https://doi.org/10.1088/1538-3873/aaeeb4

Multiband Galaxy Morphologies for CLASH: A Convolutional Neural Network Transferred from CANDELS

M. Pérez-Carrasco[1], G. Cabrera-Vives[1,2], M. Martinez-Marin[3], P. Cerulo[3], R. Demarco[3], P. Protopapas[4], J. Godoy[1], and M. Huertas-Company[5,6]

[1] Department of Computer Science, University of Concepción, Casilla 160-C, Concepción, Chile; maperezc@udec.cl
[2] Millennium Institute of Astrophysics, Santiago, Chile
[3] Department of Astronomy, Universidad de Concepción, Casilla 160-C, Concepción, Chile
[4] Institute for Applied Computational Science, Harvard University Northwest B162, 52 Oxford Street Cambridge, MA 02138, USA
[5] LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Universités, UPMC Univ. Paris 06, F-75014 Paris, France
[6] University of Paris Denis Diderot, University of Paris Sorbonne Cité (PSC), 75205 Paris Cedex 13, France

Received 2018 June 6; accepted 2018 October 25; published 2019 August 23

Abstract

We present visual-like morphologies over 16 photometric bands, from ultraviolet to near-infrared, for 8412 galaxies in the Cluster Lensing And Supernova survey with *Hubble* (CLASH) obtained using a convolutional neural network (ConvNet) model. Our model follows the Cosmic Assembly Near-IR Deep Extragalactic Legacy Survey (CANDELS) main morphological classification scheme, obtaining the probability for each galaxy at each CLASH band of being spheroid, disk, irregular, point source, or unclassifiable. Our catalog contains morphologies for each galaxy with $H_{\mathrm{mag}} < 24.5$ in every filter where the galaxy is observed. We trained an initial ConvNet model using approximately 7500 expert eyeball labels from CANDELS. We created eyeball labels for 100 randomly selected galaxies per each of the 16-filter set of CLASH (1600 galaxy images in total), where each image was classified by at least five of us. We use these labels to fine-tune the network to accurately predict labels for the CLASH data and to evaluate the performance of our model. We achieve a root-mean-square error of 0.0991 on the test set. We show that our proposed fine-tuning technique reduces the number of labeled images needed for training, as compared to directly training over the CLASH data, and achieves a better performance. This approach is very useful to minimize eyeball labeling efforts when classifying unlabeled data from new surveys. This will become particularly useful for massive data sets such as those coming from near-future surveys such as EUCLID or the LSST. Our catalog consists of prediction of probabilities for each galaxy by morphology in their different bands and is made publicly available at http://www.inf.udec.cl/~guille/data/Deep-CLASH.csv.

Key words: Galaxies – Morphology – Classification – Catalog

Online material: color figures

1. Introduction

Galaxies are complex systems and understanding their evolution represents one of the key questions in modern astrophysics. The physical properties of a galaxy, such as its baryonic content (stellar, gas and, dust masses), star formation rate (SFR), structure (morphology) and chemical abundance, change over time, and their evolution is driven by a combination of internal and environmental processes (Kauffmann et al. 2003, 2004; Peng et al. 2010; Muzzin et al. 2012). The morphology of galaxies, which is related to their structural and dynamical properties and to their star formation history (quantified by their SFR over time), represents a fundamental and powerful diagnostic to study their evolutionary changes.

Since the first attempts to understand these "nebulae," astronomers have been classifying galaxies according to their visual appearance. In 1926, Hubble proposed a tuning-fork scheme (Hubble 1936) that would constitute the basis for any morphological classification until the present. In the tuning-fork diagram, galaxies are separated into early and late morphological types, the former including galaxies without disks and spiral arms and the latter comprising all galaxies that showed spiral arms, with a spheroidal or elliptical light distribution, and in their disks. In Hubble's scheme there is also a class of lenticular galaxies, also known as S0, which correspond to systems made up of a central bulge structure surrounded by a disk component but without spiral arms. Finally, a class of irregular galaxies, which do not show any prominent morphological signature (e.g., the Magellanic clouds), have also been defined and considered as late morphological type.

This bi-modality in galaxy morphology provides us with information about the structural distribution of their stellar composition and therefore the processes that shaped them,

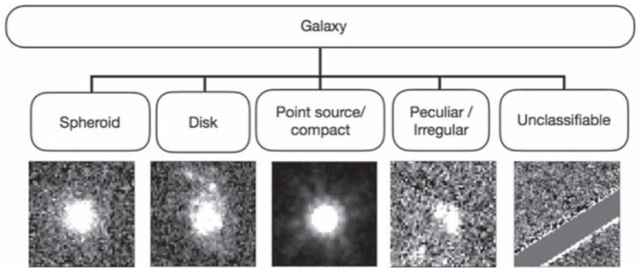


Figure 1. Morphology classification scheme used to classify galaxies in CLASH, from KA15. Images show examples for each class in the CLASH f160w filter.

including stellar mass assembly, galaxy–galaxy interactions and other environmental effects. An open question is whether there is any clear evolutionary link between early- and late-type galaxies.

To address this question, astronomers have classified galaxies in samples of varying sizes and built morphological catalogs. Traditionally, galaxies have been classified by visually inspecting high-resolution images of varying depths and sizes (e.g., de Vaucouleurs et al. 1991; Dressler 1980; Dressler et al. 1997; Couch et al. 1998; Postman et al. 2005; Desai et al. 2007; Nair & Abraham 2010). Usually, more than one classifier is employed in this task. Although the human eye is able to distinguish a great amount of detail, such as spiral arms, bulges, and bars, visual classification has gradually become impractical as the volume of data delivered by astronomical surveys has grown significantly. Astronomers have therefore passed from the classification of a few hundred galaxies to classifying tens to hundreds of thousands (Nair & Abraham 2010; Kartaltepe et al. 2015).

In the last decade, the morphological classification of galaxies has been addressed from two complementary points of view, namely, visual classification by a large number of non-expert people through crowd-sourcing platforms (e.g., Galaxy Zoo and Galaxy Zoo CANDELS (definition below); Lintott et al. 2008; Simmons et al. 2017) and automated classification through machine learning algorithms. The latter tries to find a set of physical and observational parameters that correlate with the visual morphology of a galaxy and defines the space of parameters that best characterizes a given morphological type (Abraham et al. 1996; Conselice et al. 2000; Lotz et al. 2008; Huertas-Company et al. 2008, 2011). These methods suffer from difficulties in reaching the levels of reliability required for scientific analysis and their uncertainties remain high when one tries to go beyond the early- versus late-type schemes and to distinguish elliptical from lenticular galaxies (see, e.g., the discussion in Cerulo et al. 2017).

In recent years, deep-learning methods that mimic human eye perception have been able to learn the best set of parameters for a given problem. Dieleman et al. (2015; hereafter D15) trained convolutional neural networks (ConvNets; Fukushima 1980; LeCun et al. 1998), a deep-learning model, retrieving Galaxy Zoo's visual classifications from Sloan Digital Sky Survey (SDSS; York et al. 2000) images with >99% accuracy. A similar method was later applied at higher redshifts in the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS; Grogin et al. (2011) by Huertas-Company et al. 2015) (see Section 2 for details).

Motivated by these works we decided to apply a ConvNet-based classification to galaxies in the Cluster Lensing and Supernova Survey with *Hubble* (CLASH; Postman et al. 2012), a survey of 25 clusters of galaxies at redshifts $0.2 < z < 0.8$ observed in up to 16 photometric bands with the Advanced Camera for Surveys (ACS) and the Wide Field Camera 3 (WFC) on board the *Hubble Space Telescope* (*HST*). CLASH provides us with publicly available deep and high-resolution ($0.065''$/pixel or $0.03''$/pixel) images of massive cluster cores with a wavelength coverage that spans the electromagnetic spectrum from the near-ultraviolet (0.2 μm) all the way through to the near-infrared (1.6 μm). CLASH constitutes the optimal counterpart to deep-field surveys such as CANDELS for environmental studies of galaxies at intermediate redshifts, because clusters of galaxies are the most massive, virialized large-scale structures in the universe and host a broad variety of environments, from the dense cores to the sparser and dynamically active outskirts. All this makes them observational laboratories for the study of the environmental processes that drive the evolution of galaxies. Galaxy clusters are characterized by a morphology–density relation (Dressler 1980) up to $z \approx 1.5$. More precisely, it is observed that early-type galaxies in clusters are more frequent in their inner regions, while the fraction of late-type galaxies increases towards the cluster outskirts (Dressler et al. 1997; Postman et al. 2005; Hilton et al. 2009; Muzzin et al. 2012; Mei & Stanford 2012; Holden et al. 2007). It is also observed that the fraction of blue, star-forming galaxies in clusters increases with redshift (Butcher & Oemler 1978), suggesting that clusters promote the suppression of star formation and the formation of early-type galaxies with time (see also Fasano et al. 2001).

To classify galaxies in CLASH we train a ConvNet architecture based on Inception (Szegedy et al. 2014) using CANDELS F160W (H band, 1.6 μm) images of galaxies in the GOODS-S field that were visually classified by Kartaltepe et al. (2015) (KA15 hereafter), obtaining a root-mean-square error (RMSE) of $\sim$0.125. Then, we fine-tune our model to the CLASH data using the previous knowledge acquired from CANDELS. We predict labels on CLASH galaxies in each of the 16 available *HST* photometric bands independently (F225W, F275W, F336W, F390W, F435W, F475W, F606W, F625W, F775W, F814W, F850LP, F105W, F110W, F125W, F140W, F160W) obtaining a 0.0991 RMSE over a subsample of CLASH galaxies visually labeled according to the classification scheme in Figure 1. We apply our model to 8412 galaxies from CLASH and release the first multi-band morphological catalog of CLASH galaxies. Our results are in agreement with recent results presented by Domínguez Sanchez et al. (2018), showing that transfer learning techniques can significantly reduce the number of labels needed for visual tasks.

This paper is organized as follows: in Section 2 we provide a description of the principles and functioning of ConvNets and transfer learning; in Section 3 we discuss our training sample and labeling scheme; in Section 4 we present our methodology; in Section 5 we present our results and discussion; in Section 6 we present the catalog; finally in Section 7 we summarize and conclude our work.

2. Technical Background and Preparatory Work

2.1. Convolutional Neural Networks

A ConvNet (Fukushima 1980; LeCun et al. 1998) is a deep-learning model aimed at modeling data, through different abstraction levels, using convolutional stacks followed by nonlinearities and statistical aggregation operations (e.g., maxpooling; Nagi et al. 2011). The resulting features are connected to a prediction model, such as a feedforward neural network (see Zhang 2000 and references therein), to obtain responses for a learning task. They are commonly used when the data exhibit a given kind of topological structure that needs to be preserved. This is the case of the pixels in an image.

The parameters of the model are iteratively learnt by minimizing a loss function for subsets of the data, called mini-batches. The parameters of the model are updated for each mini-batch of training examples using backpropagation (LeCun et al. 1998). Backpropagation is an efficient algorithm to compute the partial derivatives of the loss function with respect to the parameters of each layer of the model, in order to iteratively modify such parameters using a gradient descent method. Stochastic gradient descent updates the parameters in the opposite direction of the gradients modulated by a *learning rate*. Adam (Kingma & Ba 2014) is a stochastic optimization algorithm used to update weights in which the learning rate changes during training. The way in which Adam computes learning rates is through an exponentially decaying average of past gradients and the squared decaying averages of past gradients (Ruder 2016), and then controls the decay rates of these moving averages using a bias correction.

Despite ConvNets having been introduced a long time ago, the absence of computational power, enough data, and efficient regularization algorithms did not allow the successful implementation of ConvNet architectures until recently. Only in 2012, thanks to the use of ReLU nonlinearities (Nair & Hinton 2010) and dropout regularization (Hinton et al. 2012; Srivastava et al. 2014), the architecture proposed by Krizhevsky et al. (2012) obtained a considerably advantage in the ImageNet Classification Challenge, one of the most important contests in the field of computer vision, significantly improving over the previous state-of-the-art algorithms.

Since then, the most successful image classification algorithms have been based on ConvNets (Simonyan & Zisserman 2014; Szegedy et al. 2014; He et al. 2015), using deeper architectures and methods which allow better backpropagation of the errors (e.g., batch normalization; see Ioffe & Szegedy 2015 for details).

In the context of astronomical images, the first attempts to generate classification algorithms using ConvNets were made by D15 for Galaxy Zoo: The Galaxy Challenge (posted in Kaggle[7]), which tried to find the best model for morphological classification of galaxies in SDSS/Galaxy Zoo[8] labeled images. They obtained an accuracy higher than 99%. The second attempt was made by Huertas-Company et al. (2015) who trained a model to morphologically classify galaxies in five fields of the CANDELS survey. They obtained a misclassification error lower than 1%.

More recently, deep learning was used by Aniyan & Thorat (2017) to classify radio galaxies on images from the Very Large Array, while Domínguez Sánchez et al. (2018) ran a deep-learning algorithm to produce a detailed morphological classification of 670,000 SDSS galaxies. Other fields of application of deep learning in astronomy include detection of extrasolar planets, the study of transients, the characterization of pulsars, and fast radio bursts (Shallue & Vanderburg 2018; Cabrera-Vives et al. 2017; Guo et al. 2017; Connor & van Leeuwen 2018). Deep learning has also been used for the estimation of photometric redshifts in astronomical surveys with limited photometric information (e.g., Kilo Degree Survey (KiDS; Petrillo et al. 2017).

2.2. Fine-tuning ConvNet Architectures

In real-world applications, collecting enough labels to train deep-learning models is expensive and time-consuming. In some cases, a set of labeled images from a related domain (source) could be used as an initial learning set, whose obtained parameters could then be used to train a model with few labeled data (target). This is called *transfer learning* or *domain adaptation* (Pan & Yang 2010).

Deep-learning models are characterized by learning mid-level representations of the images in each convolutional layer (Yosinski et al. 2015; Zeiler & Fergus 2013) that show general abstractions in the first layers, while becoming more specific in the last layers. This allows the use of the first layers directly from the source and to train only layers which map more specific details of the images. This is called *fine-tuning* (Yosinski et al. 2014; Oquab et al. 2014).

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

[8] Galaxy Zoo is a crowd-sourcing platform in which experts and non-experts can classify jpg galaxies by their morphology.

The general suggested fine-tuning approach, when related domains are available (i.e., galaxies from different surveys), is to copy layers trained on source, keeping frozen some of these layers and training from the last layer frozen ahead on the target (Chu et al. 2016; Yosinski et al. 2014). The layers to be frozen will depend on the problem and the constructed architecture. In this paper, we evaluate the performance of our model in terms of the number of layers frozen.

3. Data

3.1. Images

We trained a baseline model on CANDELS images and transfered it to CLASH images. To train our baseline model, we used *HST* images from a CANDELS field taken with the WFC3 in the F160W band, namely GOODS-S[9] (Giavalisco et al. 2004). Notice that we are not using the following CANDELS fields: COSMOS (Scoville et al. 2007), UDS (Lawrence & Warren Almaini 2007) , and EGS (Davis et al. 2007), as the labels of KA15 come from GOOD-S. To these we added the mosaic from Hubble Legacy Fields (HLF) Data Release 1.5 for the GOODS-S region (HLF-GOODS-S) (R. Bouwens et al. 2019, in preparation), which combines exposures from the ACS/WFC and the WFC3 InfraRed Channel (WFC3/IR).

For both sets of images we used the 0.06″/pixel resolution version in the filter F160W, selecting galaxies with F160W magnitudes $H_{\rm mag} < 24.5$. This is the same limit used in KA15, who show that it corresponds to the flux limit for reliable visual morphological classifications. We created postage-stamp images from the GOODS-S mosaic setting the size to four times the Petrosian radius as reported in the catalog of Guo et al. (2013). We then performed a bi-lineal interpolation to set all images sizes to 80 × 80 pixels. We obtained a final sample of ∼7500 galaxies.

For the transfer learning sample we used images from the CLASH Multi-Cycle Treasury program (Postman et al. 2005). CLASH observed 25 clusters of galaxies at redshifts $0.15 < z < 0.9$ with WFC3 over a period of three years, in up to 16 filters, namely F225W, F275W, F336W, F390W, F435W, F475W, F606W, F625W, F775W, F814W, F850W, F105W, F110W, F125W F140W, and F160W, covering the ultraviolet, optical, and near-infrared regions of the spectrum. Molino et al. (2017) published accurate multiwavelength photometric catalogs for these clusters, which also provide the Petrosian radius. We created postage-stamp images for each filter separately following the same criterion for the magnitude cut and the size that we adopted for CANDELS, ending up with a sample of 68,531 galaxies.

3.2. Labels

In order to fine-tune and evaluate our model over CLASH data, we created a limited set of labels for CLASH images. We used a scheme similar to KA15 for the classification of galaxies in CANDELS, which considers spheroids, disks, point sources, irregulars, and unclassifiable sources (see Figure 1). We randomly selected 100 galaxies in each CLASH band (1600 in total) and classified them by eye. Each galaxy was labeled in each band by at least five of us on gray-scale images with 95% and 99.5% stretches that were uploaded on the Zooniverse web platform.[10] Notice each human annotator produced up to 16 labels per galaxy, not necessarily the same across bands, as our goal was to have a model that produces different classifications for each band. At the end of this process a probability for the galaxy to have a certain morphological type was assigned to each object. This probability is defined as

$$P_T = \frac{N_T}{N_{\rm tot}}, \tag{1}$$

where N_T is the number of people who assigned a type T to the galaxy and $N_{\rm tot}$ is the total number of people who classified that galaxy. We will call this labeled data set CL-eye hereafter.

4. Deep-learning Framework

Our model predicts the probability of each galaxy to be of each type, and it is based on the Inception model designed by Szegedy et al. (2014). We implemented it on the Keras deep learning library (Chollet et al. 2015) with a TensorFlow backend (Abadi et al. 2015). The more galaxies we use for training our ConvNet model the best performance it will achieve (Chu et al. 2016). Our goal is to have an accurate predicting model for morphology probabilities over CLASH single-filter images. However, since before the making of this catalog no labels were available for morphologies of galaxies in CLASH, we used the KA15 CANDELS morphology catalog and trained our model with this data set. We later fine-tuned the parameters of the model to a smaller subset of galaxies that we labeled for CLASH.

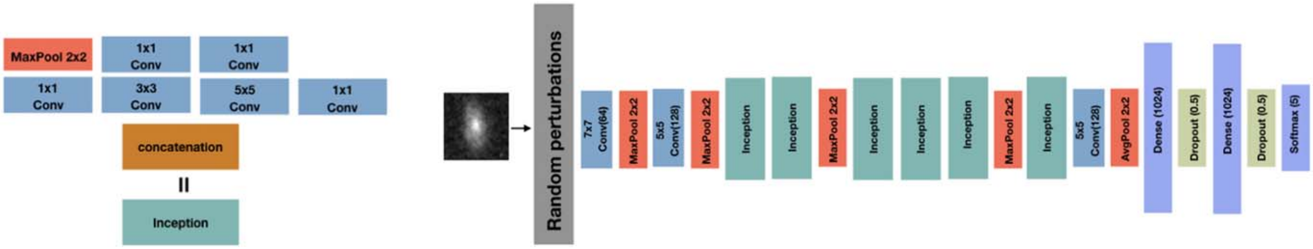
4.1. Architecture

In general, when a ConvNet architecture is used to predict image features, one of the most important hyperparameters is the filter size used to perform the convolutions. The inception model proposed by Szegedy et al. (2014) combines different convolutional filters into a single unit. It first makes a 1 × 1 convolution reducing the dimensionality of the feature space and then uses different filter sizes in the same layer paddings to maintain dimensions. Resulting operations are concatenated at the end of the inception module (see Figure 2(a)).

The model that we used is composed of six inception modules that map images of galaxies into a set of 1024 features grouped in

[9] The Great Observatories Origins Deeps Survey.

[10] https://www.zooniverse.org



(a) Inception module (b) ConvNet architecture

Figure 2. (a) Example of an inception module composed of three convolutions with different sizes and a maxpooling operation; (b) ConvNet architecture based on Inception (Szegedy et al. 2014) used in this work to train a model over the ~7500 galaxies from the GOOD-S field of CANDELS catalog of KA15 and fine-tuned over CLASH data. The algorithm starts with images of size 80 × 80 pixels which are randomly perturbed during training and convolving over six inception modules as showed, getting a reduced set of features in the final stack. After that, the low-level features are passed through a neural network with two hidden layers in order to predict the value for every possible type of galaxy using a softmax activation function.

(A color version of this figure is available in the online journal.)

Table 1
Architecture of Our Model. This Table Follows the Same Order as Figure 2(b)

Operation	Filter Size/ Depth	Output Size	1 × 1 Before 3 × 3	3 × 3	1 × 1 Before 5 × 5	5 × 5	1 × 1	1 × 1 After Maxpooling
convolution	7 × 7/64	74 × 74 × 64						
max pool		37 × 37 × 64						
convolution	6 × 6/128	32 × 32 × 128						
max pool		16 × 16 × 128						
inception		16 × 16 × 256	96	128	16	16	64	32
inception		16 × 16 × 480	128	192	32	96	128	64
max pool		8 × 8 × 480						
inception		8 × 8 × 512	96	208	16	48	192	64
inception		8 × 8 × 512	112	224	24	64	128	64
inception		8 × 8 × 512	144	288	32	64	128	64
max pool		4 × 4 × 512						
inception		4 × 4 × 528	144	288	32	64	112	64
avg pool	2 × 2	2 × 2 × 528						
convolution	1 × 1/256	2 × 2 × 256						
flatten		1 × 1 × 1024						
dense		1 × 1 × 1024						
dropout (0.5)		1 × 1 × 5						
dense		1 × 1 × 1024						
dropout (0.5)		1 × 1 × 5						
softmax		1 × 1 × 5						

the final convolutional stack. We use batch-normalization (Ioffe & Szegedy 2015) in all layers before the fourth inception unit. These features are passed through three dense layers of 1,024, 1,024, and five neurons that represent every possible value that the model can take for every type of galaxy using a softmax activation function (see Figure 2(b)). Notice that our model has ~5.5 million parameters to be fitted. Details about the hyperparameters used in this architecture are shown in Table 1.

4.2. Training Strategy

Taking advantage of the rotation-invariant property of the galaxies (see D15), during training time, in each epoch we artificially augmented the number of images in the training set four times by applying random perturbations:

- Rotations: random rotations are performed by sampling the rotation angle from a uniform probability distribution with values between 0 and 360;
- Flipping: the images are flipped horizontally and vertically with a probability of 0.5 each;
- Translation: translations are performed by sampling two values from a uniform probability distribution on the interval [−4, 4], representing the number of pixels to be translated in the x and y coordinates of the image.

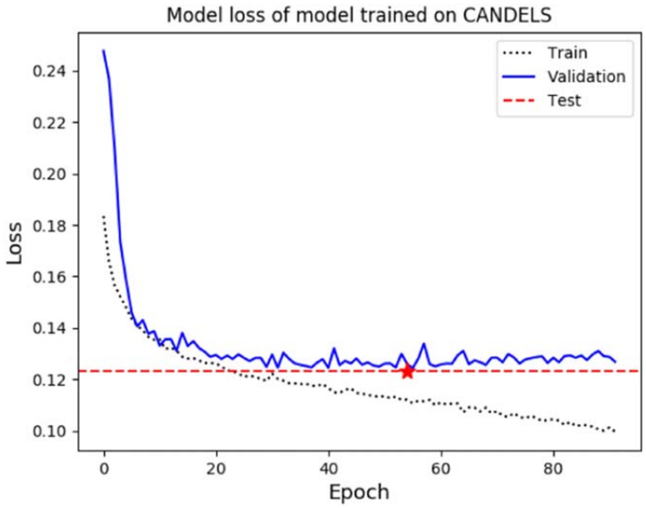


Figure 3. Learning curve of the model trained and evaluated using CANDELS data. The black dotted line shows the loss over the training set, the blue solid line shows the loss over the validation set, and the dashed red line shows the final test set loss over the best validation set model. A red star is shown at the epoch this model was obtained. The loss function shown is the RMSE between KA15 labels and predictions given by the model.

(A color version of this figure is available in the online journal.)

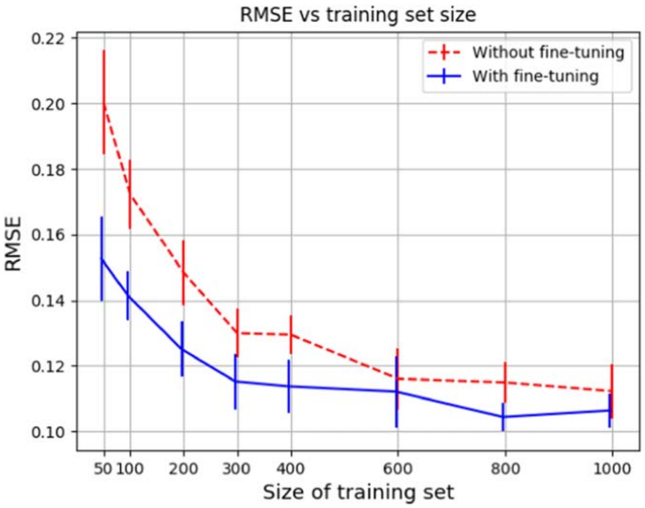


Figure 4. RMSE loss in terms of the number of galaxies in the training set. Error bars show the RMSE standard deviation for 10 random train–validation–test cross-validation splits.

(A color version of this figure is available in the online journal.)

With these perturbations it is unlikely that our model sees exactly the same image twice while training. This data augmentation helps avoiding overfitting.

We perform validation by dividing the sample in three subsets without overlapping: a training set, a validation set, and a test set. We use the training set to backpropagate the gradients and evaluate using the validation set (without random perturbations) at training time in order to test for convergence. Finally, we report the test error measured by evaluating the fully trained model over the test set (without random perturbations). This way, we have a final evaluation of our model over data never used at training time.

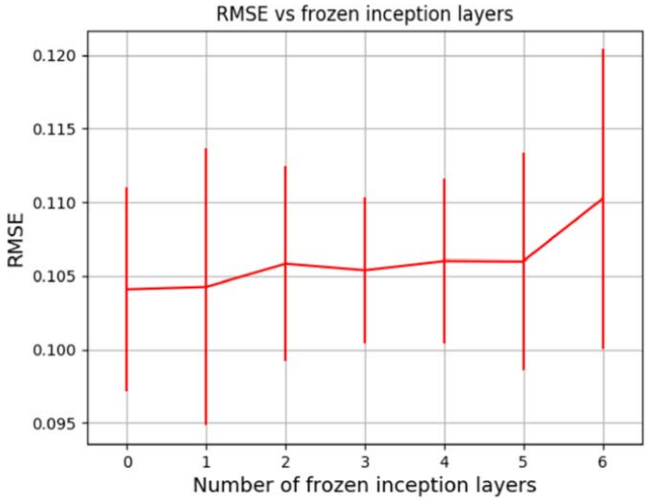


Figure 5. RMSE of our model in terms of the inception layers frozen for fine-tuning after transferring the model trained on CANDELS data. To obtain the error we freeze everything before the inception layer is indexed in the *x*-axis and fine-tuned over the layers after. We trained 20 models using a training set of 1000 galaxies, and 150 galaxies for validating and testing. Error bars show the standard deviation of the RMSE over these 20 cross-validation experiments.

(A color version of this figure is available in the online journal.)

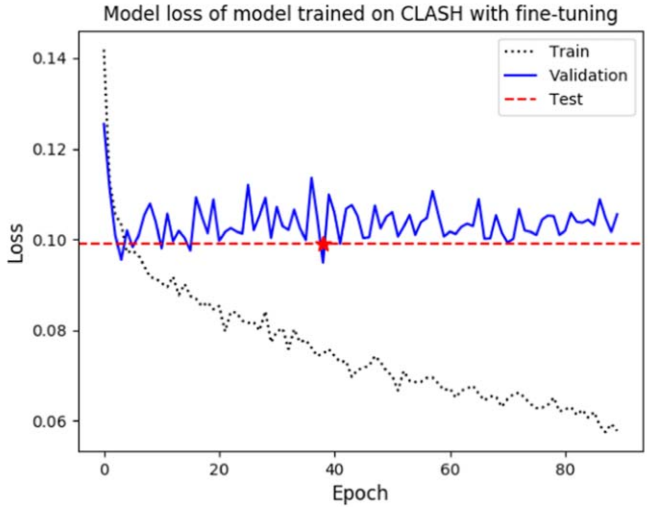


Figure 6. Learning curve of the trained CLASH using prior knowledge from CANDELS. The best model was chosen using the model in which the error on the validation set was lower. The black dotted line shows the loss over the training set, the blue solid line shows the loss over the validation set, and the dashed red line shows the final test set loss over the best validation set model. A red star is shown at the epoch this model was obtained.

(A color version of this figure is available in the online journal.)

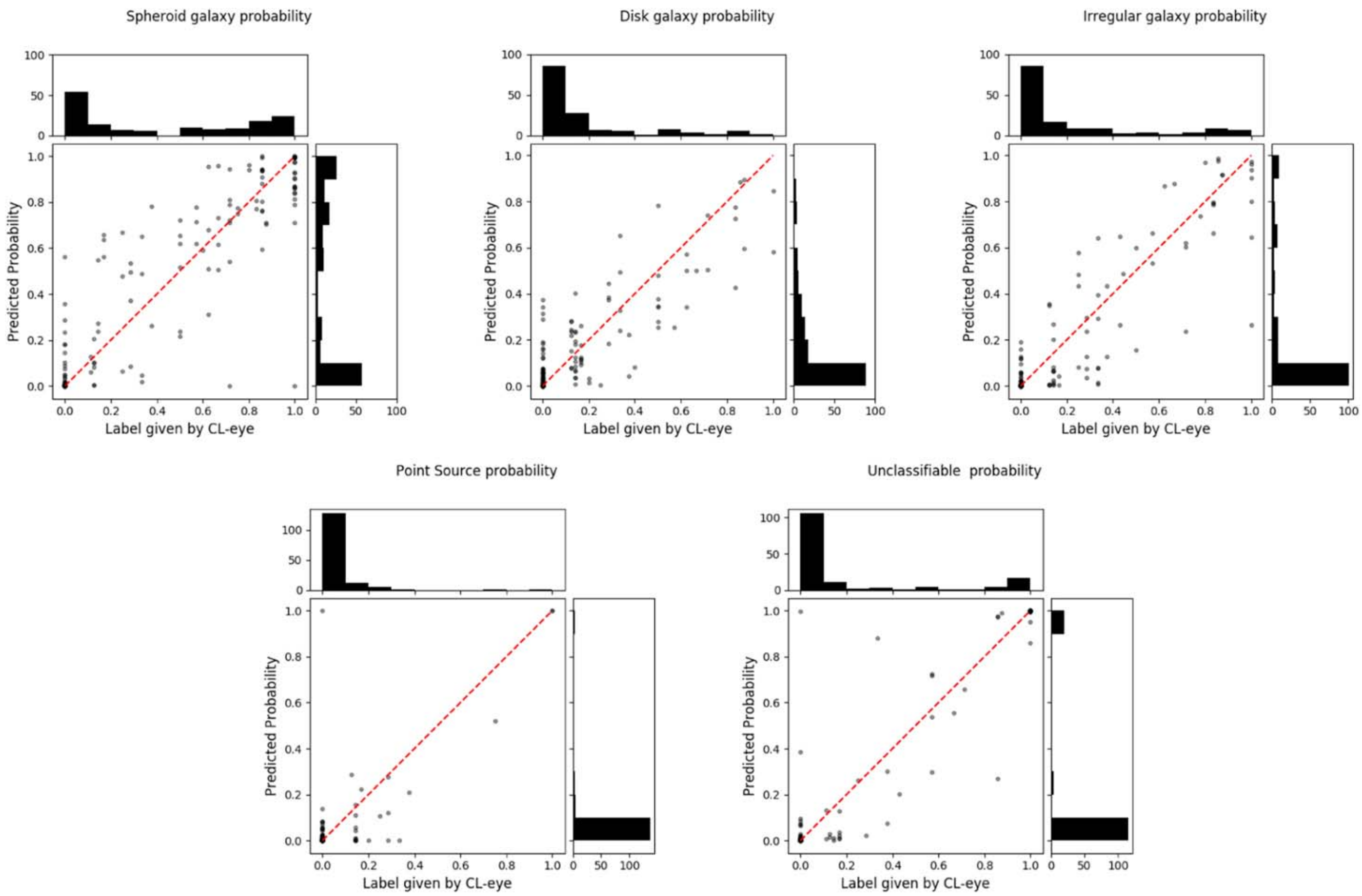


Figure 7. Predicted probabilities from our model vs. CL-eye visual labels. (A color version of this figure is available in the online journal.)

In order to compute the weights of the architecture used to predict the labels, we use the Adam method for stochastic optimization (Kingma & Ba 2014). We tried different values for the learning rate and obtained the best results in terms of cross-validation RMSE using a learning rate of 5×10^{-5}.[11] We used initial values for β_1 and β_2 of 0.9 and 0.999 respectively. In addition, the model was trained using cross-entropy as loss function, achieving statistically similar results to using RMSE in terms of test evaluation.

5. Performance and Discussion

We trained the model using $\sim$7500 galaxies from the CANDELS GOODS-S catalog of KA15 in order to acquire prior knowledge to train the model with CLASH data. Our work was divided in two steps, namely prior knowledge acquisition from CANDELS and fine-tuning over CLASH. In this section, we explain our methodology in detail.

5.1. Prior Knowledge Acquisition from CANDELS

We trained the ConvNet model described in Figure 2(b) using CANDELS data and labels from KA15. We split the data into 70% for training, 15% for validation, and 15% for testing the final model. We used 160 mini-batches of 128 images per epoch, in which each image was randomly sampled with replacement, and we applied to it the perturbations described in Section 4.2. The evaluation of the model was made in real time using images from the validation set after each epoch. We backpropagated the gradients on our training set as long as the validation test loss diminished within 35 epochs.

The learning curve of the model trained over CANDELS is shown in Figure 3. We obtain an RMSE of $\sim$0.123 on the test set, consistent with results from Huertas-Company et al. (2015). This model is used as prior knowledge to train a model over CLASH data.

[11] We evaluated for learning rates of [1×10^{-5}, 5×10^{-5}, 1×10^{-4}, 1×10^{-4}, 1×10^{-3}].

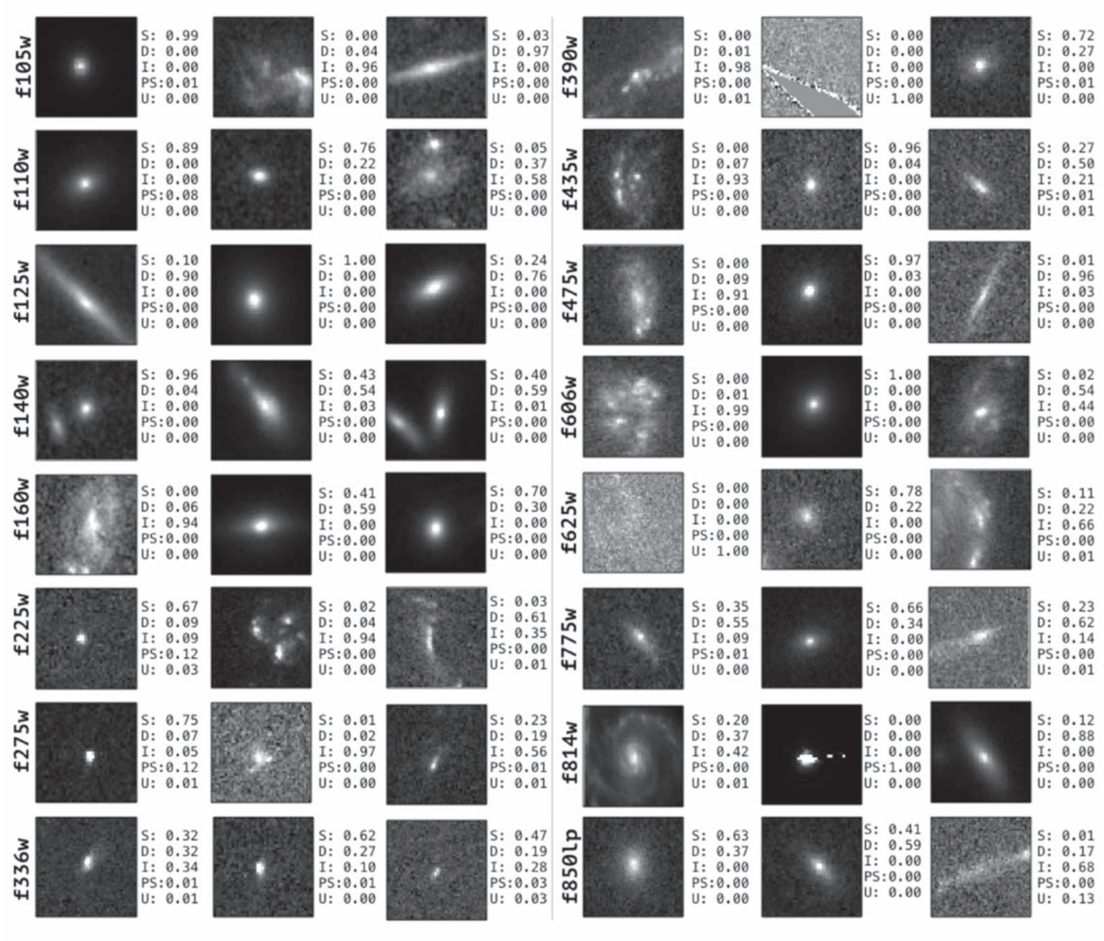

f105w
S: 0.99 D: 0.00 I: 0.00 PS:0.01 U: 0.00
S: 0.00 D: 0.04 I: 0.96 PS:0.00 U: 0.00
S: 0.03 D: 0.97 I: 0.00 PS:0.00 U: 0.00
f110w
S: 0.89 D: 0.00 I: 0.00 PS:0.08 U: 0.00
S: 0.76 D: 0.22 I: 0.00 PS:0.00 U: 0.00
S: 0.05 D: 0.37 I: 0.58 PS:0.00 U: 0.00
f125w
S: 0.10 D: 0.90 I: 0.00 PS:0.00 U: 0.00
S: 1.00 D: 0.00 I: 0.00 PS:0.00 U: 0.00
S: 0.24 D: 0.76 I: 0.00 PS:0.00 U: 0.00
f140w
S: 0.96 D: 0.04 I: 0.00 PS:0.00 U: 0.00
S: 0.43 D: 0.54 I: 0.03 PS:0.00 U: 0.00
S: 0.40 D: 0.59 I: 0.01 PS:0.00 U: 0.00
f160w
S: 0.00 D: 0.06 I: 0.94 PS:0.00 U: 0.00
S: 0.41 D: 0.59 I: 0.00 PS:0.00 U: 0.00
S: 0.70 D: 0.30 I: 0.00 PS:0.00 U: 0.00
f225w
S: 0.67 D: 0.09 I: 0.09 PS:0.12 U: 0.03
S: 0.02 D: 0.04 I: 0.94 PS:0.00 U: 0.00
S: 0.03 D: 0.61 I: 0.35 PS:0.00 U: 0.01
f275w
S: 0.75 D: 0.07 I: 0.05 PS:0.12 U: 0.01
S: 0.01 D: 0.02 I: 0.97 PS:0.00 U: 0.00
S: 0.23 D: 0.19 I: 0.56 PS:0.01 U: 0.01
f336w
S: 0.32 D: 0.32 I: 0.34 PS:0.01 U: 0.01
S: 0.62 D: 0.27 I: 0.10 PS:0.01 U: 0.00
S: 0.47 D: 0.19 I: 0.28 PS:0.03 U: 0.03
f390w
S: 0.00 D: 0.01 I: 0.98 PS:0.00 U: 0.01
S: 0.00 D: 0.00 I: 0.00 PS:0.00 U: 1.00
S: 0.72 D: 0.27 I: 0.00 PS:0.01 U: 0.00
f435w
S: 0.00 D: 0.07 I: 0.93 PS:0.00 U: 0.00
S: 0.96 D: 0.04 I: 0.00 PS:0.00 U: 0.00
S: 0.27 D: 0.50 I: 0.21 PS:0.01 U: 0.01
f475w
S: 0.00 D: 0.09 I: 0.91 PS:0.00 U: 0.00
S: 0.97 D: 0.03 I: 0.00 PS:0.00 U: 0.00
S: 0.01 D: 0.96 I: 0.03 PS:0.00 U: 0.00
f606w
S: 0.00 D: 0.01 I: 0.99 PS:0.00 U: 0.00
S: 1.00 D: 0.00 I: 0.00 PS:0.00 U: 0.00
S: 0.02 D: 0.54 I: 0.44 PS:0.00 U: 0.00
f625w
S: 0.00 D: 0.00 I: 0.00 PS:0.00 U: 1.00
S: 0.78 D: 0.22 I: 0.00 PS:0.00 U: 0.00
S: 0.11 D: 0.22 I: 0.66 PS:0.00 U: 0.01
f775w
S: 0.35 D: 0.55 I: 0.09 PS:0.01 U: 0.00
S: 0.66 D: 0.34 I: 0.00 PS:0.00 U: 0.00
S: 0.23 D: 0.62 I: 0.14 PS:0.00 U: 0.01
f814w
S: 0.20 D: 0.37 I: 0.42 PS:0.00 U: 0.01
S: 0.00 D: 0.00 I: 0.00 PS:1.00 U: 0.00
S: 0.12 D: 0.88 I: 0.00 PS:0.00 U: 0.00
f850lp
S: 0.63 D: 0.37 I: 0.00 PS:0.00 U: 0.00
S: 0.41 D: 0.59 I: 0.00 PS:0.00 U: 0.00
S: 0.01 D: 0.17 I: 0.68 PS:0.00 U: 0.13

Figure 8. Images and labels of galaxies from CLASH as labeled by our model for different filters.

5.2. Fine-tuning over CLASH

After training the model over the CANDELS data using the labels from KA15, we fine-tuned it to the CLASH data using the labels that we generated. In order to assess the need of fine-tunning, we also evaluated our model trained directly on CLASH labels.

5.2.1. Training CLASH without Fine-tuning

We started by training our model directly on CL-eye data. We used a validation set of 150 galaxies and a test set of 150 galaxies. Some galaxies in CL-eye were labeled more than once on different filters. The same galaxy was never present in more than one of these subsets. Each training epoch contained 30 mini-batches of 128 images per epoch. Recall each image in the training mini-batches was randomly sampled with replacements, and random perturbations were applied to them.

In order to assess the amount of labeled data needed to train our models, we trained different models using training sets of 50, 100, 200, 300, 400, 600, 800, and 1000 images. The results of 10 cross-validation runs for each training size are shown in Figure 4. As expected, as the training size increases, the test set loss goes down. When using more than 600 objects for training, there is no statistically significant improvement.

5.2.2. Training CLASH with Fine-tuning

Here, we explore the advantage of using CANDELS labeled data (KA15) to train a model and use the obtained parameters as initial conditions for further training using the CL-eye labeled data set.

We started by training our models using KA15 and fine-tuning some parameters using the CL-eye data set. We directly transfered all the parameters from the model trained on KA15 (shown in Figure 3) and we trained models with CLASH over copied parameters. We evaluated freezing layers before different inception units. Figure 5 shows the effect of freezing different numbers of layers. Here, fine-tuning was performed using 1000 galaxies in the training set, and 150 for both the validation and test sets. It can be seen that there is no statistically significant

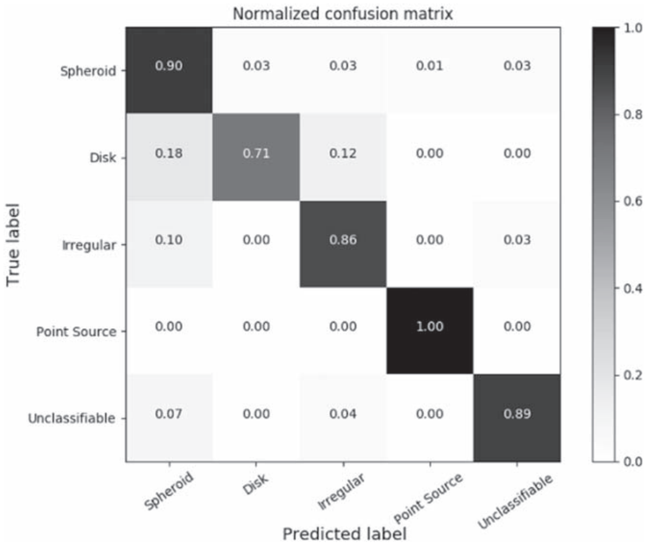


Figure 9. Confusion matrix computed on the test set using the model in Figure 6, utilizing a MAP decision rule for the probability of being spheroid, disk, irregular, point source, or unclassifiable given by the model in the *x*-axis and given by labels of CL-eye in the *y*-axis.

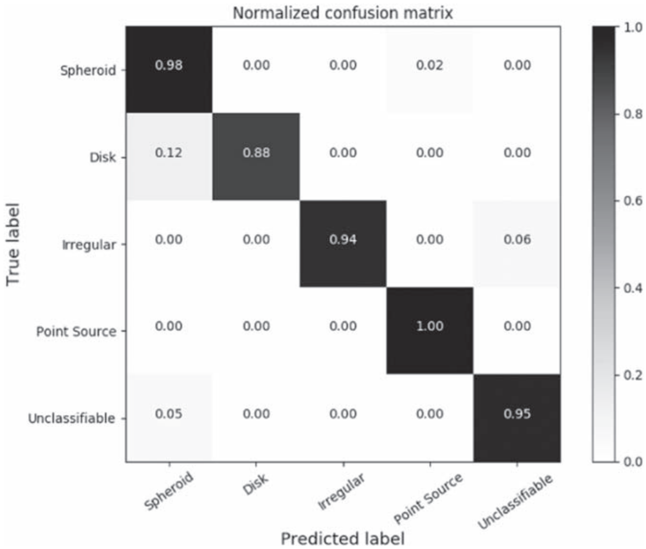


Figure 10. Confusion matrix computed on the test set using the model in Figure 6, utilizing a threshold of 0.75 on the probability of being spheroid, disk, irregular, point source, or unclassifiable given by the model in the *x*-axis and given by labels of CL-eye in the *y*-axis.

difference on the number of layers frozen. We decided to not freeze any layer, i.e., we used the model trained on KA15 as an initial condition to the model trained over CL-eye, backpropagating the error through all layers. Hereafter, all models are trained without freezing any layer.

Figure 4 shows the effect of the training set size used for fine-tuning over the model trained using CANDELS images with KA15 labels. As the training sample size increases, the

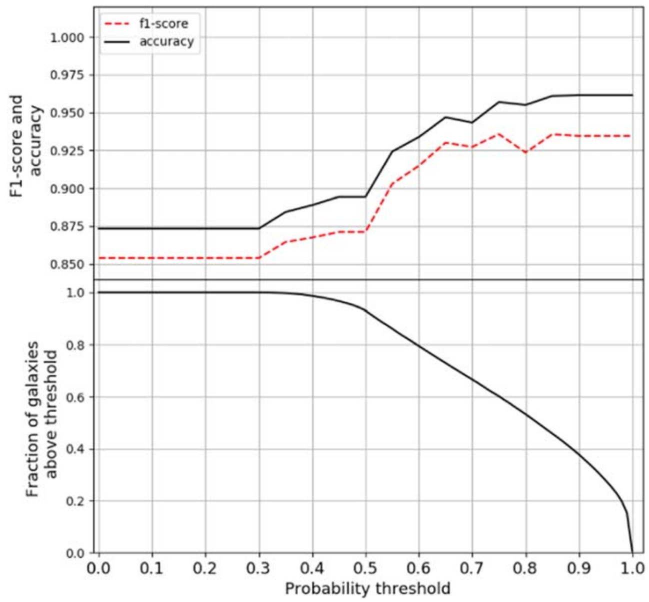


Figure 11. Accuracy, macro f1-score, and fraction of galaxies in terms of the probability threshold over our model's predictions used to define a class. As we increase the probability threshold, the accuracy and f1 score increase, and the number of galaxies with a single probability over this threshold diminishes.
(A color version of this figure is available in the online journal.)

RMSE goes down. Furthermore, the fine-tuned model outperforms the model trained over CL-eye labels with no fine-tuning. Using a training set of more than 300 galaxies does not achieve a significant improvement.

Figure 6 shows the learning curve for the model used to produce the online catalog made publicly available with this paper. This model was the one that achieved the best RMSE over the validation data set using a training set size of 1000 galaxies from CL-eye fine-tuned data and without freezing. This model obtained a validation RMSE of 0.0949 and a test RMSE of 0.0991. In Figure 7 we can see predicted probability values for 150 galaxies in the test set versus labels given by CL-eye. A random subset of galaxies labeled by our model in different filters is shown in Figure 8. In this image "S" represents spheroid, "D" disk, "I" irregular, "PS" point source, and "U" unclassifiable.

5.3. Using ConvNet Predictions for Classification

An important use of our catalog is creating labels following a classification scheme. Here, we evaluate how our model performs using maximum a posteriori (MAP) decision rule (i.e., choosing the class with the highest probability given by our model), or choosing a threshold on the probabilities.

Figure 9 shows the confusion matrix of our model following a MAP decision rule classification scheme. Here, we assign to each object the class that has the highest probability. The worst performance occurs when our model confuses ~24% of the disk galaxies with spheroids. Using this scheme, we correctly

Table 2
Visual-like Morphology Catalog of Galaxies Over CLASH Produced by the ConvNet Trained in CLASH Sample with Fine Tuning

Cluster	ID	R.A.	DECL.	F225w_PSP	F225w_PD	F225w_PI	⋯	F160w_PPS	F160w_PUN
ms2137	287	325.0698	−23.6508	0.00	0.01	0.77	⋯	0.99	0.00
macs1931	1708	292.9468	−26.5854	0.13	0.12	0.28	⋯	0.99	0.00
rxj1347	16	206.8822	−11.7663	⋯	⋯	⋯	⋯	0.00	0.00
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
rxj2248	43	342.1780	−44.5120	0.00	0.00	0.00	⋯	0.00	0.06

Note. Full version available in http://www.inf.udec.cl/~guille/data/Deep-CLASH.csv.

label 91% of spheroids, 65% of disks, 73% of irregulars, 100% of point sources, and 89% of unclassifiable sources. Huertas-Company et al. (2015) propose to select a threshold of 0.75 on the probabilities in order to have a higher certainty on the labels. This is something important to consider as human-assigned labels are not perfect. We explore this idea by evaluating our model against galaxies with high certainty on their human labels. Figure 10 shows the confusion matrix using a threshold of 0.75 in the human label probabilities to define the classes. In this experiment our model correctly labels 98% of spheroids, 100% of disks, 94% of irregulars, 100% of point sources, and 100% of the unclassifiable sources. We further explore these results by assessing the performance of our model in terms of the probability threshold used to classify sources. Figure 11 shows the accuracy and macro f1-score in terms of the probability threshold used to decide the class of each object for our test set. Galaxies with a probability lower than this threshold for all classes are not considered. As expected, as the threshold grows, the performance of our model tends to improve. There is a collateral effect to choosing this threshold: as this value increases, we are considering fewer galaxies. Figure 11 shows the impact of increasing the probability threshold on the number of selected galaxies. We keep around 80% of the galaxies for a threshold of 0.60 on the probability, corresponding to an f1-score of approximately 0.91 and an accuracy of 0.94. If we use a threshold on the probabilities of 0.90, we keep approximately 40% of the galaxies, but the f1-score and accuracy are higher than 0.93.

6. Catalog

This paper is accompanied by a visual-like catalog containing morphology probability predictions for 68,531 images of 8,532 galaxies in 16 *HST* photometric bands for galaxies in the 25 CLASH fields. Morphologies correspond to the probabilities predicted by our ConvNet model. We give the probability for each galaxy of being disk, spheroid, point source/compact, peculiar/irregular, or unclassifiable in each of the 16 different *HST* photometric bands (F225W, F275W, F336W, F390W, F435W, F475W, F606W, F625W, F775W, F814W, F850LP, F105W, F110W, F125W, F140W, F160W), with their corresponding labels, cluster name, right ascension, and declination for each image where the galaxy appears. We provide a sample table in Table 2, while the full table in machine-readable format is available at http://www.inf.udec.cl/~guille/data/Deep-CLASH.csv. The catalog provides the following information:

Cluster: cluster name to which a galaxy belongs;
ID: ID of the galaxy as reported in Molino et al.'s catalog;
R.A.: right ascension of the galaxy (J2000);
DECL.: declination of the galaxy (J2000);
F225w_PSP: probability for an object to be a spheroidal galaxy in the indicated band (from F225W to F160W);
F225w_PD: probability for an object to be a disk galaxy in the indicated band (from F225W to F160W);
F225w_PI: probability for an object to be an irregular galaxy in the indicated band (from F225W to F160W);
F225w_PPS: probability for an object to be a point source in the indicated band (from F225W to F160W);
F225w_PUN: probability for an object to be unclassifiable in the indicated band (from F225W to F160W);
...

7. Summary and Conclusions

This paper presents a visual-like morphological classification of 8532 galaxies in 16 *HST* photometric bands of CLASH, one morphology per filter, for a total of 68,531 labeled images. The catalog contains probabilities of being spheroid, disk, irregular, point source, or unclassifiable, predicted using a convolutional neural network (ConvNet) model. The model[12] was trained using ∼7500 CANDELS images and further fine-tuned using CLASH images with eyeball labels. We show that, by using more than 300 CLASH labeled images to train the model, it does not improve significantly in terms of cross-validated RMSE. We evaluate how the model performs on classification tasks in terms of the probability threshold used to define the class. The higher the threshold, the better performance of the model in terms of f1-score and accuracy, but the fewer galaxies we retain. In that sense, there is a trade-off between the certainty on the labels and the number of galaxies that can be used, which has to be taken into account when using the

[12] A demo on how to use our model to predict galaxy morphologies in CLASH is publicly available at https://github.com/mperezcarrasco/CLASH.

catalog. We believe that the fine-tuning techniques described in this work are essential to label data from new instruments such as the Large Synoptic Survey Telescope and EUCLID when new data are available with a limited amount of labeled data.

This catalog constitutes a complement to the automated morphology in CANDELS since it provides morphologies for galaxies in clusters. The comparison between cluster and field galaxies is fundamental to study the effects of the environment in the morphological transformation of galaxies (e.g., formation of early-type galaxies) and its relationships with the evolution of star formation and stellar mass build-up. Furthermore, the catalog is the first multi-wavelength morphological catalog that can be employed in the search of morphologically peculiar galaxies, which are likely undergoing interactions with their surrounding environment. It is therefore an important resource for the study of galaxy interactions.

Support for this work was provided by NASA through grant number from the Space Telescope Science Institute, which is operated by AURA, Inc., under NASA contract NAS 5-26555. This work is based on observations taken by the CANDELS Multi-Cycle Treasury Program with the NASA/ESA *HST*, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. G.C.V. acknowledges the support provided by FONDECYT postdoctoral research grant no 3160747; CONICYT-Chile and NSF through the Programme of International Cooperation project DPI201400090; the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC120009, awarded to The Millennium Institute of Astrophysics (MAS). P.C. acknowledges the support provided by FONDECYT postdoctoral research grant no 3160375. R.D. gratefully acknowledges the support provided by the BASAL Center for Astrophysics and Associated Technologies (CATA). We acknowledge Cristóbal Donoso-Oliva, Hugo Parischewsky-Zapata and Daniela Olave-Rojas who helped label CLASH images.

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/

Abraham, R. G., Tanvir, N. R., Santiago, B. X., et al. 1996, MNRAS, 279, L47

Aniyan, A. K., & Thorat, K. 2017, ApJS, 230, 20

Butcher, H., & Oemler, A., Jr. 1978, ApJ, 219, 18

Cabrera-Vives, G., Reyes, I., Förster, F., Estévez, P. A., & Maureira, J.-C. 2017, ApJ, 836, 97

Cerulo, P., Couch, W. J., Lidman, C., et al. 2017, MNRAS, 472, 254

Chollet, F. 2015, keras, https://github.com/fchollet/keras

Chu, B., Madhavan, V., Beijbom, O., Hoffman, J., & Darrell, T. 2016, in ECCV Workshops

Connor, L., & van Leeuwen, J. 2018, (arXiv:1803.03084)

Conselice, C. J., Bershady, M. A., & Jangren, A. 2000, ApJ, 529, 886

Couch, W. J., Barger, A. J., Smail, I., Ellis, R. S., & Sharples, R. M. 1998, ApJ, 497, 188

Davis, M., Guhathakurta, P., Konidaris, N. P., et al. 2007, ApJL, 660, L1

de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H. G., et al. 1991, Third Reference Catalogue of Bright Galaxies. Volume I: Explanations and references. Volume II: Data for galaxies between 0^h and 12^h. Volume III: Data for galaxies between 12^h and 24^h

Desai, V., Dalcanton, J. J., Aragón-Salamanca, A., et al. 2007, ApJ, 660, 1151

Dieleman, S., Willet, K., & Dambre, J. 2015, MNRAS, 450, 1441

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, MNRAS, 476, 3661

Dressler, A. 1980, ApJ, 236, 351

Dressler, A., Oemler, A., Jr., Couch, W. J., et al. 1997, ApJ, 490, 577

Fasano, G., Poggianti, B., Couch, W., et al. 2001, ApSSS, 277, 417

Fukushima, K. 1980, Biol. Cybern., 36, 193

Giavalisco, M., Ferguson, H. C., Koekemoer, A. M., et al. 2004, ApJL, 600, L93

Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35

Guo, P., Duan, F., Wang, P., Yao, Y., & Xin, X. 2017, (arXiv:1711.10339)

Guo, Y., Ferguson, H. C., Giavalisco, M., et al. 2013, ApJS, 207, 24

He, K., Zhang, X., Ren, S., & Sun, J. 2015, CoRR, arXiv:1512.03385

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2012, CoRR, arXiv:1207.0580

Holden, J., Shotbolt, L., Bonn, A., et al. 2007, ESRv, 82, 75

Hubble, E. P. 1936, Realm of the Nebulae

Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S., & Sánchez Almeida, J. 2011, A&A, 525, A157

Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8

Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971

Ioffe, S., & Szegedy, C. 2015, CoRR, arXiv:1502.03167

Kartaltepe, J. S., Mozena, M., Kocevski, D., et al. 2015, ApJS, 221, 11

Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 54

Kauffmann, G., White, S. D. M., Heckman, T. M., et al. 2004, MNRAS, 353, 713

Kingma, D. P., & Ba, J. 2014, CoRR, arXiv:1412.6980

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 1, NIPS'12 (USA: Curran Associates Inc.) 1097–105, http://dl.acm.org/citation.cfm?id=2999134.2999257

Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, MNRAS, 379, 1599

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proc. IEEE, 86 2278

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179

Lotz, J. M., Davis, M., Faber, S. M., et al. 2008, ApJ, 672, 177

Hilton, M., Stanford, S. A., Stott, J. P., et al. 2009, ApJ, 697, 436

Mei, S., Stanford, A., Holden, B. P., et al. 2012, ApJ, 754, 141

Molino, A., Benítez, N., Ascaso, B., et al. 2017, MNRAS, 470, 95

Muzzin, A., Wilson, G., Yee, H. K. C., et al. 2012, ApJ, 746, 188

Nagi, J., Ducatelle, F., Caro, G. A. D., et al. 2011, 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 342

Nair, P. B., & Abraham, R. G. 2010, ApJS, 186, 427

Nair, V., & Hinton, G. E. 2010, Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10 (USA: Omnipress), 807–4, http://dl.acm.org/citation.cfm?id=3104322.3104425

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. 2014, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 1717

Pan, S. J., & Yang, Q. 2010, IEEE Trans. Knowl. Data Eng., 22, 1345

Peng, Y.-j., Lilly, S. J., Kovač, K., et al. 2010, ApJ, 721, 193

Petrillo, C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129

Postman, M., Coe, D., Benítez, N., et al. 2012, ApJS, 199, 25

Postman, M., Franx, M., Cross, N. J. G., et al. 2005, ApJ, 623, 721

Ruder, S. 2016, arXiv:1609.04747

Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1

Shallue, C. J., & Vanderburg, A. 2018, AJ, 155, 94

Simmons, B. D., Lintott, C., Willett, K. W., et al. 2017, MNRAS, 464 4420
Simonyan, K., & Zisserman, A. 2014, CoRR, arXiv:1409.1556
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014, Journal of Machine Learning Research, 15, 1929
Szegedy, C., Liu, W., Jia, Y., et al. 2014, CoRR, arXiv:1409.4842
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579
Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. 2014, CoRR, arXiv:1411.1792
Yosinski, J., Clune, J., Nguyen, A. M., Fuchs, T. J., & Lipson, H. 2015, CoRR, arXiv:1506.06579
Zeiler, M. D., & Fergus, R. 2013, CoRR, arXiv:1311.2901
Zhang, G. P. 2000, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30, 451