Self-supervised learning for searching jellyfish galaxies in the ocean of data from upcoming surveys

Yash Gondhalekar
Department of CSIS
BITS Pilani, K.K. Birla Goa Campus
yashgondhalekar567@gmail.com

Rafael S. de Souza
Centre for Astrophysics Research, University of Hertfordshire
College Lane, Hatfield, AL10 9AB, UK
rd23aag@herts.ac.uk

Ana L. Chies-Santos
Instituto de Física
Universidade Federal do Rio Grande do Sul,
Av. Bento Gonçalves, 9500 - Agronomia, Porto Alegre - RS, 91501-970
Porto Alegre, RS, Brazil
ana.chies@ufrgs.br

Carolina Queiroz
Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo
Rua do Matão, 1371, CEP 05508-090, São Paulo, Brazil
c.queirozabs@gmail.com

Abstract

Human visual classification is the traditional approach to identifying jellyfish galaxies. However, this approach is unsuitable for large-scale galaxy surveys. In this study, we employ self-supervised learning on a dataset of approximately 200 images to extract semantically meaningful representations of galaxies. Despite the small dataset size, a similarity search suggests that the self-supervised representation space contains meaningful morphological information. We propose a framework for assigning JClass, a categorical disturbance measure, based on nearest-neighbor search in the self-supervised representation space to assist visual classifiers. Our pipeline is highly adaptable, allowing for the seamless identification of any rare astronomical signatures within astronomical datasets.

1 Introduction

Jellyfish galaxies are galaxies with disturbed morphologies; however, they are rare to spot [1, 4, 12, 16]. Traditionally, they have been identified only by visual inspection. A prevalent method for assessing the disturbance in morphologies of galaxies is to designate a JClass. This classification system ranges from 0 to 5, with 5 representing the most robust evidence of disturbance and 0 indicating the weakest. A distinct visual characteristic of these galaxies is the presence of elongated jellyfish-like tails extending in a preferred direction. However, such visual classification is time-consuming, and measuring the disturbance accurately based solely on visual analysis is difficult.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

Self-supervised learning (hereafter, SSL) has proven promising in recent years for learning generalized, semantically meaningful data representations. Its applications in astronomy are becoming popular. [6] demonstrated the superiority of SSL over supervised learning methods applied to multiband galaxy images from the Sloan Digital Sky Survey (SDSS), [13] showcased the improved robustness of SSL representations to non-physical properties compared to Principal Component Analysis (PCA), [14] developed an online web-based tool for fast similarity search. An application towards finding the rare strongly gravitationally lensed galaxies was demonstrated in [15].

Due to the limitations of traditional supervised learning methods in such settings (scarcity of data: $\sim$200 images, severe imbalance of classes), we leverage the end-to-end `SimCLR` contrastive learning framework to extract robust representations of galaxies. Typically, these representations form the basis for several downstream tasks, such as classification or similarity search. We introduce a downstream task of predicting the JClass of galaxies using the extracted self-supervised representations to aid the uncertain and laborious visual classification. To this end, we propose a simple framework that uses information from confident visual classifications to predict the JClass of any given galaxy by using the JClass of its n-nearest neighbors in the self-supervised representation space.

2 Methods

2.1 Data

Our dataset consists of $\sim$200 images from three nearby galaxy clusters (Fornax, Antlia, and Hydra) obtained in 12 bands (each band allowing a specific range of wavelengths) of the Southern-Photometric Local Universe Survey [10]. The dataset constitutes a main sample displaying emissions thought to be associated with jellyfish galaxies and a control sample, which did not display such emissions.

The classification was carried out in an internal project on the Zooniverse platform[1]. Given a galaxy, six visual classifiers checked for visual disturbance and assigned a JClass. Galaxies with merger evidence, i.e., two or more galaxies in interaction, were flagged. The final JClass was the median across all JClasses, excluding merger cases. This process yielded 51 candidate jellyfish galaxies across all three clusters, with a JClass distribution ranging from 4 to 1 – there are no instances of JClass 5 galaxies in our data. Since our primary aim is to improve visual JClasses without entirely relying on the visual JClasses, we focus only on galaxies with high uncertainty in the JClass among the visual classifiers (see Sect. 3.3). The extremely small size of our dataset is used to test whether (a) SSL can learn meaningful representations of galaxies and (b) SSL representations encode important features for downstream tasks such as classification, which was previously unexplored in an astronomical context for small data.

2.2 Architecture and training

We use a shallower `Resnet-34` as our encoder in `SimCLR` instead of `Resnet-50` to alleviate overfitting in our small data set. Its architecture is modified to accept our 12-channel images. The projection head is widened by using four times more neurons in the projection head [3]. The stride in convolutions is reduced from 2 to 1 in the first convolutional layer, and the amount of pooling is reduced by removing the first max pooling layer ([6], [11]). This results in a 512-dimensional representation vector for each galaxy. The representations used in this study are the outputs of the encoder instead of the projection head since such a setting is more beneficial for downstream analyses [2].

Before training, background sources from the images are removed to ensure the learning is not biased due to nearby sources. We have used the `galmask` package [5] for this purpose, and is applied independently on each band. An extensive set of augmentations are used during training to make the representations invariant to changes in noise, orientation, and the Point Spread Function (PSF). In addition to the usual random cropping and rotation augmentations, we randomly apply a custom color jitter (with 0.8 probability) to each image channel, following [7]. For invariance to PSF, we convolve the images with a 9×9 Gaussian kernel with standard deviation uniformly sampled in the range $[0.1, 2.0]$.

[1] `https://www.zooniverse.org/`

The contrastive loss function (also called the NT-Xent loss function) is defined as $l_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}(z_i, z_k)/\tau\right)}$ where $(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)$ denotes a "positive" pair (i.e., augmented versions of the image $\boldsymbol{x}$), τ is the temperature hyperparameter, the indicator function $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ evaluates to 1 if and only if $k \neq i$, $\text{sim}(\boldsymbol{p}, \boldsymbol{q}) = \frac{\boldsymbol{p} \cdot \boldsymbol{q}}{\|\boldsymbol{p}\| \|\boldsymbol{q}\|}$. During training, $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$ are passed through the network (consisting of an encoder and a non-linear projection head). The above contrastive loss is applied to the projection head's output.

Training is performed using the contrastive loss function for 1000 epochs, optimized using the Adam with decoupled weight decay (AdamW; [9]) optimizer with a weight decay of 10^{-4} and a learning rate, $lr = 10^{-4}$. The learning rate is manipulated using a cosine annealing schedule with the minimum learning rate set to $lr/50$ and without restarts [8], and the maximum number of epochs set to 1000. We set $\tau = 0.05$. All hyperparameters are tuned using K-Fold cross-validation using a combination of contrastive loss and top-5 accuracy (the number of times the desired patch is within the top 5 most similar examples to the original image in the sampled batch). For computational reasons, $K = 3$ is used instead of the common choices of $K = 5, 10$. Large batch sizes are known to improve SSL performance [2]. We thus use a batch size of 128, which is the maximum batch size feasible for our application, given our RAM limitations.

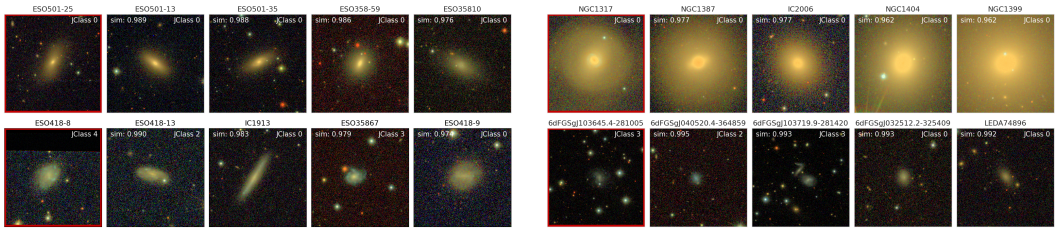

Figure 1: A few query-by-example searches using SSL. Images marked with a red border are the query images. The four closest images to the query image are shown from left to right, in decreasing order of similarity. The JClass and the cosine similarity values are shown on each image. For better clarity, we show the original images and not the ones after applying `galmask`.

The learned encoder is used to yield representations of galaxies from the test set. To highlight the robustness of SSL to the location in the sky and related systematics, we show results by training on galaxies from the Antlia and Hydra galaxy clusters and testing on images from the Fornax cluster.

3 Results

3.1 Representation quality evaluation

We use the linear evaluation protocol for quantifying self-supervised representation quality, in which we train a supervised logistic regression classifier on top of the features extracted from the encoder. The encoder is used as a fixed feature extractor. We compare it with standard supervised learning CNN. For this section, we consider binary labels (0: non-jellyfish, 1: jellyfish), which means no distinctions are made between different jellyfish strengths (JClass 1 to 4). This choice prevents the class imbalance issue from aggravating since high disturbance cases (JClass 3, 4) are extremely rare. Class imbalance is handled by using a weighted random sampler during training of the logistic regression and the supervised CNN.

The precision, recall, and accuracy on the test set for the supervised CNN and the logistic regression trained on the self-supervised representations are (72.3, 79.3), (78.8, 84.2), and (86.3, 87.95). Thus, the linear evaluation protocol outperforms traditional supervised learning. We also found that the supervised approach is susceptible to class imbalance, alleviated by the linear evaluation protocol since the self-supervised representations extracted are label-agnostic.

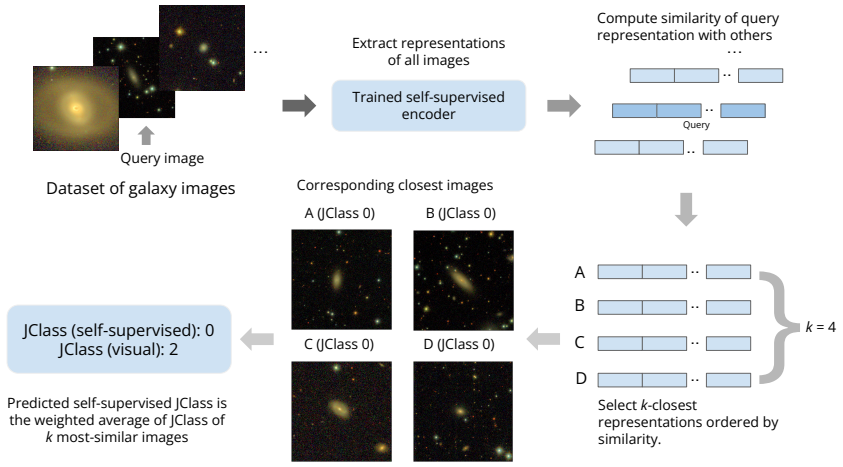

Figure 2: Workflow of our proposed approach, which conducts a similarity search and assigns a JClass to the query galaxy based on the JClasses of similar galaxies in the self-supervised representation space.

3.2 Query by example

We conduct a query by example (or similarity search) to inspect what types of galaxies are clustered closely in the self-supervised representation space. Based on cosine similarity, the top four closest images to a chosen query image are selected and shown in Fig. 1. The top row shows two cases where all closest images to the query image (with JClass 0) also have a JClass 0. The bottom row shows query galaxies with extreme disturbance (JClass 4 and 3) that have a mix of non-jellyfish (JClass 0) and jellyfish (JClass $\neq 0$) galaxies as the closest galaxies in the self-supervised space. This could happen since galaxies have many features (including or excluding the ones that distinguish jellyfish from non-jellyfish) mapped onto the same self-supervised space. Overall, it can be observed that galaxies similar in color and shape to the respective query images are present in the similarity search, thus suggesting that galaxies with similar features are clustered closely in the representation space.

3.3 Re-calibrating visual classifications

Since the JClass assigned by visual inspection is based on a subjective assessment of jellyfish-ness, a linear evaluation protocol (see Sect. 3.1) that uses these JClasses as ground-truth labels will be affected by the quality of these labels. The linear evaluation will not yield a precise disturbance strength estimate, either, since it is only a binary classification (jellyfish vs. non-jellyfish). As stated above, a multi-label classification will degrade due to the increased severity of the class imbalance. Thus, a supervised regressor trained on self-supervised representations is not ideal for improving JClass. To mitigate these issues, we develop a new downstream task to assign JClass to galaxies leveraging self-supervised representations to assist visual classifiers in their classification.

For N visual classifiers trying to assign a JClass to a given galaxy, we consider the galaxy uncertain for visual classification if there are $> \lceil N/2 \rceil$ unique visually-assigned JClasses (where $\lceil\ \rceil$ is the ceiling function). We assign a JClass for a given galaxy using the JClasses from its nearby galaxies in the self-supervised representation space. Therefore, we define the following rule: $\text{JClass}_{ss} = \frac{\sum_{i=1}^{n} s_i \, \text{JClass}_{v_i}}{\sum_{i=1}^{n} s_i}$ where s_i is the cosine similarity between the query image and the i^{th} similar image to it, n is the number of closest images considered, and JClass_{ss} and JClass_{v} are the self-supervised-predicted and visual JClasses respectively. We use $n = 4$. In our framework, we ensure that the nearby galaxies are not uncertain to classify visually, i.e., have $\leq \lceil N/2 \rceil$ unique visual JClasses.

Fig. 2 illustrates our framework to assign JClass to galaxies based on the similarity search on the representations of galaxies. Although our proposed approach uses visual JClasses for the final prediction, we note that it does not use these JClasses to learn to distinguish between jellyfish and

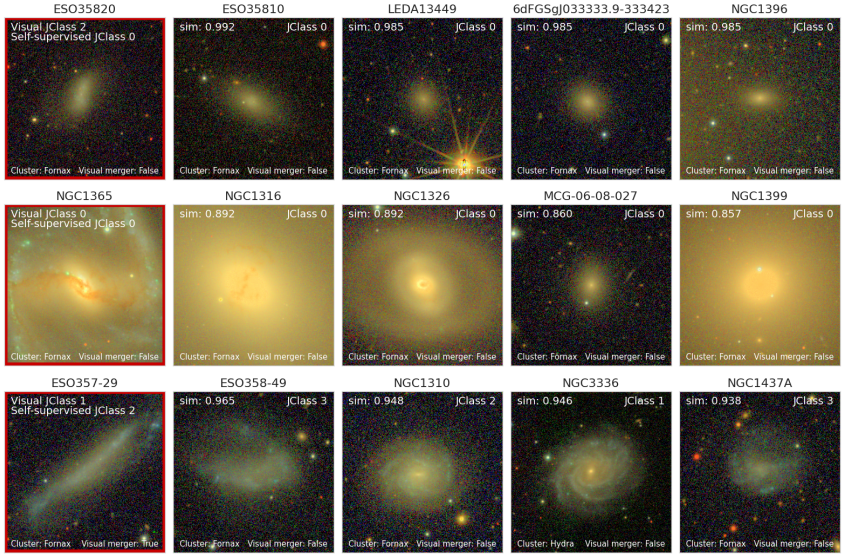

Figure 3: Three examples demonstrating the application of our framework. Galaxies to which a JClass is assigned using our approach (i.e., query galaxies) are outlined in red. Similar to Fig. 1, the corresponding similar images are shown. The visual JClass is denoted at the top-right of the images, and the JClass assigned to the query is also shown.

non-jellyfish due to the self-supervised nature of the training process, unlike supervised CNNs or supervised regressors trained on the representations.

Fig. 3 shows the application of our framework. We show three cases where the self-supervised JClass was smaller, equal, and higher (top to bottom, in that order) than the corresponding visual JClass. In the top row, the visual classification predicted the query galaxy to be a JClass 2. However, since the query image resembles galaxies with a visual JClass 0, our framework predicts it as a JClass 0 galaxy. The middle row shows an agreement between visual and self-supervised JClass assignments. All the most similar galaxies to the query have a JClass 0. Thus, the query image is assigned a JClass 0. In the bottom case, the most similar images had a stronger disturbance signature, resulting in the self-supervised approach assigning a higher JClass of 2 instead of the visual JClass 1. Here, while the self-supervised approach predicted a higher JClass for the query image, it was identified as a merger by visual classification. Hence, the self-supervised model might not clearly distinguish jellyfish from merger galaxies.

4 Conclusion

A similarity search using SSL revealed that the learned representations of our galaxies are robust to orientation and noise, which is promising considering the paucity of data. After human classification, our framework of predicting disturbance strengths of galaxies using SSL can improve the JClasses of uncertain galaxies without needing manual follow-up analysis, thus improving classification speed. SSL also alleviates human biases, so SSL predictions are expected to produce more reliable disturbance strengths. Hence, our framework can also be used as a guide to train human classifiers to assist in their visual classification. Another application is identifying false positives and negatives during follow-up analysis after human classification. Our work demonstrates the application of SSL in the low-data regime. With large astronomical datasets, more powerful semantic embeddings are expected to be obtained, thus improving the performance even further.

An extension of this work is to identify and disentangle features or learn a similarity metric (instead of fixing it to cosine similarity) that helps distinguish jellyfish from non-jellyfish galaxies, which could help differentiate between jellyfish and merger galaxies.

Acknowledgments and Disclosure of Funding

The authors acknowledge insightful discussions with Fabricio Ferrari, Amanda R. Lopes, Gabriel M. Azevedo, and Hellen Monteiro-Pereira. The authors also acknowledge financial support from the São Paulo Research Foundation (FAPESP), the Brazilian National Research Council (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES), the Carlos Chagas Filho Rio de Janeiro State Research Foundation (FAPERJ), and the Brazilian Innovation Agency (FINEP). RSS acknowledges the support from the China Manned Space Project with NO. CMS-CSST-2021-A07. ACS acknowledges funding from CNPq and the Rio Grande do Sul Research Foundation (FAPERGS) through grants CNPq-403580/2016-1, CNPq-11153/2018-6, PqG/FAPERGS-17/2551-0001, FAPERGS/CAPES 19/2551-0000696-9. This work was partially supported by the CAS PIFI programme 2021VMC0005.

References

[1] Alessandro Boselli, Matteo Fossati, and Ming Sun. Ram pressure stripping in high-density environments. *Astronomy and Astrophysics Reviews*, 30(1):3, December 2022. doi: 10.1007/s00159-022-00140-3.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL `https://arxiv.org/abs/2002.05709`.

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[4] Junia Göller, Gandhali Joshi, Eric Rohr, Elad Zinger, and Annalisa Pillepich. Jellyfish galaxies with the IllustrisTNG simulations – No enhanced population-wide star formation according to TNG50. *arXiv e-prints*, art. arXiv:2304.09199, April 2023. doi: 10.48550/arXiv.2304.09199.

[5] Yash Gondhalekar, Rafael S. de Souza, and Ana L. Chies-Santos. galmask: A python package for unsupervised galaxy masking. *Research Notes of the AAS*, 6(6):128, jun 2022. doi: 10.3847/2515-5172/ac780b. URL `https://dx.doi.org/10.3847/2515-5172/ac780b`.

[6] Md Abul Hayat, George Stein, Peter Harrington, Zarija Lukić, and Mustafa Mustafa. Self-supervised representation learning for astronomical images. *The Astrophysical Journal Letters*, 911(2):L33, apr 2021. doi: 10.3847/2041-8213/abf2c7. URL `https://doi.org/10.3847/2041-8213/abf2c7`.

[7] Svetlana Illarionova , Sergey Nesteruk , Dmitrii Shadrin, Vladimir Ignatiev , Maria Pukalchik , and Ivan Oseledets. Mixchannel: Advanced augmentation for multispectral satellite images. *Remote Sensing*, 13(11), 2021. ISSN 2072-4292. doi: 10.3390/rs13112181. URL `https://www.mdpi.com/2072-4292/13/11/2181`.

[8] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[10] C. Mendes de Oliveira, T. Ribeiro, W. Schoenell, A. Kanaan, R. A. Overzier, A. Molino, L. Sampedro, P. Coelho, and et al. The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies, and redshifts with 12 optical filters. *MNRAS*, 489(1):241–267, Oct 2019. doi: 10.1093/mnras/stz1985.

[11] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7354, 2020.

[12] Eric Rohr, Annalisa Pillepich, Dylan Nelson, Elad Zinger, Gandhali Joshi, and Mohommadreza Ayromlou. Jellyfish galaxies with the IllustrisTNG simulations – When, where, and for how long does ram pressure stripping of cold gas occur? *arXiv e-prints*, art. arXiv:2304.09196, April 2023. doi: 10.48550/arXiv.2304.09196.

[13] Regina Sarmiento, Marc Huertas-Company, Johan H. Knapen, Sebastián F. Sánchez, Helena Domínguez Sánchez, Niv Drory, and Jesus Falcón-Barroso. Capturing the physics of manga galaxies with self-supervised machine learning. *The Astrophysical Journal*, 921(2): 177, nov 2021. doi: 10.3847/1538-4357/ac1dac. URL https://dx.doi.org/10.3847/1538-4357/ac1dac.

[14] George Stein, Peter Harrington, Jacqueline Blaum, Tomislav Medan, and Zarija Lukic. Self-supervised similarity search for large scientific datasets. *arXiv e-prints*, art. arXiv:2110.13151, October 2021. doi: 10.48550/arXiv.2110.13151.

[15] George Stein, Jacqueline Blaum, Peter Harrington, Tomislav Medan, and Zarija Lukić. Mining for strong gravitational lenses with self-supervised learning. *The Astrophysical Journal*, 932 (2):107, jun 2022. doi: 10.3847/1538-4357/ac6d63. URL https://dx.doi.org/10.3847/1538-4357/ac6d63.

[16] Elad Zinger, Gandhali Joshi, Annalisa Pillepich, Eric Rohr, and Dylan Nelson. Jellyfish galaxies with the IllustrisTNG simulations – Citizen-science results towards large distances, low-mass hosts, and high redshifts. *arXiv e-prints*, art. arXiv:2304.09202, April 2023. doi: 10.48550/arXiv.2304.09202.