The Astrophysical Journal Supplement Series, 240:34 (21pp), 2019 February https://doi.org/10.3847/1538-4365/aaf9a2
© 2019. The American Astronomical Society. All rights reserved.

A Machine Learning Based Morphological Classification of 14,245 Radio AGNs Selected from the Best–Heckman Sample

Zhixian Ma[1], Haiguang Xu[2,3,4], Jie Zhu[1], Dan Hu[4], Weitian Li[4], Chenxi Shan[4], Zhenghao Zhu[4], Liyi Gu[5], Jinjin Li[6], Chengze Liu[4], and Xiangping Wu[7]

[1] Department of Electronic Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang, Shanghai 200240, People's Republic of China
mazhixian@sjtu.edu.cn, zhujie@sjtu.edu.cn
[2] School of Physics and Astronomy/Tsung-Dao Lee Institute, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang, Shanghai 200240, People's Republic of China; hgxu@sjtu.edu.cn
[3] IFSA Collaborative Innovation Center, Shanghai Jiao Tong University, Minhang, Shanghai 200240, People's Republic of China
[4] Department of Astronomy, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang, Shanghai 200240, People's Republic of China
[5] SRON Netherlands Institute for Space Research, Sorbonnelaan 2, 3584 CA Utrecht, The Netherlands
[6] Key Laboratory of Thin Film and Microfabrication Technology (Ministry of Education), Department of Micro/Nano Electronics, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China
[7] National Astronomical Observatories, Chinese Academy of Sciences, 20A Datun Road, Beijing 100012, People's Republic of China
Received 2018 May 20; revised 2018 December 7; accepted 2018 December 17; published 2019 February 12

Abstract

We present a morphological classification of 14,245 radio active galactic nuclei (AGNs) into six types, i.e., typical Fanaroff–Riley Class I/II (FRI/II), FRI/II-like bent-tailed, X-shaped radio galaxy, and ringlike radio galaxy, by designing a convolutional neural network based autoencoder, namely MCRGNet, and applying it to a labeled radio galaxy (LRG) sample containing 1442 AGNs and an unlabeled radio galaxy (unLRG) sample containing 14,245 unlabeled AGNs selected from the Best–Heckman sample. We train MCRGNet and implement the classification task by a three-step strategy, i.e., pre-training, fine-tuning, and classification, which combines both unsupervised and supervised learnings. A four-layer dichotomous tree is designed to classify the radio AGNs, which leads to a significantly better performance than the direct six-type classification. On the LRG sample, our MCRGNet achieves a total precision of ∼93% and an averaged sensitivity of ∼87%, which are better than those obtained in previous works. On the unLRG sample, whose labels have been human-inspected, the neural network achieves a total precision of ∼80%. Also, using Sloan Digital Sky Survey Data Release 7 to calculate the r-band absolute magnitude ($M_{\rm opt}$) and using the flux densities to calculate the radio luminosity ($L_{\rm radio}$), we find that the distributions of the unLRG sources on the $L_{\rm radio}$–$M_{\rm opt}$ plane do not show an apparent redshift evolution and could confirm with a sufficiently large sample that there could not exist an abrupt separation between FRIs and FRIIs as reported in some previous works.

Key words: catalogs – galaxies: statistics – methods: data analysis – radio continuum: galaxies – techniques: miscellaneous

Supporting material: machine-readable tables

1. Introduction

The study of radio galaxies and their close relatives, radio-loud quasars and blazars, began in the mid-1950s when the Third Cambridge Catalog (3C; Mackay 1971) was created. These sources, whose radio emission is nonthermal, polarized, and prodigiously powerful with luminosities up to 10^{38-39} W in the 10 MHz–100 GHz band, form an important subclass of the active galactic nucleus (AGN) population. A typical bright radio galaxy appears as an exclusively large elliptical galaxy that exhibits a distinctive radio appearance characterized by twin jets (in some sources, however, only one is visible) and twin lobes or plumes. In many cases these twin structures, which straddle a compact core, may extend out to several hundred kiloparsecs on each side of the galaxy.

By comparing the distance between the brightest regions on either side of the galaxy with the entire extent of the source, Fanaroff & Riley (1974) classified 57 3C radio-loud galaxies into two broad categories, i.e., the low-luminosity Class I sources (FRIs), which show a decreasing radio emission as the distance from the core increases, and the high-luminosity Class II sources (FRIIs), which show an inverse tendency and often have bright hotspots at the ends of the lobes. In addition to the typical Fanaroff–Riley sources, some morphological subclasses have been defined in succeeding investigations, including (1) narrow-angle tail sources (NATs; Rudnick & Owen 1976), which usually appear as FRIs bent by the ram pressure of the surrounding intracluster medium in galaxy clusters (Proctor 2011); (2) wide-angle tail sources (WATs; Rudnick & Owen 1976), which simultaneously possess FRII-like jet-hotspot structures and, connected to the ends of the jets, FRI-like large-scale plumes rather than lobes; (3) X-shaped radio galaxies (XRGs; Leahy & Parma 1992; Cheung 2007), which exhibit a pair of faint lobes (the "wings") oriented at an angle to another pair of bright active lobes; (4) ringlike radio galaxies (RRGs; Proctor 2011); and (5) others, such as hybrid-morphology radio sources (Gopal-Krishna & Wiita 2000), that show FRI and FRII lobe morphologies on opposite sides.

As of today, however, understanding of this taxonomy is inadequate and so is the relevant sample study. For example, Capetti et al. (2017a) found that although the average radio luminosity of FRIIs is higher than that of FRIs, most of the FRIIs in their sample possess a radio power lower than the FRI–FRII transition found in 3C, which is 2×10^{25} W Hz^{-1} at 178 MHz ($H_0 = 50$ km s^{-1} Mpc^{-1}; Fanaroff & Riley). Leahy & Parma (1992), Cheung (2007), and Dennett-Thorpe et al. (2002)

also reported that the radio luminosities of XRGs, the majority of which possess an FRII morphology, straddle the Fanaroff–Riley transition. These make the problem of the FRI–FRII dichotomy, one of the unresolved mysteries in radio astronomy, more puzzling.

At present the purpose of the morphological study of radio galaxies is twofold. First, the diversity of the appearances of radio galaxies is a reflection of the physics behind, with which we can investigate the central black holes and their accretion modes, interactions between radio galaxies and surrounding gas, AGN feedback mechanisms, and so on, when complemented by requisite spectral information (see reviews of McNamara & Nulsen 2007 and Fabian 2012). Second, radio galaxies make a significant contribution to the extragalactic emission in the low-frequency radio sky, causing a severe impediment to recovering the redshifted 21 cm signals of neutral hydrogen from the Epoch of Reionization (EoR), a period yet to be explored during which the universe was reionized by the emissions of the first-generation objects. According to the estimations of, e.g., Mellema et al. (2013) and Chapman et al. (2016), the total emission of radio galaxies in 50–200 MHz is more luminous than that of the redshifted EoR signals by 2–3 orders of magnitude. In direct imaging studies and studies of two-dimensional power spectra in terms of EoR window analysis (e.g., Chapman et al. 2016), the situation is further complicated by the diverse morphologies of radio galaxies, because even subpercent-level residuals in removing the radio galaxy component from the radio maps can ruin any attempt to reveal the EoR signals (e.g., Beardsley et al. 2016; Chapman et al. 2016; Procopio et al. 2017). Thus, when detecting the EoR signals with facilities such as the Square Kilometre Array (SKA; Mellema et al. 2013; Koopmans et al. 2015) and the Hydrogen Epoch of Reionization Array (HERA; DeBoer et al. 2017), the radio galaxy foreground component must be correctly identified and thoroughly removed.

A tremendously large number of radio galaxies will be detected with the high-quality radio data obtained by SKA, HERA, and their scale-down analogs. For example, from the simulation of the SKA Design Study (Wilman et al. 2008, 2010), it is estimated that about 2.5×10^4 radio sources per square degree will be detected at 1.4 GHz with a limit of 1 μJy in the SKA-MID deep survey at 350 MHz–14 GHz. This estimation agrees with the results of Padovani (2016) very well, implying a significant increase of the radio galaxy number density compared to that in Faint Images of the Radio Sky at Twenty cm (FIRST) Data Release 7 (DR7; $\sim$100 sources per square degree at 1.4 GHz) by a factor of $\sim 10^3$.

Given such a large number of sources with various morphologies, it is obvious that traditional human classification, which is primarily based on visual examination, will be impractical. Machine-learning techniques must be summoned to implement source detection and classification tasks in a correct and efficient way. In fact, machine learning has been applied to help analyze stellar spectra (e.g., Fabbro et al. 2018), estimate photometric redshifts (e.g., Mountrichas et al. 2017), search for signals from pulsars (e.g., Bethapudi & Desai 2018) and gamma-ray bursts (e.g., Ukwatta et al. 2016), classify galaxies according to their morphologies (e.g., Proctor 2011; Dieleman et al. 2015; Barchi et al. 2017; Lukic et al. 2018), predict solar flares (e.g., Benvenuto et al. 2018), model galaxy formation and evolution (e.g., Kamdar et al. 2016), and so on. As reported by Aniyan & Thorat (2017), who successfully classified 178 FRI, 284 FRII, and 254 bent-tailed (BT) galaxies, the morphological classification of radio galaxies is also a task potentially well-suited for deep machine learning when convolutional neural networks (CNNs) are applied (see also Lukic et al. 2018).

CNNs are a class of biologically inspired variants of feedforwarded multilayer perceptrons that have shown excellent performance in imaging analysis due to their distinguishing features, i.e., 3D volumes of neurons, a local receptive field, and shared weights (Goodfellow et al. 2016). By training a neural network with augmented sample images and applying a fusion classifier based on a majority-voting classifier, Aniyan & Thorat (2017) found that the neural network could classify FRI and BT images with very high precision (91% and 95%), although for FRII images the precision was relatively lower (75%). They argued that the low precision for FRII might have been caused by the fact that many BT sources exhibited an FRII appearance and were classified as FRIIs.

The primary motivation of this paper is to develop an automatic open-source tool that can classify a large sample of radio galaxies based on morphological representation learning (LeCun et al. 2015) with high computational efficiency, small resource consumption, and powerful predictive capabilities, so as to enable relevant studies on portable computers. In order to do this we have designed a convolutional autoencoder (CAE) based on a deep CNN, unsupervisedly pre-trained the CAE with 14,245 unlabeled images of radio AGNs in the Best–Heckman sample (Best & Heckman 2012, BH12 sample hereafter), and supervisedly fine-tuned the CAE with images of 1442 radio AGNs that have already been labeled in the literature (Gendre & Wall 2008; Gendre et al. 2010; Cheung 2007; Proctor 2011; Capetti et al. 2017a, 2017b; Baldi et al. 2018). Given that currently no large labeled radio galaxy (LRG) sample is available for neural network training, this approach (i.e., unsupervisedly pre-training the network with a large unlabeled sample and supervisedly fine-tuning it with a relatively small labeled sample) can best extract information hidden in the images and make the classification robust to morphological variability (e.g., Erhan et al. 2010; Hinton et al. 2012; Christodoulidis et al. 2016; see also LeCun et al. 2015 for a review). Also, the application of a large unlabeled sample helps construct a correct network by avoiding the overfitting problem (e.g., Goodfellow et al. 2016).

To support astronomical studies, besides the types of FR0, typical FRI, typical FRII, and BT (this category includes both NATs and WATs, which are labeled "FRI-like" or "FRII-like" in this work), we extend the classification to the types of XRG and RRG. As a result of this work a catalog of 14,245 radio AGNs, the labels of which have been cross-checked by human inspection, is presented.

This paper is organized as follows. In Section 2 we show the classification frame, describe our sample selection criteria, and construct the labeled and unlabeled samples. In Section 3 we describe the design of the CAE, which is based on a single small-scale CNN. In Section 4 we pre-train and fine-tune the neural network, classify the radio galaxies, and evaluate the performance of the network by visualizing the extracted features of the labeled samples with *t*-distributed stochastic neighbor embedding (*t*-SNE) and by comparing the results with those obtained with human inspection on the unlabeled samples. In Sections 5 and 6 we discuss and summarize our results. Throughout the paper we quote errors at the 90% confidence level unless otherwise stated,

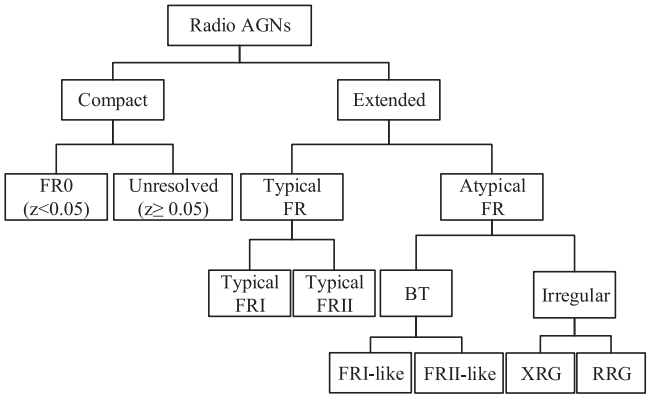


Figure 1. Tree structure used to classify radio galaxies.

and adopt cosmological parameters $H_0 = 71\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and $\Omega_m = 1–\Omega_\Lambda = 0.27$.

2. Classification Tree and Sample Selection

By consulting the morphological taxonomies of Proctor (2011), Padovani (2016), and Padovani et al. (2017), we design a four-layer dichotomous tree structure classifier (Figure 1) and use it to classify radio AGNs. Such a binary classification is expected to achieve better performance than direct multiclass classification under the same network structure and parameter settings (e.g., Bishop & Nasrabadi 2006; Fei & Liu 2006; Tsoumakas & Katakis 2007; Ma et al. 2017), because the greater the number of model outputs (i.e., classes), the higher the complexity and stability the model should possess (see, e.g., Rifkin & Klautau 2004; Tsoumakas & Katakis 2007) as shown in many practical cases, especially on machine-learning models like the support vector machine (Tsoumakas & Katakis 2007; Chang & Lin 2011). Given that a radio AGN showing a compact appearance can be either a genuine FR0 source or an extended source located at a relatively higher redshift (e.g., $z > 0.05$), we divide the compact category into FR0 (i.e., genuine FR0s located at $z \leqslant 0.05$) and unresolvable (i.e., sources with a compact appearance and located at $z > 0.05$, including genuine FR0s and extended sources whose linear extension is $\leqslant$6.8 kpc or 5″) subtypes. Considering that BTs, XRGs, and RRGs more or less exhibit FR morphological features (Proctor 2011), in the extended source category we define the atypical subtype to include BTs (further classified into FRI-like and FRII-like BTs) and irregular sources (i.e., XRGs and RRGs).

All the images of the radio AGNs included in both the unlabeled sample (unlabeled radio galaxy sample or unLRG sample) and the labeled sample (LRG sample) are drawn from the Very Large Array (VLA) FIRST survey archive at http://sundog.stsci.edu/ in FITS format with a size of 4.′5 × 4.′5 (150 × 150 pixels).

We construct the LRG sample by combining the CoNFIG catalog (Gendre & Wall 2008; Gendre et al. 2010), the FR0CAT catalog (Baldi et al. 2018), the FRICAT catalog (Capetti et al. 2017b), the FRIICAT catalog (Capetti et al. 2017a), and the catalogs of Cheung (2007) and Proctor (2011) (Tables 1 and 2). To be specific, we first include all 392 compacts, 284 FRIs, 587 FRIIs, and 430 BTs labeled in the CoNFIG, FR0CAT, FRICAT, and FRIICAT catalogs in the LRG sample. Note, however, that one source in the FRICAT catalog shows compact morphology, three sources in the FRIICAT catalog are X-shaped-like, five FRIs and eight FRIIs in the FRICAT and FRIICAT are bent, and the compact category in the CoNFIG catalog actually contains sources located at relatively higher redshifts, for which it is impossible to determine whether or not they are intrinsically compact with the spatial resolution of FIRST. To handle these special cases we either carry out visual inspection of their images or inspect their redshifts, and classify the sources into the FRI-like, FRII-like, XRG, and unresolvable categories accordingly. Next we merge 410 BTs and 65 RRGs in the Proctor (2011) catalog and 100 XRGs in the Cheung (2007) catalog into the LRG sample, thus obtaining a preliminary sample. At the final step, we filter the preliminary sample by excluding sources with particularly low peak-to-average power ratios (PAPRs) smaller than 10.0 dB, sources with ambiguous morphologies, sources that are too extended, and 37 radio galaxies labeled "uncertain" in the CoNFIG catalog. The resulting LRG sample consists of a total of 1442 radio AGNs, which span a redshift range of $0.06 < z < 3.90$ and a 1.4 GHz flux density range of 6.0 mJy $< S_{1.4\,\mathrm{GHz}} < 1.5 \times 10^4$ mJy (see Table 3).

We construct the unLRG sample, which does not overlap the LRG sample, by selecting sources directly from the large sample published by Best & Heckman (2012). In the BH12 sample the authors cataloged 18,286 radio galaxies by cross-matching the optical spectroscopic catalogs produced by the group at the Max Planck Institute for Astrophysics and Johns Hopkins University (Brinchmann et al. 2004; Tremonti et al. 2004), who

Table 1
Catalogs Used to Construct the LRG Sample

Catalog	[a]Source Number and Type	Reference
FR0CAT	108 (compact: 104)	Baldi et al. (2018)
FRICAT	233 (FRI: 192; compact: 1; BT: 5)	Capetti et al. (2017b)
FRIICAT	123 (FRII: 80; BT: 8; XRG: 3)	Capetti et al. (2017a)
Cheung (2007)	100 (XRG: 81)	Cheung (2007)
Proctor (2011)	475 (BT: 285; RRG: 32)	Proctor (2011)
CoNFIG 1–4	856 (FRI: 19; FRII: 351; BT: 9; compact: 272)	Gendre et al. (2010)

Notes. Sources marked as unconfirmed or uncertain in the CoNFIG catalog are not included in this work. Sources with low PAPRs or spanning a region larger than the 150 × 150 pixel crop box are also excluded.
[a] Source number in the original catalogs and source number included in our LRG sample with corresponding types (brackets).

Table 2
Number of Sources Included in the LRG Sample

AGN Type	[a]Label	Source Number	[b]Catalog(s)
compact	1	377	1, 2, 3
FRI	2	211	1, 3
FRII	3	431	1, 3, 4
BT	4	307	1, 3, 4, 5
XRG	5	84	6
RRG	6	32	5
Total	⋯	1442	⋯

Notes.
[a] Labels in this work of corresponding AGN types.
[b] Catalogs used to compose the LRG sample: (1) CoNFIG, (2) FR0CAT, (3) FRICAT, (4) FRIICAT, (5) Proctor (2011), and (6) Cheung (2007).

Table 3
1442 LRGs in the LRG Sample

Name (J2000)	R.A. (h)	Decl. (deg) (J2000)	z	$S_{1.4\,\mathrm{GHz}}^{\mathrm{NVSS}}$ (mJy)	[a]Catalog (Literature)	BH12	[b]Type (Literature)	[b]Type (CNN)	BT (Subtype)	Probability
J000140.17−003350.6	0.02783	−0.56406	0.2469	73.00	6	0	5	5	N/A	0.97044
J000330.73+002756.1	0.05854	0.46558			5	1	4	4	1	0.99282
J001247.57+004715.8	0.21321	0.78772	0.1480	58.60	4	1	3	3	N/A	1.00000
J002107.62−005531.4	0.35212	−0.92539	0.1080	111.90	4	1	3	3	N/A	0.99320
J002900.98−011341.7	0.48361	−1.22825	0.0830	282.80	3	1	1	1	N/A	1.00000
J003302.41−014956.6	0.55067	−1.83239	0.1301	87.00	6	0	5	5	N/A	0.93146
J003636.21+004853.4	0.61006	0.81483		280.00	6	0	5	5	N/A	0.95816
J003930.52−103218.6	0.65848	−10.53850	0.1290	23.70	3	1	2	2	N/A	1.00000
J004012.89+012542.1	0.67025	1.42836			5	0	4	3	N/A	0.95473
J004148.22−091703.1	0.69673	−9.28419	0.0530	53.00	3	1	2	2	N/A	0.96172
J004150.27−092548.8	0.69730	−9.43022			5	0	4	4	2	1.00000
J004152.10+002837.3	0.69781	0.47703			5	0	4	4	2	1.00000
J004326.80−105421.8	0.72411	−10.90606	0.1270	82.30	4	1	3	3	N/A	1.00000
J004530.46−004746.9	0.75846	−0.79636	0.1470	12.40	3	1	2	2	N/A	0.99945
J004542.10+002105.5	0.76169	0.35153		509.00	6	0	5	5	N/A	0.93325

Notes. The columns show (1) the source names (J2000), (2)–(4) the sky coordinates and redshifts (samples without redshift records are marked with N/A), (5) the 1.4 GHz flux densities obtained with the NVSS data, (6) the previous catalogs in literature used in this work, (7) flags of the sources (1 means included and 0 means not included in the BH12 sample), (8)–(9) the source types labeled in literature and by our CNN, (10) the labels for the BT sample (1 = FRI-like source, 2 = FRII-like source, and N/A = non-BT source), and (11) the probabilities of the classifications.
[a] Catalogs in the literature for radio galaxies: (1) CoNFIG, (2) FR0CAT, (3) FRICAT, (4) FRIICAT, (5) Proctor (2011), and (6) Cheung (2007).
[b] Labels of radio galaxy types in this work: (1) compact sources (FR0s are marked as 1_{F}), (2) typical FRIs, (3) typical FRIIs, (4) BTs, (5) XRGs, and (6) RRGs.

(This table is available in its entirety in machine-readable form.)

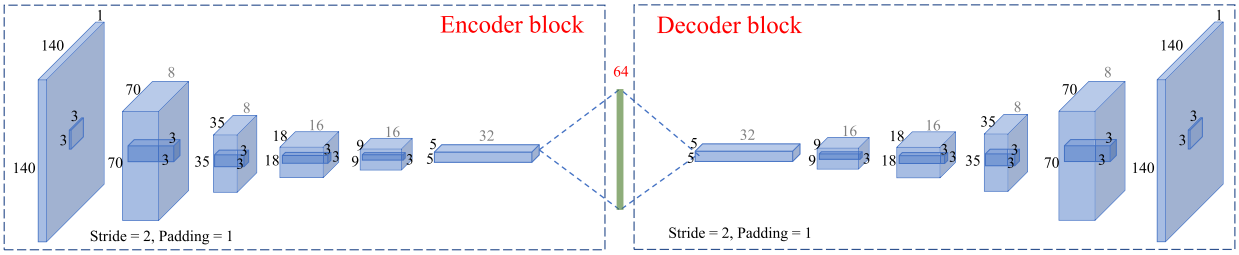
Table 4
14,245 Unlabeled Radio AGNs in the unLRG Sample

Name (J2000)	R.A. (h)	Decl. (deg) (J2000)	z	$S_{1.4\,\mathrm{GHz}}^{\mathrm{NVSS}}$ (mJy)	M_r^{SDSS} (mag)	[a]Type (CNN)	[b]BT (CNN)	[a]Type (MAN)	[b]BT (MAN)	Probability
J000001.57−092940.3	0.00044	−9.49453	0.1905	9.10	−22.45	1	N/A	1	N/A	0.99992
J000025.55−095752.8	0.00710	−9.96467	0.3540	14.10	−22.21	2	N/A	2	N/A	1.00000
J000027.89−010235.4	0.00775	−1.04317	0.4388	12.60	−23.26	1	N/A	1	N/A	1.00000
J000049.32−005042.9	0.01370	−0.84525	0.4110	22.70	−22.97	2	N/A	2	N/A	0.96863
J000052.92+003044.6	0.01470	0.51239	0.3398	15.70	−22.21	3	N/A	3	N/A	0.91637
J000052.99−011020.4	0.01472	−1.17233	0.1948	11.20	−22.26	2	N/A	3	N/A	0.95235
J000121.46−001140.3	0.02263	−0.19453	0.4615	110.90	−22.49	4	1	4	1	0.99931
J000121.53+010149.0	0.02265	1.03028	0.5520	18.10	−22.47	2	N/A	3	N/A	0.92581
J000134.89−085154.8	0.02636	−8.86522	0.1765	50.00	−22.52	3	N/A	3	N/A	0.99955
J000134.97−085727.7	0.02638	−8.95769	0.1737	5.00	−22.49	1	N/A	1	N/A	1.00000
J000140.21+010531.1	0.02784	1.09197	0.0611	6.70	−21.38	1	N/A	1	N/A	1.00000
J000140.36−010636.9	0.02788	−1.11025	0.2259	13.70	−21.97	1	N/A	1	N/A	1.00000
J000147.31−091723.5	0.02981	−9.28986	0.1726	33.30	−23.16	3	N/A	3	N/A	1.00000
J000157.24−103117.3	0.03257	−10.52147	0.2516	40.90	−22.75	1	N/A	1	N/A	0.99998
J000224.73−003252.7	0.04020	−0.54797	0.2895	17.80	−22.94	3	N/A	3	N/A	0.98935

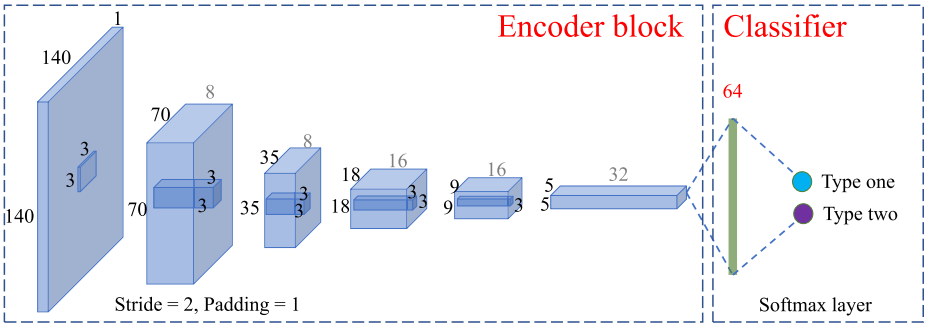
Notes. The columns show (1) the source names (J2000), (2)–(4) the sky coordinates and redshifts, (5) the 1.4 GHz flux densities obtained with the NVSS data, (6) the r-band absolute magnitudes obtained with the SDSS data, (7)–(10) the source types labeled by our CNN and our manual inspection, and (11) the probabilities of the classifications. Only the first 10 sources are listed here, and the full table is available electronically.
[a] Labels of radio galaxy types determined in this work with CNN (column 7) and manually (column 9): (1) compact sources (FR0s are marked as 1_{F}), (2) typical FRIs, (3) typical FRIIs, (4) BTs, (5) XRGs, and (6) RRGs.
[b] Labels of BT subtypes determined in this work with CNN (column 8) and manually (column 10): (1) FRI-like sources, (2) FRII-like sources, and (N/A) non-BT sources.

(This table is available in its entirety in machine-readable form.)

carried out their work based on DR7 (Abazajian et al. 2009) of the Sloan Digital Sky Survey (SDSS) spectroscopic sample, with the National Radio Astronomy Observatory VLA Sky Survey (NVSS; Condon et al. 1998) and the FIRST survey (Becker et al. 1995), after applying a flux density level of 5 mJy in the NVSS data (see also Best et al. 2005 and Donoso et al. 2009 for more



(a)



(b)

Figure 2. Schematic diagrams of (a) the CAE network used for the pre-training of the unLRG sample and (b) the neural network for the fine-tuning of the LRG sample. The convolutional layers are marked in blue, and the fully connected (FC) layer is marked in green. The image size (black), kernel size (black), and number (gray) of each convolutional layer and the length (red) of the FC layer are also marked.

technical details). We limit our selection to the BH12 sample sources by applying the following selection criteria: (1) the source is labeled an AGN according to its optical and radio properties, (2) the source is not overextended compared with the size of the image to be analyzed (140×140 pixels), (3) the 1.4 GHz flux density of the source is $\geqslant$5 mJy, and (4) the source is not included in the LRG sample. A maximum offset of $0\farcs96$ (Proctor 2011) is guaranteed when cross-matching the sky coordinates of the candidate sources with those of the LRG sources. As a result 14,245 sources, the redshifts and 1.4 GHz flux densities of which range across $0.01 < z < 0.70$ and $5.0\ \mathrm{mJy} < S_{1.4\ \mathrm{GHz}} < 2.3 \times 10^4$ mJy, respectively, are selected to construct the unLRG sample (see Table 4).

3. Autoencoder Based on a Deep CNN

3.1. Network Architecture

We have designed an autoencoder, namely MCRGNet (Morphological Classification of Radio Galaxy Network), which consists of an encoder block and a decoder block by convention (Yu & Deng 2014; Goodfellow et al. 2016), based on a deep CNN to learn the morphological features (i.e., the intrinsic representations) of radio AGN images and to precisely reconstruct the original input images by applying the learned image features.

Many of the state-of-the-art deep neural networks, e.g., AlexNet (Krizhevsky et al. 2012), VGG (Simonyan & Zisserman 2014), and GoogLeNet (Szegedy et al. 2015), have achieved good performance in image classification and object detection competitions (e.g., ILSVRC; Russakovsky et al. 2015). They have also been applied to different but related tasks through transfer learning (e.g., Pan & Yang 2010; Esteva et al. 2017). Note that although nearly all of the above networks have been trained on image sets drawn from ImageNet (Deng et al. 2009), which is composed of 14 million images of 1000 generic object classes spanning a wide range of subjects that cover many aspects of either human activity or life and natural phenomena on Earth (Esteva et al. 2017), no generic object class is directly related to celestial objects (Lukic et al. 2018). Therefore, negative transfer problems (Mahmud & Ray 2008; Pan & Yang 2010), e.g., dissimilar risks (see Rosenstein et al. 2005 for details), may be caused when these networks are directly applied to the task of radio AGN classification for transfer learning. In addition to this, all of these networks are composed of tens of layers, which contain millions of parameters; thus they must be run on workstations with multiple graphics processing units (GPUs) to be trained. In order to consume fewer device resources and enable relevant studies even on portable computers, we have decided to design a small-scale neural network instead of using the above large networks (e.g., Diamantis et al. 2018; Lukic et al. 2018).

The schematic architecture of the autoencoder is shown in Figure 2(a), which basically consists of several convolutional layers assigned with multiple 3×3 pixel kernels according to state-of-the-art neural network theories (e.g., AlexNet and VGG; Krizhevsky et al. 2012; Simonyan & Zisserman 2014; Ma et al. 2017). The encoder block functions via a feature extractor, which focuses on local image features from lower to higher orders through successive convolution operations performed in multidimensional (multichannel) convolutional layers and employs a classifier consisting of one fully connected layer to transform the derived feature maps into feature vectors to implement the classification task (see Figure 2(b)).

The decoder block is a mirror of the encoder block. It is constructed by reversing the positions of the convolutional and

Table 5
The Design of the CAE Network and the CNN Classifier

Layer	[a]Kernel Size	[b]Channels	[c]Padding	[d]Stride	[e]Subnet
Conv-En1	3 × 3	8	same	2	Encoder and CNN
Conv-En2	3 × 3	8	same	2	Encoder and CNN
Conv-En3	3 × 3	16	same	2	Encoder and CNN
Conv-En4	3 × 3	16	same	2	Encoder and CNN
Conv-En5	3 × 3	32	same	2	Encoder and CNN
Encode	64	⋯	⋯	⋯	Encoder and CNN
Softmax	2	⋯	⋯	⋯	CNN
Conv-De1	3 × 3	32	same	2	Decoder
Conv-De2	3 × 3	16	same	2	Decoder
Conv-De3	3 × 3	16	same	2	Decoder
Conv-De4	3 × 3	8	same	2	Decoder
Conv-De5	3 × 3	8	same	2	Decoder

Notes.
[a] Kernel sizes are set to 3 × 3 pixels according to AlexNet (Krizhevsky et al. 2012) and VGG (Simonyan & Zisserman 2014).
[b] Number of channels with respect to the layer.
[c] Padding modes for the convolution or deconvolution arithmetic.
[d] Stride modes for the convolution or deconvolution arithmetic.
[e] Subnet of the CAE to which the layer belongs.

fully connected layers in the encoder, and the layers in the decoder share the same structures of weights as their mirror counterparts in the encoder. The image reconstruction function of the decoder not only assists the unsupervised learning and training of the network parameters through analyzing the residuals of the model but also acts as a generator of radio AGN images when fed with proper feature vectors sampled from the probability distribution of the extracted features by the encoder. Similar architectures have been widely applied in a variety of areas, especially when simulating or generating data is important, such as in bottleneck training in speech recognition (e.g., Yu & Deng 2014), in the optimization of discriminator and generator in image simulation (e.g., Goodfellow et al. 2014), and in artistic-style learning (e.g., Radford et al. 2015).

Based on a series of tests, we embed five convolutional layers into the encoder block (Figure 2(a)) to hierarchically extract image features, so that the sizes of the input images reduce reasonably from the original 140 × 140 pixels in the first layer to 5 × 5 pixels in the fifth layer. After the fifth convolutional layer we use one fully connected layer, which is named the encode layer and is used to conduct the classifications, to further compress the derived features into a 64-dimension vector; including more fully connected layers will not significantly increase the performance of classification in this case (see Section 5.3 for a discussion). Immediately after each convolutional layer we apply a rectified linear unit to generate the activation function, which is nonsaturating and can help increase the convergence speed by truncating negative output into zero (Glorot et al. 2011). The number of kernels (i.e., the number of channels) of each convolutional layer and the corresponding kernel sizes are listed in Table 5.

Because the decoder block is designed as the reverse of the encoder block, we feed the output of the encoder (i.e., the extracted feature vector) into the decoder and compare the output of the latter, i.e., the reconstructed images, with the original images using the mean squared error to evaluate the performance of the autoencoder. Obviously, such an autoencoder can also be used to generate radio AGN images as a simulator by feeding the randomly obtained 64-length feature vectors sampled from the distributions of the extracted morphological features by the encoder into the decoder block (see, e.g., our recent work in Ma et al. 2018).

3.2. Training Strategy

Given that the size of the labeled sample is relatively small, we apply an effective training strategy that integrates the "pre-training" with the "fine-tuning" process (Hinton et al. 2012; Girshick et al. 2014; Christodoulidis et al. 2016) by adopting the following steps (Figure 3): (1) Pre-training—Apply the self-taught autoencoder to the unLRG sample (14,245 unlabeled radio AGNs) to pre-train a large number of weight and bias parameters of the deep CNN model, in order to deeply learn the representations of the radio AGN images and to avoid as much as possible the overfitting problem usually associated with learning on a small sample. (2) Fine-tuning—Append a softmax layer (e.g., Hinton et al. 2012) to the end of the encoder that has been pre-trained in the first step, and use it to classify the LRG sample (1442 radio AGNs) with an emphasis on the training of the fully connected layer, which is used to carry out classification. (3) Classification—Apply the fine-tuned encoder to classify the unLRG sample.

Our approach of training the neural network with the above three steps (PFC steps hereafter) has several advantages. First, when there is a lack of sufficiently large labeled samples (a sample containing $\gtrsim 10^4$ labeled radio AGNs may be necessary in our case, given the level of image complexity; see a similar case in LeCun et al. 2015 and the theoretical discussion in Esteva et al. 2017) for the training of a complex neural network, a straightforward training with small labeled samples does not help lower the risk of model overfitting and poor generalization capability, i.e., low predictive precision on new images. For example, the neural network is likely to overreact to minor, unimportant fluctuations in the training data and learn inessential random features. This problem is overcome by applying the PFC steps. Second, our approach enables one to perform the classification of a sample of about 15,000 radio galaxies efficiently and accurately in several seconds even on a portable computer equipped with a single GPU of moderate memory ($\lesssim$10 GB), not just on a workstation with multiple GPUs ($\gtrsim$100 GB), which other approaches usually require.

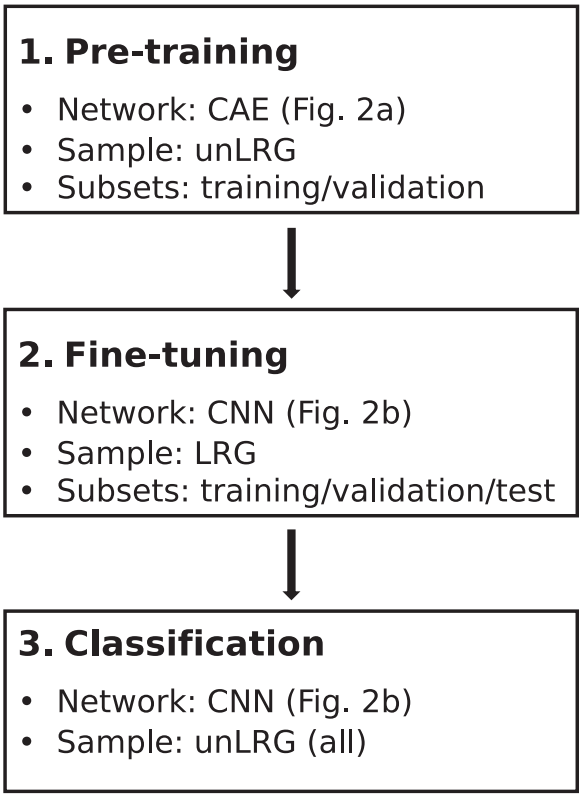


Figure 3. Framework of the PFC three-step training strategy.

This is because the resource consumption of the deep neural network in the encoder block of our autoencoder is significantly lower than that of other complex networks, e.g., GoogLeNet (Szegedy et al. 2015; Aniyan & Thorat 2017). This allows astronomers to inspect a newly obtained sample conveniently. In addition, our autoencoder can be used as a random-sample generator for morphology-based studies of radio galaxies.

3.3. Convolution and Deconvolution

3.3.1. Convolution

A convolution operation is usually a linear transformation that multiplies an input image with a kernel that slides across the input image to produce a matrix called an output feature map (Dumoulin & Visin 2016). For a CNN containing a set of L layers, each convolutional layer is assigned a set of kernels W_n^l and biases b^l to form $N_{\rm ch}^l$ channels, where $l = 1, 2, ..., L$ and $n = 1, 2, ..., N_{\rm ch}^l$ denote the ordinal numbers of the layers and channels, respectively. Thus for the nth channel in the lth layer of the network, the output feature map O_n^l can be written as

$$O_n^l = \begin{cases} W_n^l * I_0 + b_n^l, & \text{if } l = 1, \\ \sum_{m=1}^{N_{\rm ch}^{l-1}} W_n^l * I_m^l + b_n^l, & \text{if } l = 2, 3, ..., L, \end{cases} \tag{1}$$

where I_0 is the original image fed into the network, $I_m^l \equiv O_m^{l-1}$ ($l \geqslant 2$) is the mth input image ($m = 1, 2, ..., N_{\rm ch}^{l-1}$), the notation "*" is the convolution operator, and b_n^l is the bias.

Once an input is given, the corresponding output of a convolutional layer is uniquely determined by its kernels and biases and the choices of stride and zero-padding modes. For simplicity, in this work we always scale the images, as well as the kernels, as square matrices. The initial values assigned to the elements in the kernels are randomly generated with truncated normal Gaussian distribution, and corresponding biases are initialized with constants (Abadi et al. 2015). We select half-padding (i.e., same padding) $p = \lfloor k/2 \rfloor$, where k is the size of the squared kernel, to keep the shape and set the size of the input. We also set the stride, the step length to slide the kernel, to $s = 2$, which leads to a linear pooling (i.e., subsampling or downsampling), to decrease the number of free parameters and meanwhile extract the most remarkable features.

3.3.2. Deconvolution (Transposed Convolution)

The decoder block of the autoencoder can be regarded as a mirror of the encoder block, which can be derived by reversing the inputs and outputs of the CNN described in Section 3.3.1 in turn. Similar to the convolution operation (Equation (1)), deconvolution can be expressed as

$$\begin{cases} O_{n'}^{l'} = \sum_{m=1}^{N_{\rm ch}^{l'}} [W_m^{l'}]^T *^T I_m^{l'} + b_{n'}^{l'}, & \text{if } l' = 1, 2, ..., L-1, \\ O^{l'} = \sum_{m=1}^{N_{\rm ch}^{l'}} [W_m^{l'}]^T *^T I_m^{l'} + b^{l'}, & \text{if } l' = L, \end{cases} \tag{2}$$

for the n'th kernel channel ($n' = 1, 2, ..., N_{\rm ch}^{L-l'}$, and $n' \equiv 1$ when $l' = L$) in the l'th layer, where I_m^1 ($m = 1, 2, ..., N_{\rm ch}^L$) denotes the images of the first layer in the decoder block (i.e., the feature maps recovered from the Lth encode layer), $I_m^{l'} \equiv O_m^{l'-1}$ ($l' \geqslant 2$) is the mth input image of layer l' ($m = 1, 2, \ldots, N_{\rm ch}^{L-l'}$), $b_{n'}^{l'}$ is the bias, and the superscript T denotes the transpose arithmetic (Dumoulin & Visin 2016). A detailed explanation of the transpose convolution is illustrated in Appendix A with examples.

3.4. Feature Visualization with t-SNE

In order to evaluate the performance of our autoencoder, we visualize the classification of the LRG sample, which is achieved by analyzing the multidimensional feature vectors, in two-dimensional space (the visualization space) by applying the t-SNE algorithm, a dimension reduction algorithm that can preserve the similarities between vector points (i.e., distances) when mapping data from high-dimensional space onto lower dimensions (e.g., Van Der Maaten & Hinton 2008; Esteva et al. 2017).

To be specific, we use a set of multidimensional vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{N_{\rm LRG}}\}$, where $N_{\rm LRG}$ is the number of the LRG sample. These vectors form a distribution $\mathcal{P}$, in which p_{ij} ($i, j = 1, 2, ..., N_{\rm LRG}$) denotes the probability of similarity between any two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. If we further use $p_{j|i}$ to define the conditional probability of $\boldsymbol{x}_j$ when $\boldsymbol{x}_i$ is given, by assuming the Gaussian distribution, we have

$$p_{j|i} = \frac{\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2/2\sigma_i^2)}, \tag{3}$$

where σ_i^2 is the variance of $\boldsymbol{x}_i$ (see Hinton & Roweis 2002 for details), and

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N_{\rm LRG}} \tag{4}$$

by symmetry.

Similarly, in the visualization space we use a vector set $\{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{N_{\rm LRG}}\}$ to describe the distribution $\mathcal{Q}$ for the classified radio AGNs and q_{ij} to represent the probabilities of similarities between any two vectors. As shown by Van Der Maaten & Hinton (2008), q_{ij} appropriately preserves the similarities between the vectors in high-dimensional space, i.e., p_{ij}. Unlike the case of the distribution $\mathcal{P}$, for which Gaussian distribution is assumed, Van Der Maaten & Hinton (2008) employed a long-tailed Student's t-distribution with one degree of freedom to describe $\mathcal{Q}$ to avoid abnormal points, which is written as

$$q_{ij} = \frac{(1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)^{-1}}{\sum_{k\neq m}(1 + \|\boldsymbol{y}_k - \boldsymbol{y}_m\|^2)^{-1}}. \tag{5}$$

By applying the Kullback–Leibler divergence (Kullback & Leibler 1951)

$$\mathrm{KL}(\mathcal{P}||\mathcal{Q}) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{6}$$

as the objective function, we can evaluate the similarity between distributions $\mathcal{P}$ and $\mathcal{Q}$ and locate $\boldsymbol{y}_i$ by minimizing Equation (6) with respect to $\boldsymbol{y}_i$ with

$$\frac{\partial \mathrm{KL}}{\partial \boldsymbol{y}_i} = 4\sum_j (p_{ij} - q_{ij})(\boldsymbol{y}_i - \boldsymbol{y}_j)(1 + (\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2)^{-1}. \tag{7}$$

4. Application of MCRGNet to Radio AGN Samples

4.1. Preparation of the Input Images

We apply an approach that consists of the following three steps to prepare both the labeled and unlabeled images before we feed them into the autoencoder. First, we define a region of interest (ROI) in each image by cropping a box region with a size of 140 × 140 pixels (4.′2 × 4.′2) from the original 150 × 150 pixel (4.′5 × 4.′5) image. Next, in order to improve the contrast of the ROI of each image, we carry out sigma clipping to limit the image dynamic range by setting the parameter σ to 3 and the number of iterations to 50, and discard pixels dominated by background noise. Finally, in order to further avoid overfitting of the network, we perform an augmentation operation to increase the number of images in both the LRG and unLRG samples by applying the following operations to each image:

1. Randomly flip the sigma-clipped image. The flipping mode is either left-to-right, top-to-bottom, or diagonal.
2. Rotate the flipped image with an angle θ generated randomly between 0° and 360°.

For each unLRG image, the number of augmentation operations is 10, which results in a total of around 1.43×10^5 images for the pre-training. For each LRG image, however, the number of augmentation operations is determined according to the morphological type of the radio AGN, in order to balance the numbers of different morphological types in the dichotomous classification tree. Given that on the classification tree (Figure 1) the two subbranches of any branch should have nearly the same AGN numbers, the ratios of the numbers of compact, FRI, FRII, BT, XRG, and RRG types are set to 8:2:2:2:1:1 after the augmentation operation, which results in about 3.9×10^4 images in total (38,882, 19,554, 9731, 9823, 4900, and 4923 images for the binary classification branches compact–extended, typical–atypical FR, typical FRI–FRII, BT–irregular, FRI-like–FRII-like, and XRG–RRG, respectively; Table 6). One example of the flipping and augmentation operations is displayed in Figure 4.

Table 6
Augmentations of the LRG Sample

AGN Type	Label	[a]$N_{\rm Trn+Val}$	[b]$N_{\rm Tst}$	[c]$R_{\rm Aug}$	[d]$N_{\rm Aug}$	[e]Branch
compact	1	302	75	64	19,328	1
FRI	2	169	42	29	4901	1, 2, 3
FRII	3	345	86	14	4830	1, 2, 3
BT	4	245	62	20	4900	1, 2, 4, 5
XRG	5	67	17	37	2479	1, 2, 4, 6
RRG	6	26	6	94	2444	1, 2, 4, 6
Total	⋯	1154	288	⋯	38,882	⋯

Notes.
[a] Number of sources in the training (Trn) + validation (Val) subsets.
[b] Number of sources in the test (Tst) subsets.
[c] Augmentation rates for the corresponding radio galaxy types.
[d] Number of augmented images for the corresponding radio galaxy types.
[e] Branches on the dichotomous tree: (1) compact–extended, (2) typical–atypical FR, (3) typical FRI–FRII, (4) BT–irregular, (5) FRI-like–FRII-like, and (6) XRG–RRG.

The augmented unLRG sample is then divided into independent training and validation subsets, which are used to optimize the parameters of the neural network and to evaluate the learning performance (i.e., the ability to reconstruct the images from which features are learned) to avoid the overfitting risk for each epoch, respectively. The training and validation subsets are divided at the ratio of 4:1; therefore the two subsets have about 1.14×10^5 and 2.85×10^4 augmented images, respectively.

In the original LRG sample, the numbers of fine-tuning images (i.e., training and validation images) and testing images are set to 1154 (80%) and 288 (20%), respectively. Data augmentation is performed only on the images used for fine-tuning, which are divided with a ratio of 4:1 to obtain 31,106 training images and 7776 validation images (Table 6). Note that the subset of testing images is used for evaluating the classification performance (i.e., the ability to classify radio AGNs correctly) of the trained network; therefore it contains only the original LRG images that are not augmented.

4.2. Training and Evaluation of the Neural Network

Following the strategy described in Section 3.2, we first feed the unLRG training and validation subsets into the autoencoder to pre-train the weight and bias parameters, and then input the LRG training, validation, and test subsets into the neural network, which is complemented with a softmax layer at the end of the encoder block, to fine-tune the network parameters with an emphasis on the fully connected encode layer.

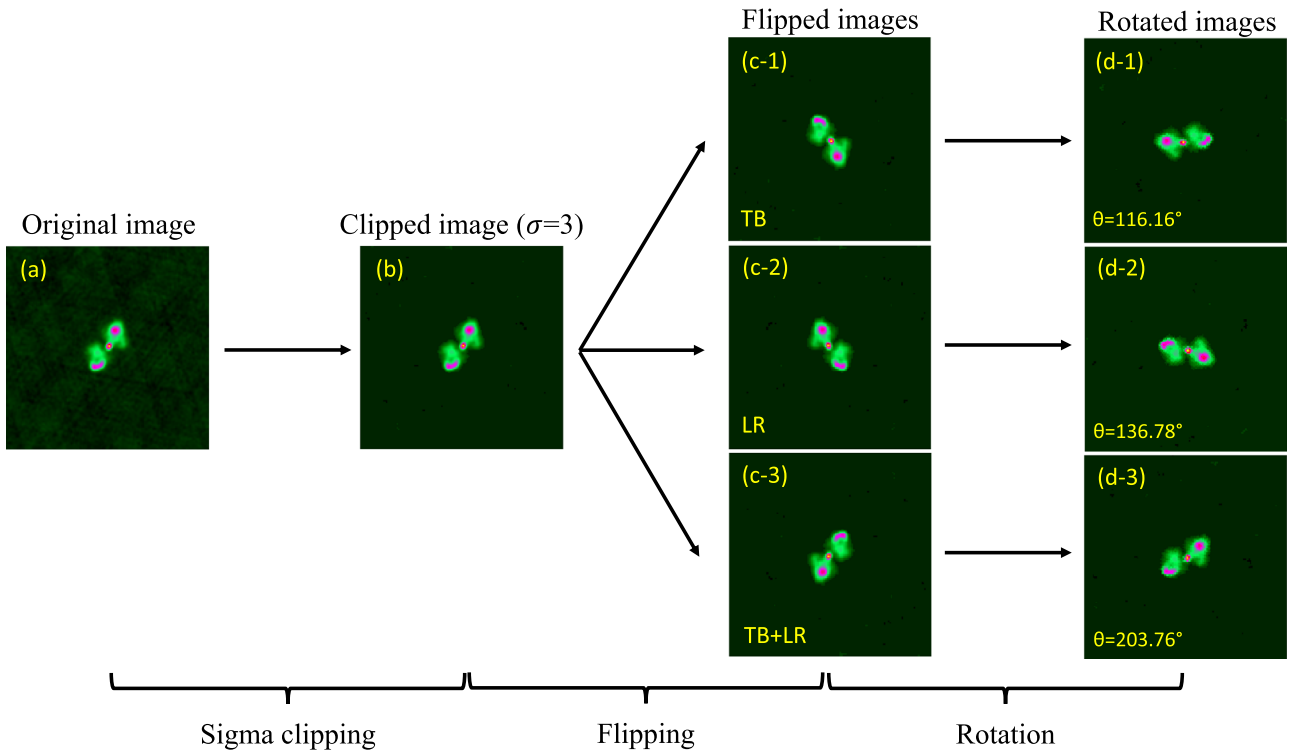


Figure 4. An example from the augmented radio galaxies, i.e., J145941.95+290331.8 (XRG), which is located at $z = 0.146$. After sigma clipping, the image is randomly flipped either top-to-bottom (TB), left-to-right (LR), or both (TB+LR) and then randomly rotated with an angle in [0, π).

4.2.1. Approach to Training and Evaluation

In this work we pre-train and fine-tune the neural network using the strategy of batch iteration (Abadi et al. 2015). Instead of using the whole training data set, we separate it into batches (100 images for each) and input them into the network batch by batch to adjust the network parameters iteratively. One round of training with all the batches is called an epoch.

For the pre-training stage the augmented images in the unLRG training or validation subset are fed into the encoder block in batches to generate 64-dimension feature vectors, which are successively fed as the input of the decoder block to reconstruct the images. We compare the reconstructed images with the corresponding original input images by calculating the mean squared residual error between them as the batch loss of pre-training, which is

$$R_{\rm P} = \frac{1}{N_{\rm B}} \sum_{i=1}^{N_{\rm B}} \sum_{j=1}^{N_{\rm row}} \sum_{k=1}^{N_{\rm col}} |I_{i,j,k}^{\rm CAE} - O_{i,j,k}^{\rm CAE}|^2, \quad (8)$$

where $N_{\rm B}$ is the number of images in one training batch, $N_{\rm row} = N_{\rm col} = 140$ are the number of rows and columns of the images, $I^{\rm CAE}$ represents the input images of the encoder block, and $O^{\rm CAE}$ represents the reconstructed images of the decoder block. According to Goodfellow et al. (2016), the loss of each batch is back-propagated to the network along the gradient decreasing direction to adjust network parameters until the loss converges to a constant value. This is achieved after about 500 epochs of pre-training are completed.

For the fine-tuning stage, a softmax layer is appended to the end of the encoder block, and the batch loss of fine-tuning $R_{\rm F}$ is defined by averaging the cross-entropy of the augmented LRG images (Goodfellow et al. 2016) as

$$R_{\rm F} = -\frac{1}{N_{\rm B}} \sum_{i=1}^{N_{\rm B}} \sum_{j=1}^{N_c} y_{i,j} \cdot \log_2 p_{i,j}^{S}, \quad (9)$$

where $N_c = 2$ is the number of subbranches for each branch on the dichotomous classification tree and $y_{i,j}$ ($i = 1, 2, ..., N_{\rm B}$, and $j = 1, 2, ..., N_c$; $y_{i,j} = 1$ if the source belongs to type j, and $y_{i,j} = 0$ if otherwise) is the one-hot probability of the source. $p_{i,j}^{S}$ represents an output of the softmax layer, which is the normalized probability of the source being classified as type j in a certain batch. The batch loss converges to a constant after about 200 epochs of the fine-tuning are completed.

In the pre-training and fine-tuning stages we have employed the following standard rules.

1. Dropout of weights with values approaching zero (Srivastava et al. 2014), which helps avoid overfitting, is performed each time after the calculation in a convolutional or a fully connected layer is completed. The dropout rate is fixed at 50% for each layer.
2. In the pre-training stage the number of epochs is set to 500, which is determined by the convergence of training curves. A high initial learning rate of 0.0005 together with a decay rate of 0.95 to form the adaptive learning rate (Zeiler 2012; Abadi et al. 2015) and an adaptive moment optimization function (Kingma & Ba 2014) are both adopted. In the fine-tuning stage the number of epochs of each subclassifier is set to 200 also according to the convergence of the training curves, and a lower initial learning rate of 0.0001 is adopted to carefully fine-tune the parameters.
3. In either the pre-training or the fine-tuning stage, the corresponding validation loss $R_{\rm P}$ or $R_{\rm F}$, which is obtained

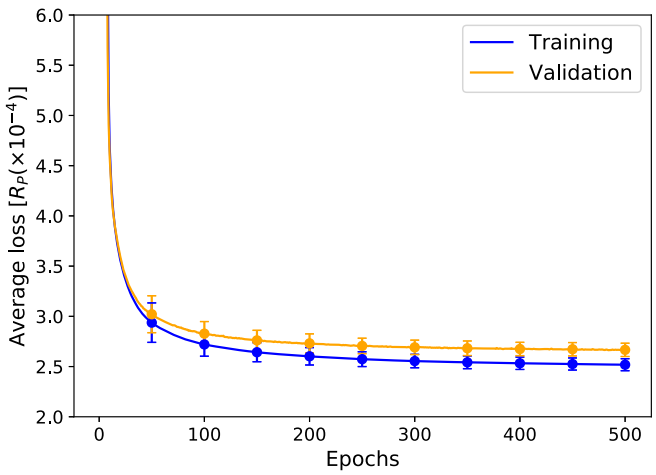


Figure 5. Average loss R_P calculated for the pre-training of the autoencoder network.

by feeding the unLRG or LRG validation subset into the autoencoder or CNN, is used to evaluate the overfitting risks for each epoch. If the validation loss does not decrease or even increase while the training loss decreases steadily to convergence, the network parameters may be overfitted. Note that the validation loss is not used for parameter optimizations.

4.2.2. Results of Training and Evaluation

In Figures 5 and 6, we illustrate the training and validation loss curves obtained in the pre-training stage and fine-tuning stage, respectively. For comparison, in the fine-tuning stage we also plot the loss curves obtained when the pre-training process is switched off for comparison as well as the classification accuracy $R_{\rm acc}$, which is defined as the ratio of the number of correctly classified sources to the actual source number (see also Equation (11)). To evaluate the robustness of the network, each training process has been conducted 10 times, and the means and 68% confidence limits of the losses and accuracies are calculated for every 25 epochs (see Figure 6). In either the pre-training stage or the fine-tuning of all the subclassifiers, both the training loss and the validation loss decrease monotonically, and the training loss becomes lower than the corresponding validation loss after about 10 epochs, and this remains unchanged until the pre-training or fine-tuning is completed. These indicate that the trained parameters in the neural network are not overfitted.

In the fine-tuning stage, except for the typical–atypical FR branch, the average fine-tuning loss $R_{\rm F}$ decreases significantly faster, and the neural network achieves better classification accuracies after pre-training is applied. As for the typical–atypical FR branch, because (1) the sources therein are comparatively easier to classify due to their distinctive morphological features and (2) the number of images in the training subset is 19,554, the pre-training becomes more or less insignificant.

In order to further evaluate the performance of the neural network, we have employed three indices, i.e., sensitivity $R^t_{\rm sen}$, accuracy $R^t_{\rm acc}$, and precision $R^t_{\rm pre}$, to describe the classifier performance for the LRG samples (Khalaf et al. 2016; Ma et al. 2017). They are defined as follows:

$$R^t_{\rm sen} = N^t_{\rm TP}/(N^t_{\rm TP} + N^t_{\rm FN}), \quad (10)$$

$$R^t_{\rm acc} = (N^t_{\rm TP} + N^t_{\rm TN})/N_{\rm LRG}, \quad (11)$$

$$R^t_{\rm pre} = N^t_{\rm TP}/(N^t_{\rm TP} + N^t_{\rm FP}), \quad (12)$$

where the superscript t denotes the morphological type ($t = 1, 2, \cdots, 6$, corresponding to compact, typical FRI, typical FRII, BT, XRG, and RRG, respectively), $N_{\rm LRG}$ represents the total number of radio AGNs in the original LRG sample, $N^t_{\rm TP}$ ($N^t_{\rm TN}$) is the number of type t (non–type t) radio AGNs that are correctly classified (T), and $N^t_{\rm FN}$ ($N^t_{\rm FP}$) is the number of type t (non–type t) radio AGNs that are misclassified (F). We list the calculated indices in Table 7 with 68% error bars and find that our neural network can achieve a satisfactory classification. The classifications of compact, typical FRI, typical FRII, and BT types are very accurate, although the classifications of XRG and RRG types are relatively poor, possibly due to insufficient training/learning caused by their small sample scale.

The results obtained by the no pre-training strategy are also listed in Table 7 for comparison. The averaged precision gain between pre-training and no pre-training is 7.77% ± 3.02%, which indicates that pre-training indeed improves the network performance. Besides, we show the confusion matrices of classifying the LRG sample with and without pre-training strategies in Figure 7 for visualizing the difference between their performance. Though the gains in precision and sensitivity after applying pre-training on the AGN types with large source numbers (e.g., compact, FRI, etc.) might be relatively insignificant, they are significant for the XRG and RRG sources of small numbers, especially the gains in sensitivity. The same conclusion could also be drawn in the classification of the unLRG sample (Figure 8; see Section 4.3 for details).

We apply MCRGNet pre-trained by the proposed PFC strategy on the LRG sample and list the results, including the corresponding 68% errors, in Table 7, along with the results obtained without applying pre-training. The corresponding averaged gains of the indices obtained after carrying out pre-training for each subtype are also shown in the same table. Compared with the results of Aniyan & Thorat (2017), who trained their neural network with a relatively small labeled sample (716 sources), the classification precisions of both our typical FRI and FRII sources are higher (95% versus 91% and 92% versus 75%), although that of BT sources is lower (87% versus 95%). Aniyan & Thorat (2017) also reported that the recall (i.e., the index $R_{\rm sen}$ in this paper), which reflected the ability to avoid misclassification due to the incorrect inclusion of sources of other types, of the BT sources obtained with their neural network was 79%. They argued that this was because some sources in the BT sample were misclassified as FRIIs, which was supported by the high recall but low precision of FRII sources. As can be seen in Table 7, the sensitivities $R_{\rm sen}$ for the BT and FRII sources derived by our approach are both >91%. This, combined with the corresponding high precisions, suggests an improved classification of BT and typical FRII radio AGNs.

4.2.3. Feature Visualization of the LRG Sample

We show the results of the dimension-reduced feature visualization of the LRG sample classification with t-SNE in

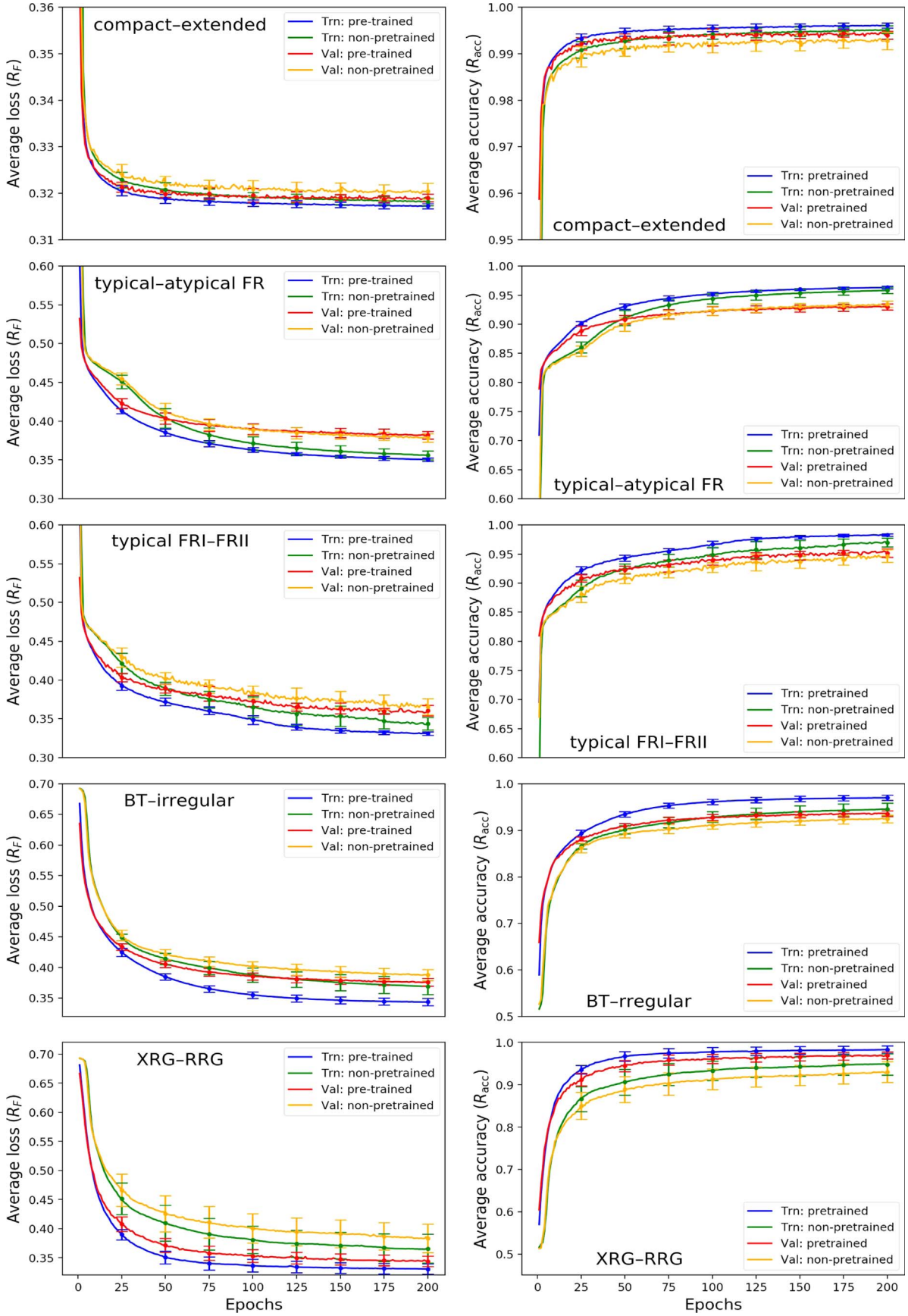


Figure 6. Average loss R_F (left) and accuracy R_{acc} (right) calculated for the five dichotomous classification branches, i.e., compact–extended, typical–atypical FR, typical FRI–FRII, BT–irregular, and XRG–RRG. Results for both training and validation subsets are presented with and without pre-training (Section 4.1).

Table 7
Evaluation of the CNN Based on the Classification of the LRG Sample

AGN Type	Pre-trained			Non-pre-trained			Gain		
	R_{sen}	R_{acc} (%)	R_{pre}	R_{sen}	R_{acc} (%)	R_{pre}	R_{sen}	R_{acc} (%)	R_{pre}
compact	98.41 ± 0.25	98.96 ± 0.16	97.88 ± 0.63	98.14 ± 0.27	98.82 ± 0.17	97.12 ± 0.67	0.27 ± 0.17	0.14 ± 0.13	0.76 ± 0.58
FRI	88.72 ± 1.69	96.94 ± 0.48	95.39 ± 2.73	83.46 ± 2.13	95.42 ± 0.56	93.94 ± 2.74	5.26 ± 1.32	1.52 ± 0.57	1.45 ± 2.80
FRII	91.07 ± 2.12	95.66 ± 0.39	92.24 ± 1.01	80.77 ± 1.97	88.49 ± 0.46	88.41 ± 1.48	10.90 ± 1.52	7.27 ± 0.33	3.83 ± 1.27
BT	94.27 ± 1.89	95.25 ± 0.46	86.56 ± 0.92	85.57 ± 2.61	89.88 ± 0.52	76.57 ± 1.88	8.70 ± 1.32	5.37 ± 0.43	9.99 ± 1.50
XRG	75.05 ± 3.54	96.72 ± 0.34	72.34 ± 2.70	64.29 ± 4.78	91.19 ± 0.53	55.76 ± 4.27	10.78 ± 2.44	5.61 ± 0.58	16.58 ± 4.32
RRG	74.62 ± 3.64	97.04 ± 0.30	65.89 ± 3.65	70.42 ± 6.47	95.15 ± 0.28	60.21 ± 5.88	4.20 ± 4.52	1.89 ± 0.45	5.68 ± 7.66
Total	⋯	⋯	92.44 ± 1.94	⋯	⋯	84.67 ± 2.82	⋯	⋯	7.77 ± 3.02

Note. R_{sen}, R_{acc}, and R_{pre} are the evaluation indices, namely the sensitivity, accuracy, and precision, of the corresponding type.

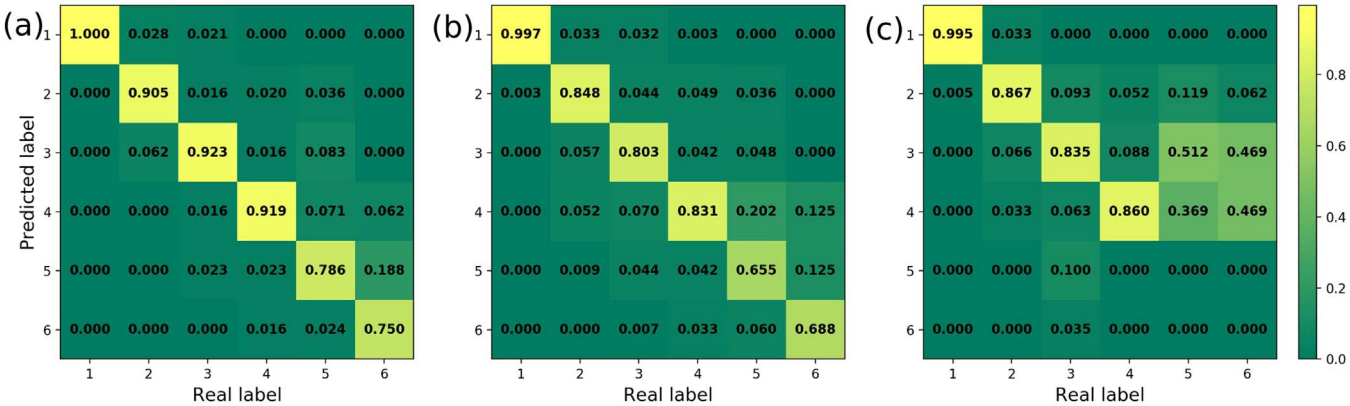


Figure 7. Confusion matrices obtained for the LRG sample by training the CNN with (a) and without (b) pre-training using the tree strategy and (c) by applying the non-tree strategy. Labels 1 to 6 are for compact, FRI, FRII, BT, XRG, and RRG sources, respectively. The numbers in each column are normalized.

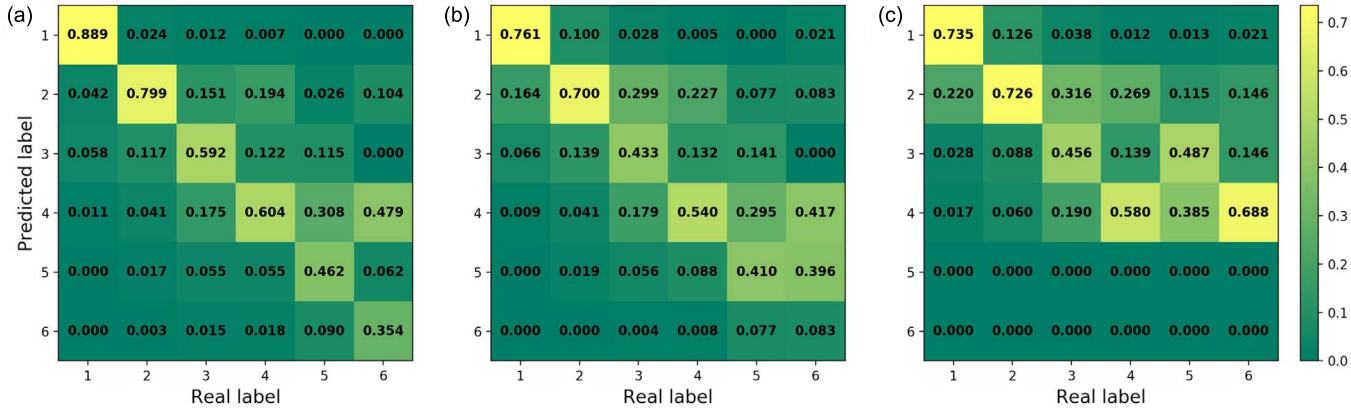


Figure 8. Confusion matrices obtained for the unLRG sample by training the CNN with (a) and without (b) pre-training using the tree strategy and (c) by applying the non-tree strategy. Labels 1 to 6 are for compact, FRI, FRII, BT, XRG, and RRG sources, respectively. The numbers in each column are normalized.

Figure 9, in which the subsamples of different types are marked with different colors. We find that the points of each type gather into a cluster that overlaps only slightly with those of other types, except for the case between FRIIs and XRGs and the case between compacts and FRIs. By carrying out visual inspection, we find that the FRII–XRG confusion is caused when the minor pair of jets in the XRG is dominated by the major pair, and the compact–FRI misclassification is caused when an FRI shows small-scale and weak jets.

To show the merit of using a dichotomous classification tree structure (Figure 1), we have trained a CNN which shares the same structure as the encoder block illustrated in Figure 2(a) but has a six-node softmax layer as a classifier and used it to directly classify the LRG images into the six types. As shown in Figure 9(c) the result is unacceptable, because the points of the different types overlap with one another severely. Based on this we may conclude that, compared with direct six-type (non-tree) classification, the application of a dichotomous tree leads to significantly better performance.

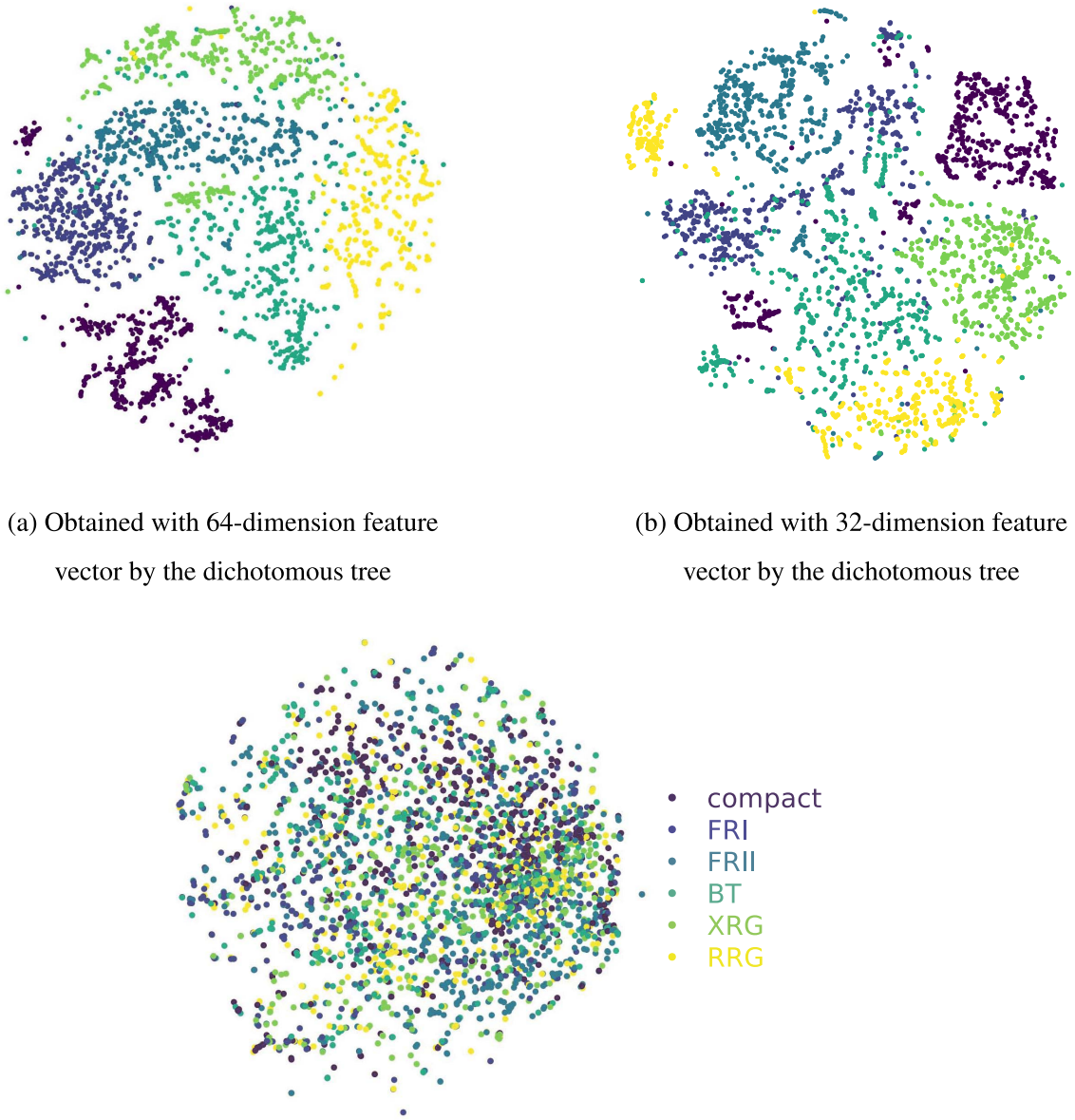
In Figure 7 we show the 6×6 confusion matrices of the LRG classification obtained with and without pre-training and that obtained when direct six-type classification is performed. Compared to the dichotomous tree based methods, the direct training leads to the worst results, especially for XRG and RRG sources, whose numbers are relatively small. The same conclusion could also be found in Figure 8 on the classification of the unLRG sample (see Section 4.3 for details).

4.3. Classification of the unLRG Sample

After deeply training the neural network with both unlabeled and labeled samples, we classify 14,245 radio AGNs in the unLRG sample and list the results in Tables 4 and 8. In addition to the CNN classification, for each source we have also estimated the joint probability of the classification by cascadingly multiplying the probabilities of the nodes decided and given by the softmax layer at the corresponding branches (Goodfellow et al. 2016; see Appendix B for details). Furthermore, if a source is classified as a BT, a subtype label (1 for FRI-like sources or 2 for FRII-like sources) is further presented, which is obtained by the CNN trained at the typical FRI–FRII branch.

In order to cross-check the results, we have also visually inspected all the unLRG images and classified the sample manually. We divide the images into three parts, and each part is inspected by at least three of the authors independently. If the three authors arrive at the same result, the visual classification is determined. Otherwise, an additional two authors will join, and the visual classification is determined when at least three authors agree on the result. The visual classification is also listed in Table 4.

We find that in general the CNN classification agrees with the human classification. About 80% of the sources have the same CNN and manual labels, and this ratio is even higher for compact sources (89%) and typical FRI sources (80%). About 3% of the small-scale FRIIs are misclassified by the neural



(a) Obtained with 64-dimension feature vector by the dichotomous tree

(b) Obtained with 32-dimension feature vector by the dichotomous tree

(c) Obtained with 64-dimension feature vector by non-tree

Figure 9. (a)–(b) Results of the *t*-SNE feature visualization for the classifications of the LRG samples using the dichotomous tree structure shown in Figure 1. (c) Direct six-type classification. Different types of sources are marked in different colors. Sixty-four-dimension length feature vectors are employed in (a) and (c), while 32-dimension length feature vectors are used in (b).

network as FRIs, due to the small distances between the two lobes and their weak fluxes. About 7% of the typical FRs are misclassified by the neural network as BTs, XRGs, and RRGs, because these FRs tend to have ambiguous morphologies and low PAPRs ($\leqslant$10.0 dB). In most cases, a low or relatively low ($\lesssim$0.90) classification probability is found for the sources misclassified by the neural network. A further discussion on this issue will be presented in Section 5.2.

5. Discussion

5.1. A Comparison between FRI and FRII Radio Galaxies

Fanaroff & Riley (1974) proposed a clear radio power transition between FRIs and FRIIs based on their study of the 3C (Mackay 1971) radio sources. Ledlow & Owen (1996) suggested that the transition was a function depending on optical luminosity and was most likely a result of the different lifetimes for sources with different radio powers. Subsequent studies (Best 2009—1083 FRIs and FRIIs; Wing & Blanton 2011—1228 FRIs and 683 FRIIs; Capetti et al. 2017a—122 FRIIs), however, do not support the claim that FRIs and FRIIs reside in distinctly different regions in the radio luminosity–optical magnitude plane ($L_{\rm radio}-M_{\rm opt}$ plane). With the large sample of radio AGNs classified in this work, which includes 5048 FRIs and 2083 FRIIs, it will be very interesting to investigate to what extent the distributions of FRIs and FRIIs differ in the $L_{\rm radio}-M_{\rm opt}$ plane.

The 1.4 GHz radio luminosities of the sources in our sample are calculated with the flux densities measured at 1.4 GHz, as listed in Table 1 of Best & Heckman (2012). The optical

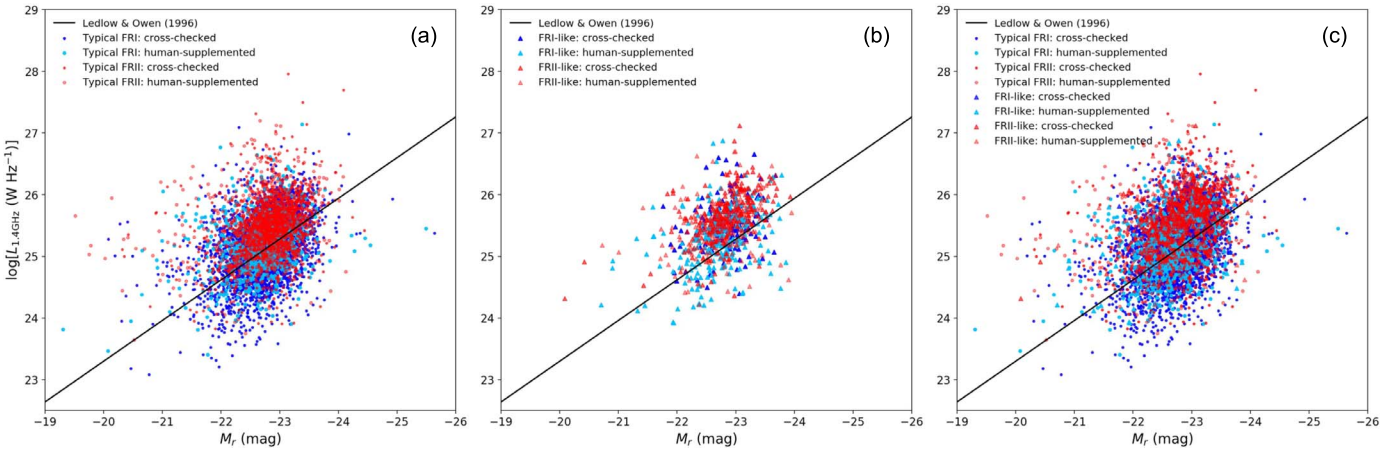


Figure 10. Joint distributions of the unLRG FR samples between 1.4 GHz luminosities (log[$L_{1.4\,\mathrm{GHz}}$]) and r-band absolute magnitudes (M_r). (a) Distributions of 4939 typical FRI sources (including 3953 sources cross-checked between the MCRGNet prediction and human inspection; the rest are only labeled by humans) and 2009 typical FRIIs (1547 sources cross-checked). (b) Same as (a) but for 420 FRI-like sources (143 sources cross-checked) and 440 FRII-like sources (238 sources cross-checked). (c) Combination of (a) and (b). The solid black line indicates the division of FRI–FRII reported by Ledlow & Owen (1996), to which corrections of magnitudes have been made following Capetti et al. (2017b).

Table 8
Evaluation of the CNN Based on Comparison between the Results Obtained with the CNN and Manual Classification of the unLRG Sample

AGN Type	[a]N_{CNN}	[b]N_{crt}	[c]N_{man}	[d]R_{pre} (%)
compact	5572	5422	6102	88.86
FRI	4781	4032	5048	79.87
FRII	2299	1233	2083	59.19
BT	1218	535	886	60.38
XRG	290	36	78	46.15
RRG	85	17	48	35.52
Total	14,245	11,275	14,245	79.15

Notes.
[a] Number of sources classified by the CNN for the corresponding types.
[b] Number of sources correctly classified by the CNN.
[c] Number of sources manually classified for the corresponding types.
[d] Precisions of the CNN classifications calculated by comparing with human inspections.

r-band luminosities are calculated by utilizing photometric parameters and spectroscopic redshifts drawn from SDSS (York et al. 2000). As recommended by SDSS, we adopt the cmodel magnitude, one of the magnitude measurements provided in the SDSS database, to estimate the optical fluxes of radio AGNs. By using the SDSS photometric pipeline to combine the de Vaucouleurs and exponential models (into a linearly composite model or cmodel) to obtain best fits of the SDSS images, we obtain the composite flux of each source as follows:

$$F_{\mathrm{composite}} = f_{\mathrm{deV}} \times F_{\mathrm{deV}} + (1 - f_{\mathrm{deV}}) \times F_{\mathrm{exp}}, \tag{13}$$

where f_{deV} is the coefficient of the de Vaucouleurs model; meanwhile F_{deV} and F_{exp} are the best-fit fluxes of the de Vaucouleurs and exponential models, respectively. With the apparent r-band magnitude m_r, which is calculated from $F_{\mathrm{composite}}$, we have the absolute r-band magnitude M_r:

$$M_r = m_r - 5\log\left(\frac{D_L}{\mathrm{pc}}\right) + 5 - A - K, \tag{14}$$

where D_L is the luminosity distance, A is the correction for the Galactic extinction derived with the map of Schlegel et al. (1998), and K is the k-correction calculated following Blanton & Roweis (2007).

In Figure 10 we show the distributions of typical FRI (blue and cyan circles), typical FRII (red and coral circles), FRI-like (blue and cyan triangles), and FRII-like (red and coral triangles) sources in our sample on the L_{radio}–M_{opt} plane, where the classification based on the cross-check between the MCRGNet prediction and human inspection is marked in blue and red and the classification purely made by humans (i.e., for sources not correctly classified by MCRGNet, which are termed "human-supplemented") is marked in cyan and coral. For the cross-checked sources, a probability threshold of 0.90 is applied to filter those sources with smaller predicted probabilities (i.e., sources with high uncertainties; see Section 5.4 for a detailed discussion).

The corresponding number distributions as a function of either $L_{1.4\,\mathrm{GHz}}$ or M_r, as well as their best Gaussian fits, are plotted in the lower panel of Figure 11 for comparison. We find that at any given M_r the distribution of typical FRIIs completely overlaps that of typical FRIs, although the latter extends to lower radio luminosities by about one order of magnitude. Meanwhile the FRI-likes and FRII-likes tend to share the same distribution. In addition to this, we find that the division line between FRIs and FRIIs that was reported in Ledlow & Owen (1996) crosses either of these distributions approximately in the middle. These results, which are obtained with a sufficiently large sample, confirm those reported in previous studies (Best 2009; Lin et al. 2010; Wing & Blanton 2011; Capetti et al. 2017a) with good consistency, indicating that an abrupt separation between FRIs and FRIIs in the L_{radio}–M_{opt} plane cannot be reproduced. A further discussion on this issue with an additional criterion on the samples will be presented in Section 5.4.

In order to search for the possible redshift evolution of the L_{radio}–M_{opt} distributions, we have divided the sources into $z \leqslant 0.3$ and $0.3 < z \leqslant 0.7$ subgroups and plot their distributions in Figure 12, where the markers and colors of the sources are the same as in Figure 10. We find that roughly the same conclusion as above can be drawn for either of the two redshift

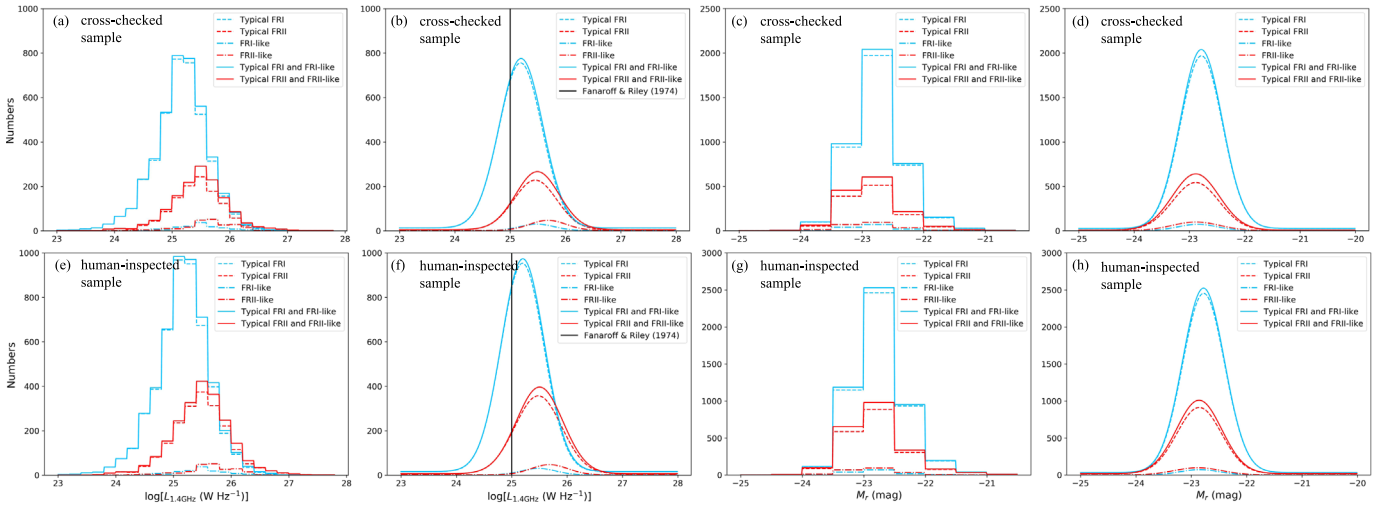


Figure 11. Binned distributions and corresponding best Gaussian models of the unLRG FR samples as a function of the 1.4 GHz luminosity (log[$L_{1.4\,\mathrm{GHz}}$]) or the r-band absolute magnitude (M_r). (a)–(d) are for the classification based on a cross-check between the MCRGNet prediction and human inspection, and (e)–(h) are for the classification based on the human-inspected classification. The black solid line indicates the division reported by Fanaroff & Riley (1974).

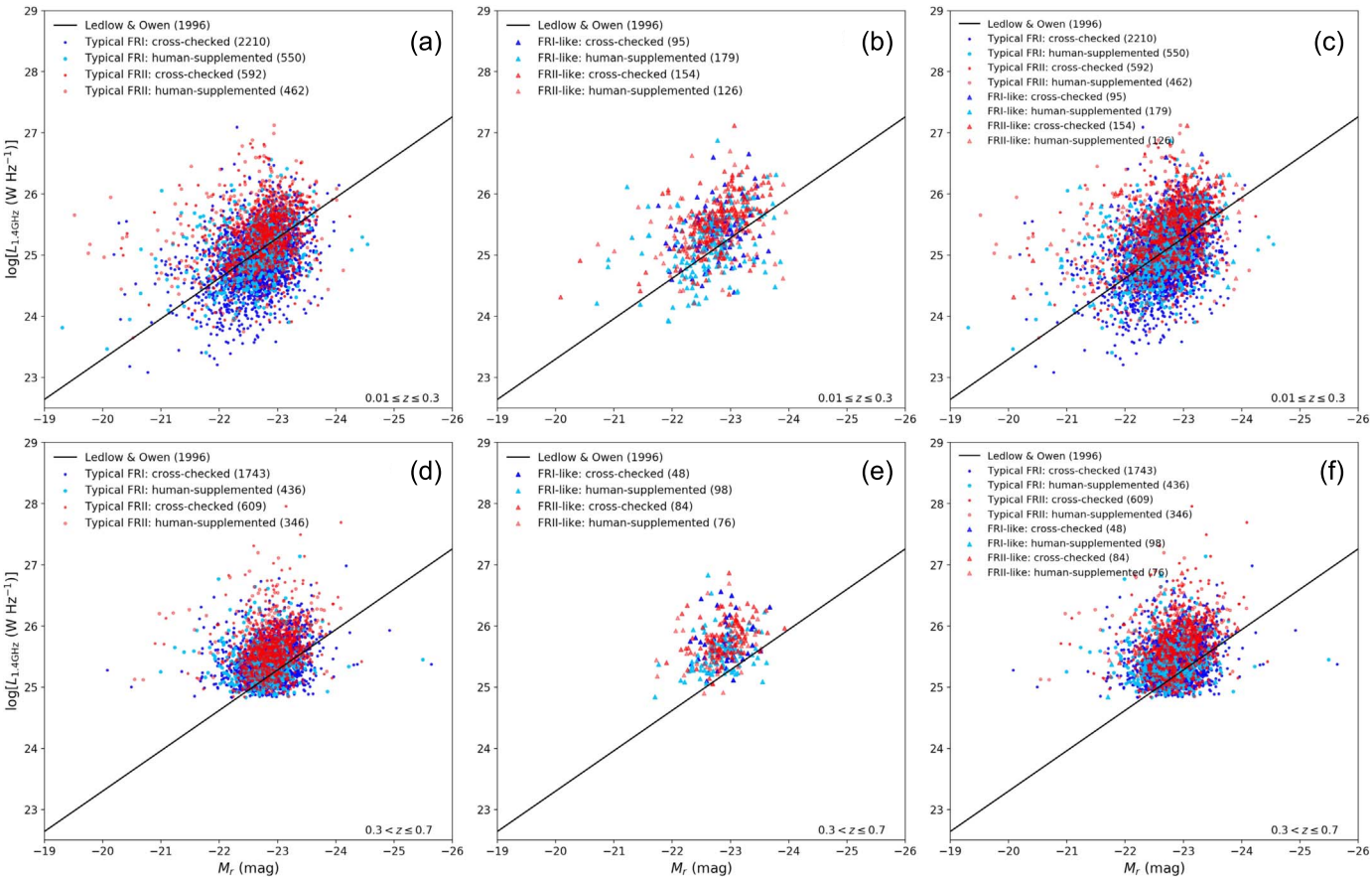


Figure 12. Joint distributions of the unLRG FR samples between 1.4 GHz luminosities (log[$L_{1.4\,\mathrm{GHz}}$]) and r-band absolute magnitudes (M_r), which are obtained in the redshift ranges $0.01 \leqslant z \leqslant 0.3$ and $0.3 < z \leqslant 0.7$, respectively. The symbols are the same as those used in Figure 10, and the corresponding source numbers are marked in the legends.

ranges. Compared with FRIs located in $z \leqslant 0.3$, FRIs in $0.3 < z \leqslant 0.7$ tend to be confined in relatively higher $L_{1.4\,\mathrm{GHz}}$ intervals, which is likely to be caused by the selection effect as fainter sources are hard to observe at higher redshifts and/or by the 5 mJy cut of NVSS to the sample. Therefore, we conclude that no redshift evolution is observed with our sample.

5.2. A Comparison between Machine and Human Classification of the unLRG Sample

As can be seen in Section 4.2.2, the fully pre-trained and fine-tuned MCRGNet has achieved very satisfying classification precision on the LRG sample. Comparing the results obtained with MCRGNet and visual classification on the

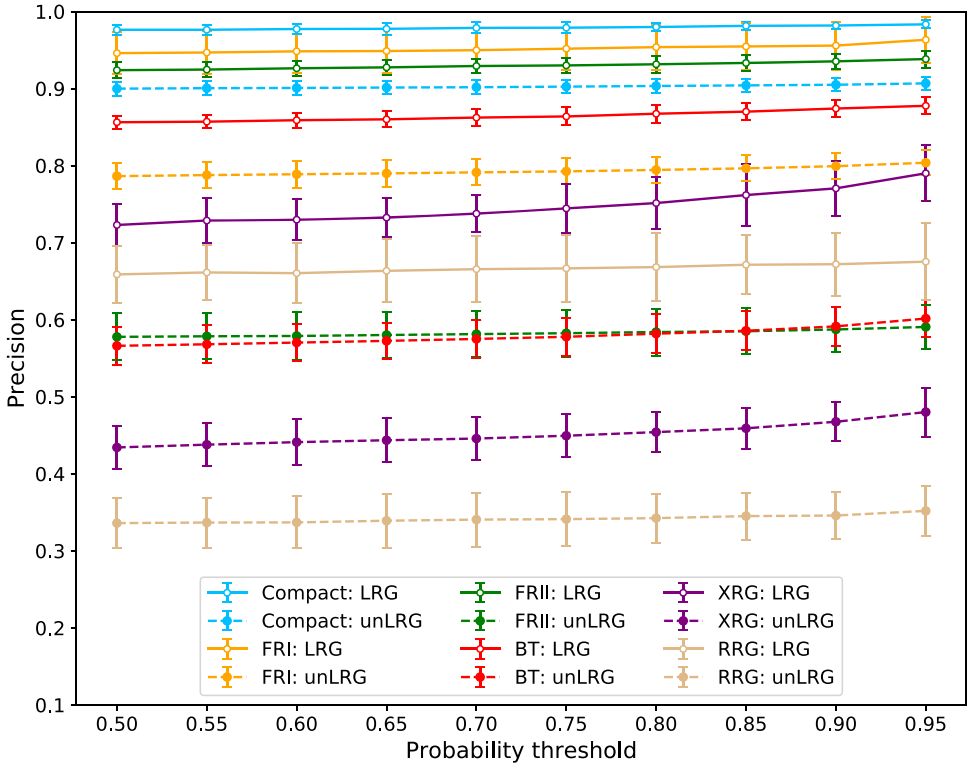


Figure 13. Precisions of the proposed MCRGNet on the LRG and unLRG samples obtained by setting different probability thresholds.

unLRG sample in Section 4.3, we find that the classification precisions for the compact and FRI sources are high ($\gtrsim$80%), but the precisions are low ($\lesssim$70%) for the FRII and BT sources and are actually worse for the XRGs and RRGs. As for the FRII and BT sources, we note that about 14% of the former (308 FRIIs) and about 12% of the latter (56 FRII-like BTs) are misclassified into each other by MCRGNet, which reproduces the result of Aniyan & Thorat (2017). In addition, about 1% of the FRIIs are misclassified into the compact category, and this can be attributed to relativistic beaming, by which a bright nearer lobe dominates its fainter counterpart in the images. As explained in Section 4.2.2 for the LRG sample classification, the low precision for the XRG and RRG sources in the unLRG sample may also be due to their small numbers (78 XRGs and 48 RRGs) in the sample (14,245 sources) and their ambiguous morphologies.

By cross-checking the misclassified sources, we find that most of them require more of us authors to join to vote them into certain AGN types. Also for these sources, most are found with low probabilities ($\lesssim$0.90) obtained by the CNN. Tentatively, we set a threshold to the probabilities of the final labels that are predicted by the dichotomous tree classifier, i.e., objects with probabilities higher than the threshold are kept to calculate the precisions. The evolution of the precision of the classification for the six AGN subtypes is illustrated in Figure 13 as a function of the probability thresholds. We find that for the compact and FRI sources, the classification precisions are not nearly affected by changes in the probability threshold, which is consistent with the corresponding high probabilities ($\gtrsim$0.95) of the labels predicted by our MCRGNet (see Table 4). However, for the XRGs and RRGs, which have relatively complex morphologies, the classification precision significantly increases as the threshold increases.

Actually, the problems described above are caused by the fact that the completeness of the labeled sample used to train the neural network is still inadequate. Although the pre-training on a large unlabeled sample (the unLRG sample) may be able to extract most of the features necessary to describe the radio galaxies, the fine-tuning process tends to enhance the features of existing instances carried in the LRG sample while weakening the uncarried. For example, no typical relativistically beamed FRII source is correctly involved in the LRG sample, as well as in other labeled FRII samples available in literature, which prohibits the recognition of the corresponding sources. This may be called a data-unseen problem, which could introduce uncertainty to the network (see Section 5.4 for a detailed discussion) and has been discussed in many machine learning based works, e.g., on image classification (e.g., Chatzilari et al. 2015) and object detection (e.g., Zhu et al. 2016). The solution to this problem is straightforward, i.e., enlarging the sample. Actually, providing astronomers with a large morphologically classified sample which is as complete as possible is part of the motivation of our work. By repeating such a process to enlarge the labeled sample data scale, the data-unseen problem can be efficiently mitigated. This is called iteratively training thinking (e.g., Ma et al. 2017).

5.3. The Use of One Fully Connected Layer and the 64-dimension Feature Vector

In our neural network (Figure 2(b)), there is only one fully connected layer, which encapsulates the extracted features of the radio galaxies in our sample. However, in typical CNNs, such as LeNet-5 (LeCun et al. 1995), AlexNet (Krizhevsky et al. 2012), and GoogLeNet (Szegedy et al. 2015), more than one fully connected layer is employed. The reason for using only one fully connected layer in this work is that in a series of

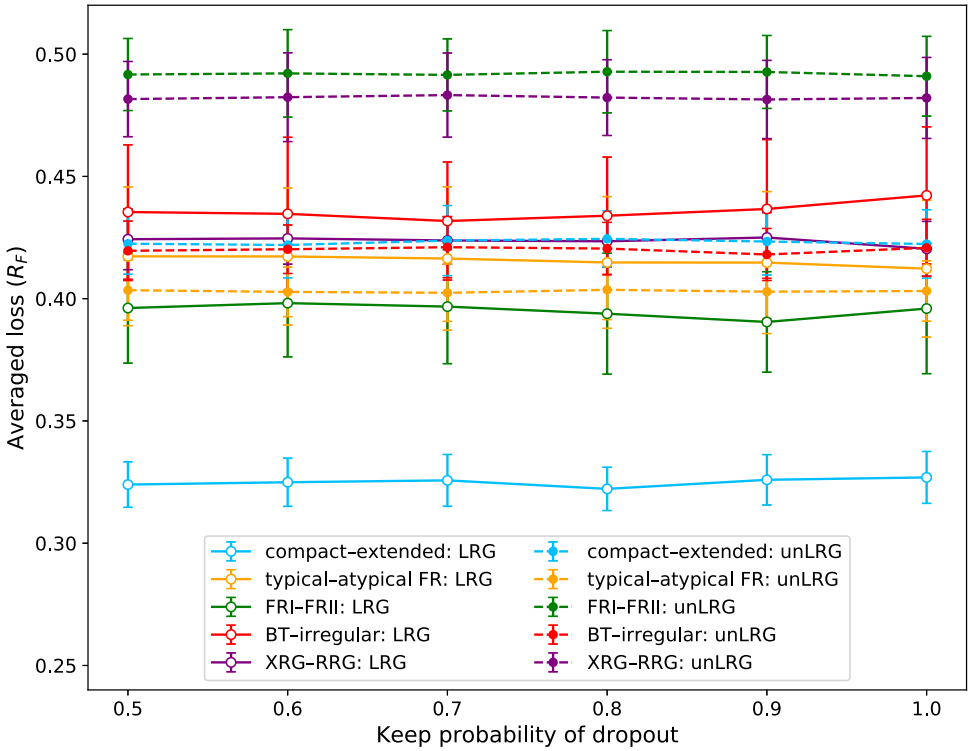


Figure 14. Uncertainties of MCRGNet on the LRG and unLRG samples as a function of the keep probability of the dropouts.

tests we find that more than one is not effective in the morphological classification task in this work. In fact we find that by adding more than one fully connected layer, the classification precision decreases sharply instead of increasing. This phenomenon has been detected in many CNN-based works and is named degradation, i.e., the performance of the network saturates or even declines gradually with more layers than needed (e.g., He et al. 2015, 2016). Thus, based on a series of tests, only one fully connected layer is adopted.

The length of the extracted features, i.e., the number of nodes in the fully connected layer, also affects the precision of the radio galaxy classification task and the complexity of the network, which has been discussed in many related machine learning based problems (e.g., Krizhevsky et al. 2012; Szegedy et al. 2015; Ma et al. 2017). In Figure 9 we illustrate the results of tree structure classification using either 64-dimension or 32-dimension vectors. We find that the degree of separation of the six subtypes (i.e., compact, typical FRI, typical FRII, BT, XRG, and RRG) obtained with 64-dimension feature vectors dominates that obtained with 32-dimension vectors, especially for the types of typical FRIs/FRIIs and BTs. This is because the shorter feature vectors squeeze the learned representations of the radio galaxy morphologies, so that some weak but useful information is lost. On the other hand, we choose not to further increase the dimensions of the features because a larger fully connected layer will exponentially increase the scale of the neural network, which will lead to overfitting risk, affect the efficiency of the network, and thus consume more resources (Han et al. 2015; Li et al. 2016).

5.4. Uncertainties in Machine Learning Based Classification

Two types of uncertainties, namely aleatoric and epistemic uncertainty, are commonly discussed in machine learning based problems (e.g., Kendall & Gal 2017; Perreault Levasseur et al. 2017). The former describes the intrinsic uncertainty related to the quality of the images (e.g., occlusions, lack of visual features, and over/underexposure; see Kendall & Gal 2017 for details), which is expected to be relieved after enough high-quality images are obtained in future astronomical observations. The latter measures the uncertainty caused by factors that the model does not take into account due to the lack of a complete training data set. We will focus on the latter by applying Bayesian inference as shown below.

In the usual way, to measure epistemic uncertainty, the prior distribution of the weight and bias parameters in the model is directly inferred, often referred to as the Bayesian neural network (Gal & Ghahramani 2015). However, such direct calculations will cost a lot of time and computing resource, especially for a large-scale network. Gal & Ghahramani (2015) and Kendall & Gal (2017) have proved that the optimization of the neural network with dropout, which zeros out neurons randomly as a Bernoulli distribution, is approximately equivalent to a form of Bayesian inference. Thus we utilize the dropout layers already appended after each of the convolutional and fully connected layers (Section 4.2.1) by switching them on using the approach of Gal & Ghahramani (2015), which is called Monte Carlo dropout. Note that the dropout layers are switched off at the final stage of the labeling of the unLRG sample (see Table 4). In Figure 14 we illustrate the test loss curves together with the 68% errors as a function of the keep probability of the dropouts for the five branches on the dichotomous tree. As expected in Section 5.2, for a given branch the uncertainty of the classification of the unLRG sample is in general larger than that of the LRG sample. When classifying the unLRG sample, we find again that the losses of the FRI–FRII (green line) and XRG–RRG (purple line) branches are significantly higher than those of the other branches by taking the uncertainty into consideration.

This may be attributed to the unLRG sources with ambiguous morphologies (e.g., the FRIIs with relativistic beaming, which is, in contrast, a clear feature under human inspection). These results agree with what we have obtained in Section 5.2. Besides, from the behavior of the error bars and the small fluctuations on the curves as the dropout keep probability decreases, we are confident that the proposed MCRGNet with dropout layers could be robust.

If we increase the threshold set to the probability, which is used to quantify how sure it is that an image belongs to a specific radio galaxy subtype (see Appendix B for details), the precision of the predictions tends to increase, especially for the XRGs and RRGs, the numbers of which are small in both the LRG and unLRG samples (see Figure 13 for details). To be on the safe side we suggest an empirical probability threshold of 0.90 for selecting sources from our unLRG sample (Table 4).

In addition, we have eventually propagated the uncertainty of our network (i.e., selecting sources with the probability threshold) to Figures 10–12, by which our conclusion remains the same as that in Section 5.1.

6. Summary

A morphological classification of 14,245 radio AGNs from the sample selected by Best & Heckman (2012) is performed by applying deep CNN classifiers, namely MCRGNet (the code is available at https://github.com/myinxd/MCRGNet). By combining and cross-checking the classifications by both MCRGNet and the authors' manual inspections, a catalog is also provided, in which we retain the original items of the 14,245 radio AGNs from Best & Heckman (2012) and append new columns hosting the morphological and optical information obtained with SDSS.

We propose to classify the radio galaxies into six types, i.e., typical FRI/FRII, FRI/II-like BT, XRG, and RRG, where we subdivide the BT radio galaxies into FRI-like and FRII-like instead of NAT and WAT as conventional. To realize the classification, a dichotomous tree classifier is designed, which is composed of cascaded CNN based subclassifiers. A three-step training strategy, namely PFC, is proposed to extract features of the radio galaxies for morphological classifications. Unsupervised pre-training of the networks with a large unlabeled sample and supervised fine-tuning with a relatively small labeled sample is proved robust to classifying radio galaxies with high precisions and sensitivities compared to results in literature (e.g., Aniyan & Thorat 2017), as well as low epistemic uncertainty by training MCRGNet with dropout-based Bayesian inference. In addition, the relatively small-scale MCRGNet can be deployed to portable computers for astronomers, which is a key motivation of this work.

By calculating the r-band absolute magnitude of the sources in the unLRG sample with the data from SDSS DR7, we find that the distributions of the unLRG sources on the $L_{\rm radio}-M_{\rm opt}$ plane, which do not show an apparent redshift evolution, indicate that there could not exist an abrupt separation between FRIs and FRIIs as reported in previous works.

We sincerely thank the referee for providing valuable comments. We also thank Dr. H. Andernach and Dr. Y. Zhu for the revision and correction of our AGN catalog. This work was supported by the National Science Foundation of China (grant Nos. 11433002, 61371147, 11835009, 11621303, 51672176, and 11673017) and the National Key Research and Discovery Plan (grant Nos. 2018YFA0404601 and 2017YFF0210903). C.L. acknowledges support from the National Key Basic Research Program of China (2015CB857002). C.L. is supported by the Key Laboratory for Particle Physics, Astrophysics and Cosmology, Ministry of Education, and the Shanghai Key Laboratory for Particle Physics and Cosmology.

Software: Vim (https://www.vim.org/), IPython (http://ipython.org/), Jupyter (https://jupyter.org/), TensorFlow (https://www.tensorflow.org/), NumPy (https://www.numpy.org/), SciPy (https://www.scipy.org), AstroPy (http://www.astropy.org/; The Astropy Collaboration et al. 2018), Pandas (https://pandas.pydata.org/).

Appendix A
Explanation of Transposed Convolution

Transposed convolution, which works by swapping the forward and backward parts of the encoder block, can be calculated by a matrix operation. As for the forwarding or the convolution of the encoder CNN, if the input I and output O were to be unrolled into vectors I_v and O_v from left to right, top to bottom, then the convolution could be represented as a sparse matrix C where the nonzero elements are the elements of the weight matrix W. For instance, supposing the shapes of I, W, and O are 3×3, 2×2, and 2×2, respectively, and the padding mode is same padding (i.e., half-padding) with a stride of 1, then the sparse matrix C is as follows:

$$C = \begin{pmatrix} w_{11} & w_{12} & 0 & w_{21} & w_{22} & 0 & 0 & 0 & 0 \\ 0 & w_{11} & w_{12} & 0 & w_{21} & w_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{11} & w_{12} & 0 & w_{21} & w_{22} & 0 \\ 0 & 0 & 0 & 0 & w_{11} & w_{12} & 0 & w_{21} & w_{22} \end{pmatrix}. \quad (15)$$

The unrolled vectorized I_v and O_v of matrices I and O are

$$I_v = (I_{11}, I_{12}, I_{13}, I_{21}, I_{22}, I_{23}, I_{31}, I_{32}, I_{33}), \quad (16)$$

$$O_v = (O_{11}, O_{12}, O_{21}, O_{22})^T. \quad (17)$$

Then the convolution operation is transformed into matrix multiplication,

$$O_v = C \cdot I_v. \quad (18)$$

Finally, by reshaping the output O_v from vector to matrix, we obtain the output convolved result O.

Similarly, based on Equations (15), (16), and (18), the deconvolution can be calculated by the arithmetic of transpose convolution as

$$I_v^T = C^T \cdot O_v^T, \quad (19)$$

where the superscript T represents the transpose operation; thus C^T shares the same shape and sparsity as the transposed matrix of C, and so do I_v^T and O_v^T. Finally, by reshaping I_v^T back into I, we can obtain the recovered deconvolved matrix.

Appendix B
How Does the Dichotomous Tree Classifier Work?

We propose a dichotomous tree structure to split the multiple classification for the six AGN morphological types into binary classification tasks (Figure 1). By feeding an AGN source into the tree from the root node, a route along the tree will be obtained, and the image will arrive at one of the leaves as a

certain AGN morphology. At the same time, a probability (or a reasonability as defined by Fei & Liu 2006) would be calculated to quantify how sure it is that the tree will classify the source into a certain AGN type. Since a CNN is trained independently at each branch as described in Section 4.2, the probability of a source being classified as an AGN type (i.e., one of the leaf nodes on the tree) by the network obeys a joint distribution composed of the branches on the tree. We explain the probability calculation algorithm in this appendix, which is based on a structured probability model defined by Goodfellow et al. (2016).

Denote $B = \{B_1, B_2, ..., B_6\}$ as the dichotomous tree, in which B_i, $i = 1, 2, \ldots, 6$ represents the classification branches compact–extended, typical–atypical FR, typical FRI–FRII, BT–irregular, FRI-like–FRII-like, and XRG–RRG, respectively. Also, we define B_i^l and B_i^r as the left and right nodes of branch i, respectively. The joint probability is then calculated by splitting it into many factors, which are the conditional probabilities of the nodes along the decision route estimated by the corresponding CNNs, and cascadingly multiplying them together.

Following Goodfellow et al. (2016), we define $Pa(n)$ as the parents of node n (i.e., the former nodes of n on the decision route $\boldsymbol{R}$), and

$$\begin{aligned} P(x) &= P(\boldsymbol{R}) \\ &= \Pi_{n\in\boldsymbol{R}} P_n(n|Pa(n)), \end{aligned} \tag{20}$$

where x is the AGN source and $\boldsymbol{R}$ is the set of nodes to be selected on the route that the source passes on the tree.

For instance, suppose we had an AGN source x which was classified as an RRG by the dichotomous tree. The route on the tree would be (root) → (extended) → (atypical FR) → (irregular) → (RRG), i.e., $\boldsymbol{R} = \{B_1^r, B_2^r, B_4^r, B_6^r\}$). Then the joint probability of x is calculated by

$$\begin{aligned} P(x) &= P(\boldsymbol{R} = \{B_1^r, B_2^r, B_4^r, B_6^r\}) \\ &= \Pi_{n\in\boldsymbol{R}} P_n(n|Pa(n)) \\ &= P(B_1^r)P(B_2^r|B_1^r)P(B_4^r|B_2^r, B_1^r)P(B_6^r|B_4^r, B_2^r, B_1^r). \end{aligned} \tag{21}$$

We provide the joint probability as the reasonability for each of the sources in both the LRG and unLRG samples obtained by classifying them with the trained tree in Tables 3 and 4. To utilize the unLRG catalog in further researches, we suggest an empirical probability threshold of 0.90 as a safe reference for users to select their AGN samples with (see Section 5.2 for a detailed discussion).

ORCID iDs

Zhixian Ma https://orcid.org/0000-0003-1263-9453
Weitian Li https://orcid.org/0000-0002-7527-380X
Chengze Liu https://orcid.org/0000-0002-4718-3428

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, http://tensorflow.org/
Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543
Aniyan, A. K., & Thorat, K. 2017, ApJS, 230, 20
Baldi, R. D., Capetti, A., & Massaro, F. 2018, A&A, 609, A1
Barchi, P. H., da Costa, F. G., Sautter, R., et al. 2017, J. Comp. Interdisciplinary Sci., 07, 0114
Beardsley, A. P., Hazelton, B. J., Sullivan, I. S., et al. 2016, ApJ, 833, 102
Becker, R. H., White, R. L., & Helfand, D. J. 1995, ApJ, 450, 559
Benvenuto, F., Piana, M., Campi, C., & Massone, A. M. 2018, ApJ, 853, 90
Best, P. N. 2009, AN, 330, 184
Best, P. N., & Heckman, T. M. 2012, MNRAS, 421, 1569
Best, P. N., Kauffmann, G., Heckman, T. M., & Ivezić, Ž. 2005, MNRAS, 362, 9
Bethapudi, S., & Desai, S. 2018, A&C, 23, 15
Bishop, C., & Nasrabadi, N. 2006, Pattern Recognition and Machine Learning (New York: Academic)
Blanton, M. R., & Roweis, S. 2007, AJ, 133, 734
Brinchmann, J., Charlot, S., Heckman, T. M., et al. 2004, arXiv:astro-ph/0406220
Capetti, A., Massaro, F., & Baldi, R. D. 2017a, A&A, 601, A81
Capetti, A., Massaro, F., & Baldi, R. D. 2017b, A&A, 598, A49
Chang, C.-C., & Lin, C.-J. 2011, ACM Trans. Intell. Syst. Tech., 2, 27
Chapman, E., Zaroubi, S., Abdalla, F. B., et al. 2016, MNRAS, 458, 2928
Chatzilari, E., Nikolopoulos, S., Kompatsiaris, Y., & Kittler, J. 2015, IEEE Conf. Image Process., 4256
Cheung, C. C. 2007, AJ, 133, 2097
Christodoulidis, S., Anthimopoulos, M., Ebner, L., Christe, A., & Mougiakakou, S. 2016, IEEE J. Biomed. Health Info., 21, 76
Condon, J. J., Cotton, W. D., Greisen, E. W., et al. 1998, AJ, 115, 1693
DeBoer, D. R., Parsons, A. R., Aguirre, J. E., et al. 2017, PASP, 129, 045001
Deng, J., Dong, W., Socher, R., et al. 2009, IEEE Conf. Comp. Vis. Pattern Recognit., 248
Dennett-Thorpe, J., Scheuer, P. A. G., Laing, R. A., et al. 2002, MNRAS, 330, 609
Diamantis, D., Iakovidis, D., & Koulaouzidis, A. 2018, ITIP, 18287573
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Donoso, E., Best, P. N., & Kauffmann, G. 2009, MNRAS, 392, 617
Dumoulin, V., & Visin, F. 2016, arXiv:1603.07285
Erhan, D., Bengio, Y., Courville, A., et al. 2010, J. Mach. Learn. Res., 11, 625, http://www.jmlr.org/papers/v11/erhan10a.html
Esteva, A., Kuprel, B., Novoa, R. A., et al. 2017, Natur, 542, 115
Fabbro, S., Venn, K., O'Briain, T., et al. 2018, MNRAS, 475, 2978
Fabian, A. C. 2012, ARA&A, 50, 455
Fanaroff, B. L., & Riley, J. M. 1974, MNRAS, 167, 31P
Fei, B., & Liu, J. 2006, ITNN, 17, 696
Gal, Y., & Ghahramani, Z. 2015, arXiv:1506.02158
Gendre, M. A., Best, P. N., & Wall, J. V. 2010, MNRAS, 404, 1719
Gendre, M. A., & Wall, J. V. 2008, MNRAS, 390, 819
Girshick, R., Donahue, J., Darrell, T., & Malik, J. 2014, in IEEE Conf. Comp. Vis. Pattern Recognit., 580
Glorot, X., Bordes, A., & Bengio, Y. 2011, in Proc. 14th Int. Conf. Artificial Intelligence Stat., ed. G. Gordon, D. Dunson, & D. Miroslav (Sheffield: PMLR), 315, http://proceedings.mlr.press/v15/glorot11a.html
Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (Cambridge, MA: MIT Press) http://www.deeplearningbook.org/
Goodfellow, I. J., Pougetabadie, J., Mirza, M., et al. 2014, Adv. Neural Info. Process. Syst., 27, 2672, http://papers.nips.cc/paper/5423-generative-adversarial-nets
Gopal-Krishna, & Wiita, P. J. 2000, A&A, 363, 507
Han, S., Pool, J., Tran, J., & Dally, W. 2015, Adv. Neural Info. Process. Syst., 28, 1135, http://papers.nips.cc/paper/5784-learning-both-weights-and-connections-for-efficient-neural-network
He, K., Zhang, X., Ren, S., & Sun, J. 2015, IEEE Int. Conf. Comp. Vis., 1026
He, K., Zhang, X., Ren, S., & Sun, J. 2016, IEEE Conf. Comp. Vis. Pattern Recognit., 770
Hinton, G., Deng, L., Yu, D., et al. 2012, ISPM, 29, 82
Hinton, G., & Roweis, S. 2002, Adv. Neural Info. Process. Syst., 15, 833, http://papers.nips.cc/paper/2276-stochastic-neighbor-embedding
Kamdar, H. M., Turk, M. J., & Brunner, R. J. 2016, MNRAS, 455, 642
Kendall, A., & Gal, Y. 2017, arXiv:1703.04977
Khalaf, A. F., Yassine, I. A., & Fahmy, A. S. 2016, IEEE Int. Conf. Image Processing, 385
Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
Koopmans, L., Pritchard, J., Mellema, G., et al. 2015, in Proc. Advancing Astrophysics with the Square Kilometre Array (AASKA14), ed. T. L. Bourke et al. (Trieste: PoS), 1
Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, Adv. Neural Info. Process. Syst., 25, 1097, http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks
Kullback, S., & Leibler, R. A. 1951, Ann. Math. Stat., 22, 79

Leahy, J. P., & Parma, P. 1992, in Proc. 7th I.A.P. Meeting, Extragalactic Radio Sources. From Beams to Jets, ed. J. Roland, H. Sol, & G. Pelletier (Cambridge: Cambridge Univ. Press), 307
LeCun, Y., Bengio, Y., & Hinton, G. 2015, Natur, 521, 436
LeCun, Y., Jackel, L., Bottou, L., et al. 1995, NN, 261, 276
Ledlow, M. J., & Owen, F. N. 1996, AJ, 112, 9
Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H. P. 2016, arXiv:1608.08710
Lin, Y.-T., Shen, Y., Strauss, M. A., Richards, G. T., & Lunnan, R. 2010, ApJ, 723, 1119
Lukic, V., Brüggen, M., Banfield, R., et al. 2018, MNRAS, 476, 246
Ma, Z., Li, W., Wang, L., Xu, H., & Zhu, J. 2017, IEEE Conf. Signal Process., 1021
Ma, Z., Zhu, J., Li, W., & Xu, H. 2018, IEEE Conf. Image Process., 3044
Ma, Z., Zhu, J., Li, W., & Xu, H. 2017, IEITI, 100, 2578
Mackay, C. D. 1971, MNRAS, 154, 209
Mahmud, M., & Ray, S. 2008, Adv. Neural Info. Process. Syst., 985, http://papers.nips.cc/paper/3228-transfer-learning-using-kolmogorov-complexity-basic-theory-and-empirical-evaluations
McNamara, B. R., & Nulsen, P. E. J. 2007, ARA&A, 45, 117
Mellema, G., Koopmans, L. V. E., Abdalla, F. A., et al. 2013, ExA, 36, 235
Mountrichas, G., Corral, A., Masoura, V. A., et al. 2017, A&A, 608, A39
Padovani, P. 2016, A&ARv, 24, 13
Padovani, P., Alexander, D. M., Assef, R. J., et al. 2017, A&ARv, 25, 2
Pan, S. J., & Yang, Q. 2010, IEEE Trans. Knowl. Data Eng., 22, 1345
Perreault Levasseur, L., Hezaveh, Y. D., & Wechsler, R. H. 2017, ApJL, 850, L7
Procopio, P., Wayth, R. B., Line, J., et al. 2017, PASA, 34, e033
Proctor, D. D. 2011, ApJS, 194, 31
Radford, A., Metz, L., & Chintala, S. 2015, arXiv:1511.06434
Rifkin, R., & Klautau, A. 2004, J. Mach. Learn. Res., 5, 101, http://www.jmlr.org/papers/v5/rifkin04a.html
Rosenstein, M., Marx, Z., Kaelbling, L., & Dietterich, T. 2005, in NIPS 2005 Workshop on Transfer Learning (Vancouver: NIPS), 1, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.7909&rep=rep1&type=pdf
Rudnick, L., & Owen, F. N. 1976, ApJL, 203, L107
Russakovsky, O., Deng, J., Su, H., et al. 2015, J. Comp. Vis., 115, 211
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Simonyan, K., & Zisserman, A. 2014, arXiv:1409.1556
Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014, J. Mach. Learn. Res., 15, 1929, https://dl.acm.org/citation.cfm?id=2670313
Szegedy, C., Liu, W., Jia, Y., et al. 2015, IEEE Comp. Vis. Pattern Recognit., 1
The Astropy Collaboration, Price-Whelan, A. M., Sipöcz, B. M., et al. 2018, arXiv:1801.02634
Tremonti, C. A., Heckman, T. M., Kauffmann, G., et al. 2004, ApJ, 613, 898
Tsoumakas, G., & Katakis, I. 2007, J. Data Warehous. Min., 3, 1
Ukwatta, T. N., Woźniak, P. R., & Gehrels, N. 2016, MNRAS, 458, 3821
Van Der Maaten, L., & Hinton, G. 2008, J. Mach. Learn. Res., 9, 2579, http://www.jmlr.org/papers/v9/vandermaaten08a.html
Wilman, R. J., Jarvis, M. J., Mauch, T., Rawlings, S., & Hickey, S. 2010, MNRAS, 405, 447
Wilman, R. J., Miller, L., Jarvis, M. J., et al. 2008, MNRAS, 388, 1335
Wing, J. D., & Blanton, E. L. 2011, AJ, 141, 88
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579
Yu, D., & Deng, L. 2014, Automatic Speech Recognition: A Deep Learning Approach (Berlin: Springer)
Zeiler, M. D. 2012, arXiv:1212.5701
Zhu, X., Vondrick, C., Fowlkes, C. C., & Ramanan, D. 2016, J. comp. Vis., 119, 76