Deep-CEE I: Fishing for Galaxy Clusters with Deep Neural Nets

Matthew C. Chan[1] and John P. Stott[1]
E-mails: m.c.chan@lancaster.ac.uk and j.p.stott@lancaster.ac.uk
[1] *Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We introduce Deep-CEE (**Deep** Learning for Galaxy **C**luster **E**xtraction and **E**valuation), a proof of concept for a novel deep learning technique, applied directly to wide-field colour imaging to search for galaxy clusters, without the need for photometric catalogues. This technique is complementary to traditional methods and could also be used in combination with them to confirm existing galaxy cluster candidates. We use a state-of-the-art probabilistic algorithm, adapted to localise and classify galaxy clusters from other astronomical objects in SDSS imaging. As there is an abundance of labelled data for galaxy clusters from previous classifications in publicly available catalogues, we do not need to rely on simulated data. This means we keep our training data as realistic as possible, which is advantageous when training a deep learning algorithm. Ultimately, we will apply our model to surveys such as LSST and *Euclid* to probe wider and deeper into unexplored regions of the Universe. This will produce large samples of both high redshift and low mass clusters, which can be utilised to constrain both environment-driven galaxy evolution and cosmology.

Key words: galaxies: clusters: general – methods: statistical – methods: data analysis – techniques: image processing

1 INTRODUCTION

Galaxy clusters are the largest gravitationally bound systems in the Universe. We study galaxy clusters to understand the environmental effects on galaxy evolution and to determine the cosmological parameters that govern the growth of large scale structure in the Universe. In order to do this, large well understood samples of clusters across a range of masses and redshifts are required.

Throughout the 1950s to 1980s the astronomer George Abell created the 'Abell catalogue' containing 4073 galaxy clusters (Abell et al. 1989), which we now refer to as Abell galaxy clusters. George Abell used a magnifying glass to manually examine photographic plates and looked specifically for over-dense regions of galaxies. He could then measure or estimate properties such as distance, richness and galaxy magnitudes for each cluster (Cudnik 2013). That would be the last time that a wide-field cluster search would be conducted manually by eye.

Since then a variety of techniques have been developed and used to search for galaxy clusters. The primary technique for extracting clusters from imaging data is red sequence fitting (e.g. Gladders & Yee 2000, Gladders & Yee 2005, Koester et al. 2007, Rykoff et al. 2014a, Rykoff et al. 2014b and Rykoff et al. 2016). Unlike the Abell method, this technique is applied to photometric catalogue data extracted from imaging, as opposed to the images themselves. It identifies clusters via the distinctive red sequence slopes containing red, passive galaxies found in colour magnitude diagrams. In the charge-coupled device (CCD) era this catalogue-based technique has proven to be an efficient alternative to by-eye searches.

Both X-ray emission (e.g. Sarazin 1986, Ebeling et al. 1998, Ebeling et al. 2000a, Ebeling et al. 2000b, Böhringer et al. 2004, Takey et al. 2011, Mehrtens et al. 2012 and Adami, C. et al. 2018) and the Sunyaev-Zeldovich (SZ, Sunyaev & Zeldovich 1972) effect (e.g. E. Carlstrom et al. 2009, Staniszewski et al. 2009, Vanderlinde et al. 2010, Planck Collaboration et al. 2014, Bleem et al. 2015, Planck Collaboration: et al. 2015, Planck Collaboration et al. 2016, Hilton et al. 2018 and Hurier et al. 2017) reveal the presence of galaxy clusters through the properties of the hot intracluster medium (ICM). The ICM emits strongly at X-ray wavelengths, so X-ray telescopes such as *XMM-Newton* and *Chandra* can be used to search for clusters, which appear as extended sources. Low energy photons from the Cosmic Microwave Background (CMB) radiation experience Compton scattering when they interact with the high energy electrons of the ICM. This is the SZ effect, in which the presence of a galaxy cluster leaves a shadow on the CMB itself at the galaxy clusters location.

Finally, as clusters are massive structures their presence can also be inferred via weak gravitational lensing (e.g. Wittman et al. 2006 and Umetsu 2010). Statistical techniques are applied to wide-field galaxy surveys to search for minute signatures of the convergence and alignment of the shears produced by gravitational lensing.

However X-ray, SZ and weak lensing techniques need to optically confirm their candidate clusters as there are contaminants (e.g. active galactic nuclei [AGN] or nearby galaxies) and line-of-sight coincidences (e.g. unrelated low mass groups at different redshifts) that can conspire to give a false detection. This confirmation has to be done by eye, which is time inefficient, or by relying on a red sequence selection again. This can introduce biases or result in an uncertain selection function. Therefore an approach that can produce fast and precise analysis of imaging data would be advantageous to both search for or confirm galaxy clusters.

The Large Synoptic Survey Telescope (LSST, Ivezić et al. 2019) is currently under construction in Chile and engineering first-light is expected to be in 2020. LSST will be the deepest wide-field optical survey ever conducted, performing multiple scans of the entire Southern sky over ten years, with an estimated 15 TB of data generated per night. *Euclid* (Amiaux et al. 2012) is a wide-field space telescope that is due to commence operation in 2022. It will conduct a weak lensing survey to probe the nature of Dark Matter and Dark Energy, with an estimated 1PB of data generated per year. Data mining techniques such as deep learning will be required to analyse the enormous outputs of these telescopes. LSST and *Euclid* will observe thousands of previously unknown galaxy clusters across a wide range of masses and redshifts but cataloguing them presents a significant challenge.

During the last two decades computing power has significantly improved (Roser & Ritchie 2019), as such deep learning techniques have become an increasingly popular approach to replace repetitive manual tasks. In particular, convolutional neural networks (CNN) have been successful in the field of computer vision, where CNNs are designed to mimic the human brain at learning to perceive objects by activating specific neurons upon visualising distinctive patterns and colours. We can train and utilise CNNs to process high-dimensional features directly from digital images into a meaningful understanding with minimal human input (Huang 1996). LeCun et al. (1999) first introduced a deep learning approach using CNNs to classify uniquely handwritten digits in images from the MNIST dataset achieving error rates of less than 1 per cent.

Deep learning is very applicable in astronomy due to the abundance of imaging data available from modern telescopes. This makes it preferable when conducting data mining tasks such as classification, regression and reconstruction. These include determining the morphology of galaxies (e.g. Khalifa et al. 2018, Walmsley et al. 2019 and Zhu et al. 2019), identifying gravitational lenses (e.g. Hezaveh et al. 2017, Schaefer, C. et al. 2018, Pearson et al. 2018, Davies et al. 2019, Ribli et al. 2019 and Petrillo et al. 2019), predicting galaxy cluster mass (e.g. Ntampaka et al. 2015, Ntampaka et al. 2016, Ho et al. 2019, Armitage et al. 2019 and Ntampaka et al. 2019), predicting photometric redshift (e.g. Hoyle 2016, Pasquet et al. 2019 and Chong & Yang 2019), generating synthetic surveys (Smith & Geach 2019), denoising of astronomical images (Flamary 2016) and astronomical object classification (e.g. Hála 2014 and Kim & Brunner 2017). However, a deep learning approach has yet to be developed to detect and determine intrinsic properties of galaxy clusters from wide-field imaging data.

Conventional CNN classifiers are adept at distinguishing learned features in images but rather naive at determining their positions in an image. For this paper we want to apply a deep learning approach that can efficiently localise and classify objects in images. Szegedy et al. (2013) demonstrates a deep learning approach to perform object detection in images by modifying the architecture of CNNs into modules that are specific to classification and localisation tasks, where objects with importance are classed as 'foreground' whilst everything else is considered as 'background'.

TensorFlow (Abadi et al. 2015) is an open source data science library that provide many high level application programming interfaces (API) for machine learning. The object detection API[1] (Huang et al. 2016) contains multiple state-of-the-art deep learning algorithms, which are designed to enhance the speed or accuracy of a model. These include SSD (Single Shot Detection, Liu et al. 2015) and Faster-RCNN (Faster Region-based CNN, Ren et al. 2015). Huang et al. (2016) finds that the Faster-RCNN algorithm returns high precision for predictions on the COCO dataset and is suitable for large input images during training and testing. However, the algorithm can take a long time to train and be slow at generating predictions. Whilst the SSD algorithm is quick to train and produces fast predictions, but the overall precision of predictions is lower compared to Faster-RCNN. We choose the Faster-RCNN algorithm as we prefer accuracy over speed.

We organise this paper in the following format. We split §2 into two subsections to outline our methodology. In §2.1 we explain the concept behind the Deep-CEE model[2] and in §2.2 we describe the procedure to create the training and test sets. In §3 we analyse the performance of our model with the test set (see §3.1) and we also assess our model on an unseen dataset (see §3.2). In §4 we discuss the limitations and future applications of our model. Finally, in §5 we summarise this paper.

Throughout this paper, we adopt a ΛCDM cosmology with $H_0 = 71$ km s^{-1} Mpc^{-1}, $\Omega_m = 0.27$ and $\Omega_\Lambda = 0.73$.

2 METHOD

2.1 Deep Learning Method

We use a supervised learning approach (Kotsiantis 2007) to train the Faster-RCNN algorithm by providing it labelled data. The architecture of the algorithm can be seen in Figure 1. It is comprised of three different individual networks that

[1] The full list of object detection algorithms can be found at: `https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md` (Huang et al. 2016).

[2] We use the term 'model' to represent the trained 'Deep-CEE' Faster-RCNN algorithm.

work collectively during the training phase. These three networks are called the Feature Network (FN), Region Proposal Network (RPN) and Detection Network (DN). The convolution layers in the RPN and the fully-connected (FC) layers in the DN require training as the other layers in the RPN and the DN have no trainable parameters. To train our model, we use a joint training (Ren et al. 2015) approach, which means we allow the outputs from all the networks to be generated before all trainable layers are updated. Throughout this section we adopt a similar methodology and hyper-parameters as described in Ren et al. (2015). We set a learning rate of 0.0002, momentum of 0.9, gradient clipping threshold of 10 and a mini-batch size of one[3]. We perform random initialisation of the weights from a zero-mean truncated Gaussian distribution with a standard deviation of 0.01 for the RPN and we apply variance scaling initialisation (Glorot & Bengio 2010) from a uniform distribution for the weights in the DN. We also initialise bias values for the trainable layers in the RPN and DN to be zero.

2.1.1 *Feature Network*

The FN is found at the beginning of the Faster-RCNN algorithm and takes an image as the input. We apply transfer learning (Torrey & Shavlik 2009) by using a pre-trained CNN called Inception-v2 (Szegedy et al. 2015) as the architecture of the FN. Inception-v2 consists of convolution, ReLU (rectified linear unit) activation and pooling layers[4] in Inception-v2 are not included for the FN since we do not want to perform classification or regression in this network. (Schmidhuber 2014). The convolution and ReLU activation layers are responsible for extracting non-linear features from an image such as straight lines, edges and curves. The pooling layers then down-sample the convolved image to form a feature map. The reason we choose the Inception-v2 architecture as opposed to other architectures such as VGG16 (Simonyan & Zisserman 2014) and AlexNet (Krizhevsky et al. 2012) is that it has been designed to reduce the number of parameters needed in the network. This means less computational power is required to train the algorithm, while still retaining high accuracy. Inception-v2 has been pretrained to recognise objects from the COCO (The Common Objects in Context) dataset (Lin et al. 2014), which contains images of commonly found objects in daily life such as vehicles and animals. This means we do not have to fully retrain the weights and biases in the network since they are sufficiently optimised at finding generic structures, as retraining every single weight and bias from scratch is computationally expensive. Furthermore we do not alter the architecture of Inception-v2, as it has already been tuned for object detection.

2.1.2 *Region Proposal Network*

The RPN is found after the FN and consists of a shallow architecture of three convolution layers with a ReLU activation layer specific only to the first convolution layer, see Figure 1. The weights and biases in the the first convolution layer are shared for classification and localisation tasks whilst the remaining convolution layers are separated into two parallel convolution layers, with independent weights and biases for each task. The RPN takes the feature map output from the FN as its input. The role of the RPN is to generate possible positions at which an object could be located within an image. In the first convolution layer, a 3×3 pixel sliding window is used with zero-padding[5] and a pixel stride of one, which translates to every sixteenth pixel in the original image. At the centre of each sliding window an 'anchor' is placed. Each anchor has a set number of different sized boxes generated around it. The dimensions and number of boxes is dependent on the scaling and aspect ratios. We set scaling ratios of 0.25, 0.5, 1.0 and 2.0 and aspect ratios of 0.5, 1.0 and 2.0. We choose these values to reflect the dimensions of the ground truth boxes in all of the images. A scaling ratio of 1.0 relates to a box of 256×256 pixels in the original image, such that setting other values for the scaling ratio would generate additional larger or smaller boxes at each anchor. The aspect ratio produces boxes around each anchor that have adjusted widths and heights with respect to each scaling ratio. This means at every anchor there are twelve boxes of different sizes. In the final convolution layers, a 1×1 pixel sliding window is used with no-padding[6] and a pixel stride of one. This ensures fixed dimensions for the output of this layer.

Any box that overlaps by more than 70 per cent with the ground truth box is assigned as a positive 'foreground' label or otherwise a negative 'background' label if less than 30 per cent, whilst boxes between these overlap thresholds are ignored. From which 128 positive labelled boxes and 128 negative labelled boxes are randomly chosen for each image to update the weights and biases, so that the RPN learns to distinguish important objects as 'foreground' and irrelevant objects as 'background'. If there are fewer than 128 positive labelled boxes in an image then additional negative labelled boxes with the next highest percentage overlapping are chosen to represent positive labels. Therefore the RPN will learn two outputs: whether a box is likely to contain a ground truth object and whether a box is not likely to contain a ground truth object, based on percentage overlap with the ground truth box. During the testing phase, if a box has a high probability of containing an object then this box will be passed onto the next stage in the Faster-RCNN algorithm. However, if a box has a high probability of not containing an object then the box is disregarded. Backpropogation (BP, Rumelhart et al. (1986)) and stochastic gradient descent (SGD, Bottou (2010)) is used to train the weights and biases in RPN[7].

Two additional steps are applied to limit the number of boxes for faster computation. Firstly, any box which extends outside the image borders are disregarded after the boxes are generated. Non-Maximum Suppression (NMS, Hosang et al.

[3] The definitions for each hyper-parameter is beyond the scope of this paper but is explained in Ruder (2016).

[4] The fully-connected layers and softmax function

[5] Additional layers of pixels are added around the edge of an image with values of zero. This helps to preserve the dimensions of the input as it passes through the layer.

[6] No additional pixels are added around the edge of an image.

[7] Back-propogation with respect to box coordinates is disabled as it causes the training to be unstable (Huang et al. 2016).

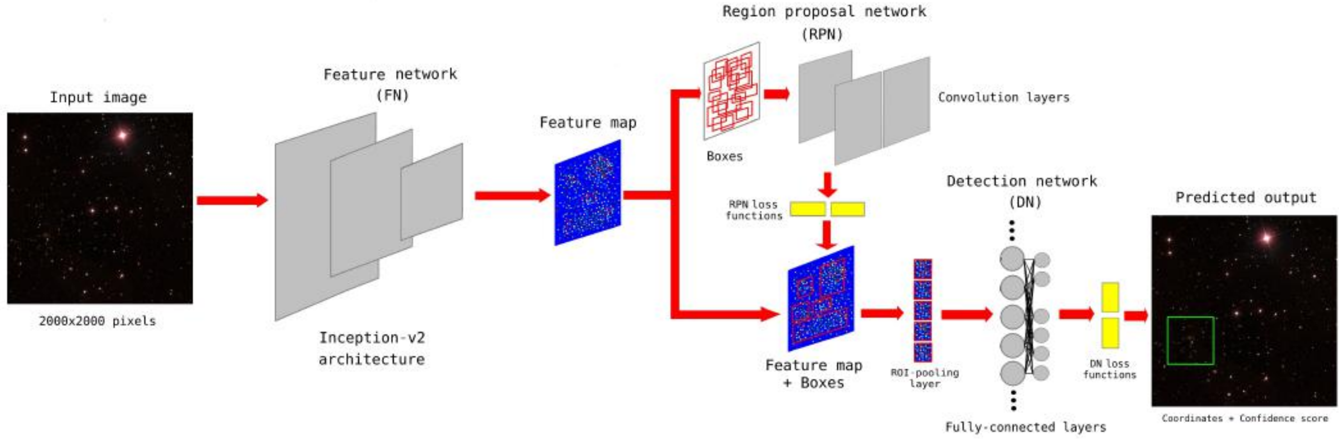

Figure 1. High-level overview of the architecture for the Faster-RCNN algorithm which contains the Feature Network, Region Proposal Network and Detection Network. The output from each network is used as the input for the next network. The algorithm's architecture is similar to the system demonstrated in Figure 2 from Ren et al. (2015). For simplicity, the Inception-v2 architecture is not displayed fully. It should be noted that '*Mixed_4e*' is used as the final layer of the Feature Network (Huang et al. 2016). The full details of the Inception-v2 architecture can be found in Ren et al. (2018). The RPN and the DN loss functions are only active during the training phase. A softmax function (Nwankpa et al. 2018) is used with the RPN and DN loss functions for classification of proposed boxes.

(2017)) is used on the boxes after the losses are calculated. NMS keeps the highest overlapping box with the ground truth box and disregards any remaining boxes that overlap by more than 70 per cent with this box. These steps are repeated on the remaining boxes, such that the next box with the highest overlap is kept and any other box with more than 70 per cent overlap with this box are also disregarded. This procedure continues until there are fewer than 300 boxes for each image.

We utilise two loss functions (categorical cross-entropy loss and smooth L1-loss) to calculate prediction errors in the RPN, these two loss functions are associated to separate convolution layers in the final layer of the RPN. The categorical cross-entropy loss layer (Martinez & Stiefelhagen 2018) creates a probability distribution for all the proposed boxes, where the sum of all probabilities equals one. This function is described in Equation 1:

$$L_{cls}(p_i, p_i^*) = -p_i^* log(pi_i) - (1 - p_i^*)log(1 - p_i), \tag{1}$$

where p_i is the predicted probability of a box and p_i^* is zero or one depending on whether the ground truth box is correctly classified. The categorical cross-entropy loss function calculates the objectness error for boxes being predicted as 'foreground' and 'background'.

The smooth L1-loss layer (Girshick 2015) only considers positive labelled boxes in the training phase. It takes into account the distance between the centre coordinates of the ground truth box and predicted boxes, and also the difference in size of the boxes compared to the ground truth box. This function is seen in Equation 2:

$$L_{reg}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise}, \end{cases} . \tag{2}$$

where $x = (t_i - t_i^*)$ is the error between the ground truth and the predicted boxes. The smooth L1-loss function penalises localisation error by taking the absolute value (behaves like L1-loss) for large errors and the square value (behaves like L2-loss) for small errors (Ng 2004). This encourages stable regularisation of the weights and biases during training.

The proposed boxes from the RPN are merged with the feature maps from the FN, so that each box is overlaid on an 'object' in a feature map. An ROI-pooling (region-of-interest) layer (Dai et al. 2015) is then applied, which divides each box into the same number of sections[8]. The largest value in each section is extracted to generate new cropped feature maps associated to each box from the previous feature map. ROI-pooling speeds up computation later on in the Faster-RCNN algorithm, as having fixed sized feature maps leads to faster convergence (Girshick 2015).

2.1.3 Detection Network

The DN is found after the ROI-pooling layer at the end of the Faster-RCNN algorithm. The DN is composed of FC layers, see Figure 1. The purpose of an FC layer is to combine all the outputs from the previous layer, this allows for the algorithm to make decisions. The FC layers are run in parallel, such that the weights and biases are split between classification and localisation. One of these two FC layers consists of 2 neurons to categorise the outputs for classification, the other FC layer consists of 4 neurons to predict the properties for box regression. Similar to the procedure for RPN, 16 positive and 48 negative labelled boxes are randomly chosen in each image to train the weights and biases in the DN, where boxes that overlap by more than 50 per cent with the ground truth box are assigned as positive 'foreground' labels or otherwise negative 'background' labels. Additional negative boxes with the next highest overlap are assigned as positive labels if there are fewer than 16 positive labelled boxes.

[8] Tensorflow's '*crop_and_resize*' operation is used to replicate the ROI-pooling layer in Ren et al. (2015) (Huang et al. 2016).

At the end of the DN, another categorical cross-entropy loss and smooth-L1 loss layer is used to calculate the classification and localisation errors, where each loss function is also associated to its own FC layer in the DN. NMS is applied again using a 60 per cent threshold to reduce the number of overlapping boxes until fewer than 100 boxes remain. The weights and biases in the FC layers are also trained via BP and SGD[7]. Classification loss is measured by comparing the cropped feature maps of each box with the ground truth box. We determine box regression loss by calculating the difference between the pixel coordinates, height and width of the positive labelled boxes with the ground truth box.

Finally, the loss functions of the RPN and the DN are combined into one multi-tasking loss function (Ren et al. 2015 and Huang et al. 2016) to train the algorithm. It takes into account the classification loss and box regression loss, by comparing all of the predicted properties for the boxes with the ground truth box in each image. Therefore the total loss of the algorithm is the weighted sum of the RPN and DN losses. The multi-tasking loss function is described in Equation 3:

$$\begin{aligned} L(\{p_i\},\{t_i\}) = \alpha \frac{1}{N}\sum_i L_{cls}(p_i, p_i^*) \\ +\beta \frac{1}{N}\sum_i p_i^* L_{reg}(t_i, t_i^*), \end{aligned} \tag{3}$$

where i is the box index number in each mini-batch, p_i is the predicted probability of a labelled box, t_i represents the height, width, center x and y coordinates of a labelled box, α and β are balancing weights for the classification and regression terms, p_i^* signifies positive/negative box labels, t_i^* is the ground truth box properties and N is the number of anchors/proposals selected to compute the loss function per mini-batch. L_{cls} is classification loss and $p_i^* L_{reg}$ is box regression loss for only positive labelled boxes.

2.2 Catalogue and Image Pre-Processing

Wen et al. (2012) applied the Friends-of-Friends cluster detection algorithm (Huchra & Geller 1982) on photometric data taken from the Sloan Digital Sky Survey III (SDSS-III, Eisenstein et al. 2011) Data Release 8 (DR8, Aihara et al. 2011). Wen et al. (2012) identified 132,684 galaxy clusters in the redshift range $0.05 \le z < 0.8$. The resultant catalogue has a completeness ratio of > 95 per cent for detecting galaxy clusters with mass greater than 1.0×10^{14} $M_\odot$ inside R_{200}[9] and in the redshift range of $0.05 \le z < 0.42$. We use the Abell galaxy clusters identified in the Wen et al. (2012) catalogue, to obtain the labelled data needed to create the training set. We choose the Abell galaxy clusters because our technique uses visual inspection of images in a similar manner to that performed by George Abell and is therefore appropriate to this proof of concept paper.

We do not train the algorithm on the entire Wen et al. (2012) catalogue, as this is a pilot study to test the performance of the Faster-RCNN algorithm at identifying galaxy clusters from a sample set. We limit the photometric redshift range of galaxy clusters to $0.1 < z < 0.2$, as we want to maximise the signal-to-noise available and avoid nearby galaxy clusters that could fill the field of view. See Figure 2 for the photometric redshift distribution in the training set. We set a threshold of 20 $\ge$ galaxy members inside a R_{200} radius, as poorly populated galaxy clusters may have a lower signal-to-noise. Applying these constraints results in a sample set of 497 Abell galaxy clusters. Figure 2 shows the richness distribution of the galaxy clusters in the training set within R_{200}. Richness is defined by Equation 4:

$$R_{L^*} = \frac{L_{200}}{L^*}, \tag{4}$$

where R_{L^*} is the galaxy cluster richness, L_{200} is the total r-band luminosity within R_{200} and L^* is the typical luminosity of galaxies in the r-band (Wen et al. 2012).

The brightest cluster galaxy (BCG) is a giant elliptical galaxy that is usually located in the vicinity of the spatial and kinematic centre of a galaxy cluster (Stott et al. 2008). We convert the right ascension (RA) and declination (Dec) of the BCG in each image to pixel coordinates. We adopt these pixel coordinates as the centre coordinates for the ground truth boxes in both training and test sets. Figure 2 shows the distribution of the r-band magnitudes for the BCGs in the training set. We also restrict Dec to greater than 0 degrees to reduce the amount of data near the galactic plane, due to a higher concentration of stars that may introduce significant foreground contamination.

The imaging camera on the SDSS telescope has a pixel size scaling of 0.396 arcsec pixel^{-1}. The SDSS telescope has five broadband imaging filters, referred to as u, g, r, i, z covering a wavelength range of 3543 to 9134 Å (Eisenstein et al. 2011). We used the i, r, g filters but not the u and z filters as the sensitivity of the SDSS telescope is poorer at these wavelengths. Each image in the training set contains one Abell galaxy cluster labelled as a ground truth. We fix each image size to 2000×2000 pixels (approximately 1443×1443 kpc at redshift $z = 0.1$ and 2588×2588 kpc at redshift $z = 0.2$) to capture the wider context but we lower the resolution of the images during training to a fixed dimension of 1000×1000 pixels for computational efficiency. We apply a random offset from a uniform distribution to the input coordinates since we do not want the algorithm to be biased towards specific positions in an image. This offset results in a uniform spread of the positions of the galaxy clusters in all images, where a galaxy cluster could be found anywhere within 270 arcseconds from the x and y plane of the image centre. See Figure 3 for the distribution of the positions in the training and test sets.

To make colour images we ensure that the individual images taken from the publically available SDSS-III Data Release 9 (DR9, Ahn et al. 2012)[10] for the i, r, g filters are set to the same scaling and aspect ratios. We stack the three filter images to RGB channels and apply a non-linear transformation to 'stretch' each image channel. We experimented with linear, square root and logarithmic transfor-

[9] R_{200} is the radii at which the mean density of the galaxy cluster is 200 times greater than the critical density of the Universe (Carlberg et al. 1997).

[10] The imaging data for SDSS-III DR9 can be found via NASA's SkyView (`http://skyview.gsfc.nasa.gov`) online database (McGlynn et al. 1998).

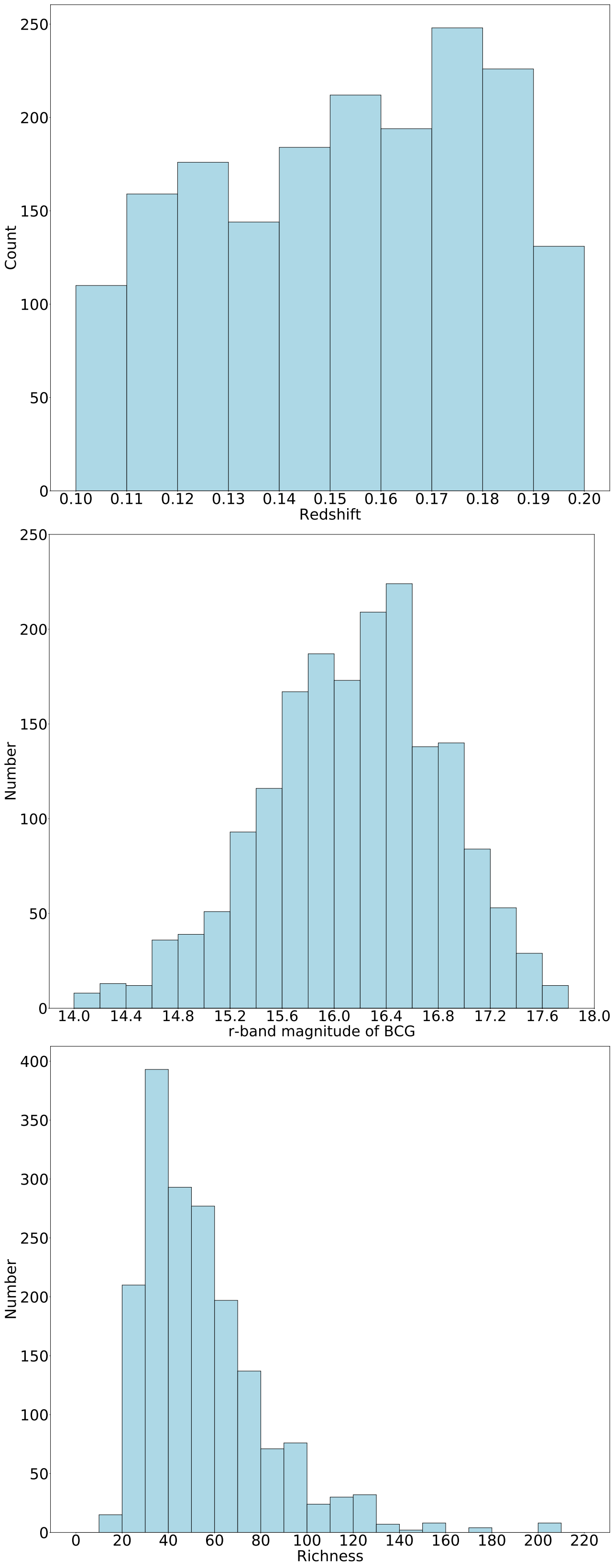

Figure 2. The distributions of the properties for all of the galaxy clusters in the training set. The histograms display the photometric redshift, r-band magnitude of the BCG and richness (from top to bottom).

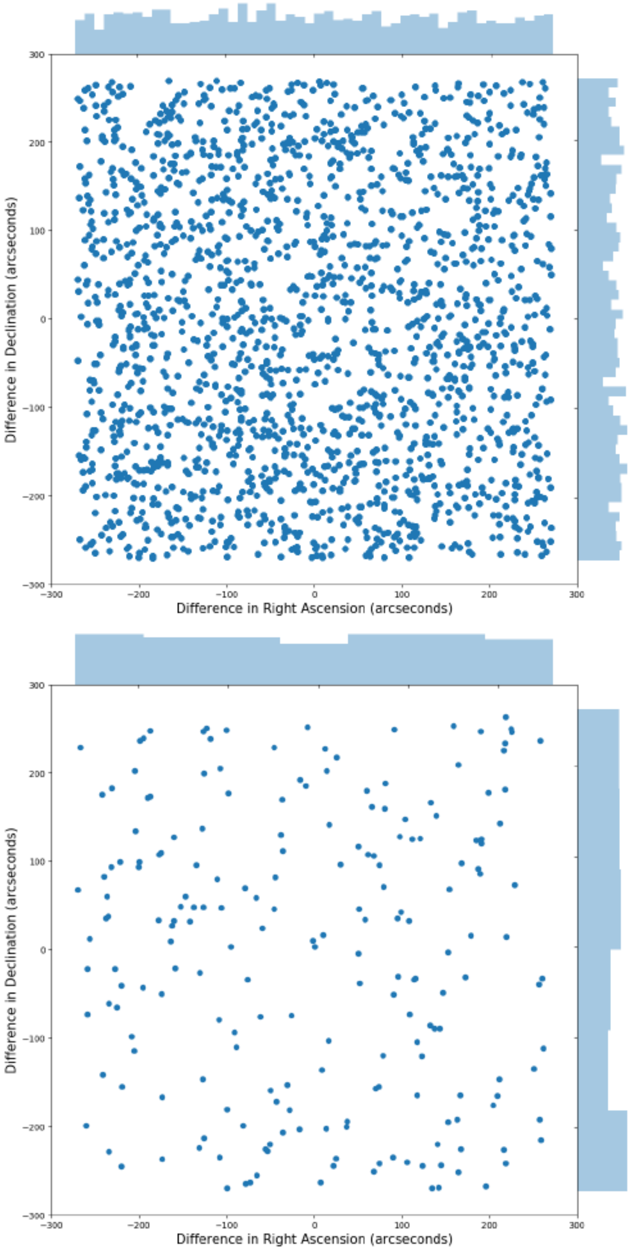

Figure 3. The distribution of image positions for the galaxy clusters in the training set (top) and test set (bottom). The positions are determined by calculating the difference in arcseconds between the coordinate offset and the true coordinates of the galaxy cluster at its respective photometric redshift, where the coordinate offsets are sampled from a uniform distribution.

mations. These transformations adjust the contrast of the image, as we define lower and upper flux limits by mapping the image onto a luminosity scale. We do this to reduce background noise, dim extremely bright objects and make stars and galaxies easily distinguishable. We find the square root function to be better for visualizing galaxy structure and colour. This benefits the algorithm by decreasing the learning complexity of the features.

We increase the amount of variance in the sample set by applying image augmentation techniques. In Ren et al.

(2015), it is stated that one of the properties of the Faster-RCNN algorithm is translational invariance, which means the algorithm is robust at learning translated objects. We train the algorithm to recognise that an object could appear at any location in an image. Since our method applies a random offset to the input coordinates via translation we augment the sample set three additional times, which boosts the sample size to 1988. This also introduces additional negative boxes for the algorithm to learn, as each image contains a different background. We randomly shuffle the sample set and perform simple random sampling to split the sample set into a training and test set, which are approximated representations of the full population. The training set is made up of ∼90 per cent of the sample set consisting of 1784 labelled galaxy clusters and the test set is made up of the remaining ∼10 per cent consisting of 204 labelled galaxy clusters. Figure 4 displays the astronomical coordinates for the galaxy clusters in the training and test sets compared to all the galaxy clusters in the Wen et al. (2012) catalogue. We also horizontally flip the images in the training set, where each image has a 50 per cent chance of being flipped. This approach can double the size of the training set to 3568 if all images are flipped once but this does not affect the size of the test set. Since galaxy clusters can be observed from any orientation we find these augmentation techniques to be appropriate during training.

3 RESULTS

3.1 Model Analysis with Test Set

We train our model with graphics processing unit (GPU) support for a maximum of 25,000 steps to ensure the algorithm has enough training time to sufficiently minimise prediction errors. The number of steps is a tunable hyperparameter that can shorten or extend the run-time of training a model. In Figure 5, we find that the algorithm generalises well as the total loss stabilises at approximately 3,000 steps, where a step represents one iteration per mini-batch from the dataset through the algorithm. For a competent model, the total loss should not fluctuate significantly during training.

We monitor the performance of the RPN and the DN via their respective loss functions, where we measure the objectness and box regression loss in the RPN and the classification and box regression loss in the DN. Objectness loss measures whether a box is likely to contain a ground truth object, box regression loss measures the exactness of the dimensions between the positive labelled boxes and ground truth box, and classification loss compares the resemblance of the features in a box with the features of the ground truth box. A lower loss value means that the prediction is almost identical to the ground truth. In Figure 6, we find that each of the losses also eventually stabilise at approximately 3,000 steps. Figure 6 suggests that the RPN is better generalised than the DN as there are fewer fluctuations. This could be explained by two possible reasons: either the training set is suited for localising objects in an image but improvements could be made to enhance the feature classification of objects in the training set, or we need to allow our model to train for more steps.

The 204 Abell galaxy clusters in the test set appear in the training set at least once but since the positions of the galaxy clusters and their surrounding image environments are different, we can assume the images to be unique. This is a useful test of the localisation performance of our model.

To evaluate our model we use common metrics such as precision, recall and F1-score (Goutte & Gaussier 2005). We do not account for true negatives, as there may be many other astronomical objects aside from galaxy clusters to consider. We expect that finding a galaxy cluster would be relatively rare as stars, galaxies and image artifacts populate the field-of-view. Our approach only searches for clusters rather than attempting to classify all objects. The final output of our model is a ‘confidence’ score generated for every predicted box, where a high confidence score means a high probability of an object being a ‘real’ galaxy cluster. We want to determine a threshold for the confidence score that returns high precision and high recall ratios. We re-run our model on the test set using different confidence score thresholds to examine the number of true positives (TP), false positives (FP) and false negatives (FN) returned.

We define a distance threshold by calculating a linear distance between the predicted and ground truth centre coordinates, where the predicted cluster centre is assumed to be at the same redshift as the ground truth cluster centre. We only apply this distance threshold during the model analysis, as we want to distinguish whether a predicted object is considered as a TP or FP detection to assess classification and localisation.

For our model, TP refers to the number of predicted boxes that score greater than the confidence score threshold and has a predicted centre within the distance threshold of the ground truth centre. FP refers to the number of predicted boxes that score greater than the confidence score threshold but does not have a predicted centre within the distance threshold of the ground truth centre. FN refers to the number of ground truth galaxy clusters without a positive prediction above the confidence score threshold within the distance threshold.

We calculate the precision and recall ratios using the number of TP, FP and FN at each confidence score threshold. Precision (also known as purity) is a ratio that effectively determines the number of ground truth objects returned by our model compared with the number of new predictions, see Equation 5:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{5}$$

Recall (also known as completeness) determines the number of ground truth objects returned by our model but compared with the number of ground truth objects that our model failed to predict, see Equation 6:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{6}$$

Precision-Recall (PR) curves are used as visual representations to examine the performance of a model, especially when a class population imbalance exists in the dataset (Davis & Goadrich 2006). Each point on the PR curve refers to the precision and recall ratio at a specific cut-off threshold. We explore eleven cut-off thresholds for

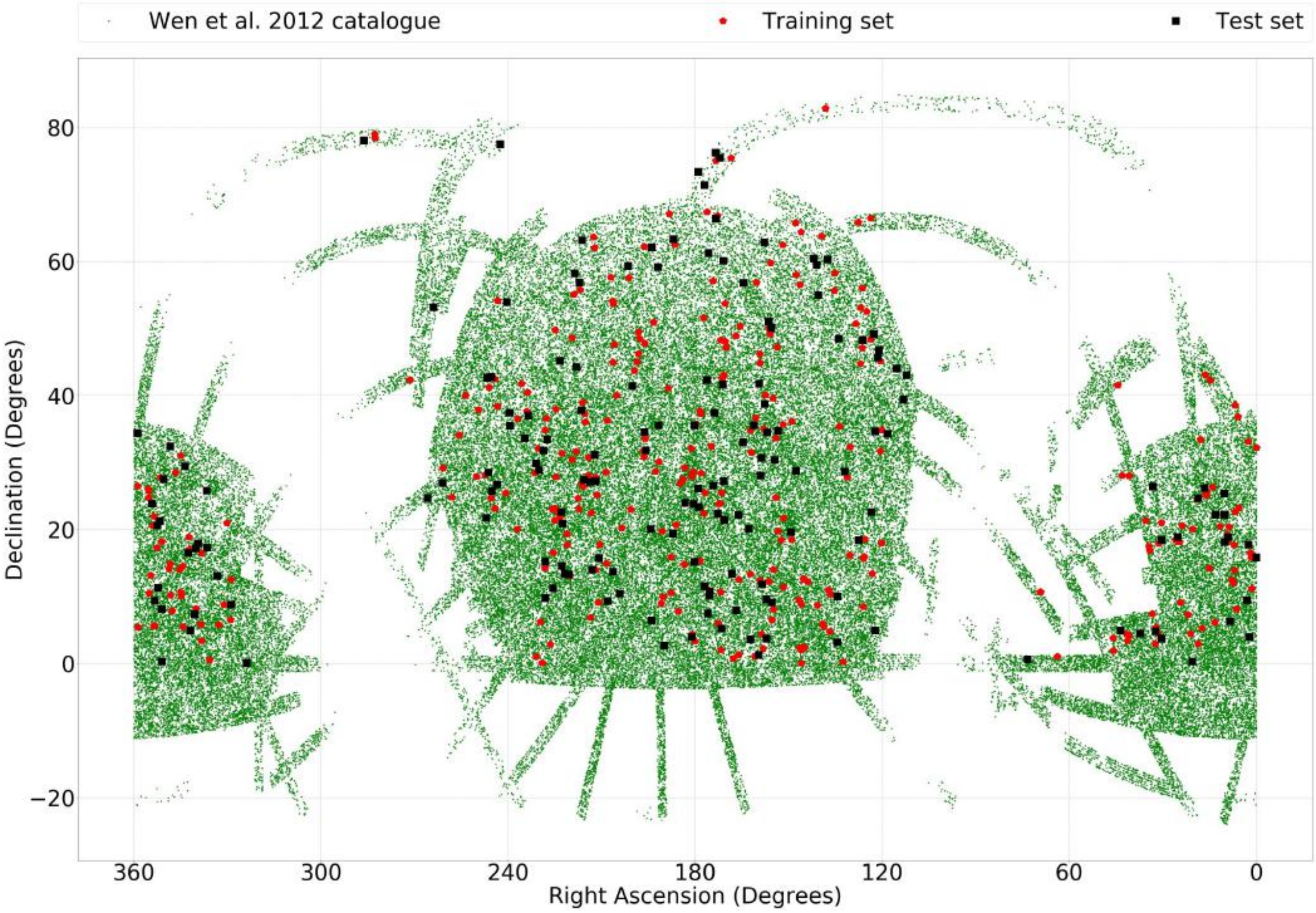

Figure 4. A map of astronomical coordinates using the J2000 epoch system of the galaxy clusters in the training set, test set and Wen et al. (2012) catalogue.

confidence scores ranging from 0 to 100 per cent and calculate the corresponding F1-score at each confidence score threshold.

F1-score is the harmonic mean between the precision and recall ratios at each confidence score threshold (Chase Lipton et al. 2014). We want to maximise the F1-score for our model to find the optimal balance between precision and recall. F1-score is described in Equation 7:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{7}$$

We analyse three galaxy clusters (a), (b) and (c) in the test set that have contrasting predicted confidence scores. These galaxy clusters can be seen in Figure 7. From Table 1, we find that (c) has the lowest confidence score, whilst (a) has the highest. This is because even though (c) has a high richness value, its higher redshift means it has fainter galaxies. Additionally (b) has a lower richness than (c) but is estimated to have a higher confidence score, again because it is at lower redshift. This demonstrates that galaxy clusters at lower redshift with brighter galaxies will receive higher confidence scores than their fainter counterparts at higher redshift. However, a galaxy cluster would also need high richness to achieve a very high confidence score. As a demonstration of our model, we choose a confidence score threshold of 80 per cent for the remainder of this paper.

We investigate how the environment (actual or contaminants) surrounding a galaxy cluster in an image can affect the detections generated by our model. We visually inspect multiple high-scoring candidate galaxy clusters from four different images in the test set. In Figure 8, we examine a candidate galaxy cluster that lies just outside the distance threshold at 88 kpc from the ground truth centre. We observe that multiple possible candidate galaxies could be classified as the BCG of the galaxy cluster. Since we train our model to predict a BCG as the galaxy cluster centre we would expect one of the galaxies in Figure 8 to be chosen. However, we find that our model is unable to definitively determine the ground truth cluster centre if there are multiple BCG-like galaxies close together in an image, such as in the event of an on-going cluster merger. Instead it finds an average centre, which is likely more appropriate for such systems. Wen & Han (2014) developed an approach to determine the dynamical state of a cluster based on photometric data to quantify a relaxation parameter ‘Γ’, which considers factors such as morphological asymmetry, ridge flatness and normalised deviation. They define the relaxation parameter as $\Gamma \geq 0$ representing as a relaxed state and $\Gamma < 0$ as a unrelaxed state, such that a larger positive/negative value

Table 1. The predicted confidence scores and properties of each galaxy cluster in Figure 7.

ID	Confidence Score (%)	Photometric Redshift	r-band Magnitude of the BCG	Richness
(a)	98	0.1474	14.99	78.49
(b)	51	0.1303	16.19	35.90
(c)	20	0.1875	16.73	58.44

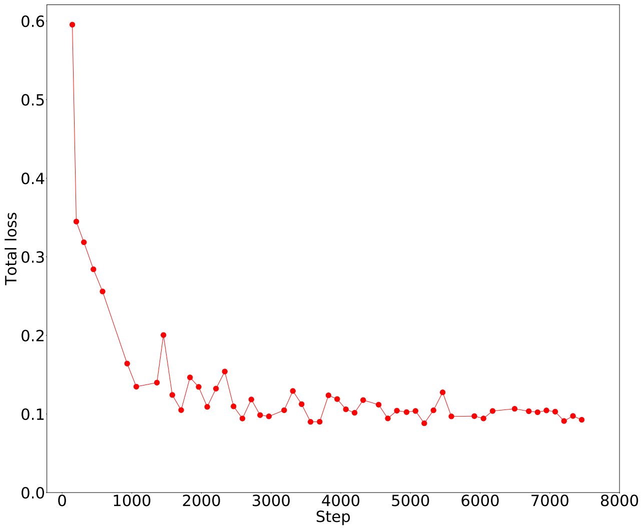

Figure 5. The total loss (see Equation 3) considers the errors from the RPN and the DN during training. We stop the the training of the model after evaluation at 7458 steps as the loss stabilises. Each point represents the total loss recorded at different step intervals. The values of these points can be found in Table A1.

implies a more relaxed/unrelaxed dynamical state. Wen & Han (2014) suggests that the cluster seen in Figure 8 is in a very unrelaxed dynamical state with a relaxation parameter value of $\Gamma = -1.43 \pm 0.08$, indicating a possible on-going merger.

In Figure 9, we examine another candidate galaxy cluster that lies 158 kpc from the ground truth centre. We use SDSS-III's Baryon Oscillation Spectroscopic Survey (BOSS, Eisenstein et al. 2011) to identify the spectroscopic redshift of the cluster. We identify the spectroscopic redshift of the predicted 'BCG' at z = 0.15765 ± 0.00003 and the spectroscopic redshift of the ground truth BCG at z = 0.19282 ± 0.00003. This means that while the galaxies are in the same line-of-sight they are not part of the same gravitationally-bound system. Our model is unable to determine the ground truth BCG since the predicted 'BCG' has stronger visual features and is at a lower redshift.

We find that our model identifies two BCG-like galaxies in Figure 10 as potentially two separate galaxy cluster centres but are within each others respective optical mean core radii defined in §2.2. One of the two candidate objects lies within the distance threshold of the ground truth centre whilst the other object is 220 kpc away. We find the spectroscopic redshift of the predicted 'BCG' is z = 0.15974 ± 0.00003 whilst the ground truth BCG has a spectroscopic redshift of z = 0.15989 ± 0.00002. Wen & Han (2014) suggests the cluster has a relaxation parameter value of $\Gamma = -0.61 \pm 0.09$. This value implies the cluster is in an unrelaxed dynamical state but to a lesser extent than the cluster seen in Figure 8, suggesting that this cluster is not experiencing an on-going merger but is likely in a pre-merger or post-merger state.

From Figure 11, we find that our model is able to detect a potential galaxy cluster which is extremely far from the ground truth centre at 1163 kpc distance, assuming it is at a similar redshift. We find that the predicted centre is located on top of a BCG-like object. We again use the BOSS survey to find the spectroscopic redshift of the predicted 'BCG' at z = 0.10608 ± 0.00002 and the ground truth BCG spectroscopic redshift of z = 0.14551 ± 0.00003. This shows that the two objects are clearly physically unrelated, therefore we identify the candidate object as a galaxy cluster. We cross-match the new galaxy cluster candidate as 'MSPM 08519' with heliocentric redshift of z = 0.1055 in the Smith et al. (2012) galaxy groups and clusters catalogue.

We decide to set an appropriate distance threshold value based on Figures 8 and 9 to be between 88 and 158 kpc. At these distances, we identify multiple galaxy clusters, which are considered far enough apart that cluster mergers and line-of-sight overlap clusters are distinguishable. Since we want to differentiate between cases of TP and FP in our model analysis, we choose a distance threshold of 100 kpc for the remainder of this paper.

In Figures 10 and 11, we observe secondary positive detections made by our model. We categorise these as FP detections, since the detections do not meet our set threshold criteria. However, we know that FP detections could still be actual galaxy clusters, such as is the case for Figure 11. This implies that FP detections consist of candidate galaxy clusters that need further verification or require cross-matching to existing galaxy cluster catalogues. We would label these detections as new galaxy cluster candidates. Whilst in Figure 9, we find that no positive detections are made on the ground truth galaxy cluster within the distance threshold. We categorise the ground truth galaxy cluster as a FN and the detection as a FP. This suggests that FN's consist of galaxy clusters that do not appear to have an obvious overdensity of galaxies and do not contain distinctive BCG-like galaxies. These visual features contribute to how distinguishable the galaxy cluster appears at the output of the final layer in the feature network.

From Figure 12, we observe that high precision diminishes recall and high recall diminishes precision. A low precision ratio results in a large number of candidate objects, whilst a low recall ratio means many real galaxy clusters are not predicted by our model. Table 2 shows that an 80 per cent confidence score threshold has the highest F1-score, which suggests that this confidence score threshold is the most effective at balancing precision and recall.

We analyse the distance between all of the predicted

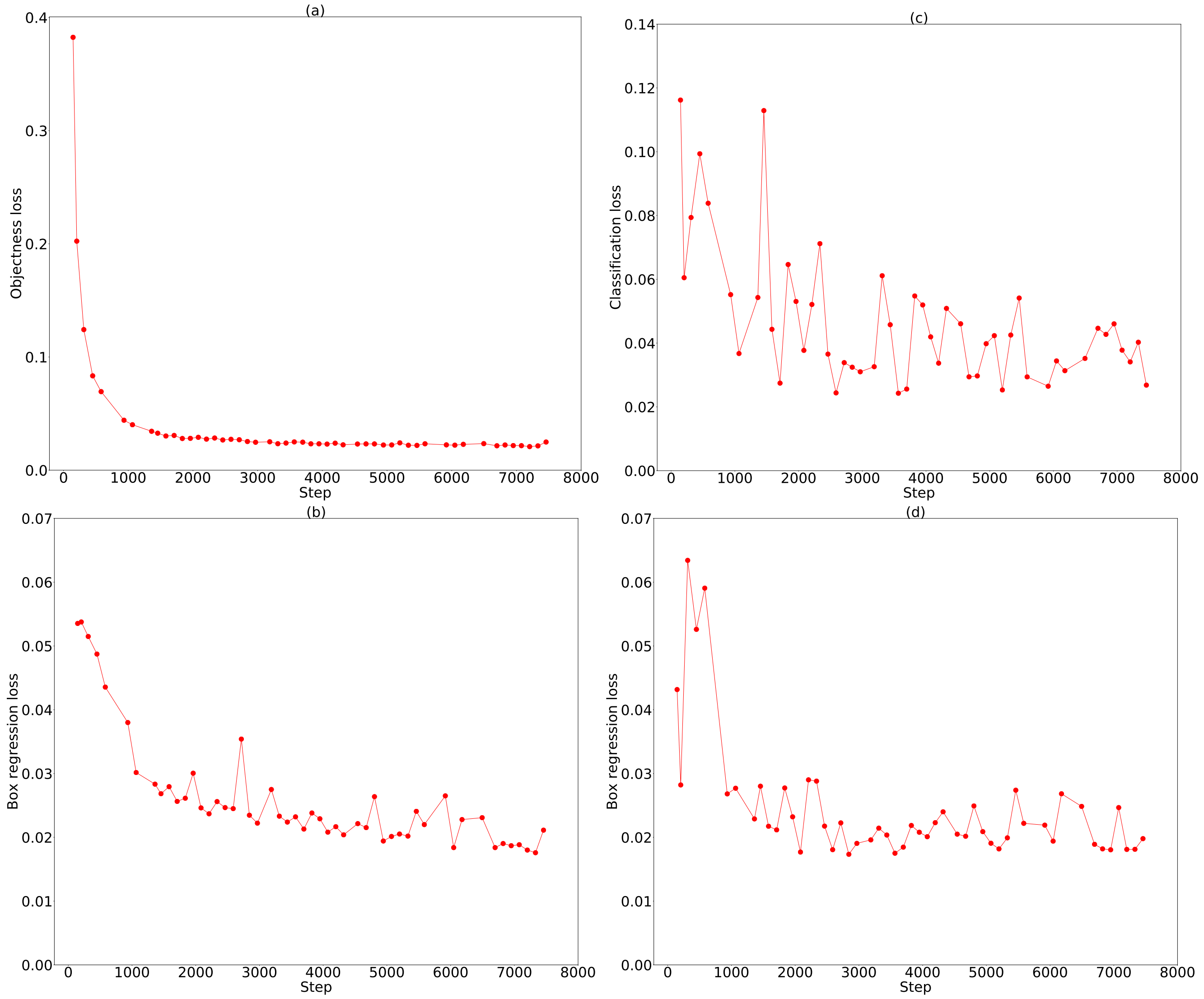

Figure 6. Training losses of the RPN and the DN are represented in (a), (b), (c) and (d). Where (a) displays the RPN objectness loss, (b) displays the RPN box regression loss, (c) displays the DN classification loss and (d) displays the DN box regression loss. The training of the model is stopped after evaluation at 7458 steps when the total loss no longer fluctuates, see Figure 5. The values for each point in (a),(b),(c) and (d) can be found in Table A1.

Table 2. The total number of true positives, false positives and false negatives returned by our model on the test set, where the precision, recall and F1-score ratios are calculated for each confidence score threshold.

Confidence score threshold (%)	# TP	# FP	# FN	Precision	Recall	F1-score
0	203	60997	1	0.003317	0.9951	0.006612
10	198	538	6	0.2690	0.9706	0.4213
20	197	391	7	0.3350	0.9657	0.4975
30	193	302	11	0.3899	0.9461	0.5522
40	191	243	13	0.4401	0.9363	0.5987
50	188	202	16	0.4821	0.9216	0.6330
60	181	163	23	0.5262	0.8873	0.6606
70	177	119	27	0.5980	0.8676	0.7080
80	165	72	39	0.6962	0.8088	0.7483
90	136	29	68	0.8242	0.6667	0.7371
100	0	0	204	0.00	0.00	0.00

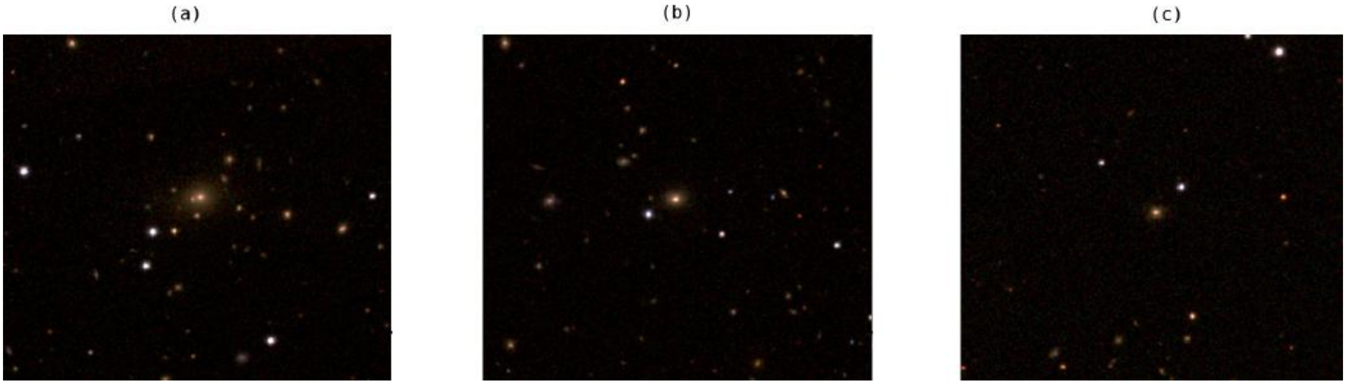

Figure 7. Colour images (a), (b) and (c) contain three different Abell galaxy clusters from the test set. The J2000 coordinates for each galaxy cluster are as follows. (a) RA: 222.78917 and Dec: 14.61203, (b) RA: 180.19902 and Dec: 35.58229 and (c) RA: 137.49464 and Dec: 60.32841. The predicted confidence scores and properties for the galaxy clusters in (a), (b) and (c) can be found in Table 1.

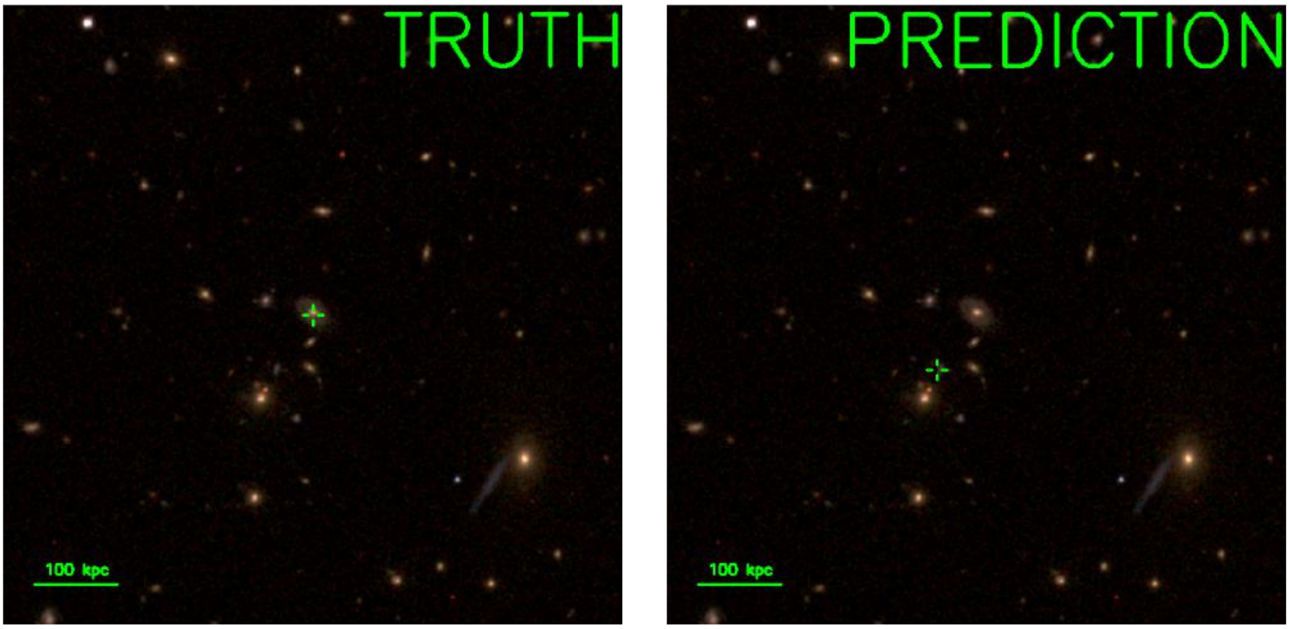

Figure 8. The linear distance between the ground truth and predicted centre coordinate is 88 kpc in respect to the photometric redshift z = 0.1788 of the ground truth galaxy cluster. The J2000 coordinates of the ground truth galaxy cluster is RA: 191.85623 and Dec: 35.54509.

centre coordinates from the ground truth centre using an 80 per cent confidence score threshold. In Figure 13, we find that the distance threshold of 100 kpc contains 70 per cent of all of the predictions and returns 81 per cent of the total ground truth clusters in the test set. We disregard any detection further than 250 kpc from the ground truth centre from being considered as a TP prediction, since the prediction would lie outside the optical mean core radii stated in §2.2. We calculate the standard error of coordinate regression of our model to determine the average distance of the predicted centre coordinate from the ground truth centre coordinate using Equation 8:

$$\sigma_{estimate} = \sqrt{\frac{\sum (Y - Y')^2}{N}}, \tag{8}$$

where $\sigma_{estimate}$ is the standard error of regression, Y is the ground truth value, Y' is the predicted value and N is the sample size (L. McHugh 2008). We obtain a standard error of 17.42 kpc for predictions considered as TP's in the test set using an 80 per cent confidence score threshold. Therefore we can estimate a 95 per cent confidence interval for all predicted centre coordinates to be approximately within ±1.96 × standard error of a ground truth centre coordinate (Altman & Bland 2005).

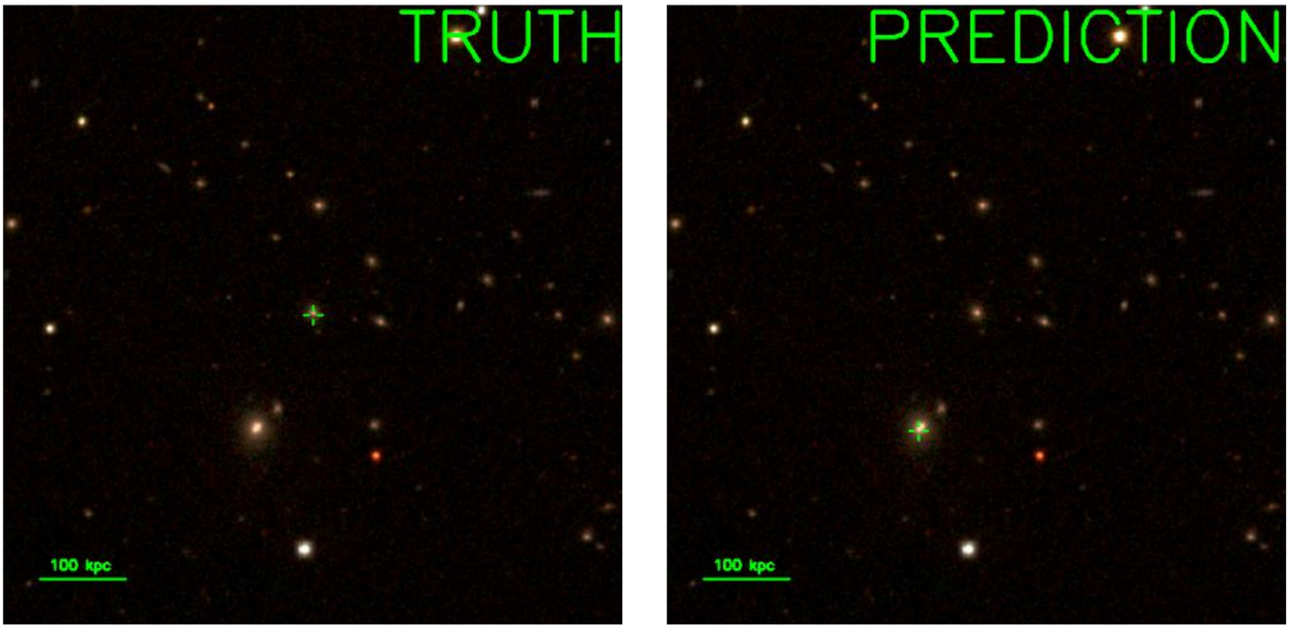

Figure 9. The linear distance between the ground truth and predicted centre coordinate is 158 kpc in respect to the photometric redshift z = 0.1618 of the ground truth galaxy cluster. The J2000 coordinates of the ground truth galaxy cluster is RA: 161.20657 and Dec: 35.54042.

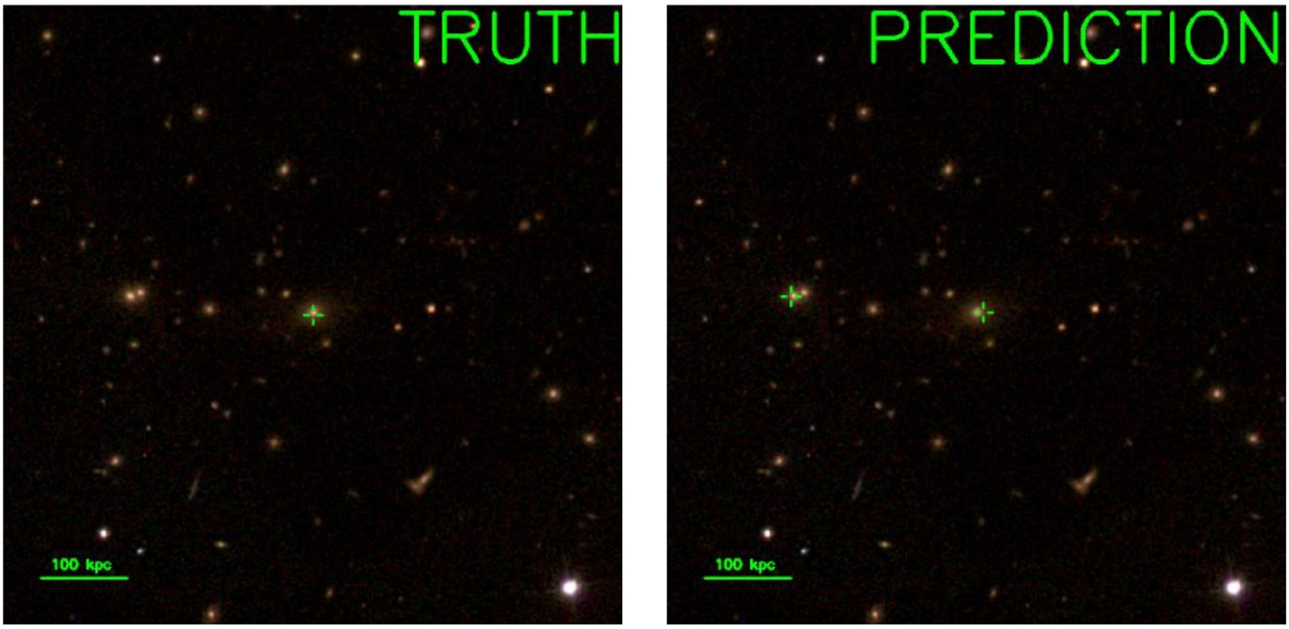

Figure 10. The linear distance between the ground truth and predicted centre coordinate (the one that does not overlap directly with the ground truth) is 220 kpc in respect to the photometric redshift z = 0.1603 of the ground truth galaxy cluster. The J2000 coordinates of the ground truth galaxy cluster is RA: 353.35867 and Dec: 9.42395.

In Figure 14, we visually examine the positions of returned ground truth galaxy clusters in the test set with their original locations as shown in Figure 3. We find that our model does not show bias towards any particular location in an image. This suggests that the random offset we apply in §2.2 is effective at reducing location bias during training.

We also compare the photometric redshift, BCG r-band magnitude and richness distributions of all the galaxy clusters returned by our model with their original distributions in the test set. Figure 15 shows that our model has no clear prediction bias towards any of these properties. We perform a two sample Kolmogorov-Smirnov (KS) test (Smirnov 1939) to test whether the original and returned distributions violate the null hypothesis. Since the KS test is non-parametric,

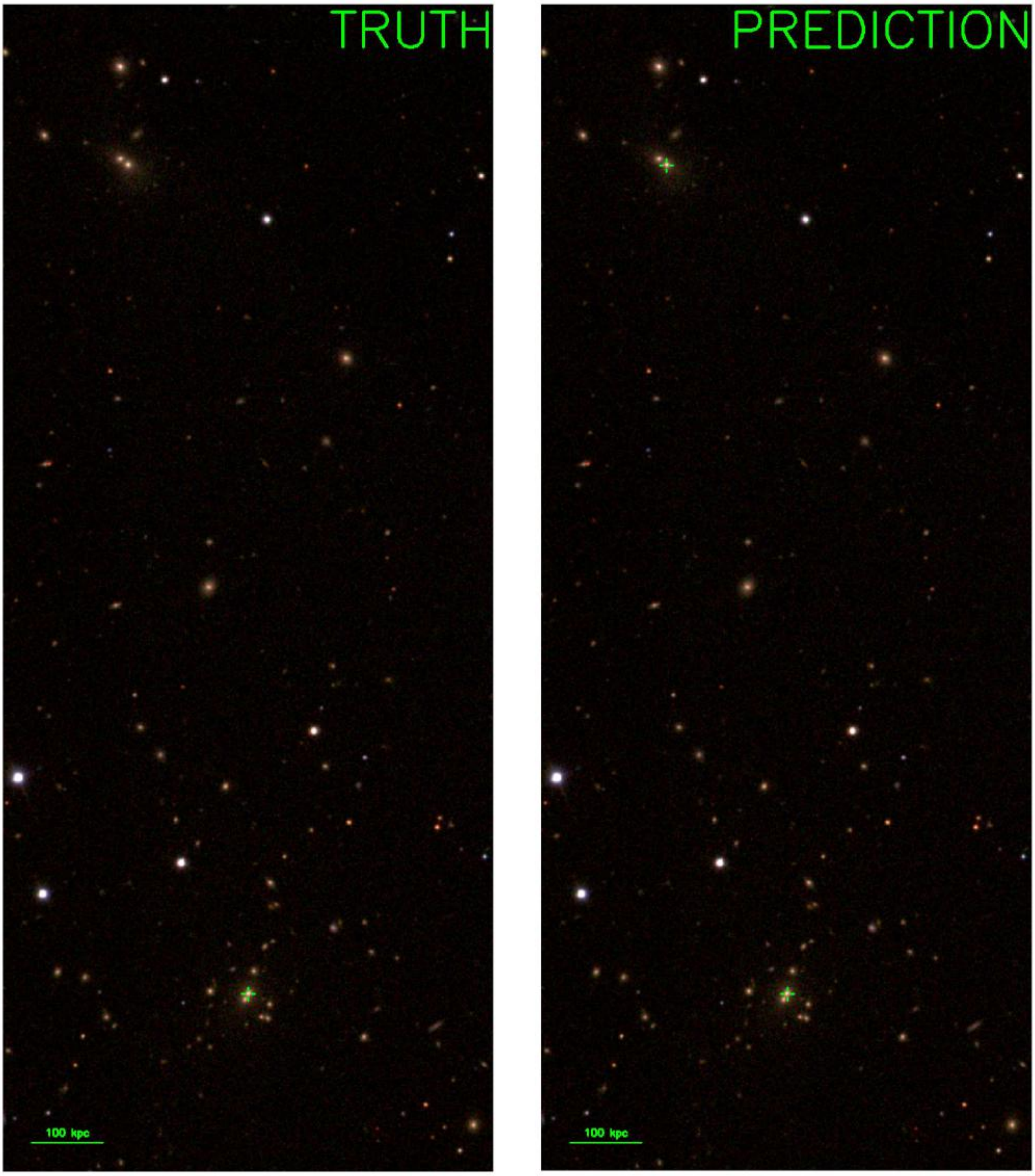

Figure 11. The linear distance between the ground truth and predicted centre coordinate (the one that does not overlap directly with the ground truth) is 1163 kpc in respect to the photometric redshift $z = 0.1368$ of the ground truth galaxy cluster. The J2000 coordinates of the ground truth galaxy cluster is RA: 186.96341 and Dec: 63.38483.

the distributions do not need to have normality. We calculate test statistic values of 0.06275, 0.07335 and 0.02193 for photometric redshift, BCG r-band magnitude and richness respectively. We set $\alpha = 0.05$ as the level of significance to obtain a critical value of 0.1281 (Gail & Green 1976). Since the test statistic values are smaller than the critical value at $\alpha = 0.05$, we cannot reject that the original and returned distributions are statistically the same.

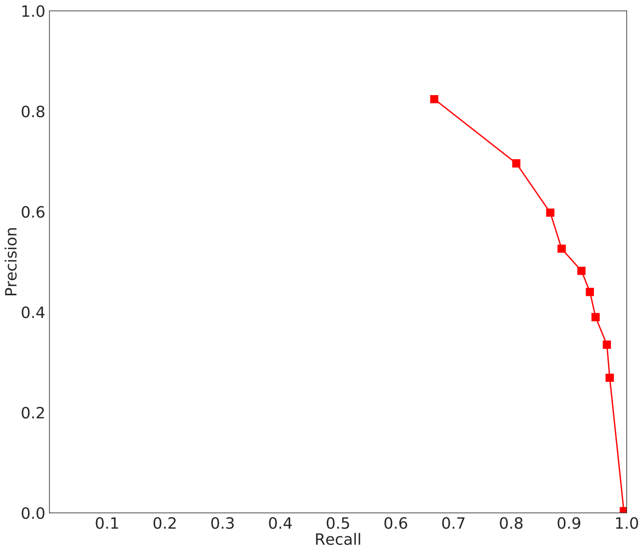

Figure 12. Precision vs Recall ratios from the test set, where each point represents the ratios at a confidence score threshold. The values of each point can be found in Table 2. We do not include the precision and recall ratio for the 100 per cent confidence score threshold, as it provides no conclusive evaluation of the performance of the model.

3.2 Comparison To redMapper galaxy clusters

The redMapper algorithm predicts galaxy clusters using the red sequence fitting technique and probabilistic percolation of galaxies based on their photometric redshift. Rykoff et al. (2014b) apply their algorithm to SDSS DR8 (Aihara et al. 2011), to create a catalogue of ~25, 000 candidate galaxy clusters in the photometric redshift range of $0.08 < z < 0.55$. We apply the same testing constraints used in §2.2 on the redMapper galaxy clusters, where galaxy clusters must be in the photometric redshift range of $0.1 < z < 0.2$. We do not need to apply a galaxy member count constraint as the redMapper algorithm by default only recognises galaxy clusters with greater than 20 member galaxies[11]. In Figure 16, we locate a 105 square degree region that contains 31 galaxy clusters identified by the redMapper algorithm. We found that this region did not contain any galaxy clusters from our training or test sets in the photometric redshift range of $0.1 < z < 0.2$ but it did contain clusters found via redMapper in the same redshift range. We use these redMapper clusters to create a new sample test set and examine the localisation and classification performance of our model on unseen clusters.

We adopt the same procedure from §2.2 to generate new redMapper test set images. We re-run our model on the redMapper test set and apply the evaluation metrics of precision, recall and F1-score again. In Figure 17, we observe a precision and recall trade-off similar to Figure 12, where precision increases with confidence score threshold whilst recall decreases. From Table 3, we find the confidence score with the highest F1-score is 70 per cent. This suggests that our model has not overfit since it performs better on an unseen dataset, where the confidence score threshold has lowered from 80 per cent in §3.1. A lower confidence score threshold increases the number of objects detected while still retaining high precision and recall. From using a 70 per cent confidence score threshold, we obtain a standard error of 12.33 kpc for predictions considered as true positives in the redMapper test set.

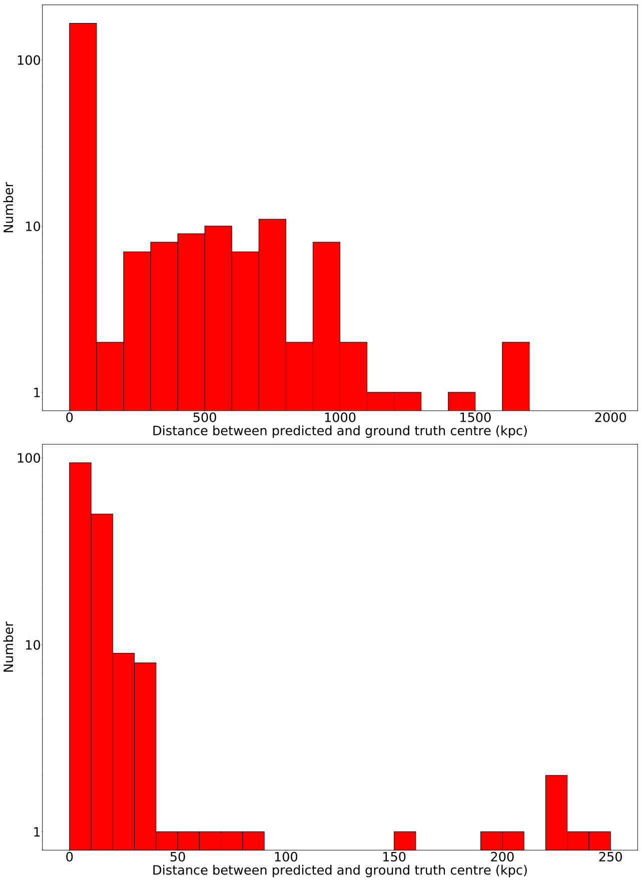

Figure 13. The distribution of the linear distance between the ground truth and predicted centre coordinates in test set images for all predictions (top) and predictions within the distance threshold (bottom) using an 80 per cent confidence score threshold.

We follow-up the FP detections made on the redMapper test set images by cross-matching with the redMapper and Wen et al. (2012) catalogues with relaxed redshift constraints. We use the 70 per cent confidence score threshold, as this threshold has the highest F1-score (see Table 3), from which we obtain eight new candidate galaxy clusters. We apply a redshift constraint of $0.05 < z < 0.4$ on the galaxy clusters in the redMapper and Wen et al. (2012) catalogues. We utilise a ~1.61 arcminute radius of the predicted RA and Dec coordinates, which corresponds with the optical mean core radii of 250 kpc for a galaxy cluster at redshift $z = 0.15$, to search through the catalogues. We find that none of the detections match with existing clusters in the redMapper catalogue. However, we identify three FP detections as existing galaxy clusters in the Wen et al. (2012) catalogue. This

[11] Note that Rykoff et al. (2014b) does not define galaxy members within R_{200} but from an optical radius cutoff that scales with the number of galaxies found via percolation.

Table 3. The total number of true positives, false positives and false negatives returned by our model on the redMapper test set where the precision, recall and F1-score ratios are calculated for each confidence score threshold.

Confidence score threshold (%)	# TP	# FP	# FN	Precision	Recall	F1-score
0	30	9270	1	0.003226	0.9677	0.006430
10	29	50	2	0.3671	0.9355	0.5273
20	29	29	2	0.50	0.9355	0.6517
30	29	23	2	0.5577	0.9355	0.6988
40	28	18	3	0.6087	0.9032	0.7273
50	28	14	3	0.6667	0.9032	0.7671
60	28	11	3	0.7179	0.9032	0.80
70	27	8	4	0.7714	0.8710	0.8182
80	23	4	8	0.8519	0.7419	0.7931
90	17	0	14	1.00	0.5484	0.7083
100	0	0	31	0.00	0.00	0.00

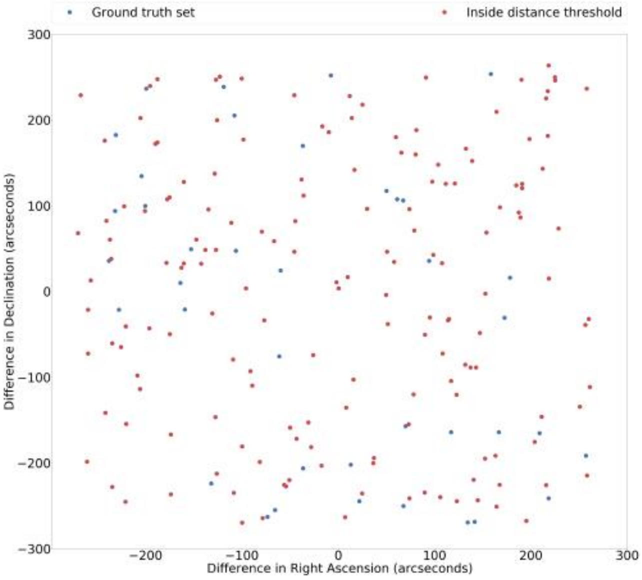

Figure 14. A comparison of the centre coordinate offset between the ground truth galaxy clusters returned by our model and the full list of offset values in the test set (see Figure 3) using an 80 per cent confidence score threshold.

shows that our model is generalised enough to detect galaxy clusters that have not been drawn from the same sample as the training set. We find five detections that do not match with any known galaxy cluster in the redMapper and Wen et al. (2012) catalogues but these may exist in other galaxy cluster catalogues.

4 DISCUSSION

4.1 Limitations of our model

Feature engineering is an important process for improving computational efficiency and the performance of a deep learning model. We apply constraints to the training set to reduce the complexity of the features. For example, we use Abell galaxy clusters which contain a minimum of 50 galaxies within a 1.5 h^{-1} Mpc radii of the galaxy cluster centre (Abell et al. 1989). This means that these galaxy clusters would have strong signal-to-noise and are very likely to be real gravitationally bound clusters. However, not all Abell galaxy clusters have been verified so one limitation of our approach is that our model is reliant on the catalogue from Wen et al. (2012) for training data. In Wen et al. (2012), Monte Carlo simulations are used to determine a false detection rate of less than 6 per cent for the entire catalogue. False detections lower the overall precision of the predictions because the training data could be contaminated with objects that should not be classified as galaxy clusters. Similarly, the test set will also suffer from this. Since we have a large training set, it is impractical to directly check for contaminants in every image with spectroscopic follow-up of all ground truth cluster members. We must therefore assume that all of the galaxy clusters in this catalogue are real. Additionally we must account for the errors in RA and Dec coordinates from Wen et al. (2012), as we use these coordinates for the ground truth centre coordinates.

As with all deep learning algorithms, there are hyper-parameters that require either minor or extensive fine tuning, depending on the task at hand. The main training hyper-parameters of Faster-RCNN include the learning rate, momentum, gradient clipping threshold, mini-batch size, number of layers, number of neurons in each layer and architecture. We have shown that the values set for these hyper-parameters are capable of being adapted to perform generalised object detection of galaxy clusters. However, to fully optimise the value of every hyper-parameter is computationally expensive, so we rely on the use of transfer learning for partial optimisation of the hyper-parameters in our deep learning model.

We adopt a specific methodology to generate all of the images by applying the same contrasting, image aspect and image scaling ratios for computational efficiency. However, this may create a trade-off between computational efficiency and bias from image pre-processing. This means that all future input images to our model are somewhat restricted to using the same pre-processing techniques that we apply in §2.2 to obtain maximum performance.

We perform hold-out validation on the sample set to form the training and test sets. However, this approach is limited to a simple approximation since we observe a population imbalance in the sample in Figure 2. For example, we find that there are fewer low redshift galaxy clusters compared with high redshift galaxy clusters. This means our

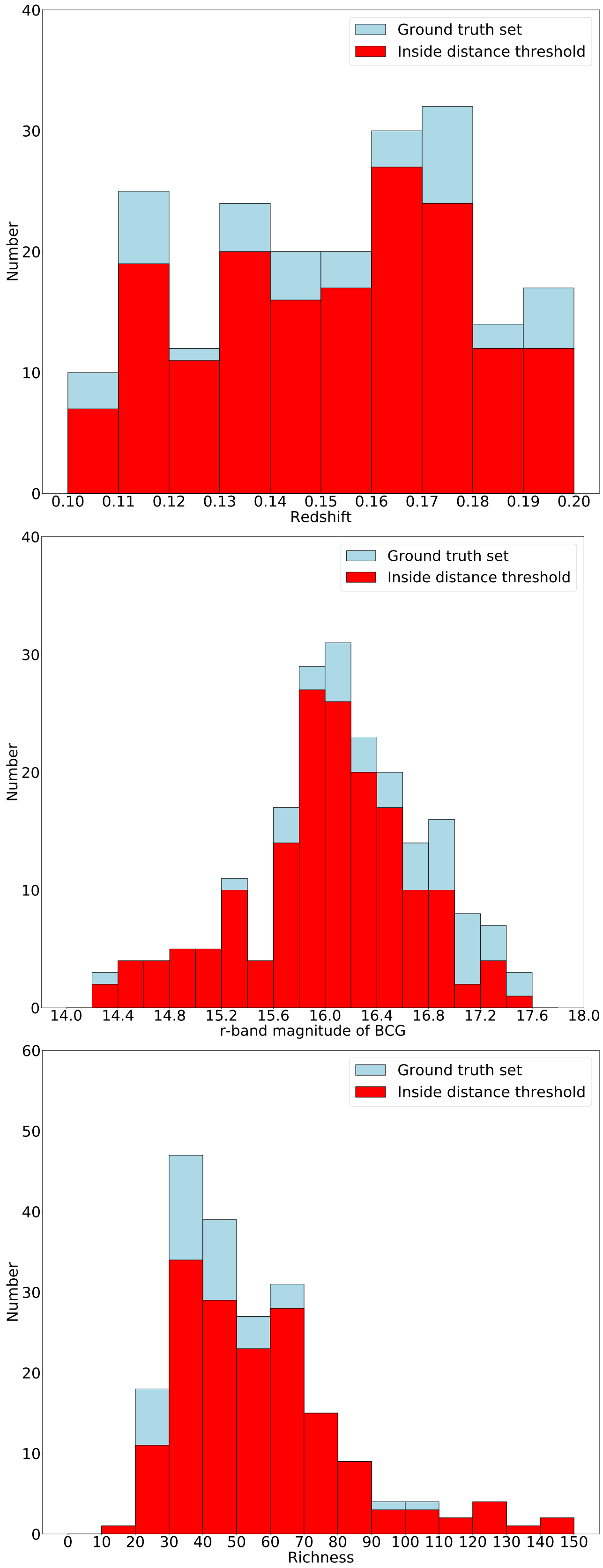

Figure 15. The distributions of the properties from the original and ground truth galaxy clusters returned by our model in the test set using an 80 per cent confidence score threshold. The histograms display the photometric redshift, r-band magnitude of the BCG and richness (from top to bottom).

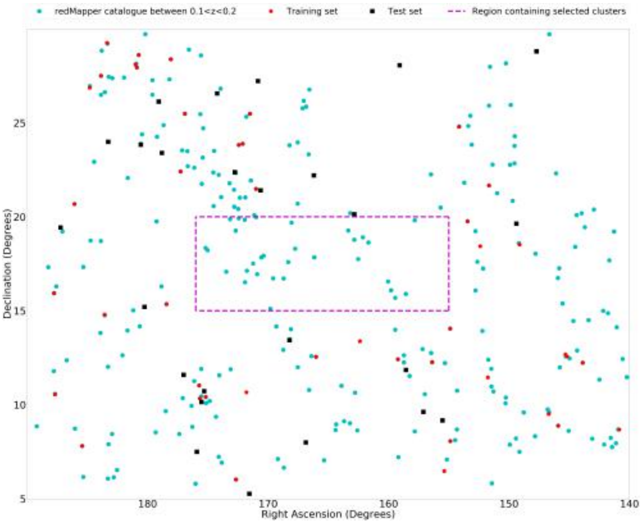
redMapper catalogue between 0.1<z<0.2
Training set
Test set
Region containing selected clusters
Declination (Degrees)
5
10
15
20
25
180
170
160
150
140
Right Ascension (Degrees)

Figure 16. A map of astronomical coordinates using the J2000 epoch system for the galaxy clusters in the training set, test set and constrained redMapper catalogue (Rykoff et al. 2014b). We highlight the region of galaxy clusters identified by the redMapper algorithm, which are not already part of the training set or test set.

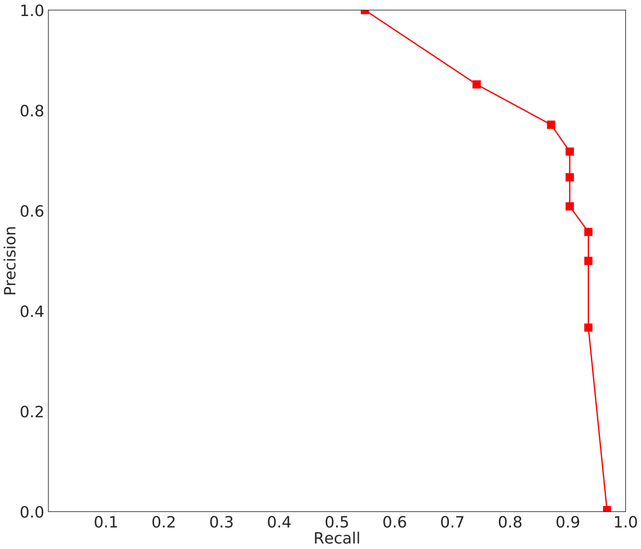

Figure 17. Precision vs Recall ratios on the redMapper test set, where each point represents the ratios at a confidence score threshold. The values of each point can be found in Table 3. Again we do not include the precision and recall ratio for the 100 per cent confidence score threshold, as it provides no conclusive evaluation of the performance of the model.

model could overfit to populations that appear more frequently. To reduce the chance of overfitting from population bias in the training set we could perform k-fold cross validation (Yadav & Shukla 2016) when splitting the sample set. K-fold validation splits the sample set into an arbitrary 'k' number of folds where one fold becomes the test set whilst the remaining folds are merged to form the training set. This is repeated until every fold has been used as the test set. Then all the folds are compared and the fold with the best performance is chosen to represent the training and test set.

4.2 Future Applications of this Technique

LSST and *Euclid* are ideal surveys to apply our deep learning model to, as they will be wider and deeper than any survey conducted before them. This will produce detections of many thousands of candidate high redshift or low mass clusters that are currently undiscovered. This may be an iterative process in practice with these large datasets also being used to improve the training of our model.

The Deep-CEE method will be of great use for confirming candidate galaxy clusters detected by X-ray or SZ surveys, as they often have many interlopers. We also intend to use training sets based on galaxy clusters selected from traditional techniques. For example, it will be interesting to compare the galaxy clusters predicted by a deep learning algorithm based on a training set of X-ray selected galaxy clusters compared with those trained on a red sequence fitting sample. This may be a good way to test the various biases of cluster detection methods, which can filter through to any cosmological predictions made with them.

Our deep learning model can also be adapted and applied to both optical imaging and other cluster detection methods at the same time. For example, a related algorithm could be shown a red sequence fit and/or an X-ray image at the same time as visual imaging to create a robust sample of galaxy clusters.

To improve the model itself, we will investigate whether applying additional image augmentation techniques such as rotation and vertical flipping can improve the performance of our model. Perez & Wang (2017) shows that using simple image transformations can result in a more robust model. We will also train the model to predict intrinsic properties of galaxy clusters such as redshift and richness, which will be vital for the thousands of galaxy clusters discovered by wide-field surveys such as LSST and *Euclid*. This would improve our criteria for categorising TP's.

To prepare for LSST and *Euclid* we aim to run our model on the entirety of SDSS and cross-match all FP detections with existing catalogues. This would be beneficial in constraining the resultant purity ratio of our model, as we expect many FP detections to be actual galaxy clusters.

5 CONCLUSION

We present Deep-CEE a novel deep learning model for detecting galaxy clusters in colour images and returning their respective RA and Dec. We use Abell galaxy clusters from the Wen et al. (2012) catalogue as the ground truth labels in colour images to create a training set and test set. We determine an optimal confidence score threshold based on the threshold value with the greatest F1-score. We initially find that a 80 per cent confidence score threshold is optimal for finding galaxy clusters in our original test set, as we achieve a precision ratio of 70 per cent and a recall ratio of 81 per cent[12]. We find that a 70 per cent confidence score threshold is the optimal threshold for detecting galaxy clusters in the redMapper test set, as we obtain a precision ratio of 77 per cent and a recall ratio of 87 per cent. We consider any detections with a predicted confidence score greater than the confidence score threshold as galaxy cluster candidates. It should be noted that our precision ratios are specific to the test sets, since we know that some false positive detections are actual galaxy clusters but we do not classify these as true positives during the analysis. We show that our model has not overfit to galaxy clusters in the training set, as we obtain a lower optimal confidence score threshold when we run the model on unseen galaxy clusters. This suggests that the features in the training set and the hyper-parameters used for our model are suitable for generalised object detection of galaxy clusters.

By applying Deep-CEE to wide-deep imaging surveys such as LSST and *Euclid*, we will discover many new higher redshift and lower mass galaxy clusters. Our approach will also be a powerful tool when combined with catalogues or imaging data from other wavelengths, such as X-ray (e.g. Sarazin 1986, Ebeling et al. 1998, Ebeling et al. 2000a, Ebeling et al. 2000b, Böhringer et al. 2004, Takey et al. 2011, Mehrtens et al. 2012 and Adami, C. et al. 2018) and SZ surveys (e.g. E. Carlstrom et al. 2009, Staniszewski et al. 2009, Vanderlinde et al. 2010, Planck Collaboration et al. 2014, Bleem et al. 2015, Planck Collaboration: et al. 2015, Planck Collaboration et al. 2016, Hilton et al. 2018 and Hurier et al. 2017). It is hoped that the future cluster samples produced by Deep-CEE alone or in combination with other selection techniques will be well-understood and therefore applicable to constraining cosmology, as well as environmental galaxy evolution research. We will build upon this model by including methods to estimate intrinsic properties of galaxy clusters such as redshift and richness in a similar manner to George Abell many years ago.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referee for their thorough feedback which has improved the clarity of our paper.

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 web site is `http://www.sdss3.org/` (Eisenstein et al. 2011).

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,

[12] An ideal confidence score threshold would have a precision and recall ratio of 100 per cent. (Tharwat 2018)

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.

We acknowledge studentship funding by the Science and Technology Facilities Council (STFC), travel support provided by STFC for UK participation in LSST: All Hands Meeting through grant ST/N002512/1 and NASAâĂŹs SkyView facility located at NASA Goddard Space Flight Center.

We also would like to thank Min-Su Shu at the University of Michigan, James Schombert at the University of Oregon, Edward L. Wright at the University of California, Los Angeles (Wright 2006) and Dat Tran (https://github.com/datitran) for allowing free distribution of their code.

REFERENCES

Abadi M., et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, http://tensorflow.org/

Abell G. O., Corwin Jr. H. G., Olowin R. P., 1989, ApJS, 70, 1

Adami, C. et al., 2018, A&A, 620, A5

Ahn C. P., et al., 2012, ApJS, 203, 21

Aihara H., et al., 2011, ApJS, 193, 29

Altman D. G., Bland J. M., 2005, BMJ, 331, 903

Amiaux J., et al., 2012, in Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave. p. 84420Z (arXiv:1209.2228), doi:10.1117/12.926513

Armitage T. J., Kay S. T., Barnes D. J., 2019, Monthly Notices of the Royal Astronomical Society, 484, 1526

Bleem L. E., et al., 2015, ApJS, 216, 27

Böhringer H., et al., 2004, A&A, 425, 367

Bottou L., 2010, in Lechevallier Y., Saporta G., eds, Proceedings of COMPSTAT'2010. Physica-Verlag HD, Heidelberg, pp 177–186

Carlberg R. G., Yee H. K. C., Ellingson E., 1997, ApJ, 478, 462

Chase Lipton Z., Elkan C., Narayanaswamy B., 2014, preprint (arXiv:1402.1892)

Chong K., Yang A., 2019, EPJ Web of Conf., Vol. 206, EDP Sciences

Cudnik B., 2013, Faint Objects and How to Observe Them. Astronomers' Observing Guides, Springer, New York, doi:10.1007/978-1-4419-6757-2, http://cds.cern.ch/record/1493245

Dai J., He K., Sun J., 2015, preprint (arXiv:1512.04412)

Davies A., Serjeant S., Bromley J. M., 2019, Monthly Notices of the Royal Astronomical Society, 487, 5263

Davis J., Goadrich M., 2006, in Proceedings of the 23rd International Conference on Machine Learning. ICML '06. ACM, New York, NY, USA, pp 233–240, doi:10.1145/1143844.1143874, http://doi.acm.org/10.1145/1143844.1143874

E. Carlstrom J., et al., 2009, Publications of The Astronomical Society of The Pacific - PUBL ASTRON SOC PAC, 123, 568

Ebeling H., Edge A. C., Bohringer H., Allen S. W., Crawford C. S., Fabian A. C., Voges W., Huchra J. P., 1998, MNRAS, 301, 881

Ebeling H., Mullis C. R., Tully R. B., 2000a, in Kraan-Korteweg R. C., Henning P. A., Andernach H., eds, Astronomical Society of the Pacific Conference Series Vol. 218, Mapping the Hidden Universe: The Universe behind the Mily Way - The Universe in HI. p. 79

Ebeling H., Edge A., P. Henry J., 2000b, The Astrophysical Journal, 553, 668

Eisenstein D. J., et al., 2011, AJ, 142, 72

Flamary R., 2016, preprint (arXiv:1612.04526)

Gail M. H., Green S. B., 1976, Journal of the American Statistical Association, 71, 757

Girshick R., 2015, preprint (arXiv:1504.08083)

Gladders M. D., Yee H. K. C., 2000, AJ, 120, 2148

Gladders M. D., Yee H. K. C., 2005, ApJS, 157, 1

Glorot X., Bengio Y., 2010, Journal of Machine Learning Research - Proceedings Track, 9, 249

Goutte C., Gaussier E., 2005, in Proceedings of the 27th European Con- ference on Advances in Information Retrieval Research. ECIRâĂŹ05, Vol. 3408, Springer-Verlag, Berlin, Heidelberg, p. 345 doi:10.1007/978-3-540-31865-1_25, http://dx.doi.org/10.1007/978-3-540-31865-1_25

Hála P., 2014, preprint (arXiv:1412.8341)

Hezaveh Y. D., Perreault Levasseur L., Marshall P. J., 2017, Nature, 548, 555

Hilton M., et al., 2018, ApJS, 235, 20

Ho M., Rau M. M., Ntampaka M., Farahi A., Trac H., Poczos B., 2019, preprint (arXiv:1902.05950)

Hosang J., Benenson R., Schiele B., 2017, preprint (arXiv:1705.02950)

Hoyle B., 2016, Astronomy and Computing, 16, 34

Huang T., 1996, International conference; 5th, High technology: Imaging science and technology, Chiba, Japan, p. 13, https://doi.org/10.5170/cern-1996-008.21

Huang J., et al., 2016, preprint (arXiv:1611.10012)

Huchra J. P., Geller M. J., 1982, ApJ, 257, 423

Hurier G., Singh P., Hernández-Monteagudo C., 2017, Astronomy & Astrophysics, 625

Ivezić Ž., et al., 2019, ApJ, 873, 111

Khalifa N. E., Hamed Taha M., Hassanien A. E., Selim I., 2018, in 2018 International Conference on Computing Sciences and Engineering (ICCSE)., IEEE, pp 1–6, doi:10.1109/ICCSE1.2018.8374210

Kim E. J., Brunner R. J., 2017, MNRAS, 464, 4463

Koester B. P., et al., 2007, ApJ, 660, 239

Kotsiantis S. B., 2007, in Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. Vol. 31, IOS Press, Amsterdam, The Netherlands, The Netherlands, pp 3–24, http://dl.acm.org/citation.cfm?id=1566770.1566773

Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Pereira F., Burges C. J. C., Bottou L., Weinberger K. Q., eds, , Advances in Neural Information Processing Systems 25. Curran Associates, Inc., New York, USA, pp 1097–1105, http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-netwo pdf

L. McHugh M., 2008, Biochemia Medica, 18, 7

LeCun Y., Haffner P., Bottou L., Bengio Y., 1999, in Forsyth D., ed., Feature Grouping. Springer, London, UK

Lin T.-Y., et al., 2014, preprint (arXiv:1405.0312)

Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.-Y., Berg A. C., 2015, preprint (arXiv:1512.02325)

Martinez M., Stiefelhagen R., 2018, preprint (arXiv:1810.05075)

McGlynn T., Scollick K., White N., 1998, in Brian J. M., Daniel A. G., Jeffrey J. E. H., Harry E. P., eds, New Horizons from Multi-Wavelength Sky Surveys, Proc. 179th Symposium of the International Astronomical Union, Vol. 179, Kluwer Academic Publishers, Baltimore, USA, p. 465

Mehrtens N., et al., 2012, MNRAS, 423, 1024

Ng A. Y., 2004, in Proceedings of the Twenty-first International Conference on Machine Learning. ICML '04. ACM, New York, NY, USA, pp 78–, doi:10.1145/1015330.1015435, http://doi.acm.org/10.1145/1015330.1015435

Ntampaka M., Trac H., Sutherland D. J., Battaglia N., Póczos B., Schneider J., 2015, The Astrophysical Journal, 803, 50

Ntampaka M., Trac H., Sutherland D. J., Fromenteau S., Póczos B., Schneider J., 2016, The Astrophysical Journal, 831, 135

Ntampaka M., et al., 2019, The Astrophysical Journal, 876, 82

Nwankpa C., Ijomah W., Gachagan A., Marshall S., 2018, preprint (arXiv:1811.03378)

Pasquet J., Bertin E., Treyer M., Arnouts S., Fouchez D., 2019, A&A, 621, A26

Pearson J., Pennock C., Robinson T., 2018, Emergent Scientist, 2, 1

Perez L., Wang J., 2017, preprint (arXiv:1712.04621)

Petrillo C. E., et al., 2019, MNRAS, 482, 807

Planck Collaboration et al., 2014, A&A, 571, A29

Planck Collaboration: et al., 2015, A&A, 582, A30

Planck Collaboration et al., 2016, Astronomy and Astrophysics, 594, A13

Ren S., He K., Girshick R., Sun J., 2015, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39

Ren Y., Zhu C., Xiao S., 2018, Mathematical Problems in Engineering, 2018, 1

Ribli D., Ármin Pataki B., Zorrilla Matilla J. M., Hsu D., Haiman Z., Csabai I., 2019, Monthly Notices of the Royal Astronomical Society, Vol. 490, Oxford University Press, p. 1843

Roser M., Ritchie H., 2019, Our World in Data

Ruder S., 2016, preprint (arXiv:1609.04747)

Rumelhart D. E., Hinton G. E., Williams R. J., 1986, Nature, 323, 533

Rykoff E. S., et al., 2014a, ApJ, 785, 104

Rykoff E. S., et al., 2014b, ApJ, 785, 104

Rykoff E. S., et al., 2016, ApJS, 224, 1

Sarazin C. L., 1986, Rev. Mod. Phys., 58, 1

Schaefer, C. Geiger, M. Kuntzer, T. Kneib, J.-P. 2018, A&A, 611, A2

Schmidhuber J., 2014, Neural Networks, 61

Simonyan K., Zisserman A., 2014, preprint (arXiv:1409.1556)

Smirnov N. V., 1939, Bull. Math. Univ. Moscou, 2, 3

Smith M. J., Geach J. E., 2019, Monthly Notices of the Royal Astronomical Society, Oxford University Press, Vol. 490, p. 4985

Smith A. G., Hopkins A. M., Hunstead R. W., Pimbblet K. A., 2012, MNRAS, 422, 25

Staniszewski Z., et al., 2009, ApJ, 701, 32

Stott J. P., Edge A. C., Smith G. P., Swinbank A. M., Ebeling H., 2008, MNRAS, 384, 1502

Sunyaev R. A., Zeldovich Y. B., 1972, Comments on Astrophysics and Space Physics, 4, 173

Szegedy C., Toshev A., Erhan D., 2013, in Proceedings of the 26th International Conference on Neural Information Processing Systems. NIPSâĂŹ13, Vol. 2, Curran Associates Inc., New York, USA, p. 2553, http://dl.acm.org/citation.cfm?id=2999792.2999897

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z., 2015, preprint (arXiv:1512.00567)

Takey A., Schwope A., Lamer G., 2011, Astronomy and Astrophysics, 534, A120

Tharwat A., 2018, Applied Computing and Informatics

Torrey L., Shavlik J. W., 2009, in Transfer Learning.

Umetsu K., 2010, preprint (arXiv:1002.3952)

Vanderlinde K., et al., 2010, ApJ, 722, 1180

Walmsley M., et al., 2019, Monthly Notices of the Royal Astronomical Society, Oxford University Press

Wen Z. L., Han J. L., 2014, Substructure and dynamical state of 2092 rich clusters of galaxies derived from photometric data, Monthly Notices of the Royal Astronomical Society, Vol. 743, Oxford University Press, p. 275

Wen Z. L., Han J. L., Liu F. S., 2012, ApJS, 199, 34

Wittman D., Dell'Antonio I. P., Hughes J. P., Margoniner V. E., Tyson J. A., Cohen J. G., Norman D., 2006, ApJ, 643, 128

Wright E. L., 2006, PASP, 118, 1711

Yadav S., Shukla S., 2016, in 2016 IEEE 6th International Conference on Advanced Computing (IACC). IEEE, New York, USA, pp 78–83, doi:10.1109/IACC.2016.25

Zhu X.-P., Dai J.-M., Bian C.-J., Chen Y., Chen S., Hu C., 2019, Astrophysics and Space Science, 364, 55

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

Table A1. The Total, RPN and DN loss values at different steps during the training of our model evaluated on the test set.

Step	Total loss	RPN objectness loss	RPN box regression loss	DN classification loss	DN box regression loss
147	0.5954	0.3825	0.0535	0.1162	0.0432
204	0.3448	0.2023	0.0538	0.0605	0.0282
313	0.3185	0.1242	0.0515	0.0794	0.0634
450	0.2841	0.0834	0.0487	0.0994	0.0526
581	0.2559	0.0694	0.0436	0.0838	0.0591
934	0.1641	0.0441	0.0380	0.0552	0.0268
1065	0.1347	0.0401	0.0302	0.0367	0.0277
1361	0.1399	0.0344	0.0283	0.0543	0.0229
1455	0.2004	0.0327	0.0268	0.1129	0.0280
1582	0.1242	0.0302	0.0279	0.0443	0.0217
1709	0.1049	0.0307	0.0256	0.0274	0.0212
1836	0.1465	0.0280	0.0261	0.0646	0.0277
1960	0.1344	0.0281	0.0301	0.0531	0.0232
2083	0.1090	0.0290	0.0246	0.0377	0.0177
2209	0.1322	0.0274	0.0237	0.0521	0.0290
2335	0.1540	0.0284	0.0256	0.0712	0.0288
2461	0.1096	0.0267	0.0247	0.0365	0.0218
2588	0.0942	0.0273	0.0245	0.0244	0.0181
2716	0.1184	0.0269	0.0354	0.0339	0.0223
2842	0.0986	0.0253	0.0235	0.0324	0.0173
2968	0.0969	0.0247	0.0222	0.0310	0.0190
3187	0.1048	0.0251	0.0275	0.0326	0.0196
3312	0.1293	0.0234	0.0233	0.0611	0.0214
3438	0.1125	0.0240	0.0224	0.0458	0.0204
3566	0.0900	0.0250	0.0232	0.0243	0.0175
3697	0.0901	0.0247	0.0213	0.0256	0.0185
3823	0.1237	0.0233	0.0238	0.0548	0.0219
3948	0.1189	0.0233	0.0229	0.0520	0.0208
4073	0.1059	0.0231	0.0208	0.0420	0.0201
4198	0.1015	0.0239	0.0217	0.0337	0.0223
4321	0.1177	0.0224	0.0204	0.0509	0.0240
4543	0.1118	0.0231	0.0221	0.0461	0.0205
4675	0.0943	0.0232	0.0215	0.0294	0.0202
4805	0.1042	0.0232	0.0264	0.0297	0.0249
4944	0.1024	0.0222	0.0194	0.0398	0.0209
5072	0.1038	0.0223	0.0201	0.0423	0.0191
5198	0.0881	0.0241	0.0205	0.0253	0.0182
5330	0.1047	0.0220	0.0202	0.0425	0.0199
5462	0.1275	0.0219	0.0241	0.0541	0.0274
5587	0.0969	0.0233	0.0220	0.0294	0.0222
5917	0.0972	0.0224	0.0265	0.0265	0.0219
6048	0.0943	0.0221	0.0184	0.0344	0.0194
6179	0.1038	0.0228	0.0228	0.0313	0.0268
6495	0.1066	0.0235	0.0231	0.0352	0.0248
6698	0.1035	0.0216	0.0184	0.0446	0.0189
6824	0.1022	0.0223	0.0190	0.0427	0.0182
6951	0.1045	0.0217	0.0187	0.0460	0.0180
7077	0.1029	0.0217	0.0188	0.0378	0.0247
7204	0.0910	0.0208	0.0180	0.0341	0.0181
7332	0.0974	0.0215	0.0176	0.0403	0.0181
7458	0.0926	0.0249	0.0211	0.0268	0.0198