Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **482,** 1211–1230 (2019) doi:10.1093/mnras/sty2646
Advance Access publication 2018 October 23

Radio Galaxy Zoo: CLARAN – a deep learning classifier for radio morphologies

Chen Wu,[1★] Oiwei Ivy Wong,[1★] Lawrence Rudnick,[2★] Stanislav S. Shabala,[3] Matthew J. Alger,[4,5] Julie K. Banfield,[4,6] Cheng Soon Ong,[5] Sarah V. White,[7] Avery F. Garon,[2] Ray P. Norris,[8,9] Heinz Andernach,[10] Jean Tate,[11] Vesna Lukic,[12] Hongming Tang,[13] Kevin Schawinski[14] and Foivos I. Diakogiannis[1,15]

[1]*International Centre for Radio Astronomy Research (ICRAR), The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia*
[2]*School of Physics and Astronomy, University of Minnesota, Minneapolis, MN 55455, USA*
[3]*School of Natural Sciences, University of Tasmania, Private Bag 37, Hobart, Tasmania 7001, Australia*
[4]*Research School of Astronomy and Astrophysics, The Australian National University, Canberra, ACT 2611, Australia*
[5]*Data61, CSIRO, Canberra, ACT 2601, Australia*
[6]*ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO), Building A28, School of Physics, The University of Sydney, NSW 2006, Australia*
[7]*International Centre for Radio Astronomy Research (ICRAR), Curtin University, Bentley, WA 6102, Australia*
[8]*Western Sydney University, Locked Bag 1797, Penrith South, NSW 1797, Australia*
[9]*CSIRO Astronomy and Space Science, Australia Telescope National Facility, PO Box 76, Epping, NSW 1710, Australia*
[10]*Departamento de Astronomía, DCNE, Universidad de Guanajuato, Apdo. Postal 144, Guanajuato, CP 36000, Gto., Mexico*
[11]*Zooniverse Citizen Scientist, c/o Oxford Astrophysics, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK*
[12]*Hamburger Sternwarte, University of Hamburg, Gojenbergsweg 112, D-21029 Hamburg, Germany*
[13]*School of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK*
[14]*Institute for Particle Physics and Astrophysics, ETH Zürich, Wolfgang-Pauli-Str. 27, CH-8093 Zürich, Switzerland*
[15]*Data61, CSIRO, Floreat, WA 6014, Australia*

Accepted 2018 September 21. Received 2018 August 17; in original form 2018 May 29

ABSTRACT

The upcoming next-generation large area radio continuum surveys can expect tens of millions of radio sources, rendering the traditional method for radio morphology classification through visual inspection unfeasible. We present CLARAN — **Cla**ssifying **R**adio sources **A**utomatically with **N**eural networks — a proof-of-concept radio source morphology classifier based upon the Faster Region-based Convolutional Neutral Networks method. Specifically, we train and test CLARAN on the FIRST and WISE (Wide-field Infrared Survey Explorer) images from the Radio Galaxy Zoo Data Release 1 catalogue. CLARAN provides end users with automated identification of radio source morphology classifications from a simple input of a radio image and a counterpart infrared image of the same region. CLARAN is the first open-source, end-to-end radio source morphology classifier that is capable of locating and associating discrete and extended components of radio sources in a fast (<200 ms per image) and accurate (≥90 per cent) fashion. Future work will improve CLARAN's relatively lower success rates in dealing with multisource fields and will enable CLARAN to identify sources on much larger fields without loss in classification accuracy.

Key words: methods: numerical – methods: statistical – techniques: image processing – galaxies: active – radio continuum: galaxies.

1 INTRODUCTION

Understanding the growth and evolution of active galactic nuclei (AGNs) is a fundamental area of research in the field of galaxy evolution as the pre-Square Kilometre Array (pre-SKA) experiments are now beginning their surveys. Radio AGN can be classed as 'jetted' or 'non-jetted' (Padovani 2017). On larger angular scales, radio jets can extend to great distances away from their host galaxies depending on their intrinsic mechanical energy and the environment into which they are launched. Over time, a bipolar jet may fade into two distinct radio lobes that are no longer connected to the host galaxy where it originated. Therefore, while approximately

★ E-mail: chen.wu@uwa.edu.au (CW); ivy.wong@uwa.edu.au (OIW); larry@umn.edu (LR)

90 per cent of radio sources are compact in structure, the remaining radio galaxy morphologies are extended with multiple radio source components and a rich set of structures.

Until now the cross-identification of associated radio source components as well as the originating host galaxies are made via visual inspection. Currently, the most efficient form of visual identification is via citizen science projects such as Radio Galaxy Zoo (RGZ; Banfield et al. 2015). RGZ is based on large-area radio surveys and the efficacy of this project is demonstrated by the science results and recent discoveries of extreme classes of radio source morphologies (Banfield et al. 2016; Kapińska et al. 2017; Contigiani et al. 2017).

On the other hand, it is clear that we have reached even the limitations of citizen science since the number of complex, extended sources expected from the next-generation radio surveys such as the Evolutionary Map of the Universe (EMU; Norris et al. 2011) will be far too great for a standalone citizen science project to be an efficient method. Therefore, automated methods of classification are necessary. Simple automated methods based upon source position matching can be effective for a significant fraction of radio sources (e.g. Kimball & Ivezić 2008). However, complex extended radio sources with multiple discrete components and morphology will require more sophisticated methods. Therefore, deep learning methods provide one such avenue for the specific task of radio source identification and classification. Recently, Wright et al. (2017) demonstrated that a combination of citizen science and deep learning methods will maximize the science output of a data set and outperform the capabilities of each method individually.

The main purpose of this paper is to present a proof-of-concept, publicly available,[1] deep learning-based method known as Classifying Radio sources Automatically using Neural networks (CLARAN). CLARAN takes as input a pair of World Coordinate System-aligned radio and infrared (IR) images. It finds all radio sources and classifies them into one of the six morphology classes based on RGZ. The six classes of morphologies are not defined in the traditional manner of Fanaroff–Riley (FR) classes – FR-I and FR-II (Fanaroff & Riley 1974; Owen & Ledlow 1994) – but in terms of source associations and identifications that are produced by RGZ's Data Release 1 (Wong et al., in preparation) represented as the number of components and peaks . Therefore, a single radio galaxy or radio source can be composed of one or more components and/or peaks. This paper builds upon RGZ's earlier exploration in combining the results from RGZ with advanced machine learning algorithms such as Lukic et al. (2018) and Alger et al. (2018).

We briefly introduce advanced machine learning (also known as deep learning) methods in Section 2. The RGZ citizen science project and data pre-processing for feature fusion is described in Section 3. In the spirit of open source reproducibility, Section 4 provides a complete technical description of CLARAN. Section 5 details the error analysis and metrics-based evaluation commonly used in the field of machine learning. Section 6 describes an example of the simplest automated application of CLARAN from the perspective of an astronomer and its reliability verification analysis. This ensures the accuracy of the classifications and provides additional information on the presence of multiple radio sources within the same image. Implications of our work and future research are briefly discussed in Section 7 and we provide a summary of our results in Section 8.

2 DEEP LEARNING METHODS

Deep learning methods (LeCun, Bengio & Hinton 2015), particularly Convolutional Neural Networks (CNNs, Krizhevsky, Sutskever & Hinton 2012), have recently achieved recognition capabilities that are comparable to or even better than humans in several visual recognition tasks, such as understanding traffic signs (Ciregan, Meier & Schmidhuber 2012), identifying faces (Taigman et al. 2014), and classifying general images (He et al. 2016). CNNs have recently been explored to address a number of astrophysical problems such as: (1) effective identification of exoplanet candidates (Shallue & Vanderburg 2018; Pearson, Palafox & Griffith 2018); (2) the identification of gravitational lenses (Schaefer et al. 2018) and the estimation of strong gravitational lensing parameters (Hezaveh, Levasseur & Marshall 2017); (3) automatic visual detection of galaxy structures such as galactic bars and mergers (Abraham et al. 2018; Ackermann et al. 2018); (4) the determination of physical stellar parameters from optical stellar spectra (Fabbro et al. 2018); and (5) the identification of transients in real-time via image differencing (Sedaghat & Mahabal 2018).

Despite many successful applications of CNNs, automated deep learning methods for localizing and classifying multicomponent, multipeak radio sources are still in their infancy. This has motivated our work in this paper. The winning solution (Dieleman, Willett & Dambre 2015) of the *galaxy challenge*[2] did utilize CNNs for accurate (>90 per cent) galaxy morphology classification . However, our work solves a very different problem from the *galaxy challenge*: we need to determine the number of radio sources in a given field of view (FoV) or *subject* (as is referred to within the RGZ project), each of which may contain multiple discrete source components. Such a determination is estimated from the combination of a radio continuum image and an IR map in the same position. Moreover, we need to localize each detected radio source with a bounding box, and finally to predict the morphology class for each detected source with some probability. Our problem is also different from radio continuum source finders, which typically involve identifying individual source components that are above a certain signal-to-noise threshold (Hancock et al. 2012). We need to group these components into one or more radio sources, and provide the morphology classification for each radio source.

The CNN method developed in Aniyan & Thorat (2017) accurately classifies a FIRST (Faint Images of the Radio Sky at Twenty-centimeters) radio source into FR and bent-tailed (BT) morphology classes. Although CLARAN is closely related to Aniyan & Thorat (2017), our research problem and method differ from Aniyan & Thorat (2017). CLARAN performs two tasks – *source identification* in a given field and *morphology classification* for each identified source. These two tasks address very different issues, and CLARAN is trained to solve both tasks simultaneously in a single, end-to-end training pipeline. During testing, CLARAN finds both compact and extended radio sources in all possible locations on an image, and classifies each one of them into some morphology. In contrast, the Aniyan & Thorat (2017) CNN classifier is trained to perform morphology classification *only*. As such the input image is cut out from the main image during pre-processing, and is *centred* at a known, given source. Moreover, while both CLARAN and Aniyan & Thorat (2017) use radio images, CLARAN can also use IR signals to significantly improve classification performance as shown in Section 5. The ability to integrate multiwavelength data sets for automated

[1] https://github.com/chenwuperth/rgz_rcnn/

[2] https://kaggle.com/c/galaxy-zoo-the-galaxy-challenge

source identification and morphology classification is unique to CLARAN.

2.1 CLARAN overview

In this work, we use Faster R-CNN model (Ren et al. 2017) as the basis to develop CLARAN for identifying multicomponent/peak radio sources from DR1. This is because Faster R-CNN is intuitive to understand, flexible to augment, and most importantly, offers optimal trade-offs between robust accuracy and execution latency (Huang et al. 2017). As a result, CLARAN includes an end-to-end data pipeline that enables fast identification and classification of radio sources with a mean Average Precision[3] (mAP, which is formally defined in Section 5.2) of 83.6 per cent and an empirical accuracy above 90 per cent. In particular, we make several contributions to deep learning-based methods for RGZ:

(i) We develop and evaluate several methods to combine radio emission and near-IR maps for source identification. This paves the way for future work on optimal (e.g. adaptive, learning-based) integration of multiwavelength data sets for automated source-matching and identification.

(ii) We tailor and fine-tune the Faster R-CNN (Ren et al. 2017) – a state-of-the-art object detection deep learning model – for effective radio source detection. To the best of our knowledge, latest research in object detection and computer vision has not yet been explored and utilized for radio source identification.

(iii) We augment the Faster R-CNN model by replacing its Region-of-Interest (RoI) cropping layer (`RoI pooling`) with differentiable affine transformations (`ST pooling`) based on the Spatial Transformer Network (STN, Jaderberg et al. 2015). Compared to the original Faster R-CNN model, training CLARAN becomes truly end to end – all training errors are accounted for by the learning model within a single data pipeline.

(iv) We develop a transfer learning (Yosinski et al. 2014; Ackermann et al. 2018) strategy – loading weights pre-trained on the ImageNet (Deng et al. 2009) data set and selectively controlling low-level convolutional kernels – to significantly accelerate the training error convergence.

(v) We demonstrate that CLARAN can distinguish between six distinct classes of radio source morphologies using both machine learning metrics and empirical accuracy evaluation performed by radio astronomers.

(vi) We evaluate CLARAN's scalability by showing its ability to identify radio sources with plausible classifications when the angular size in each direction of its input field is five times greater than what is available in the training set.

Taken together, our study provides an excellent starting platform for developing future machine learning-based methods for wide-area radio continuum surveys.

3 USING RADIO GALAXY ZOO CLASSIFICATIONS

The citizen science project RGZ obtains visual identification of radio sources from over 12 000 volunteers, who have collectively completed over two million classifications to date. Upon completion, RGZ will result in a catalogue of associated radio components and cross-matched host galaxies for over 170 thousand radio sources from the FIRST (Becker, White & Helfand 1995) survey and over 2000 sources from the Australia Telescope Large Area Survey (Norris et al. 2006). Currently, the cross-identification of extended radio sources and sources with disconnected radio lobes is through the visual inspection of radio sky maps with near-IR maps from the WISE (Wide-field Infrared Survey Explorer) telescope (Wright et al. 2010). Therefore, the method of crowd sourcing is used in RGZ to create one of the largest catalogues of extended radio galaxies with associated source components and host galaxy identifications.

[3]It should be noted that *precision* here differs from the definition (Bevington & Robinson 2003) in physical sciences.

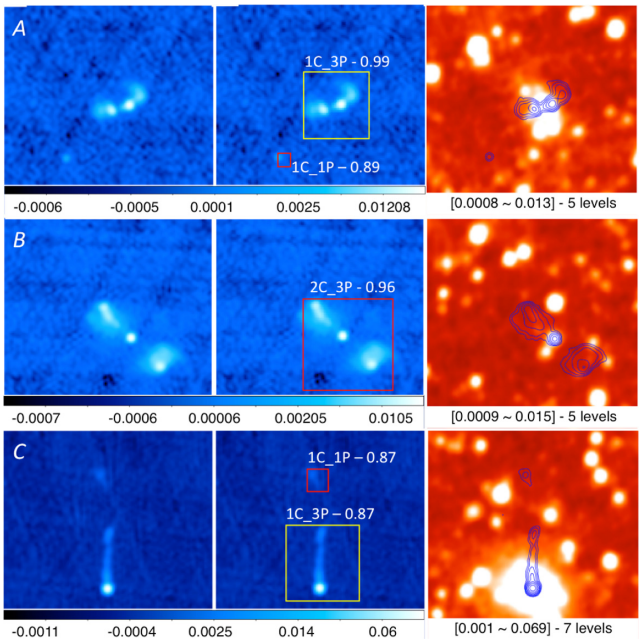


Figure 1. Three classification examples (A, B, and C) on RGZ subjects – each of them 3 arcmin × 3 arcmin in size – `FIRST J081700.6 + 571626`, `FIRST J070822.2 + 414905`, and `FIRST J083915.7 + 285125`. The first column shows the FIRST radio emission. The second column shows the CLARAN output – a box encompassing each identified source, and its morphology is labelled as ***i*C_*j*P**, where *i* and *j* denote the number of radio components and the number of radio peaks, respectively. Each morphology label is associated with a score between 0 and 1, indicating the probability of the quoted morphology class. The first two columns share the same colour bar at the bottom, denoting flux density values in Jy beam^{-1}. The last column shows the corresponding WISE IR image overlaid with 5σ radio contours. The contour levels (Jy beam^{-1}) are shown at the bottom of each IR image.

3.1 Classification examples

Before discussing the data set used for this study, we first present some classification examples shown in Fig. 1. Given a pair of FIRST and WISE images, CLARAN directly outputs the following in approximately 200 ms when measured on a single Tesla K40c GPU with 12GB GPU memory.

(i) the location and size of each detected radio source shown as a bounding box predicted by CLARAN during testing,

(ii) the morphology *m* of each detected source labelled as '*i*C_*j*P', where *i* is the number of components, and *j* is the number of flux-density peaks, and

(iii) the probability (*P*-value) of *m* for each detected radio source.

Following the definitions from the RGZ project (Banfield et al. 2015) and (Wong et al., in preparation), each RGZ *subject* is a 3 ar-

cmin by 3 arcmin FoV inspected by the citizen scientists, and the term *component* refers to discrete individual radio source components identified at the 4σ flux-density threshold level, and the term *peak* refers to the number of resolved peaks that are identifiable within each class of objects. For example, a double-lobed radio galaxy with small angular extent and no radio core may be identified as a source with one component-two peaks (1C-2P) or a two component-two peaks (2C-2P) if the two lobes appear disconnected in the radio image.

In example A of Fig. 1, CLARAN correctly identifies two radio sources – the large source has one component with three peaks, and the small one has one component with one peak. Both detections are given probabilities (0.99 and 0.89) much higher than 0.8. This example shows CLARAN is able to identify sources at different scales in the same image. In example B, CLARAN correctly locates a source with two radio components and three peaks (as per DR1) with a probability of 0.96. This example shows that CLARAN is able to identify extended sources.

In example C, CLARAN detects two independent sources, and assigns the same probability (0.87) to both of them. Although the real radio source is much larger based on the NRAO VLA Sky Survey (Condon et al. 1998), extending beyond the RGZ subject and including both red and yellow boxes as its internal components, CLARAN's prediction is still highly plausible considering its view is completely restricted within the 3 arcmin by 3 arcmin RGZ subject.

It should be noted that all radio and IR images in Fig. 1 are taken from the testing set (cf. Section 3.3), which CLARAN does not see during training.

3.2 Consensus level

We use two criteria to select fields from DR1 in order to create the training set and the testing set for CLARAN. First, for each selected subject *f*, we ensure *all* radio sources within *f* have a user-weighted consensus level (CL) no less than 0.6. CL measures the relative agreement levels of classification among citizen scientists and is defined in Banfield et al. (2015) as the largest fraction of the total classifications for a radio source that have been agreed upon. This is to ensure most radio sources exposed to CLARAN are morphologically human-resolvable.

Second, we ensure *every* radio source within *f* has fewer than four components and four peaks. This is because radio sources that (1) have a CL $\geq$0.6 and (2) have more than three components or peaks are rare as shown in Table 1. Inclusion of these sources into our study leads to highly unbalanced training and testing sets. Although dealing with unbalanced data sets is an ongoing machine learning research topic (He & Garcia 2009), in this paper we focus solely on the main demographic of multicomponent/peak sources, and leave for future work the issue of tackling unbalanced data sets with rarer sources.

Upon applying the above two selection criteria on DR1, we obtain a data set *E* that has 10 744 RGZ subjects. Fig. 2 shows the CL distribution of sources in *E* across the six morphology classes. Most one-component sources have high CL (with medians of 1C-1P and 1C-3P reaching the maximum CL value of 1.0) due to their relative simplicity. In particular, 1133 out of 1412 (80 per cent) 1C-3P sources have CLs equal to 1.0, which explains why its box in Fig. 2 is collapsed to a line when the first and third quartiles are both 1.0. 1C-2P has a slightly lower median CL (0.98) than that of 1C-1P or 1C-3P, but its third quartile also reaches 1.0. On the other hand, multicomponent/peak sources have much lower CLs in general. For example, most CLs of both 2C-2P and 2C-3P are distributed between 0.69 and 0.85 with 0.76 as their medians. CLs of 3C-3P sources have a similar median of 0.73 and a distribution between 0.66 and 0.81. Although CLs vary between these two groups of single-/multicomponent sources, reaching consensus naturally becomes harder with increasing morphological complexity associated with multicomponent sources. Given the above reasons we define the morphology classes listed in Table 3 as *ground-truth morphology* for both training and testing.

Table 1. The number of DR1 radio sources (CL $\geq$0.6) for each morphology class. The number of components and peaks for each source in this table is determined by RGZ DR1. Sources with more than three components/peaks are rare, and are excluded from this study to avoid unbalanced data sets. Sources with a morphology in the bold face (i.e. 1C-1P, 1C-2P, 1C-3P, 2C-2P, 2C-3P, and 3C-3P) are included in the training and testing sets for this study.

Morph	Count	Morph	Count	Morph	Count
1C-1P	49 766	2C-5P	36	4C-6P	7
1C-2P	14 242	2C-6P	7	4C-7P	5
1C-3P	1412	2C-7P	2	5C-5P	28
1C-4P	191	**3C-3P**	1347	5C-6P	11
1C-5P	28	3C-4P	163	5C-7P	1
1C-6P	12	3C-5P	20	6C-6P	2
1C-7P	3	3C-6P	13	6C-7P	1
2C-2P	9772	3C-7P	2	7C-7P	2
2C-3P	1220	4C-4P	99	7C-10P	1
2C-4P	181	4C-5P	18		

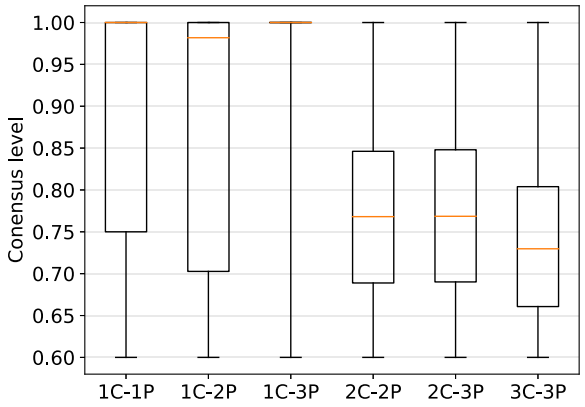


Figure 2. The distribution of the CL across six morphology classes in the data set that consists of 10 744 RGZ subjects selected from DR1. The whiskers above and below the box represent the maximum and minimum CL (fixed at 0.6 by the first criterion). The box itself spans the third and the first quartile CL. Note that since 80 per cent of 1C-3P sources have a CL of 1.0, its box is reduced to a single horizontal line when its interquarter range becomes 0. The horizontal (orange) line inside each box is the median.

3.3 Training and testing sets

We randomly split the data set *E* described in Section 3.2 into two subsets – the training set that contains 6141 subjects, and the testing set that contains 4603 subjects. Their basic properties are summarized in Table 2. Table 3 shows the morphology distribution of radio sources across six combinations of components and peaks. Although the number of 1C_1P sources is far greater than sources of other morphology classes in Table 3, the evaluation in Section 5 will show that CLARAN is not biased towards 1C_1P sources.

Table 2. Basic properties of the training and testing data sets used by CLARAN. One subject may contain multiple sources. One source may contain multiple components and radio peaks.

Set	Subjects	Sources	Components	Peaks
Training	6141	6978	9566	12 441
Testing	4603	4858	7397	9885
Total	10 744	11 836	16 963	22 326

Table 3. Number of radio sources (CL ≥0.6) for each morphology class in the training and testing data sets. A morphology class is represented as a combination of the number of components *C* and the number of peaks *P*. CL is discussed in Section 3.2 and further illustrated in Table 1.

Set	1C-1P	1C-2P	1C-3P	2C-2P	2C-3P	3C-3P
Training	3518	810	728	647	609	666
Testing	1782	521	684	604	599	668
Total	5300	1331	1412	1251	1208	1334

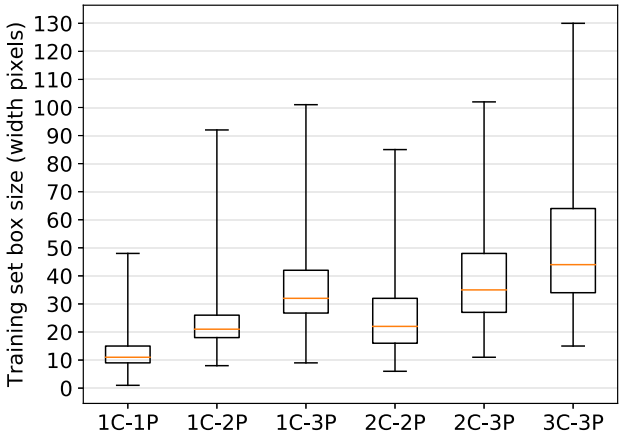


Figure 3. The distribution of bounding box sizes (width or height) in the training set for each morphology class. Note that the FIRST image pixel size is 1.375 arcsec, therefore the 3 arcmin × 3 arcmin angular size of each subject corresponds to 132 × 132 pixels, which sets the maximum possible value of the box size.

To generate the *ground-truth location* – both *location* and *size* of each known source within a given subject – we produce a square bounding box for each source based on its physical attributes defined in the RGZ data set. We use its central location `RA` and `DEC` as the box centre, and calculate the sky coordinates S_c of the box's four corners using the RGZ DR1 `max_angular_extent` parameter, which is an estimate of the source's angular size for all RGZ consensus sources as detailed in Banfield et al. (2015) and Wong et al. (in preparation). We then convert S_c into pixel coordinates P_c that can be processed by imaging software libraries. An extra step is taken to ensure the first element of P_c represents the top left corner as required by formats such as PNG or JPEG rather than the bottom left corner as in the FITS format.

Fig. 3 shows the size distribution of generated ground-truth boxes (i.e. radio sources) in the training set. The median size of the box appears positively correlated with the number of peaks, and if two sources have the same number of peaks, the one with more components has a slightly bigger size. Several extraordinarily large three-component sources almost cover the entire image.

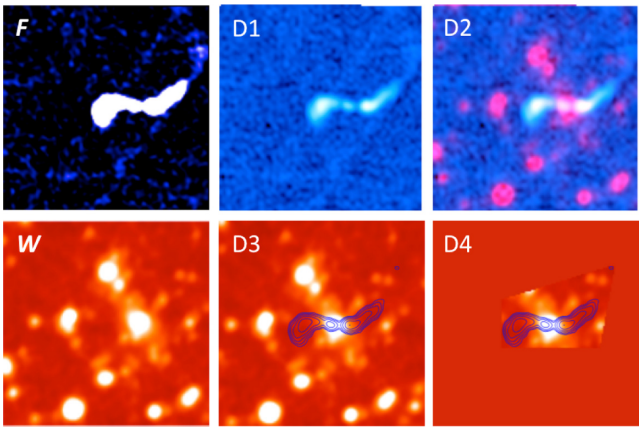


Figure 4. Based on the input FIRST image `FIRST J014110.8 + 121353`, examples of derived data sets are shown as D1, D2, D3, and D4. These maps are discussed in more detail in Section 3.4.

3.4 Derived data sets

The original RGZ data set contains FIRST radio images (in both FITS and PNG formats) and WISE IR PNG images. While the beam size of the FIRST survey is 5 arcsec, the size of each FITS pixel is about 1.375 arcsec. Therefore, the angular size of a 132 by 132 pixel RGZ subject is ∼ 3 arcmin × 3 arcmin. An example RGZ subject with the radio source `FIRST J014110.8 + 121353` is shown as a PNG image *F* in Fig. 4, and its WISE IR counterpart is shown as image *W* underneath *F*. Note that *F* is exported from the original FITS format as a three-channel (RGB) image under the 'cool' colourmap using DS9 (Joye & Mandel 2003). To effectively train CLARAN, we derive four additional data sets – D1, D2, D3, and D4 – from *F* and *W*. While both *F* and D1 display radio emission only, *F* uses the `DS9 linear-zscale` scale to represent flux values in the PNG format, whereas D1 uses the `DS9 log-min-max` scale. The rationale of creating D1 is to reveal the internal structures, but potentially at the cost of exposing more background noise. In this example, three separate radio peaks can be identified in D1 by eye but they appear blended together in *F*. It should be noted that training and testing on data sets F or D1 do not involve any IR images.

Similar to D1, D2 also uses the `DS9 log-min-max` scale. However, it increases the intensity of D1's red channel by corresponding pixel values in *W* while keeping D1's blue and green channels unchanged. This essentially overlays IR sources as red blobs on top of radio sources. The intention is to let CLARAN learn interaction patterns between the host galaxy (if detected in WISE) and its surrounding radio emission. D3 aims to achieve the same goal but operates in the opposite direction. It generates 5σ contours[4] based on surface brightness as recorded in the FIRST FITS file, and then overlays the radio contours on top of *W*. The RGZ Web user interface allows citizen scientists to transition between *F* and D3 (with a different level of sigma and contour colours) via a slider. Detailed descriptions of the RGZ interface can be found in Banfield et al. (2015).

[4]Unlike the RGZ Web interface which uses 4σ contours, we selected 5σ to reduce potential contamination from noise artefacts that are present in some fields.

We notice that there are numerous IR sources in W that are not directly related to the overlaid radio contours/sources. Their existence may mislead CLARAN to learn patterns from noise rather than features. To alleviate this issue, D4 generates a convex hull[5] over (sample points on) all radio contours in D3. The convex hull here denotes the union area enclosed by all radio contours on the IR image. For each channel c, D4 masks pixels outside the convex hull with the mean pixel value of c over all images in the training set. As a result, we remove all the IR signals that do not fall within the convex hull. Since the convex hull covers all radio contours, it should expose sufficient IR signals to capture the interplay between all radio sources/components. However, this cannot deal with certain special cases where a host galaxy is situated outside the union area formed by all radio source components within a subject. Such examples include remnant radio galaxies (there is no core) or there are faint, compact, separate (i.e. disconnected) lobes on opposite sides of WISE objects in the RGZ subjects. For these cases, D3 is perhaps more appropriate. Future research should investigate more optimal and generalizable data fusion techniques that, for example, have the advantages of both D3 and D4.

4 DATA PIPELINE

In this section, we introduce our dual-task, end-to-end data pipeline based on the Faster R-CNN method. By dual task, we mean the pipeline trains a detector to learn two separate tasks – *localization* and *recognition*. While both tasks share the same input features derived from the convolutional layer, the learning outcome of the first task will directly affect the learning performance of the second task. By end to end, we mean the entire training pipeline has only a single step of optimization, and the two tasks are trained simultaneously in a single training iteration. It also means little human involvement is needed for deriving hand-crafted features, and feature extraction is driven primarily by convolutional kernels learned from training sets rather than prior assumptions imposed by experts. Fig. 5 shows the data pipeline during the training stage, which we explain in detail below.

4.1 Pre-processing

In the first phase, three pre-processing operations – zero centring, size scaling, and horizontal flipping – are performed on-the-fly in a streaming mode on each input image.

Zero-centring involves (1) calculating the mean μ_C for each channel C across the entire training set, and (2) subtracting μ_C from each pixel of C in a given input image I. Since the subsequent convolutional filters are also initialized as truncated Gaussians centred at zero with a small standard deviation (0.01 in our training pipeline), filter response R from I is also zero-centred with a small variance. R is then transformed by the subsequent Rectified Linear Unit (ReLU, Nair & Hinton 2010) activation function defined as $A(x) = \max(0, x)$ to output the feature map. It has been reported (Krizhevsky et al. 2012) that ReLU, while simple and efficient to compute, accelerates the convergence of the optimization procedure such as the Stochastic Gradient Descent (SGD) by a factor of six. Moreover, it often results in superior solutions (Glorot, Bordes & Bengio 2011; Zeiler et al. 2013) than more traditional, sigmoid-like activation functions. During SGD, if all Rs are closely centred around zero, given a fixed pixel p, it is highly likely p in some R becomes positive to activate

[5] http://mathworld.wolfram.com/ConvexHull.html

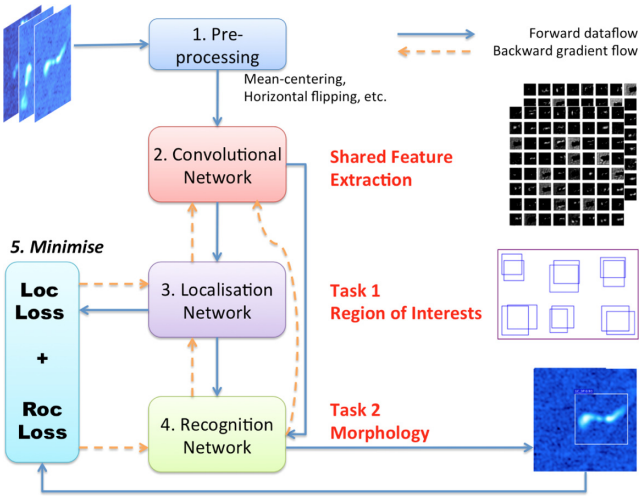


Figure 5. The end-to-end training pipeline that learns two related tasks simultaneously. The solid arrow denotes forward dataflow, in which a list L_f of parametrized functions are computed consecutively on each image batch. The output from L_f, known as 'prediction', is fed to the loss function (step 5) to calculate the error between ground truths and predictions. The error is converted to the global gradient, and propagated (via local gradient updates) backward to each function in L_f so that they can adjust their parameters to reduce the errors. The alteration of forward dataflow and backward gradient flow is repeated for each image batch, iteratively minimizing the loss function until the error converges below a threshold.

ReLU (for non-zero gradient descent), which will be less likely without zero-centring.

The largest receptive field[6] of a neuron in the last shared convolutional feature map is 228. Fig. 3 shows that the median box size of 3C_3P is slightly below 50. Therefore, we scale up the image size by a factor of $228/50 = 4.56$ to match the median box size to the final receptive field size. This involves increasing the height and width of the (fused) image from 132×132 pixels to 600×600 pixels using the bilinear interpolation. Moreover, we scale up coordinates of each ground-truth box by the same factor 4.56. It should be noted that scaling up the image size does not scale pixel intensities, which is a useful pre-processing technique (Stark et al. 2018) that we will explore in our future work for CLARAN.

During training, we use horizontal flipping to create a symmetric counterpart for a given input image I by appending an extra image I' that reverses the pixels order along the horizontal axis of I. This allows CLARAN to expect different source orientations other than provided in the original training set. We also create horizontally flipped ground-truth boxes to match the flipped image I'.

4.2 Convolutional network

The Convolutional Network (*ConvNet*) – including layers 1–17 in Table 4 – performs feature extraction in order to produce feature maps shared by both tasks and their associated networks. The basic two-dimensional convolution operation at each layer can be

[6] Section 4.2 describes the concept of *receptive field* and equation (2) defines its calculation.

Table 4. The Faster R-CNN model used by CLARAN, which consists of three networks – the *ConvNet* (layers 1–17), the *LocNet* (layers 18–22), and the *RecNet* (layers 23–29). Functions in *ConvNet* are either convolution operations (e.g. `Conv1_1`) or max pooling operations (e.g. `MaxPool_1`). Functions in *LocNet* and *RecNet* are explained in Sections 4.3 and 4.4, respectively. The Filter/Input tensor size in *ConvNet* refers to the convolutional/pooling filter size, whose first three dimensions (left to right) are the height, width, and depth of the filter. For convolutional filters, the fourth dimension denotes the number of filters. The Output tensor size in *ConvNet* refers to the height, width, and depth of the output feature map. Convolution operations in *LocNet* – `RPN_Conv`, `Anchor_Cls_Conv`, and `Anchor_Reg_Conv` – also have the same four-dimension filter sizes. Input and output tensor sizes for other functions are explained in Sections 4.3 and 4.4. All activations associated with convolution and dense-layer functions (i.e. `FC_6` and `FC_7`) are ReLU. The model in total consists of 136 777 443 'trainable' parameters, which are summed over all rows of the last column.

Layer	Function	Filter/Input tensor size	Activation	Stride	Output tensor size	Number of parameters
0	Input	$600 \times 600 \times 3$	–	–	–	0
1	Conv1_1	$3 \times 3 \times 3 \times 64$	ReLU	1	$600 \times 600 \times 64$	1,728
2	Conv1_2	$3 \times 3 \times 64 \times 64$	ReLU	1	$600 \times 600 \times 64$	36,864
3	MaxPool_1	$2 \times 2 \times 64$	–	2	$300 \times 300 \times 64$	0
4	Conv2_1	$3 \times 3 \times 64 \times 128$	ReLU	1	$300 \times 300 \times 128$	73,728
5	Conv2_2	$3 \times 3 \times 128 \times 128$	ReLU	1	$300 \times 300 \times 128$	147,456
6	MaxPool_2	$2 \times 2 \times 128$	–	2	$150 \times 150 \times 128$	0
7	Conv3_1	$3 \times 3 \times 128 \times 256$	ReLU	1	$150 \times 150 \times 256$	294,912
8	Conv3_2	$3 \times 3 \times 256 \times 256$	ReLU	1	$150 \times 150 \times 256$	589,824
9	Conv3_3	$3 \times 3 \times 256 \times 256$	ReLU	1	$150 \times 150 \times 256$	589,824
10	MaxPool_3	$2 \times 2 \times 256$	–	2	$75 \times 75 \times 256$	0
11	Conv4_1	$3 \times 3 \times 256 \times 512$	ReLU	1	$75 \times 75 \times 512$	1,179,648
12	Conv4_2	$3 \times 3 \times 512 \times 512$	ReLU	1	$75 \times 75 \times 512$	2,359,296
13	Conv4_3	$3 \times 3 \times 512 \times 512$	ReLU	1	$75 \times 75 \times 512$	2,359,296
14	MaxPool_4	$2 \times 2 \times 512$	–	2	$37 \times 37 \times 512$	0
15	Conv5_1	$3 \times 3 \times 512 \times 512$	ReLU	1	$37 \times 37 \times 512$	2,359,296
16	Conv5_2	$3 \times 3 \times 512 \times 512$	ReLU	1	$37 \times 37 \times 512$	2,359,296
17	Conv5_3	$3 \times 3 \times 512 \times 512$	ReLU	1	$37 \times 37 \times 512$	2,359,296
18	RPN_Conv	$3 \times 3 \times 512 \times 512$	ReLU	1	$512 \times 37 \times 37$	2,359,296
19	Anchor_Cls_Conv	$1 \times 1 \times 512 \times 12$	–	1	$12 \times 37 \times 37$	6,144
	Anchor_Cls_Conv_RS	$12 \times 37 \times 37$	–	–	$(6 \times 37) \times 37 \times 2$	0
20	Anchor_Cls_Softmax	$(6 \times 37) \times 37 \times 2$	–	–	$(6 \times 37) \times 37 \times 2$	0
	Anchor_Cls_Softmax_RS	$(6 \times 37) \times 37 \times 2$	–	–	$37 \times 37 \times 12$	0
20	Anchor_Target	12×37^2, gt_box $\times$ 5	–	–	$37^2 \times 12$, $37^2 \times 24$	0
19	Anchor_Reg_Conv	$1 \times 1 \times 512 \times 24$	–	1	$24 \times 37 \times 37$	12,288
21	RoI_Proposal	$37^2 \times 12$, 24×37^2	–	–	NMS_TopN $\times$ (4 + 1)	0
22	RoI_Proposal_Target	NMS_TopN $\times$ 5, gt_box $\times$ 5	–	–	RoI_batch $\times$ 1, RoI_batch $\times$ 28	0
23	ST_RoI_Pool	$37 \times 37 \times 512$, RoI_batch $\times$ 5	–	–	RoI_batch $\times 7 \times 7 \times 512$	0
24	FC_6	RoI_batch $\times 7 \times 7 \times 512$	ReLU	–	RoI_batch $\times$ 4096	102,764,544
25	DropOut_6	RoI_batch $\times$ 4096	–	–		
	RoI_batch $\times$ 4096	0				
26	FC_7	RoI_batch $\times$ 4096	ReLU	–	RoI_batch $\times$ 4096	16,781,312
27	DropOut_7	RoI_batch $\times$ 4096	–	–	RoI_batch $\times$ 4096	0
28	FC_Cls_Score	RoI_batch $\times$ 4096	–	–	RoI_batch $\times$ 7	28,679
28	FC_Reg_Pred	RoI_batch $\times$ 4096	–	–	RoI_batch $\times$ (7 $\times$ 4)	114,716
29	Cls_SoftMax	RoI_batch $\times$ 7	–	–	RoI_batch $\times$ 7	0

expressed as:

$$Y(m,n) = A\left(\sum_{k=1}^{C}\sum_{j=1}^{W}\sum_{i=1}^{H} X(k,i,j)K(k,m-i,n-j) + B\right) \quad (1)$$

In equation (1) X is an input image or an intermediary tensor with C planes (or 'channels' for RGB images), height H, and width W. Y is the output of the convolution, i.e. the *feature map*. $Y(m, n)$ denotes Y's value at row m and column n. K is a centre-originated kernel with channel C, height and width s, and $K(k, a, b) = 0$ if $|a|$ or $|b| > \frac{s}{2}$. Note that a feature map of one convolutional layer becomes the input (i.e. X) of the next convolutional layer. B is the *bias* tensor that has the same dimensions as the feature map Y. A is the element-wise ReLU activation function. Only K and B have learnable parameters that are updated during backpropagation through SGD.

We use the first 17 layers (13 weight layers and four pooling layers) from the VGG-16 (Configuration D) network (Simonyan & Zisserman 2015) as the architecture of the ConvNet. This is shown in Table 4 from layers 1 to 17). Compared to other ConvNet, a neuron in a VGG-16 convolution feature map has a smaller local FoV – the *receptive field* (Hubel & Wiesel 1962) – a 3 $\times$ 3 region from its input layer. However, stacking multiple convolutional layers gradually increases the global receptive field – i.e. the region in the input image. Neurons in the final feature map (i.e. layer 17 in Table 4) has a receptive field of size 228 $\times$ 228 when k is set to 17

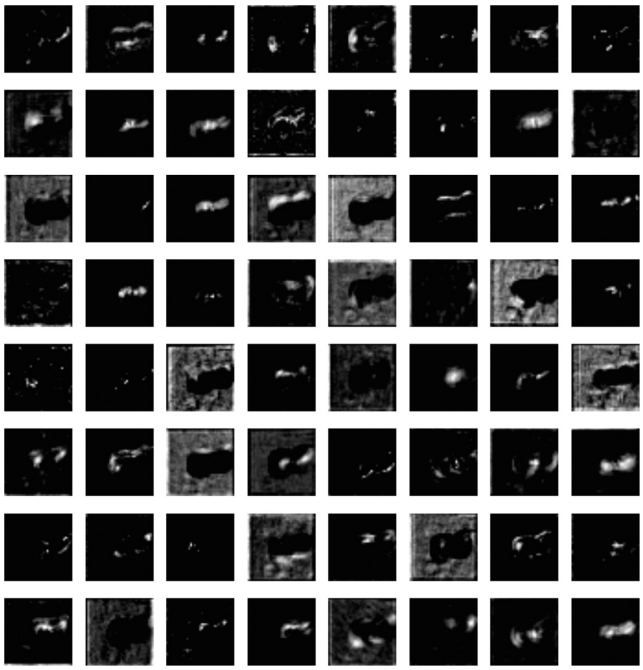


Figure 6. 64 feature maps produced by layer 18 (Table 4) given the input image `FIRST J014110.8 + 121353` (i.e. the example D1 data set in Fig. 4). The first 64 of the total 512 channels are shown, and each channel is visualized as a (37 by 37) grey-scale matrix, and white colours denote matrix elements with higher values. The feature map at row 3 column 5 (R3C5, channel 21) appears to be the outcome of cutting out the entire radio emission, revealing the overall contour of the source. R2C1 and R2C3 (channels 9 and 11) appear to represent the top and bottom parts of the source respectively as if they were separated by a gap tilting along the direction of the jet. More interestingly and importantly, we always find similar features at the same channel for different input images. This shows that the convolution kernels have learned something intrinsic and constant across different subjects.

in equation (2):

$$r_k = r_{k-1} + \left[(f_k - 1) \times \prod_{i=1}^{k-1} s_i \right] \quad (2)$$

where r_{k-1} denotes the size of the receptive field of neurons at layer $k-1$, f_k is the filter width/height (the third column of Table 4) at layer k, and s_i is the stride of layer i (the fourth column of Table 4). More importantly, stacking increases the number of non-linear activations since each convolutional layer has its own ReLU non-linearities. It is these non-linearities that ultimately offer the network discriminative capabilities for feature extraction. It should be noted that the size S of the receptive field of a single neuron does not limit CLARAN from detecting sources larger than S. This is because a feature map consists of multiple neurons, which collectively can detect much larger objects. Fig. 6 shows feature maps produced by the last shared convolution network layer (i.e. layer 17 *Conv5_3* in Table 4). The features are extracted from the input image `FIRST J014110.8 + 121353` shown in Fig. 4. The extraction were performed after the completion of training, which consists of 80 000 iterations of forward computation and backward propagation in order to find optimal values for all the kernel weights in the ConvNet. Visual inspection reveals some resemblance between the input image and each one of the 64 feature maps that capture the shape of the radio jets. However, each feature map exposes distinct features produced by a different set of kernels, each of which has learned to find and match a unique set of patterns from its input tensors. Collectively, these feature maps provide input for the two tasks to learn.

The parameters in the 13 weight layers are essentially shared by all following layers starting from layer 18, and are learned jointly by both task 1 (*localization*) and task 2 (*recognition*). To initialize weights in layers 1–17, we load public VGG-16 model weights[7] pre-trained from the ImageNet (Russakovsky et al. 2015). We then freeze the weights of the first four convolutional layers (1, 2, 4, and 5) by assuming low-level features learned by these filters remain constant across different domains, and set free weights in higher layers in order for them to learn higher level structures and patterns unique to radio galaxy morphology. We choose these four layers because their neurons have relatively small receptive fields – 5×5, 6×6, 14×14, and 16×16 pixels on the scaled 600×600 pixel image – well suited to capture low-level, local features.[8] Compared to learning these weights from scratch, we find that using pre-trained weights significantly improves the detection performance given the same amount of training time.

4.3 Localization network

The localization network (`LocNet`) – layers 18–22 in Table 4 – is trained to propose a set $\boldsymbol{R}$ of RoI proposals (boxes) given a subject, and each RoI proposal $r \in \boldsymbol{R}$ represents a potential radio source.

4.3.1 Regional proposal network

The `LocNet` starts with a mini-network – the regional proposal network (RPN, Ren et al. 2017), which consists of two layers of three convolutional functions. Layer 18 slides 512 filters over Layer 17 `Conv5_3`. Each filter outputs a $[37 \times 37]$ matrix, and all filters in total produce a $[512 \times 37 \times 37]$ feature map – `RPN_Conv`. Reshaping it to $[37 \times 37 \times 512]$, we treat `RPN_Conv` as a grid of 37×37 pixels, and each pixel x_i (where $i = 1 \ldots 37^2$) has 512 values.

The first step of the RPN is to construct k *anchors*, which are boxes of different sizes and aspect ratios affixed at the centre of each x_i. These k anchors act as 'prior boxes', some of which have the potential to grow into RoI proposals. Since anchors are stationary and input invariant, they constitute a fixed reference grid to locate radio source candidates across the entire feature map in parallel. All that is left to figure out is *which* and *how* anchors could be shifted and scaled in order to become RoI proposals.

We set $k = 6$ to cover scales [1, 2, 4, 8, 16, 32] and aspect ratio [1.0]. Since the total number of *strides* on `Conv5_3` after four layers of 2×2 *max poolings*[9] (i.e. Layer 3, 6, 10, and 14 in Table 4) is $2^4 = 16$, the anchor sizes projected back on the 600×600 subject are [16, 32, 64, 128, 256, 512]. We keep the anchor aspect ratio to 1 since all ground-truth boxes are squares although the proposed RoI may not be fully square due to the spatial offset described later. As a result, `RPN_Conv` corresponds to a set $\boldsymbol{A}$ of $6 \times 37^2 = 8214$ anchors.

[7] http://www.deeplearningmodel.net/

[8] The next convolutional layer 7 has a receptive field of 32×32 pixels (thus 7×7 pixels on the original image), which equate to the first quartile size of 1C_1P sources, and therefore too large for low-level feature extraction.

[9] *Stride* controls the offset by which the convolutional filter shifts across the input tensor, and *max pooling* downsamples the input tensor by selecting the maximum pixel in every subregion convolved with the pooling filter.

For each anchor a of each x_i, layer 19 maps a to two vectors. `Anchor_Cls_Conv` transforms a into the *objectness* score $\boldsymbol{o_p}$ = [`bkg_score`, `source_score`]. `Anchor_Reg_Conv` transforms a into the *anchor-source offset* $\boldsymbol{d} = [d_1, d_2, d_3, d_4]$. Given anchor a's spatial extent (a_x, a_y, a_w, a_h), equation 3 (Girshick et al. 2014) takes $\boldsymbol{d}$ as input, and outputs the spatial extent – centre coordinates, width, and height – of the RoI proposal . Therefore, $\boldsymbol{d}$ essentially predicts how a ought to be shifted and scaled to become an RoI proposal – surrounding some source inside its bounding box.

$$S(\boldsymbol{d}\,; a) = \left(d_1 a_w + a_x,\ d_2 a_h + a_y,\ e^{d_3} a_w,\ e^{d_4} a_h\right) \tag{3}$$

Both transformations in layer 19 can be expressed by a fully connected layer, performing dot products between its weight vector $\boldsymbol{w}_j$ and x_i, where $j = 1 \ldots 6m$, and $|\boldsymbol{w}_j| = 512$. We let $m = 2$ for `Anchor_Cls_Conv` and $m = 4$ for `Anchor_Reg_Conv`. In practice, these two transformations are implemented using $6m$ filters of $1 \times 1 \times 512$ convolutions for improved performance and efficiency. This is shown in layer 19 (`Anchor_Cls_Conv` and `Anchor_Reg_Conv`) in Table 4.

To train `Anchor_Cls_Conv` and `Anchor_Reg_Conv`, the RPN relies on `Anchor_Target` to dynamically generate ground truths for each anchor $a \in \boldsymbol{A}$. The ground truth for the *objectness* score vector is a scalar $o_{\rm g}$, denoting a negative anchor by 0 or a positive by 1. It indicates whether a matches a nearby ground-truth box (generated in Section 3.3) b, and its quantity is determined by the intersection-over-union (IoU) overlap $\frac{a \cap b}{a \cup b}$. a is positive if either (1) it has an IoU higher than a threshold τ with any ground-truth boxes or (2) it has the highest IoU if no anchors are positive. We set τ to 0.7 as a reasonable balance between loose (e.g. 0.5) and tight (e.g. 0.9) overlap values. An anchor is negative if its highest IoU overlap (with some ground-truth box) is less than $1 - \tau$, i.e. 0.3 in our tests. Anchors that are neither positive nor negative are excluded from training. Random selection is used to ensure the total number of negative and positive anchors is equal to the batch size $B = 256$ for each subject. Moreover, efforts were made to keep the ratio between the positive and the negative roughly at 1 : 1 to avoid unbalanced training sets. The loss function for training `Anchor_Cls_Conv` against each batch is defined as:

$$L_{\rm ac} = \frac{1}{B} \sum_{i=1}^{B} \{-[\log\left(\texttt{softmax}(\boldsymbol{o_{p_i}})\right) \cdot \texttt{one\_hot}(o_{g_i})]\} \tag{4}$$

where function `softmax`$(\cdot)$ converts $\boldsymbol{o_{p_i}}$ into a probability distribution, and function `one_hot`$(\cdot)$ encodes the scalar o_{g_i} into a vector.

The ground truth for the predicted anchor-source offset vector $\boldsymbol{d}$ is calculated using the inverse of S defined as:

$$\begin{aligned} S^{-1}(\boldsymbol{b}\,; a) &= \left(\frac{b_x - a_x}{a_w},\ \frac{b_y - a_y}{a_h},\ \log\frac{b_w}{a_w},\ \log\frac{b_h}{a_h}\right) \\ &= (g_1, g_2, g_3, g_4) = \boldsymbol{g} \end{aligned} \tag{5}$$

Given anchor a and its spatial extent (a_x, a_y, a_w, a_h), equation (5) takes as input the spatial extent vector $\boldsymbol{b}$ of a ground-truth box b, with which a has the highest IoU among all ground-truth boxes, and outputs the true (actual) anchor-source offset $\boldsymbol{g} = [g_1, g_2, g_3, g_4]$. The loss function for training `Anchor_Reg_Conv` is defined as:

$$L_{\rm ar} = \frac{1}{|A'|} \sum_{i=1}^{|A'|} \left(\sum_{j=1}^{4} [\texttt{smooth\_L1}(d_j - g_j)\, o_{g_i}] \right) \tag{6}$$

in which $A' \subset \boldsymbol{A}$, and $|A'| = 5241$. $\boldsymbol{A} \setminus A'$ includes anchors that lie (partially) outside the subject, and function `smooth_L1` is a Huber loss (Huber et al. 1964).

4.3.2 RoI proposal

In the second step of `LocNet`, the `RoI_Proposal` layer shifts every anchor $a \in \boldsymbol{A}$ by $\boldsymbol{d}$ based on equation (3), yielding 6×37^2 candidate RoI proposals. After excluding unreasonably small candidates (i.e. less than 4×4 pixels in the subject), it sorts remaining proposals by their *source objectness* scores `softmax`$(\boldsymbol{o_p})$[1] in a descending order, and selects the top M proposals (M is a hyper-parameter set to 6000) for pruning using the Non-Maximum Suppression (NMS) algorithm (Neubeck & Van Gool 2006). Iterating over the sorted list of M proposals, NMS accepts a proposal p' with the highest *source objectness* score, then discards all subsequent proposals whose IoU overlap with p' is greater than a threshold (a hyper-parameter set to 0.7) and repeats the procedure with the remaining proposals until the end of the list. Finally, only the top P scoring proposals are kept after NMS pruning, where P is a hyper-parameter set to 2000 and 5 for training and testing, respectively.

During testing, each one of the five proposals $p \in P$ is directly fed to the recognition network (`RecNet`, cf. Section 4.4) to predict (1) the proposal-source offset $\boldsymbol{u}$, by which p ought to be shifted and scaled in order to become a nearby ground-truth box, and (2) the morphology class m (cf. Table 3) of p. However, to train `RecNet` to perform such prediction during training, each one of the 2000 $p \in P$ goes through the `RoI_Proposal_Target` layer, which aims to produce ground truths for both $\boldsymbol{u}$ and m. For each $p \in P$ and given a set T of ground-truth boxes associated with the subject, the ground-truth box $t \in T$ that has the highest IoU with p is the *target* of p. The ground truth of $\boldsymbol{u}$ for p is then calculated as:

$$\begin{aligned} S^{-1}(\boldsymbol{t}\,; p) &= \left(\frac{t_x - p_x}{p_w},\ \frac{t_y - p_y}{p_h},\ \log\frac{t_w}{p_w},\ \log\frac{t_h}{p_h}\right) \\ &= (q_1, q_2, q_3, q_4) = \boldsymbol{q} \end{aligned} \tag{7}$$

The ground truth of m is a scalar $v \in \{0 \ldots 6\}$ denoting six morphology classes (1–6) plus the background class (0). However, since each $t \in T$ contains a radio source with a given morphology defined in Table 3, v cannot possibly take the value of 0 to represent the background *target*. To address this, the Faster R-CNN model treats as *background* proposals the set $G \not\subset P$ of proposals whose IoUs with their targets are within the range of [0.1, 0.5], and the ground truth of m for each $g \in G$ is manually set to 0. Similarly, a proposal is *foreground* if its IoU with its *target* is greater than 0.5. Random selection is used to (1) adjust the number of *foreground* and *background* proposals such that the ratio between the two is approximately 1 : 3, and (2) to further reduce the total number of RoI proposals from 2000 to 128, thus $|P| = 128$. The output of the `RoI_Proposal_Target` layer – P, and the ground-truth $\boldsymbol{q}$ and v associated with each $p \in P$ – is fed to `RecNet` for training.

4.4 Recognition network

For each subject, `RecNet` accepts two inputs – (1) the feature map F produced by the convolution network layer `Conv5_3` and (2) the set of RoI proposals P produced by either the `RoI_Proposal` layer during testing or the `RoI_Proposal_Target` layer during training. For each $p \in P$, the first layer of `RecNet` – `ST_RoI_Pool` – crops the RoI r out of F based on p, and downsamples r into a feature map f of size $512 \times 7 \times 7$. The original Faster R-CNN

(Ren et al. 2017) study uses RoI pooling (Girshick 2015) for downsampling. It works by evenly partitioning each channel of r into a 7×7 grid of subsections, each of which has an approximate size $37/7 \times 37/7$, and max pooling the values from each subwindow to form a single channel of f. However, the issue with RoI pooling is that while it accepts both F and P as input during forward pass, only the gradient with respect to F is calculated during backpropagation via max pooling. The gradients with respect to P are completely ignored. In other words, training errors caused by P are not sufficiently accounted for, resulting in an approximate optimization solution at most. To overcome this limitation, we use two tensor operations defined in the STN (Jaderberg et al. 2015) to crop and downsample r – the affine transformation $\mathcal{T}_\theta$ and the bilinear sampling $\mathcal{B}$. Since $\mathcal{T}_\theta$ is differentiable with respect to P, and $\mathcal{B}$ is differentiable with respect to both F and the output of $\mathcal{T}_\theta$, the error gradients are able to flow back not only to F but also to coordinates of each $p \in P$. Given the coordinates $[x_1, x_2, y_1, y_2]$ of $p \in P$, the affine transformation is defined as:

$$\mathcal{T}_\theta(G_i) = \begin{bmatrix} \frac{x_2 - x_1}{w_\mathrm{F}} & 0 & \frac{x_1 + x_2 - w_\mathrm{F}}{w_\mathrm{F}} \\ 0 & \frac{y_2 - y_1}{h_\mathrm{F}} & \frac{y_1 + y_2 - h_\mathrm{F}}{h_\mathrm{F}} \end{bmatrix} \begin{pmatrix} u_i^f \\ v_i^f \\ 1 \end{pmatrix} = \begin{pmatrix} u_i^F \\ v_i^F \end{pmatrix} \tag{8}$$

where $w_\mathrm{F} = 37$ and $h_\mathrm{F} = 37$ are the width and height of F, and $G_i = (u_i^f, v_i^f) \forall i \in \{0, 1, \ldots, 7^2 - 1\}$ are coordinates of the regular grid on f, and (u_i^F, v_i^F) are coordinates of the sample points on F.

The output from `ST_RoI_Pool` is a set $\boldsymbol{R}$ of RoI feature maps of size $512 \times 7 \times 7$ and $|M| = 128$ and 5 for training and testing, respectively. The next fully connected layer `FC_6` reshapes $\boldsymbol{R}$ as a matrix $\boldsymbol{R'}$ of size $|M| \times 25,088$, and uses a weight matrix of size $25\,088 \times 4096$ to linearly transform $\boldsymbol{R'}$ into a $|M| \times 4096$ matrix F_1. During training, a dropout layer (Srivastava et al. 2014) `Dropout_6` is added such that for a given element el of F_1, `Dropout_6` either resets the value of el to 0 with a probability of $1 - k$ or scales up the value of el by a factor of $\frac{1}{k}$ with a probability of k, $0 \leq k \leq 1$. Compared to conventional regularization methods, dropout is more effective and computationally efficient to prevent overfitting for layers with a large number of parameters – 102 million weights in the case of `FC_6`. After dropout updates, F_1 is transformed by another fully connected layer `FC_7` followed by another dropout layer `Dropout_6`, producing a matrix F_2 of size $|M| \times 4096$. It should be noted that dropout layers – `Dropout_6` and `Dropout_7` – are only used during training, and are skipped during testing as shown in Fig. 7. Both `FC_6` and `FC_7` use ReLU as their internal activation function to output F_1 and F_2.

The first output of `RecNet` contains scores of each RoI $r \in \boldsymbol{R}$ against morphology classes defined in the first row of Table 3. To produce such output, a fully connected layer `FC_Cls_Score` takes F_2 as input, and produces as output an $|M| \times 7$ matrix F_3, whose value at row i and column j denotes the score of the ith RoI in $\boldsymbol{R}$ being an instance of class j, and $1 \leq i < |M|$, $0 \leq j \leq 6$. During training, F_3 is used as the input of the classification log-loss function `RoI_Cls_Loss` shown as the grey rectangle at the bottom of Fig. 7. The formal expression of `RoI_Cls_Loss` L_rc is defined as:

$$L_\mathrm{rc} = \frac{1}{|M|} \sum_{i=1}^{|M|} \{-[\log(\texttt{softmax}(F_3[i])) \cdot \texttt{one\_hot}(v_i)]\} \tag{9}$$

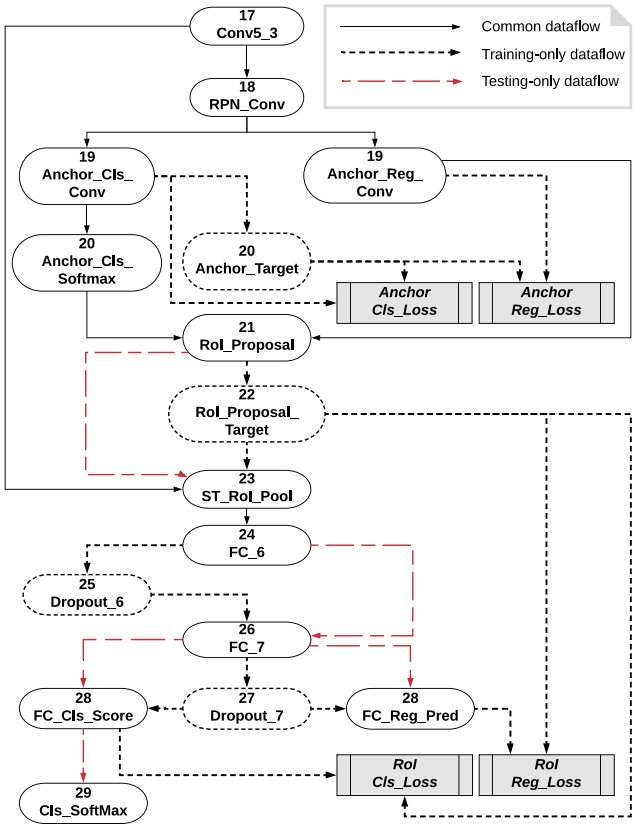


Figure 7. A dataflow diagram for `LocNet` and `RecNet`. Each ellipse represents a `Function` defined in the second column of Table 4. Solid ellipses appear in both training and testing, but dotted ones are used for training only. For example, `Anchor_Target` and `RoI_Proposal_Target` dynamically generate ground truths for training given a subject – i.e. positive and negative anchors in `Anchor_Target` or proposal-source offsets and morphology class for each proposal in `RoI_Proposal_Target`. These two operations are only used during training, and are removed during testing. Similarly, solid arrows, which denote the dataflow between two data transformations, appear in both training and testing, and dotted ones are used only for training, and dashed ones represent dataflows for testing only. The four grey rectangles denote the four loss functions – equations (4), (6), (10), and (9) – in a clockwise order. Since loss functions are minimized during training, dataflows that provide inputs to these functions are all dotted arrows.

where scalar $v_i \in \{0 \ldots 6\}$ denotes the ground-truth class for the ith RoI in $\boldsymbol{R}$, and is provided by the `RoI_Proposal_Target` layer as described in Section 4.3.2. The `softmax` function in the `Cls_SoftMax` layer converts the ith row of F_3 into a discrete probability distribution vector $\boldsymbol{d}$, whose jth element represents the probability of RoI i being an instance of class j. In practice, the morphology class $\hat{m}$ with the highest probability is often chosen as the output classification result.

The second output of `RecNet` contains the proposal-source offsets of each $r \in \boldsymbol{R}$ for each morphology class. To produce such output, the `FC_Reg_Pred` layer takes F_2 as input, and produces as output an $|M| \times 28$ matrix F_4, whose values at row ith and between columns $[4j, 4j + 4]$ denote the proposal-source offsets of the ith RoI for class j, and $1 \leq i < |M|$, $0 \leq j \leq 6$. During training, F_4 is used as the input of the regression loss function `RoI_Reg_Loss`

(the rectangle at the bottom right of Fig. 7), which is defined as:

$$L_{\rm rr} = \frac{1}{|M|}\sum_{i=1}^{|M|}([\texttt{smooth\_L1}(\boldsymbol{d_{ij}} - \boldsymbol{g_{ij}})]) \quad (10)$$

where $\boldsymbol{d_{ij}} = [d_x, d_y, d_w, d_h]$ is the predicted proposal-source offset of RoI i corresponding to its true morphology class j, $\boldsymbol{g_{ij}} = [g_x, g_y, g_w, g_h]$ is the ground-truth proposal-source offset of i for the same true class j, and `smooth_L1` is the Huber loss function (Huber et al. 1964).

5 QUANTIFYING CLASSIFICATION PRECISION

We implement the data pipeline described in Section 4 using TENSORFLOW (Abadi et al. 2016). Both training and testing require GPU resources, and we deploy the pipeline to run on both Tesla K40c (12 GB device RAM) and Tesla P100 (16 GB device RAM) GPUs. For training, we use the momentum optimizer to update network weights, and set the initial learning rate to 0.001 with a decay rate of 0.1 for every 50 000 iterations. The training speed is about 0.52 and 0.11 s per iteration on K40 and P100, respectively. Thus, a pipeline instructed to execute 80 000 iterations requires 3–12 h of training time on provisioned GPU resources. For testing, it takes the learned model 220 and 45 ms per subject on K40c and P100, respectively to generate detected radio sources, their associated morphology and probabilities.

5.1 Training error

The efficiency and effectiveness of the training pipeline is largely determined by the *training error*, which is the sum of the four losses defined in equations (4), (6), (9), and (10):

$$\text{Training error} = L_{\rm ac} + L_{\rm ar} + L_{\rm rc} + L_{\rm rr} \quad (11)$$

The goal of training is to reduce the training error on the training set using various optimization techniques without compromising the model generality on future unseen data sets. To examine the change of training error, we compare two learning curves in Fig. 8, where the Y-axis denotes training errors and the X-axis represents the number of iterations. As training proceeds on data set D4, the average training error becomes smaller in both cases, reduced from 0.35 to 0.05 for the bottom learning curve, and from 0.7 to 0.28 for the top curve. Both curves exhibit a sharp plunge within the first 5000 iterations, and turn into a more steady descent afterwards. The downwards trend appears to plateau out after 65 000–75 000 iterations for both curves, suggesting the model has reached its learning capacity given current network architecture and data sets.

Training errors in the bottom learning curve in Fig. 8 are significantly smaller than those in the top curve. The bottom learning curve was generated by the training process in which low-level (i.e. layer 1, 2, 4, and 5 in Table 4) convolutional kernels were set to read-only once loaded from the pre-trained VGG-16 model, and were never updated throughout training. The training process that produced the top curve, on the other hand, continuously updates these low-level kernels during training. Since these low-level kernels have been pre-trained using much larger data sets for an extended period of time (e.g. several weeks), we believe they capture features common enough to be shared across different domains.

Fig. 8 suggests that freezing these low-level kernels in effect reduces the training error with a much higher efficiency. This is because pre-trained low-level weights become fine-tuned and optimal

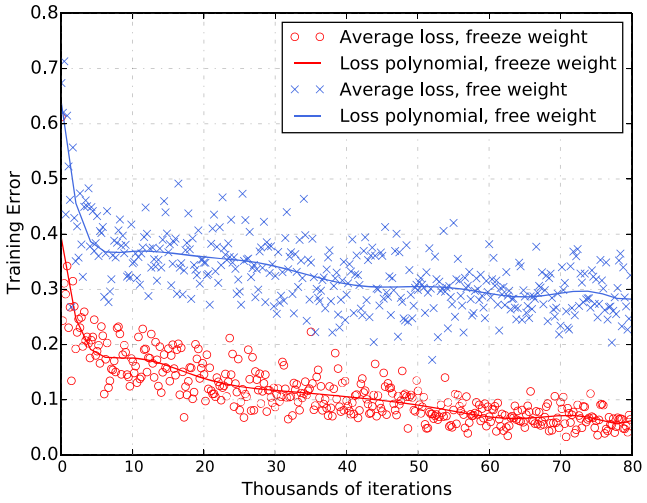


Figure 8. Learning curves (on data set D4) monitor the change of training losses (Y-axis) as the training progresses by some number of iterations (X-axis). The top part denotes the case where the low-level (i.e. layer 1, 2, 4, and 5 in Table 4) kernel weights ($N = 259\,716$) are trainable – i.e. free to be updated during the training process. The bottom part shows when those low-level kernels weights freeze and are thus not updated during training. The losses are sampled every 10 iterations during training, and are only shown every 200 iterations for visualization. However, both polynomials are plotted based on all collected loss samples.

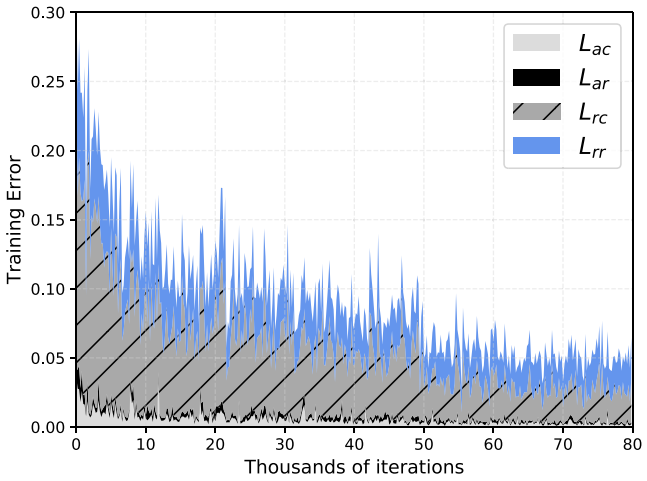
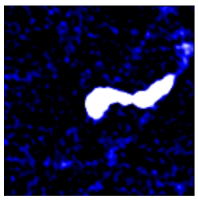
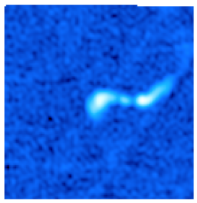
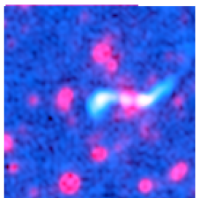
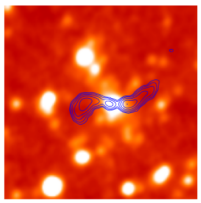
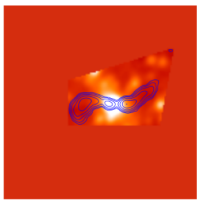


Figure 9. Training errors are decomposed into four losses (equation 11) stacked on top of one another every 10 iterations as the training progresses. Areas covered by dark diagonal lines denote the RoI classification loss $L_{\rm rc}$.

in extracting low-level features *common* to generic object detection tasks including those in CLARAN. If not retained during retraining (particularly given the high initial learning rate and different loss functions), they are subject to gradient updates much higher than those received towards the end of the ImageNet pre-training. Consequently, they quickly diverge from the current optimal region in the high-dimensional parameter space.

Since the training error defined in equation (11) is the sum of four loss terms, we visually breakdown the training error as a stack plot shown in Fig. 9. Initially, about 60 per cent of the training error was attributed to the RoI classification loss $L_{\rm rc}$. While the overall training error declines as training progresses, the portion of $L_{\rm rc}$ is gradually diminishing, reaching to 35 per cent in the end. On the other hand,

Table 5. Evaluation of five data pre-processing methods using AP and mAP. Each row represents APs achieved by all five methods for a given morphology class. The highest AP for each morphology class is highlighted in the bold face. Each column denotes APs achieved by a particular method over all six morphology classes and the overall mAP. Method D4 has achieved the highest mAP, highest APs for morphology 1C_1P and 2C_2P, and second highest AP for 3C_3P.

Methods	F	$D1$	$D2$	$D3$	$D4$
					
1C_1P	0.8087	0.8580	0.8242	0.8485	**0.8784**
1C_2P	0.6376	0.6882	0.6843	0.6746	**0.7074**
1C_3P	0.8250	0.8816	0.8561	0.8876	**0.8941**
2C_2P	0.7474	0.7014	0.7231	0.7983	**0.8200**
2C_3P	**0.8087**	0.7099	0.6989	0.8047	0.7916
3C_3P	0.7708	0.8636	0.8561	**0.9424**	0.9269
mean AP	78.5%	78.4%	77.4%	82.6%	**83.6%**

the portion of $L_{\rm rr}$ is increasing to above 55 per cent. This suggests that training of morphology classification is slightly more efficient than that of localization regression. We find that the correlation coefficients between Anchor errors ($L_{\rm ac}$ and $L_{\rm ar}$) and $L_{\rm rr}$ are slightly higher than those between Anchor errors ($L_{\rm ac}$ and $L_{\rm ar}$) and $L_{\rm rc}$, suggesting RoI regression is more sensitive to errors caused by the region proposal network. Moreover, there is a moderate positive correlation (0.508) between $L_{\rm rc}$ and $L_{\rm rr}$, since these two tasks share a large number of weights in the fully connected layers 24 and 26, which contain 87.4 per cent of the parameters stored in the model.

5.2 Testing metrics and evaluation

To evaluate CLARAN against the testing set, we use a single evaluation metric – the mAP. The Average Precision (AP) is a function of both *reliability* and *completeness*, which are referred to as *precision* and *recall*, respectively, in machine learning. *Precision* measures the fraction of identified sources that are correct according to the RGZ ground truth and *Recall* refers to the fraction of RGZ ground-truth radio sources that have been identified. Given a morphology class $m \in 1\ldots6$, let L_m denote a list of radio sources detected by CLARAN as 'class m sources' from all subjects in the testing set, and let T_m denote a set of radio sources that are truly of morphology m contained in the testing set. Sources in L_m are ranked by their morphology class probabilities (P-values) in a descending order. The AP_m for morphology class m is calculated as:

$$\mathrm{AP}_m = \frac{\sum_{k=1}^{|L_m|}(P(k) \times \mathrm{tp}(k))}{|T_m|} \tag{12}$$

where $\mathrm{tp}(k)$ is an indicator function equaling 1 if $L_m[k]$ is a true positive detection, 0 otherwise, and $P(k)$ denotes the *precision* calculated up to element $L_m[k]$:

$$P(k) = \frac{\sum_{i=1}^{k}(\mathrm{tp}(i))}{k}, \quad 1 \le k \le |L_m| \tag{13}$$

A detected source $K \in L_m$ in subject S is true positive if and only if the IoU (defined in Section 4.3.1) between K and some ground-truth sources of class m in S is greater than 0.5.

Finally, the mAP is calculated as:

$$\mathrm{mAP} = \frac{\sum_{m=1}^{6}(\mathrm{AP}_m)}{6} \tag{14}$$

We apply equations (12) and (14) to evaluate the testing set detection results produced by five different data pre-processing methods – F, D1, D2, D3, and D4 as discussed in Fig. 4. The result of both AP and mAP for each method is presented in Table 5.

The results of F and D1 – pure radio emission – are slightly better than D2, which simply places spatially aligned radio and IR planes in different channels of the input subject. This suggests that radio source detection from multiwavelength data sets requires different data fusion techniques than those used for object detection from common RGB images. We therefore explore several alternative data fusion methods, and found methods D3 and D4 have consistently achieved better AP and mAP than other methods. On the other hand, not all fusion methods worked as expected. For example, in one method, we prepend to the network a $1 \times 1 \times 3$ convolutional layer (Szegedy et al. 2015), which is then trained to learn optimal weighted averages of fluxes from different channels in the original subject input. However, this method is merely 0.5 per cent better than D2, achieving an mAP of 77.9 per cent. We suspect the reason D3 and D4 perform better is because their fusion method visually resembles the RGZ Web interface, through which citizen scientists have collectively produced the 'RGZ truth' for training CLARAN. However, we note that visual classification may not always reflect the 'true' ground truth as the accuracy of the classifications may be limited by the angular resolution, frequency, or sensitivity of the observations. However, the purpose of our work is to be able to replicate the accuracy standards set by visual classifications in an automated fashion.

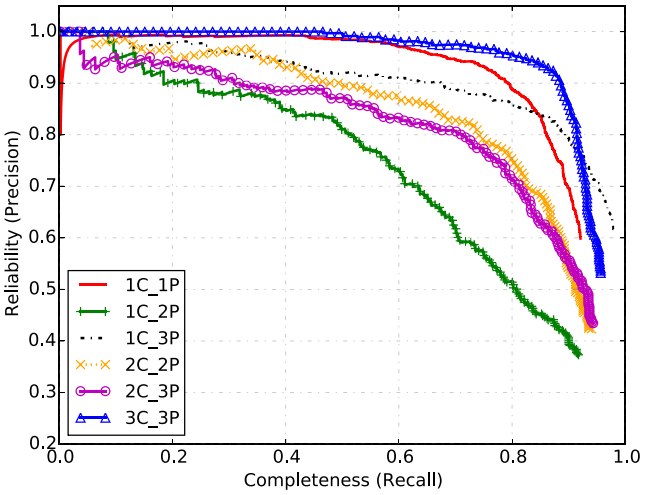


Figure 10. The trade-off between reliability and completeness is shown by the PR curves against the unseen test data set with 4603 subjects using the pre-processing method D4. Each morphology class has its own PR curve, which records the reliability (*Y*-axis) achieved by CLARAN (using method D4) at each level of completeness (*X*-axis). The area under a PR curve is known as AP, which has a discrete form expressed in equation (12). Therefore, the top right curves, with larger area beneath them, have greater APs.

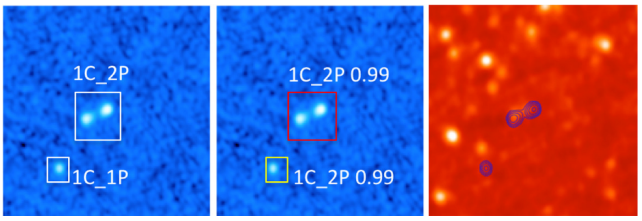


Figure 11. A 'misclassified' source 1C_2P (bottom left) in subject FIRST J131100.4 + 034608 selected from the training set. From left to right are: RGZ truth (white boxes), sources detected by CLARAN (coloured boxes), and 5σ radio contours overlaid on the IR map. The bottom left source has a high probability (99 per cent) of being 1C_2P, which should have been 1C_1P according to the RGZ truth. False positive detections such as this one (with a high *P*-value) will cause the sudden drop of the 1C_2P PR curve shown in Fig. 10.

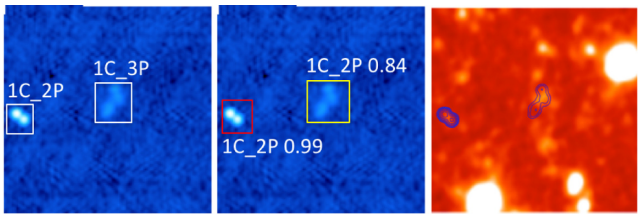


Figure 12. A 'misclassified' 1C_2P source in the middle of subject FIRST J110148.2 + 252746 with a relatively high probability of 84 per cent. According to the RGZ truth, it should have been 1C_3P. But this mistake is more likely due to differences in the con tour level between the RGZ DR1 pipeline and the CLARAN data preparation.

5.3 Reliability versus completeness

The Precision–Recall (PR) curves plotted in Fig. 10 shows how CLARAN deals with the trade-off between these two metrics for different morphology classes. In general, PR curves closer to the top right corner (e.g. 3C_3P) have better mAPs than those further away from it. The 3C_3P PR curve starts with a horizontal line (at the reliability level of 1.0) until the completeness level reaches 0.6. This suggests that, if we put all predicted 3C_3P sources into a list L sorted by P-values in descending order, and let C be the set of ground-truth 3C_3P sources in the testing set (where $|C| = 668$ as per Table 3), then 60 per cent of C ($N = 400$) are also the *first* 400 sources in L, and 80 per cent of C ($N = 534$) are in the *first* 561 elements of L.

In contrast, in the PR curve for 1C_2P, the reliability quickly drops immediately after 30 per cent of the true 1C_2P sources have been detected, and by the time the completeness reaches 80 per cent, nearly half of the detected 1C_2P sources are false positives. This is consistent with the relatively poor mAP results shown in Table 5. In particular, the wiggle section between Completeness 0.1 and 0.2 of the PR curve is caused by some top-ranked yet false positive 1C_2P detections. In general, false positives have lower P-values because most PR curves in Fig. 10 are smoothly bent downwards to the right.

To identify potential causes for this, we show several false positive 1C_2P examples taken from the training set. Fig. 11 shows CLARAN outputs for two sources: a true positive 1C_2P with a high P-value of 99 per cent at the centre, and a false positive 1C_2P at the lower left with an equally high probability. It appears that this source is slightly elongated, but it should be noted that 'ground-truth peaks' did not come from RGZ user consensus but were automatically produced by the RGZ DR1 pipeline. The false detection in Fig. 12 could be caused by the difference in the contour level (4σ) used in DR1 and that (5σ) used for training CLARAN. This difference may prevent CLARAN from distinguishing the two peaks at the top right. However, we find that laying 4σ contours to train CLARAN exposes more unrelated noise in general, jeopardizing the overall detection performance. Our tests show that the D4 method could only achieve an mAP of 78 per cent when using 4σ contours. These two examples show that resolving double peaks from a relatively small single-component source (1C_2P) poses challenges to CLARAN, which could potentially confuse a star-forming galaxy with an AGN. Identifying triple peaks from a double-component source (2C_3P) also appears challenging to CLARAN.

Although CLARAN does not agree with the RGZ truth in terms of the number of peaks for certain 1C_2P and 2C_3P sources, we hypothesize that CLARAN is able to correctly identify their components as exemplified in Figs 11 and 12. To verify this hypothesis, we re-organize sources in the testing set into three morphology classes based on their ground-truth 'number of components' regardless of their 'number of peaks'. We then recategorize sources detected by CLARAN from six classes (as in the first column of Table 5) into three classes based solely on their 'number of components'. For example, sources of classes 1C_1P, 1C_2P, and 1C_3P are merged into a single class denoted by 1C_[1P or 2P or 3P]. Finally, we use equations (12) and (14) to evaluate D3 and D4 against these three classes instead of the original six classes. The result is shown in Table 6. All metrics in Table 6 are higher than those in Table 5 (except for 3C_3P that remains unchanged), particularly for 1C_2P and 2C_3P. This indicates that CLARAN is able to produce correct components for most of the 1C_2P and 2C_3P sources, increasing overall mAPs by nearly 8 per cent for both D3 and D4. In practice, we can recover ground-truth peaks by rerunning the same peak calculation algorithm used in DR1 on each RoI detected by CLARAN.

Table 6. Evaluation of D3 and D4 based on a three-class scheme, in which only the ground truth 'number of components' is used to determine the classification of each radio source in the testing set.

Morphology class	D3	D4
1C_[1P or 2P or 3P]	0.8644	0.9054
2C_[2P or 3P]	0.8699	0.8946
3C_3P	0.9424	0.9269
mAP	89.2 per cent	90.9 per cent

Table 7. Evaluation of D3 and D4 in a 'small' (250 subjects) testing set *T*, in which each subject has at least 2 RGZ DR1 sources within its 3-arcmin by 3-arcmin FoV. The first column denotes the number of sources with the corresponding number of components.

Source count	Morphology	*D*3	*D*4
487	1C_	0.7452	0.8394
13	2C_	0.2800	0.3892
5	3C_	0.8850	0.2709
505	mAP	63.7 per cent	50.0 per cent

Since this paper focuses on the development and evaluation of a deep learning method, we leave for future work the optimal integration of CLARAN with other RGZ data reduction and analysis algorithms.

5.4 Multisource subjects

A key problem that RGZ aims to address is to distinguish multiple unrelated sources from multiple components of single sources. CLARAN demonstrates this capability in Fig. 1(A), and Figs 11 and 12 (regardless of peaks). However, a statistical measure is needed to quantify this capability. Since 94 per cent of the subjects in the testing set (4858 sources in 4603 subjects) have only one radio source, mAPs in Tables 5 and 6 do not effectively measure CLARAN's performance in separating multiple sources. Therefore, we create a 'small' data set *T*, which includes *every* subject in the testing set that has at least two sources. In total, *T* contains 505 such sources, excluding 4353 single-source subjects (i.e. 4353 sources) from the original testing set.

Table 7 presents mAPs that are significantly lower than those in Table 6. Although D3 achieves a reasonably good AP (0.88) on three-component (3C) sources, it performs very poor on two-component (2C) sources (0.28). While D4 has a marginally improved 2C AP (0.38), its 3C AP is low. This shows that identification of multicomponent sources from multisource subjects still poses a challenge to CLARAN. However, it is worth noting that the median CL of 2C and 3C sources in *T* is merely 0.64. Moreover, given the low number of sources (18) of classes 2C and 3C in *T*, their APs do not constitute reliable statistical measures, and this is particularly true for 3C. Given that the RGZ DR1 (with more than 11 000 multicomponent sources with CL >0.6) has the largest set of multicomponent radio sources that have been visually classified and labelled to date, we need to obtain additional data sets with far more multiple-source subjects to obtain quantitative measures. This will be the main focus for our future work, which will update Table 7 based on a larger number of multisource subjects in the testing set.

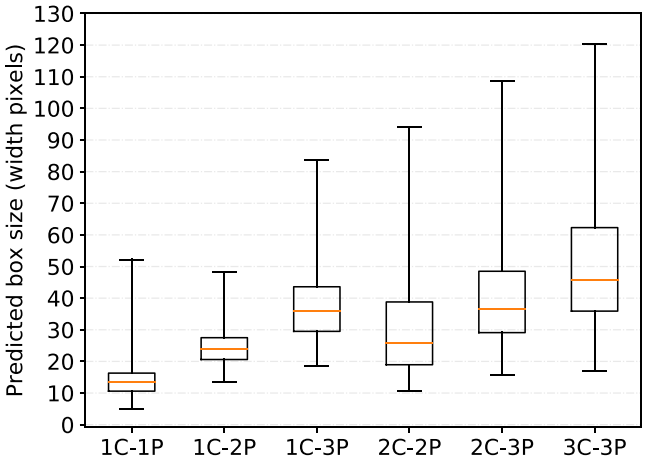


Figure 13. The distribution of detected box sizes for each class in the testing data set. Similar to Fig. 3, each box spans the third and the first quartile size for a morphology.

Table 8. Correlation coefficients between sizes of DR1 (ground-truth) bounding boxes and sizes of boxes predicted by CLARAN for subjects in the testing set.

Morphology	D3	D4
1C_1P	0.9718	0.9712
1C_2P	0.9877	0.9866
1C_3P	0.9933	0.9946
2C_2P	0.9940	0.9952
2C_3P	0.9934	0.9916
3C_3P	0.9939	0.9927

5.5 Predicted box sizes

Since the RoI regression loss contributes 55 per cent of the total training error as shown in Fig. 9, we compare the box sizes detected by CLARAN and the box sizes specified in the RGZ truth in the testing set. Fig. 13 shows the size distributions of detected boxes for each morphology class in the testing set. They appear visually consistent in terms of medians and interquartile ranges (IQR) with Fig. 3. But how do they compare to the testing-set ground truth? We calculate the correlation coefficients between the size (width) of each CLARAN-generated box and the size of its matching ground-truth (DR1) box. Table 8 shows that the correlation coefficients are high (> 0.97) across all six morphology classes for both D3 and D4. This suggests that box sizes predicted by CLARAN are very close to ground-truth values for all six morphologies.

5.6 *P*-value versus consensus level

In order to ascertain whether RGZ CLs might have affected CLARAN's performance, we examine the distribution of classification probabilities (*P*-values) of radio sources based on their RGZ CLs as shown in Fig. 14. Intuitively, a higher level of consensus corresponds to an easier case, which in turn should result in a more 'confident' classification result. This is indeed the case for simple morphology 1C_1P, as CL increases from 0.6 to 1.0, the IQR becomes much smaller, thus producing more stable and robust classifications, although the increase of median *P*-value is negligible: ≤ 1 per cent. However, the reduction of IQR is because 50 per cent

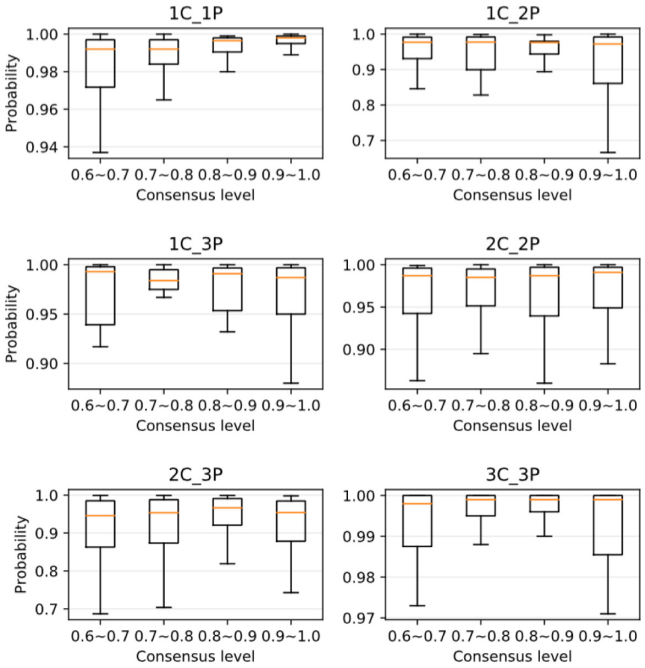


Figure 14. Comparisons between RGZ CLs (*X*-axis) and CLARAN classification probabilities (*Y*-axis) in the D4 testing set. The CL is segmented into four bins. Given a morphology class *m* within each bin, the box plot shows the distribution of classification probabilities of true positive detections whose morphology is *m*.

of 1C_1P sources have a CL close to 1 (as shown in Fig. 2) and the total number of 1C_1P sources is substantially greater than other classes (as shown in Table 3).

It is worth noting that CLARAN is not given any CL information whatsoever during both training and testing, and it treats each ground-truth subject and source equally without any CL-induced bias. This could explain the relatively flat yet high median *P*-values across all morphology classes. This suggests the CL-filtered sampling process described in Section 3.2 is appropriate and does not introduce systematic bias correlated to CLs as far as training CLARAN is concerned.

5.7 Model capacity and overfitting

To investigate the impact of the large number of trainable parameters (over 136 million) on model overfitting given the relatively small training set (6141 subjects), we conduct two experiments. In the first experiment, we reduce the number of model parameters from 136 to 23 million, and in the second one, we reduce it further to 18 million. This is achieved by reducing the dimension of the two fully connected layers (layers 24 and 26 in Table 4) from 4096 to 256 and 64, respectively. We retrain these two 'small-capacity' networks using the same training set (6141 subjects), and test them against the same testing set (4603 subjects). Their test accuracy – mAPs of 82.9 per cent and 81.7 per cent – is slightly poorer than CLARAN (mAP 83.6 per cent). This suggests that the CLARAN model is not in the overfitting zone (Goodfellow, Bengio & Courville 2016), in which higher model capacities correspond to higher test errors (thus lower test accuracy). We contend that the following factors mitigate overfitting in CLARAN.

First, the sole purpose of the two dropout layers (layers 25 and 27 in Table 4) is to prevent CLARAN from overfitting during training, which is discussed in Section 4.4. Second, Section 4.3 shows that, for each training subject, CLARAN dynamically generates thousands of RoI proposals and anchors to train the morphology classifier and the RoI regressor. This means the actual number of training examples going through the classification and regression loss functions (i.e. equations 4, 6, 9, and 10) is on the order of millions rather than thousands. Moreover, transfer learning (discussed in Section 4.2) ensures that all parameters in the convolutional layers have been trained on the ImageNet data set with millions of training images. This is particularly relevant to those 'frozen' parameters in the low-level convolutional kernels.

For the above reasons, it is not essential to use other data augmentation techniques (such as image rotation) to enlarge the training set. More importantly, rotating an image around the source centre, as is done in Aniyan & Thorat (2017), is not directly applicable to CLARAN. This is because it is CLARAN's job to find sources on an image. The pre-processing step cannot possibly 'reveal' a source *S*, and rotate/crop the image around *S* since *S* (and its location) is the very target CLARAN needs to predict. It is possible to blindly rotate the entire image/field around its own centre regardless of the location of the sources. However, doing so may place some components of an extended source out of the field if we do not resize the rotated image based on the rotation angle. Moreover, coordinates of the 'new' corners of all boxes on the image need to be recalculated and updated in the training set. Considering the above overheads, we will instead use CLARAN's STN layer to support rotation invariant feature extraction as discussed in Section 7 for our future work.

5.8 Comparison with Aniyan & Thorat (2017)

The classifier in Aniyan & Thorat (2017) produces a catalogue, which consists of 187 radio sources and their associated FR and BT morphology classifications (for both ground truths and predictions) and spatial locations. A direct comparison between this catalogue and the CLARAN output is not feasible since the morphology categories used in Aniyan & Thorat (2017) and CLARAN are different as described in Section 2. However, if we consider all FRII sources having two radio components, it is possible to make an indirect comparison between the 57 FRII sources (out of the 187 sources) and all two-component sources (i.e. 2C_2P and 2C_3P) predicted by CLARAN.

Out of the 4603 subjects in the CLARAN testing set, the D3 method identifies 904 two-component sources, and the D4 method identifies 1031 two-component sources. All the identified sources have *P*-values above 80 per cent. For each identified two-component source, we calculate its centre sky location from its bounding box coordinates predicted by CLARAN. This produces two location lists L_3 and L_4 for D3 and D4, respectively. We then perform spatial cross-matchings between $L_{3/4}$ and the 57 FRII sources predicted by the Aniyan & Thorat (2017) classifier. When setting the maximum match radius to 20 arcsec, the cross-matching finds one match between the 57 FRII sources and L_3, and one match between the 57 FRII sources and L_4. Both matches are under a 3.5 arcsec radius, and both matches refer to the same pair: source `3C 251` in the Aniyan & Thorat (2017) catalogue, and source `FIRST J110836.2 + 385854` in RGZ DR1. While the Aniyan & Thorat (2017) classifier predicts it as an FRII source with a probability of 99.9 per cent, both D3 and D4 methods predicts its morphology as 2C_3P with *P*-values of 95.7 per cent and 97.9 per cent, respectively.

6 DIRECTIONS AND RECOMMENDATIONS FOR USE OF CLARAN

We encourage interested astronomers to use CLARAN for their own research projects, because it can provide useful results even in its initial incarnation, and because experimentation and feedback on CLARAN will improve its performance. Access to CLARAN's source code from the `GitHub` repository is described in footnote 1. Given the results to date, we recommend the use of either method D3 or D4. Therefore, data would need to be provided in those forms, which can be obtained by following the descriptions on the `GitHub` repository. Also available in the repository are software modules that convert pairs of radio and IR maps to these forms.

In Section 6.1, we describe how CLARAN could be implemented in a simple automated manner for radio source classifications. In Section 6.2, we describe a variety of limitations in the current implementation, and in particular, how that would affect interpretation of the results.

6.1 Classifying radio sources automatically with CLARAN

6.1.1 How to use CLARAN?

For each input field, CLARAN detects and classifies the detected radio sources into the six RGZ morphology classes discussed in Section 3. Each classification generated will have a *P*-value which approximates the probability the identified source belongs to the identified morphology class. Therefore, CLARAN may provide more than one morphology classification for each radio source in the field. An additional post-processing filtering algorithm is then recommended for deciding how to handle multiple classifications for a single radio source, as well as dealing with fields with more than one radio source.

The simplest filtering algorithm that a user can implement is to make two simple assumptions: (1) reject all classifications with *P*-values below 0.8 unless the classification with the highest *P*-value is below 0.8; and (2) that there is only *one* radio source per field. While multiple sources exist within a test subject, our experience suggests that the source classification with the highest *P*-value is likely the correct classification as determined by CLARAN. The assumption of one radio source per field is not unreasonable because 98.5 per cent of RGZ DR1 fields contain only one radio source (Wong et al., in preparation). Further discussion on the impact of these assumptions can be found in Section 6.2.2.

6.1.2 Does this work?

The reliability analysis in Section 5 does not include the filtering method described in Section 6.1.1. From the perspective of an astronomer, the analysis in Section 5 may not be sufficient because it is crucial for an astronomer to identify the correct classification from the multiple classifications produced by CLARAN. As such, we will describe, in this subsection, the accuracy and reliability of CLARAN in combination with the simple filtering method described in Section 6.1.1.

To demonstrate that CLARAN (plus filtering) yields accurate and reliable classifications of resolved radio morphologies, we visually inspect an arbitrary sample of 500 test fields (from the entire testing set of 4603). We then apply the filtering method described in Section 5 to this sample. This arbitrary sample was selected via a simple Monte Carlo method that stops after a sample of 500 is reached. A plot that includes both RGZ DR1 classifications and CLARAN predictions is generated for each one of the 500 fields, which are then inspected and evaluated by a radio astronomer (OIW). 367 of the 500 verification fields contain extended, non-compact radio sources, and 133 fields contain compact unresolved radio sources.

Table 9. Visual inspection results for the 500 verification fields for CLARAN's D3 and D4 training methods. We refer to the independent visual verification conducted by the radio astronomer which includes the plausibility factor, as 'astronomer' in this table.

Compared to	D3	D4
RGZ DR1	447.0	465.2
Astronomer	465.5	477.2

A mismatch between CLARAN and RGZ DR1 does not necessarily mean that one or the other is incorrect for two main reasons. First, both CLARAN and RGZ classifications are limited by observational factors such as surface brightness sensitivity and resolution. In addition, a mismatch in number of peaks can also be due to the limitation of the DR1 pipeline. Therefore, a direct comparison between the classifications from CLARAN and those from DR1 is not a fair assessment of CLARAN's true performance. As such, we compare the results from CLARAN using the simple method described in Section 6.1.1 to RGZ DR1, and to a plausibility factor that is determined by an astronomer. The main idea for the plausibility factor is to determine whether a classification from CLARAN can be deemed plausible by an expert astronomer given the radio and IR maps presented, irrespective of the classification from the DR1 catalogue. For example, a field containing two unresolved radio source components with no IR counterpart in between, or at the positions of the radio components, can be plausibly classified as either one 2C_2P source or as two 1C_1P sources.

We use a simple scoring method for quantifying the efficacy of CLARAN. A score of one is awarded to each correct radio source classification. The total number of correct classifications is then divided by the total number of sources within the field. Hence, a field with multiple source classifications will require a correct classification for each source to recover the total score of one for that field. In this verification process, we ask two questions: (1) does CLARAN reproduce the RGZ DR1 classification?; and (2) if CLARAN provides a classification *C* different from that of RGZ DR1, is *C* still plausible given the radio and IR observations?

Table 9 lists the recovered verification scores for the 500 fields. Comparing the results from the D3 and D4 training methods to RGZ DR1, we find D4 to outperform D3 in a consistent manner. While this is not surprising, it confirms that this scoring method works. Taking into consideration the plausibility factor, our results show that CLARAN is likely to produce accurate source classifications at the optimal accuracy level above 93.1 per cent and 95.4 per cent using the training methods of D3 and D4. Hence, we can expect reliable results from the current D4 version of CLARAN in combination with the simple post-processing filtering method described in Section 6.1.1.

6.2 Limitations and insights

While Section 6.1 shows that CLARAN is a relatively accurate and reliable prototype classifier, we caution the reader and users of CLARAN that the current version does include a number of limitations that we discuss in more details in this subsection. Previously in Section 6.1.2, we noted that a mismatch between the two does

not necessarily mean that either CLARAN or RGZ is incorrect. In this subsection, we explore and describe the limitations and lessons learnt from the implementation of CLARAN, from the perspective of an astronomer.

There are several reasons why a mismatch between the two methods may still result in a plausible source classification. For many complex radio sources, further follow-up observations may be required to ascertain the precise source component associations and host galaxy. Furthermore, the determination of the number of peaks is an approximation by the DR1 pipeline that is based upon the contour levels. Hence, we discuss in Section 6.2.2, CLARAN's reliability from an astronomer's perspective based on the often-used method of visual inspection.

6.2.1 *Source angular size*

Similar to the RGZ project, CLARAN will not be able to provide accurate classifications for radio sources which extend beyond the 3-arcmin FoV. RGZ DR1 found the median angular size of multi-component radio sources to be 43.1 arcsec and that 95.2 per cent of the DR1 multicomponent sources have an angular size that is smaller than 97 arcsec (Wong et al., in preparation). However, there is a small fraction of sources which may be limited by the current FoV size. Fig. 15 a illustrates one example field within the verification set of 500 that encounters the limitations of the 3-arcmin FoV, whereby the field presented in RGZ only encapsulates three of the four radio components. The northernmost radio component lies beyond the top edge of the field. Consequently, both the classifications from RGZ DR1 and CLARAN are incorrect (Fig. 15b). Enlarging the field by five times to a 15 arcmin by 15 arcmin field (Fig. 15c), we reveal that the central radio source has a double–double morphology (4C-4P), for which CLARAN was not trained to identify. When running directly on this larger field, CLARAN ends up breaking this double–double source into two smaller sources – 3C-3P and 1C-1P (Fig. 15e). On the other hand, the host galaxy captured inside the 3C-3P bounding box is still correct.

Of the 500 verification fields, we find two classifications in which CLARAN estimated a significantly larger angular source size (by a factor of a few) relative to that reported by RGZ DR1. Two most likely reasons exist for such an estimation: either CLARAN is confused by the synthesis imaging artefacts that remain in some fields, or that CLARAN is capable of detecting low-level diffuse emission. We will investigate these specific aspect of CLARAN in future studies as it is beyond the scope of this proof-of-concept paper to provide an in-depth investigation into this specific area.

6.2.2 *Assumption of one source per field*

Of the two assumptions recommended for the simple filtering method in Section 6.1.1, the second assumption of one source per field, may not be necessary for some studies to obtain individual classifications. Also, this assumption of one source per field may be invalidated for two main reasons. First, multicomponent radio sources with large angular sizes can result in multiple plausible classifications as discussed in the previous subsection. Secondly, the classifications of multiple radio sources in the 8 per cent of verification fields are not distinguishable from multiple classifications of a single multicomponent source. Hence, this subsection investigates the reliability of CLARAN when we remove the single-source assumption.

To this end, we examine the classification degeneracies that become inherent (when we do not assume one source per field) using a completeness ratio. We define a *completeness ratio* to be the ratio of the total number of radio sources to the total number of correct classifications per subject ($N_{\rm TRUE}/N_{\rm CLASSIFIED}$). A ratio of 1.0 indicates that every source within a field has been correctly classified. Ratios greater than 1.0 suggest that an individual field contains more DR1 sources than classified. Likewise, a ratio below 1.0 indicates that CLARAN found more than one classification per source within a field. Fig. 16 presents the distribution of completeness ratios for the 500 verification fields using the D3 (grey shaded) and D4 (striped) training methods. As shown in Table 10, we find that a ratio of 1.0 is obtained for 83.4 per cent and and 86.4 per cent (using the D3 and D4 methods, respectively) of the verification sample. This result is consistent with the precision of the classifications quantified in Section 5. However, since out of the 500 verification fields, only 36 are in fact multisource subjects, and the majority (464) of them still contain only one source per FoV. Therefore, Table 10 may not generalize well beyond this particular 500-subject sample to reflect the effect of 'removing the single-source assumption' on CLARAN's reliability. To address this issue, we create another special subset S ($|S| = 36$) from the 500 verification fields, in which each subject has at least two sources. We recalculate the completeness ratios against S, and report the updated result in Table 11. Compared to Table 10, Table 11 essentially examines classification degeneracies under the 'worst-case' scenario where every subject has multiple sources. We find that the updated results of 55.6 per cent and 66.8 per cent (for a completeness ratio of 1.0) are largely consistent with Table 7.

7 DISCUSSION

This paper builds upon earlier machine learning explorations that use RGZ classifications as a training set (Lukic et al. 2018; Alger et al. 2018). Following Alger et al. (2018) who found that compact radio source classifications do not benefit significantly from using advanced machine learning convolution methods, we specifically train and test CLARAN on a large sample of extended non-compact radio sources. Our work demonstrates the feasibility of applying modern deep learning methods, which originate from generic object detection and computer vision, for cross-matching complex radio sources of multiple components with IR catalogues. The promising results of this study have implications for further development of fully automated cross-wavelength source identification, matching, and morphology classifications for pre-SKA surveys.

The data fusion methods and their performance evaluations described in this paper provide a good starting point to train machines to appropriately incorporate and integrate numerous deep multi-wavelength catalogues and other information (e.g. redshifts) for radio source identification and morphology classification.

We adapt the STN for cropping out RoI proposals from feature maps in order to obtain a differentiable loss function for end-to-end training. The adaptation takes place in the affine transformation matrix (equation 8) where we fix the rotation angle to zero degree (thus no rotation). By running a fully fledged STN that allows rotation angles to be learned from the feature map, CLARAN could perform source-dependent, rotation-invariant morphology classification within a single end-to-end pipeline. This approach will differ from random rotations of the entire image for feature augmentation (Dieleman et al. 2015) because it is trained to rotate each potential source by a distinct angle for optimal morphology classification. The assumption that all sources within a subject rotate simultaneously by a pre-defined angle does not always hold.

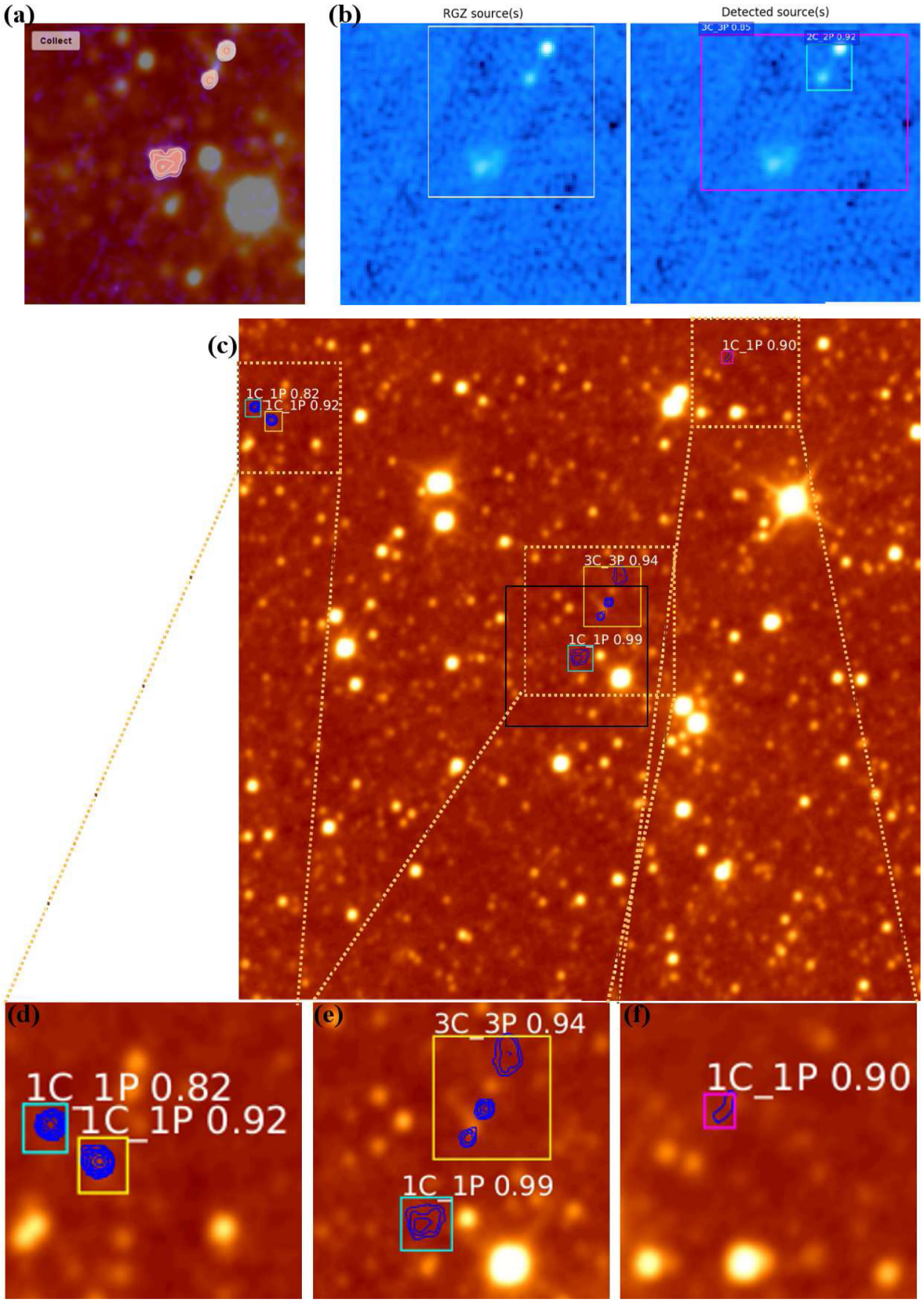


Figure 15. An example of the limitation inherent to the 3-arcmin FoV whereby radio sources are misclassified by both RGZ DR1 and CLARAN due to the missing radio component that lies beyond the northernmost edge of the field. North and East are to the top and left of the page, respectively. Panel (a) shows the 3-arcmin by 3-arcmin RGZ subject presented to the participants for RGZ J080837.0 + 170804. Panel (b) is a pair of verification maps showing the DR1 classification (left) and CLARAN's classification (right). Panel (c) shows the expanded 15-arcmin by 15-arcmin FoV for RGZ J080837.0 + 170804 where the original RGZ image size is marked by the black box. For added visual clarity, zoomed-in maps of the three radio sources found by CLARAN within panel (c) are shown in panels (d)–(f).

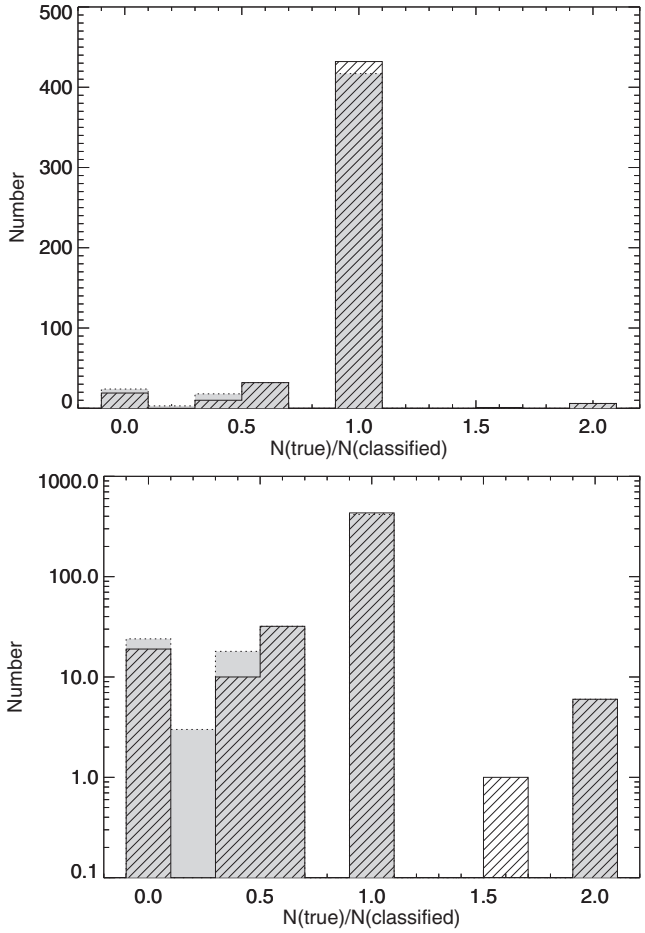


Figure 16. Fraction of total number ($N_{\rm true}$) to classified number ($N_{\rm classified}$) of radio sources in the verification sample of 500. The top figure shows a linear *y*-axis and the bottom figure shows a logarithmic *y*-axis for the same distribution. Verification results from the D3 and D4 methods are represented by the grey-shaded and striped distributions, respectively.

Table 10. The fraction of verification fields within three divisions of completeness ratios.

$N_{\rm TRUE}/N_{\rm CLASSIFIED}$	D3	D4
=1	0.834	0.864
<1	0.154	0.122
>1	0.012	0.014

Table 11. The fraction of verification fields on a set of 36 subjects, each of which has at least two sources in its FoV.

$N_{\rm TRUE}/N_{\rm CLASSIFIED}$	D3	D4
=1	0.556	0.668
<1	0.278	0.166
>1	0.166	0.166

Despite the great difference between common images and RGZ subjects, we demonstrate that the CNN weights thoroughly trained on the comprehensive, well-labelled ImageNet provide far better initial conditions than random weight initialization with respect to training efficiency and evaluation metrics. However, as shown in Fig. 8, appropriate control of these pre-loaded weights is equally important in order to achieve a desirable level of efficiency and precision. The fact that freezing low-level weights leads to a much smaller training error suggests that high-level feature extractions (such as shapes, texture, structure, etc.) should be prioritized after transfer learning.

On the other hand, low-level feature learning should be carried out at a much slower pace to avoid overfitting. This implies that we may need different learning rates for different parts of the neural network. In this example, freezing weights is equivalent to reducing the learning rate to 0, which simply gives up opportunities to learn any low-level features unique to the RGZ data set. Therefore, a more fine-grained learning rate distribution applied across the network can take advantage of the benefits from transfer learning.

8 CONCLUSIONS

Cross-identification of radio source components is currently done through visual inspection by expert astronomers or citizen scientists. However, such a labour-intensive method is not scalable even for the pre-SKA radio surveys such as EMU (Norris et al. 2011). In this paper, we describe a machine learning-based method for automated localization and identification of multicomponent, multipeak radio sources with associated morphological information. Drawing on the latest models developed in object detection and deep learning, our method has achieved efficient identification of radio galaxies on unseen RGZ data sets with an mAP of 83.6 per cent and an empirical plausibility accuracy of above 90 per cent. CLARAN is able to distinguish between six of the most common distinct classes of radio source morphologies. These six classes of morphologies are defined in terms of the number of components and peaks that describe source associations and identifications produced by the RGZ Data Release 1 (Wong et al., in preparation). CLARAN also works reasonably well on fields much larger than those provided in the training set.

For future work, we will focus on improving CLARAN's capability of separating unrelated multiple sources from multiple components of single sources. To begin with, we will incorporate more multi-source subjects into the testing set as suggested by Table 7. Targeted plausibility analysis of the confusion between 2C and 3C classifications and reliable statistical measures will help us develop robust feature augmentation schemes needed to address this key problem.

ACKNOWLEDGEMENTS

We acknowledge the contribution of more than 12 000 volunteers in the RGZ project. The International Centre for Radio Astronomy Research is a joint venture between Curtin University and The University of Western Australia with support and funding from the State Government of Western Australia. This work was supported by computing resources provided by the Pawsey Supercomputing Centre (with funding from the Australian Government and the Government of Western Australia) and CSIRO (Commonwealth Scientific and Industrial Research Organisation). Partial support for LR's and AG's work at the University of Minnesota comes from grant AST-1714205 from the U.S. National Science Foundation. HA benefited from grant DAIP no. 66/2018 of Universidad de Guanajuato.

REFERENCES

Abadi M. et al., 2016, in Keeton K., Roscoe T., eds, 12th USENIX Symposium on Operating Systems Design and Implementation, Vol. 16. USENIX Association, Savannah, GA, USAp. 265

Abraham S., Aniyan A. K., Kembhavi A. K., Philip N. S., Vaghmare K., 2018, MNRAS, 477, 894

Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, MNRAS, 479, 415

Alger M. J. et al., 2018, MNRAS, 478, 5547

Aniyan A. K., Thorat K., 2017, ApJS, 230, 20

Banfield J. K. et al., 2015, MNRAS, 453, 2326

Banfield J. K. et al., 2016, MNRAS, 460, 2376

Becker R. H., White R. L., Helfand D. J., 1995, ApJ, 450, 559

Bevington P. R., Robinson D. K., 2003, Data Reduction and Error Analysis for the Physical Sciences, 3rd edn. McGraw-Hill Education, New York

Ciregan D., Meier U., Schmidhuber J., 2012, in Chellappa R., Kimia B., Zhu S. C., eds, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Providence, RI, USA, p. 3642

Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., 1998, AJ, 115, 1693

Contigiani O. et al., 2017, MNRAS, 472, 636

Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., 2009, in Huttenlocher D., Medioni G., Rehg J., eds, Proceedings of the IEEE Computer Vision and Pattern Recognition. IEEE Computer Society, Miami, FL, USAp. 248

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

Fabbro S., Venn K. A., O'Briain T., Bialek S., Kielty C. L., Jahandar F., Monty S., 2018, MNRAS, 475, 2978

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

Girshick R., 2015, in Bajcsy R., Hager G., Ma Y., eds, Proceedings of the IEEE International Conference on Computer Vision. IEEE Computer Society, Washington, DC, USA, p. 1440

Girshick R., Donahue J., Darrell T., Malik J., 2014, in Dickinson S., Metaxas D., Turk M., eds, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Columbus, OH, USA, p. 580

Glorot X., Bordes A., Bengio Y., 2011, in Gordon G., Dunson D., Dudik M., eds, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Vol. 15. PMLR, Ft. Lauderdale, FL, USA, p. 315

Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press, Cambridge, MA

Hancock P. J., Murphy T., Gaensler B. M., Hopkins A., Curran J. R., 2012, MNRAS, 422, 1812

He H., Garcia E. A., 2009, IEEE Trans. Knowl. Data Eng., 21, 1263

He K., Zhang X., Ren S., Sun J., 2016, in Bajcsy R., Li F.-F., Tuytelaars T., eds, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Las Vegas, NV, USA, p. 770

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

Huang J. et al., 2017, in Chellappa R., Zhang Z., Hoogs A., eds, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Honolulu, HI, USA, p. 7310

Hubel D. H., Wiesel T. N., 1962, J. Physiol., 160, 106

Huber P. J. et al., 1964, Ann. Math. Stat., 35, 73

Jaderberg M., Simonyan K., Zisserman A. et al., 2015, in Cortes C., Lawrence N., Lee D., Sugiyama M., Garnett R., eds, Advances in Neural Information Processing Systems, Vol. 28. NIPS Foundation, Inc., Montreal, Quebec, Canadap. 2017

Joye W., Mandel E., 2003, in Payne H. E., Jedrzejewski R. I., Hook R. N., eds, ASP Conf. Ser. Vol. 295, Astronomical Data Analysis Software and Systems XII. Astron. Soc. Pac., San Francisco, p. 489

Kapińska A. D. et al., 2017, AJ, 154, 253

Kimball A. E., Ivezić Ž., 2008, AJ, 136, 684

Krizhevsky A., Sutskever I., Hinton G. E., 2012, in Pereira F., Burges C., Bottou L., Weinberger K., eds, Advances in Neural Information Processing Systems, Vol. 25. NIPS Foundation, Inc., Stateline, NV, USA, p. 1097

LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436

Lukic V., Brüggen M., Banfield J. K., Wong O. I., Rudnick L., Norris R. P., Simmons B., 2018, MNRAS, 476, 246

Nair V., Hinton G. E., 2010, in Wrobel S., ed., Proceedings of the 27th International Conference on Machine Learning. International Machine Learning Society, Haifa, Israel, p. 807

Neubeck A., Van Gool L., 2006, in Tang Y., Wang P., Lorette G., Yeung D., eds, IEEE International Conference on Pattern Recognition, Vol. 18. IEEE Computer Society, Hong Kong, China, p. 850

Norris R. P. et al., 2006, AJ, 132, 2409

Norris R. P. et al., 2011, PASA, 28, 215

Owen F. N., Ledlow M. J., 1994, in Bicknell G. V., Dopita M. A., Quinn P. J., eds, ASP Conf. Ser., Vol. 54, The Physics of Active Galaxies. Astron. Soc. Pac., San Francisco, p. 319

Padovani P., 2017, Nat. Astron., 1, 0194

Pearson K. A., Palafox L., Griffith C. A., 2018, MNRAS, 474, 478

Ren S., He K., Girshick R., Sun J., 2017, IEEE Trans. Pattern Anal. Mach. Intell., 39, 1137

Russakovsky O. et al., 2015, Int. J. Comput. Vis., 115, 211

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

Sedaghat N., Mahabal A., 2018, MNRAS, 476, 5365

Shallue C. J., Vanderburg A., 2018, AJ, 155, 94

Simonyan K., Zisserman A., 2015, Bengio Y., LeCun Y., eds, International Conference on Learning Representations. ICRL Organising Committee, San Diego, p. 1150

Srivastava N., Hinton G. E., Krizhevsky A., Sutskever I., Salakhutdinov R., 2014, J. Mach. Learn. Res., 15, 1929

Stark D. et al., 2018, MNRAS, 477, 2513

Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., 2015, in Bajcsy R., Hager G., Ma Y., eds, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Washington, DC, USA, p. 1

Taigman Y., Yang M., Ranzato M., Wolf L., 2014, in Dickinson S., Metaxas D., Turk M., eds, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Columbus, OH, USA, p. 1701

Wright E. L. et al., 2010, The Astronomical Journal, 140, 1868

Wright D. E. et al., 2017, MNRAS, 472, 1315

Yosinski J., Clune J., Bengio Y., Lipson H., 2014, in Ghahramani Z., Welling M., Cortes C., eds, Advances in Neural Information Processing Systems, Vol. 27. NIPS Foundation, Inc., Montreal, Quebec, Canada, p. 3320

Zeiler M. D. et al., 2013, in Ward R., Deng L., eds, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE Signal Processing Society, Vancouver, Canada, p. 3517

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