

Estimating Cluster Masses from SDSS Multi-band Images with Transfer Learning

Sheng-Chieh Lin,¹★ Yuanyuan Su,¹ Gongbo Liang,² Yuanyuan Zhang,^{3,4} Nathan Jacobs⁵ and Yu Zhang⁵

¹Department of Physics and Astronomy, University of Kentucky, 505 Rose Street, Lexington, KY 40506, USA

²Department of Computer Science, Eastern Kentucky University, 521 Lancaster Avenue, Richmond, KY 40475, USA

³Fermi National Accelerator Laboratory, PO Box 500, Batavia, IL 60510, USA

⁴Mitchell Institute for Fundamental Physics and Astronomy, and Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843-4242, USA

⁵Department of Computer Science, University of Kentucky, 329 Rose Street, Lexington, KY 40506, USA

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The total masses of galaxy clusters characterize many aspects of astrophysics and the underlying cosmology. It is crucial to obtain reliable and accurate mass estimates for numerous galaxy clusters over a wide range of redshifts and mass scales. We present a transfer-learning approach to estimate cluster masses using the *ugriz*-band images in the SDSS Data Release 12. The target masses are derived from X-ray or SZ measurements that are only available for a small subset of the clusters. We designed a semi-supervised deep learning model consisting of two convolutional neural networks. In the first network, a feature extractor is trained to classify the SDSS photometric bands. The second network takes the previously trained features as inputs to estimate their total masses. The training and testing processes in this work depend purely on real observational data. Our algorithm reaches a mean absolute error (MAE) of 0.232 dex on average and 0.214 dex for the best fold. The performance is comparable to that given by redMaPPer, 0.192 dex. We have further applied a joint integrated gradient and class activation mapping method to interpret such a two-step neural network. The performance of our algorithm is likely to improve as the size of training dataset increases. This proof-of-concept experiment demonstrates the potential of deep learning in maximizing the scientific return of the current and future large cluster surveys.

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

1 INTRODUCTION

Clusters of galaxies, residing at the knots of the cosmic web filaments, are the most massive gravitationally-bound systems in the Universe. The understanding of galaxy clusters is built upon the standard cosmology model and complicated baryon physics (e.g., Abell et al. 1989, Navarro et al. 1996, Ebeling et al. 1998, Lin et al. 2004, Berlind et al. 2006, Mantz et al. 2010a). The abundance of massive halos across cosmic time is sensitive to the matter density, the normalization of the primordial perturbation, and the density of state. Halo mass functions constructed from samples of galaxy clusters provide a unique way to examine the cosmological models. In practice, the galaxy cluster mass estimates are subject to a variety of systematic effects particularly associated with the complicated substructures within a cluster (Rasia et al. 2006). As a booming number of cluster cosmology studies emerge, it is paramount to obtain unbiased mass estimates through a careful handling of systematics.

As multi-component systems, galaxy clusters are detectable at different parts of the electromagnetic spectrum. In optical and near-infrared, cluster galaxies are luminous and can be effectively searched by ground-based observatories. The intra-cluster medium (ICM),

on the other hand, can be observed in X-ray through the thermal Bremsstrahlung and from the spectral distortion of the cosmic microwave background (CMB) through the Sunyaev–Zeldovich effect (Sunyaev & Zeldovich 1972). Multi-wavelength observations have been utilized to probe the total masses of galaxy clusters. An ideal mass estimate can be obtained through weak lensing, the distortion of the galaxy shape caused by the foreground gravitational potential field. Cluster masses can also be measured with the ICM pressure profile based on the assumption of hydrostatic equilibrium. Such hydrostatic mass estimates can be biased by non-thermal processes in the ICM related to gas motions, although *Hitomi* observations reveal a remarkably quiescent atmosphere for the central region of the Perseus cluster (Aharonian et al. 2016). The measurements of weak lensing mass or hydrostatic mass require deep and high quality observations that are yet achievable for a large number of galaxy clusters. It is thus practical to adopt mass proxies inferred from mass-observable relations for large cluster surveys. Common mass proxies are X-ray luminosity (L_X , Maughan 2007), spectral temperature (T_X), gas mass (M_{gas} , Mantz et al. 2016), and their product ($Y_X = T_X M_{\text{gas}}$, Kravtsov et al. 2006), the SZ flux (Y_{SZ} , Planck Collaboration et al. 2016b), optical richness (Rykoff et al. 2014), and the velocity dispersion of member galaxies (Bocquet et al. 2015). Since 90% of the baryon content in galaxy clusters is in the form of ICM (Laganá et al. 2013, Eckert et al. 2016), observables that trace the ICM, such as diffuse

★ E-mail: sli346@uky.edu

X-ray and SZ emissions, are relatively low-scatter mass proxies compared to optical observables such as cluster richness (Angulo et al. 2012).

The scatter in mass at a fixed mass proxy depends critically on the mass range and dynamical state of the cluster sample. Generally speaking, M_{gas} and T_x have a scatter of 10% – 15%. L_x is the noisiest mass proxy from X-ray and requires a small number of photons to measure. It has a scatter of 20% – 30%, while that of richness is 25% (Vikhlinin et al. 2009, Mantz et al. 2010b, Lovisari et al. 2015; Rozo & Rykoff 2014). The systematic uncertainties can be further reduced in these measurements when mass proxies are calibrated with weak-lensing measurements (e.g., Applegate et al. 2016, Chiu et al. 2020). Y_{SZ} has been shown to be a robust mass proxy against different systematic uncertainties such as pressure profile variations and the asymmetric beam effects. It is also unaffected by cosmological dimming and allows us to probe high redshift clusters. The calibrated $Y_{\text{SZ}} - M_{500}$ relation has a remarkably low scatter of $\sim 6\%$ (Planck Collaboration et al. 2016b).

Observationally, X-ray and certain SZ signals are mostly accessible to space observatories, such as *ROSAT*, *XMM-Newton*, and *Planck*, which are not comparable to ground-based observatories in terms of efficiency and costs. In general, optical-based mass proxies are more available and applicable for a larger number of clusters with a wider range of masses than ICM-based estimates. It is therefore essential to build a connection between the cluster masses derived from X-ray or SZ observables and their optical properties. We aim to achieve this goal using modern data-processing techniques.

Machine learning methods have been widely applied in astronomy and astrophysics, e.g., to estimate the photometric redshift (Sadeh et al. 2016), to identify strong lensing signal (Lanusse et al. 2018), and to probe the core properties of galaxy clusters (Su et al. 2020). In particular, Convolutional Neural Networks (CNN) have been shown to be a promising technique to study cluster masses. For example, Ntampaka et al. (2019) used a CNN algorithm to predict cluster masses from synthetic Chandra X-ray images derived from the IllustrisTNG simulations. Meanwhile, several studies have been conducted to learn the dynamical cluster masses using the phase-space distribution of member galaxies. For example, Ho et al. 2019 constructed two CNN models with different dimensions to infer cluster masses from the line-of-sight (LOS) velocities of member galaxies and their projected radial distance to the cluster center. Kodi Ramanah et al. (2021) adopted a 3D CNN on the 3D phase-space information (LOS and 2D projected spatial distribution) of member galaxies to estimate cluster masses. Moreover, to achieve robust estimates of cluster masses with uncertainties, various techniques have been employed in the training of CNN. Kodi Ramanah et al. 2020 adopted a CNN with normalising flow mechanism to estimate the conditional probability distribution of cluster masses. Ho et al. 2021 applied the variational Bayesian inference on the dynamical observables of member galaxies to estimate the cluster masses. However, to our best knowledge, previous machine learning studies of cluster masses all rely on simulated data since real observations often lack known ground truth and involve non-trivial noise and contamination. In this work, we present a novel transfer-learning approach in the framework of CNN to estimate ICM-based mass proxies using the photometric images from the Sloan Digital Sky Survey (SDSS, Blanton et al. 2017).

With the on-going and upcoming large-area surveys, such as the Dark Energy Survey (DES, Abbott et al. 2018) and the Vera C. Rubin Observatory (previously known as LSST), machine learning techniques are expected to play an important role in enhancing the scientific returns of the tremendous amount of in-flowing data. Our

pioneering study using purely real observations can pave the way for this promising future. This paper is structured as follows. We describe the data preparation of the images and catalogues in Section 2. Section 3 is devoted to the model design and training strategy. We present the result and compare it with the performance of traditional methods in Section 4. We interpret the models in Section 5, and discuss our findings in Section 6. Throughout this work we adopt a flat Λ -CDM cosmology from Planck Collaboration et al. (2016a): $\Omega_m = 0.3089$, $\Omega_\Lambda = 0.691$, $\Omega_b = 0.0486$, $H_0 = 67.74 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

2 DATA PREPARATION

We aim to directly predict the ICM-based mass proxy for a cluster from its multi-band photometric images. In this section, we describe the construction of our samples.

2.1 SDSS images and redMaPPer catalogue

To acquire a sizable cluster sample from a single survey, we use the *ugriz*-band image data from SDSS Data Release 12 (DR12, Alam et al. 2015). The survey covers roughly 31000 deg^2 of the sky with the depth of *r*-band magnitude to 22.2, providing a large sample of galaxy clusters. The broad-band images used in training are background subtracted and calibrated by the SDSS imaging pipeline (Lupton et al. 2001; for an updated version, Aihara et al. 2011).

To select a parent sample of galaxy clusters, we make use of the cluster catalogue constructed by the red-sequence Matched-filter Probabilistic Percolation (redMaPPer) algorithm using SDSS DR8 data (Rykoff et al. 2014). In short, the algorithm detects galaxy clusters using the colours and the spatial distribution of the red sequence galaxies. The probability of being a member of a cluster is assigned to each individual galaxy. The sum of these probabilities within a given radius, called richness (λ), approximates the number of red-sequence galaxies within a cluster. Richness is therefore correlated with the optical content of a cluster and can be treated as an optical-based mass proxy for galaxy clusters. Based on the catalogue, we have selected ~ 25000 clusters with richness greater than 20.

The SDSS photometric images of the clusters are obtained by applying the following searching criteria. For each cluster, we select all the images within a square with a length of 4 Mpc centred at the cluster centre listed in the catalogue. These images are further mosaic-ed using an image processing package, SWarp (Bertin et al. 2002). When mosaicing, the images are co-added by the median pixel value, and re-sampled using the LANCZOS3 filter. The resulting images are central-cropped to a physical size of 2 Mpc.

2.2 Image pre-processing

A global constant was added to all the input images to avoid negative pixels in the images due to the background subtraction adopted by the SDSS imaging pipeline. This treatment may introduce a bias to the magnitude measurements of individual pixels. However, it does not affect the relative differences among pixels and images on which our model focuses. The pixel values are converted to a log scale and normalized to the range of 0 and 1. All the input images over a physical size of $2 \times 2 \text{ Mpc}^2$ centred on each cluster are resized to have a dimension of $256 \text{ pixel} \times 256 \text{ pixel}$. Each single band image is triply-stacked to produce a 3-channel image with identical channels to match the number of input channels used by the network. The input

images are augmented during the training process by a random combination of a horizontal/vertical flip and/or a 0/90/180/270 degree rotation.

2.3 Mass estimates

We use M_{500} to characterize the total mass of each cluster, defined as the enclosed mass within R_{500} , a radius within which the average density is 500 times the critical density of the Universe at that redshift. It is desirable to obtain ICM-based mass estimates for as many clusters as possible, whereas the availability of their SZ or X-ray measurements is limited. We make use of the SZ source catalogue from the *Planck* mission 2015 data release for SZ detected clusters. The MCXC, XCS, and RASS-DR7 catalogues are used for X-ray detected clusters that are not included in the *Planck* SZ catalogue already.

Planck: The cluster catalogue built from the *Planck* 2015 data release (Planck Collaboration et al. 2016b) contains 1653 clusters with redshifts spanning from 0.01 to 0.9. Three algorithms have been implemented to detect clusters in the CMB map: two of them are based on the matched multi-filtering technique and the third one uses Bayesian inference. The cluster mass is estimated from the Compton Y -parameter using the $Y_{500} - M_{500}$ relation presented in Beteoule et al. (2014). The redshifts of these clusters are determined through the counterparts matched in external data sets, including MCXC, SDSS redMaPPer catalogue, ACT, and SPT cluster catalogue.

MCXC: The Meta-catalogue of the compiled properties of X-ray detected Clusters of galaxies (MCXC) is composed from nine publicly available *ROSAT* All-sky survey-related cluster catalogues and seven supplemental cluster catalogues from other *ROSAT* surveys, such as The 160 Square Degree *ROSAT* Survey. The data has been homogenised in terms of coordinates, names, soft X-ray luminosities, and masses. Duplicated clusters have been removed. The cluster masses are derived using the $L_X - M_{500}$ relation (Arnaud et al. 2010). Piffaretti et al. (2011) provides more details on the construction of the catalogue.

XCS: The *XMM* Cluster Survey (XCS) catalogue is constructed from the publicly available database, XMM-Newton Science Archive, for the purpose of studying cosmology and the evolution of cluster X-ray properties. It contains 503 optically confirmed X-ray clusters with redshift ranging from 0.06 to 1.46. The detailed construction of this catalogue is given by Mehrrens et al. (2012). We derive the cluster masses directly from the cluster radius R_{500} provided in the catalogue.

RASS-DR7: This catalogue is constructed by Wang et al. (2014) based on the SDSS galaxy group catalogue of Yang et al. (2007). The X-ray properties of these optically-selected clusters are taken from the *ROSAT* broad-band X-ray images. We select clusters with $\text{SNR} > 1$ that are not included in other X-ray or SZ catalogues. Their masses are estimated using the $L_X - M_{500}$ scaling relation calibrated for galaxy groups from Mantz et al. 2010b.

We cross-match the redMaPPer catalogue with the above X-ray and SZ cluster catalogues. We require the difference of the cluster redshifts to be within $\delta(z) < 0.05$ and the projected separation of the cluster centres to be within 10 arcmin. Spectroscopic redshift is used if available in these catalogues; otherwise, photometric redshift is adopted. To avoid assigning multiple X-ray/SZ sources to the same cluster, we further impose that the 3D spatial separation between the cluster centres is within 500 kpc. The matching criteria applied here are slightly stricter than that applied in the cylindrical matching in Rozo & Rykoff (2014). A similar projected matching criteria can also

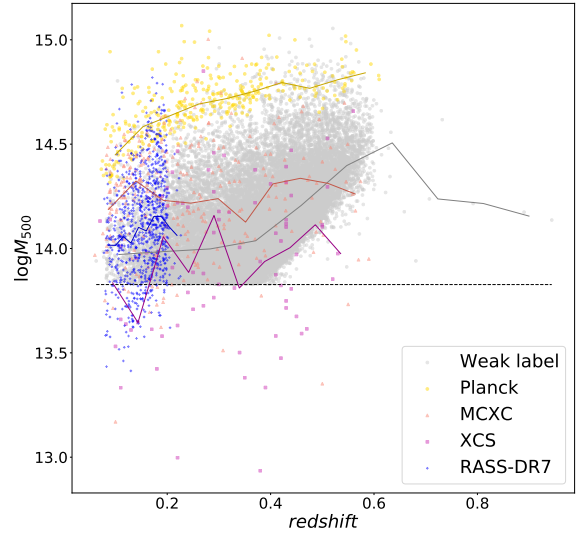


Figure 1. Cluster mass as a function of redshift for clusters in our sample. The distributions of different source catalogues are colour-coded separately. Those without ICM-based mass estimates are denoted in grey. The solid lines represent the median of $\log M_{500}$ in each redshift bin width of $\delta_z = 0.01$. The dashed line shows the richness cut $\lambda = 20$ applied in the redMaPPer catalogue.

be found in Sadibekova et al. (2014). The cross-matching criteria are summarized below:

- (i) $\delta(z) < 0.05$
- (ii) $\delta(\theta) < 10'$
- (iii) $\delta(r) < 500 \text{ kpc}$

The cluster sample is divided into two subsets, strong and weak labels, based on the availability of their X-ray/SZ counterparts. For strong labels, we obtain 1515 SDSS redMaPPer clusters for which the ICM-based masses are available. Of these clusters, 340 are from the *Planck* catalogue, 271 are from MCXC, 78 are from XCS, and 826 are from RASS-DR7. Cluster properties other than the halo mass estimates such as centre, redshift, richness are taken from the redMaPPer catalogue. The mass-redshift distribution of all the clusters in our sample is shown in Fig. 1.

The rest of the 24611 redMaPPer clusters without X-ray or SZ counterparts are used as weak labels with the cluster masses derived from the richness-mass relation given by Jimeno et al. 2018, which takes the form of a powerlaw:

$$\langle M_{500} \rangle(\lambda) = M_0 \left(\frac{\lambda}{\lambda_0} \right)^{\alpha_{M|\lambda}} \quad (1)$$

where M_0 is the reference mass defined at $\lambda = \lambda_0$. We follow the best-fit values given in the paper where $M_0 = 10^{14.434} M_\odot$, $\alpha = 1.23$, and a fixed $\lambda_0 = 60$. These values are obtained by comparing the mass bias measurement and two richness-mass relations obtained from the weak-lensing analysis in Melchior et al. (2017) and Simet et al. (2017).

Table 1. Layer parameters of the regression network. The input size of 2D Conv 1 is $1 \times 512 \times 5$, which is the size of the stacked feature vectors from the feature extractor. The outputs from the last 2D Conv layer and the 1D Conv layer are flattened and concatenated to be fed into FC 1 layer.

| Layer | kernel | output size |
|---------------------|--------------|--------------------------|
| Input: feature maps | | |
| 2D Conv 1 | 1×1 | $1 \times 512 \times 5$ |
| BN | – | $1 \times 512 \times 5$ |
| ReLU | – | $1 \times 512 \times 5$ |
| 2D Conv 2 | 3×3 | $16 \times 510 \times 3$ |
| BN | – | $16 \times 510 \times 3$ |
| ReLU | – | $16 \times 510 \times 3$ |
| 2D Conv 3 | 3×3 | $64 \times 508 \times 1$ |
| BN | – | $64 \times 508 \times 1$ |
| ReLU | – | $64 \times 508 \times 1$ |
| Input: catalog | | |
| 1D Conv | 1×1 | $1 \times 1 \times 1$ |
| BN | – | $1 \times 1 \times 1$ |
| ReLU | – | $1 \times 1 \times 1$ |
| FC 1 | – | 1024 |
| FC 2 | – | 512 |
| FC 3 | – | 1 |

3 METHODS

To build the mass prediction model, we adopt a two-step semi-supervised approach to utilize both the large weak label sample and the small strong label sample. Conventionally, semi-supervised learning refers to a category of machine learning using both unlabeled and labeled data to achieve better predictions than those solely based on labeled data (Zhu & Goldberg 2009, Van Engelen & Hoos 2020). The need of this type of learning is due to lack of sufficient labeled data to train a deep network. Our model falls into this category because a larger set of weak-labeled data (although it is not unlabeled) has been used to facilitate the training.

In the first step, we perform a 5-class classification task to predict the photometric band of each image in the combined sample including both weak and strong label data. In the second step, we perform a regression task to estimate the ICM-based mass, M_{500} , for the strong label data. The neural networks of these two tasks are connected through transfer learning. These networks are implemented in PyTorch (Paszke et al. 2019). The model architecture and the training procedure are described in detail in the following subsections.

3.1 Neural network architecture

We use a residual network (ResNet; He et al. 2015), in particular the well-known ResNet-18 architecture, as the core feature extraction architecture. Residual networks are a class of CNNs that use residual connections to facilitate faster training. Such networks have been widely used for image classification applications. Like other CNNs, they are composed of sets of trainable convolutional filters, which learn to extract features based on the task the network is being optimized for. Such networks can learn to detect low-level features, such as curves and edges, and higher-level features, such as human faces and vehicles. The convolutional filters are arranged roughly in sequence, and are interspersed with pooling layers, which are used to reduce the spatial resolution, batch normalization (BN) layers, and nonlinear activation layers (e.g., the rectified linear unit), which enable the learning of nonlinear mappings. Two types of pooling are

used: max pooling and average pooling, which respectively compute the maximum value and the average value in local windows across the image. These two pooling operations can also be applied to the entire image, called global max pooling and global average pooling (GAP). The GAP layer is typically near the end of the network, and marks the transition from extracting spatial features to computing properties of the full image. In addition, it provides flexibility by allowing images of different sizes.

Neural network-based approaches are often considered uninterpretable, *black box* methods. However, there is often useful structure that can be observed in the extracted features. The convolutional layers that are closer to the input image typically extract the low-level features and the layers closer to the output extract the high-level features, which are based on the low-level features. In addition, it can be useful to view the network as being composed of two major modules: one is a linear classifier or regression model, which is just the final layer of the network, and the other is a feature extraction module that is optimized to translate raw inputs into a suitable feature representation for the linear model.

The residual connection is the unique architecture component that was popularized by the introduction of residual networks. It is realised by adding the input of a module f_i , which can consist of several convolutional and other layers, to the output of the module as follows:

$$y_i = \phi(f_i(x_i; \Theta_i) + x_i), \quad (2)$$

where x_i is the input, $f_i(\cdot)$ represents the mapping with trainable parameters Θ_i , $\phi(\cdot)$ is the non-linear activation function, and y_i is the output.

The full architecture of our model is shown in Fig. 2. In the band classification task, ResNet-18 is used as the feature extractor. All layers before the GAP layer in the ResNet-18 are used and followed by an GAP layer, and three fully connected (FC) layers. The network is initialized with ImageNet weights, which have been shown to be widely useful for transfer learning. Note that, as mentioned in Sec. 2.2, we stack a single band image into a 3-channel image to avoid changing the original ResNet-18 architecture and to allow us to initialize the network with the ImageNet weights. As for the regression task, we use a relatively simple architecture consisting of three 2D Conv layers, a 1D Conv layer, and three FC layers at the end. A BN layer and a ReLU layer are connected after each of the Conv layers. The first 2D Conv layer takes the feature vectors from the classification network as the inputs. The 1D Conv layer is used to include the information on redshift z , and its output is directly concatenated at the end of the output vectors of the 2D Conv layer, together to be fed to the FC layers. The layer information of the regression network is shown in Table 1. The number of parameters of the feature extractor is $\sim 1.25 \times 10^7$ and that of the regression network is $\sim 3.38 \times 10^7$.

We note that while we used the ResNet-18 architecture, there are many alternatives that could be used. We selected it by empirical evaluation over alternatives, including VGG-16 (Simonyan & Zisserman 2014), GoogLeNet (Szegedy et al. 2014), and ResNet-50, which is a deeper variant of ResNet-18.

3.2 Training procedure

The semi-supervised transfer learning model consists of two CNNs which perform the classification task on SDSS photometric bands and the regression task on estimating masses, respectively. The band classification training is performed on the feature extractor, which outputs the predicted photometric bands and the feature vectors for

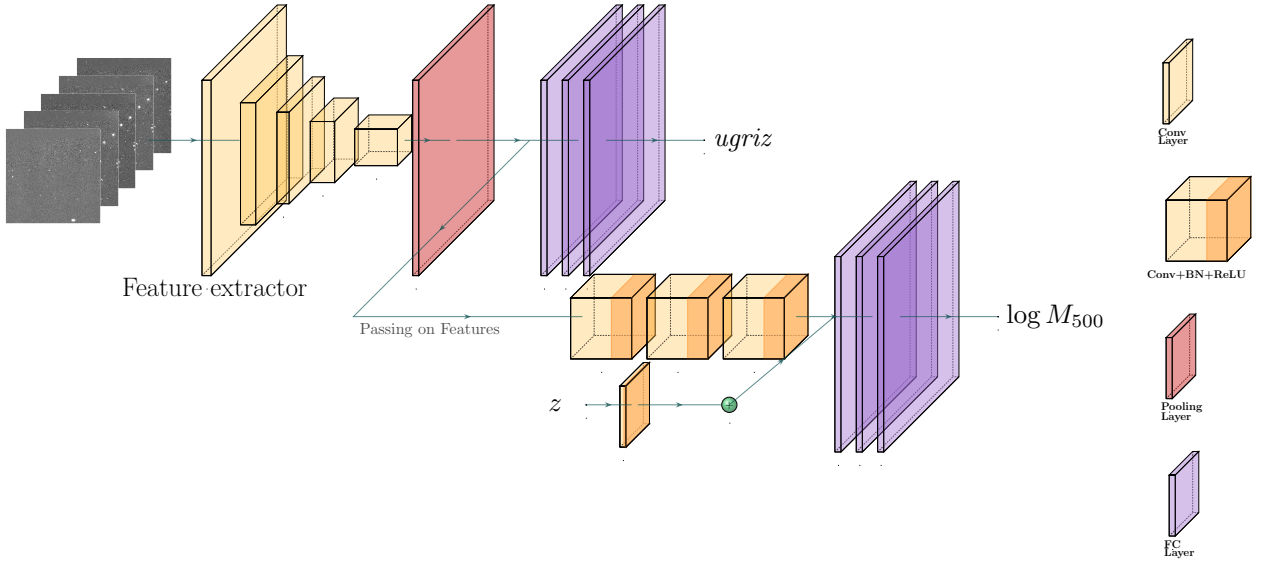


Figure 2. The full architecture of our model. The feature extractor is solely based on the ResNet-18 architecture. After the GAP layer, we divide the outputs into two branches: one is directly connected to three FC layers which outputs the photometric bands and the other branch leads to the next training of the mass estimation. The green plus sign after the 1D Conv layer indicates the concatenation of the features from both types of the Conv layers in the regression network.

later use. The learning procedure is carried out by minimizing the cross-entropy loss function. The performance of feature extractor is evaluated by employing a train-test split with 80% of data used for training and the rest for testing. The training iterates for 100 epochs and reaches a convergence within 50 epoch. We run the training for longer epoch to make sure that over-fitting does not occur. As a result, the best accuracy of band predictions on test set is 92%. This high accuracy implicitly indicates the feature extractor has built up knowledge about the photometric images.

For the regression training, the second network takes the feature vectors obtained from the GAP layer of the extractor as the input to be trained to predict the cluster masses. As a pretraining strategy, we first train the regression network on weak label data. Initialized with the pretrained weights, the network is further trained on strong label data to produce the final results. We also feed the network with the cluster redshift z using a 1D Conv layer to help break the degeneracy between cluster sizes and distances. Redshift is provided with a trainable parameter that the network by itself can decide to take it into account or not. The loss function used in the regression training is defined as follows,

$$L = \frac{1}{N} \sum_{i=1}^N w_i (y_i - \hat{y}_i)^2 + 100\sigma_{\Delta}, \quad (3)$$

where y_i is the output of the model, \hat{y}_i is the ground truth, σ_{Δ} denotes the 1σ standard deviation of the difference between y_i and \hat{y}_i , and w_i is the weight accounting for the imbalance of the data, which is inversely proportional to the number of clusters in a mass range. This loss is specifically designed to encourage the network to minimize the difference between the ground truth and the predictions, and to constrain the distribution of the predictions. For the latter term in the loss function, the scalar multiplied on σ_{Δ} is empirically chosen. The Adam optimizer (Kingma & Ba 2014) with a learning rate of $L_r = 0.0001$ is adopted to update the network for 100 epochs, which is sufficient for the training to converge. Additionally, a 10-fold cross-validation is used to guarantee the generality of the model. In short, the strong label data is evenly divided into 10 subsets with no repetition. Eight subsets are used as the training set, one subset as the

validation set and one as the testing set for each fold. This process is repeated 10 times to cycle through all the data.

4 RESULTS

We compare the predicted cluster masses and their target values in Fig.3. The following statistics are measured to evaluate the performance of the model: the mean absolute error (MAE) is defined as

$$\text{MAE} = \frac{1}{N} \sum_{i=0}^N |\log M_{500} - \log \hat{M}_{500}|, \quad (4)$$

where $\log M_{500}$ is the predicted mass, $\log \hat{M}_{500}$ is the target ICM-based mass, and N is the batch size; the standard deviation $\sigma_{\Delta \log M}$ of the mass residual in log scale, which is defined as

$$\Delta \log M = \log M_{500} - \log \hat{M}_{500}. \quad (5)$$

The best fold has an MAE of 0.214 dex. The MAE averaged over 10 cross-validation folds is 0.232 dex. The cross-validation standard deviation of MAE is 0.012 dex showing a stable result across 10 folds. The standard deviation of the residual averaged over 10 folds is 0.292 dex.

4.1 Compared with richness-based mass estimates

For comparison, we also derive the expected cluster mass for each cluster based on the mass-richness relation shown in Eq.1. The redMaPPer algorithm applied to the SDSS catalogue data reaches an MAE of 0.192 dex and a standard deviation of the residual $\sigma_{\Delta \log M} = 0.249$ dex, which is superior but comparable to our model directly applied to SDSS photometric images.

To understand the dependence of the performance on cluster mass scales, we compare the values of MAE and $\sigma_{\Delta \log M}$ as a function of the target value with those derived from redMaPPer. As shown in Fig.4, the performance of our model in terms of both MAE and the scattered residual is comparable to that of redMaPPer at all mass

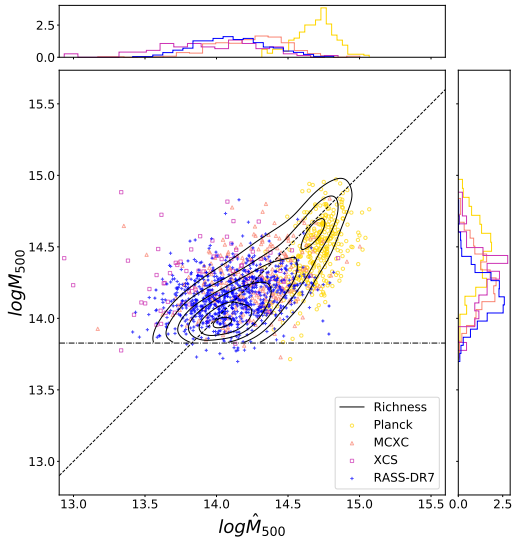


Figure 3. Comparison between the predicted masses $\log M_{500}$ and the target values $\log \hat{M}_{500}$. Four sets of masses obtained from different source catalogues are colour-coded separately. The richness-based mass as a function of target value is shown as the coloured contour. The horizontal dashed-dotted line shows the richness cut of 20 applied in the original redMaPPer catalogue. The probability distributions of predicted and target masses of four subsets (weak label excluded) are shown in the above and right panels, respectively.

scales. Both our model and redMaPPer estimate the masses accurately for galaxy clusters with masses above $10^{14} M_{\odot}$. The accurate estimates are likely due to a higher purity, the ratio of the member galaxies to that of the observed galaxies within a given FOV, for massive clusters, such that the effects of interlopers and background noise are reduced. On the other hand, a clear overestimating trend at the low-mass end can be seen in the lower panel of Fig.4 for both methods. The overestimate in redMaPPer sample is mainly caused by the richness cut at $N = 20$ which removes potential low-mass clusters. In our case, the characterization of low-mass clusters with fewer member galaxies are more vulnerable to interlopers and background noise. We note a slightly worse performance of our model at the very high-mass end (above $10^{14.5} M_{\odot}$) which is not observed in redMaPPer. This may be related to the mean-reversion edge effect as the model prediction for the end of the target value range can be biased towards the mean (Ho et al. 2021; Kodi Ramanah et al. 2021). This bias might be mitigated once we obtain sufficient number of clusters above $10^{14.5} M_{\odot}$. In general, it is desirable to have more labeled data at both ends to improve the performance even though our approach has already greatly reduced the need of strong label data.

Cluster redshift is another important factor since it affects the apparent size and SNR of a cluster. In Fig.5, we show the dependence of the predicted mass on redshift and compare it with redMaPPer. For both methods, a gradually increasing trend of MAEs toward higher redshift is observed. The performance of our model has a higher MAE than that of redMaPPer at all redshifts as shown in the lower panel of Fig.5.

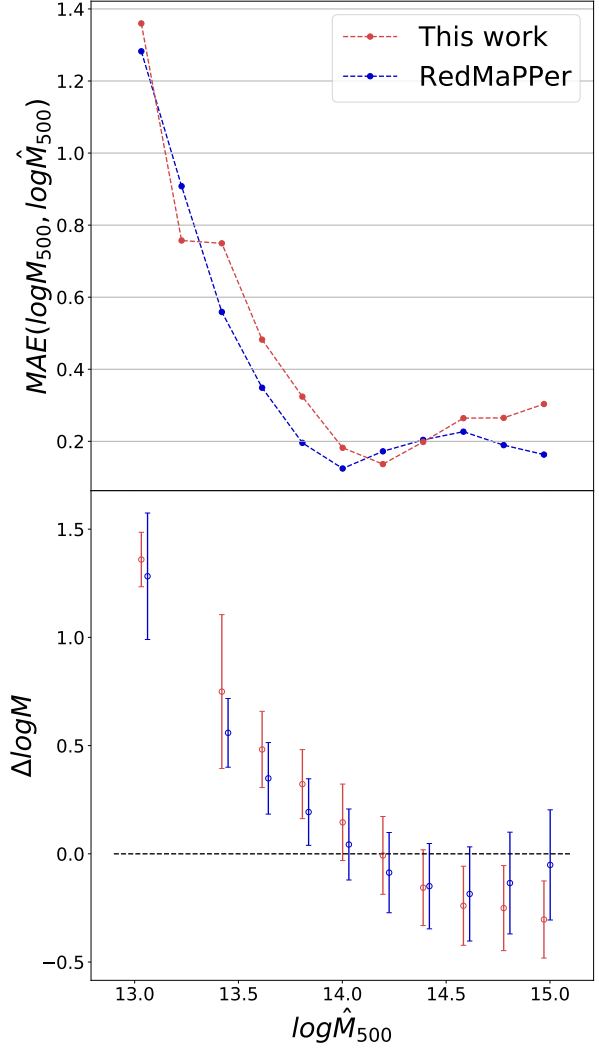


Figure 4. The performance of our deep learning approach in comparison with that of redMaPPer as a function of cluster mass scale. *Upper panel:* the comparison of the MAE defined in Eq.4. *Lower panel:* the comparison of the residue defined in Eq.5. The errorbar denotes the 1σ error. The redMaPPer result is slightly shifted to the right for clarity.

4.2 Compared with plain ResNet-18 model

In this subsection, we demonstrate the necessity of the transfer learning approach for the task of estimating cluster masses from optical images. We compare the performance of our semi-supervised transfer learning model (S.S.+Trans.) with a baseline model using the ResNet-18 architecture which is trained on 5-channel (photometric bands are treated as channels) 256×256 images to estimate the ICM-based masses. We apply the same training procedure and regression loss function given in Sec.3; however, we do not use the ImageNet

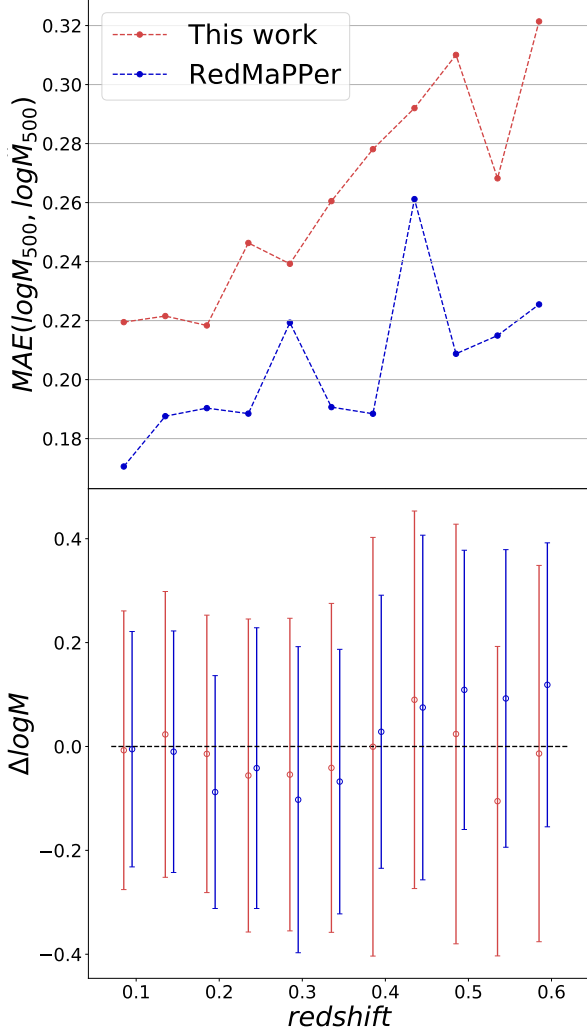


Figure 5. The dependence of performance on redshift for our deep learning approach and the redMaPPer algorithm. The format is the same as Fig. 4. *Upper panel:* the MAE values as a function of redshift. *Lower panel:* the mass residual as a function of redshift with 1σ errorbars.

weights to initialize the baseline model due to the modification of the input layer. To show the improvement given by the semi-supervised transfer learning, we set up two experiments: (1) training baseline on strong label data and (2) semi-supervised training (baseline+S.S.) on weak label data and strong label data.

Table 2 shows the cross-validation MAEs of three models: baseline, baseline+S.S., and our approach, S.S.+Trans. In Fig. 6, we compare the distributions of ground truth as well as the results from these three models. As expected, the baseline model trained directly on strong label data gives the worst performance since the amount of data is extremely insufficient for a deep network like

Table 2. Cross-validation MAEs of the two baselines and our model.

| Models | cv-MAE Ave. | cv-MAE 1σ |
|---------------|-------------|------------------|
| Baseline | 1.069 | 0.302 |
| Baseline+S.S. | 0.28 | 0.018 |
| S.S.+Trans. | 0.232 | 0.012 |

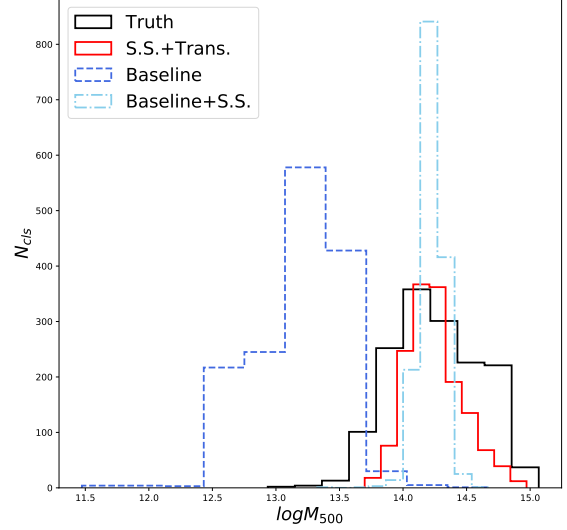


Figure 6. The distribution of predicted masses from several different models and the target values (ICM-based masses). Here we show the results from ResNet-18 baseline (Baseline), ResNet-18 baseline with semi-supervised learning (Baseline+S.S.), and our approach (S.S.+Trans.). The target value, $\log \bar{M}_{500}$, is denoted in black. Both baseline models are shown in the dashed lines and the red solid line is our approach.

ResNet-18 to learn meaningful features. We also find that in Fig. 6 the baseline+S.S. results are distributed sharply around the median of $\log M_{500} = 14.218$. It shows that the model does not acquire enough information about masses, even though a much larger dataset with ~ 25000 clusters is used in the pretraining. By comparing these three models, we conclude that the S.S.+Trans. model learns better from the data and performs well on the given task.

5 MODEL INTERPRETATION

To interpret the mechanism, we utilize the integrated gradients (IG, Sundararajan et al. 2017) method and the Class Activation Mapping (CAM, Zhou et al. 2015) method to quantify the importance of the input features. The two-step transfer learning approach involves two networks – the feature extractor and the regression network, which take different forms of inputs and outputs. We therefore divide the model interpretation into two parts accordingly.

The interpretation method used in both networks is IG, which attributes the output of a network to the input features by integrating the path along the gradients from a baseline input to the original input. The choice of baseline can be vastly different with distinct impacts on the interpretation (Sturmfels et al. 2020). The formal

definition of IG is as follows:

$$IG_i ::= (x_i - x_i^b) \int_0^1 \nabla F(\alpha x_i + (1 - \alpha)x_i^b) d\alpha \quad (6)$$

where $\nabla F(\cdot)$ is the gradients of the model, x_i is the i -th feature of the original input, and x_i^b is the baseline at the same spatial location of x_i . This method has an advantage that it overcomes the issue of the saturate gradient problem commonly occurred in the gradient-based attribution methods. In general, one should avoid using a zero baseline if zero represents certain features of the inputs as suggested by Sturmfels et al. (2020). In our case, we use a zero baseline, which is a set of black images, because the re-scaled input images do not contain pixels with the value of zero across five bands. Additionally, the zero baseline allows us to examine whether the network focuses on the background noise or not.

CAM, the other interpretation method used for the regression network, gives the highest ranked features by weighted-summing the last Conv layer at different spatial locations:

$$M(x, y) = \sum_k w_k f_k(x, y), \quad (7)$$

where the subscript k denotes a neuron unit in the last Conv layer, w_k is the weight of that neuron, and $f_k(x, y)$ is the activation function of that neuron at spatial position (x, y) in a given image. In principle, the weight w_k indicates the importance of such feature with respect to a given class for the classification training. Given the design of our model, we adopt a slightly different version where the IG maps are treated as weights instead of the class weights in traditional CAM.

5.1 Feature extractor

For the feature extractor, the interpretation is based solely on the IG method (Eq.6). In principle, brighter objects in images should gain more attention from the network since they are easier to capture. A set of examples obtained by IG are shown in the middle row of Fig.7. We find different patterns in IG maps across different bands showing that the network has learnt to extract information about the photometry accordingly. However, there are only a certain amount of galaxies have been noticed by the network. This is possibly due to a bright star distracting the network from searching for galaxies. Therefore, a treatment for these bright stars such as incorporating their prior known coordinates would be a possible way to mitigate this bias.

We observe that high impact features distribute at the central parts of the images, especially in *gri*-bands. This could be related to the more noticeable contrast between the central region and the background, providing more information for the network. It is also encouraging to observe that bad columns and artefacts shown in *riz*-bands are ignored by the extractor, which reduces the need to perform image pre-processing to remove these patterns. Additionally, we find that the network has extracted complicated features in fainter bands, such as *u*-band shown in the first row of Fig.7, demonstrating its potential of detecting faint objects in optical images. Based on the interpretation using IG, it is reassuring that in the classification task the network has learnt to extract meaningful features in the images.

5.2 Regression network

For the regression network, we combine IG with CAM given that the feature vectors have different dimensions from single-band 3-channel images. For details, we build the IG maps from the feature vectors with the corresponding regression outputs, de-project the maps back

to the input images by treating the maps as the weights, and apply CAM on those weights. The resulting CAM+IG maps are shown in the bottom row of Fig.7.

The CAM+IG maps depict more abstract features compared to the patterns in the pure IG maps. This is because the inputs of the regression network are the products of higher layers of the feature extractor in which the spatial information has been compressed in the return of the increasing depth. The central regions in some of the bands but not all are highlighted. We do not find a significantly better performance for the cases with their central regions highlighted. This indicates that our network treats these features in a distinct way compared to the traditional algorithms that use the information of photometry and colour. Overall, we do find a consistent result from the visualizations of both methods, IG and CAM+IG. The impact of bright stars still presents in the regression training and no bad features or artifacts have been focused on by the regression network. The interpretations of both networks have touched on the direction of improving network abilities in the future.

6 DISCUSSION AND SUMMARY

In this paper, we presented a transfer learning model consisting of two CNNs, the feature extractor and the regression network. We have utilized a semi-supervised approach to make full use of the entire cluster sample including those without ICM-based mass estimates. The combination of the transfer learning and the semi-supervised learning approaches overcomes the issue of lacking sufficient amount of labeled data. The improvement is confirmed by comparing the performance of our model with the plain ResNet-18 model.

The first network captures the important features of the photometric images by learning to distinguish *ugriz* bands. It correctly predicts photometric bands with an accuracy of 92%. The learnt features are then passed on to the second network which outputs the cluster masses. The mean absolute error of the predicted masses reaches 0.232 dex. Cluster richness derived through the sophisticated catalogue processing adopted by redMaPPer is strongly correlated with ICM-based mass proxies. Our model is at the same level of accuracy as the redMaPPer. It is worth noting that our algorithm takes photometric images as the input directly while redMaPPer uses SDSS categorical data of galaxy.

Most deep learning algorithms require a large amount of labeled data for training. In astronomy, this is mostly available from the computationally expensive simulations. It also takes heavy efforts to mimic real observations, including adding realistic astrophysical and instrumental backgrounds. Compared to other cluster mass estimates using deep learning algorithms, our model overcomes the data hungry issue and can be trained directly on the real data. This work therefore comes closer to the real world.

It is crucial yet challenging to explain the predictions made by deep learning models. We make use of the IG method to interpret the important features obtained by the extractor from the input images. We use the CAM method to de-project the IG maps from the regression network back to the inputs. The IG maps show that the galaxies in the images are important when determining the bands, especially those at the central regions of the field. However, we find that not all member galaxies are noticed by the feature extractor due to the bright stars appearing in the images. On the other hand, the CAM+IG method shows more abstract features compared to the IG maps. We observe that for the regression network, only certain bands have their central regions highlighted, indicating a complicated process of handling the information of photometric bands is adopted by

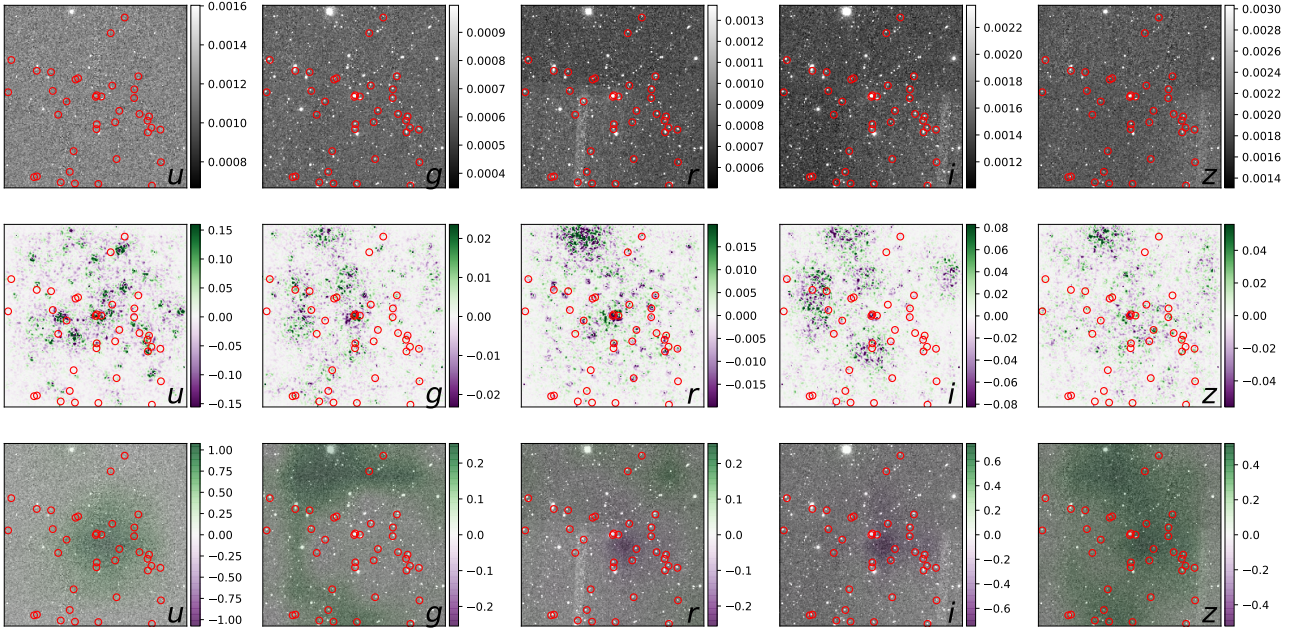


Figure 7. An example of the visualizations of the interpretations given by the IG and CAM methods across five bands. The red open circles denote the cluster members given by the redMaPPer algorithm. The letter at the lower-right corner shows the photometric band of the image. *Upper panel:* the original image from SDSS as the input of the extractor. *Middle panel:* the resulting maps given by the IG applied on the extractor. *Lower panel:* the resulting maps given by the CAM+IG applied on the regression network, and the former is overlapped on the original image.

the network. It is worth noting that our algorithm has learnt to ignore most background, bad pixels, and artefacts, which potentially can be utilized to facilitate imaging processing and cataloguing.

Here we point out a few directions to improve the model. Based on the model interpretation, we have shown that the bright stars have a noticeable impact on the feature extractor which undermines the performance. Therefore, certain treatment of bright stars is expected to improve the results. Deep learning results are known to be positively correlated with the size of the data set. Labeled data at the low mass end is particularly desirable in our case. Presumably, a more advanced model, such as ConvNeXt (Liu et al. 2022), may give better results via extracting more representative features of galaxy clusters or handling the bright stars in a more creative way. Our model currently provides the point estimate of cluster mass without uncertainty measurements. It is crucial for future work to incorporate the uncertainty quantification into our model to provide robust estimation on cluster masses. A few techniques, including the variational inference via Bayesian neural network (Blundell et al. 2015) and the ensemble learning (Lakshminarayanan et al. 2016), have been proposed to quantify the uncertainties, and can be employed alongside our regression network. We stress that the performance of our model is likely to scale up in the near future given the in-flowing high resolution photometric images (e.g., DES, LSST) as well as the more accurate and consistent ICM-based mass estimates (e.g., eROSITA, CMB-S4), promised by the current and future multi-wavelength observations.

ACKNOWLEDGEMENTS

We would thank the University of Kentucky Center for Computational Sciences and Information Technology Services Research Computing for their support and use of the Lipscomb Compute Cluster and associated research computing resources. S.-C. L. and Y. S.

were supported by Chandra X-ray Observatory grant GO1-22126X, NASA grant 80NSSC21K0714, and NSF grant 2107711.

Funding for the Sloan Digital Sky Survey has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions.

SDSS acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

- Abbott T. M. C., et al., 2018, *ApJS*, **239**, 18
- Abell G. O., Corwin Harold G. J., Olowin R. P., 1989, *ApJS*, **70**, 1
- Aharonian F., et al., 2016, *Nature*, **535**, 117–121
- Aihara H., et al., 2011, *ApJS*, **193**, 29
- Alam S., et al., 2015, *ApJS*, **219**, 12
- Angulo R. E., Springel V., White S. D. M., Jenkins A., Baugh C. M., Frenk C. S., 2012, *MNRAS*, **426**, 2046
- Applegate D. E., et al., 2016, *MNRAS*, **457**, 1522
- Arnaud M., Pratt G. W., Piffaretti R., Böhringer H., Croston J. H., Pointecouteau E., 2010, *A&A*, **517**, A92
- Berlind A. A., et al., 2006, *ApJS*, **167**, 1
- Bertin E., Mellier Y., Radovich M., Missonnier G., Didelon P., Morin B., 2002, in Bohlender D. A., Durand D., Handley T. H., eds, *Astronomical Society of the Pacific Conference Series Vol. 281, Astronomical Data Analysis Software and Systems XI*. p. 228
- Betoule M., et al., 2014, *A&A*, **568**, A22
- Blanton M. R., et al., 2017, *AJ*, **154**, 28
- Blundell C., Cornebise J., Kavukcuoglu K., Wierstra D., 2015, in *International Conference on Machine Learning*. pp 1613–1622
- Bocquet S., et al., 2015, *ApJ*, **799**, 214
- Chiu I. N., Umetsu K., Murata R., Medezinski E., Oguri M., 2020, *MNRAS*, **495**, 428
- Ebeling H., Edge A. C., Böhringer H., Allen S. W., Crawford C. S., Fabian A. C., Voges W., Huchra J. P., 1998, *MNRAS*, **301**, 881
- Eckert D., et al., 2016, *A&A*, **592**, A12
- He K., Zhang X., Ren S., Sun J., 2015, arXiv e-prints, [p. arXiv:1512.03385](https://arxiv.org/abs/1512.03385)
- Ho M., Rau M. M., Ntampaka M., Farahi A., Trac H., Póczos B., 2019, *ApJ*, **887**, 25
- Ho M., Farahi A., Rau M. M., Trac H., 2021, *ApJ*, **908**, 204
- Jimeno P., Diego J. M., Broadhurst T., De Martino I., Lazkoz R., 2018, *MNRAS*, **478**, 638
- Kingma D. P., Ba J., 2014, arXiv e-prints, [p. arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
- Kodi Ramanah D., Wojtak R., Ansari Z., Gall C., Hjorth J., 2020, *MNRAS*, **499**, 1985
- Kodi Ramanah D., Wojtak R., Arendse N., 2021, *MNRAS*, **501**, 4080
- Kravtsov A. V., Vikhlinin A., Nagai D., 2006, *ApJ*, **650**, 128
- Laganá T. F., Martinet N., Durret F., Lima Neto G. B., Maughan B., Zhang Y. Y., 2013, *A&A*, **555**, A66
- Lakshminarayanan B., Pritzel A., Blundell C., 2016, arXiv preprint [arXiv:1612.01474](https://arxiv.org/abs/1612.01474)
- Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018, *MNRAS*, **473**, 3895
- Lin Y.-T., Mohr J. J., Stanford S. A., 2004, *ApJ*, **610**, 745
- Liu Z., Mao H., Wu C.-Y., Feichtenhofer C., Darrell T., Xie S., 2022, arXiv e-prints, [p. arXiv:2201.03545](https://arxiv.org/abs/2201.03545)
- Lovisari L., Reiprich T. H., Schellenberger G., 2015, *A&A*, **573**, A118
- Lupton R., Gunn J. E., Ivezić Z., Knapp G. R., Kent S., 2001, in Harnden F. R. J., Primini F. A., Payne H. E., eds, *Astronomical Society of the Pacific Conference Series Vol. 238, Astronomical Data Analysis Software and Systems X*. p. 269 ([arXiv:astro-ph/0101420](https://arxiv.org/abs/astro-ph/0101420))
- Mantz A., Allen S. W., Rapetti D., Ebeling H., 2010a, *MNRAS*, **406**, 1759
- Mantz A., Allen S. W., Ebeling H., Rapetti D., Drlica-Wagner A., 2010b, *MNRAS*, **406**, 1773
- Mantz A. B., et al., 2016, *MNRAS*, **463**, 3582
- Maughan B. J., 2007, *ApJ*, **668**, 772
- Mehrtens N., et al., 2012, *MNRAS*, **423**, 1024
- Melchior P., et al., 2017, *MNRAS*, **469**, 4899
- Navarro J. F., Frenk C. S., White S. D. M., 1996, *ApJ*, **462**, 563
- Ntampaka M., et al., 2019, *ApJ*, **876**, 82
- Paszke A., et al., 2019, in , *Advances in Neural Information Processing Systems* **32**. Curran Associates, Inc., pp 8024–8035, <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-1.pdf>
- Piffaretti R., Arnaud M., Pratt G. W., Pointecouteau E., Melin J. B., 2011, *A&A*, **534**, A109
- Planck Collaboration et al., 2016a, *A&A*, **594**, A13
- Planck Collaboration et al., 2016b, *A&A*, **594**, A27
- Rasia E., et al., 2006, *MNRAS*, **369**, 2013
- Rozo E., Rykoff E. S., 2014, *ApJ*, **783**, 80
- Rykoff E. S., et al., 2014, *ApJ*, **785**, 104
- Sadeh I., Abdalla F. B., Lahav O., 2016, *PASP*, **128**, 104502
- Sadibekova T., Pierre M., Clerc N., Faccioli L., Gastaud R., Le Fevre J. P., Rozo E., Rykoff E., 2014, *A&A*, **571**, A87
- Simet M., McClintock T., Mandelbaum R., Rozo E., Rykoff E., Sheldon E., Wechsler R. H., 2017, *MNRAS*, **466**, 3103
- Simonyan K., Zisserman A., 2014, arXiv e-prints, [p. arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
- Sturmfels P., Lundberg S., Lee S.-I., 2020, *Distill*
- Su Y., et al., 2020, *MNRAS*, **498**, 5620
- Sundararajan M., Taly A., Yan Q., 2017, arXiv e-prints, [p. arXiv:1703.01365](https://arxiv.org/abs/1703.01365)
- Sunyaev R. A., Zeldovich Y. B., 1972, *Comments on Astrophysics and Space Physics*, **4**, 173
- Szegedy C., et al., 2014, arXiv e-prints, [p. arXiv:1409.4842](https://arxiv.org/abs/1409.4842)
- Van Engelen J. E., Hoos H. H., 2020, *Machine Learning*, **109**, 373
- Vikhlinin A., et al., 2009, *ApJ*, **692**, 1033
- Wang L., et al., 2014, *MNRAS*, **439**, 611
- Yang X., Mo H. J., van den Bosch F. C., Pasquali A., Li C., Barden M., 2007, *ApJ*, **671**, 153
- Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A., 2015, arXiv e-prints, [p. arXiv:1512.04150](https://arxiv.org/abs/1512.04150)
- Zhu X., Goldberg A. B., 2009, *Synthesis lectures on artificial intelligence and machine learning*, **3**, 1

This paper has been typeset from a \LaTeX file prepared by the author.