行复制补全；对两侧紧邻正常值的记录，取正常值的平均进行补全；对多条连续缺省记录，若缺省的两端正常值差异较小，则以两端数值为边界，按均匀分布随机采样进行补全，否则舍弃。最后，因 1 min 内的紫外极光卵图像极其相似，结构化相似度（Structural Similarity，SSIM）为 0.993 左右，取平均图像与参数记录对应即可。

最终构建的实验数据集包括一个大小为 5161×6 的参数数据集以及一个大小为 5161×242×242 的图像数据集，参数与紫外极光卵图像按时刻一一对应。

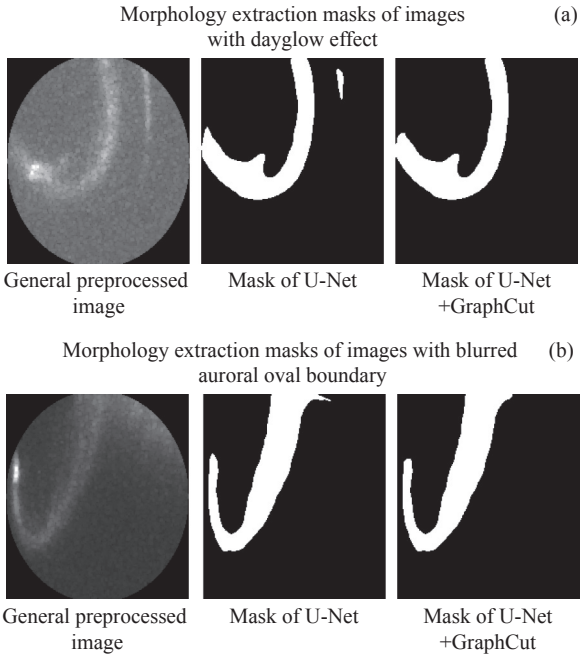

图 4 不同模型对日晖干扰图像和极光卵边界模糊图像的形态提取掩膜对比

Fig. 4 Comparison of morphology extraction masks obtained by different models from images with dayglow effect and images with blurred boundary

3 实验及结果分析

3.1 MoCo-GMM 紫外极光卵图像聚类实验

实验按图 1 所述流程进行，使用国家空间科学中心公共技术服务中心空间科学数据融合计算平台提供的计算服务（CPU：Intel(R) Xeon(R) CPU E5-2660v3@ 2.60 GHz，GPU：Tesla K80）。紫外极光卵图像特征提取模块使用改进的 MoCo 模型，后文均简写作 MoCo。训练使用紫外极光卵图像数据集（2.3 节），MoCo 并行训练使用 4 GPU，学习率为 0.015，批量大小为 128，动量为 0.999，温度参数为 0.2，动态队列大小为 5120，学习率按余弦衰减，优化器使用随机梯度下降（Stochastic Gradient Descent，SGD）。

由于 MoCo 的编码器选择与图像数据集规模以及图像的尺寸和复杂度有关，需调优 MoCo 使之适配实验数据。实验以 ResNet-18 和 ResNet-34 为编码器的 MoCo，两种 MoCo 训练时的 top-1 和 top-5 准

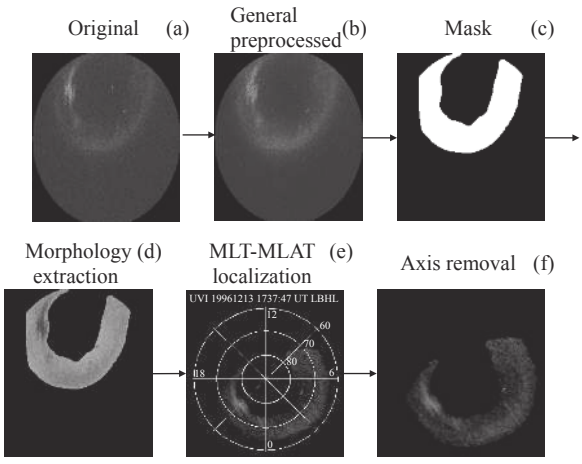

图 5 紫外极光卵图像数据预处理管线及效果

Fig. 5 Pipeline and results of UV auroral oval image preprocess

确率变化曲线如图 6 所示, 编码器为 ResNet-34 的 MoCo 最终的 top-1 和 top-5 准确率分别为 62% 和 90% 左右, 均优于以 ResNet-18 为编码器的 MoCo 且达到预期。学习完成后, 固定 MoCo 查询编码器 ResNet-34 的模型参数, 输入紫外极光卵图像数据, 得到 5161×512×8×8 的 4 维张量, 表示每幅图像对应 512×8×8 的特征图, 进而经过自适应平均池化 (Adaptive Average Pooling)得到 5161×512 的紫外极光卵图像特征数据。

随后使用 GMM 对紫外极光卵图像特征数据进行聚类, 得到每个特征的类别标签。聚类前, 首先需估计最佳聚类簇数 K。cost-K 肘部曲线如图 7(a) 所示, 没有明显拐点。进一步利用间隔统计量法进行估计, 为便于直观判断 K 值, 此处计算间隔统计量法中的作差二级指标[14], 差值为正的最小 K 即为最佳。由于间隔统计量法基于蒙特卡洛采样, 每次的计算结果不尽相同, 故只能缩小最佳 K 值的范围而无法确定唯一的 K 值。某次差值计算结果如图 7(b) 所示, 表示最佳 K 为 15。在 100 次有结果实验中, 最佳 K 值出现次数最多的为 10, 概率为 35%, 其次是 15, 16, 18 和 12。本文利用高斯混合聚类模型进行特征 10-聚类。

3.2 实验结果

由于每个特征样本的维度为 512, 难以直接进行可视化, 故利用 t-SNE[21] 技术将特征样本降至 2 维, 并予以原 512 维样本空间的聚类标签, 得到可视化散点图如图 8 所示。可见, 尽管部分簇间存在重叠, 但样本点基本能够分明地聚集和划分为不同簇。需要说明的是, 2 维空间的簇分布并不与 512 维空间的簇分布完全一致, 可视化结果只能作为辅助示意, 聚类结果的定量评估见 3.3.2 节。

因图像特征与图像一一对应, 可将特征标签赋予图像从而得到紫外极光卵图像聚类结果。其簇中心图像及簇内部分图像如图 9 所示, 相似图像被划分至一类, 不同簇的极光卵亮度、大小、带宽、弧度等有所不同。特别说明, 缺口极光卵图像大多源于卫星成像的观测不全, 少量源于形态提取误差, “缺口”本身不反映物理背景, 但图像仍可体现其他极光卵形态特征, 故予以保留。

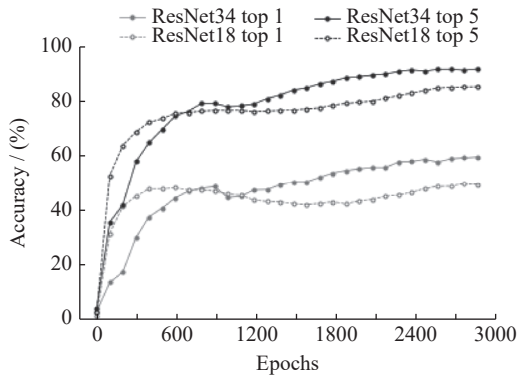

图 6 不同编码器结构的 MoCo 表征学习的准确率变化曲线

Fig. 6 Accuracy curve of MoCo representation learning with different encoders

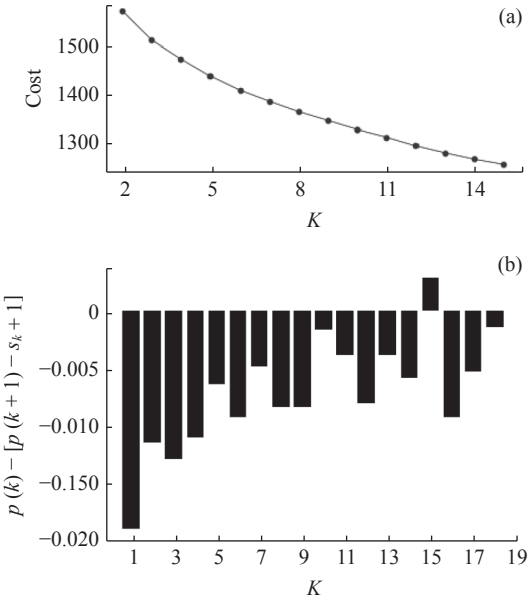

图 7 预估聚类簇数 K。(a) Cost-K 曲线, (b) 间隔统计量法条状图

Fig. 7 Estimate K, number of clusters. (a) Cost-K curve. (b) Gap statistic bar graph

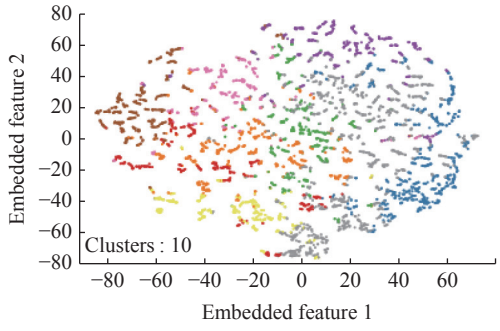

图 8 GMM 10-聚类结果可视化

Fig. 8 Visualization of GMM 10-clustering results

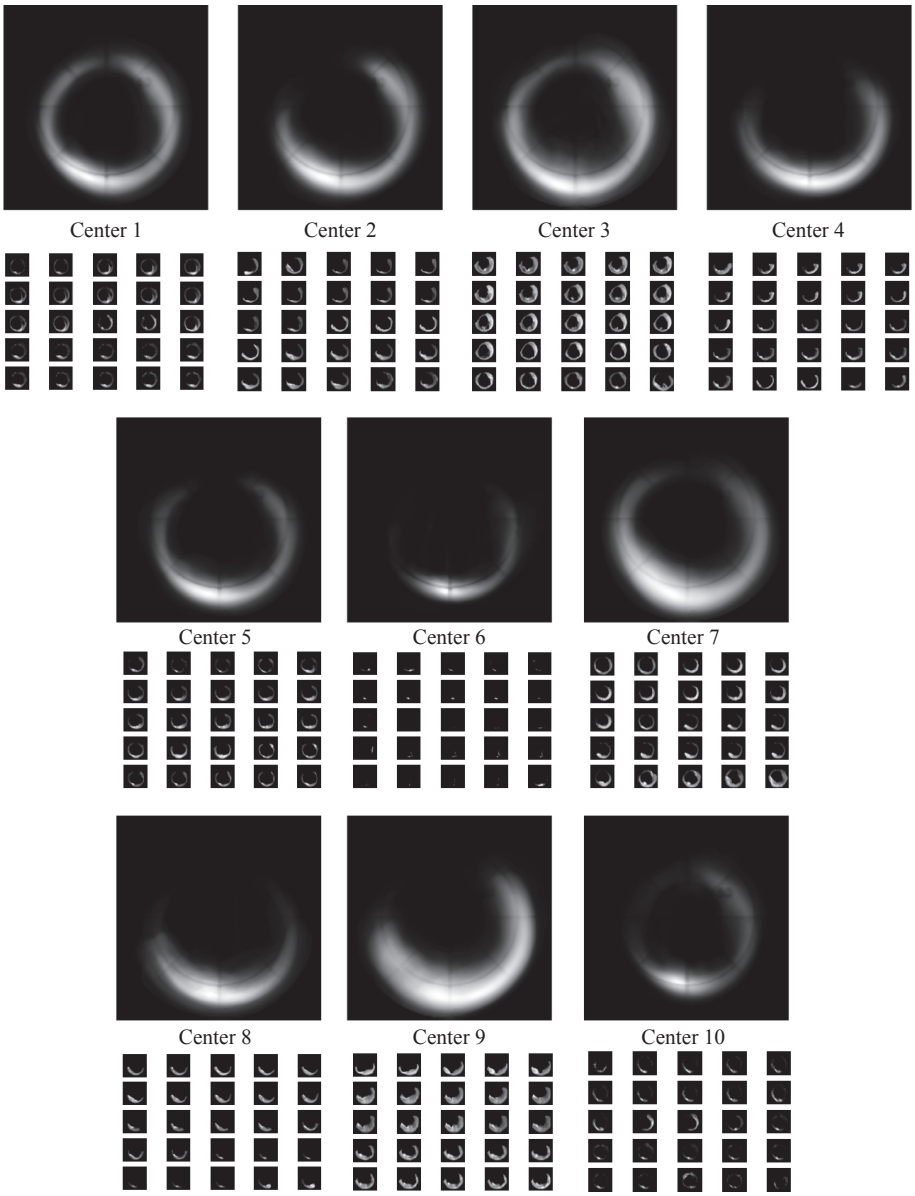

图 9 POLAR 紫外极光卵图像 MoCo-GMM 10-聚类结果

Fig. 9 MoCo-GMM 10-clustering results of POLAR UV auroral oval images

3.3 结果分析

3.3.1 紫外极光卵图像特征提取结果分析

模型的特征提取效果越优,图像特征对图像的表征能力越强,则相似特征对应的图像间应越相似。因此,提出计算每个特征对应图像及其 10-最近邻对应图像的平均相似度,并将所有平均相似度的均值记为表征度。表征度越高,则该模型所得特征的图像表征能力越强。

图像相似度可用 M_{SSIM} 表示,公式如下:

$$M_{\mathrm{SSIM}}(x,y)=l(x,y)^{\alpha}c(x,y)^{\beta}s(x,y)^{\gamma},$$

$$l(x,y)=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1},$$

$$c(x,y)=\frac{2\sigma_x\sigma_y+C_2}{\sigma_x^2+\sigma_y^2+C_2},$$

$$s(x,y)=\frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3}. \tag{4}$$

其中, x, y 表示两幅图像; $l\,(x,y)$, $c\,(x,y)$, $s\,(x,y)$ 分别表示 x 和 y 的亮度、对比度和结构相似度; α, β, $\gamma>0$ 用于调整三个模块的重要性; μ_x, μ_y, σ_x, σ_y 分别表示 x, y 的均值和标准差; σ_{xy} 表示 x 和 y 的协方差;

C_1, C_2, C_3 为常数。通常情况下,α=β=γ=1, C_3=0.5 C_2。SSIM 取值范围为 0～1,取值越大表明两幅图像越相似。

选用传统图像特征提取算法中的方向梯度直方图(Histogram of Oriented Gradient, HOG)、局部二值模式(LBP)、ORB(Oriented FAST and Rotated BRIEF)和 PCA 的特征提取结果与 MoCo 进行对比,结果列于表 1。ORB 和 LBP 无法得到统一维度的特征,MoCo 所得特征的表征度高于 HOG 和 PCA,其特征提取效果更佳。图 10 给出了 PCA, HOG 与 MoCo 提取同一幅图像特征的 10-最近邻对应图像。其中,第 1 列为实验图像,第 2～11 列为特征 10-最近邻从近到远的对应图像。可见,MoCo 提取特征的 10-最近邻图像与实验图像更相似。

3.3.2 紫外极光卵图像特征聚类结果分析

对于没有真实标签的紫外极光卵图像数据,其特征聚类结果无法使用计算聚类标签与真实标签相似度的外部评估指标,而只能使用内部指标从聚类的簇内凝聚性和簇间分散性,即簇结构质量,来评估聚类的优劣。簇内不相似度越小,簇内凝聚性越高;簇间不相似度越大,簇间分散性越强。若簇内不相似度小于簇间不相似度,则表明聚类质量良好。二者相差越大,聚类质量越高。

表 1 各算法紫外极光卵图像特征提取结果对比
Table 1 Comparison of UV auroral oval image feature extraction results with different algorithms

Algorithm	Result	Representation degree
ORB	No unified dimensions	\
LBP	No unified dimensions	\
PCA	15 dimensions	0.865
HOG	3600 dimensions	0.854
MoCo	**512 dimensions**	**0.892**

注 字体加黑组表示模型效果更优。

3.3.2.1 轮廓系数

轮廓系数(Silhouette Coefficient, SC)能够度量样本与所属簇的相似度,即内聚性,以及与其他簇的分散性。计算步骤如下。

步骤 1 计算样本 i 与同簇其他样本的平均距离 $a(i)$,为样本 i 的簇内不相似度。$a(i)$ 越小,i 与所属簇的关联越强。簇 C 中所有样本的 $a(\cdot)$ 均值即为该簇的不相似度,不相似度越小,簇内凝聚性越高。不同类的样本可使用不同的距离度量,例如向量可使用欧式距离等。

步骤 2 计算样本 i 与其他簇 $C(k)$ 中所有样本的平均距离 $b(i, k)$,为样本 i 与簇 $C(k)$ 的不相似度。i 与所有其他簇的不相似度的最小值 $b(i)$ 为 i 与其他簇的分散度,$b(i)$ 越大,i 与其他簇越分散。

步骤 3 所有样的本轮廓系数 $s(i)$ 的均值即为聚类的 SC, $s(i)$ 的定义如下:

$$s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i);\\ 0, & a(i)=b(i);\\ \dfrac{b(i)}{a(i)}-1, & a(i)>b(i).\end{cases} \tag{5}$$

SC 的取值范围为–1～1,以 0 为界。若 SC>0,则

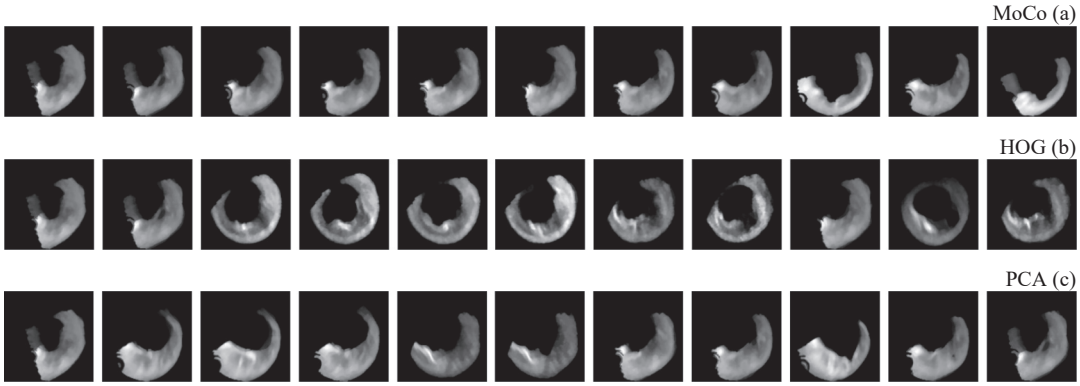

图 10 不同模型提取同一幅紫外极光卵图像特征的 10-最近邻对应图像

Fig. 10 UV auroral oval images corresponding to 10-nearest neighbors of feature extracted from the same image with different models

说明聚类质量良好; 若接近于 1, 则说明聚类质量极高。SC 对凸簇结构敏感, 例如基于密度的 DBSCAN 的聚类结果, 其 SC 通常更高, 不能证明该算法聚类效果更优。反之, 则一定表明该算法聚类效果更差。

3.3.2.2 卡林斯基–哈拉巴斯指数

利用卡林斯基–哈拉巴斯指数(Calinski-Harabasz index, CH)计算聚类结果簇内协方差矩阵的迹与簇间协方差矩阵的迹的比值, 比值越大, 聚类质量越高, 有

$$M_{\mathrm{CH}}=\frac{\mathrm{tr}\,(B_K)}{\mathrm{tr}\,(W_K)}\times\frac{(N-K)}{(K-1)}. \tag{6}$$

其中, tr(·)表示矩阵的迹, B_K为簇间散布矩阵, W_K为簇内散布矩阵, K 为聚类簇数, N 为数据大小。簇数较少的聚类通常 CH 更高, 不能说明聚类质量更高; 反之, 则一定说明聚类质量更差。CH 也对凸簇结构敏感。

GMM 与其余 5 种聚类算法在 MoCo 所得紫外极光卵图像特征数据上的聚类质量评估结果列于表 2。其中, 因 DBSCAN 5-聚类质量为其自身各簇数聚类质量的最高值, 而其次的 11-聚类可与其他算法的 10-聚类结果进行对比, 故均作记录。由表可知, GMM 的 CH 和 SC 均最高, 对紫外极光卵图像特征数据的聚类效果最佳, 其次是 K-means++和谱聚类。而 DBSCAN 的 CH 最低, SC 始终为负, 说明其聚类效果并不理想, 可能是因为 DBSCAN 算法聚类较高维数据存在较大困难, 难以划分本身相近的特征样本。

尽管 GMM 在两项指标上仅略胜于 K-means++, 但由于 K-means++可视为 GMM 的特例, 且 GMM 能拟合更复杂形状的簇, 本文选取 GMM 算法用于特征聚类模块。

表 2 各特征聚类算法效果对比

Table 2 Comparison of six feature clustering algorithms

Model	clusters	CH	SC
GMM	**10**	**183.14**	**0.057**
K-means++	10	182.79	0.054
DBSCAN	5	132.77	–0.028
	11	67.06	–0.031
Spectral clustering	10	152.37	0.044
BIRCH	10	151.63	0.035
AHC	10	151.05	0.034

注 字体加黑组表示模型效果更优。

3.3.3 紫外极光卵图像聚类结果合理性分析

分别计算了单个及 2～6 维组合的共 63 个空间环境参数向量与图像 10-聚类结果的关联度, 由强至弱, 如图 11 所示。结果表明, 单参数与紫外极光卵图像聚类结果的关联强弱顺序为 AE>N_{p}>B_z>B_x>B_y>v_{p}; 2 维组合参数关联度最高的为 (N_{p}, AE), (v_{p}, N_{p}) 的关联度高于 v_{p} 单参数; 3 维组合参数关联度最高的为 (B_y, B_z, AE), IMF 三分量组合的关联度高于单独或 2 维组合的 IMF 分量, (B_z, v_{p}, N_{p}) 的关联度高于 (B_x, v_{p}, N_{p}) 和 (B_y, v_{p}, N_{p}); 4 维组合参数关联度最高的为 (B_x, B_y, B_z, AE), 也是所有组合参数中的最高; 5 维参数关联最强的为 (B_x, B_y, B_z, N_{p}, AE), 高于 (B_x, B_y, B_z)、单独的 N_{p} 和 (v_{p}, N_{p}); 6 维组合参数的关联度处于所有组合参数的中值附近。

关联度排序的几个重要结论均符合科学认知。首先, AE 指数本身表征极光带全球电急流的活性, 作为指示极光区磁场活动的指数, 与极光卵形态实时关联, 在单参数中与极光卵形态关联度最高是合理的。其次, 本文主要考虑太阳风对极光的短时效影响, 即太阳风挤压磁层和对磁重联的影响, 在这两种过程中, 影响极光的因素主要为磁场及其方向和太阳风垂直动压, 因此 (v_{p}, N_{p}) 的关联度高于 v_{p} 是合理的。此外, B_z 影响太阳风与磁场相互作用的方式, 影响磁重联发生的位置和效率, 在 IMF 三分量中关联度最高是合理的。最后, 文献 [1] 的研究表明, 极光卵形态变化在 IMF 南向时对太阳风动压的响应最强烈, 即 (N_{p}, v_{p}) 与 B_z 的组合关联度高于其与其他 IMF 分量的组合是合理的。因此, MoCo-GMM 紫外极光卵图像聚类结果符合极光动力学理论预期, 具有一定的物理可解释性。其他组合参数的关联度排序结论也可作为探索极光卵形态与空间环境物理背景关联的实验性参考依据。

4 结论

建立了基于深度表征学习的紫外极光卵图像聚类模型 MoCo-GMM, 实现了对极光卵形态的客观归类, 主要结论如下。

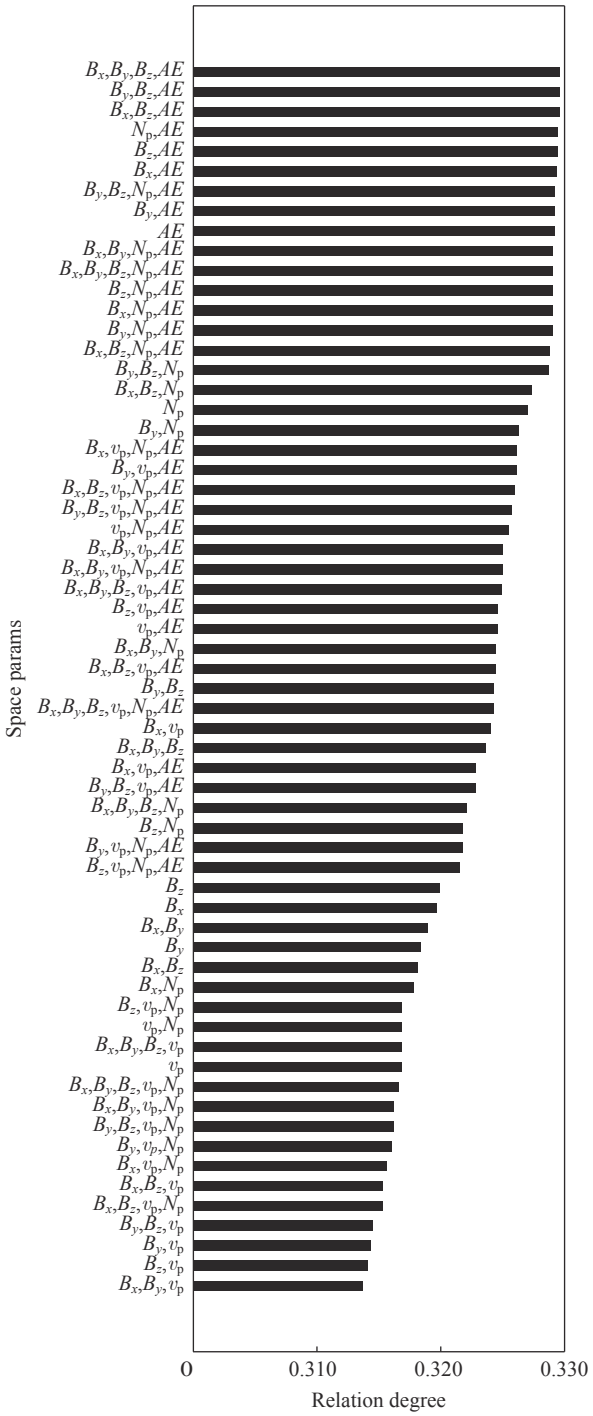

图 11 不同空间环境参数组合与紫外极光卵图像10-聚类结果的关联度

Fig. 11 Relation degree between different combinations of space environment parameters and UV auroral oval image 10-clustering results

引入 GraphCut 算法对图像进行极光卵形态提取预处理, 进一步提升了 U-Net 模型对有日晖干扰等图像的提取效果, 去除了图像中包含光学干扰的背景区域, 为 MoCo-GMM 利用极光卵形态特征进行聚类奠定了数据基础。

MoCo-GMM 紫外极光卵图像聚类模型无监督地实现了极光卵形态的客观归类, 比 PCA 等传统模型得到了更有效的图像特征, 比 DBSCAN 等传统算法得到更优的特征聚类效果, 聚类结果具有良好的簇内凝聚性和簇间分散性以及一定的物理可解释性, 避免了人为设计极光卵形态特征进行归类产生的主观偏差, 初步探索解决了因无明确分类标准而难以利用有监督分类模型归类极光卵及其形态的研究困境。

提出了基于空间环境参数的紫外极光卵图像聚类合理性评估方法, 为物理应用场景下图像聚类模型的评估提供了新思路。

需要说明的是, 本文使用的成熟的聚类簇数估计方法只能缩小特征聚类的簇数范围, 而非确定唯一的值, 可能对聚类结果产生影响, 具有一定的局限性。未来, 可进一步改进聚类模型为全深度网络, 自动确定最佳簇数, 例如最新的 Ronen 等[22] 等。

本文提出的图像聚类模型可在拟于 2024 年发射的 SMILE 卫星获取的更高质量的紫外极光卵图像上实现其应用价值; 未来可进一步融合图像聚类结果与物理判据, 建立基于多源数据的更科学的磁层动力学过程分类标准; 聚类后的细分图像数据集可进一步开展空间环境参数与极光卵形态参数的回归研究, 并将回归结果作为聚类模型调优的超参约束, 进一步探索聚类模型的物理意义。

致谢 POLAR 卫星紫外极光卵图像数据由美国国家航空航天局(NASA)提供, 数据分析环境、计算服务、应用平台支持及软硬件项目由国家科技资源共享服务平台–国家空间科学数据中心(https://www.nssdc. ac.cn)提供资助。方少峰和钟佳为本研究提供了帮助并提出了意见。

参考文献

[1] KEIKA K, NAKAMURA R, BAUMJOHANN W, *et al*. Substorm expansion triggered by a sudden impulse front propagating from the dayside magnetopause[J]. *Journal of Geophysical Research: Space Physics*, 2009, **114**(52): A00C24

[2] BOUDOURIDIS A, ZESTA E, LYONS R, *et al*. Effect of solar wind pressure pulses on the size and strength of the auroral oval[J]. *Journal of Geophysical Research: Space Physics*, 2003, **108**(A4): 8012

[3] YANG Q J, WANG Y Y, REN J. Auroral image classifica-

tion with very limited labeled data using few-shot learning[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, **19**: 6506805

[4] KVAMMEN A, WICKSTRØM K, MCKAY D, *et al*. Auroral image classification with deep neural networks[J]. *Journal of Geophysical Research: Space Physics*, 2020, **125**(10): e2020JA027808

[5] YANG Q J, ZHOU P H. Representation and classification of auroral images based on convolutional neural networks[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2020, **13**: 523-534

[6] AKASOFU S I. Auroral morphology: a historical account and major auroral features during auroral substorms[M]//KEILING A, DONOVAN E, BAGENAL F, *et al*. Auroral Phenomenology and Magnetospheric Processes: Earth and Other Planets. Washington DC: American Geophysical Union, 2012: 29-38

[7] HENDERSON M G. Auroral substorms, poleward boundary activations, auroral streamers, omega bands, and onset precursor activity[M]//KEILING A, DONOVAN E, BAGENAL F, *et al*. Auroral Phenomenology and Magnetospheric Processes: Earth and Other Planets. Washington DC: American Geophysical Union, 2012: 39-54

[8] GRODENT D, BONFOND B, YAO Z, *et al*. Jupiter's aurora observed with HST during Juno orbits 3 to 7[J]. *Journal of Geophysical Research: Space Physics*, 2018, **123**(5): 3299-3319

[9] NICHOLS J D, KAMRAN A, MILAN S E. Machine learning analysis of Jupiter's far-ultraviolet auroral morphology[J]. *Journal of Geophysical Research: Space Physics*, 2019, **124**(11): 8884-8892

[10] HE K M, FAN H Q, WU Y X, *et al*. Momentum contrast for unsupervised visual representation learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020

[11] HE K M, ZHANG X Y, REN S Q, *et al*. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016

[12] CHEN X L, FAN H Q, GIRSHICK R, *et al*. Improved baselines with momentum contrastive learning[OL]. arXiv preprint arXiv: 2003.04297, 2020

[13] MCLACHLAN G J, BASFORD K E. Mixture models: inference and applications to clustering[M]. New York: Marcel Dekker, 1988

[14] TIBSHIRANI R, WALTHER G, HASTIE T. Estimating the number of clusters in a data set via the gap statistic[J]. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 2001, **63**(2): 411-423

[15] HU Zejun, HAN Bing, LIAN Huifang. Modeling of ultraviolet auroral intensity based on Generalized Regression Neural Network associated with IMF/solar wind and geomagnetic parameters[J]. *Chinese Journal of Geophysics*, 2020, **63**(5): 1738-1750 (胡泽骏, 韩冰, 连慧芳. 基于广义回归神经网络的行星际/太阳风参数和地磁指数的紫外极光强度建模[J]. 地球物理学报, 2020, **63**(5): 1738-1750)

[16] HAN Bing, LIAN Huifang, HU Zejun. Modeling of ultraviolet auroral oval boundaries based on neural network technology[J]. *Scientia Sinica Technologica*, 2019, **49**(5): 531-542 (韩冰, 连慧芳, 胡泽骏. 基于神经网络模型的紫外极光卵边界建模[J]. 中国科学: 技术科学, 2019, **49**(5): 531-542)

[17] WANG Qian, MENG Qinghu, HU Zejun, *et al*. A method for extracting auroral ovals in UVI images and its evaluation[J]. *Chinese Journal of Polar Research*, 2011, **23**(3): 168-177 (王倩, 孟庆虎, 胡泽骏, 等. 紫外极光图像极光卵提取方法及其评估[J]. 极地研究, 2011, **23**(3): 168-177)

[18] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2014

[19] BOYKOV Y Y, JOLLY M P. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images[C]//Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. Vancouver: IEEE, 2001

[20] WANG Zihan, TONG Jizhou, ZOU Ziming, *et al*. Auroral oval morphology extraction based on u-net from ultraviolet aurora observation[J]. *Chinese Journal of Space Science*, 2021, **41**(4): 667-675 (王梓涵, 佟继周, 邹自明, 等. 基于U-net的紫外极光观测极光卵形态提取[J]. 空间科学学报, 2021, **41**(4): 667-675)

[21] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. *Journal of Machine Learning Research*, 2008, **9**(11): 2579-2605

[22] RONEN M, FINDER S E, FREIFELD O. DeepDPM: Deep clustering with an unknown number of clusters[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 9851-9860