Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **515,** 5285–5305 (2022) https://doi.org/10.1093/mnras/stac2105
Advance Access publication 2022 July 29

Photometric redshifts from SDSS images with an interpretable deep capsule network

Biprateep Dey [1]★ Brett H. Andrews,[1] Jeffrey A. Newman,[1] Yao-Yuan Mao [2]† Markus Michael Rau[3,4] and Rongpu Zhou [5]

[1]*Department of Physics and Astronomy and PITT PACC, University of Pittsburgh, Pittsburgh, PA 15260, USA*
[2]*Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA*
[3]*Department of Physics, McWilliams Center for Cosmology, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[4]*High Energy Physics Division, Argonne National Laboratory, Lemont, IL 60439, USA*
[5]*Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*

Accepted 2022 July 22. Received 2022 July 22; in original form 2021 December 7

ABSTRACT

Studies of cosmology, galaxy evolution, and astronomical transients with current and next-generation wide-field imaging surveys like the Rubin Observatory Legacy Survey of Space and Time are all critically dependent on estimates of photometric redshifts. Capsule networks are a new type of neural network architecture that is better suited for identifying morphological features of the input images than traditional convolutional neural networks. We use a deep capsule network trained on *ugriz* images, spectroscopic redshifts, and Galaxy Zoo spiral/elliptical classifications of ∼400 000 Sloan Digital Sky Survey galaxies to do photometric redshift estimation. We achieve a photometric redshift prediction accuracy and a fraction of catastrophic outliers that are comparable to or better than current methods for SDSS main galaxy sample-like data sets ($r \leq 17.8$ and $z_{\rm spec} \leq 0.4$) while requiring less data and fewer trainable parameters. Furthermore, the decision-making of our capsule network is much more easily interpretable as capsules act as a low-dimensional encoding of the image. When the capsules are projected on a two-dimensional manifold, they form a single redshift sequence with the fraction of spirals in a region exhibiting a gradient roughly perpendicular to the redshift sequence. We perturb encodings of real galaxy images in this low-dimensional space to create synthetic galaxy images that demonstrate the image properties (e.g. size, orientation, and surface brightness) encoded by each dimension. We also measure correlations between galaxy properties (e.g. magnitudes, colours, and stellar mass) and each capsule dimension. We publicly release our code, estimated redshifts, and additional catalogues at https://biprateep.github.io/encapZulate-1.

Key words: methods: data analysis – methods: statistical – galaxies: distances and redshifts.

1 INTRODUCTION

Wide-field extragalactic sky surveys collect photometric or spectroscopic measurements to create three-dimensional maps of the Universe by measuring on-sky positions and redshifts of a variety of astronomical objects. These maps help us study the growth of the Universe and its large-scale structure over time by measuring various observable quantities as a function of redshift. For example, Hubble (1929) studied distances to nearby galaxies as a function of redshift to discover the expansion of the Universe and more recently, Riess et al. (1998) and Perlmutter et al. (1999) studied the relationship between luminosity distances of Type Ia supernovae and their redshifts to discover cosmic acceleration and hence dark energy. Detection of baryon acoustic oscillations (BAO) using large redshifts surveys (Cole et al. 2005; Eisenstein et al. 2005) similarly gave us another independent measurement of the cosmic acceleration and other parameters of the concordance model of cosmology. Cosmological redshifts are a proxy for the distance to extragalactic objects thereby allowing us to measure their intrinsic properties (like luminosity, mass, star formation rate, etc.) and enabling studies of the formation and evolution of galaxies. Accurate redshift measurements of satellite galaxies in the nearby Universe allow us to study the nature and distribution of dark matter and help us constrain models of galaxy formation and evolution. Redshift measurements also help with rapid identification of host galaxies of transient sources for follow-up as made evident by the recent discovery of gravitational wave sources with electromagnetic counterparts (Abbott et al. 2017).

Given the long exposure times required and the limited multiplexing of spectroscopic instruments, high precision spectroscopic redshifts (spec-z's) can only be measured for a tiny fraction of galaxies for which we have images. For example, it will be possible to measure spectroscopic redshifts for less than 1 per cent of the galaxies that will be used in the Rubin Observatory Legacy Survey of Space and Time (*LSST*) studies of galaxy evolution and cosmology (LSST Science Collaboration 2009). Because of this limitation, it will be necessary to infer redshift information using imaging data alone; the resulting measurements are called photometric redshifts or photo-z's. Accurate photo-z estimates along with well-

★ E-mail: biprateep@pitt.edu
† NASA Einstein Fellow.

calibrated uncertainties will be crucial to achieve the ambitious science goals set for the next generation of photometric surveys like *LSST*.

Most photo-z estimation methods involve finding a non-linear mapping between photometrically observed properties of galaxies (like apparent magnitudes and colours) and redshift. This is achieved either by fitting the observed photometry with redshifted templates of galaxy spectral energy distributions (SEDs; e.g. LePhare, Arnouts et al. 1999; Ilbert et al. 2006, 2009; BPZ, Benítez 2000; ZEBRA, Feldmann et al. 2006; EAZY, Brammer, van Dokkum & Coppi 2008; Phosphoros, Apostolakos et al. 2019; MAGPHYS, Battisti et al. 2019; Lee & Chary 2020) or using a machine learning (ML) based model trained on galaxies with spectroscopic redshifts to approximate this relationship. The optimal method generally depends on the amount and quality of data available and the scientific questions to be addressed. Template-based methods work well for deep, high-redshift surveys where the faintness of the galaxies and the broad redshift range covered makes it prohibitively expensive to collect large data sets. However, SED templates often rely on assumptions on galaxy physics (like star formation history or initial mass function), have incomplete coverage of the entire wavelength range and model dust attenuation poorly, all of which are significant sources of errors (Salvato, Ilbert & Hoyle 2019).

In contrast, in regimes with more complete training data like that provided by shallow low-redshift spectroscopic surveys, common statistical methods like linear regression (e.g. Connolly et al. 1995; Beck et al. 2016) or classical ML techniques such as decision trees and random forests (e.g. Carliles et al. 2010; Dalmasso et al. 2020; Zhou et al. 2021; Li et al. 2022), support vector machines (e.g. Wadadekar 2005; Jones & Singal 2017), K-nearest neighbours (e.g. Ball et al. 2008; Graham et al. 2018), self-organized mapping (e.g. Geach 2012; Carrasco Kind & Brunner 2014; Wright et al. 2020; Myles et al. 2021), Gaussian processes (e.g. Way et al. 2009; Almosallam, Jarvis & Roberts 2016), and simple neural networks (e.g. Firth, Lahav & Somerville 2003; Tagliaferri et al. 2003; Collister & Lahav 2004; Cavuoti et al. 2017; Razim et al. 2021) tend to outperform template-based methods (Hildebrandt et al. 2010; Euclid Collaboration 2020; Schmidt et al. 2020).

A challenge for photo-z estimation methods that take magnitudes and colours as inputs is that there is not enough information available to break various degeneracies in the colour–redshift relation. One way to break these degeneracies is to include information about morphology, orientation, surface brightness, ratios of magnitudes, or visual appearance in general (e.g. Stabenau, Connolly & Jain 2008; Jones & Singal 2017; D'Isanto et al. 2018; Gomes et al. 2018; Nakoneczny et al. 2021). A galaxy may appear red not just because its stellar population is intrinsically red but because it is a dusty edge-on spiral galaxy. Moreover, the fact that farther objects appear to be smaller and fainter to an observer also give us an additional piece of information to help break degeneracies. Most traditional methods for quantifying galaxy morphology, like ellipticity and Sérsic index, cannot fully encode all of the visual information that an image of a galaxy provides and hence methods that use images of galaxies directly as inputs (e.g. Pasquet et al. 2019; Hayat et al. 2021; Schuldt et al. 2021; Henghes et al. 2022) and rely on artificial neural networks are the current state-of-the-art.

Artificial neural networks are mathematical models, originally developed to mimic the logical operations of the human brain. The simplest unit of such a model (also called an artificial neuron) is a linear transformation of an input followed by some non-linear function (also called an activation function). Successive layers of such transformations arranged together form a deep neural network. The process of training such a model involves finding a set of parameters (also called weights) for these transformations which will minimize a loss function. The optimization is generally done using the back propagation algorithm (Lecun 1985; Rumelhart, Hinton & Williams 1986) or some optimizer based on it like the Adam optimizer (Kingma & Ba 2015). The simplest deep neural network architecture called multilayer perceptrons or fully connected (FC) networks use successive matrix and non-linear transformations to connect every input feature to an output. A sufficiently deep or wide fully connected network can be used to approximate any function (Cybenko 1989; Hornik, Stinchcombe & White 1989; Hornik 1991) and hence can be used to effectively predict photometric redshifts.

If the input data are images, then the number of trainable weights required for a fully connected neural network architecture becomes very large, making them very inefficient to train and prone to overfitting to the training data. Convolutional neural networks (CNNs; Fukushima & Miyake 1982; LeCun et al. 1989), on the other hand, perform convolution operations using filters whose parameters are learned. Since the same set of filters are reused by stepping across the input images, it reduces the number of trainable parameters. Moreover, each successive convolution layer can extract more complex features which in turn increases the model accuracy while reducing the complexity of the model. Various multilayered neural network architectures (i.e. deep neural networks; LeCun, Bengio & Hinton 2015) built using CNNs have been used to make state-of-the-art photo-z prediction algorithms as they can leverage the pixel level data to extract additional information thereby achieving even better prediction accuracy. Hoyle (2016) modified the ImageNet challenge-winning AlexNet (Krizhevsky, Sutskever & Hinton 2012) to *griz* images of ∼64 000 SDSS galaxies, finding comparable accuracy to the best tree-based classical ML algorithms. D'Isanto & Polsterer (2018) combined a CNN and a mixture density network to produce photo-z probability density functions (PDFs) generated using Gaussian mixture models and achieved comparable performance to existing efforts in the literature. As larger training data sets become available along with advances in graphical processing unit (GPU) hardware and associated software, training CNNs have become very efficient and currently form the backbone of most state-of-the-art photo-z algorithms. Pasquet et al. (2019) produced the current best photo-z's using a supervised algorithm for the SDSS Main Galaxy Sample, which consists of ∼500 000 *ugriz* images with spec-z's in the range $z = 0$–0.4. They applied an innovative deep CNN that included five inception modules (Szegedy et al. 2015, 2016) which use multiple filter sizes within the CNN operating at the same level rather than being stacked sequentially to capture information on different scales efficiently. Recently, self-supervised learning-based approaches have shown promising results on astronomical data sets (e.g. Sarmiento et al. 2021; Stein et al. 2021). Hayat et al. (2021) used a self-supervised training scheme paired with a ResNet50-based CNN (He et al. 2016) to achieve similar results but with less data. They pre-trained their network on a very large unlabelled data set to find similarities between different augmentations of the inputs and then fine-tuned the network to predict photometric redshifts. When fine-tuned using the whole SDSS main galaxy sample, they achieve state-of-the-art results.

Deep neural network-based methods are continuing to improve but have substantial limitations in terms of the interpretability of the features learnt from images and efficiency in training. Models with larger number of trainable parameters not only require more data and computational resources to train but also are prone to

overfitting. To alleviate some of these issues, we explore the use of a modern deep learning method called capsule networks (Hinton, Krizhevsky & Wang 2011) to jointly predict photo-z's and basic morphological types of galaxies (spiral/elliptical). Capsule networks are robust to rotations and invariant to viewpoint – a useful quality for analysing randomly oriented galaxies and require less training data and trainable parameters than CNNs because they generalize much better to novel viewpoints (Mazzia, Salvetti & Chiaberge 2021). Capsule networks also learn a low-dimensional representation of the input images, which provides us with a way to interpret the features learnt by the model.

In this work, we will focus on predicting photo-z point estimates but ideally we would like to quantify the uncertainty in our estimates by predicting full photo-z PDFs. However, producing properly calibrated photo-z PDFs remains extremely challenging. PDFs predicted by neural networks are often poorly calibrated (see e.g. Guo et al. 2017) and provide very misleading uncertainty estimates. Moreover, most methods currently used to check the quality of photo-z PDFs (like distributions of probability integral transform, etc.) focus on checking the calibration of the entire sample of PDFs (i.e. global calibration) rather than focusing on the calibration of individual PDFs (i.e. local/individual calibrations). Amaro et al. (2019) and Schmidt et al. (2020) show that such metrics can be optimized by pathological but non-physical photo-z PDFs. Zhao et al. (2021) show that global calibration of PDFs does not imply local calibration and proposes new diagnostics which may be used to check for local calibration. In a future paper, we plan to extend our methods and produce locally calibrated PDFs following the procedure described in Dey et al. (2021, 2022). That being said, the prediction errors on our photo-z point estimates are sufficiently small that we can safely use these estimates for studies of the evolution of galaxies, their connection with dark matter haloes, and the localization of transient sources. where photo-z PDFs are not strictly required.

The paper is organized as follows. In Section 2, we discuss the various data sets used in this work. In Section 3, we introduce the concept of capsule networks and explain our network architecture. In Section 4 we describe the process of training a multitask capsule network. In Section 5, we present our results for photo-z point estimates and compare our results with other similar works. We also provide interpretations of the features learnt by the capsule network in order to predict photo-z's. Lastly, in Section 6 we summarize our results.

2 DATA

2.1 SDSS imaging and spectroscopic redshifts

To train and test our models, we use the same pre-processed images and spectroscopic redshifts that were used by Pasquet et al. (2019) for their CNN-based photo-z estimation method and were generously made publicly available by the authors.[1] The data set contains 516 525 galaxies with de-reddened r-band petrosian magnitudes, $r \leq 17.8$, and spectroscopic redshifts, $z \leq 0.4$ selected from the 12th Data Release (DR12) of the Sloan Digital Sky Survey (SDSS; Gunn et al. 1998, 2006; York et al. 2000; Smee et al. 2013; Alam et al. 2015). The sample is mainly defined by the magnitude limit as the redshift limit removes only a few tens of galaxies. The 12th data release of SDSS was used as that was the most recent data release available when this work began. Moreover, there has not been any changes to the data for this particular set of galaxies since DR8, making all SDSS data releases since DR8 equivalent for our purposes.

For this set of galaxies, Pasquet et al. (2019) pre-processed the raw five band images after obtaining them from the 8th Data Release of SDSS (Aihara et al. 2011). They stacked and re-sampled the images to a common $64 \times 64 \times 5$ pixel grid centred on the spectroscopic target. The images were background subtracted and photometrically calibrated with the same zero-point (Padmanabhan et al. 2008; Blanton et al. 2011). No foreground/background objects were removed. Most of the galaxies had only one or two imaging frames per band, whereas galaxies in Stripe 82 (Jiang et al. 2014) had up to 64 imaging frames. So, the Stripe 82 galaxies which satisfy our magnitude and redshift cuts defining the parent sample have significantly less noise than the other images. The Stripe 82 galaxies form less than 4 per cent of the entire data set and can be used to check how amount of noise in the images affect our methods (see Section 5.2). All the galaxies in the data set are spatially resolved, so their sizes, surface brightnesses, morphologies in each band, and the presence of neighbouring and background galaxies provide additional information not captured in spatially integrated photometry. A detailed description of the image processing steps can be found in section 2.1 of Pasquet et al. (2019). The processed images along with their spectroscopic redshifts used in this work are publicly available.

2.2 Galaxy Zoo-1 morphological class labels

We use a deep capsule network to jointly predict the basic morphology of a galaxy along with its redshift. We use crowd-sourced morphological class labels of galaxies from the Galaxy Zoo-1 project (Lintott et al. 2011) to train our capsule network. Galaxy Zoo-1 labels galaxies as spirals (with various sub-classes), ellipticals, mergers, or stars-and-artefacts. The classifications are considered 'confident' only if the de-biased fraction of votes received for a class is greater than 0.8. Since the numbers of mergers and artefacts in the images of the SDSS-MGS are very low, we use the spiral and elliptical classes only. This gives us high-quality morphological classifications for 177 442 of the galaxies in our parent data set. We generate morphological class labels for the remaining 339 083 galaxies in our data set, using an iterative semisupervised system where we train a deep capsule network using the confident class labels and use it to generate the labels for all other galaxies (see Section 4.1 for details). Out of these 339 083 galaxies that do not have a confident classification, we obtain the fraction of votes received in Galaxy Zoo-1 for each class for 296 767 galaxies which we use to cross-check our results. For the remaining 42 316 galaxies, no morphology information was available since they did not pass some of the quality cuts imposed by Galaxy Zoo-1. We do not use these galaxies to asses the quality of our morphological class prediction and only their deep capsule network generated class labels are used for redshift prediction.

2.3 Catalogue of galaxy properties

To interpret the features learnt by our deep capsule network, we measure correlations between the low-dimensional encodings of the input images produced by the capsules and various other galaxy properties (see Section 5.3.3). For this purpose, we created a cross-matched catalogue of various observed and estimated physical properties for the galaxies in our data set.

[1] https://deepdip.iap.fr/#item/60ef1e05be2b8ebb048d951d

For all galaxies, we query their model magnitudes, composite model (cmodel) magnitudes,[2] and extinction due to Milky Way dust from Schlegel, Finkbeiner & Davis (1998) for each of the five SDSS photometric bands from the SDSS DR12 data base (Alam et al. 2015). We also query the velocity dispersion (σ_v) measured from the spectra. We use the extinction corrected cmodel magnitudes as a measure of the galaxy magnitudes and use extinction corrected model magnitudes to calculate the colours of the galaxies. We also query measurements of stellar mass ($M_\star$), star formation rate (SFR), and specific star formation rate (sSFR) from the Max Planck Institute for Astrophysics and the Johns Hopkins University (MPA-JHU) value-added catalogue[3] available as a part of SDSS DR12. These measurements are based on the methods developed in Kauffmann et al. (2003), Brinchmann et al. (2004), and Tremonti et al. (2004). For estimates of absolute magnitudes ($M_{u/g/r/i/z}$), we use the measurements from the New York University Value Added Galaxy Catalog[4] (NYU-VAGC; Blanton et al. 2005) for objects common between our data set and the NYU-VAGC within a tolerance of 1 arcsec. We also use measurements of Sérsic-index in the r band (n_r) and the corresponding 90 per cent light radius ($R_{90,r}$) from the NYU-VAGC as a proxy for a galaxy's size.

A small number of objects in our data set do not have matches with the external catalogues and there are also some measurements in these catalogues that are problematic. We only use the objects in our data set that have cross-matches with the external catalogues for each of the galaxy properties. We also remove measurements of any property which are more than five units of median absolute deviation (scaled to replicate Gaussian standard deviation) away from the median of that property. This step is done to remove the small number (< 1 per cent) of problematic measurements of galaxy properties that can affect our analysis.

3 CAPSULE NETWORKS

CNNs (Fukushima & Miyake 1982; LeCun et al. 1989) are currently the de facto standard for neural network architectures when the input data are images. They work by learning weights for a set of convolutional filters which extract useful features from the images. As the filters are reused by translating them across the input, CNNs have fewer trainable parameters compared to their fully connected counterparts and also invariant to small translations of the object of interest with respect to the background. Each successive layer of the deep network extracts more and more complex features in an hierarchical fashion. CNNs have been immensely successful in solving problems in computer vision (e.g. Krizhevsky et al. 2012; Szegedy et al. 2015; Liu et al. 2022) and have been used extensively for predicting photometric redshifts from images (e.g. Hoyle 2016; D'Isanto & Polsterer 2018; Pasquet et al. 2019; Hayat et al. 2021; Henghes et al. 2022).

Though CNNs are invariant to translations by design (LeCun et al. 1998; Lee et al. 2009), they use pooling layers (i.e. replacing the input with the local maximum or average value) to locally combine the signal and reduce dimensionality (Ranzato et al. 2007). This comes at the cost of losing precise location and pose information (see e.g. Hinton et al. 2011; Hinton 2021). To solve this problem, Hinton et al. (2011) proposed that artificial neural networks should be organized as local groups that perform complex computations on their inputs and encapsulates the results into highly informative output vectors. These vector counterparts of artificial neurons are called capsules and the entire computational chain is termed as a capsule network. Each capsule vector should learn to recognize the presence of a visual entity irrespective of its orientation, viewing conditions, etc. They should not only encode the probability of the object being present but also encode a set of 'instantiation parameters' for the entity (e.g. location, size, orientation, colour, etc.). For an ideal capsule network, the encoded probability of an object being present should stay the same but the instantiation parameters should change when the input image goes through some transformation (like, rotation, translation, occlusion, etc.).

Though Hinton et al. (2011) introduced the idea of a capsule network, a concrete architecture and training methodology was not proposed. More recently, Sabour, Frosst & Hinton (2017) proposed a training method called the dynamic routing algorithm which made capsule networks viable. Their architecture encodes the 'probability' of an object being present using the length of the capsule vectors. During the training process, information from each capsule is weighted before passing it on to the next layer of capsules via the dynamic routing algorithm (Sabour et al. 2017). The elements of the transformation matrices between two successive capsules are determined by the gradient descent algorithm whereas the routing weights are determined so as to maximize the cosine similarity (i.e. vector dot product) between the capsule vectors of the two consecutive layers in an iterative fashion. Dynamic routing allows capsule networks to focus on specific sections or traits of the input data while making decisions. After each routing step, the capsules are scaled using the non-linear squashing function, $f(\mathbf{v}) = \frac{\|\mathbf{v}\|^2}{1+\|\mathbf{v}\|^2}\frac{\mathbf{v}}{\|\mathbf{v}\|}$ which re-scales the length of each capsule to be between 0 and 1 and acts as the non-linear activation function for the layer.

The original implementation of capsule networks in Sabour et al. (2017) was geared towards the classification of grey-scale handwritten digits. The same implementation was adapted for an astronomical application by Katebi et al. (2019) for morphological classification of galaxies, both of which are easier problems compared to photo-z estimation. Consequently, they got state-of-the-art results while using only a single layer of capsules and a routing algorithm that does not train efficiently if multiple capsule layers are present. To do well in more complicated tasks, it is helpful to have multiple layers of capsules (i.e. a deep capsule network). For this work, we adopt the deep capsule network architecture and dynamic routing algorithm as proposed in Rajasegaran et al. (2019). They propose convolution operation based capsule network layers and a 3D-convolution based routing algorithm which reduces the number of trainable parameters and makes the routing process significantly more efficient thereby making deep capsule networks possible. They also use skip connections (He et al. 2016) which add outputs of earlier layers with the outputs of layers ahead of it to improve the convergence of the training process by preventing the gradients from vanishing and allowing information from earlier capsules to flow efficiently to later ones. Rajasegaran et al. (2019) also introduced an improved class independent decoder network which reconstructs the input image from the final layer capsules and thereby enforces that the components of the capsule vectors form a low-dimensional encoding of the input image. The class-independent nature of the decoder ensures that the capsule dimensions encode the same properties for both morphological classes. A mathematical description of the capsule network layers and routing algorithms mentioned in this section is given in Appendix A.

[2] https://www.sdss.org/dr12/algorithms/magnitudes/
[3] https://www.sdss.org/dr12/spectro/galaxy_mpajhu/
[4] http://sdss.physics.nyu.edu/vagc/

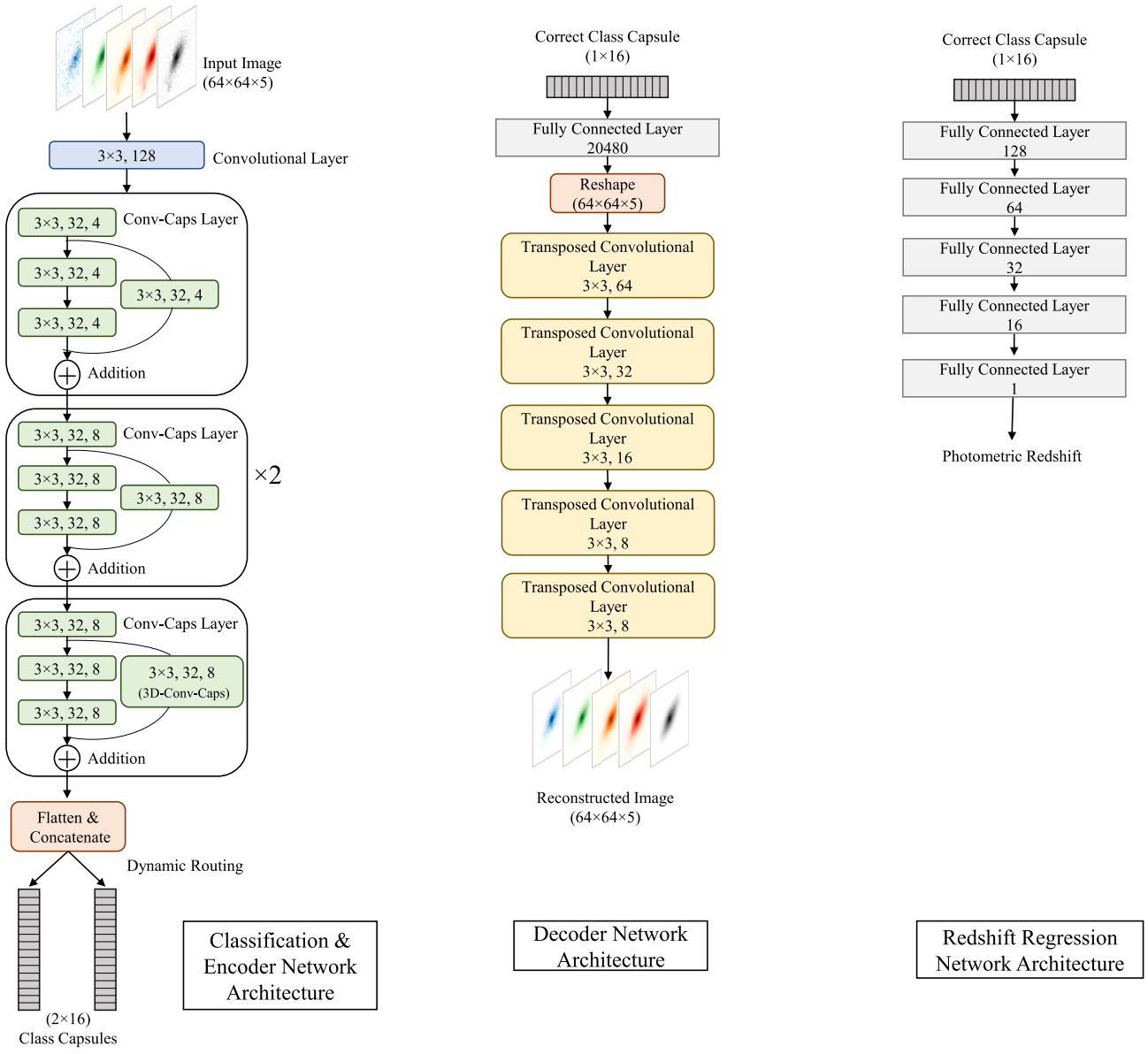


Figure 1. Schematic representation of the neural network architecture we use. The design of the classification and encoder network is based on Rajasegaran et al. (2019). The classification network takes *ugriz* images as inputs and produces two 16-dimensional capsule vectors as outputs, each representing a morphological class (spiral or elliptical). During training, the capsule corresponding to the correct morphological class is used as an input for the decoder and redshift regression networks whereas during inference the capsule vector with the largest magnitude (i.e. highest class probability) is used as the input for the subsequent networks. The numbers in each box represent the shape of the layer being used. For convolutional capsule layers (i.e. `Conv-Caps` and `3D-Conv-Caps` layers), they stand for the width × height, the number of capsules and the total number of dimensions for each capsule, respectively. For convolutional or transposed convolutional layers, they represent the width × height of the convolution filter kernel followed by the number of such filters being used. For fully connected layers, the number represents the number of nodes in the layer. We use a combination of the classification-and-encoding network and decoder network to generate morphological class labels for all the galaxies as a preliminary step and then use a combination of the three networks to predict redshifts. Details of the mathematical operations performed by the various kinds of capsule layers can be found in Appendix A.

3.1 Our capsule network architecture

The network architecture we use has three main components: a deep capsule network-based classification-and-encoding network, a class independent decoder network, and a redshift prediction network. We use a combination of classification-and-encoding network and the decoder network to generate morphological class labels for the entire data set as a preliminary step and then use a combination of all three networks to jointly predict the morphology and photo-z as described below and shown in Fig. 1.

The classification-and-encoding network (Fig. 1, left-hand column) inherits its architecture from Rajasegaran et al. (2019). It takes the 5 band 64 × 64 pixel images of a galaxy as inputs and uses a set of convolutional filters to convert the image into capsules. Next four blocks of skip connected convolutional capsule cells are used. The convolutional capsule layers were introduced in Rajasegaran et al. (2019) and use 3D-convolution operations to perform routing between two capsule layers more efficiently. Skip connections refer to the element-wise summing of outputs of an earlier layer with the output of a non-consecutive layer ahead of it.

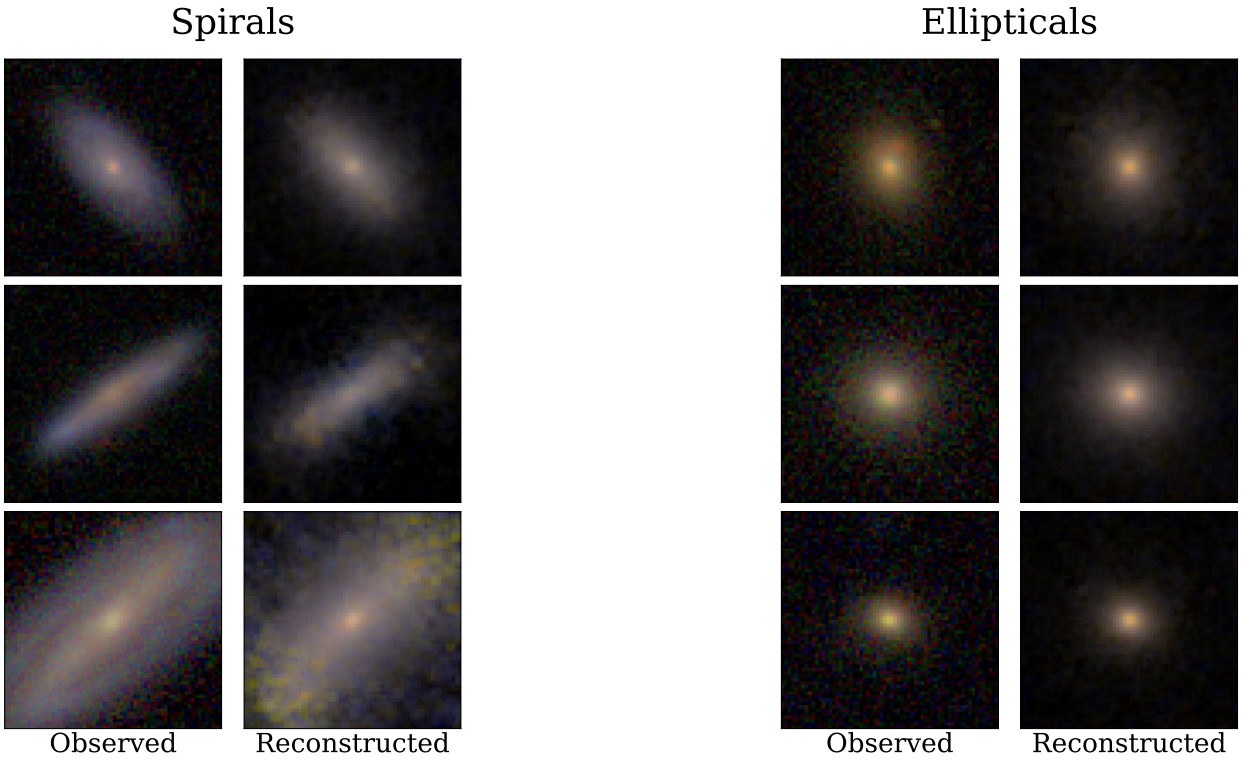


Figure 2. Comparison of the observed and reconstructed *grz* images of a few randomly selected spirals (left) and ellipticals (right) from the test set. The reconstructions were produced by the decoder network using the 16-dimensional capsule corresponding to the predicted morphological type. We see that the reconstructions capture basic properties of the input like shape, orientation, and colour.

This improves the convergence of the training process by preventing the gradients from vanishing and allowing information from earlier capsules to flow efficiently to later ones. The output of the final layer is a set of two 16-dimensional capsule vectors that we use to represent the spiral or elliptical morphological class of a galaxy. The Euclidean lengths of these capsules denote the probability of the input image being a spiral or elliptical. The individual dimensions of the vectors encode information about the input image, which can be used to predict the photometric redshift and reconstruct the input image. This part of the network has about 7.5 million trainable weights.

The class independent decoder network (Fig. 1 middle column) is composed of successive transposed convolutional layers (also called de-convolution layers) which take one of the capsule vectors as input and try to reconstruct the input image as its output. Transposed convolution layers are mathematically similar to convolution layers except their input and outputs are switched. During the training process, we use the capsule representing the correct morphological class as the input of this network. During inference, the capsule with the largest length (i.e. the capsule representing the most probable class) is passed as the input to the decoder network. The decoder network acts as a regularizer and enforces that each dimension of the capsule vector represents a low-dimensional encoding of the input. The decoder network also helps us visually interpret the features encoded by the capsules. Using the same decoder network for both capsules (i.e. class independent decoder) makes the dimensions for both capsules represent similar properties. The decoder network has 0.88 million trainable weights. Some examples of the input and reconstructed images of galaxies are shown in Fig. 2.

The redshift regression network (Fig. 1 right column) is a set of five fully connected neural network layers for redshift estimation. It takes as input the capsule corresponding to the correct morphological class during training and the capsule with the highest class probability during inference. This network has about 13 000 trainable weights.

3.2 Loss functions

The weights of the networks are obtained by minimizing a composite loss function which is a weighted sum of the losses calculated from the outputs of the three networks. The outputs from each of the networks are used to calculate a different loss function, a weighted sum of which is minimized depending on the task we are trying to solve. Following Sabour et al. (2017), we use the output of the classification-and-encoding network to calculate the margin loss (also called the Hinge loss) defined as

$$L_{\rm margin} = \sum_{j=1}^{2} T_j \max(0, m^+ - \|\mathbf{v}_j\|)^2) + \lambda(1 - T_j) \max(0, \|\mathbf{v}_j\| - m^-)^2, \quad (1)$$

where T_j represent the class labels and $T_j = 1$ when a galaxy corresponding to class j is present in the input image and $T_j = 0$ otherwise, $m^+ = 0.9$, $m^- = 0.1$ and $\lambda = 0.5$. The parameters $m^{+/-}$ define a threshold for the length of the capsule above which the classification is considered correct/incorrect. The λ parameter down-weights the margin loss for an absent morphological class, preventing the lengths of all the capsules from shrinking during the initial learning phase. The loss is summed over each class (two in our case). This loss function is optimized to ensure that the length of

one of the capsules is close to 1 and the other one close to 0 when the input is a spiral galaxy and vice versa when the input is an elliptical galaxy.

We use the output of the decoder network to calculate the sum of squared errors between the input and reconstructed image pixels defined as

$$L_{\rm decoder} = \sum_{k=1}^{5}\sum_{j=1}^{64}\sum_{i=1}^{64}(x_{ijk} - \hat{x}_{ijk})^2, \quad (2)$$

where $\mathbf{x}$ and $\hat{\mathbf{x}}$ denote the input and reconstructed images, respectively, and the summation is carried out over all the 64×64 pixels and 5 imaging bands.

Similarly, we use the output of the redshift regression network to calculate the squared error between the spectroscopic redshift and the predicted photometric redshift defined as

$$L_{{\rm photo}-z} = (z_{\rm spec} - z_{\rm phot})^2. \quad (3)$$

All the losses are then averaged over the number of objects present in the training batch. The exact weighting of these losses will be discussed in the next two sections.

4 TRAINING PROCEDURE

4.1 Generating morphological class labels

Morphological class labels from Galaxy Zoo-1 are available for only 34 per cent of the galaxies in our data set (see Section 2.2). We follow a fully supervised learning approach, and our capsule network design relies on the availability of morphological class labels. Therefore, we need to generate morphological class labels for the remainder of the data set to train the network to predict redshifts. To achieve this, we train a deep capsule network that is a combination of the classification-and-encoding network and the decoder network. The decoder network acts as a regularizer. We minimize the weighted sum of the margin loss for classification and the total squared error for reconstruction with a weight of 1 on the margin loss and 0.005 on the reconstruction loss. So, for this task we the loss function (L) given by

$$L = L_{\rm margin} + 0.005 \times L_{\rm decoder}. \quad (4)$$

We divide the set of 177 442 galaxies with good morphological class labels into a training set (80 per cent), validation set (10 per cent), and test set (10 per cent). We train the network to classify the galaxies as spirals or ellipticals and achieve over 99 per cent classification accuracy on the test set. We then use this network to predict morphology labels for the galaxies that do not have a label from Galaxy Zoo-1. We then calibrate the predicted class probabilities with isotonic regression (Zadrozny & Elkan 2001, 2002) using the validation set for training the isotonic regression model and the test set to verify the calibration. This step ensures that the class probabilities predicted by the network are statistically consistent. We then select galaxies with calibrated class probabilities over 0.8, assign them to their corresponding class label and merge them with the initial training set. We train the same network again with this new training set and follow the same procedure above to assign labels and extend the training set. We do this step one more time and find that 99.6 per cent of the galaxies in our parent set has a class label with more than 0.8 class probability. For the remaining 0.4 per cent of the galaxies, we assign a label corresponding to the class with the highest probability.

We are generating morphological class labels for 339 083 galaxies based on a human labelled training set of 177 442 galaxies. The bulk of the galaxies do not have a confident morphological class label in Galaxy Zoo-1 as a strong consensus was not achieved among the human volunteers. This either might be because the shape of the galaxy is ambiguous or there were some artefacts in the image. A visual inspection of the galaxies in the test set which do not have a label from Galaxy Zoo-1 shows that the objects can almost always be classified into a spiral or elliptical galaxy by the authors and the predictions of our model for those objects matches with the judgement of the authors. The number of images which have ambiguous morphology or where an artefact or merger makes the morphology difficult to infer are negligibly small (0.1 per cent). Since our main goal is to improve photo-z prediction performance, we are comfortable with using the smaller training set with only good classifications to generate class labels for the entire data set and ignoring the very small number of ambiguous cases. As a separate cross-check, we compared the class labels generated by our method for the galaxies which do not have a confident label from Galaxy Zoo-1 with the most voted Galaxy Zoo-1 class label and find that they are in agreement for over 70 per cent of the objects.

4.2 Training for photo-z estimation

Once we have morphological class labels for all the galaxies in our data set, we now train a neural network that is a combination of the classification-and-encoding network, the redshift regression network, and the class independent decoder network. The classification-and-encoding network gives us a low-dimensional representation of the input image which is then used by the redshift regression network to predict the photometric redshift. Although the decoder network does not directly help with redshift prediction, it has been shown to have a regularization effect on capsule networks (Sabour et al. 2017). The decoder network also ensures that the low-dimensional encoding learnt has physically meaningful information, which can be used to reconstruct the input image. In Section 5.3.2, we use the decoder network to interpret the features learnt by the capsule network.

During the training process, the capsule corresponding to the correct morphological class is used as an input for both the decoder and redshift regression networks whereas during inference the capsule vector with the largest Euclidean length (i.e. highest class probability) is used as their inputs. To find the optimum set of weights for the network, we minimize a composite loss function which is a weighted sum of the losses from each of the three networks. Similar to Section 4.1, we use the weighted sum of the margin loss and total squared error for the classification and reconstruction tasks, but now we also add the squared error of the predicted redshift to the total loss (L):

$$L = L_{\rm margin} + 0.005 \times L_{\rm decoder} + L_{{\rm phot}-z}. \quad (5)$$

The classification, reconstruction, and redshift regression losses are given the weights of 1, 0.005, and 1 so that they contribute an equal amount towards the total value of the loss. This allows us to put equal importance on each of the individual tasks as all of them help to improve the accuracy of photometric redshifts. Some examples of reconstructed images of galaxies obtained after training the network are shown in Fig. 2.

Instead of directly predicting the redshifts, we scale the redshifts using the logistic transformation defined as

$$h(z) = \log\left(\frac{z - z_{\rm min}}{z_{\rm max} - z}\right). \quad (6)$$

For our data set $z_{\rm min} = 0$ and $z_{\rm max} = 0.4$. We find that performing this transformation gives us better performance especially at very low redshifts ($z < 0.05$). This is because the logistic transformation makes the distribution of the target variable (redshift in our case) fall gradually at the boundaries, thereby alleviating the problem of attenuation bias.

We randomly split our data into three subsets; the training set which is used to train the network, the validation set, which is used to tune the hyperparameters of the network and decide when to stop training and a test set which is used to check the final performance. All results quoted in this work use a training set that is 80 per cent the size of the parent data set and have been calculated on the test set which is 10 per cent the size of the parent set (unless stated otherwise). The remaining 10 per cent of the data is used as the validation set. We also check the performance of our photo-z prediction as a function of the size of the training set as shown in Fig. 6.

To randomly initialize the weights for the networks, we use the He-Normal initializer (He et al. 2015). We use the PReLU (He et al. 2015) activation function for all the hidden layers and a linear activation function for the output layers of the decoder and redshift regression networks. To train all the networks, we use the Adam optimizer (Kingma & Ba 2015) with an initial learning rate of 0.001. After each epoch the learning rate is decreased following the rule: learning rate = initial learning rate $\times$ $0.95^{\rm epoch}$. We also augment the training set by randomly rotating the images in steps of 90° or flipping them along the horizontal or vertical axis before passing them to the networks for training. The same setup is used for both the morphological label prediction and redshift estimation tasks.

We train the networks for 100 epochs but the training generally converges within 70 epochs. We choose the epoch which has the best performance – i.e. the highest classification accuracy when generating morphology labels and the lowest average redshift prediction error on the validation set. Since the model is initialized randomly, each training run can result in a different set of optimal weights. Hence we run the training process 5 times and take the average of their output as our photo-z prediction. For this reason, we also select epochs that have a low bias and moderate variance since bias stays roughly the same whereas variance decreases when averaged.

The models are defined in Keras with Tensorflow 1.15 as the back end. The training is done on an Alienware Area 51 PC with an Intel Core i7 9800X processor, 2 RTX 2080Ti GPUs and 64GB of RAM. We use a batch size of 400 which takes about 8 h to train for 100 epochs. The model is copied on to the two GPUs and the training is parallelized by sending half of the batch to each GPU.

5 RESULTS

5.1 Photo-z evaluation metrics

In this work, we are focusing only on photo-z point estimates and not full PDFs. We will therefore assess the performance of our photo-z estimates by measuring how much the spectroscopic and photometric redshifts for each galaxy in the test set differ. We use the following three common metrics:

(i) **Prediction bias** defined as $\langle \frac{\Delta z}{1+z_{\rm spec}} \rangle$, i.e. the average value of the prediction error.

(ii) **Normalized Median Absolute Deviation (σ_{NMAD})** defined as $1.4826 \times {\rm Median}(|\ \frac{\Delta z}{1+z_{\rm spec}} - {\rm Median}(\frac{\Delta z}{1+z_{\rm spec}})\ |)$. This is a robust measure of the spread of prediction errors.

(iii) **Fraction of Outliers ($\mathbf{f_{outlier}}$)** defined as the fraction of photo-z predictions for which $|\ \frac{\Delta z}{1+z_{\rm spec}}\ | > 0.05$, i.e. the fraction of cases where the prediction error is very high. We chose the threshold of 0.05 to easily compare our results with other similar works.

The specific choice of the metrics and the threshold to define an outlier is based on convention and allows us to easily compare our results with recent similar work.

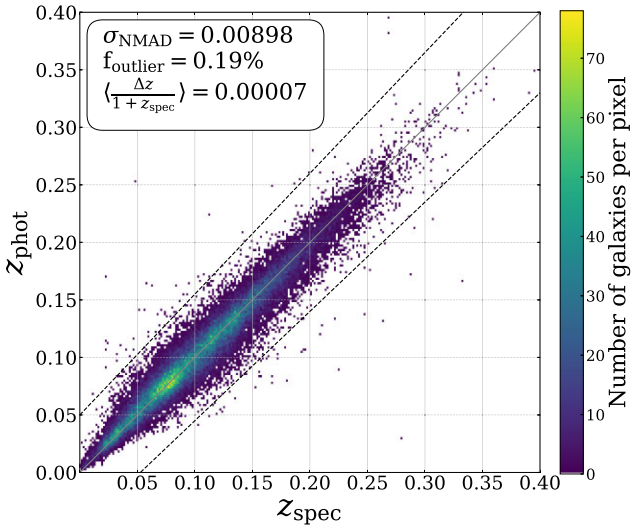


Figure 3. Comparison of photometric redshift point estimates predicted by our capsule network with the corresponding spectroscopic redshifts for galaxies in the test set. The central grey line shows $z_{\rm phot} = z_{\rm spec}$, i.e. a perfect photo-$z$ estimate. The outer dashed lines mark $|\ \frac{\Delta z}{1+z_{\rm spec}}\ | = 0.05$. Any point lying outside these limits (i.e. $|\ \frac{\Delta z}{1+z_{\rm spec}}\ | > 0.05$) is considered to be an outlier. The colour on the scatter plot shows the number of data points present in each pixel of the figure. We see that the scatter is tight and symmetrically distributed about the $z_{\rm phot} = z_{\rm spec}$ line and with a negligible bias. The scatter looks random and shows no visible patterns at the limits of training data ($z_{\rm spec} \approx 0$ and $z_{\rm spec} > 0.3$) indicating stable performance across the redshift range.

5.2 Photo-z point estimate predictions

When trained on 80 per cent and tested on 10 per cent (with the remaining 10 per cent used as validation set) of the parent data set and results averaged over an ensemble of 5 models, our photo-z estimates have $\sigma_{\rm NMAD} = 0.00898$, $f_{\rm outlier} = 0.19$ per cent, and $\langle \frac{\Delta z}{1+z_{\rm spec}} \rangle = 7 \times 10^{-5}$. For comparison, other deep learning based methods which take images as inputs like Pasquet et al. (2019) achieve $\sigma_{\rm NMAD} = 0.00912$, $f_{\rm outlier} = 0.31$ per cent and $\langle \frac{\Delta z}{1+z_{\rm spec}} \rangle = 1 \times 10^{-4}$ when trained on the same data set and Hayat et al. (2021) achieves $\sigma_{\rm NMAD} = 0.00825$, $f_{\rm outlier} = 0.21$ per cent and $\langle \frac{\Delta z}{1+z_{\rm spec}} \rangle = 1 \times 10^{-4}$, by first pre-training on a large unlabelled data set (about twice as big as our data set) and then fine-tuning on a data set similar to ours. Both of them use models with about 3 times as many trainable parameters compared to ours ($\sim$24 million versus $\sim$8 million). Our algorithm has comparable $\sigma_{\rm NMAD}$ and better $f_{\rm outlier}$ performance among these deep learning based methods.

We show a comparison between the photometric and the spectroscopic redshifts for the test set in Fig. 3. We see that the scatter is tight and distributed symmetrically about the $z_{\rm phot} = z_{\rm spec}$ line. The scatter in the points and distribution of outliers look random and show no visible patterns of a sudden change in performance at the limits of training data ($z_{\rm spec} \approx 0$ and $z_{\rm spec} > 0.3$) indicating stable performance across the redshift range. We also see no evidence of attenuation bias (i.e. almost constant predictions for a subset of inputs; see Freeman et al. 2009 for a discussion on attenuation bias in

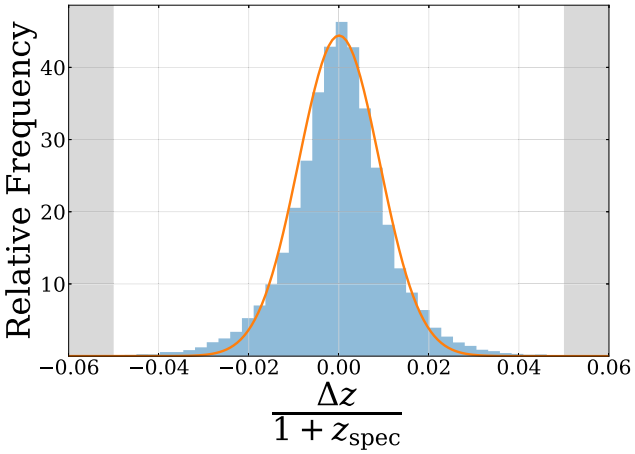


Figure 4. Normalized distribution of the redshift prediction errors. The blue histogram shows the distribution of redshift prediction errors of our algorithm on the test set. The orange line shows a Gaussian distribution with the location and scale parameters set as the prediction bias and $\sigma_{\rm NMAD}$, respectively. The distributions are normalized to have unit area under the curves. The shaded region marks the threshold for outliers. The distribution of the prediction errors is symmetric, centred around 0 and closely resembles a Gaussian distribution, indicating little if any systematic preference for overestimation or underestimation.

photo-z algorithms). The images used to train the networks include observations of Stripe 82 (Jiang et al. 2014), which are about 2 mag deeper and have less noise than the rest of the images. Since Stripe 82 is a small fraction (< 4 per cent) of the whole data set, we do not account for this varying depth by weighing data points differently. We find a significantly smaller spread in the predictions ($\sigma_{\rm NMAD}$ = 0.00741) and a fraction of outliers consistent with the rest of the sample ($f_{\rm outlier}$ = 0.35 per cent), given the small number of Stripe 82 objects in the test set. Galaxies in the test set outside of Stripe 82 produce photo-z's with $\sigma_{\rm NMAD} = 0.00906$ and $f_{\rm outlier} = 0.19$ per cent. This shows that having images with a higher signal-to-noise ratio improves the quality of photo-z predictions.

When the test set is split into subsets based on morphology, we find that the photo-z predictions have a lower spread for ellipticals than spirals ($\sigma_{\rm NMAD}$ = 0.00844 versus 0.00956) with a comparable fraction of outliers (0.18 per cent versus 0.20 per cent). This might be because elliptical galaxy populations have similar rest-frame colours as older stellar populations tend to change very little in colour with time. The observed colours and magnitudes (or any other measure of flux) therefore trace the redshift well making it is easier to predict redshifts of elliptical galaxies than spirals. When we split the test set based on the availability of human labelled morphology, we find that photo-z prediction performance is better when human labelled morphology is available ($\sigma_{\rm NMAD}$ = 0.00815, $f_{\rm outlier}$ = 0.11 per cent versus $\sigma_{\rm NMAD}$ = 0.00948, $f_{\rm outlier}$ = 0.23 per cent). Although human labelled morphology improves the performance, the lack of it does not reduce the performance drastically.

We performed a visual inspection of the images of the galaxies which were photo-z prediction outliers. We find that around 18 per cent of these outliers have bad or missing photometry. Removing these objects from our test set reduces our outlier fraction to $f_{\rm outlier}$ = 0.16 per cent. We kept these rare objects in the parent data set for easy comparisons with Pasquet et al. (2019).

The distribution of prediction errors is shown in Fig. 4. They follow a symmetric distribution centred about 0 indicating little if any systematic preference for overestimation or underestimation. Since the fraction of outliers is very small and we see that the distribution of prediction errors closely resembles a Gaussian distribution, $\sigma_{\rm NMAD}$ can be treated as the 1σ Gaussian uncertainty around each prediction up to a good approximation.

We also check the performance (prediction bias and $\sigma_{\rm NMAD}$) of our photo-z estimates as a function of the spectroscopic redshift and r-band Petrosian magnitude of galaxies as shown in Fig. 5. We use the Petrosian magnitude as it was used to define the faintness cut of the data set we are using. As a function of redshift, the absolute magnitude of the bias is small though it is positive at low redshifts and negative at high redshifts with the inflection point being at the median redshift (≈0.1) of our data set. This kind of pattern is common for ML-based algorithms. When seen as a function of r-band magnitude, the bias is almost constant and negligibly small in magnitude throughout the entire range of magnitudes. $\sigma_{\rm NMAD}$ tends to increase both as we go to higher redshifts and fainter magnitudes. This can be attributed to the fact that there is less training data and increased noise in the images at these regimes. We also see that $\sigma_{\rm NMAD}$ (∼0.006) is significantly lower than the global value at low redshifts ($z < 0.05$) even though the number of training samples available is small in this regime due to lower survey volume. We suspect this is because at very low redshifts resolved information in the images, like morphology, size, and surface brightness, contains rich information about galaxy distances. Better photo-z performance at very low redshifts can aid in the identification of satellite galaxies that require a massive spectroscopic effort to get redshifts (e.g. Geha et al. 2017, Mao et al. 2021).

Obtaining spectroscopic redshifts is often an expensive process, so it is important that ML-based methods can perform well when the training data sets are smaller. To see how the photo-z performance of our algorithm changes, we train our capsule network-based model using varying sizes of training data by random sub-sampling of the parent data set (after obtaining morphological labels) into smaller subsets while keeping everything else the same in the training process. The results are shown in Fig. 6 and also compared with other similar works like Pasquet et al. (2019), Hayat et al. (2021), and Beck et al. (2016). The data for Pasquet et al. (2019) and Beck et al. (2016) were obtained from table 2 in Pasquet et al. (2019), the data for Hayat et al. (2021) were obtained from their fig. 4 using the WebPlotDigitizer (Rohatgi 2020). The metrics for Beck et al. (2016) provided here are calculated on their photo-z estimates of the same objects as ours. They train on a much larger data set spread over a larger redshift range compared to ours which maybe one reason for higher prediction errors. We always use 10 per cent of the parent data set as the validation set and use the remaining amount of data to test the performance. We observe that we outperform Beck et al. (2016), which is a widely used source of SDSS photo-z estimates using just 2 per cent of the parent sample (or $\sim 10^4$ galaxies) as a training set. Many surveys of the high-redshift Universe like CANDELS (Grogin et al. 2011; Koekemoer et al. 2011) have spectroscopic observations for a similar number of galaxies, albeit across a larger redshift range and our method could potentially be used to improve the photo-z estimates for them. We see that our method has performance comparable to other deep learning-based photo-z estimation methods like Pasquet et al. (2019) or Hayat et al. (2021) when both are trained on random subsets of data.

5.3 Interpreting the features learnt by the capsule network

As ML-based methods have started replacing more traditional physics-based methods to model astrophysical phenomena and make predictions that reduce the need for making extra observations, it

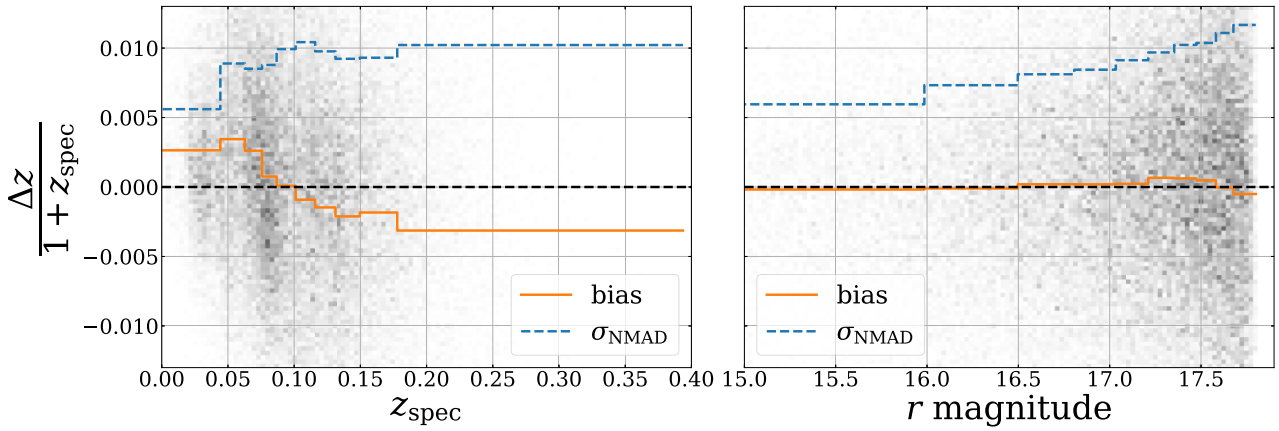


Figure 5. Prediction bias ($\langle \frac{\Delta z}{1+z_{spec}} \rangle$) and σ_{NMAD} of our photometric redshift estimates as a function of spectroscopic redshift (z_{spec}, left) and r-band Petrosian magnitude (right). The metrics have been calculated for 10 bins of equal population. We use bins with varying widths but equal populations so that the standard errors on the binned statistics are comparable across all bins. The grey points show the distribution of individual galaxies. Due to the relatively large number of samples in each bin, the standard errors on the statistics are very small. We see that σ_{NMAD} increases at higher redshifts (where we have less training data) and for fainter galaxies (where the signal-to-noise ratio of the images are lower). Though the bias on average is very small, it is higher at the lowest and highest redshift bins but with opposite signs with an inflection at the median z_{spec} (≈0.1). The bias is constant and negligible in magnitude over the entire range of r-band Petrosian magnitudes.

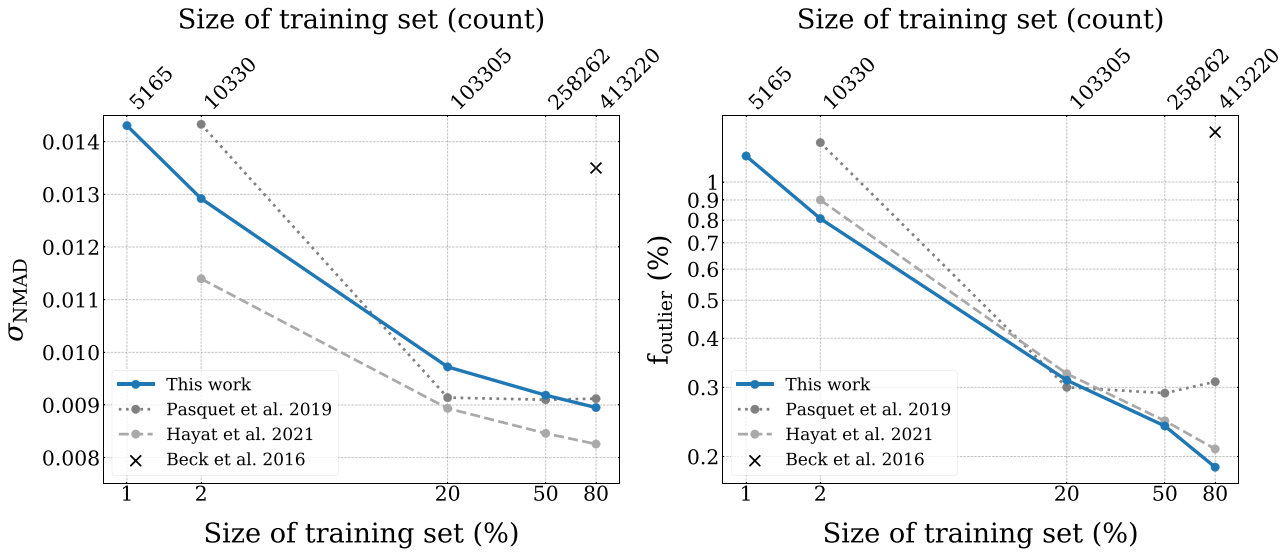


Figure 6. Performance of photometric redshift prediction algorithm as a function of the size of training data. The standard errors on the statistics are negligibly small and hence not shown. Our algorithm has comparable σ_{NMAD} and better $f_{outlier}$ performance to the two deep learning-based efforts (Pasquet et al. 2019; Hayat et al. 2021) and significantly better performance than the classical ML-based technique (Beck et al. 2016) while requiring less training (or pre-training) data and fewer trainable parameters (∼8 million versus ∼23 million).

is becoming increasingly important to peer inside these complex mathematical models to identify what physical features they are learning. This will not only help us to validate what the algorithms are predicting but also help us bridge the gap between the traditional physics-driven and the newer data-driven approaches.

In our work, we use the capsule vectors along with the decoder network to shed some light on the features learnt by the network. Since the capsules composing the output layer of the morphology classification network are trained to represent a morphological class of galaxies along with holding enough information to predict the redshift and a reconstruction of the input image, we expect the components of the capsule vector to learn a low-dimensional encoding of the input galaxy image. Moreover, we expect that each of the component will learn properties so that all capsule dimensions combined can effectively predict the morphology, redshift and a reconstruction of the input image.

The features learnt by the networks are not constrained to be easily identifiable visual properties or commonly used physical quantities derived from images. We will therefore perform both visual exploration of these features and also measure how well these features correlate with galaxy properties.

5.3.1 *Visualizing the capsule encoded space*

We first take a look at how the capsules corresponding to each galaxy in the test set are organized in their manifold. We use Uniform

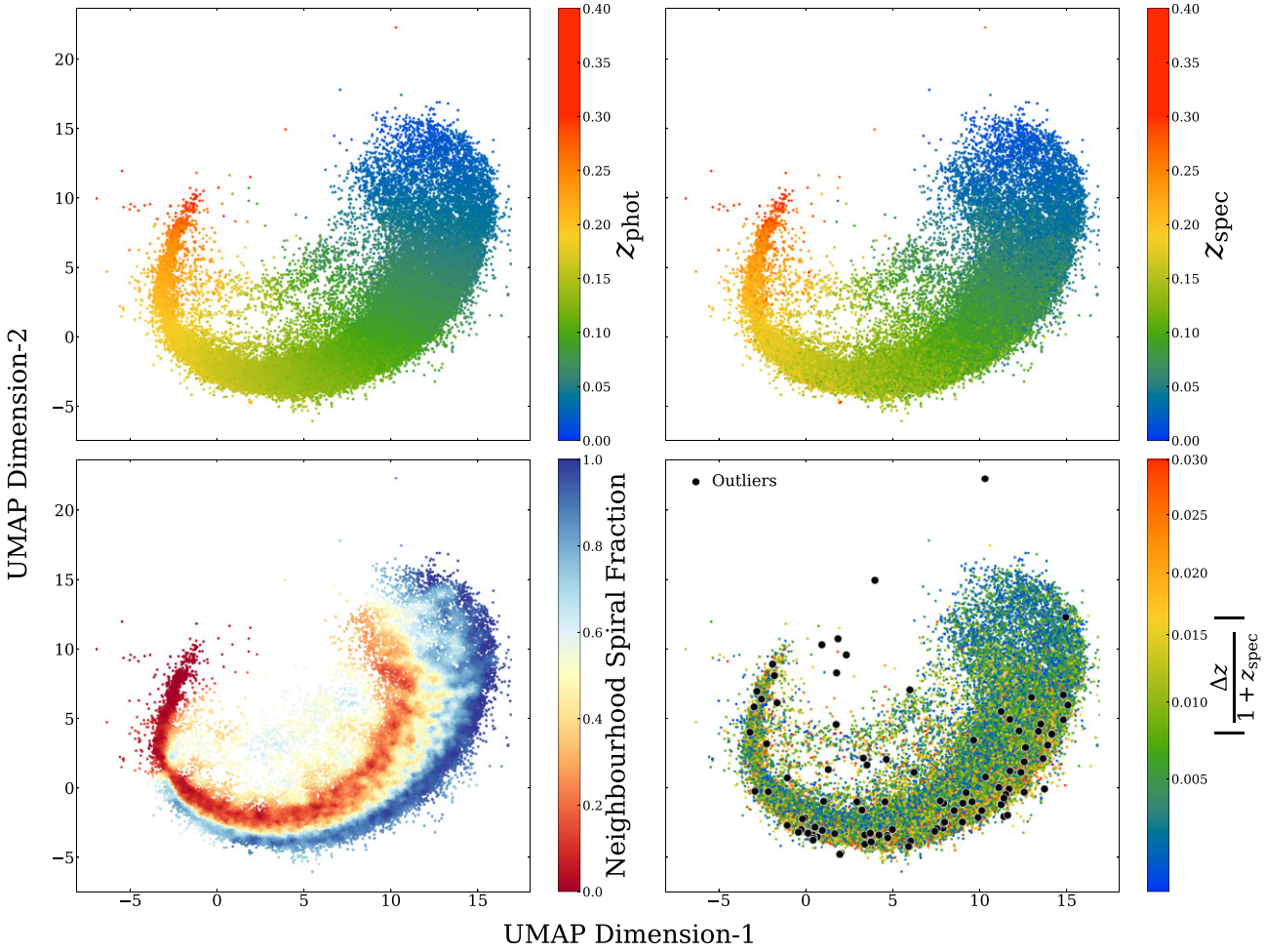


Figure 7. Two-dimensional UMAP embedding of the 16-dimensional capsules colour-coded by photometric redshift (top left), spectroscopic redshift (top right), fraction of spiral galaxies in the neighbouring region (bottom left), and redshift prediction error (bottom right). The photo-z outliers are shown in black in the bottom right panel. The UMAP embedding of the capsules creates a nearly perfect redshift sequence indicating that the capsules learn a good representation of redshift. Spirals and ellipticals tend to occupy separate regions though there is a region with overlap with morphology producing a gradient almost perpendicular to the redshift sequence. We notice that regions dominated by spirals tend to have slightly higher redshift prediction errors compared to regions dominated by ellipticals. Though spirals and ellipticals have almost the same fraction of outliers, visually it may seem that there are more outliers which are spirals than ellipticals. Many of those outliers are actually ellipticals which lie close to the region dominated by the spiral galaxies in the 2D UMAP representation. An interactive version of this figure showing galaxy image thumbnails is available online[5].

Manifold Approximation and Projection (UMAP; McInnes & Healy 2018) to embed the 16-dimensional capsules into a two-dimensional space to visualize and interpret any structures, if present. UMAP is a non-linear dimensionality reduction method that uses techniques from manifold learning and topological data analysis to embed a high-dimensional data set into a low-dimensional manifold. To ensure that the relative local density of data is preserved when we project the capsules on to a two-dimensional space, we use DensMAP (Narayan, Berger & Cho 2020), which computes the estimates of local density and uses them as a regularizer in the optimization of the 2D UMAP representation. UMAP with the DensMAP regularizer preserves the local structure of the data while capturing global structure better than many other similar algorithms and is also computationally efficient.

Fig. 7 shows the two-dimensional UMAP embedding of the 16-dimensional capsules colour coded by various properties. When coloured by photometric or spectroscopic redshift (top row), the embedding shows a nearly perfect redshift sequence. As UMAP places nearby capsules in the high-dimensional space close together in their two-dimensional projection, we can infer that the capsules track a smooth redshift sequence. This is in contrast to the representations generated by self-organizing maps (SOMs; Kohonen 1981, 1982), which group galaxies with similar spectral energy distributions together using their photometry but impose a geometry that can force adjacent cells to have wildly different redshifts (Masters et al. 2015). Currently, SOMs are widely used to determine regions with incomplete spectroscopic data (e.g. Masters et al. 2015, 2019), but dimensionality reduced capsules may perform better at this task due to its smooth redshift distribution.

If we colour the points based on the fraction of spirals among the 80 nearest neighbours in the 2D space (bottom left), we see that the spirals and ellipticals tend to occupy separate regions of the space although there is a significant overlap. The fraction of spirals exhibits a gradient almost perpendicular to the redshift sequence thereby effectively encoding both redshift and morphology, properties the capsules were trained to learn. When colour-coded by the redshift prediction errors (bottom right) and compared with the plot showing the fraction of neighbouring spirals, we notice that regions dominated

by spirals tend to have slightly higher redshift prediction errors compared to regions dominated by ellipticals. This was quantified in Section 5.2 where we noted that spirals have slightly higher value of $\sigma_{\rm NMAD}$ compared to ellipticals but equivalent $f_{\rm outlier}$. Visually from Fig. 7 it may seem that there are more outliers which are spirals than ellipticals but many of those outliers are ellipticals which lie close to the region dominated by the spiral galaxies in the 2D UMAP representation.

Most of the galaxies in the 2D UMAP representation lie on the large crescent shaped sequence. A small number (about 1–2 per cent) of galaxies deviate from this sequence forming a smaller sequence encircled by the larger crescent. These galaxies all have higher values for dimension 10 of their capsules. Synthetic images generated by perturbing capsule dimensions (see Appendix B) shows that higher values of dimension 10 tend to increase the extended component of the galactic disc. Some of the dimension 10 outliers in the main redshift sequence clearly have stars in the image. However, our investigation of the dimension 10 outliers in the smaller sequence has yet to yield a clear interpretation. These galaxies are a 50/50 mix of spirals and ellipticals, and the majority do not have neighbouring stars, galaxies, or artefacts. A systematic study of these outlier galaxies will be done in a future work. The other galaxies that randomly scatter away from the two large sequences almost always have a neighbouring star, galaxy, or an artefact.

5.3.2 *Generating synthetic images by perturbing capsule dimensions*

To check whether the components of the capsules represent any visually identifiable properties of the galaxies, we take the capsule corresponding to the predicted morphology of a galaxy and add a small perturbation to one of the components keeping all the others fixed. The perturbation is added in units of standard deviation of the values of the components in our test set. We pass on this perturbed capsule vector to the decoder network to see how the reconstructed image of the input changes.

Fig. 8 shows the synthetic galaxy images generated from the perturbed capsule vectors for two galaxies (the first instance of each morphological type from Fig. 2). We can see that perturbing specific components change properties like size (i.e. the angular size of the galaxy and how fast the light profile falls off), orientation, amount of central bulge, and surface brightness. This shows that some of the features learnt by the capsule network correspond to physical properties of galaxies. Visual properties like size and surface brightness change with the distance of the galaxies and can help to break degeneracies in the colour–redshift relation and provide better redshift inference. Fig. 8 shows the synthetic images from perturbed capsules for only a subset of dimensions for which the change in the images is easily identifiable visually. Appendix B shows the synthetic images generated by perturbing all 16 of the dimensions individually.

5.3.3 *Correlations of capsule dimensions with physical properties*

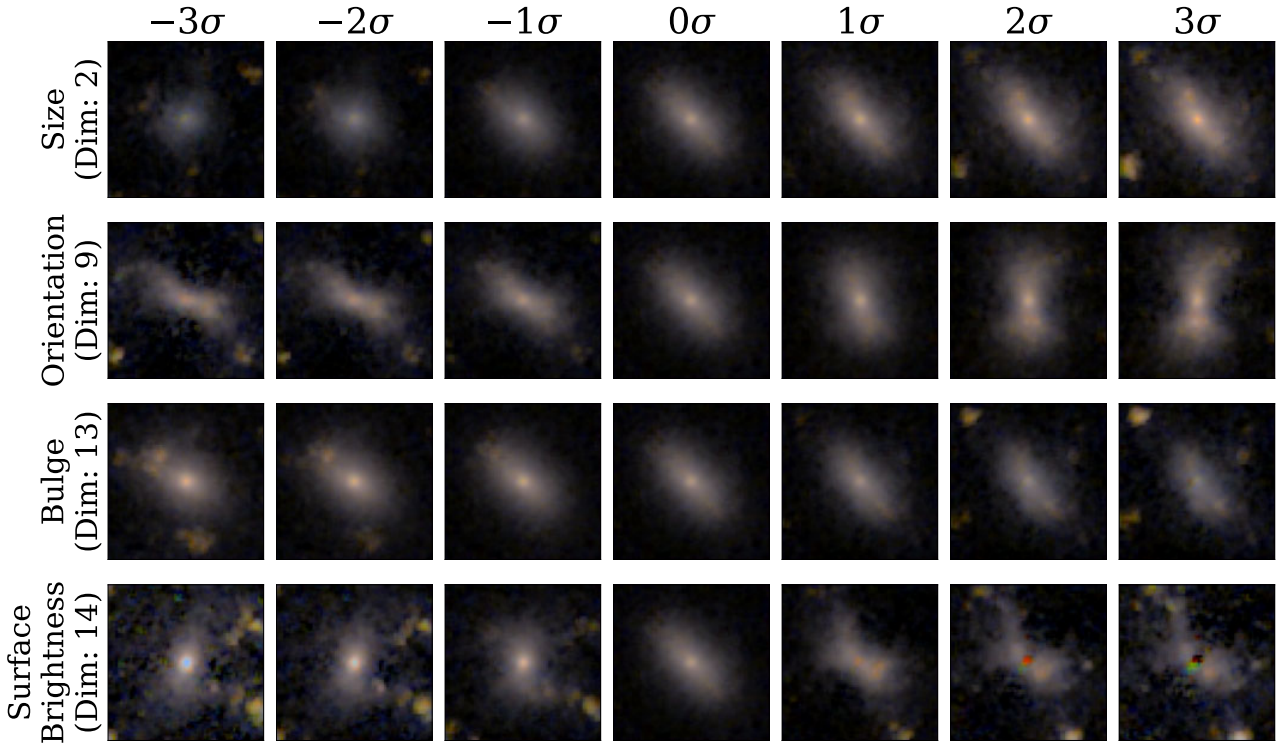
To check whether any physical properties of the galaxies are encoded by the capsules that cannot be identified by simply looking at synthetic images generated from perturbed capsules, we measure the correlations between each dimension of the capsules and various global galaxy properties. Since we expect the correlations to be non-linear in nature, we use the distance correlation (Székely, Rizzo & Bakirov 2007) to measure them. The distance correlation quantifies the dependence between two random variables by measuring how much the Euclidean distance between two samples of one random variable changes for a given change in distance between two samples of another random variable. This makes the distance correlation sensitive to any kind of dependence between two random variables, unlike Pearson or Spearman correlations which measure linear and strictly monotonic relationships respectively. The distance correlation has a value between 0 and 1, where 0 would mean that the random variables are independent whereas a value of 1 would mean the linear sub-spaces spanned by the two random variables are almost equal, indicating a very high degree of dependence.

Fig. 9 shows values of distance correlation between each of the components of the capsule vector corresponding to the predicted morphology and global properties of galaxies in the test set. Unsurprisingly, we find that many of the capsule components have strong correlations with the spectroscopic redshift, with dimensions 8, 14, and 3 being the strongest. The capsule dimensions that show strong correlations with spectroscopic redshift also show strong correlations with observed frame galaxy colours and apparent magnitudes which are known to be good predictors of photometric redshift. Given this pattern, we also expect them to be well correlated with galaxy absolute magnitudes ($M_{u/g/r/i/z}$) which we can also verify from Fig. 9. Sérsic index (n_r) correlates the most with dimension 13 which we saw controls the amount of a galaxy's central bulge (see Fig. 8). Similarly, dimension 2 which we saw control the visual size of the galaxy image has the strongest correlation with the 90 per cent light radius ($R_{90,\,r}$) among all capsules and also correlates well with Sérsic index which are the two quantities which together quantify the visual size of the galaxy on the sky. We can therefore infer that the capsules successfully encode almost all of the photometric properties of the galaxy image. A few illustrative examples of these correlations in form of scatter plots can be found in Appendix C.

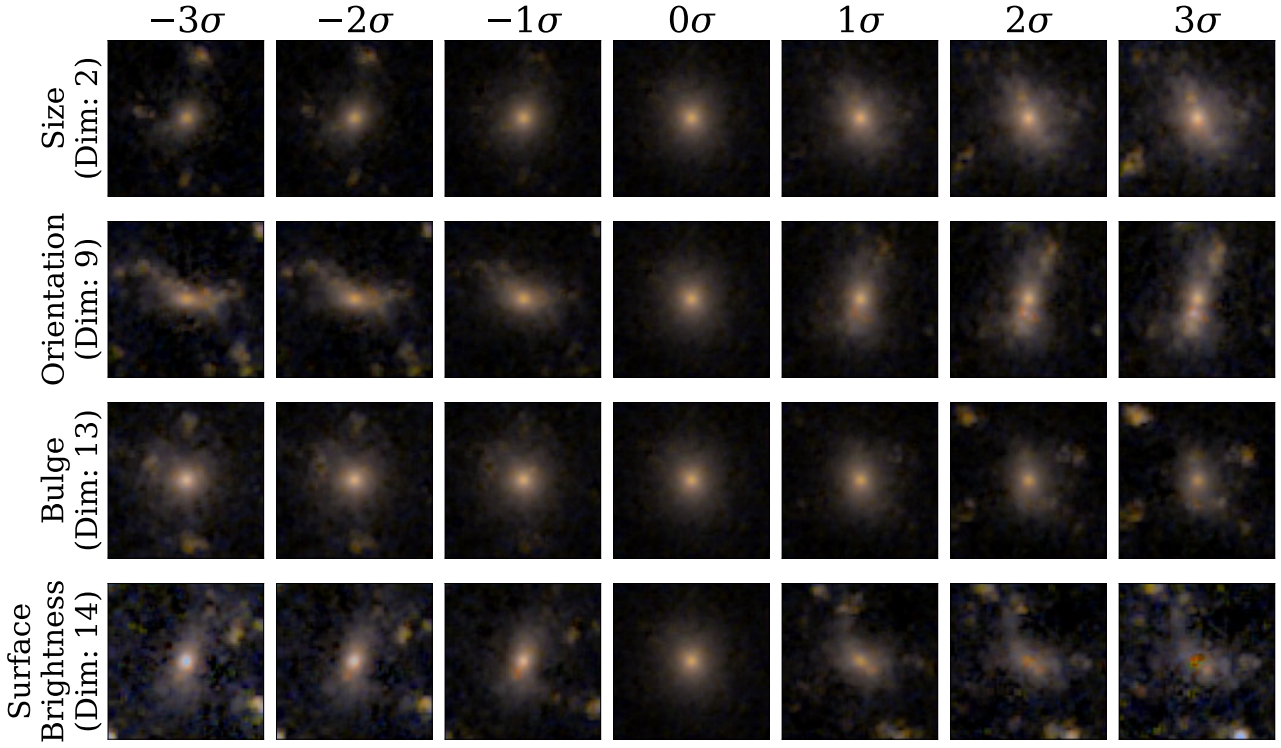
Many capsule dimensions show correlation with physical properties like stellar mass ($M_\star$) and velocity dispersion of the spectra (σ_v) and a small number of dimensions show strong correlations with SFR and sSFR. Most likely, these correlations arise because SFR and sSFR depend on galaxy magnitudes and spectroscopic redshifts which the capsules efficiently encode, but the capsules may also encode some physical properties of the galaxies. Even though we focus on predicting photometric redshifts in this work, we expect that capsule-based encodings can be used to create a general purpose image-based inference methodology for physical properties of galaxies and will be explored in a future work.

5.3.4 *Feature importance using SHAP values*

As shown in the Sections 5.3.2 and 5.3.3, each capsule dimension tends to encode a somewhat different property of the input image, so we would like to see which of the dimensions are most useful in predicting photo-z's. To quantify this, we calculate the SHapley Additive exPlanations (SHAP; Lundberg & Lee 2017) values for each of the capsule dimensions that are used by the redshift prediction network using the test data. SHAP is a method to explain a prediction by computing the contribution of each feature. It takes a game theory approach to optimally distribute credit to each feature for a given prediction using Shapley Values (Shapley 1953). The Shapley value for a feature is defined as the average marginal contribution of a feature across its all possibilities for a given prediction. The SHAP value is then calculated via a weighted sum of Shapley values to ensure that the contribution of each feature to a prediction add up to the value of the prediction. Since it would be prohibitively



(a) The first spiral galaxy from Fig. 2



(b) The first elliptical galaxy from Fig. 2

Figure 8. Synthetic galaxy images generated by perturbing capsule dimensions. Each column shows the decoded image when one of the 16 dimensions of the capsule vector is perturbed in units of its standard deviation (keeping all the others fixed). The 0σ column shows the decoded image from the unperturbed capsule and are identical for each row. We show a subset of the dimensions here for which the perturbations have a clear interpretation (see Appendix B for a version with all the dimensions). We see that some of the capsule dimensions, encode physical features like size, orientation, amount of central bulge and surface brightness of the galaxies.

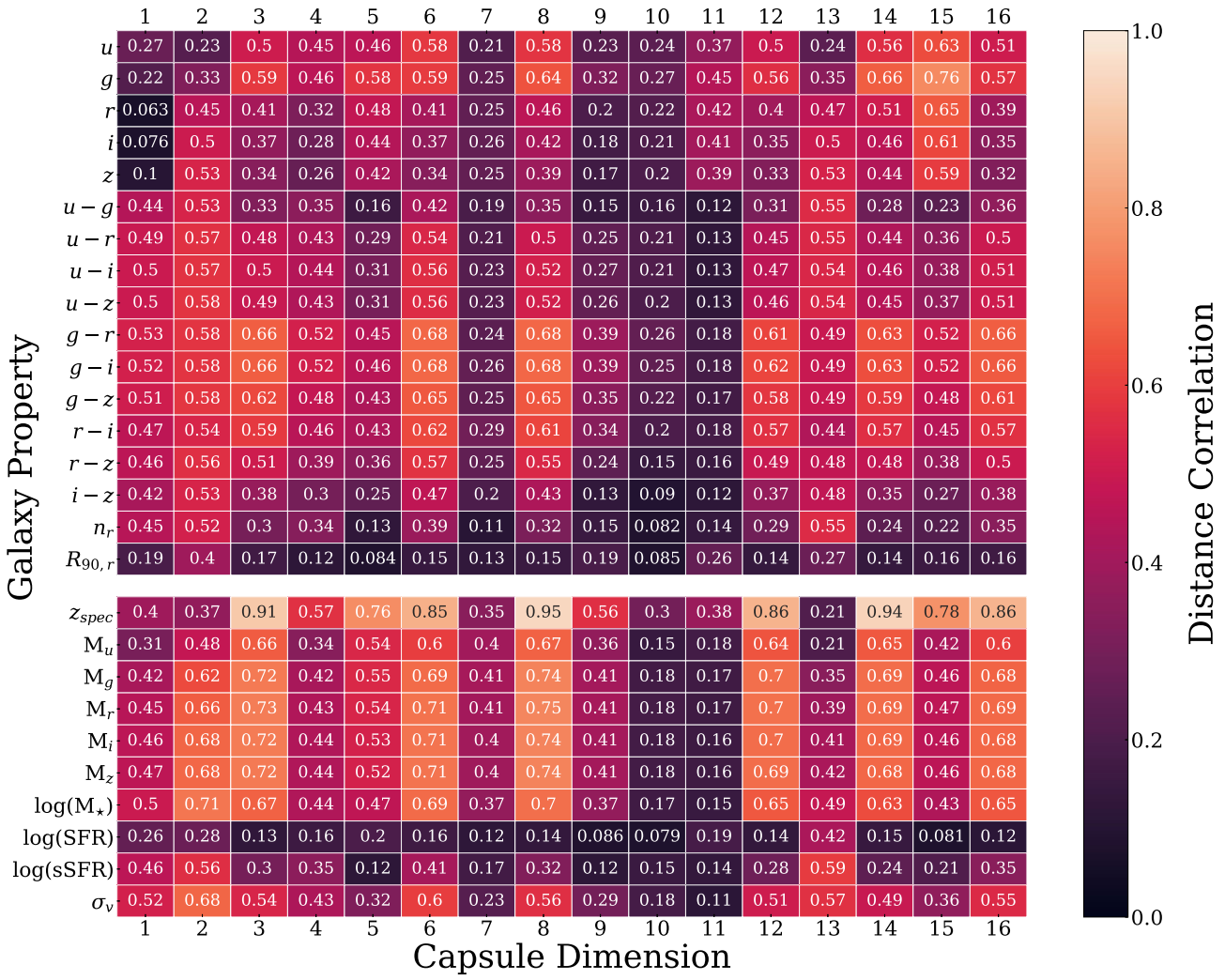


Figure 9. Measurements of the distance correlation between the capsule dimensions corresponding to the predicted morphological class and global galaxy properties (as described in Section 2.3). The values have a range between 0 and 1 where a value of 0 means the two random variables being compared are independent and a value of 1 indicates a high level of dependence. We have grouped the galaxy properties into two sets: properties which solely depend on photometry (top) and properties which include knowledge of the spectroscopic redshift along with photometry (bottom). u, g, r, i, z represent the extinction corrected cmodel magnitudes. $u - g$, $g - r$, etc. represent galaxy colours calculated using extinction corrected model magnitudes. n_r and $R_{90,r}$ represent the Sérsic index and the 90 per cent light radius obtained from the Sérsic profile fit to r-band photometry and are used as a proxy for a galaxy's size. $z_{\rm spec}$ denotes the spectroscopic redshift; $M_{u/g/r/i/z}$ represent the absolute magnitudes in each of the five bands. $M_\star$ stands for the stellar mass, SFR stands for the star formation rate, and sSFR stands for the specific star formation rate. σ_v represents the velocity dispersion of the spectra. We see that the components of the capsule vectors are not only correlated with the spectroscopic redshift but also correlated with the apparent magnitudes and colours, measurements that are traditionally used by photometric redshift prediction algorithms. We also see that they are well correlated with parameters of a Sérsic fit which are indirect indicators of morphology as well as physical properties of the galaxies that would traditionally require spectroscopic measurements.

expensive to calculate contributions across all possibilities of the feature space, we use the expected gradients method which combines ideas from Integrated Gradients (Sundararajan, Taly & Yan 2017), SHAP (Lundberg & Lee 2017), and SmoothGrad (Smilkov et al. 2017) to approximately calculate the SHAP values for a neural network. A positive SHAP value indicates that the particular value of the feature increases the value of the output, a negative SHAP value indicates that the output is decreased, whereas a value of zero means that the feature does not contribute towards the output for that specific prediction. We then rank the features (i.e. capsule dimensions) based on their magnitude of SHAP values averaged over all predictions in the test set. Thus, a capsule dimension is deemed to be the most important if it influences the output most across all the predictions.

We show the SHAP values for each prediction in the test set in the summary plot shown in Fig. 10. The capsule dimensions are listed in decreasing order of their importance (i.e. average magnitude of SHAP values). The points are also colour coded as per the value of the feature which helps us to qualitatively identify how much the prediction changes based on a change in the value of the dimension. We see that capsule dimensions 8 and 14 are the most important, followed by dimensions 3 and 6. The next four capsule dimensions still contribute significantly to the prediction as dimensions 9, 7, 10, 12 have relatively high SHAP values. All the other dimensions contribute to the prediction significantly only a small number of times. For many of the dimensions, we see a pile up of SHAP values around 0. This indicates that the particular feature does not contribute much towards the prediction for that specific case. This can

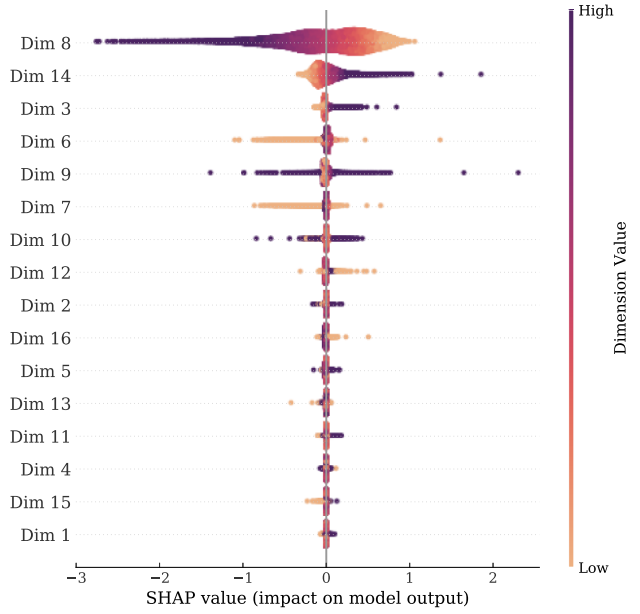


Figure 10. A SHAP summary plot showing the SHAP values of each capsule dimension for the entire test set. The capsule dimensions are listed in decreasing order of their importance (i.e. average magnitude of SHAP values). The points are colour-coded as per the value of the capsule dimension. We see that dimensions 8 and 14 are the most important followed by dimensions 3 and 6. The pile-up of points at a SHAP value of zero indicates that the dimension does not contribute towards the prediction for this specific case and the network gets similar information from another capsule dimension. This can happen when features are correlated. We do see that all the dimensions have some non-zero SHAP values, indicating that all the dimensions contribute towards the prediction at least sometimes.

happen if the input features are correlated and the model gets similar information from a different dimension for that specific prediction. This is also evident from the fact that the 2D UMAP projection of the capsules form a nearly perfect redshift sequence (see Fig. 7) suggesting that the data do not fully span the 16-dimensional latent space. We therefore define the importance ranking of a capsule dimension as an average over the entire test set and the ranking may be different for a specific prediction.

Dimension 8 has the highest SHAP feature importance. Although we cannot clearly discern what physical property it represents from the synthetic images generated from perturbed capsules, we can see from the figures in Appendix B that perturbing this dimension causes the image to morph from an elliptical galaxy to a spiral galaxy. We hypothesize that dimension 8 learns a representation which is a combination of the morphological type, colour, and orientation of the galaxy which helps it to distinguish between an elliptical galaxy which is intrinsically red and an edge-on spiral galaxy which appears to be reddened because of dust. This helps the capsule network to learn representations of galaxy colour while being aware of the morphology and orientation which can be very useful to break degeneracies in the colour–redshift relation. We also see that dimension 14 is the second most important feature. Fig. 8 shows that dimension 14 encodes information about the surface brightness of the observed galaxy. A lower value of dimension 14 corresponds to a brighter object. From Fig. 10 we see that a lower value of dimension 14 reduces the redshift prediction since they have a negative SHAP value. This shows that the neural network assigns a lower redshift to objects with higher surface brightness. Surface brightness is a very good proxy to the distance of a galaxy (and therefore redshift) since objects farther away appear fainter at a fixed luminosity. Learning a representation of surface brightness hence helps the network to better predict redshifts.

6 SUMMARY AND DISCUSSION

In this paper, we use a deep capsule network to produce photometric redshift point estimates from images of galaxies and provide interpretation of the features learnt by the network. We use ∼400 000 SDSS *ugriz* images, their spectroscopic redshifts, and morphological class labels from Galaxy-Zoo-1 (see Section 2) to train our deep capsule network. Capsule networks are a new type of neural network architecture that are better suited for identifying morphological features than traditional CNNs. We use a deep capsule network architecture that uses 3D convolution based routing mechanisms and skip connections to efficiently train the network (see Section 3 and Fig. 1).

We achieve a photometric redshift prediction accuracy comparable to or better than current methods while requiring less data and fewer trainable parameters (see Figs 3 and 6). The performance of our algorithm is stable across the brightness and redshift range of our data set (see Fig. 5). Moreover, the decision-making of our capsule network is easier to interpret as capsules act as a low-dimensional encoding of the input image and can be used to produce reconstructed images (see Fig. 2). We use UMAP, a non-linear dimensionality reduction method to embed the capsules in two-dimensional space and show that the capsules produce an almost perfect redshift sequence with the fraction of spirals in a region exhibiting a gradient roughly perpendicular to the redshift sequence (see Fig. 7). We then perturb the encodings of real galaxy images to generate synthetic galaxy images that demonstrate the image properties (e.g. size, orientation, and surface brightness) encoded by each capsule dimension (see Fig. 8). We calculate the feature importance of each capsule dimension using their SHAP values to rank them based on their usefulness towards predicting photo-z's (see Fig. 10). We also demonstrate that galaxy properties (e.g. magnitudes, colours, and stellar mass) correlate strongly with each capsule dimension (see Fig. 9). This tells us that the capsule dimensions encode and use visual and morphological properties of galaxy images (like surface brightness, orientation) in addition to measures of amount of light (like colours and magnitudes) to infer the photometric redshift.

Here, we have presented photo-z point estimates, though for many science cases photo-z PDFs are more desirable and sometimes necessary for meaningful analyses. However, current ML-based photo-z PDF estimation efforts suffer from poor calibration (Schmidt et al. 2020). In future work, we plan to incorporate methods described in Dey et al. (2021, 2022) to properly calibrate ML-based photo-z PDFs based on a galaxy's position in input space with capsule network photo-z PDFs serving as a natural example to demonstrate the expected improvements.

More generally, the future of capsule network-based photo-z estimation looks bright. Their high training efficiency will allow for deeper and wider models with greater capacity to handle the massive training sets from current and future spectroscopic surveys like DESI (DESI Collaboration 2016) and PFS (Takada et al. 2014) that extend to higher redshifts, span a wider redshift range, and probe to fainter magnitudes. Specifically, we plan to enable early DESI science by estimating photo-z's for objects in the DESI Legacy Imaging Surveys (Dey et al. 2019) before the DESI spectroscopic survey is complete. At even higher redshifts, we are optimistic that capsule networks can leverage morphology – especially the evolution

of galaxy morphologies from $z \sim 2$ to $z < 0.5$ – from space-based high-resolution imaging to help break the SED degeneracies that plague template-fitting methods at high-z. With growing high-z spectroscopic training sets and rapidly progressing capsule network architecture development, we are optimistic that capsule networks will provide complementary constraints or even superior photo-z's to template-based methods at high-z.

ACKNOWLEDGEMENTS

BD, BHA, and JAN acknowledge the support of the National Science Foundation under Grant No. AST-2009251. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Support for YYM was provided by NASA through the NASA Hubble Fellowship grant no. *HST*-HF2-51441.001 awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. RZ is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Argonne National Laboratory's work was supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357.

The pre-processed galaxy images and their spectrscopic redshifts used in this work were compiled by Emmanuel Bertin and sent to us by Johanna Pasquet and Marie Treyer. We are indebted to them for providing us early access to the data set and providing relevant documentation.

This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. We specifically acknowledge the assistance of Barry Moore II.

The authors would like to thank Rachel Bezanson, Sonja Cwik, Scott Dodelson, Ayres Freitas, Yasha Kaushal, Ann Lee, Christine Mazzola Daher, Alan Pearl, and David Setton for helpful discussions and suggestions during the course of this work. The authors would also like to thank the anonymous reviewer for their helpful comments and suggestions.

This research made use of the following software packages: ASTROPY (Astropy Collaboration 2013, 2018), COLORCET (Kovesi 2015), KERAS (Chollet et al. 2015), MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), PANDAS (Wes McKinney 2010; The pandas development team 2020), SCIKIT-LEARN (Pedregosa et al. 2011), SCIPY (Virtanen et al. 2020), SEABORN (Waskom 2021), SHAP (Lundberg & Lee 2017), TENSORFLOW (Abadi et al. 2015), UMAP (McInnes et al. 2018), and WEBPLOTDIGITIZER (Rohatgi 2020).

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III website is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

DATA AVAILABILITY

The images, spectroscopic redshifts, and associated value added catalogues from Pasquet et al. (2019) used in this work are available at https://deepdip.iap.fr/#item/60ef1e05be2b8ebb048d951d. PYTHON code used to train and test the models, additional value added catalogues and redshift predictions by our model along with interactive visualizations are available at https://biprateep.github.io/encapZulate-1/.

REFERENCES

Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems
Abbott B. P. et al., 2017, ApJ, 848, L12
Aihara H. et al., 2011, ApJS, 193, 29
Alam S. et al., 2015, ApJS, 219, 12
Almosallam I. A., Jarvis M. J., Roberts S. J., 2016, MNRAS, 462, 726
Amaro V. et al., 2019, MNRAS, 482, 3116
Apostolakos N., Degaudenzi H., Dubath F., Dubath P., Morisset N., Paltani S., Schefer M., 2019, in Molinaro M., Shortridge K., Pasian F., eds, ASP Conf. Ser. Vol. 521, Astronomical Data Analysis Software and Systems XXVI. Astron. Soc. Pac., San Francisco, p. 169
Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., 2008, ApJ, 683, 12
Battisti A. J. et al., 2019, ApJ, 882, 61
Beck R., Dobos L., Budavári T., Szalay A. S., Csabai I., 2016, MNRAS, 460, 1371
Benítez N., 2000, ApJ, 536, 571
Blanton M. R. et al., 2005, AJ, 129, 2562
Blanton M. R., Kazin E., Muna D., Weaver B. A., Price-Whelan A., 2011, AJ, 142, 31
Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503
Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, ApJ, 712, 511
Carrasco Kind M., Brunner R. J., 2014, MNRAS, 438, 3409
Cavuoti S., Amaro V., Brescia M., Vellucci C., Tortora C., Longo G., 2017, MNRAS, 465, 1959
Chollet F. et al., 2015, Keras
Cole S. et al., 2005, MNRAS, 362, 505
Collister A. A., Lahav O., 2004, PASP, 116, 345
Connolly A. J., Csabai I., Szalay A. S., Koo D. C., Kron R. G., Munn J. A., 1995, AJ, 110, 2655
Cybenko G., 1989, Math. Control. Signals Syst., 2, 303
DESI Collaboration, 2016, preprint (arXiv:1611.00036)
D'Isanto A., Polsterer K. L., 2018, A&A, 609, A111
D'Isanto A., Cavuoti S., Gieseke F., Polsterer K. L., 2018, A&A, 616, A97
Dalmasso N., Pospisil T., Lee A. B., Izbicki R., Freeman P. E., Malz A. I., 2020, Astron. Comput., 30, 100362
Dey A. et al., 2019, AJ, 157, 168
Dey B., Newman J. A., Andrews B. H., Izbicki R., Lee A. B., Zhao D., Rau M. M., Malz A. I., 2021, preprint (arXiv:2110.15209)
Dey B., Zhao D., Newman J. A., Andrews B. H., Izbicki R., Lee A. B., 2022, preprint (arXiv:2205.14568)
Eisenstein D. J. et al., 2005, ApJ, 633, 560
Euclid Collaboration, 2020, A&A, 644, A31

Feldmann R. et al., 2006, MNRAS, 372, 565
Firth A. E., Lahav O., Somerville R. S., 2003, MNRAS, 339, 1195
Freeman P. E., Newman J. A., Lee A. B., Richards J. W., Schafer C. M., 2009, MNRAS, 398, 2012
Fukushima K., Miyake S., 1982, Pattern Recognit., 15, 455
Geach J. E., 2012, MNRAS, 419, 2633
Geha M. et al., 2017, ApJ, 847, 4
Gomes Z., Jarvis M. J., Almosallam I. A., Roberts S. J., 2018, MNRAS, 475, 331
Graham M. L. et al., 2018, AJ, 155, 1
Grogin N. A. et al., 2011, ApJS, 197, 35
Gunn J. E. et al., 1998, AJ, 116, 3040
Gunn J. E. et al., 2006, AJ, 131, 2332
Guo C., Pleiss G., Sun Y., Weinberger K. Q., 2017, in Precup D., Teh Y. W., eds, Proceedings of Machine Learning Research Vol. 70, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. PMLR. p. 1321
Harris C. R. et al., 2020, Nature, 585, 357
Hayat M. A., Harrington P., Stein G., Lukic Z., Mustafa M., 2021, CoRR, abs/2101.04293
He K., Zhang X., Ren S., Sun J., 2015, Proceedings of the IEEE International Conference on Computer Vision (ICCV)
He K., Zhang X., Ren S., Sun J., 2016, 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, p. 770
Henghes B., Thiyagalingam J., Pettitt C., Hey T., Lahav O., 2022, MNRAS, 512, 1696
Hildebrandt H. et al., 2010, A&A, 523, A31
Hinton G., 2021, preprint (arXiv:2102.12627)
Hinton G. E., Krizhevsky A., Wang S. D., 2011, in Honkela T., Duch W., Girolami M. A., Kaski S., eds, Lecture Notes in Computer Science Vol. 6791, Artificial Neural Networks and Machine Learning - ICANN 2011-21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I. Springer, p. 44
Hornik K., 1991, Neural Netw., 4, 251
Hornik K., Stinchcombe M. B., White H., 1989, Neural Netw., 2, 359
Hoyle B., 2016, Astron. Comput., 16, 34
Hubble E., 1929, Proc. Natl. Acad. Sci., 15, 168
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ilbert O. et al., 2006, A&A, 457, 841
Ilbert O. et al., 2009, ApJ, 690, 1236
Jiang L. et al., 2014, ApJS, 213, 12
Jones E., Singal J., 2017, A&A, 600, A113
Katebi R., Zhou Y., Chornock R., Bunescu R., 2019, MNRAS, 486, 1539
Kauffmann G. et al., 2003, MNRAS, 341, 33
Kingma D. P., Ba J., 2015, in Bengio Y., LeCun Y., eds, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings
Koekemoer A. M. et al., 2011, ApJS, 197, 36
Kohonen T., 1981, Proceedings of the 2nd Scandinavian Conference on Image Analysis
Kohonen T., 1982, Biol. Cybern., 43, 59
Kovesi P., 2015, preprint (arXiv:1509.03700)
Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Bartlett P. L., Pereira F. C. N., Burges C. J. C., Bottou L., Weinberger K. Q., eds, Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States. p. 1106
LSST Science Collaboration, 2009, preprint (arXiv:0912.0201)
Lecun Y., 1985, Proceedings of Cognitiva 85, Paris, France. p. 599
LeCun Y., Boser B. E., Denker J. S., Henderson D., Howard R. E., Hubbard W. E., Jackel L. D., 1989, Neural Comput., 1, 541
LeCun Y., Bottou L., Bengio Y., Haffner P., 1998, Proc. IEEE, 86, 2278
LeCun Y., Bengio Y., Hinton G. E., 2015, Nature, 521, 436
Lee B., Chary R.-R., 2020, MNRAS, 497, 1935
Lee H., Grosse R. B., Ranganath R., Ng A. Y., 2009, in Danyluk A. P., Bottou L., Littman M. L., eds, ACM International Conference Proceeding Series Vol. 382, Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009. ACM, p. 609
Li C. et al., 2022, MNRAS, 509, 2289
Lintott C. et al., 2011, MNRAS, 410, 166
Liu Z., Mao H., Wu C.-Y., Feichtenhofer C., Darrell T., Xie S., 2022, preprint (arXiv:2201.03545)
Lundberg S. M., Lee S., 2017, in Guyon I., von Luxburg U., Bengio S., Wallach H. M., Fergus R., Vishwanathan S. V. N., Garnett R., eds, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. p. 4765
McInnes L., Healy J., 2018, CoRR, abs/1802.03426
McInnes L., Healy J., Saul N., Grossberger L., 2018, J. Open Source Softw., 3, 861
Mao Y.-Y., Geha M., Wechsler R. H., Weiner B., Tollerud E. J., Nadler E. O., Kallivayalil N., 2021, ApJ, 907, 85
Masters D. et al., 2015, ApJ, 813, 53
Masters D. C. et al., 2019, ApJ, 877, 81
Mazzia V., Salvetti F., Chiaberge M., 2021, CoRR, abs/2101.12491
Myles J. et al., 2021, MNRAS, 505, 4249
Nakoneczny S. J. et al., 2021, A&A, 649, A81
Narayan A., Berger B., Cho H., 2020, bioRxiv
Padmanabhan N. et al., 2008, ApJ, 674, 1217
Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Perlmutter S. et al., 1999, ApJ, 517, 565
Rajasegaran J., Jayasundara V., Jayasekara S., Jayasekara H., Seneviratne S., Rodrigo R., 2019, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
Ranzato M., Huang F. J., Boureau Y., LeCun Y., 2007, 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA. IEEE Computer Society
Razim O., Cavuoti S., Brescia M., Riccio G., Salvato M., Longo G., 2021, MNRAS, 507, 5034
Riess A. G. et al., 1998, AJ, 116, 1009
Rohatgi A., 2020, Webplotdigitizer: Version 4.4
Rumelhart D. E., Hinton G. E., Williams R. J., 1986, Nature, 323, 533
Sabour S., Frosst N., Hinton G. E., 2017, in Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., eds, Advances in Neural Information Processing Systems 30. Curran Associates, Inc., p. 3856,
Salvato M., Ilbert O., Hoyle B., 2019, Nat. Astron., 3, 212
Sarmiento R., Huertas-Company M., Knapen J. H., Sánchez S. F., Domínguez Sánchez H., Drory N., Falcón-Barroso J., 2021, preprint (arXiv:2104.08292)
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Schmidt S. J. et al., 2020, MNRAS, 499, 1587
Schuldt S., Suyu S. H., Cañameras R., Taubenberger S., Meinhardt T., Leal-Taixé L., Hsieh B. C., 2021, A&A, 651, A55
Shapley L. S., 1953, in Kuhn H., Tucker A., eds, Contributions to the Theory of Games II, A Value for n-Person Games. Princeton University Press, Princeton, p. 307
Smee S. A. et al., 2013, AJ, 146, 32
Smilkov D., Thorat N., Kim B., Viégas F., Wattenberg M., 2017, preprint (arXiv:1706.03825)
Stabenau H. F., Connolly A., Jain B., 2008, MNRAS, 387, 1215
Stein G., Harrington P., Blaum J., Medan T., Lukic Z., 2021, preprint (arXiv:2110.13151)
Sundararajan M., Taly A., Yan Q., 2017, preprint (arXiv:1703.01365)
Szegedy C. et al., 2015, IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society, p. 1

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z., 2016, 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, p. 2818
Székely G. J., Rizzo M. L., Bakirov N. K., 2007, Ann. Stat., 35, 2769
Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C., Giordano G., 2003, Neural Networks for Photometric Redshifts Evaluation. p. 226
Takada M. et al., 2014, PASJ, 66, R1
The pandas development team, 2020, pandas-dev/pandas: Pandas
Tremonti C. A. et al., 2004, ApJ, 613, 898
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Wadadekar Y., 2005, PASP, 117, 79
Waskom M. L., 2021, J. Open Source Softw., 6, 3021
Way M. J., Foster L. V., Gazis P. R., Srivastava A. N., 2009, ApJ, 706, 623
Wes M., 2010, Stéfan van der Walt Jarrod Millman eds, Proceedings of the 9th Python in Science Conference. p. 56
Wright A. H., Hildebrandt H., van den Busch J. L., Heymans C., 2020, A&A, 637, A100
York D. G. et al., 2000, AJ, 120, 1579
Zadrozny B., Elkan C., 2001, in Brodley C. E., Danyluk A. P., eds, Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001. Morgan Kaufmann. p. 609
Zadrozny B., Elkan C., 2002, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada. ACM, p. 694
Zhao D., Dalmasso N., Izbicki R., Lee A. B., 2021, preprint (arXiv:2102.10473)
Zhou R. et al., 2021, MNRAS, 501, 3309

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Figure S1. Reconstructions from perturbed capsule vectors.
Figure S2. A few examples of strong correlations between capsule dimensions and physical properties of galaxies visualized using scatter plots.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: CAPSULE NETWORKS AND ROUTING MECHANISMS

To construct the classification-and-encoding network, we first use a set of convolutional filters, the outputs of which are reshaped into a set of tensors which are treated as the initial set of capsules. We then use two main kind of capsule layers, dynamic routing-based class capsules and convolution routing-based capsules (i.e. `Conv-Caps` and `3D-Conv-Caps` layers in Fig. 1). As shown in Fig. 1, the convolution routing based capsules are used to construct the hidden layers whereas the dynamic routing based capsules are used to construct the output layer of the classification-and-encoding network where each capsule represents a morphological type. In this section, we give a brief overview of the mathematical aspects of the capsule layer architectures used in this work. This is intended to be a short summary and interested readers are recommended to refer to Sabour et al. (2017) for a detailed discussion on capsules with dynamic routing and Rajasegaran et al. (2019) for convolutional capsules. We have tried to follow the same mathematical notation used by these two works for easy reference.

A.1 Dynamic routing (i.e. routing by agreement)

Let $\mathbf{u}_i$ denote the ith capsule vector in layer l of the network and $\mathbf{v}_j$ denote the jth capsule vectors in layer $l+1$. To obtain the capsules in layer $l+1$ from the ones in layer l we define an intermediate 'prediction' vector ($\hat{\mathbf{u}}_{j|i}$) as

$$\hat{\mathbf{u}}_{j|i} = \mathbf{W}_{ij}\mathbf{u}_i, \tag{A1}$$

where $\mathbf{W}_{ij}$ is a weight matrix learnt by gradient descent. The capsules in the following layer ($\mathbf{v}_j$) are calculated using a weighted sum of these prediction vectors after being passed through a non-linear activation function called the squashing function defined as

$$\mathbf{v}_j = \frac{\|\mathbf{s}_j\|^2}{1+\|\mathbf{s}_j\|^2}\frac{\mathbf{s}_j}{\|\mathbf{s}_j\|}, \tag{A2}$$

where $\mathbf{s}_j$ is the weighted sum given by

$$\mathbf{s}_j = \sum_i c_{ij}\hat{\mathbf{u}}_{j|i}, \tag{A3}$$

where c_{ij} are the coupling coefficients determined by an iterative process. To ensure that they always add up to 1, they are defined in terms of the softmax transformed variables b_{ij} as

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}. \tag{A4}$$

The variables b_{ij} can be treated as the log prior probability that the capsule i in layer l is coupled to the capsule j in layer $l+1$. In a single pass of back propagation, we begin with $b_{ij}=0$ to provide equal weights to all the capsules initially, and then the coupling coefficients are iteratively updated by measuring the agreement between the current output of each capsule in layer $l+1$, i.e. $\mathbf{v}_j$ and the prediction made by the capsules in layer l, i.e. $\hat{\mathbf{u}}_{j|i}$. The agreement is defined as the scalar product $\mathbf{v}_j.\hat{\mathbf{u}}_{j|i}$ and is added to b_{ij} before computing the coupling coefficients. So, for each step in the iteration:

$$b_{ij} \leftarrow b_{ij} + \mathbf{v}_j.\hat{\mathbf{u}}_{j|i}. \tag{A5}$$

The number of iterations is a tunable hyperparameter. Larger number of iterations will provide better estimates of the coupling coefficients at the cost of increasing the number of computations. We use three iterations as it was found to work reasonably well by Sabour et al. (2017) who proposed this algorithm. Since these capsules ($\mathbf{v}_j$) form the final layer of the classification-and-encoding network, we calculate their Euclidean norms which are used as a measure of the class probabilities the capsules represent. These predicted class probabilities are then used as inputs to the margin loss function (equation 1).

A.2 Convolution based routing

One of the drawbacks of the dynamic routing algorithm described in Section A1 is that the computations are done in a way analogous to fully connected neural networks. This means that the number of trainable weights increase dramatically for a deep network architecture required for complex tasks like predicting photo-z's. To solve this problem, Rajasegaran et al. (2019) proposed capsule network layers that use computationally efficient convolutional operations. We use them as the intermediate layers of our classification-and-encoding network. The weights of the convolution filters are determined using gradient descent whereas the coupling coefficients for routing are determined by an iterative process. In the initial layers, the feature maps obtained from convolution operations is large and iterative routing can be expensive. So, following Rajasegaran et al. (2019),

we use a mix of two kinds of convolutional capsule layers, one which does one routing iteration (viz. `Conv-Caps`) and another one doing three routing iterations (viz. `3D-Conv-Caps`) in our network architecture (see Fig. 1).

To facilitate convolution operations, the capsules start out as 3D tensors which are flattened into 1D capsule vectors when we reach the final layer in our architecture. Let the output of the convolutional capsule layer l be $\Phi^l \in \mathbb{R}^{(w^l, w^l, c^l, n^l)}$, where w^l denotes the height and width, c^l the depth, and n^l the number of 3D capsule tensors. Similarly, let $\Phi^{l+1} \in \mathbb{R}^{(w^{l+1}, w^{l+1}, c^{l+1}, n^{l+1})}$ represent the output of the layer $l+1$.

The `Conv-Caps` layer first reshapes Φ^l into a tensor of shape $(w^l, w^l, c^l \times n^l)$ and convolves with $(c^{l+1} \times n^{l+1})$ number of filters, producing $(c^{l+1} \times n^{l+1})$ number of feature maps of shape (w^{l+1}, w^{l+1}). They are then reshaped into a tensor of shape $(w^{l+1}, w^{l+1}, c^{l+1}, n^{l+1})$. This 3D tensor (S_{pqr}) is then used as the input to a non-linear squashing function defined by

$$\hat{S}_{pqr} = \frac{\|S_{pqr}\|^2}{1 + \|S_{pqr}\|^2} \frac{S_{pqr}}{\|S_{pqr}\|}. \tag{A6}$$

Since we will use just one iteration of routing for this layer, the output of the squashing function is treated as the output of the layer (i.e. $\Phi^{l+1} = \hat{\mathbf{S}}$).

For the `3D-Conv-Caps` layer, we first reshape $\Phi^{\mathbf{l}}$ into a tensor of shape $(w^l, w^l, c^l \times n^l, 1)$. Then, it is convolved with $(c^{l+1} \times n^{l+1})$ number of 3D convolution kernels of appropriate shape so as to produce a tensor of shape $(w^{l+1}, w^{l+1}, c^l, c^{l+1} \times n^{l+1})$. It is then reshaped into a tensor, $\tilde{\mathbf{V}}$ of shape $(w^{l+1}, w^{l+1}, c^l, n^{l+1}, c^{l+1})$ which acts as the intermediate 'prediction' tensor. The capsules of the following layer are then calculated via the weighted sum of tensors given by

$$S_{pqr} = \sum_s k_{pqrs} . \tilde{V}_{pqrs}. \tag{A7}$$

Then $\mathbf{S}$ is used as an input to the tensor squashing function defined in equation A6 to obtain the squashed tensor, $\hat{\mathbf{S}}$ which after the iterative updates will be treated as the output capsules($\hat{\Phi}^{l+1}$). The coupling coefficients for the weighted sum (k_{pqrs}) are determined by an iterative process. To ensure that they are normalized they are defined in terms of softmax transformed variable $\mathbf{B} \in \mathbb{R}^{(w^{l+1}, w^{l+1}, c^{l+1}, c^l)}$ given by

$$k_{pqrs} = \frac{\exp(b_{pqrs})}{\sum_x \sum_y \sum_z \exp(b_{xyzs})}. \tag{A8}$$

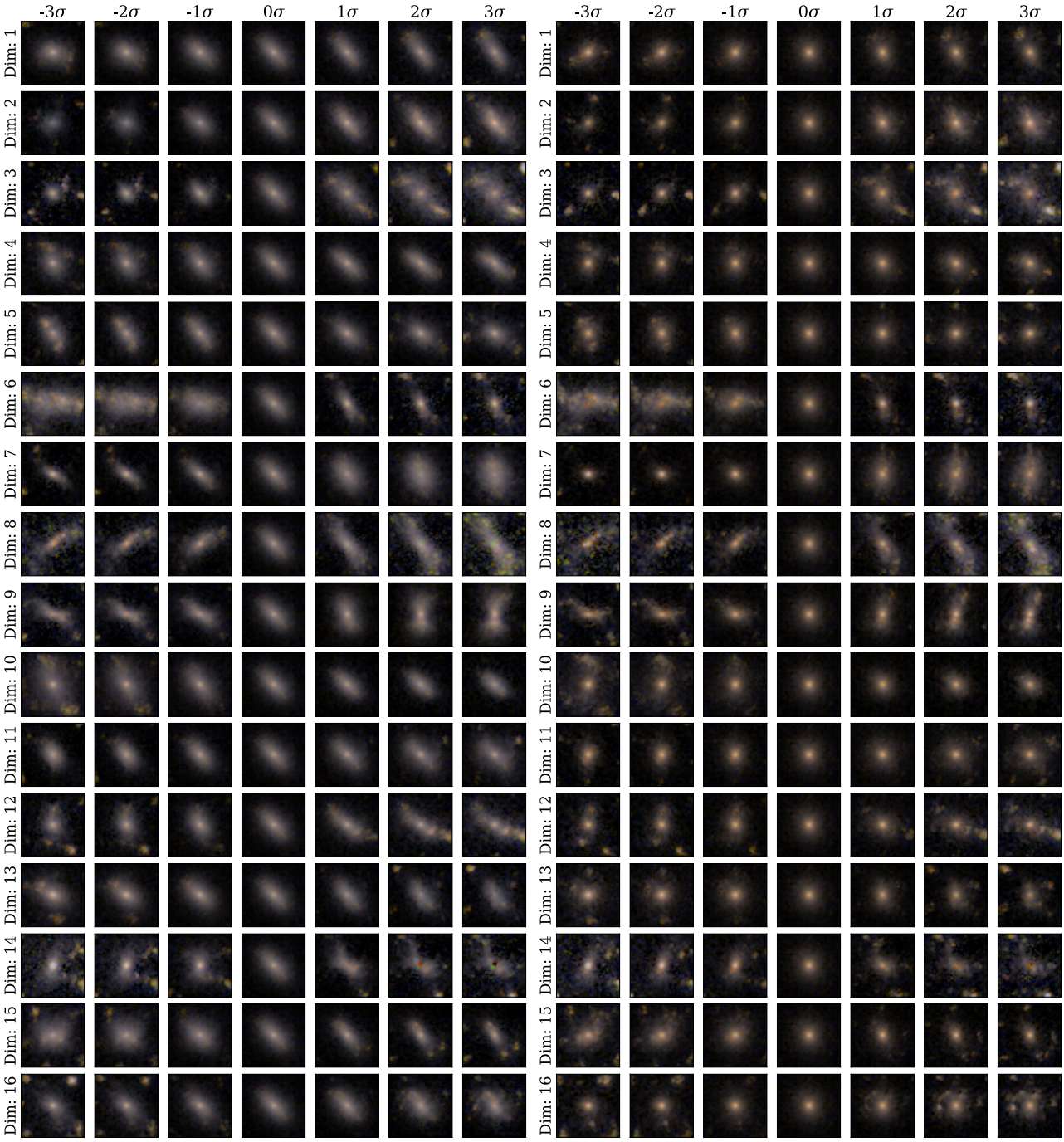
In a single pass of back propagation we begin with $b_{pqrs} = 0$ to provide equal weight to all capsules initially and then the coupling coefficients are iteratively updated 3 times by measuring the agreement (via the scalar product) between the current output of the capsules and the intermediate prediction tensors in each iteration i.e.

$$b_{pqrs} \leftarrow b_{pqrs} + \hat{S}_{pqr} . \tilde{V}_{pqrs}. \tag{A9}$$

Finally, when the output of the convolutional capsules are used as inputs to the capsules with dynamic routing, the tensors in a layer l of shape (w^l, w^l, c^l, n^l) are flattened to the shape $(w^l \times w^l \times c^l, n^l)$, i.e. we get n^l number of capsule vectors each with $w^l \times w^l \times c^l$ number of dimensions.

APPENDIX B: SYNTHETIC IMAGES FROM PERTURBED CAPSULE COMPONENTS

Here, we show an extended version of Fig. 8 with synthetic galaxy images generated from perturbing all 16 of the dimensions individually. Each column shows the decoded image when one of the 16 dimensions of the capsule vector is perturbed in units of its standard deviation (keeping all the others fixed). The 0σ column shows the decoded image from the unperturbed capsule and are identical for each row. Since the capsule network training process does not disentangle the features learnt by each dimension, not all the dimensions control a single easily identifiable feature. A subset of the dimensions for which the features are easily identifiable are shown in Fig. 8.



(a) The first spiral galaxy from Fig. 2

(b) The first elliptical galaxy from Fig. 2

Figure B1. Reconstructions from perturbed capsule vectors. Each column shows the reconstructions when one of the 16 components of the capsule vector is perturbed in units of their standard deviation (keeping all the others fixed). This is an extended version of the Fig. 8 and shows reconstructions from perturbations of all the dimensions.

APPENDIX C: CORRELATIONS OF CAPSULE DIMENSIONS WITH PHYSICAL PROPERTIES

A few illustrative examples of strong correlations between capsule dimensions and physical properties of galaxies have been visualized using scatter plots in Fig. C1. We observe that the value of the capsule dimensions varies with the galaxy property indicating some correlation.

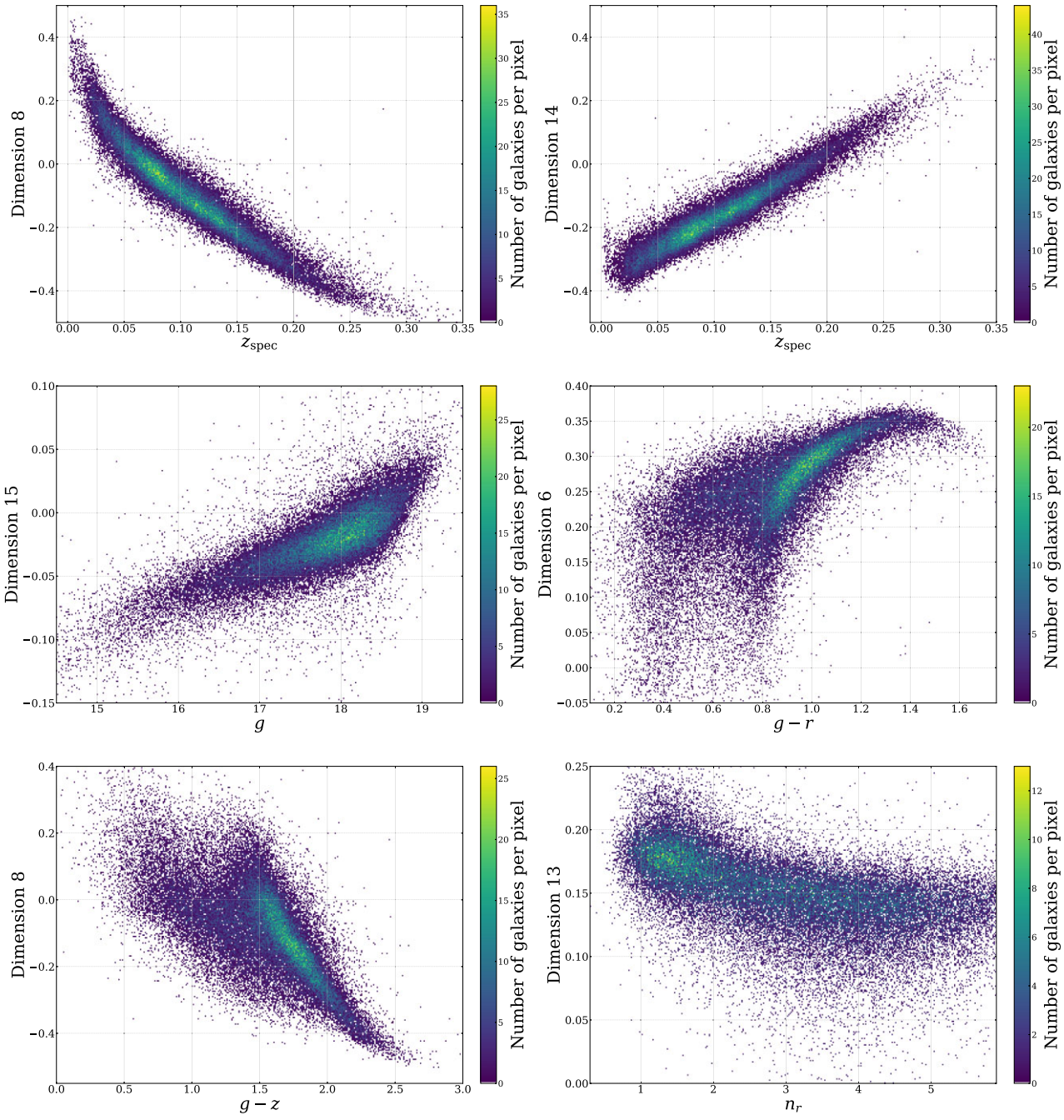


Figure C1. A few examples of strong correlations between capsule dimensions and physical properties of galaxies visualized using scatter plots. g represents the extinction corrected SDSS g-band cmodel magnitude. $g-r$ and $g-z$ represent galaxy colours calculated using extinction corrected model magnitudes. n_r represents the Sérsic index obtained from a Sèrsic profile to the r-band photometry. We observe that the value of the capsule dimensions varies with the galaxy property indicating some correlation.

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