Comparing simulations and observations with deep generative models

Lorenzo Zanisi[1,2]★, Marc Huertas-Company[2,3], François Lanusse[4], Connor Bottrell[5], Joel Primack[6]

[1] *Department of Physics and Astronomy, University of Southampton, Highfield, SO17 1BJ, UK*
[2] *Instituto de Astrofísica de Canarias (IAC); Departamento de Astrofísica, Universidad de La Laguna (ULL), E-38200, La Laguna, Spain*
[3] *LERMA, Observatoire de Paris, CNRS, PSL, Université Paris Diderot, France*
[4] *Berkeley Center for Cosmological Physics, Department of Physics, University of California, Berkeley, California, USA*
[5] *Department of Physics and Astronomy, University of Victoria, Victoria, British Columbia V8P 1A1, Canada*
[6] *Department of Physics, University of California at Santa Cruz, Santa Cruz, CA 95064, USA*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We use PixelCNN, an autoregressive model for image generation with an explicit, probabilistically interpretable likelihood, to assess the capabilities of state-of-the-art hydrodynamical cosmological simulations of galaxy formation and evolution in reproducing the optical morphologies of a local galaxy sample extracted from the Sloan Digital Sky Survey. As a proof of concept, we apply such framework to mock observations of the Illustris Project and the Illustris TNG Project.

We find that PixelCNN is able to assess the widely accepted improvement of Illustris TNG with respect to the previous Illustris run. PixelCNN can also identify the simulated galaxies whose morphologies are not realistic. We dissect the correlations between likelihood and galaxy properties in SDSS, finding that larger, more irregular galaxies tend to have lower values of likelihood. We also find that having realistic background in simulated images is fundamental to our purpose. To conclude, we outline potential improvements to the current framework and discuss its applications.

1 INTRODUCTION

In the recent years, cosmological hydrodynamical simulations of galaxy formation and evolution have reached unprecedented accuracy. Early efforts (e.g. Croft et al. 2009, Crain et al. 2009, Schaye et al. 2010, Nuza et al. 2010, Di Matteo et al. 2012, Vogelsberger et al. 2014) have paved the way to state-of-the art simulations (Schaye et al. 2015, Dubois et al. 2014,Davé et al. 2019, Pillepich et al. 2018a), which broadly agree with a number of observations (only to mention a few, Genel et al. 2018, Donnari et al. 2019, Pillepich et al. 2018b, Rodriguez-Gomez et al. 2019, Trayford et al. 2017, Furlong et al. 2015, Bignone et al. 2019, Dubois et al. 2016, Kaviraj et al. 2017). As a matter of fact, however, assessing the level of agreement between the morphologies of the full populations of observed and simulated galaxies is a harder task. Some authors (Rodriguez-Gomez et al. 2019, Bignone et al. 2019) made use of integrated, parametric and nonparametric quantities as diagnostics (such as the popular $C - A - S - G - M_{20}$ statistics, Conselice 2003, Lotz et al. 2004), with the aim of describing galaxy morphologies with only a few numbers. Such an approach may still not grasp the full complexity of a galaxy image. In fact, although technically all the pixels of a galaxy image are used to retrieve these quantities, their choice may be incomplete (i.e. the $C - A - S - G - M_{20}$ spatial diagnostics may in principle be extended, see for instance Freeman et al. 2013, Wen et al. 2014, Pawlik et al. 2016, Rodriguez-Gomez et al. 2019), and, for this reason, limited in power (i.e. the similarity of these statistics between observed and simulated galaxies, although informative, is no guarantee of the overall quality of simulated galaxy properties). The key point is that all the precious information contained in the pixels of an image may not be fully accessible with standard techniques, which may be a major shortcoming when comparing the morphologies of observed and simulated galaxies. Moreover, the different statistics provide separate pieces of information, while it would be desirable to be able to assess the quality of a simulation using a single-valued metric, something that has not yet been done in previous work.

An alternative approach is using Neural Networks. In particular Convolutional Neural Networks (CNNs) are able to learn the complex, fine grained structure of an image, since they use the information contained in pixels much more efficiently without requiring any explicit choice of spatial diagnostics or any simplified fit to the light profile. Therefore, they constitute a more general framework than that provided by (non) parametric diagnostic tools. CNNs are now being extensively used for image recognition tasks (He et al. 2015), and have also been successfully applied to astronomy in a number of works. For example, CNNs have been employed to classify the morphology of galaxies in large galaxy surveys (Domínguez Sánchez et al. 2018, Dieleman et al. 2015, Huertas-Company et al. 2015) as well as the well known Fanaroff & Riley (1974) morphological dichotomy of radio jets of Active Galactic Nuclei (Lukic et al. 2019). Walmsley et al. (2019) have used CNNs to identify faint tidal features in galaxies from the CFHTLS-Wide Survey, while Dimauro et al. (2018) and Tuccillo et al. (2018) have proposed the use of a CNN to improve the fits to the light profiles of galaxies. In Huertas-Company et al. (2019) a CNN was trained on images from Nair & Abraham (2010) (where galaxies were assigned labels in the form of TType by means of eyeball classification by the authors) and then applied to both the Sloan Digital Sky Survey (SDSS, Abazajian et al. 2009, Meert et al. 2015) and the Illustris TNG simulation (Nelson et al. 2019).

All these works have used CNNs under the assumption that the training and the test data come from the same set. This is a *critical assumption*, which is not necessarily true in the case where training is performed on observations but then the CNN is applied to simulations, since we do not know a priori whether simulations agree with observations. In fact, a test image will always be assigned a class by the CNN, even though it looks nothing like any of the images in the training set. In Huertas-Company et al. (2019) this issue was addressed by using Monte Carlo Dropout (Gal & Ghahramani 2016), which consists in randomly setting to zero a number of weights in the CNN, with the aim of selecting objects for which the network finds a high variance in the output label. This technique allowed the authors to identify galaxies in Illustris TNG which do not look realistic.

A major step forward in the field of Machine Learning has been made in the very recent years, when Deep Generative Models proved able to generate from scratch new, extremely realistic samples. This is for instance the case of Generative Adversarial Networks (GANs, Goodfellow et al. 2014) and Variational Autoencoders (VAEs, Kingma & Welling 2014). Both GANs and VAEs rely on the definition of a lower dimensional latent space from which random numbers are drawn and fed to the generator (for GANs) or the decoder (for VAEs) which eventually will produce mock samples the realism of which is assessed via the minimization of a loss function. GANs and VAEs have been successfully used in astronomy with various purposes (only to mention a few, Reiman & Göhre 2019, Schawinski et al. 2017, Glaser et al. 2019, Zingales & Waldmann 2018, Karmakar et al. 2018). GANs are extremely powerful, but they lack an explicit likelihood that can be evaluated on a single image, which is the single-valued metric we would hope to use to compare simulations and observations. VAEs do have such feature but their likelihood may not be easily interpretable.

Here we propose the use of PixelCNN, a deep autoregressive generative model, as a novel tool to efficiently compare simulations and observations. PixelCNNs (van den Oord et al. 2016a, van den Oord et al. 2016b) explicitly learn the probability distribution of the pixel values of an image (i.e. from 0 to 255 in a png image) in an autoregressive fashion (i.e. the value of each pixel is conditioned to that of previously processed pixels). The appeal of PixelCNN is that it features an explicit, tractable likelihood with probabilistic meaning. A PixelCNN network trained on images from galaxy surveys provides a framework to compare the likelihood of real and mock observations. Such likelihood is a well defined metric that may be used to assess to which extent current competing hydrodynamical simulations of galaxy formation and evolution reproduce galaxy morphologies and colors.

The outline of this work is as follows. In Section 2 we present the training sample from observations (Section 2.1) and the test set coming from simulations (Section 2.2). In Section 3 we describe the implementation of PixelCNN and we give details about the training procedure. In Section 4 we give the main result of our work, while in Section 5 we describe a few tests that we have performed to better assess the validity of our results. Section 6 is a summary of the main critical points that need to be addressed in the near future and some practical applications and extensions of the PixelCNN framework are also discussed.

2 DATA

2.1 Observations

In the following we will use the SDSS DR7 (Abazajian et al. 2009) spectroscopic sample as presented in Meert et al. (2015), Meert et al. (2016). The Meert et al. catalogues consist of 670722 objects the photometry of which benefits of substantial improvement both in background subtraction and fits to the light profiles. The galaxy stellar masses are computed adopting `Sérsic+Exponential` photometric fits and the mass-to-light ratio M_{star}/L by Mendel et al. (2014). Although the spectral energy distribution of galaxies contains information which is critical to understand the physical processes that galaxy formation, in this exploratory work we choose to adopt only single band images (specifically r-band) as a proof of concept. We plan to expand our work to multi-band photometry in the imminent future. We also match the Meert et al. catalogues with the Domínguez Sánchez et al. (2018) CNN-based morphological classification which will be useful in the following.

As for the training sample, we use the images of SDSS galaxies that have a stellar mass $M_{star} > 10^{10} M_{\odot}$. An important issue that must be dealt with when choosing the training sample is that of the redshift evolution of the angular diameter distance driven by cosmology. Indeed, the pixel physical scale[1] is a strong function of redshift, which means that the training sample must be chosen so that the average pixel scale is as close as possible to the pixel scale at the redshift of the snapshot that we use for the simulations (i.e. $z \sim 0.045$, see Section 2.2). Hence, we also limit the redshift range of the SDSS training sample to $0.02 < z < 0.08$, which gives a median pixel scale $\lesssim 30\%$ larger than the pixel scale at $z \sim 0.045$. This redshift cut leaves us with ~100000 galaxies in SDSS. We found this to be a good compromise between the size of the training sample and the effect of the strongly decreasing angular diameter distance with increasing redshift. This choice may be critical and subject to change in future work.

2.2 Simulations

We here use the Illustris Simulation (Vogelsberger et al. 2014, Genel et al. 2014, Sijacki et al. 2015) and its successor Illustris TNG (Pillepich et al. 2018a, Nelson et al. 2019). Both are hydrodynamical cosmological simulations , run with the *AREPO* solver (Springel 2010). The Illustris simulation has been proved capable of reproducing several observables, but presents major shortcomings (both features are summarized in Nelson et al. 2015). In the Illustris TNG simulation significant changes have been made with respect to the Illustris simulation (see Pillepich et al. 2018a for a detailed summary). These include a better numerical resolution, modelling of magnetic fields, the substitution of the *bubble mode* AGN feedback at low accretion rates (Sijacki et al. 2007) with a kinetic AGN feedback (Weinberger et al. 2017), a modification of the implementation of galaxy-wide winds, updated mass yield from star particles, and injection of $r - process$ material from neutron star-neutron star mergers (Naiman et al. 2018).

[1] i.e. kpc/pix

We here aim to compare the Illustris and Illustris TNG simulations with available observations by means of a novel technique based on a deep learning framework to assess the widely recognised improvements featured by Illustris TNG with respect to the original Illustris simulation (e.g. Pillepich et al. 2018b, Nelson et al. 2018, Rodriguez-Gomez et al. 2019, Donnari et al. 2019). In both simulations we select galaxies with $M_{star} > 10^{10} M_{\odot}$[2] in the snapshot number 95 at $z \sim 0.045$, for a total of $\sim$ 12000 galaxies each. The images are processed with a joint use of the radiative transfer code `SKIRT` (Baes et al. 2011, Camps & Baes 2015), the nebular modelling code `MAPPINGS-III` (Groves et al. 2008) and the Bruzual & Charlot (2003) `GALAXEV` stellar population synthesis code and are originally presented in Rodriguez-Gomez et al. (2019). Briefly, each stellar particle in either simulation (which represents a coeval stellar population) is modelled with `GALAXEV` for stellar particles older than 10 Myr, while younger stellar particles are treated as a starbursting population with `MAPPINGS-III`. To model dust, it is assumed that the diffuse dust content of each galaxy is traced by the star-forming gas, that the dust-to-metal mass ratio is constant and equal to 0.3 (Camps et al. 2016), and that dust is a mix of graphite grains, silicate grains, and polycyclic aromatic hydrocarbons (Zubko et al. 2004). Galaxies are observed along a random line of sight. We refer the reader to Rodriguez-Gomez et al. (2019) for further details.

2.2.1 *Realistic images from simulations*

When comparing images from simulations and observations, it is essential that the mock observations are performed with the same level of realism that is found in galaxy surveys. Bottrell et al. (2017a) and Bottrell et al. (2017b) presented *RealSim*, an algorithm that enables such procedure. Briefly, with *RealSim* it is possible to place a galaxy from a given simulation, processed with radiative transfer as explained above, in a real SDSS field. The mock galaxy will also be convolved with the Point Spread Function of that particular field; the effects of shot noise and cosmological surface brightness dimming are also included. For more details about *RealSim*, we refer the reader to the original papers.

3 PIXELCNN

PixelCNN (van den Oord et al. 2016a, van den Oord et al. 2016b) is an autoregressive generative model with an explicit likelihood, namely, the likelihood a given pixel is assigned is conditioned on all the previous pixels of the image (which sometimes are collectively called "context"), so that

$$P(X) = \Pi_{i=1}^{N^2} P(X_i | X_{1\ldots i-1}). \tag{1}$$

Here $P(X_i | X_{1\ldots i-1})$ is the conditional probability distribution function of pixel i evaluated at X_i. Eq. 1 models explicitly the likelihood of the training sample. In the following we will use the negative log-likelihood, which is less prone to floating point limitations,

$$\mathcal{L} \equiv -log(P(X)) = -\sum_{i=1}^{N^2} log(P(X_i | X_{1\ldots i-1})) \tag{2}$$

The above ansatz imposes the choice of an ordering for the pixels. We follow a prescription according to which the image is scanned from top left to bottom right, row by row. This is a standard implementation of PixelCNN that takes advantage of the way convolutions are typically implemented in deep learning frameworks such as `TensorFlow` and `PyTorch`. The autoregressive nature of PixelCNN is achieved by means of a particular type of convolutions that mask the pixels to the right and bottom of the current pixel, so that the network is forced to learn the relationship between each pixel and the previous context only (van den Oord et al. 2016a,van den Oord et al. 2016b).

We here adopt the `PixelCNN++` architecture proposed by Salimans et al. (2017)[3] interfaced with a higher level `Tensorflow` API[4]. Briefly, Salimans et al. adopt a fully convolutional autoencoder-like architecture, with three downsampling and three upsampling stages respectively, where downsampling and upsampling are implemented using strided convolutions [5]. Each stage consists of an adjustable number of Gated Resnet layers (van den Oord et al. 2016a, He et al. 2015), which entail zero-padding convolutions to preserve dimensionality. Stages in the downsampling and upsampling parts of the network with the same dimensionality are connected with shortcut connections as in Ronneberger et al. (2015), to ensure that part of the information lost in the downsampling is recovered. We refer the reader to Salimans et al. (2017) and van den Oord et al. (2016b), van den Oord et al. (2016a) for further details of the implementation.

Obviously, not all images will have the same likelihood. Rather, PixelCNN maps a distribution of images into a distribution of likelihoods. This feature is extremely powerful, since it allows to collapse the complexity that characterizes images into a single-valued function. Such property is particularly amenable to compare two different datasets. Indeed, a trained PixelCNN produces a distribution of likelihood values for any given test set $\mathcal{L}_{test}(X_{test})$. If these images come from the same underlying distribution of the training set Q, $X_{test} \sim Q$, then $\mathcal{L}_{train}(X_{train}) = \mathcal{L}_{test}(X_{test})$ where $\mathcal{L}_{train}(X_{train})$ is the likelihood distribution for the training set. If, on the other hand, the test set is not a realization of Q, then $\mathcal{L}_{train}(X_{train}) \neq \mathcal{L}_{test}(X_{test})$. PixelCNN may therefore be used in principle as a tool to assess whether two datasets are consistent. Moreover, if the distributions from which the datasets under exam come from are similar but not identical, our framework is able to identify candidate outliers in the test dataset. In our specific case, the training set comes from observations, while the test set can come from either observations or simulations.

3.1 Training the network

The images in the training sample (in the simulations), which originally were of size of 256x256 (128x128) pixels, are cropped to 64x64 and degraded to reach the size of 32x32 pixel[6] in order to meet memory and time constraints. This however might be dangerous as we are interested in probing the fine morphological structure of the simulated galaxies, which might get lost in the downsampling. Such procedure might play either in favour or against simulations, as peculiar features might be washed out (if they are isolated) or enhanced (if they are more clustered). We also recall from Section

[2] This is the same mass cut performed in SDSS.

[3] Available at https://github.com/openai/pixel-cnn

[4] Available at https://github.com/pmelchior/scarlet-pixelcnn

[5] Transposed convolutions in the case of upsampling.

[6] We use the publicly available `scipy.ndimage` library.

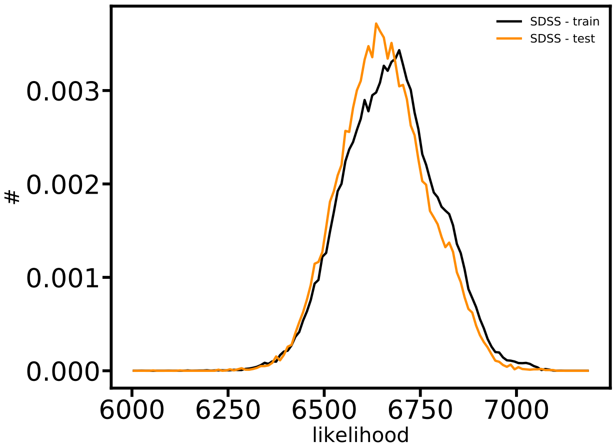


Figure 1. Likelihood distribution of training and test sets (black and orange lines respectively). The distribution for the training set peaks at slightly higher likelihood values than the test set, however this is expected as in any machine learning framework the performance of the training set is always higher than in the test set.

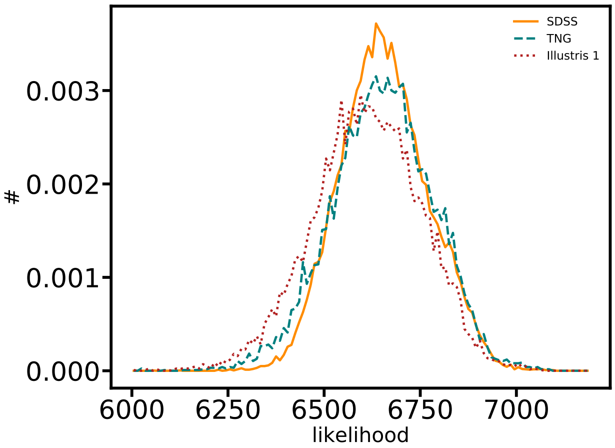


Figure 2. Likelihood distribution for the test set of SDSS (orange solid line), Illustris TNG (teal dashed line) and Illustris (red dotted line). It is visually clear that the distribution for Illustris TNG is much more similar to that of SDSS than the distribution of Illustris, which peaks at a lower likelihood and is significantly broader.

2.1 that the average pixel scale in the training set is roughly 30% larger that in the mock observations of Illustris and Illustris TNG, which might bias the network towards smaller objects, as they would be shown more frequently during the training. As a test to address these last issues, we plan to train the network in a narrower redshift range and retaining the full resolution of the original images.

To train PixelCNN we use 75000 galaxies randomly extracted from our SDSS sample, which are augmented ten times via random rotations. The remaining objects are used as a test set to evaluate the model. Figure 1 shows the distribution of likelihood values of the training and test set once the model has converged. It can be seen that the two distributions are very similar, with the training set peaking at slightly higher likelihood than the test. This is expected, as the performance of a trained machine on the training set is always higher than on the test set.

4 RESULTS

Figure 2 shows the likelihood distributions of images coming from SDSS, Illustris and Illustris TNG. This figure constitutes the main result of our work. It can be seen that the likelihood distribution of Illustris peaks at a lower likelihoods and has a higher variance than those of SDSS and Illustris TNG. The better agreement of the likelihood distribution of Illustris TNG to that of SDSS, compared to Illustris, is a clear sign that Illustris TNG performs significantly better that Illustris, as widely recognised in a number of works (see Section 1). Our results support the use of the deep learning framework adopted here to compare multiple state-of-the-art simulations and set the benchmark for the future generation of hydrodynamical cosmological simulations.

Clearly, an eyeball assessment of Figure 2, although informative, is not yet scientifically appealing. We are now working on a metric that may be able to formally quantify the distance between distributions. As a preliminary test we found that the Kullback-Leibler divergence is not sensitive enough, while a more classical two sample Kolmogorov-Smirnov test is too sensitive and would return $p - value = 0$. An alternative solution would be to use an extension of the method outlined in Sautter & Barchi (2017) to compute the geometric distance between histograms, which is currently under development (P. Barchi, private communication).

4.1 The weirdest objects in Illustris and Illustris TNG

An interesting feature of our likelihood-based framework is that it allows to identify potential outliers in the simulations. As a preliminary proof-of-concept, we identify outliers as the objects that lie at more that three standard deviations from the peak of the likelihood distribution of SDSS. Figures 3 and 4 show galaxies in Illustris and Illustris TNG selected in such a way. Visually, the improvement of the morphological features in the latter simulation compared to the former is clear. In the following we discuss these results.
First of all, some of these low-likelihood cutouts have very bright field stars that the network may recognise as a rare feature. We recall however that simulated galaxies are placed in real fields, so it is possible that the low likelihood values for the thumbnails in Figures 3 and 4 stem from a combination of an unfortunate background and unrealistic galaxy morphologies. Furthermore, we recall that images lying three σ below the mean of the SDSS are *candidate* outliers and not necessary *true* outliers. Indeed, with our selection criterion the candidate outliers may still have a likelihood consistent with that of the low likelihood tail of SDSS galaxies. It would instead be desirable to assess the probability of a galaxy being an outlier given its likelihood value, but due to the limited time we have not been able to explore this further. Another potential way of improving the current strategy would be to place images from the simulations in a range of real SDSS fields. Given that for each field the likelihood will be different, we will select outliers as objects that systematically lie at low likelihood.

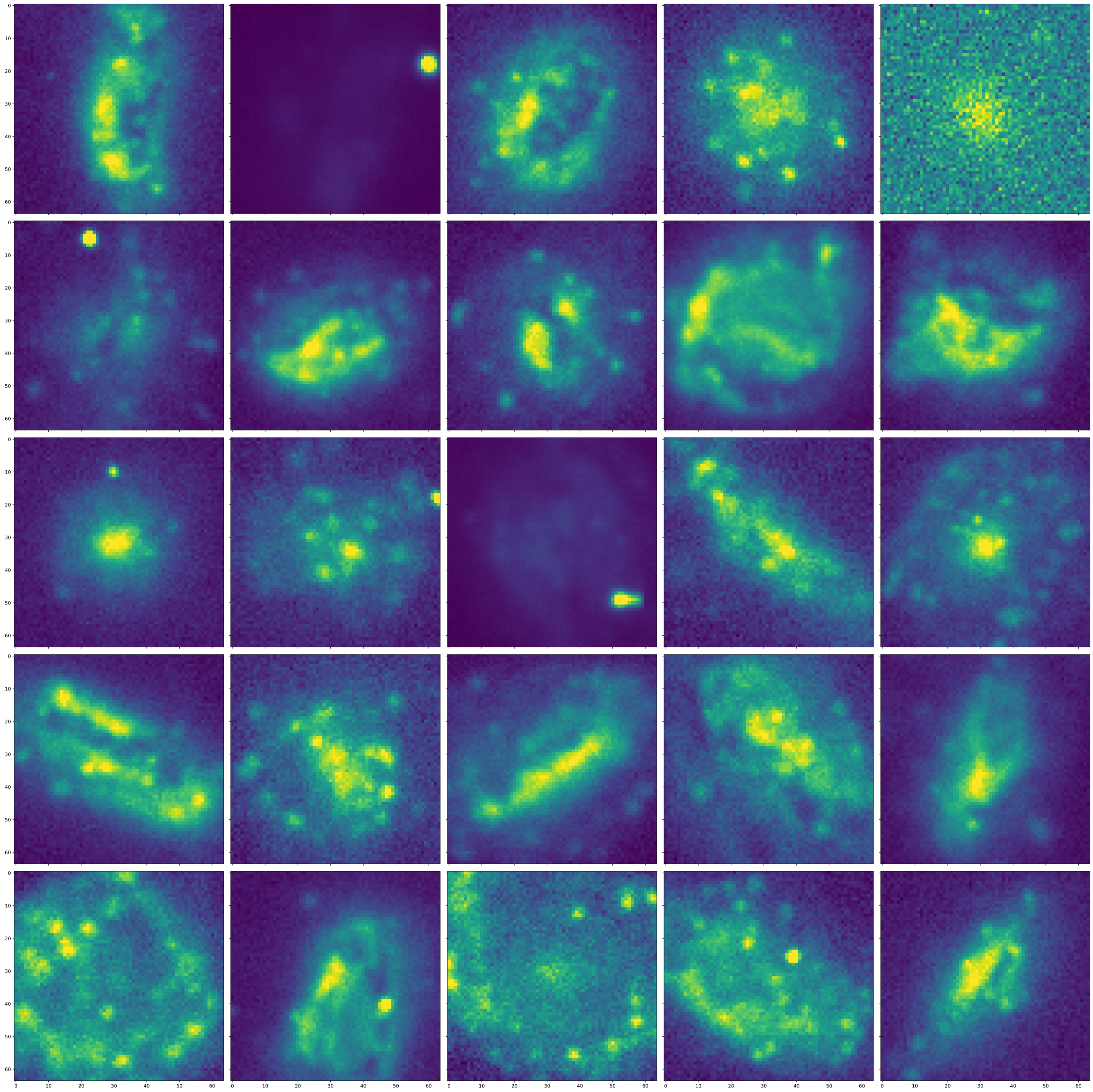


Figure 3. Illustris galaxies with values of likelihood three standard deviations below the mean of the likelihood distribution of SDSS.

5 WHAT DOES THE NETWORK LEARN?

The success of our PixelCNN framework is surely promising, but a more thorough understanding of its inner working is needed in order to avoid using it as a black box. We have therefore performed some tests which we summarize below.

5.1 UMAP representation

What drives the broadness of the likelihood distributions? In other words, why does the network attribute a lower or higher likelihood to certain galaxies? To address this issue, we can take a step back and look at correlations between the likelihood and the (non) parametric diagnostics that are available in the Meert et al. (2015) catalogues and the matched Domínguez Sánchez et al. (2018) morphological catalogue. Specifically, we use stellar mass, effective radius, axis ratio, redshift, TType, Sérsic index and the $C-A-S-G-M_{20}$ parameters. Due to the high dimensionality of the problem, it is very well possible that the different parameters have higher order correlations, which would be hard to read from a traditional pair plot. A valid alternative is to make use of UMAP (Uniform Manifold Approximation and Projection, McInnes et al. 2018), a dimensionality reduction algorithm thanks to which it is possible to find lower di-

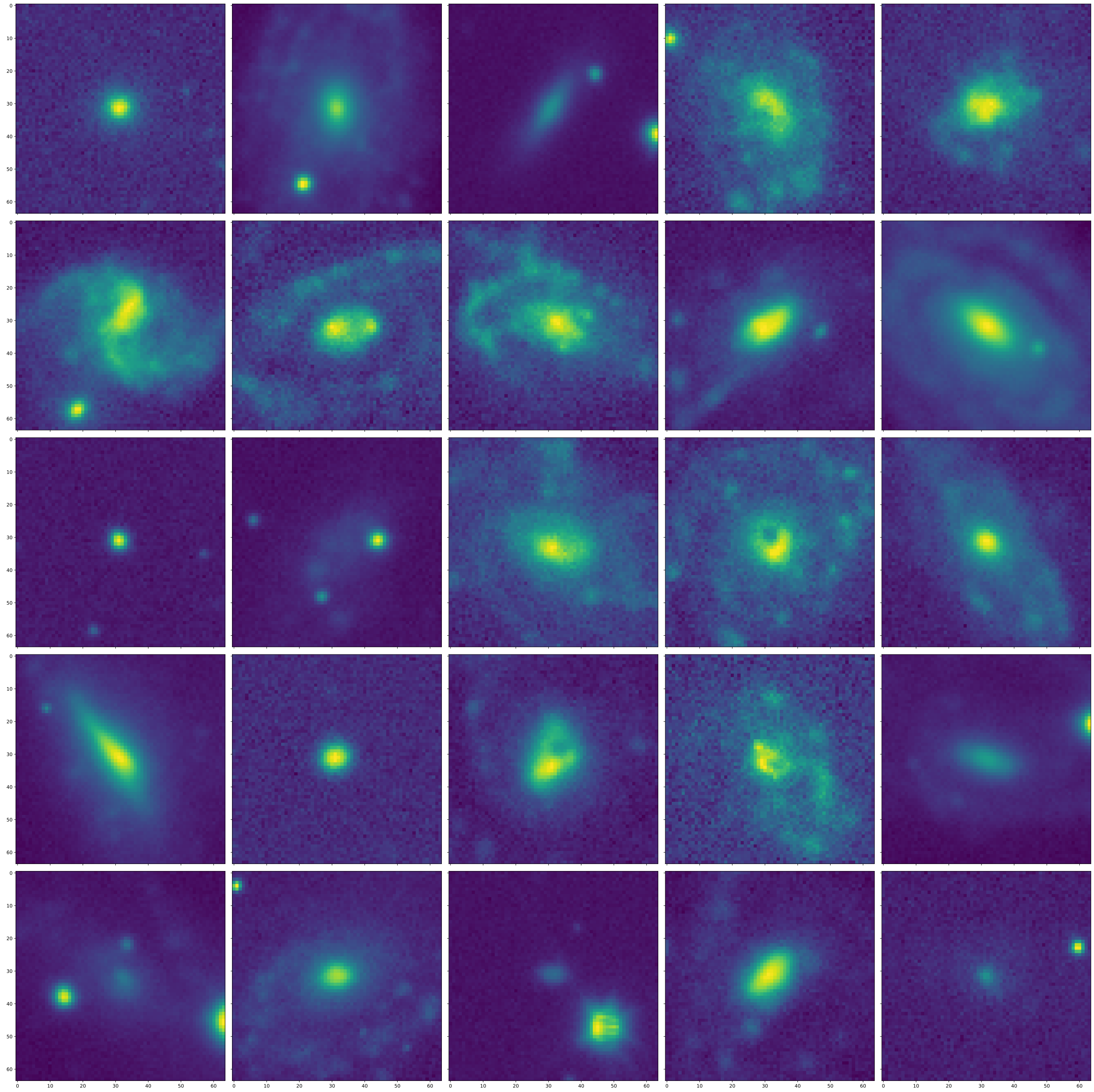


Figure 4. Illustris TNG galaxies with values of likelihood three standard deviations below the mean of the likelihood distribution of SDSS

mensional projections of a high dimensional spaces. UMAP, which is constructed from a theoretical framework based in Riemannian geometry and algebraic topology, searches for a low dimensional projection of the data that has the closest possible equivalent topology. We apply UMAP to our (non) parametric catalogue of galaxy properties and color code such embedding according to each entry in the catalogue.

Figure 5 shows the result of the exercise outlined above for galaxies in SDSS with $10^{10.5} M_{\odot} < M_{star} < 10^{11} M_{\odot}$. We show here only the results for such mass cut because it has a richer morphological mix than higher or lower masses (Zanisi et al. 2019 submitted), which turns out to be very important (see below). In any case, the results are qualitatively similar for galaxies with lower and higher mass.

The correlations between the likelihood and the parameters that we are interested in can be found by looking at the *spatial* correlation between the different panels of Figure 5. The most striking feature is the anticorrelation between likelihood and galaxy morphology (first row, first panel and second row, second panel respectively). Indeed, the network assigns systematically lower likelihood values

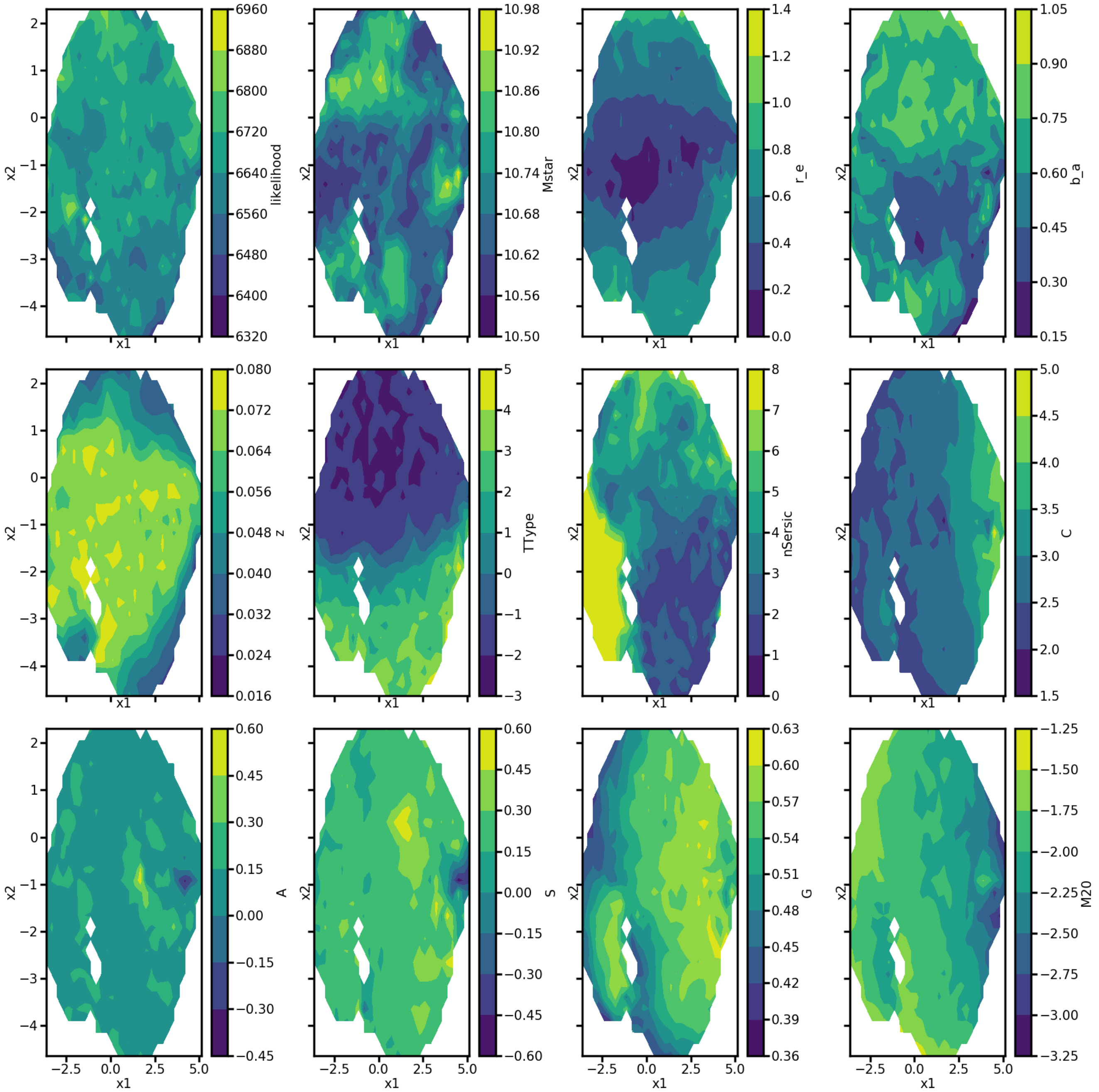


Figure 5. UMAP representation of SDSS with likelihood.

to progressively more irregular galaxies[7], while earlier types tend to be recognised as more likely than average. We show a compilation of cutouts from SDSS with low and high likelihood in Figure 6. Extremely interesting is also the *lack* of correlation between the likelihood and *any* of the classical $C-A-S-G-M_{20}$ non-parametric diagnostics, which, along with the already noted anticorrelation between likelihood and galaxy morphology, is proof that our approach is more general than that of, e.g., Rodriguez-Gomez et al. (2019). Moreover, the UMAP approach can reveal more intricate, higher order, multi-variate correlations that may be extremely useful to understand the preferences of the network. As an example, it is possible to examine the UMAP region of galaxies with high Sérsic index (second row, third column, left part of the panel in yellow) and explore which values of the other parameters typically characterize galaxies in that region. It can be seen that low redshift, larger, more massive galaxies with irregular morphology (of which the high Sérsic index signals a failed photometric fit) that sit at the bottom of the region under exam tend to have lower values of likelihood compared to galaxies at the top of that region of the plot, with a higher redshift, more regular morphology, and lower stellar mass

[7] Which have an increasing TType.

and effective radius (which are fit with a high Sérsic index because they truly are Early Type Galaxies). On the other hand, not all large galaxies are unlikely. Indeed, it can be seen that earlier types with lower redshift and large sizes have the highest likelihoods, possibly because they also entail axis ratios close to one (i.e. they are perfect spheroids).
We have given only a couple of examples of how to read Figure 5, but we encourage the reader to explore the multiple, interesting correlations that are shown there.

5.2 Additional tests

It is instructive to feed the network images that the network is not supposed to recognise well. Below we present two experiments which allow us to dive deeper in the details of the inner working of PixelCNN.

5.2.1 Feeding sky background to the network

As a first experiment, instead of evaluating the likelihood of galaxies, we evaluate that of sky background. To do so, we use the top left corner of all our original (i.e. not cropped) SDSS images. As an additional test, we get rid of background sources in a significant amount of "corners" by requiring that less than 1% of the pixels in an image feature values above the 99.7 percentile of the average noise of SDSS, which was computed from a sample of visually inspected sourceless "corners". A
Figure 7 shows the result of such experiment. It is interesting to see that the distribution of likelihood of the sky background $\mathcal{L}_{sky}$ has the same width of that of galaxies, but it is shifted to lower likelihoods. We recall that the likelihood mentioned here and throughout the paper is really the log-likelihood, which is the sum of the log-likelihoods of all the pixels (see eq. 2). Therefore, most of the overall value of the likelihood will come from the sky background, since it constitutes the bulk of the pixels of an astronomical image. Also interesting is that $\mathcal{L}_{sky}$ is roughly as wide as that of galaxies, which suggests that the sky background is important in determining $\mathcal{L}_{gal}$. This would be supported by the fact that $\mathcal{L}_{sky}$ considerably shrinks if only the cutouts without background sources are retained (red dotted line in Figure 7). This test shows that at least some of the width of the likelihood distribution of galaxies must come from the diversity of the sky backgrounds a galaxy may be found in, in addition to the galaxy properties (see Section 5.1). Indeed, Figure 6 shows that most low-likelihood galaxies lie in more crowded fields, while the opposite is true for galaxies at the high end of the likelihood spectrum.

5.2.2 Feeding idealistic light profiles

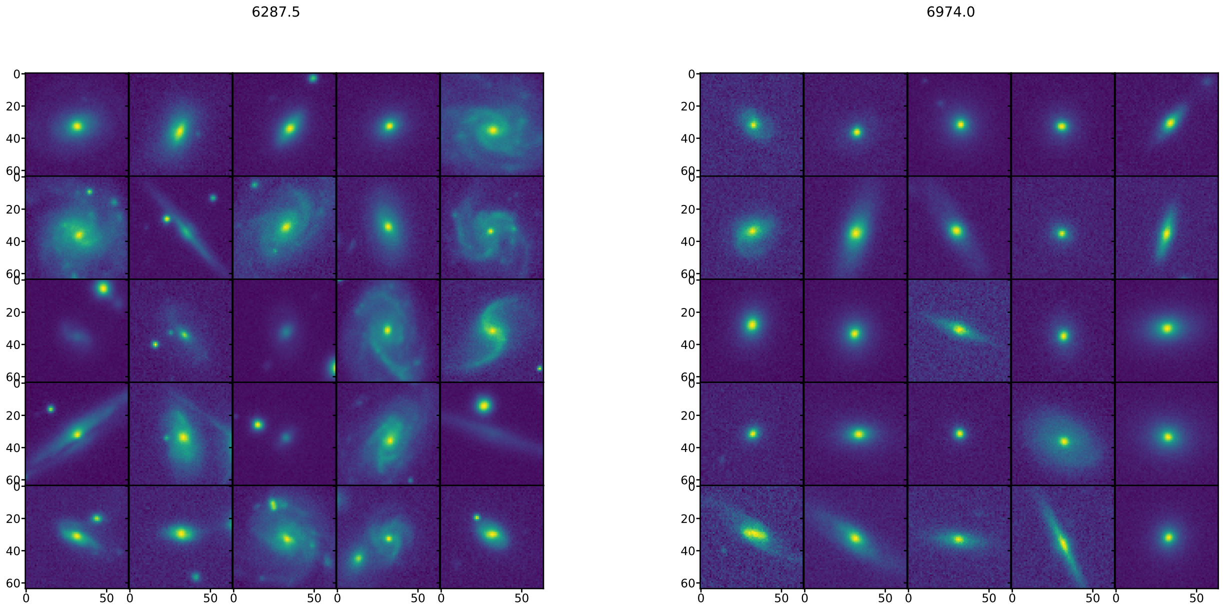
In previous work CNNs have been trained to find the best photometric fit to a galaxy's light profile (e.g., Dimauro et al. (2018), Tuccillo et al. (2018)). Does PixelCNN recognise better a galaxy or its best photometric fit? To answer this question, we use *GalSim* (Rowe et al. 2015) to produce images of the best photometric fits from the Meert et al. (2015) catalogue. We use the `finalflag` in the file `UPenn_Phot_Dec_Models_rband.fits`, ensuring that only good fits are retained. In particular, a galaxy is fitted either by a pure Sérsic profile, a disk profile or a Sérsic+Exponential profile and only the best fit between the three is used. To the resulting images we also add SDSS realism with *RealSim*, as we do for images from simulations (Section 2.2.1). The likelihood distribution of the resulting images is compared to that of real galaxies in Figure 8. Perhaps not surprisingly, the network assigns slightly higher values of likelihood to the best fit of galaxies than to real galaxies. This may be expected, given that the best fit of a galaxy is an idealized, smooth profile which the network surely finds simpler to recognize. This may seem trivial, but it is a hot topic in the machine learning community. Indeed, it has been found that in likelihood-based approaches, as the one adopted here, similar trends in the likelihood distributions to the one shown in Figure 8 arise when comparing two datasets that do not come from the same underlying distribution. For example, **?** find that several networks (including a PixelCNN) trained on the CIFAR dataset (Krizhevsky 2012) would assign a higher likelihood to images from the SVHN dataset (Netzer et al. 2011)[8]. This result means that, for instance, a network assigns a higher likelihood to an image of a number, than to that of a cat, despite being trained on a set of images that includes cats and not numbers (Shafaei et al. 2018). This is of course similar to what we see in Figure 8, where $\mathcal{L}_{best}$ peaks at higher likelihoods than $\mathcal{L}_{gal}$ despite our network was trained on galaxies and not on their best fit. Yet our framework comes with the advantage that there is a correspondence between galaxy images and simpler, idealized objects such as the best fit to their light profile, which is something that cannot be done for natural images. Our work suggests that simpler objects (such as Sérsic profiles instead of galaxies, and numbers instead of animals) tend to be assigned a higher likelihood by likelihood-based generative models such as PixelCNN. However we are also able to show higher order complexity. In Figure 9 we show the quantity $\mathcal{L}_{gal} - \mathcal{L}_{best}$ as a function of $\mathcal{L}_{gal}$. There is a clear correlation, in the sense that galaxies that have lower $\mathcal{L}_{gal}$ (typically Late Type Galaxies) also have higher $\mathcal{L}_{best}$, that is the network "prefers" smooth, idealized objects to more irregular morphologies. Indeed, Late Type Galaxies are also characterized by an overall high value of reduced chi square (Figure 9, right panel). What is more surprising is that observed Early Type Galaxies are preferred to their best fit light profile, despite having typically better chi square values. A potential interpretation of this result is that PixelCNN is able to pick finer details in the light profiles of Early Type Galaxies that a classical χ^2 analysis is blind to. This is clearly a matter worth exploring in future work.

6 CAVEATS AND FUTURE WORK

Future work includes both a refinement of the methodology and further tests, and obviously more practical applications of the framework outlined here:

- We plan to train the network in a narrower redshift range, retaining the full resolution of the original images, to better tackle the issue of the evolution of the angular diameter distance and potential biases due to the currently adopted image degradation.
- Moreover, we need a metric to assess the simliarity between observations and simulations. Indeed, while it is visually clear from Figure 2 that Illustris TNG performs better than Illustris, it might not be as clear when adding the comparison with other simulations,

[8] CIFAR and SVHN are popular datasets used in the machine learning community to validate the architecture of neural networks. CIFAR is a collection of natural images, such as animals, plants and objects in various flavours, while SVHN is a collection of house numbers gathered from Google Street View images.



(a)

(b)

Figure 6. Left: galaxies with low likelihood ($\mathcal{L}_{gal} \approx 6287.5$) in SDSS. Right: galaxies with high likelihood ($\mathcal{L}_{gal} \approx 6974$) in SDSS. It can be seen that galaxies with low likelihood tend to be more disky and irregular, but also have significant contamination from bright field stars. Galaxies with high values of likelihood tend to be smaller and less irregular, and are in less crowded fields. However, it can also be noted that some galaxies with low likelihood are very regular (e.g. top left thumbnail in left panel), while some late type galaxies are assigned higher likelihoods (e.g. bottom left thumbnail in right panel)

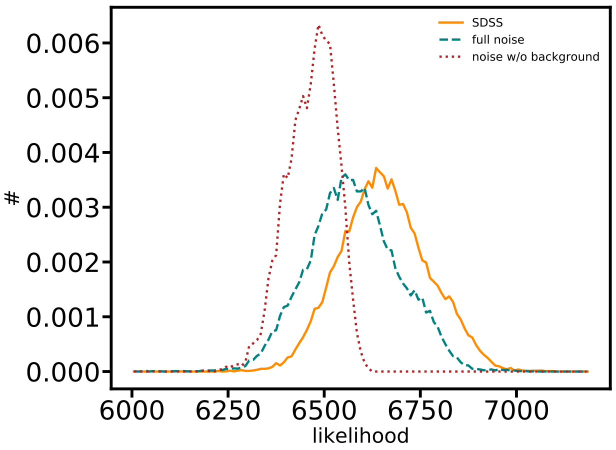


Figure 7. Likelihood distribution of SDSS (orange solid line) compared to that of the sky background with sources (teal dashed line) and with a reduced number of sources (red dotted line).

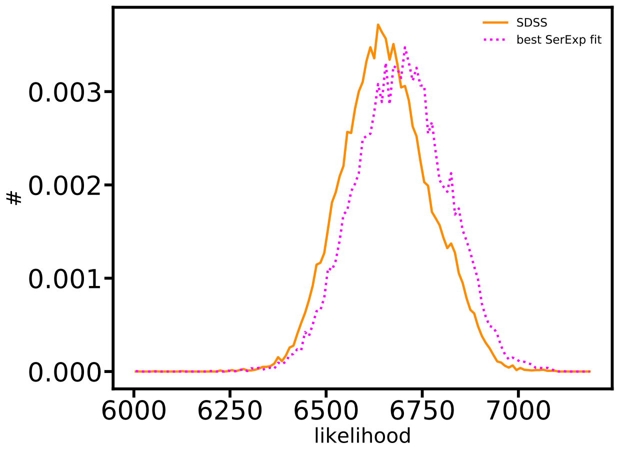
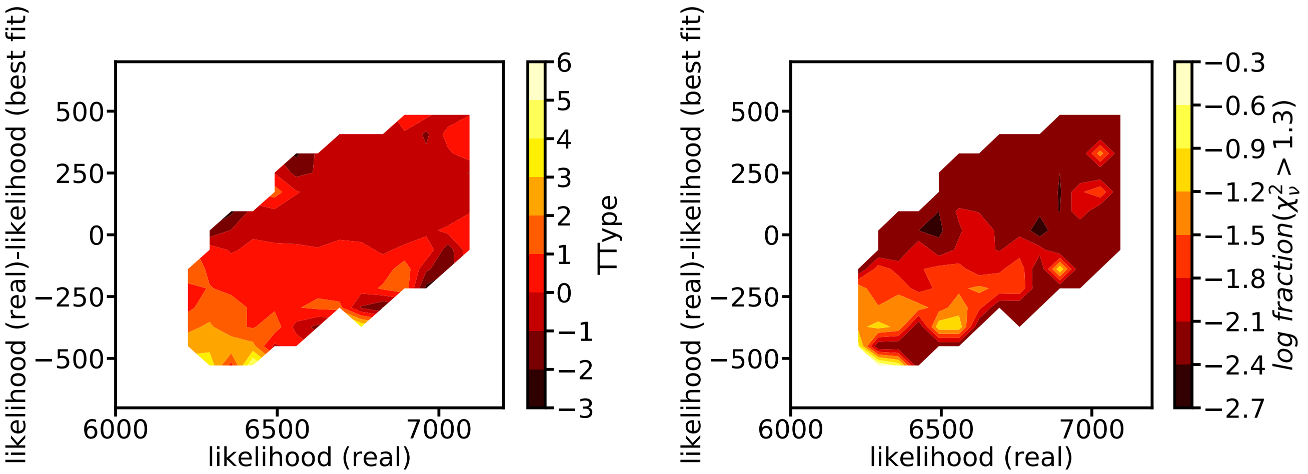


Figure 8. Likelihood distribution of the fit to the galaxy light profile as detailed in Meert et al. (2015) catalogue (magenta dotted line) compared to the likelihood distribution of observed galaxies (orange solid line).

especially at higher redshift, where the predictions from simulations may differ more significantly from observations.

- The selection of outliers may not be very efficient yet, as we are currently excluding potentially interesting objects that are not too far away from the mean of SDSS and including objects that may be consistent with observations. Indeed, for example, at a likelihood of ~ 6400 (which is below the threshold for our outlier selection) there is a much higher number of Illustris galaxies compared to SDSS which as of now are completely missed by our outlier selection procedure. We plan to adopt a more robust methodology.
- A related issue is that of the sky background, which is fundamental in determining the likelihood of simulated galaxies. As a consequence, some non realistic galaxies may be assigned a higher likelihood simply because of a fortunate background, and galaxies the morphology of which would agree with observations are pushed at a lower likelihood because of, e.g., a bright star in the field or a



(a) (b)

Figure 9. Correlation between the difference of the likelihood of real galaxies $\mathcal{L}_{gal}$ and that of their best fit $\mathcal{L}_{best}$ as a function of $\mathcal{L}_{gal}$ and color coded by TTYpe (lef) or the fraction of fits with $\chi^2_\nu > 1.3$ (right), that corresponds roughly to $p - value = 0.25$.

peculiar noise feature. We plan to address this issue by placing each simulated galaxy in a collection of 100 fields and select the objects that are systematic outliers.

- An obvious extension of the present work would be to apply the PixelCNN framework to high redshift data, where competing simulations make substantially different predictions. We plan to compare the zoom-in suites of simulations VELA (Ceverino et al. 2014) and FIRE (Hopkins et al. 2014) to Illustris TNG50, which entails a comparable resolution but also a much larger cosmological box.

- We plan to explore the potential of our framework in selecting better fits to the light profile of galaxies.

- We also aim to expand our framework to include multi-band images. In the original `PixelCNN++` paper (Salimans et al. 2017) it is proposed that the three filters that characterize a color image be modelled autoregressively by imposing that the green color be linearly dependent from the red color with the blue color also depending linearly from green and red. Given that the relationship between the bands of an astronomical image (i.e. the Spectral Energy Distribution, SED) is more complicated than a simple linear scaling, we would like to have more freedom in the modelling of the colors for our application. A possible solution would be to use a Masked Autoencoder for Distribution Estimation (MADE, Germain et al. 2015). A MADE is an autoencoder the weights of which are suitably masked so that the input image is reconstructed autoregressively at the output of the network, with a loss function similar to eq. 1. Such approach has the advantage that the relationship between the filters of a colour image is learned and not forced to be linear as in Salimans et al. (2017). Our plan is to use a MADE to empirically model the SED of each pixel in an autoregressive fashion. This will allow us to perform a more detailed comparison between observations and simulations that includes also multi-band information. A preliminary version of our `PixelMADE`, where we couple a MADE to PixelCNN, is available at `https://github.com/lorenzozanisi/Kavli2019`.

- Of course our results may be somewhat dependent on the implementation of radiative transfer described in Section 2.2, and this issue may become even more important when dealing with multi-band information. However, while this may be an issue when selecting outliers from the simulations, it should not be too critical when comparing different simulations processed exactly in the same way.

- Finally, but not in order of importance, we will be able to look at the formation histories of outliers in the simulations. This will allow us to understand which physical processes (or numerical artifacts) have generated such galaxies.

ACKNOLEDGEMENTS

L.Z. warmly thanks Pascale Gauraud, Brant Robertson, Xavier Prochaska, Alexie Leauthaud, Shirley Ho, David Kirkby, Brice Ménard, Michelle Lochner and Marc Huertas-Company for organizing the Kavli Summer Program in Astrophysics 2019 and UCSC for hosting the school. We thank the Kavli Foundation and the National Science Foundation for their generous support. L.Z. also wishes to thank all the participants for their help and support during the school.

REFERENCES

Abazajian K. N., et al., 2009, ApJS, 182, 543
Baes M., Verstappen J., De Looze I., Fritz J., Saftly W., Vidal Pérez E., Stalevski M., Valcke S., 2011, ApJS, 196, 22
Bignone L. A., Pedrosa S. E., Trayford J. W., Tissera P. B., Pellizza L. J., 2019, arXiv e-prints, p. arXiv:1908.10936
Bottrell C., Torrey P., Simard L., Ellison S. L., 2017a, MNRAS, 467, 1033
Bottrell C., Torrey P., Simard L., Ellison S. L., 2017b, MNRAS, 467, 2879
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Camps P., Baes M., 2015, Astronomy and Computing, 9, 20
Camps P., Trayford J. W., Baes M., Theuns T., Schaller M., Schaye J., 2016, MNRAS, 462, 1057
Ceverino D., Klypin A., Klimek E. S., Trujillo-Gomez S., Churchill C. W., Primack J., Dekel A., 2014, MNRAS, 442, 1545
Conselice C. J., 2003, ApJS, 147, 1
Crain R. A., et al., 2009, MNRAS, 399, 1773
Croft R. A. C., Di Matteo T., Springel V., Hernquist L., 2009, MNRAS, 400, 43

Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, MNRAS, 486, 2827

Di Matteo T., Khandai N., DeGraf C., Feng Y., Croft R. A. C., Lopez J., Springel V., 2012, ApJ, 745, L29

Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441

Dimauro P., et al., 2018, MNRAS, 478, 5410

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

Donnari M., et al., 2019, MNRAS, 485, 4817

Dubois Y., et al., 2014, MNRAS, 444, 1453

Dubois Y., Peirani S., Pichon C., Devriendt J., Gavazzi R., Welker C., Volonteri M., 2016, MNRAS, 463, 3948

Fanaroff B. L., Riley J. M., 1974, MNRAS, 167, 31P

Freeman P. E., Izbicki R., Lee A. B., Newman J. A., Conselice C. J., Koekemoer A. M., Lotz J. M., Mozena M., 2013, MNRAS, 434, 282

Furlong M., et al., 2015, MNRAS, 450, 4486

Gal Y., Ghahramani Z., 2016, in Proceedings of the 33rd International Conference on Machine Learning (ICML-16).

Genel S., et al., 2014, MNRAS, 445, 175

Genel S., et al., 2018, MNRAS, 474, 3976

Germain M., Gregor K., Murray I., Larochelle H., 2015, arXiv e-prints, p. arXiv:1502.03509

Glaser N., Wong O. I., Schawinski K., Zhang C., 2019, MNRAS, 487, 4190

Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, arXiv e-prints, p. arXiv:1406.2661

Groves B., Dopita M. A., Sutherland R. S., Kewley L. J., Fischera J., Leitherer C., Brandl B., van Breugel W., 2008, ApJS, 176, 438

He K., Zhang X., Ren S., Sun J., 2015, CoRR, abs/1512.03385

Hopkins P. F., Kereš D., Oñorbe J., Faucher-Giguère C.-A., Quataert E., Murray N., Bullock J. S., 2014, MNRAS, 445, 581

Huertas-Company M., et al., 2015, ApJS, 221, 8

Huertas-Company M., et al., 2019, arXiv e-prints, p. arXiv:1903.07625

Karmakar A., Mishra D., Tej A., 2018, arXiv e-prints, p. arXiv:1809.01434

Kaviraj S., et al., 2017, MNRAS, 467, 4739

Kingma D. P., Welling M., 2014, in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. `http://arxiv.org/abs/1312.6114`

Krizhevsky A., 2012, University of Toronto

Lotz J. M., Primack J., Madau P., 2004, AJ, 128, 163

Lukic V., Brüggen M., Mingo B., Croston J. H., Kasieczka G., Best P. N., 2019, MNRAS, 487, 1729

McInnes L., Healy J., Saul N., Grossberger L., 2018, The Journal of Open Source Software, 3, 861

Meert A., Vikram V., Bernardi M., 2015, MNRAS, 446, 3943

Meert A., Vikram V., Bernardi M., 2016, MNRAS, 455, 2440

Mendel J. T., Simard L., Palmer M., Ellison S. L., Patton D. R., 2014, ApJS, 210, 3

Naiman J. P., et al., 2018, MNRAS, 477, 1206

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

Nelson D., et al., 2015, Astronomy and Computing, 13, 12

Nelson D., et al., 2018, MNRAS, 475, 624

Nelson D., et al., 2019, Computational Astrophysics and Cosmology, 6, 2

Netzer Y., Wang T., Coates A., Bissacco A., Wu B., Ng A. Y., 2011

Nuza S. E., Dolag K., Saro A., 2010, MNRAS, 407, 1376

Pawlik M. M., Wild V., Walcher C. J., Johansson P. H., Villforth C., Rowlands K., Mendez-Abreu J., Hewlett T., 2016, MNRAS, 456, 3032

Pillepich A., et al., 2018a, MNRAS, 473, 4077

Pillepich A., et al., 2018b, MNRAS, 473, 4077

Reiman D. M., Göhre B. E., 2019, MNRAS, 485, 2617

Rodriguez-Gomez V., et al., 2019, MNRAS, 483, 4140

Ronneberger O., Fischer P., Brox T., 2015, CoRR, abs/1505.04597

Rowe B. T. P., et al., 2015, Astronomy and Computing, 10, 121

Salimans T., Karpathy A., Chen X., Kingma D. P., 2017, in ICLR.

Sautter R., Barchi P., 2017, Journal of Computational Interdisciplinary Sciences, 8

Schawinski K., Zhang C., Zhang H., Fowler L., Santhanam G. K., 2017, MNRAS, 467, L110

Schaye J., et al., 2010, MNRAS, 402, 1536

Schaye J., et al., 2015, MNRAS, 446, 521

Shafaei A., Schmidt M., Little J. J., 2018, CoRR, abs/1809.04729

Sijacki D., Springel V., Di Matteo T., Hernquist L., 2007, MNRAS, 380, 877

Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, MNRAS, 452, 575

Springel V., 2010, MNRAS, 401, 791

Trayford J. W., et al., 2017, MNRAS, 470, 771

Tuccillo D., Huertas-Company M., Decencière E., Velasco-Forero S., Domínguez Sánchez H., Dimauro P., 2018, MNRAS, 475, 894

Vogelsberger M., et al., 2014, MNRAS, 444, 1518

Walmsley M., Ferguson A. M. N., Mann R. G., Lintott C. J., 2019, MNRAS, 483, 2968

Weinberger R., et al., 2017, MNRAS, 465, 3291

Wen Z. Z., Zheng X. Z., An F. X., 2014, ApJ, 787, 130

Zingales T., Waldmann I. P., 2018, AJ, 156, 268

Zubko V., Dwek E., Arendt R. G., 2004, ApJS, 152, 211

van den Oord A., Kalchbrenner N., Vinyals O., Espeholt L., Graves A., Kavukcuoglu K., 2016a, in NIPS.

van den Oord A., Kalchbrenner N., Vinyals O., Espeholt L., Graves A., Kavukcuoglu K., 2016b, in Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16. Curran Associates Inc., USA, pp 4797–4805, `http://dl.acm.org/citation.cfm?id=3157382.3157633`

This paper has been typeset from a TEX/LATEX file prepared by the author.