Galaxy Zoo Evo: 1 million human-annotated images of galaxies

Mike Walmsley
University of Toronto
m.walmsley@utoronto.ca

Steven Bamford
University of Nottingham

Hugh Dickinson
Open University

Tobias G ron
University of Toronto

Alexander J. Gordon
University of Edinburgh

Annette M. N. Ferguson
University of Edinburgh

Lucy Fortson
University of Minnesota

Sandor Kruk
European Space Agency

Natalie Lines
University of Portsmouth

Chris J. Lintott
University of Oxford

Karen L. Masters
Haverford College

Robert G. Mann
University of Edinburgh

James Pearson
Open University

Hayley Roberts
University of Minnesota

Anna M.M. Scaife
University of Manchester

Stefan Schuldt
Università degli Studi di Milano

Brooke Simmons
University of Lancaster

Rebecca Smethurst
University of Oxford

Josh Speagle
University of Toronto

Kyle Willett
Amazon

Abstract

We introduce Galaxy Zoo Evo, a labeled dataset for building and evaluating foundation models on images of galaxies. GZ Evo includes 104M crowdsourced labels for 823k images from four telescopes. Each image is labeled with a series of fine-grained questions and answers (e.g. 'featured galaxy, two spiral arms, tightly wound, merging with another galaxy'). These detailed labels are useful for pretraining or finetuning. We also include four smaller sets of downstream labels (167k galaxies in total) for downstream tasks of specific interest to astronomers, including finding strong lenses and describing galaxies from the new space telescope *Euclid*. We hope GZ Evo will serve as a real-world benchmark for computer vision topics such as domain adaption (from terrestrial to astronomical, or between telescopes) or learning under uncertainty from crowdsourced labels. We also hope it will support a new generation of foundation models for astronomy; such models will be critical to future astronomers seeking to better understand our universe.

1 Introduction

The development of computer vision methods has historically been driven by longstanding benchmarks like ImageNet [1]. But such benchmarks may no longer be good proxies for real-world performance [2, 3]. Further, purpose-built benchmarks may (by design) elide important complexities of real-world tasks such as label uncertainty [4, 5]. This paper aims to support new practical vision methods by sharing a large-scale crowdsourced dataset, created with scientific rigor, carefully doc-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

umented, and free of privacy, safety, or commercial concerns, to serve as a challenging real-world benchmark.

Galaxy Zoo Evo (Fig. 1) is made up of a 'Core' dataset of galaxy images taken by four telescopes, where volunteers have answered a broad series of questions about the galaxy in each image, plus four smaller 'Downstream' datasets of specific scientific interest.

Figure 2 shows three illustrative workflows for using GZ Evo. One might train a supervised model on our Core dataset and then finetune it on our Downstream datasets (e.g. [6]). Or one might take an existing model and separately apply supervised finetuning to evaluate generalization (zero-shot similarity search, supervised finetuning, etc.) on each of our Core and Downstream datasets (e.g. [7–10]. Or one might use continual learning to finetune an existing model first on our Core datasets and then further finetune it on our Downstream datasets (see our ViT-SO400 baseline, Sec. 5.1).

Our Core dataset is composed of 823k labelled images of galaxies from the Galaxy Zoo project [11]. Galaxy Zoo asks members of the public to volunteer their time annotating images. These annotations – and models trained on these annotations – help us learn how galaxies evolve; measuring the appearance of millions of galaxies, each with a different history, allows one to identify how galaxies grow, merge, and age. [12]. We collate sixteen years of annotations by 334k volunteers.

Our Downstream datasets are composed of four smaller sets of images with labels of specific scientific interest. Three of the Downstream datasets are searches for a particular type of galaxy: strong lenses (where the light from a background galaxy is bent by the gravity of a foreground galaxy); galaxies with rings; and galaxies with faint debris (often indicating a recent collision with another galaxy). The fourth dataset is GZ Euclid, where volunteers answered the same series of broad questions as Core but for images from a new and state-of-the-art space telescope, *Euclid*. Generalizing to images from *Euclid* is a real and urgent test for foundation models in astronomy.

No astronomy knowledge is required. We include minimal and reproducible code for training and evaluating baseline models. Our framework is extensible to new data, allowing Galaxy Zoo 'Evo' to 'evolve' as new telescopes and new labels become available.

Galaxy Zoo Evo is available at huggingface.co/mwalmsley with supporting code at github.com/mwalmsley/gz-evo.

2 Related Work

This paper is inspired by the 2014 'Galaxy Challenge' Kaggle competition which asked competitors to predict volunteer votes to an early Galaxy Zoo iteration (GZ2, below). The winning submission [13] introduced CNNs to astronomy and sparked a transition to deep learning methods throughout the discipline (see [14] for a review). The underlying dataset of 62k training images has also been widely-used in computer science as a real-world benchmark (see e.g. [15–18]). Galaxy Zoo Evo is intended as a ten year update on the Galaxy Challenge dataset, bringing an order-of-magnitude increase in scale, a new focus on multi-task learning and and downstream generalization, and an extensible framework to allow the dataset to grow over the next ten years.

We are also motivated by Fang et al. 2 's call that 'researchers should explicitly search for methods that improve accuracy on real-world non-web-scraped datasets, rather than assuming that methods that improve accuracy on ImageNet will provide meaningful improvements on real-world datasets as well'. This is especially relevant for domains such as medical imaging[19–21] and manufacturing [22, 23] with a substantial shift in features vs. ImageNet. Recent progress in vision-language pretraining, e.g., training models on extreme-scale web-scraped images, shows impressive few-shot gains on other web-scraped tasks [24–26]. We hope Galaxy Zoo Evo will help support a growing theme of work investigating vision-language pretraining for improved performance on scientific and technical images (e.g. [26–31]).

We build on a series of papers by the Galaxy Zoo collaboration ([32–34, 6]) that cumulatively motivate the design of our baselines. Our labelled dataset complements recent work to create an unlabelled multi-modal (i.e. images, time series, spectra) dataset for astronomy [35]; together, these datasets form a comprehensive resource for self-supervised pretraining, supervised finetuning, and downstream evaluation.

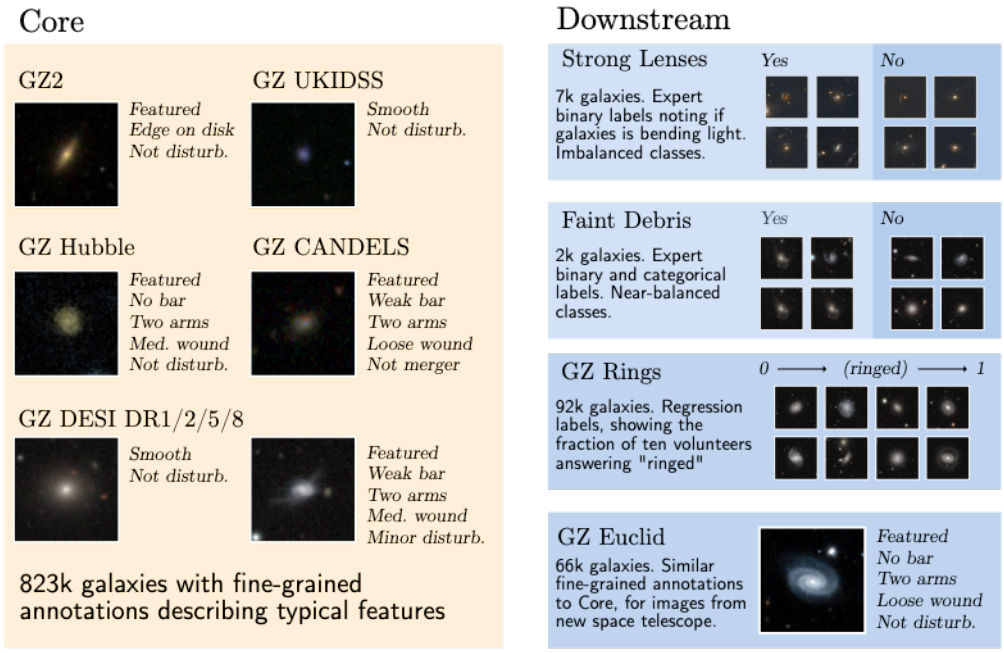

Figure 1: Overview of GZ Evo. GZ Evo includes a large-scale 'Core' dataset with fine-grained annotations describing the typical features of 823k galaxies, and four smaller 'Downstream' datasets with annotations of specific scientific interest.

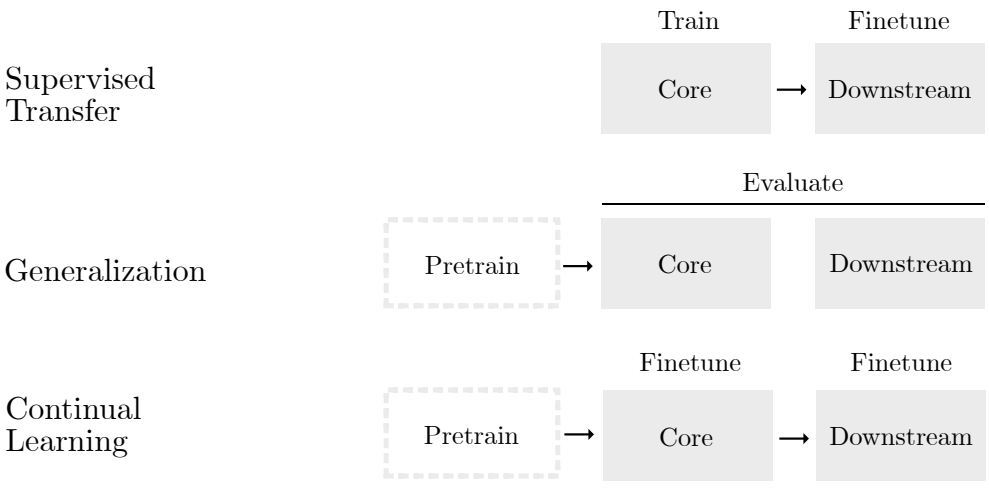

Figure 2: Three illustrative workflows using GZ Evo. Supervised transfer: train a supervised model on 'Core' (galaxy images with general descriptions) and finetune on 'Downstream'. Generalization evaluation: test an existing model on 'Core' and 'Downstream'. Continual learning: finetune an existing model on 'Core', and then finetune again to maximise performance on 'Downstream'

Table 1: Subset sizes. GZ Evo size is slightly smaller than naive sum to avoid train/test contamination from galaxies that appear in multiple campaigns. Logged-in volunteers are unique per subset.

Name	Galaxies	Logged-in Volunteers	Votes	Start Year	Data Release
GZ2	209k	84k	35.8M	2009	[36]
GZ Hubble	96k	88k	17.6M	2010	[37]
GZ UKIDSS	71k	25k	9.5M	2013	[38]
GZ CANDELS	48k	45k	5.2M	2012	[39]
GZ DESI	399k	125k	39.1M	2015	[33]
GZ Evo	823k	334k	104M	-	(here)

3 Core Dataset

3.1 Context and Construction

Galaxy Zoo asks members of the public to label galaxy images. Volunteers are presented with a galaxy image and initial question. As they answer, they are led down a decision tree where the next question asked depends on the previous answer. This breaks down the process of identifying detailed galaxy features into a series of intuitive judgements. Volunteers are guided by a tutorial, illustrative icons, expandable help text, and a comprehensive 'Field Guide'. They can share discoveries and compare notes on an integrated forum. We invite the reader to take part at www.galaxyzoo.org.

The questions and images both change over time, reflecting advances in telescopes and scientific knowledge. Galaxy Zoo publications therefore divide sets of questions and images into homogeneous labelling 'campaigns'. Adjusting a question or sourcing images from a new telescope is considered the start of a new campaign. The labels from each campaign were published separately, in disparate formats, in astronomical journals. Galaxy Zoo Evo brings together the separate Galaxy Zoo campaigns in a consistent and machine-learning-friendly format. We include all[1] published Galaxy Zoo campaigns, covering four telescopes and a wide diversity of galaxies. Table 1 shares key statistics for each campaign. The decision trees asked in each campaign are visualised at https://data.galaxyzoo.org/gz_trees/gz_trees.html.

3.2 Labels, Loss Functions, and Metrics

Each galaxy image in the Core dataset is annotated in a format like:

```
{
    smooth-or-featured_gz2_smooth: 30
    smooth-or-featured_gz2_smooth_fraction: 0.75
    smooth-or-featured_gz2_featured-or-disk: 10
    smooth-or-featured_gz2_featured-or-disk_fraction: 0.25
    ...
    summary: 'smooth'
}
```

The number of volunteers voting for an answer to a given question is reported in the columns `{question}_{campaign}_{answer}`. The total number of volunteers answering each question varies (e.g., because fewer volunteers reach questions further down the decision tree). Training directly on these counts using a multinomial loss[2] is a simple approach for incorporating label uncertainty. We train our baselines on Core using a multinomial loss summed across all questions and all campaigns (multi-task learning). We also report RMSE metrics[3].

[1] Excluding the first campaign, Galaxy Zoo 1, which predates the decision tree and asked only one question

[2] i.e., for each question, predict the odds $\hat{P} = f \quad x)$ that each volunteer selects each answer, and penalize the model according to the likelihood that we therefore observe $y = \vec{k}$ volunteers selecting each answer

[3] We precalculate the fraction of volunteers giving each answer to each question in the columns `{question}_{campaign}_{answer}_{fraction}`

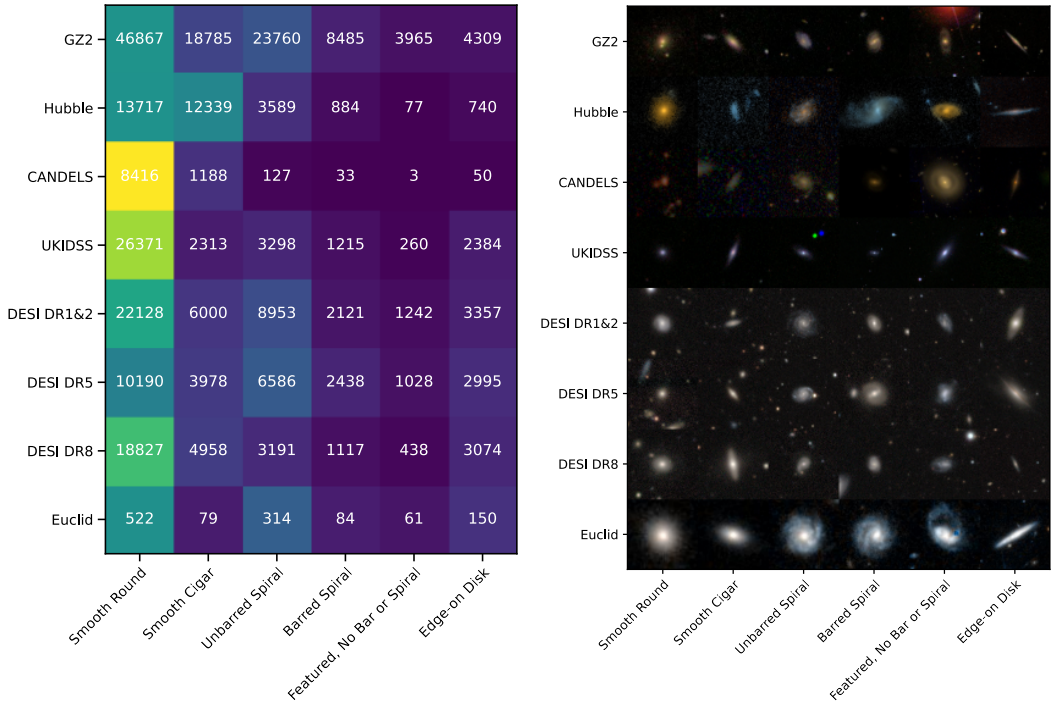

Figure 3: Left: Class counts by subset, when using our aggregated high-confidence class labels. Colours normalized by subset (row) to show the varying class distribution within each subset. This is a real(istic) test scenario for domain adaption. Right: an example galaxy image for each survey and class.

For researchers who prefer to work in a classification context, we also aggregate our count labels into six non-overlapping classes. The class label is reported in the `summary` column. Galaxies are labelled as smooth and round, smooth and cigar-shaped, edge-on disks, featured with spiral arms and a bar, featured with spiral arms and no bar, or featured with another characteristic (i.e. neither spiral arms nor a bar). We only provide a class label for the subset (287k) of galaxies where a class label can be confidently assigned. We consider as high confidence galaxies where a single answer received over half of all volunteer votes (an absolute majority), for all of the questions needed to derive our summary label, and where the galaxy received at least 30 volunteer votes in total (20 for GZ Euclid). We include code for reproducing our aggregation of votes to class labels. Figure 3 shows the distribution of classes in each subset and an example of every class for every campaign.

We anticipate the class labels will be useful for methods that assume a classification context (common in e.g. active learning, domain adaption) and where label uncertainty is not the research focus. We anticipate the count labels will be useful as a challenging multi-task finegrained learning problem, for pretraining and finetuning (Fig. 2), and for learning under realistic uncertainty from crowds.

4 Downstream Datasets

We complement our Core dataset – galaxy images with finegrained labels from volunteers – with four downstream datasets intended for evaluating the generalization of trained foundation models. Three downstream datasets (Strong Lenses, GZ Rings, and Faint Debris) ask models to answer a new task using images already seen. The remaining dataset (GZ Euclid) asks models to answer a comparable set of questions using images from the new *Euclid* space telescope.

4.1 GZ Euclid

The *Euclid* space telescope is beginning to take images of one third of the sky. When compared to images from the ground, *Euclid*'s images are approximately 10 times sharper[4] and promise a revolution in our understanding of distant galaxies [40]. Galaxy Zoo volunteers were asked to annotate these new images using the same tree structure as for previous campaigns extensively annotated in Core [41]. This downstream task requires the model to predict those volunteer responses. This directly tests the ability of models to generalise from GZ Evo Core to new telescopes. Additionally, because *Euclid* is new, training galaxies only have 5 volunteer votes (test galaxies have 25) and so models must generalise from noisy training labels.

4.2 Strong Lenses

Mass bends light. When two galaxies line up, the foreground galaxy can bend and magnify the light of the background galaxy into striking visual arcs. This rare alignment allows astronomers to measure the mass of the invisible dark matter in the foreground galaxy, and, potentially, to measure the rate of the expansion of the universe [42]. This downstream task uses expert labels from a recent search for strong lenses in new *Euclid* images [43–47] and requires models to predict whether 10 experts judged each image to contain a strong lens or a similar-looking 'imposter' galaxy. Strong lenses are rare and hence the dataset is highly imbalanced (6.7% lens).

4.3 GZ Rings

Ringed galaxies have (as one expects) a circular feature within or around the galaxy. Their formation mechanism is controversial [48]. To identify ringed galaxies, we use labels collected via Galaxy Zoo Mobile over two years. Volunteers were shown a galaxy image and asked to swipe according to whether it included a ring. This downstream task requires models to predict the fraction of ten volunteers who responded 'Ring'.

4.4 Faint Debris

Galaxies grow partly by colliding and merging with one another. These collisions leave faint but visible remnants: 'low surface brightness features', in astronomy jargon, or 'faint debris' here. Identifying these remnants allows astronomers to measure the rate of galaxy growth, but remnants are hard to distinguish automatically [49–51]. Gordon et al. 52 annotated 2000 images from DECaLS [53, 32] according to whether each image showed faint debris and, if so, assigned a subclass. Subclasses are 'arm', 'stream', 'shell', or 'diffuse' [54]. We use these labels for two downstream tasks: predict if the image has debris ('Is Debris') and (where positive) the subclass ('Which Debris').

5 Core Dataset Results

We train popular architectures (ConvNeXt [55], EfficientNetv2 [56], MaxViT [57], ResNet [58, 59], SoVit-400m [60]) with a simple standard recipe: schedule-free AdamW with weight decay (0.05), [61], stochastic depth [62] (values per the original papers), and standard rotations/flips/crops for augmentations. We select architecture variants from 10 A100-hours (ConvNext-Nano) to 80 A100-hours (SoViT-400m/14). We use a learning rate of 1^{-4}, reducing to 1^{-5} for the largest models.

Table 2 reports performance metrics on our Core Galaxy Zoo subsets. We report classification metrics for models trained with cross-entropy loss against our class labels, and regression metrics for models trained with multinomial loss against our vote counts ('k of N volunteers'). No single architecture is clearly strongest; EfficientNetV2 and Max-ViT are strong at classification and multi-task regression, respectively, while ConvNeXt is strong overall. The largest models perform poorly due to overfitting. All modern architectures significantly outperform ResNet50.

Regression metrics show a relatively small spread vs. other datasets because many answers have similar vote fractions for most galaxies. For example, due to physical symmetries, galaxies typically have 0, 2, or 4 spiral arms, and very few have 1 or 3. Any model will achieve a low RMSE/loss at

[4] Astronomers measure image sharpness using the 'point spread function full-width-half-maximum', which is calculated as the angular distance by which the apparent light recorded from a point-like object falls by half

Table 2: Metrics for supervised baselines training on GZ Evo Core dataset.

	Classification			Regression	
Architecture	Accuracy	Balanced Acc.	CE Loss	RMSE	Multinomial Loss
ResNet50	0.8872	0.719	1.1961	0.0960	17.602
ConvNeXT Nano	0.8791	0.595	1.7019	0.0874	17.173
ConvNeXT Base	0.9300	**0.813**	0.6863	**0.0842**	16.967
ConvNeXT Large	0.8936	0.612	1.3430	0.0864	17.271
EfficientNetV2 S	**0.9310**	0.810	**0.6222**	0.0878	17.150
EfficientNetV2 M	0.9220	0.784	0.6837	0.0879	17.100
EfficientNetV2 L	0.9259	0.799	0.6237	0.0869	17.105
MaX-ViT Tiny	0.9275	0.804	0.7639	0.0858	**16.965**
MaX-ViT Base	0.9013	0.696	0.9237	0.0852	17.004

predicting the fraction of volunteers voting for 1 or 3 spiral arms (almost none), and this reduces the variation in RMSE/loss when averaging across all answers. This does not imply that predicting vote fractions for all answers is easy to solve. Below, we finetune the regression-trained models on our Downstream datasets, and show they outperform the equivalent ImageNet-trained models.

5.1 Downstream Dataset Results

Table 3 shows test loss when training on the GZ Evo Core dataset and then finetuning on each of the GZ Evo Downstream datasets (mean of six seeds). For finetuning, we adapt all layers with a weight decay of 0.2 and an aggressive[5] exponential learning rate reduction by block (layer group) of $_{i\ 1} = \ _{i}{}^{2}$. In addition to test loss by dataset, we report an *overall score* calculated by rescaling the test loss for each dataset from 0 (worst model) to 1 (best) and then taking a mean across datasets.

Overall, ConvNeXt models perform strongest downstream, closely followed by MaxViT models. SoViT-400m/14 excels at identifying strong lenses, possibly because identifying multiple separated arcs of lensed light benefits from an architecture designed for nonlocal features. ConvNeXt-Nano performs best at the smallest two datasets (Is Debris and Which Debris) while ConvNeXt-Base performs best at the two largest Downstream datasets (GZ Euclid, GZ Rings).

For comparison, we also include several models directly-finetuned from their non-domain-specific pretrained versions (ImageNet-1k [1] for ResNet50, ImageNet-12k[6] for ConvNeXt-Nano, and webli [63] via SigLIP 2 [64] for SoViT-400m). These perform worse in every case, suggesting a continual learning approach is useful for maximising downstream performance. Notably, when considering only the three models finetuned directly on our downstream datasets without intermediate Core training, the SigLIP-trained model (SoViT-400m) is more succesful than the ImageNet-1k/12k models (ResNet/ConvNexT), which may suggest that features from SigLIP (i.e. vision-language training) may generalise better to our astronomy images than features learned from supervised ImageNet, consistent with similar observations in other domains [28].

Figure 4 investigates label-efficient learning. We again show test loss when training on Core and then finetuning on each Downstream dataset, but artificially restrict the size of each Downstream dataset to include fewer images for finetuning. Consistent with Table 3, all models tested outperform their directly-finetuned equivalents at all dataset sizes (not shown for clarity). All models show consistent and predictable improvement when scaling dataset size, suggesting that, first, data is a bottleneck to scaling, and second, more efficient strategies to learn from limited data (or from unlabelled data) should improve on our baselines.

6 Limitations

GZ Core, GZ Rings, and GZ Euclid are annotated by volunteers, and Strong Lensing is annotated by professional astronomers. Volunteer annotations of galaxy images have been consistently shown

[5]With the exception of SoViT-400m, where we use $_{i\ 1} = \ _{i}{}^{.8}$ due to the atypically large number of blocks

[6]`https://github.com/rwightman/imagenet-12k`

Table 3: Test loss when training on the GZ Evo Core dataset and then finetuning on the GZ Evo Downstream datasets. Overall score is normalised from 1 (best model at every dataset) to 0 (worst model at every dataset). Mean of six seeds. '*' indicates otherwise-identical models without intermediate training on GZ Evo Core; these perform worse in every case. SoViT-400m/14 performs best for strong lenses, while ConvNeXt models are strongest overall.

Model	Strong Lenses	GZ Euclid	GZ Rings	Is Debris	Which Debris	Overall Score
ConvNeXt-Base	0.1372	**2.1271**	**0.0201**	0.0734	0.6709	**0.94**
ConvNeXt-Nano	0.1431	2.1500	0.0203	0.0713	**0.6334**	0.92
MaxViT-Base	0.1336	2.1457	0.0203	0.0777	0.6918	0.91
MaxViT-Tiny	0.1409	2.1568	0.0204	**0.0698**	0.6631	0.91
SoViT-400m/14	**0.1110**	2.1764	0.0211	0.0944	0.6973	0.87
EfficientNetV2-M	0.1596	2.1843	0.0207	0.0773	0.7511	0.80
SoViT-400m/14*	0.1309	2.1921	0.0216	0.0974	0.8088	0.77
ResNet50	0.1699	2.2195	0.0218	0.0839	0.9118	0.66
ConvNeXt-Nano*	0.1720	2.2251	0.0223	0.1291	0.9879	0.56
ResNet50*	0.2463	2.3208	0.0274	0.2811	1.1844	0.00

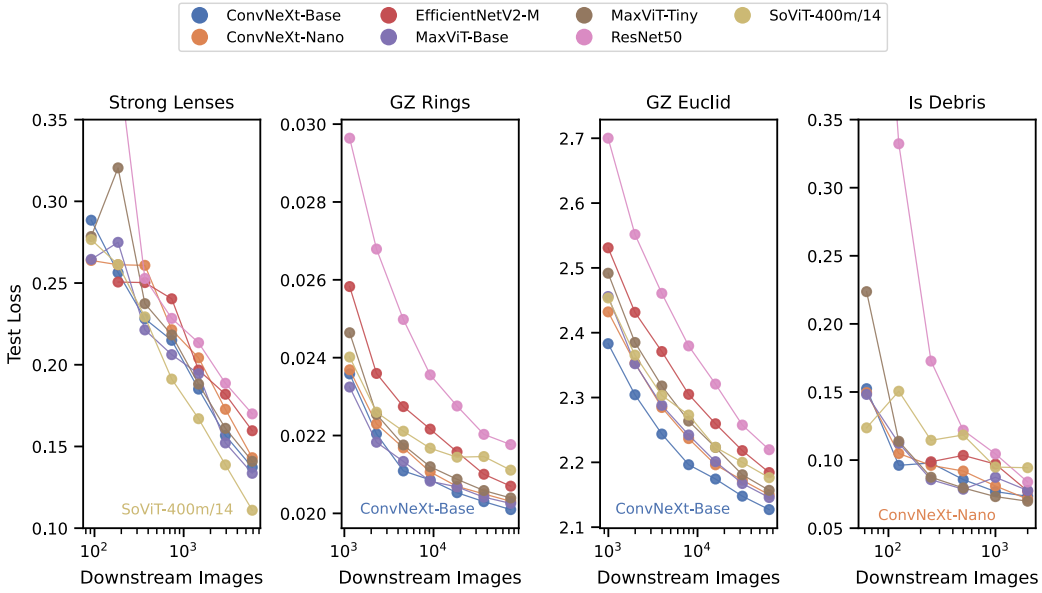

Figure 4: Test loss vs. dataset size when training on the GZ Evo Core dataset and then finetuning on the GZ Evo Downstream datasets. Mean of six seeds. ConvNeXt-Base performs best at GZ Rings, GZ Euclid, and overall. SoViT-400m/14 performs best at finding strong lenses. Small modern models (MaxViT-Tiny and ConvNeXt-Nano) perform best at our smaller Is Debris dataset.

to closely approximate professional annotations. However, for both volunteer and professional annotators, inter-annotator disagreement is expected. See Appendix C. Real scientific data often cannot be annotated perfectly by humans and operating under ambiguity is necessary for real-world machine learning systems.

For annotations using a decision tree structure (Core and GZ Euclid), the number of volunteers answering each question varies. Fine-grained questions deeper in the tree collect fewer votes[7]. This introduces heteroskedastic uncertainty and motivates our choice to share only high-confidence labels for our classification setting. The benefit of a decision tree is that annotators are only asked relevant questions; we hope our decision tree labels will support research into efficient crowdsourcing where finegrained labels are necessary but annotation is expensive (e.g. medical imaging).

We broadly find that the largest variants of the models tested tend to underperform their moderate-size variants, particularly for our smaller datasets. In principle, large variants should at least match the performance of small variants (e.g. by trivially replicating the small variant) and, where data is plentiful, are often shown to exceed the performance of smaller models. In practice, properly regularizing large models on small datasets is nontrivial, and our finetuning recipe likely causes these large models to overfit. Evaluating the effectiveness of new approaches for few-shot finetuning of large models is a key goal of GZ Evo.

GZ Evo is not suitable for research requiring raw telescope data. GZ Evo provides RGB images to support researchers without an astronomy background.

7 Data Access

Data is available via the HuggingFace Hub. You can download them with e.g.

```python
from datasets import load_dataset
dataset = load_dataset("mwalmsley/gz_evo", split="train")
# your framework of choice e.g. numpy, tensorflow, jax, etc
dataset.set_format("torch")
```

7.1 Temporary Note for Reviewers

Our code is public at `github.com/mwalmsley/gz_evo`. GZ Evo Core is public on HuggingFace. We temporarily apply gated access to the Downstream datasets while this paper is under review. Reviewers can anonymously access the gated datasets by logging in to the HuggingFace account:

Email: `yucehl1h0@mozmail.com`. Password: `Reviewer2!`

8 Conclusion

Galaxy Zoo Evo brings together sixteen years of effort to annotate one million images of galaxies. GZ Evo is large-scale, meticulously collected, and free of privacy, safety, or commercial concerns. It is designed for building and evaluating foundation models, and also suits broader research into practical methods for crowdsourcing, label uncertainty, multi-task learning, and active learning.

GZ Evo is a living dataset. New telescopes drive new discoveries; our modular multi-task format makes it straightforward to add new labelled galaxy images from any new telescopes. Researchers can then seamlessly apply the same training code to new GZ Evo versions, creating new telescope-ready finetuned models in weeks instead of years.

New telescope data is also ideal for testing the generalisation of foundation models. It is difficult to cleanly evaluate such models because any internet data, including standard benchmarks, is in-distribution (or worse, memorized training data). We know for certain that current foundation models could not have been trained on data from new telescopes, regardless of engineering resources, because the data did not previously exist. Generalization performance to new telescopes, as measured with GZ Evo, is a cast-iron test for existing models.

[7] For example, if we ask 40 people 'Is this galaxy smooth or featured?', and 25 answer 'Featured', then 15 further select 'Spiral', only those 15 people would be asked 'How many spiral arms?'

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848. ISSN: 1063-6919.

[2] Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on ImageNet transfer to real-world datasets? *dvances in Neural Information Processing Systems*, 36:25050–25080, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4eb33c53ed5b14ce9028309431f565cc- bstract-Datasets_and_Benchmarks.html.

[3] Lukas Tuggener, Jürgen Schmidhuber, and Thilo Stadelmann. Is it enough to optimize CNN architectures on ImageNet? *Frontiers in Computer Science*, 4:1041703, November 2022. ISSN 2624-9898. doi: 10.3389/fcomp.2022.1041703. URL http://arxiv.org/abs/2103.09108. arXiv:2103.09108 [cs].

[4] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet?, June 2019. URL http://arxiv.org/abs/1902.10811. arXiv:1902.10811 [cs, stat].

[5] Laurence Aitchison. a Statistical Theory of Cold Posteriors in Deep Neural Networks. *ICLR 2021 - 9th International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Rd138pWXMvG. arXiv: 2008.05912.

[6] Mike Walmsley, Micah Bowles, Anna M. M. Scaife, Jason Shingirai Makechemu, Alexander J. Gordon, Annette M. N. Ferguson, Robert G. Mann, James Pearson, Jürgen J. Popp, Jo Bovy, Josh Speagle, Hugh Dickinson, Lucy Fortson, Tobias Géron, Sandor Kruk, Chris J. Lintott, Kameswara Mantha, Devina Mohan, David O'Ryan, and Inigo V. Slijepevic. Scaling Laws for Galaxy Images, April 2024. URL https://arxiv.org/abs/2404.02973v1.

[7] George Stein, Jacqueline Blaum, Peter Harrington, Tomislav Medan, and Zarija Lukić. Mining for Strong Gravitational Lenses with Self-supervised Learning. *The strophysical Journal*, 932(2):107, 2022. ISSN 0004-637X. doi: 10.3847/1538-4357/ac6d63. URL http://arxiv.org/abs/2110.00023. arXiv: 2110.00023.

[8] Michael J. Smith, Ryan J. Roberts, Eirini Angeloudi, and Marc Huertas-Company. AstroPT: Scaling Large Observation Models for Astronomy, May 2024. URL http://arxiv.org/abs/2405.14930. arXiv:2405.14930 [astro-ph].

[9] Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, Ruben Ohana, Mariel Pettee, Bruno Régaldo-Saint Blancard, Kyunghyun Cho, Shirley Ho, and The Polymathic AI Collaboration. AstroCLIP: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal stronomical Society*, 531(4): 4990–5011, July 2024. ISSN 0035-8711. doi: 10.1093/mnras/stae1450. URL https://doi.org/10.1093/mnras/stae1450.

[10] Euclid Collaboration, M. Siudek, M. Huertas-Company, M. Smith, G. Martinez-Solaeche, F. Lanusse, S. Ho, E. Angeloudi, P. A. C. Cunha, H. Domínguez Sánchez, M. Dunn, Y. Fu, P. Iglesias-Navarro, J. Junais, J. H. Knapen, B. Laloux, M. Mezcua, W. Roster, G. Stevens, J. Vega-Ferrero, N. Aghanim, B. Altieri, A. Amara, S. Andreon, N. Auricchio, H. Aussel, C. Baccigalupi, M. Baldi, S. Bardelli, P. Battaglia, A. Biviano, A. Bonchi, E. Branchini, M. Brescia, J. Brinchmann, S. Camera, G. Cañas-Herrera, V. Capobianco, C. Carbone, J. Carretero, S. Casas, F. J. Castander, M. Castellano, G. Castignani, S. Cavuoti, K. C. Chambers, A. Cimatti, C. Colodro-Conde, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, F. Courbin, H. M. Courtois, M. Cropper, A. Da Silva, H. Degaudenzi, G. De Lucia, A. M. Di Giorgio, J. Dinis, C. Dolding, H. Dole, F. Dubath, C. A. J. Duncan, X. Dupac, S. Dusini, S. Escoffier, M. Farina, R. Farinelli, F. Faustini, S. Ferriol, F. Finelli, S. Fotopoulou, M. Frailis, E. Franceschi, S. Galeotta, K. George, B. Gillis, C. Giocoli, J. Gracia-Carpio, B. R. Granett, A. Grazian, F. Grupp, S. Gwyn, S. V. H. Haugan, W. Holmes, I. M. Hook, F. Hormuth, A. Hornstrup, K. Jahnke, M. Jhabvala, E. Keih nen, S. Kermiche, A. Kiessling, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, Q. Le Boulc'h, A. M. C. Le Brun, D. Le Mignant, S. Ligori, P. B. Lilje, V. Lindholm, I. Lloro, G. Mainetti, D. Maino, E. Maiorano, O. Mansutti, S. Marcin, O. Marggraf, M. Martinelli, N. Martinet, F. Marulli, R. Massey, S. Maurogordato, H. J. McCracken, E. Medinaceli, S. Mei, M. Melchior, Y. Mellier, M. Meneghetti, E. Merlin, G. Meylan, A. Mora, M. Moresco, L. Moscardini, R. Nakajima, C. Neissner, S.-M. Niemi, J. W. Nightingale, C. Padilla, S. Paltani, F. Pasian, K. Pedersen, W. J. Percival, V. Pettorino, S. Pires, G. Polenta, M. Poncet, L. A. Popa, L. Pozzetti, F. Raison, A. Renzi, J. Rhodes, G. Riccio, E. Romelli, M. Roncarelli, R. Saglia, Z. Sakr, A. G. Sánchez, D. Sapone, B. Sartoris, J. A. Schewtschenko, P. Schneider, T. Schrabback, M. Scodeggio, A. Secroun, G. Seidel, M. Seiffert, S. Serrano, P. Simon, C. Sirignano, G. Sirri, L. Stanco,

J. Steinwagner, P. Tallada-Crespí, A. N. Taylor, I. Tereno, S. Toft, R. Toledo-Moreo, F. Torradeflot, I. Tutusaus, L. Valenziano, J. Valiviita, T. Vassallo, G. Verdoes Kleijn, A. Veropalumbo, Y. Wang, J. Weller, A. Zacchei, G. Zamorani, F. M. Zerbi, I. A. Zinchenko, E. Zucca, V. Allevato, M. Ballardini, M. Bolzonella, E. Bozzo, C. Burigana, R. Cabanac, A. Cappi, D. Di Ferdinando, J. A. Escartin Vigo, L. Gabarra, J. Martín-Fleitas, S. Matthew, N. Mauri, R. B. Metcalf, A. Pezzotta, M. Pöntinen, C. Porciani, I. Risso, V. Scottez, M. Sereno, M. Tenti, M. Viel, M. Wiesmann, Y. Akrami, I. T. Andika, S. Anselmi, M. Archidiacono, F. Atrio-Barandela, C. Benoist, K. Benson, D. Bertacca, M. Bethermin, L. Bisigello, A. Blanchard, L. Blot, M. L. Brown, S. Bruton, A. Calabro, B. Camacho Quevedo, F. Caro, C. S. Carvalho, T. Castro, Y. Charles, F. Cogato, A. R. Cooray, O. Cucciati, S. Davini, F. De Paolis, G. Desprez, A. Díaz-Sánchez, J. J. Diaz, S. Di Domizio, J. M. Diego, P.-A. Duc, A. Enia, Y. Fang, A. G. Ferrari, P. G. Ferreira, A. Finoguenov, A. Fontana, A. Franco, K. Ganga, J. García-Bellido, T. Gasparetto, V. Gautard, E. Gaztanaga, F. Giacomini, F. Gianotti, G. Gozaliasl, M. Guidi, C. M. Gutierrez, A. Hall, W. G. Hartley, S. Hemmati, C. Hernández-Monteagudo, H. Hildebrandt, J. Hjorth, J. J. E. Kajava, Y. Kang, V. Kansal, D. Karagiannis, K. Kiiveri, C. C. Kirkpatrick, S. Kruk, J. Le Graet, L. Legrand, M. Lembo, F. Lepori, G. Leroy, G. F. Lesci, J. Lesgourgues, L. Leuzzi, T. I. Liaudat, A. Loureiro, J. Macias-Perez, G. Maggio, M. Magliocchetti, E. A. Magnier, F. Mannucci, R. Maoli, C. J. A. P. Martins, L. Maurin, M. Miluzio, P. Monaco, C. Moretti, G. Morgante, C. Murray, K. Naidoo, A. Navarro-Alsina, S. Nesseris, F. Passalacqua, K. Paterson, L. Patrizii, A. Pisani, D. Potter, S. Quai, M. Radovich, S. Sacquegna, M. Sahlén, D. B. Sanders, E. Sarpa, A. Schneider, D. Sciotti, D. Scognamiglio, E. Sellentin, L. C. Smith, K. Tanidis, G. Testera, R. Teyssier, S. Tosi, A. Troja, M. Tucci, C. Valieri, A. Venhola, D. Vergani, G. Verza, P. Vielzeuf, N. A. Walton, and J. G. Sorce. Euclid Quick Data Release (Q1) Exploring galaxy properties with a multi-modal foundation model, March 2025. URL http://arxiv.org/abs/2503.15312. arXiv:2503.15312 [astro-ph].

[11] Karen L. Masters. Twelve years of Galaxy Zoo. In *Proceedings of the International stronomical Union*, volume 14, pages 205–212, October 2019. doi: 10.1017/S1743921319008615. URL http://arxiv.org/abs/1910.08177. arXiv: 1910.08177 Issue: S353 ISSN: 17439221.

[12] Ronald J. Buta. Galaxy morphology. In *Planets, Stars and Stellar Systems: Volume 6: Extragalactic stronomy and Cosmology*, pages 1–89. Springer Netherlands, January 2013. ISBN 978-94-007-5609-0. doi: 10.1007/978-94-007-5609-0_1. URL https://link.springer.com/referenceworkentry/10.1007/978-94-007-5609-0_1. arXiv: 1304.3529.

[13] S. Dieleman, K. W. Willett, and J. Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal stronomical Society*, 450(2):1441–1459, 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv632. URL http://arxiv.org/abs/1503.07077. arXiv: 1503.07077.

[14] M. Huertas-Company and F. Lanusse. The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys. *Publications of the stronomical Society of ustralia*, 40, October 2023. ISSN 14486083. doi: 10.1017/pasa.2022.55. URL http://arxiv.org/abs/2210.01813. arXiv: 2210.01813.

[15] Reza Katebi, Yadi Zhou, Ryan Chornock, and Razvan Bunescu. Galaxy morphology prediction using Capsule Networks. *Monthly Notices of the Royal stronomical Society*, 486(2):1539–1547, June 2019. ISSN 13652966. doi: 10.1093/mnras/stz915. URL https://academic.oup.com/mnras/article/486/2/1539/5424782. arXiv: 1809.08377 Publisher: Narnia.

[16] Yevonnael Andrew. Galaxy Classification Using Transfer Learning and Ensemble of CNNs With Multiple Colour Spaces. 2023. doi: 10.48550/ARXIV.2305.00002. URL https://arxiv.org/abs/2305.00002. Publisher: arXiv Version Number: 1.

[17] Guangping Li, Tingting Xu, Liping Li, Xianjun Gao, Zhijing Liu, Jie Cao, Mingcun Yang, and Weihong Zhou. Galaxy morphology classification using multiscale convolution capsule network. *Monthly Notices of the Royal stronomical Society*, 523(1):488–497, May 2023. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stad854. URL https://academic.oup.com/mnras/article/523/1/488/7179754.

[18] Sneh Pandya, Purvik Patel, Franc O, and Jonathan Blazek. E(2) Equivariant Neural Networks for Robust Galaxy Morphology Classification. 2023. doi: 10.48550/ARXIV.2311.01500. URL https://arxiv.org/abs/2311.01500. Publisher: arXiv Version Number: 1.

[19] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. In *dvances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91- bstract.html.

[20] Keno K. Bressem, Lisa C. Adams, Christoph Erxleben, Bernd Hamm, Stefan M. Niehues, and Janis L. Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *Scientific Reports*, 10(1):13590, August 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-70479-z. URL https://www.nature.com/articles/s41598-020-70479-z. Publisher: Nature Publishing Group.

[21] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y. Ng, and Pranav Rajpurkar. CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, pages 116–124, New York, NY, USA, April 2021. Association for Computing Machinery. ISBN 978-1-4503-8359-2. doi: 10.1145/3450439.3451867. URL https://dl.acm.org/doi/10.1145/3450439.3451867.

[22] Piyapoj Kasempakdeepong, Pondsulee Ponchaiyapruek, Pattamon Viriyothai, Anuwat Songchumrong, Pittipol Kantavat, and Prasertsak Pungprasertying. Sugarcane Classification for On-Site Assessment Using Computer Vision. *2022 17th International Joint Symposium on rtificial Intelligence and Natural Language Processing (iS I-NLP)*, pages 1–6, November 2022. doi: 10.1109/iSAI-NLP56921.2022.9960252. URL https://ieeexplore.ieee.org/document/9960252/. Conference Name: 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) ISBN: 9781665457279 Place: Chiang Mai, Thailand Publisher: IEEE.

[23] Ahmad Mohamad Mezher and Andrew E. Marble. A Novel Strategy for Improving Robustness in Computer Vision Manufacturing Defect Detection. 2023. doi: 10.48550/ARXIV.2305.09407. URL https://arxiv.org/abs/2305.09407. Publisher: arXiv Version Number: 1.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *Open I*, page 47, 2021. URL http://arxiv.org/abs/2103.00020. arXiv: 2103.00020.

[25] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling Vision Transformers, June 2022. URL http://arxiv.org/abs/2106.04560. arXiv:2106.04560 [cs].

[26] Aishwarya Agrawal, Damien Teney, and Aida Nematzadeh. Vision-language pretraining: Current trends and the future. In *nnual Meeting of the ssociation for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:248780422.

[27] Shuai Xiao, Yangyang Zhang, Liming Jiang, and Zhengxia Wang. Asimsa: Advanced semantic information guided multi-scale alignment framework for medical vision-language pretraining. *2024 IEEE 9th International Conference on Computational Intelligence and pplications (ICCI)*, pages 99–103, 2024. URL https://api.semanticscholar.org/CorpusID:273534365.

[28] Siyuan Yan, Xieji Li, Ming Hu, Yiwen Jiang, Zhen Yu, and Zongyuan Ge. Make: Multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment. 2025. URL https://api.semanticscholar.org/CorpusID:278602553.

[29] Bo Liu, Donghuan Lu, Dong Wei, Xianming Wu, Yan Wang, Yu Zhang, and Yefeng Zheng. Improving medical vision-language contrastive pretraining with semantics-aware triage. *IEEE Transactions on Medical Imaging*, 42:3579–3589, 2023. URL https://api.semanticscholar.org/CorpusID:259856982.

[30] Yeongjae Cho, Taehee Kim, Heejun Shin, Sungzoon Cho, and Dongmyung Shin. Pretraining vision-language model for difference visual question answering in longitudinal chest x-rays. *rXiv*, abs/2402.08966, 2024. URL https://api.semanticscholar.org/CorpusID:267657514.

[31] Dimitrios Tanoglidis and Bhuvnesh Jain. At first sight! zero-shot classification of astronomical images with large multimodal models. *Research Notes of the S*, 8, 2024. URL https://iopscience.iop.org/article/10.3847/2515-5172/ad887a.

[32] Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W. Willett, Steven Bamford, Lee S. Kelvin, Lucy Fortson, Yarin Gal, William Keel, Karen L. Masters, Vihang Mehta, Brooke D. Simmons, Rebecca Smethurst, Lewis Smith, Elisabeth M. Baeten, and Christine MacMillan. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *Monthly Notices of the Royal stronomical Society*, 509(3):3966–3988, December 2022. ISSN 13652966. doi: 10.1093/mnras/stab2093. URL https://doi.org/10.1093/mnras/stab2093. arXiv: 2102.08414.

[33] Mike Walmsley, Tobias Géron, Sandor Kruk, Anna M. M. Scaife, Chris Lintott, Karen L. Masters, James M. Dawson, Hugh Dickinson, Lucy Fortson, Izzy L. Garland, Kameswara Mantha, David O'Ryan, Jürgen Popp, Brooke Simmons, Elisabeth M. Baeten, and Christine Macmillan. Galaxy Zoo DESI: Detailed Morphology Measurements for 8.7M Galaxies in the DESI Legacy Imaging Surveys. *MNR S*, 000:1–18, September 2023. URL http://arxiv.org/abs/2309.11425. arXiv: 2309.11425 ISBN: 2309.11425v1.

[34] Mike Walmsley, Campbell Allen, Ben Aussel, Micah Bowles, Kasia Gregorowicz, Inigo Val Slijepcevic, Chris J. Lintott, Anna M. M. Scaife, Maja Jabłońska, Kosio Karchev, Denise Lanzieri, Devina Mohan, David O'Ryan, Bharath Saiguhan, Crisel Suárez, Nicolás Guerra-Varas, and Renuka Velu. Zoobot: Adaptable Deep Learning Models for Galaxy Morphology. *Journal of Open Source Software*, 8(85):5312, 2023. doi: 10.21105/joss.05312. URL https://doi.org/10.21105/joss.05312. Publisher: The Open Journal.

[35] Eirini Angeloudi, Jeroen Audenaert, Micah Bowles, Benjamin M. Boyd, David Chemaly, Brian Cherinka, Ioana Ciucă, Miles Cranmer, Aaron Do, Matthew Grayling, Erin E. Hayes, Tom Hehir, Shirley Ho, Marc Huertas-Company, Kartheik G. Iyer, Maja Jablonska, Francois Lanusse, Henry W. Leung, Kaisey Mandel, Juan Rafael Martínez-Galarza, Peter Melchior, Lucas Meyer, Liam H. Parker, Helen Qu, Jeff Shen, Michael J. Smith, Connor Stone, Mike Walmsley, and John F. Wu. The Multimodal Universe: Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *dvances in Neural Information Processing Systems*, volume 37, pages 57841–57913. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6a57493d35fefea59d06396c7cb69228-Paper-Datasets_and_Benchmarks_Track.pdf.

[36] Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, Karen L. Masters, Brooke D. Simmons, Kevin R V Casteels, Edward M. Edmondson, Lucy F. Fortson, Sugata Kaviraj, William C. Keel, Thomas Melvin, Robert C. Nichol, M. Jordan Raddick, Kevin Schawinski, Robert J. Simpson, Ramin A. Skibba, Arfon M. Smith, and Daniel Thomas. Galaxy Zoo 2: Detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal stronomical Society*, 435(4):2835–2860, 2013. ISSN 00358711. doi: 10.1093/mnras/stt1458. arXiv: 1308.3496v2 ISBN: doi:10.1093/mnras/stt1458.

[37] Kyle W. Willett, Melanie A. Galloway, Steven P. Bamford, Chris J. Lintott, Karen L. Masters, Claudia Scarlata, B. D. Simmons, Melanie Beck, Carolin N. Cardamone, Edmond Cheung, Edward M. Edmondson, Lucy F. Fortson, Roger L. Griffith, Boris H ußler, Anna Han, Ross Hart, Thomas Melvin, Michael Parrish, Kevin Schawinski, R. J. Smethurst, and Arfon M. Smith. Galaxy Zoo: Morphological classifications for 120 000 galaxies in HST legacy imaging. *Monthly Notices of the Royal stronomical Society*, 464(4):4176–4203, February 2017. ISSN 13652966. doi: 10.1093/mnras/stw2568. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stw2568. arXiv: 1610.03068 Publisher: Oxford University Press.

[38] Karen L. Masters, Melanie Galloway, Lucy Fortson, Chris Lintott, Mike Read, Claudia Scarlata, Brooke Simmons, Mike Walmsley, and Kyle Willett. Galaxy Zoo: Morphologies based on UKIDSS NIR Imaging for 71,052 Galaxies. *Research Notes of the S*, 8(8):198, August 2024. ISSN 2515-5172. doi: 10.3847/2515-5172/ad6f10. URL http://arxiv.org/abs/2408.10160. arXiv:2408.10160 [astro-ph].

[39] B. D. Simmons, Chris Lintott, Kyle W. Willett, Karen L. Masters, Jeyhan S. Kartaltepe, Boris H ußler, Sugata Kaviraj, Coleman Krawczyk, S. J. Kruk, Daniel H. McIntosh, R. J. Smethurst, Robert C. Nichol, Claudia Scarlata, Kevin Schawinski, Christopher J. Conselice, Omar Almaini, Henry C. Ferguson, Lucy Fortson, William Hartley, Dale Kocevski, Anton M. Koekemoer, Alice Mortlock, Jeffrey A. Newman, Steven P. Bamford, N. A. Grogin, Ray A. Lucas, Nimish P. Hathi, Elizabeth McGrath, Michael Peth, Janine Pforr, Zachary Rizer, Stijn Wuyts, Guillermo Barro, Eric F. Bell, Marco Castellano, Tomas Dahlen, Avishai Dekel, Jamie Ownsworth, Sandra M. Faber, Steven L. Finkelstein, Adriano Fontana, Audrey Galametz, Ruth Grützbauch, David Koo, Jennifer Lotz, Bahram Mobasher, Mark Mozena, Mara Salvato, and Tommy Wiklind. Galaxy Zoo: Quantitative visual morphological classifications for 48 000 galaxies from CANDELS. *Monthly Notices of the Royal stronomical Society*, 464(4):4420–4447, 2017. ISSN 13652966. doi: 10.1093/mnras/stw2587. arXiv: 1610.03070v1.

[40] H. Euclid Collaboration: Aussel, I. Tereno, M. Schirmer, et al. Euclid quick data release (q1): Data release overview. *stronomy and strophysics, submitted*, 2025.

[41] M. Euclid Collaboration: Walmsley, M. Huertas-Company, L. Quilley, et al. Euclid quick data release (q1): First visual morphology catalogue. *stronomy and strophysics, submitted*, 2025.

[42] Shajib, A.J., Vernardos, G., Collett, T.E. Strong lensing by galaxies. *Space Science Reviews*, 220(87), 2024.

[43] M. Euclid Collaboration: Walmsley, P. Holloway, N.E.P. Lines, et al. Euclid quick data release (q1): The strong lensing discovery engine a -o system overview and first lens sample. *stronomy and strophysics, submitted*, 2025.

[44] K. Euclid Collaboration: Rojas, T. Collett, J. Acevedo Barroso, et al. Euclid quick data release (q1): The strong lensing discovery engine b – early strong lens candidates from visual inspection of high velocity dispersion galaxies. *stronomy and strophysics, submitted*, 2025.

[45] N. E. P. Euclid Collaboration: Lines, T. E. Collett, M. Walmsley, et al. Euclid quick data release (q1): The strong lensing discovery engine c – finding lenses with machine learning. *stronomy and strophysics, submitted*, 2025.

[46] T. Euclid Collaboration: Li, T. Collett, M. Walmsley, et al. Euclid quick data release (q1): The strong lensing discovery engine d – double-source-plane lens candidates and cosmological forecast. *stronomy and strophysics, submitted*, 2025.

[47] P. Euclid Collaboration: Holloway, A. Verma, M. Walmsley, et al. Euclid quick data release (q1): The strong lensing discovery engine e – ensemble classification of strong gravitational lenses: lessons for data release 1. *stronomy and strophysics, submitted*, 2025.

[48] Ronald J. Buta. Galactic rings revisited - I. CVRHS classifications of 3962 ringed galaxies from the Galaxy Zoo 2 Database. *Monthly Notices of the Royal stronomical Society*, 471(4):4027–4046, November 2017. ISSN 13652966. doi: 10.1093/MNRAS/STX1829. URL `http://s3.amazonaws.com/zoo2/filen.jpg,`. arXiv: 1707.06589 Publisher: Oxford University Press.

[49] M. M. Pawlik, V. Wild, C. J. Walcher, P. H. Johansson, C. Villforth, K. Rowlands, J. Mendez-Abreu, and T. Hewlett. Shape asymmetry: A morphological indicator for automatic detection of galaxies in the post-coalescence merger stages. *Monthly Notices of the Royal stronomical Society*, 456(3):3032–3052, 2016. ISSN 13652966. doi: 10.1093/mnras/stv2878. arXiv: 1512.02000.

[50] Connor Bottrell, Maan H. Hani, Hossen Teimoorinia, Sara L. Ellison, Jorge Moreno, Paul Torrey, Christopher C. Hayward, Mallory Thorp, Luc Simard, and Lars Hernquist. Deep learning predictions of galaxy merger stage and the importance of observational realism. *Monthly Notices of the Royal stronomical Society*, 490(4):5390–5413, October 2019. ISSN 13652966. doi: 10.1093/mnras/stz2934. URL `http://arxiv.org/abs/1910.07031`. arXiv: 1910.07031.

[51] W. J. Pearson, L. Wang, J. W. Trayford, C. E. Petrillo, and F. F.S. van der Tak. Identifying Galaxy Mergers in Observations and Simulations with Deep Learning. *stronomy & strophysics*, 626(49), February 2019. ISSN 0004-6361. doi: 10.1051/0004-6361/201935355. arXiv: 1902.10626 Publisher: arXiv.

[52] Alexander J. Gordon, Annette M. N. Ferguson, and Robert G. Mann. Uncovering tidal treasures: Automated classification of faint tidal features in DECaLS data. URL `http://arxiv.org/abs/2404.06487`.

[53] Arjun Dey, David J. Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, Xiaohui Fan, Joseph R. Findlay, Doug Finkbeiner, David Herrera, Stéphanie Juneau, Martin Landriau, Michael Levi, Ian McGreer, Aaron Meisner, Adam D. Myers, John Moustakas, Peter Nugent, Anna Patej, Edward F. Schlafly, Alistair R. Walker, Francisco Valdes, Benjamin A. Weaver, Christophe Yèche, Hu Zou, Xu Zhou, Behzad Abareshi, T. M. C. Abbott, Bela Abolfathi, C. Aguilera, Shadab Alam, Lori Allen, A. Alvarez, James Annis, Behzad Ansarinejad, Marie Aubert, Jacqueline Beechert, Eric F. Bell, Segev Y. BenZvi, Florian Beutler, Richard M. Bielby, Adam S. Bolton, César Briceño, Elizabeth J. Buckley-Geer, Karen Butler, Annalisa Calamida, Raymond G. Carlberg, Paul Carter, Ricard Casas, Francisco J. Castander, Yumi Choi, Johan Comparat, Elena Cukanovaite, Timothée Delubac, Kaitlin DeVries, Sharmila Dey, Govinda Dhungana, Mark Dickinson, Zhejie Ding, John B. Donaldson, Yutong Duan, Christopher J. Duckworth, Sarah Eftekharzadeh, Daniel J. Eisenstein, Thomas Etourneau, Parker A. Fagrelius, Jay Farihi, Mike Fitzpatrick, Andreu Font-Ribera, Leah Fulmer, Boris T. G nsicke, Enrique Gaztanaga, Koshy George, David W. Gerdes, Satya Gontcho A Gontcho, Claudio Gorgoni, Gregory Green, Julien Guy, Diane Harmer, M. Hernandez, Klaus Honscheid, Lijuan (Wendy) Huang, David J. James, Buell T. Jannuzi, Linhua Jiang, Richard Joyce, Armin Karcher, Sonia Karkar, Robert Kehoe, Jean-Paul Kneib, Andrea Kueter-Young, Ting-Wen Lan, Tod R. Lauer, Laurent Le Guillou, Auguste Le Van Suu, Jae Hyeon Lee, Michael Lesser, Laurence Perreault Levasseur, Ting S. Li, Justin L. Mann, Robert Marshall, C. E. Martínez-Vázquez, Paul Martini, Hélion du Mas des Bourboux, Sean McManus, Tobias Gabriel Meier, Brice Ménard, Nigel Metcalfe, Andrea Muñoz-Gutiérrez, Joan Najita, Kevin Napier, Gautham Narayan, Jeffrey A. Newman, Jundan Nie, Brian Nord, Dara J. Norman, Knut A. G. Olsen, Anthony Paat, Nathalie Palanque-Delabrouille, Xiyan Peng, Claire L. Poppett, Megan R. Poremba, Abhishek Prakash, David Rabinowitz, Anand Raichoor, Mehdi Rezaie, A. N. Robertson, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Gregory Rudnick, Sasha Safonova, Abhijit Saha, F. Javier Sánchez, Elodie Savary, Heidi Schweiker, Adam Scott, Hee-Jong Seo, Huanyuan Shan, David R. Silva, Zachary Slepian, Christian Soto, David Sprayberry, Ryan Staten, Coley M. Stillman, Robert J. Stupak, David L. Summers, Suk Sien Tie, H. Tirado, Mariana Vargas-Magaña, A. Katherina Vivas, Risa H. Wechsler, Doug Williams, Jinyi Yang, Qian Yang, Tolga Yapici, Dennis Zaritsky, A. Zenteno, Kai Zhang, Tianmeng Zhang, Rongpu Zhou, and Zhimin Zhou. Overview of the DESI Legacy Imaging Surveys. *The stronomical Journal*, 157(5):168, April 2019. ISSN 00046256. doi: 10.3847/1538-3881/ab089d. URL `http://arxiv.org/abs/1804.08657`. arXiv: 1804.08657.

[54] Adam M. Atkinson, Roberto G. Abraham, and Annette M N Ferguson. Faint tidal features in galaxies within the Canada-France-Hawaii telescope legacy survey wide fields. *strophysical Journal*, 765(1), 2013. ISSN 15384357. doi: 10.1088/0004-637X/765/1/28. arXiv: 1301.4275.

[55] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. January 2022. URL http://arxiv.org/abs/2201.03545. arXiv: 2201.03545.

[56] Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller Models and Faster Training. April 2021. URL http://arxiv.org/abs/2104.00298. arXiv: 2104.00298.

[57] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis Vision Transformer. In *Lecture Notes in Computer Science (including subseries Lecture Notes in rtificial Intelligence and Lecture Notes in Bioinformatics)*, volume 13684 LNCS, pages 459–479, April 2022. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_27. URL https://arxiv.org/abs/2204.01697v4. arXiv: 2204.01697 ISSN: 16113349.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778. IEEE, June 2016. ISBN 978-1-4673-8850-4. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/. arXiv: 1512.03385 ISSN: 10636919.

[59] Ross Wightman, Hugo Touvron, and Hervé Jégou. ResNet strikes back: An improved training procedure in timm, October 2021. URL http://arxiv.org/abs/2110.00476. arXiv:2110.00476 [cs].

[60] Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting ViT in Shape: Scaling Laws for Compute-Optimal Model Design, January 2024. URL http://arxiv.org/abs/2305.13035. arXiv:2305.13035 [cs].

[61] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019. URL http://arxiv.org/abs/1711.05101. arXiv:1711.05101 [cs, math].

[62] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep Networks with Stochastic Depth, July 2016. URL http://arxiv.org/abs/1603.09382. arXiv:1603.09382 [cs].

[63] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A Jointly-Scaled Multilingual Language-Image Model, June 2023. URL http://arxiv.org/abs/2209.06794. arXiv:2209.06794 [cs].

[64] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, February 2025. URL http://arxiv.org/abs/2502.14786. arXiv:2502.14786 [cs].

[65] Kevork N. Abazajian, Jennifer K. Adelman-McCarthy, Marcel A. Agüeros, Sahar S. Allam, Carlos Allende Prieto, Deokkeun An, Kurt S. J. Anderson, Scott F. Anderson, James Annis, Neta A. Bahcall, C. A. L. Bailer-Jones, J. C. Barentine, Bruce A. Bassett, Andrew C. Becker, Timothy C. Beers, Eric F. Bell, Vasily Belokurov, Andreas A. Berlind, Eileen F. Berman, Mariangela Bernardi, Steven J. Bickerton, Dmitry Bizyaev, John P. Blakeslee, Michael R. Blanton, John J. Bochanski, William N. Boroski, Howard J. Brewington, Jarle Brinchmann, J. Brinkmann, Robert J. Brunner, Tamás Budavári, Larry N. Carey, Samuel Carliles, Michael A. Carr, Francisco J. Castander, David Cinabro, A. J. Connolly, István Csabai, Carlos E. Cunha, Paul C. Czarapata, James R. A. Davenport, Ernst de Haas, Ben Dilday, Mamoru Doi, Daniel J. Eisenstein, Michael L. Evans, N. W. Evans, Xiaohui Fan, Scott D. Friedman, Joshua A. Frieman, Masataka Fukugita, Boris T. G nsicke, Evalyn Gates, Bruce Gillespie, G. Gilmore, Belinda Gonzalez, Carlos F. Gonzalez, Eva K. Grebel, James E. Gunn, Zsuzsanna Györy, Patrick B. Hall, Paul Harding, Frederick H. Harris, Michael Harvanek, Suzanne L. Hawley, Jeffrey J. E. Hayes, Timothy M. Heckman, John S. Hendry, Gregory S. Hennessy, Robert B. Hindsley, J. Hoblitt, Craig J. Hogan, David W. Hogg, Jon A. Holtzman, Joseph B. Hyde, Shin-ichi Ichikawa, Takashi Ichikawa, Myungshin Im, Zeljko Ivezic, Sebastian Jester, Linhua Jiang, Jennifer A. Johnson, Anders M. Jorgensen, Mario Jurić, Stephen M. Kent, R. Kessler, S. J. Kleinman, G. R. Knapp, Kohki Konishi, Richard G. Kron, Jurek Krzesinski, Nikolay Kuropatkin, Hubert Lampeitl, Svetlana Lebedeva, Myung Gyoon Lee, Young Sun Lee, R. French Leger, Sébastien Lépine, Nolan Li, Marcos Lima, Huan Lin, Daniel C. Long, Craig P. Loomis, Jon Loveday, Robert H. Lupton, Eugene Magnier, Olena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Bruce Margon, John P. Marriner, David Martínez-Delgado, Takahiko Matsubara, Peregrine M. McGehee, Timothy A. McKay, Avery Meiksin, Heather L. Morrison, Fergal Mullally, Jeffrey A. Munn, Tara Murphy, Thomas Nash, Ada Nebot, Eric H. Neilsen, Heidi Jo Newberg, Peter R. Newman, Robert C. Nichol,

Tom Nicinski, Maria Nieto-Santisteban, Atsuko Nitta, Sadanori Okamura, Daniel J. Oravetz, Jeremiah P. Ostriker, Russell Owen, Nikhil Padmanabhan, Kaike Pan, Changbom Park, George Pauls, John Peoples, Will J. Percival, Jeffrey R. Pier, Adrian C. Pope, Dimitri Pourbaix, Paul A. Price, Norbert Purger, Thomas Quinn, M. Jordan Raddick, Paola Re Fiorentin, Gordon T. Richards, Michael W. Richmond, Adam G. Riess, Hans-Walter Rix, Constance M. Rockosi, Masao Sako, David J. Schlegel, Donald P. Schneider, Ralf-Dieter Scholz, Matthias R. Schreiber, Axel D. Schwope, Uroš Seljak, Branimir Sesar, Erin Sheldon, Kazu Shimasaku, Valena C. Sibley, A. E. Simmons, Thirupathi Sivarani, J. Allyn Smith, Martin C. Smith, Vernesa Smolčić, Stephanie A. Snedden, Albert Stebbins, Matthias Steinmetz, Chris Stoughton, Michael A. Strauss, Mark SubbaRao, Yasushi Suto, Alexander S. Szalay, István Szapudi, Paula Szkody, Masayuki Tanaka, Max Tegmark, Luis F. A. Teodoro, Aniruddha R. Thakar, Christy A. Tremonti, Douglas L. Tucker, Alan Uomoto, Daniel E. Vanden Berk, Jan Vandenberg, S. Vidrih, Michael S. Vogeley, Wolfgang Voges, Nicole P. Vogt, Yogesh Wadadekar, Shannon Watters, David H. Weinberg, Andrew A. West, Simon D. M. White, Brian C. Wilhite, Alainna C. Wonders, Brian Yanny, D. R. Yocum, Donald G. York, Idit Zehavi, Stefano Zibetti, and Daniel B. Zucker. the Seventh Data Release of the Sloan Digital Sky Survey. *The strophysical Journal Supplement Series*, 182(2):543–558, 2009. ISSN 0067-0049. doi: 10.1088/0067-0049/182/2/543. URL `http://stacks.iop.org/0067-0049/182/i=2/a=543?key=crossref.bc07496fb06b943bcf82755687fa84b4`.

[66] Mike Walmsley, Lewis Smith, Chris Lintott, Yarin Gal, Steven Bamford, Hugh Dickinson, Lucy Fortson, Sandor Kruk, Karen Masters, Claudia Scarlata, Brooke Simmons, Rebecca Smethurst, and Darryl Wright. Galaxy Zoo: Probabilistic Morphology through Bayesian CNNs and Active Learning. *Monthly Notices of the Royal stronomical Society*, 491(2):1554–1574, January 2020. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stz2816. URL `http://arxiv.org/abs/1905.07424`. arXiv:1905.07424 [astro-ph].

[67] A. Lawrence, S. J. Warren, O. Almaini, A. C. Edge, N. C. Hambly, R. F. Jameson, P. Lucas, M. Casali, A. Adamson, S. Dye, J. P. Emerson, S. Foucaud, P. Hewett, P. Hirst, S. T. Hodgkin, M. J. Irwin, N. Lodieu, R. G. McMahon, C. Simpson, I. Smail, D. Mortlock, and M. Folger. The UKIRT infrared deep sky survey (UKIDSS). *Monthly Notices of the Royal stronomical Society*, 379(4):1599–1617, August 2007. ISSN 13652966. doi: 10.1111/j.1365-2966.2007.12040.x. URL `https://academic.oup.com/mnras/article-lookup/doi/10.1111/j.1365-2966.2007.12040.x`. arXiv: astro-ph/0604426 Publisher: Oxford University Press ISBN: 0035-8711.

[68] M. Jordan Raddick, Georgia Bracey, Pamela L. Gay, Chris J. Lintott, Carie Cardamone, Phil Murray, Kevin Schawinski, Alexander S. Szalay, and Jan Vandenberg. Galaxy Zoo: Motivations of Citizen Scientists, March 2013. URL `http://arxiv.org/abs/1303.6886`. arXiv:1303.6886 [astro-ph, physics:physics].

[69] Corey Jackson, Liz Dowthwaite, Ellie Jeong, Laura Trouille, Lucy Fortson, Chris Lintott, Brooke Simmons, and Grant Miller. Unleashing the Power of the Zooniverse: The 2021 Survey of Volunteers, May 2024. URL `https://papers.ssrn.com/abstract=4830179`.

[70] Eunmi (Ellie) Jeong, Corey Jackson, Liz Dowthwaite, Cliff Johnson, and Laura Trouille. How Personal Value Orientations Influence Behaviors in Digital Citizen Science. *Proceedings of the CM on Human-Computer Interaction*, 8(CSCW1):64:1–64:25, April 2024. doi: 10.1145/3637341. URL `https://dl.acm.org/doi/10.1145/3637341`.

[71] Edwin Simpson, Stephen Roberts, Ioannis Psorakis, and Arfon Smith. Dynamic bayesian combination of multiple imperfect classifiers. *Studies in Computational Intelligence*, 474:1–35, 2013. ISSN 1860949X. doi: 10.1007/978-3-642-36406-8_1. arXiv: 1206.1831 ISBN: 9783642364051.

[72] Edith Law, Krzysztof Z. Gajos, Andrea Wiggins, Mary L. Gray, and Alex Williams. Crowdsourcing as a Tool for Research: Implications of Uncertainty. *Proceedings of the 2017 CM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1544–1561, February 2017. doi: 10.1145/2998181.2998197. URL `https://dl.acm.org/doi/10.1145/2998181.2998197`. Conference Name: CSCW '17: Computer Supported Cooperative Work and Social Computing ISBN: 9781450343350 Place: Portland Oregon USA Publisher: ACM.

[73] Constance Thierry, Jean-Christophe Dubois, Yolande Le Gall, and Arnaud Martin. Modeling Uncertainty and Inaccuracy on Data from Crowdsourcing Platforms: MONITOR. *2019 IEEE 31st International Conference on Tools with rtificial Intelligence (ICT I)*, pages 776–783, November 2019. doi: 10.1109/ICTAI.2019.00112. URL `https://ieeexplore.ieee.org/document/8995266/`. Conference Name: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) ISBN: 9781728137988 Place: Portland, OR, USA Publisher: IEEE.

[74] Asim B. Khajwal and Arash Noshadravan. An uncertainty-aware framework for reliable disaster damage assessment via crowdsourcing. *International Journal of Disaster Risk Reduction*, 55:102110, March 2021. ISSN 22124209. doi: 10.1016/j.ijdrr.2021.102110. URL `https://linkinghub.elsevier.com/retrieve/pii/S2212420921000765`.

[75] Viet-An Nguyen, Peibei Shi, Jagdish Ramakrishnan, Narjes Torabi, Nimar S. Arora, Udi Weinsberg, and Michael Tingley. Crowdsourcing with Contextual Uncertainty. *Proceedings of the 28th CM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3645–3655, August 2022. doi: 10.1145/3534678.3539184. URL https://dl.acm.org/doi/10.1145/3534678.3539184. Conference Name: KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining ISBN: 9781450393850 Place: Washington DC USA Publisher: ACM.

[76] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets, December 2021. URL http://arxiv.org/abs/1803.09010. arXiv:1803.09010 [cs].

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This is a dataset submission. Our claims relate only to the content of our dataset, which is accurately described.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include a Limitations section (Sec. 6).

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is a dataset paper. It includes no theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our baselines are intended for reproduction. All code and data is public. All experiments are seeded (42) where not deliberately repeated for uncertainty marginalization.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All data and code is public. We include simple working code with documentation. We hope authors will build on our baselines.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Train and test splits are included in our HuggingFace datasets. The (many) baselines are described to an appropriate level of detail and full code is provided.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All downstream experiments are run with six seeds and our uncertainties are reported (separately, for clarity). Our upstream training (GZ Evo Core) is only run once and error bars are not reported; Core is intended as an intermediate step towards improving downstream performance, and is far more computationally expensive.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We report the compute requirements in the main text (80 A100-hours for our largest models).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/ thicsGuidelines`?

Answer: [Yes]

Justification: We have carefully reviewed the Guidelines and believe we fully conform.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: Our dataset introduces astronomical images of galaxies and therefore has no obvious direct societal impacts.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our baselines create models for interpreting images of galaxies and so have no obvious potential for misuse.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We carefully list all licenses in Sec. A.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We carefully list all licenses in Sec. A.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: via the publications in which the human labels were originally shared with atronomers

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not use study participants (for completeness, we note that we do hold an IRB approval for our crowdsourcing work).

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: LLMs are not part of this research.