Learning useful representations for radio astronomy "in the wild" with contrastive learning

Inigo Val Slijepcevic [1] **Anna M. M. Scaife** [1 2] **Mike Walmsley** [1] **Micah Bowles** [1]

Abstract

Unknown class distributions in unlabelled astrophysical training data have previously been shown to detrimentally affect model performance due to dataset shift between training and validation sets. For radio galaxy classification, we demonstrate in this work that removing low angular extent sources from the unlabelled data before training produces qualitatively different training dynamics for a contrastive model. By applying the model on an unlabelled data-set with unknown class balance and sub-population distribution to generate a representation space of radio galaxies, we show that with an appropriate cut threshold we can find a representation with FRI/FRII class separation approaching that of a supervised baseline explicitly trained to separate radio galaxies into these two classes. Furthermore we show that an excessively conservative cut threshold blocks any increase in validation accuracy. We then use the learned representation for the downstream task of performing a similarity search on rare hybrid sources, finding that the contrastive model can reliably return semantically similar samples, with the added bonus of finding duplicates which remain after pre-processing.

1. Introduction

To date in radio galaxy classification, supervised CNNs have largely been used to classify sources into the Fanaroff-Riley I (FRI) and FRII categories (Figure 1), which has persisted as the canonical morphological distinction since it was established over 40 years ago (Fanaroff & Riley, 1974). However, such supervised classification requires many labelled samples. Acquiring labels is costly due to the labelling expertise required and may introduce selection biases training data are often chosen subject to observational factors such as brightness and distance constraints, Hardcastle & Croston 2020. Since our current understanding of the physics of radio galaxies is incomplete, there is also a question of whether the FR classification scheme is itself enforcing an implicit bias which may limit future scientific insight.

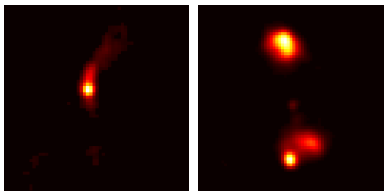
New radio sky surveys enabled by the construction of next-generation telescopes such as the Square Kilometre Array (SKA) will produce data on much larger scales than previous instruments, with significantly higher sensitivities expected to reveal a more morphologically diverse population of samples than are currently known. These data volumes will inevitably contain samples that challenge the current FR classification paradigm, and may be undetected or misclassified if our models remain tailored to the FR distinction. Furthermore, the volume of data, which is expected to be on the *exabyte* scale, will likely be intractable even for citizen science projects such as Radio Galaxy Zoo (Banfield et al., 2015) to label for supervised training.

Semi-supervised learning has been explored as a potential solution to address these issues, as it leverages unlabelled data samples and allows for more data-driven separation in classification outputs (Sohn et al., 2020; Berthelot et al., 2019; Tarvainen & Valpola, 2017; Sellars et al., 2021). Semi-supervised learning without pre-training (consistency regularization Tarvainen & Valpola (2017); pseudo-labelling Pham et al. (2021), etc.) can make use of unlabelled data but is brittle to differences in the labelled and unlabelled data distributions, as has been shown both from a theoretical perspective (van Engelen & Hoos, 2020) and empirically with astronomical data (Slijepcevic & Scaife, 2021; Slijepcevic et al., 2022), making it less well-suited to training "in the wild" on real astronomical data.

Self-supervised representation learning has recently driven improvements in state-of-the-art performance on image classification tasks by leveraging large unlabelled data-sets for pre-training, with methods including SimCLR (Chen et al., 2020a), MoCo (He et al.), and BYOL (Grill et al., 2020); Jaiswal et al. (2020) provide a survey. Adapting pre-training for effective representation learning in the radio astronomy domain may allow us to use all of the available unlabelled

*Equal contribution [1]Department of Physics and Astronomy, University of Manchester, Manchester, UK [2]Che Alan Turing Institute, 96 Euston Rd, London, UK. Correspondence to: Inigo Val Slijepcevic <inigo.slijepcevic@postgrad.manchester.ac.uk>.

ICML 2022 Workshop on Machine Learning for Astrophysics, Baltimore, Maryland, USA, 2022.



(a) Fanaroff-Riley I (b) Fanaroff-Riley II

Figure 1. A typical example of each class cherry-picked from the MiraBest data-set. (Miraghaei & Best, 2017).

data for training, rather than just a vanishingly small labelled fraction. Furthermore, it may help to provide flexibility for scientists to easily test hypotheses (such as alternate classification schema) with a significantly lower labelling cost by fine-tuning the pre-trained representation. Lastly, the use of techniques such as similarity search (Stein et al., 2021; Walmsley et al., 2022) highlight the usefulness of a good representation space for applications beyond classification that require no labels at all.

This work: We focus on using contrastive learning without negative pairs (BYOL) to learn a semantically meaningful representation of radio galaxy images. We use the separation of the FRI/II classes as measured by the accuracy of a kNN classifier applied to the representation space to assess the quality of the representation, showing that our model separates the classes in a labelled validation data-set *even when trained only on unlabelled data with an unknown distribution of labels and different selection cuts*. We compare the quality of this representation to the feature space learned by a supervised classifier with the same backbone architecture (Resnet18; He et al. (2016)). We show that unresolved and low angular extent sources sources in the training data of the contrastive model negatively impact the learned representation, highlighting the importance of domain-specific data curation for self-supervised models, which in our case can be done automatically before training. We show that by feeding the encoder a single data-point we can extract similar data from the unlabelled data through a similarity search, highlighting the scientific usefulness of a semantically meaningful representation space beyond just classification.

2. Data

The Mirabest data-set contains 1256 radio galaxies, which have been selected for classification based on a variety of criteria (Miraghaei & Best, 2017). The data-set is publicly available (Porter, 2020) and has been widely used to train and evaluate machine learning models for radio galaxy classification. The images are classified into FRI/II classes and also contain *Confident* and *Uncertain* tags which are aligned with the confidence in the respective classification of the expert labellers. This is the dataset that we use for evaluating our model performance and to train the supervised baseline. The data-set also contains some *hybrid* radio galaxy samples which we use only as input for a similarity search.

The Radio Galaxy Zoo Data Release 1 (RGZ DR1; Wong et al. in prep) contains $\sim 100,000$ sources which have been obtained from the FIRST survey (Becker et al., 1995) for the RGZ citizen science project. Each image is paired with an arc-second extension (calculated algorithmically), which gives the projected angular size of the source in the sky. The RGZ DR1 data-set has an unknown FRI/II class balance and the abundance of sub-populations is also unknown. Since both the MiraBest and RGZ DR1 data-sets are drawn from the FIRST survey, we remove any shared samples from RGZ DR1.

3. Method

Self-supervised learning attempts to learn a representation, $\boldsymbol{y}_\theta$, given data, $\boldsymbol{x}_i \in \boldsymbol{X}$, by giving the model a task without labels but forces the model to learn a representation space which is useful for not-yet-known downstream tasks. The backbone of the model is an encoder, with the rest of model being discarded at inference: the set up is analogous to transfer learning without any initial labels. In our contrastive learning framework, the model is trained by being shown pairs of augmented images and rewarding augmentations of the same image being placed close together in the representation space of the encoder. Some algorithms, such as SimCLR (Chen et al., 2020a), also teach the model to move augmentations of other images further away. However, recent work has shown that this is not required to achieve state-of-the-art accuracy, with Bootstrap Your Own Latent (BYOL) (Grill et al., 2020) exceeding SimCLR classification performance on ImageNet. BYOL has significantly less computation time than algorithms with negative pairs and a much lower memory footprint, which allows us to iterate faster given our limited computation budget. Furthermore, our images have a fuzzier semantic meaning, with less clear cut differences between classes and a more fine-grained nature than computer vision data-sets, which may cause negative pairs to confound the model. For these reasons, we use BYOL as the core of our model.

BYOL uses a momentum encoder to calculate positive pair losses, which helps to avoid representation collapse, although it has been shown that this is not strictly required (Chen & He, 2021). An exponential moving average of the online network parameters, θ, gives the parameters of the momentum encoder, ξ, which are updated at each step such that

$$\xi_i \leftarrow \tau\xi_{i-1} + (1-\tau)\theta, \tag{1}$$

using Equation 1, where the hyperparameter τ gives the decay rate.

Both the momentum and online encoder have a fully connected projection head which output $\boldsymbol{z}_\xi$ and $\boldsymbol{z}_\theta$, respectively, with the online network also having an additional fully connected prediction head, q_θ. These extra layers separate the loss from $\boldsymbol{y}_\theta$, helping to learn a more generalisable representation space by preventing the model from overfitting to the loss function. This has been empirically shown to improve performance in the contrastive learning setting (Chen et al., 2020a).

Two augmentations, t and t', are drawn from the same augmentation pool, $\mathcal{T}$, giving views $\boldsymbol{v}$ and $\boldsymbol{v}'$ of the original image, $\boldsymbol{x}$. $\boldsymbol{v}$ ($\boldsymbol{v}'$) is passed through the online (momentum) encoder to give q_θ ($\boldsymbol{z}_\theta$). A simple mean squared error between the momentum projection and online prediction of each augmentation gives the loss

$$\mathcal{L} = ||q_\theta(\boldsymbol{z}_\theta) - \boldsymbol{z}_\eta||_2^2, \tag{2}$$

which is back-propagated to update only the online parameters, θ. The parameters of the momentum encoder, ξ, are updated using the exponential moving average in Equation 1. Figure 2 illustrates the flow of tensors through the model.

Training setup: During training of BYOL, we only provide samples from the RGZ DR1 data-set, without showing the model any samples from MiraBest, and evaluate on *Confident* samples from MiraBest. The supervised model is trained on MiraBest only, and evaluated on held out *Confident* test samples. We use the representation learned by BYOL once the **training** loss has converged, rather than using early stopping to take the peak of the validation accuracy (which would likely give a higher test accuracy). We do this because we wish to emulate the setting where we are training a foundation model on unlabelled data, in which case there will be no labels with which to perform validation. During contrastive learning we modify the augmentations used in Chen et al. (2020a) by reducing the size of the blurring kernel to 3 pixels and reducing the probability of blurring to 0.05, which empirically improves our results. We hypothesise that this is because of the sparseness of the radio galaxy images, where strong blurring results in distortion of semantically different fine-grained images and edges into similar looking blurs. We also reduce the range of sizes possible when cropping, to avoid cropping out important features of the image (e.g. jets). Both BYOL and the supervised baseline use the same ResNet-18 backbone with a fully connected layer at the end to downscale the feature space to a size of 100. BYOL uses multiple fully connected heads as explained above (details can be found in Grill et al. 2020), while the supervised model uses a simple fully connected classification head. **Hyperparameters:** cosine scheduler with SGD optimizer, learning rate: 0.6, momentum: 0.9, weight decay: 0.0005, batch size: 256.

Evaluation: we use a simple kNN classifier with 20 neighbours applied directly to the representation to evaluate the separation of the classes on a held out test set of *Confident* MiraBest samples.

Similarity search: we perform a similarity search in the representation space by calculating the cosine similarity

$$S^{j,input} = \frac{\boldsymbol{y}_\theta^{input} \cdot \boldsymbol{y}_\theta^{j}}{|\boldsymbol{y}_\theta^{input}||\boldsymbol{y}_\theta^{j}|} \tag{3}$$

of the input sample representation $\boldsymbol{y}_\theta^{input}$ with the representation, $\boldsymbol{y}_\theta^{j}$, of each data-point in the unlabelled data. The top k images with the largest similarity scores are returned.

4. Results & discussion

4.1. Model performance

We find that without seeing any MiraBest samples, BYOL learns a meaningful representation that separates out the *Confident* MiraBest samples into FRI and FRII clusters. Figure 3 illustrates the representation quality compared with the supervised baseline, showing that the model is able to generalise from RGZ DR1 to MiraBest, which has been shown *not* to be the case for a semi-supervised model using consistency regularization (Slijepcevic et al., 2022). Although there is a gap between the supervised and self-supervised models (92.42% vs. 85.25%), it is surprisingly small given that the supervised model is *explicitly trained to separate these classes on the same data-set being used in evaluation and with the benefit of labels*, whereas the self-supervised model has been trained on a different data-set with no labels. Furthermore, the representation learned by the contrastive model is likely to generalise better to a variety of downstream tasks due to the task-agnostic nature of the training. For example, in Figure 4 we see that our learned representation is structured with respect to source extension, which the model has no knowledge of during training. Lastly, any labels available post-hoc for a specific downstream classification task can still be used to fine-tune the network, which has been shown to be more data-efficient than training a supervised model from scratch (Chen et al., 2020b).

Figure 3 shows that choosing a suitable cut threshold is important, with the model's kNN classification performance suffering by over 10% in the most extreme examples when low angular extent sources are included. Figure 5 shows that there is a qualitative difference in the way the model trains when the cut threshold is too low: at low cut thresholds the model is unable to improve its validation score even as the training loss decreases.

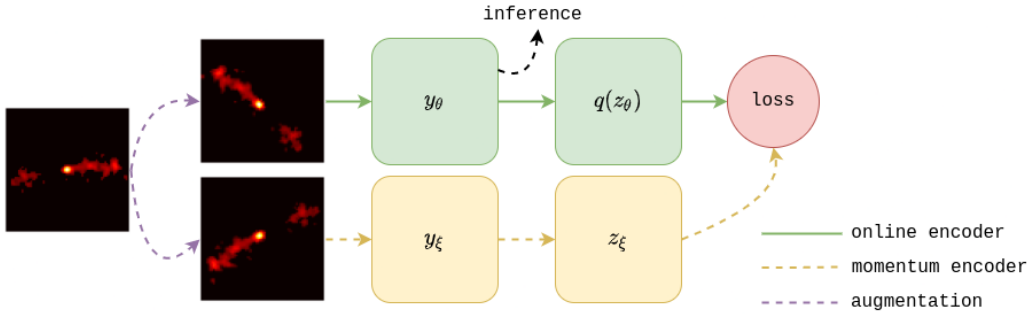


Figure 2. Diagram showing the flow of tensors in the BYOL algorithm. Dashed lines indicate that gradients are **not** back-propagated

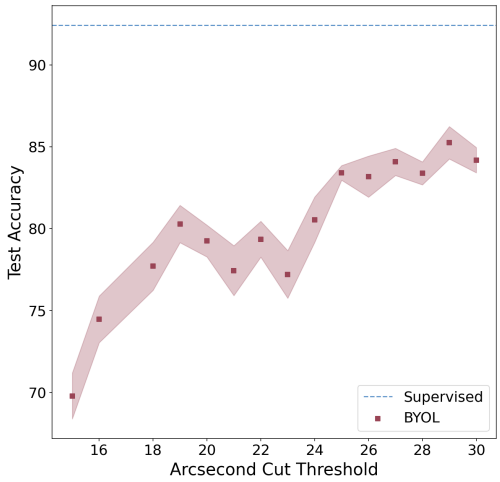


Figure 3. kNN accuracy as a function of cut threshold. Shaded area shows the standard error over 5 seeded runs.

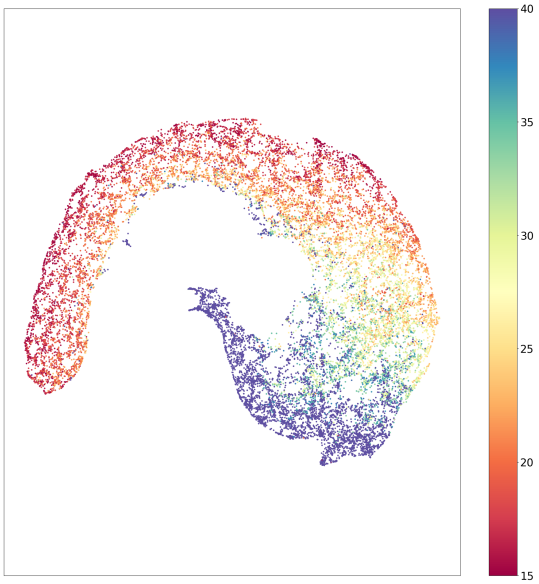


Figure 4. UMAP projection of the encoded representation RGZ DR1 data-set, with colorbar showing arcsecond extension.

Figure 5. BYOL validation accuracy curves at different cut thresholds. Shaded area shows the standard error over 5 seeded runs, smoothed with an EMA constant of 0.7.

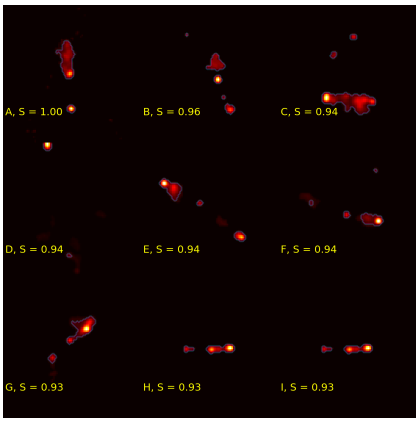


Figure 6. A similarity search performed on A (a rare hybrid galaxy from the MiraBest data-set) in the RGZ DR1 data-set. **Similar features:** (i) One extended diffuse jet. (ii) One bright spot from jet. (iii) Features extend to a similar angular scale. H and I are duplicates which evaded duplicate removal during pre-processing.

4.2. Similarity search

In Figure 6 we feed our model with a hybrid image with both FRI and FRII properties from MiraBest, which would be considered out-of-distribution in the supervised case, and extract 3 nearest neighbours in the representation space. We see that we are able to extract semantically similar data-points from the RGZ DR1 data-set. We also find a duplicate data-point in the RGZ DR1 catalogue, see Figure 6, even though the data have been (naively) pre-processed to remove duplicates. This shows that similarity search could also be used to correct errors where the same object has been included twice due to cataloging differences, e.g. slightly offset cutouts, which are difficult to catch in pre-processing.

5. Conclusion

We find that we are able to learn a semantically meaningful representation of radio galaxies with self-supervised learning. We we approach performance of the supervised case without any labels. This verifies that the representation learned is robust to data-set shift, indicating that we can use the vast quantities of available astronomical unlabelled data for representation learning *without* the requirement of costly human classification and curation. We also find that some domain specific pre-processing is necessary for the radio galaxy case. While in our case this is easily applied, future work could investigate ways of automating this in a domain-agnostic way, which may also prove beneficial for training contrastive models in domains outside of astronomy.

We believe this work is a step towards developing large scale astronomical foundation models which can be shared and used by scientists for a variety of downstream tasks. Given the self-supervised nature of the learning task, the FRI/II separation and source extension structure in the representation space is indicative of a more general semantic structure of the vector space, and we expect that it may perform well on other physically meaningful tasks on similar data.

We show an example downstream task where we find images with similarly unusual physical properties to an out-of-distribution image by feeding it to the model and performing a similarity search in the representation space. This method could be used to find further examples of specific types of astronomical objects in a large unlabelled data-set and also to remove difficult-to-detect duplicates.

6. Acknowledgements

We thank the anonymous reviewer whose comments improved this work.

IVS, AMS, MB & MW gratefully acknowledge support from the UK Alan Turing Institute under grant reference EP/V030302/1. IVS gratefully acknowledges support from the Frankopan Foundation. HT gratefully acknowledges the support from the Shuimu Tsinghua Scholar Program of Tsinghua University.

This work has been made possible by the participation of more than 12,000 volunteers in the Radio Galaxy Zoo Project. The data in this paper are the result of the efforts of the Radio Galaxy Zoo volunteers, without whom none of this work would be possible. Their efforts are individually acknowledged at http://rgzauthors.galaxyzoo.org.

References

Banfield, J. K., Wong, O. I., Willett, K. W., Norris, R. P., Rudnick, L., Shabala, S. S., Simmons, B. D., Snyder, C., Garon, A., Seymour, N., Middelberg, E., Andernach, H., Lintott, C. J., Jacob, K., Kapińska, A. D., Mao, M. Y., Masters, K. L., Jarvis, M. J., Schawinski, K., Paget, E., Simpson, R., Klöckner, H. R., Bamford, S., Burchell, T., Chow, K. E., Cotter, G., Fortson, L., Heywood, I., Jones, T. W., Kaviraj, S., López-Sánchez, R., Maksym, W. P., Polsterer, K., Borden, K., Hollow, R. P., and Whyte, L. Radio Galaxy Zoo: Host galaxies and radio morphologies derived from visual inspection. *Monthly Notices of the Royal Astronomical Society*, 453(3):2326–2340, 2015. ISSN 13652966. doi: 10.1093/mnras/stv1688. URL http://www.sepnet.ac.uk.

Becker, R. H., White, R. L., Helfand, D. J., Becker, R. H., White, R. L., and Helfand, D. J. The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters. *ApJ*, 450:559, 9 1995. ISSN 0004-637X. doi: 10.1086/176166. URL https://ui.adsabs.harvard.edu/abs/1995ApJ...450..559B/abstract.

Berthelot, D., Carlini, N., Goodfellow, I., Oliver, A., Papernot, N., and Raffel, C. MixMatch: A holistic approach to semi-supervised learning. Technical report, 2019. URL https://github.com/google-research/mixmatch.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. 2020a. URL https://github.com/google-research/simclr.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 2020-Decem, 2020b. ISSN 10495258. URL https://github.com/google-research/simclr.

Chen, X. and He, K. Exploring Simple Siamese Representation Learning. pp. 15745–15753, 2021. doi: 10.1109/cvpr46437.2021.01549. URL https://github.com/facebookresearch/simsiam.

Fanaroff, B. L. and Riley, J. M. The Morphology of Extragalactic Radio Sources of High and Low Luminosity. *Monthly Notices of the Royal Astronomical Society*, 1974. ISSN 0035-8711. doi: 10.1093/mnras/167.1.31p.

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 2020-Decem, 2020. ISBN 2006.07733v3. URL https://github.com/deepmind/deepmind-research/tree/master/byol.

Hardcastle, M. J. and Croston, J. H. Radio galaxies and feedback from AGN jets. *New Astronomy Reviews*, 88, 2020. ISSN 13876473. doi: 10.1016/j.newar.2020.101539. URL http://www.jb.man.ac.uk/atlas/;.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. URL https://github.com/facebookresearch/moco.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pp. 770–778, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90. URL http://image-net.org/challenges/LSVRC/2015/.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. A Survey on Contrastive Self-Supervised Learning. *Technologies*, 9(1):2, 2020. doi: 10.3390/technologies9010002.

Miraghaei, H. and Best, P. N. The nuclear properties and extended morphologies of powerful radio galaxies: The roles of host galaxy and environment. *Monthly Notices of the Royal Astronomical Society*, 466(4):4346–4363, 2017. ISSN 13652966. doi: 10.1093/mnras/stx007.

Pham, H., Dai, Z., Xie, Q., Luong, M.-T., and Le, Q. V. Meta Pseudo Labels. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. ISSN 2331-8422. URL http://arxiv.org/abs/2003.10580.

Porter, F. A. M. MiraBest Batched Dataset. *https://zenodo.org/record/4288837*, 11 2020. doi: 10.5281/ZENODO.4288837. URL https://zenodo.org/record/4288837.

Sellars, P., Aviles-Rivero, A. I., and Schönlieb, C.-B. LaplaceNet: A Hybrid Energy-Neural Model for Deep Semi-Supervised Classification. *CoRR*, 2021. URL http://arxiv.org/abs/2106.04527.

Slijepcevic, I. V. and Scaife, A. M. M. Can semi-supervised learning reduce the amount of manual labelling required for effective radio galaxy morphology classification? *NeurIPS 2021: Machine Learning and the Physical Sciences Workshop*, 11 2021. URL https://arxiv.org/abs/2111.04357v2http://arxiv.org/abs/2111.04357.

Slijepcevic, I. V., Scaife, A. M. M., Walmsley, M., Bowles, M., Wong, I., Shabala, S. S., and Tang, H. Radio Galaxy Zoo: Using semi-supervised learning to leverage large unlabelled data-sets for radio galaxy classification under data-set shift. *MNRAS*, (May), 2022. URL http://arxiv.org/abs/2204.08816.

Sohn, K., Berthelot, D., Li, C. L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 2020-Decem, 2020.

Stein, G., Harrington, P., Blaum, J., Medan, T., and Lukic, Z. Self-supervised similarity search for large scientific datasets. 2021. URL http://arxiv.org/abs/2110.13151.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 2017-Decem: 1196–1205, 2017. ISSN 10495258.

van Engelen, J. E. and Hoos, H. H. A survey on semi-supervised learning. *Machine Learning*, 109(2), 2020. ISSN 15730565. doi: 10.1007/s10994-019-05855-6.

Walmsley, M., Scaife, A. M. M., Lintott, C., Lochner, M., Etsebeth, V., Géron, T., Dickinson, H., Fortson, L., Kruk, S., Masters, K. L., Mantha, K. B., and Simmons, B. D. Practical galaxy morphology tools from deep supervised representation learning. *Monthly Notices of the Royal Astronomical Society*, 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac525.