An AI-Assisted Labeling Tool for Cataloging High-Resolution Images of Galaxies

Gustavo Perez[1,*], **Sean Linden**[2], **Timothy McQuaid**[2], **Matteo Messa**[3], **Daniela Calzetti**[2], **and Subhransu Maji**[1]

[1]Manning College of Information and Computer Sciences, University of Massachusetts Amherst
[2]Department of Astronomy, University of Massachusetts Amherst
[3]Department of Astronomy, Stockholm University
[*]gperezsarabi@umass.edu

Abstract

The *Hubble Space Telescope* (HST), the recently launched *James Web Space Telescope* (JWST), and many earth-based observatories collect data allowing astronomers to answer fundamental questions about the Universe. In this work we focus on an ecosystem of AI tools for cataloging bright sources within galaxies, and use them to analyze young star clusters – groups of stars held together by their gravitational fields. Their ages and masses, among other properties provide insights into the process of star formation and the birth and evolution of galaxies. Significant domain expertise and resources are required to discriminate star clusters among tens of thousands of sources that may be extracted for each galaxy. To accelerate this step we propose: 1) a web-based annotation tool to label and visualize high-resolution astronomy data, encouraging efficient labeling and consensus building; and 2) techniques to reduce the annotation cost by leveraging recent advances in unsupervised representation learning on images. We present case studies where we work with astronomy researchers to validate the annotation tool and find that the proposed tools can reduce the annotation effort by $3\times$ on existing HST catalogs, while facilitating accelerated analysis of new data.

1 Introduction

The origin of the universe, the existence of extraterrestrial life, and numerous other questions motivate the study of the cosmos. More than three decades of data from the Hubble Space Telescope (HST), and new data from the recently launched James Webb Space Telescope (JWST), are expected to provide fundamental insights into many of these questions. One such question is the process of birth, spatial and temporal evolution, and chemical properties of *star clusters* within galaxies. Star clusters are groups of stars, typically consisting of thousands to millions of stars, held together by their mutual gravitational attraction. Since the majority of star formation in a galaxy occurs in clustered structures [30], the distribution of their ages, masses, and other properties provide insights into the formation, evolution, and appearance of galaxies [30, 7, 37, 3].

The HST has produced high-resolution images of many nearby galaxies, and the data is of sufficient quality to resolve these galaxies into their individual sources. Astronomers often collaborate in large teams to produce a *survey* or a *catalog* documenting individual sources within these images. For instance, The HST Legacy ExtraGalactic UV Survey (LEGUS) [11] consists of more than 15000 human-labeled stars and star clusters, while the USNO-B Catalog [39] comprises one billion cataloged

NeurIPS 2021 AI for Science Workshop.

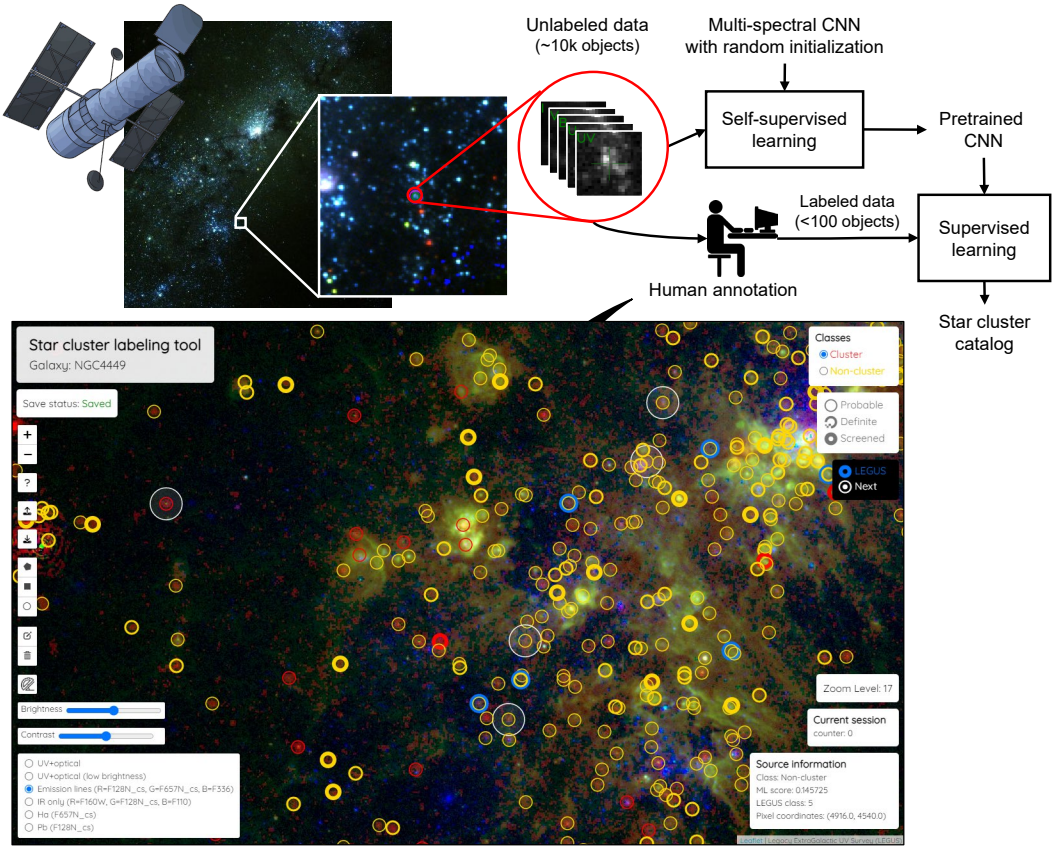

Figure 1: **The proposed tool for cataloging sources within galaxies.** (Top) An AI model assists data labeling by using the few labels provided initially to guide further labeling. (Bottom) A web-based user interface to label sources. The UI allows annotators to zoom into different regions dynamically and can be customized to support different labels (e.g., center coordinates, bounding boxes, irregular polygons) and spectral measurements as layers. In this figure we show a customized version of the UI for IR images showing annotations overlaid on HST observations of galaxy NGC 4449.

objects. Yet, producing catalogs can be a challenging and time-consuming task – a nearby galaxy might contain ten of thousands of bright sources and classifying these requires considerable domain expertise. LEGUS represents about 9000 person-hours of estimated manual annotation effort, which is hard to replicate as finer-grained categorization will be enabled as new measurements become available, e.g., from JWST.

Our work focuses on constructing star-cluster catalogs, though the tools are general and can be applied to construct other surveys. The process typically involves the identification of sufficiently bright sources within the galaxy using a tool such as the Source Extractor [8], followed by heuristics to remove sources that are unlikely to be clusters based on their brightness profile (to a rough approximation clusters are more spatially extended than individual stars). Subsequently, these are manually inspected to identify the star clusters. The process requires careful inspection of the spectral and spatial properties of their brightness distributions of several tens of thousands of sources extracted in the previous step. The task is so challenging and cumbersome that the existing LEGUS star-cluster catalogs have about a $\approx$87% consensus among humans for the binary (cluster/non-cluster) classification [43, 35].

To address these challenges we propose an ecosystem of AI tools along two major thrusts (§ 3). First is a *web-based annotation tool* allowing visualization of high-resolution images and supporting various forms of annotations (See Figure 1–bottom).This serves as a platform for collaborative labeling, visualization of existing labels, and consensus building. Projects such as the Galaxy Zoo [36] has

led to the labeling of 1.5 million galaxies into their morphological types by crowdsourcing on the Zooniverse platform [50]. However, similar tools are lacking for distributed labeling of star clusters due to the unique visualization and labeling needs. Second, are *techniques to produce catalogs with less labeling effort* by utilizing unsupervised learning (See Figure 1–top). A unique aspect of this problem is that astronomy data is heterogeneous, where the spatial resolution, wavelength coverage and number of filters vary across tasks. Frequently, there are no existing networks that transfer well making learning from little data challenging. We find that domain-specific representations trained using contrastive learning are effective even when they are trained on a *single* high-resolution image of a galaxy. This allows us to integrate heterogeneous data without significant labeling effort.

We pilot the proposed tool in two case studies. First, we show that existing catalogs in LEGUS can be constructed with a near-human agreement with only 32% of the labeling effort (§ 4). Second, we collaborate with astronomy researchers to construct new catalogs involving new modalities to understand and improve the entire workflow from data acquisition, to initial labeling, AI-assisted catalog construction, to enabling basic science (§ 5).

2 Related work

Astronomy image analysis Computer vision tools have been applied to numerous astronomy applications, including galaxy classification [18, 28, 6], galaxy morphology [16, 54], and galaxy mergers [1, 66]. More recently, machine learning has been applied to the classification of young star clusters [21, 57, 9, 10, 43]. In [9, 10], the authors select a subset of channels based on domain knowledge to classify star cluster images using transfer learning [60] from pretrained networks on color images. Later, [57] incorporates all the channels from the multi-spectral data to classify star clusters. The authors use an ensemble [32] of pretrained networks in parallel and split the image into subsets of three channels. More recently, [43] trains a multi-scale CNN from scratch using all channels in the star cluster images. The success of these methods is underscored by the availability of significant training data, while our work focuses on learning with limited data.

Labeling and visualization tools Tools for labeling and visualizing training data are indispensable for training and validating modern AI systems. While a large number of tools exist for labeling images or videos (e.g.,[52, 51, 19, 49]), similar tools are lacking for Astronomy imagery. They are similar to satellite imagery data (e.g., they are multi-spectral and of high resolution) or medical images and tools in these domains can be repurposed for our goals. For instance, in [27] and [47], the authors propose web-based labeling tools with dynamic tiling to handle high-resolution remote sensing data. In medicine, multiple labeling interfaces have been proposed [61, 45] for 3D medical data. Our web-based labeling tool for high-resolution galaxy images is based on [48] – a web-based framework for land cover labeling in satellite imagery. This allowing dynamic tiling for high-resolution imagery such as HST and JWST images, and supports tools for point, bounding box, and segmentation annotations.

Unsupervised representation learning The labeled data bottleneck can also be approached from the perspective of learning useful representations from few labeled samples using semi-supervised learning [33, 31, 64, 53, 46] methods, or with only unlabeled images using self-supervised learning [17, 56, 40, 63, 20, 42, 26, 65, 5]. In particular, contrastive learning [14, 25, 13, 22, 62] have performed particularly well in learning features from unlabeled data that are useful for downstream tasks and have been applied to other domains such as medical imaging [12, 59], remote sensing [34, 4, 58], and astronomy [23, 38].

3 An AI-assisted Labeling Tool for Building Catalogues of Galaxies

Thrust 1: Labeling interface Our labeling interface is based on Microsoft's AI4Earth satellite imagery labeling tool [48], an open-source web-based tool that supports dynamic tiling for high-resolution remote sensing imagery (see Figure 1). We modified and added new features based on

the goals of astronomy researchers. The tool supports different renditions which are helpful for annotating objects (e.g., ultra-violet, optical, and infrared bands). These renditions are saved as a cloud-optimized GeoTIFF (COG) and hosted on a web server. The tool supports dynamic tiling, allowing handling of high-resolution images. Labels are managed using GeoJSON files, so previously created catalogs (e.g., uncurated catalogs generated with SExtractor) can be imported and modified to later export them as a new GeoJSON file. The web-based interface allows distributed labeling and visualization on browsers without the need to install custom software packages.

The labeling interface can be customized to work with any high-resolution image, and to annotate different types of astronomical sources, as it allows the user to quickly mark objects by their center coordinates, bounding box, or select irregular regions. Furthermore, the implementation only requires a standard web server to host the data renditions and a tiling application for dynamically rendering the high-resolution images.

Thrust 2: Learning with limited data Our framework supports different aspects of learning with astronomy data:

- **Deep networks, pre-training and transfer** Our framework supports various forms of pre-trained models (e.g., ImageNet [15] pretrained) and self-supervised contrastive learning customized for astronomy data. One challenge is that astronomy data is of a different shape than color images (e.g., there are more channels) and some modification of the network is necessary (e.g., splitting the channels into groups of three or an adaptor [44]). Pre-training might improve generalization when a few labels are available for the initial training. See Appendix C for how contrastive learning from natural images can be adapted to astronomy data.
- **Active and semi-supervised learning** Active and semi-supervised learning can be used to efficiently label the unlabeled samples. For active learning we explore the effectiveness of entropy sampling (i.e., sampling points that maximize the entropy of the predicted probabilities) over random sampling, while for semi-supervised learning we explore a label propagation method based on pseudo-labels [33].

4 Case Study 1: Building Star Cluster Catalogs with LEGUS

The HST Treasury Program Legacy ExtraGalactic UV Survey (LEGUS) [11] consist of 50 galaxies spanning distances from 3.5 Mpc and ≈16 Mpc using HST imagery in five bands: NUV (F275W filter), U (F336W filter), B (either F438W or F435W filter), V (either F555W or F606W filter) and I (F814W filter). When building the star cluster catalogs for each galaxy, the raw images are initially preprocessed and aligned. Then, a preliminary set of detections is generated with SExtractor [8] using a set of basic selection functions to remove the majority of stars and artifacts. A set of distribution fits are applied to rule-out additional non-cluster objects before a final visual inspection on objects brighter than V=−6 mag [1] by at least three human annotators. The total number of objects in LEGUS catalog is 15471, and each one is classified into four classes based on morphology [2] (See Figure 2):

- **Class 1:** compact objects with symmetric density profile.
- **Class 2:** compact objects with asymmetric or elongated density profile.
- **Class 3:** multi-peak associations with a shared diffuse underlying emission.
- **Class 4:** foreground stars, background galaxies, artifacts, and any other spurious detection.

Cluster/non-cluster variant For most applications, a binary classification of the sources to *cluster/non-cluster* is sufficient. From LEGUS, classes 1 and 2 are combined into the *cluster* class, and classes 3 and 4 are combined into the *non-cluster* class. LEGUS contains a total of 5573 (36%) *cluster* sources and 9898 (64%) *non-cluster* sources. We use the binary classification scheme in our case studies (§ 4–5).

[1] Measure of the brightness of a star or other celestial body.

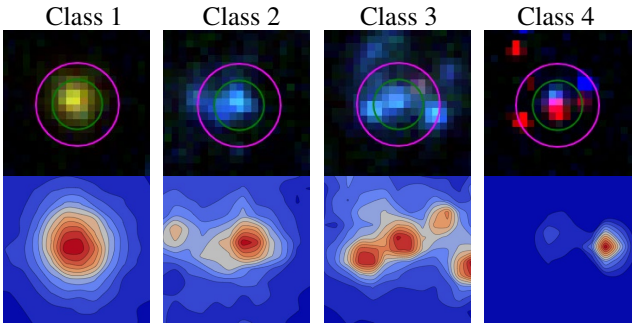

Figure 2: **LEGUS morphological classes.** Prototypical examples of LEGUS classes. Class 1 are compact symmetric sources, class 2 compact asymmetric sources, class 3 multi-peak associations, and class 4 spurious detections. Class 1 and 2 correspond to star clusters. The top row shows the color renditions, while the bottom row shows the contour and heat maps of the V band of each example.

Human agreement Each source was labeled by several annotators. We calculate human agreement as the fraction of annotations that agree with the final classification (mode) from all the annotators. The agreement is 75.0% for four-class task and 87.2% for binary task.

4.1 Problem formulation

Initially a small labeled dataset $D_L = (x_i, y_i)_{i=1}^{N_L}$ and a unlabeled dataset $D_U = (x_i)_{i=1}^{N_U}$ where $N_U \gg N_L$ are available, which can inform choices for pre-training, active learning, and self-supervised learning. We also assume a separate validation set $D_V = (x_i, y_i)_{i=1}^{N_V}$, such that $D_V \cap D_U \cap D_L = \varnothing$. We use the training-validation split proposed in [43], which consists of ≈11000 and ≈1200 training and validation objects respectively, and report the mean result for each experiment after four runs with four different seeds. See training hyper-parameters in Appendix B.

4.2 Results

Role of representation learning. Training a deep network for star cluster classification that matches human agreement (≈85%) requires 5000 labeled samples, amounting to 32% of the entire LEGUS catalog, about a 3× reduction in the labeling effort. All baselines perform similarly down to 500 labeled images, as shown in Figure 3. On the other hand, for very low data regimes (less than 100 labeled images) unsupervised learning with contrastive pre-training (using BYOL [22]) leads to significantly higher performance initially (≈75% with 20 labeled samples) compared to training a network from scratch or transfer learning from an ImageNet pretrained ResNet50 [24] (62-66% which is close to chance performance).

On this dataset the performance of the linear classifier is comparable to fine-tuning the entire network (light green). Training the linear classifier is significantly faster and can be done interactively with little computational resources within the labeling tool.

Applying semi-supervised learning on top of self-supervision increases performance slightly, as shown in Figure 3(black). On the active learning side, sampling based on entropy of the predictions did not improve accuracy compared to random sampling for any number of labeled images.

5 Case Study 2: Building Catalogs of Young Star Clusters from IR data

Stars are born inside dense dust and gas concentrations known as molecular clouds. When these young and massive stars ionize the surrounding gas clouds with high-energy UV radiation, ionized hydrogen (HII) regions are formed. These very young regions produce a sufficient amount of ionizing radiation that their high-redshift counterparts could be responsible for the reionization at redshifts 6-9 (age of the Universe: 0.5-1 Gyr), the last phase transition our Universe underwent. For this to be the case, these regions need to emerge quickly from their dense and opaque natal clouds, so that enough

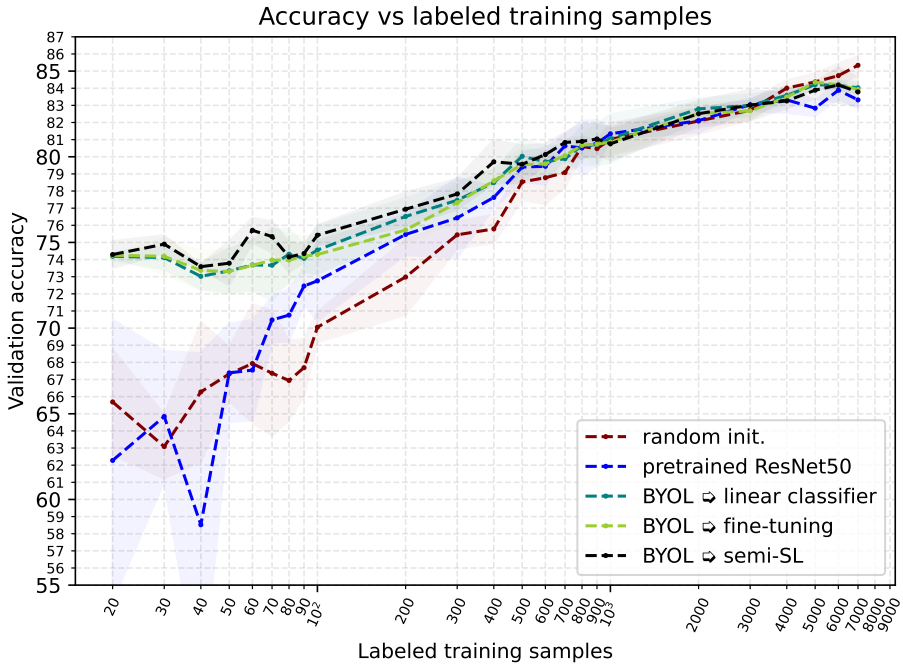

Figure 3: **Case study 1: performance as a function of training data.** Unsupervised learning using contrastive pre-training (light green) leads to significantly higher performance than training a network from scratch (red) or transfer learning (blue) on a low data regime (<100 labeled samples) indicated with orange color. Furthermore, the performance of the linear classifier (teal) is comparable to fine-tuning the entire network (light green), and applying semi-supervised learning on top of self-supervision increases performance slightly (black). Finally, all methods perform similarly with 500 labeled images or more, and performance saturates (close to human agreement) after 5000 labeled samples (approximately 32% of the entire LEGUS objects).

massive stars are still alive to contribute to the reionization. The timescale for this 'emergence' remains highly debated, and is one of the questions the JWST is posed to address. Because these very young star clusters reside in very dense regions, optical light cannot emerge and they need to be investigated using infrared (IR) light. Figure 4(a) provides an example, from HST imaging, of a young star cluster surrounded by its HII region.

In preparation for the upcoming JWST images, we design a pilot study to classify young star clusters in HII regions using the near-IR filters from the HST [2]. Our pilot consists of five rounds of annotations with 3-4 human annotators on each round. We perform the study on galaxy NGC4449 due to its high rate of star formation and the presence of a population of massive and young star clusters.

For this case study, we use a preliminary catalog of 1119 sources generated with SExtractor on galaxy NGC4449, similar to the procedure in Section 4. These initial sources generally focus on the peaks in the emission line image [3] and are marked with yellow rings in the labeling interface, as shown in Figure 4(c). In addition to the near-UV and optical bands used in LEGUS (i.e., F275W, F336W, F435W, F555W, F657N, and F814W), we include near-infrared channels to our image data: F1100W, F1280W, and F1600W.

5.1 Labeling interface for young star clusters in near-IR images

We customize the tool to support emission lines and near IR renditions to the visualization (See Figure 4(a)). Also, we added a marker to sources that are found in LEGUS catalogs; given that

[2]HST near-IR images are lower resolution than optical and near-UV ones due to its mirror size. JWST filters are at longer wavelengths than HST near-IR filters, but JWST will provide IR images at a higher resolution.

[3]The emission line image is built with F657N and F1280N bands where atomic hydrogen emits radiation in the Balmer and Paschen series.

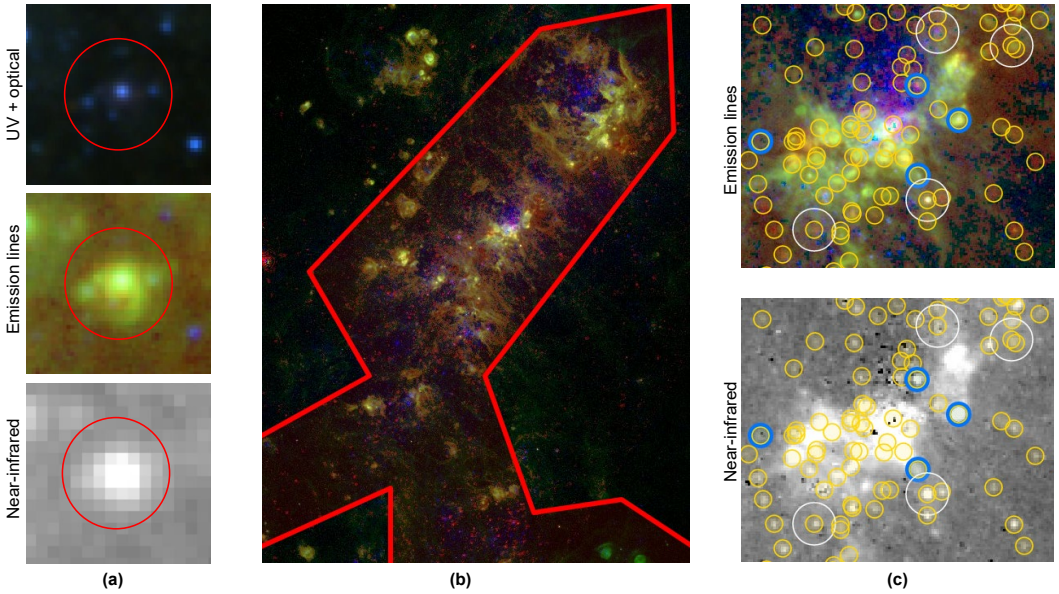

Figure 4: **(a)** Example of a young cluster inside an HII region. The cluster is shown in UV + optical bands (top), Emission lines (middle), and low-resolution near-IR (bottom). **(b)** Region with high HII concentration to sample sources for annotation. **(c)** Crops of our labeling tool showing SExtractor sources (yellow circle markers), sources also in the final LEGUS catalog for NGC 4449 (blue circle marker), and sources sampled for annotation (white large circle marker).

LEGUS clusters are already visually inspected (sources with blue circle markers in Figure 4(c)). Finally, we display specific information about each source, including the prediction score and the class the human annotators gave in the previous round with a majority voting consensus (See Figure 1).

We sample and mark the sources to be annotated in the following rounds to measure the agreement between human annotators (sources marked with a large white circle marker in Figure 4(c)). These sources are sampled randomly from high-informative areas selected manually by a user (i.e., from high HII concentration regions as shown in Figure 4(b)).

5.2 Results

Contrastive learning for near-IR images We use the entire set of 1119 sources obtained with SExtractor to pre-train the network using contrastive learning. In parallel, we annotate 658 from the 1119 sources where 398 (60.5%) sources were labeled as *clusters* and 260 (39.5%) as *non-clusters* and split the labeled samples into 90% for training and 10% for validation. Finally, using the same problem formulation as in Section 4.1, the same hyperparameters from Appendix B, and following the evaluation procedure described in Appendix C, we get a classification accuracy of 82.7%, demonstrating the generalizability of our approach to additional modalities and different image resolutions.

Human agreement Following the procedure in Section 4, we calculate a human-to-mode agreement of 93.3% averaged over the five rounds.

6 Conclusion

The analysis of star clusters has proven critical in answering scientific questions regarding stellar evolution and how galaxies are born. In this work we collaborate with astronomy researchers to propose an ecosystem of AI tools for cataloging bright sources in high-resolution images of nearby galaxies, reducing to 32% the labeling effort on the star cluster classification task with the LEGUS dataset. In addition, we demonstrate the robustness of our approach to new modalities and different image resolutions anticipating the upcoming images from the JWST.

References

[1] Ackermann, S. et al. (2018) Using transfer learning to detect galaxy mergers. *Monthly Notices of the Royal Astronomical Society* **479**1, pp. 415–425.

[2] Adamo, A. et al. (2017) Legacy ExtraGalactic UV Survey with The Hubble Space Telescope: Stellar Cluster Catalogs and First Insights Into Cluster Formation and Evolution in NGC 628. *The Astrophysical Journal* **841**(2)

[3] Adamo, A. et al. (2020) Star Clusters Near and Far; Tracing Star Formation Across Cosmic Time. *Space Science Reviews* **216**(4)

[4] Agastya, C. et al. (2021) Self-supervised Contrastive Learning for Irrigation Detection in Satellite Imagery. *CoRR abs/2108.05484*.

[5] Assran, M. et al. (2022) Masked Siamese Networks for Label-Efficient Learning. *CoRR abs/2204.07141*.

[6] Barchi, P. et al. (2020) Machine and Deep Learning applied to galaxy morphology - A comparative study. *Astronomy and Computing* **30**.

[7] Bastian, N. (2008) On the star formation rate - brightest cluster relation: estimating the peak star formation rate in post-merger galaxies*Monthly Notices of the Royal Astronomical Society* **390**(790).

[8] Bertin, E. & Arnouts, S. (1996) SExtractor: Software for source extraction. *Astronomy and Astrophysics, Supplement* **117** pp. 393-404.

[9] Bialopetravičius, J. et al. (2019) Deriving star cluster parameters with convolutional neural networks. *Astronomy Astrophysics* 621:A103.

[10] Bialopetravičius, J. & Narbutis D. (2020) Deriving star cluster parameters with convolutional neural networks. *Astronomy Astrophysics* 633:A148.

[11] Calzetti, D. et al. (2015) Legacy Extragalactic UV Survey (LEGUS) With the Hubble Space Telescope. I. Survey Description *The Astronomical Journal* **149**(2) pp. 25.

[12] Chaitanya, K. et al. (2020) Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems* **33**.

[13] Chen, T. et al. (2020) A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of Machine Learning Research*. pp. 1597–1607.

[14] Chopra, S. et al. (2005) Learning a similarity metric discriminatively, with application to face verification *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[15] Deng, J. et al. (2009) ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[16] Dieleman, S. et al. (2015) Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society* **450**2, pp. 1441–1459.

[17] Doersch, C. et al. (2015) Unsupervised visual representation learning by context prediction. *International Conference on Computer Vision (ICCV)*

[18] Domínguez Sánchez, H. et al (2018) Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society* **476**3, pp. 3661–3676.

[19] Dutta, A. & Zisserman, A. (2019) The VIA Annotation Software for Images, Audio and Video. *In Proceedings of the 27th ACM International Conference on Multimedia*

[20] Gidaris, S. et al. (2018) Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations (ICLR).*

[21] Grasha, K. et al. (2019) The spatial relation between young star clusters and molecular clouds in M51 with LEGUS. *Monthly Notices of the Royal Astronomical Society* **483**4, pp. 4707–4723.

[22] Grill, J. et al. (2020) Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. *Advances in Neural Information Processing Systems*. pp. 21271–21284.

[23] Hayat, A. et al. (2021) Self-supervised Representation Learning for Astronomical Images. *The Astrophysical Journal Letters* **911**(2).

[24] He, K. et al. (2015) Deep Residual Learning for Image Recognition. *arXiv 1512.03385*

[25] He, K. et al. (2020) Momentum Contrast for Unsupervised Visual Representation Learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] He, K. et al. (2022) Masked Autoencoders Are Scalable Vision Learners. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[27] Hou, D. et al. (2019) V-RSIR: An Open Access Web-Based Image Annotation Tool for Remote Sensing Image Retrieval. *IEEE Access* **7** pp. 83852-83862.

[28] Khan, A. et al. (2019) Deep learning at scale for the construction of galaxy catalogs in the Dark Energy Survey. *Physics Letters B* **795**, pp. 248-258.

[29] Kingma, D. & Ba, J. (2015) Adam: A method for stochastic optimization. *3rd International Conference for Learning Representations*.

[30] Lada, C.& Lada, E. (2003) Embedded Clusters in Molecular Clouds. *Annual Review of Astronomy and Astrophysics* **41**(57).

[31] Laine, S. & Aila, T. (2017) Temporal ensembling for semi-supervised learning. *International Conference on Learning Representations (ICLR)*

[32] Lakshminarayanan, B. et al. (2016) Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *10.48550/ARXIV.1612.01474*.

[33] Lee, D. (2013) Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. *ICML Workshop Challenges in Representation Learning (WREPL)*.

[34] Li, H. et al. (2021) Remote Sensing Images Semantic Segmentation with General Remote Sensing Vision Model via a Self-Supervised Contrastive Learning Method. *CoRR abs/2106.10605*.

[35] Linden, S. et al. (2022) Star Cluster Formation and Evolution in M101: An Investigation with the Legacy Extragalactic UV Survey. *The Astrophysical Journal* **935**(2).

[36] Lintott, C. et al. (2008) Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* **389**(3) pp. 1179-1189

[37] Longmore, S. et al. (2014) The Formation and Early Evolution of Young Massive Clusters. *Protostars and Planets VI*. Jan. pp. 291.

[38] Martinazzo, A. et al. (2020) Self-supervised Learning for Astronomical Image Classification. *CoRR abs/2004.11336*.

[39] Monet, B. et al. (2003) The USNO-B Catalog. *The Astronomical Journal* **125**(2) pp. 984-993.

[40] Noroozi, M. & Favaro, P. (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. *European Conference on Computer Vision (ECCV)*.

[41] Paszke, A. et al. (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32* pp. 8024–8035.

[42] Pathak, D. et al. (2017) Learning features by watching objects move. *IEEE Computer Vision and Pattern Recognition (CVPR)*.

[43] Perez, G. et al. (2021) StarcNet: Machine Learning for Star Cluster Identification. *The Astrophysical Journal* **907**(2) pp. 100.

[44] Perez, G. & Maji, S. (2022) Domain Adaptors for Hyperspectral Images. *26TH International Conference on Pattern Recognition (ICPR)*.

[45] Philbrick, K. et al. (2019) RIL-Contour: a Medical Imaging Dataset Annotation Tool for and with Deep Learning. *Journal of Digital Imaging* **32** pp. 571–581.

[46] Rasmus, A. et al. (2015) Semi-Supervised Learning with Ladder Networks. *The Conference and Workshop on Neural Information Processing Systems (NeurIPS)*.

[47] Robinson, C. et al. (2020) Human-Machine Collaboration for Fast Land Cover Mapping. *Thirty-fourth AAAI Conference on Artificial Intelligence (AAAI)*.

[48] Robinson, C. et al. (2022) Fast building segmentation from satellite imagery and few local labels. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. pp. 1463-1471.

[49] Sekachev, B. et al. (2020) opencv/cvat: v1.1.0. *https://doi.org/10.5281/zenodo.4009388*

[50] Simpson, R. & Page, K. & De Roure, D. (2014) Zooniverse: Observing the World's Largest Citizen Science Platform. *Proceedings of the 23rd International Conference on World Wide Web* **04** pp. 1049–1054.

[51] Smailis, C. & Iakovidis, D. (2012) Ontology-Based Automatic Image Annotation Exploiting Generalized Qualitative Spatial Semantics. *Lecture Notes in Computer Science* **7297** pp. 205-214.

[52] Tzutalin. (2015) LabelImg. *GitHub repository*. https://github.com/tzutalin/labelImg

[53] Verma, V. et al. (2019) Interpolation Consistency Training for Semi-Supervised Learning. *International Joint Conferences on Artificial Intelligence Organization*.

[54] Walmsley, M. et al. (2019) Identification of low surface brightness tidal features in galaxies using convolutional neural networks. *Monthly Notices of the Royal Astronomical Society* **483**3 pp. 2968–2982.

[55] Wang, P. (2022) byol-pytorch. *GitHub repository*. https://github.com/lucidrains/byol-pytorch.

[56] Wang, X. & Gupta, A. (2015) Unsupervised learning of visual representations using videos. *International Conference on Computer Vision (ICCV)*.

[57] Wei, W. et al. (2020) Deep transfer learning for star cluster classification: I. application to the PHANGS-HST survey. *Monthly Notices of the Royal Astronomical Society* **493**(3) pp. 3178–3193.

[58] Wei, Z. et al. (2021) Large-Scale River Mapping Using Contrastive Learning and Multi-Source Satellite Imagery. *Remote Sensing* **13** pp. 2893.

[59] Wu, Y. et al. (2021) Federated Contrastive Learning for Volumetric Medical Image Segmentation. *International Conference on Medical Image Computing and Computer Assisted Intervention*.

[60] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014) How transferable are features in deep neural networks?. *In Advances in Neural Information Processing Systems 27* pp. 3320–3328.

[61] Yushkevich, P. et al. (2006) User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **31**(3) pp. 1116-28.

[62] Zbontar, J. et al. (2021) Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *CoRR abs/2103.03230* .

[63] Zhang, R. et al. (2016) Colorful image colorization. *European Conference on Computer Vision (ECCV)*.

[64] Zhang, H. et al. (2018) Mixup: Beyond Empirical Risk Minimization. *International Conference on Learning Representations (ICLR)*.

[65] Zhou, J. et al. (2022) iBOT: Image BERT Pre-Training with Online Tokenizer. *International Conference on Learning Representations (ICLR)*.

[66] Ćiprijanović, A. et al. (2020) DeepMerge: Classifying high-redshift merging galaxies with deep neural networks. *Astronomy and Computing* **32**C.

A Color renditions for astronomical data

Astronomical multi-spectral images are often combined to form color renditions. For visualization purposes, in this work we merge the two reddest channels into a single red channel and the two bluest channels as a single blue channel as described below:

- R $= (\gamma_v V + \gamma_i I)/2$
- G $= (\gamma_b B)$
- B $= (\gamma_n NUV + \gamma_u U)/2$

where γ is the inverse gain of the filters used by the HST. This ensures the images we use are calibrated according to their relative fluxes, in units of $erg/s/cm^2/Angstrom$. The input *NUV*, *U*, *B*, *V*, and *I* images from HST are calibrated in $electrons/s$.

B Supervised and sel-supervised training hyper-parameters

Contrastive learning pretraining We use BYOL PyTorch [41] implementation by [55]. We pretrain our network with the adapted version of BYOL for 170,000 iterations (≈1000 epochs) with a batch size of 128 samples and a learning rate of 3E-4. We train for approximately 7 hours using a single GPU `1080ti`.

Supervised training We use the CNN optimized for star cluster classification proposed by [43]. We train for ten epochs and a batch size of 64 images using Adam optimizer [29]. We use a learning rate of η=1E-04 with a learning schedule $\eta' = \eta * 0.1$ in the 6^{th} epoch.

C BYOL adaptation for multi-spectral images and ablation

For our self-supervised learning experiments, we use an adapted version of BYOL [22] for multi-spectral star cluster data, which achieves good performance without needing negative pairs. In addition, BYOL is more robust to lower batch sizes, which allows training models with good performance using a single GPU.

Off-the-shelf BYOL (with the backbone input adapted to work with multi-spectral star cluster images) does not converge during training. This is not surprising since BYOL was designed with ResNet encoders for color images. Our approach uses an encoder that works on multi-spectral images.

Empirically, we find that the projector's size affects the network's training considerably in our approach (See Table C.1a). This is expected since the size of our CNN's output features $f_\theta(v)$ is 256 (compared to 2048 of ResNet). Using the default projector size of BYOL with our encoder and input data yields poor performance. Therefore, we reduce the projector hidden size while still emulating the dimensional increase from $f_\theta(v)$ to the projector hidden size (See second row in Table C.1a). This configuration improves over the default settings, but the performance is still low. We get the best results using a projector size of (256-256). See Table C.1a for more configurations.

Contrastive learning methods are sensitive to the choice of image augmentations [13, 22], and star cluster data is very different from color images like ImageNet. We apply modifications to the data augmentations that are also key for the model to converge during training. First, we limit the lower-bound scale of the random resize crop augmentation (See Table C.1c). The scale specifies the lower and upper bounds for the random area of the crop before resizing. Astronomical object images contain the object of interest centered in the image with multiple other objects in the close surroundings. For that reason, we hypothesize that small-scale values might result in two different centered objects signaled as a positive pair, thus confusing the model.

For the BYOL ablation experiments, we divide the LEGUS dataset into a 90-10% training-validation split. From the training set, we select 100 samples as our 'labeled dataset', the remaining samples as the 'unlabeled dataset', and evaluate over the complete validation set to compare performances between models. We follow the experimental setup as the main experiments in Figure 3. First, we pretrain a network from scratch with the 'unlabeled dataset' using BYOL. Then, we finetune the model with the 'labeled dataset'. After the training is complete, we evaluate the validation set.

Table C.1: **Ablation experiments with BYOL**. Ablation on different projector sizes and data augmentation choices. (Prob.) indicates probability of occurrence of the augmentation. (-) indicates the model was not able to train. We show the result with the default hyper-parameters in gray and the best result in **bold**. We perform all experiments by varying a single hyper-parameter from the best model.

(a) Projector hidden & output size

(hidden-output)	Acc.
(4096-256)	-
(512-256)	59.2%
(256-256)	**76.0%**
(256-128)	71.3%

(b) Color jitter

Prob.	Bri.	Con.	Sat.	Hue	Acc.
0.3	0.8	0.8	0.8	0.2	73.4%
0.3	**0.8**	**0.8**	**0.8**	**0.3**	**76.0%**
0.4	0.8	0.8	0.8	0.2	73.5%
0.3	0.8	0.8	0.8	0.5	73.4%
0.3	0.5	0.5	0.5	0.2	-

(c) Random crop scale

Scale	Acc.
(0.08, 1.0)	-
(0.3, 1.0)	-
(0.4, 1.0)	71.9%
(0.5, 1.0)	**76.0%**
(0.6, 1.0)	-

(d) Grayscale augm. prob.

Probability	Acc.
0.1	-
0.2	74.0%
0.3	**76.0%**
0.4	73.6%
0.5	-