

Quantifying non-parametric structure of high-redshift galaxies with deep learning

C. TOHILL,¹ L. FERREIRA,¹ C. J. CONSELICE,^{1,2} S. P. BAMFORD,¹ AND F. FERRARI³

¹*University of Nottingham, School of Physics & Astronomy, Nottingham, NG7 2RD, UK*

²*Jodrell Bank Centre for Astrophysics, University of Manchester, Oxford Road, Manchester UK*

³*Instituto de Matemática Estatística e Física, Universidade Federal do Rio Grande, 96203-900 Rio Grande, Brasil*

(Received December 16, 2020; Revised March 31, 2021; Accepted May 18, 2021)

Submitted to AJ

ABSTRACT

At high redshift, due to both observational limitations and the variety of galaxy morphologies in the early universe, measuring galaxy structure can be challenging. Non-parametric measurements such as the CAS system have thus become an important tool due to both their model-independent nature and their utility as a straightforward computational process. Recently, convolutional neural networks (CNNs) have been shown to be adept at image analysis, and are beginning to supersede traditional measurements of visual morphology and model-based structural parameters. In this work, we take a further step by extending CNNs to measure well known non-parametric structural quantities: concentration (C) and asymmetry (A). We train CNNs to predict C and A from individual images of $\sim 150,000$ galaxies at $0 < z < 7$ in the CANDELS fields, using Bayesian hyperparameter optimisation to select suitable network architectures. Our resulting networks accurately reproduce measurements compared with standard algorithms. Furthermore, using simulated images, we show that our networks are more stable than the standard algorithms at low signal-to-noise. While both approaches suffer from similar systematic biases with redshift, these remain small out to $z \sim 7$. Once trained, measurements with our networks are $> 10^3$ times faster than previous methods. Our approach is thus able to reproduce standard measures of non-parametric morphologies and shows the potential of employing neural networks to provide superior results in substantially less time. This will be vital for making best use of the large and complex datasets provided by upcoming galaxy surveys, such as Euclid and Rubin-LSST.

Keywords: Astronomy, Galaxies — Machine Learning — Deep Learning — High Redshift

1. INTRODUCTION

A galaxy's morphology is a useful indicator of its assembly, interaction and star-formation history. Morphological studies have therefore proven invaluable for tracing the evolution of the galaxy population over cosmic time. However, the faintness and small angular size of galaxies at high redshift ($z \gtrsim 1$) makes them difficult to classify in the same manner as those nearby. Cosmological dimming causes more subtle features, such as spiral arms, to rapidly disappear with increasing redshift,

leaving only the brightest galaxy components detectable (Barden et al. 2008).

Furthermore, the traditional Hubble sequence is of limited applicability at high redshift. At early times, higher rates of star-formation and merging increase the prevalence of more varied and irregular morphologies (Abraham et al. 1996a; Elmegreen et al. 2005; Conselice & Arnold 2009; Mortlock et al. 2013). For studies of distant galaxies we need to consider more general and robust approaches to characterising galaxy structure.

Galaxy structure can be studied using both parametric and non-parametric methods. Parametric approaches fit analytic models, such as the Sérsic profile (Sérsic 1963), to a galaxy's light distribution (e.g. Peng et al. 2002; Buitrago et al. 2008; Simard et al. 2011; Häußler et al. 2013; Robotham et al. 2017). Such para-

metric methods are valuable for classifying symmetrical Hubble-type galaxies. However, they break down for more irregular, peculiar-type galaxies, as they assume a smooth light distribution. Non-parametric methods make no such assumptions. They are therefore more applicable to the variety of galaxies seen in the more distant universe, such as those with ‘clumpy’ morphologies (Abraham et al. 1994; Noguchi 1998; Bershady et al. 2000).

Motivated by these considerations, a number of authors, including Abraham et al. (1994, 1996b), Schade et al. (1995), and Conselice (1997) focused on two such non-parametric parameters, the concentration (C) and asymmetry (A) of a galaxy’s light distribution. It has been shown that the concentration parameter correlates with the bulge-to-disk ratio (B/D) of a galaxy, while the asymmetry parameter is a good indicator of the merger history of the galaxy (Conselice 2003; Lotz et al. 2008; Nevin et al. 2019). Using these parameters they were able to separate galaxies into their morphological type based on their position in this $C - A$ plane. Conselice (2003) expanded on this by introducing a third parameter, the smoothness (S) of a galaxy’s light distribution, creating the CAS system which has become one of the most common non-parametric measures of galaxy structure. This system has since been used in many investigations of galaxy structure across a wide range of redshifts (e.g. Yagi et al. 2006; Hoyos et al. 2012). A variety of similar non-parametric statistics are also in use (e.g. Lotz et al. 2004b; Freeman et al. 2013).

With the imminent arrival of large imaging surveys from new facilities, such as the Euclid, Rubin and Roman telescopes, it is of paramount importance to look into the efficacy of existing methods for measuring galaxy structure. For example, parametric structural measurements are often very time-consuming to apply to large surveys. Non-parametric measurements are generally faster, but the algorithms are still typically applied to individual galaxies in series. While the problem is ‘embarrassingly parallel’, significant computational resources are required to measure large numbers of galaxies in a timely fashion. With the future of extragalactic astronomy moving to extremely large surveys, it is useful to explore more computationally efficient approaches.

One increasingly popular technique, which has already proved useful in a number of areas of astronomy (Frontera-Pons et al. 2017; D’Isanto & Polsterer 2018; Pearson et al. 2019), is machine learning. In particular, deep learning, utilizing neural networks, can apply sophisticated analyses to large datasets at a much faster rate than conventional methods (e.g. Tuccillo et al.

2018). Deep learning has been applied to the morphological classification of both nearby (Dieleman et al. 2015; Cheng et al. 2020) and distant galaxies (Huertas-Company et al. 2015; Ferreira et al. 2020). It has also been shown to be very effective at reproducing parametric structural measurements (Tuccillo et al. 2018). However, as yet, deep learning has not been applied to determine the non-parametric CAS parameters. Given the arguments above, this could be a highly valuable tool for studying the local and high-redshift galaxy population in the next generation of surveys.

In this work we therefore create neural networks capable of predicting concentration and asymmetry parameters from a galaxy’s image. (For now we neglect the smoothness parameter as it is more difficult to measure at high redshifts and needs a separate treatment.) We show that our networks are consistent with conventional algorithms in their output, and demonstrate reliable behaviour down to very low signal-to-noise ratios. Furthermore, we find that our trained network is able to analyse $\sim 10,000$ galaxies in under 1.5 seconds, much faster than convention methods, making it well-suited to the large number of galaxies in future surveys.

This paper is organised as follows. In §2 we introduce the imaging data used in this work and describe how the conventional CAS parameters are measured using the MORFOMETRYKA software (Ferrari et al. 2015). The pre-processing of the data and all data augmentation is detailed in §3.1 and 3.2. In §3.3 and 3.4 we describe the architecture and optimization of our neural networks. The resulting performance of these networks is demonstrated through a number of tests in §4, concluding with a brief summary in §5.

2. DATA

2.1. CANDELS Fields

All of the images used in this project were taken with the Wide Field Camera 3 (WFC3) of the Hubble Space Telescope (HST) as part of the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS). We use data from all 5 CANDELS fields: the Great Observatories Origins Deep Survey (GOODS)-North and GOODS-South fields, COSMOS, Extended Groth Strip (EGS) and Ultra-Deep Survey (UDS).

The CANDELS/Deep survey (5σ point-source limit $H = 27.7$ mag) covers an area of ~ 125 arcmin² with a resolution of $0.06''$ per pixel (Grogan et al. 2011; Koekoer et al. 2011). In total we have $\sim 150,000$ galaxy postage-stamp images. These galaxies have photometric redshifts covering $z = 0-7$, with many parameters already calculated, including star formation rates (SFR) (Duncan et al. 2019) and CAS values. The apparent

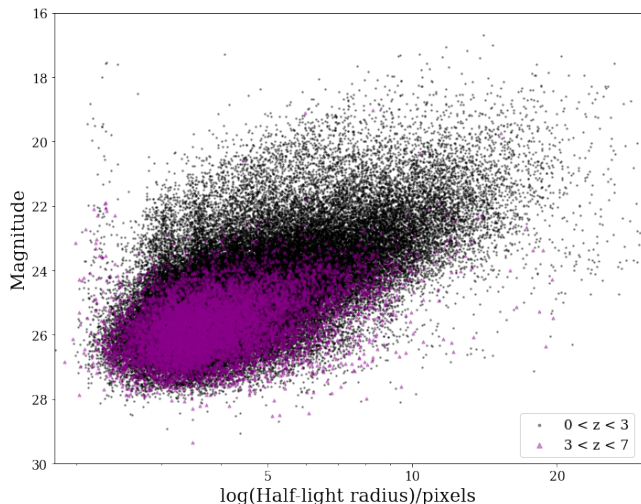


Figure 1. Distribution of the apparent magnitude in the H-band (F160W) vs half light radius, with a resolution of $0.06''$ per pixel, for the CANDELS galaxy sample used in this work.

magnitude–size distribution of our sample is shown in Fig. 1. In this paper we use imaging from the H -band (F160W), as it provides the most complete deep-coverage over all five CANDELS fields.

2.2. Concentration and asymmetry

As mentioned in the introduction, non-parametric methods have been used for many years to analyse the light distributions of distant galaxies, in order to better understand their structure (Conselice 2003; Lotz et al. 2004a; Sazonova et al. 2020). Such methods make very few assumptions, and so can be applied to peculiar and irregular galaxies as well as to classic Hubble types.

In this paper we utilise a subset of the CAS (Concentration, Asymmetry and Clumpiness) system as defined in Conselice (2003). This is a robust, non-parametric method for classifying galaxy structure, in a manner that is sensitive to their ongoing and past formation modes. In this paper, only concentration and asymmetry are considered. The concentration (C) is based on the measurement first established by Bershady et al. (2000), which was found to correlate with both galaxy bulge-to-disk ratio (B/D) and the effective radius of the bulge. This quantity is defined as

$$C = 5 \log_{10} \left(\frac{r_{80}}{r_{20}} \right). \quad (1)$$

where r_{20} and r_{80} are the radii containing 20% and 80% of the total light of the galaxy, respectively. The value of C is simply a measure of how concentrated the light in the central region is relative to the galaxy’s overall size. Galaxies with higher concentrations are typically

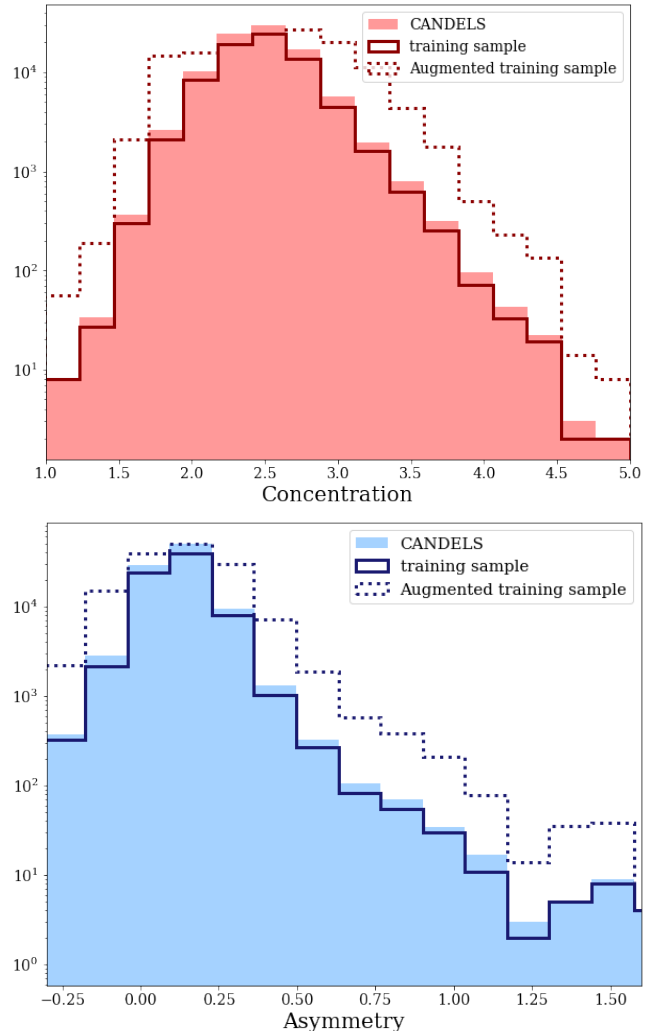


Figure 2. Distributions of asymmetry and concentration values for our selected sample of CANDELS galaxies, measured with MORFOMETRYKA. The solid line shows the selected training sample and the dotted line shows the training sample after our augmentation process.

ellipticals, early-type disks, and edge-on disks. In this manner, it shares similarities with the Sérsic index (Graham et al. 2005).

Galaxy asymmetry was first used in a basic form by Schade et al. (1995), when trying to classify distant galaxies imaged with HST. Asymmetry (A) is determined by rotating a galaxy 180° about its center and then subtracting from the original image. The centre of rotation is determined by an iterative process that finds the minimum asymmetry. Further algorithmic details are described in Conselice (2000) and Conselice (2003). The absolute values of the residuals are summed and normalized by the original galaxy flux. The resulting asymmetry contains a contribution from the background noise. This is accounted for by subtracting a background

term, determined by computing the asymmetry for small areas of sky near the galaxy. The basic calculation for the asymmetry is therefore given by

$$A = \frac{\sum |I - I_{180}|}{I} - A_{\text{bkg}}, \quad (2)$$

where I is the original galaxy image, I_{180} is the rotated galaxy image, and A_{bkg} is the background asymmetry (discussed further below).

Asymmetry can be used to identify a number of interesting galaxy classes, such as mergers and starburst galaxies (Conselice 1997, Conselice 2000, Bluck et al. 2012). These types of galaxies have a higher A value than regular ellipticals and disk galaxies, due to distributed areas of increased star formation.

2.3. *Morfometryka*

CAS measurements were originally obtained using IRAF. However, a more modern implementation, in Python, is provided by MORFOMETRYKA¹ (Ferrari et al. 2015). MORFOMETRYKA extracts a number of features from astronomical images, such as non-parametric morphology (including the CAS parameters) and Sérsic profiles. Full details of the software can be found in Ferrari et al. (2015), however we will briefly describe how the parameters used in this paper were calculated.

MORFOMETRYKA calculates the concentration, C , as explained in 2.2, with the exception that the factor of 5 in Eq. 1 is omitted. However, in order to remain consistent with previous studies, this factor was re-applied to our concentration values.

The asymmetry, A , is also determined as described in Section 2.2, by applying Eq. 2 within a $1.5 \times$ Petrosian-radius elliptical aperture centred on the galaxy. However, the background term in Eq. 2 is computed in a way that slightly deviates from the original CAS implementation. The standard approach utilises a single background region. Originally MORFOMETRYKA did not include the background asymmetry correction term. For our measurements, we construct a 10×10 pixel grid over the image area outside the galaxy segmentation map. We then measure the asymmetry for each cell in the grid, according to the first term of Eq. 2. Finally, we select the median asymmetry across all the cells as our background term, A_{bkg} . This ensures a robust and accurate background correction, improving upon the original background subtraction by eliminating the bias inherent in choosing only one background area. This is now incorporated into MORFOMETRYKA.

The errors on the concentration values are derived from those of the individual size measurements, which assume Poisson distributed fluxes. The typical error on C is ± 0.23 . The error on the asymmetry values were calculated using the method described in Conselice (2003). We find that the typical error on A is ± 0.072 for our sample.

We applied MORFOMETRYKA to all of the images in our dataset. We then select suitable galaxies for our analysis based upon the steps described in §3.1. The distributions of concentration and asymmetry values for our selected sample are shown in Fig. 2.

A subset of galaxy images were inspected to check that the measurements correspond to visual expectations. As can be seen from the top row of Fig. 3, galaxies with high C values appear compact and spheroidal. Such galaxies typically have low A values, reflecting a broad anti-correlation between C and A for normal Hubble types. Galaxies with high asymmetries are shown in the bottom row of Fig. 3. The contrast between the two sets of galaxies is clear, with high A galaxies appearing disrupted, or possessing features associated with merging, such as tidal tails and multiple bright sources. Note that high asymmetry galaxies span a range of concentrations. This is a reassuring reconfirmation of how these parameters have been seen to behave in past studies (Conselice et al. 2008, 2011).

3. METHOD

3.1. *Pre-Processing*

The initial images used in this analysis are 101×101 pixel cutouts, with the target galaxy in the center of each stamp. As we are only interested in training the network to predict the A and C values for the target galaxy, we need to remove any other sources. In order to remove neighbouring sources from the cutouts, the galclean algorithm (Ferreira et al. 2018) was utilised. This algorithm removes any non-central sources at a certain threshold above the background level. These masked areas are replaced with values sampled randomly from the background distribution to ensure they do not leave shapes which could be picked up by the network.

The majority of our galaxies have a half light radius of ~ 10 pixels. For computational efficiency, the individual galaxy images are therefore further reduced in size to 60×60 pixels, centered on the galaxy.

Since we are interested in measuring structure irrespective of overall galaxy brightness, we individually normalize our images. The pixel values of each image are rescaled so that the maximum pixel value of each image has a value of 1. This is also a standard pre-processing procedure for deep learning. It improves learning effi-

¹ The results in this paper are based on MORFOMETRYKA version 8.2

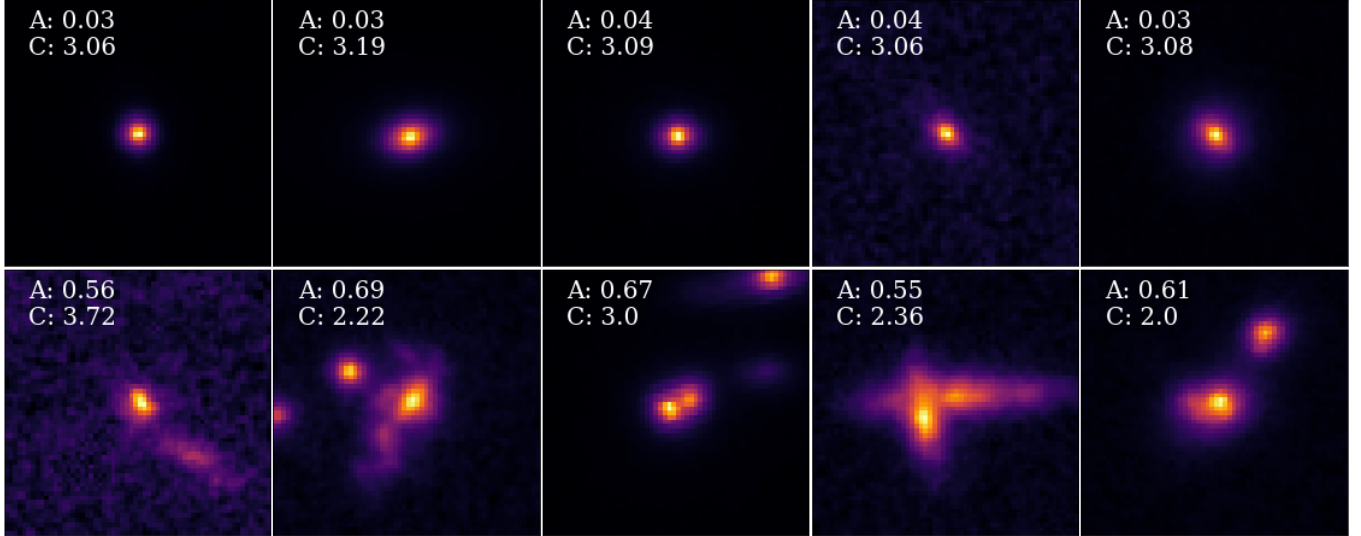


Figure 3. *Top row:* Images of the galaxies with high concentration values. These galaxies appear to be compact, spheroidal and have no close neighbours. These galaxies also typically have low asymmetry values, although there are some objects with high concentration and asymmetry. *Bottom row:* Images of galaxies with high asymmetry values. Many of these galaxies appear to be undergoing mergers and have tidal effects present, although there are occasional cases of line-of-sight projection. The concentration (C) and asymmetry (A) are indicated above each galaxy stamp.

ciency by ensuring that the inputs to the networks are compatible with the domain of the activation functions used within the model.

We wish to consider only reliable galaxy detections, for which structural parameters can reasonably be obtained. We therefore limit our sample to galaxies above a minimum signal-to-noise. We define the average signal-to-noise per pixel for each galaxy as

$$SNR_p = \frac{L_{\text{tot}}}{q\pi R_p^2 \sigma_{\text{sky}}} \quad (3)$$

where L_{tot} is the total integrated flux within the Petrosian region (with semi-major axis R_p), q is the axis ratio measured from the intensity distribution using the image moments, and σ_{sky} is the standard deviation of the sky background. By visual inspection we define a selection for our galaxy sample of $SNR_p > 2$. We also limit our sample to galaxies with $R_p > 5$ to ensure they are properly resolved.

Once these steps have been completed, we are left with 94,192 galaxy images with a median $SNR_p \sim 4.5$. These images were then split randomly into training (80%), testing (10%) and validation (10%) datasets to apply to our machine learning methods. With over 9,000 galaxies in each of our testing and validation sets, our performance estimates will be both accurate and precise.

3.2. Data Augmentation

Unbalanced datasets, whereby there are many more galaxies at one particular value compared with others,

can cause issues when dealing with both regression and classification problems in machine learning. The relative frequency of classes in the training set acts as a prior; the network may therefore be biased against identifying rare cases. In extreme circumstances, the network may fail to learn to identify rare cases at all. One way to combat this issue is by data augmentation (Shorten & Khoshgoftaar 2019).

Data augmentation is primarily used as a way of creating a larger training sample, which more finely samples the space of possible inputs. It is a form of regularisation and hence helps to prevent overfitting. By selectively expanding the size of the potential training set, augmentation can also help to balance the prevalence of different classes, while still using all of the input data.

Looking at Fig.2, there is a large imbalance in the CAS values for our sample, such that very high asymmetries are not common, nor are very low or high concentrations. As we want our model to be accurate across all concentrations and asymmetries, we selectively apply augmentation to create a more balanced training set. That is, we need to supplement the images that occupy the parameter space where there are few galaxies. For the range of C or A values where there are around half the number of images compared to the median value, we rotated each image by 90° once. Where there are relatively fewer images, we apply a greater variety of augmentations: rotating by 90° 3 times and mirroring along both axes. These images were then shuffled and added to the training set. After data augmentation, our

training sample increases in size from 75,353 to 141,453 images.

3.3. Convolutional Neural Networks

The purpose of this project is to efficiently and robustly predict CAS values of a galaxy from an image. We chose to implement a Convolutional Neural Network (CNN), as these are known to perform well when dealing with spatial structured data. CNNs are made up of convolutional layers, which are able to extract features from images by applying multiple filters (convolutional kernels) to the image. Individually, these filters can detect simple features. However, successive layers act hierarchically, identifying increasingly complex patterns. One major advantage of CNNs for image classification problems is the fact that they are able to exploit the spatial structure of the data which in turn reduces the number of parameters and allows the recognition of location invariant features.

CNNs were first popularised for image recognition/classification problems with the creation of LeNet-5 (Lecun et al. 1998), a network trained to classify handwritten digits. From this, CNNs have been applied in a range of fields, addressing a number of different problems and are becoming increasingly popular in astronomy.

CNNs were first utilised for galaxy classification by Dieleman et al. (2015) using data from the Galaxy Zoo project (Willett et al. 2013). While many others had applied different machine learning (ML) techniques to address this problem (e.g., Storrie-Lombardi et al. 1992; Naim et al. 1995; Huertas-Company et al. 2008; Banerji et al. 2010), these all required an earlier step of extracting features (often including CAS parameters or similar) from the images. The advent of CNNs provided a technique for efficiently extracting high-quality information directly from images. CNNs have since seen wide usage in extra-galactic astronomy, including morphological classification (e.g., Domínguez Sánchez et al. 2018; Cheng et al. 2020; Barchi et al. 2020), performing photometry (Tuccillo et al. 2018; Boucaud et al. 2020), and estimating merger rates (Ferreira et al. 2020).

There are many factors to consider when choosing the optimum architecture for a network. Many early studies based their architecture on previous studies (Huertas-Company et al. 2015; Domínguez Sánchez et al. 2018; Aniyon & Thorat 2017), trial-and-error (Dieleman et al. 2015; Feinstein et al. 2020), and arbitrary choices. However, there are a number of optimisation techniques that allow these choices to be optimised in a more satisfactory manner for the problem at hand. The variety of network

architectures we consider, and our method for selecting from these, are described in the following section.

To evaluate how well our networks are performing we compute the mean absolute error (MAE) and root mean squared error (RMSE) of the network’s predictions. The RMSE metric also serves as our loss function. The MAE is simply a measure of the average magnitude of error between the network’s prediction and the expected result,

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (4)$$

where n is the number of samples, y_i is the expected value and \hat{y}_i is the network’s prediction. The RMSE is similar to the MAE, but it is more sensitive to large errors and so can indicate if there are many outliers present. It is calculated as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (5)$$

3.4. Bayesian Optimization

The various choices that must be made before training a network can be considered as hyperparameters. These include aspects of the network architecture, such as the number of convolutional layers and number of filters in each layer, and of the training, such as the update algorithm, learning rate and batch size. Varying these choices can significantly alter the performance of the trained network. The problem of determining which combination of hyperparameters will be best suited to a given problem typically involves a trial and error process, which is often only partially explored, or entirely neglected, resulting in a non-optimal solution.

To avoid this, many optimization techniques have been developed, from simplistic random or grid-based searches (Bergstra et al. 2011), to more advanced techniques such as random forests (Hutter et al. 2011). The aim of these techniques is to find the optimum hyperparameters that will minimise the average loss. Traditionally, these techniques can be computationally expensive, as each variation in the hyperparameters results in a new version of the network which must be trained and then evaluated. Bayesian Optimisation (Snoek et al. 2015) provides a more efficient solution: a record of past evaluation results are kept and used to form a probabilistic model, which the method builds upon, reducing the time to converge on a optimal model.

Our networks comprise a number of convolutional blocks, between 1 and 3, with each block having either 1 or 2 convolutional layers. Each convolutional layer in a block has the same number of filters between 8 and 256 in powers of 2. The kernel sizes are all

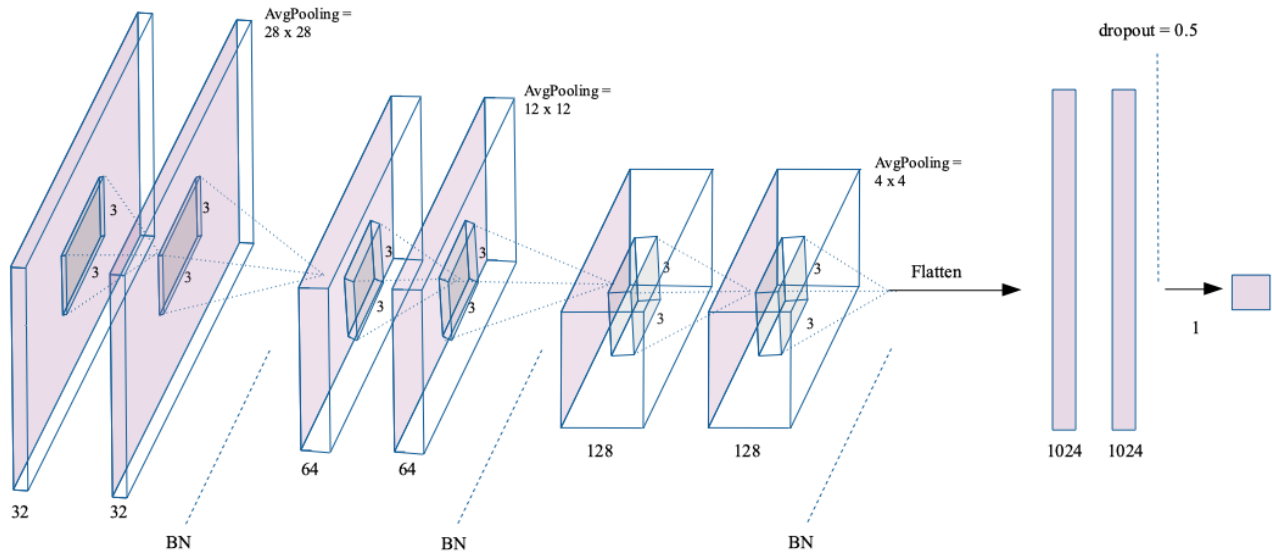


Figure 4. Architecture of the asymmetry network after optimization of its hyperparameters. This network takes input images of size 60×60 pixels, followed by 3 blocks, each containing 2 convolutional layers with 32, 64 and 128 features, respectively. Each block is followed by an average pooling layer of size 2 and a batch normalisation layer. Finally there are 2 fully-connected layers of size 1024 each, which is followed by dropout of 0.5 before the output value, i.e the network’s prediction of the asymmetry.

fixed to 3×3 . Each convolutional block is followed by a `BatchNormalisation` layer and an `AveragePooling` layer of fixed (2×2) size. Originally we started with `MaxPooling` layers, however we found that the networks’ performance improved when using `AveragePooling` layers. (Similar behaviour was found by [Pasquet et al. 2019](#) when analysing SDSS images.) Following the convolutional blocks we add some fully connected layers, with their number and size as hyperparameters. The number of fully connected layers ranges from 1 to 4, with each layer having the same number of filters between 128 and 1024 in powers of 2. We include a dropout layer before our output layer as a form of regularisation, allowing the dropout rate to vary as another hyperparameter. The dropout rate is allowed to vary continuously between 0.25 and 0.60. The activation function is fixed to the common `ReLU` ([Nair & Hinton 2010](#)) non-linearity.

When training a network, an optimization algorithm adjusts the weights to minimise the cost function. With a plethora of optimizers now available, we have included the choice as a hyperparameter, selecting from a pool of those most commonly used, we include Adam, Adadelata, RMSprop, SGD and Adamax. We also set the learning rate as a hyperparameter, where we evaluate 5 values, 0.001, 0.005, 0.01, 0.05 and 0.1.

The parameters we defined as hyperparameters and their optimised values are displayed in Table (1).

Hyperparameter	Optimum value	
	Asymmetry	Concentration
batch size	512	512
convolutional blocks	3	3
conv. layers per block	2	2
fully-connected layers	2	2
fully-connected layer size	1024	512
number of filters	32	64
optimization	Adamax	Adam
learning rate	0.001	0.001
dropout	0.50	0.55

Table 1. Summary of the hyperparameters selected by the Bayesian Optimization technique.

Each network was trained for a maximum of 300 epochs, but we applied ”early stopping” to halt the training when the validation loss had converged, which was typically after ~ 100 epochs.

Our Bayesian Optimization was carried out using the `GPyOpt` python package ([The GPyOpt authors 2016](#)), with the aim to minimise the RMSE of the networks. Each network created during the optimization was trained and validated using the samples defined in §2. The MAE, RMSE and the Pearson coefficient were monitored for each iteration in the optimization. The network that had the lowest MAE and RMSE was selected as the optimum architecture for our network.

The architecture of the CNN selected for our asymmetry network is shown in Fig.4. To ensure that the choice of optimum architecture is robust, we retrain multiple times, and compare the variation in the loss to the variation observed between different networks. The variation in MAE for the optimum asymmetry network is quite stable and varies by ~ 0.001 . Comparing the top 10 network architectures, we find that the MAE varies by ~ 0.002 . The hyperparameters of these networks are quite similar, although the number of fully connected layers varies between 1 and 2, the dropout rate between 0.47 and 0.56, and the optimizer varies between Adamax and Adadelta. These parameters are not as significant in determining the optimum network.

The selected concentration network has a similar architecture, with some slight variations. The MAE loss variation across different training runs is ~ 0.002 , and the MAE of the top 10 architectures vary by ~ 0.003 , very similar to above. Looking at the variation in the architectures which give equivalent performance, we see that the number of fully connected layers varies between 1 and 2 layers, the batch size between 256 and 512, the dropout rate from 0.3 to 0.6, and the number of convolutional blocks varies between 2 and 3.

Following our use of Bayesian Optimisation and the above tests, we can be confident that our final selected networks are well-optimised. However, it is also reassuring that the performances we report below are robust to minor variations in network architecture and training.

4. RESULTS

As explained above, we train our networks on a subset (80%) of the images and select our optimal model by its performance on a validation set (10%). To then evaluate our selected, trained network, we use an additional independent test set (10%; 9,420 galaxies). This ensures that the metric used to evaluate the network’s performance is not biased by over-fitting the hyperparameters. We find that the networks perform similarly on both the test and validation sets: another indication that the selected network architecture is robust.

4.1. Model performance

The basic results of the concentration network after the hyperparameter optimisation can be seen in Fig.5. The network’s predictions correlate strongly with the MORFOMETRYKA measurements, with a MAE value of 0.15 with a RMSE of 0.21. This error is lower than the average error on the concentration measurements. This shows that our machine learning regression can measure these parameters just as well as the direct measurement method. Hence, the values from the network can be

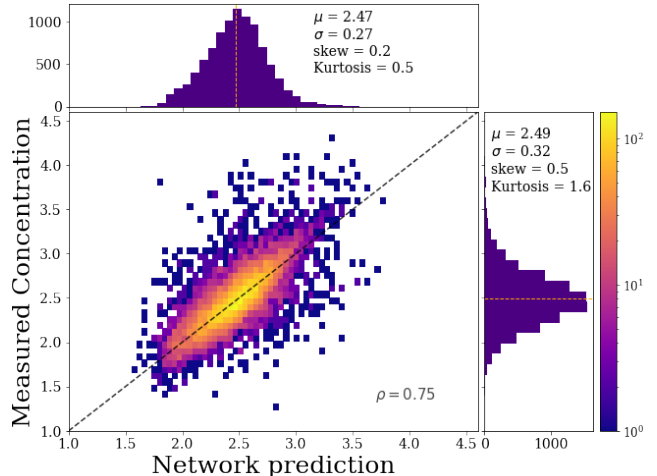


Figure 5. Our final network’s predictions of concentration values for the test set versus those measured by MORFOMETRYKA. The network does not see any of the images used in this comparison during its training. The MAE of the network was 0.15, with a RMSE of 0.21 (see §4). The Pearson correlation coefficient of 0.75 indicates that there is a strong correlation between the two.

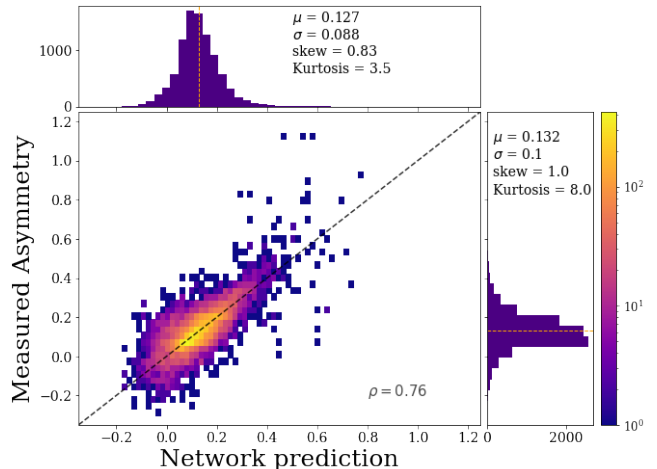


Figure 6. Our final network’s predictions of the asymmetries for the test set versus those measured by MORFOMETRYKA. The network did not use any of these images during its training. The MAE of the network was 0.045, with a RMSE of 0.065 (see §4). The Pearson correlation coefficient of 0.76, indicates that there is a strong correlation between the two.

utilised with a similar level of confidence as the original algorithm. However, the scatter does get larger at parameter values where there are fewer galaxies.

The results for the asymmetry network, again after the hyperparameter optimisation, can be seen in Fig.6. The network’s predictions for the asymmetries have a MAE of 0.045 with a RMSE of 0.065. As before, this error is lower than the average error on the asymmetry

measurements, showing that our networks can be used to reliably measure both the concentration and asymmetry values for a galaxy.

Overall, both networks perform well, achieving low residuals between the measured values and the network’s predictions. Looking at the images of galaxies where there was a large difference between our networks and the MORFOMETRYKA-measured CAS values, we find that they are quite noisy, with $SNR_p < 4$. From this we decided to further investigate the impact of noise on both our network predictions and the directly-measured CAS values.

4.2. Impact of noise

Our networks’ ability to accurately predict CAS values is potentially dependent on the noise level in a given galaxy image. To investigate this, we consider how the residuals (network prediction – MORFOMETRYKA value) of each network depend on the signal-to-noise per pixel, SNR_p . This is shown in Fig.7. Confirming the results from Figs.5 & 6, we see that for galaxies with moderate and high SNR_p the residuals are close to zero. The random scatter is also fairly constant with SNR_p , indicating that our networks are reliable across a broad range of SNR_p . There are a small number of galaxies with large deviations at the higher SNR_p however, when inspecting these images we find that most contain another source in the image that was not removed by the galeclean algorithm. This could explain why the measurements for these galaxies from MORFOMETRYKA and our networks varied. Within the low SNR_p regime we find a slight bias where the networks, on average, under-predict the values measured by standard algorithms.

The origin of this systematic trend at low signal-to-noise is interesting. Our networks have been trained to reproduce the measured values, and are clearly doing so in the majority of cases. So why the deviation at low SNR_p ? This could be seen as a failure of our model to capture the details of the measurements. On the other hand, we apply regularisation and optimise the hyperparameters to avoid over-fitting, with the aim of producing a generally applicable model, capable of accurate measurements for a wide variety of images. One optimistic possibility is that our networks are able to learn a model which is more robust than the regular methods. This is not inconceivable, since the regular methods must make a series of algorithmic ‘decisions’ (masking, fitting elliptical isophotes, recentering, etc.). The networks, instead, consider all of these issues within a single ‘holistic’ calculation.

In order to determine if this low signal-to-noise trend is a bias in our networks or in the standard algorithm (as

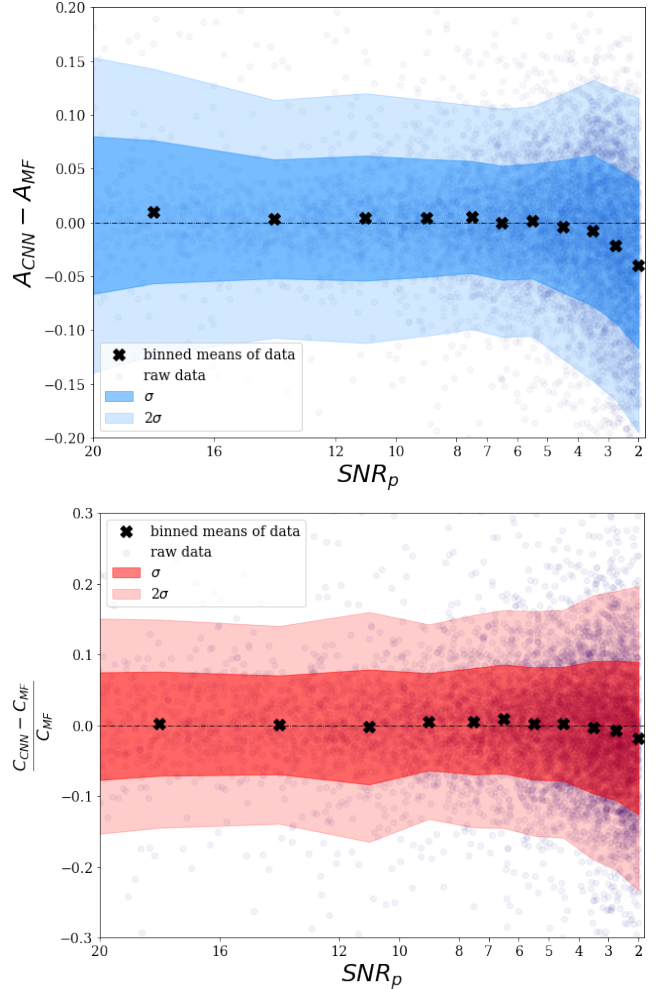


Figure 7. *Top:* The residuals of the asymmetry network and the measured values from MORFOMETRYKA versus the SNR_p of each galaxy image. *Bottom:* The fractional residuals between the concentration network and the measured values from MORFOMETRYKA versus the SNR_p of each galaxy image. In both panels, black points represent the means plotted for each bin with the darker shaded region representing ± 1 standard deviation and the lighter shaded region shows the area containing 2σ . The trend indicates that there is a bias at low SNR_p , where the networks will, on average, slightly under-predict the values measured by the standard algorithms.

implemented in MORFOMETRYKA), we investigate how noise impacts these two approaches in an independent manner. For this test, we select a sub-sample of 622 high SNR_p galaxies ($SNR_p > 10$), with low asymmetry residuals ($|\Delta A| < 0.01$) from the validation set. These galaxies also have low residuals in their concentration values. These galaxies are chosen as both the network and MORFOMETRYKA predicted these galaxies to have similar parameters, and hence we can assume these to be the true values for the purpose of this test.

We then produce versions of each galaxy image with varying SNR_p values. To do so, we first measure the mean and standard deviation in the background of the original galaxy image, then create an image with corresponding Gaussian noise. To this simulated background image, we add the original image with the overall flux scaled, such that we achieve our desired SNR_p . Finally, the image is normalized in the usual manner (§3.1). Except for the variation in SNR_p , each galaxy image remains identical to its original version. An example of these simulated noisy images can be seen in Fig.8.

The asymmetry and concentration values for these galaxies are then re-measured at each SNR_p , using both MORFOMETRYKA and our trained networks. The variation from the values measured in the $SNR_p = 10$ image is plotted as a function of decreasing SNR_p in Figs.9 & 10.

For asymmetry, it can be seen that at both high and moderate SNR_p the values recovered by our network are very similar to the ‘true’ values. For $SNR_p > 5$ the recovered values vary with an average standard deviation of 0.025, reflecting the uncertainties due to shot noise.

Furthermore, this scatter is significantly lower for our network than the standard algorithm. The average scatter in the asymmetry measurements, at $SNR_p \geq 5$, is 0.025 compared to 0.037 for MORFOMETRYKA. Since we have already seen that our networks accurately recover MORFOMETRYKA measurements, this suggests that our network is using information in these moderately-noisy images that is not utilized by the MORFOMETRYKA algorithm.

At low SNR_p , we find a bias present in both the network and MORFOMETRYKA, such that the A values are, on average, overestimated. However, it can be seen from Fig.9 that MORFOMETRYKA has a larger bias at these low SNR_p , indicating that our network is slightly more accurate in the low SNR_p regime. At $SNR_p = 3$ the bias in the network’s asymmetry values is 0.016 compared to 0.028 for MORFOMETRYKA.

While there is scatter in the individual measurements, on average our network is able to accurately estimate asymmetry, with little bias from the ‘true’ value, at a lower SNR_p than the original algorithm. This is useful for merger fraction estimates, especially at high redshift, as we can now include galaxy images down to a SNR_p as low as 3 while still retrieving unbiased measurements with our network. This means that we are able to measure reliable CAS parameters for more galaxies using deep learning and then we can with a direct measurement.

For concentration, plotted in Fig.10, we again see a difference in the variation of the measurements for

moderately-noisy images, with our network producing a significantly lower scatter than the standard algorithm. The average scatter in the concentration measurements at $SNR_p \geq 5$ are 0.04 for the network compared to 0.07 for MORFOMETRYKA. Both the network and MORFOMETRYKA slightly overestimate the concentration at $SNR_p \leq 5$.

We also investigate the ‘catastrophic’ fraction (f_c) of both MORFOMETRYKA and our network, that is the number of galaxies that fall outside of 2 sigma deviation from the mean. Again the network performs marginally better, with f_c at $SNR_p \geq 5$ being 4.3% for MORFOMETRYKA compared to 3.3% for our network.

Based on these results, we conclude that our deep-learning approach is performing at least as well as traditional measurements of non-parametric structure.

4.3. Impact of redshift effects

While the previous test examined how our networks fare with respect to noise alone, here we combine the effects of signal-to-noise and resolution to determine our networks’ performance for galaxies at high redshifts.

It is known that at higher redshifts, cosmological dimming and decreasing apparent size result in galaxies appearing more symmetric and less concentrated than they otherwise would (Conselice 2003). However, through the use of simulations, we can model and correct for this variation. We thus quantify the extent to which C and A values estimated by our networks are biased by these issues, and hence the level of any correction which should be applied when comparing galaxies at different redshifts.

To investigate how the performance of our networks varies with redshift, we take a sample of nearby galaxies, with reliably measured concentration and asymmetry values, and simulate how they would appear at higher redshift. For this test, we selected objects from the Frei catalogue of nearby galaxies (Frei et al. 1996). These are a well studied sample of regular, nearby galaxies, containing all Hubble types and with previously measured A and C values (Conselice 2003). We simulate the appearance of these galaxies as if they were observed at a range of redshifts, from $z = 0.1-7$.

There are a number of effects that need to be considered when artificially redshifting galaxies. The first effect we address is geometric scaling, whereby the apparent size of the galaxy will decrease when viewed at a higher redshift. We follow the same procedure as described in Conselice (2003) and de Albarnaz Ferreira & Ferrari (2018) to reduce the sizes of the galaxies to how they would appear at higher redshifts. Previous simulation work has kept the physical size of the galaxies

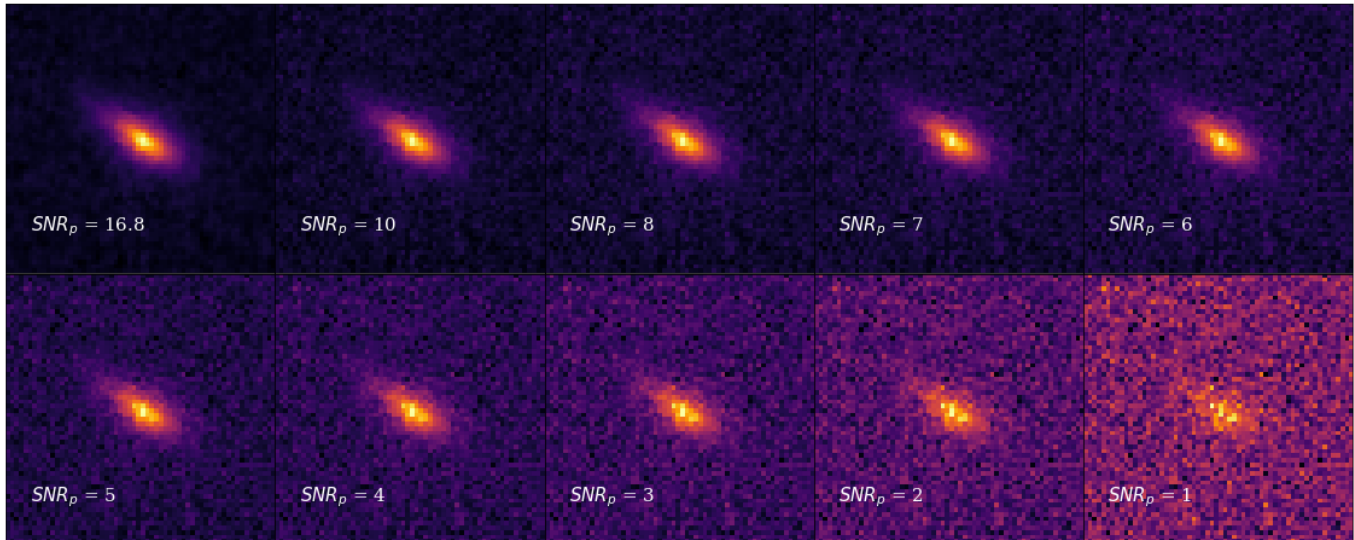


Figure 8. An example of our simulated noisy galaxy images. The top-left panel shows the original image, while the remainder show the same galaxy at different simulated SNR_p .

constant. However, it is well known that galaxies of a given stellar mass are intrinsically smaller at higher redshifts, reducing in size by a factor of ~ 5 between $z = 0$ and $z = 3$ (Trujillo et al. 2007; Buitrago et al. 2008). We therefore introduce size evolution to better represent the properties of high-redshift galaxies.

We use the size evolution determined by Whitney et al. (2019), which is based on the same H -band data from CANDELS GOODS North and South fields that we use in our training sample. They measured how the average physical Petrosian radius, R_p , changes with redshift for a mass selected sample, finding that it varies according to $R_p(z) = \frac{R_p(z=0)}{(1+z)^\beta}$, with $\beta = -0.97$. We therefore multiply the geometric scaling factor by this value to correct for the size evolution in our simulation.

After the (flux-preserving) geometrical scaling, we apply cosmological dimming Tolman (1930), according to

$$I(z) = \frac{I(z=0)}{(1+z)^4} \quad (6)$$

where $I(z)$ is the observed intensity. This is one of the major issues when detecting high redshift galaxies, as it introduces a bias such that only the brightest, most compact galaxies are detectable. The intrinsic brightness of galaxies, with a given stellar mass, varies with redshift. We therefore implement an evolution in the surface brightness of the galaxies as outlined in Whitney et al. (2020). They found that the correction for the intrinsic surface brightness follows

$$\mu(z) = \mu(z=0)(1+z)^\alpha \quad (7)$$

where $\mu = -2.5 \log(I)$ and $\alpha = -0.13$. The value of α was found to vary from -0.09 to -0.18 , but this will

not result in much variation in our results. The value of -0.13 was the value found for their size corrected sample.

To complete our simulations, the galaxies are convolved with the HST PSF in the H_{160} -band filter and placed in an actual CANDELS background.

We do not account for morphological or magnitude corrections. Instead, we test how a galaxy image would vary in restframe optical wavelengths, i.e. choosing appropriate observed filters for different redshifts. While we are currently only able to probe restframe optical up to $z = 3$ with HST, future surveys, such as JWST, will be able to probe up to $z = 7$. Furthermore, it has been found that the CAS parameters do not vary much between the UV and optical for star forming galaxies (Conselice 2003).

We select the brightest galaxies for this test, such that they are above $SNR_p > 2$ in all images out to $z = 7$, as this was the cut off used in our training sample. This leaves us with 100 out of the original 112 galaxies in the Frei sample. The original asymmetry values of the sample range from $0.01 < A < 0.48$, while the concentration measurements vary from $2.4 < C < 5.0$. The A and C parameters of each galaxy were remeasured at each redshift. We only consider $z = 0.2$ onward, since we need the whole galaxy to fit within a 60×60 pixel image, for input to our networks. An example of one of the redshifted galaxies is shown in Fig. 11.

The variations of the CAS values measured by our networks and MORFOMETRYKA are plotted against redshift in Fig. 12, with the full distributions at a sample of redshifts shown in Fig. 13 for comparison. As expected, at higher redshifts both methods measure the galaxies to be more symmetric and less concentrated than at

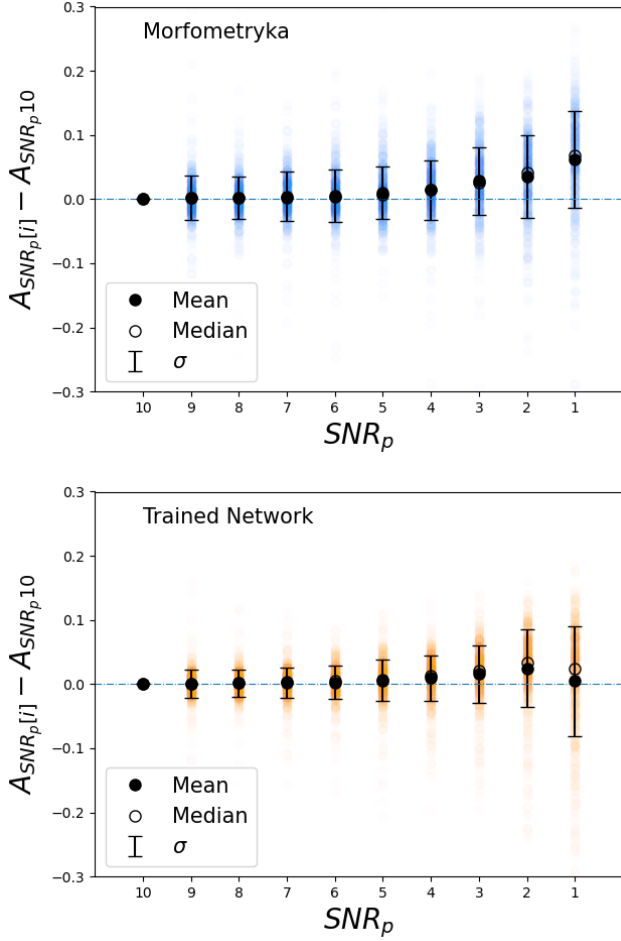


Figure 9. *Top:* Deviation in the asymmetry measurements from $SNR_p = 10$ for MORFOMETRYKA as the signal-to-noise is decreased in bins of SNR_p . At $SNR_p \leq 5$ there is a consistent bias, such that, on average, MORFOMETRYKA overestimates the values. Black points indicate the mean residual at each bin, with error-bars showing ± 1 standard deviation. The median deviations are shown by open points.

Bottom: As the top panel, but for our trained network. A similar trend can be seen at low SNR_p , but the scatter and systematic deviation is lower. This shows that the network is more stable and accurate than MORFOMETRYKA at these low SNR_p values.

$z = 0.2$. As the outer regions of a galaxy fade below the background noise, and their apparent size approaches the resolution limit of the PSF, they appear more symmetric, as has been found previously (Conselice 2003).

While the Frei sample of galaxies appear somewhat different to those the networks were trained on, we see that the networks still perform well, measuring values similar to MORFOMETRYKA. The average change in asymmetry at $z = 1.0$ compared to $z = 0.2$ is 0.074, which is similar to the average error on the high- z asymmetry measurements. This is important for merger esti-

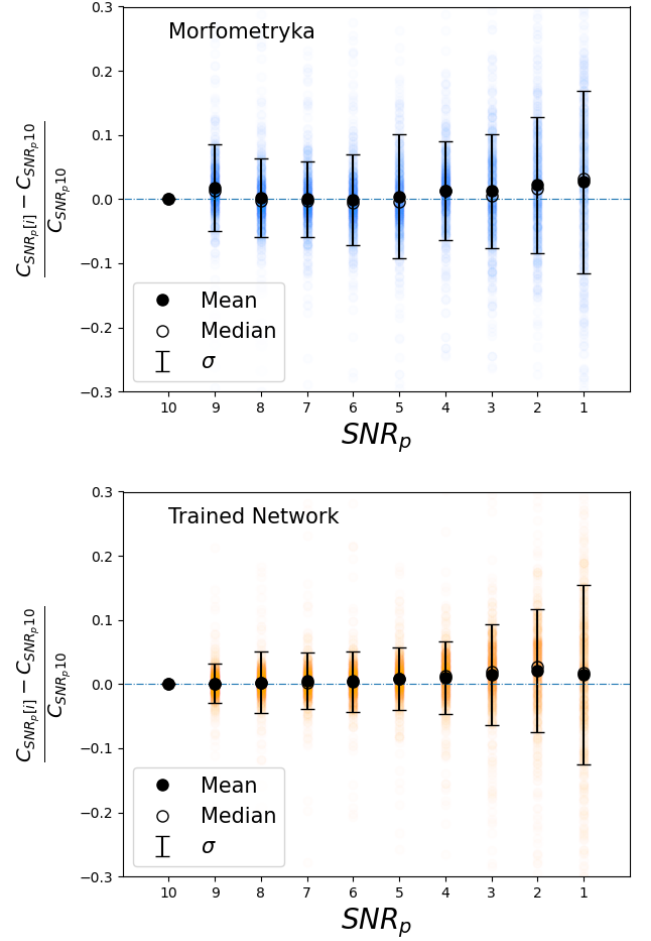


Figure 10. *Top:* Deviation in the concentration measurements from $SNR_p = 10$ for MORFOMETRYKA as the signal-to-noise is decreased in bins of SNR_p . Black points indicate the mean residual at each bin, with error-bars showing ± 1 standard deviation. The median deviations are shown by open points.

Bottom: As the top panel, but for our trained network. At $SNR_p \leq 5$ there is a slight bias, such that, on average, MORFOMETRYKA and the network overestimate the values of the sample. It can be seen that the C measurements are quite stable for $SNR_p \geq 3$. It can also be seen that the network produces a significantly lower scatter than MORFOMETRYKA.

mates, as this variation is small enough to avoid a merger appearing as a non-merger and vice versa. At redshifts higher than $z = 1$ the average variation is around twice the average error on the measurements. While we cannot be sure what the equivalent $z \sim 0$ A value of an individual galaxy would be, if investigating the galaxy population at high redshift, the average A value could be corrected.

We have included the equivalent results from Conselice (2003) in Fig.13. In their redshift test they in-

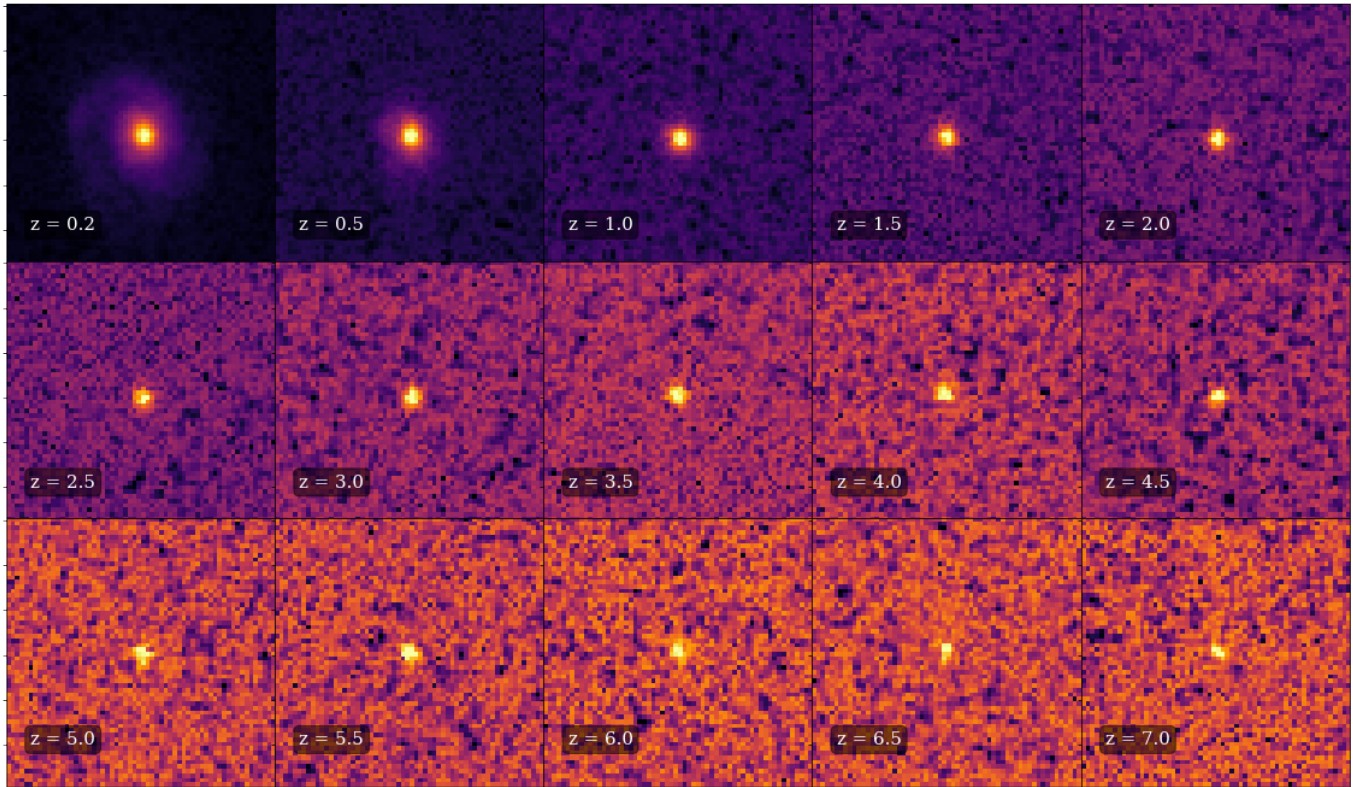


Figure 11. An example of our redshifted Frei galaxy images. In our redshifting technique we include effects due to cosmological dimming, luminosity evolution, size evolution and geometric scaling. Top left panel shows the reference image used in our tests. We start at $z = 0.2$ so that the whole galaxy fits within the 60×60 crop which is the input for our networks. The simulated redshift is indicated in each image.

cluded 82 galaxies from the Frei sample and investigated redshift effects up to $z = 3$. They implement no size correction in their redshifting procedure. Nevertheless, we see similar behaviour in both the bias and scatter, illustrating that the qualitative behaviour is insensitive to the details of our simulations.

These results indicate that our networks’ measurements at high redshift are better behaved than MORFOMETRYKA, with MORFOMETRYKA having a broader range of values. While both methods show systematic biases, the reduced scatter and lower prevalence of outliers suggests one could more confidently correct high- z C and A values based on the trained networks.

The differences seen in this section are greater than might be anticipated purely due to SNR from the results in Sec. 4.2. This indicates that there are other factors affecting MORFOMETRYKA more than the networks. In the SNR tests, simple uncorrelated Gaussian noise was added to the original images. However, in the artificial-redshifting procedure, the galaxies are placed in an apparently empty region of a real CANDELS image. The resulting background is more realistic, containing low-level structure due to pixel covariances introduced during the reduction and faint background

galaxies. This could affect the asymmetry and background calculations by MORFOMETRYKA, especially at higher redshifts where the background subtraction becomes more significant. Our networks, which we have shown to be less susceptible to noise, appear to be less sensitive to these effects. The result is more stable measurements, which may be applied at high-redshift with greater confidence.

Looking at the concentration measurements, we see that the network performs significantly better than the standard algorithm at recovering the original C values. The network variations are, on average, a factor of two lower than those measured by MORFOMETRYKA. The average change in the concentration measurements between $z = 1-3$ is 15% compared to 26% for MORFOMETRYKA. However, as mentioned above, such systematic trends could be corrected. More importantly, the scatter in the network’s measurements is somewhat lower and more consistent than those measured by the standard algorithm.

Our networks have been trained on CANDELS data, but successfully applied to data with a simple noise-degradation, and to artificially-redshifted, ground-based data. The individual images cover a very wide variety of

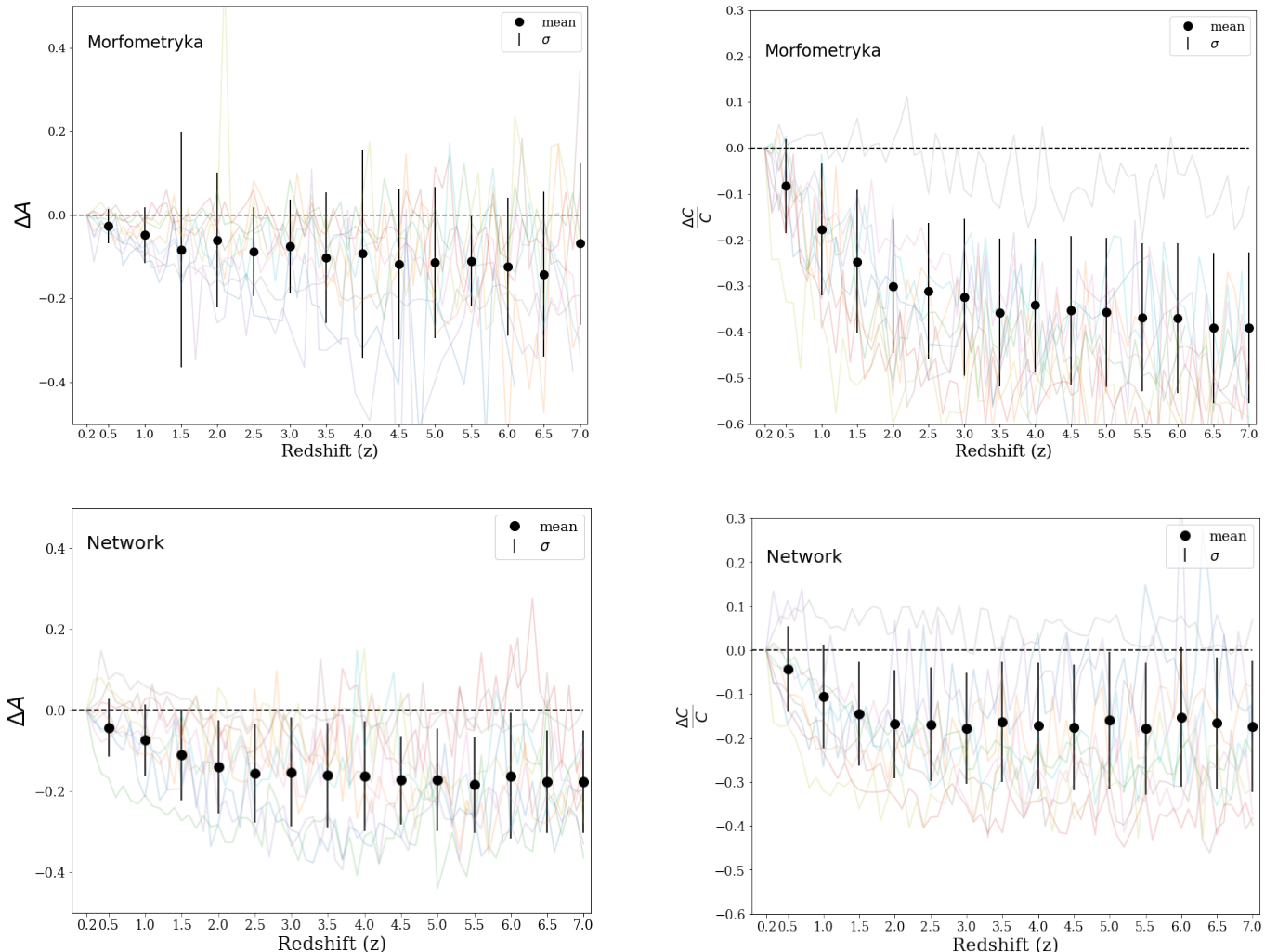


Figure 12. Variations in the C and A measurements from our trained networks and MORFOMETRYKA as a function of simulated redshift. Our simulations include geometric scaling (with accompanying reduction in resolution), size evolution, cosmological dimming and luminosity evolution. The original asymmetry measurements range from $0.01 < A < 0.48$, while the concentration measurements vary from $2.4 < C < 5.0$. There are a total of 100 galaxies from the Frei sample in this analysis. The variations for a sample of individual galaxies in the sample are shown by faint lines. These span a range of morphological types and initial asymmetry and concentration values. Black points indicate the mean variation and the error bars show the standard deviation of the individual measurements.

appearances, and yet we recover reliable measurements across our test set. This indicates that our networks are not particularly sensitive to the details of the observations. They can be applied to roughly comparable datasets with similar performance to standard methods, without the need for retraining.

The primary reason for this flexibility, is that both C and A measurements are determined from an image alone, without requiring any other information, such as the PSF, noise characteristics, etc. The network has learned to calculate a statistic from the image pixel values, irrespective of the observational details. It is therefore expected, but still pleasing to see, that the networks

remain accurate when applied to a wide variety of images.

4.4. Computational efficiency

We now briefly turn to the efficiency of our CNNs, compared to CAS measurements using conventional measurements. Running both our trained networks and MORFOMETRYKA on a single computational core, for comparison, our CNNs are able to produce measurements $\sim 3,000$ times faster. However, for a modern workstation, containing a single high-end consumer GPU (e.g. an NVIDIA GeForce GTX 1080 Ti) and 16 CPU cores, the results are even more striking. On such a system, our trained networks can analyse $\sim 10,000$

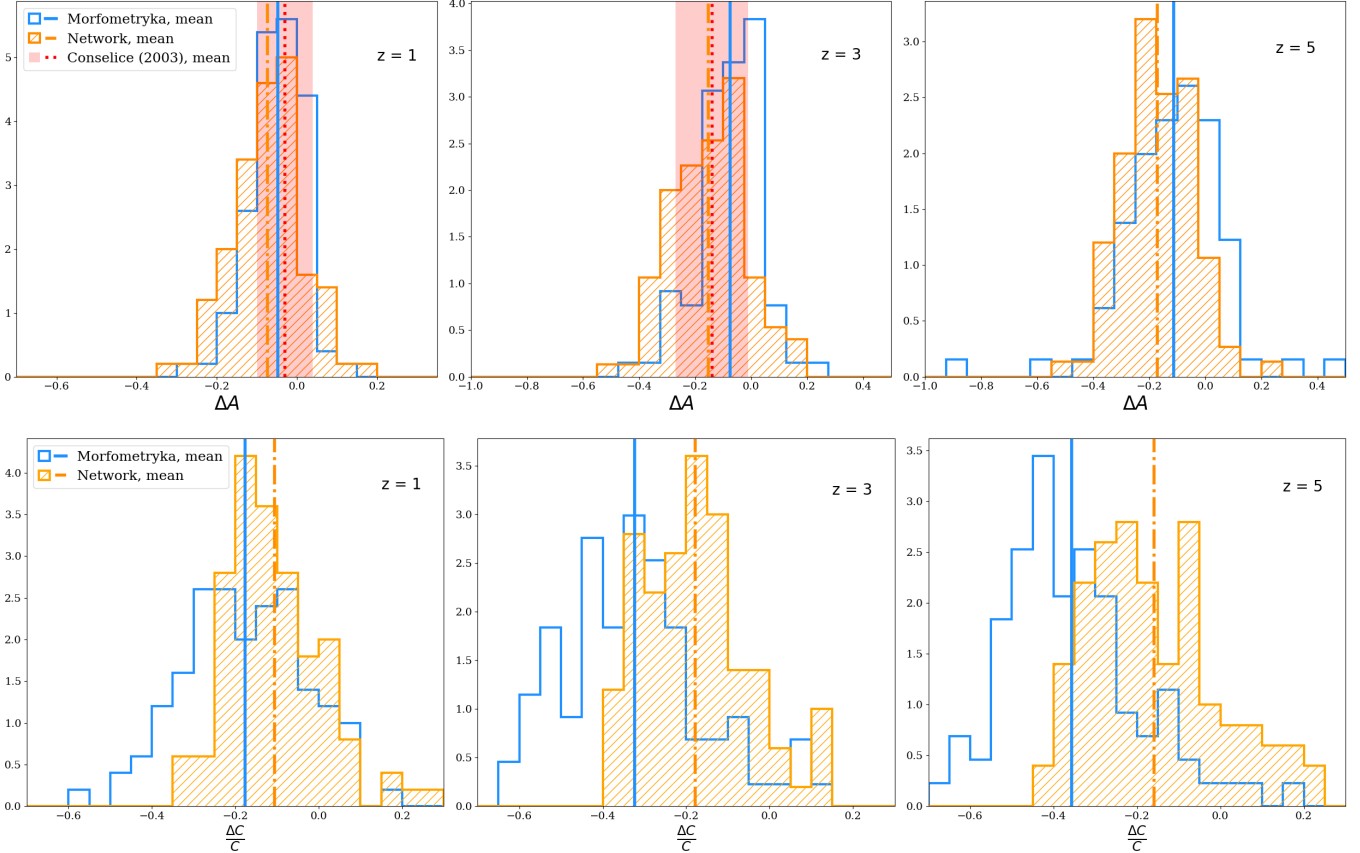


Figure 13. Investigating the effect of redshift on the asymmetry (*top*) and concentration (*bottom*) measurements from both our networks and MORFOMETRYKA. Plots show the distributions of the deviations in the measurements for a sample of redshifts. It can be seen that there are some extreme outliers in the measurements made by Morfometryka, especially with the asymmetry measurements. The range in the first asymmetry plot is reduced to better see both distributions. The original asymmetry measurements range from $0.01 < A < 0.48$, while the concentration measurements vary from $2.4 < C < 5.0$. There are a total of 100 galaxies from the Frei sample in this analysis. We also show the results from Conselice (2003) (the shaded region) demonstrating that our trends agree well with those found previously.

galaxies in under 1.5 seconds, while it would take 2 hours to perform these measurements using the MORFOMETRYKA code. Thus, our networks could measure all 1.5 billion resolved galaxies in the Euclid survey (Lauzeijs et al. 2011) on a single machine in a little over an hour. To do the same with MORFOMETRYKA would take several weeks on a 1000-CPU cluster! Even with highly-optimized software, using conventional algorithms would require significant time on a computing cluster.

In the previous section we have argued that our networks may be applied to other datasets without needing to retrain, providing the data characteristics are reasonably similar (which will be the case for any intermediate- to high-redshift galaxy surveys). However, should retraining be deemed necessary, this need not be an onerous process. In Sec. 3.4 we show that the network performance is consistent for moderate variations around the optimum. We expect that the selected hyperparameters will be a suitable choice for a variety of datasets. There

should be no need to rerun the Bayesian Optimisation process again.

Given the performance we see for our networks, only a few tens of thousands of galaxies would be required to retrain the network. Using the optimal architecture, fully training the network with 75,000 training examples takes only around 30 minutes. Transfer learning is also a possibility, but the training time should be no longer. In any case, the network training time is short compared to that required to prepare the training set, which itself is substantially faster than applying conventional methods to a large dataset.

A more general argument in favour of moving towards deep learning techniques for these kinds of calculations, is that there is potential for many, currently required, preparatory steps to be avoided. Pre-processing steps such as creating segmentation maps and cleaning neighbouring objects could, in principle, be performed by the network itself. In this paper we have not explored this,

and have instead applied our networks to the data prepared for Morfometryka. However, an indication of the networks’ robustness is provided by its stability when applied to artificially-redshifted galaxies. This could significantly reduce the computational and human time spent preparing the data to run these measurements.

It should be noted that MORFOMETRYKA performs a number of additional measurements that are complementary to those discussed in this work. However, one could train a network, in the same manner presented here, to predict these parameters. Indeed, this has already been done for Sérsic profiles (Tuccillo et al. 2018).

A further outstanding issue is that of uncertainties. We have not attempted to produce uncertainties on individual measurements output by our networks, beyond examining the scatter relative to Morfometryka. However, Pearson et al. (2021) give a detailed explanation of how the estimation of uncertainties can be incorporated into CNNs.

Extending our network to measure a wider variety of parameters, with uncertainties, is beyond the scope of this present work. However, we hope that this paper demonstrates the benefits for upcoming ‘Big Data’ surveys, such as Euclid and Rubin-LSST. Deep-learning has the potential to improve over conventional approaches in terms of the efficiency, accuracy and flexibility with which the next generation of surveys can be analysed.

5. SUMMARY

In this paper we trained two convolutional neural networks to perform concentration (C) and asymmetry (A) measurements based on individual galaxy input images. Our trained networks reproduce measurements by standard algorithms with an average absolute error on the C and A values of 0.15 and 0.045, respectively. These are lower than the average uncertainties on those measurements using conventional methods. Our networks can therefore be used to measure these quantities with a similar level of confidence to existing algorithms. Analysing these quantities for large samples of galaxies can provide an estimate of the merger fraction, and help us understand the transition from peculiar/irregular galaxies at high redshift to the well-defined Hubble sequence we observe locally.

We have shown how both our networks’ and MORFOMETRYKA’s measurements are impacted by noise, but find that our networks’ estimates are more stable in the low signal-to-noise regime, in terms of both lower scatter and systematic bias. By artificially-redshifting a sample of local galaxies from the Frei catalogue, we investigate trends in the measurements due to redshift effects. Again, we find that our networks produce measurements

with a lower level of random variation, compared to the conventional algorithms. While the measured A and C values are slightly biased at high-redshift, our networks and MORFOMETRYKA are both affected in similar manner, and consistent with behaviour seen previously (Conselice 2003). Furthermore, the systematic offsets are comparable to the random uncertainty on individual galaxy measurements, and so relatively minor.

Our trained networks are up to several thousand times faster than previous non-parametric measurement algorithms, presenting a substantial advantage for upcoming surveys. Our trained networks are made public with this work ². The future of extragalactic astronomy consists of ‘Big Data’ surveys, which will image billions of galaxies. Current state of the art computational methods for analysing these surveys will become impractical due to the computational resources and time they need. While detailed analyses will be required for certain measurements, machine learning techniques can replace many current algorithms. CNN-based approaches are more efficient and, as we have shown for measuring CAS parameters, can be more accurate and reliable than traditional measurements. Measuring non-parametric morphologies in upcoming galaxy surveys, including those by the Euclid, Rubin, and Roman observatories, will greatly benefit from the methods presented in this paper. In addition, the high accuracy of our CNN-based measurements make them equally suitable for use on smaller samples from deeper surveys, such as those by JWST.

Software: Astropy (Astropy Collaboration et al. 2018), MORFOMETRYKA (Ferrari et al. 2015), Tensorflow (Abadi et al. 2016), GPyOpt (The GPyOpt authors 2016)

ACKNOWLEDGMENTS

The authors would like to thank the Centre for Astronomy and Particle Theory of the University of Nottingham for providing the computational infrastructure needed to produce the networks used in this paper. CT acknowledges funding from the Science and Technology Facilities Council (STFC). LF acknowledges funding from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We thank the anonymous referee for their thorough review, which helped to significantly improve the presentation of this work.

² <https://github.com/cbtohill/CASNET>

REFERENCES

- Abadi, M., Agarwal, A., Barham, P., et al. 2016, arXiv e-prints, arXiv:1603.04467.
<https://arxiv.org/abs/1603.04467>
- Abraham, R. G., Tanvir, N. R., Santiago, B. X., et al. 1996a, MNRAS, 279, L47, doi: [10.1093/mnras/279.3.L47](https://doi.org/10.1093/mnras/279.3.L47)
- Abraham, R. G., Valdes, F., Yee, H. K. C., & van den Bergh, S. 1994, ApJ, 432, 75, doi: [10.1086/174550](https://doi.org/10.1086/174550)
- Abraham, R. G., van den Bergh, S., Glazebrook, K., et al. 1996b, ApJS, 107, 1, doi: [10.1086/192352](https://doi.org/10.1086/192352)
- Aniyan, A. K., & Thorat, K. 2017, ApJS, 230, 20, doi: [10.3847/1538-4365/aa7333](https://doi.org/10.3847/1538-4365/aa7333)
- Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123, doi: [10.3847/1538-3881/aabc4f](https://doi.org/10.3847/1538-3881/aabc4f)
- Banerji, M., Lahav, O., Lintott, C. J., et al. 2010, MNRAS, 406, 342, doi: [10.1111/j.1365-2966.2010.16713.x](https://doi.org/10.1111/j.1365-2966.2010.16713.x)
- Barchi, P. H., de Carvalho, R. R., Rosa, R. R., et al. 2020, Astronomy and Computing, 30, 100334, doi: [10.1016/j.ascom.2019.100334](https://doi.org/10.1016/j.ascom.2019.100334)
- Barden, M., Jahnke, K., & Häußler, B. 2008, ApJS, 175, 105, doi: [10.1086/524039](https://doi.org/10.1086/524039)
- Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. 2011, in Advances in Neural Information Processing Systems 24, ed. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Curran Associates, Inc.), 2546–2554. <http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf>
- Bershady, M. A., Jangren, A., & Conselice, C. J. 2000, AJ, 119, 2645, doi: [10.1086/301386](https://doi.org/10.1086/301386)
- Bluck, A. F. L., Conselice, C. J., Buitrago, F. o., et al. 2012, ApJ, 747, 34, doi: [10.1088/0004-637X/747/1/34](https://doi.org/10.1088/0004-637X/747/1/34)
- Boucaud, A., Huertas-Company, M., Heneka, C., et al. 2020, MNRAS, 491, 2481, doi: [10.1093/mnras/stz3056](https://doi.org/10.1093/mnras/stz3056)
- Buitrago, F., Trujillo, I., Conselice, C. J., et al. 2008, ApJL, 687, L61, doi: [10.1086/592836](https://doi.org/10.1086/592836)
- Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. 2020, MNRAS, doi: [10.1093/mnras/staa501](https://doi.org/10.1093/mnras/staa501)
- Conselice, C. J. 1997, PASP, 109, 1251, doi: [10.1086/134004](https://doi.org/10.1086/134004)
- . 2000, arXiv e-prints, astro.
<https://arxiv.org/abs/astro-ph/0012454>
- . 2003, ApJS, 147, 1, doi: [10.1086/375001](https://doi.org/10.1086/375001)
- Conselice, C. J., & Arnold, J. 2009, Monthly Notices of the Royal Astronomical Society, 397, 208
- Conselice, C. J., Bluck, A. F. L., Ravindranath, S., et al. 2011, MNRAS, 417, 2770, doi: [10.1111/j.1365-2966.2011.19442.x](https://doi.org/10.1111/j.1365-2966.2011.19442.x)
- Conselice, C. J., Rajgor, S., & Myers, R. 2008, MNRAS, 386, 909, doi: [10.1111/j.1365-2966.2008.13069.x](https://doi.org/10.1111/j.1365-2966.2008.13069.x)
- de Albernaz Ferreira, L., & Ferrari, F. 2018, MNRAS, 473, 2701, doi: [10.1093/mnras/stx2266](https://doi.org/10.1093/mnras/stx2266)
- Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441, doi: [10.1093/mnras/stv632](https://doi.org/10.1093/mnras/stv632)
- D’Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111, doi: [10.1051/0004-6361/201731326](https://doi.org/10.1051/0004-6361/201731326)
- Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, MNRAS, 476, 3661, doi: [10.1093/mnras/sty338](https://doi.org/10.1093/mnras/sty338)
- Duncan, K., Conselice, C. J., Mundy, C., et al. 2019, ApJ, 876, 110, doi: [10.3847/1538-4357/ab148a](https://doi.org/10.3847/1538-4357/ab148a)
- Elmegreen, D. M., Elmegreen, B. G., Rubin, D. S., & Schaffer, M. A. 2005, ApJ, 631, 85, doi: [10.1086/432502](https://doi.org/10.1086/432502)
- Feinstein, A. D., Montet, B. T., Ansdell, M., et al. 2020, AJ, 160, 219, doi: [10.3847/1538-3881/abac0a](https://doi.org/10.3847/1538-3881/abac0a)
- Ferrari, F., de Carvalho, R. R., & Trevisan, M. 2015, ApJ, 814, 55, doi: [10.1088/0004-637X/814/1/55](https://doi.org/10.1088/0004-637X/814/1/55)
- Ferreira, L., Conselice, C. J., Duncan, K., et al. 2020, arXiv e-prints, arXiv:2005.00476.
<https://arxiv.org/abs/2005.00476>
- Ferreira, L., Ferrari, F., & Griffiths, A. 2018, galclean: v1.0.0, v1.0.0, Zenodo, doi: [10.5281/zenodo.4004571](https://doi.org/10.5281/zenodo.4004571)
- Freeman, P. E., Izbicki, R., Lee, A. B., et al. 2013, MNRAS, 434, 282, doi: [10.1093/mnras/stt1016](https://doi.org/10.1093/mnras/stt1016)
- Frei, Z., Guhathakurta, P., Gunn, J. E., & Tyson, J. A. 1996, AJ, 111, 174, doi: [10.1086/117771](https://doi.org/10.1086/117771)
- Frontera-Pons, J., Sureau, F., Bobin, J., & Le Floch, E. 2017, A&A, 603, A60, doi: [10.1051/0004-6361/201630240](https://doi.org/10.1051/0004-6361/201630240)
- Graham, A. W., Driver, S. P., Petrosian, V., et al. 2005, AJ, 130, 1535, doi: [10.1086/444475](https://doi.org/10.1086/444475)
- Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35, doi: [10.1088/0067-0049/197/2/35](https://doi.org/10.1088/0067-0049/197/2/35)
- Häußler, B., Bamford, S. P., Vika, M., et al. 2013, MNRAS, 430, 330, doi: [10.1093/mnras/sts633](https://doi.org/10.1093/mnras/sts633)
- Hoyos, C., Aragón-Salamanca, A., Gray, M. E., et al. 2012, MNRAS, 419, 2703, doi: [10.1111/j.1365-2966.2011.19918.x](https://doi.org/10.1111/j.1365-2966.2011.19918.x)
- Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971, doi: [10.1051/0004-6361:20078625](https://doi.org/10.1051/0004-6361:20078625)
- Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8, doi: [10.1088/0067-0049/221/1/8](https://doi.org/10.1088/0067-0049/221/1/8)
- Hutter, F., Hoos, H. H., & Leyton-Brown, K. 2011, in Proceedings of the 5th international conference on Learning and Intelligent Optimization, LION’05 (Springer-Verlag), 507–523, doi: [10.1007/978-3-642-25566-3_40](https://doi.org/10.1007/978-3-642-25566-3_40)
- Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, 197, 36, doi: [10.1088/0067-0049/197/2/36](https://doi.org/10.1088/0067-0049/197/2/36)

- Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193
<https://arxiv.org/abs/1110.3193>
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proceedings of the IEEE, 86, 2278
- Lotz, J. M., Jonsson, P., Cox, T. J., & Primack, J. R. 2008, MNRAS, 391, 1137, doi: [10.1111/j.1365-2966.2008.14004.x](https://doi.org/10.1111/j.1365-2966.2008.14004.x)
- Lotz, J. M., Madau, P., Giavalisco, M., & Primack, J. 2004a, in American Astronomical Society Meeting Abstracts, Vol. 205, American Astronomical Society Meeting Abstracts, 163.07
- Lotz, J. M., Primack, J., & Madau, P. 2004b, AJ, 128, 163, doi: [10.1086/421849](https://doi.org/10.1086/421849)
- Mortlock, A., Conselice, C. J., Hartley, W. G., et al. 2013, MNRAS, 433, 1185, doi: [10.1093/mnras/stt793](https://doi.org/10.1093/mnras/stt793)
- Naim, A., Lahav, O., Sodre, L., J., & Storrie-Lombardi, M. C. 1995, MNRAS, 275, 567, doi: [10.1093/mnras/275.3.567](https://doi.org/10.1093/mnras/275.3.567)
- Nair, V., & Hinton, G. 2010, Proceedings of the 27th international conference on machine learning (ICML-10), Omnipress USA
- Nevin, R., Blecha, L., Comerford, J., & Greene, J. 2019, ApJ, 872, 76, doi: [10.3847/1538-4357/aafd34](https://doi.org/10.3847/1538-4357/aafd34)
- Noguchi, M. 1998, Nature, 392, 253, doi: [10.1038/32596](https://doi.org/10.1038/32596)
- Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26, doi: [10.1051/0004-6361/201833617](https://doi.org/10.1051/0004-6361/201833617)
- Pearson, J., Li, N., & Dye, S. 2019, MNRAS, 488, 991, doi: [10.1093/mnras/stz1750](https://doi.org/10.1093/mnras/stz1750)
- Pearson, J., Maresca, J., Li, N., & Dye, S. 2021, arXiv e-prints, arXiv:2103.03257.
<https://arxiv.org/abs/2103.03257>
- Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266, doi: [10.1086/340952](https://doi.org/10.1086/340952)
- Robotham, A. S. G., Taranu, D. S., Tobar, R., Moffett, A., & Driver, S. P. 2017, MNRAS, 466, 1513, doi: [10.1093/mnras/stw3039](https://doi.org/10.1093/mnras/stw3039)
- Sazonova, E., Alatalo, K., Lotz, J., et al. 2020, ApJ, 899, 85, doi: [10.3847/1538-4357/aba42f](https://doi.org/10.3847/1538-4357/aba42f)
- Schade, D., Lilly, S. J., Crampton, D., et al. 1995, ApJL, 451, L1, doi: [10.1086/309677](https://doi.org/10.1086/309677)
- Sérsic, J. L. 1963, Boletín de la Asociación Argentina de Astronomía La Plata Argentina, 6, 41
- Shorten, C., & Khoshgoftaar, T. 2019, Journal of Big Data, 6, doi: [10.1186/s40537-019-0197-0](https://doi.org/10.1186/s40537-019-0197-0)
- Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., & McConnachie, A. W. 2011, ApJS, 196, 11, doi: [10.1088/0067-0049/196/1/11](https://doi.org/10.1088/0067-0049/196/1/11)
- Snoek, J., Rippel, O., Swersky, K., et al. 2015, arXiv e-prints, arXiv:1502.05700.
<https://arxiv.org/abs/1502.05700>
- Storrie-Lombardi, M. C., Lahav, O., Sodre, L., J., & Storrie-Lombardi, L. J. 1992, MNRAS, 259, 8P, doi: [10.1093/mnras/259.1.8P](https://doi.org/10.1093/mnras/259.1.8P)
- The GPyOpt authors. 2016, GPyOpt: A Bayesian Optimization framework in Python,
<http://github.com/SheffieldML/GPyOpt>
- Tolman, R. C. 1930, Proceedings of the National Academy of Science, 16, 511, doi: [10.1073/pnas.16.7.511](https://doi.org/10.1073/pnas.16.7.511)
- Trujillo, I., Conselice, C. J., Bundy, K., et al. 2007, MNRAS, 382, 109, doi: [10.1111/j.1365-2966.2007.12388.x](https://doi.org/10.1111/j.1365-2966.2007.12388.x)
- Tuccillo, D., Huertas-Company, M., Decenci re, E., et al. 2018, MNRAS, 475, 894, doi: [10.1093/mnras/stx3186](https://doi.org/10.1093/mnras/stx3186)
- Whitney, A., Conselice, C. J., Bhatwadekar, R., & Duncan, K. 2019, ApJ, 887, 113, doi: [10.3847/1538-4357/ab53d4](https://doi.org/10.3847/1538-4357/ab53d4)
- Whitney, A., Conselice, C. J., Duncan, K., & Spitler, L. R. 2020, ApJ, 903, 14, doi: [10.3847/1538-4357/abb824](https://doi.org/10.3847/1538-4357/abb824)
- Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835, doi: [10.1093/mnras/stt1458](https://doi.org/10.1093/mnras/stt1458)
- Yagi, M., Nakamura, Y., Doi, M., Shimasaku, K., & Okamura, S. 2006, MNRAS, 368, 211, doi: [10.1111/j.1365-2966.2006.10144.x](https://doi.org/10.1111/j.1365-2966.2006.10144.x)