Identifying Galaxy Mergers With Self-supervised Deep Learning

Jasmine Guo
High Technology High School, Lincroft, NJ 07738

Abstract

A galaxy merger in the universe occurs when two or more galaxies collide with combined mass and material. Galaxy mergers can trigger the formation of new stars, and studying galaxy mergers is important to understand the evolution of the universe. With the huge amount of galaxy pictures captured by telescopes, identifying galaxy mergers becomes the first challenging task to facilitate future research.

In this research project, an approach based on self-supervised deep learning is proposed to identify galaxy mergers from large scientific datasets. Self-supervised deep learning is an evolving machine learning technique which does not require human-labeled data, and it has demonstrated remarkable performance in extracting feature representations from data. In order to discover the characteristics of galaxy mergers in this research project, a self-supervised deep learning method was used to train a model from the images of galaxy mergers, and the learned model can be further utilized to extract feature representations from any galaxy image. The galaxy images which are semantically similar to known galaxy mergers can be identified by finding similar galaxy images in the feature representation space.

The dataset used in this research project consists of 61578 galaxy images from Galaxy Zoo and 3.5 million galaxy images from Dark Energy Spectroscopic Instrument (DESI) Legacy Imaging Surveys. The comparison with the baseline method showed significant improvement. A large number of galaxy mergers can be identified with the proposed method, which can significantly save human effort to find galaxy mergers.

1 Introduction

With each passing day, telescopes around and above the Earth (there are 27 telescopes in orbit) capture more and more images of distant galaxies and the datasets begin to explode in size. In order to better understand how the different shapes of galaxies relate to the underlying physics that create them, it is crucial to classify and organize these images.

A galaxy merger in the universe occurs when two or more galaxies collide with combined mass and material. Galaxy mergers play an important role in the formation and evolution of galaxies. Galaxy mergers can trigger the formation of new stars, and

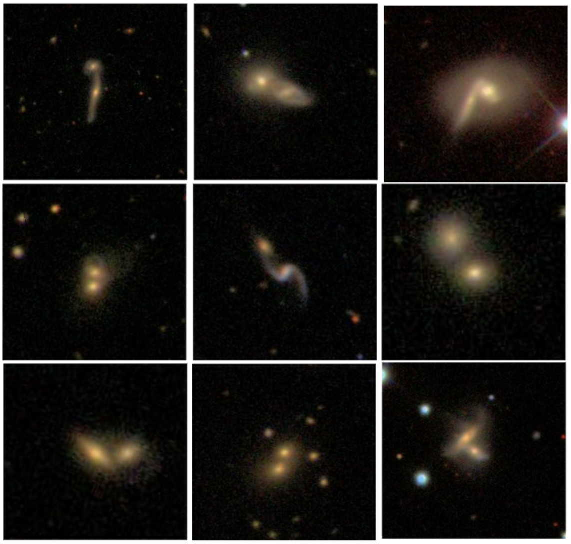


Figure 1: Examples of galaxy mergers

studying galaxy mergers is important to understand the evolution of the universe as a whole. With billions of galaxy images captured by telescopes, identifying galaxy mergers becomes the first challenging task to facilitate the future research on galaxy mergers.

2 Related Work

Some galaxies have already been classified through the help of hundreds of thousands of volunteers, who collectively classified the shapes of these images (including galaxy mergers) by eye in a successful citizen science crowdsourcing project.

In Galaxy Zoo [2, 6], volunteers are shown images of galaxies and asked to classify them based on their visual appearance using a set of simple questions. The responses are then aggregated to generate a classification for each galaxy. This crowdsourcing approach allows for the efficient classification of large numbers of galaxies and has been instrumental in advancing our understanding of galaxy morphology and evolution. Some examples of galaxy mergers are shown in Figure 1. However, this approach becomes less scalable as datasets grow to consist of billions of galaxies.

In recent years, machine learning has proven particularly powerful for various tasks, and they can handle large datasets quickly and accurately. Galaxies are complex systems with many variables such as size, shape, color, brightness, and spectral features. Machine learning algorithms are capable of detecting patterns and correlations in these variables, which can help classify galaxies based on their physical properties.

There are many different kinds of machine learning methods, each with their own strengths. Some types of machine learning methods are:

- Supervised learning which requires the human-labeled training data to make predictions. Deep learning is a supervised learning technique based on the artificial neural networks, and it can be used for various learning tasks.
- Unsupervised learning which cluster and organize the data into different categories without any human-labeled data.
- Semi-supervised learning which use the huge amount of unlabeled data to improve the performance of supervised learning because human-labeled data is very expensive.
- Dimensionality reduction which transform the data from high-dimensional space into a low-dimensional space while preserving the useful properties of the original data.

There has been some research work on applying the machine learning methods to analyze galaxy images. However, there is not much existing work in identifying galaxy mergers. In this research, an approach based on self-supervised deep learning is proposed to identify galaxy mergers from large scientific datasets. Specifically, the galaxy images are transformed from a high-dimensional space into a low-dimensional space and the galaxy mergers can be identified in the low-dimensional space more easily and accurately.

3 Self-supervised Learning

Self-supervised deep learning is based on CNN (convolutional neural network), which takes an image x as input and produces a lower dimensional representation $\mathbf{z}$ and it does not require the human-labeled data. CNN learns to make meaningful representations by associating augmented views of the same image as similar, and views of different images as dissimilar. The augmented views of the image are the new transformed versions of images from the original image to increase the diversity, and the normal operations are rotation, scaling, and addition of some random noises. This research project followed the self-supervised model in [3], and used the public code in [5, 4] for implementation, which is based on PyTorch Lightning.

The architecture and the training procedure of the self-supervised learning proposed in this research projects is shown in Figure 2. The same set of augmentation operations as in [5] are applied to each image in the training dataset, including rotation, size rescaling, and Gaussian noise. Each of this augmentation generates a new image. After these augmentation operations, a sequence of images $q_{ik}, 1 \leq k \leq m$ are generated for each image $p_i, 1 \leq i \leq n$, where n is the number of images in the training dataset and m is the number of augmentation operations. The augmented

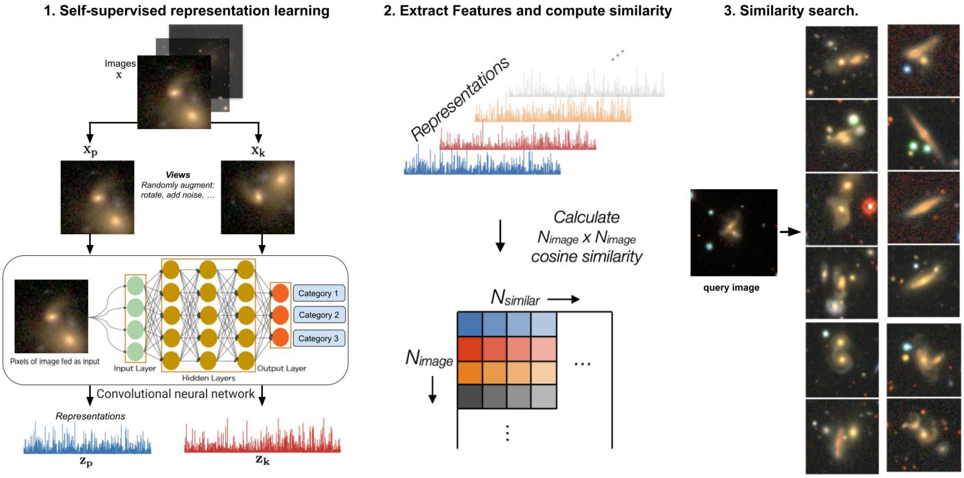


Figure 2: Self-supervised Learning

views of the same images are considered as similar, which means that two images in the image pair (p_i, q_{ik}) are similar. The augmented views of the different images are considered as dissimilar, which means that two images in the image pair $(p_j, q_{ik}), i \neq j$ are dissimilar.

The goal of the self-supervised learning is to learn meaningful representations in a low-dimensional space by minimizing the distance between two images in the image pair (p_i, q_{ik}) and maximizing the distance between two images in the image pair $(p_j, q_{ik}), i \neq j$. This is achieved by training a convolutional neural network. The training dataset used in this research project is a set of galaxy merger images which are labeled by human in Galaxy Zoo [2]. The labeled data is not necessary for self-supervised learning, but the learned CNN model will better capture the characteristics of galaxy mergers by only using the galaxy merger images as the training dataset.

The dimension of the original galaxy images from Galaxy Zoo is 424 × 424, so the original image is in a very high dimension space (424 × 424 × 3) because each pixel has 3 colors in RGB space. This high-dimensional space data is fed as input to the self-supervised learning algorithm to learn the underlying representation of galaxy mergers, which is represented as the feature representation in the low-dimensional space (1024 dimensions). The learned self-supervised learning model can be used for calculating the feature representation in the low-dimensional space for any image. The galaxy images which are semantically similar to known galaxy mergers can be identified by calculating the cosine similarity score of the feature vector.

$$similarity\left(\mathbf{z_i}, \mathbf{z_j}\right) = \frac{\mathbf{z_i} \cdot \mathbf{z_j}}{||\mathbf{z_i}||||\mathbf{z_j}||} \tag{1}$$

4 Experimental Result

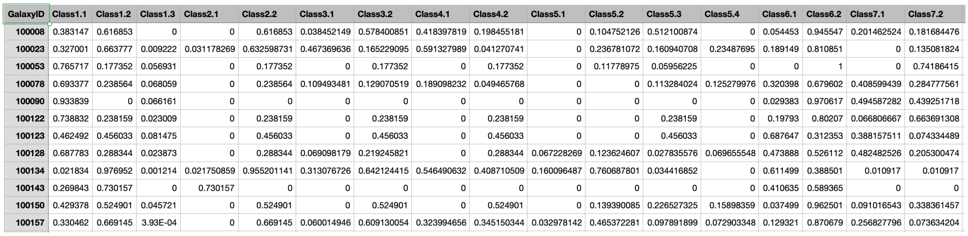
4.1 Galaxy Zoo Dataset

Task	Question	Responses	Next
01	*Is the galaxy simply smooth and rounded, with no sign of a disk?*	smooth	07
		features or disk	02
		star or artifact	**end**
02	*Could this be a disk viewed edge-on?*	yes	09
		no	03
03	*Is there a sign of a bar feature through the centre of the galaxy?*	yes	04
		no	04
04	*Is there any sign of a spiral arm pattern?*	yes	10
		no	05
05	*How prominent is the central bulge, compared with the rest of the galaxy?*	no bulge	06
		just noticeable	06
		obvious	06
		dominant	06
06	*Is there anything odd?*	yes	08
		no	**end**
07	*How rounded is it?*	completely round	06
		in between	06
		cigar-shaped	06
08	*Is the odd feature a ring, or is the galaxy disturbed or irregular?*	ring	**end**
		lens or arc	**end**
		disturbed	**end**
		irregular	**end**
		other	**end**
		merger	**end**
		dust lane	**end**
09	*Does the galaxy have a bulge at its centre? If so, what shape?*	rounded	06
		boxy	06
		no bulge	06
10	*How tightly wound do the spiral arms appear?*	tight	11
		medium	11
		loose	11
11	*How many spiral arms are there?*	1	05
		2	05
		3	05
		4	05
		more than four	05
		can't tell	05

Figure 3: The Galaxy Zoo decision tree, comprising 11 tasks and 37 responses

In Galaxy Zoo, volunteers are shown images of galaxies and asked to classify them based on their visual appearance using a set of simple questions [6], as show in Figure 3. The responses are then aggregated to generate a classification for each galaxy.

In Galaxy Zoo dataset used in this research project, there are 61,578 human-labeled galaxy images with 37 labels. For each of 37 labels, each image is associated with a floating point number between 0 and 1 inclusive, which represents a probability for each category; a high number (close to 1) indicates that many users identified this category for the galaxy with a high level of confidence and a low number (close to 0) indicate the galaxy image is likely not in the category. Among 61578 images, there are 1022 images associated with the label ”galaxy merger” with a probability large than 0.5. Some examples of label information are shown in Figure 4, which only show a subset of labels because of space limitation.



GalaxyID	Class1.1	Class1.2	Class1.3	Class2.1	Class2.2	Class3.1	Class3.2	Class4.1	Class4.2	Class5.1	Class5.2	Class5.3	Class5.4	Class6.1	Class6.2	Class7.1	Class7.2
100008	0.383147	0.616853	0	0	0.616853	0.038452149	0.578400851	0.418397819	0.198455181	0	0.104752126	0.512100874	0	0.054453	0.945547	0.201462524	0.181684476
100023	0.327001	0.663777	0.009222	0.031178269	0.632598731	0.467369636	0.165229095	0.591327989	0.041270741	0	0.236781072	0.160940708	0.23487695	0.189149	0.810851	0	0.135081824
100053	0.765717	0.177352	0.056931	0	0.177352	0	0.177352	0	0.177352	0	0.11778975	0.05956225	0	0	1	0	0.74186415
100078	0.693377	0.238564	0.068059	0	0.238564	0.109493481	0.129070519	0.189098232	0.049465768	0	0	0.113284024	0.125279976	0.320398	0.679602	0.408599439	0.284777561
100090	0.933839	0	0.066161	0	0	0	0	0	0	0	0	0	0	0.029383	0.970617	0.494587282	0.439251718
100122	0.738832	0.238159	0.023009	0	0.238159	0	0.238159	0	0.238159	0	0	0.238159	0	0.19793	0.80207	0.066806667	0.663691308
100123	0.462492	0.456033	0.081475	0	0.456033	0	0.456033	0	0.456033	0	0	0.456033	0	0.687647	0.312353	0.388157511	0.074334489
100128	0.687783	0.288344	0.023873	0	0.288344	0.069098179	0.219245821	0	0.288344	0.067228269	0.123624607	0.027835576	0.069655548	0.473888	0.526112	0.482482526	0.205300474
100134	0.021834	0.976952	0.001214	0.021750859	0.955201141	0.313076726	0.642124415	0.546490632	0.408710509	0.160096487	0.760687801	0.034416852	0	0.611499	0.388501	0.010917	0.010917
100143	0.269843	0.730157	0	0.730157	0	0	0	0	0	0	0	0	0	0.410635	0.589365	0	0
100150	0.429378	0.524901	0.045721	0	0.524901	0	0.524901	0	0.524901	0	0.139390085	0.226527325	0.15898359	0.037499	0.962501	0.091016543	0.338361457
100157	0.330462	0.669145	3.93E-04	0	0.669145	0.060014946	0.609130054	0.323994656	0.345150344	0.032978142	0.465372281	0.097891899	0.072903348	0.129321	0.870679	0.256827796	0.073634204

Figure 4: Examples of human labeled data in Galaxy Zoo

4.2 Experimental Result on Galaxy Zoo

The experiments are conducted on the Galaxy Zoo to measure the effectiveness of self-supervised learning. Two-thirds (681 images) of the galaxy merger images were used as the training dataset to train the self-supervised learning model. The learned model is used to calculate the feature representation in the lower dimensional space for the test dataset (all other 60897 galaxy images including the rest 341 galaxy merger images).

For each of galaxy images in the test dataset, the similarity score is calculated as the average of all the similarity scores from the galaxy mergers in the training dataset:

$$\underset{\mathbf{z_i}\in S}{Average}\left(similarity\left(\mathbf{z},\mathbf{z_i}\right)\right) \tag{2}$$

where S is the training dataset. The self-supervised learning is compared with the baseline method where the original high-dimensional feature representation in RGB space is directly used as the feature representation. The top k images with highest similarity scores are considered to calculate the precision and recall. Precision is calculated by dividing the true positives by k and recall is calculated by dividing the true positives by the number of galaxy mergers in the test dataset.

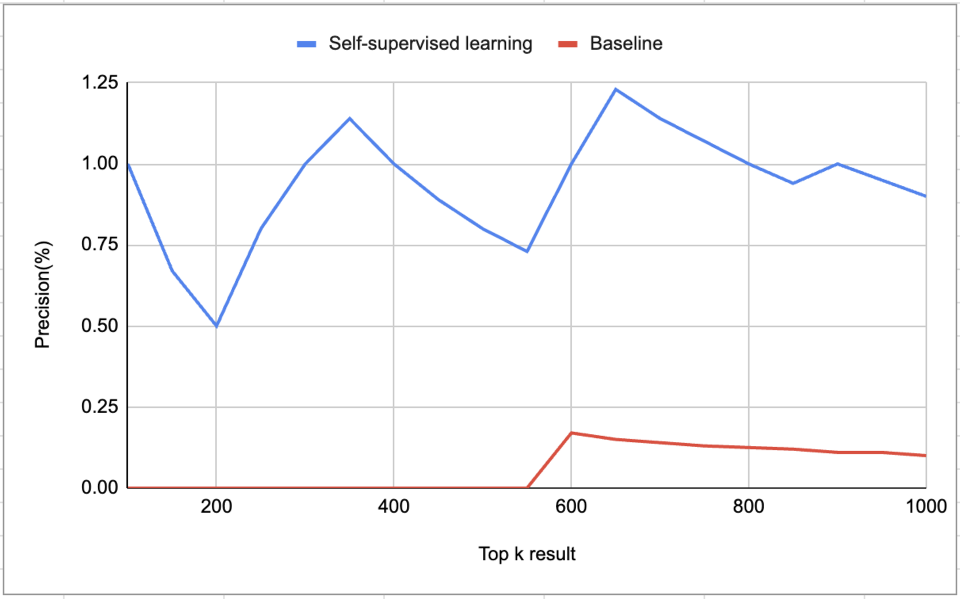


Figure 5: Precision on Galaxy Zoo dataset

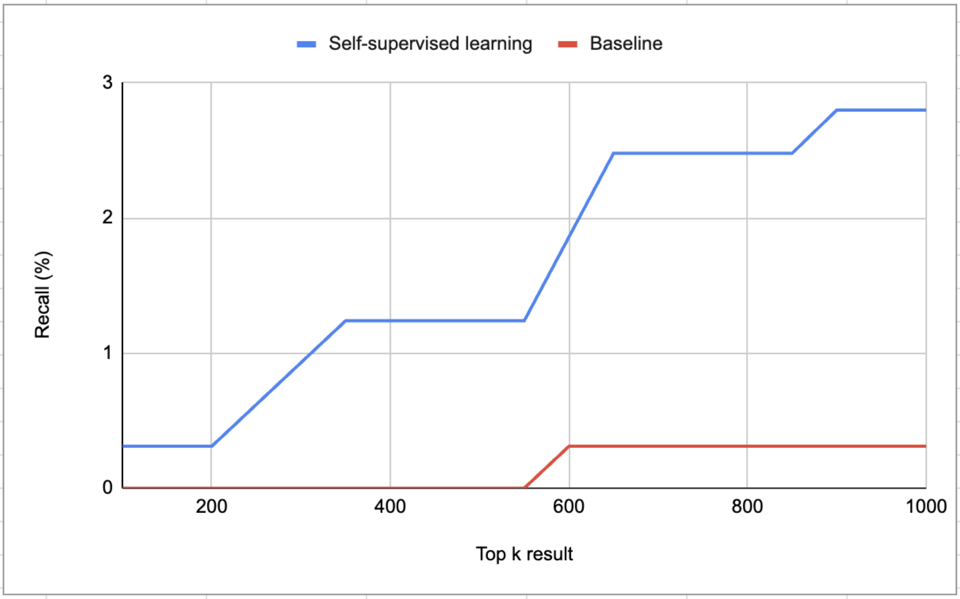


Figure 6: Recall on Galaxy Zoo dataset

A two tailed t-test for two samples of unequal variance was performed to test for significance in favor of the self-supervised learning for all data points in Figures 5 and 6. As shown in Figures 5 and 6 and Table 1, the self-supervised learning shows significant improvement over the baseline method.

Statistical Data Table for Precision and Recall		
P-value for precision	P-value for recall	alpha
0	1.10237495E-6	0.05

Table 1: Statistical Data Table for Precision and Recall

4.3 Experimental Results on DESI Legacy Imaging Surveys

Dark Energy Spectroscopic Instrument (DESI) [1] Legacy Imaging Surveys are a combination of three public projects, and it includes billions of galaxy images. A subset of 3.5 million galaxy images was used for experiments in this research project to evaluate the performance of self-supervised learning. All the galaxy images in DESI Legacy Imaging Surveys are not human-labeled, and to identify galaxy mergers from such a large dataset would require significant human effort. The set of 1022 galaxy merger images from Galaxy Zoo is used as the training dataset to train a CNN model for transforming the image from high-dimensional space to low-dimensional space, and then the learned model is used to extract the low-dimensional feature representations for all 3.5 million images in DESI. The Equation (2) is used to identify the galaxy mergers from DESI dataset.

In Figure 7, the first image shows the top 16 results and the second image shows the result with the ranking after top 10000, which shows that a large number of galaxy mergers can be identified with the self-supervised learning.

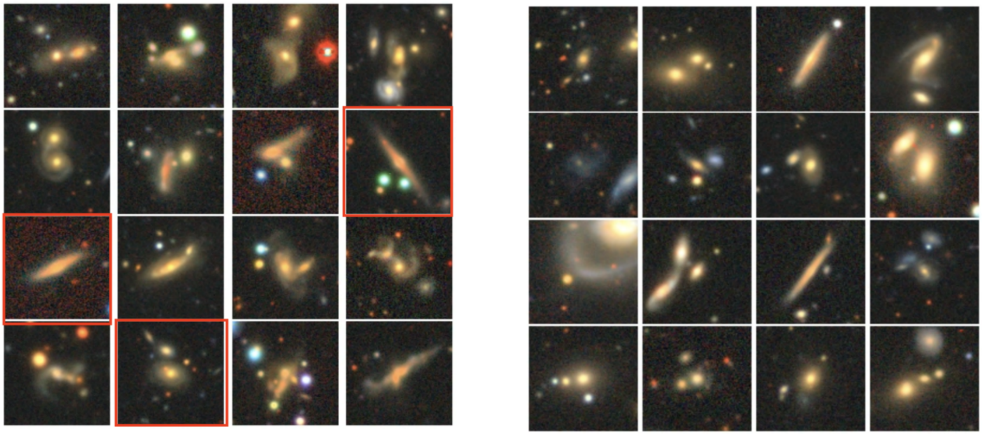


Figure 7: Results on DESI Legacy Imaging Surveys

4.4 Training process

The goal of self-supervised learning is to minimize a loss function by optimizing parameters in convolutional neural network. The graph in Figure 8 shows how the value of the loss function is decreased as the number of the iterations increased. During each iteration, the self-supervised learning goes through the entire training data to minimize the loss function.

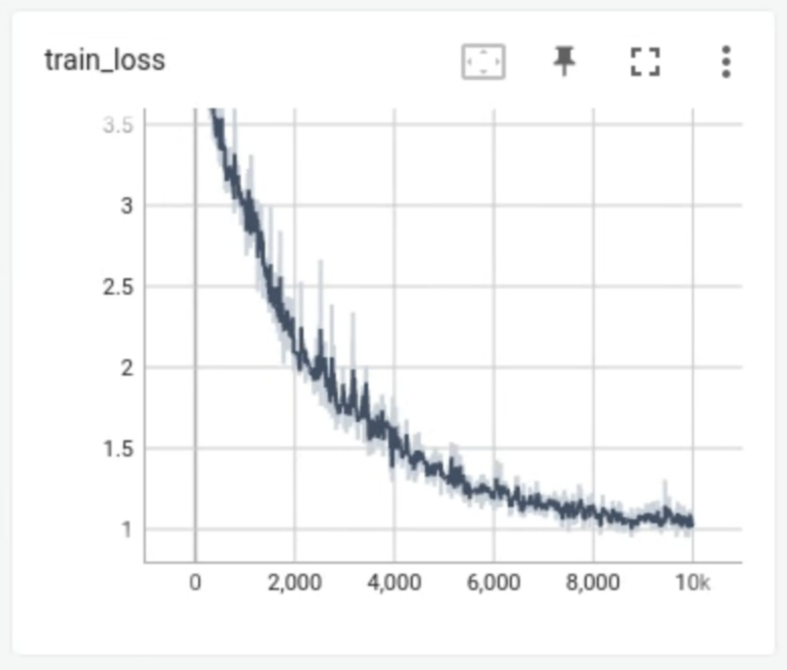


Figure 8: The loss function

5 Conclusion

Self-supervised deep learning is a very powerful technique in processing huge amounts of galaxy images, which can significantly save human efforts to find galaxy mergers. Like all other machine learning methods, self-supervised learning is not always reliable, as some false positives marked in red in the first image in Figure 7. However, self-supervised learning is very important in selecting very good candidates for the human to validate. It is impossible for humans to manually process billions of images without machine learning methods. The results in this research shows promising performance of self-supervised learning.

6 Future Research

Future studies will focus on exploring other machine learning tasks for galaxy images.

- To train a self-supervised learning model on a larger dataset should give a better performance, which will be a future direction for investigation.
- Explore machine learning methods to classify galaxy images.
- Extend the self-supervised learning to identify other categories of galaxy.

References

[1] `http://www.legacysurvey.org/`.

[2] `http://www.zooniverse.org/projects/zookeeper/galaxy-zoo/`.

[3] X. Chen, H. Fan, R. Girshick, , and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[4] M. A. Hayat, G. Stein, P. Harrington, Z. Lukic, and M. Mustafa. Self-supervised representation learning for astronomical images. *The Astrophysical Journal Letters*, 2021.

[5] G. Stein, P. Harrington, J. Blaum, T. Medan, and Z. Lukic. Self-supervised similarity search for large scientific datasets. *Fourth Workshop on Machine Learning and the Physical Sciences*, 2021.

[6] K. W. Willett. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.