A Licensing

A.1 Licensing for Existing Assets

Public astronomy data releases are intended for broad re-use by the research community. GZ2 uses Sloan Digital Sky Survey data released under a Creative Commons Attribution license[8] GZ Hubble and GZ CANDELS use images from the Hubble Space Telescope, which is operated by NASA. NASA images are not generally subject to copyright[9] GZ DESI (plus downstream datasets) uses Legacy Survey data published via NOIRLab. NOIRLab uses a Creative Commons Attribution 4.0 International License[10]. GZ UKIDSS uses images from the UKIDSS survey; they do not include a specific license, but make clear that the data is intended is for open use[11]. GZ Euclid (downstream) uses images from the *Euclid* space telescope, which is operated by ESA. ESA releases the images under a CC BY-NC 3.0 IGO license plus certain additional terms (e.g., no liability)[12].

A.2 Licensing for New Assets

Galaxy Zoo volunteers freely contribute their time to advance science. It would inappropriate to exploit their efforts for commercial gain. We are therefore releasing the dataset with a non-commercial license (Creative Commons Attribution Non Commercial Share Alike 4.0).

We share the volunteer labels in the hope of advancing open foundation models. We therefore explicitly require that researchers training models on this dataset make the source code for constructing and training such models public by publication, such that other researchers can build upon their work.

B Core Dataset Subsets

This appendix provides a narrative summary of each subset of our Core dataset. Each subset asks volunteers a series of questions and answers (a decision tree). Each tree is visualized at `data.galaxyzoo.org/gz_trees/gz_trees.html`. These questions and answers aim to identify the features of a galaxy. Galaxies often have many features; for example, many galaxies with spiral arms also have bars. Features are better understood as attributes (e.g. this animal has four legs) rather than classes (e.g. this animal is a cat). The images used and the questions asked vary between subsets. This is ideal for building models that generalize to new images and new questions. We summarize these differences below, and refer the reader to the original astronomical publications for full details.

Figure 5 shows the distribution of the number of volunteer annotators per galaxy, per campaign. Across all campaigns (our Core dataset) volunteers completed 28.7M decision trees and 104M individual questions.

B.1 Galaxy Zoo 2

Galaxy Zoo 2 [36] includes images from the Sloan Foundation Telescope in New Mexico [65]. This was the first Galaxy Zoo campaign to use a decision tree; we will describe the other trees with reference to this tree. In short, volunteers are asked if a galaxy is smooth (i.e. a 'blob'), featured (i.e. anything else of interest, typically a disk), or an artifact (i.e. an image problem). Volunteers selecting 'smooth' are then asked to describe the shape of the galaxy and of any central 'bulge' (i.e. a bright core). Volunteers selecting 'featured' are instead asked to describe those features. These include spiral arms (swirling streams of stars, like milk in coffee), bars (a straight line of stars through the center of the galaxy), and bulges (above).

B.2 Galaxy Zoo Hubble and Galaxy Zoo CANDELS

Galaxy Zoo Hubble [37] and Galaxy Zoo CANDELS [39] use images from the Hubble Space Telescope. Hubble images are sharper than ground-based telescopes (i.e., the images have a narrower

[8] https://www.sdss.org/collaboration/image-use-policy/

[9] https://www.nasa.gov/nasa-brand-center/images-and-media/

[10] https://noirlab.edu/public/copyright/

[11] http://www.ukidss.org/archive/archive.html

[12] https://www.cosmos.esa.int/web/esdc/terms-and-conditions

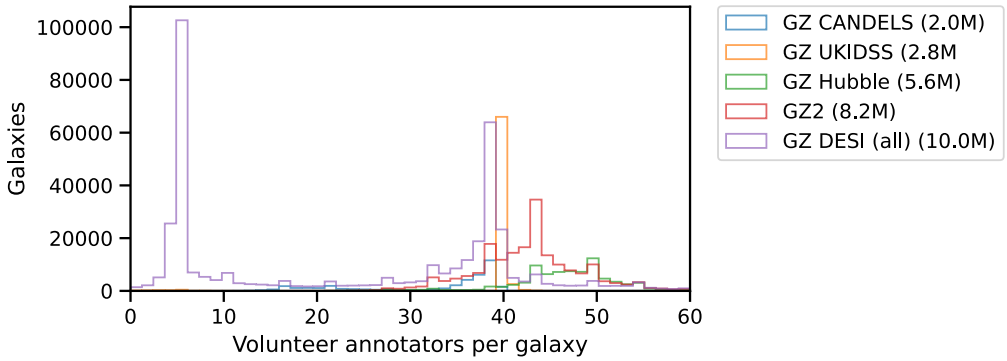

Figure 5: Distribution of the number of volunteer annotators per galaxy, per campaign. Most campaigns have approximately 40 annotators per galaxy. GZ DESI has a bimodal distribution with approximately half of galaxies receiving approximately 40 votes, and the remainder receiving approximately 5; see below. Legend shows the total number of completed decision tree annotations (i.e. the sum of the above) per campaign.

point-spread function than the effective ground-based point-spread function after accounting for atmospheric distortion). However, the galaxies in GZ Hubble and GZ CANDELS are more distant than in other campaigns, and so images in GZ Hubble and GZ CANDELS ultimately appear *less* sharp.

B.3 Galaxy Zoo DESI

Galaxy Zoo DESI [33] uses images primarily from the Blanco Telescope in Chile. GZ DESI is itself made up of four subcampaigns: GZD-1, GZD-2, GZD-5, and GZD-8. GZD-1 and GZD-2 used a question tree similar to GZ Hubble and GZ CANDELS, with only minor modifications to the number of possible answers to some questions (e.g. three vs. four possible bulge sizes). GZD-1 and GZD-2 ask identical questions on identically-processed images and so we jointly refer to them as GZD-1&2. GZD-5 added a new question asking if galaxies are merging (this may be an important mechanism for galaxies to grow). GZD-5 also used an active learning strategy where volunteers were asked to provide 40 annotations for galaxies selected by the acquisition function (BALD, [66]) and 5 otherwise [67]. GZD-8 asked the same questions as GZD-5 for images taken from MzLS and BASS, two telescopes in the Northern Hemisphere. We present the class balance between various DESI campaigns grouped by decision tree (GZD-1/2, GZD-5, and GZD-8) in Fig. 3.

B.4 Galaxy Zoo UKIDSS

GZ UKIDSS [38] includes images taken by the UK Infrared Telescope (UKIRT); the survey is described in [68]. Compared to images with optical wavelengths (e.g., those of GZ2 and GZ DESI), infrared images show more light from cooler sources – primarily older stars. GZ UKIDSS used the same questions as GZ2.

C Comparison between Volunteer and Professional Astronomers

Surveys suggest that Galaxy Zoo volunteers are strongly motivated by a willingness to contribute to original scientific research [69–71]. The aggregated responses from GZ volunteers have been repeatedly shown to agree well with those of professional astronomers and with automated measurements [36, 39, 72–74]. Figures 6 and 7 show the agreement between Galaxy Zoo 2 aggregated volunteer responses and two independent teams of professional astronomers [75, 76] for two questions answered by all three groups.

We can also compare the aggregated responses of volunteers in our GZ Rings downstream dataset with professional annotations from Buta et. al. [48]. 3735 of 3840 (97.6%) of galaxies annotated as ringed by Buta et. al. were identified as ringed by a majority of GZ volunteers.

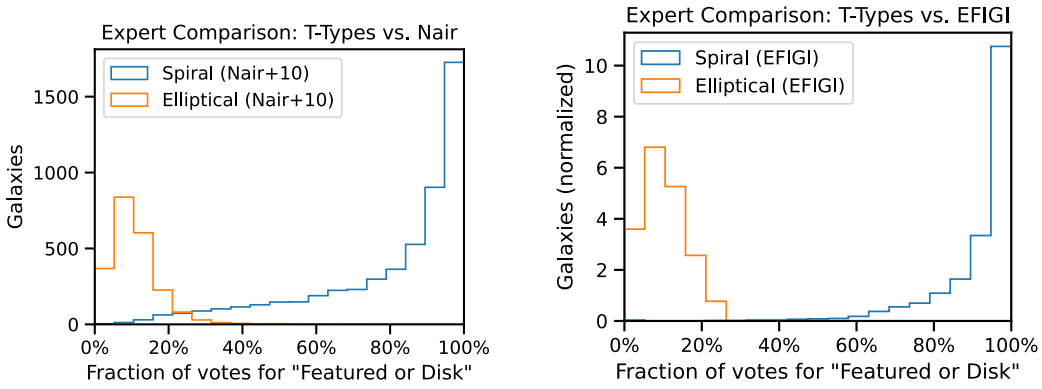

Figure 6: Agreement between GZ2 volunteers [36] and professional astronomers (left: Nair et. al. [75], right: Baillard et. al. [76]). Agreement shown as the fraction of volunteers answering 'Featured or Disk' to the question 'Is this galaxy smooth, or featured?' vs. the professional label.

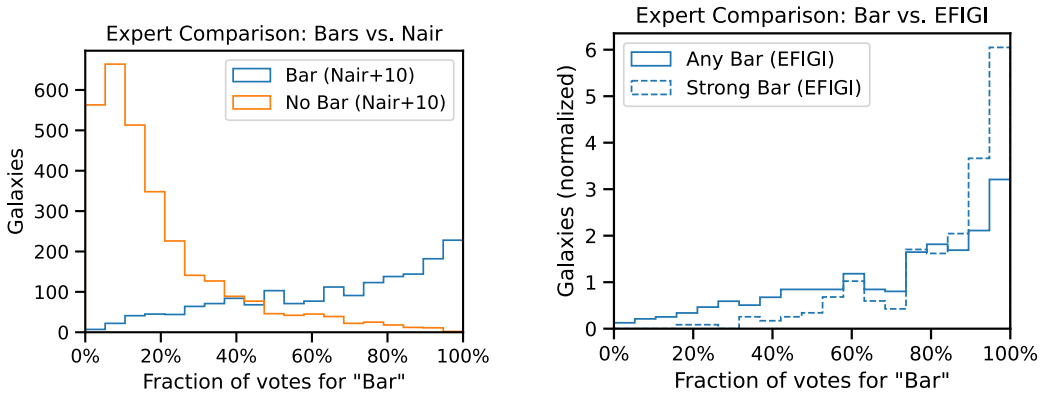

Figure 7: Agreement between GZ2 volunteers [36] and professional astronomers (left: Nair et. al. [75], right: Baillard et. al. [76]). Agreement shown as the fraction of volunteers answering 'Yes' to the question 'Is there a bar' vs. the professional label. For Baillard et. al., we group the professional annotations of ('Short', 'Long', 'Prominent') as 'Any' and we group ('Long' 'Prominent') as 'Strong'. Volunteer vote fraction is a good proxy for both. Volunteers give higher vote fractions for 'Strong' bars, as expected.

Finally, we note that it is not necessary for volunteers to agree with professional annotators to agree for our dataset to be useful for training; we only need volunteers to be self-consistent. We estimate the consistency in our vote fractions by taking advantage of a natural experiment. During GZ DESI, 690 galaxies were accidentally uploaded twice, two years apart. Figure 8 shows the change in vote fraction for the 'Featured' answer of the first decision tree question. 90% of vote fractions changed by less than 20%. We compare this with the change expected when simulating the volunteer responses as biased coin flips, using the observed vote fractions as the coin bias and tossing once per annotator. We find that the noise level in the aggregated responses is close to the irreducible aleatoric counting error.

D Datasheet

Each dataset is shared on the HuggingFace Hub with a concise practical summary. For completeness, we share a full datasheet below, following the 'Datasheets for Datasets' framework [77]. The datasheet answers show that Galaxy Zoo Evo is well-suited to open sharing for computer science research; it is an academic scientific dataset of galaxy images with no personal information, commercial motivations, terrestrial implications, etc.

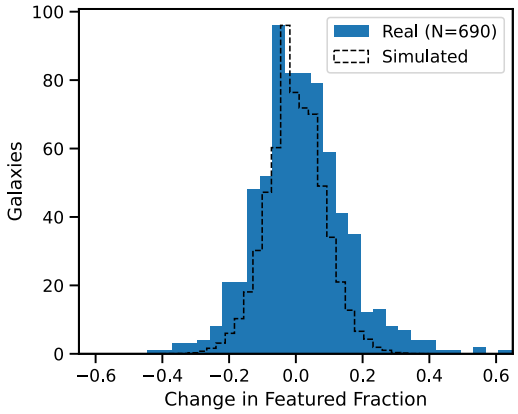

Figure 8: Change in 'Featured' vote fraction for 690 galaxies accidentally uploaded twice, two years apart. 90% of vote fractions changed by less than 20%. This change is close to the irreducible aleatoric counting error expected when modeling each response as a biased coin flip.

D.1 Motivation

1. *For what purpose was the dataset created?* The dataset was created to encourage development of new models for measuring the visual appearance of galaxies, as a benchmark for general-purpose foundation models, and as a real-world playground for open vision research problems e.g. efficient crowdsourcing, active learning, learning under label uncertainty, etc.
2. *Who created the dataset and on behalf of which entity*? The dataset was compiled by Dr Mike Walmsley (MW) on behalf of the Galaxy Zoo collaboration. The collection of the original telescope images and volunteer labels was organised by the researchers named on the papers for each subset. The volunteer labels were contributed by members of the public, to whom we are deeply grateful.
3. *Who funded the creation of the dataset?* The dataset was funded by MW's postdoctoral fellowship at the University of Toronto (see Acknowledgements). The underlying data collection was funded by numerous grants over 15 years, listed individually in the papers for each subset. These grants were primarily from governments, academic institutions, and individual philanthropy (e.g. the Sloan Foundation). The website platform used to collect the votes has additionally received funding from Google (see Acknowledgements).

D.2 Composition

1. *What do the instances that comprise the dataset represent?* The instances are pairs of (galaxy, volunteer labels).
2. *How many instances are there in total?* There are 990k instances in total in this initial release (823k in Core and 167k in Downstream). We intend Galaxy Zoo Evo as a living dataset with new galaxy images and labels being collected and added.
3. *Does the dataset contain all possible instances or is it a sample of instances from a larger set?* Each GZ Evo subset contains all instances of galaxies imaged by a specific telescope operational campaign (a.k.a. 'survey', in astronomical language) that meet campaign-specific criteria for relevance e.g. the galaxy must be bright and large enough to see detailed features, not suffer from obvious imaging artifacts, etc.
4. *What data does each instance consist of?* The images are processed to represent the raw light (flux) collected by the telescope in an RGB format suitable for human viewing. The exact details vary by subset and are carefully documented in each associated paper. In short, the images have light collected by each telescope filter assigned to a channel (i.e. they are in false color) and use a compressed dynamic range to ensure faint features remain visible.
5. *Is there a label or target associated with each instance?* Each label is a vector of counts of volunteer answers to a series of questions (the Galaxy Zoo decision tree).

6. *Is any information missing from individual instances?* A small (less than 1%) set of instances may suffer from image problems due to the challenging nature of taking telescope images of distant galaxies.
7. *Are relationships between individual instances made explicit?* The instances are related by position on the sky. This is shared as part of the dataset (right ascension and declination coordinates, analogous to latitude and longitude, measured in degrees).
8. *Are there recommended data splits?* Each subset has a canonical random train/test split, accessible via HuggingFace. The combined Galaxy Zoo Evo set has a train/test split consistent with the splits of each subset.
9. *Are there any errors, sources of noise, or redundancies in the dataset?* As with all human annotations, the volunteer labels are imperfect and hence we expect label noise. See Sec. 6.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources?* The dataset is self-contained.
11. *Does the dataset contain data that might be considered confidential?* No.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?* No. MW has found the galaxy images to be an effective antidote to anxiety.

D.3 Collection Process

1. *How was the data associated with each instance acquired?* The images were acquired by professional astronomers using international telescope facilities. The labels were acquired primarily by presenting the images to volunteers at galaxyzoo.org.
2. *How were these mechanisms or procedures validated?* Publications for the underlying data have passed peer review by professional astronomers. See each associated paper for full validation details.
3. *Who was involved in the data collection process and how were they compensated?* Data annotation was primarily performed by volunteers at galaxyzoo.org. Volunteers were not compensated; they freely contributed their time to help advance our scientific understanding of the universe, for which we are grateful. Approximately 334k unique logged-in volunteers participated.
4. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances?* The data was collected over a roughly 15-year period, and collection continues today. The timeframe associated with the instances is of the order 100 million to 10 billion years.
5. *Were any ethical review processes conducted (e.g., by an institutional review board)?* Our annotation approach was approved by an institutional review board at the University of Oxford, where the annotation platform was first developed.

D.4 Preprocessing/Cleaning/labelling

1. *Was any preprocessing/cleaning/labeling of the data done?* The labels were acquired by presenting the images to online volunteers at galaxyzoo.org.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?* The original telescope images remain publicly available, although access by non-specialists may be challenging in some instances. See the associated paper for each subset for telescope details.
3. *Is the software that was used to preprocess/clean/label the data available?* The annotation platform is zooniverse.org and the specific project is galaxyzoo.org.

D.5 Uses

1. *Has the dataset been used for any tasks already?* Galaxy Zoo labels have been used to support machine learning research since the 2014 Galaxy Challenge Kaggle competition, and remain the primary source of labelled data within this domain. An early internal version of the Galaxy Zoo Evo dataset has been used in a series of papers by the Galaxy Zoo

collaboration investigating self-supervised learning and neural scaling laws for building foundation models. See Sec. 2. Pre-release versions of this dataset were used for [6] and [78].

2. *Is there a repository that links to any or all papers or systems that use the dataset?* No, as the dataset is not yet published.
3. *What (other) tasks could the dataset be used for?* We suggest research opportunities throughout the main text.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?* The GZ DESI subset used an active learning approach to select which galaxies to label and so is not a random set of possible galaxies.
5. *Are there tasks for which the dataset should not be used?* We recommend against using the dataset for precise benchmarking (in the style of e.g. 'we achieve +0.1% on ImageNet') as the volunteer labels have limited precision. In general, we feel very small accuracy improvements are less consequential for astronomy than methods addressing broader problems e.g. multi-task learning, transfer learning, uncertainty, etc.

D.6 Distribution

Distribution (e.g. access, licensing, etc.) is addressed in the appendices above. We do not anticipate any regulatory restrictions.

D.7 Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?* The dataset will be hosted by HuggingFace and maintained by the Galaxy Zoo Collaboration.
2. *How can the owner/curator/manager of the dataset be contacted (e.g., email address)?* Please email Dr Mike Walmsley at (as of the time of writing) `m.walmsley@utoronto.ca`.
3. *Will the dataset be updated?* The dataset will be updated as new images and labels become available. We anticipate updates on a frequency between several months and several years. Updates will be communicated via HuggingFace README pages.
4. *Will older versions of the dataset continue to be supported/hosted/maintained?* Older versions will be maintained where practical via HuggingFace versioning (extending git LFS).
5. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?* Contributions from astronomers with annotated galaxy images would be deeply welcome and credited appropriately. Please reach out to MW (above).

E Reproducibility

All benchmark results reported in this work can be reproduced using the GitHub repository github.com/mwalmsley/gz-evo. We include:

1. A Dockerfile to ensure a consistent environment
2. Utility code to download the Galaxy Zoo Evo dataset
3. Scripts to run each baseline
4. Configuration files documenting the exact hyperparameters used

ConvNeXT-Nano can be reproduced on a consumer GPU. We recommend 2xA100-40GB GPUs or better for reproducing the larger models.