Publications of the Astronomical Society of the Pacific, 134:114508 (15pp), 2022 November https://doi.org/10.1088/1538-3873/aca04e
© 2022. The Astronomical Society of the Pacific. All rights reserved.

Unsupervised Galaxy Morphological Visual Representation with Deep Contrastive Learning

Shoulin Wei[1,2], Yadi Li[1], Wei Lu[1], Nan Li[3], Bo Liang[1,5], Wei Dai[1], and Zhijian Zhang[4]
[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, People's Republic of China
liangbo@astrolab.cn
[2] Computer Technology Application Key Lab of Yunnan Province, Kunming University of Science and Technology, Kunming, 650500, People's Republic of China
[3] National Astronomical Observatories Chinese Academy of Sciences, Beijing, 100012, People's Republic of China
[4] Faculty of Science, Kunming University of Science and Technology, Kunming, 650500, People's Republic of China
Received 2022 June 11; accepted 2022 November 2; published 2022 November 29

Abstract

Galaxy morphology reflects structural properties that contribute to the understanding of the formation and evolution of galaxies. Deep convolutional networks have proven to be very successful in learning hidden features that allow for unprecedented performance in the morphological classification of galaxies. Such networks mostly follow the supervised learning paradigm, which requires sufficient labeled data for training. However, the labeling of a million galaxies is an expensive and complicated process, particularly for forthcoming survey projects. In this paper, we present an approach, based on contrastive learning, with aim of learning galaxy morphological visual representation using only unlabeled data. Considering the properties of low semantic information and contour dominated of galaxy images, the feature extraction layer of the proposed method incorporates vision transformers and a convolutional network to provide rich semantic representation via the fusion of multi-hierarchy features. We train and test our method on three classifications of data sets from Galaxy Zoo 2 and SDSS-DR17, and four classifications from Galaxy Zoo DECaLS. The testing accuracy achieves 94.7%, 96.5% and 89.9%, respectively. The experiment of cross validation demonstrates our model possesses transfer and generalization ability when applied to new data sets. The code that reveals our proposed method and pretrained models are publicly available and can be easily adapted to new surveys.[6]

Unified Astronomy Thesaurus concepts: Galaxy evolution (594); Surveys (1671); Astrostatistics techniques (1886)

1. Introduction

Galaxy morphology provides the most intuitive features to inspect the structure of galaxies, which has a strong correlation with galactic star formation history. Massive early type galaxies tend to appear as elliptical, while spiral and irregular galaxies show evidence for on-going star formation. Galaxy morphological classification can lead to important astrophysical insights that help to explore the formation and evolution of galaxies (Wilman & Erwin 2012). On the flip side, any theory of galaxy formation and evolution also has to explain the structures of dazzling galaxies.

The classification of galaxies dates back to the most famous Hubble sequence (Hubble 1926). The De Vaucouleurs system (Vaucouleurs 1959) extends the Hubble sequence to form four main categories: Elliptical, Lenticular, Spiral and Irregular. In particular, the Spiral type is subdivided into Bars, Rings and Spiral Arm, which can still be divided into subtler classifications according to the tightness of spiral arms.

The traditional methods of morphological classifications based on visual inspection and fitting light profiles have achieved remarkable success. However, these methods are not unfeasible in the era of large data volume. Moreover, forthcomming survey projects will reach increasing depths and widths over nearly the entire extragalactic sky and produce massive numbers of galaxies. The Large Synoptic Survey Telescope (LSST) will produce billions of galaxies covering the entire southern sky, in multiple bands (Robertson et al. 2019). The space-borne telescopes, e.g., Euclid (Racca et al. 2016) and the Chinese Space Station Telescope (CSST) (Zhan 2021) will cover 15,000 and 17,500 square degrees of the celestial sky respectively and are expected to detect billions of galaxies.

In response to the coming exponential growth in the data volume of galaxies, automated morphological classification has become an inevitable approach. In the past few years, we have witnessed the tremendous success of the Convolutional Neural Network (CNN) based Deep Learning (DL) in various computer vision tasks. Meanwhile, CNNs also have been extensively studied for the morphological classification of galaxies. In a Kaggle competition of galaxy morphological classification on Galaxy Zoo 2 (Willett et al. 2013), a CNN based model finished in first place out of 326 participants (Dieleman et al. 2015). Cheng et al. (2020b) carry out a comparison between CNN and the

[5] Author to whom any correspondence should be addressed.
[6] https://github.com/kustcn/galaxy_contrastive

traditional machine learning methods, including K-nearest neighbor, logistic regression, Support Vector Machine, Random Forest, and Neural Networks, over the Dark Energy Survey (DES) data and Galaxy Zoo 1. Their results show that CNN has the best performance. Khalifa et al. (2018) proposed a Deep Galaxy V2 based on Deep Convolutional Neural Networks for galaxy classification. The architecture was trained over 4238 images taken from the EFIGI catalog (Baillard et al. 2011) and achieved a 97.772% testing accuracy on three classifications (Elliptical, Spiral, and Irregular). Zhu et al. (2019) proposed a variant of residual network, Resnet-26, which classifies 28790 galaxy image samples from Galaxy Zoo 2 into five categories. The results show that Resnet-26 achieves the best performance with 95.2083% accuracy, compared with AlexNet, VGG, Inception and Resnet. Katebi et al. (2019) used Capsule Network (CapsNet) for regression and predicted probabilities for all of the questions in the Galaxy Zoo project and trained a CapsNet classifier that outperforms the baseline CNN by a 36.5% error reduction. Gupta et al. (2022) introduced a continuous depth version of the Residual Network called Neural Ordinary Differential Equations (NODE) which obtained an accuracy between 91% and 95% depending on the classifications. DL methods used for galaxy morphological classifications are discussed in Tuccillo et al. (2016), Khan et al. (2019), Ghosh et al. (2020), Bhambra et al. (2022), Zhang et al. (2022) and Vavilova et al. (2022).

The machine learning methods in galaxy morphological classification are still dominated by supervised learning at present. However, acquisition of large training sets with labeled data is one of the key challenges, especially when approaching a new survey project. Transfer learning, to some extent, allows us to deal with a new domain without labeled data, by leveraging the already existing trained model with labeled data of some related task. Domínguez Sánchez et al. (2018) proposed a transfer learning method with a CNN model trained with Sloan Digital Sky Survey (SDSS) data. When applying the models directly to unlabeled Dark Energy Survey (DES) data, the accuracy reaches approximately 90%. In Variawa et al. (2022), the proposed method used the pre-trained ResNet50 based model, fined-tuned on the Galaxy Zoo 2 and EFIGI with expert labeled data. Using nine Hubble classes, this model achieved an accuracy of 30.50% on the Revised Shapley-Ames (RSA) catalog as test data. Although these results have proved the effectiveness of transfer learning from one survey to another, the models still need to be fine-tuned to generalize well to a data set with similar distribution, but would fail catastrophically on unexpected or atypical inputs, like a totally different redshift distribution.

Compared with supervised learning, unsupervised learning that aims to uncover the internal structure of a data set without the need for any label will be the solution for automated morphological classification of the new and forthcoming surveys. Schutter & Shamir (2015) used unsupervised analysis to quantitatively obtain similarities between the different morphological types using merely the galaxy images, whereas the purpose of the method is not to automatically classify galaxies, and the method relies heavily on feature extraction and analysis. Martin et al. (2020) constructed an unsupervised model based on patches for the automatic segmentation and labeling of galaxy images, where Growing Neural Gas (GNG) is used for feature extraction, Hierarchical Clustering (HC) for a hierarchical representation and Connected Component Labeling for construction of object feature vectors. Galaxy morphological classification is obtained by comparing the similarity between feature vectors. Cheng et al. (2020a) combined feature extraction with a vector-quantized variational autoencoder (VQ-VAE) and HC for unsupervised machine learning to explore galaxy morphological analysis. The test accuracy reached 87% when performing a binary classification over two large preliminary clusters.

Inspired by the success of contrastive learning in computer vision (Hjelm et al. 2019; He et al. 2020 and van den Oord et al. 2018), we are interested in exploring whether it could also be used for unsupervised galaxy morphological visual representation. Hayat et al. (2021) have applied CL to perform galaxy morphological classification and photometric redshift estimation on multiband galaxy photometry from SDSS. Their results achieve the accuracy of supervised models while using $2-4$ times fewer labels for training. Sarmiento (2020) adopted the Simple framework for Contrastive Learning of visual Representations (SimCLR) to galaxy morphological classification on SDSS galaxy images with $z < 0.15$. The accuracy achieved 85%, which was similar to the results of their supervised learning approach.

Based on contrastive learning, we propose an unsupervised learning network that includes a well-designed encoder, to learn multi-hierarchy feature representation. We carried out training on the data sets of Galaxy Zoo 2, SDSS-DR17 and Galaxy Zoo DECaLS, with test accuracy of classification reaching 94.7%, 96.5%, and 89.9%, respectively. In order to verify the robustness of the model, we performed cross validation in that the model was trained with one of the three data sets, and tested on the other two data sets. The accuracies of cross validation are reduced less and still achieved approximately 90%. The results show that the model possesses high transfer and generalization ability, and could be applied to the new surveys. The results of experiments performed with hierarchical clustering demonstrate that the visual representation learned include detailed features in which clustering can identify subtler classifications.

The remainder of this paper is organized as follows: Section 2 describes the morphological catalog used in this work and data preprocessing. Section 3 describes the architecture of the proposed network and our methodology. Section 4 reports the experiments performed and discusses the results. Finally, Section 5 concludes the paper and describes ideas for future work.

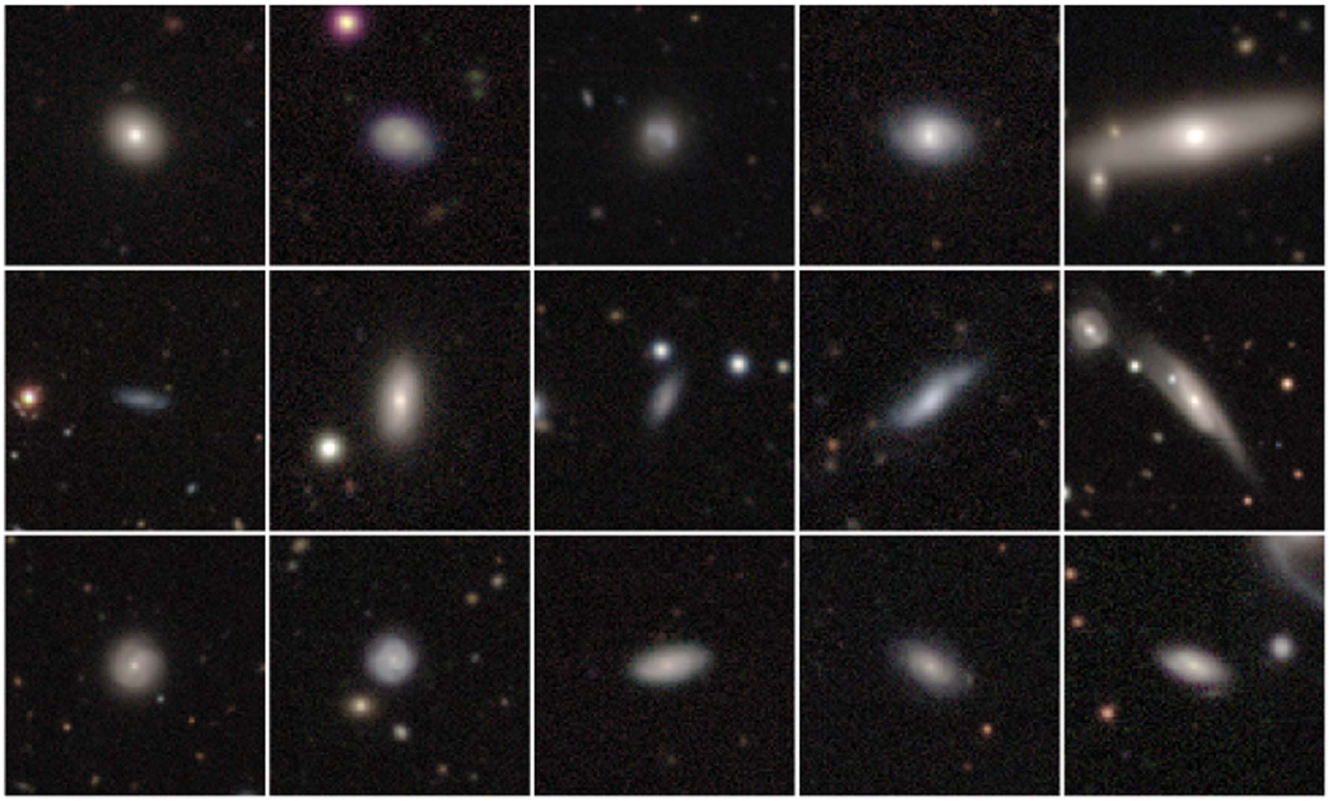


Figure 1. Examples of a sample of galaxies from GZ2. The first row shows five images labeled as Round type. but it can be seen that the first two have little difference in contour information from elliptical galaxies, the third and fourth have slightly spiral features, and the last one is more similar to the lateral and cigar shape; but these five images are all classified as round galaxies. Similarly, the labels in the second row are lateral and cigar-shaped. The first three are similar to elliptical galaxies, while the last two have spiral arms and are more suitable for spiral galaxies. The third row is labeled spiral galaxies, but because the spiral arms are less distinctive, the first two are more circular and the last three are more elliptical.

2. Dataset

In this paper, we train the proposed model and test the performance with sample images from GZ2, GZ DECaLS and SDSS-DR17. In this section, we describe the galaxy catalogs, the sampling strategy and threshold configurations of the constructed data sets used for this study.

2.1. Galaxy Zoo

Galaxy Zoo (GZ) is an online crowdsourcing project which invites volunteers to visually inspect and classify galaxies via the internet (zooniverse.org) (Lintott et al. 2008). Since 2007, volunteers involved have performed a billion classification tasks. Their work has greatly promoted the development of galaxy morphological classification, and led to a robust database of galaxy classifications (Lintott et al. 2011) including GZ2, GZ Hubble, GZ CANDELS and others. At present, GZ has gone to the 4th phase launched in 2012 (Simmons et al. 2016).

In this work, the two data sets used are drawn from the GZ2 and Dark Energy Camera Legacy Survey (DECaLS) in the GZ catalogs. The GZ2 catalog presents morphological classification of more than 300,000 nearby galaxies from SDSS (Willett et al. 2013). The GZ DECaLS catalog, taken from the Victor M. BLanco 4M telescope, has published classifications of 314,000 galaxies so far (Walmsley et al. 2022). The GZ classifications are based on the volunteers' answers to questions of binary decision trees (The details of decision trees are depicted at https://data.galaxyzoo.org/gz_trees/gz_trees.html). Volunteers are asked various questions, such as "How rounded is it?" and "Could this be a disk viewed edge-on?" Various volunteers' answers for one image produce a vote fraction for each feature. For example, a image with $F_{\mathrm{edge-on}} = 0.7$ represents that 70% of volunteers answered "Yes" to the question "Could this be a disk viewed edge-on?" of this image. Unsupervised learning works on unlabeled data when training models. Without this powerful correction, it is necessary to enhance the differences between different classes of data sets. Selection with moderate thresholds such as $F_{\mathrm{edge-on}} > 0.5$, $F_{\mathrm{features,disk}} > 0.5$ and $F_{\mathrm{not-clumpy}} > 0.5$ may lead to fewer differences between classes (as shown in Figure 1). Therefore, we prune these data sets and merely select those images that are classified with high vote fractions in their respective classes. The details of vote fraction thresholds for

Table 1
The Vote Fraction Threshold and Selection Criteria of Samples from GZ2, GZ DECaLS and SDSS

Type	GZ2		SDSS		GZ DECaLS	
	Selection	Count	Selection	Count	Selection	Count
Round					$F_{\text{smooth}} \geqslant 0.9$ $F_{\text{completelyround}} \geqslant 0.9$	2926
Elliptical	$F_{\text{smooth}} \geqslant 0.7$ $F_{\text{in-between}} \geqslant 0.7$	2203	zns.p_el $\geqslant$0.8	3000	$F_{\text{smooth}} \geqslant 0.95$ $F_{\text{in-between}} \geqslant 0.95$	5263
Edge	$F_{\text{features,disk}} \geqslant 0.7$ $F_{\text{edge-on/yes}} \geqslant 0.7$	2058	zns.p_edge $\geqslant$0.8	2994	$F_{\text{features,disk}} \geqslant 0.79$ $F_{\text{edge-on/yes}} \geqslant 0.79$	5871
Spiral	$F_{\text{features,disk}} \geqslant 0.9$ $F_{\text{edge-on/no}} \geqslant 0.9$ $F_{\text{spirals/yes}} \geqslant 0.9$	2907	zns.p_cw $\geqslant$0.8 zns.p_acw $\geqslant$0.8	1919 2001	$F_{\text{features,disk}} \geqslant 0.9$ $F_{\text{edge-on/no}} \geqslant 0.9$ $F_{\text{spirals/yes}} \geqslant 0.9$	2224
Count		7168		9914		16284

selection are listed as Table 1. After pruning, we have a total of 7168 images from GZ2 and 16,284 images from GZ DECaLS. The samples from GZ2 are in the classifications of Elliptical, Edge-on and Spiral. Each image of GZ DECaLS is divided into four classes (Round, Elliptical, Edge-on, and Spiral).

2.2. SDSS

This work uses the third data set acquired from the latest 17th public data release of SDSS Phase IV (SDSS-IV) (Domínguez Sánchez et al. 2022). The Catalog Archive Server (CAS) of the SDSS provides a Structured Query Language (SQL) web-based interface for researchers to retrieve galaxies of interest. The full SQL we used to retrieve data for each sample, can be found at https://github.com/kustcn/galaxy_contrastive/blob/master/prepdata/sdss_pdr17_query.sql, and can be run through the SDSS data access site at http://skyserver.sdss.org/dr17/SearchTools/sql (accessed 2022 January 7).

The detailed selection criteria for SDSS are described in Table 1. zns.p_el represents the probability that the sample will be considered as elliptical, zns.p_cw represents the probability of it having a clockwise spiral shape, and zns.p_acw represents the probability of it having a counterclockwise spiral. In total, this sample from SDSS includes 9914 galaxies classified into the same three classes as the samples from GZ2. In addition, the global conditions for all SDSS samples include g. clean > 1, g.scale $= 0.2$, and zns. nvote > 20, which indicate selecting those objects with a degree of clean greater than 1, where each image has a scale factor of 0.2, and each image has more than 20 classified votes, respectively.

2.3. Pre-processing

Each image from GZ2 and GZ DECaLS is an RGB composite image of 424×424 pixels in size. The galaxy of interest is generally located at the center of the image. The image quality in GZ2 is relatively poor, with an average size of a dozen KB, while the average size of a single image in GZ DECaLS reaches 300 KB. This means that the resized pixels with small scale in GZ2 result in seriously distorted images. All images are fed to the training model as 84×84 by 3-channel RGB values, which is a reasonable compromise between image quality and computational efficiency. In practice, we centrally crop and scale the images to that size. Figure 2 shows the steps of the preprocessing procedure for GZ2 and GZ DECaLS, which allows the main information to be contained in the center of the image and eliminates all random noise, such as some other secondary object. In the selection of SDSS, we uniformly define scale $= 0.2$ and cutout size as 128×128, so that the original downloaded images have the size of 128×128. Therefore, the preprocessing procedure for SDSS only includes scaling from 128×128 to 84×84.

Note that unsupervised learning does not require labels in the training phase, we only use labels in the testing phase to verify the accuracy of our proposed method.

3. Methodology

In this section, we first briefly introduce the overall architecture of our proposed method, as shown in Figure 3. We then present the strategies of data augmentation. Next, we discuss the construction of the encoder that converges the multi-hierarchy features with deep learning. Finally, we describe the definition of loss function.

3.1. Architecture

The architecture of our method is mainly based on a strong framework for contrastive learning SimCLR (Chen et al. 2020). First, for each image in a batch, we apply two random data augmentations to obtain two different views. For one image X,

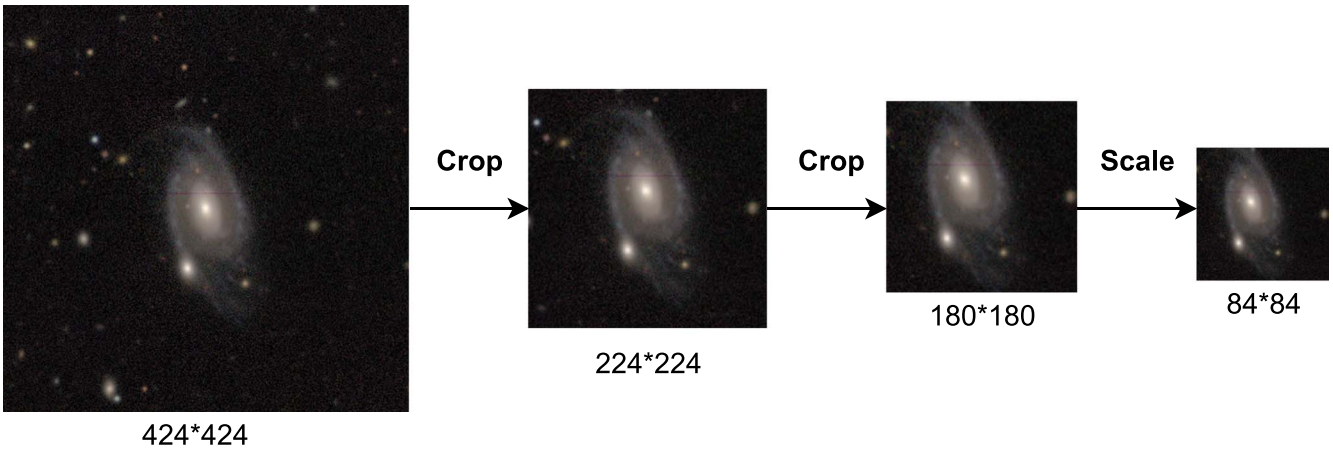


Figure 2. The steps of the preprocessing procedure for GZ2 and GZ DECaLS. The first crop acts on the center of the image. In order to increase randomness and disturbance, the 180 × 180 image is randomly cropped out of the 224 × 224 image. Scaling is performed on GZ2 and GZ DECaLS from 180 × 180 to 84 × 84.

the augmented views (X^{+1} and X^{+2}) are called the positive pair of this image. While the views of other images are called corresponding negative pairs. Next, the Encoder learns the feature representation that best describes each view and produces a one-dimensional vector of size 128 for each view. Accordingly, f^{+1} and f^{+2} are the representation for X^{+1} and X^{+2}, and f^{-} for negative pairs. Finally, the similarity, Sim, between feature representations is measured with cosine similarity. The goal of contrastive learning is to increase the similarity between instances from the positive pairs and decrease the similarity between the positive and negative pairs. Instance-wise contrastive learning achieves this objective by optimizing a contrastive loss function (discussed in Section 3.4) and back-propagation through the Encoder to update the network parameters.

3.2. Data Augmentation

In computer vision tasks of supervised learning, data augmentation generally is used to expand the size of a training set and to increase variability by creating modified data from the existing data. Data augmentation has played a core role of creating pretext tasks for training a contrastive learning network. For this purpose, we construct the positive pair and negative pairs by randomly augmenting all the images in a batch twice. Figure 4 shows the original image and the images after applying these augmentations:

1. HorizontalFlip and VerticalFlip: an image is flipped with a probability of 0.5 horizontally and vertically.
2. Crop and Resize: random crop with a scale factor of 0.8, and then scale to the original size (84 × 84).
3. ColorJitter: random adjustment of brightness, contrast, saturation and hue of the image. The corresponding values are 0.2, 0.5, 0.5 and 0.4. Taking 0.2 as an example, this means to randomly change the brightness to between 80% (1–0.2) and 120% (1+0.2) of the original image brightness.
4. Rotation: random rotation with an angle sampled uniformly between 0° and 360°.
5. Composed: composition with the series of transformations mentioned above.

3.3. Encoder

The goal of the Encoder is to identify latent structures by using pixel data alone. Galaxy images, especially, are relatively lacking in semantic information, with monotonous colors and single contour features. Some galaxies with strong luminosity even have no obvious contours, but only scattered halos. Therefore, the feature extraction of galaxy images is more susceptible to noise interference than common data sets used in computer vision tasks. We choose current state-of-the-art transformers (Vaswani et al. 2017) combined with CNN for learning the multi-hierarchy features representation of the input images. Instead of using only coarse semantic features, as conventional CNNs do, we resort to the capability of the Vision Transformer (ViT) on global spatial information and CNN on local information for capturing the rich representation of galaxies. As shown in Figure 5, we use ResNet-18 (He et al. 2016) as the backbone network. As discussed in Raghu et al. (2021), ViT incorporates more global information than ResNet at lower layers. The training data is fed into the ResNet-18 base and the ViT sub-networks in parallel. Thus, multi-hierarchy features depends on fusing the learned features from these two sub-networks.

Considering that galactic images have less semantic information and the classification mainly relies on contour features, we reduce the depth of the network model (remove the fourth layer of ResNet-18), in order to quickly extract contour features. We also remove the maximum pooling in the first convolutional

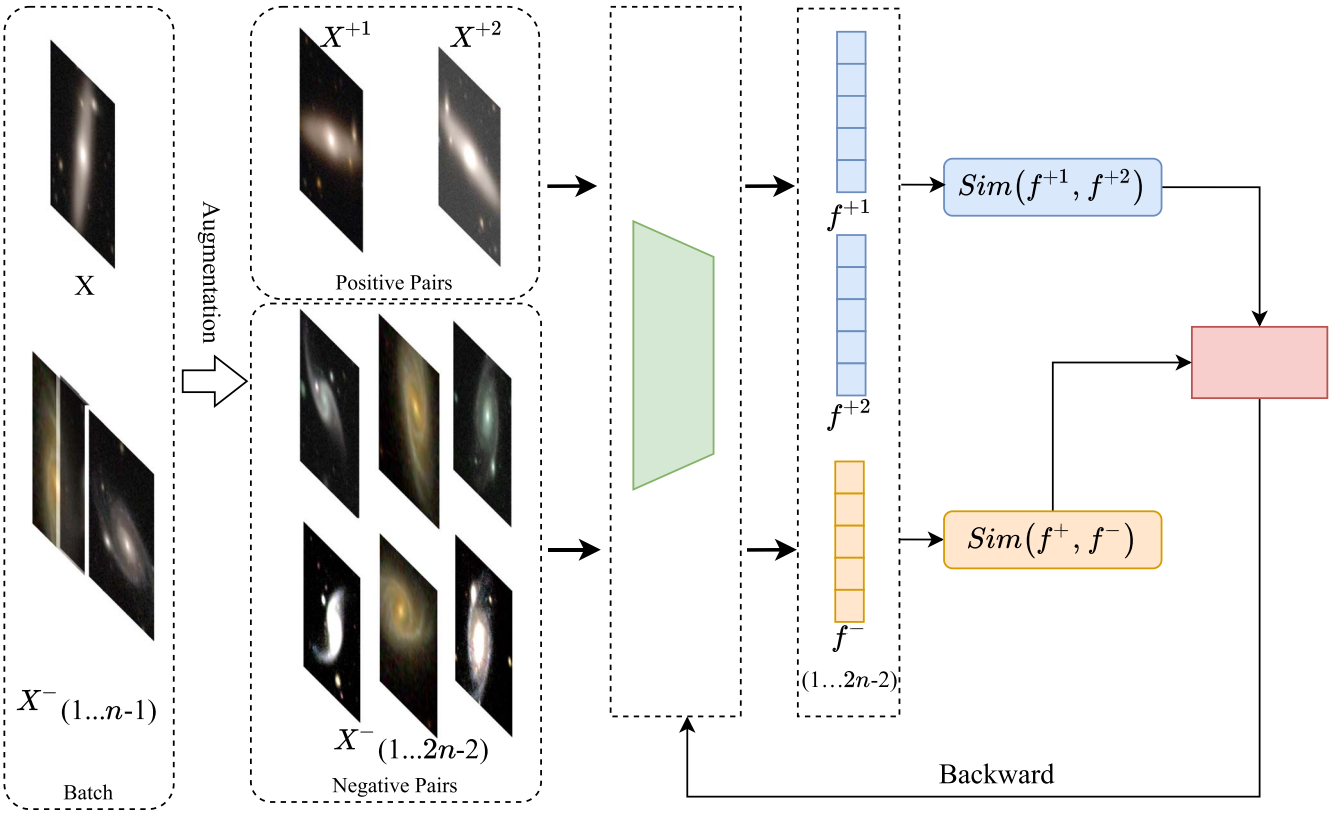


Figure 3. The overall architecture of the proposed network, which learns representation with multi-hierarchy features through the Encoder and backward feedback, with the loss calculated from the similarities between the positive pair and negative pairs.

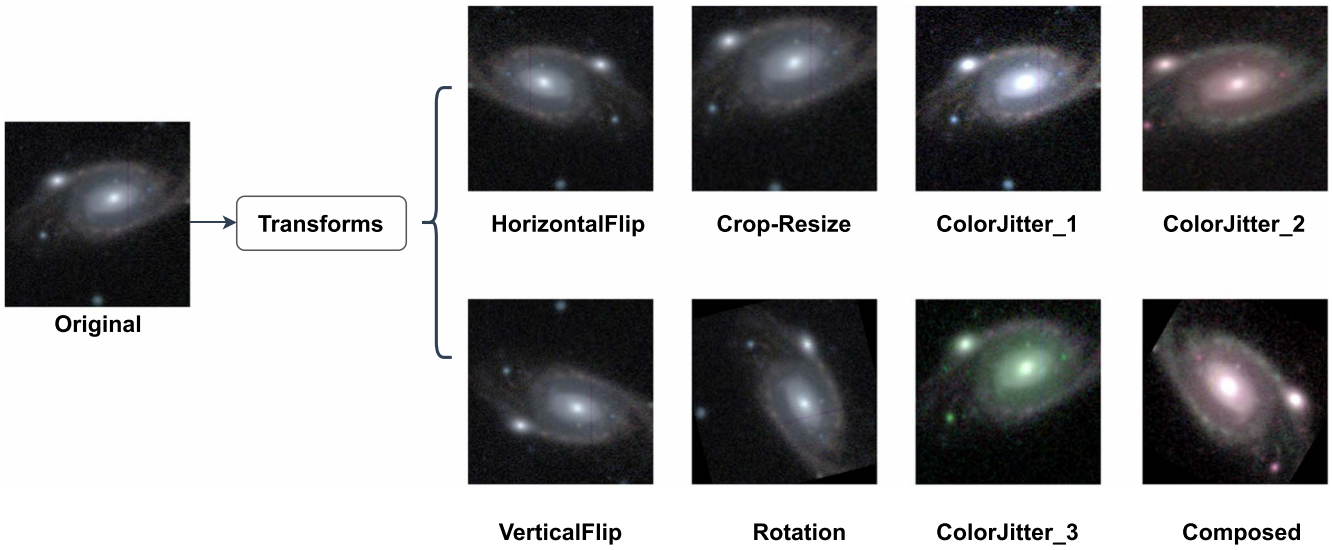


Figure 4. The original image and the eight sample results after applying the data augmentation used in this work.

layer, as maximum pooling may filter out the slightly darker structure information of galaxies with bright core structures, and increase the interference of galaxies with higher luminosity. The rectified linear units (RELU) (Agarap 2018) activation is used after each convolution layer, except for the last one.

The output size of the ResNet-18 based sub-network, after an adapted-average pooling, is 1×256. The output size of ViT is of the same size, 1×256. Before being fed into a Multilayer Perceptron (MLP) block, the fusing of features into a size of 1×512 is performed by concatenating the features from these

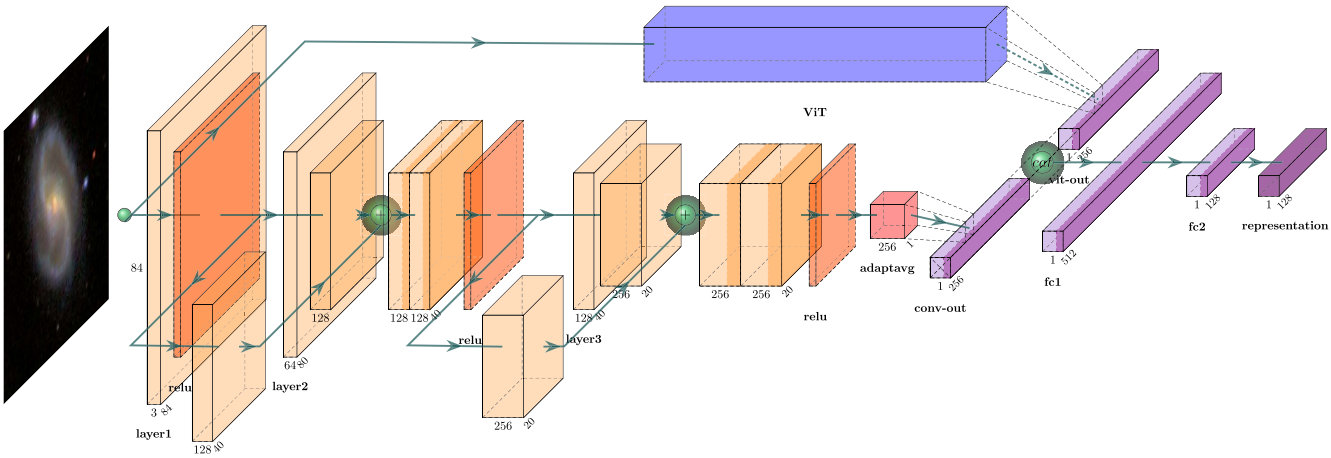


Figure 5. An illustration of the components in the Encoder, which uses ResNet-18 as the backbone network fusing with ViT to learn multi-hierarchy features.

two sub-networks. The MLP block with two fully-connected layers is used to map the features extracted by the backbone to the latent space with a size of 128×1. Instead of a linear layer, an MLP block will retain detailed information related to data augmentation, while the role of a linear layer is to remove this information (Chen et al. 2020). We include the layer details of the Encoder in Table 2.

3.4. Loss Function

InfoNCE (van den Oord et al. 2018) loss was used as a contrastive loss, which aims to pull positive examples closer and push negative examples farther apart. We take Cosine similarity as a distance metric between the representation vectors. The similarity function is defined as Equation (1),

$$\mathrm{Sim}(f_i, f_j) = \frac{f_i^T f_j}{||f_i||_2 ||f_j||_2}, \tag{1}$$

where $||f_i||_2$ and $||f_j||_2$ are the L2 regularization of the representation vector f_i and f_j. (Wang & Isola 2020) argued that L2 regularization of the representation vector can improve the training model. For the i-th instance, the loss function is defined as Equation (2),

$$\mathcal{L}_i = -\log \frac{\exp\left(\mathrm{Sim}(f_i^{+1}, f_i^{+2})/\tau\right)}{\sum_{j=0}^{K} \exp\left(\mathrm{Sim}(f_i, f_j)/\tau\right)}, \tag{2}$$

where f_i^{+1}, f_i^{+2} represent the corresponding representation vectors of the positive pair. Within a batch of n images, the number of negative samples K is $2n-2$. τ, called temperature, is a hyper-parameter that plays a role in controlling the strength of penalties on negative samples (Wang & Liu 2021). The final loss is an arithmetic mean of the loss for each sample in the batch. It can be defined as Equation (3).

$$\mathcal{L} = \frac{1}{N} \sum_{k=1}^{N} \mathcal{L}_k. \tag{3}$$

Table 2
The Layer Details of the Decoder, which Consists of Three Convolution Blocks, a Visual Transformers Block, and Two Fully Connected Layers

Block	Layer Type	Input Shape	Output Shape	Parameters
layer1	Covn2D(7*7)	(3, 84, 84)	(64, 80, 80)	9408
layer2	Covn2D(3*3)	(64, 80, 80)	(128,40,40)	73,728
	Covn2D(3*3)	(128,40,40)	(128,40,40)	147,456
	Covn2D(1*1)	(64, 80, 80)	(128,40,40)	8192
	Covn2D(3*3)	(128,40,40)	(128,40,40)	147,456
	Covn2D(3*3)	(128,40,40)	(128,40,40)	147,456
layer3	Covn2D(3*3)	(128,40,40)	(256, 20, 20)	294,912
	Covn2D(3*3)	(256, 20, 20)	(256, 20, 20)	589,824
	Covn2D(1*1)	(128,40,40)	256, 20, 20)	32,768
	Covn2D(3*3)	(256, 20, 20)	(256, 20, 20)	589,824
	Covn2D(3*3)	(256, 20, 20)	(256, 20, 20)	589,824
	Avg pooling	(256, 20, 20)	(256, 1, 1)	0
	Flatten()	(256, 1, 1)	(256)	0
VIT	Rearrange	(3, 84, 84)	(7*7432)	0
	Linear(embed)	(49+1, 432)	(50,128)	16,384
	6*Attention	(50,128)	(50,128)	1,179,648
	Linear(mlphead)	(128)	(256)	16,384
Projector	Linear	(512)	(512)	262,144
	Linear	(512)	(128)	65,536
	Total:			4,187,328

Note. The input shape, output shape and the number of parameters in each layer are also shown.

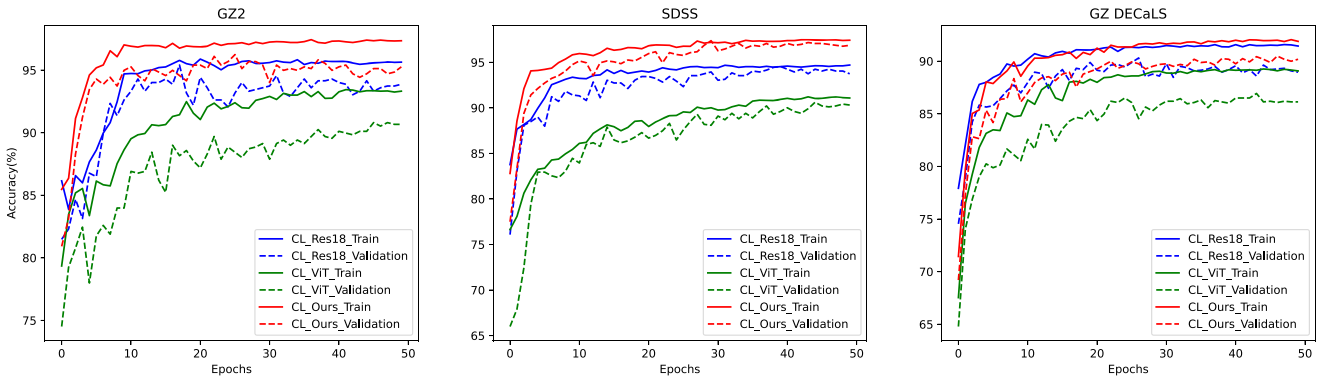


Figure 6. Performance comparison of the accuracy of training and validation along with the epochs. CL_Res18, CL_ViT and CL_Ours represents the CL methods using ResNet-18, ViT and our proposed model as encoder.

The objective of model learning is to minimize the loss function. It can be seen that the molecular part of Equation (2) encourages the higher similarity between the positive samples. In the denominator part, the similarity between positive and negative samples is encouraged to be as low as possible. In this way, during the optimization process, the model can be trained by the loss function guidance to achieve the desired goal.

4. Experiments

4.1. Training

We use Stochastic Gradient Descent (SGD) as our optimizer. The size of a batch is set to 256, a momentum is 0.9, and the maximum number of epochs is 50. The temperature τ is 0.1, and weight decay is 0.0001. We warm-up the network in the first 20 epochs by only using the InfoNCE loss. Following the suggestions of the Learning Rate Finder (Smith 2017), the initial learning rate is set to 0.01, and reduced with a cosine curve during the 50 epochs. We implemented our model on Pytorch. We train the model on the three data sets of GZ2, GZ DECaLS and SDSS-DR17 respectively. Each of data set is split into three different sets, a training, validation, and test set, which are randomly drawn from the total sample of galaxies. The training set consists of 80%, the validation set is 10%, and the test set is 10% of the data. The model takes about 1 – 2.5 hr for training 50 epochs on an NVIDIA TITAN V GPU. All of the experiments were performed on a CentOS 7.7 server configured with dual Intel Xeon CPU E5-2660 v4, 3.4 GHz maximum frequency, and 512 GB of RAM.

4.2. Evaluation Metric

To evaluate the classification performance of the proposed model, we adopt the k-nearest neighbor's algorithm(*k*NN) provided by faiss (Johnson et al. 2019) for efficient classification and agglomerative hierarchical clustering by scikit-learn (Pedregosa et al. 2011) for clustering. *k*NN performs a voting mechanism to determine the class of a data point. For an unseen data point, *k*NN computes its the nearest k neighbors and performs a voting mechanism to determine the class. Among the k neighbors, one of the classes accounts for the highest proportion; this class is then regarded as the predicted label of the data point. Accuracy is defined as the percentage of the number of correct predicted classifications compared to the true labels from the total samples.

NMI (Danon et al. 2005) and ARI (Santos & Embrechts 2009) are introduced to evaluate the performance of the clustering results. To some extent, NMI and ARI reflect the level of information correlation between the clustering results and the real labels. We use the implementations of NMI and ARI via the functions normalized_mutual_info_score and adjusted_rand_score in scikit-learn (Pedregosa et al. 2011). These two functions measure the similarity of the two assignments (predicted labels and real labels). The value range of NMI is [0, 1], and the value range of ARI is [−1, 1]. The closer the value is to 1, the closer the clustering results are to the real labels.

4.3. Performance

We performed the comparative performance tests of using ResNet-18, ViT and our proposed model as encoder. The tests were conducted over 50 epochs on the three data sets. The results are shown in Figure 6. Compared with only using ResNet-18 and ViT, our improved model achieves about a 0.44% – 2.73% increase in training accuracy and about a 1.22% – 3.13% increase in validation accuracy.

To explore a comparison of performance with supervised learning (SL), we performed the tests on the different sizes of training samples extracted from the three data sets. For the training subsets of GZ2 and SDSS, the samples are uniformly distributed in three categories (Elliptical, Edge-on and Spiral). For GZ DECaLS, the subsets are uniformly sampled from four categories(Round, Elliptical, Edge-on, and Spiral). To facilitate a fair comparison, the SL approach is trained from scratch with

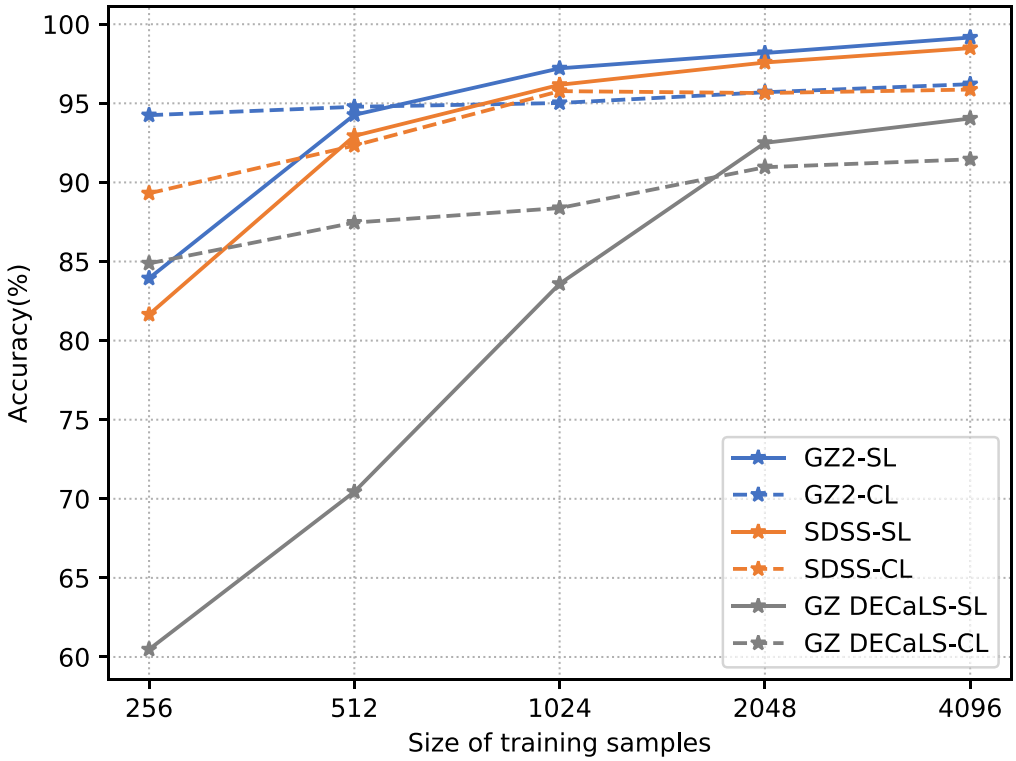


Figure 7. Performance comparison between our CL-based method and the SL approach. The panels demonstrate the testing accuracy along with the different size of training samples from GZ2, SDSS and GZ DECaLS respectively.

Table 3
The Classification Accuracy when Using *k*NN on the Representation of the Test Sets

Dataset	*k* (%)				
	1	5	10	15	20
GZ2	94.55	95.67	94.97	94.69	94.55
SDSS-DR17	93.24	95.66	96.37	96.51	95.96
GZ DECaLS	85.62	88.32	89.18	89.92	89.92

the same architecture as our encoder. As shown in Figure 7, when the size of the training subset is less than 512, our CL-based method achieves better performance than SL. For GZ DECaLS, especially, trained on 256 samples, our CL-based method can achieve an accuracy of about 85%, while the accuracy of SL is only 60%. Moreover, when the size of the training samples increases to nearly 2000, our CL-based method is still ahead of SL. For the subsets from GZ DECaLS that include four classifications, that means SL requires more samples to learn sufficient features. When sufficient samples are fed in, the performance of our CL-based method is within a few percentage points short of SL.

We also perform *k*NN under $k = 1$, 5, 10, 15 and 20 directly on the learned representations from our trained model. The accuracies on testing data sets are described in Table 3. Note that the testing accuracy is the median accuracy calculated from five different runs. In order to analyze the specific classification of the test set in the three data sets in detail, we illustrate the confusion matrixes upon the three test sets in Figure 8. Taking DECaLS as a example, there are 12 spiral galaxies that are mistakenly predicted as edge galaxies. We pick out these 12 images as shown in Figure 9. It can be seen that the model has poor discrimination ability for samples with indistinct spiral arms due to the high central brightness, and one galaxy will cause great interference to the identification of the others in a merged galaxy.

In order to measure the clustering performance, we compute ARI and NMI on validation sets after each epoch. The two value is almost close to 1 as shown in Figure 10. This proves that the predicted labels well match the real labels.

4.4. Cross Validation

In order to further validate the generalization ability of our model, we perform cross validation between the three data sets. That is, applying the model trained upon one data set to the other two test data sets. The test accuracy values obtained only decline slightly, as shown in Figure 11.

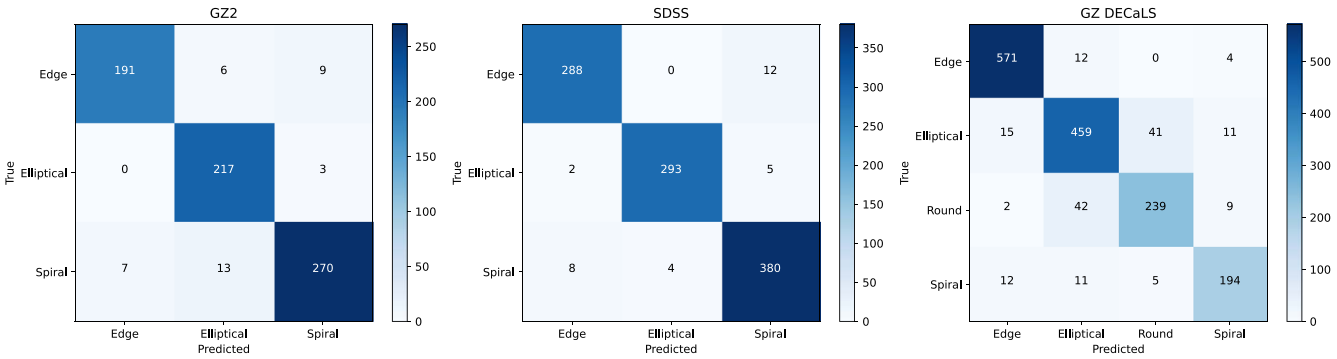


Figure 8. The confusion matrixes of the testing results on GZ2, DECaLS and SDSS-DR17, where the *x*-axis corresponds to the predicted label from the network and the *y*-axis to the visually assigned label.

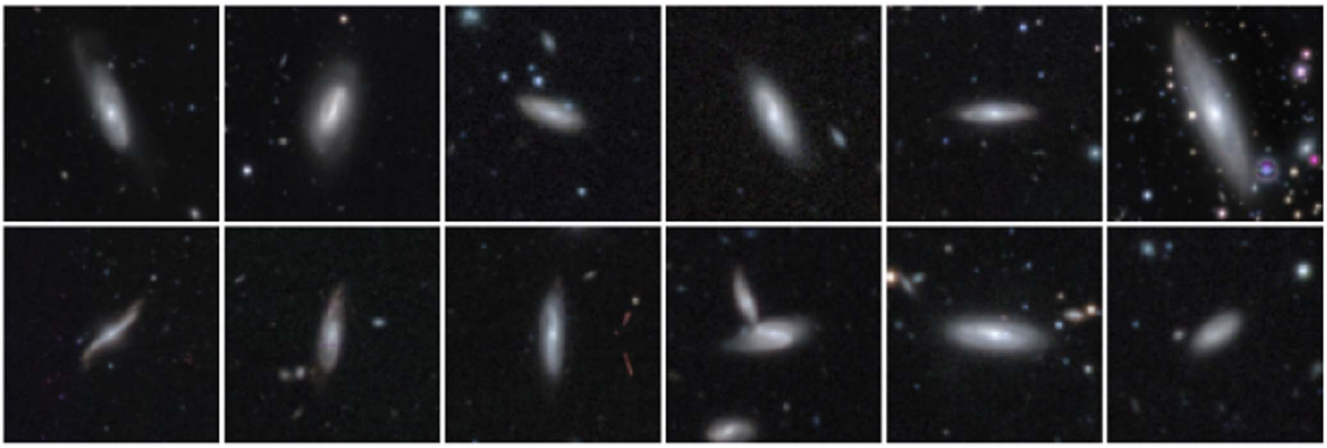


Figure 9. The samples mistakenly predicted as Edge-on galaxies from the test set of DECaLS, which were originally voted as spiral galaxies.

Note that the accuracy of M-GZ2 and M-SDSS on DECaLS (4) achieve only 79.4% and 80%. This is because the test set of DECaLS(4) contains round classification, while the two models M-GZ2 and M-SDSS do not contain round classification during training. When the round classification is excluded from DECaLS(3), the accuracy increased to 91.2% and 89.2% respectively. The decrease in number of classification types leads to a reduction of the error rate; even with its own M-DECaLS, the accuracy also increased from 89.9% to 91.3%.

4.5. Discussion of the Result

4.5.1. Hierarchical Clustering

The visual differences between images of galaxies are not obvious, which leads to the ambiguity of the boundary between the classification results. Thus, the accuracy based on the labels from volunteers votes does not fully reflect the performance of the model. In order to get rid of the constraints of labels, we use the hierarchical clustering algorithm (Murtagh & Contreras 2012) to cluster the obtained sample feature representations. Then, we evaluate the clustering results to verify whether the model can truly cluster the most similar samples.

The process of hierarchical clustering is to initially treat each sample as a cluster (i.e., leaf node). At each iteration, the two closest distance clusters polymerize to form a new cluster, until one cluster is left. This idea coincides with the pairwise calculation of the cosine similarity between positive and negative samples in contrastive learning.

Note that in a hierarchical clustering tree, the closer to the root node, the greater the difference between the clusters. Although the number of clusters is decreasing, it seems that this is for clustering, and it does not reflect the differences and similarities between each cluster. On the other hand, the closer to the leaf node, the more scattered the clusters are, which cannot reflect the general characteristics of a certain classification. Therefore, we set an upper limit threshold of the distance between clusters, and select appropriate clustering results in the inter-layer of the tree, to obtain a more detailed and representative classification. Taking the test set of GZ2 as an example, we set this threshold to 3.3, and obtained 20 clusters (as shown in Figure 12).

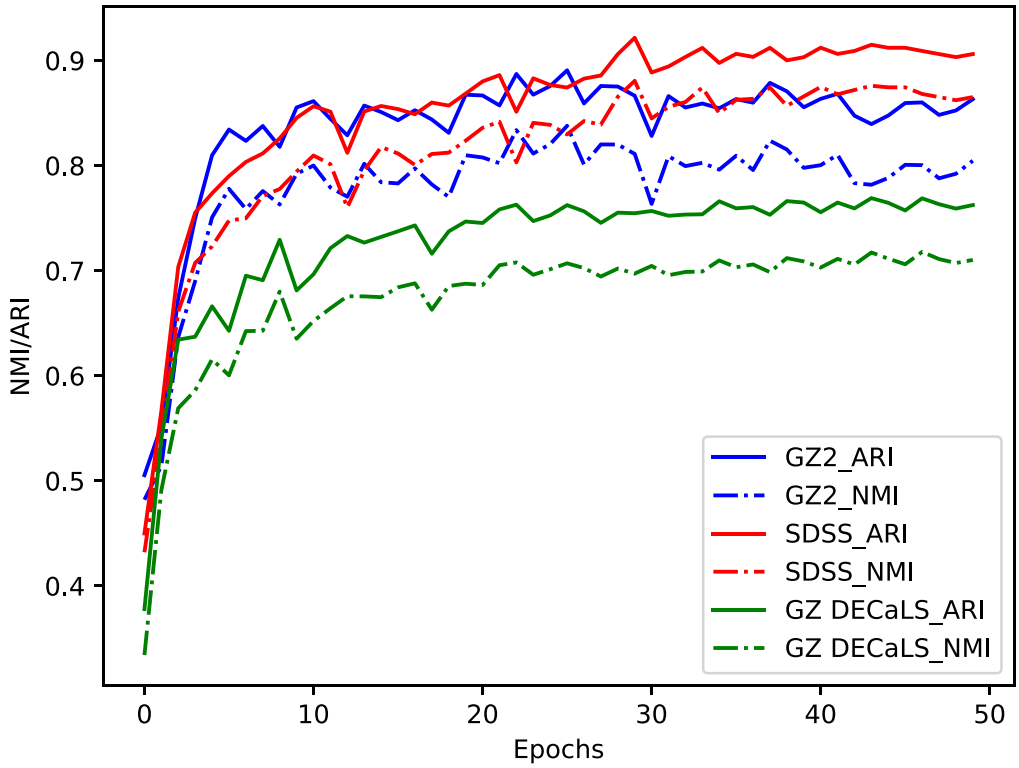


Figure 10. The trend diagram of ARI and NMI on validation sets as a function of the training epoch. The value gradually becomes stable as the training progresses, indicating that the clustering results also tend to be stable.

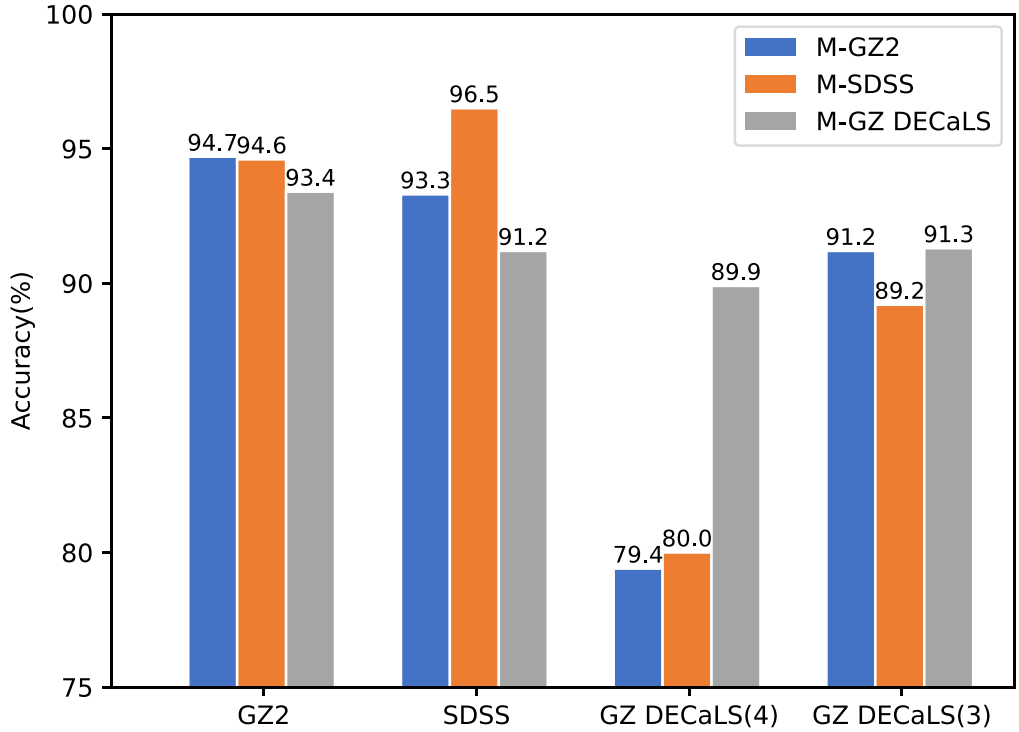


Figure 11. The accuracy of cross validation between GZ2, DECaLS and SDSS-DR17. The *x*-axis represents the test data set used, DECaLS(4) with 4 categories (Round, Elliptical, Edge-on, and Spiral), and DECaLS(3) with 3 categories (Round, Edge-on, and Spiral). M-GZ2, M-SDSS and M-GZ DECaLS represents the model trained on the data sets of GZ2, SDSS, and GZ DECaLS respectively.

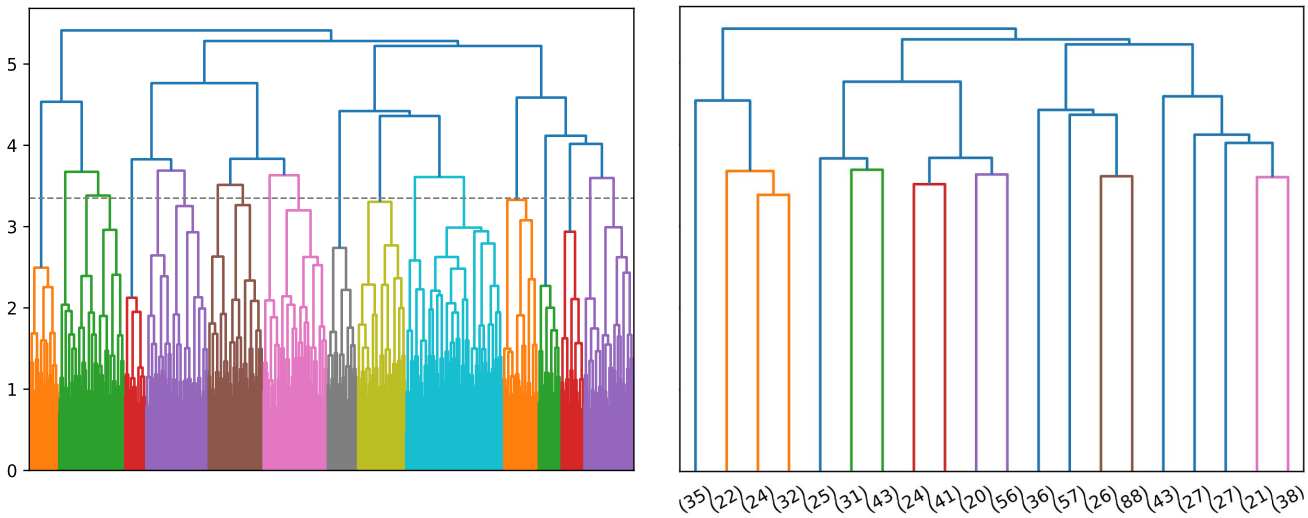


Figure 12. The results of hierarchical clustering on GZ2. The left panel represents the actual results of hierarchical clustering on all the test data from GZ2. We set the threshold of the distance between clusters to 3.3, as shown by the gray dotted line in the left panel, so as to obtain the schematic diagram of the right panel. The *X*-axis of the right panel indicates the number of samples in each cluster.

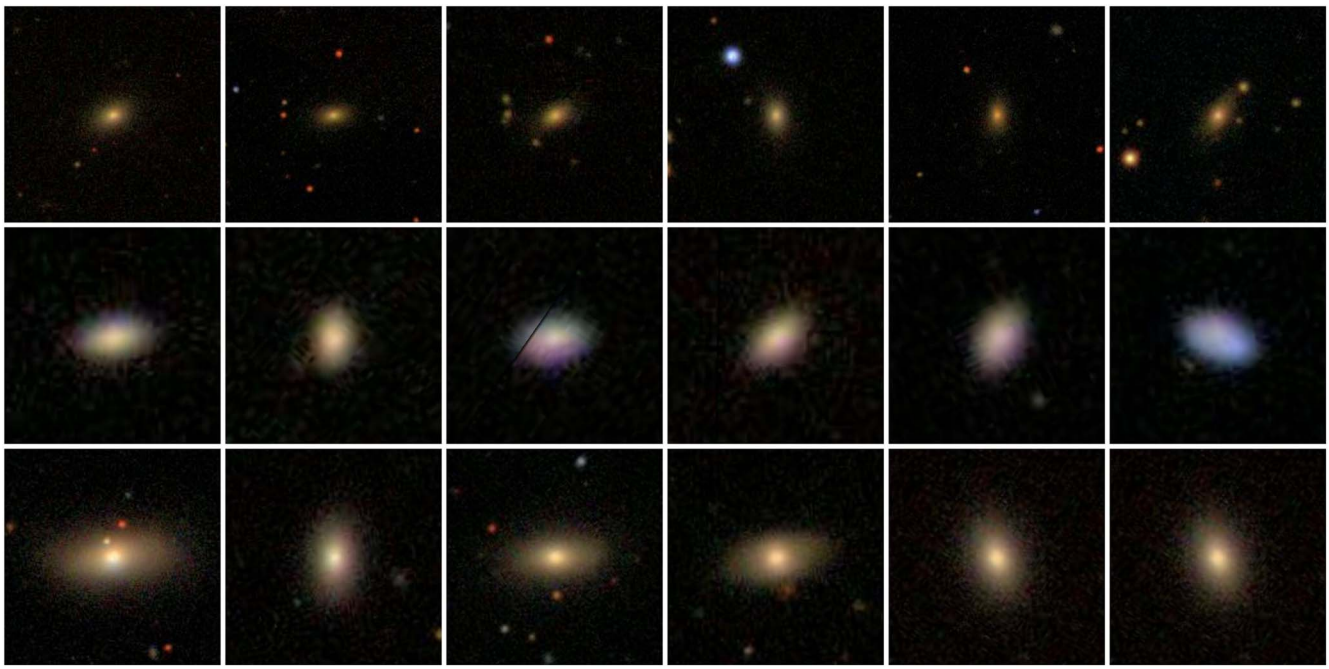


Figure 13. The selected 3 clusters after hierarchical clustering on the test data set of GZ2. All samples of the 3 clusters are all elliptical galaxies.

We have selected 3 clusters that all are elliptical galaxies, shown as Figure 13, where the galaxies in the first row are smaller in shape, with a nucleus but less luminosity. The galaxies in the second row are slightly larger in shape, but with blurred outlines and the nuclei are less clear. The galaxies from the third row are the most luminous and with the most clear nuclei. That is, the elliptical galaxies can be divided into 3 clusters according to the morphological detailed features. It proves that our model has learned subtle morphological representations.

4.5.2. Comparison with Morphological Measures

To explore whether clustered galaxies also have intrinsic similarities on the quantified measures of galactic morphology, we make statistical studies on these measures after clustering.

Table 4
The 6 Selected Quantified Measures of Galactic Morphology (Pawlik 2014)

Measure	Description
A	Asymmetry with the image rotating by 180°.
G	Gini index is a measure of the inequality of the distribution of light within the galaxy.
M_{20}	Measuring the second-order moment of the brightest 20% of the galaxy pixels.
SB_0	Sérsic model best-fit parameter: the central surface brightness.
R_{eff}	Sérsic model best-fit parameter: the effective radius.
n	Sérsic model best-fit parameter: Sérsic Index.

This test data set was retrieved from the table PawlikMorph in the SDSS skyserver at http://skyserver.sdss.org/dr17/SearchTools/sql (accessed 2022 May 7), which contains more than 4200 galaxies and their measures, such as A, G and M_{20}, etc. After filtered the null data, the used test set consists of 3228 galaxies. We have selected six common measures, as described in Table 4.

These test galactic images were fed into our trained model upon SDSS to obtain the feature representations on which hierarchical clustering was performed. With the condition of the distance threshold as 1, we get 17 clusters, from which we randomly selected seven clusters for analysis. Table 5 lists the mean of six measures of galactic morphology on the seven selected clusters, from which Figure 14 shows one representative image from each cluster.

Among seven selected clusters, there is the biggest distance between C1 and C7. We compare the histogram of six measures between these two clusters in Figure 15. The M_{20} reflects the strength of luminosity convergence. The lower value represents the higher luminosity and the stronger convergence. Although the C1 has a slightly higher central luminosity, the convergence is reduced due to the larger shape. The overall luminosity of C7 is slightly weaker, the smaller shape increases the convergence. Therefore, the M_{20} distribution of C7 is generally leftward, as shown in Figure 15(c). In the same way, the C1 cannot concentrate the luminosity in the central area, due to the existence of spiral arms. Hence the C1 is generally with small G, as shown in Figure 15(b). The R_{eff} reflects the size of shape; the R_{eff} of C1 is larger than that of C7, as shown in Figure 15(e). For the similarities of asymmetry, central surface brightness and Sérsic index between Spiral and Round, it also can been seen that there were no significant differences on A, SB_0 and n. From the above analysis, it is shown that our clustering results are well correlated with the structural measures of galaxies.

4.5.3. Visualization Results

Visualization of feature maps provide insight into the internal representation of a neural network. As discussed in Section 3.3, our model adopted a fusing network, with a Resnet-18 based convolutional network and a ViT block. For the convolutional network, we use Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al. 2017) to extract the weight coefficients and gradients of each convolution layer to generate a heat map, which is then superimposed with the original image to generate a visual image. For the ViT block, we use the method proposed by Chefer et al. (2021) to extract the attention map, then scale it to the original image size and overlay it with the original image to generate a visual image as shown in Figure 16.

The feature maps of the first convolutional layer can response low-level patterns. As expected, Figure 16(b) shows obvious spiral contour. While the second and third convolutional layer provides high-level filters to show more detailed patterns, as shown in Figures 16(c) and (d). As can be seen from Figure 16(e), the attention weights of the ViT block are mainly distributed on the spiral arms of the host spiral galaxy. This proves that our ViT block also pays attention to the backbone structure of the image.

5. Conclusions

In this paper, we present an unsupervised method for learning galaxy visual representations and galaxy morphological classification. In order to learn contour-information-dominated images of galaxies, we design an encoder that fuses high-level and low-level features maps via ViT and CNN. We train the proposed model and evaluate the performance with sample images from GZ2, DECaLS and SDSS-DR17. The testing results of accuracy and clustering quality prove that our method has better performance. Even if it cannot surpass supervised learning techniques for now, it yet could be a powerful method that can be used in galaxy morphological classification where new surveys can produce massive unlabeled galactic images.

We performed cross validation to evaluate the robustness of the model. The results indicated the accuracy of cross validation with only a slight decrease. It demonstrates the transfer ability of our model from one survey to another. Analysis of the results through hierarchical clustering proves that our model can perform more detailed classification, which may also be helpful for correcting mislabeled images and discovering galaxies with new morphologies. We also study the distribution of typical structural measures of the clustering results with our method, and demonstrate the consistent correlation between the morphological classification results and the structural measures. The visualization of feature maps shows that the Encoder could capture the features of multiple-levels to provide rich semantic representation.

While the data sets used in this paper are mainly based on limited samples, the forthcoming surveys such as LSST and CSST will produce billions of galaxies which may include peculiar galaxies that cannot be associated with the known morphological classifications of the Hubble sequence. For future work, we would like to train this network on even larger data sets, to investigate the

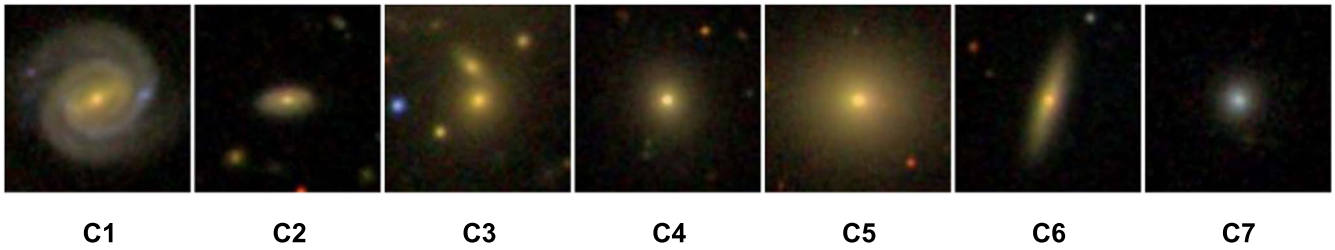


Figure 14. The representative sample from each cluster of 7 selected clusters.

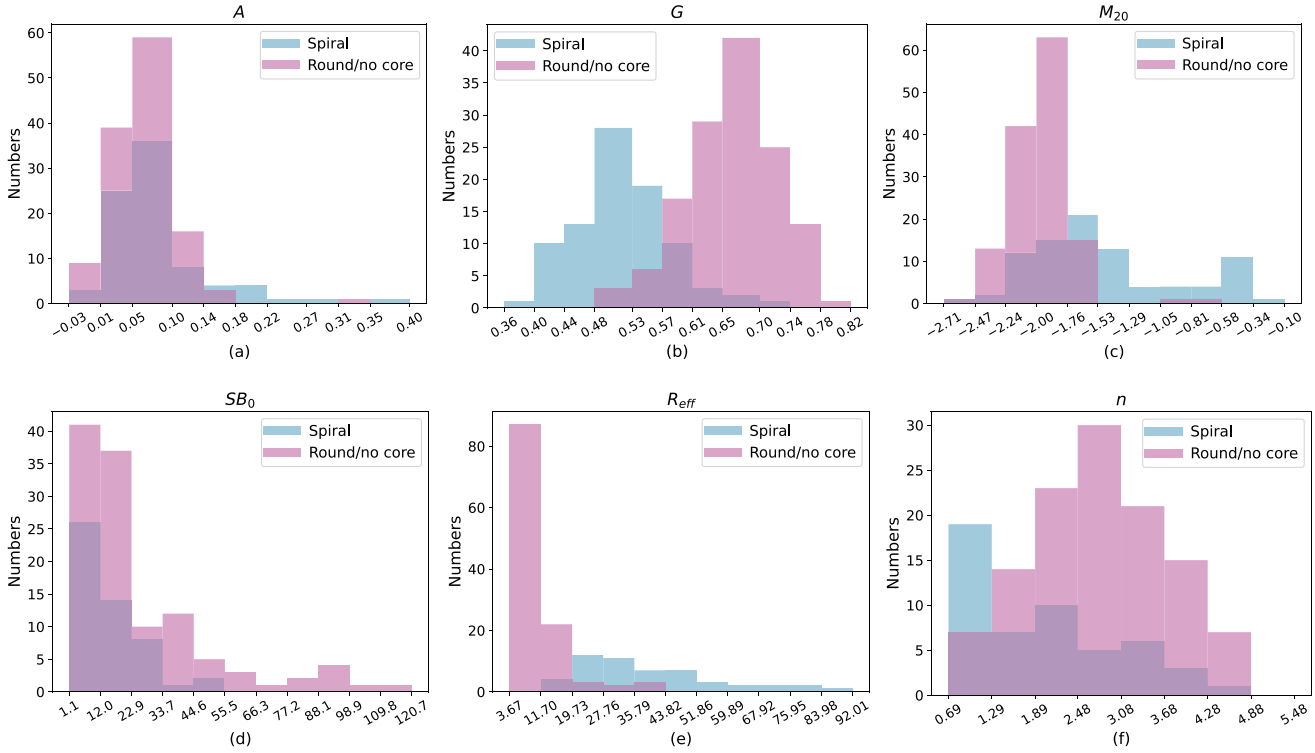


Figure 15. The comparison histogram of the 6 measures between C1 and C7, where the blue and pink histogram indicate C1 and C7 respectively.

Table 5
The Mean of 6 Measures of Galactic Morphology on 7 Selected Clusters

Cluster ID	A	G	M_{20}	SB_0	R_{eff}	n	Size	Shape
C1	0.0782	0.5135	−1.5042	14.8824	38.7451	1.9755	89	Spiral
C2	0.0671	0.6514	−1.8845	29.5304	7.5086	2.1487	135	Short-edge
C3	0.1335	0.6367	−1.5368	6.6808	27.0769	3.6885	148	Merge
C4	0.0637	0.6606	−2.0805	15.6735	17.4898	3.2908	240	Bright nucleus
C5	0.0830	0.6620	−1.7738	16.1455	28.0536	3.4286	71	Bright nucleus/halo
C6	0.0516	0.6377	−2.0500	18.8785	12.7593	2.5856	131	Long-edge
C7	0.0623	0.6622	−1.9596	24.7966	10.3644	2.7669	137	Round/no nucleus

efficacy of detecting peculiar galaxies. We also plan to take FITS as the input of training, since multiple band observational data intrinsically come with richer semantic information than RGB images. Although the current accuracy is above 90%, it cannot compete with supervised learning. We expect the improvement of performance through using the state-of-the-art networks. We also plan to carry out a study on the subtler classification of the spiral type, according to the tightness of spiral arms.

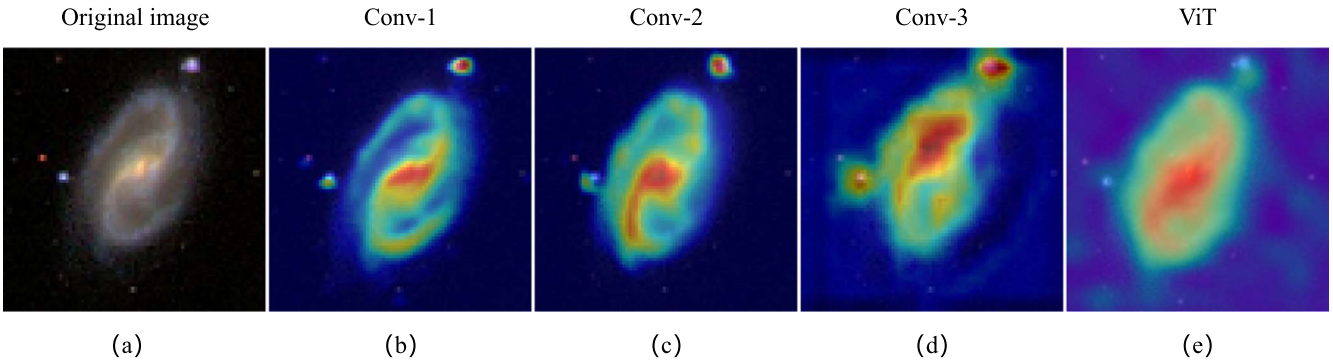


Figure 16. The original image and its generated feature maps through our model, where the sub-plots of Conv1-3 are generated from the 3 convolutional blocks of a Resnet-18 based network, and the sub-plot of ViT is from the ViT block.

We thank the referee for useful comments that helped to improve the paper. This work was supported by the National Key Research and Development Program of China (2020SKA0110300), Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (11961141001), and the Joint Research Fund in Astronomy (U1831204, U1931141) under a cooperative agreement between the National Natural Science Foundation of China (NSFC) and the Chinese Academy of Sciences (CAS), National Natural Science Foundation of China (No.11903009). The authors acknowledge financial support from the Yunnan Ten Thousand Talents Plan Young & Elite Talents Project.

ORCID iDs

Shoulin Wei https://orcid.org/0000-0002-3547-4025

References

Agarap, A. F. 2018, arXiv:1803.08375
Baillard, A., Bertin, E., de Lapparent, V., et al. 2011, A&A, 532, A74
Bhambra, P., Joachimi, B., & Lahav, O. 2022, MNRAS, 511, 5032
Chefer, H., Gur, S., & Wolf, L. 2021, in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) (Nashville, TN: IEEE), 782
Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, in Proc. Machine Learning Research, 119, Proc. 37th Int. Conf. Machine Learning (PMLR) (Vienna: PMLR), 1597
Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. 2020b, MNRAS, 493, 4209
Cheng, T.-Y., Li, N., Conselice, C. J., et al. 2020a, MNRAS, 494, 3750
Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. 2005, JSMTE, 2005, P09008
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2018, MNRAS, 484, 93
Domínguez Sánchez, H., Margalef, B., Bernardi, M., & Huertas-Company, M. 2022, MNRAS, 509, 4024
Ghosh, A., Urry, C. M., Wang, Z., et al. 2020, ApJ, 895, 112
Gupta, R., Srijith, P. K., & Desai, S. 2022, A&C, 38, 100543
Hayat, M. A., Stein, G., Harrington, P., Lukic, Z., & Mustafa, M. 2021, ApJL, 911, L33
He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. 2020, in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (Seattle, WA: IEEE), 9729
He, K., Zhang, X., Ren, S., & Sun, J. 2016, in Proc. IEEE Conf. Computer Vision and Pattern Recognition (Las Vegas, NV: IEEE), 770
Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., et al. 2019, arXiv:1808.06670
Hubble, E. P. 1926, ApJ, 64, 321
Johnson, J., Douze, M., & Jégou, H. 2019, IEEE Transactions on Big Data, 7, 535
Katebi, R., Zhou, Y., Chornock, R., & Bunescu, R. 2019, MNRAS, 486, 1539
Khalifa, N. E., Taha, M. H., Hassanien, A. E., & Selim, I. 2018, in Int. Conf. Computing Sciences and Engineering (ICCSE) (Wuhan: IEEE), 1
Khan, A., Huerta, E. A., Wang, S., et al. 2019, PhLB, 795, 248
Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166
Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179
Martin, G., Kaviraj, S., Hocking, A., Read, S. C., & Geach, J. E. 2020, MNRAS, 491, 1408
Murtagh, F., & Contreras, P. 2012, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2, 86
Pawlik, M. 2014, A&G, 55, 13
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
Racca, G. D., Laureijs, R., Stagnaro, L., et al. 2016, Proc. SPIE, 9904, 99040O
Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. 2021, arXiv:2108.08810
Robertson, B. E., Banerji, M., Brough, S., et al. 2019, NatRP, 1, 450
Santos, J. M., & Embrechts, M. 2009, in Artificial Neural Networks – ICANN 2009, ed. C. Alippi et al. (Berlin: Springer), 175
Sarmiento, R. 2020, XIV.0 Scientific Meeting (virtual) of the Spanish Astronomical Society, 86
Schutter, A., & Shamir, L. 2015, A&C, 12, 60
Selvaraju, R. R., Cogswell, M., Das, A., et al. 2017, in Proc. IEEE Int. Conf. Computer Vision (ICCV), 618
Simmons, B. D., Lintott, C., Willett, K. W., et al. 2016, MNRAS, 464, 4420
Smith, L. N. 2017, in IEEE Winter Conf. Applications of Computer Vision (WACV) (Santa Rosa, CA: IEEE), 464
Tuccillo, D., Huertas-Company, M., Decenciere, E., & Velasco-Forero, S. 2016, Proc. Int. Astron. Union, 12, 191
van den Oord, A., Li, Y., & Vinyals, O. 2018, arXiv:1807.03748
Variawa, M. Z., Van Zyl, T. L., & Woolway, M. 2022, IEEE Access, 10, 19539
Vaswani, A., Shazeer, N., Parmar, N., et al. 2022, in NIPS'17, 30, Proc. 31st Int. Conf. Neural Information Processing Systems, 19539
Vaucouleurs, G. 1959, General Physical Properties of External Galaxies (Berlin: Springer), 311
Vavilova, I., Khramtsov, V., Dobrycheva, D., et al. 2022, arXiv:2203.06373
Walmsley, M., Lintott, C., Geron, T., et al. 2022, MNRAS, 509, 3966
Wang, F., & Liu, H. 2021, in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) (Nashville, TN: IEEE), 2495
Wang, T., & Isola, P. 2020, Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, Proc. 37th Int. Conf. Mach. Learn. (Vienna: JMLR), 2495
Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835
Wilman, D. J., & Erwin, P. 2012, ApJ, 746, 160
Zhu, X.-P., Dai, J.-M., Bian, C.-J., et al. 2019, Astrophysics and Space Science, 364, 1
Zhan, H. 2021, Chinese Science Bulletin, 66, 1290
Zhang, Z., Zou, Z., Li, N., & Chen, Y. 2022, RAA, 22, 055002