11-2015

A Catalog of Visual-Like Morphologies in the 5 CANDELS Fields using Deep Learning

M. Huertas-Company
Observatoire de Paris, France

R. Gravet
Université Paris Diderot, France

G. Cabrera-Vives
University of Chile, Chile

Pablo G. Pérez-González
Universidad Complutense de Madrid, Spain

J. Kartaltepe
Rochester Institute of Technology

See next page for additional authors

A Catalog of Visual-Like Morphologies in the 5 CANDELS Fields using Deep Learning

Digital Object Identifier (DOI)
https://doi.org/10.1088/0067-0049/221/1/8

Notes/Citation Information

Published in *The Astrophysical Journal Supplement Series*, v. 221, no. 1, article 8, p. 1-23.

Authors

M. Huertas-Company, R. Gravet, G. Cabrera-Vives, Pablo G. Pérez-González, J. Kartaltepe, Guillermo Barro, M. Bernardi, S. Mei, F. Shankar, P. Dimauro, E. F. Bell, Dale D. Kocevski, David C. Koo, Sandra M. Faber, and Daniel H. McIntosh

The Astrophysical Journal Supplement Series, 221:8 (23pp), 2015 November doi:10.1088/0067-0049/221/1/8

A CATALOG OF VISUAL-LIKE MORPHOLOGIES IN THE 5 CANDELS FIELDS USING DEEP LEARNING

M. Huertas-Company[1], R. Gravet[1], G. Cabrera-Vives[2,3], P. G. Pérez-González[4], J. S. Kartaltepe[5], G. Barro[6], M. Bernardi[7], S. Mei[1], F. Shankar[8], P. Dimauro[1], E. F. Bell[9], D. Kocevski[10], D. C. Koo[6], S. M. Faber[6], and D. H. Mcintosh[11]

[1] GEPI, Observatoire de Paris, CNRS, Université Paris Diderot, 61, Avenue de l'Observatoire F-75014, Paris, France
[2] Center for Mathematical Modeling and Department of Computer Science, University of Chile, Santiago, Chile
[3] AURA Observatory in Chile, La Serena, Chile
[4] Departamento de Astrofísica, Facultad de CC. Físicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain
[5] School of Physics and Astronomy, Rochester Institute of Technology, 84 Lomb Memorial Drive, Rochester, NY 14623, USA
[6] UCO/Lick Observatory, Department of Astronomy and Astrophysics, University of California, Santa Cruz, CA 95064, USA
[7] Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA
[8] School of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK
[9] Department of Astronomy, University of Michigan, 500 Church Street, Ann Arbor, MI 48109, USA
[10] Department of Physics and Astronomy, University of Kentucky, Lexington, KY 40506, USA
[11] Department of Physics & Astronomy, University of Missouri-Kansas City, 5110 Rockhill Road, Kansas City, MO 64110, USA
Received 2015 March 17; accepted 2015 September 4; published 2015 October 26

ABSTRACT

We present a catalog of *visual-like* H-band morphologies of ∼50.000 galaxies ($H_{f160w} < 24.5$) in the 5 CANDELS fields (GOODS-N, GOODS-S, UDS, EGS, and COSMOS). Morphologies are estimated using Convolutional Neural Networks (ConvNets). The median redshift of the sample is $\langle z \rangle \sim 1.25$. The algorithm is trained on GOODS-S, for which visual classifications are publicly available, and then applied to the other 4 fields. Following the CANDELS main morphology classification scheme, our model retrieves for each galaxy the probabilities of having a spheroid or a disk, presenting an irregularity, being compact or a point source, and being unclassifiable. ConvNets are able to predict the fractions of votes given to a galaxy image with zero bias and ∼10% scatter. The fraction of mis-classifications is less than 1%. Our classification scheme represents a major improvement with respect to *Concentration-Asymmetry-Smoothness-based* methods, which hit a 20%–30% contamination limit at high z. The catalog is released with the present paper via the Rainbow database (http://rainbowx.fis.ucm.es/Rainbow_navigator_public/).

Key words: catalogs – galaxies: high-redshift – galaxies: structure – surveys

1. INTRODUCTION

Since the pioneering works in the first half of the twentieth century by E. Hubble, galaxies have been classified according to their visual aspect (see, e.g., Hubble 1926, 1936). This very first optical classification revealed that galaxies in the local universe are broadly bimodal, with or without a stellar disk (*Hubble Fork*). Understanding the physical processes that lead to such a bimodality—i.e., how bulges and disks form and evolve—is one of the major challenges in the field of galaxy evolution and the main goal of deep field surveys. The classification of galaxies at different cosmic epochs is therefore a key step toward understanding how the progenitors of today's Hubble Fork were shaped. The main difficulty is that it is hampered by the impressive amount of data which are and will be available from large galaxy surveys.

A question naturally arises: can human classifiers be replaced by automatic techniques? Different groups have conducted studies in that direction using existing visual morphologies on a smaller data set to train automated machine learning algorithms (e.g., Ball et al. 2004; Huertas-Company et al. 2008; Shamir & Wallin 2014). The basic idea behind these approaches is to find a set of parameters that correlates with the visual morphology of a galaxy and defines the parameter space that best characterize a given morphological type (e.g., Abraham et al. 1996; Conselice et al. 2000; Lotz et al. 2008). In astronomy, the parameters defining morphology traditionally include concentrations, asymmetries, clumpiness (or smoothness), gini coefficient, moments of light, etc.

In recent years, we proposed a generalization of this approach with the development of galSVM (Huertas-Company et al. 2008, 2009, 2011), which enables an n-dimensional classification with optimal nonlinear boundaries in the parameter space as well as a quantification of the errors following a probabilistic approach (see also Scarlata et al. 2007; Peth et al. 2015). These *Concentration-Asymmetry-Smoothness* (*CAS*)*-based* methods have been proven to be relatively useful, but are also affected by several limitations. The values of the parameters strongly depend on the data quality and redshift, and they only provide rough morphological classifications in two or three classes. The most evident shortcoming of such techniques is that the fraction of mis-classifications is high, especially at high redshifts (∼20%–30%, Huertas-Company et al. 2014). The latter could be the main reason why their popularity among the astronomical community is still quite low (see the review by Ball & Brunner 2010).

The problem might reside in the parameters which people traditionally adopt. Concentrations, asymmetries, etc., and by extension principal components, are useful because they reduce the complexity of the problem by globally describing a galaxy with just a few parameters. However, at the same time, this approach neglects an enormous amount of information contained in the pixels themselves. Consequently, CAS-based methods might not be suitable to actually represent the ability

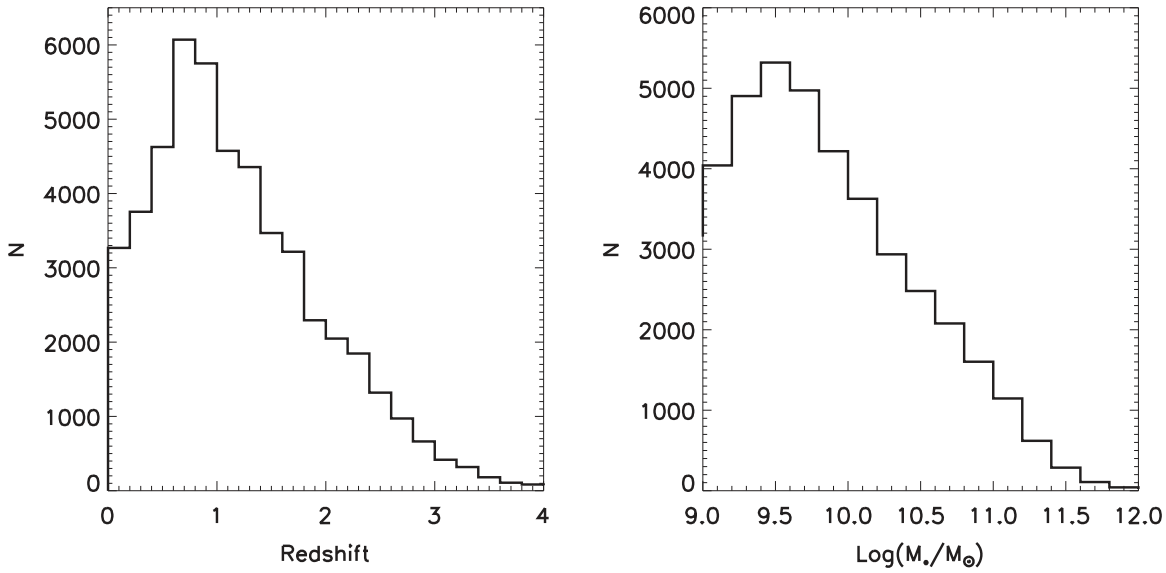

Figure 1. Redshfit (left) and stellar mass (right) distributions of the selected sample for morphological classifications. The data set contains more than 20,000 galaxies at $z > 1$ where the CANDELS fields probe the optical rest-frame morphologies.

of the human brain to capture the full, complex distribution of light. Using all of the pixels as the parameter space is now possible with the advent of powerful computing resources such as Graphic Processor Units (GPUs). At the same time, very powerful machine learning algorithms exist which are suited to mimicing human perception (such as *deep learning*) and which are able to learn the best set of parameters for a given problem. This new approach was first used in astronomy at low redshift earlier this year, in the framework of an online competition led by the Galaxy Zoo team (see Section 3 for more details), yielding very promising results (Dieleman et al. 2015, hereafter D15).

In this paper, we extend this new methodology to high redshift by classifying ∼50,000 galaxies with median redshift $\langle z \rangle \sim 1.25$ in the CANDELS fields where detailed visual classifications are available for a subsample of ∼8000 objects (Kartaltepe et al. 2014). We show that the use of deep learning yields a classification that is almost free-of-contamination and closely mimics human perception. We release the resulting catalog of the 5 CANDELS fields (GOODS-S, GOODS-N, UDS, EGS, and COSMOS) with the present work.

The paper is structured as follows. In Section 2, we describe the data set. In Section 3, we describe the method and how the CANDELS data are pre-processed before feeding the algorithm. In Sections 4 and 5, we discuss the performance and accuracy of the resulting classification, and in Section 6 we describe the properties of the catalog which is released. We conclude with a summary of the main results (Section 7).

2. DATA SET

We use the CANDELS public photometric catalogs for UDS (Galametz et al. 2013) and GOODS-S (Guo et al. 2013) as our starting point. Preliminary CANDELS catalogs were used for COSMOS, EGS, and GOODS-N (CANDELS team 2015 private communication). We select all those galaxies in the F160W filters with F160W < 24.5 mag (AB system), which is the magnitude limit imposed by Kartaltepe et al. (2014) to perform reliable visual morphological classifications. Since our goal is to provide a morphological classification as close as possible to the visual classification, we restrict our selection to the same criteria in all of the considered fields.

The resulting sample consists of 50,000 galaxies, which increases by a factor of 5 the visual catalog published in CANDELS to date. Approximately 50% of the sources are in the range $1 < z < 3$ (Figure 1) where the CANDELS filters probe optical rest-frame morphologies. As was extensively discussed in Kartaltepe et al. (2014), the sample is ∼80% complete down to $\log(M_*/M_\odot) \sim 10$ (see their Figure 1).

3. CANDELS MORPHOLOGICAL CLASSIFICATION WITH DEEP LEARNING

3.1. Convolutional Neural Network (ConvNet) Configuration

In this work, we mimic human perception with *deep learning* using convolutional neural networks (ConvNets). Although it is clearly beyond the scope of the present paper to provide a complete description of how convolutional neural networks work, we provide a brief introduction below. We refer the interested reader to D15 for more details.

Deep learning is a methodology to automatically learn and extract the most relevant features (or parameters) from raw data for a given classification problem through a set of nonlinear transformations.

Though deep learning architectures have existed since the early 80s (Fukushima 1980), they involve complex technological problems which only allowed their use in massive data sets in the last decade. Several factors have contributed to the rise in their popularity: (i) the availability of much larger training sets with millions of labeled examples[12]; (ii) powerful GPU implementations, making the training of very large models practical; and (iii) improved model regularization algorithms, which helped to reduce computing time.

[12] ConvNets are particularly sensitive to this since the risk of over-fitting is large given the complexity of the models.

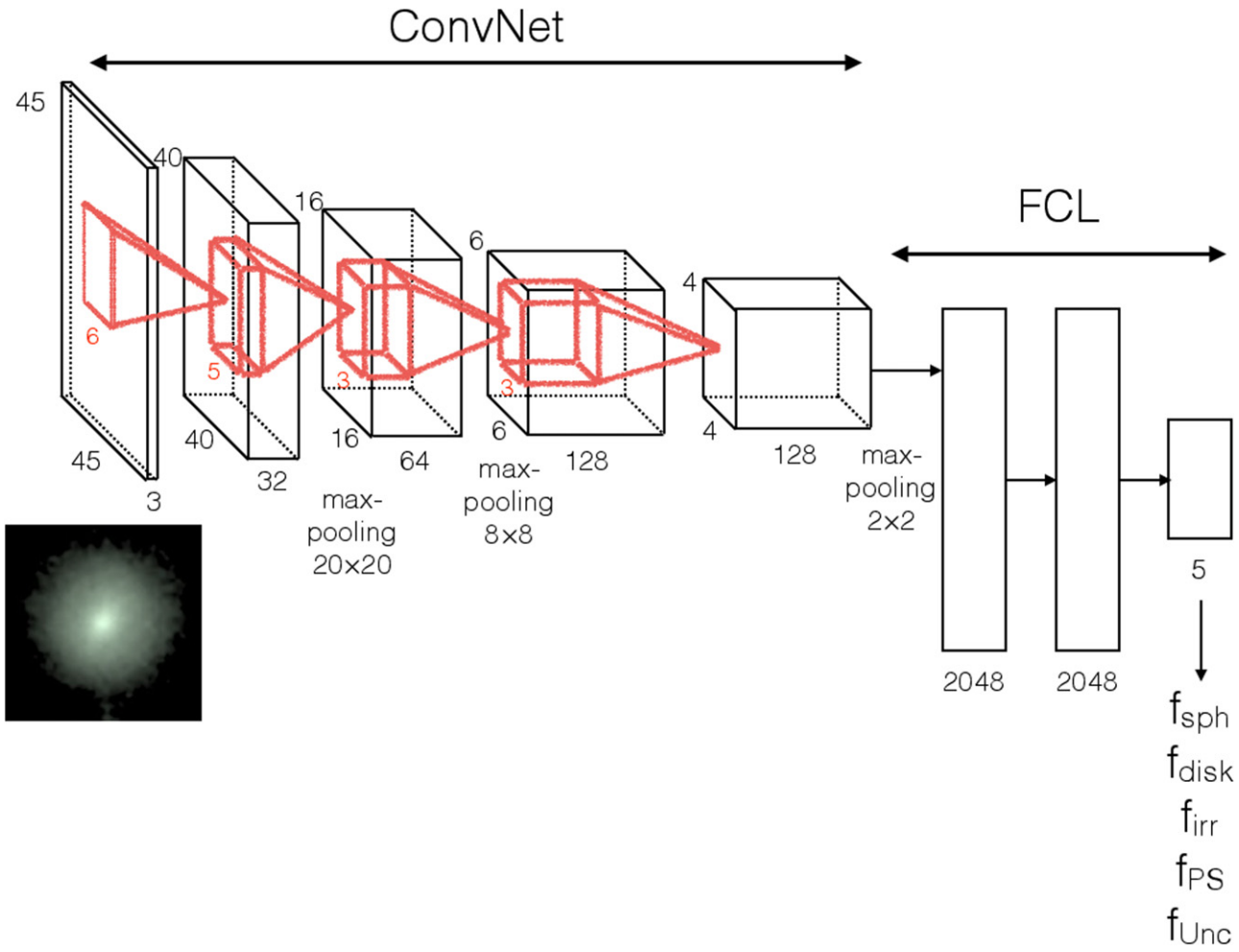

Figure 2. Configuration of the Convolutional Neural Network used in this paper. The Network is based on the one used by Dieleman et al. (2015) on SDSS galaxies. It is made of 5 convolutional layers followed by 2 fully connected perceptron layers. In the convolutional part there are also 3 max-pooling steps of different sizes. The input are *SDDSized* CANDELS galaxies as explained in the text and the output (for this paper) is made of 5 real values corresponding to the fractions defined in the CANDELS classification scheme.

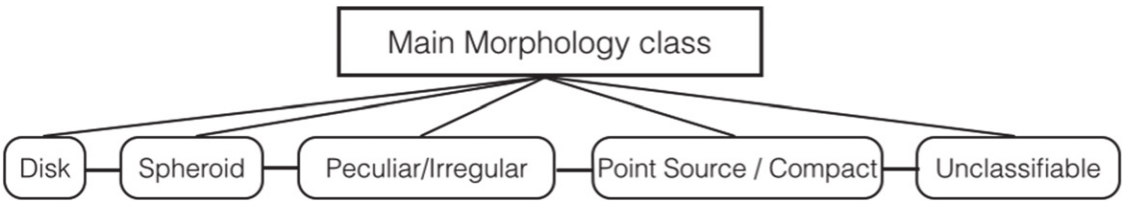

Figure 3. CANDELS *Main Morphology* visual classification scheme as described in Kartaltepe et al. (2014). Each classifier (3–5 per galaxy on average) is asked to provide 5 flags for each galaxy corresponding to the main morphological properties of the galaxy as labeled in the figure. The flags are then combined to produce the fractions of people that voted for a given feature.

ConvNets have been proven to perform extremely well in image recognition tasks. For example, they have achieved an error rate of 0.23% for the MNIST database, which is a collection of manuscript numbers considered as a standard test for all new machine learning algorithms (Ciresan et al. 2012). When applied to facial recognition, they achieve a 97.6% recognition rate on 5600 images of more than 10 subjects (Matusugu et al. 2003). The ImageNet Large Scale Visual Recognition Challenge is a benchmark in object classification and detection, with millions of images and hundreds of object classes. In Krizhevsky et al. (2012), ConvNets were able to achieve an error rate of 15.3% compared to the rate of 26.2% achieved by the second best competitors (non-deep). Also, the performance of convolutional neural networks on the ImageNet tests is now close to a purely human-based classification (Russakovsky et al. 2014).

ConvNets were first applied to galaxy morphological classification earlier this year in the framework of the Galaxy Zoo Challenge on the Kaggle platform.[13] The goal of the challenge was to find an algorithm able to predict the 37 votes of the Galaxy Zoo 2 release. The winner of the competition

[13] https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

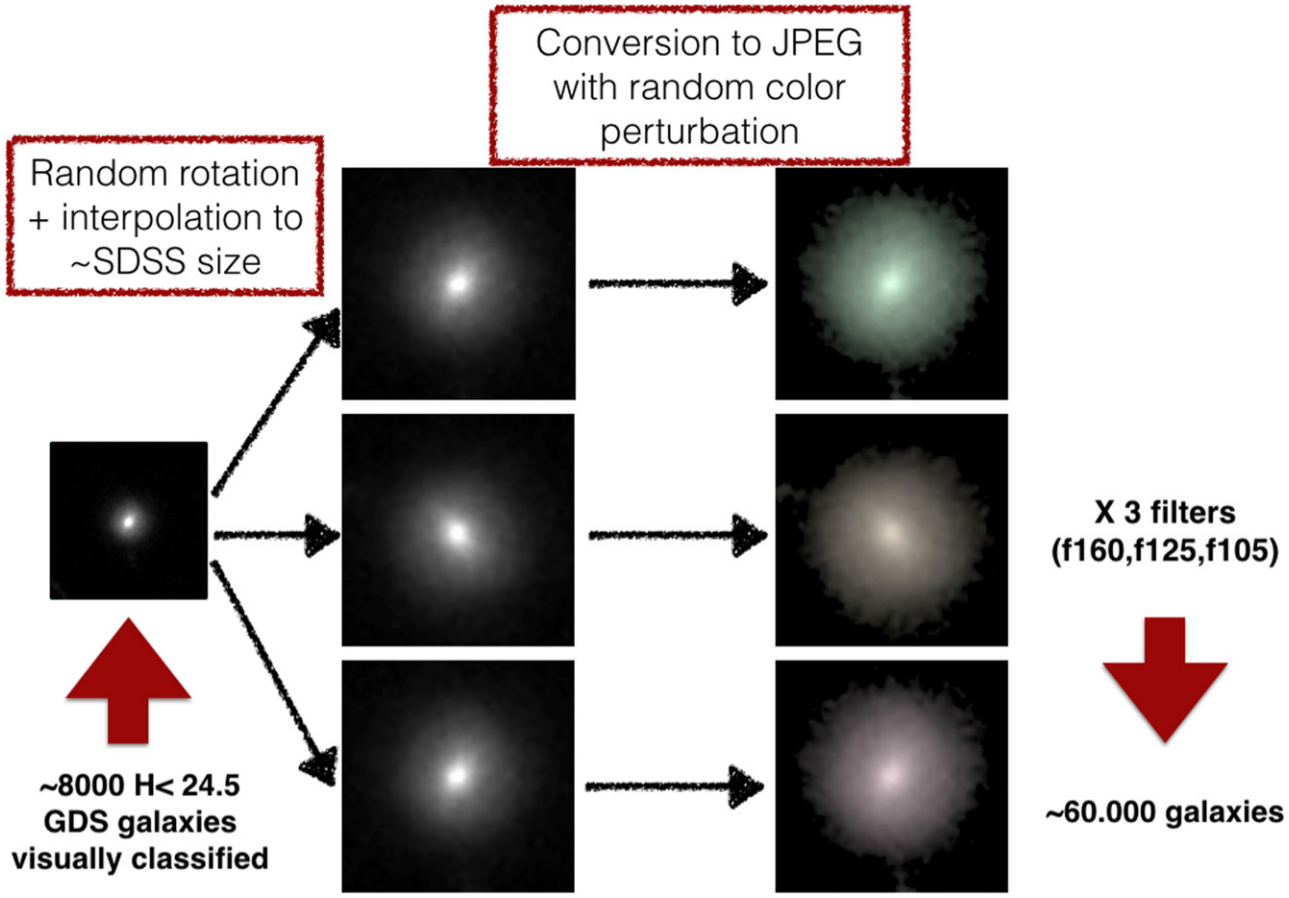

Figure 4. Pre-processing of the CANDELS stamps before being fed into the convolutional neural network. Galaxies are first interpolated so that they all have similar sizes. In a second step, we add some redundancy to the data by performing random rotations in order to avoid over-fitting, and finally converted the images to JPEG. This is repeated for three CANDELS filters. See text for details.

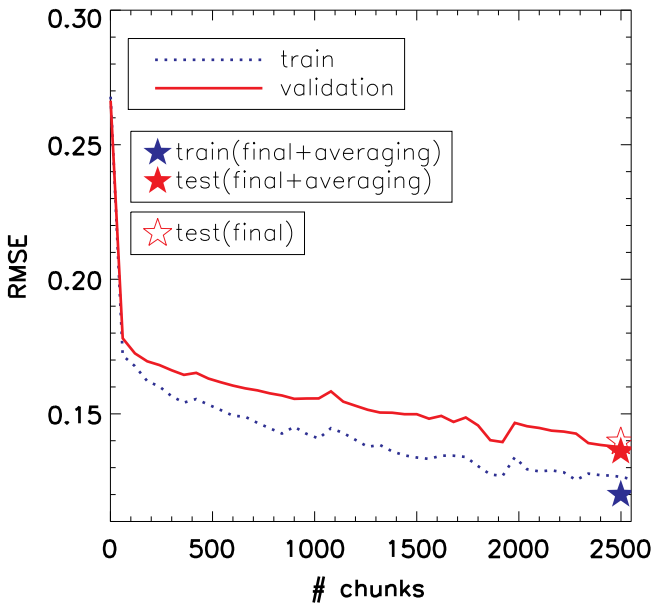

Figure 5. Time trajectories for the training (dotted blue line) and validation (red solide line) sets (see text for details). The rmse is computed every 60 chunks. The blue/red stars indicate the values computed with the final model (2500 chunks) on the training and test samples, respectively, after averaging and reported in Table 2. The empty star shows the rmse on the test sample before averaging.

used ConvNets to obtain a final rms of ∼7% on the parameters (Dieleman et al. 2015). This work clearly showed that ConvNets are a very promising tool for automated morphological classifications.

There is no clear methodology for finding the optimal convolutional neural network for a given problem, except for trying different configurations and comparing the outputs. The methodology used for the Galaxy Zoo challenge provided excellent results for a problem similar to ours (Figure 2). We therefore decided to use the D15 configuration to classify the CANDELS sample. Given the different nature of the SDSS and CANDELS images, our methodology, by design, requires specific pre-processing steps, as discussed in Section 3.3. This is certainly not the cleanest approach, but it is sufficient for our classification purposes as discussed in the subsequent sections.

3.2. Training Set

The ConvNet is trained to reproduce the CANDELS visual morphological classification defined in Kartaltepe et al. (2014). This classification is based on the efforts of 65 individual classifiers who contributed to the visual inspection of all of the galaxies in the GOODS-S field (the average number of classifiers per galaxy being 3–5). The classifiers were asked to provide a number of flags related to the galaxy structure, morphological k-correction, interaction status, and

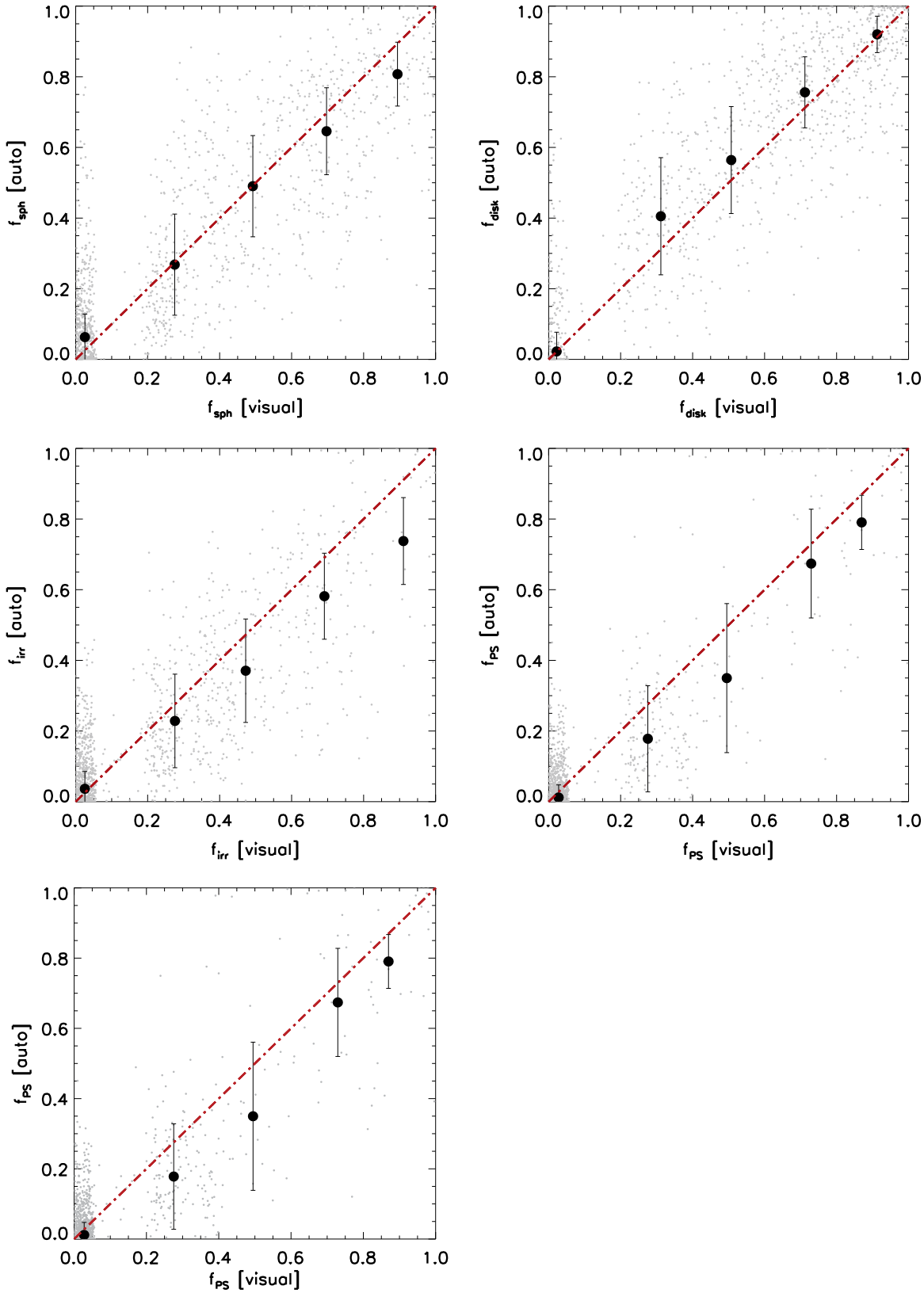

Figure 6. Correlation between the fractions of classifiers voting for a given feature (spheroid (top left), disk (top right), irregular (middle left), point source (middle right), and unclassifiable (bottom left)) and the predictions of the ConvNet based classification on a test data set. Detailed quantifications of the bias and the dispersion are shown in Table 1.

clumpiness. As a result, each galaxy in the catalog has a number of flags which measure the fraction of classifiers who selected a morphological feature. Classification was mainly performed in the H band (F160W), even though each classifier had access to images of the same galaxy in other wavelengths.

Table 1
Median Bias ($\Delta_f = (f_{\rm auto} - f_{\rm visu})$), Root Mean Square Error (rmse), and Scatter as a Function of the Visual Morphological Frequencies for the Test (Top) and the Training (Bottom) Sets

Test Sample					
	$0 < f_{\rm sph} < 0.2$	$0.2 < f_{\rm sph} < 0.4$	$0.4 < f_{\rm sph} < 0.6$	$0.6 < f_{\rm sph} < 0.8$	$0.8 < f_{\rm sph} < 1.0$
Bias	0.03	−0.01	0.00	−0.05	−0.10
rmse	0.09	0.15	0.15	0.17	0.16
Scatter	0.07	0.14	0.14	0.12	0.09
	$0 < f_{\rm disk} < 0.2$	$0.2 < f_{\rm disk} < 0.4$	$0.4 < f_{\rm disk} < 0.6$	$0.6 < f_{\rm disk} < 0.8$	$0.8 < f_{\rm disk} < 1.0$
Bias	−0.00	0.11	0.06	0.06	−0.00
rmse	0.09	0.17	0.16	0.13	0.09
Scatter	0.05	0.17	0.15	0.10	0.05
	$0 < f_{\rm irr} < 0.2$	$0.2 < f_{\rm irr} < 0.4$	$0.4 < f_{\rm irr} < 0.6$	$0.6 < f_{\rm irr} < 0.8$	$0.8 < f_{\rm irr} < 1.0$
Bias	0.01	−0.06	−0.10	−0.12	−0.14
rmse	0.06	0.13	0.16	0.20	0.23
Scatter	0.05	0.13	0.15	0.12	0.12
	$0 < f_{\rm PS} < 0.2$	$0.2 < f_{\rm PS} < 0.4$	$0.4 < f_{\rm PS} < 0.6$	$0.6 < f_{\rm PS} < 0.8$	$0.8 < f_{\rm PS} < 1.0$
Bias	−0.01	−0.11	−0.10	−0.04	−0.09
rmse	0.04	0.14	0.21	0.19	0.16
Scatter	0.04	0.15	0.21	0.15	0.08
	$0 < f_{\rm Unc} < 0.2$	$0.2 < f_{\rm Unc} < 0.4$	$0.4 < f_{\rm Unc} < 0.6$	$0.6 < f_{\rm Unc} < 0.8$	$0.8 < f_{\rm Unc} < 1.0$
Bias	−0.02	−0.17	−0.07	0.19	−0.03
rmse	0.03	0.16	0.12	0.23	0.09
Scatter	0.03	0.21	0.07	0.22	0.02
Training Sample					
	$0 < f_{\rm sph} < 0.2$	$0.2 < f_{\rm sph} < 0.4$	$0.4 < f_{\rm sph} < 0.6$	$0.6 < f_{\rm sph} < 0.8$	$0.8 < f_{\rm sph} < 1.0$
Bias	0.03	−0.02	−0.02	−0.01	−0.07
rmse	0.08	0.13	0.15	0.13	0.12
Scatter	0.06	0.13	0.13	0.10	0.07
	$0 < f_{\rm disk} < 0.2$	$0.2 < f_{\rm disk} < 0.4$	$0.4 < f_{\rm disk} < 0.6$	$0.6 < f_{\rm disk} < 0.8$	$0.8 < f_{\rm disk} < 1.0$
Bias	0.01	0.07	0.08	0.05	−0.00
rmse	0.09	0.15	0.14	0.12	0.08
Scatter	0.06	0.13	0.12	0.09	0.05
	$0 < f_{\rm irr} < 0.2$	$0.2 < f_{\rm irr} < 0.4$	$0.4 < f_{\rm irr} < 0.6$	$0.6 < f_{\rm irr} < 0.8$	$0.8 < f_{\rm irr} < 1.0$
Bias	0.00	−0.06	−0.08	−0.08	−0.11
rmse	0.05	0.12	0.15	0.16	0.18
Scatter	0.05	0.12	0.13	0.12	0.10
	$0 < f_{\rm PS} < 0.2$	$0.2 < f_{\rm PS} < 0.4$	$0.4 < f_{\rm PS} < 0.6$	$0.6 < f_{\rm PS} < 0.8$	$0.8 < f_{\rm PS} < 1.0$
Bias	−0.01	−0.11	−0.16	−0.07	0.01
rmse	0.04	0.13	0.18	0.19	0.13
Scatter	0.03	0.15	0.18	0.14	0.08
	$0 < f_{\rm Unc} < 0.2$	$0.2 < f_{\rm Unc} < 0.4$	$0.4 < f_{\rm Unc} < 0.6$	$0.6 < f_{\rm Unc} < 0.8$	$0.8 < f_{\rm Unc} < 1.0$
Bias	−0.02	−0.10	−0.11	−0.01	0.03
rmse	0.03	0.14	0.19	0.22	0.22
Scatter	0.03	0.14	0.17	0.15	0.09

In this work, we focus on the main *classification tree*, which defines the main morphological class (Figure 3). For each galaxy there are therefore five parameters, $f_{\rm spheroid}$, $f_{\rm disk}$, $f_{\rm irr}$, $f_{\rm PS}$, and $f_{\rm Unc}$ which refer, respectively, to the frequency at which human classifiers flagged a given galaxy as *having a spheroid*, *a disk*, *some irregularities*, being a point source (or unresolved), and *unclassifiable*. It is important to note that one flag does not exclude the other (except for the Unc one), i.e., a galaxy can obviously have both a disk and a spheroid, or have a disk and be irregular, and so the sum of all of the frequencies for a given object is not one.

The main purpose of this work is to mimic human behavior. In other words, we want the machine to be able to predict how many people will vote for a given feature given the galaxy image. Recall that the objective we consider here is to replace humans by computers, not to find the *correct* morphology of a galaxy, which actually depends on the definition one adopts. Hence, if the visual classification is intrinsically biased, then the machine-based one also will be.

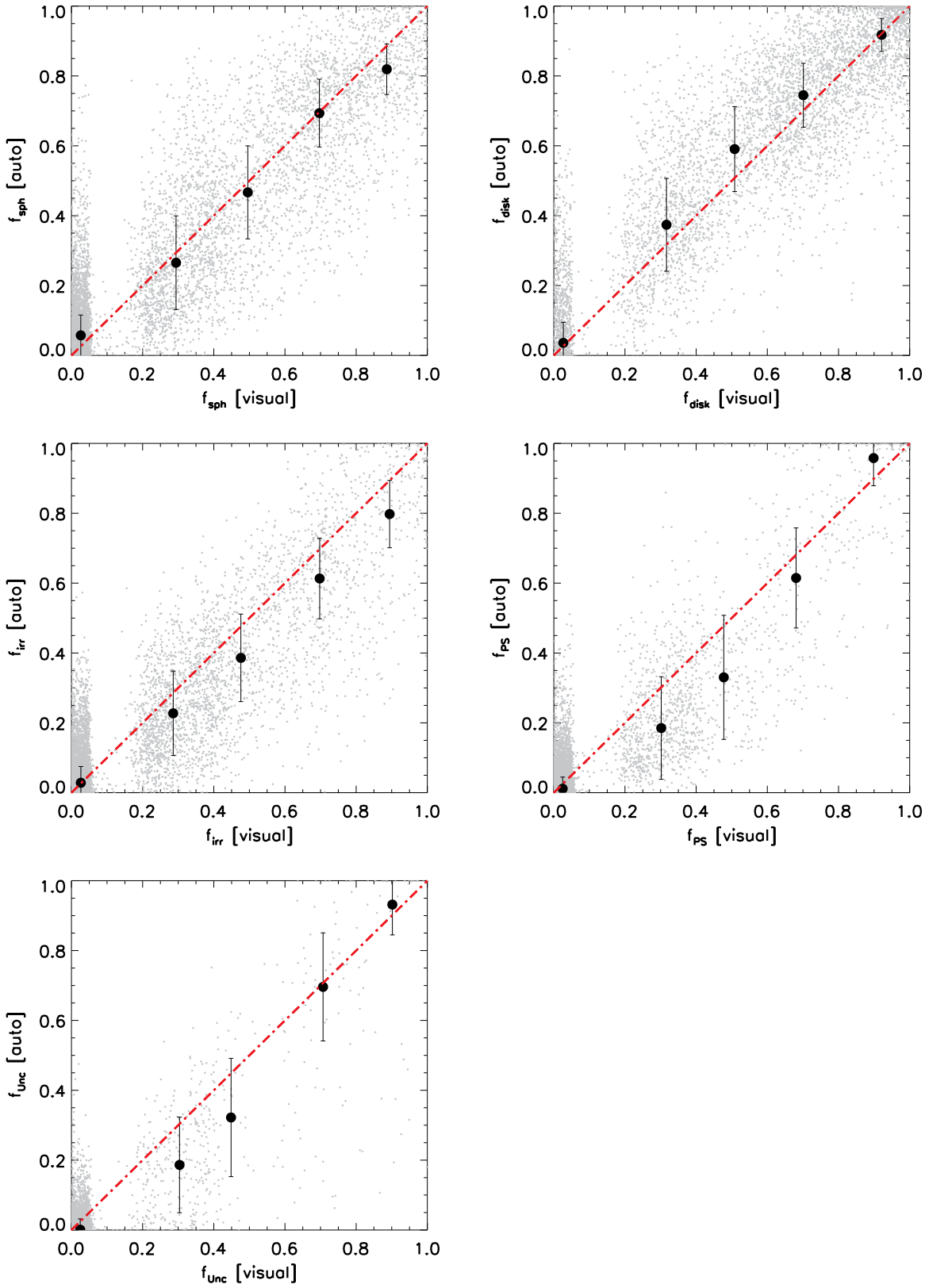

Figure 7. Same as Figure 6 but for objects used for the training.

The classification in GOODS-S contains ∼8000 galaxies for which we know the visual classification performed by (expert) humans, and so we can use part of this sample to train the machine learning algorithm and keep a fraction for an independent test. Also note that during the preparation of the present work, the UDS field was also finalized, and hence it also represents an independent test for the classification as discussed in Section 5. In the following, we describe the pre-processing done to images before being fed into ConvNet.

Table 2
Median Bias ($\Delta_f = (f_{auto} - f_{visu})$) and Scatter for Each Visual Morphological Frequency for the Test and Training Samples

Test Sample			
Parameter	Bias	Scatter	rmse
$f_{spheroid}$	0.03	0.09	0.17
f_{disk}	0.03	0.08	0.15
f_{irr}	−0.01	0.07	0.14
f_{PS}	−0.01	0.04	0.10
f_{Unc}	−0.02	0.03	0.07
ALL	0.00	0.05	0.13
Training Sample			
Parameter	Bias	Scatter	rmse
$f_{spheroid}$	0.02	0.08	0.15
f_{disk}	0.02	0.08	0.14
f_{irr}	−0.01	0.06	0.12
f_{PS}	−0.01	0.04	0.09
f_{Unc}	−0.02	0.03	0.05
ALL	−0.01	0.05	0.12

3.3. Pre-processing

As previously discussed, for this work, we will use the ConvNet design shown in D15 optimized for the SDSS. There are some obvious problems related to this approach, since galaxies at high redshift are intrinsically smaller[14] and fainter. Also, the training set is made of only ∼8000 galaxies from GOODS-S with visual parameters, compared to the 60×10^3 galaxies used for the SDSS training. This last point is particularly critical since training ConvNet with a significantly smaller sample can easily lead to over-fitting issues, i.e., too many parameters in the model we want to build compared with the number of data points.

To overcome the latter potential issues, we pre-processed the training set before feeding it to ConvNet by applying the following steps (see Figure 4).

1. All of the galaxies in the GOODS-S visual morphology catalog are interpolated to the typical SDSS size (i.e., ∼40 pixels). This is performed using a classical cubic interpolation. The procedure obviously introduces some redundancy in the data since we artificially reduce the pixel size, but ensures that the network *sees* the same ratio of background versus galaxy pixels as for the SDSS. This is important because the size of the convolution box is fixed. An alternative approach would have been to adapt the network size to the typical size of CANDELS images. In any case, some interpolation is required given the wide redshift range probed by the CANDELS data ($z \sim 0.1$ to $z \sim 3$), which means that the length scale changes by more than a factor of 4. Therefore, even if the interpolation factor could be decreased, it is required at some level. In this work, since we are interested in broad morphologies, the impact of interpolation is not a major issue, and therefore we decided to keep the original network.
2. Each galaxy is randomly rotated three times before being fed into the net. Since our data set is significantly smaller than the one used in the GZOO competition, there is a clear risk of over-fitting in the classification process. We therefore introduce additional redundancy in the training set to increase the number of training points, taking advantage of the fact that morphological classifications should be rotationally invariant (Dieleman et al. 2015). As explained in D15, the algorithm itself will introduce additional redundancy by performing two more 90° rotations.
3. We then introduce some random Gaussian noise to each of the rotated images so that the pixel values of each realization are not exactly the same. The added noise is small enough so as not to affect the visual aspect of the galaxy, but it slightly changes the pixel values. This ensures that the redundancy is actually efficient and that the network considers each rotated galaxy as a different object with very similar morphological parameters, just as the human eye does. Finally, each of the rotated images is converted to JPEG with a power-law stretching optimized for astronomy[15] (Bertin 2012) and a 10% compression. This is important to keep the number of possible pixel values reasonable and also to obtain a similar normalization for all of the galaxies. We again stress that since here we are interested in broad morphologies (disk versus bulge, irregular, compact), the impact of compression is not critical, as shown in subsequent sections. For more detailed morphologies (e.g., LSB features, bars, etc.), especially at high redshift, a careful investigation of optimal compression will certainly be required.
4. The previous steps were repeated in three CANDELS filters (f105, f125, and f160) to reach a final training set of ∼58,000 galaxies (8000 × 3(rotations) × 3(filters)), very close to the 60,000 SDSS objects for which the net was designed. Note that the spatial coverage of all of the filters is not exactly the same, which explains why we only reach ∼60,000 galaxies. The size of the data set is enough to avoid over-fitting and reach satisfactory results, as shown in the next sections. The use of the same galaxies in three different filters might introduce some biases since the morphology might look slightly different from one filter to another. However, Kartaltepe

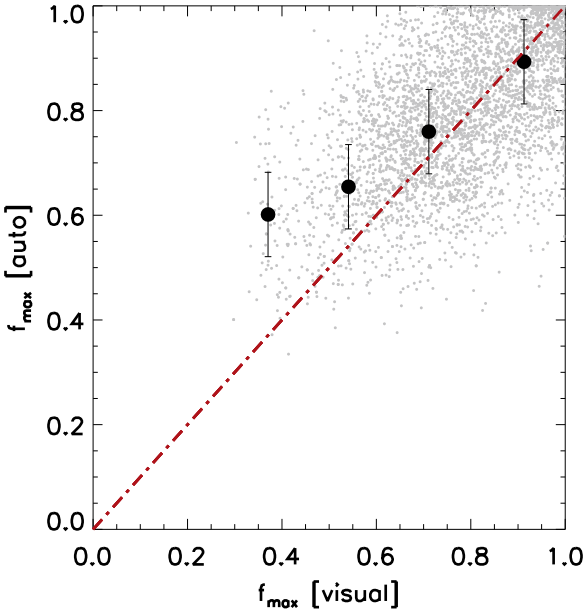

Figure 8. Relation between the maximum fraction in the visual and the automatic classifications.

[14] Typically 5–10 pixels—∼0.″3–compared to 40 pixels—∼10″—for the SDSS galaxies.

[15] http://www.astromatic.net/software/stiff

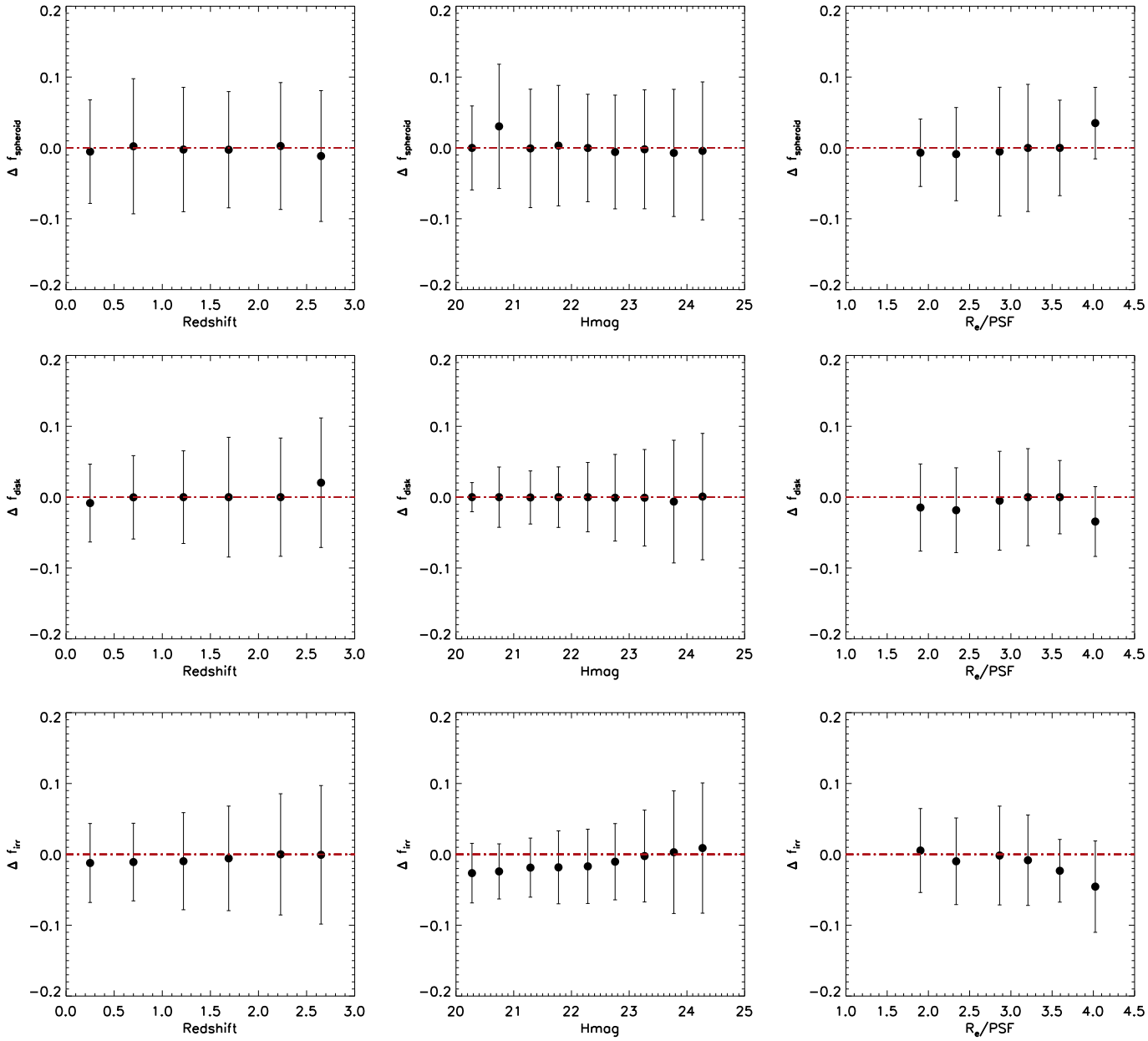

Figure 9. Mean Bias ($\Delta f = f_{\rm auto} - f_{\rm visu}$) and scatter ($\sqrt{\rm VAR(\Delta_f)}$) of the three main morphological fractions (spheroid, disk, and irregular from top to bottom) as a function of redshfit, magnitude, and resolution (from left to right).

et al. (2014) show that the fraction of galaxies that actually change their morphology between these three filters is very small. In any case, we also tried the algorithm using only f160 images (reducing the training set by a factor of three), leading to no significant changes in the final results ($\sim$0.01 change in the final root mean square error (rmse) value).

5. We finally introduce some *noise* in the visual parameters of each galaxy ($f_{\rm spheroid}$, $f_{\rm disk}$, $f_{\rm irr}$, $f_{\rm PS}$, and $f_{\rm Unc}$) by adding a random Gaussian 10% scatter. This is done, first, to make sure that ConvNet does not see exactly the same data points for different redundant images and force optimization. Second, because the CANDELS fractions are very discretized since the actual number of classifiers per galaxy is rather small and therefore the full range of values from 0 to 1 is not covered. The 10% value is calibrated empirically and is of the order of the magnitude of the intrinsic noise of the labels (assuming that they follow a binomial distribution—see Section 5). Below this value, the effect is almost negligible; meanwhile, above this value, the original signal is diluted. As we will show in Section 5, this also has some important consequences on the final output.

The final data set used for classification thus contains $\sim$58,000 redundant JPEG images, of which 47,700 are used for training the machine (i.e., finding the best model), 5300 are used for real-time evaluation during model training (validation data set), and 5000 galaxies are used to assess the final accuracy with the best final model (test data set). These 5000 galaxies constitute the test sample and are not used at all during

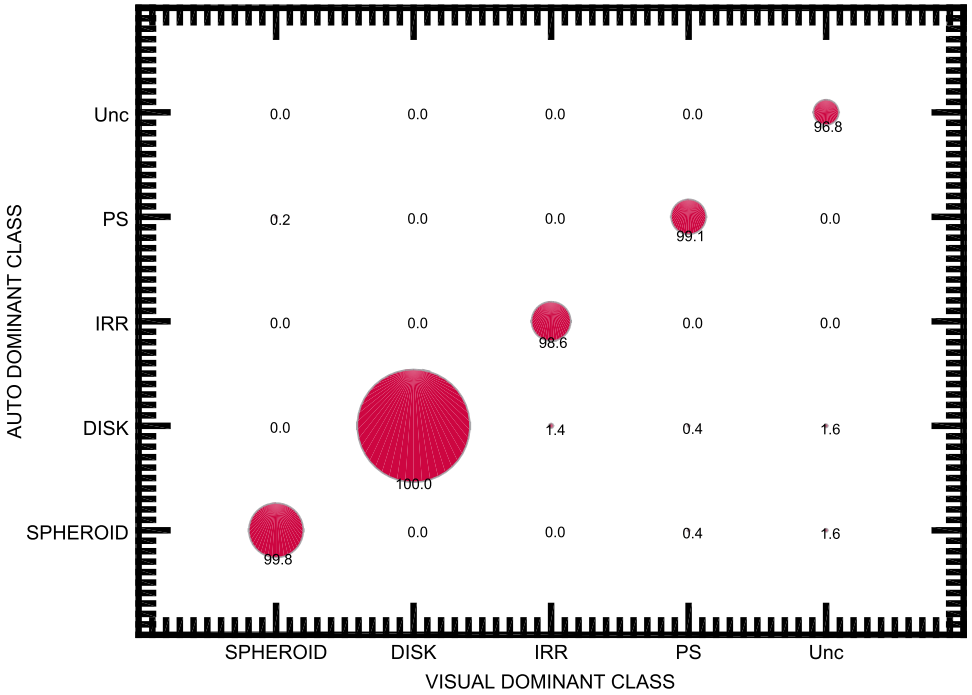

Figure 10. Relation between the visual and automatic dominant morphological classes for well-defined objects. The sizes of the symbols are proportional to the number of objects. The level of agreement is >95%.

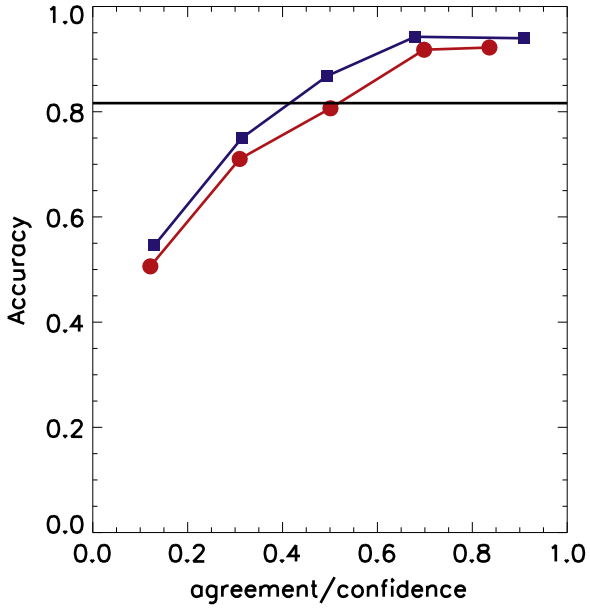

Figure 11. Classification accuracy as a function of the level of agreement between classifiers (*a*). The red line shows the relation when *a* is computed using visual classification. The blue line indicates the same relation but a computed form the automated classification. The horizontal line indicates the average accuracy.

the training process (but their visual morphology is known), and so they can be used independently to study the behavior of the best trained model on an unknown data set. The final model is taken at 2500 chunks. As described in Dieleman et al. (2015), to further improve the classification accuracy, averaging of 17 variants of the best model is applied as post-processing. These variants include modifications such as the removal of dense layers, different filter size configurations, and a different number of filters, among others. We refer to Dieleman et al. (2015) for more details. The best model followed by the averaging process is then used to classify the other four CANDELS fields for which visual morphology is not yet available. The classification is done at a rate of ∼1000 galaxies/hour on a TESLA M2090 GPU, which is compatible with the treatment of massive data sets expected in the near future (e.g., EUCLID, WFIRST).

The evolution of the rmse during the final learning process for the training and validation data sets is shown in Figure 5. The difference in rmse for the validation data set in the last 10 iterations is of the order of 10^{-4}, confirming that the algorithm has converged. There is no significant over-fitting given the convergence of the validation set's rmse. As expected, the rmse for the training set is slightly smaller (∼0.01), as this is the data directly used to fit our ConvNet model (recall that the validation data set is used for real-time evaluation of the model on unseen data). Also, in Figure 5, we show the values of the rmse for the test sample before and after averaging. As explained above, this third data set is needed to assess the final rmse of the model, as it may happen that the 2500 chunks we use for convergence are over-fitted to the validation data set. The rmse over the test set is very consistent with that obtained for the validation data set. Averaging slightly reduces the rmse by $\sim 10^{-3}$, which is consistent with the values reported in Dieleman et al. (2015).

We made sure that the different pre-processing steps described above always result in a decrease of the average rmse on the validation and test samples. More precisely, before any pre-processing, the average rmse is ∼0.25. Adding noise to the labels decreases the error to ∼0.22. Interpolation makes it reach ∼0.17, and finally redundancy, together with noise addition, brings it to a final value of ∼0.13 (Figure 5).

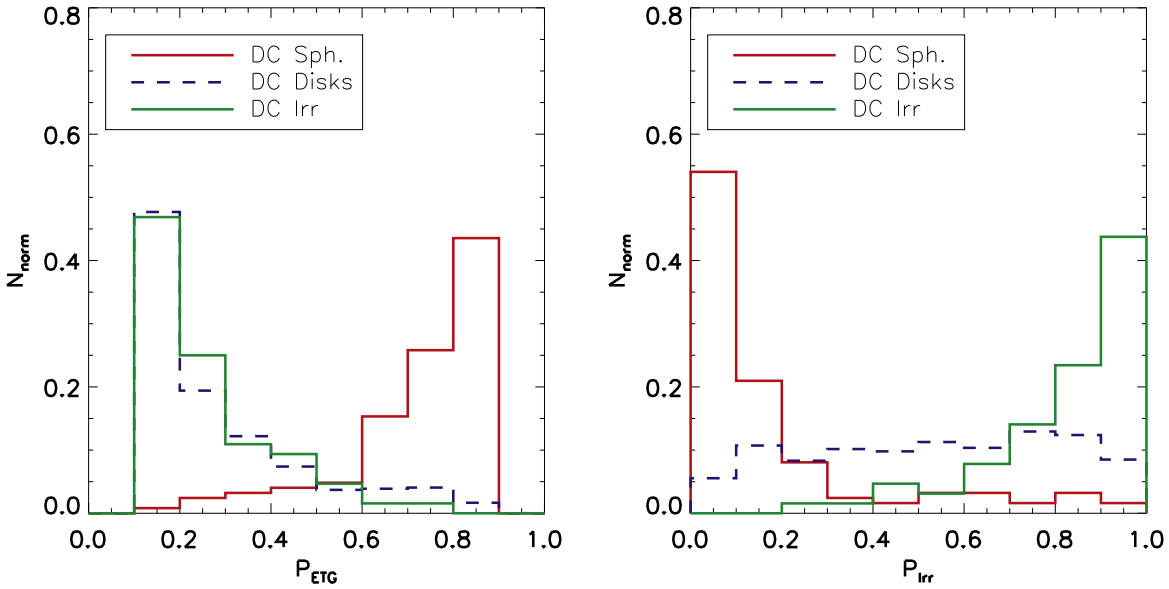

Figure 12. Probability distributions of being early-type (left panel) and irregular (right panel) estimated by galSVM (see Huertas-Company et al. 2014) for three dominant classes in the CANDELS visual classification as labelled. Dominant disks cannot reliably be separated from dominant irregulars using this approach.

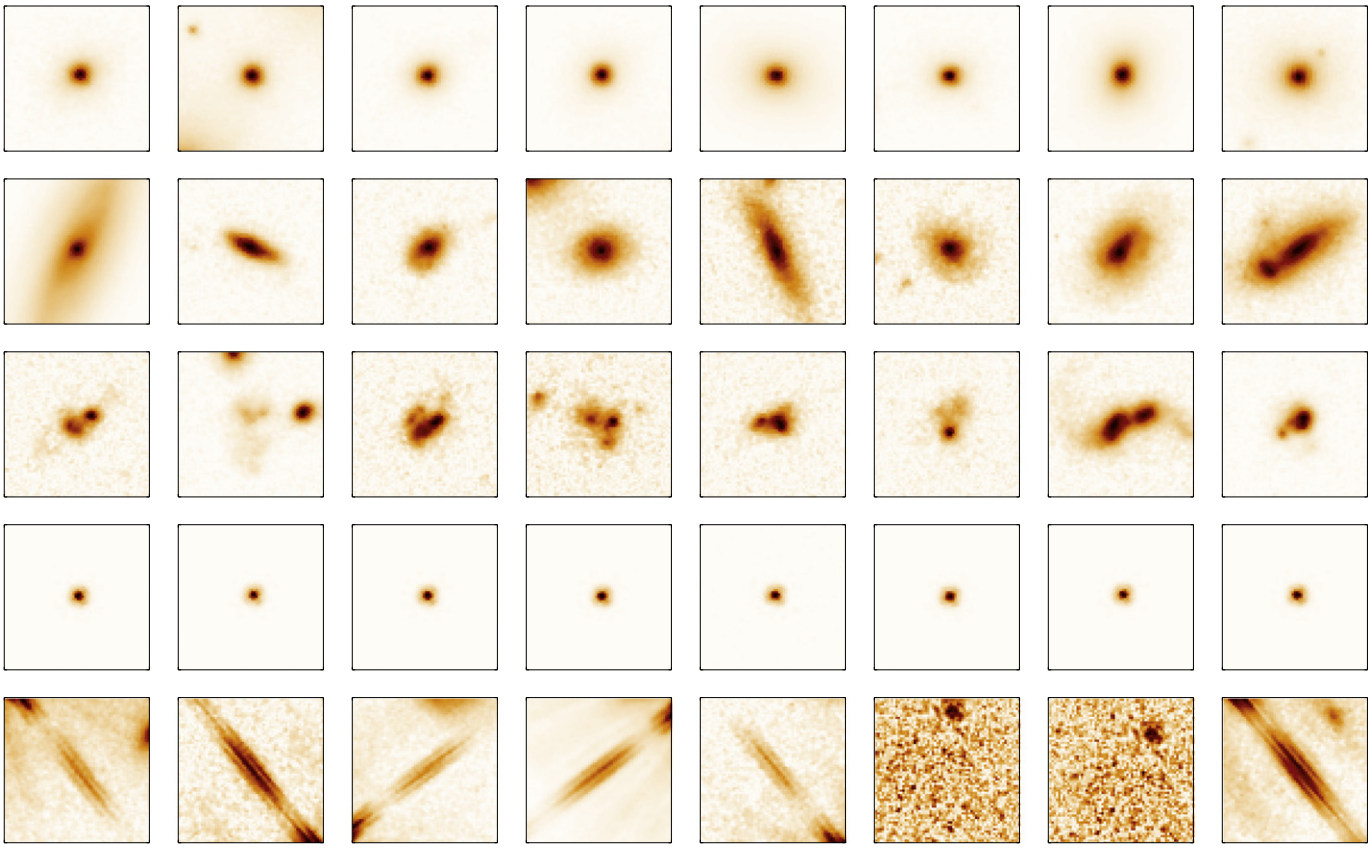

Figure 13. Examples stamps of the five *dominant* morphological classes in the COSMOS/CANDELS field. From top to bottom we show dominant spheroids, dominant disks, dominant irregulars, dominant point/sources-compact, and dominant unclassifiable. The selection of these stamps is performed fully randomly. Recall that COSMOS galaxies have not been used for training the algorithm, and therefore they are completely new for the best model. The size of the stamps is $3\farcs8 \times 3\farcs8$.

4. ACCURACY

4.1. Recovering Votes

Figure 6 shows the relation between the visual fractions for each galaxy provided in Kartaltepe et al. (2014) once the random shifts have been applied, and the predicted values for the main classification tree ($f_{\rm spheroid}$, $f_{\rm disk}$, $f_{\rm irr}$, $f_{\rm PS}$, and $f_{\rm Unc}$). In Figure 6, we only plot those objects in the test sample (5000 objects) which were not used for training in order to assess the behavior of the machine with an unknown data set. Results in

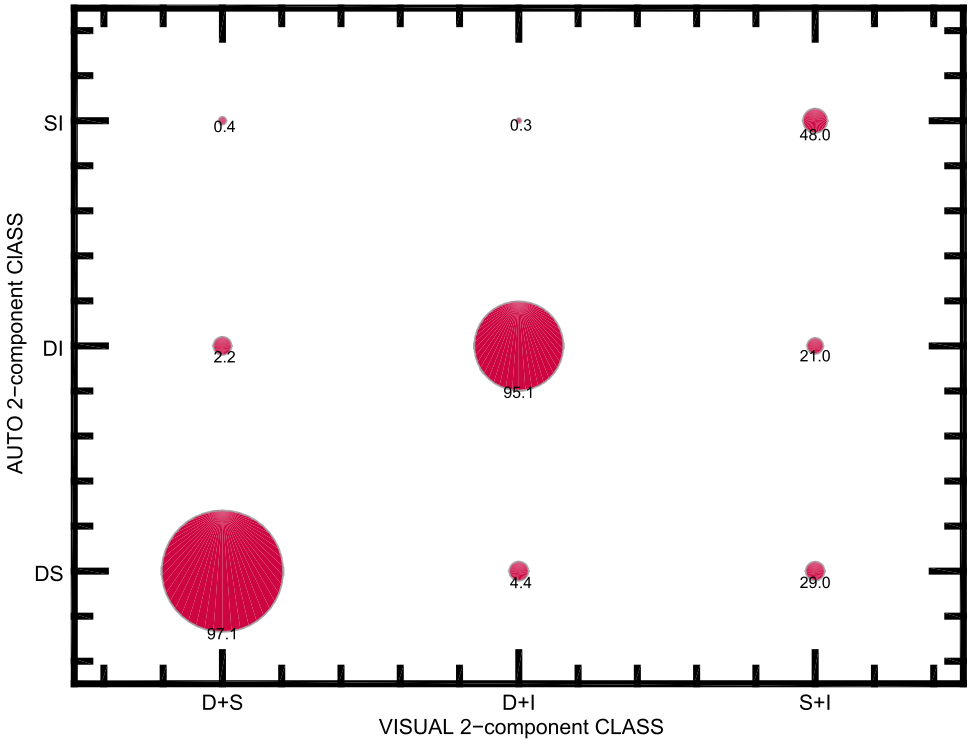

Figure 14. Relation between the visual and automatic two-component classes. The level of agreement is >95%.

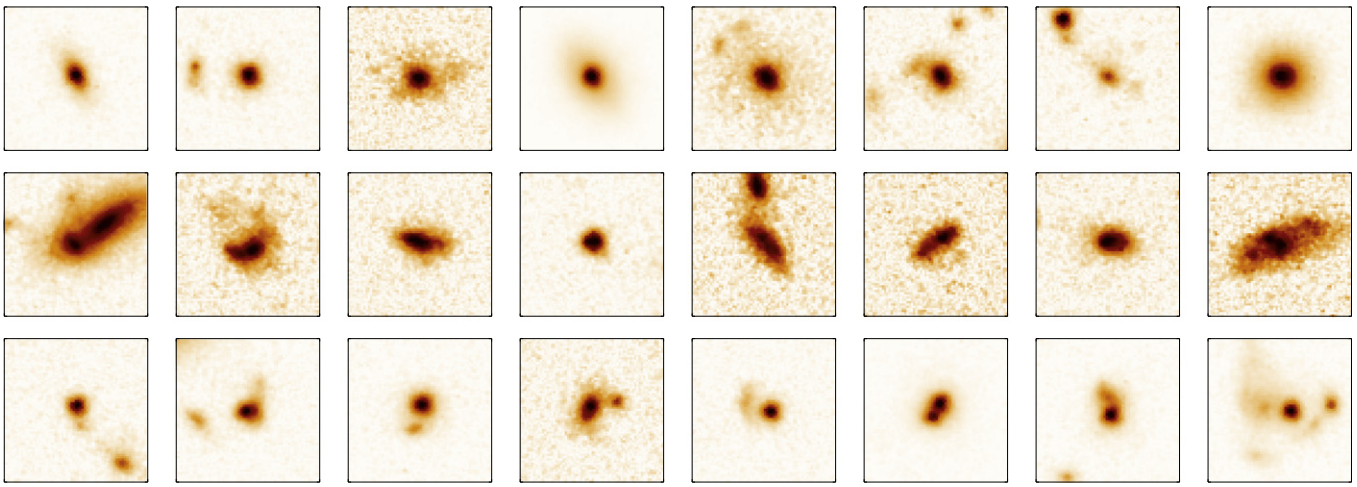

Figure 15. Example stamps of objects with two main morphological classes in the COSMOS/CANDELS field. From top to bottom, we show spheroids+disks, disks +irregular, and spheroids+irregular. The selection of these stamps is done fully randomly. Recall that COSMOS galaxies have not been used for training the algorithm, and therefore they are completely new for the best model. The size of the stamps is $3\farcs8 \times 3\farcs8$.

terms of bias and scatter are also tabulated in Table 1. There is a clear one-to-one correlation between the automatically derived quantities and the visual ones. Table 1 shows that the typical bias and dispersion are lower than 10%. It is important to keep in mind that the distribution of frequencies is not homogenous between 0 and 1 (there are bins in which there are very few objects) and the machine is therefore optimized to minimize the global bias. In fact, the median bias and scatter for all of the morphological frequencies are even smaller and range between 0–0.02 and 0.03–0.1, respectively, as shown in Table 2. If we instead plot galaxies in the training set, then the scatter is almost the same, as expected from the learning histories shown in Figure 5. This confirms that the model is well-optimized and that there is no over-fitting (Figure 7).

Despite the scatter, it is important to note that the tails in the distribution seen in Figure 6 do not necessarily imply misclassifications as we currently define them, i.e., galaxies that clearly fall in the wrong morphological class after visual inspection. As a matter of fact, a galaxy that might have a slightly larger bulge probability in the automated scheme than in the purely visual classification will, however, clearly be classified as a disk since its probability is much higher. Figure 8 shows the relation between the maximum visual frequency, defined as the maximum frequency irrespective of the morphology for each galaxy, and the maximum automatic

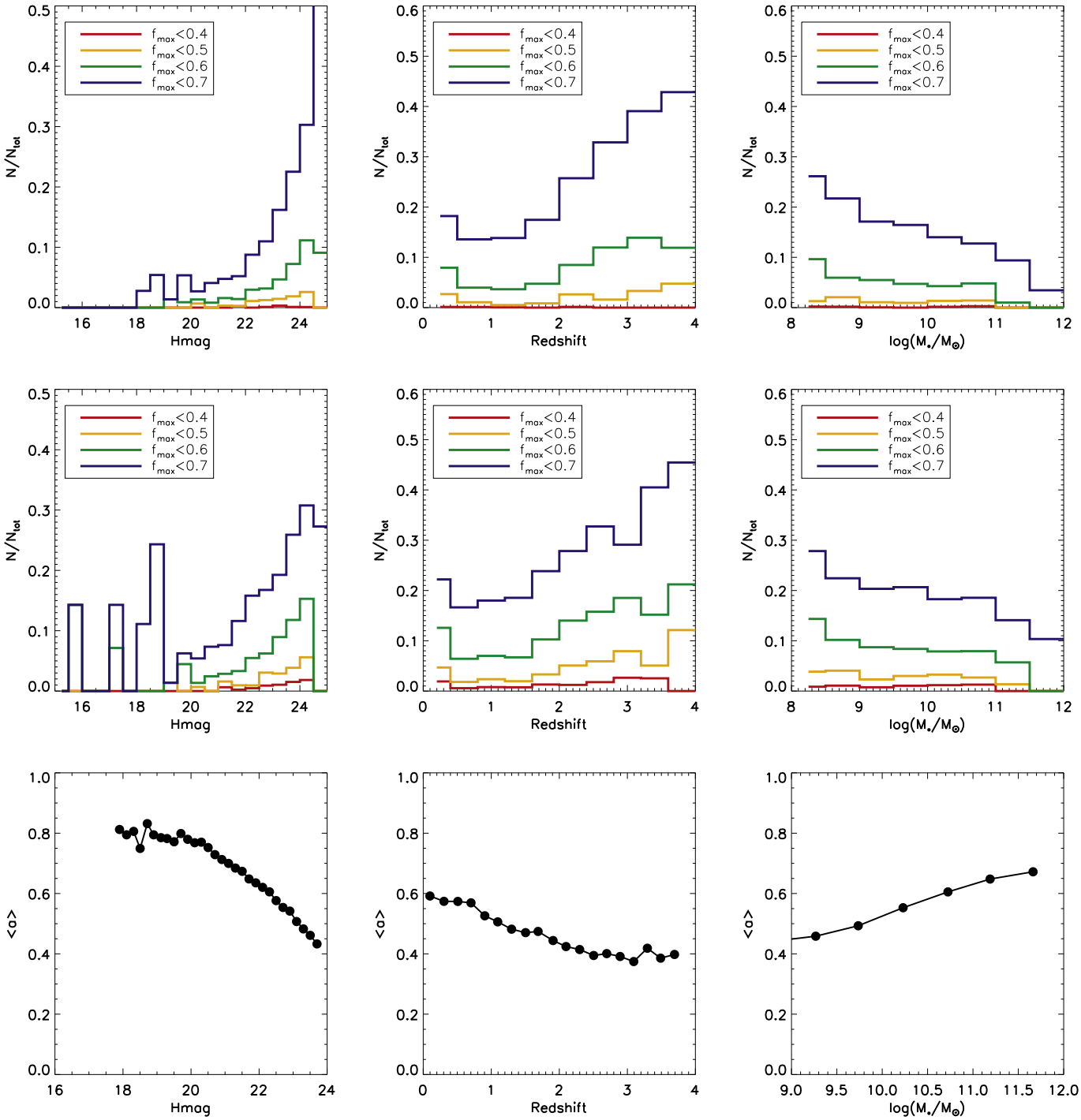

Figure 16. Fraction of uncertain objects defined for different $f_{\rm max}$ thresholds as labelled in the automatic (top) and visual (middle) classifications. The fraction of uncertain objects increase for fainter objects, high redshifts, and low masses. Similar trends are recovered in both classifications. The bottom line shows the relation between the level of agreement a (see text) and magnitude (left), redshift (middle), and stellar mass (right).

frequency. Both quantities are correlated with the expected scatter with no tails, even though there seems to be an increasing bias at low frequencies ($f_{\rm max} < 0.5$). This is not surprising since those are the most unclear objects of the visual catalog.

Also, in Figure 9, we explore how the performance of the classification depends on physical properties such as redshift, magnitude, and size relative to the point-spread function (PSF) FWHM. Interestingly, we do not observe any particular trend on the bias or scatter with magnitude and redshift. The bias in the morphological fractions stays at <0.05, and the scatter is rather constant at 0.1 for all magnitudes and redshifts spanned by our sample. Only very small objects, close to the size of the PSF, or very large (>4 times the PSF size) have a larger bias ($\sim$0.05–0.1). For large objects, this could be explained by the fact that part of the wings might be lost during the interpolation process at fixed size. Recall that this does not necessarily mean that the morphology can be assessed equally independently of

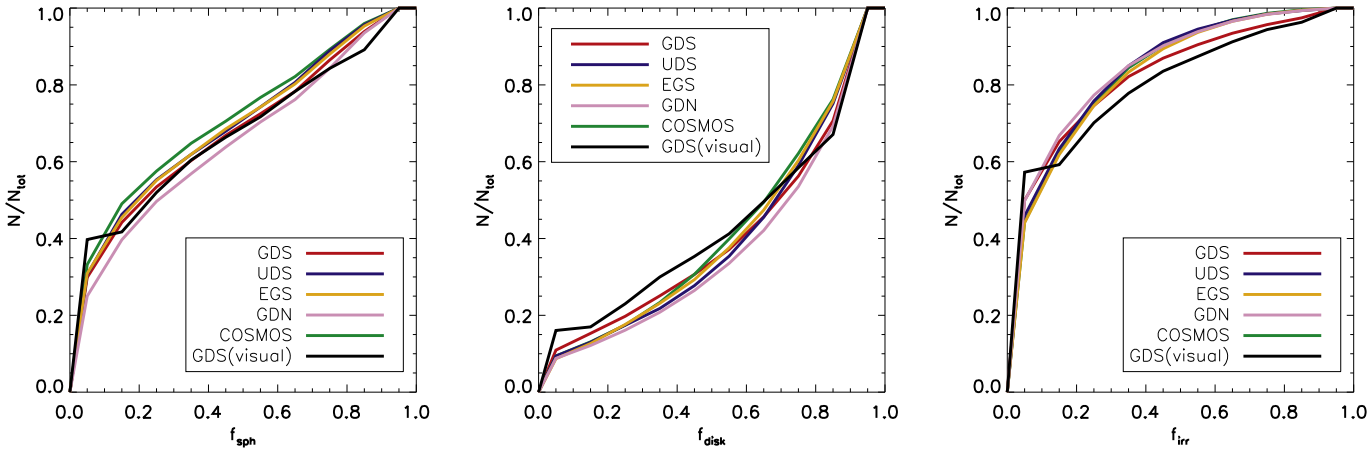

Figure 17. Cumulative distribution functions (CDFs) of $f_{\rm sph}$ (left), $f_{\rm disk}$ (middle), and $f_{\rm irr}$ (right) derived in all 5 CANDELS fields as labelled. We also show in black the CDF of visual classifications in GDS (after addition of random noise). There are no major differences between the fields, and the distributions follow the distributions of the visual classification.

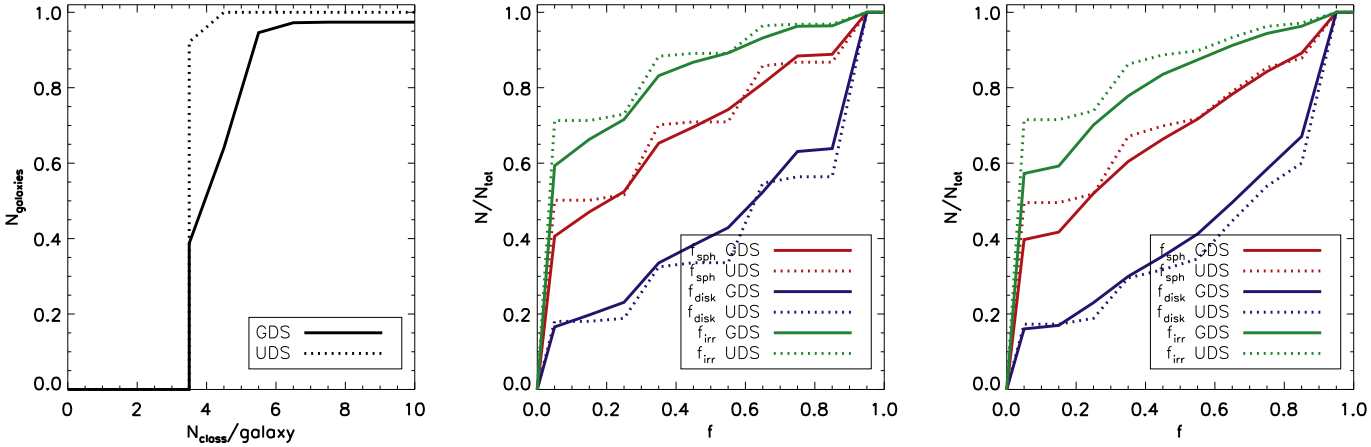

Figure 18. Left: number of visual classifiers per galaxy in the UDS and the GOODS-S fields. 90% of the galaxies are classified by only 3 people in UDS. Middle: CDFs of the main morphological parameters in UDS and GOOD-S. Right: same CDFs after addition of Gaussian noise.

brightness, redshift, or size, but that the algorithm is able to reproduce the visual classification (with its eventual biases) with the same accuracy.

4.2. Recovering Dominant Classes and Mis–classifications

An important measurement in any automated classification scheme is the fraction of objects which are mis-classified, i.e., objects that will fall in a different morphological class in the automated classification compared to the visual one. Since both classifications are continuous in the sense that each galaxy has five real numbers associated with it, the answer to this question will strongly depend on the *boxes* one considers and on how these boxes are defined.

In order to provide an estimate of this mis-classification rate that can be compared to previous classification methods, we select objects that have a clearly dominant class (DC) in the automatic and visual classifications. We define a galaxy with a DC if at least one frequency is considerably larger than the other four. We then compare how both DCs match.

Here, we adopt a conservative offset value of 0.5 between the highest frequency and the second highest, i.e., if $f_{\rm max} > 0.75$, then the second largest probability has to be smaller than 0.25, as a criterion to identify galaxies with a clear dominant morphology. Therefore, there are five DCs, i.e., *dominant spheroid*, *dominant disk*, *dominant irregular*, *dominant point source,* and *dominant unclear*. The results of such a comparison are shown in Figure 10. The degree of agreement in the identification of the main morphology of a galaxy is ∼97%–100%.

More generally, we can also investigate how the global classification accuracy depends on the level of agreement between the classifiers. As shown in Dieleman et al. (2015) for the SDSS classification, objects for which a high number of people provided the same classification are better recovered than those that present a uniform distribution in their frequencies. This simply reflects the fact that galaxies that are not easily classified by humans are also hardly recovered by the classification model. Following the same approach as D15, we define the level of agreement a between classifiers for a five-class problem:

$$a = 1 - H(f)/\log(5),$$

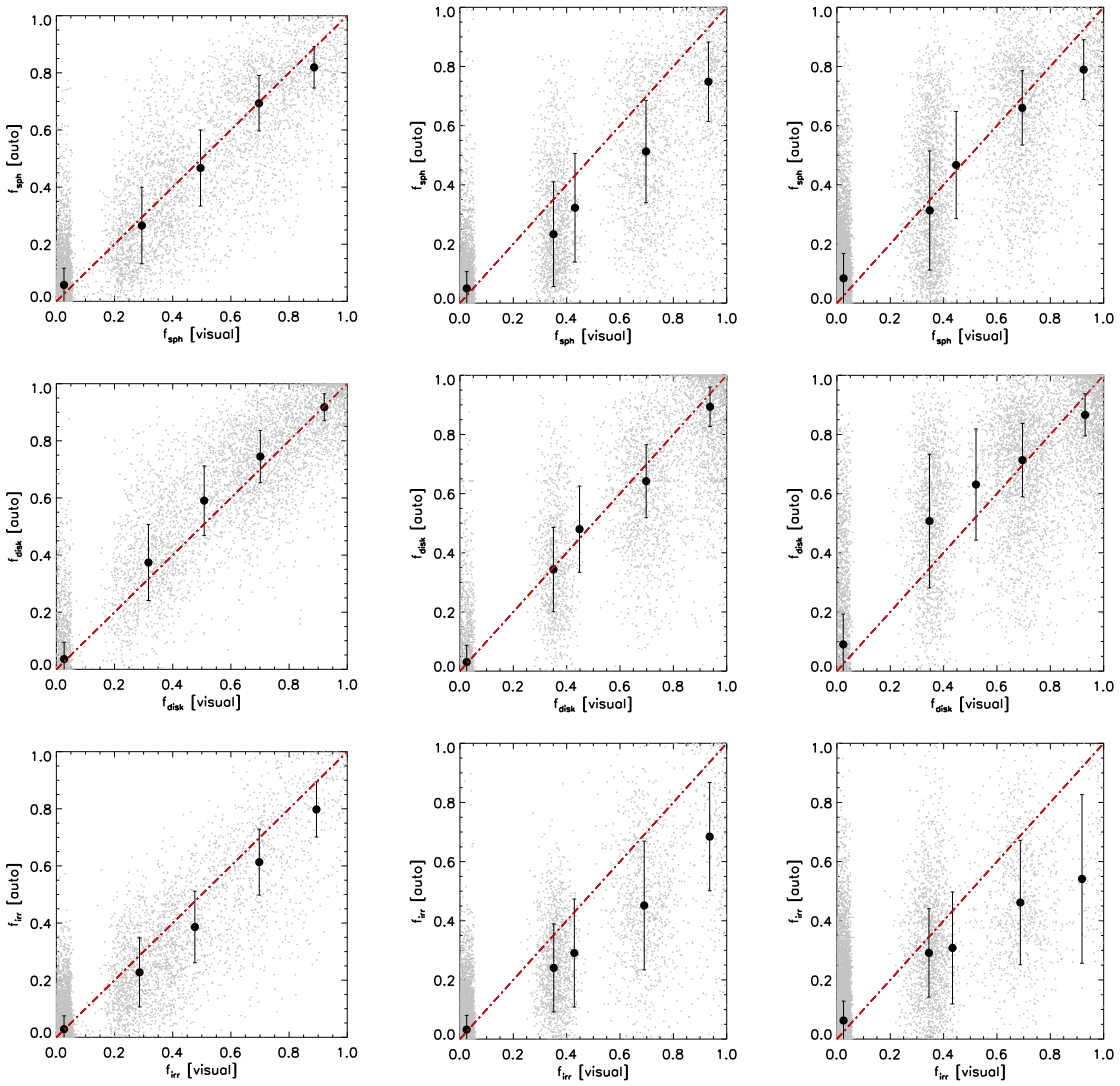

Figure 19. Correlation between the fractions of classifiers voting for a given feature. Left: GOOD-S when all classifiers are considered. Middle: GOOD-S with only three classifiers. Right: UDS where 90% of galaxies are classified by three people. The trends observed in the middle and right columns for all parameters are similar, suggesting that the worsening of the results observed in the UDS are due to a difference in the input catalog.

where $H(f)$ is the entropy defined as

$$\begin{aligned} H(f) = &- f_{\rm spheroid}\ \log\left(f_{\rm spheroid}\right) \\ &- f_{\rm disk}\ \log\left(f_{\rm disk}\right) \\ &- f_{\rm irr}\ \log\left(f_{\rm irr}\right) \\ &- f_{\rm PS}\ \log\left(f_{\rm PS}\right) \\ &- f_{\rm Unc}\ \log\left(f_{\rm Unc}\right). \end{aligned} \quad (1)$$

The agreement parameter a ranges between 0 and 1, with large values indicating high levels of agreement (most of the classifiers selected the same class) and low values associated with objects with low levels of agreement (the votes are distributed uniformly between the different classes).

Figure 11 reports the mean classification accuracy defined as the match between the automatic DC and the visual DC, as a function of a. The agreement parameter a is computed using the automatic and visual classifications. As expected, the accuracy increases when the level of agreement increases. Well-defined objects reach an accuracy >90%, but this drops to ∼50% for galaxies with $a < 0.2$. This behavior is very similar to that reported in Figure 9 of D15, which confirms the similar behavior of the classifier at high redshift.

The results above clearly represent a major step forward compared to other CAS-based methods. First, CAS methods are not able to clearly distinguish between unclassifiable objects and galaxies since the morphological parameters for unclassifiable objects can have any unpredictable value. ConvNets identify them without ambiguity.

A similar issue affects point/compact sources which will usually fall in the early-type galaxy (ETG) class in CAS methods, unless a previous cleaning is performed. The most important thing, however, is that, even for the distinction of dominant spheroids from dominant disks, advanced CAS-based methods such as galSVM do show a tail of dominant disks with high ETG probability and vice versa (Figure 12), yielding a ∼20% mis-classification rate (Huertas-Company et al. 2014). The situation is more dramatic for the distinction between dominant irregulars and dominant disks. It is almost impossible with CAS-based approaches, given that at high redshift many of the disks presents high asymmetric values (Huertas-

Table 3
Sample of the Morphological Catalog Released with the Paper

ID	IAU_NAME	R.A.	decl.	Filter	$f_{\rm spheroid}$	$f_{\rm disk}$	$f_{\rm irr}$	$f_{\rm PS}$	$f_{\rm Unc}$	$f_{\rm max}$	Δ_f	DOM_CLASS	a
1	HCPG J142112.26+5303004.5	215.3011017	53.051239	f160	0.1	0.1	0.17	0.0	0.72	0.72	0.54	4	0.38
1000	HCPG J142051.15+5300016.8	215.2131348	53.0046539	f160	0.73	0.12	0.08	0.37	0.0	0.73	0.36	0	0.34
10001	HCPG J141955.98+5253037.2	214.9832611	52.8936768	f160	0.11	1.0	0.01	0.0	0.0	1.0	0.89	1	0.82
10002	HCPG J142044.89+5301059.4	215.187027	53.0331574	f160	0.57	1.0	0.0	0.01	0.0	1.0	0.43	1	0.78
10003	HCPG J142013.52+5256044.1	215.0563202	52.9455872	f160	0.25	0.88	0.22	0.01	0.03	0.88	0.63	1	0.4
10004	HCPG J141924.91+5248004.0	214.8538055	52.8011017	f160	0.84	0.16	0.06	0.24	0.01	0.84	0.59	0	0.39
10005	HCPG J142025.18+5258045.7	215.1049042	52.9793701	f160	0.34	0.92	0.16	0.0	0.0	0.92	0.58	1	0.53
10010	HCPG J141906.89+5244043.3	214.778717	52.7453613	f160	0.34	1.0	0.09	0.0	0.0	1.0	0.66	1	0.64
10015	HCPG J141859.26+5243018.4	214.746933	52.7217865	f160	0.19	0.97	0.18	0.0	0.0	0.97	0.78	1	0.59
10017	HCPG J142009.87+5256005.7	215.0411224	52.934906	f160	0.33	0.95	0.09	0.02	0.0	0.95	0.62	1	0.55
10018	HCPG J141927.56+5248031.8	214.8648376	52.8088379	f160	0.0	0.95	0.14	0.0	0.0	0.95	0.81	1	0.78
10019	HCPG J141952.59+5253001.8	214.9691162	52.8838196	f160	0.05	0.16	0.98	0.0	0.0	0.98	0.82	2	0.71
10020	HCPG J142037.78+5301000.3	215.1574097	53.0167541	f160	0.34	0.62	0.42	0.1	0.0	0.62	0.2	1	0.22
10024	HCPG J141917.09+5246040.1	214.8211975	52.7778015	f160	0.84	1.0	0.0	0.01	0.0	1.0	0.16	1	0.89
10026	HCPG J141922.45+5247042.5	214.8435364	52.7951355	f160	0.47	0.94	0.13	0.0	0.01	0.94	0.47	1	0.55
10027	HCPG J141938.69+5250035.2	214.9111938	52.8431091	f160	0.11	0.9	0.16	0.04	0.05	0.9	0.74	1	0.44
10029	HCPG J142055.91+5304013.0	215.2329407	53.070282	f160	0.78	0.13	0.02	0.34	0.01	0.78	0.44	0	0.42
1003	HCPG J142011.48+5253015.9	215.0478363	52.8877411	f160	0.58	0.85	0.15	0.04	0.01	0.85	0.27	1	0.45
10032	HCPG J142027.07+5259005.7	215.112793	52.9849091	f160	0.18	0.66	0.56	0.0	0.0	0.66	0.1	1	0.44
10035	HCPG J141938.21+5250030.9	214.9091949	52.841919	f160	0.22	0.91	0.17	0.04	0.0	0.91	0.69	1	0.48
10036	HCPG J141939.83+5250048.2	214.9159393	52.8467102	f160	0.44	0.21	0.02	0.44	0.25	0.44	0.0	0	0.08

Note. In addition to the five main morphological indicators, for each galaxy we provide two measurements of the level of agreement between classifiers: a, linked to the entropy—see text for details; and Δ_f, the difference between the two largest frequencies. DOM_CLASS provides the dominant class (class which has the maximum frequency), being 0, spheroid, 1, disk, 2, irregular, 3, point-source, and 4 unclassifiable. The catalog can be downloaded form the rainbow database: http://rainbowx.fis.ucm.es/Rainbow_navigator_public/.

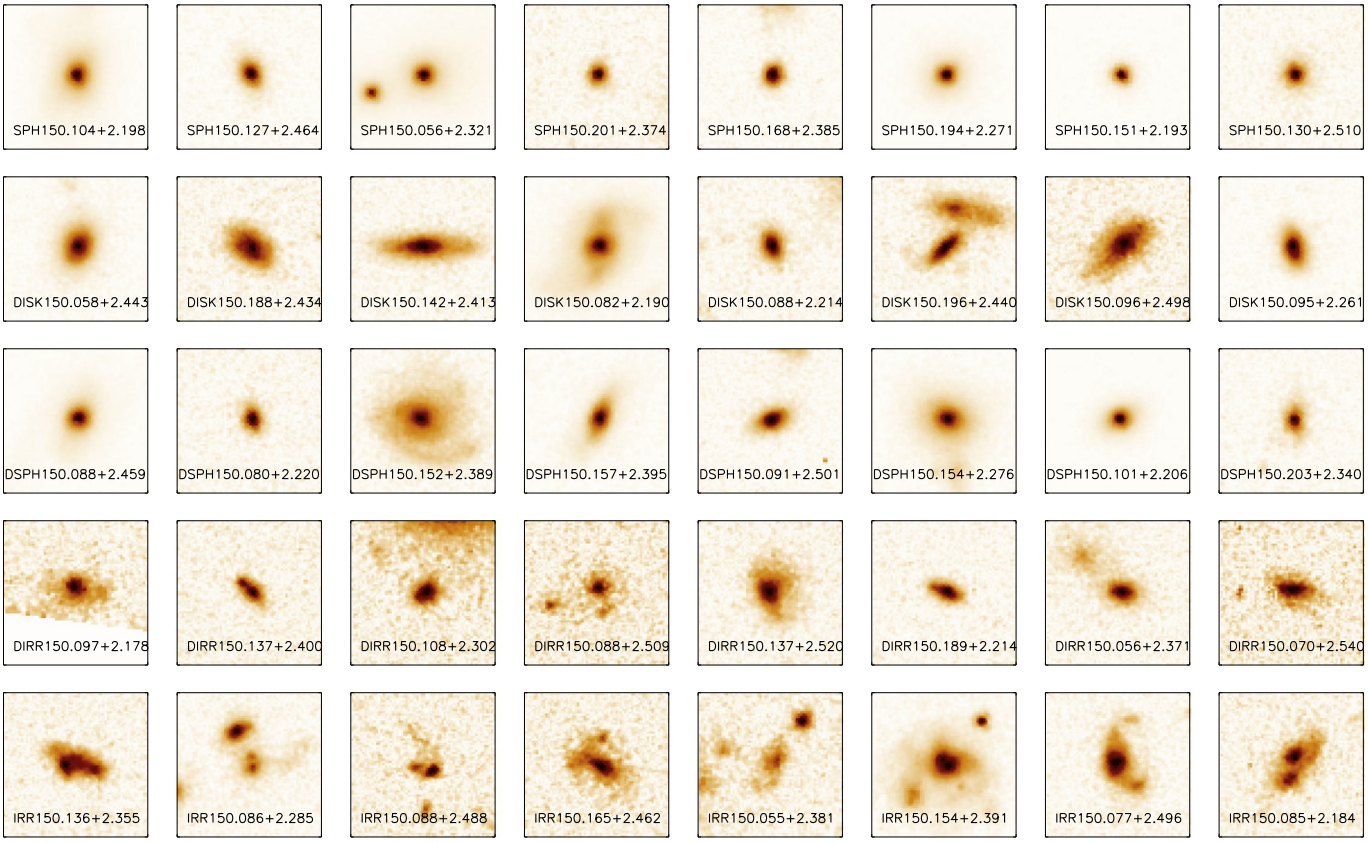

Figure 20. Examples stamps of the five morphological classes defined for illustration in the COSMOS/CANDELS field. From top to bottom, we show spheroids, disks, disk+spheroids. irregular disks and irregulars. The selection of these galaxies is done fully randomly. Recall that COSMOS galaxies have not been used for training the algorithm, and therefore they are completely new for the best model. The size of the stamps is $3\farcs8 \times 3\farcs8$.

Company et al. 2014). This is clearly shown in the right panel of Figure 12 where dominant disks have a very wide irregular probability distribution. Here, ConvNets provide a huge improvement by perfectly separating both classes.

Figure 13 shows some example stamps of these five DCs selected in the COSMOS field where no visual morphologies are available. Objects are fully randomly selected. Clearly, the visual aspect of all objects matches the DC in which they fall in the ConvNet classification, confirming the low mis-classification rate estimated in Figure 10 for GOODS-S.

4.3. Secondary Classes—Multi-component Objects

Also important are those galaxies composed of different structures. We use two parameters to identify these objects, which are simply the value of the maximum frequency ($f_{\rm max}$) and the difference between the largest and the second largest frequency (Δ_{f1-f2}). A galaxy with a fairly high $f_{\rm max}$ value and a low Δ_{f1-f2} should be a galaxy with two clear components. For the purpose of this test, we define these galaxies as those that have $f_{\rm max} > 0.5$ and a $\Delta_{f1-f2} < 0.5$.

We then look for the three different possible combinations of primary and secondary classes (Disk+Spheroid (DS), Disk +Irregular (DI), Spheroid+Irregular (SI)). Figure 14 shows the relation between the three defined two-component classes from the visual and the automatic classifications. The agreement is again close to 95% for DSs and DIs, which means that the algorithm is not only able to identify the primary class but also the secondary one whenever the galaxy has two clear morphological components. The agreement for the SI class is poor. However, this is a very marginal class since very few objects have a dominant bulge with an irregular structure. They are usually associated with bulges with some kind of structure in the surroundings in the automatic classification (Figure 15).

4.4. Uncertain Objects—Limitations

A galaxy with none of the five associated frequencies large enough (none of the available flags was clearly selected by the majority of the classifiers) should correspond to an object which has an uncertain morphology. The identification of these objects can help in understanding the limits of the morphological classification.

Figure 16 shows how the fraction of uncertain objects changes with magnitude, redshift, and stellar mass for different $f_{\rm max}$ thresholds, starting at $f_{\rm max} < 0.4$ and finishing at $f_{\rm max} < 0.7$, i.e., objects for which their maximum frequency is less than 0.4 and 0.7, respectively.

The number of objects with $f_{\rm max}$ lower than 0.4–0.5 is very small (<5%) for both the visual and automatic classifications which reflects the fact that the magnitude limit imposed ($H < 24.5$) allows us to identify a main morphology in most of the cases.

When the threshold is increased, the expected trends are observed, i.e., the number of defined *uncertain* objects increases with magnitude and redshift, and is also higher for lower stellar masses. Interestingly, the trends are very similar

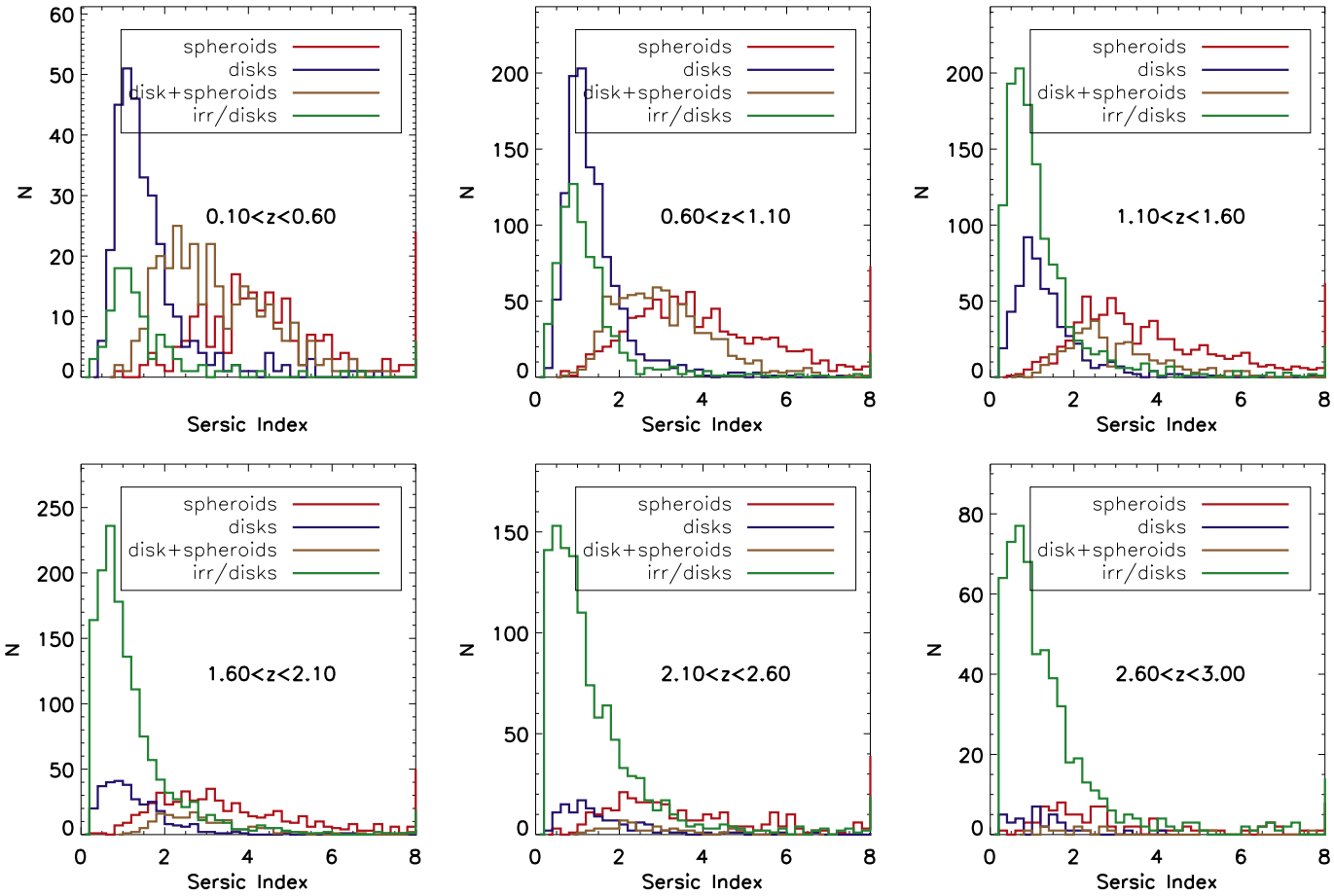

Figure 21. Sérsic index distribution for different morphological types as labeled. We show galaxies with $M_*/M_\odot > 10^{10}$. Each panel shows a different redshift bin. The expected trends are observed, i.e., bulge-dominated systems tend to have high Sérsic indices while more disky galaxies peak at lower values.

for the visual and automatic morphologies. The automated classification is therefore reproducing the same uncertainties that the human eye encounters when classifying a galaxy.

In the bottom row of Figure 16, we also show the median value of a, the level of agreement between classifiers, in bins of magnitude, redshift, and stellar mass. The level of agreement of the classification decreases for faint, distant, and low-mass objects as expected. The strongest correlation, however, is with magnitude, indicating that the main limitation to properly classify a galaxy is the signal-to-noise ratio. Notice also that the median level of agreement is always >0.4 which, according to Figure 11, corresponds to an accuracy >80% for all objects.

5. ACCURACY IN ALL CANDELS FIELDS

All previous results are based on GOODS-S where visual classifications are available for training and testing. The main purpose of the present work is to extend the classification to all CANDELS fields where visual inspection is not yet available. It is therefore important to provide an estimate of how the algorithm behaves in these *blank* fields.

5.1. Field-to-field Homogeneity

One quick sanity check consists of making sure that there are no significant statistical differences among the morphological distributions in the different fields. We do expect that all of the fields should have similar fractions of all morphologies within cosmic variance since they have similar depths and are selected randomly. It is true that the CANDELS surveys has some *deep* and *wide* areas which are observed at different depths. However, in this work, we impose a magnitude cut much brighter than the magnitude limit of the survey, and so our classification should not be affected by these different depths. Therefore, eventual significant differences could be a sign of biases in the derived morphological classifications in a given field and an eventual signature of over-fitting problems.

Figure 17 shows the cumulative distribution functions (CDFs) of the different frequencies ($f_{\rm sph}$, $f_{\rm disk}$, $f_{\rm irr}$) in the five fields. We do not observe significant differences from field to field in the distribution of frequencies, suggesting that the algorithm behaves in a similar way independently of the field. Recall, however, that the machine tends to *smooth* the distribution compared to the visual one. In other words, it removes any gap or abrupt changes. Gaps are instead present in the visual classifications given the reduced number of classifiers per object (even after noise addition).

5.2. UDS Visual Classification

During the production of the automated classification presented in this work, the visual classification for the UDS field was finalized using the same classification scheme. Comparing the resulting parameters with the automated results on this field is therefore a fully independent test of the morphologies released in this work and a definitive test to rule out any over-fitting issues.

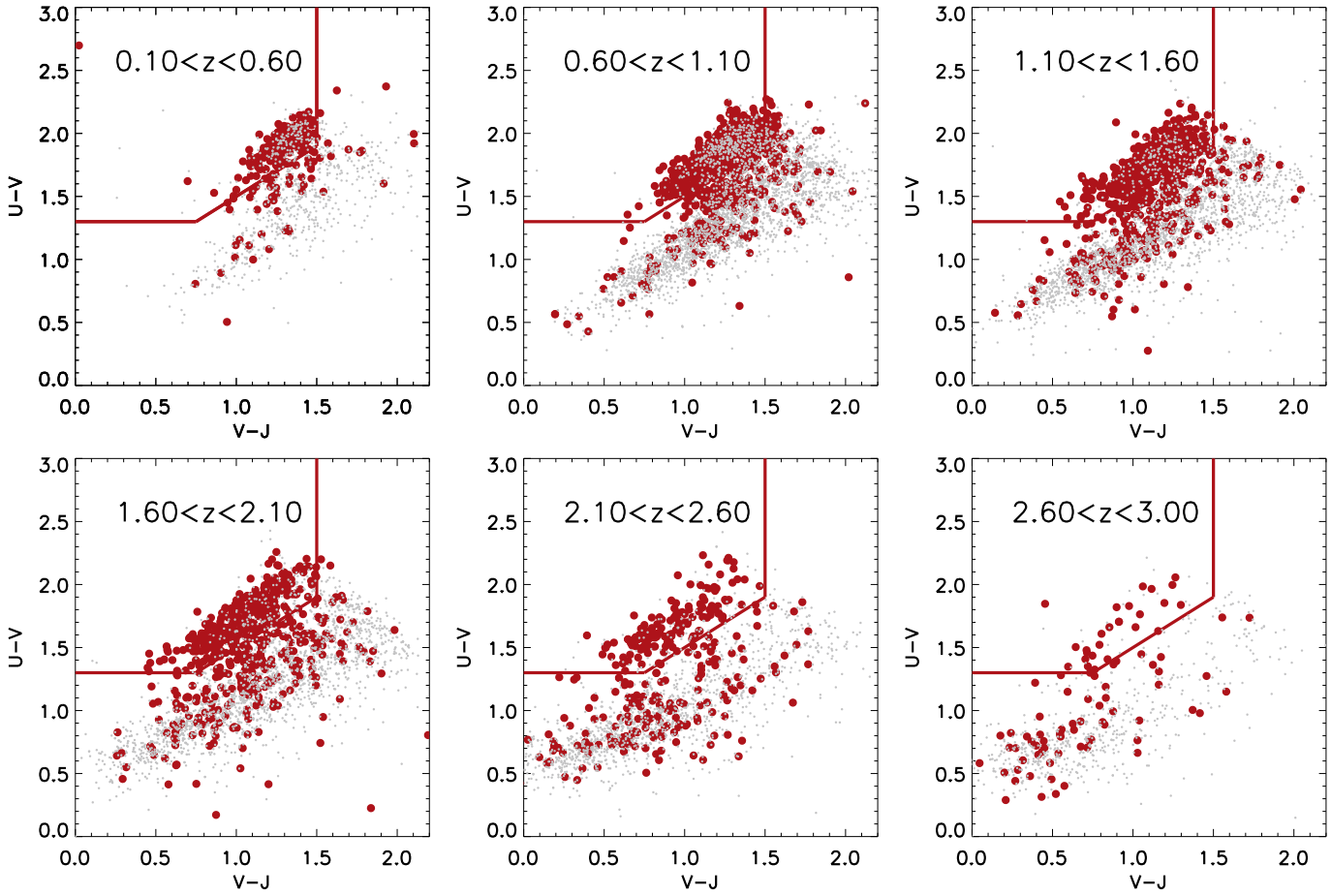

Figure 22. UVJ plane for $M_*/M_\odot > 10^{10}$ galaxies in different redshfit bins as labeled. Red dots show spheroids and gray points show all other galaxies. The red lines show the location of passive galaxies according to Whitaker et al. (2012).

There are unfortunately important differences between the visual classifications in GOOD-S and UDS that need to be taken into account before performing a fair comparison.

As a matter of fact, as shown in Figure 17, the distribution of the morphological parameters for the ConvNets classification is similar in all fields and mimics the distribution of the visual GOODS-S classification, as expected. The problem is that while in GOODS-S the number of classifiers per galaxy is roughly homogeneously distributed between 3 and 5 with some galaxies classified by ∼50 people, in UDS ∼90% of the galaxies are only classified by 3 people and the remaining 5% by 4 (see Figure 18). This difference results in a different distribution of the visual morphological frequencies between UDS and GOODS-S (i.e., frequencies in UDS only have 4 possible values for most of the galaxies) which persists even after the addition of random noise for smoothing (Figure 18). Since the automated classification necessarily follows the distribution for which it was trained, the comparison with UDS visual classifications will have a larger scatter which is not due to a failure in the algorithm but to a difference in the inputs.

In order to estimate how much this will affect the comparison in the UDS, we recomputed the GOOD-S frequencies by randomly taking only three classifiers per galaxy (i.e., ignoring the classifications whenever there are more than three classifiers) and compared with the automated classification as done in Figure 7.

The results of such an exercise are shown in Figure 19. In the left column, we plot the comparison when all classifiers are taken into account (as in Figure 7) and in the middle column the same comparison but only with three classifiers. There is a clear increase of the scatter and the bias which is only caused by the change of the distribution of the input values (the output is exactly the same). Interestingly, the trends are very similar to what is observed in the comparison with the UDS (right column), which suggests that the worsening of the results in the UDS is not due to a bad behavior of the algorithm for this field, but is simply due to a different distribution of the inputs.

The latter effect can also be understood if we consider that, at the first level, the process of having n classifiers visually selecting between two labels (binary classification) follows a binomial distribution. Let us assume, for example, that an image has an intrinsic probability p of being classified as a spheroid. It follows that the variance of the distribution of the number of people labeling it as "yes" from a total of n is $np(1-p)$. Therefore, the deviation of the visually classified fractions is $\sqrt{p(1-p)/n}$. The deviation of the fractions will depend on the intrinsic probability p and the number of annotations. The fewer annotators we have, the higher the variance on the fractions, i.e., the less reliable the probabilities of each class will become (compared to the intrinsic one). Therefore, training a machine with a noisier training set will also result in a noisier classification.

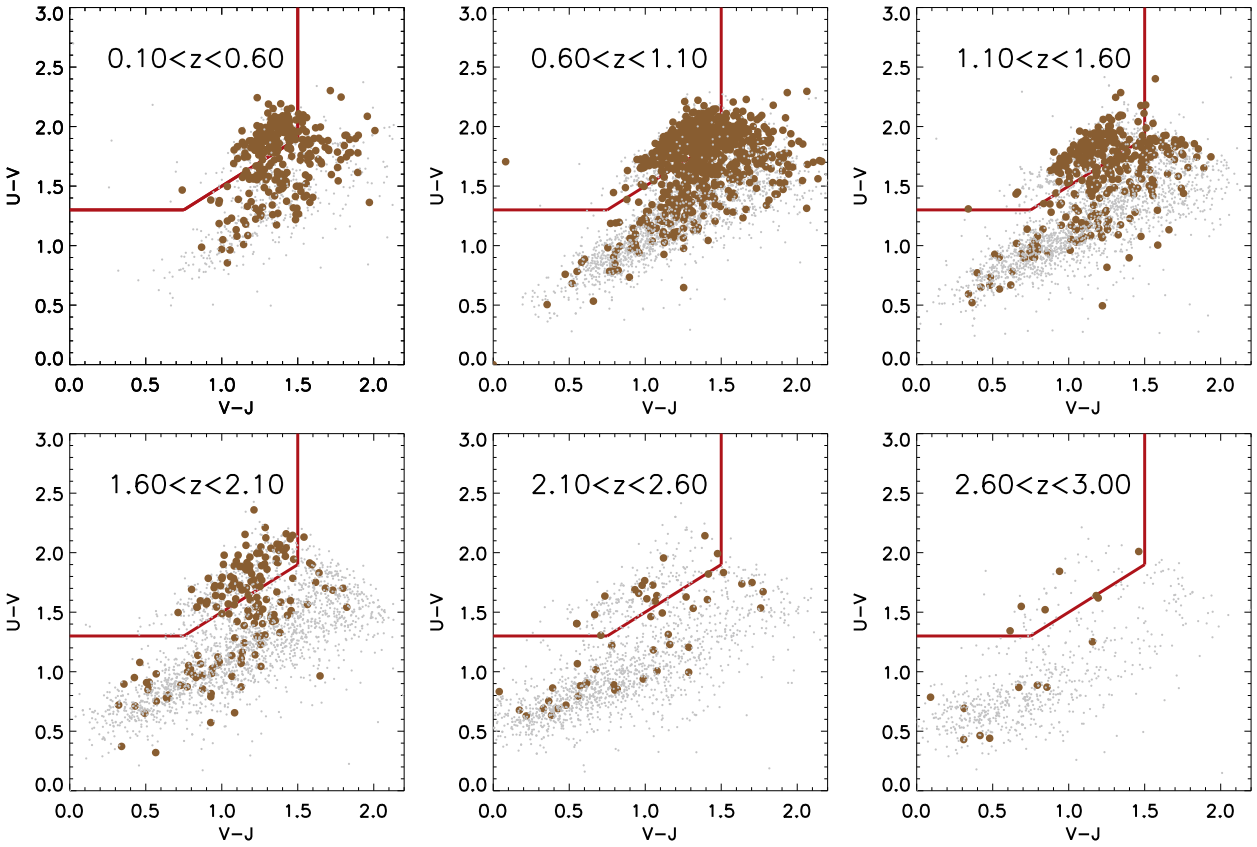

Figure 23. UVJ plane for $M_*/M_\odot > 10^{10}$ galaxies in different redshfit bins as labeled. Brown dots show disk+spheroids systems and gray points show all other galaxies. The red lines show the location of passive galaxies according to Whitaker et al. (2012).

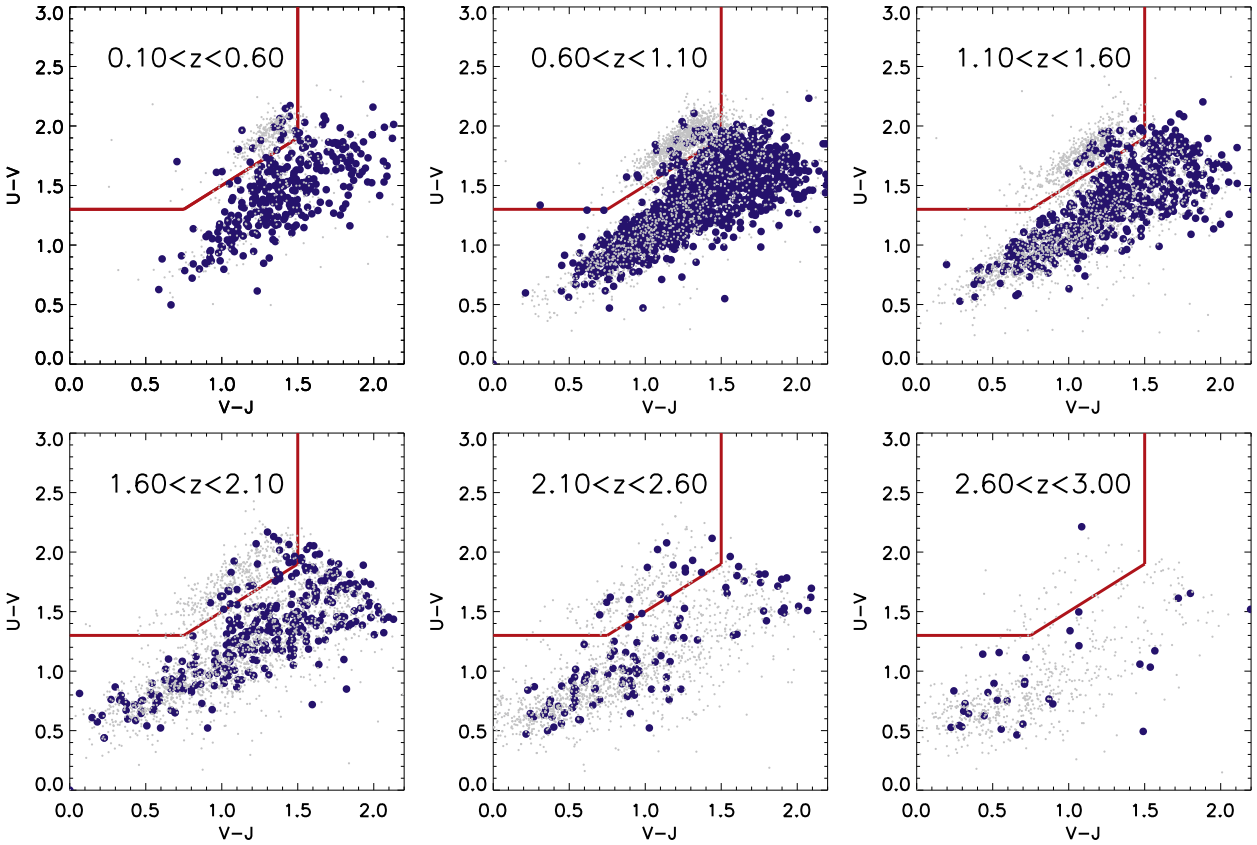

Figure 24. UVJ plane for $M_*/M_\odot > 10^{10}$ galaxies in different redshfit bins as labeled. Blue dots show disks and gray points show all other galaxies. The red lines show the location of passive galaxies according to Whitaker et al. (2012).

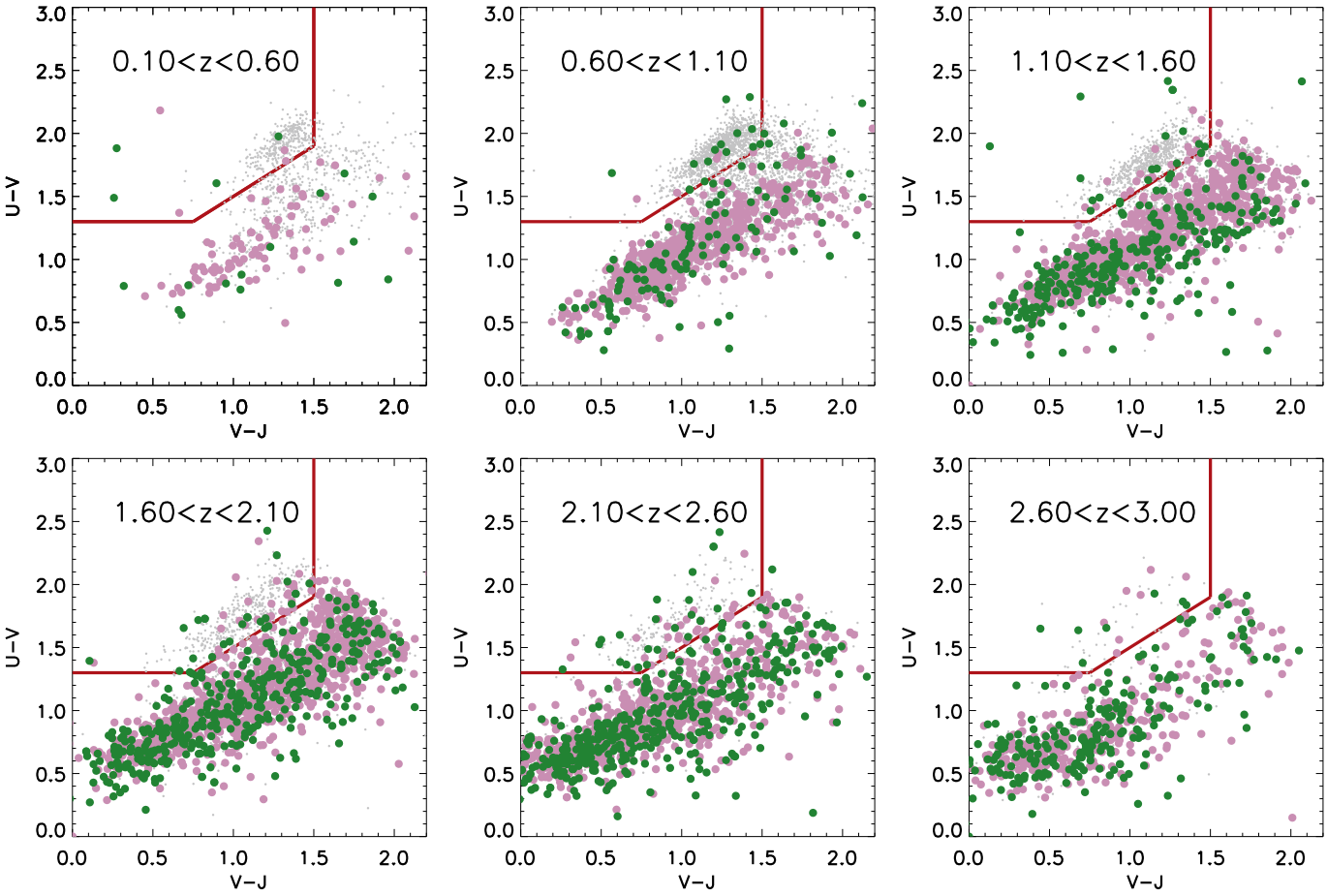

Figure 25. UVJ plane for $M_*/M_\odot > 10^{10}$ galaxies in different redshfit bins as labeled. Green and violet dots show irregular and disk/irregular galaxies, respectively, and gray points show all other galaxies. The red lines show the location of passive galaxies according to Whitaker et al. (2012).

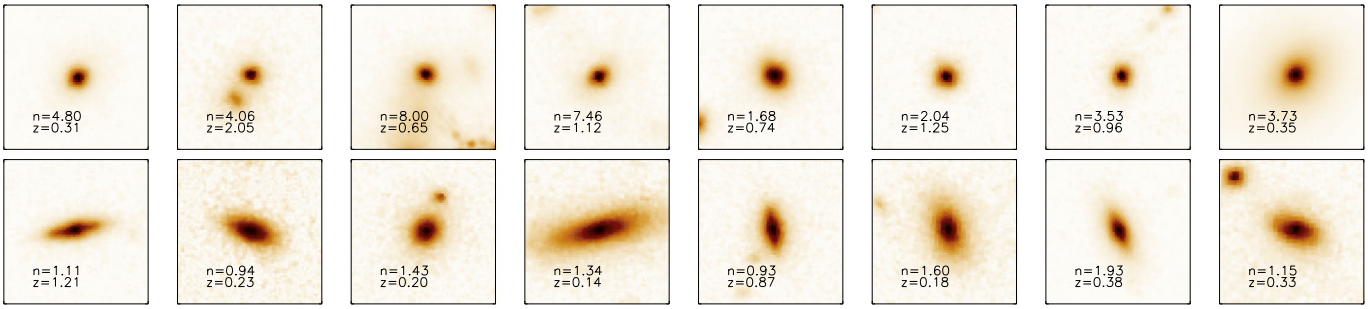

Figure 26. Example stamps of star-forming spheroids (top row) and passive disks (bottom row). For each galaxy we show the sersic index and the redshift.

This issue emphasizes one main advantage of the automated classifications with respect to the visual when a small number of classifiers is involved. Namely, the results are by definition homogeneous for all data sets. The fact that the UDS and the GOODS-S with only three classifiers look very similar also suggests that the algorithm has a similar accuracy in both fields, confirming that the classification is not severely affected by over-fitting.

6. CATALOG

This paper is accompanied by the public release of the morphology of all of the galaxies in the CANDELS fields brighter than $H_{\rm F160W} = 24.5$. In addition to the five morphological parameters, we also provide in the catalog two measurements of the quality of the classification discussed in the text (a and $\Delta_{f_1-f_2}$) as well as the DC and the maximum frequency $f_{\rm max}$. Table 3 shows the first few lines of the catalog. The catalog is released through the Rainbow database: http://rainbowx.fis.ucm.es/Rainbow_navigator_public/.

The classification provided is by definition continuous, since each galaxy has five parameters spanning from 0 to 1. The use of these parameters to actually define morphological classes strongly depends on the science purposes and the galaxy properties one would like to highlight. Establishing thresholds in the different fractions necessarily implies a trade-off between pure and complete samples.

For illustration purposes on how to use the catalog, we propose one possible classification in five different morphological classes based on establishing thresholds in the different frequencies (see Huertas-Company et al. 2015a):

1. *pure bulges [SPH]:* $f_{\rm sph} > 2/3$ AND $f_{\rm disk} < 2/3$ AND $f_{\rm irr} < 1/10$;
2. *pure disks [DISK]:* $f_{\rm sph} < 2/3$ AND $f_{\rm disk} > 2/3$ AND $f_{\rm irr} < 1/10$;
3. *disk+sph [DISKSPH]:* $f_{\rm sph} > 2/3$ AND $f_{\rm disk} > 2/3$ AND $f_{\rm irr} < 1/10$;
4. *irregular disks [DISKIRR]:* $f_{\rm disk} > 2/3$ AND $f_{\rm sph} < 2/3$ AND $f_{\rm irr} > 1/10$;
5. *irregulars/mergers [IRR]:* $f_{\rm disk} < 2/3$ AND $f_{\rm sph} < 2/3$ AND $f_{\rm irr} > 1/10$.

The thresholds are obviously arbitrary but have been calibrated through visual inspection to make sure that they result in different morphological classes. The smoothed particle hydrodynamics (SPH) class contains galaxies fully dominated by the bulge component with little or no disk at all. The DISK class is made of galaxies in which the disk component dominates over the bulge. Between both classes, lies the DISKSPH class in which we put galaxies with no clear dominant component. Then, we distinguish 2 types of irregulars: DISKIRR, i.e., disk-dominated galaxies with some asymmetric features; and IRR, which are irregular galaxies with no clear dominant disk component (including mergers).

Some random example stamps in the COSMOS field are shown in Figure 20. Also, for illustration purposes, in Figures 21–25 we show the Sérsic index distributions and UVJ planes for galaxies with $M_*/M_\odot > 10^{10}$ split in different morphological types and for several redshift bins. The expected trends are observed in both figures and are also very similar to the distributions shown by Kartaltepe et al. (2014) on which our classification is based.

We observe that the different morphological types have very different Sérsic index distributions. Objects with a clear bulge component according to their visual inspection (spheroids and bulge+disk systems) tend to have larger Sérsic indices and also tend to be located in the passive zone of the UVJ plane. Disk-dominated objects peak at $n \sim 1$ and are star-forming based on their locus on the UVJ plane.

One interesting class is the bulge+spheroid class (i.e., objects with no clear dominant disk or spheroidal component) because they do not have a clear locus in the UVJ diagram. Roughly half of these are passive and the other half are star-forming. Any selection based on star formation activity will therefore split this population into two groups. Having a pure morphological classification enables us to isolate objects that are difficult to identify with colors and/or single profile fitting. It is also interesting to note that the large morphological catalog put together in this paper allows to study objects which deviate from the general trends (i.e., passive disks, star-forming bulges) with reasonable statistics (see Figure 26).

7. SUMMARY AND CONCLUSIONS

This work presents a *visual-like* morphological classification of ∼50,000 galaxies ($H < 24.5$) in 5 CANDELS fields (GOODS-S, GOODS-N, UDS, COSMOS, and EGS) in the H band, which probes optical rest-frame morphologies in the redshift range $1 < z < 3$. The sample is ∼80% complete down to $\log(M_*/M_\odot) \sim 10$.

Morphologies are estimated with a five-layer Convolutional Neural Network (ConvNet) followed by two layers of fully connected perceptrons trained to reproduce the visual morphologies of ∼8000 galaxies in GOODS-S published by the CANDELS collaboration (Kartaltepe et al. 2014). ConvNets are a particular family of neural networks that take advantage of the image stationarity to mimic the way human brain cells behave to recognize specific patterns.

Following the approach in CANDELS, we associate five real numbers, $f_{\rm spheroid}$, $f_{\rm disk}$, $f_{\rm irr}$, $f_{\rm PS}$ and $f_{\rm Unc}$, with each galaxy corresponding, respectively, to the frequency at which *expert classifiers* flagged a galaxy as having a bulge, having a disk, presenting an irregularity, being compact or point-source, and being unclassifiable. Galaxy images are interpolated to a fixed size, rotated, and randomly perturbed before feeding the network to (i) avoid over-fitting and (ii) reach a comparable ratio of background versus galaxy pixels in all images.

ConvNets are able to predict the *votes* of expert classifiers with a <10% bias and a ∼10% scatter. This makes the classification almost equivalent to a visual-based classification. The training took 10 days on a GPU and the classification is performed at a rate of 1000 galaxies/hour. As opposed to generalized CAS methods (i.e., galSVM), ConvNets are able to identify without ambiguity (<1% mis–classifications) objects that are not galaxies (high $f_{\rm Unc}$ values), distinguish irregulars from disks at all redshifts, and spheroids from disks.

The catalog of ∼50,000 galaxies is released with the present paper through the Rainbow database: http://rainbowx.fis.ucm.es/Rainbow_navigator_public/. The catalog actually increases by a factor of five the existing (public) morphologies in the CANDELS fields and is intended to be used for many diverse scientific applications (i.e., evolution of merger rates, morphological evolution from $z \sim 3$, morphology-density/environment relation, morphology-active galactic nucleus connection, etc.).

Future efforts will be focused on optimizing deep-learning-based approaches like the one presented here for EUCLID/WFIRST/LSST like data, analyzing deeper data such as the Hubble Frontier Fields, and providing more detailed morphological descriptors in CANDELS (i.e., tidal features etc.).

We thank the two anonymous referees for contributing to significantly improve this work. M.H.C acknowledges D. Gratadour for kindly giving us access to the GPU cluster at LESIA. G.C.V gratefully acknowledges financial support from CONICYT-Chile through its doctoral scholarship and grant DPI20140090. S.M. acknowledges financial support from the Institut Universitaire de France (IUF), of which she is senior member. G.B., D.C.K., and S.M.F. acknowledge support from NSF grant AST-08-08133 and NASA grant HST-GO-12060.10A.

REFERENCES

Abraham, R. G., van den Bergh, S., Glazebrook, K., et al. 1996, ApJS, 107, 1
Ball, N. M., & Brunner, R. J. 2010, IJMPD, 19, 1049
Ball, N. M., Loveday, J., Fukugita, M., et al. 2004, MNRAS, 348, 1038
Bernardi, M., Meert, A., Vikram, V., et al. 2012, arXiv:1211.6122
Bertin, E. 2012, adass XXI, 461, 263
Ciresan, D., Meier, U., Schmidhuber, J., et al. 2012, in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 3642
Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Fukushima, K. 1980, Biological Cybernetics, 36, 193
Galametz, A., Grazian, A., Fontana, A., et al. 2013, ApJS, 206, 10
Guo, Y., Ferguson, H. C., Giavalisco, M., et al. 2013, ApJS, 207, 24
Hubble, E. P. 1926, ApJ, 64, 321
Hubble, E. P. 1936, Realm of the Nebulae (New Haven, CT: Yale Univ. Press)
Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S., & Sánchez Almeida, J. 2011, A&A, 525, A157

Huertas-Company, M., Kaviraj, S., Mei, S., et al. 2014, arXiv:1406.1175
Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971
Huertas-Company, M., Tasca, L., Rouan, D., et al. 2009, A&A, 497, 743
Kartaltepe, J. S., Mozena, M., Kocevski, D., et al. 2014, arXiv:1401.2455
Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, Advances in Neural Information Processing Systems (Cambridge, MA: MIT Press), 1097
Lotz, J. M., Davis, M., Faber, S. M., et al. 2008, ApJ, 672, 177
Matusugu, M., Mori, K., Mitari, Y., & Kaneda, Y. 2003, NN, 16, 555
Peth, M. A., Lotz, J. M., Freeman, P. E., et al. 2015, arXiv:1504.01751
Russakovsky, O., Deng, J., Su, H., et al. 2014, ImageNet Large Scale Visual Recognition Challenge, arXiv:1409.0575
Scarlata, C., Carollo, C. M., Lilly, S., et al. 2007, ApJS, 172, 406
Shamir, L., & Wallin, J. 2014, MNRAS, 443, 3528
Whitaker, K. E., van Dokkum, P. G., Brammer, G., & Franx, M. 2012, ApJL, 754, L29