Draft version April 9, 2021
Typeset using LaTeX **twocolumn** style in AASTeX63

Self-Supervised Representation Learning for Astronomical Images

Md Abul Hayat,[1,2,*] George Stein,[2,3,*] Peter Harrington,[2] Zarija Lukić,[2] and Mustafa Mustafa[2]

[1] *University of Arkansas, Fayetteville, AR 72701*
[2] *Lawrence Berkeley National Laboratory, Berkeley, CA 94720*
[3] *Berkeley Center for Cosmological Physics, University of California, Berkeley, CA 94720*

Submitted to The Astrophysical Journal Letters

ABSTRACT

Sky surveys are the largest data generators in astronomy, making automated tools for extracting meaningful scientific information an absolute necessity. We show that, without the need for labels, self-supervised learning recovers representations of sky survey images that are semantically useful for a variety of scientific tasks. These representations can be directly used as features, or fine-tuned, to outperform supervised methods trained only on labeled data. We apply a contrastive learning framework on multi-band galaxy photometry from the Sloan Digital Sky Survey (SDSS), to learn image representations. We then use them for galaxy morphology classification, and fine-tune them for photometric redshift estimation, using labels from the Galaxy Zoo 2 dataset and SDSS spectroscopy. In both downstream tasks, using the same learned representations, we outperform the supervised state-of-the-art results, and we show that our approach can achieve the accuracy of supervised models while using 2-4 times fewer labels for training. The codes, trained models, and data can be found at https://portal.nersc.gov/project/dasrepo/self-supervised-learning-sdss.

1. INTRODUCTION

Observing and imaging objects in the sky has been the main driver of the scientific discovery process in astronomy, because doing controlled experiments is not a viable option. The rapid advance of digital sky surveys in the 1990s, spearheaded by SDSS (Gunn et al. 1998, 2006), has rendered obsolete the approach of manual inspection of images by an expert. Instead, computational analysis methods are constantly being developed and applied (Ivezić et al. 2019a). Additionally, "citizen science" like the Galaxy Zoo[1] project (GZ, Lintott et al. 2008) plays an important role for tasks which are too complex to describe algorithmically, yet are heuristically quite comprehensible to humans, such as classification of galaxies based on their morphological types (Lintott et al. 2013). In recent years, machine learning methods have proven particularly useful for both classification and regression tasks (see Stein 2020 for a comprehensive list), but the majority of published works rely on the quantity and quality of (manually assigned) image labels.

Serendipitous discovery of an ionization echo from a recently faded quasar (Lintott et al. 2009), and the cumbersome search for similar systems that followed (Keel et al. 2012), showcases other big data challenges. It demonstrates the need for methods which allow for the discovery of truly unusual and previously unseen objects, and also the need to perform semantic (or feature) similarity searches on images in situations when the number of labels is as low as one. In the near future, incoming sky surveys such as the Vera Rubin Observatory (Ivezić et al. 2019b), Euclid (Laureijs et al. 2011), Nancy Grace Roman Space Telescope (Spergel et al. 2013), or the Square Kilometer Array[2] will open yet another research epoch, where datasets are of sizes which completely overwhelm even the most ambitious citizen science concepts. It is fair to say that the vast majority of images from these observatories will never be seen by a human eye. Thus, the capability to organize images without labels and programmatically search for semantic similarity or for interesting outliers will be

mahayat@uark.edu, gstein@berkeley.edu, pharrington@lbl.gov, zarija@lbl.gov, mmustafa@lbl.gov

* Equal contribution first authors.

[1] https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/

[2] https://www.skatelescope.org/

essential to maximize the scientific output of these missions.

This capability is heavily dependent on image *representations* — low-dimensional mappings of images which preserve their inherent information. Finding good representations is crucial to scientific *downstream* tasks such as clustering and classification of images, but is often elusive, partly due to the difficulty of gathering enough high-quality labels. Unsupervised machine learning methods aim to learn semantically meaningful representations of the data without relying on any labels (see, e.g. Alloghani et al. 2020). Many such methods have already been applied to studies of galaxy morphology (Hocking et al. 2018; Martin et al. 2020; Cheng et al. 2020a; Spindler et al. 2020), identification of strong lenses (Cheng et al. 2020b), and anomaly detection (Xiong et al. 2018; Margalef-Bentabol et al. 2020). Unfortunately, across most computer vision applications, the utility of unsupervised representations for downstream tasks has historically lagged behind that of the representations coming from supervised training (Caron et al. 2018).

However, very recent progress in self-supervised learning has now closed the gap with supervised learning in computer vision (He et al. 2020; Chen, T. et al. 2020a; Chen, X. et al. 2020; Chen, T. et al. 2020b). Self-supervised methods learn representations by training models to solve contrived tasks (e.g., filling in empty regions of data samples, or identifying different versions of the same object as a pair) where the labels are generated algorithmically from an unlabeled dataset. The aim is to design models and tasks that yield semantically meaningful representations which are useful for a variety of downstream tasks, and can be directly used or fine-tuned for these applications. Self-supervised pre-training is vital to state-of-the-art natural language models (Devlin et al. 2019; Radford et al. 2018; Nayak 2019); now that this method has undoubtedly crossed over into the computer vision domain, it has exciting prospects for broad scientific use.

In this paper, we demonstrate that self-supervised learning indeed has great utility for large astronomical surveys, using $\sim$ 1.2 million SDSS *ugriz* galaxy images with 64 $\times$ 64 pixels as a proof of concept dataset (full details of data acquisition and selection are given in Appendix A). In section 2, we review the method of contrastive self-supervised learning and propose data augmentations that induce good representations for sky survey images. This approach allows us to build powerful representations which we showcase in section 3.1. In sections 3.2 and 3.3 we use the self-supervised representations to quickly outperform supervised learning at two very common downstream tasks: morphological classification and inference of photometric redshifts, respectively.

2. METHOD

Recent self-supervised works (Bachman et al. 2019; Goyal et al. 2019; He et al. 2020; Chen, T. et al. 2020a; Chen, X. et al. 2020; Chen, T. et al. 2020b) use contrastive losses (Hadsell et al. 2006) to minimize the distance between different *views* (augmentations) of the same image in a learned representation space, while maximizing the distance between the representations of different images. The randomized augmentations producing these views should be semantic-preserving transformations of the input images, and the goal is to make the final representation invariant to these transformations (Tian et al. 2020; Xiao et al. 2020). This key design choice is application-dependent and requires prior knowledge. For example, in a galaxy survey, changing colors of galaxies could be detrimental for the downstream task of inferring photometric redshifts, even though color augmentation may be useful when classifying cats and dogs. For a base set of image augmentations that would be useful to the vast majority of downstream applications in sky surveys we propose the following:

- **Galactic extinction**. We want features to be invariant to the galactic latitude and object's position on the celestial sphere. To model the effects of foreground galactic dust, we introduce artificial reddening by sampling a $E(B-V)$ reddening value from $\mathcal{U}(0, 0.5)$ and applying the corresponding per-channel extinction according to the photometric calibration from Schlafly & Finkbeiner (2011).
- **Point Spread Function** (PSF). Due to a variety of factors over the time span of a galaxy survey, images do not have a consistent PSF. To be invariant to this we experiment with a PSF augmentation, modeled as wavelength-dependent Gaussian smoothing with a standard deviation in r-band drawn from $\mathcal{N}(0, 0.13'')$ and scaled appropriately to the other channels using $\lambda^{-0.3}$ (Xin et al. 2018).
- **Rotation**. To be invariant to the apparent orientation of each galaxy, we sample the angle of random rotation of each image from $\mathcal{U}(0, 2\pi)$.
- **Random jitter & crop**. We also desire invariance to the image centering. Thus, two integers are sampled from $\mathcal{U}(-7, 7)$ to move (jitter) the center of the image (of size 107^2) along each respective axis, then the jittered image is center-cropped to size 64^2.
- **Gaussian noise**. Finally, to be invariant to the instrumental noise, we sample a scalar from $\mathcal{U}(1, 3)$ and

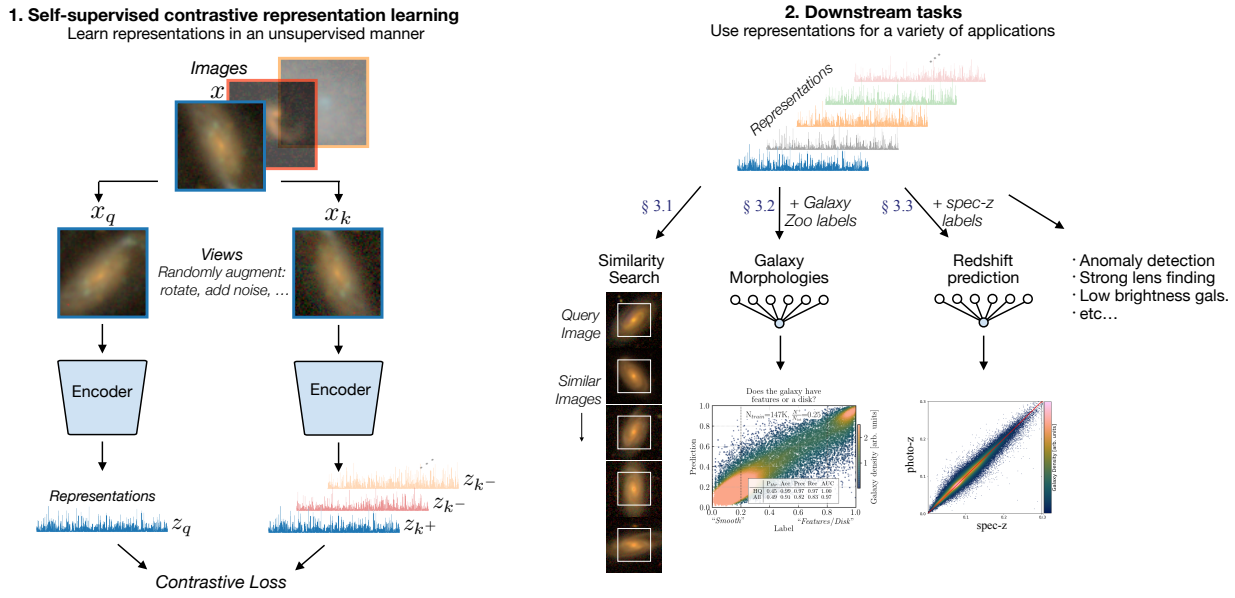

Figure 1. (Left) A schematic of the contrastive self-supervised framework. (Right) Examples of downstream tasks that can be implemented on the learned representations.

multiply it with the aggregate median absolute deviation (MAD) of each channel (pre-computed over all training examples) to get a fixed per-channel noise scale γ_c. Then, we introduce Gaussian noise sampled from $\mathcal{N}(0, \gamma_c)$ for each color channel.

The relative importance of these augmentations for producing good representations depends on both the dataset and the implementation of each augmentation. We evaluate representation quality by fine-tuning our representations for the task of redshift estimation under limited data labels (see Appendix D for details), finding Gaussian noise to be our strongest data augmentation and PSF the weakest, likely because pooling layers in our convolutional neural networks (CNNs) are robust to small-scale smearing. Best quality is achieved when we apply all augmentations except PSF. Note that these findings will not necessarily generalize to other surveys with different resolutions, signal-to-noise ratios, or target objects. This base set of image augmentations was chosen to remain as task-agnostic as possible, and additional augmentations could be added to target specific applications. For example, in tasks where the angular extent of a galaxy is irrelevant, an augmentation to change the apparent galaxy size (via image rescaling/interpolation) would be useful.

A schematic of the self-supervised pre-training framework used is shown in Figure 1 (Left). Applying our augmentations to samples $\mathbf{x}$, we get a pair of views that are denoted "positive" ($\mathbf{x_q}$, $\mathbf{x_{k^+}}$) when the two come from different transformations of the same image, and "negative" ($\mathbf{x_q}$, $\mathbf{x_{k^-}}$) otherwise. For each of the views, an encoder network extracts a 2048 dimensional representation $\mathbf{z} = \text{encoder}(\mathbf{x})$, and is trained to make positive pairs have similar representations while making negative pairs have dissimilar representations via a contrastive loss function:

$$L_{q,k^+,\{k^-\}} = -\log\left(\frac{\exp(\text{sim}(\mathbf{z_q}, \mathbf{z_{k^+}}))}{\exp(\text{sim}(\mathbf{z_q}, \mathbf{z_{k^+}})) + \sum_{\mathbf{k^-}} \exp(\text{sim}(\mathbf{z_q}, \mathbf{z_{k^-}}))}\right), \quad (1)$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a} \cdot \mathbf{b}/(\tau \|\mathbf{a}\| \, \|\mathbf{b}\|)$ is the cosine similarity measure between vectors $\mathbf{a}$ and $\mathbf{b}$, normalized by a tunable "temperature" hyper-parameter τ. This loss (InfoNCE, Oord et al. 2018) is minimized when positive pairs have high similarity, while negative pairs have low similarity. We have closely followed Chen, X. et al. (2020) in our self-supervised learning setup, and more implementation details are given in Appendix E.

3. RESULTS

We first visualize how the model has organized the image representation space, and explore how morphological characteristics from the Galaxy Zoo 2 project (GZ2, Willett et al. 2013) and spectroscopic redshifts from SDSS map onto this representation space. Then, using the labels from these two sources, we evaluate the utility of our self-supervised representations in actually performing the downstream tasks of morphology classification and photometric redshift estimation.

3.1. *Self-supervised learning visualization*

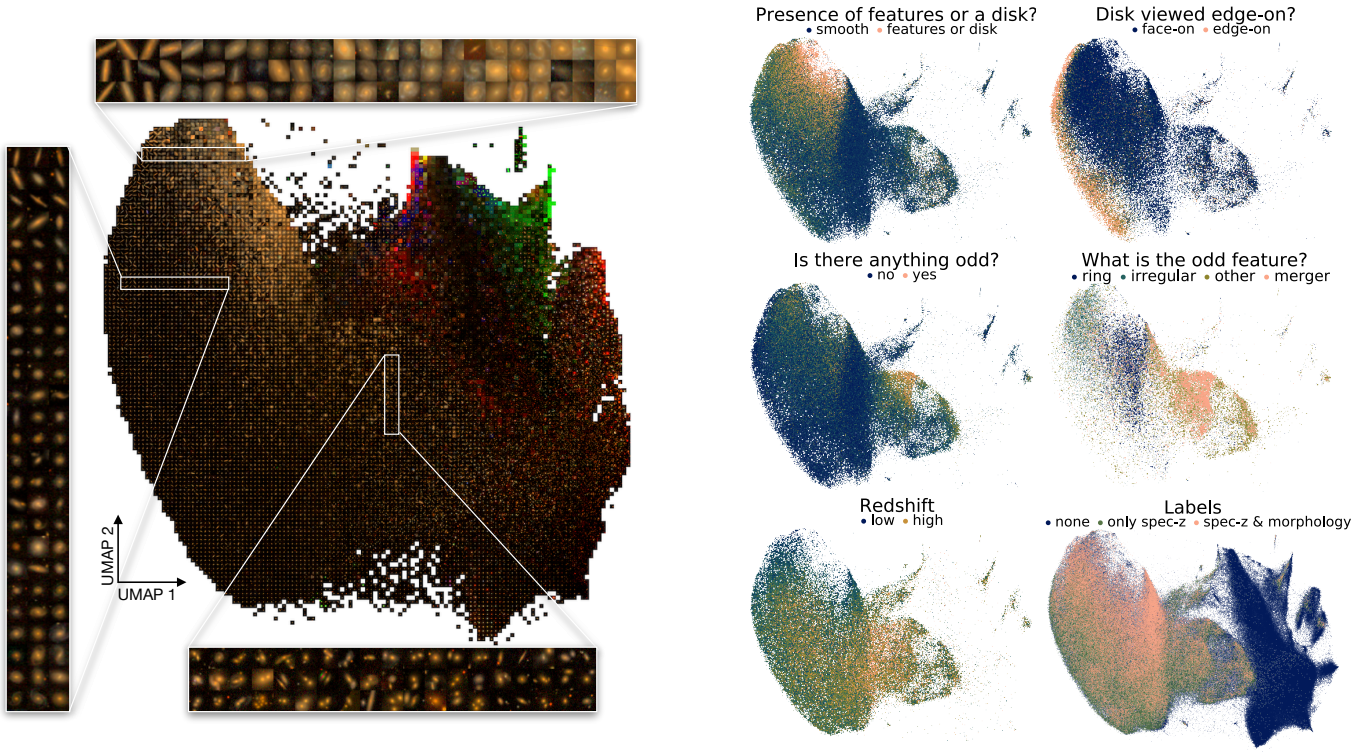

Figure 2. Visualizing the two dimensional UMAP projection of the self-supervised representations. The left panel shows randomly sampled representative images at each point in the space, while the right colors the space using answers to morphological classification questions from Galaxy Zoo 2, SDSS spectroscopic redshifts, or by labels.

To visualize the information contained in the self-supervised representations we use Uniform Manifold Approximation and Projection (UMAP, McInnes et al. 2018) to reduce the 2048 dimensional representations to a more manageable 2, while preserving structure information on both local and global scales. We want to emphasize that although UMAP can produce meaningful clusters when trained directly on image data, we are using it here only for visualization purposes of the representation space. The fact that the morphological classification tasks described in the next section can achieve a high performance through only a linear transformation of the representations, with no fine-tuning, means that the galaxies are organized in a semantically meaningful way in the representation space.

In Figure 2 we investigate this 2D projection. The left panel was created by binning the space into 128×128 cells, randomly selecting a sample that resides within each cell, and plotting its corresponding rgb mapped galaxy image at that location[3]. Around the edges we show zoom-ins to a variety of hand-selected areas, in which it is clear that images are grouped by their visual similarity (e.g., spiral or not, edge-on or not, etc).

The following six panels color each point using the redshift and morphology labels, and confirm that clustering is not only along visual characteristics. Distinct clusters as a function of morphological type and redshift are immediately apparent, to the level where decision boundaries for a number of GZ2 questions can be drawn by eye. Morphological labels are uncertain, so we illustrate them as continuous colors representing the fraction of votes for one answer over the other. Interestingly, we see that a large number of unlabeled samples are separated from any that have either redshift or morphology labels, but as we show below, using them for self-supervised learning still proves beneficial for the downstream tasks.

Appendix B further examines this 2D space in context of galaxy size and magnitude, and shows the advantage over the equivalent UMAP representations derived instead directly from the pixel space. We also demonstrate how a sample of galaxies under simple augmentations move drastically through this plane when the UMAP is derived from the images, but remain stationary when using the self-supervised representations.

When displayed through an interactive data portal (e.g. Reis et al. 2021), such visualizations built upon

[3] rgb images are obtained by 'luptonizing' (Lupton et al. 2004) the *gri* photometric bands

self-supervised representations can be invaluable to the broader astronomical community.

3.2. *Galaxy morphology classification*

Morphological classification of galaxies into subclasses based on the presence of visual characteristics such as spiral arms, central bars, or odd features, is key in order to study galaxy formation and evolution. The Galaxy Zoo project (Lintott et al. 2008) has been fundamental in this endeavour by crowd-sourcing morphological classifications for a significant number of galaxies. GZ2 (Willett et al. 2013), the successor to GZ, is focused on more fine-grained features, and in total achieved morphological classifications of 304,122 SDSS galaxies. Shown most prominently by the winners of the "Galaxy Challenge" (Dieleman et al. 2015) and numerous subsequent works since (Domínguez Sánchez et al. 2018, 2019; Khan et al. 2019; Walmsley et al. 2020; Spindler et al. 2020; Vega-Ferrero et al. 2020), CNNs excel at this task.

Here, by treating each question as a separate binary classification task, we predict answers to the subset of GZ2 questions that are most commonly undertaken by ML methods. We train three separate classifiers. The first is a CNN trained from scratch in a supervised setting with the same architecture of the encoder, the second is a linear classifier applied *directly on the self-supervised representations*, and for the third we fine-tune the self-supervised encoder for a few epochs. We note that the linear classifier requires only $\sim 0.5 - 10$ seconds on a GPU to train depending on the number of training samples used, while the fully supervised training takes up to 2 hours on 8 GPUs.

Figure 3 demonstrates the quality of the morphological predictions for the first GZ2 question, and shows the predicted label against the true label for the three classifiers as a function of the number of labels used for training. We quantify the accuracy, precision, recall, area under the receiver operator characteristic curve (AUC), and the outlier percentage η in the inlaid tables (see Appendix C for definitions: 1.00 is the ideal value for the first four, and 0 is ideal for η). Our results should be viewed most closely in relation to Domínguez Sánchez et al. (2018, DS+18) and Walmsley et al. (2020, W+20), as both used SDSS images and focused on GZ2 questions. The performance metrics shown are calculated on "high quality" labels with $P > 0.80$ or $P < 0.2$, although the networks were trained using all labels for a given GZ2 question with at least 5 votes regardless of the vote fraction for one answer over the other.

We find that both the linear classifier from the representations and the fine-tuned self-supervised model far outperform the supervised network when training with a limited number of labels, and are both able to achieve accurate classifications even in the regime where the supervised network fails to converge. Comparing the supervised network to the fine-tuned self-supervised for this GZ question, we find that roughly a factor of 16 more labels are required in the supervised setting to achieve the same performance as fine-tuned. When increasing to 65k labels we find that the supervised and fine-tuned networks are approaching optimal classification performance on this dataset given the high level of label uncertainty introduced by ambiguous class boundaries and crowd-sourced labeling. We note that an exact quantitative comparison of our performance metrics to DS+18 and W+20 are not possible due to a lack of consistency in data sets. In Appendix C we present additional discussion, methods, and morphological results on other GZ2 questions.

These results demonstrate that the self-supervised representations are extremely valuable for morphological classification. (1) They are essential to make accurate predictions when restricted by the number of available labels; (2) they improve accuracy metrics beyond what was achieved by pure supervised learning in the non-optimal classification performance regime; (3) they provide avenues to investigate and isolate imaging artifacts and anomalies as shown in the Appendix C; (4) they can be used in pipelines to speed up crowd-sourced classification tasks: determine the next galaxy to be classified, perform a very computationally inexpensive similarity search[4] to find N other similar galaxies, and classify them all at once; and (5) they can reduce the barrier to entry when analysing survey data by achieving high classification accuracy through linear classifications on the representations which requires minimal machine learning experience and compute resources.

3.3. *Photometric Redshift Estimation*

Determining the redshifts, and hence distances, to the billions of galaxies imaged in cosmological surveys is crucial to studying large-scale-structure, but taking high-precision spectroscopic redshift (spec-z) measurements for each galaxy is infeasible. Thus, a task of great importance to sky surveys is photometric redshift (photo-z) estimation (for a recent review, see Salvato et al. 2019). Photo-z models take multi-band galaxy photometry and use template fitting (Loh & Spillar 1986), machine learning (Connolly et al. 1995), or hybrid methods to produce estimates for the galaxy's redshift z and its associated probability density function. Traditionally, such models

[4] we use faiss: https://github.com/facebookresearch/faiss

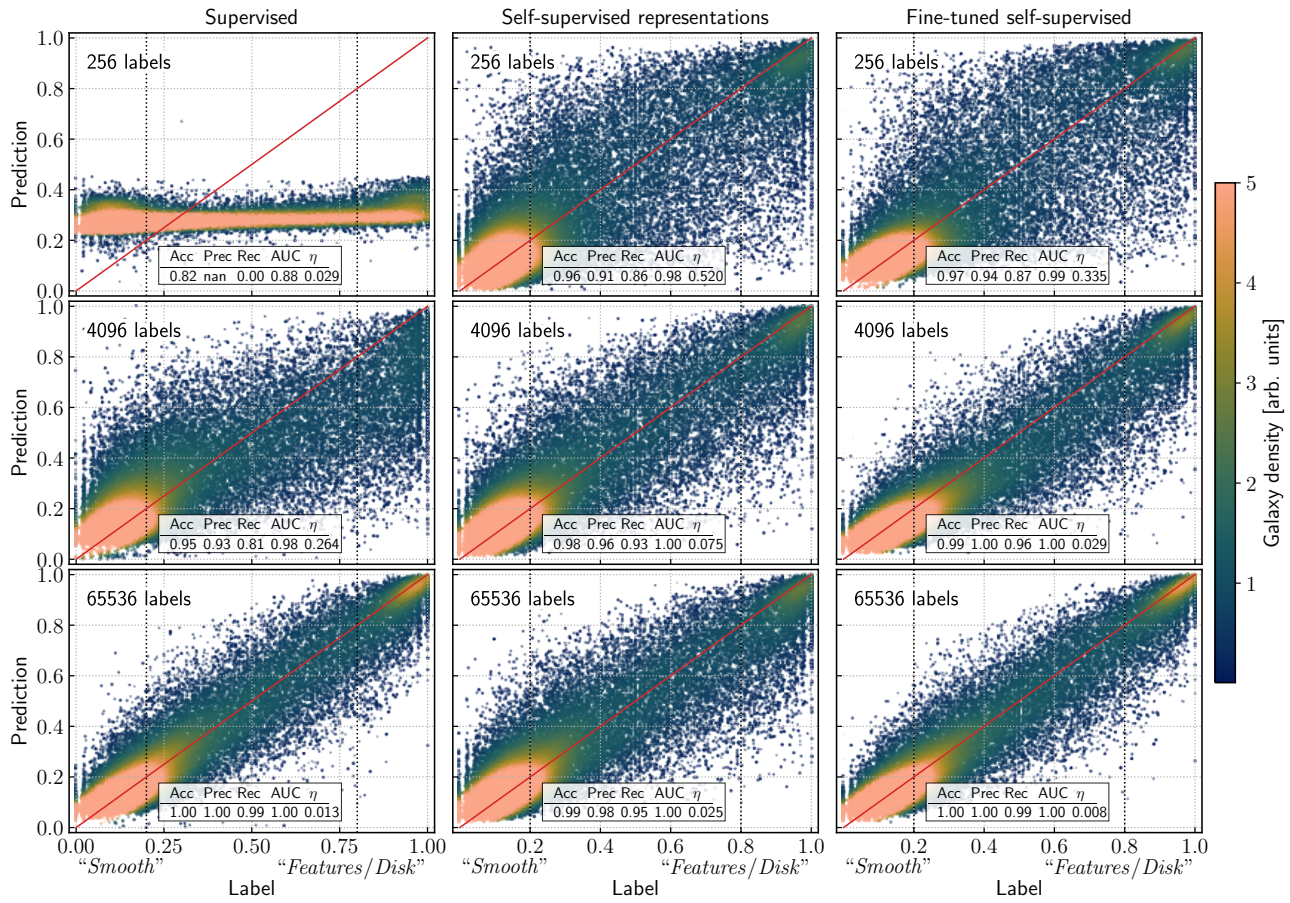

Figure 3. Predicted labels compared to crowd-sourced answers for the first GZ2 question: "Is the galaxy simply smooth and rounded, with no sign of a disk?". The three columns show the classification performance in a supervised setting (left), a linear classifier *directly on the self-supervised representations* (center), and when fine-tuning the self-supervised encoder for a few epochs (right). We report the accuracy, precision, recall, AUC, and outlier percentage η for "high quality" labels (those with $P > 0.80$ or $P < 0.2$). The size and opacity of the points is proportional to the number of crowd-sourced labels they received.

relied on "hand-crafted" features extracted from the observations, but recent work has also successfully applied CNNs directly to the images themselves (Hoyle 2016; D'Isanto & Polsterer 2018; Pasquet et al. 2019). While promising, these supervised methods are inherently constrained by the limited size (and galaxy features, such as relative size or brightness) of the training dataset. Therefore, self-supervised representations derived from a larger body of unlabeled samples are a promising venue for improving accuracy and robustness.

For easier comparison against established baselines on SDSS data, we closely follow the setup of Pasquet et al. (2019), whose CNN achieved significantly lower dispersion than previous image-based ML models. Their network is trained as a classifier over a discrete set of 180 redshift bins spanning $0 \leq z \leq 0.4$, where the photo-z estimate z_p is computed as the expectation $\mathbb{E}(z)$ over the probabilities predicted in each bin. We adapt this design into our ResNet50 model, and establish our own supervised baseline (with identical architecture as our self-supervised model) by training on the available spec-z labels. We use the following standard metrics to evaluate the accuracy of photo-z estimates:

- The prediction residual $\Delta z = (z_p - z_s)/(1 + z_s)$, where z_p and z_s correspond to the photometric and spectroscopic redshifts, respectively.
- The dispersion or MAD deviation, $\sigma_{\rm MAD} = 1.4826 \times {\rm MAD}(\Delta {\rm z})$, where ${\rm MAD} = {\rm median}(|\Delta z - {\rm median}(\Delta z)|)$.
- η, the percent of "catastrophic" outliers with $|\Delta z| > 0.05$.

Results of our supervised training study are shown in the three left panels of Figure 4. As shown, we also test the improvement in the model accuracy as we increase the training dataset size. Similar to the results of Pasquet et al. (2019), the prediction bias $\langle \Delta z \rangle$ (a noise-

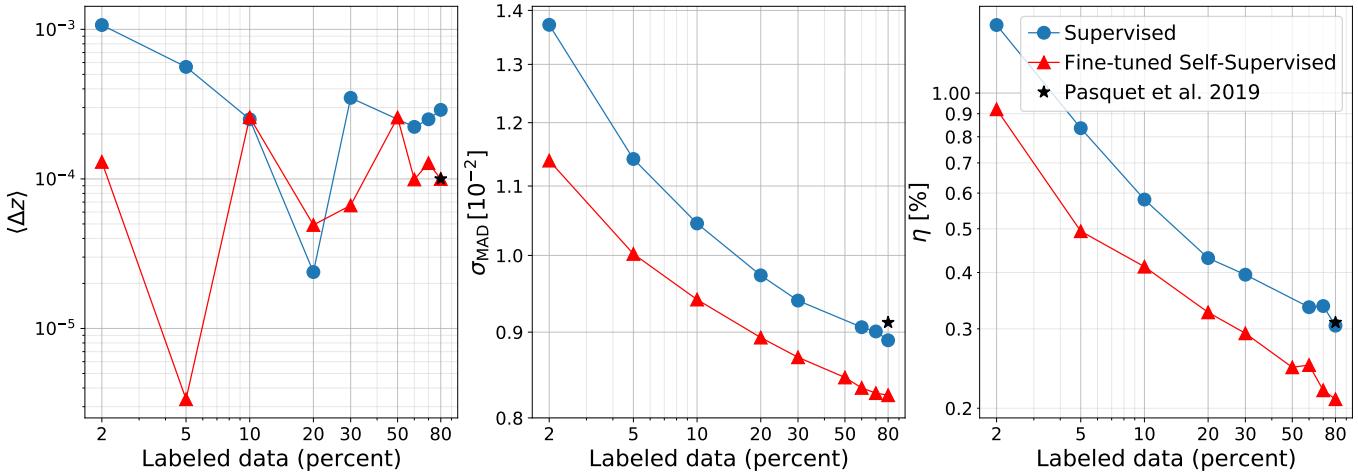

Figure 4. The prediction bias $\langle \Delta z \rangle$, dispersion $\sigma_{\rm MAD}$, and outlier percentage η of photo-z estimates on test data, from our fine-tuned representations compared to reference fully-supervised networks. Trained on increasing fractions of the spec-z dataset, the fine-tuned self-supervised models perform best.

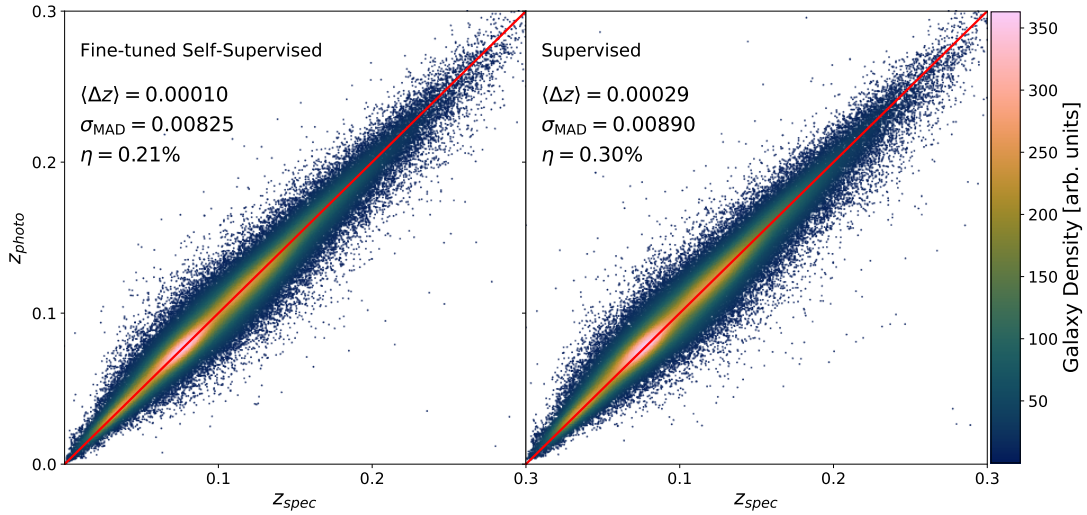

Figure 5. The photo-z estimates of our best model on test data, arising from fine-tuned self-supervised model (left), compared to the equivalent supervised network (right).

dominated metric) is negligibly small, and our supervised baseline model matches the accuracy of Pasquet et al. (2019) in $\sigma_{\rm MAD}$ and η.

We then move to evaluating the utility of our self-supervised representations for photo-z estimation. To fine-tune them on the spec-z labels, we train a linear classifier on the representations while allowing the encoder weights to train with a $10\times$ smaller learning rate than that of the classifier (more fine-tuning details are in Appendix E). Results are shown in Figure 4.

At all fractions of training data used, the fine-tuned self-supervised representations achieve superior performance in $\sigma_{\rm MAD}$ and η compared to the equivalent supervised network. Impressively, the accuracy gained from pre-training on unlabeled data is equivalent to supervised training on $2-4\times$ more spec-z labels, with no modifications in architecture size or complexity. Consequently, our model, fine-tuned on the full training dataset, achieves a new state-of-the-art accuracy for CNN-based photo-z prediction on SDSS galaxies, as shown in Figure 5. This is an exciting result, as it suggests any existing supervised network developed for similar tasks on this type of data could get an immediate performance benefit from a self-supervised pre-training stage.

4. DISCUSSION AND CONCLUSIONS

In this letter we have demonstrated that self-supervised representation learning on unlabeled data yields notable performance gains over supervised learn-

ing for multiple tasks. These performance gains are achieved even when the self-supervised model is limited to have the same size as the baseline CNNs in downstream tasks. However, results from ML literature show that the best performance (i.e., the best representation quality) is achieved when self-supervised models are much larger (Radford et al. 2019; Chen, T. et al. 2020b). Thus, the possibility of training a large self-supervised model on massive photometry databases and "serving" the model for usage by the larger community, much like the operation of existing state-of-the-art language models (Devlin et al. 2018; Radford et al. 2019), is an exciting new direction for ML applications in sky-surveys.

A major issue with all machine learning studies on labeled sky survey data is not just the limited size of the training data set, but also the selection bias imposed by gathering labels. For example, due to the series of flux and quality cuts applied when selecting spectroscopic targets (Strauss et al. 2002), galaxies with spec-z labels have a distinct bias towards nearby bright objects with low galactic extinction. Thus, galaxies selected for spectroscopic measurement are not representative of all those with photometry, as can clearly be seen in the representation space of Figure 2. Although ML methods can train on this labeled data and achieve a good test accuracy within this subset of galaxies, there are few robustness guarantees for photo-z estimation beyond the labeled distribution of galaxies. It has been shown, recently, that self-supervised pre-training (Hendrycks et al. 2019a,b) improves model robustness and uncertainty quantification, and that self-supervised models outperform their supervised counterparts in out-of-distribution detection on difficult outliers. This means self-supervised models have excellent prospects for mitigating distributional differences between (labeled) training and (unlabeled) inference data in sky surveys. We believe that self-supervised representation learning opens the door to leveraging the vast amounts of unlabeled, existing and future, sky survey data, promising a new era for ML applications in precision and discovery astrophysics.

ACKNOWLEDGMENTS

Authors would like to thank François Lanusse, Peter Melchoir, Evan Racah, and Edward Schlafly for helpful discussions. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231. Md.H.'s work was supported by the NERSC's summer internship program. G.S. and Z.L. were partially supported by the DOE's Office of Advanced Scientific Computing Research and Office of High Energy Physics through the Scientific Discovery through Advanced Computing (SciDAC) program.

REFERENCES

Alam, S., Albareti, F. D., Prieto, C. A., et al. 2015, The Astrophysical Journal Supplement Series, 219, 12

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. 2020, A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science (Springer International Publishing), 3–21, doi: 10.1007/978-3-030-22475-2_1

Bachman, P., Hjelm, R. D., & Buchwalter, W. 2019, in Advances in Neural Information Processing Systems, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett, Vol. 32 (Curran Associates, Inc.), 15535–15545. https://proceedings.neurips.cc/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, MNRAS, 460, 1371, doi: 10.1093/mnras/stw1009

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. 2018, arXiv e-prints, arXiv:1807.05520. https://arxiv.org/abs/1807.05520

Cheng, T.-Y., Huertas-Company, M., Conselice, C. J., et al. 2020a, arXiv e-prints, arXiv:2009.11932. https://arxiv.org/abs/2009.11932

Cheng, T.-Y., Li, N., Conselice, C. J., et al. 2020b, MNRAS, 494, 3750, doi: 10.1093/mnras/staa1015

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020a, arXiv preprint arXiv:2002.05709

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. 2020b, arXiv preprint arXiv:2006.10029

Chen, X., Fan, H., Girshick, R., & He, K. 2020, arXiv preprint arXiv:2003.04297

Connolly, A. J., Csabai, I., Szalay, A. S., et al. 1995, Astronomical Journal, 110, 2655, doi: 10.1086/117720

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018, arXiv e-prints, arXiv:1810.04805. https://arxiv.org/abs/1810.04805

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2019, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Minneapolis, Minnesota: Association for Computational Linguistics), 4171–4186, doi: 10.18653/v1/N19-1423

Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441, doi: 10.1093/mnras/stv632

D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111, doi: 10.1051/0004-6361/201731326

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, MNRAS, 476, 3661, doi: 10.1093/mnras/sty338

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2019, MNRAS, 484, 93, doi: 10.1093/mnras/sty3497

Fruchter, A. S., & Hook, R. N. 2002, PASP, 114, 144, doi: 10.1086/338393

Goyal, P., Mahajan, D., Gupta, A., & Misra, I. 2019, in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 6390–6399, doi: 10.1109/ICCV.2019.00649

Gunn, J. E., Carr, M., Rockosi, C., et al. 1998, AJ, 116, 3040, doi: 10.1086/300645

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, AJ, 131, 2332, doi: 10.1086/500975

Hadsell, R., Chopra, S., & LeCun, Y. 2006, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, 1735–1742, doi: 10.1109/CVPR.2006.100

Hart, R. E., Bamford, S. P., Willett, K. W., et al. 2016, Monthly Notices of the Royal Astronomical Society, 461, 3663–3682, doi: 10.1093/mnras/stw1588

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. 2020, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9729–9738

He, K., Zhang, X., Ren, S., & Sun, J. 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778

Hendrycks, D., Lee, K., & Mazeika, M. 2019a, arXiv e-prints, arXiv:1901.09960. https://arxiv.org/abs/1901.09960

Hendrycks, D., Mazeika, M., Kadavath, S., & Song, D. 2019b, arXiv e-prints, arXiv:1906.12340. https://arxiv.org/abs/1906.12340

Hocking, A., Geach, J. E., Sun, Y., & Davey, N. 2018, MNRAS, 473, 1108, doi: 10.1093/mnras/stx2351

Hoyle, B. 2016, Astronomy and Computing, 16, 34, doi: 10.1016/j.ascom.2016.03.006

Ivezić, Ž., Connelly, A. J., Vanderplas, J. T., & Gray, A. 2019a, Statistics, Data Mining, and Machine Learning in Astronomy (Princeton University Press)

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019b, ApJ, 873, 111, doi: 10.3847/1538-4357/ab042c

Keel, W. C., Chojnowski, S. D., Bennert, V. N., et al. 2012, MNRAS, 420, 878, doi: 10.1111/j.1365-2966.2011.20101.x

Khan, A., Huerta, E. A., Wang, S., et al. 2019, Physics Letters B, 795, 248, doi: 10.1016/j.physletb.2019.06.009

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193. https://arxiv.org/abs/1110.3193

Lintott, C., Masters, K., Simmons, B., Bamford, S., & Kaviraj, S. 2013, Astronomy and Geophysics, 54, 5.16, doi: 10.1093/astrogeo/att162

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179, doi: 10.1111/j.1365-2966.2008.13689.x

Lintott, C. J., Schawinski, K., Keel, W., et al. 2009, MNRAS, 399, 129, doi: 10.1111/j.1365-2966.2009.15299.x

Loh, E. D., & Spillar, E. J. 1986, The Astrophysical Journal, 303, 154, doi: 10.1086/164062

Lupton, R., Blanton, M. R., Fekete, G., et al. 2004, PASP, 116, 133, doi: 10.1086/382245

Margalef-Bentabol, B., Huertas-Company, M., Charnock, T., et al. 2020, MNRAS, 496, 2346, doi: 10.1093/mnras/staa1647

Martin, G., Kaviraj, S., Hocking, A., Read, S. C., & Geach, J. E. 2020, MNRAS, 491, 1408, doi: 10.1093/mnras/stz3006

McInnes, L., Healy, J., & Melville, J. 2018, arXiv e-prints, arXiv:1802.03426. https://arxiv.org/abs/1802.03426

Nayak, P. 2019, "Understanding searches better than ever before". "https://blog.google/products/search/search-language-understanding-bert/"

Oord, A. v. d., Li, Y., & Vinyals, O. 2018, arXiv preprint arXiv:1807.03748

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, Astronomy & Astrophysics, 621, A26

Paszke, A., Gross, S., Massa, F., et al. 2019, in Advances in Neural Information Processing Systems 32, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Curran Associates, Inc.), 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-libr pdf

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. 2018, Improving language understanding by generative pre-training

Radford, A., Wu, J., Child, R., et al. 2019, Language Models are Unsupervised Multitask Learners

Reis, I., Rotman, M., Poznanski, D., Prochaska, J., & Wolf, L. 2021, Astronomy and Computing, 34, 100437, doi: https://doi.org/10.1016/j.ascom.2020.100437

Salvato, M., Ilbert, O., & Hoyle, B. 2019, Nature Astronomy, 3, 212

Schlafly, E. F., & Finkbeiner, D. P. 2011, The Astrophysical Journal, 737, 103

Spergel, D., Gehrels, N., Breckinridge, J., et al. 2013, arXiv e-prints, arXiv:1305.5422. https://arxiv.org/abs/1305.5422

Spindler, A., Geach, J. E., & Smith, M. J. 2020, MNRAS, doi: 10.1093/mnras/staa3670

Stein, G. 2020, georgestein/ml-in-cosmology: Machine learning in cosmology, v1.0, Zenodo, doi: 10.5281/zenodo.4024768

Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810, doi: 10.1086/342343

Tian, Y., Sun, C., Poole, B., et al. 2020, arXiv e-prints, arXiv:2005.10243. https://arxiv.org/abs/2005.10243

Vega-Ferrero, J., Domínguez Sánchez, H., Bernardi, M., et al. 2020, arXiv e-prints, arXiv:2012.07858. https://arxiv.org/abs/2012.07858

Walmsley, M., Smith, L., Lintott, C., et al. 2020, MNRAS, 491, 1554, doi: 10.1093/mnras/stz2816

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835, doi: 10.1093/mnras/stt1458

Xiao, T., Wang, X., Efros, A. A., & Darrell, T. 2020, arXiv e-prints, arXiv:2008.05659. https://arxiv.org/abs/2008.05659

Xin, B., Ivezić, Ž., Lupton, R. H., et al. 2018, AJ, 156, 222, doi: 10.3847/1538-3881/aae316

Xiong, L., Poczos, B., Connolly, A., & Schneider, J. 2018, Anomaly Detection for Astronomical Data, Carnegie Mellon University, doi: https://doi.org/10.1184/R1/6475475.v1

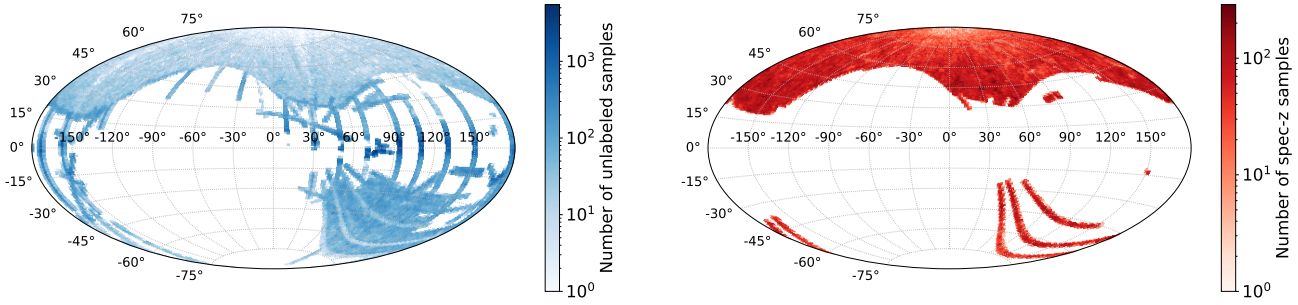

Figure 6. The spatial distribution of the full unlabeled (left) galaxy image dataset, and the spec-z labeled (right) dataset.

APPENDIX

A. DATASET DETAILS

Our database of galaxies is assembled from Data Release 12 (DR12; Alam et al. 2015) of the SDSS. To pull samples with spectroscopic redshift labels, we follow the process of Pasquet et al. (2019) in pulling from the Main Galaxy Sample to enable direct comparison to their results. Their SQL query filters for objects classified as `'GALAXY'` with dereddened petrosian magnitudes $r \leq 17.8$ and spectroscopic redshifts $z \leq 0.4$. For us, the query returns 547,224 objects, and after removing some duplicates, we are left with 517,190 to use as labeled training examples. When fine-tuning our image representations for the photo-z estimation task, we use 400,000 images for training and 103,000 as validation dataset.

To build our larger set of galaxies with no spectroscopic labels, we filter for objects with dereddened petrosian magnitudes $r \leq 17.8$, on the `'PhotoObjAll'` full photometric catalog of the SDSS. In the resulting set of galaxies, we remove duplicates which were already included in our spectroscopic sample, and exclude samples with an estimated photometric redshift (as estimated by Beck et al. (2016)) $z_{phot} > 0.8$. This eliminates objects which are too distant compared to the spectroscopic sample, but decreases the possibility that we are unnecessarily excluding samples whose true redshift is less than 0.4 (the cutoff for our spectroscopic sample) due to incorrect photo-z estimates. After imposing these cuts, we were able to successfully pull 845,254 unlabeled images. The spatial distributions of our labeled and unlabeled galaxy datasets are shown in Figure 6

SDSS photometric images contain data in 5 passbands ($ugriz$), and come background-subtracted but are not dereddened to account for galactic extinction. To pull images for our datasets, we use the Montage[5] tool to query the imagery catalog in SDSS Data Release 9 (DR9), based on the tabulated equatorial coordinates for each object in our dataset. For each set of object coordinates, we sample a patch of sky of size $(0.012°)^2$, centered on the object, and project onto a 2D image with 107^2 pixels (this ensures the resulting pixel scale is as close as possible to the native pixel scale in the SDSS, 0.396 arcsec). In each image, we store the u, g, r, i, and z passbands as 5 color channels.

The Montage pipeline uses Variable-Pixel Linear Reconstruction (Fruchter & Hook 2002) during the projection process to appropriately transform source input pixel values into the output pixel space. With this setup, a few samples were returned containing corrupted values at the edges of the image, so we crop all images to 107^2 pixels to eliminate such issues. Note that during training of the self-supervised model, we impose random rotations and random jitter to each image before cropping out the central portion as a data augmentation, so while our photometric images contain 107 pixels per side, the CNNs in this work only view samples of size 64^2 pixels. This input size of CNN is consistent with the photo-z CNN model of Pasquet et al. (2019).

B. SIMILARITY SEARCH AND UMAP

Section 3.1 visualized the information contained within the self-supervised representations derived from the SDSS dataset in context of a subset of the available labels. We found that the representations were organized to a high-

[5] http://montage.ipac.caltech.edu/

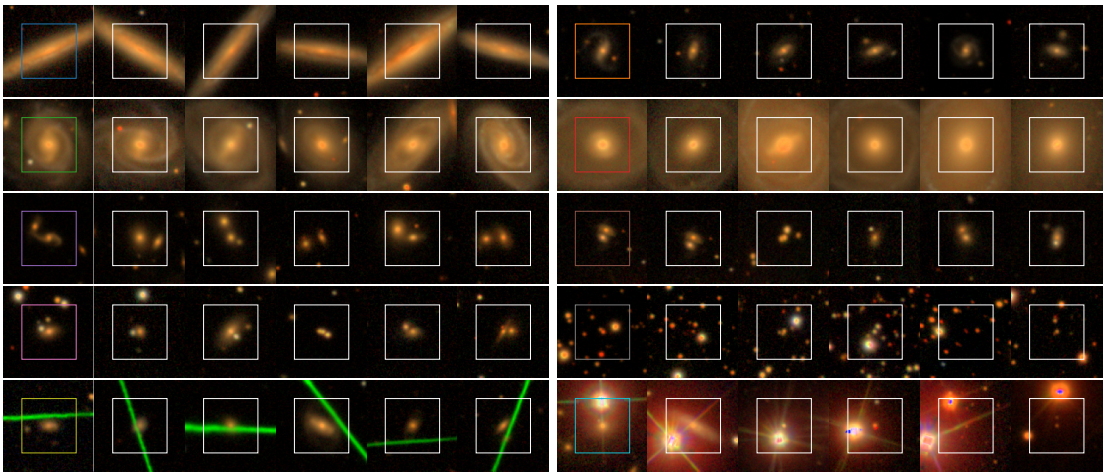

Figure 7. Reference SDSS galaxies (leftmost panels) and the most similar galaxies (following 5 panels) identified through a self-supervised similarity search. Squares outline the 64^2 pixels that are "seen" by the network.

degree by the semantic information contained within the images. Here we further show that the UMAP performed on the representations surpasses the utility of a UMAP performed directly on the images, and demonstrate how the self-supervised representations allow for simple and effective similarity searches.

B.1. *Similarity search*

As previously mentioned, the self-supervised representations provide a venue to perform similarity searches. The contrastive framework that the encoder was trained under worked to encode images that are semantically similar to nearby points in the representation space. Therefore, given any desired query image, finding semantically similar images that may exist in the dataset corresponds to finding nearby data points in the 2048 dimensional representation space. This can be achieved by taking the cosine similarity (i.e. the normalized dot product) of the query vector and all other representations, and sorting by decreasing value. Alternative similarity measures to the cosine similarity can also be used, but were not studied here.

Figure 7 demonstrates a similarity search performed on ten different examples of SDSS images. The images returned by this simple similarity search, in a completely unsupervised fashion, visually appear extremely similar, and are seemingly agnostic to rotations and jitter, as desired.

B.2. *Image-based UMAP*

Similar to the UMAP analysis on the self-supervised representations, UMAP can also be applied directly in pixel space by flattening the 5 band images into a vector. Due to computational difficulties with such large vectors we randomly sampled 5% of the galaxies in the total dataset to determine the UMAP transformation, then used the transformation on the entire set of $\sim$1.3 million flattened images (labeled and unlabeled).

Figure 8 shows the equivalent of Figure 2, but directly derived from the images. Here we only show galaxies that have labels. In the left panel we find that the images have been separated by the size and brightness of the galaxy, with bright and large galaxies residing in the top-right of the two dimensional space and dim small galaxies along the left. In the right panels we color each point by a subset of available labels and find that the first morphological classification - "presence of features or a disk" - shows a level of separation between the classes, while more detailed morphological features do not. Separation is also seen by redshift and label type, which are both strongly correlated with galaxy size and magnitude.

While an unsupervised dimensionality reduction technique such as UMAP performed on the images can provide some level of clustering based on the semantic information within the images, it by definition does not attempt to map semantically similar images to similar points in the compressed space. To explicitly illustrate this undesirable outcome of an unsupervised clustering method we select a sample of galaxies and apply three sets of augmentations to each. The first rotates each image between 0 and 360 degrees in 45 degree increments, the second jitters each image from (-7,-7) to (7,7) pixels in 7 linearly spaced values and then crops the central 64^2 pixels, and the third adds Gaussian noise with a standard deviation ranging from 0 to 3 times the aggregate median absolute deviation measured over

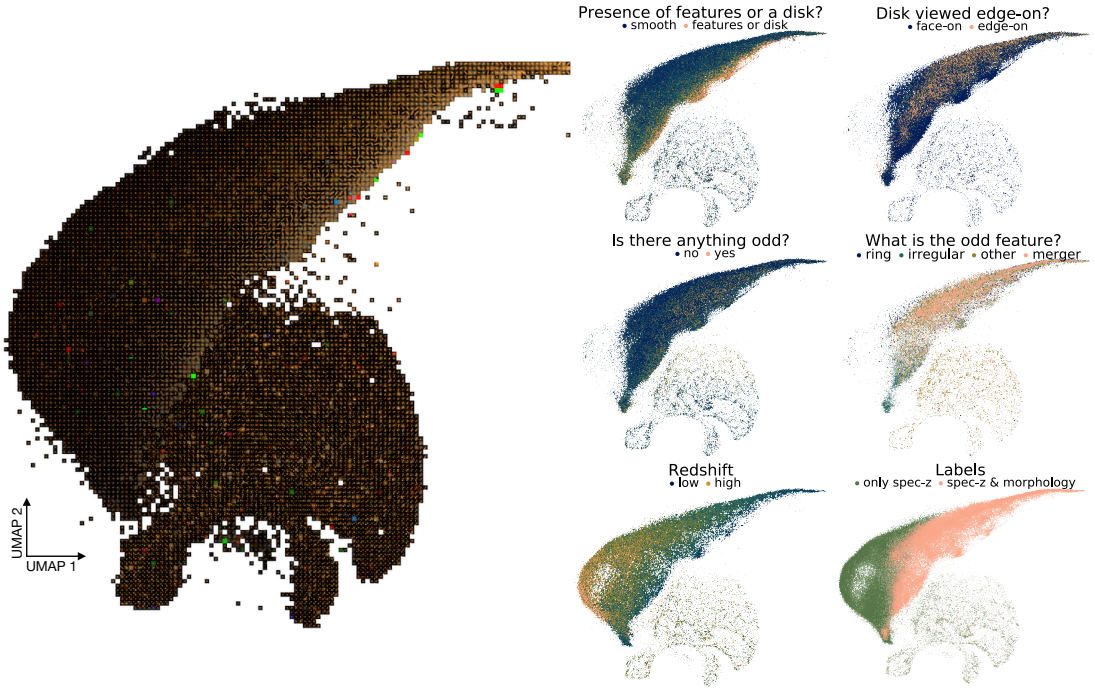

Figure 8. Visualizing the two dimensional UMAP projection of the image pixel space. The left panel shows randomly sampled representative images at each point in the space, while the right colors the space using answers to morphological classification questions from Galaxy Zoo 2, SDSS spectroscopic redshifts, or by labels.

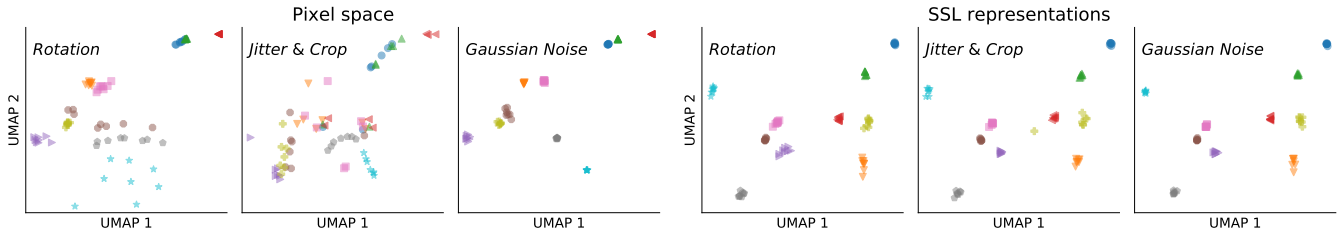

Figure 9. Stationarity of selected galaxies under data augmentations. Rotations were chosen uniformily between 0 and 360 deg, jitter from (-7,-7) to (7,7), and Gaussian noise with a standard deviation ranging from 0 to 3 times the aggregate median absolute deviation measured over the dataset. Colors match those of the left-most panels of Figure 7, as we used the same 10 galaxies.

the dataset. Both the pixel space and representation space UMAP transforms were trained on 5% of the dataset to facilitate a fair comparison.

Figure 9 shows how the 10 galaxies used in the similarity search figure move through the UMAP plane under three different types of data augmentation. The marker color is consistent with the color of the square around the query image in Figure 7. The left three panels are created from the pixel space, while the right three panels are from the self-supervised representations. For the pixel space UMAP we see that the simple augmentations of rotation and jitter/crop cause the test galaxies move through a large fraction of the two dimensional plane, even though the semantic qualities of the galaxy remain unchanged. For the self-supervised representations we instead find that they are nearly invariant to the augmentations, which is expected due to the contrastive learning framework used to train the encoder. Only one augmented instance of a single galaxy shows any significant movement in the plane - the yellow plus symbol under the jitter/crop augmentation. This is the image shown at the bottom left of Figure 7, which is contaminated by a large green streak across the top third of the image. Under a large negative jitter, this streak no longer appears in the 64^2 pixel image fed to the encoder, so the representation changes. This demonstrates the advantage of contrastive self-supervised representation learning over common unsupervised clustering techniques.

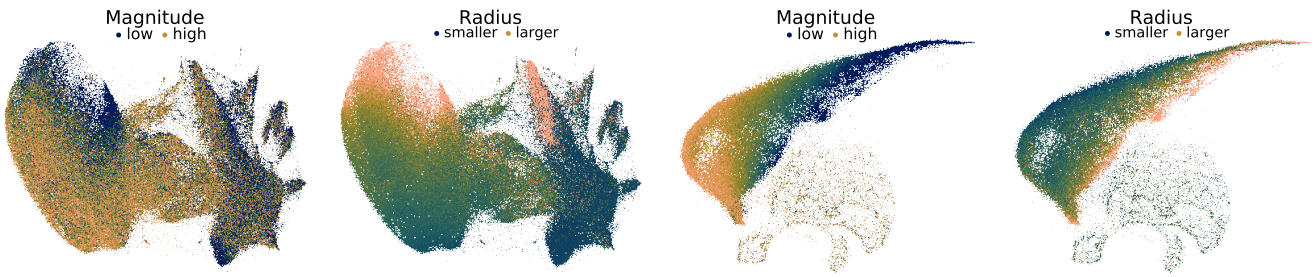

Figure 10. Representation (left) and image (right) UMAPs colored by r-band petrosian magnitude and radius.

B.3. *Additional Augmentations*

To remain as downstream task-agnostic as possible we only implemented the base set of augmentations described in Section 2. This has consequences on the information contained within the final self-supervised representations, and for specific targeted downstream applications additional augmentations may be useful. For example, morphological classifications should not depend on the size of a galaxy, and it is common to include a re-scaling augmentation when training a supervised classification network for this task. However, galaxy size is a useful indicator of redshift for nearby galaxies and is desirable in similarity search applications. Therefore task-specific augmentations beyond the base set can instead be added during the fine-tuning process.

Figure 10 colors each point in the UMAP space by the r-band petrosian magnitude and radius to shows that the representations respond to both the size and magnitude of the galaxy. This is expected as we did not include a size scaling augmentation, and no normalization was performed on the images to remove magnitude information. When comparing this figure to the previous UMAP visualizations it is apparent that there are strong correlations between the radius, magnitude, redshift, presence of features or a disk, and the available labels. When pushing the photometric redshift prediction to higher redshifts in future works, where the size and magnitude of an object are not as good of indicators of redshift, a further study on the utility of size and magnitude augmentations for this task will be required.

C. GALAXY MORPHOLOGY CLASSIFICATION

In Section 3.2 we displayed morphological predictions for the first Galaxy Zoo 2 (GZ2) questions. Here, we describe the training data and detail the training procedure for the linear classifier, denonstrate the results on two additional galaxy zoo questions, and put in further context the accuracy of the predictions with respect to previous methods.

C.1. *Training methodology*

Galaxy Zoo 2 (Willett et al. 2013) achieved 16 million morphological classifications of 304,122 galaxies drawn from SDSS with r < 17 and petroR90_r[6] > 3 arcsec, collected via a web-based interface. For each galaxy users were shown a 424×424 pixel image scaled to 0.02petroR90_r arcseconds per pixel, and were led down a multi-step decision tree to answer increasingly detailed questions about the visual appearance of the galaxy. See Figure 2 of (Hart et al. 2016) for a clear illustration of the decision tree. For example, users were first asked "Is the galaxy simply smooth and rounded, with no sign of a disk?", and had the option of selecting responses "smooth", "features or disk", or "star or artifact". Their selection of a response is referred to as their *vote*. Based on their response to a question they are lead down different branches of the decision tree - if "smooth" was selected, they are next asked "how rounded is it", but if "features or disk" was selected they will be asked a series of questions about spiral arms, bulge shapes, and a variety of other morphological features.

There are a total of 11 *classification tasks* (questions) that have the potential of being asked in the GZ2 decision tree, and the number of possible responses for each classification ranges from 2 to 7. We focus on binary tasks, with the exception of the first question where the response "star or artifact" occurs so infrequently (0.08% of responses select this as the majority) that it can be effectively neglected. All tasks beyond the first one depend on responses to previous tasks in the decision tree. For example, "could this be a disk viewed edge on?" is only asked if the user responded "features or disk" on the first task. Thus, a response for "edge-on" over "face-on" is not a binary classification of the

[6] petroR90_r is the Petrosian radius which contains 90% of the r-band flux

total galaxy population, but only a sub-classification of galaxies that were already considered to have features or a disk.

The nature of data collection (non-expert labelled), coupled with the uncertain class boundaries for galaxies with faint features, result in individual GZ2 users voting for different answers, and therefore uncertain morphological classifications are given for each galaxy. We consider the "consensus weighted vote fractions" - the fraction of users who voted for an answer[7] - as the true probability of a galaxy belonging to one class over the other, and we predict an equivalent class probability between (0, 1). Rather than a binary prediction, the returned probability represents the uncertainty of the morphology of the galaxy as seen in an SDSS image, whether this uncertainty stems from faint features, mislabelling, or imaging artifacts. Other works use the "de-biased estimate", which estimates how the galaxy would have been classified if viewed at $z = 0.03$ (Hart et al. 2016). By using the consensus weighted fractions we estimate what the image actually shows, not the "true" morphology, and debiasing can be performed after prediction.

We use the GZ2 main sample with spectroscopic redshifts which includes morphological classifications for 243,500 galaxies. The main sample without photometric redshifts of 42,462 galaxies, and the stand-alone Stripe 82 catalogue of 17,787 galaxies, were not included. We cross match the GZ2 table[8] with our SDSS database and search for pairs whose equatorial coordinates overlap within within 5 pixels (1.98 arcsec). This returns a final sample of 183,929 galaxies for which we have both *ugriz* images and crowd-sourced morphological classifications. This sample also contains redshift information which was used separately for photo-z prediction. For each question we required a minimum of 5 votes for each galaxy, which results in questions that are less frequently answered having smaller number of labelled samples than the full 183,929.

C.2. *Network & training*

We predict the vote weighted user response using three different classifiers. The first is a CNN trained from scratch in a supervised setting with the same ResNet50 architecture of the encoder, the second is a linear classifier *directly on the self-supervised representations*, and for the third we fine-tune the self-supervised encoder for a few epochs. Classifcation is performed by the addition of a fully connected layer on the 2048 dimensional output of the ResNet50. This maps the 2048 dimensional representation to 1 dimension with 2048 trainable weights and one bias parameter, followed by a softmax function to ensure the predicted probability is within the range (0, 1). For each GZ2 question networks are trained separately using the subset of galaxies that have at least five total votes for that question. Thus, the predicted probability should be interpreted as follows: regardless of the vote fraction of responses to previous GZ2 questions, what is the consensus vote *of the users that ended up at this question on the decision tree*. If only 10% of users selected "features or a disk" for the first question, the galaxy most likely is smooth and does not have a disk. Yet we still use the vote fraction of that 10% of users that were then subsequently asked "could this be a disk viewed edge-on?" when training a classifier for the edge-on question. Classifiers can then be used in conjunction after training.

Training was conducted in PyTorch (Paszke et al. 2019) through a binary cross entropy loss on the soft labels (consensus weighted vote fractions). A random sample of 20% of the data was set aside for testing and was not used for training any of the classifiers, and 10% was set aside for validation. For the supervised and fine-tuned networks we augmented images at each training epoch with jitter/crop and random rotations. Many questions have a high degree of class imbalance which reflects the occurrence of galaxy morphologies in the nearby universe. We found that class-balanced class weights (each instance of the class weighted by the overall occurrence fraction of that class) did not improve the classification performance. We weighted each instance by the total number of votes it received, although this resulted in negligible performance differences.

The supervised network was trained on 8 NVIDIA Tesla V100 GPUs for 100 epochs with a batch size of 128 using the SGD optimizer with a learning rate of 0.01, which we reduced by a factor of 10 at 60 and 90 epochs. The fine tuned-network was trained similarly, but with the pre-trained weights having a learning rate 10x smaller than the linear classifier layer. Optimization for the linear layer directly on the representations was performed using Limited-memory BFGS (LBFGS) and a learning rate of 0.05, which was decreased by a factor of 10 at 10 and 25 epochs. For all networks we used the epoch that produced the highest accuracy on "high quality" labels in the validation set. Supervised ResNets and fine-tuning required between a few minutes and a few hours of training time on the 8 GPUs,

[7] consensus weighted fractions are slightly different than the true vote fraction, they are the result of re-weighting users votes based on their overall consensus with others who looked at the same image

[8] Galaxy Zoo data is located at https://data.galaxyzoo.org/

depending on the number of training samples. For each linear classifier on the representations training generally concluded within a few epochs and $\sim 0.5 - 60$ seconds of compute time on a GPU. Note that this does not take into account the compute time required to learn the self-supervised representations, which this is shared between all downstream tasks, and does not need to be undertaken for each one separately.

As evidenced by Figure 2, the self-supervised representations have achieved a high degree of separation between numerous types of galaxy morphologies. Likewise, we found that classification became a straightforward task when using any subset of the labels. Unlike DS+18, we found no need to separate uncertain labels from high probability ones. Limiting our training set to only 'high quality' (HQ) labels (those with $P < 0.2$ or $P > 0.8$), resulted in nearly equivalent performance on HQ labels in the test set, but significantly decreased the performance on uncertain labels (those with $0.2 < P < 0.8$). We also found no need to impose higher minimum cuts on the number of votes (10) for some questions, as W+20 found was needed to improve classification performance from random initialization.

C.3. *Performance metrics & additional results*

Figure 3 demonstrated the quality of the morphological predictions for the first GZ2 question, and here we include the results on the second and third questions. To measure the performance of a binary classifier, we quantify the accuracy (Acc), precision (Prec), recall or true positive rate (Rec/TPR), false positive rate (FPR), area under the receiver operator characteristic curve (AUC), and the outlier fraction η:

$$\begin{aligned}
\mathrm{Acc} &= \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{FP}) + (\mathrm{TN} + \mathrm{FN})} \\
\mathrm{Prec} &= \frac{\mathrm{TP}}{\mathrm{TN} + \mathrm{FP}} \\
\mathrm{Rec/TPR} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \\
\mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \\
\mathrm{AUC} &= \int \mathrm{TPR} d(\mathrm{FPR}) \\
\eta &= \frac{[(P_{true} \leq 0.2) \cap (P_{predict} \geq 0.8)] + [(P_{true} \geq 0.8) \cap (P_{predict} \leq 0.2)]}{N_{HQ}} \times 100,
\end{aligned}$$

Where TP and TN are the number of true positives and true negatives, and FP and FN are the number of false positives and false negatives, respectively, P_{true} and $\mathrm{P}_{predict}$ are the true and predicted labels, and N_{HQ} is the total number of labels considered high quality. We use a straight probability cut of 0.5 for both the true labels and predictions ($\mathrm{P} < 0.5$ belongs to class 0 and $\mathrm{P} \geq 0.5$ belongs to class 1), although quoted results can be slightly improved by allowing the probability cut on the predicted labels to vary.

Figures 11 and 12 show the classification results in the same style as Figure 3, but for the second and third GZ questions, respectively. For the second question we find that 256 labels are sufficient to train the linear classifier on the representations and the fine tuned network to a high degree of accuracy, while the supervised network does not converge. This question, "edge-on or face-on" is likely the 'easiest' of all GZ2 questions with obvious differences between the classes, and we find that 4096 labels is sufficient to train all three networks to a high accuracy. The third question, 'bar or no-bar', is one of the most difficult from GZ2 due to uncertainty in human labelling and ambiguity for all but the most-obvious bars. It also suffers from the largest class imbalance of the three, and 256 training samples only contains 17 high quality positive samples, and 38 with label probability ranging from 0.5 to 0.8. Nonetheless, accurate results are still achieved with limited training samples. In all cases the fine-tuned self supervised model outperforms its supervised counterpart, and with a limited number of labels a simple linear layer on the representations outperforms a fully supervised network.

We note that an exact quantitative comparison of our performance metrics to DS+18 and W+20, or to other automated morphological classification works, is not possible due to a lack of consistency in data sets. The samples used from the full set of GZ2 answers are not consistent: images can be from different SDSS data releases and have inconsistent pixel sizes or number of pixels, they use different observation bands (we use 5 here, while most works use 3) and different image normalization/re-sizing is applied, and different GZ2 vote fraction definitions are used as the "ground truth" morphological classification. Most importantly, the class imbalance used in training/testing is

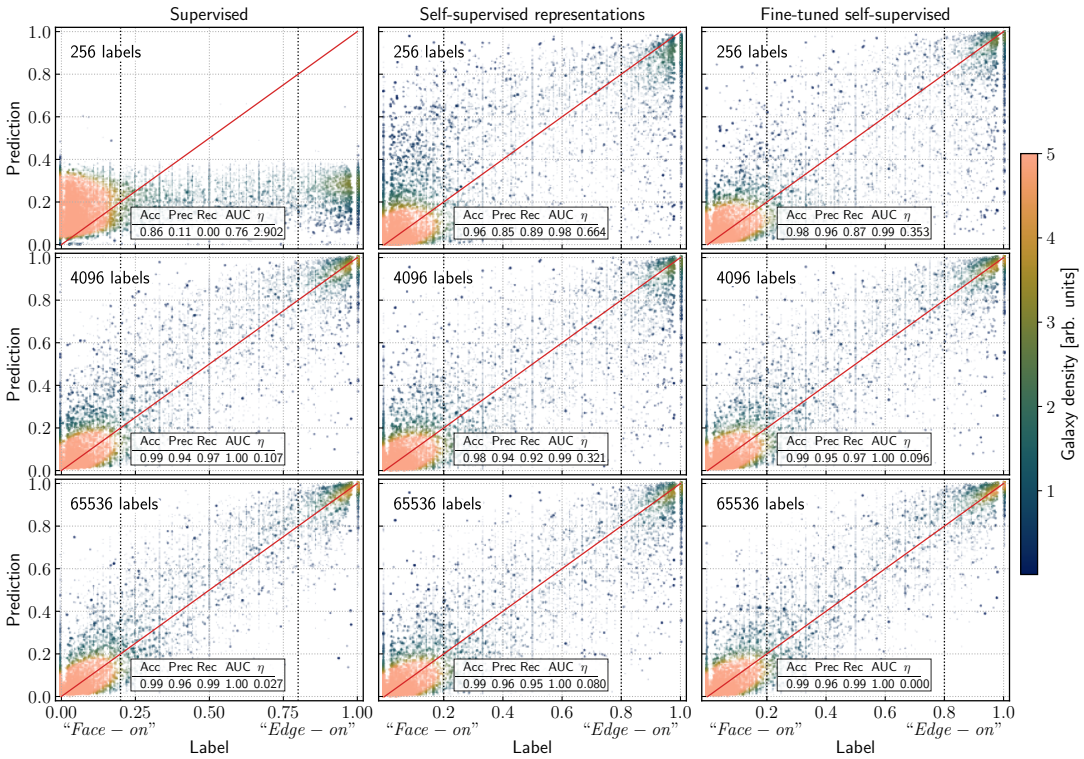

Figure 11. Predicted labels compared to crowd-sourced answers for the first GZ2 question: "Could this be a disk viewed edge-on?". The size and opacity of the points is proportional to the number of crowd-sourced labels they received.

significant (generally lower in our work as we use all samples), and more imbalanced datasets will more easily show up as higher metric scores. Nevertheless, the results achieved here are extremely promising, and show that the use self-supervised representations allows us to push beyond the limits of supervised learning for automated morphological classifications when limited by the number of available labels.

We have performed a number of tests to ensure that the networks are not basing their classification on secondary characteristics. The results shown are from un-normalized versions of the images, which include color information based on the relative brightness in each channel. We performed the same suite of classification exercises on images where we first normalized each channel of each image by the maximum pixel value, and found negligible differences to the results shown here. As seen in both UMAP figures, at least the first GZ2 question 'features or disk" has labels that are correlated with both galaxy size and magnitude. To test if a simple model using mainly these pieces of information can achieve high classification results; we trained a UMAP on images from the test set to reduce to 2048 dimension representations in an unsupervised fashion. Using these UMAP representations as inputs for a linear classifier, equivalent to how we trained out linear classifier on the self-supervised representations, we found that performance never increased above initialization. These tests confirm that a simple combination of size/brightness/color is not enough to achieve accurate predictions for these classifications.

Misclassified galaxies, specifically those that are large outliers, have three main failure modes which can be addressed in future work to further improve results beyond those achieved here. Firstly, our images span 25.3 arcseconds, which is not large enough to cover the entire angular extent of very nearby and large galaxies. In contrast, GZ2 participants were shown images scaled to ensure the entire galaxy was always visible, as do other automated classifications. Classifying these limited number of these galaxies is easily achievable by human means, so targeting these few samples is not of top priority. But, for the specific downstream task of classifying galaxy morphologies an additional augmentation which scales the angular extent of galaxies may prove beneficial. Second, a number of misclassified galaxies have imaging artifacts, which fine tuning the network would likely help improve the classification. A similarity search on the self-supervised representations provides a very valuable tool to isolate these artifacts. Finally, some 'misclassified' galaxies are the result of label uncertainty, which is especially rampant when the true label is outside of the high quality range.

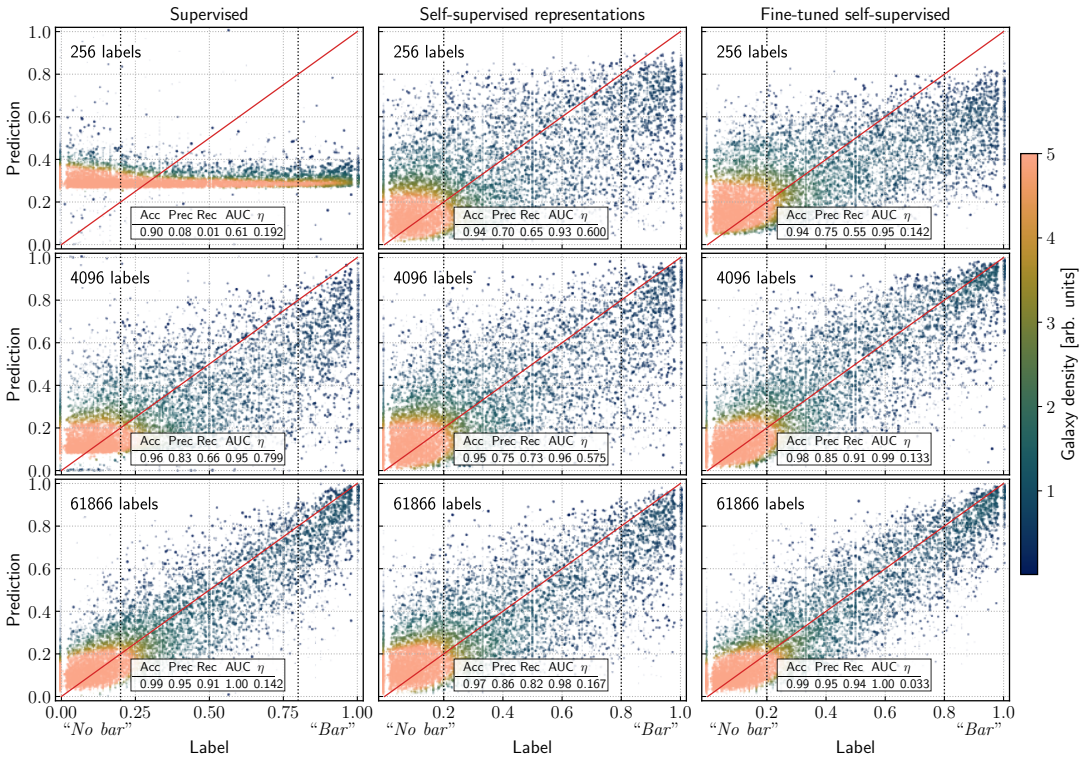

Figure 12. Predicted labels compared to crowd-sourced answers for the third GZ2 question: "Is there a sign of a bar feature through the centre of the galaxy?". The size and opacity of the points is proportional to the number of crowd-sourced labels they received.

D. DATA AUGMENTATION ABLATION STUDY

In Section 2 we have listed a basic set of augmentations which make intuitive sense for use in a contrastive learning framework, as each augmentation is associated with some source of observational variability within the images that we want the learned representations to be invariant to. A relevant task in developing the self-supervised model is deciding which combinations of these augmentations are most effective in producing high-quality, semantically useful image representations. This can be answered in a number of ways, but is traditionally done in computer vision by taking a sample downstream task, like image classification, and training a linear classifier on top of the learned representations to perform it. Doing so evaluates the quality of the learned representations by measuring, e.g., how easily different classes are linearly separated in the representation space.

Such an approach is straightforward for the task of morphology classification, but is slightly ill-conceived for something more challenging like photo-z estimation, since the "classes" output by the network represent consecutive redshift bins which should not necessarily be linearly separable. Instead, we evaluate our representations by fine-tuning them for the photo-z estimation task and using the $\sigma_{\rm MAD}$ of predictions on test data as our quality metric. To ensure this metric is more closely tied to the representation quality rather than the supervised fine-tuning process, we only consider the performance of models which are fine-tuned on 10%, 20%, and 30% of the labeled data.

The results of our evaluation are shown in Figure 13. We find that on our dataset, the Gaussian noise augmentation seems to be the strongest, but that the best performance is achieved when we use all augmentations except the PSF smoothing. An important note here is that this finding depends both on our dataset as well as the way we have implemented each augmentation. Using contrastive learning with other surveys would require different implementations and possibly different augmentations, and could produce different "hierarchies" of augmentation strengths. In general, careful thought needs to be put into the data augmentations and which transformations one wants the representations to be invariant against (Xiao et al. 2020).

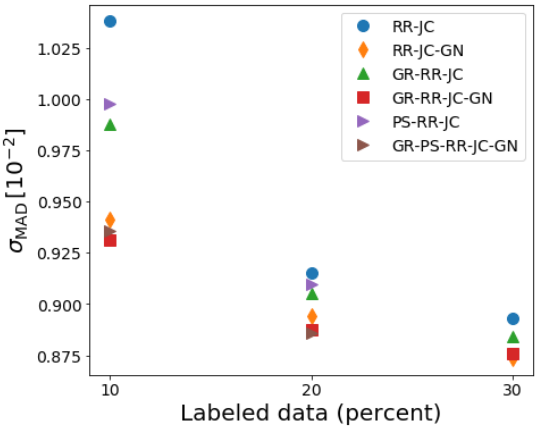

Figure 13. The $\sigma_{\rm MAD}$ of photo-z predictions from models fine-tuned on various fractions of the labeled data, showing how different combinations of augmentations affect downstream performance. Augmentations shown are random rotation (RR), jitter crop (JC), Gaussian noise (GN), galactic reddening (GR), and point spread function (PS).

E. ADDITIONAL ARCHITECTURE, HYPERPARAMETER, & TRAINING DETAILS

E.1. *Self-supervised framework*

Encoder: We follow (He et al. 2020; Chen, T. et al. 2020a) in using a ResNet50 architecture (He et al. 2016) as our encoder. Specifically, we use the implementation of the TorchVision library (`torchvision.models.resnet50`) which is a part of the PyTorch project (Paszke et al. 2019). The standard ResNet50, however, is designed to work on wider images than our 64×64 ones. To maintain reasonably wide representations by the end of the 50 layers, we change the first `Conv2d` layer to have `stride=1` instead of the default `stride=2`, we also remove the first `MaxPool2d` layer. This gives us $4\times$ wider activations throughout the network than what we would get with the defaults of ResNet50. The output of the `AdaptiveAvgPool2d` layer of the network is our representation, a 2048 dimensional vector. We also change the number of input channels to 5 to work with our 5 *ugriz* passbands data.

Following (Chen, T. et al. 2020a; Chen, X. et al. 2020), we don't use the representation $\mathbf{z}$ directly in the loss in Eq. 1, instead, we use a two layer MLP projection head which maps the representations to a space where the contrastive loss is applied. This has been shown to improve the learned representations. The output of the projection head is a 128 dimensional vector. The head is discarded after the self-supervised training process is completed.

Momentum encoder: In contrastive learning setups, in order to make the task of identifying positive examples non-trivial, it is crucial to have a large set of negative examples. For this we use the momentum encoder idea from (He et al. 2020); we maintain a queue of size 62k representations ($\sim 5\%$ of the training dataset size) that is continuously being updated during the training process. The representations in the queue are encoded using a momentum encoder; a second encoder with same architecture whose weights are an exponentially moving average of the main encoder weights. The parameters θ_k of the momentum encoder network are updated using the encoder parameters θ_q with momentum parameter m via

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q. \tag{E1}$$

The momentum hyper-parameter is set to $m = 0.999$. The momentum encoder helps maintain some consistency between the representations in the negative examples queue during the training. We use temperature parameter $\tau = 0.1$ in the contrastive loss.

During self-supervised pre-training, we use stochastic gradient descent with a cosine learning rate schedule, having an initial learning rate of 0.03. We pre-trained our network for 12 hours using 8 NVIDIA V100 GPUs on $\sim$1.3M images to complete $\sim$50 epochs and this proved to be good enough for learning useful features used in this study. We have used `DistributedDataParallel` of PyTorch to leverage distributed training.

E.2. *Photometric redshift estimation networks*

We closely follow the setup of Pasquet et al. (2019), whose CNN is trained as a classifier over a discrete set of 180 redshift bins of size $\delta z = 2.2 \times 10^{-3}$ spanning $0 \leq z \leq 0.4$, where the photo-z estimate z_p is computed as the

expectation $\mathbb{E}(z)$ over the probabilities predicted in each bin. We train models from scratch to establish a baseline with the ResNet50 architecture (with the same modifications made for the self-supervised pre-training encoder, see E.1). During training, images are de-reddened by using the tabulated $E(B-V)$ value with the photometric calibration in Schlafly & Finkbeiner (2011), then augmented with random rotations, and random jitter & crop. Only de-reddening and cropping is applied at testing or evaluation.

For the fine-tuned networks, we take the pre-trained encoder and replace the projection head with a single MLP layer. Here, parameters up to `AdaptiveAvgPool2d` are initialized with the pre-trained weights and the MLP layer has random initialization. The whole network is then trained on labels, with the pre-trained weights having a learning rate of 0.0001, and the MLP classifier layer having a learning rate of 0.001. We train for 100 epochs and reduce the learning rate by a factor of 10 at 60, 90-th epochs. In all cases we have a batch size of 256.