Draft version May 17, 2016
Preprint typeset using LaTeX style AASTeX6 v. 1.0

GALAXY ZOO: MORPHOLOGICAL CLASSIFICATIONS FOR 150,000 GALAXIES IN HST LEGACY IMAGING

Kyle W. Willett[1,2], Melanie A. Galloway[1], Karen L. Masters[3,4], B.D. Simmons[5,6,7], Chris J. Lintott[5], Steven P. Bamford[8], Edmond Cheung[9], Tom Melvin[3], Lucy F. Fortson[1], Melanie Beck[1], Claudia Scarlata[1], Roger L. Griffith[10,11], Kevin Schawinski[12], Edward M. Edmondson[3], Arfon M. Smith[5,13,14], Michael Parrish[13], Anna Han[15]

ABSTRACT

We present the data release paper for the Galaxy Zoo: Hubble (GZH) project.[a] This is the third phase in a large effort to measure reliable, detailed morphologies of galaxies by using crowdsourced visual classifications of color composite images. Images in GZH were selected from various *Hubble Space Telescope* Legacy programs using the Advanced Camera for Surveys (AEGIS, COSMOS, GEMS, GOODS-N, and GOODS-S), with filters that probe the rest-frame optical emission from galaxies out to $z \sim 1$. The galaxies selected for GZH classifications go down to magnitude limits of $m_{I814W} < 23.5$ and have a median redshift of $\langle z \rangle = 0.9 \pm 0.6$, with a tail extending out to $z \simeq 4$. The GZH morphological data include measurements of both bulge- and disk-dominated galaxies, details on spiral disk structure that relate to the Hubble type, bar identification, and numerous measuremenst of clump identification and geometry. This paper also describes a new method for calibrating morphological bias by using artificially-redshifted galaxy images as a baseline. The GZH catalog contains both raw and calibrated morphological vote fractions for 150,771 galaxies, providing the largest data set to date suitable for large-scale studies of galaxy evolution out to $z \sim 1$.

Keywords: galaxies:structure — galaxies:high redshift — galaxies:evolution — methods:data analysis — catalogs

1. INTRODUCTION

The morphology of galaxies encodes information on the orbital parameters and assembly history of their contents, including gas, dust, stars, and the central black hole. The morphology is also closely related to the local environment of a galaxy, as interactions such as tides, shocks in cluster environments and direct mergers can all influence a galaxy's internal structure. For $M^\star$ galaxies in the local Universe, this interplay between the physical development of a galaxy and its external appearance typically manifests at the most basic level as the difference between bulge-dominated systems with no/little spiral structure (early-types) and disk-dominated, rotationally-supported galaxies (late-types) frequently exhibiting spiral arms. This dichotomy has been used to explore much of the astrophysics governing galaxy formation and evolution, and has been shown to be closely linked with other galactic properties such as stellar mass, halo mass, bolometric luminosity, black hole activity, effective radius, and the relative ages of the stellar populations.

The advent of larger telescopes in an increasing range of observing wavelengths has revealed that the distribution and properties of galaxy morphology have strongly evolved over the lifetime of the Universe. At redshift $z \simeq 1$ (roughly 6 Gyr after the Big Bang), many galaxies are still in the process of assembling the baryonic and stellar mass required to reproduce the mature, coherent structures seen at the present day. This growth occurs via a variety of pathways, including accretion from large-scale galactic filaments onto halos via streaming, mergers of individual halos, conversion of gas into stars via gravitational collapse, etc. The accrual of baryons can be slowed or even reversed via feedback from stellar winds, supernovae, and active black holes. Each of these processes affects galaxy morphology in different ways, and so an accurate measurement of morphological demographics as a function of redshift provides an extremely powerful observational constraint on the physics involved (for recent reviews see Buta 2013; Conselice 2014).

Theoretical predictions for the morphology of galaxies as a function of redshift are primarily computed within the ΛCDM cosmological framework. Full treatments model gravitational interactions between baryons and dark matter, hydrodynamics of the gas, and baryonic physics related to star formation and evolution. The most advanced simulations now span volumes up to

[a] This publication has been made possible by the participation of more than 200,000 volunteers in the Galaxy Zoo project. Their contributions are individually acknowledged at http://authors.galaxyzoo.org/authors.html.

~ 100 Mpc^3 while simultaneously resolving the smaller (< 1 kpc) scales necessary to reproduce the influence of baryonic physics (Vogelsberger et al. 2014; Schaye et al. 2015). Such simulations predict clustering of galaxies on large scales in a hierarchical assembly model (Silk & Mamon 2012). The structure of individual galaxies is affected by their merger history (Toomre & Toomre 1972; Steinmetz & Navarro 2002; Hopkins et al. 2010), local environment (such as the morphology-density relation; Dressler 1980), initial dark halo mass, secular evolution rate, and many other factors. Morphologies of individual simulated high-mass galaxies at $z \sim 2-3$ commonly show kpc-scale "clumpy" structures, with few galaxies that are either smooth or well-ordered spirals; asymmetric profiles with strong density contrasts dominate simulated populations in the early Universe until at least $z \sim 1$ (Bell et al. 2012; Genel et al. 2014).

Observational studies of galaxies at high-redshift also display a wide range of morphological types, many of which are rare or absent at $z \sim 0$. Although spheroids and disks are present, they are typically much more compact than those in the local Universe. There is also a significant population of more irregular, massive galaxies, including mergers, tadpoles, chains, double-clumps, and clump-clusters (Elmegreen et al. 2005, 2007; Cameron et al. 2011; Förster Schreiber et al. 2011; Kartaltepe et al. 2015). In contrast, while grand-design spirals have been observed as far back as $z = 2.18$ (Law et al. 2012a), their spatial density suggests that they are exceedingly rare, with a very low fraction of undisturbed disks (Mortlock et al. 2013). Current observational data thus strongly suggests that the classical Hubble sequence/tuning fork (Hubble 1936) is not a suitable framework for characterizing high-redshift morphology.

Space-based observatories, particularly the *Hubble Space Telescope* (*HST*), have been responsible for the bulk of imaging studies of high-redshift galaxies. Observations of fixed fields with very deep imaging (eg, Williams et al. 1996; Giavalisco et al. 2004; Beckwith et al. 2006; Davis et al. 2007; Scoville et al. 2007; Grogin et al. 2011) give the photometric sensitivity necessary to detect $L^\star$ galaxies at $z > 1$, while also providing the angular resolution to distinguish internal structure and characterize the morphology. While these measurements are helped by the fact that the angular diameter distance is relatively flat beyond $z > 1$ in a flat ΛCDM cosmology, the relevant angular scales are only of the order $\sim 5-10$ kpc/$''$ (Wright 2006). *HST* can thus resolve much of the structure for a Milky Way-sized galaxy (at least for distinguishing a disk from a bulge), but will be limited for more compact structures. Since the size of galaxies evolves as roughly $r \propto (1+z)^{-1}$ (Mao et al. 1998; Law et al. 2012b), the compact sizes of high-redshift galaxies make detailed morphologies a challenge even for *HST* (Chevance et al. 2012). The public availability of more than 10^5 galaxies in archival imaging across various studies gives a data sample with the potential for high statistical significance.

One of the major difficulties in studying the morphologies of galaxies lies in the techniques used for measurement. Visual classification by experts has been used for many decades (eg, Hubble 1926; de Vaucouleurs 1959; Sandage 1961; van den Bergh 1976; Nair & Abraham 2010; Baillard et al. 2011; Kartaltepe et al. 2015). These methods have the advantage of using the significant processing power of the human brain to identify patterns, but suffer from issues such as lack of scaling to large surveys and potential issues with replicability and calibration. Automated measurements, both parametric (Peng et al. 2002; Simard et al. 2011; Lackner & Gunn 2012) and non-parametric (Abraham et al. 2003; Conselice 2003; Lotz et al. 2004; Scarlata et al. 2007; Bamford et al. 2008; Freeman et al. 2013), scale well to very large sample sizes, but do not always fully capture the relevant features, especially for asymmetric galaxies that become increasingly common at high redshifts. The Galaxy Zoo project (Lintott et al. 2008) utilizes crowdsourced visual classifications to measure galaxies in color-composite images. The efforts of more than 2×10^5 classifiers allow for multiple independent classifications of each image which are combined and calibrated to give a distribution of vote fractions proportional to the probability of a feature being visible. While crowdsourced data require extensive calibration (Bamford et al. 2009; Willett et al. 2013), they have a proven reliability and have been used for a wide variety of scientific studies (eg, Land et al. 2008; Bamford et al. 2009; Darg et al. 2010; Masters et al. 2011; Skibba et al. 2012; Simmons et al. 2013; Schawinski et al. 2014; Willett et al. 2015).

This paper presents the classifications collected from the Galaxy Zoo Hubble (GZH) project.[1] GZH was the third phase of Galaxy Zoo, following its initial results classifying $\sim 900,000$ SDSS images into primarily early/late types (Lintott et al. 2011) and Galaxy Zoo 2, which covered $\sim 250,000$ SDSS images using a more detailed classification scheme that included bars, spiral arms, and galactic bulges (Willett et al. 2013). GZH used a similarly detailed classification scheme, but focused for the first time on images of high-redshift galaxies taken with *HST*. The Galaxy Zoo: CANDELS project has also classified morphologies of galaxies at high redshift using ACS and WFC3 imaging (Simmons et al., submitted).

The sample selection and creation of the images used

[1] http://zoo3.galaxyzoo.org/

for GZH is described in Section 2. Section 3 describes the GZH interface and the collection of classifications. Section 4 outlines the process used to calibrate and correct the crowdsourced vote fractions for redshift-dependent bias. Section 5 gives the main catalog of results, with several examples of how the data may be queried in Section 6. Section 7 gives a short overview of the observed morphological demographics and compares them to several other catalogs, with a summary in Section 8.

This paper assumes the WMAP9 cosmological parameters of $(\Omega_m, \Omega_\Lambda, h) = (0.282, 0.718, 0.697)$ (Hinshaw et al. 2013).

2. SAMPLE AND DATA

2.1. *Hubble Legacy Surveys*

The GZH project contains images drawn from a number of different dedicated surveys and sample selection criteria. The majority of the data (as implied by the project name) were taken directly from *HST* Legacy Surveys, all of which primarily used imaging from the Advanced Camera for Surveys (ACS). Results from the individual surveys have been combined into a single photometric and morphological database, the Advanced Camera for Surveys General Catalog (ACS-GC; Griffith et al. 2012). A summary of the key parameters is given in Table 1.

The properties of the individual surveys are as follows:

- The All-Wavelength Extended Groth Strip International Survey (AEGIS; Davis et al. 2007) covers a strip centered at $\alpha = 14^{\rm h}17^{\rm m}, \delta = +52°30'$. This area of the sky was selected for a deep survey due to a combination of low extinction and low Galactic/zodiacal emission. The ACS images covered 63 separate tiles over a total area of ~ 710 arcmin^2. The two ACS bands for AEGIS had exposure times of 2300 seconds in F606W (V_{606W}) and 2100 seconds in F814W (I_{814W}). The final mosaic images were dithered to a resolution of 0.03 ″/pixel. For extended objects, the limiting magnitude of sources was 26.23 (AB) in V_{606W} and 25.61 (AB) in I_{814W}.

- The Great Observatories Origins Deep Survey (GOODS; Giavalisco et al. 2004) covered two separate fields in the northern and southern hemispheres: the Hubble Deep Field-North ($\alpha = 12^{\rm h}36^{\rm m}, \delta = +62°14'$) and the Chandra Deep Field-South ($\alpha = 03^{\rm h}32^{\rm m}, \delta = -27°48'$). The *HST* ACS imaging data from the two fields are referred to as GOODS-N and GOODS-S, respectively. ACS imaging in GOODS fields used 4 filters – F435W (B_{435W}), V_{606W}, F775W (i_{775W}), and F850LP (z_{850LP}). The mean exposure times for each epoch varied by band, from 1050–2100 seconds. The B_{435W} images were completed in a single epoch at the beginning of the survey, but the V_{606W}, i_{775W}, and z_{850LP} images were taken in five separate epochs separated by 40–50 days each. Images were dithered to a pixel scale of 0.03 ″/pixel and covered a total area of ~ 320 arcmin^2 (160 arcmin^2 per north/south field). The 5σ limiting magnitude for extended sources was 25.7 for V_{606W} and 25.0 for i_{775W}.

- The Cosmic Evolution Survey (COSMOS; Scoville et al. 2007) covered an area of ~ 1.8 deg^2 centered at $\alpha = 10^{\rm h}00^{\rm m}, \delta = +02°12'$. Its location near the celestial equator was designed to enable coverage by ground-based telescopes in both the Northern and Southern Hemispheres, in addition to space-based observatories. The ACS data for COSMOS consisted of 1 orbit per pointing with an exposure time of 2028 seconds in I_{814W}; 590 total pointings were used to cover the entire field. The image resolution was dithered to 0.05 ″/pixel. The 50% completeness magnitude for a galaxy with a half-light radius of 0.50″ in I_{814W} was 24.7 mag.

- The Galaxy Evolution from Morphologies and SEDs (GEMS; Rix et al. 2004; Caldwell et al. 2008) survey was also centered on the Chandra Deep Field-South. The GEMS data covered ~ 800 arcmin^2, completely surrounding the area covered by GOODS-S. Images from ACS in GEMS had 1 orbit per pointing for a total of 63 pointings. The exposure times were 2160 and 2286 seconds in V_{606W} and z_{850LP}, respectively. The image resolution had a pixel scale of 0.03 ″/pixel. The 5σ limiting magnitude for source detection was 25.7 AB in V_{606W} and 24.2 AB in z_{850LP}.

2.2. *Galaxy selection*

In the ACS-GC (Griffith et al. 2012), individual galaxies were identified using a combination of SExtractor (Bertin & Arnouts 1996) and the galaxy-profile fitting framework Galapagos (Barden et al. 2012). GZH included all galaxies with $m < 23.5$, where m is in the I_{814W}, z_{850LP}, or i_{775W} for the AEGIS + COSMOS, GEMS + GOODS-S, and GOODS-N surveys, respectively. This yielded a total of 113,166 images (Table 1).

Single-epoch images from SDSS Stripe 82 were selected using the same criteria from Willett et al. (2013), which required limits of `petroR90_r`$> 3''$ (where `petroR90_r` is the radius containing 90% of the r' Petrosian flux) and a magnitude brighter than $m_r < 17.77$. The images used were three-color composites using the

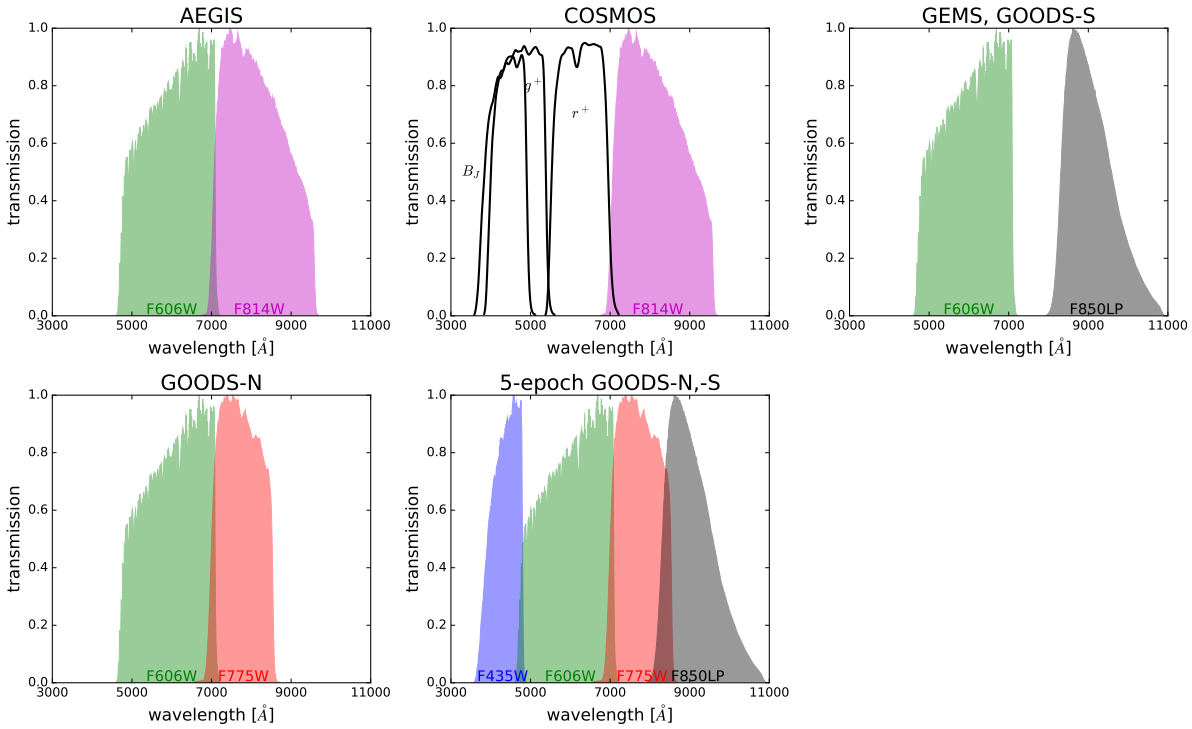


Figure 1. Transmission curves of the filters used by *HST* Advanced Camera for Surveys (ACS) in wide-field channel mode for the various surveys in GZH. The unfilled black curves show the filters for the Suprime Camera on *Subaru*, which was used to create color gradients in the GZH COSMOS images.

Table 1. Summary of GZH imaging

Survey	Total $t_{\rm exp}$ [sec]	Filters	Resolution [″/pix]	Area [arcmin^2]	$N_{\rm galaxies}$
AEGIS	2100−2300	V_{606W}, I_{814W}	0.03	710	8157
COSMOS	2028	I_{814W}	0.05	6480	88530
GEMS	2160−2286	V_{606W}, z_{850LP}	0.03	800	9143
GOODS	–	–	–	–	–
GOODS-N 2 epoch	2100−4200	V_{606W}, i_{775W}	0.03	320	2551
GOODS-S 2 epoch	2100−4200	V_{606W}, z_{850LP}	0.03	320	3593
GOODS-N 5 epoch	5100−10500	B_{435W}, V_{606W}, i_{775W}, z_{850LP}	0.03	″	6015
GOODS-S 5 epoch	5100−10500	B_{435W}, V_{606W}, i_{775W}, z_{850LP}	0.03	″	5142
total	–	–	–	8630	123131

g', r', and i' filters (Nieto-Santisteban et al. 2004). 21,522 galaxies in SDSS met these criteria. Co-added images from Stripe 82 were selected from the union of galaxies with co-added magnitudes brighter than 17.77 mag, and the galaxies detected in the single-depth images and matched to a co-add source. This resulted in a total set of 30,339 images. Of the images in the co-added sample, 5144 (17 percent) were dimmer than the initial magnitude cut of 17.77.

2.3. *Image creation*

The images used for classification in GZH were color-composite JPEGs made from multi-band data. These were created followed the method of Lupton et al. (2004), which preserves colour information irrespective of intensity. An asinh intensity mapping was applied to enhance the appearance of faint features while avoiding saturating galaxy centers. The relative scalings of the filter bands were chosen while trying to reproduce the color appearance of the SDSS images in previous iterations of Galaxy Zoo.

Many of the Legacy surveys described in Section 2.1 provided images in only two filters. For these, the

shorter-wavelength band was mapped to the blue channel, the longer-wavelength band to the red channel, and the green channel created by taking the arithmetic mean of the two. The bands used in each of the surveys are listed in Table 1. Although four bands were available for the GOODS survey, only two bands were used to create the original 2-epoch images, for consistency with AEGIS and GEMS. The 2-epoch GOODS-N and GOODS-S images were created using different filters — this was a deliberate choice made so that the GEMS images could be directly compared with the overlapping coverage of GOODS-S (Figure 1).

Only 2-epoch GOODS images were included at the launch of GZH. Deeper, 5-epoch GOODS images were added into GZH later (in March 2015). These images made use of the full four-band data by using the arithmetic mean of B_{435W} and V_{606W} in the blue channel, I_{814W} in the green channel, and z_{850LP} in the red channel.

The COSMOS survey provided only I_{814W} HST imaging at the time of the GZH launch. These galaxies used "pseudocolor" images created by using the ACS I_{814W} data as an illumination map and ground-based imaging from the *Subaru* telescope in B_J, r^+, and i^+ filters to provide color information (see Griffith et al. 2012 for further details). This resulted in images with the angular resolution of *HST* ($\sim$ 0.05 $''$/pixel) for the overall intensity, but color gradients at ground-based resolution, with seeing between 0.95$''$ and 1.05$''$ (Taniguchi et al. 2007).

Stripe 82 single-epoch images were taken directly from the DR7 SDSS Skyserver, which combined g', r', and i' exposures into the RGB channels. The co-added Stripe 82 images were assembled from runs 106 and 206 in DR7 and processed into color composites in a corresponding manner.

In some cases, we found that attempting to emphasize faint features in the images resulted in the sky noise taking the appearance of brightly colored speckles. This impaired the aesthetics of the images and might have been a distraction to visual classification. To counteract this, a soft-edged object mask was applied to the color images and a desaturation operation performed. This masked procedure was effective in preserving the color balance for galaxies, retaining the visibility of faint features, but reduced the color contrast in the sky noise and greatly improved the appearance of the images. This solution was applied to the coadded Stripe 82, COSMOS and 5-epoch GOODS images.

2.4. *Images with simulated nuclear point sources*

GZH also included a set of images designed to measure the effect of active galactic nuclei (AGN) on morphological classifications. Since galactic nuclei can have bright, unresolved optical emission, AGN have the potential to mimic or distort the identification of a bulge component. The presence of an AGN was simulated by modeling the point spread function (PSF) of the telescope and then inserting a bright source near the center of a real galaxy. For each image, the simulated AGN was assigned one of three colors – either blue, red, or flat (white) as seen in the color images – and a range of brightnesses such that $L_{\rm ratio} \equiv L_{\rm galaxy}/L_{\rm AGN}$ is in $(0.2, 1.0, 2.0, 5.0, 10.0, 50.0)$. Combining these parameters generated 15 images with different simulated AGN for each host.

Two sets of simulated AGN were generated in GZH. The first set (version 1) was assembled from 95 galaxies from GOODS-S imaging and empirical PSFs made by combining stars in the GOODS fields using the PSF creation tools in `daophot`. The second set (version 2) was assembled from 96 galaxies in GOODS-S; this version used simulated PSFs from `TinyTim` (Krist 1993), drizzled using the same procedures as those used in reduction of the GOODS-S images (Koekemoer et al. 2002, 2003; Giavalisco et al. 2004). The use of these 2 versions facilitates comparisons between these different PSF creation methods, which are widely used in AGN host galaxy morphology studies (e.g., Sánchez et al. 2004; Simmons & Urry 2008; Pierce et al. 2010; Simmons et al. 2011). Each PSF creation method has advantages and disadvantages: the empirical PSFs better represent the nuances of the PSF in the specific data being used and look more realistic at lower luminosities, but the extended features of the noiseless `TinyTim` PSFs are visually more realistic at higher luminosities.

Images with simulated AGN were classified in the main interface in an identical manner and evenly distributed with unaltered images of the galaxies. Classifiers were not explicitly told that the images had been altered during classification, as the goal was to measure the effect on normal classifications in as unbiased a manner as possible. Following classification, a classifier could view a page with additional details about each galaxy they had classified; where applicable these pages contained further information regarding image modifications.

2.5. *Galaxy metadata*

Photometric data for the bulk of the GZH sample were largely drawn from the tables in Griffith et al. (2012). This included photometric parameters such as the fluxes, magnitudes, radii, ellipticities, position angles, and positions drawn from both SExtractor and GALFIT. GALFIT also provided the parametric Sérsic index and effective half-light radius for the best-fit model. All parameters were measured in both bands of the ACS imaging, with the exception of the single-band COSMOS images.

Redshifts for the GZH catalog were compiled from a variety of sources. For each galaxy, the primary redshift used for debiasing calculations (Section 4) is in the `Z_BEST` column of Table 5. The redshift type (spectroscopic: `SPEC_Z`, photometric: `PHOTO_Z`, or grism: `GRISM_Z`) is listed in the column `Z_BEST_TYPE`, and the source catalog of the redshift is included as `Z_BEST_SOURCE`.

For galaxies which have published redshifts from multiple sources, the following algorithm was used to select the `Z_BEST` quantity. The first priority is spectroscopic redshifts; these were taken from the ACS-GC (Griffith et al. 2012), 3DHST (Momcheva et al. 2015), and MUSYC (Cardamone et al. 2010) catalogs. A high-quality spectroscopic redshift in the ACS-GC is the primary option; if none is available, then `Z_BEST` uses the spectroscopic redshifts in 3DHST, and then MUSYC. For galaxies with multiple spectroscopic redshifts, more than 98% are consistent ($\Delta z < 0.001$), and so the order of selection made no practical difference. Galaxies with inconsistent spectroscopic redshifts between any pair of catalogs are marked with a flag in Table 5. If no spectroscopic redshifts were available, the 1-σ errors of the photometric (ACS-GC, 3DHST, MUSYC, UltraVISTA; Ilbert et al. 2013) and UltraVISTA grism data were used to select the redshift with the smallest error.

Photometric and spectroscopic data for the SDSS Stripe 82 galaxies were taken from the CasJobs DR7 tables. This included *ugriz* Petrosian magnitudes and fluxes, as well as the relative de Vaucouleurs and exponential fits from the model magnitudes. All redshifts used for SDSS galaxies were spectroscopic. 82.6% of galaxies in the single-depth images and 65.1% of galaxies in the co-add images had a measured DR7 spectroscopic redshift.

3. GZH INTERFACE AND CLASSIFICATIONS

Below we describe the classification structure of GZH, including the software interface and the hierarchical structure of a classification. Section 3.2 describes the process of combining individual classifications into vote fractions for each galaxy.

3.1. *Interface and decision tree*

Classifications for GZH were made using a web-based interface (Figure 2), similar in design to Galaxy Zoo and Galaxy Zoo 2. The front-end runs on a Ruby on Rails framework with classifications stored in a MySQL backend. Classifiers were shown a randomly-selected color composite image from the GZH sample; the default showed the image with a black sky background, although they had the option to invert the color palette if desired. The questions and responses for morphology appeared on the right side of the image as a panel, including both text and icons. There was no tutorial required for participation, although classifiers could access an extensive "Help" section containing example images and descriptive text for all the morphological labels.

The procedure for classifying an image in GZH followed a hierarchical decision tree (Figure 3). Every classification began with the step of identifying whether the object at the center of the image was a smooth galaxy, a galaxy with a disk or other features, or a star/artifact. Subsequent questions in the tree depended on the previous answer(s) given by the classifier; the decision tree was designed so that every question relevant to the morphology in the process of being identified was answered. Questions that were not answered were implicitly assumed to be absent in the image — for example, if the classifier identified a galaxy as being smooth, they were not asked to count the number of spiral arms. For every task, the classifier chose a single answer before continuing to the next question; they also had the option to restart any classification in-progress.

The GZH decision tree had four broad sets of morphologies. The first option designated stars or image artifacts (the result of either bad data or incorrect classification by the ACS pipeline as a galaxy); in this case, the classification process ended and no further questions were asked. The second set were smooth galaxies, intended to select ellipticals/early-types; classifiers also measured the relative axial ratio (roundness) for these galaxies. The third set was for disk/late-type galaxies, which measured the features necessary to place a galaxy on the standard Hubble tuning fork (bars, spiral arm, strength of the central bulge). The final set, which was new in this phase of Galaxy Zoo and designed for high-redshift targets, identified objects dominated by clumpy morphologies. Further annotations for clumpy galaxies included assessing the number, arrangement, relative brightness, and location of the clumps within the galaxy. Finally, every classification of a galaxy gave the option of identifying "odd" features within the image; these labels were for relatively rare ($\lesssim 1\%$) phenomena, including dust lanes, gravitational lenses, and mergers.

The number of independent classifications per subject collected by GZH was on average higher than GZ1 or GZ2, due to both the increased complexity of the decision tree and the relative difficulty of classifying images of small and distant galaxies. Images from the main AEGIS, GEMS, and GOODS data sets had a median of 122 independent classifications per image. The remainder of the images either had a later activation date (COSMOS, simulated AGN) or a lower retirement limit (the low-redshift SDSS Stripe 82 galaxies). Images from these samples had a median of 46–48 classifications per image (Figure 4).

The GZH project was launched on 23 Apr 2010 with

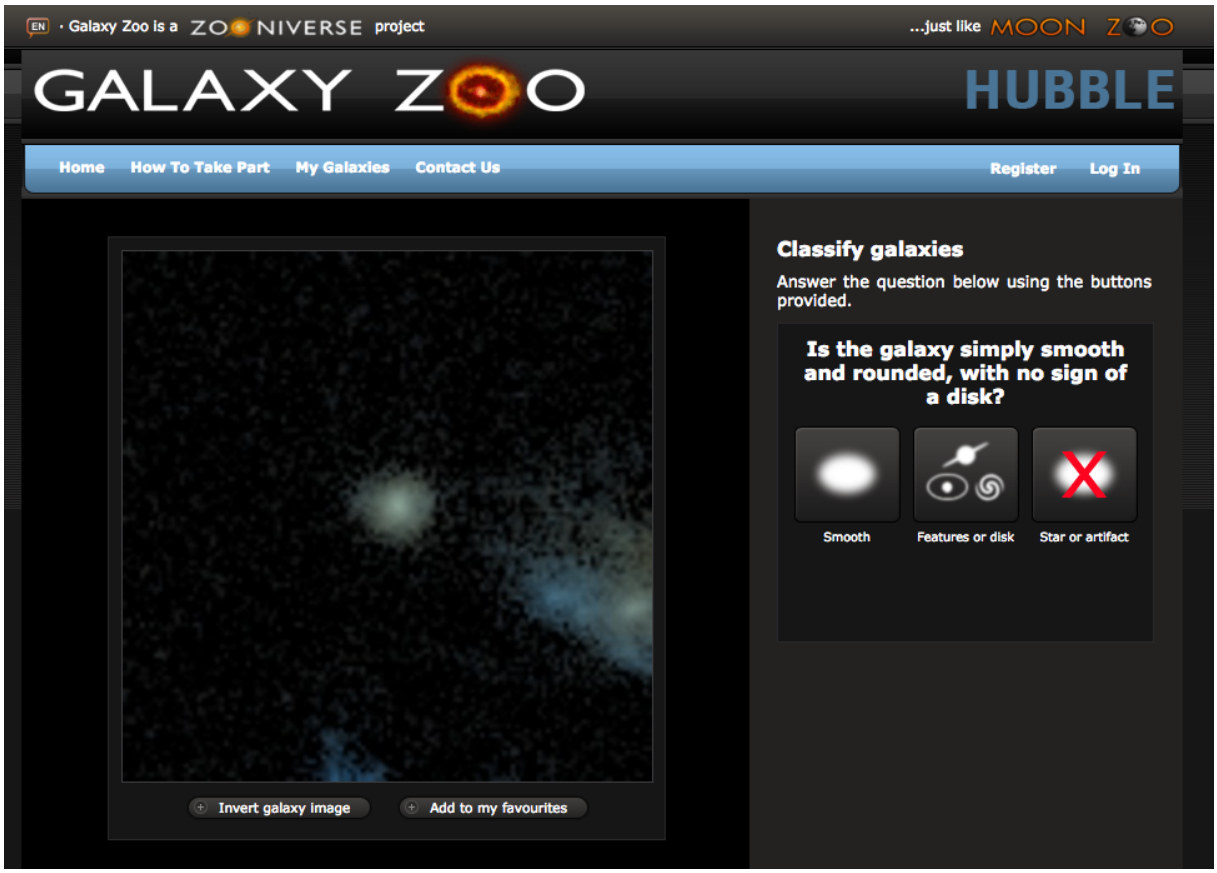


Figure 2. Screenshot of the GZH interface (http:/zoo3.galaxyzoo.org) at the beginning of a classification, with the classifier ready to select an answer for the first question in the decision tree.

the inclusion of the AEGIS, GEMS, GOODS 2-epoch, and SDSS Stripe 82 images. Images from COSMOS and the simulated AGN were activated in Dec 2010, as well as a small sample of images from AEGIS, GEMS, and GOODS that were previously excluded from the original sample due to cuts on blended and/or saturated objects and subsequently confirmed as classifiable galaxies. The main GZH site collected data until its replacement, the fourth phase of Galaxy Zoo[2] (including data from both the *HST* CANDELS survey and SDSS DR8) began on 10 Sep 2012. Classifications for the GOODS 5-epoch images were separately obtained from Mar-Jun 2015 on the fourth version of the Galaxy Zoo site. The GZH project had a total of 10,349,357 classifications from 93,898 registered participants.

3.2. *Classifier weighting*

As a first step to producing consensus classifications of each galaxy, the votes of individual classifiers in GZH were combined to make a vote fraction for each response ($f_{response}$) to a question in the classification tree. Votes were subsequently weighted and re-combined in an iterative method similar that in previous versions of Galaxy Zoo (Land et al. 2008; Willett et al. 2013), using a method chosen to be as egalitarian as possible while also identifying and downweighting classifiers who frequently disagreed with others. The weighting factor (w) was 1 for the top 95% of classifiers, as ranked by consistency. For the bottom 5% of classifiers, w was designed to drop smoothly and was effectively zero for the bottom 1%. Since this only affected a tiny percentage of the classifiers (and an even smaller percentage of the classifications) the overall effect on the GZH dataset was minimal. The method was effective, however, at filtering out contributions from random or deliberately malicious classifiers.

Classifications were aggregated and weighted only if the classifier was logged into GZH under their username (which was encouraged, but not required for participation). Classifications by participants who were not logged in were marked as "Anonymous" and assumed to have $w = 1$.

4. CORRECTING FOR REDSHIFT-DEPENDENT CLASSIFICATION BIAS

The previous versions of Galaxy Zoo morphology classifications (Lintott et al. 2008; Willett et al. 2013) were

[2] http://zoo4.galaxyzoo.org

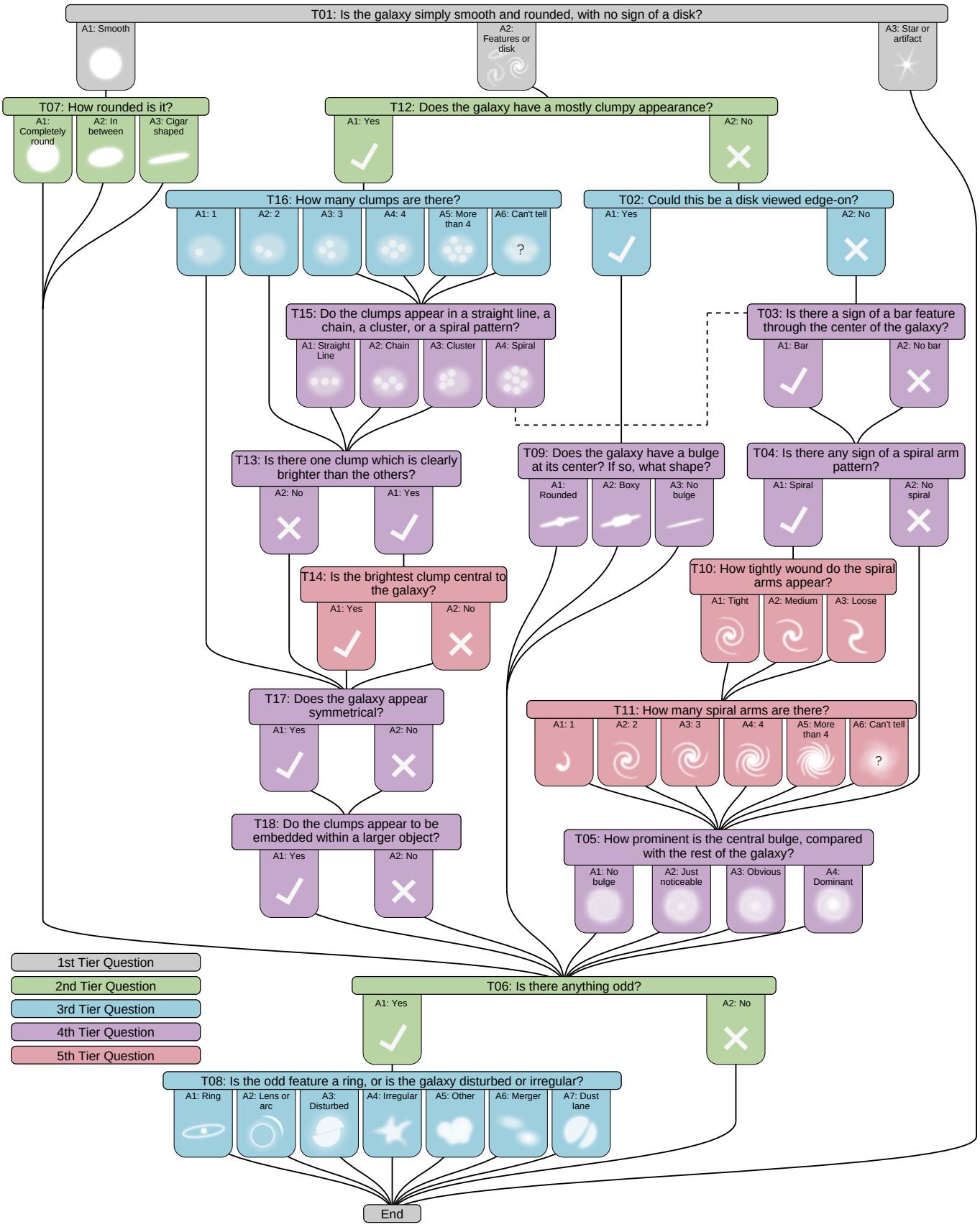


Figure 3. Flowchart of the questions presented to GZH classifiers, labeled with the corresponding Task numbers. Tasks in the decision tree are color-coded by tier level: Gray-colored Tasks are 1st-tier questions which are asked in every classification. Tasks colored green, blue, purple, and pink (respectively) are one, two, three, or four steps below branching points in the decision tree.

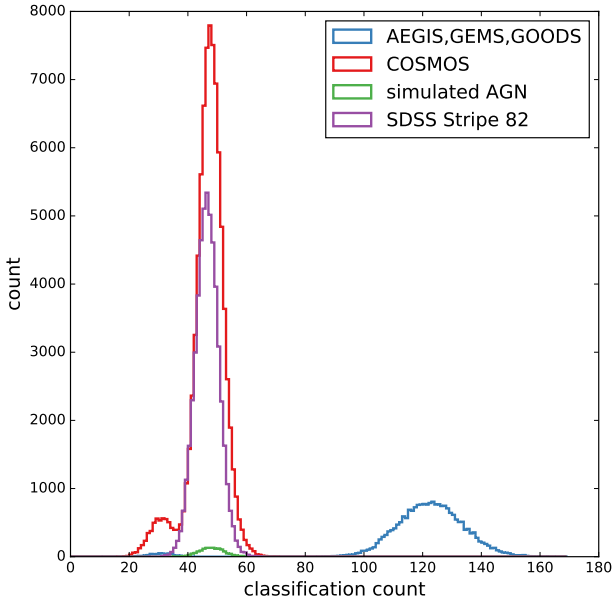


Figure 4. Distribution of the total number of classifications per image for GZH, split by survey.

based on observations of galaxies in the Sloan Digital Sky Survey (SDSS) which have a median redshift of $z < 0.2$. In these cases, it was assumed that there was no cosmological evolution of the morphologies of galaxies and therefore any observed changes in the distribution of galaxies with different consensus morphologies was due to the effects of redshift on the image quality (*i.e.*, the reduction in physical resolution, surface brightness dimming, etc). For both previous releases of GZ morphologies, a correction for redshift-dependent bias was applied based on matching the classification fractions at the highest redshifts with those at the lowest redshift. Bamford et al. (2009) and Willett et al. (2013) provide complete descriptions of the process for GZ1 and GZ2, respectively.

In the GZH samples, the redshift range is large enough that cosmological evolution of the types and morphologies of galaxies is expected for galaxies in the sample. As a result, the previous methods of correcting for redshift dependent bias do not work. In addition, the effects of band shifting will change the images even more across these redshift ranges.

In order to test and correct for the effects of redshift, GZH includes a set of calibration images. These images consist of the same galaxy as it would appear over a variety of redshifts. The input images are from the SDSS (York et al. 2000; Strauss et al. 2002) and are processed using the FERENGI code (Barden et al. 2008) to match the observational properties of the *HST* surveys out to $z = 1$. These images were classified in the GZH interface using the same classification scheme as the original *HST* images.

Table 2. Summary of FERENGI artificial redshifting

$z_{\rm sim}$	$N_{z\rm bins}$	$N_{\rm evolution}$	$e_{\rm max}$	$N_{\rm galaxies}$	$N_{\rm images}$
0.3	8	7	−3.0	72	4032
0.5	6	4	−1.5	72	1728
0.8	3	3	−1.0	72	648
1.0	1	3	−1.0	72	216

4.1. *Generating images of artificially-redshifted galaxies*

The GZH classifications include 288 unique galaxies originally generated from SDSS imaging and processed with the FERENGI code. The selection spanned a variety of galaxy morphologies (as selected by GZ2 classifications) and r'-band surface brightnesses, and also spanned the redshift range of SDSS targets (in $N_z = 4$ bins) in order to be optimized for different target minimum redshifts in *HST* imaging.

The surface brightness selection ($N_\mu = 3$) for the FERENGI galaxies was (1) low: $\mu > 21.5$ mag arcsec^{-2}; (2) mid: $20.5 < \mu < 21.5$ mag arcsec^{-2}; and (3) high: $\mu < 20.5$ mag arcsec^{-2}. For each of the four "target redshifts" ($z_{\rm sim} = 0.3, 0.5, 0.8$ and 1.0), the images were redshifted in $\Delta z = 0.1$ bins up to $z_{\rm sim} = 1.0$.

In addition to the physical parameters of the input images, the FERENGI output depends on assumptions of the global galaxy evolution model. This evolution is a crude mechanism that mimics the brightness increase of galaxies with increasing redshift (out to at least $z \sim 1 - 2$). The effect on the redshifted images is simply an empirical addition to the magnitude of a galaxy of the form $M' = e \times z + M$, where M' is the corrected magnitude, and e is the evolutionary correction in magnitudes (i.e., $e = -1$ essentially brightens the galaxy by 1 magnitude by $z = 1$). FERENGI was run on the images for values of e starting from $e = 0$ and decreasing to $e = -3.5$ in increments of $\Delta e = 0.5$. Figure 5 shows several examples of the effects of "losing" spiral/disc features with increasing redshift for two galaxies with $e = 0$.

The final number of FERENGI images produced for each galaxy is ultimately a function of galaxy's redshift (since the new images cannot be resampled at better angular resolution than the original SDSS data), as well as the number of e values selected. Table 2 summarizes the redshifted images produced for GZH.

4.2. *Correcting morphologies for classification bias*

The approach used in GZH for correcting the weighted classifications for redshift bias rests on the assumption that the *amount* of bias is a function of the apparent size and brightness of the image as seen on screen. This is

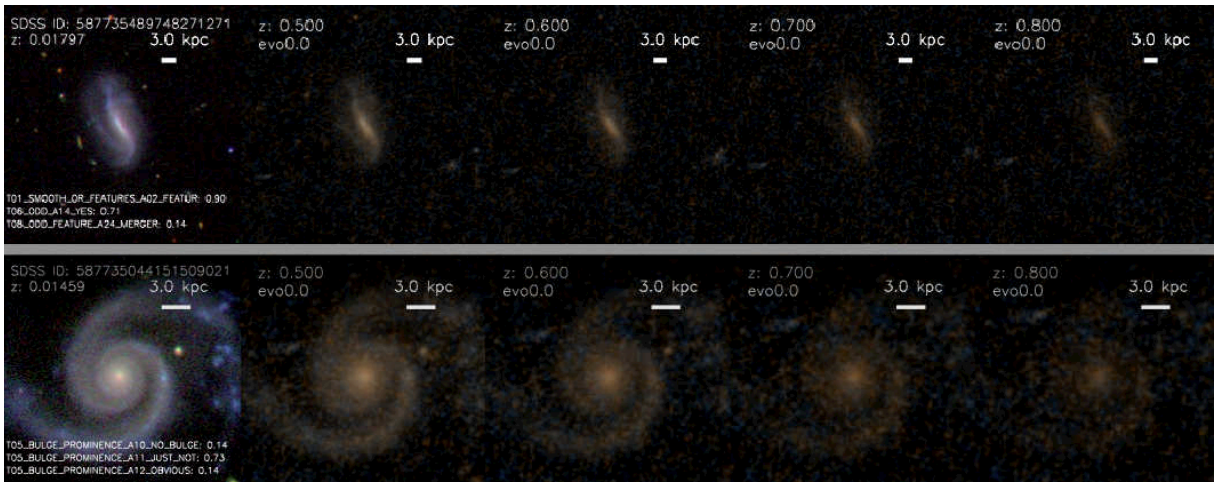


Figure 5. Examples of two galaxies which have been run through the FERENGI code to produce simulated *HST* images. The measured value of $f_{\rm features}$ from GZH for the images in each panel are (1) Top row: $f_{\rm features}$ = (0.900, 0.625, 0.350, 0.350, 0.225) and (2) Bottom row: $f_{\rm features}$ = (1.000, 0.875, 0.875, 0.625, 0.375).

controlled by two types of parameters: **intrinsic** properties of the galaxy itself, such as its physical diameter and luminosity, and **extrinsic** properties, such as the distance (redshift) of the galaxy and its relative orientation. There may also be other parameters that affect classifier accuracy, such as the proximity of close companions ("distraction bias"; see Johnson et al. 2015) or bias as a function of individual classifier ability and skill. The combination of all such parameters forms a high-dimensional space, and it is not clear how to separate this into individual effects. Instead, the method used here employs only two parameters intended to capture the bulk of the change in bias: the r'-band surface brightness (μ_r; intrinsic) and redshift (z; extrinsic).

The change in bias as a function of μ_r and $z_{\rm sim}$ is measured using the FERENGI images over all the evolutionary correction factors. It is assumed that the "true" (ie, debiased) vote fraction $f_{\mu,z}$ for a galaxy can be expressed as:

$$f_{\mu,z} = (f_{\mu,z=0.3}) \times e^{\frac{z-z_0}{\zeta}}, \quad (1)$$

where $f_{\mu,z=0.3}$ is the "calibrated" vote fraction at the lowest redshift in the FERENGI bins ($z = 0.3$) and ζ is a positive parameter that controls the rate at which f decreases with increasing redshift. This formula fits the data relatively well (with very few exceptions, the vote fractions for featured galaxies decrease monotonically with increasing redshift), and the exponential function bounds the observed vote fractions between $f_{\mu,z=0.3}$ and zero. Figure 6 shows the change in vote fraction and the fit to Equation 1 for a random selection of galaxies in the FERENGI images.

GZH uses the values of ζ for *all* sets of artificially redshifted galaxies to fit the overall distribution as a function of surface brightness, since the correction being applied is expected to vary as a function of the intrinsic galaxy properties. The galaxies that can be used to measure the calibration are restricted to those with a non-zero $f_{\rm features}$ at $z_{\rm sim} = 0.3$ and with a reasonable fit ($\Delta\chi^2 > 3.0$) to the exponential model for the measured change in $f_{\rm features}$.

Figure 7 shows the results of fitting the FERENGI images with Equation 1; the correction is only a weak function of surface brightness (μ). Higher-surface brightness galaxies have stronger average corrections, likely because these galaxies are more likely to have larger $f_{\rm features}$ values at high redshifts. Low surface brightness galaxies are more likely to begin low and remain low; the bounded nature of the dropoff (and variance among the individual voters) means that the average magnitude of ζ will be lower.

The data in Figure 7 are fit with a linear function such that:

$$\log_{10}(\hat{\zeta}) = \zeta_0 + (\zeta_1 \times \mu), \quad (2)$$

where $\hat{\zeta}$ is the correction factor applied to each galaxy as a function of surface brightness. The best-fit parameters to the linear fit (from least-squares optimization) are $\zeta_0 = 0.1$, $\zeta_1 = 1.4$. To make the final debiased correction, the simple exponential form of Equation 1 is modified to bound the debiased vote fractions between f and 1:

$$f_{\rm features,debiased} = 1 - (1 - f_{\rm features,weighted}) e^{\frac{z-z_0}{\hat{\zeta}}}. \quad (3)$$

4.3. *Effects of morphological debiasing*

Figure 9 shows the change in $f_{\rm features}$ for the FERENGI galaxies relative to their lowest simulated redshift. In this analysis, only galaxies whose lowest simulated redshift image was $z_{\rm sim} = 0.3$ were used (see Table 2), and

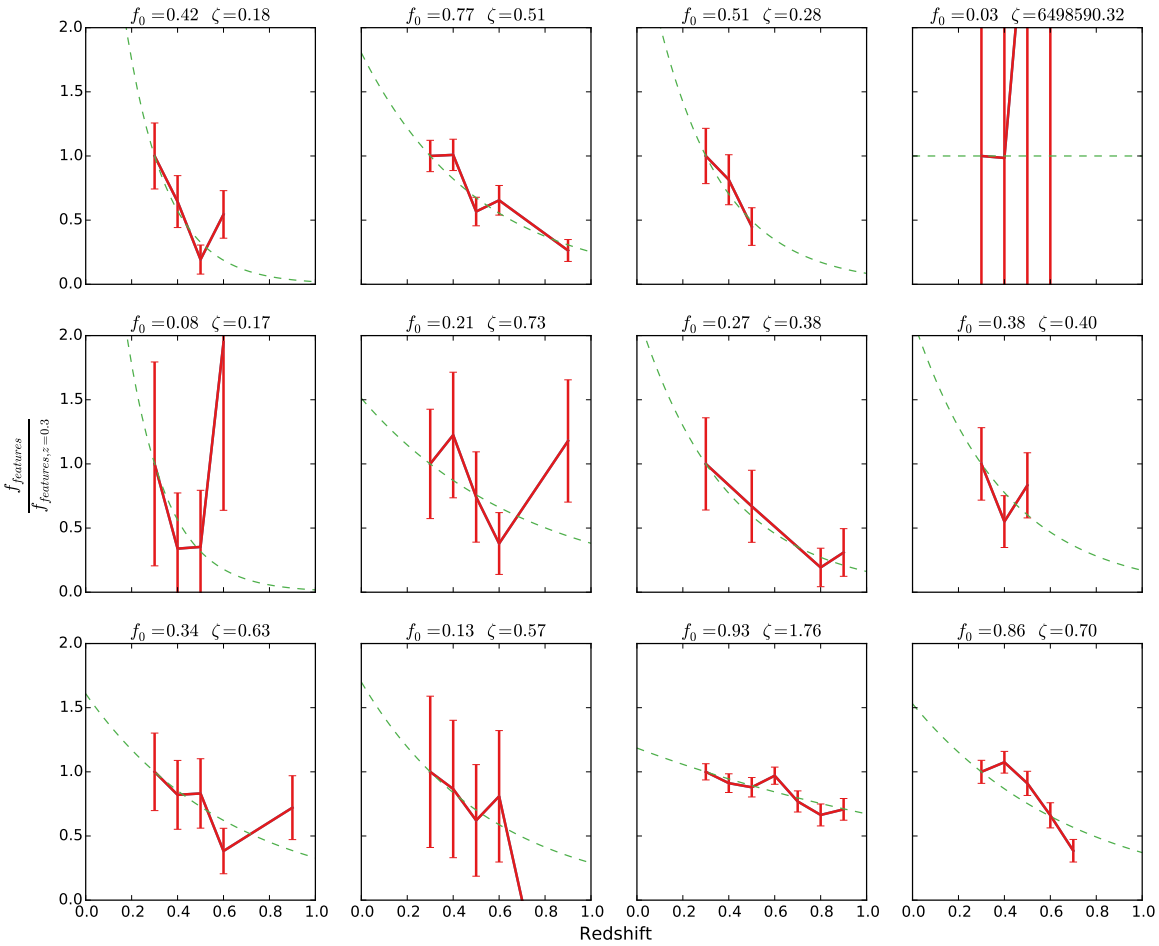


Figure 6. Behavior of the normalized, weighted vote fractions of features visible in a galaxy ($f_{\rm features}$) as a function of redshift in the artificial FERENGI images. Galaxies in this plot were randomly selected from a distribution with $e = 0$ and at least three detectable images in redshift bins of $z \geq 0.3$. Measured vote fractions (red points) are fit with an exponential function (Equation 1); the best-fit parameters are given above each plot. Error bars are Poissonian, assuming a median of 40 votes per galaxy.

only those which had detectable surface brightness measurements in SEXTRACTOR; this includes 3,950 of the total 6,466 images. For each simulated redshift value $z_{\rm sim}$ at a fixed surface brightness μ, $f_{{\rm features},z}$, the value measured at that simulated redshift, is plotted against $f_{{\rm features},z=0.3}$, the value measured for the same galaxy at $z_{\rm sim}$= 0.3.

The objective is to use these data to predict, for a galaxy with a measured $f_{{\rm features},z}$ value, what its $f_{\rm features}$ value *would have been* if it had been viewed at $z = 0.3$. This predicted value is defined as the debiased vote fraction $f_{\rm features,debiased}$, and is calculated by applying a correction to the measured value of $f_{\rm features}$, determined by the ζ function described in the previous section. A reliable predicted value can be obtained so long as the relationship between $f_{{\rm features},z}$ and $f_{{\rm features},z=0.3}$ is single-valued; that is, for a given $f_{{\rm features},z}$, there is exactly one corresponding value of $f_{\rm features}$ at $z = 0.3$.

Figure 9 shows that the relationship between $f_{{\rm features},z}$ and $f_{{\rm features},z=0.3}$ is *not* always single valued; hence, it is not appropriate to correct galaxies that lie in certain regions of surface brightness/redshift/$f_{\rm features}$ space. These regions tend to have low $f_{\rm features}$ values at high redshift, but a wide range of values at $z = 0.3$. These regions contain two morphological types of galaxies: the first set are genuine ellipticals, which have low values of $f_{\rm features}$ at both high and low redshift. The second group are disks whose features become indistinct at high redshift; hence their $f_{\rm features}$ value at $z = 0.3$ may be quite high, while the value observed at high redshift

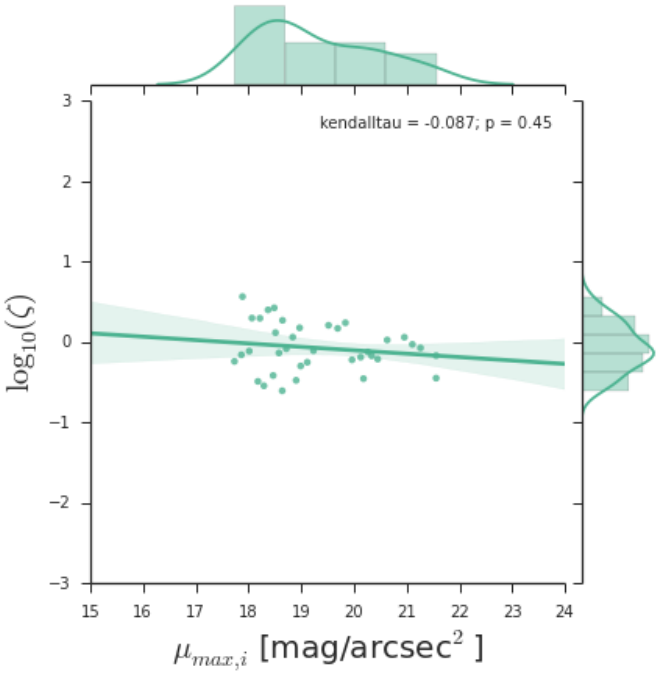


Figure 7. All fits for the vote fraction dropoff parameter ζ for $f_{\rm features}$ in the FERENGI galaxies as a function of surface brightness. This includes only those galaxies with a reasonably bounded range on the dropoff ($-10 < \log(\zeta) < 10$) and sufficient points to fit each function (37 original galaxies, each artificially redshifted to create a total of 296 images

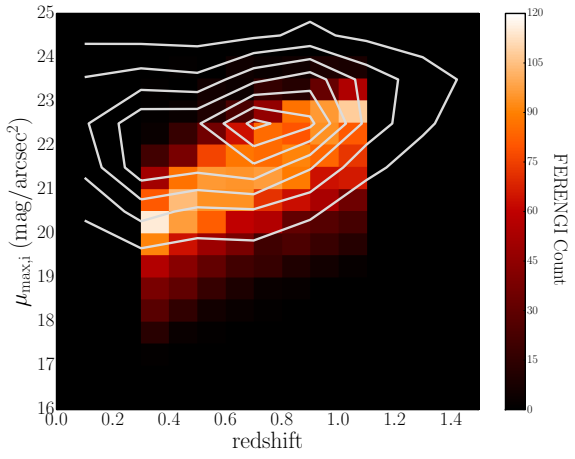


Figure 8. Surface brightness as a function of redshift for 3,950 FERENGI images and the 102,548 galaxies in the ACS sample with measured μ and z values. The color histogram shows the number of FERENGI images as a function of μ and $z_{\rm sim}$. White contours show counts for the full ACS sample, with contours starting at $N = 1000$ and separated by intervals of 1000 up to 7000.

is very low. This effect is strongest at high z and low μ, where features become nearly impossible to discern in the images.

The criteria for determining whether a region of this space is single-valued, and therefore correctable, is as follows: In each surface brightness and redshift bin, the relationship between $f_{\rm features,z}$ and $f_{\rm features,z=0.3}$ is modeled by fitting the data with a polynomial of degrees $n = 3$, 2, and 1, and using the best formal fit out of the three as measured by the sum of the residuals. These fits are shown as the dashed black lines in Figure 9(a). Any flat regions of the polynomial fits are areas in which there is not a clear single-valued relationship between $f_{\rm features,z}$ and $f_{\rm features,z=0.3}$; this is quantified by setting a minimum slope cut of 0.4. Any data in which the polynomial fit has a slope less than this value is considered *not* one-to-one, and therefore $f_{\rm features,z}$ cannot be boosted to its $f_{\rm features,z=0.3}$ value. Galaxies in this region are referred to as the *lower limit* sample, because the most stringent correction available is that the weighted $f_{\rm features}$ is a lower limit to the true value. These regions are highlighted in blue in Figure 9(a). Uncolored (white) regions of the plot have sufficiently high slopes to consider the relationship as single-valued; galaxies in these regions are considered "correctable", and only these are used in measuring the parameters for the ζ function (Section 4.2). Only surface brightness/redshift bins with at least 5 galaxies were considered; regions with fewer than 5 galaxies are considered to have "not enough information" to determine the $f_{\rm features,z}$ and $f_{\rm features,z=0.3}$ relationship, colored gray in Figure 9(a).

The unshaded regions in Figure 9(a) define discrete ranges of redshift, surface brightness, and $f_{\rm features}$ a galaxy must have in order for the debiased correction to be confidently applied to a galaxy in the GZH sample. While the appropriate correctable regions were defined discretely, the true correctable region is assumed to be a smooth function of z, μ, and $f_{\rm features}$. To define this smooth space, we calculate the shape of the convex hull that encloses the correctable and lower-limit FERENGI galaxies in z-μ-$f_{\rm features}$ space. The boundaries are then adjusted until the contamination from both groups is minimized. The resulting hulls define the correctable and lower-limit regions for categorizing the *HST* galaxies. The results of this method and final categorization of the *HST* sample are in Table 4. Of the galaxies at redshift higher than $z = 0.3$, 17% can be debiased using the ζ method, 27% cannot be debiased, and 56% cannot have the potential for a correction determined, due either to an unknown redshift or insufficient FERENGI images corresponding to the relevant z-μ values.

For the "lower-limit" galaxies for which a single debiased $f_{\rm features}$ value cannot be confidently assigned, the

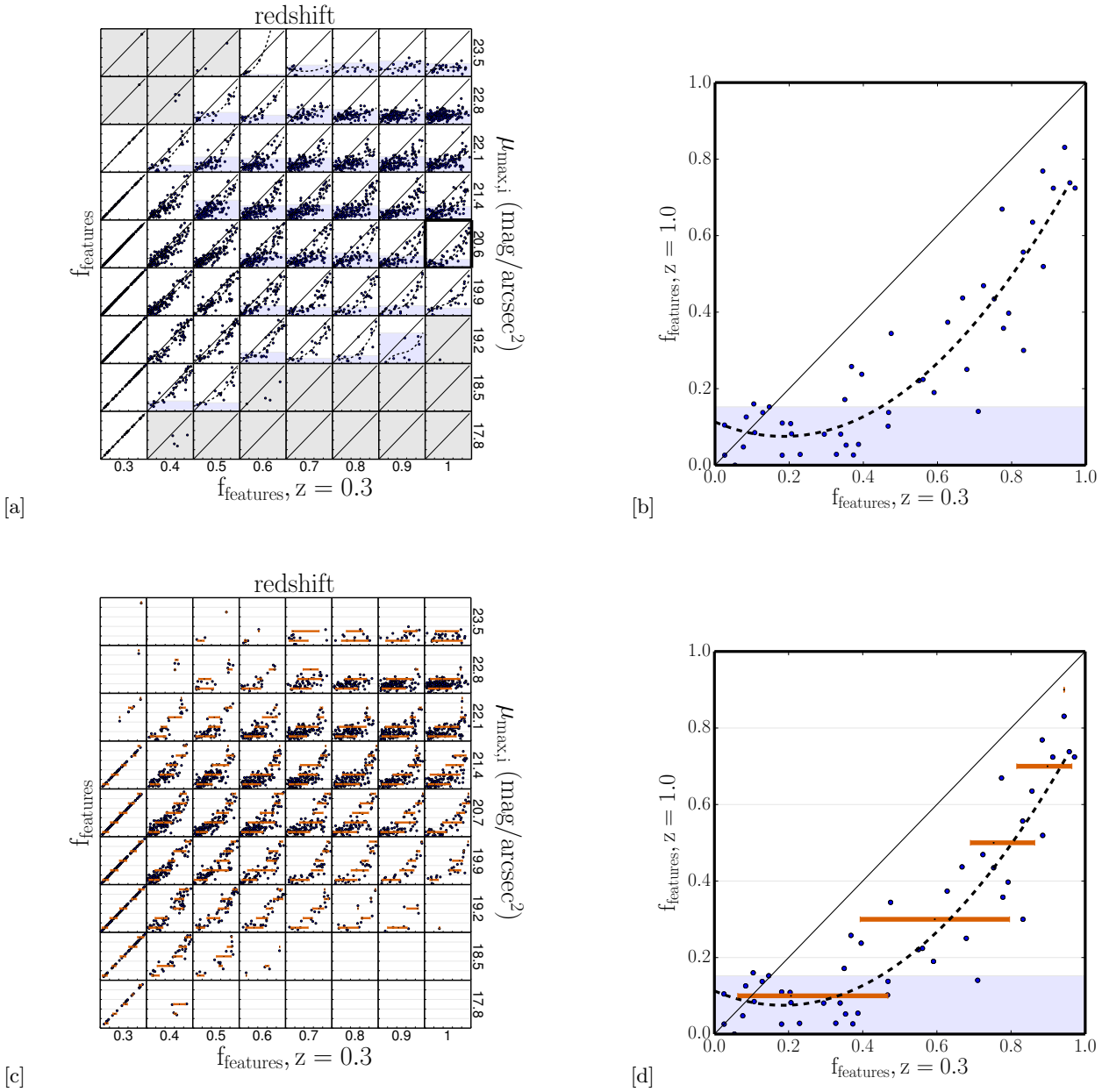


Figure 9. Effects of redshift bias in 3,950 images in the FERENGI sample. [a]: Each point in a given redshift and surface brightness bin represents a unique galaxy. On the y-axis in each bin is the $f_{\rm features}$ value of the image of that galaxy redshifted to the value corresponding to that redshift bin. On the x-axis is the $f_{\rm features}$ value of the image of the same galaxy redshifted to $z = 0.3$. The dashed black lines represent the best-fit polynomials to the data in each square. The solid black line represents $f_{\rm features,z}$=$f_{\rm features,z=0.3}$. Regions in which there is a single-valued relationship between $f_{\rm features}$ at high redshift and at $z = 0.3$ are white; those in which there is not are blue, and those with not enough data ($N < 5$) are gray. [b]: A larger version of the dark-outlined square in [a], containing FERENGI galaxies that have been artificially redshifted to $z = 1.0$ and have surface brightnesses between $20.3 < \mu < 21.0$ (mag/arcsec^2). [c]: The same data as [a] is shown. Each z, μ bin is divided into 4 sub-bins to determine the range of intrinsic $f_{\rm features,z=0.3}$ for a given range of observed $f_{\rm features,z}$ values. In each sub-bin, the orange bars represent the inner 80$^{\rm th}$ percentiles of the data, the boundaries of which are the lower and upper limits of the debiased values. [d]: The same data as [b], but highlighting the upper and lower limit regions.

Table 3. Distribution of FERENGI images analyzed in Figure 9. Correctable images have a single-valued relationship between their measured $f_{\rm features}$ values at high and low redshifts (white regions in Figure 9). Images with only a lower-limit on $f_{\rm features}$ have a non single-valued relationship (blue regions). NEI ("not enough information") images have undetermined relationships due to a lack of data ($N < 5$) in their corresponding z-μ bins (gray regions).

	N	%
Correctable	1,884	48%
Lower-limit	1,986	50%
NEI	80	2%
Total	3,950	100%

range of debiased values is estimated using a method visualized in Figure 9(c). This uses the FERENGI simulated data to analyze the range of intrinsic $f_{\rm features,z=0.3}$ values for any given observed $f_{\rm features}$ value, again as a function of surface brightness and redshift. Each z,μ bin, shows the spread of intrinsic values of $f_{\rm features,z=0.3}$ for four ranges of observed $f_{\rm features}$. The range of intrinsic values is defined as the inner 80% of the data, represented by the orange bars in Figure 9(c). For any galaxy which cannot be directly debiased by the ζ method, these ranges are used to denote the upper and lower limits on the expected values $f_{\rm features,z=0.3}$ as a function of the observed $f_{\rm features}$.

Few galaxies in the sample have sufficiently high corrections to completely change them from being confidently smooth to featured following the bias correction (Figure 10). As a check, highly boosted galaxies are compared to the expert classifications in CANDELS (Kartaltepe et al. 2015). There are only eight galaxies that were strongly boosted ($f_{\rm features} < 0.2$ and $f_{\rm features,best} > 0.5$) in GZH and also appears in the CANDELS expert sample. Of those eight galaxies, Kartaltepe et al. (2015) classify 5 as spheroids, 2 as disks, and 1 as irregular/disky. The $f_{\rm features,best}$ values for GZH are all between 0.5 and 0.6, making them intermediate disk candidates. One of the spheroids has a hint of extended emission to the south that may have been missed by the CANDELS experts, but almost all images are dim and with very low surface brightnesses in the ACS imaging, making positive identification challenging. Since the surveys use different rest-frame filters and the overlap sample is tiny, though, detailed comparisons between the overall morphologies are highly difficult.

4.4. *Challenges of debiasing questions beyond "smooth or features"*

As with the HST images, each FERENGI subject had a varying number of classifiers answering the various questions in the hierarchical decision tree. Every classifier answers the first question, *"Is the galaxy smooth and rounded, with no sign of a disk?"*; as such the vote fractions $f_{\rm smooth}$, $f_{\rm features}$, and $f_{\rm artifact}$ are all computed with the minimum statistical error for any question, with between 40 and 140 responses (see Section 3). The number of participants answering subsequent questions, however, is always equal to or less than the number who answered the preceding question. The average number of responses per task for fourth- or fifth-tier questions such as spiral arm structure (Tasks 12–14) is only 4 ± 4 for the FERENGI sample; while this is distribution is strongly bimodal (reflecting the true morphologies of selected galaxies), the very low absolute numbers of votes introduce very high variance when attempting to calculate a statistical correction.

In the FERENGI data, these numbers severely limit the amount of information that can be extracted for the higher-tier questions. The debiasing technique used (Section 4.2) requires that at least 10 classifiers answer each question for a galaxy with $z_{\rm sim} = 0.3$ *and* the corresponding image at higher redshift. This requirement is (by default) met by all galaxies for the smooth/features question. However, this is often *not* met for questions beyond Task 01; on average. $60\% \pm 24\%$ of the galaxies do not have sufficient data to measure a correction, as compared to 2.0% achieved for Task 01 (Table 3). Of the remaining FERENGI galaxies which did meet this requirement, polynomials were fit to the data in each redshift-surface brightness bin, as was done for Task 01. The goodness-of-fit was evaluated in each bin, for each Task, using a normalized χ^2 metric. The mean χ^2 for all higher-order tasks was 0.10 ± 0.04, significantly larger than the Task 01 value of 0.04. Based on this statistic, the data cannot not be modeled with sufficient confidence for the higher order tasks. An example of the debiasing technique applied to $f_{\rm bar}$ is shown in Section A.2; visual inspection of Figure A5 confirms that the points in each bin are scattered to a degree at which no strong correlation is present. For these reasons, debiased vote fractions are only offered in the GZH catalog for Task 01 (smooth/features).

5. THE GALAXY ZOO: HUBBLE CATALOG

The data release for GZH includes morphological data for 181,101 images (generated from a total of 150,771 unique galaxies). The full table can be accessed at `http://data.galaxyzoo.org`. The online data also includes a secondary metadata table, which is drawn from a variety of sources detailed in Section 2.

Each image is listed under a unique project ID (eg, AHZ000001); the actual galaxy in the image is identified by the combination of the OBJNO and original survey. For each of the 55 responses in the GZH decision tree, the following classification data is provided: for each question, $\texttt{N}_{\texttt{votes}}$ is the number of classifiers

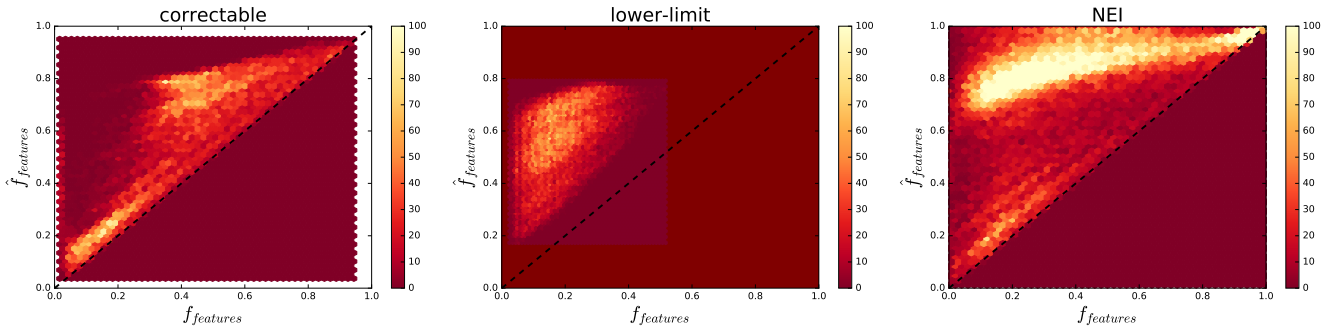


Figure 10. Debiased $f_{\rm features}$ corrected to $z = 0.3$ vs weighted $f_{\rm features}$ for the correctable (left), lower-limit (middle), and NEI (right) galaxies in the GZH sample.

Table 4. Correctable fractions for the top-level task in GZH, split by survey.

	Correction type	AEGIS	COSMOS	GEMS	GOODS-N	GOODS-S	SDSS	Total
Correctable	0	1,654	15,170	1,837	993	835	0	20,489
Lower-limit	1	1,917	26,113	2,423	1,385	1,282	0	33,120
No Correction Needed ($z \leq 0.3$)	2	955	11,926	1,175	415	400	37,545	52,416
NEI	3	2,847	34,511	3,308	2,535	2,523	0	45,724
No Redshift Information	4	1,134	5,088	561	687	102	14,316	21,888
Total		8,507	92,808	9,304	6,015	5,142	51,861	173,637

who answered that question. For each unique answer, `fraction` is the fraction of classifiers who selected that answer (N_{answer}/N_{votes}, and `weighted` is the weighted fraction, which takes into account overall consistency (Section 3.2).

The GZH vote fractions can be largely dependent on the resolution of the image. Two otherwise morphologically identical galaxies which differ significantly in redshift, brightness, or size may result in very different vote fractions for any given question, given that many features of a galaxy are difficult to discern in less-resolved images (bars, spiral arms, disk structure, etc). For this reason, caution must be used when taking vote fractions as cut-offs to determine morphological structure; guidelines for careful classification are given in Section 6.

The GZH catalog is corrected for classification bias only for the first question of the GZH decision tree (Section 4), which asks *"Is the galaxy smooth and round, with no sign of a disk?"* For this question, the catalog provides the additional parameters `debiased`, `lower limit`, `upper limit`, and `best` vote fractions. The `best` fraction for $f_{\rm features}$ is chosen based on the categorization of the galaxy: if it is "correctable", `best` = `debiased`; if it is a lower limit, `best` = `lower limit`; if neither condition applies, then `best` = `weighted`. Flags are also given for the three responses of Task 01 to identify clean samples of galaxies with high likelihoods of being smooth, featured, or an artifact. These are set as `smooth/featured/artifact_flag` = `1` if $f_{\rm smooth/features/artifact\_best} > 0.8$. For a galaxy to be flagged as "smooth", an additional criterion of `correction_type` = 0, 1, or 2 is applied. This is to account for the uncertainty in distinguishing between genuine ellipticals and disks whose features have been washed out due to surface brightness and redshift effects, as described in Section 4.

The debiased and best vote fractions for $f_{\rm smooth}$ are calculated on the criteria that vote fractions for all answers must sum to unity:

$$f_{\rm smooth} \equiv 1 - f_{\rm features,best} - f_{\rm artifact}. \quad (4)$$

In rare cases ($< 1\%$ of the sample), this requirement resulted in negative vote fractions for $f_{\rm smooth}$; these were cases in which the $f_{\rm features}$ vote fraction was boosted to a high value relative to $f_{\rm artifact}$. In these cases, the constraint of Equation 4 is met by setting $f_{\rm smooth} = 0.0$ and decreasing $f_{\rm features,best}$ accordingly. This correction is typically small, with a median decrease/increase of $f_{\rm features}/f_{\rm smooth}$ of $\Delta f = 0.04$.

Data products for GZH are split by the type of image being classified. Table 5 contains the classifications for the *HST* images from the AEGIS, COSMOS, and GEMS surveys, as well as 5-epoch deep imaging from the GOODS-N and GOODS-S surveys. This contains 118,425 galaxies and is the primary output from the GZH project. The next two tables have data for a small subset of 3,927 COSMOS images that were reprocessed to study the effect of color balance on mor-

phological classification. Table 6 has images that are desaturated to minimize the color contrast; Table 7 has images with the red and blue color channels inverted. Table 8 contains data for 6,144 galaxies with 2-epoch images from GOODS. These have been mostly supplanted in the main table with deeper 5-epoch GOODS imaging; however, there are 1,683 galaxies in the shallower imaging that were not classified in the deeper mosaics. This data can also be compared to the counterparts in Table 5 to study the effect of depth on morphological classification. Tables 9 and 10 contain data for the SDSS Stripe 82 single-depth and co-added images, respectively, that were classified using the GZH interface and decision tree. Finally, Table 11 contains classifications for images with artificial point sources intended to simulate the effect of a bright AGN (see Simmons et al. 2014 for an example).

Table 5. GZH morphological classifications for *HST* images from AEGIS, COSMOS, GEMS, and GOODS

			t01_smooth_or_features_		t01_smooth_or_features_a01_smooth_						
Zooniverse ID	Survey ID	Imaging	Correction[1]	$N_{\rm votes}$	fraction	weighted	debiased	best	lower limit	upper limit	flag
AHZ100002g	10010842	AEGIS	0	127	0.118	0.128	0.085	0.085	0.226	0.226	0
AHZ100002h	10010870	AEGIS	4	127	0.567	0.592	0.927	0.592	–	–	0
...											
AHZ20004kd	20014731	COSMOS	3	44	0.682	0.675	0.147	0.675	–	–	0
AHZ20004ke	20014732	COSMOS	2	45	0.689	0.756	0.893	0.756	–	–	0
...											
AHZ400043g	90022729	GEMS	1	121	0.702	0.733	0.487	0.734	0.483	0.800	0
AHZ4000416	90022735	GEMS	1	127	0.646	0.698	0.508	0.698	0.171	0.727	0
...											
AGZ0007z47	10014	GOODS-N-FULLDEPTH	1	40	0.475	0.475	0.197	0.475	0.011	0.496	0
AGZ0007z48	10017	GOODS-N-FULLDEPTH	3	40	0.675	0.675	0.048	0.675	0.168	0.669	0
...											
AGZ00083jb	8869	GOODS-S-FULLDEPTH	1	40	0.425	0.425	0.109	0.425	0.070	0.548	0
AGZ00083jc	8878	GOODS-S-FULLDEPTH	0	40	0.205	0.205	0.048	0.048	−0.005	0.287	0
...											

[1] Flag indicating how the vote fractions for this galaxy were corrected through debiasing (Section 4.3), if possible. 0 = correctable, 1 = lower-limit (f_{raw} — f_{adj} single-valued), 2 = uncorrected ($z_{\rm gal} < 0.3$), 3 = uncorrected (insufficient FERENGI galaxies in this z-μ bin), 4 = uncorrected (no galaxy redshift available).

NOTE—The full version of this table is available in electronic form, as well as at http://data.galaxyzoo.org. The complete version includes data for 118,425 gala morphological information for all tasks in the tree. A subset of the information is shown here to illustrate form and content.

Table 6. GZH morphological classifications for color-faded Hubble images

			t01_smooth_or_features_		t01_smooth_or_features_a01_smooth_							...
Zooniverse ID	Survey ID	Imaging	Correction	$N_{\rm votes}$	fraction	weighted	debiased	best	lower limit	upper limit	flag	
AHZF000001	20000002	COSMOS	1	48	0.708	0.755	0.228	0.754	0.325	0.829	0	
AHZF000003	20000004	COSMOS	3	49	0.367	0.379	0.100	0.379	0.198	0.198	0	
AHZF000004	20000006	COSMOS	3	49	0.265	0.271	0.010	0.270	–	–	0	
AHZF00000z	20000102	COSMOS	1	44	0.727	0.78	0.233	0.780	0.316	0.820	0	
AHZF000010	20000104	COSMOS	2	53	0.811	0.849	0.904	0.848	–	–	0	
...												

NOTE—The full version of this table is available in electronic form, as well as at http://data.galaxyzoo.org. The complete version includes data for 3,927 galaxies and morphological information for all tasks in the tree. A subset of the information is shown here to illustrate form and content.

Table 7. GZH morphological classifications for color-inverted Hubble images

			t01_smooth_or_features_		t01_smooth_or_features_a01_smooth_						...
Zooniverse ID	Survey ID	Imaging	Correction	N_{votes}	fraction	weighted	debiased	best	lower limit	upper limit	flag
AHZC000001	20000002	COSMOS	1	168	0.615	0.664	0.160	0.663	0.271	0.775	0
AHZC000003	20000004	COSMOS	0	235	0.333	0.364	0.002	0.002	0.063	0.063	0
AHZC000004	20000006	COSMOS	3	316	0.235	0.252	−0.011	0.252	–	–	1
AHZC00000z	20000102	COSMOS	1	207	0.755	0.757	0.272	0.756	0.249	0.796	0
AHZC000010	20000104	COSMOS	2	158	0.843	0.882	0.936	0.881	–	–	1
...											

NOTE—The full version of this table is available in electronic form, as well as at http://data.galaxyzoo.org. The complete version includes data for 3,927 galaxies and morphological information for all tasks in the tree. A subset of the information is shown here to illustrate form and content.

Table 8. GZH morphological classifications for GOODS 2-epoch images

			t01_smooth_or_features_		t01_smooth_or_features_a01_smooth_						...
Zooniverse ID	Survey ID	Imaging	Correction	N_{votes}	fraction	weighted	debiased	best	lower limit	upper limit	flag
AHZ3000001	50000000	GOODS-N	0	123	0.390	0.415	0.090	0.090	–	–	0
AHZ3000002	50000001	GOODS-N	2	126	0.341	0.355	0.356	0.356	0.220	0.279	0
AHZ3000003	50000005	GOODS-N	1	129	0.760	0.826	0.633	0.825	0.596	0.834	0
AHZ3000004	50000008	GOODS-N	1	120	0.758	0.787	0.639	0.787	0.658	0.834	0
AHZ3000005	50000010	GOODS-N	1	123	0.854	0.890	0.611	0.889	0.597	0.914	0
...											

NOTE—The full version of this table is available in electronic form, as well as at http://data.galaxyzoo.org. The complete version includes data for 6,144 galaxies and morphological information for all tasks in the tree. A subset of the information is shown here to illustrate form and content.

Table 9. GZH morphological classifications for SDSS Stripe 82 single-epoch images

			t01_smooth_or_features_		t01_smooth_or_features_a01_smooth_				...
Zooniverse ID	Survey ID	Imaging	Correction	N_{votes}	fraction	weighted	debiased	best	flag
AHZ5000001	587730845812064684	SDSS	2	41	0.585	0.595	0.759	0.594	0
AHZ5000002	587730845812065247	SDSS	2	46	0.609	0.651	0.897	0.651	0
AHZ5000003	587730845812196092	SDSS	2	51	0.039	0.044	0.067	0.043	0
AHZ5000004	587730845812196825	SDSS	2	35	0.514	0.605	0.928	0.605	0
AHZ5000005	587730845812524122	SDSS	2	47	0.766	0.812	1.038	0.810	1
AHZ5000006	587730845812654984	SDSS	2	42	0.5	0.542	0.680	0.541	0
AHZ5000007	587730845812655541	SDSS	2	41	0.488	0.526	0.697	0.525	0
AHZ5000008	587730845812720365	SDSS	2	53	0.792	0.84	1.050	0.839	1
AHZ5000009	587730845812720640	SDSS	4	43	0.0	0.0	0.0	0.0	0
AHZ500000a	587730845812720699	SDSS	2	40	0.425	0.478	0.588	0.477	0
...									

NOTE—The full version of this table is available in electronic form, as well as at http://data.galaxyzoo.org. The complete version includes data for 21,522 galaxies and morphological information for all tasks in the tree. A subset of the information is shown here to illustrate form and content.

Table 10. GZH morphological classifications for SDSS Stripe 82 co-added images

			t01_smooth_or_features_		t01_smooth_or_features_a01_smooth_					...
Zooniverse ID	Survey ID	Imaging	Correction	N_{votes}	fraction	weighted	debiased	best	flag	
AHZ6000001	8647474690312306978	SDSS	4	40	0.275	0.289	0.762	0.289	0	
AHZ6000002	8647474690312307154	SDSS	2	43	0.605	0.634	0.858	0.635	0	
AHZ6000003	8647474690312307877	SDSS	2	51	0.608	0.627	0.906	0.627	0	
AHZ6000004	8647474690312308301	SDSS	4	52	0.038	0.038	0.723	0.038	0	
AHZ6000005	8647474690312308318	SDSS	2	44	0.614	0.632	0.776	0.631	0	
AHZ6000006	8647474690312308880	SDSS	2	36	0.667	0.683	0.901	0.683	0	
AHZ6000007	8647474690312372644	SDSS	4	48	0.646	0.674	1.145	0.674	0	
AHZ6000008	8647474690312372789	SDSS	4	45	0.489	0.571	0.964	0.570	0	
AHZ6000009	8647474690312372931	SDSS	4	47	0.553	0.587	0.926	0.587	0	
AHZ600000a	8647474690312373190	SDSS	4	47	0.574	0.559	1.008	0.559	0	
...										

NOTE—The full version of this table is available in electronic form, as well as at http://data.galaxyzoo.org. The complete version includes data for 30,339 galaxies and morphological information for all tasks in the tree. A subset of the information is shown here to illustrate form and content.

Table 11. GZH morphological classifications for *HST* images with simulated AGN

											t01_smooth_or_features_a01_smooth_		
Zooniverse ID	Survey ID	Imaging	Correction	Version	L_{ratio}	AGN color[1]	N_{votes}	fraction	weighted	debiased	best	lower limit	upper
AHZ7000001	90024700	GEMS	1	1	0.2	1	42	0.238	0.239	−0.110	0.238	−0.113	
AHZ7000002	90024700	GEMS	1	1	1.0	1	51	0.255	0.265	−0.107	0.264	−0.128	
AHZ7000003	90024700	GEMS	0	1	5.0	1	47	0.170	0.167	−0.018	−0.018	−0.049	
AHZ7000004	90024700	GEMS	0	1	10.0	1	41	0.195	0.195	0.045	0.045	0.044	
AHZ7000005	90024700	GEMS	0	1	50.0	1	47	0.170	0.178	0.067	0.067	0.146	
...													
AHZ700013m	90024700	GEMS	0	2	0.0	0	35	0.171	0.136	0.011	0.011	0.029	
AHZ700013n	90024700	GEMS	1	2	0.2	1	20	0.150	0.158	−0.278	0.049	−0.351	
AHZ700013o	90024700	GEMS	1	2	1.0	1	32	0.281	0.300	−0.086	0.281	−0.119	
AHZ700013p	90024700	GEMS	0	2	5.0	1	29	0.103	0.115	−0.098	−0.098	−0.152	−
AHZ700013q	90024700	GEMS	0	2	10.0	1	35	0.171	0.181	0.027	0.027	0.023	
AHZ700013r	90024700	GEMS	0	2	50.0	1	34	0.206	0.206	−0.005	−0.005	−0.056	
...													

[1] Flag indicating the color of the PSF in the simulated AGN. 0 = no simulated AGN, 1 = blue, 2 = flat, 3 = red.

NOTE—The full version of this table is available in electronic form, as well as at http://data.galaxyzoo.org. The complete version includes data for 2,961 galaxie information for all tasks in the tree. A subset of the information is shown here to illustrate form and content.

6. USING THE CATALOG

The intended purpose of the GZH catalog is to provide a simple, yet flexible, means of identifying samples of galaxies with a desired morphological type. This section provides instructions for creating such samples using the vote fractions corresponding to the tasks shown in Figure 3. We stress that the selection process will vary based on the particular science case. More conservative cuts can be applied to create pure, but not necessarily complete, samples. These are useful for selecting galaxies exhibiting unique morphologies for individual case studies or observing proposals. Looser cuts can be applied to obtain samples with a higher level of completeness, which will increase the sample size but can potentially decrease purity. Population studies would make use of such large samples, in which statistical sig-

nificance is a crucial factor in evaluating results. Masters et al. (2011); Melvin et al. (2014); Cheung et al. (2015), and Galloway et al. (2015) are examples of GZ papers which select morphologies in this way.

In GZH, participants answered questions about a galaxy's morphology in a decision tree format. With this structure, questions shown to a classifier were dependent on their answers to the previous questions (Section 3.1). Figure 3 shows the possible paths available for answering the questions offered in GZH. The colors represent the tier level of Tasks, which indicate the number of previous Tasks the given Task is dependent upon. The arm-number question, Task 11, is a $5^{\rm th}$-tier Task, meaning that whether this question is seen by a classifier was dependent on four Tasks preceding it. In this case, it as only shown to classifiers who identified the galaxy as featured/disk-shaped in Task 01, not clumpy in Task 12, not edge-on in Task 02, and having spiral structure in Task 10.

To select galaxies of a morphological type identified with a particular Task, a cut is placed on the vote fraction for that Task ($f_{\rm task}$), *as well as* the vote fractions for the Tasks preceding it, because of the dependency induced by the decision-tree structure. For example, to select barred galaxies, a cut may be placed on $f_{\rm bar}$ such that only galaxies where a high fraction of votes for this task voted for the *bar, yes* answer. This is not the only necessary cut, however, since not all classifiers answer this question; only those who have previously selected "features" in Task 01, "not clumpy" in Task 12, and "not edge-on" in Task 02 will have the opportunity to vote on the bar question, Task 03. To ensure that $f_{\rm bar}$ is well-sampled, cuts on all previous tasks must be applied.

The flexibility of this catalog allows users to set their own selection criteria for vote fraction thresholds to create a morphologically pure sample. Table 12 provides baseline cuts for selecting galaxies of a variety of morphologies. These thresholds are determined by a visual inspection of subsamples. For each Task, at least 20 classifiers are required to have voted on the given question along with a cut on the vote fraction is applied to the previous task. We visually analyzed subsamples of 50 galaxies meeting both criteria, as well as a control sample of galaxies which had 20 classifiers vote on the task, but did not meet the threshold cut set for the previous task. The threshold cut was adjusted and new subsamples were inspected until both the original and control samples achieved $> 80\%$ purity.

As an example of how to use Table 12 to create a sample of 3-armed spirals galaxies, we suggest selecting objects with $N_{arm\ number} \geq 20$, $f_{\rm features} > 0.23$, $f_{\rm clumpy,no} > 0.30$, $f_{edgeon,no} > 0.25$, and $f_{spiral,yes} > 0.25$. These cuts define a sample of galaxies of "arm number candidates"; *i.e.*, galaxies for which answering the arm number question makes physical sense and the vote fraction $f_{\rm arm\ number}$ is well-sampled. In this example, such galaxies are featured, non-clumpy, non-edge on spirals. At this point a cut can be made on $f_{\rm arm\ number=3}$ to select spirals with three arms.

Tasks 03, 04 and 05 have an additional possible pathway; as shown in Figure 3, a classifier might also be shown this question if they select "featured/disk" in Task 01, "clumpy" in Task 12, two or more clumps in Task 16, and "spiral arrangement" in Task 15. After applying the appropriate thresholds for this path, $< 0.5\%$ of the galaxies which have ≥ 20 answers to these questions used this pathway to arrive at these Tasks. Further, of these subjects, none were found to exhibit disk structure, although the clumps within were arranged in a spiral pattern.

This section described how to use the previous Task thresholds in Table 12 to select well-sampled galaxies which may or may not contain the feature associated with a unique Task. The following two examples show to use the vote fractions for a particular Task to obtain a sample of galaxies with a certain morphological type.

6.1. *Example 1: Selecting barred galaxies*

A sample of barred disk galaxies is created by applying cuts on the previous tasks as listed in Table 12. 11,049 "bar candidates", galaxies for which asking the bar question is meaningful, were selected by applying the cuts $N_{bar} \geq 20$, $f_{\rm features} > 0.23, f_{\rm clumpy,no} > 0.30$, and $f_{edgeon,no} > 0.25$. These galaxies are featured, non-clumpy, non-edge on galaxies. Of these, a pure sample of 730 barred disks was identified by applying a cut of $f_{\rm bar} > 0.7$. A subsample of 50 galaxies were visually inspected and 94% were found to contain strong bars. A complete sample of strong and weak bars was created by applying a cut of $f_{\rm bar} > 0.3$. This sample contained 3,218 galaxies, 86% of which were found to contain weak or strong bars through visual inspection in a subsample of 50.

The resulting bar sample can be used to estimate the redshift evolution of bar fraction and find a steady decrease of $f_{\rm bar} \sim 0.32$ at $z = 0.4$ to $f_{\rm bar} \sim 0.24$ at $z = 1.0$. The decrease in bar fraction agrees with Melvin et al. (2014), although they report a lower overall bar fraction going from $f_{\rm bar} = 0.22$ at $z = 0.4$ to $f_{\rm bar} = 0.11$ at $z = 1.0$. The difference in total bar fraction is expected, as this analysis used a looser cut on $f_{\rm bar}$, there is no luminosity cut, and the use of debiased values for $f_{\rm features}$ increases the total amount of disks in the sample compared to Melvin et al. (2014). For a more thorough analysis of the bar fraction, a cut on $f_{\rm bar}$ that evolves with redshift may yield more accurate results, since there are no explicit debiased values of $f_{\rm bar}$ that would take redshift induced bias into account.

6.2. *Example 2: Identifying clump multiplicity*

Clumps are known to be a characteristic feature of galaxies outside the local universe, and there is evidence they play a crucial role in the evolution of modern spirals, particularly in the formation of central bulges (Elmegreen et al. 2005; Elmegreen & Elmegreen 2014; Guo et al. 2015; Behrendt et al. 2016). Simulations show clumps migrate from the outer disk to the galactic center in only a few orbital periods (Mandelker et al. 2015) and observations show increasing bulge to clump mass and density ratios as the universe evolves since $z \sim 1.5$ (Elmegreen et al. 2009), suggesting that clumps coalesce over time to form the modern bulges of disk galaxies. GZH added a "clumpy" path to the decision tree for the purpose of both identifying clumps and investigating their evolution with redshift.

For galaxies identified as "clumpy" in GZH, the number of clumps can be determined using Task 16. Table 12 can be used to select 8,444 galaxies measured as "clumpy" using $f_{\rm features} > 0.23$ and $f_{\rm clumpy,yes} > 0.80$ to ensure the vote fractions for Task 16 are well-sampled. The clump number can be reasonably identified for 1,112 of the clumpy galaxies; for the remainder, the unique clumps were less distinguishable from each other and the exact number of clumps could not be deduced without careful visual inspection. In the 1,112 which did have distinguishable clumps, there are 61 one-clump galaxies using $f_{1\ \rm clump} > 0.50$, 442 two-clump galaxies using $f_{2\ \rm clumps} > 0.80$, 275 three-clump galaxies using $f_{3\ \rm clumps} > 0.75$, 71 four-clump galaxies using $f_{4\ \rm clumps} > 0.70$, and 263 galaxies with more than four clumps using $f_{>4\ \rm clumps} > 0.70$. Alternatively, these data may be used to create more general samples of clumpy galaxies with few clumps and many clumps. A sample of 989 "few clumps" galaxies can be made using $(f_{1\ \rm clump} + f_{2\ \rm clumps}) > 0.5$ and 2,910 "many clumps" galaxies using $(f_{3\ \rm clumps} + f_{4\ \rm clumps} + f_{>4\ \rm clumps}) > 0.5$.

Table 12. Suggested thresholds for selecting morphological samples from GZH.

No.	Task	Previous task(s)	Vote fraction threshold $N_{\rm task} \geq 20$
01	smooth or features	–	–
02	edge on	01,12	$f_{\rm clumpy,no} > 0.30$
03	bar	01,12,02	$f_{\rm edgeon,no} > 0.25$
		01,12,16,15	$f_{\rm clumpy\ spiral} > 0.65$
04	spiral arms	01,12,02	$f_{\rm edgeon,no} > 0.25$
		01,12,16,15	$f_{\rm clumpy\ spiral} > 0.65$
05	bulge prominence	01,12,02	$f_{\rm edgeon,no} > 0.25$
		01,12,16,15	$f_{\rm clumpy\ spiral} > 0.65$
06	odd yes/no	–	–
07	rounded	01	$f_{\rm smooth} > 0.70$
08	odd feature	06	$f_{\rm odd,yes} > 0.50$
09	bulge shape	01,12,02	$f_{\rm edgeon,yes} > 0.40$
10	arms winding	01,12,02,04	$f_{\rm spiral,yes} > 0.25$
11	arms number	01,12,02,04	$f_{\rm spiral,yes} > 0.25$
12	clumpy	01	$f_{\rm features} > 0.23$
13	bright clump	01,12,16	$f_{\rm one\ clump} < 0.40$
14	bright central clump	01,12,16,13	$f_{\rm bright\ clump,yes} > 0.50$
15	clump arrangement	01,12,16	$f_{\rm multiple\ clumps} > 0.45$
16	clump count	01,12	$f_{\rm clumpy,yes} > 0.80$
17	clumps symmetrical	01,12	$f_{\rm clumpy,yes} > 0.80$
18	clumps embedded	01,12	$f_{\rm clumpy,yes} > 0.80$

7. ANALYSIS

7.1. *Demographics of morphology*

Any analysis of the morphological distribution of galaxies must properly consider morphology with respect to other physical properties of the sample, such as color, mass, size, environment, and redshift. We defer such analyses to later papers, and comment here only on a few broad characteristics of the overall GZH sample.

Figure 11 shows the breakdown of morphologies for the combined surveys in GZH as a flow diagram, where each node represents a task and the flows between the nodes represent the possible responses. The width of the flows is proportional to the number of galaxies associated with that response; Figure 11 uses a simple plurality vote to assign the combined set of labels for each galaxy. This method is likely to undersample some of the categories; visual inspection of images suggests that $f_{\rm features} \gtrsim 0.3$ identifies disk-dominated or irregular galaxies reliably, for example. This method also does not use the suggested thresholds in Table 12.

Given these caveats, the overall distribution of galaxy types is significantly different than the low-redshift sample classified in SDSS imaging from GZ1 and GZ2. Lintott et al. (2011) found that elliptical galaxies exceeded spiral galaxies by a factor of $\sim 2:1$ in the main spectroscopic sample if using a plurality vote criterion (although Bamford et al. 2009 show that this strongly depends on the selection method; spirals are the dominant population in a volume-limited sample at $z < 0.088$, for example). The weighted votes for GZH have smooth galaxies outnumbering disks and clumpy galaxies by a factor of $\sim 3:1$, however. The fraction of objects identified as stars or artifact is also much higher in the Hubble imaging; by plurality votes, these encompass only $\sim 0.1\%$ of images in SDSS (Willett et al. 2013), but 6% of images in GZH.

Within the sample of galaxies identified as "not smooth", it is clear that the addition of the clumpy branch is necessary to describe a large fraction of

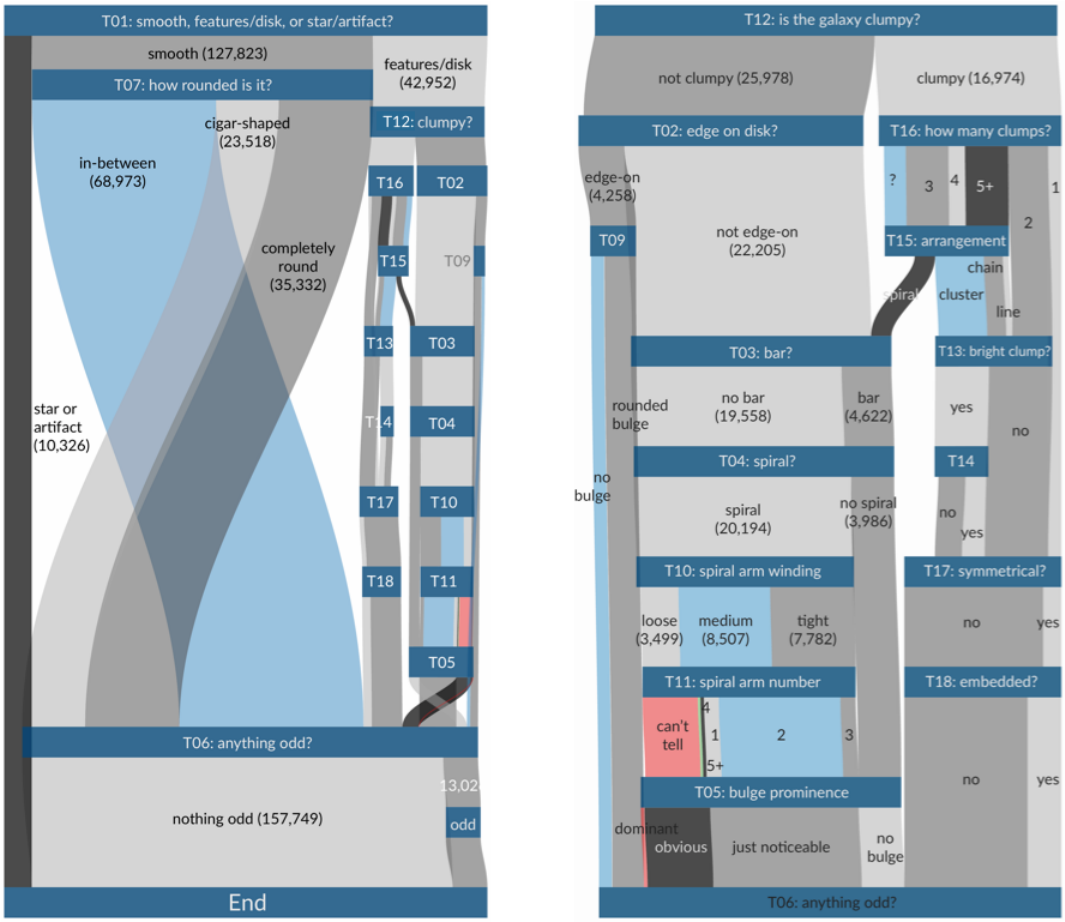


Figure 11. Demographics of the morphologies for all galaxies in GZH (including both *HST* and SDSS imaging). Each node in each diagram (dark blue horizontal bars of uniform height) represents a task in the tree. The left diagram shows the full decision tree; the right diagram zooms in on the features/clumpy tasks, which are otherwise difficult to see. The paths between tasks represent each possible answer to the task; these flow from top to bottom between their origin question and the subsequent task in the tree. Labels are assigned to each galaxy based on the plurality answer for each task. A galaxy is assigned only one label at each node. Widths of the paths are proportional to the number of galaxies assigned to that path; the widths of the nodes are proportional to the number of galaxies for which the question was reliably answered.

the sample; disk-dominated galaxies outnumber clumpy morphologies by less than a factor of 2. Disk galaxies are primarily unbarred Melvin et al. (2014) and possessing two visible spiral arms over a flat distribution of pitch angles and bulge prominence. Clumpy galaxies are identified across the full range of clump multiplicities, with the exception of 1-clump galaxies (which would be difficult to differentiate from compact spheroids). Roughly half of the galaxies have at least 1 clump identified as the brightest; the clumps are most commonly asymmetrically arranged in clusters and are not often seen embedded in larger structures.

7.2. *Comparing GZH morphologies to other catalogs*

All of the Legacy surveys included in the GZH imaging have had morphological catalogs previously published; these catalogs have significant differences in the number of galaxies, size and magnitude limits, classification scheme, and the methods used for measuring morphology. These catalogs have been cross-matched to GZH to compare the results; this is not presented as an endorsement of any particular method, but as an exploration of the strengths and weaknesses of the GZ crowdsourced catalogs as compared to products made with machine learning, automatic fits, and expert visual classification.

The types and accuracy of morphological classification strongly depend on the sample and methods being used. In an attempt to make a consistent comparison between different techniques, galaxies are broadly grouped into three categories: bulge-dominated/elliptical/smooth, disk-dominated/spiral, and irregular/clumpy. These categories are compared to two parameters in GZH: $f_{\rm features,best}$, which is designed to identify smooth (elliptical) galaxies, and $f_{\rm odd}$, which is designed to identify deviations from well-formed spirals or S0s and which constitutes a "catch-all" for the variety of asymmetric morphologies that can constitute an irregular galaxy.

Morphologies for GEMS galaxies were measured by

Häußler et al. (2007), who used single-component Sérsic fits to the F850LP imaging. This analysis used parameters from the GALFIT code, which Häußler et al. (2007) evaluated as more reliable than GIM2D for GEMS due to the ability to fit multiple galaxies in crowded fields. A 1″ positional match for the GEMS galaxies in Table 9 in Häußler et al. (2007) to the GZH sample gives 8,846 galaxies in both Häußler et al. (2007) and GZH. The main morphological parameter in the automated catalog is the Sérsic index n defining the radial surface brightness profile. "Elliptical" galaxies are selected by $n > 2.5$ and "spirals" by $n > 2.5$. There is no automatic measurement of irregular or clumpy structure in this catalog.

AEGIS galaxies were morphologically classified using non-parametric measurements by Lotz et al. (2008). This method used a combination of the Gini coefficient (G), which measures the relative inequality in pixel brightness, and M_{20}, the second-order moment of the brightest 20% of the light (Lotz et al. 2004). A linear combination of G and M_{20} delineates three broad categories of galaxy morphology: E/S0/Sa ("elliptical"), Sb/Sc/Ir ("spiral"), and mergers ("irregular"). A 1″ positional match on AEGIS and GZH yields 4,031 galaxies with reliably-measured morphologies and $S/N > 3$ in both V- and I-bands.

Galaxies in both of the GOODS fields down to a limit of $z_{AB} = 22.5$ were visually classified by a single expert (R.S. Ellis), inspecting both z-band and composite Viz color images (Bundy et al. 2005). These morphologies are assigned a numerical value based on categories in Brinchmann et al. (1998); the corresponding morphologies used are "elliptical" (classes 0,1,2), "spirals" (classes 3,4,5), and "irregular" (classes 6,7,8). A 0.5″ matched radius yields 2,435 galaxies (1300 in GOODS-N, 1135 in GOODS-S) in both Bundy et al. (2005) and GZH.

COSMOS galaxies have multiple published datasets automatically classifying morphology, all using a variation of non-parametric measurements. Cassata et al. (2007) used a combination of concentration (C), asymmetry (A), G, and M_{20} (Cassata et al. 2005) to classify all galaxies with $m_{\rm I,petro} < 25$, and used an empirical division based on several hundred training images to assign galaxies to discrete morphological categories. A similar method is employed by Tasca (2011), using the same non-parametric indices but with a different method of calculating the Petrosian radius and total light profile. They employ a nearest-neighbors method is used to assign morphological categories for the entire sample based on a small training set. Scarlata et al. (2007, ZEST) used C, A, G, M_{20}, the galaxy ellipticity (ϵ), and Sérsic index (n) to quantify the galaxy shapes; a principal component analysis is used to assign galaxies to discrete morphological categories. The categories in all three methods are split into ellipticals, spirals, and peculiar/irregulars. A 0.5″ radius search on GZH matches to 76,241 galaxies in Cassata et al. (2007), 76,608 in Tasca (2011), and 77,425 in Scarlata et al. (2007).

Figure 12 shows the proportion of galaxies as split by their automated/expert visual morphologies for each of the six catalogs matched to GZH. The relation between GZH and the automated/expert morphologies is shown to vary significantly for different surveys — this could be due to intrinsic differences in the galaxies, in the techniques used to measure morphology, or both. The distribution of $f_{\rm features,best}$ for AEGIS galaxies is essentially flat for spirals, showing almost no difference in proportion at the highest and lowest ends. Early-type galaxies are at their highest proportion at $f_{\rm features,best} \sim 0$, but constitute at most only 40% of galaxies. In contrast, GOODS galaxies at low $f_{\rm features,best}$ are dominated by early-types, although spiral and irregular galaxies have comparable sizes as low as $f_{\rm features,best} \sim 0.2$. In GEMS galaxies, the impact of irregular or peculiar galaxies is not measured; splitting on $n = 2.5$ gives an essentially flat proportion of ellipticals to spirals at $f_{\rm features,best} > 0.3$. Visual inspection of images of galaxies where both $f_{\rm features,best}$ and n are high show that the majority are obvious spirals but with prominent bulges, indicating that the single-component Sérsic fit is likely choosing too small of an effective radius and missing the extended disk structure for a large population of galaxies.

Even within the same sample of galaxies in COSMOS, different techniques of measuring morphology show significant differences. Both Cassata et al. (2007) and Tasca (2011) have higher proportions of early-type galaxies at $f_{\rm features,best} \sim 0$, but constitute less than half the sample in both cases. The majority of all galaxies in Scarlata et al. (2007), in contrast, are spirals; oddly, the *highest* proportions of spirals are found at the lowest values of $f_{\rm features,best}$. The decrease in spiral galaxies at higher $f_{\rm features,best}$ is almost entirely balanced by an increase in irregulars. Early-types have a small bump near $f_{\rm features,best} \sim 0.15$, but never constitute more than 10% of the sample.

The second set of plots in Figure 12 indicates that $f_{\rm odd}$ is largely an effective method for separating irregular and/or peculiar galaxies from spirals. Despite the somewhat vague wording of the *"anything odd?"* question in GZH, the AEGIS, COSMOS, and GOODS all show a marked increase in the irregular fraction increasing with $f_{\rm odd}$. Any threshold value for distinguishing between the two, however, depends strongly on survey/redshift range.

As the data shown in Figure 12 cover a wide range of redshifts and sizes, a volume-limited sample of galaxies (or at least binning by redshift) is likely a much more appropriate comparison. While detailed analysis is left

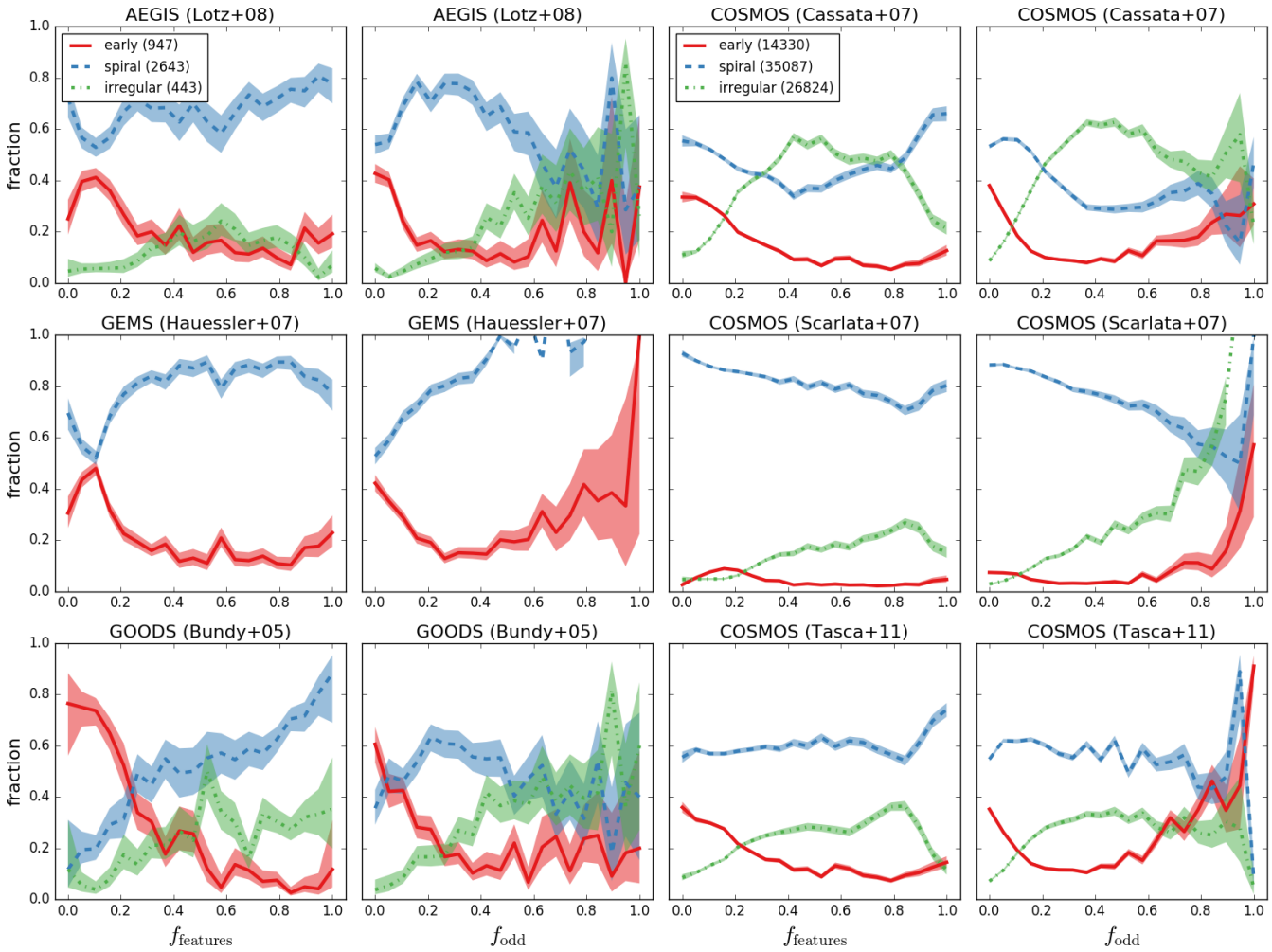


Figure 12. Distributions of morphological parameters for galaxies matched in GZH and other published catalogs, split by survey (AEGIS, COSMOS, GEMS, and GOODS). The first and third columns show the fraction of overall galaxies in the matched samples, split by their published morphologies as a function of the GZH $f_{\rm features,best}$. The corresponding second and fourth columns show data for the same galaxies as a function of $f_{\rm odd}$. Confidence intervals for each binned fraction in the bin are calculated for a binomial population (Cameron 2011).

to a further paper, we note that there is no strong change in the proportion of galaxies when binning by redshift intervals of $\Delta z = 0.2$ out to $z = 1.0$.

Finally, there are 7,681 galaxies in the GOODS-S field with morphological classifications in both GZH and the Galaxy Zoo: CANDELS project (Simmons et al., submitted). Since both the sensitivity and filters for the two sets of images differ (and there is no debiased correction applied to GZC), there is no prior reason to expect a perfect correlation between the separate vote fractions for the projects. Briefly, we note that the $f_{\rm features}$ value for GZH is on average higher than GZC; the effect is strongest at $f_{\rm features,GZC} < 0.3$, for which roughly half the galaxies have $f_{\rm features,GZH} > 0.5$. However, the correlation between vote fractions is single-valued (although not linear, with a Pearson correlation coefficient of $r = 0.6$), and should be possible to calibrate using a similar approach to that described in Section 4; the correlation between other tasks, such as edge-on galaxies is significantly stronger ($r = 0.9$). So while the raw vote fractions are not directly comparable, the initial analysis indicates that the broad morphologies are at least consistent.

8. SUMMARY

This paper presents the catalog release for the Galaxy Zoo: Hubble project, which uses crowdsourced visual classifications to measure galaxy morphologies. The first two phases of Galaxy Zoo (Lintott et al. 2011; Willett et al. 2013) used images of low-redshift galaxies from SDSS; this is the first result of the project with space-based images of high-redshift targets (in addition to the Galaxy Zoo: CANDELS collaboration; Simmons et al., submitted). The final sample includes classifications for 181,101 images generated from 150,771 unique galaxies.

Galaxies were selected from a brightness-limited sample from multiple Legacy surveys using the Advanced Camera for Surveys on the Hubble Space Telescope, including AEGIS, GEMS, GOODS-N, GOODS-S, and COSMOS. The catalog also includes classifications for 51,861 Sloan Digital Sky Survey images in Stripe 82 at relatively low redshift; these serve both as a low-redshift anchor for cosmological studies and a potential comparison for the different epochs of classification between GZH and Galaxy Zoo 2 (Willett et al. 2013).

The data for the GZH catalogs has been extensively tested and reduced. The dominant effect is a known bias against identifying disky and asymmetric sub-structures at either low resolution or surface brightness. This can be the result either of genuinely small or dim galaxies, or a perceived effect from observing galaxies at further distances (higher redshift). To calibrate this *without* potentially overcorrecting for the genuine morphological evolution of galaxies over cosmic time, the GZH project uses SDSS images of low-redshift galaxies, processes them to appear as if they were at higher redshift, and classifies them through the GZH interface in an identical fashion. The resulting change in f_{features} as a function of z and μ is applied as a multiplicative correction to the top-level vote fractions for $\sim 50\%$ of the GZH galaxies.

Galaxies in GZH show significant changes in the disk/elliptical fraction as a function of redshift, along with an increasing number of galaxies dominated by smaller clumps and presumed to be in the process of building up their baryonic mass through a combination of hierarchical merging and in-situ star formation. While the majority of scientific interpretation is left to future work, this paper confirms the decrease in observed bar fraction with increasing redshift (Melvin et al. 2014) and identifies a new way for selecting clumpy galaxies as a function of clump multiplicity.

The full data tables for the catalogs can be accessed in machine-readable form from both the journal and at `http://data.galaxyzoo.org`. All the code and data tables used to generate this manuscript can be found at `https://github.com/willettk/gzhubble`.

We thank Meg Schwamb and the ASIAA for hosting the "Citizen Science in Astronomy" workshop, 3-7 Mar 2014 in Taipei, Taiwan, at which some of this analysis was done. We thank Jennifer Lotz for sharing her G-M_{20} measurements for the AEGIS sample. We thank Coleman Krawczyk for his assistance in producing Figure 3.

This project made heavy use of the Astropy packages in Python (Astropy Collaboration et al. 2013), the `seaborn` plotting package (Waskom et al. 2015), astroML (Vanderplas et al. 2012), and TOPCAT (Taylor 2005, 2011). Modified code from Nick Wherry and David Schlegel was used to create the JPG images. Figure 11 was generated with `http://sankeymatic.com/`.

KS gratefully acknowledges support from Swiss National Science Foundation Grant PP00P2_138979/1.

This work is based on (GO-10134, GO-09822, GO-09425.01, GO-09583.01, GO-9500) program observations with the NASA/ESA Hubble Space Telescope, obtained at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-26555.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS website is `http://www.sdss.org/`.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory and the University of Washington.

REFERENCES

Abraham R. G., van den Bergh S., Nair P., 2003, ApJ, 588, 218
Astropy Collaboration et al., 2013, A&A, 558, A33
Baillard A. et al., 2011, A&A, 532, A74
Bamford S. P. et al., 2009, MNRAS, 393, 1324
Bamford S. P., Rojas A. L., Nichol R. C., Miller C. J., Wasserman L., Genovese C. R., Freeman P. E., 2008, MNRAS, 391, 607
Barden M., Häußler B., Peng C. Y., McIntosh D. H., Guo Y., 2012, MNRAS, 422, 449
Barden M., Jahnke K., Häußler B., 2008, ApJS, 175, 105
Beckwith S. V. W. et al., 2006, AJ, 132, 1729
Behrendt M., Burkert A., Schartmann M., 2016, ApJL, 819, L2
Bell E. F. et al., 2012, ApJ, 753, 167
Bertin E., Arnouts S., 1996, A&AS, 117, 393

Brinchmann J. et al., 1998, ApJ, 499, 112
Bundy K., Ellis R. S., Conselice C. J., 2005, ApJ, 625, 621
Buta R. J., 2013, Galaxy Morphology, Oswalt T. D., Keel W. C., eds., Springer, pp. 1–89
Caldwell J. A. R. et al., 2008, ApJS, 174, 136
Cameron E., 2011, PASA, 28, 128
Cameron E., Carollo C. M., Oesch P. A., Bouwens R. J., Illingworth G. D., Trenti M., Labbé I., Magee D., 2011, ApJ, 743, 146
Cardamone C. N. et al., 2010, ApJS, 189, 270
Cassata P. et al., 2005, MNRAS, 357, 903
Cassata P. et al., 2007, ApJS, 172, 270
Cheung E. et al., 2015, MNRAS, 447, 510
Chevance M., Weijmans A.-M., Damjanov I., Abraham R. G., Simard L., van den Bergh S., Caris E., Glazebrook K., 2012, ApJL, 754, L24
Conselice C. J., 2003, ApJS, 147, 1
Conselice C. J., 2014, ARA&A, 52, 291
Darg D. W. et al., 2010, MNRAS, 401, 1552
Davis M. et al., 2007, ApJL, 660, L1
de Vaucouleurs G., 1959, Handbuch der Physik, 53, 275
Dressler A., 1980, ApJ, 236, 351
Elmegreen D. M., Elmegreen B. G., 2014, ApJ, 781, 11
Elmegreen D. M., Elmegreen B. G., Ferguson T., Mullan B., 2007, ApJ, 663, 734
Elmegreen D. M., Elmegreen B. G., Marcus M. T., Shahinyan K., Yau A., Petersen M., 2009, ApJ, 701, 306
Elmegreen D. M., Elmegreen B. G., Rubin D. S., Schaffer M. A., 2005, ApJ, 631, 85
Förster Schreiber N. M., Shapley A. E., Erb D. K., Genzel R., Steidel C. C., Bouché N., Cresci G., Davies R., 2011, ApJ, 731, 65
Freeman P. E., Izbicki R., Lee A. B., Newman J. A., Conselice C. J., Koekemoer A. M., Lotz J. M., Mozena M., 2013, MNRAS, 434, 282
Galloway M. A. et al., 2015, MNRAS, 448, 3442
Genel S. et al., 2014, MNRAS, 445, 175
Giavalisco M. et al., 2004, ApJL, 600, L93
Griffith R. L. et al., 2012, ApJS, 200, 9
Grogin N. A. et al., 2011, ApJS, 197, 35
Guo Y. et al., 2015, ApJ, 800, 39
Häußler B. et al., 2007, ApJS, 172, 615
Hinshaw G. et al., 2013, ApJS, 208, 19
Hopkins P. F. et al., 2010, ApJ, 715, 202
Hubble E. P., 1926, ApJ, 64, 321
Hubble E. P., 1936, Realm of the Nebulae. Yale University Press
Ilbert O. et al., 2013, A&A, 556, A55
Johnson L. C. et al., 2015, ApJ, 802, 127
Kartaltepe J. S. et al., 2015, ApJS, 221, 11
Koekemoer A., Fruchter A., Hack W., 2003, Space Telescope European Coordinating Facility Newsletter, Volume 33, p.10, 33, 10
Koekemoer A. M., Fruchter A. S., Hook R. N., Hack W., 2002, in The 2002 HST Calibration Workshop : Hubble after the Installation of the ACS and the NICMOS Cooling System, Proceedings of a Workshop held at the Space Telescope Science Institute, Baltimore, Maryland, October 17 and 18, 2002. Edited by Santiago Arribas, Anton Koekemoer, and Brad Whitmore. Baltimore, MD: Space Telescope Science Institute, 2002., p.339, pp. 339–+
Krist J., 1993, in Astronomical Society of the Pacific Conference Series, Vol. 52, Astronomical Data Analysis Software and Systems II, R. J. Hanisch, R. J. V. Brissenden, & J. Barnes, ed., pp. 536–+
Lackner C. N., Gunn J. E., 2012, MNRAS, 421, 2277
Land K. et al., 2008, MNRAS, 388, 1686
Law D. R., Shapley A. E., Steidel C. C., Reddy N. A., Christensen C. R., Erb D. K., 2012a, Nature, 487, 338
Law D. R., Steidel C. C., Shapley A. E., Nagy S. R., Reddy N. A., Erb D. K., 2012b, ApJ, 745, 85
Lintott C. et al., 2011, MNRAS, 410, 166
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Lotz J. M. et al., 2008, ApJ, 672, 177
Lotz J. M., Primack J., Madau P., 2004, AJ, 128, 163
Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004, PASP, 116, 133
Mandelker N., Dekel A., Ceverino D., DeGraf C., Guo Y., Primack J., 2015, ArXiv e-prints, 1512.08791
Mao S., Mo H. J., White S. D. M., 1998, MNRAS, 297, L71
Masters K. L. et al., 2011, MNRAS, 411, 2026
Melvin T. et al., 2014, MNRAS
Momcheva I. G. et al., 2015, ArXiv e-prints, 1510.02106
Mortlock A. et al., 2013, MNRAS, 433, 1185
Nair P. B., Abraham R. G., 2010, ApJS, 186, 427
Nieto-Santisteban M. A., Szalay A. S., Gray J., 2004, in Astronomical Society of the Pacific Conference Series, Vol. 314, Astronomical Data Analysis Software and Systems (ADASS) XIII, Ochsenbein F., Allen M. G., Egret D., eds., p. 666
Peng C. Y., Ho L. C., Impey C. D., Rix H.-W., 2002, AJ, 124, 266
Pierce C. M. et al., 2010, MNRAS, 405, 718
Rix H.-W. et al., 2004, ApJS, 152, 163
Sánchez S. F. et al., 2004, ApJ, 614, 586
Sandage A., 1961, The Hubble atlas of galaxies. Carnegie Institute of Washington
Scarlata C. et al., 2007, ApJS, 172, 406
Schawinski K. et al., 2014, MNRAS, 440, 889
Schaye J. et al., 2015, MNRAS, 446, 521
Scoville N. et al., 2007, ApJS, 172, 1
Silk J., Mamon G. A., 2012, Research in Astronomy and Astrophysics, 12, 917
Simard L., Mendel J. T., Patton D. R., Ellison S. L., McConnachie A. W., 2011, ApJS, 196, 11
Simmons B. D. et al., 2013, MNRAS, 429, 2199
Simmons B. D. et al., 2014, MNRAS, 445, 3466
Simmons B. D., Urry C. M., 2008, ApJ, 683, 644
Simmons B. D., Van Duyne J., Urry C. M., Treister E., Koekemoer A. M., Grogin N. A., The GOODS Team, 2011, ApJ, 734, 121
Skibba R. A. et al., 2012, MNRAS, 423, 1485
Steinmetz M., Navarro J. F., 2002, NewA, 7, 155
Strauss M. A. et al., 2002, AJ, 124, 1810
Taniguchi Y. et al., 2007, ApJS, 172, 9
Tasca L. A. M., 2011, VizieR Online Data Catalog, 7265
Taylor M., 2011, TOPCAT: Tool for OPerations on Catalogues And Tables. Astrophysics Source Code Library
Taylor M. B., 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, Shopbell P., Britton M., Ebert R., eds., p. 29
Toomre A., Toomre J., 1972, ApJ, 178, 623
van den Bergh S., 1976, ApJ, 206, 883
Vanderplas J., Connolly A., Ivezić Ž., Gray A., 2012, in Conference on Intelligent Data Understanding (CIDU), pp. 47 –54
Vogelsberger M. et al., 2014, MNRAS, 444, 1518
Waskom M. et al., 2015, seaborn: v0.6.0 (june 2015)
Willett K. W. et al., 2013, MNRAS, 435, 2835
Willett K. W. et al., 2015, MNRAS, 449, 820
Williams R. E. et al., 1996, AJ, 112, 1335
Wright E. L., 2006, PASP, 118, 1711
York D. G. et al., 2000, AJ, 120, 1579

Table A1. Correctable fractions for the top-level task in GZH in the GOODS shallow-depth (2-epoch) images.

	GOODS-N	GOODS-S	Total
Correctable	748	514	1,262
Lower limit	526	1,143	1,669
No Correction Needed ($z \leq 0.3$)	267	267	534
NEI	851	2,670	3,521
No Redshift Information	159	319	478
Total	2,551	4,913	7,464

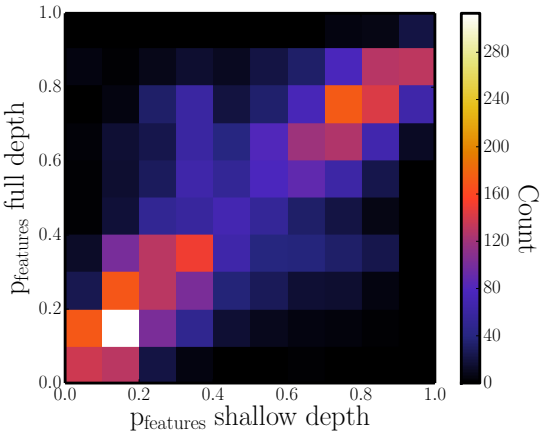


Figure A1. Distribution of $f_{\rm features}$ for the 4,461 GOODS galaxies with both shallow (2-epoch) and full-depth (5-epoch) images morphologically classified in GZH. For most galaxies, the value of $f_{\rm features}$ is consistent ($\Delta f_{\rm features} < 0.2$) between depths. Examples of galaxies with sharp changes in $f_{\rm features}$, as well as those with little to no change are shown in Figures A2-A4.

APPENDIX

A. GOODS SHALLOW DEPTH DATA

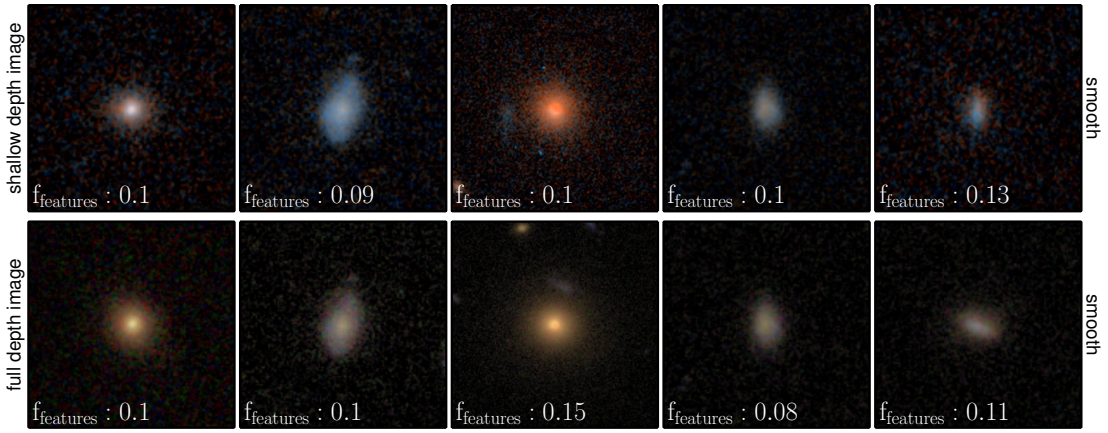
GZH used both 5-epoch and 2-epoch sets of data to construct the GOODS set of images. The 11,157 full depth 5-epoch images are used in the main catalog; the classifications for the 7,464 shallow depth 2-epoch images are provided as a supplementary table. This section analyzes the effect of image depth on the ability of the GZ classifiers to identify features or disk structure in the images.

A.1. *Comparing shallow and full depth morphologies*

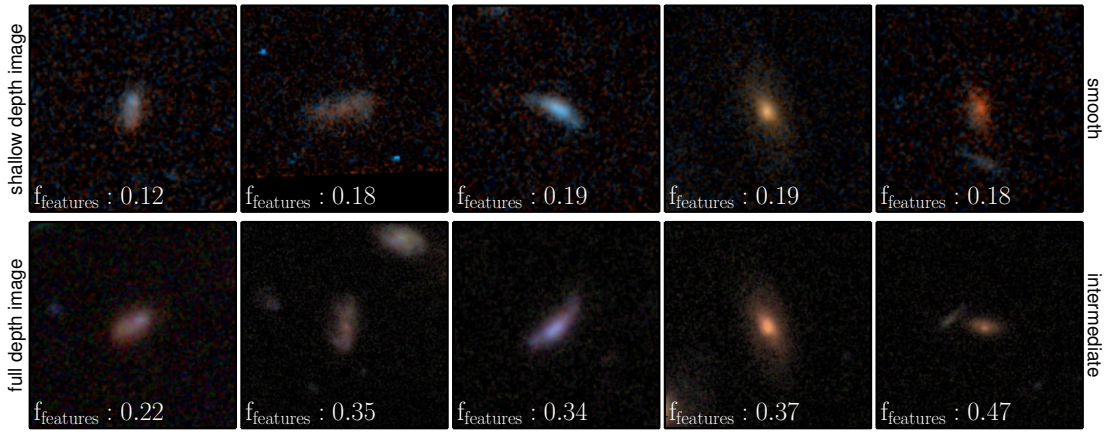
Of the 11,157 galaxies in the GOODS-N and GOODS-S full depth sample, 4,461 of these are in the shallow-depth sample. Figure A1 shows a strong correlation between $f_{\rm features}$ for both sets of images. The mean change in $f_{\rm features}$ from the shallow to full depth images $f_{\rm features,full} - f_{features,shallow} \equiv \Delta f = 0.00$, with a standard deviation of $\sigma = 0.17$. While there is some variance in Δf in the whole sample, the change is usually small and not often significant enough to change a morphological classification. Defining a clean sample of disk galaxies as those with $f_{\rm features,best} > 0.8$, elliptical galaxies as those with $f_{\rm smooth,best} < 0.2$, and intermediate as those in between, 75% of the sample would not change morphology. Of the remaining 25% that would change morphology, only 0.3% (representing 10 galaxies total) drastically change morphology from smooth to featured or vice versa, while the rest would transition to or from the "intermediate" morphology. Details can be seen in Table A2 and examples of images representing the 9 possible changes (or lack of) in morphology are shown in Figures A2,A3, and A4.

A.2. *Debiasing higher-order tasks: $f_{\rm bar}$*

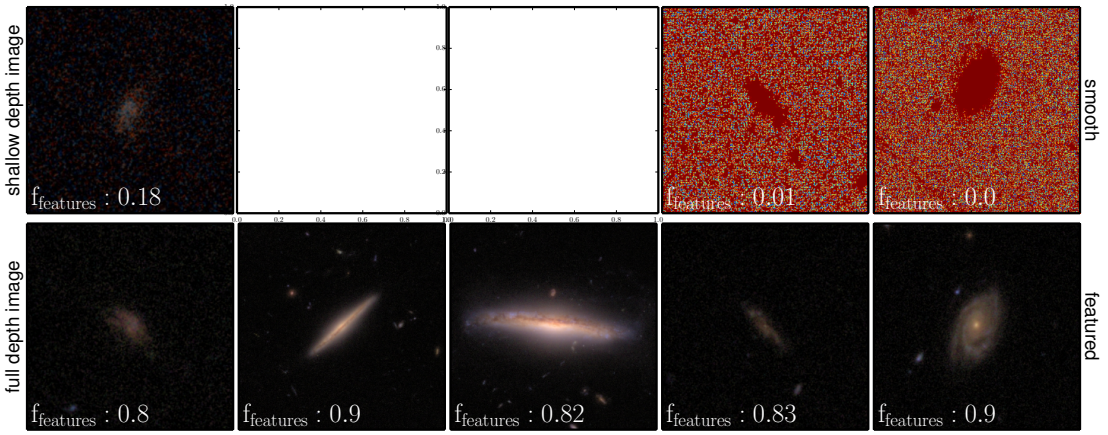
To determine the rate at which the vote fraction for any Task decreases with redshift or surface brightness, data from simulated FERENGI images were modeled with functions in discrete redshift and surface brightness bins (Section 4.3).



[a]

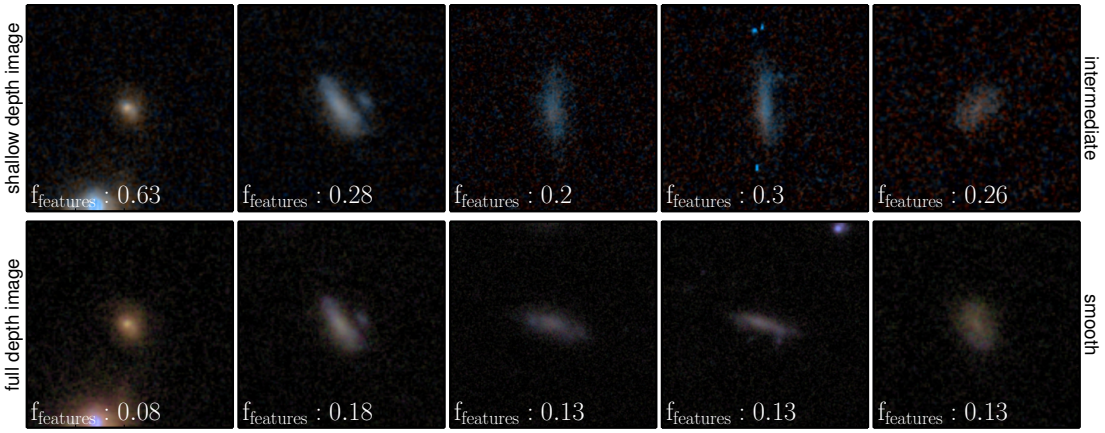


[b]

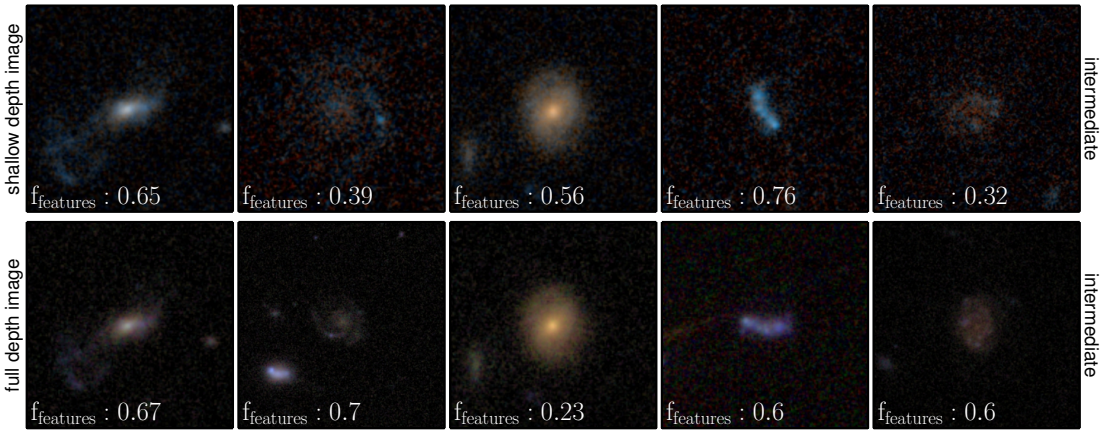


[c]

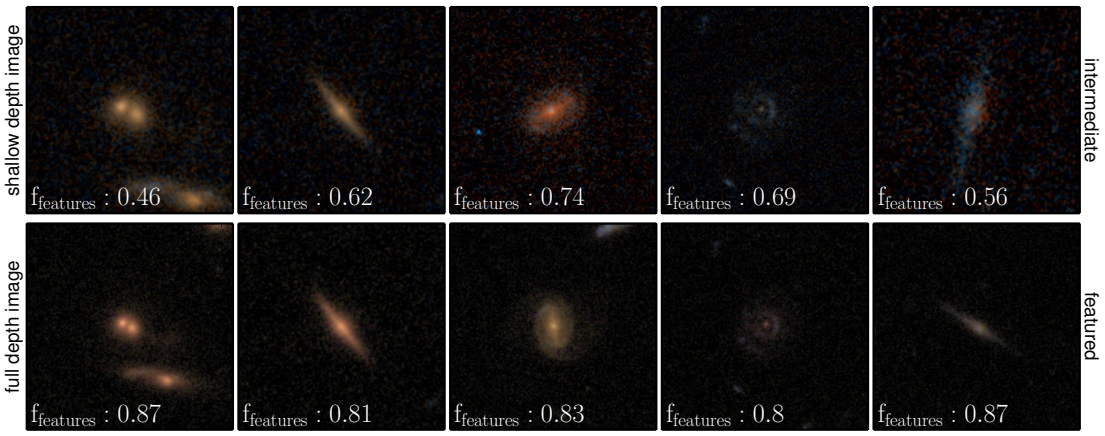
Figure A2. Galaxies whose shallow images were classified as smooth and full depth images were classified as smooth, intermediate, or featured.



[b]

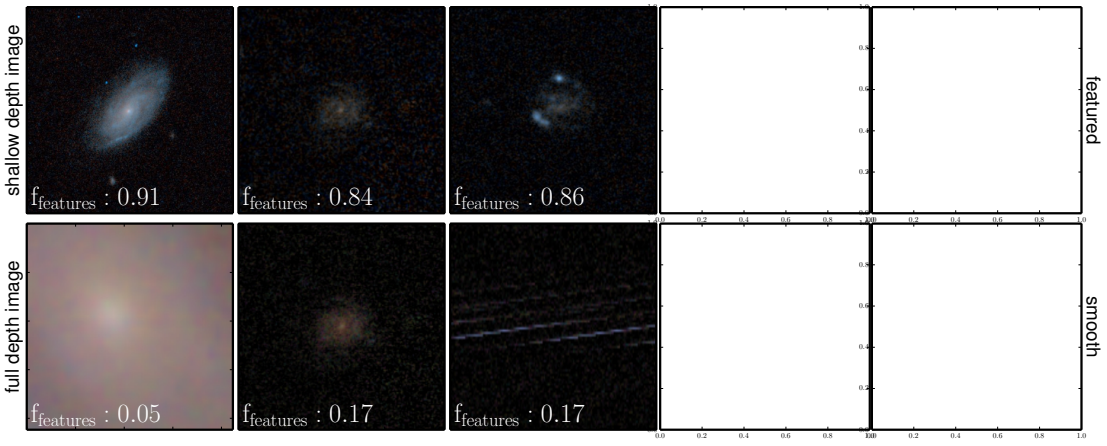


[b]

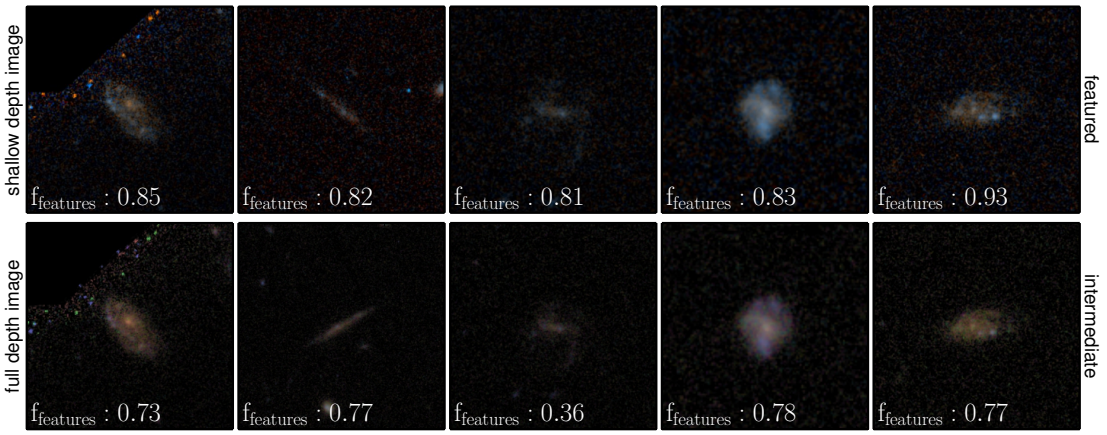


[b]

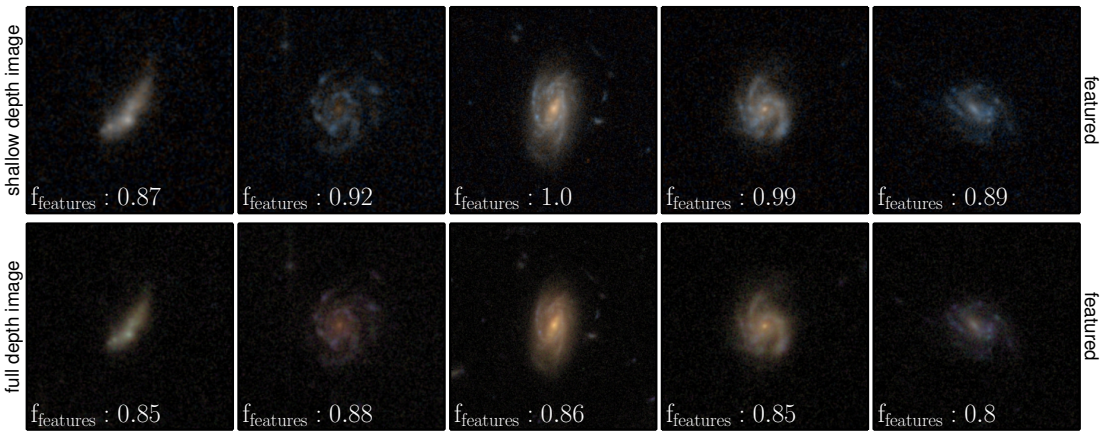
Figure A3. Galaxies whose shallow images were classified as intermediate and full depth images were classified as smooth, intermediate, or featured.



[b]



[b]



[b]

Figure A4. Galaxies whose shallow images were classified as featured and full depth images were classified as smooth, intermediate, or featured.

Table A2. Properties of galaxies whose morphologies changed or stayed the same in the shallow vs full images. Featured here is defined as $f_{\rm features,best} > 0.8$, intermediate = $0.2 < f_{\rm features,best} < 0.8$, smooth = $f_{\rm smooth,best} < 0.2$.

shallow to full morphology	N	%	$< \Delta f >$	$< z >$
smooth to smooth	758	17.0	-0.00	0.69
smooth to intermediate	367	8.2	0.18	0.69
smooth to featured	7	0.2	0.76	0.57
intermediate to smooth	214	4.8	-0.18	0.65
intermediate to intermediate	2,303	51.6	0.01	0.78
intermediate to featured	168	3.8	0.19	0.83
featured to smooth	3	0.1	-0.74	0.71
featured to intermediate	337	7.6	-0.18	0.68
featured to featured	301	6.8	-0.05	0.71
Total	4,461	100		

Table A3. Distribution of FERENGI images analyzed in Figure A5. Correctable images had a single-valued relationship between their measured $f_{\rm bar}$ values at high and low redshifts (white regions in Figure A5). Galaxies with a lower-limit on $f_{\rm bar}$ had a non single-valued relationship (blue regions). NEI images had undetermined relationships due to a lack of data ($N < 5$) in their corresponding z-μ bins (gray regions). Only 17% (maximum) of FERENGI galaxies in the sample were considered "correctable", which is not sufficient to compute a ζ function applicable to the Hubble data.

	N	%
Correctable	664	17%
Lower-limit	483	12%
NEI	2,803	71%
Total	3,950	100%

A cut of $N > 10$ was placed on the number of votes to reduce the error in the vote fractions and ensure they were well-sampled before applying a fit to the data. This requirement was only met for Task 01, which significantly reduced the goodness-of-fit for the model functions as compared with to the smooth/features task. Here we show an example of results obtained for the bar Task. Figure A5 shows that both the low sample rate and the degree of scatter in each bin inhibit fitting a reliable function that predicts $f_{\rm bar,z=0.3}$. Table A5 summarizes the correction types for the FERENGI data; 71% of the galaxies simulated did not meet the GZH requirements for well-sampled vote fractions for $f_{\rm bar}$, as compared to 2% for $f_{\rm features}$. Similar results were obtained for all higher-order Tasks in GZH (Section 4.4).

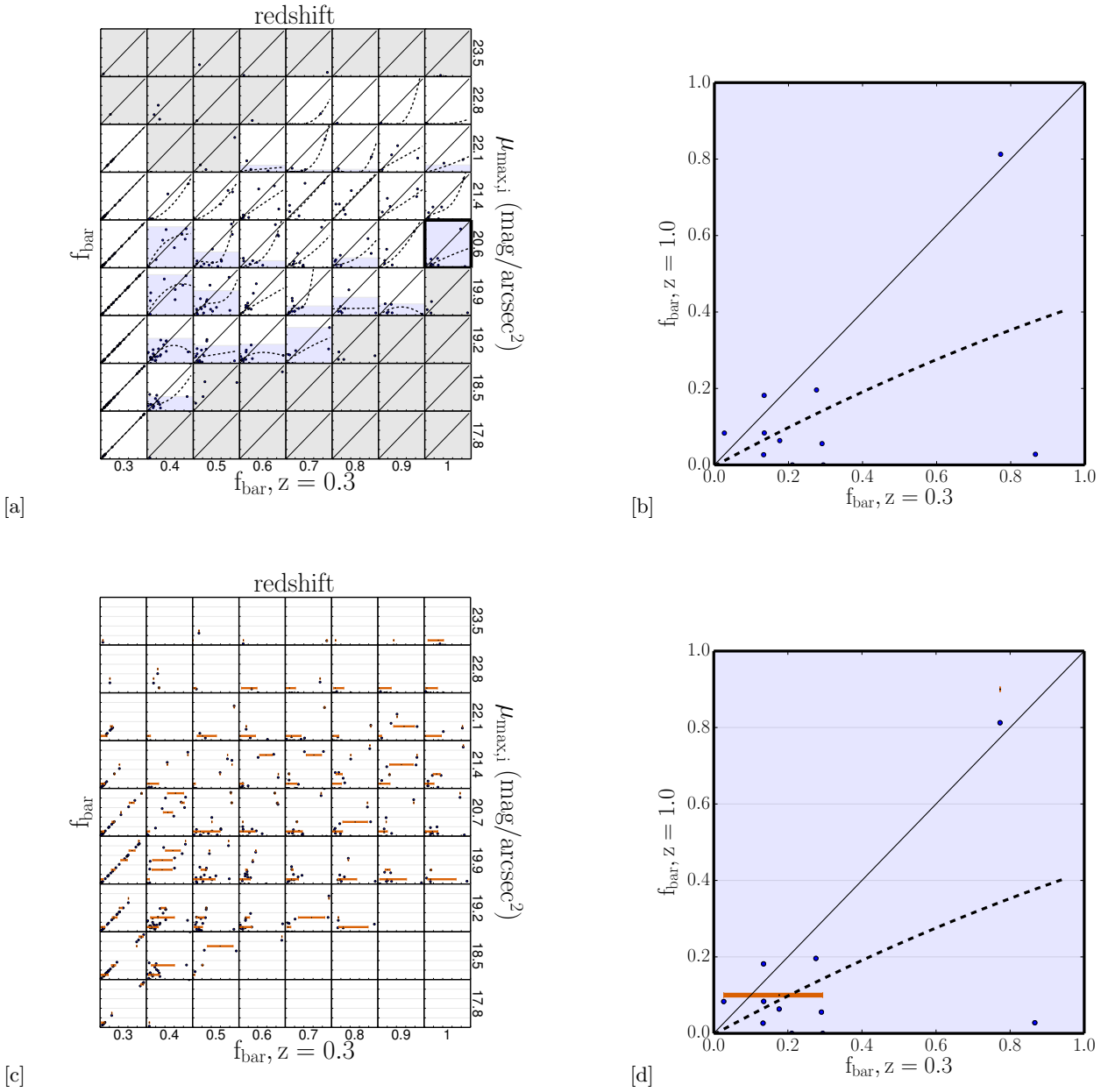


Figure A5. Similar to Figure 9, showing effects of redshift bias for 3,950 images in the FERENGI sample. [a]: Each point in a given redshift and surface brightness bin represents a unique galaxy. The y-axis in each bin is the $f_{\rm features}$ value of the image of that galaxy redshifted to the value corresponding to that redshift bin. The x-axis is the $f_{\rm features}$ value of the image of the same galaxy redshifted to $z = 0.3$. The dashed black lines represent the best-fit polynomials to the data in each square. The solid black line represents $f_{\rm features,z}$=$f_{\rm features,z=0.3}$. Regions in which there is a single-valued relationship between $f_{\rm features}$ at high redshift and at $z = 0.3$ are marked in white; those in which there is not are blue, and those with not enough data ($N < 5$) are gray. The mean normalized χ^2 of the change in vote fraction is 0.08. None of the correlations for higher-order tasks (including bars) were applied to galaxies in the GZH catalog. [b]: A larger version of the dark-outlined square in [a], containing FERENGI galaxies artificially redshifted to $z = 1.0$ and have surface brightnesses between $20.3 < \mu < 21.0$ (mag/arcsec^2). [c]: The same data as [a]. Each z, μ bin is divided into 4 sub-bins to determine the range of intrinsic $f_{\rm features,z=0.3}$ for a given range of observed $f_{\rm features,z}$ values. In each sub-bin, the orange bars represent the inner 80$^{\rm th}$ percentiles of the data, the boundaries of which are the lower and upper limits of the debiased values. [d]: The same data as [b], but highlighting the upper and lower limit regions.