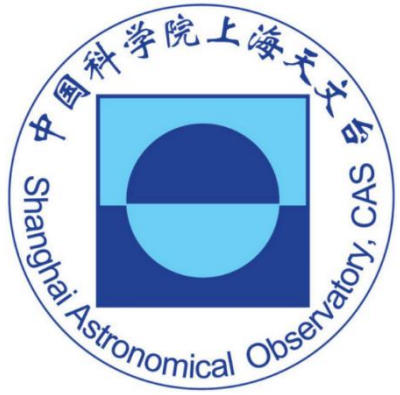


星系形态分类中的隐变量特征研究

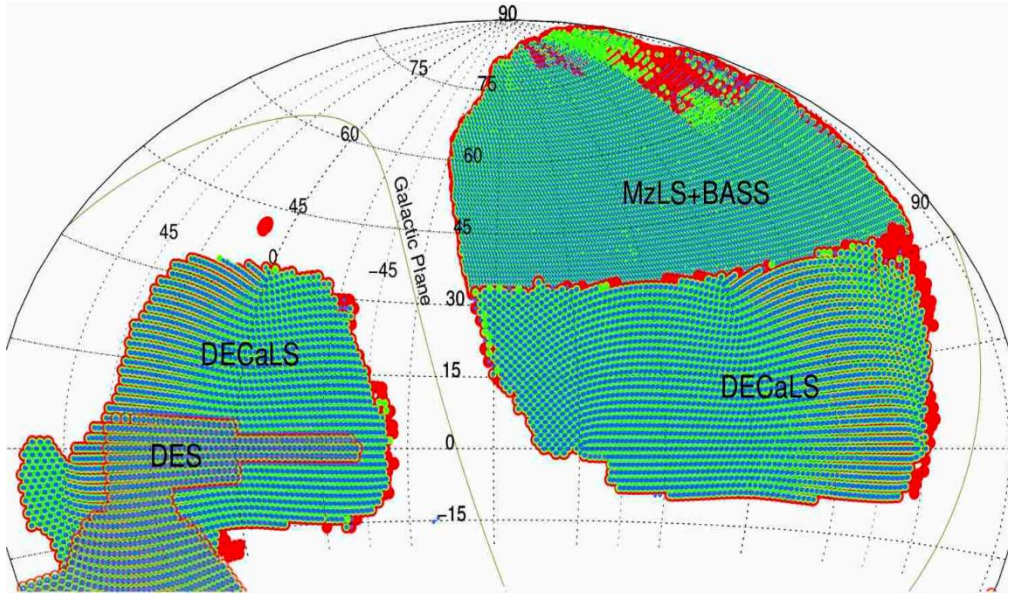
报告人：徐权峰　　指导教师：沈世银　　日期：2022年7月20日

研究背景和意义

- 研究背景
- 研究意义

研究背景

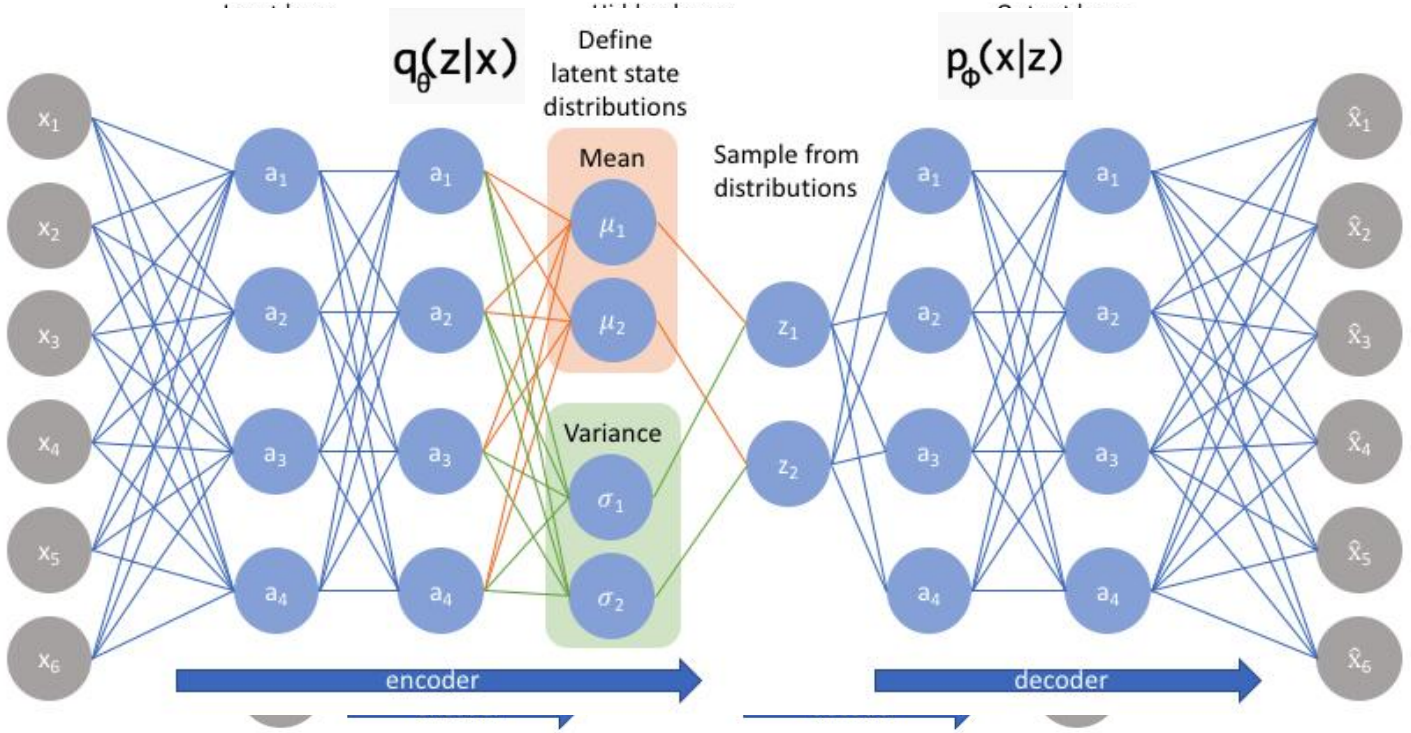
暗能量巡天(Dark Energy Spectroscopic Instrument, DESI)项目已完成14000平方度观测天区内的测光巡天，共发布了北京-亚利桑那巡天（BASS)、暗能量相机遗珍巡天(DECaLS)和Mayall z波段巡天(MzLS)三个测光项目。



传统的统计学方法无法很好将星系图像数据有效提取隐空间变量，同时也无法处理海量数据。星系图像中包含星系的化学成分、物理性质以及轮廓状态等丰富的信息。

深度学习理论背景

变分自编码器



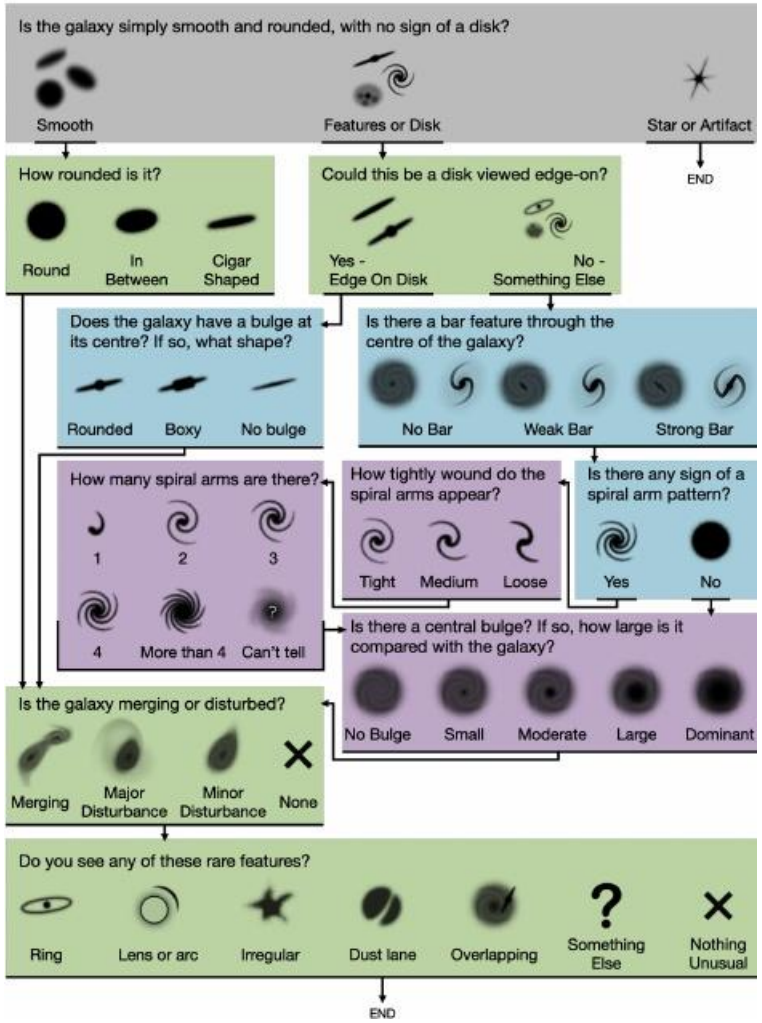
$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$$

$$D_{\mathrm{KL}}(q||p) = \int q(z) \log\left(\frac{q(z)}{p(z)}\right) dz.$$

$$\frac{1}{n}\sum_{i=1}^{n} \|x_i - \hat{X}_i\|^2$$

构建了图像的重构损失，变分自编码器中为了将生成器中采样的分布更加趋近于星系图像隐变量的先验概率分布，还使用了一个KL散度衡量采样隐变量分布与和图像的隐变量中的先验概率分布之间的差异，将其加上重构损失合成总损失，经过误差反向传播后可将高维数据提取出隐变量特征。

研究意义



暗能量相机遗珍巡天计划发布的图像比以前斯隆数字巡天(Sloan Digital Sky Survey，SDSS)图像深度更深，更好显示斯隆数字巡天中不可见的旋臂、弱柱和潮汐等视觉特征。

隐空间特征探究以星系形态为例，根据GZD-5分类决策树的各个特征标签，隐变量空间中对星系成分是否包含核球、盘和棒，星系结构和成分，星系的悬臂数量等各个视觉特征的物理性质的解释。

大型巡天计划参数海量天文数据，未来用计算机算法快速检索大型巡天项目图像成为可能。

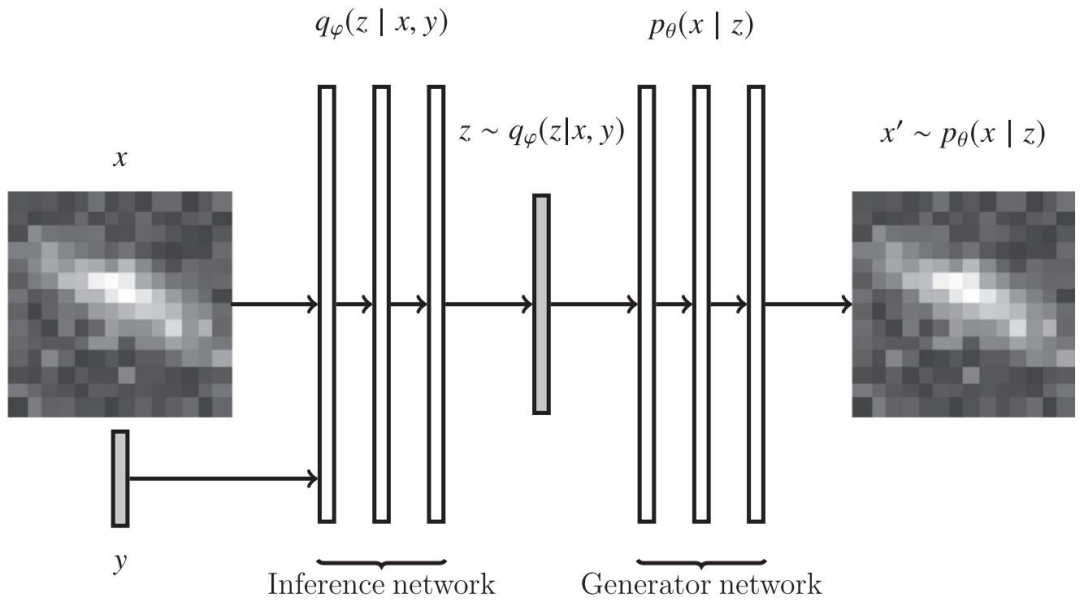
相关工作

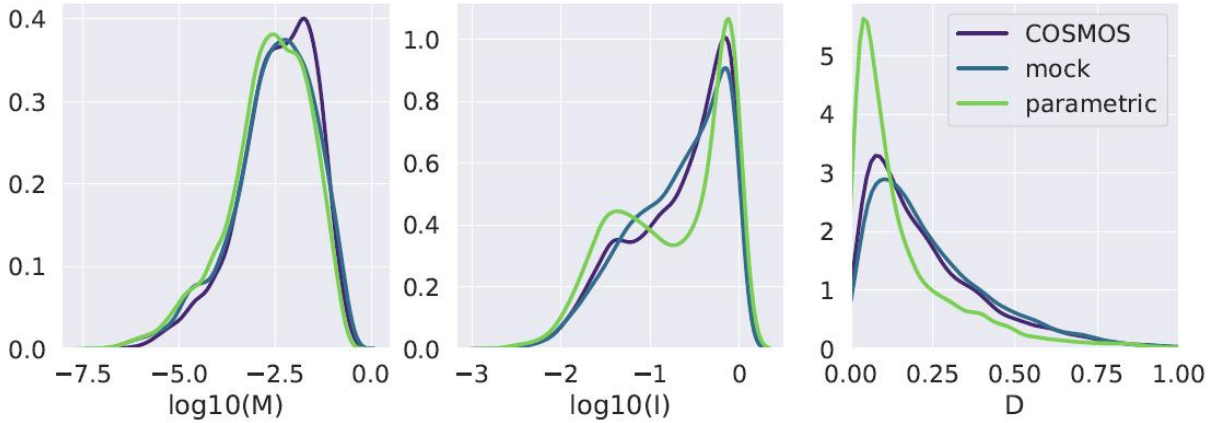
相关工作一

Deep generative models for galaxy image simulations

François Lanusse [1★] Rachel Mandelbaum [2] Siamak Ravanbakhsh

Lanusse et al.(2021):探索机器学习对产生真实星系的能力，提出了一种基于深度生成模型的方法来创建复杂的星系形态模型。

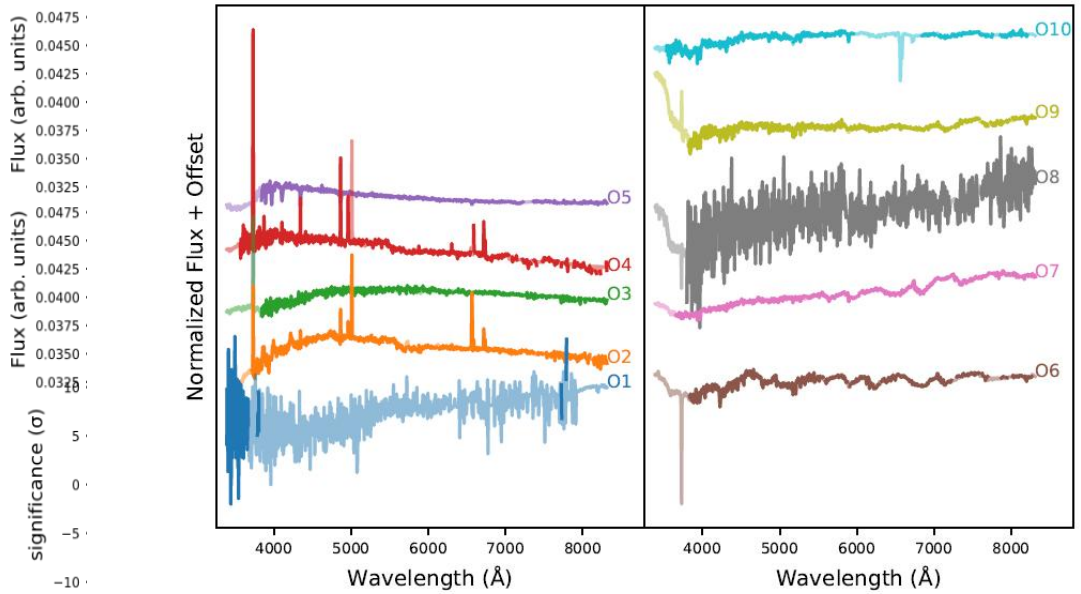




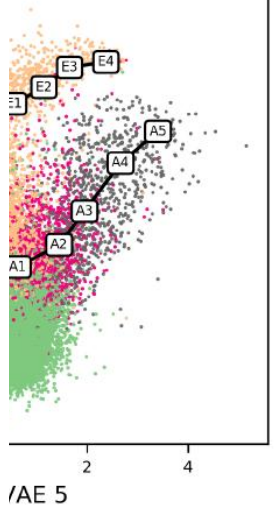
multi-mode (M), intensity (I), and Deviation (D)

相关工作二

SKN Portillo et al.(2020):对SDSS光谱不同波段的光谱数据应用变分自编码器提取光谱特征进行相关分析研究。对宁静星系、恒星形成星系、窄线AGN和宽线AGN的四种光谱进行PCA和VAE降维分析。



pectrum	plate	MJD	fiber	explanation
O1	445	51873	68	low SNR
O2	334	51993	203	bad calibration
O3	480	51989	77	close to bright star
O4	454	51908	607	bad calibration
O5	424	51893	587	A star
O6	305	51613	299	M star
O7	352	51694	340	M star
O8	529	52025	200	low SNR
O9	276	51909	2	M star
O10	414	51869	296	missing data

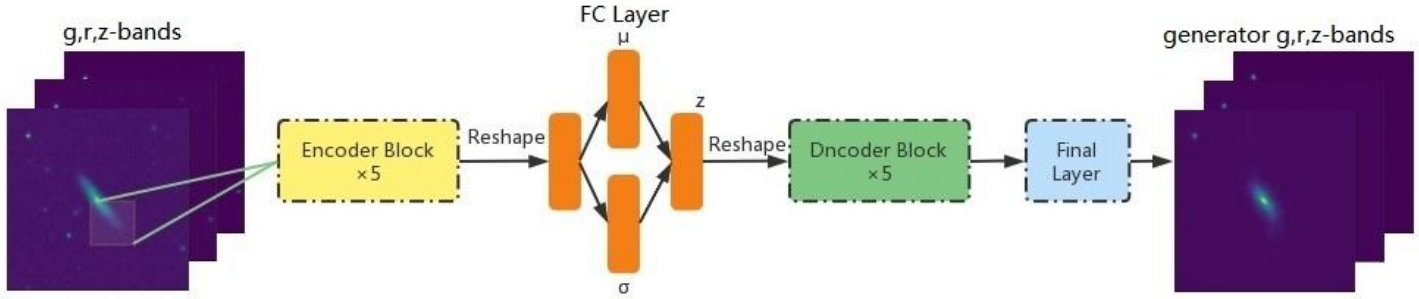


er plot of the first, second, and fifth VAE components, with the four tracks discussed in subsection 3.2 overlaid. he (S) star formation track, the (E) extreme line emitters, the (P) post-starburst track, and the (A) active

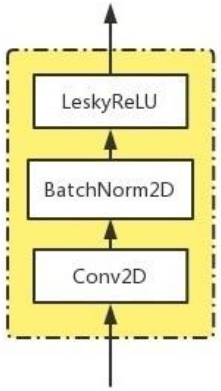
研究方法与进展

- 研究方法
- 研究进展

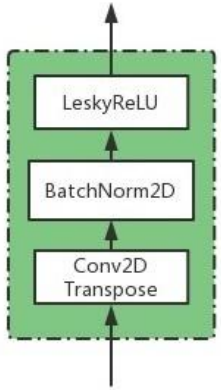
变分自编码器网络的搭建



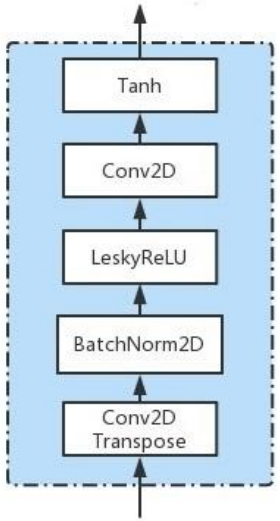
(a) VAE网络结构图



(b) Encoder模块



(c) Decoder模块

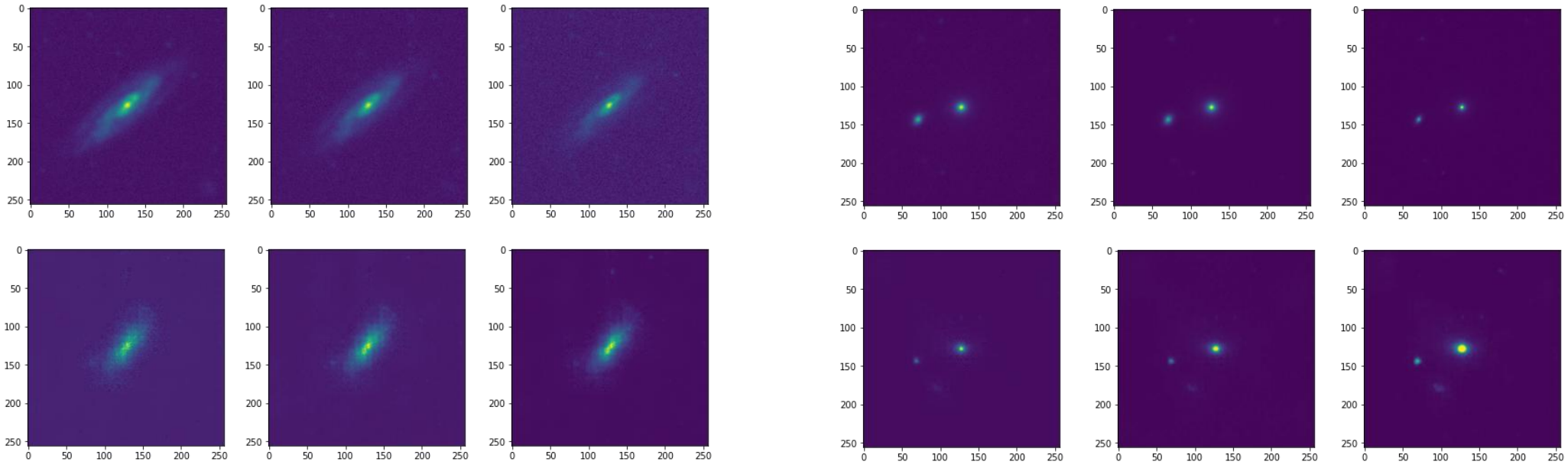


(d) 编码器输出后，对反卷积输出数据维度大小调整为原数据维度大小

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{rec}} + k \cdot \mathcal{L}_{\mathrm{KL}} = ||x - \overline{x}|| + k \cdot D_{\mathrm{KL}}(q(\mathbf{z}|x)||p(\mathbf{z})).$$

研究进展

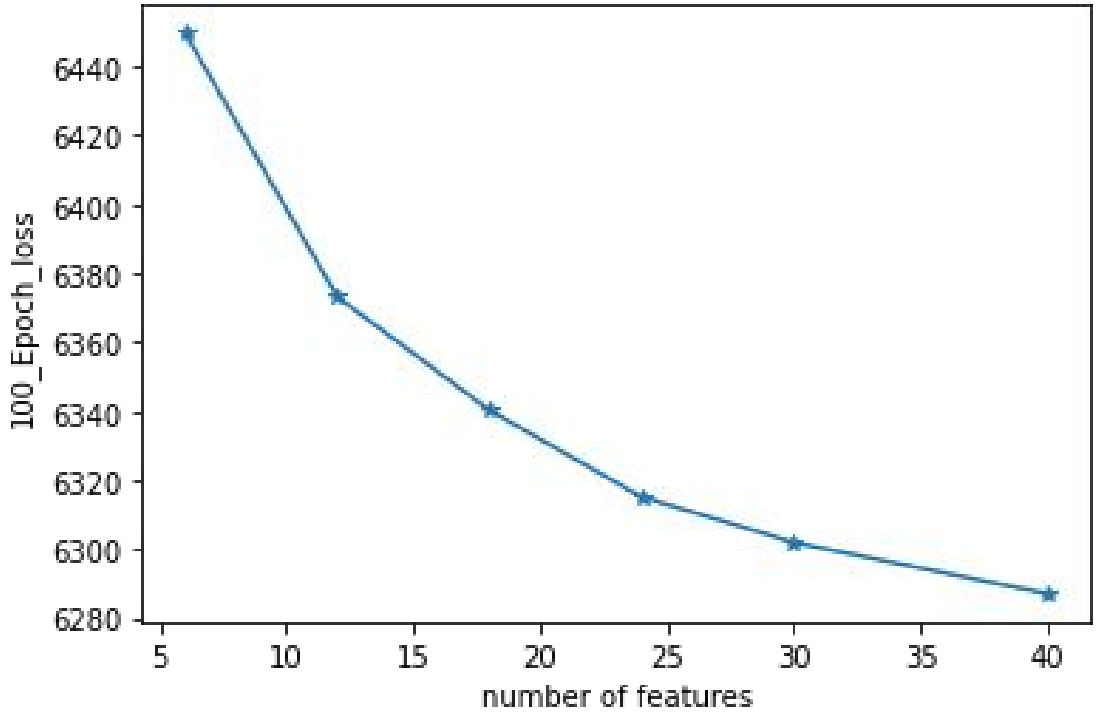
变分自编码器网络对星系图像进行特征提取与还原，隐变量可还原与原图特征类似的星系图像，结果表示隐变量确实可以对星系形态学特征进行描述。



研究进展

根据VAE提取不同维度变量的重构图与原图的MSE损失值。损失值评价学习过程中生成图像与原图的偏差，无法定量表示生成图像与原图之间相似度。

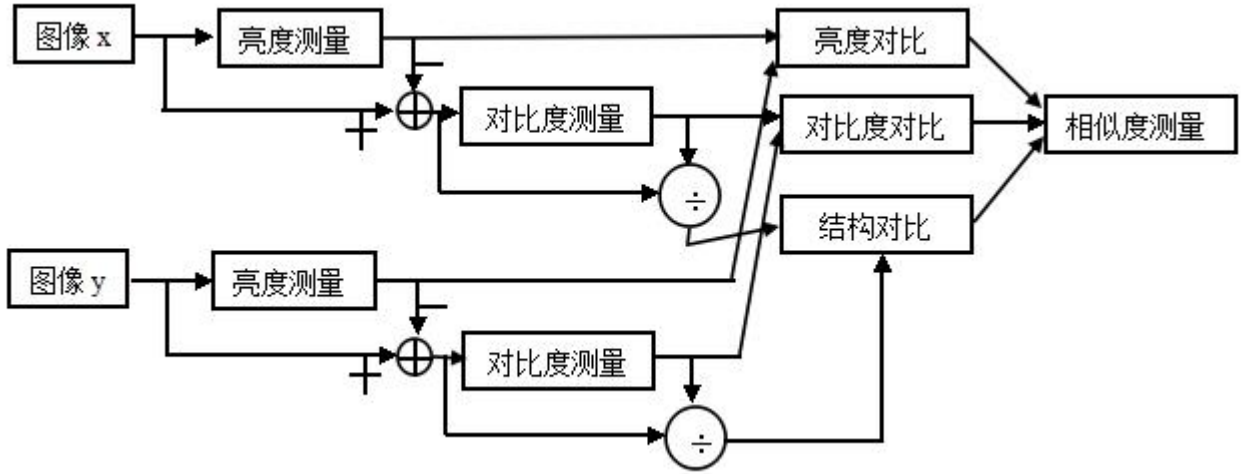
评价重构图像与原图的相似度用SSIM ，结构相似性指标 (structural similarity index measure, SSIM)，是衡量生成图像和原始图像相似程度的指标。



研究方法

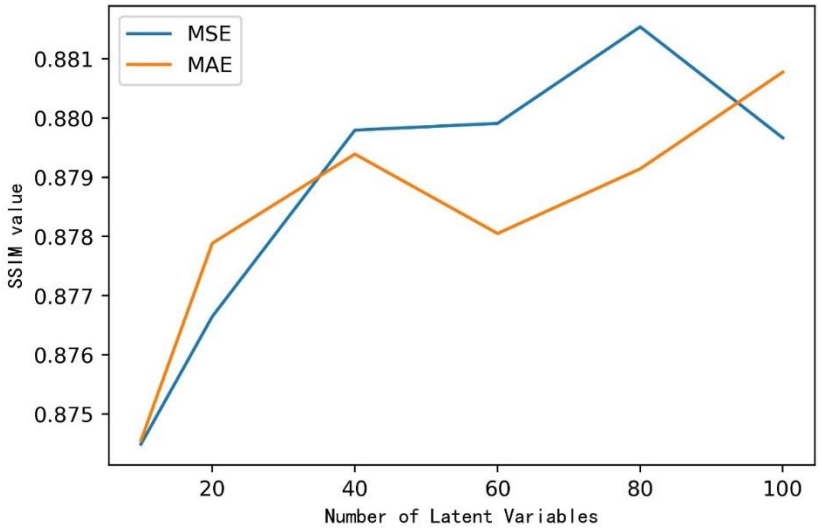
结构相似性指标 (structural similarity index measure, SSIM)是从原图和生成图像同一位置取一个 N × N 的窗口，根据不断滑动窗口计算该窗口的相似指标，最后取所有窗口的平均值作为全局的结构相似性指标。

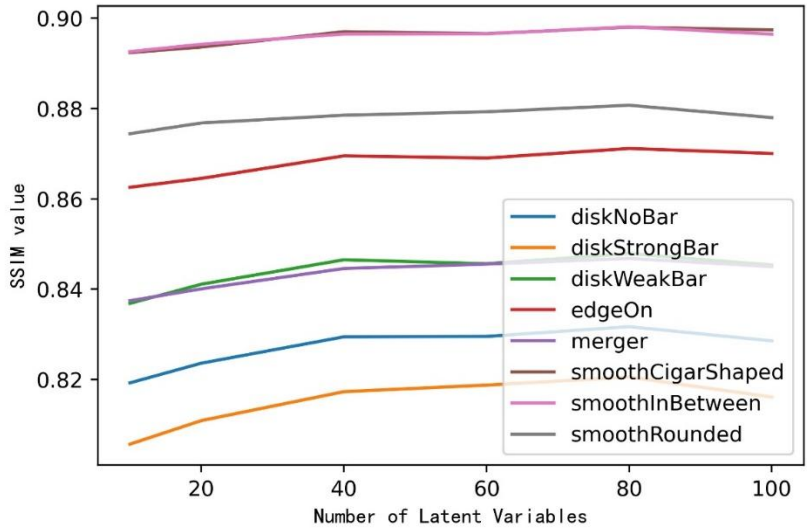
相似指标由原图和生成图像的某窗口对亮度 (luminance)、对比度 (contrast) 和结构(structure) 三个值进行度量。



研究进展

通过各类重构评价指标的SSIM值，对比得到MSE和40维隐变量来解释DECaLS图像较为合理。

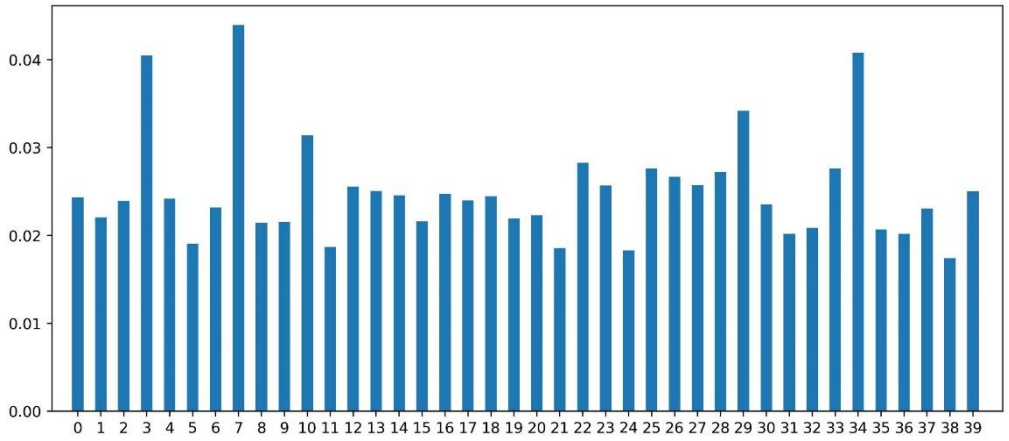




研究进展

选取合适GZD-5分类树的标签与40维隐变量特征，用随机森林探索40维隐变量的各个分量在投票各个标签的比重。

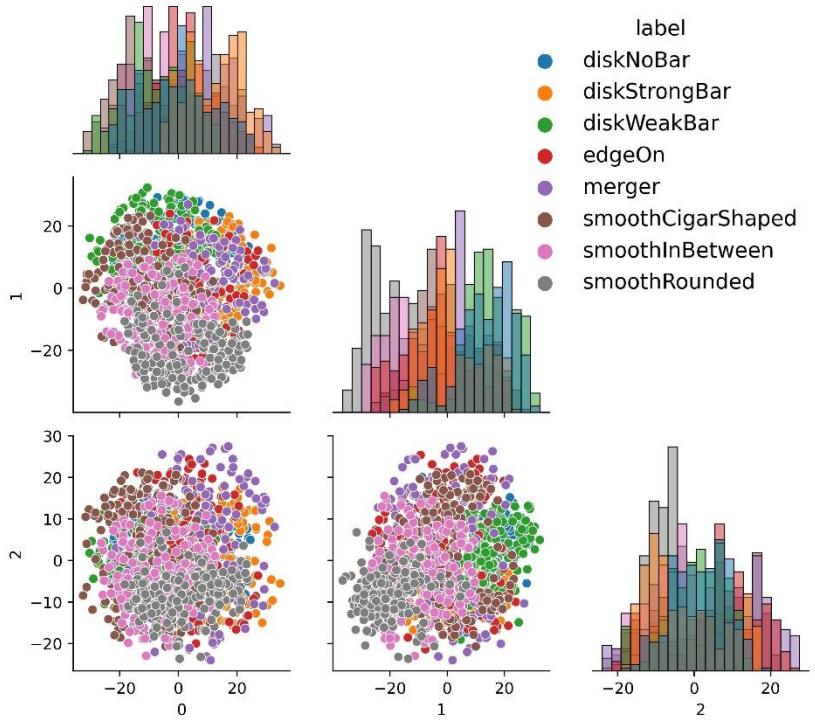
类别	形态特征	阈值筛选
Smooth	Round	smooth or featured smooth fraction > 70%
		how rounded round fraction > 80 %
	In Between	smooth or featured smooth fraction > 70%
		how rounded in between fraction > 85 %
	Cigar	smooth or featured smooth fraction > 50%
		how rounded cigar shaped fraction > 60 %
Edge-on	Edge on Disk	smooth or featured featured or disk fraction > 50%
		disk edge on yes fraction > 70 %
	Something Else	smooth or featured featured or disk fraction > 50%
		disk edge on no fraction > 70 %

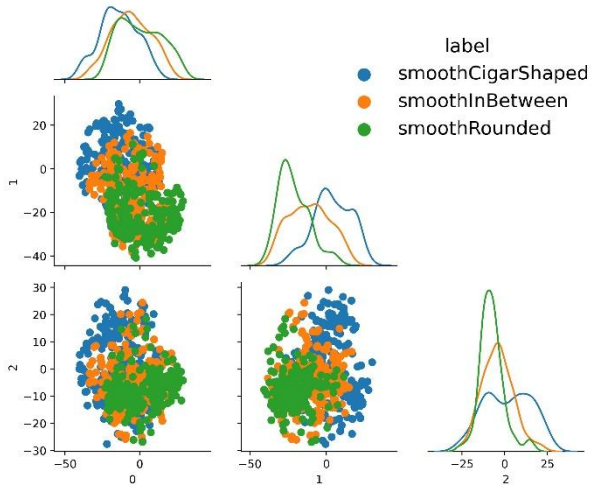


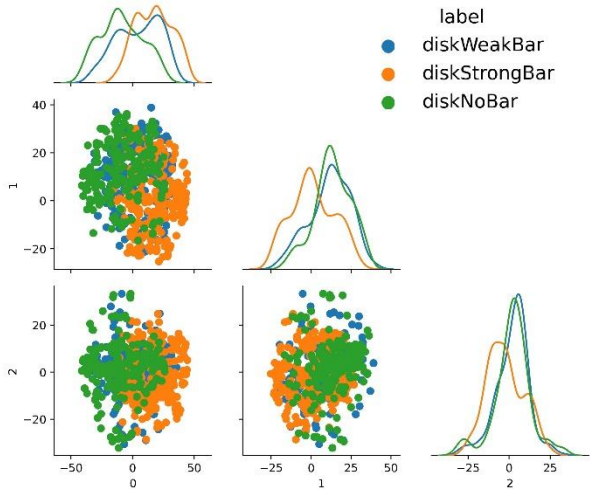
特征	类别数量	训练精度	测试精度
Smooth	3	99.85%	86.92 %
Edge-on	2	99.72%	92.00 %
Bulge Shape	3	95.35%	82.44 %
Bar	3	98.43%	72.08 %
Arm	2	97.83%	82.10 %
Arm Tightly	3	85.38%	68.78 %
Arm Number	5	87.69%	74.65 %
Bulge Size	5	86.03%	75.78 %
Merger	2	98.68%	94.65 %

隐变量特征t-SNE可视化分析

40维隐变量通过t-SNE降维可视化，看到其隐变量在低维空间的分布具有明显区分，其隐变量的分布具有区域性，符合其物理结构度量。

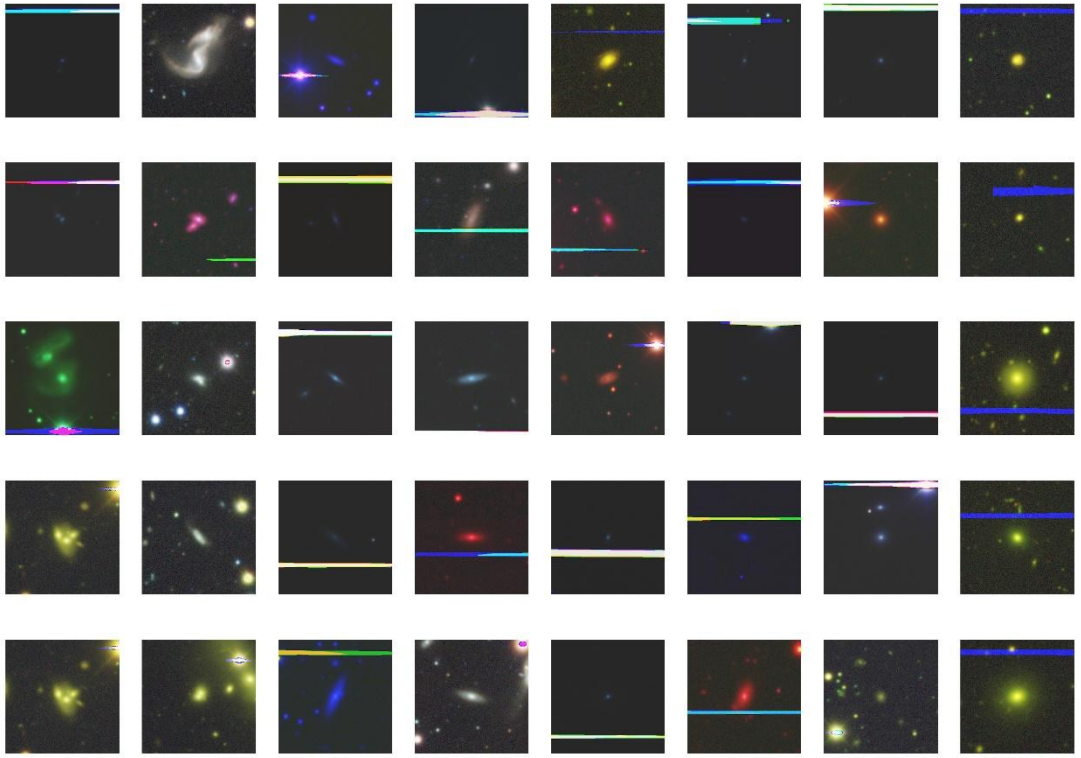






离群点分析

40维隐变量在隐空间中，选出各个类别中距离聚类中心较远的隐变量称为异常点，异常点可能是具有不常见形态特征的星系、有巨大扰动的星系或者不能很好重构的星系。



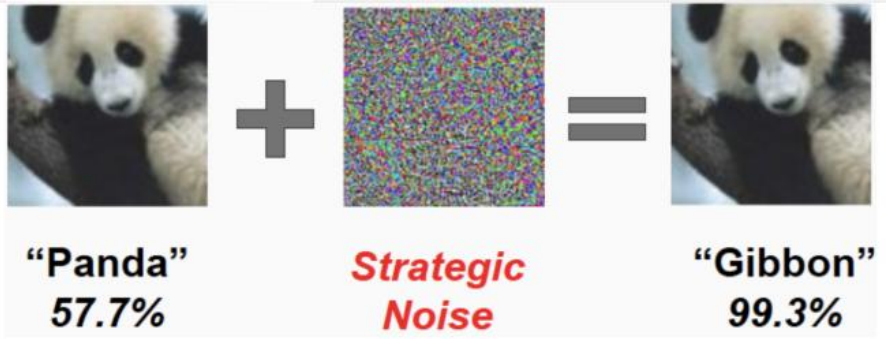
进行数据清洗，去除DECaLS数据集卫星等噪声存在对学习任务的干扰。

找到一些特征罕见的星系形态特征。

迁移学习

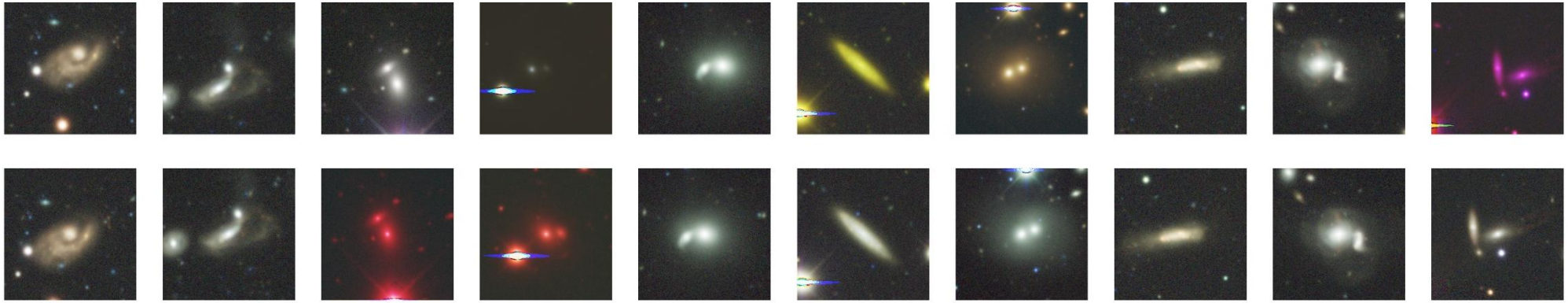
实现未来巡天计划快速检索，VAE模型需要起码做到：

- VAE模型能够识别各类星系图像形态特征
- 不同巡天项目由于psf(Point Spread Function)、望远镜口径等差异在特定天区得到不同星系图像，VAE模型识别隐变量蕴含的形态学特征也应相同。



DECaLS

BASS+MzLS



迁移学习--MMD

Max Mean Discrepancy 最大均值差异。MMD常被用来度量两个分布之间的距离，是迁移学习中常用的损失函数。

最大均值差异（MMD）：

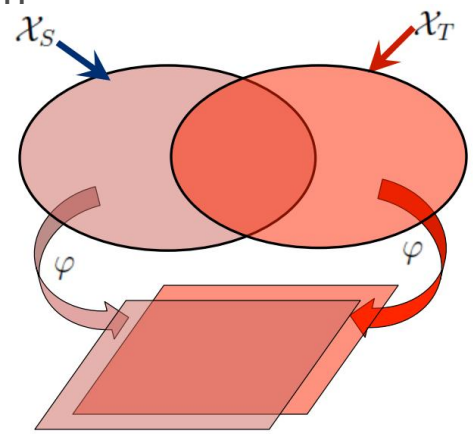
- p分布生成样本空间P
- q分布生成样本空间Q

在泛函空间$\mathcal{F}$寻找一个映射f使得$|mean(f(P))-mean(f(Q))|$有最大值，即为p、q分布的MMD

$$\mathrm{MMD}[\mathcal{F},p,q]:=\sup_{f\in\mathcal{F}}\left(\mathbf{E}_{x\sim p}[f(x)]-\mathbf{E}_{y\sim q}[f(y)]\right)$$

$$mean(f(P)) == mean(f(Q))$$

代表p和q是同一分布

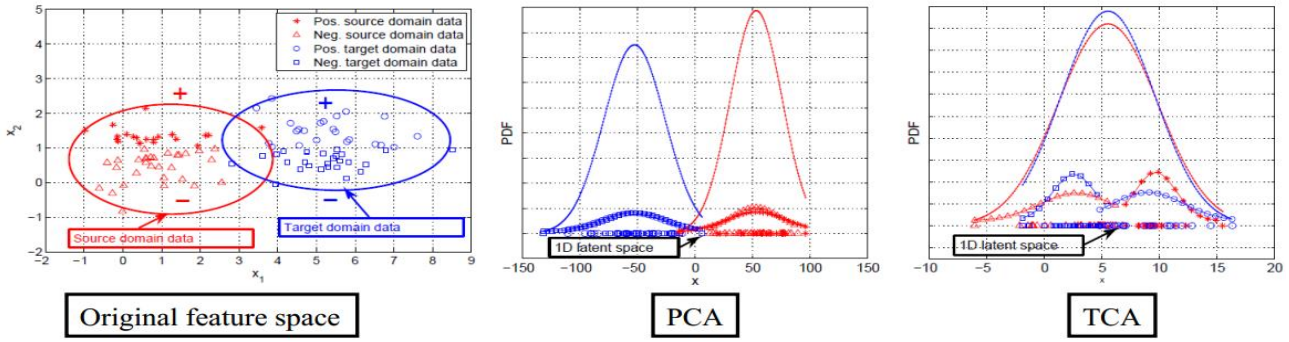


迁移学习--MMD

MMD算法得益于TCA（Transfer Componet Analysis）算法的提出。

TCA和PCA算法有点像，可以实现降维，两个高维的大矩阵（源域和目标域矩阵）进去，得到两个低维的矩阵（降维后的源域和目标域矩阵）。

TCA可以将分布不同的源域和目标域数据映射到高维再生核希尔伯特空间中，然后不断缩小源域和目标域的距离并最大程度的保留其内部属性。

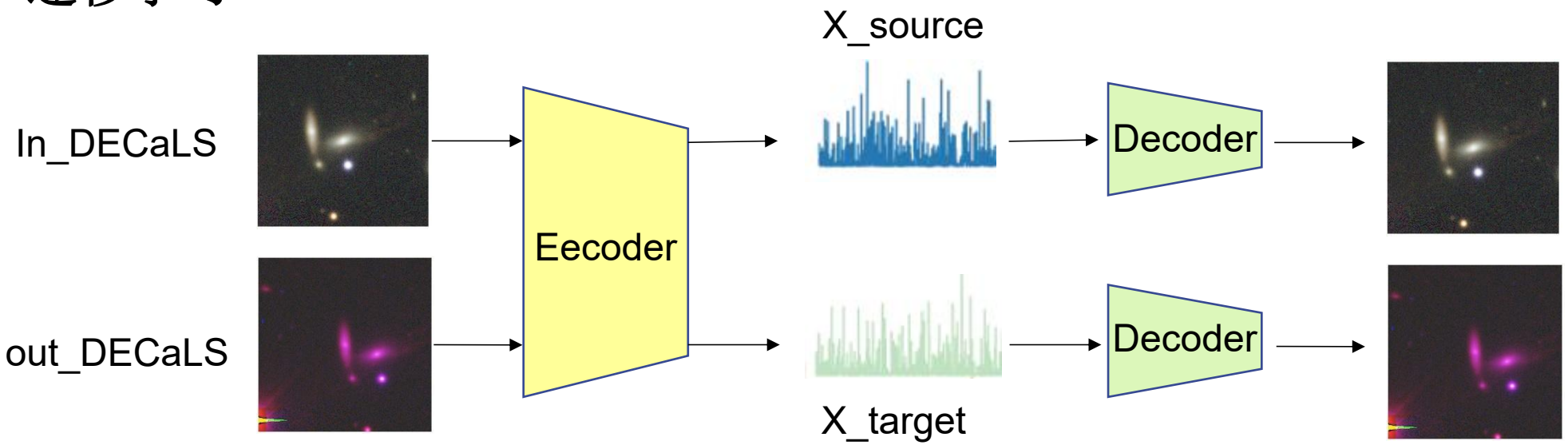


Original feature space　　PCA　　TCA

$$\mathrm{MMD}[\mathcal{F},p,q] := \sup_{f\in\mathcal{F}} \left(\mathbf{E}_{x\sim p}[f(x)] - \mathbf{E}_{y\sim q}[f(y)]\right) \Rightarrow MMD\left(X_S, X_T\right) = \left\| \frac{1}{|X_S|}\sum_{x_s\in X_S}\phi(x_s) - \frac{1}{|X_T|}\sum_{x_t\in X_T}\phi(x_t)\right\|$$

研究方法

迁移学习



$$\mathcal{L} = \mathcal{L}_{VAE} + \lambda \cdot MMD^2(X_S, X_T)$$

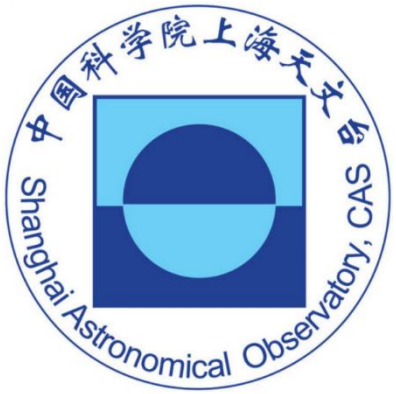
VAE具有识别各类星系图像形态特征能力。同一天区在不同巡天计划的星系图像在VAE下识别到形态学特征一样，将VAE的编码器结构进行迁移学习的域迁移。

用BASS+MzLS巡天计划与DECaIS重合天区的数据进行同构迁移学习，调整VAE模型在其他天区的学习状态。

迁移学习结果

特征	类别数量	迁移前精度	迁移后精度
Smooth	3	67.57%	80.42 %
Edge-on	2	74.07%	91.05 %
Bar	3	64.71%	79.41 %
Merger	2	78.62%	86.23 %

以上用线性分类器——随机森林所得结果，表明VAE模型可以从DECaLS天区迁移到非DECaLS天区进行特征分类，变分自编码器对广大天区特征无监督的提取，有望广泛应用于未来大型巡天项目的快速检索、分类和回归。



Thanks!

报告人：徐权峰　　指导教师：沈世银　　日期：2022年7月20日