

OPTIMIZATION OF DEEP LEARNING MODELS FOR RADIO GALAXY CLASSIFICATION

PHILIPP DENZEL¹, MANUEL WEISS¹, ELENA GAVAGNIN¹, AND FRANK-PETER SCHILLING^{1*}

¹Zurich University of Applied Sciences ZHAW, 8400 Winterthur, Switzerland

Version January 9, 2026

ABSTRACT

Modern radio telescope surveys, capable of detecting billions of galaxies in wide-field surveys, have made manual morphological classification by humans impracticable. This applies in particular when the Square Kilometre Array Observatory (SKAO) becomes operable in 2027, which is expected to close an important gap in our understanding of the Epoch of Reionization (EoR) and other areas of astrophysics. To this end, foreground objects, contaminants of the 21-cm signal, need to be identified and subtracted. Source finding and identification is thus an important albeit challenging task, especially in radio continuum images. In this work, we investigate the ability of AI and deep learning (DL) methods that have been previously trained on other data domains to localize and classify radio galaxies with minimal changes to their architectures. Various well-known pretrained neural network architectures for image classification and object detection are trained and fine-tuned and their performance is evaluated on a public radio galaxy dataset derived from the Radio Galaxy Zoo. A comparison between convolutional neural network (CNN)- and transformer-based algorithms is performed. The best performing architecture is systematically optimized, the influence of data preprocessing is studied, and an uncertainty estimation is performed by means of an ensemble analysis. Radio source classification performance nearly comparable to the current leading customized models in the literature can be quite easily obtained using existing standard pretrained deep learning architectures, without modification and increase in complexity of the model architectures but rather adaptation of the data. Combining various transformations on replicated image channels seems particularly effective in this regard. Using an ensemble of models can also further improve performance to over 90% accuracy, on par with top-performing models in the literature. The results of this study can be transferred to the analysis of other survey data, e.g. from the Murchison Wide-field Array (MWA), and in the future be used to study the EoR by enabling galaxy foreground subtraction in continuum images from the SKAO.

Keywords: radio galaxy zoo, deep learning

1. INTRODUCTION

The classification of galaxies is a fundamental task in astronomy, serving as a cornerstone for various astrophysical studies (Braun et al. 2015). The morphology of galaxies provides valuable insights into the formation and evolution of galaxies over cosmic time (Conselice 2014), the formation of large-scale structure and the cosmic distribution of dark matter (Reddick et al. 2013), or the accelerated expansion of the Universe (Pesce et al. 2020). Upcoming surveys by, e.g., the Euclid mission¹ or the Square Kilometre Array Observatory (SKAO)² will resolve billions of galaxies (Scaramella et al. 2022; Santos et al. 2015). Their increased sensitivity yields better quality and farther-reaching observations, leading to more and higher resolution detections at redshifts up to ~ 5 (Tingay et al. 2013; Trott et al. 2020; Pritchard et al. 2015; Koopmans et al. 2015). In wide-field surveys, telescopes generally sacrifice angular resolution for aperture and distance, which complicates a precise classification

due to the loss in image quality. Radio low-frequency telescopes such as the SKA also mainly probe cold gas emissions, pulsars, quasars, and other radio-loud objects which are less noticeable in optical or near-optical observations (Shimabukuro et al. 2022). Thus, the shape of galaxies in these bandwidths is quite different from their typical Hubble-esque appearance in the optical, further increasing difficulty of the galaxy finding and classification tasks, e.g. (Hancock et al. 2018).

Manual galaxy finding has long been abandoned in favour of algorithmic approaches based on classical computer-vision techniques such as thresholding (Hopkins et al. 2002), edge detection (Mouhcine et al. 2005), watershed segmentation (Berry 2015), wavelet transform (Ellien et al. 2021), or clustering (Johnston et al. 2014). Correspondingly, more and more galaxy discoveries await analysis or validation. Several strategies to approach this challenge have already been explored such as crowd-sourcing labour-intensive tasks in citizen science projects where non professional volunteers perform the initial data selection and preparation (Lintott et al. 2008, 2010). However, while citizen science projects harness human intuition and the ability to quickly identify and

frank-peter.schilling@zhaw.ch

¹ https://www.esa.int/Science_Exploration/Space_Science/Euclid

² <https://www.skao.int/>

process interesting features, they will struggle to scale to the unprecedented volume of data expected from next-generation surveys.

Machine learning and especially deep learning have been recognized as promising solutions to manage these vast amounts of data. Neural networks have the ability to overcome the curse of dimensionality and generalize to more abstract concepts, thereby imitating human-level attention while leveraging machine-level processing speeds. The proliferation of DL models and, in particular, pretrained vision models (PVMs) for representation learning such as DINO (Zhang et al. 2022a), EfficientFormer (Li et al. 2022) or YOLOv8 (Jocher et al. 2023) has facilitated the application of deep learning in natural sciences. Still, PVMs are underutilized in many scientific disciplines due to the often considerable distribution shift between domain-specific data and the training data of these pretrained models. The overabundance, diversity, and inherent complexity pose additional hurdles in the adoption of such models by scientists. In radio astronomy, it has been shown that PVMs often fall short of traditional supervised models trained from scratch solely on domain data (Lastufka et al. 2024). ResNets (He et al. 2015) pretrained on ImageNet (Deng et al. 2009) seem to be an exception here, as a common backbone of more complex architectures. So, domain-specific custom DL architectures, e.g. Wu et al. (2019a); Burke et al. (2019), have been successfully employed as proof-of-concept in radio galaxy detection and classification, demonstrating performance still comparable to the latest top-performing models in terms of precision and average accuracy. On the other hand, PVMs typically offer more robust predictions by leveraging learned representations across many domains. They are particularly advantageous in optimizing model performance with limited labelled data (in a weakly supervised regime), often comprising only a few hundred or thousand images, which is a common scenario in astronomical applications.

Thus, this work focuses on training and evaluation of widely-used DL models established within the DL literature and which have demonstrated outstanding performance on standard computer vision (CV) benchmark datasets, applied in the context of radio galaxy classification and detection. The focus is on identifying best practices for adapting these models to domain-specific datasets, and vice versa. Notably, the choice of data preprocessing techniques can significantly impact the effectiveness of a pretrained model in adapting to a new data domain. Finally, most DL models give little to no feedback on the confidence in their decision and, if at all, only implicitly take into account the inherent spread in the data distribution. We employ a practical approach to these issues in the form of model ensembling.

In the following sections, we discuss related work on radio galaxy classification (section 2) and present the dataset, DL architectures as well as the data processing methods used (section 3). The results for radio galaxy detection and classification are discussed in section 4, followed by a summary and outlook (section 5).

2. RELATED WORK

Machine learning has been integral to the field of radio astronomy, enabling significant advancements in data analysis and interpretation. Methods like Decision

Trees (Fayyad et al. 1993; Ball et al. 2006), Support Vector Machines (Y. & Y. 2004; Sadeghi et al. 2021), or Random Forests (Cheng et al. 2020) used to outperform simple neural networks for radio source classification (Alger et al. 2018; Becker & Grobler 2019). Convolutional neural networks (CNNs) were the first to reveal a clear advantage over these more traditional methods (Alhasan et al. 2018). So, astronomers started adapting well-known architectures, such as LeNet (Mohan et al. 2022) or AlexNet (Tang et al. 2019), modifying them for their specific data domain (Lukic et al. 2018); for a detailed comparison, see Becker et al. (2021).

A common problem in radio astronomy is a heavy skew in the data distribution towards object classes exhibiting Gaussian-like brightness distributions barely above the noise level. This is often caused by a lack of resolution and leads to less robust models. Aniyan & Thorat (2017) and Maslej-Krešnáková et al. (2021) mitigated this problem of class imbalance using data augmentation techniques, including rotation or mixing of synthetically generated data, e.g. using generative adversarial neural networks (Hosenie 2018). More recent investigations go beyond classical supervised learning methods into the field of semi-supervised or unsupervised learning approaches to overcome the shortage of accurately labelled data (Polsterer et al. 2015; Galvin et al. 2020; Tang et al. 2022; Slijepcevic et al. 2022, 2023; Hossain et al. 2023)

The classification and detection of radio galaxies with more complex DL models has been attempted many times, although almost exclusively trained from scratch. Wu et al. (2019a) presented *CLARAN*, a DL model designed to classify radio galaxy morphologies. It is based on the Faster R-CNN method, an object detection network based on region proposals (Ren et al. 2017), trained on (a subset of) the Radio Galaxy Zoo Data Release 1 catalogue (RGZ DR1; see section 3.1 or Banfield et al. (2015); Wong et al. (2024)). The study demonstrates the potential to handle the large volume of data expected from next-generation radio surveys, achieving high accuracy.

Riggi et al. (2022) proposed another R-CNN variant, Mask R-CNN, named *caesar-mrcnn* (Compact And Extended Source Automated Recognition) for the detection of radio continuum sources. Their model was trained on data surveys from the Australian SKA Pathfinder (ASKAP) telescope. It aims to reduce the false detection rate and improves the association of multiple disjoint islands into physical objects, as is the case for strong active galactic nucleus outflows. The model achieved a classification precision and recall above 90% but only a 65% reliability. Riggi et al. (2023); Mostert et al. (2022); Zhang et al. (2022b); Wang et al. (2021) tried to improve on these efforts using the more performant one-stage object detection networks YOLO (Redmon et al. 2016; Jocher et al. 2022) which are favoured for many applications in industrial sectors.

Sortino et al. (2023) provides a comparative study of these and similar object detection models such as Masked R-CNN (He et al. 2017), Detectron2 (Wu et al. 2019b), DETR (Carion et al. 2020), and EfficientDet (Tan et al. 2020), finding promising results of slightly above 90% accuracy for radio astronomical images. Their benchmarks yield YOLO as the overall best performing model,

even when compared against transformer-based models (Vaswani et al. 2017) such as DETR.

Finally, Lastufka et al. (2024) identifies the potential of using pretrained vision foundation models such as DINOv2 (Oquab et al. 2023) or AM-RADIO (Ranzinger et al. 2024) in optical and radio astronomy applications. They demonstrate a good adaption of these foundation models for classification and detection in the optical spectrum, but observe a degradation in performance especially for the classification task of radio galaxies. This study highlights the challenges of distribution shift and the need for careful consideration to model selection, fine-tuning, and data handling when adapting PVMs to the domain of radio astronomy. Moreover, Cecconello et al. (2024) recently provided a benchmark of pretrained models via semi-supervised and self-supervised approaches and demonstrated that domain-specific pretraining can be even more effective in comparison to traditional pre-training.

With exception of the latter studies, pretrained models are still not even considered for application in radio astronomy on the presumption of incompatibility of models trained on natural images with the domain of radio astronomy. To identify best practices for the adoption of pretrained models, we therefore aim to investigate how DL model performance changes for different architectures with various (low-effort) data preprocessing, data augmentation, and fine-tuning techniques.

3. METHODS

3.1. Dataset

For this work, the publicly available Radio Galaxy Zoo (RGZ) (Banfield et al. 2015) dataset was employed. RGZ is one of the largest radio galaxy classification datasets with around 100,000 radio source images taken from the surveys *Faint Images of the Radio Sky at Twenty Centimeters* (FIRST) (Becker et al. 1995) and *Australia Telescope Large Area Survey* (ATLAS) (Franzen et al. 2015).

RGZ is organized as an online citizen science project that enlists volunteers to classify radio sources and their host galaxies. It bases its methodology on the original Galaxy Zoo citizen science project (Lintott et al. 2008, 2010). After 5.5 years of operation, more than 12,000 registered users contributed to the first data release (RGZ DR1; Wong et al. (2024)) by labelling over 2.29 million queries, yielding 99,146 source classifications for FIRST and 583 for ATLAS.

However, to train a (supervised) object detection model, not only a dataset with labelled classes must be available but also annotated bounding boxes, which specify the coordinates of the recognised objects in an image. Such annotations are available only for a small subset of RGZ DR1, produced by Wu et al. (2019a), here referred to as RGZ OD.

The RGZ DR1 dataset is separated into classes in terms of number of peaks and number of components. The number of components is defined as the number of discrete radio components that a source encompasses, identified at the 4σ flux-density threshold. The number of peaks refers to the count of bright peaks in a radio source that are detected. This results in classes that are named after the combination of these two characteristics and written as $[\# \text{components}]_{-}[\# \text{peaks}]$, e.g.

1_1. For instance, a double-lobed radio galaxy with small angular extent and no radio core may be identified as a source with one component-two peaks (1_2) or a two component-two peaks (2_2) if the two lobes appear disconnected in the radio image, which directly translates to the Fanaroff-Riley classification (FRI and FRII; Fanaroff & Riley (1974)).

RGZ OD is the result of two filter rules that were applied to the RGZ DR1 dataset. First, only samples that exceed a user-weighted consensus level (CL) of ≥ 0.6 were selected, which should ensure that most radio sources are morphologically recognizable by humans. Second, samples with more than four components and four peaks were disregarded to reduce class imbalance. After these two rules, the RGZ OD contained in total 11,836 samples in six classes: 1_1, 1_2, 1_3, 2_2, 2_3 and 3_3 (see figure 1).

The RGZ OD image pixels represent aperture fluxes stored in a normalized range in units of $\mu\text{Jy}/\text{beam}$, from negative to positive values directly derived from the original cleaned and gridded images (Lukic et al. 2018). Bounding boxes with respect to location and size were generated from the metadata considering the physical attributes defined in the RGZ DR1 dataset and converting sky coordinates into pixel coordinates. In this work, the images provided in FITS file format (Greisen 2012) were used with the annotation information by Wu et al. (2019a), to retain the accuracy of the intensity values before preprocessing.

3.2. Deep Learning Architectures

The selected models for this study can roughly be grouped into two categories for each task (classification and object detection) respectively, convolutional and transformer-based networks. Pretrained weights are also available for the selected transformer-based networks, usually trained on ImageNet. From each category, we tested at least one well-established architecture to yield results independent from specific model features.

The following DL image classification architectures were considered: ResNet-50 (He et al. 2015) (here simply ResNet) is the 50-layer variant of the residual CNN architecture which won the ImageNet (Deng et al. 2009) Large Scale Visual Recognition Challenge (ILSVRC) in 2015. Its layers optimize gradient flow through residual functions with respect to the layer inputs. While ResNet is not considered state-of-the-art (SOTA) anymore, it is used as backbone in many other architectures. Thus, any performance enhancement on ResNet is likely to translate to other networks containing it as backbone. EfficientNet (Tan & Le 2019) (EffNet) is a family of CNNs for computer vision published in 2019. Its key innovation is compound scaling, which uniformly scales all dimensions of depth, width, and resolution using a single parameter. Here, the variant EfficientNetV2-S (Tan & Le 2021) is used. It yields superior performance to ResNet in many applications, and can be considered a SOTA among CNN architectures. Vision Transformer (Dosovitskiy et al. 2020) (ViT) and EfficientFormer (Li et al. 2022) (EffFor) are recent transformer models for computer vision. Analogous to their CNN-variants, they represent baseline and optimized variants of established transformer models.

For the object detection task, two architectures were used: YOLOv8-S (Jocher et al. 2023) (Yolov8) is a recent

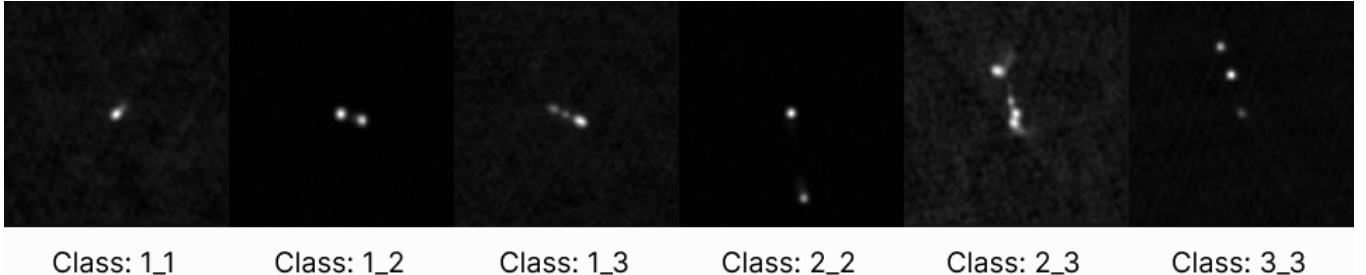


Figure 1. Example images from the RGZ-OD dataset (Wu et al. 2019a). The dataset is separated into classes in terms of number of peaks and number of components, where the former refers to the count of bright peaks that are detected, while the latter is defined as the number of discrete radio components. The resulting classes are named after the combination of the two as [<#components]-[#peaks].

variant of the YOLO (Redmon et al. 2016) family of single shot object detection models. It frames object detection as a regression task by spatially separating bounding boxes and associating probabilities to each detected image using a single neural network. DINO (Zhang et al. 2022a) (here: DINO; with a Swin, shifted window transformer, backbone (Liu et al. 2021)) is a transformer-based object detection model based on DETR (Carion et al. 2020) with various improvements.

3.3. Preprocessing

Radio source images exhibit a significant amount of noise and show a narrow distribution of distinct intensity values. To ensure stable and fast training and increase performance, appropriate preprocessing of radio source images is essential (Sharma 2022).

Scaling allows efficient and robust training and can prevent exploding or vanishing gradients. It can be used to constrain values within defined boundaries, e.g. between zero and one, or to distribute them more Gaussian-like. Here, two well-known scaling techniques are used: z-scale and min-max scale (normalization). The z-scale algorithm was originally developed by the National Optical Astronomy Observatory and is implemented in the IRAF Framework (Tody 1986). It is designed to display the intensity values near the median image, which is especially valuable for astronomical images, as they typically exhibit a peaked histogram in comparison to the background sky.

Sigma clipping is a common method used in astronomy to reduce noise during preprocessing (Lukic et al. 2018), where values that deviate more than a certain number of standard deviations from the median or mean are discarded.

Stretching can be advantageous because, while the dynamic range in astronomical images may be large, the majority of values are near the noise level (White et al. 2007; Dumitrescu et al. 2022). This also normalizes values across samples from different distributions and datasets, e.g., from different telescopes or standards on how the data is stored. If this distinction is not a discriminatory factor, it can hinder the performance of a model. The following stretching methods are used: square root, power, power dist, linear, log, and sinh stretch.

Dynamic Weight Conversion takes the mean, standard deviation and skew of an image to dynamically extend a greyscale channel to three artificial colour channels (Alrubai & Hameed 2018), in contrast to simply replicating it.

The preprocessing techniques were evaluated by means of a grid search in order to determine their influence on the performance of a basic YOLOv3 (Redmon & Farhadi 2018) model without augmentation. All techniques were implemented using the Astropy library (Astropy Collaboration 2022). We observed a substantial increase in model performance from the z-scaling technique. Further, the experiments show that sigma clipping as well as dynamic weights conversion can have a positive effect. On the other hand, the various stretching methods did not have a measurable impact on the results. Thus, only z-scaling, sigma clipping, and dynamic weights conversion are reported for the experiments in section 4.

3.4. Augmentation

Data augmentation is a widely used strategy when data is scarce or imbalanced to mitigate overfitting or to increase the robustness of models. It includes applying label-preserving transformations to an input dataset in order to add more invariant examples (Taylor & Nitschke 2018). It alleviates the acute imbalance of the RGZ D1 dataset in the single-component, single-peak class compared to the other classes. Images are generally cropped to lie approximately in the centre of the image which creates an unfavourable weighting during training for offset source images as it is the case in the other classes. The following augmentation techniques were evaluated for radio source detection and classification: Rotation (flipping in 90° increments), translation, scaling, and shearing.

4. RESULTS

The models introduced in section 3.2 are trained and evaluated with a few basic variations of the training setup, including min-max (MM) scaling and z-scaling (ZS), the decision whether or not to use pretrained weights (PT and NPT, respectively) on ImageNet (Deng et al. 2009) for classification models and on COCO (Lin et al. 2014) for object detection models. Finally, a distinction is made between training on the imbalanced dataset or on a reduced but balanced dataset using undersampling (US).

Fine-tuning of the pretrained models (PT) was done by replacing the classification/projection heads, training the latter with frozen backbones first, and subsequently continue fine-tuning with unfrozen backbones after 5 epochs. All models were trained for 300 epochs (or until convergence) and a batch size of up to 32. The performance of the various DL model architectures is then evaluated for the radio source RGZ OD dataset (Wu et al.

Model	Configuration	Acc.	Top-1	Acc.	Top-2
ResNet	ZS, NPT	83.38	94.53		
ResNet	ZS, NPT, US	81.42	93.90		
ResNet	ZS, PT, US	81.42	94.47		
EffNet	ZS, NPT	79.23	92.30		
EffNet	ZS, PT	78.42	92.71		
EffNet	ZS, PT, US	76.74	90.21		
EffFor	ZS, NPT	76.60	89.46		
EffFor	ZS, PT	75.48	89.56		
EffFor	ZS, NPT, US	73.05	90.64		
ViT	ZS, PT	59.47	77.91		
ViT	ZS, NPT	58.46	77.41		
ViT	MM, PT	52.99	70.72		

Table 1

Performance of classification models: Top-1 accuracy (Acc. Top-1) takes only the prediction with the highest score (output of the last hidden layer) into account, whereas Acc. Top-2 uses the two most likely predictions. ZS refers to the z-scaling preprocessing technique, MM to min-max scaling, US to undersampling to balance classes in the dataset, and (N)PT whether the model uses weights from pretraining (or not). All values are reported in per-centages.

2019a). Note that a limited-size dataset affects model performance of a CNN- in comparison to a transformer-based architecture differently but reflects the real-world use case in astronomy where labels are a rarity.

4.1. Classification Task

Table 1 presents the three out of eight best working configurations of each examined classification model. To provide comparability with previous works (see section 2), both top-1 and top-2 accuracies are provided. Moreover, since it is common to observe a confusion by most models (as well as human experts) between two pairs of classes (the class pairs 1_2, 1_3 and 2_2, 2_3, roughly corresponding to FR-1 and FR-2 galaxies, respectively), the top-2 accuracies show the general performance in absence of these confusions. To account for imbalanced classes in the dataset, the performance metrics are macro-weighted.

Surprisingly, ResNet-50 surpasses the performance of EfficientNet notably, with a maximum top-1 accuracy performance of 83.38%. On the other hand, unsurprisingly, the results confirmed our intuition that on the limited RGZ OD dataset transformer-based architectures cannot leverage their large number of parameters as an advantage. Z-scaling showed to be more suitable for radio source images in comparison to min-max scaling for all applied models. The experiments further indicate that from-scratch training slightly helps to optimise model performance, as opposed to fine-tuning pretrained models on ImageNet (Deng et al. 2009), which aligns with previous studies (see section 1). The model performance did generally not benefit from applying an undersampling filter to the imbalanced dataset. Through inspection of misclassified samples, it can be observed that most confusion exists between class 1_2 and 1_3 as well as 2_2 and 2_3, which is also a distinction with which even human experts struggle most.

4.2. Object Detection Task

Similar to the classification model benchmark experiments, the object detection models were trained on the RGZ OD dataset, with one difference in the preprocess-

Model	Configuration	mAP	IOU	Acc.	<i>mAP</i> ₅₀	<i>mAP</i> ₇₅
DINO	MM, PT	70.92	66.19	80.28	83.90	79.90
DINO	ZS, PT	68.88	65.73	79.11	80.80	77.40
DINO	MM, PT, US	67.80	64.83	76.60	80.90	77.00
DINO	ZS, PT, US	66.02	64.76	74.47	78.30	74.90
DINO	MM, NPT	64.71	67.32	78.70	79.20	74.70
DINO	ZS, NPT	64.20	68.42	78.86	78.10	73.90
DINO	ZS, NPT, US	63.05	70.88	73.01	76.80	72.70
DINO	MM, NPT, US	61.18	67.52	72.21	75.10	70.10
YOLOv8	ZS, PT, US	55.75	88.08	71.68	69.80	66.00
YOLOv8	ZS, PT	52.72	89.08	70.09	66.10	62.00
YOLOv8	MM, PT, US	49.96	85.66	67.02	63.90	58.50
YOLOv8	MM, PT	47.41	85.97	69.09	62.40	55.50
YOLOv8	ZS, NPT	43.27	79.60	68.50	57.30	50.00

Table 2

Performance of object detection models evaluated on the most common CV metrics (see subsection 4.2). All values are reported in per-centages.

ing: Samples with multiple classes, which were excluded when training the classification models, are now considered. The performance results of the object detection benchmark experiments are listed in table 2.

The results show a different characteristic to the classification benchmark experiments. Firstly, in terms of mean average precision (mAP) (Padilla et al. 2020), the YOLOv8 models could not outperform transformer-based models such as DINO. However, YOLOv8 achieved the best performance in terms of Intersection over Union (IoU). This could be interpreted as DINO surpassing its CNN-based counterparts for the combination of object detection and source classification, while YOLOv8 performs best considering only the detection task. On the other hand, as the bounding boxes in the RGZ OD dataset were generated automatically and always have a quadratic shape, the IoU has a systematic shift. Lower IoU values could therefore also result from a deviation between inaccurate ground truth (GT) labelling and predicted boxes that are well-matched to the actual object.

In contrast to the classification experiments, using pre-trained weights is beneficial. This could be explained by the comparatively higher complexity of object detection models and the ability to better generalize. In fact, the use of pretrained weights was the most influential factor leading to a higher mAP. In addition, training the DINO models on images with min-max scaling resulted in higher performance compared to z-scaling. Lastly, contrary to YOLOv8, DINO could not benefit from balancing the dataset with undersampling.

Inspecting samples misclassified by DINO models indicates a similar result as for the classification models with a strikingly higher confusion between class 1_2 and 1_3 as well as between 2_2 and 2_3. Moreover, a frequent mistake involved misclassifying an actual source emission as background turning multi-component sources into 1_1. This effect is most likely explained by the combination of an over-representation of the 1_1 class and general, inherent difficulty in dealing with noise in radio astronomy observations.

In summary, object detection models are generally more proficient at detecting radio sources, whereas the optimization of their classification performance is more challenging, leading to mixed results.

As the localization of radio sources can be compara-

tively accomplished with non-machine learning algorithmic finders, more emphasis was placed on further improving the classification task performance in the subsequent experiments.

4.3. Model Uncertainty and Ensembling

The ensemble method is applied to examine the (epistemic) uncertainty of the classification model predictions depending on the data quantity and quality at hand (Rahaman & Thiery 2020; Stahl et al. 2020). The softmax scores are used as the collaborative ensemble result from differently initialized and trained models, calculating their variance as statistical measure per class.

The best performing classification model (ResNet) and settings from section 4.1 are used, training an ensemble of 30 randomly initialized models on z-scaled radio sources, without pretrained weights. The resulting uncertainty per class is in the range 0.041–0.062, with the exception of class 3.3, for which it is 0.022. The average uncertainty is 0.0466.

Furthermore, the number of standard deviations of the difference between the mean value of the predictions and the ground truth was calculated for the test set. 83.1% of the samples are within one standard deviation. However, 4.6% (45 samples) are above the six-sigma interval. An excerpt of those high-variance samples is displayed in figure 2. Two phenomena can be observed: The first cluster of samples are actual misclassifications of the model. However, if the z-scaled representation on the left is compared with the unscaled representation on the right, it can be understood that some confusion of the model is attributable to different scaling methods. So, z-scaling seems to be advantageous for most of the samples but not for all. The second cluster of misclassified images could have been mislabelled examples or at least difficult cases to which even humans cannot precisely assign a unique class. These two clusters of errors can especially be observed for samples on which all 30 models agreed on the same different class from the ground truth.

Over all samples in the test set, on average, 27.66 models out of 30 agree on the predicted class with a standard deviation of 4. Further analysis shows that with an ensemble of three individually trained models, a possible top-1 accuracy of 90.68% and a top-2 accuracy of 98.58% can be achieved.

As discussed above, a major contributing cause to the model uncertainty is the inherent difficulty in distinguishing two pairs of classes (FR-I and FR-II galaxies, 1_2,1_3 and 2_2,2_3) which even domain experts have issues distinguishing at times. Moreover, image quality limitations induce an additional bias towards the trivial class (1_1).

4.4. Parameter Tuning

The most promising model of the experiments in section 4.1 (ResNet-50) is further fine-tuned with respect to specific choices for scaling, augmentation, optimizer, and preprocessing.

Scaling: The previous experiments (section 4.3) revealed that scaling has a significant impact on the performance of the classifier. Examples were identified where the model classified noise as an additional component or peak due to the increase in brightness from z-scaling, or

Model	Configuration	Acc. Top-1	Acc. Top-2
ResNet	Sigma Median Clipping	89.67	97.47
ResNet	Optimiser	89.36	97.57
ResNet	Augmentation	88.35	96.86
ResNet	ZMZStack	86.12	94.43
ResNet	Benchmark	83.38	94.53

Table 3
Performance of ResNet-50 after Parameter Tuning (reported analogous to figure 1).

side lobes of weak intensity could not be distinguished from noise by the classifier without scaling. To remedy this shortcoming, two simple approaches were tested. Firstly, tuning the contrast parameter of the z-scale implementation and secondly, the use of a z-scaled and a min-max scaled representation as different image channels. The resulting preprocessed sample then contained three channels, two z-scaled and one min-max scaled in the centre, subsequently referred to as ZMZStack. A comparison with z-scaled and min-max scaled images is visualised in figure 3. The highest performance increase was achieved when using ZMZStack, which increased the top-1 accuracy from 83.38% to 86.12% without further tuning.

Augmentation: To account for possible distortions of the real distribution, rotation, scaling, translation and shearing were applied by a random factor p between 0 and 0.5. Each method was analysed individually and in combination. While scaling and translation were able to further increase performance, rotation had a neutral effect and shearing even a negative one. Ultimately, the highest improvement was achieved by a combination of scaling and translation with $p = 0.25$, which further increased the top-1 accuracy to 88.35%.

Optimizer: Tuning the optimiser by strategy (SGD (Ruder 2016), AdamW (Loshchilov & Hutter 2017), RMSProp (Tieleman et al. 2012)), learning rate and learning rate scheduler could supplementary raise the performance level to 89.36%.

Preprocessing: Sigma median clipping could slightly improve the performance to 89.67% top-1 accuracy. However, in terms of top-2 accuracy, the performance declined to 97.47%. Sigma median clipping was subsequently only applied to the z-scaled channels of ZMZStack.

Finally, employing deeper models, e.g. ResNet-101 or ResNet-152, did not yield a significant increase in classification performance. A summary of the incrementation steps is shown in table 3.

5. SUMMARY AND OUTLOOK

Well-known DL architectures, which have demonstrated excellent performance in general computer vision tasks, were applied to the task of detecting and classifying radio galaxies from the RGZ OD dataset, derived from the Radio Galaxy Zoo (RGZ D1) dataset. The focus of this study was to investigate the data domain adaptation of pretrained DL models, which up until recently has been believed to be inherently incompatible with the domain of radio astronomy data. We found preprocessing the image data by means of scaling (z-scaling or min-max scaling) to be the most influential factors for model performance in classification. Performance could be fur-

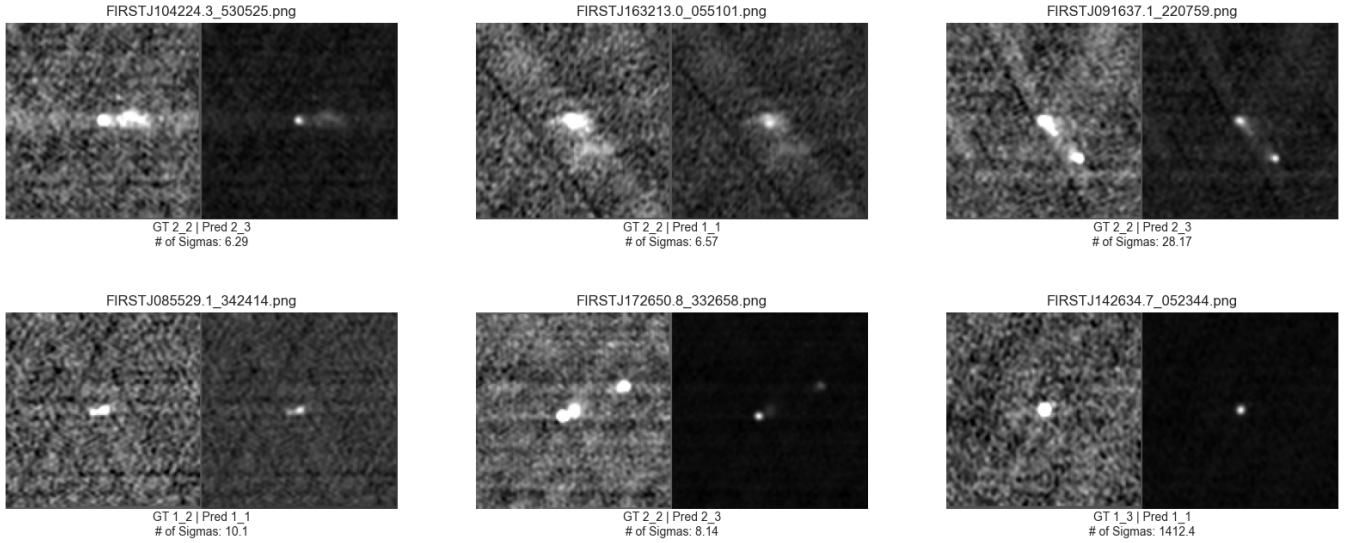


Figure 2. High variance examples (for each panel, left: Z-scaled, right: unscaled).

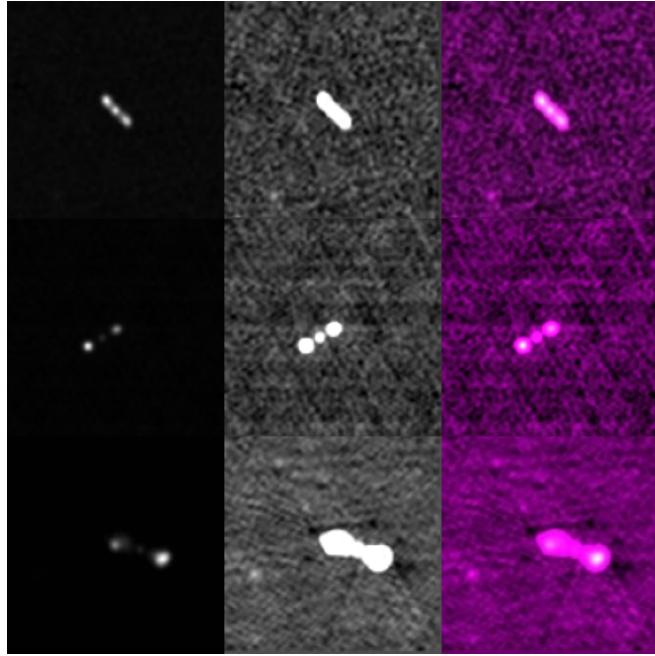


Figure 3. Visual comparison of Min-Max scale (left), Z-scale (middle) and ZMZStack (right) for three example images.

ther improved by stacking image channels with two different scalings. This suggests that further improvements in data adaption of pretrained models may be possible with suitable combinations of channel scalings.

While transformer-based PVMs such as DINO outperform other model architectures in terms of object detection, we obtained the best classification performance with lower-complexity, non-pretrained model architectures such as the ResNet-50. This confirms the results by Lastufka et al. (2024) who observed similar behaviour. They have also demonstrated that this is not only due to the limited size of the dataset, as this behaviour is also observed with other, more diverse (but smaller) galaxy datasets in the radio spectrum. There, another limiting factor is the data quality, e.g. rather high noise lev-

els due to instrumental limits imposed by current radio telescopes.

The performance of a ResNet-50 model could also be further increased in terms of average accuracy from 83.38% to 89.67% by means of scaling, data augmentation, tuning, and hyperparameter optimization. Non-pretrained ResNet models exhibited marginally better performance, suggesting that learned representations on ImageNet have only favourable impact for models with a larger number of parameters. In future studies, the domain adaption of models pretrained on smaller datasets should be investigated. Using an ensemble of 30 models, we explored the model prediction's statistical uncertainty which was relatively low ($< 1\sigma$) for the majority of the test set examples (83%). The analysis of failure cases revealed actual misclassification, as well as potentially mislabelled training data.

We conclude that radio source classification performance nearly comparable to the current leading models in the literature (with an accuracy of above 90%, Becker et al. (2021)) can be quite easily obtained using existing DL architectures, without modification and increase in complexity of the model architectures but rather adaptation of the data. Indeed, while most models applied to the radio galaxy classification task in the literature are trained from scratch (no pre-trained weights) with elaborate, time-consuming modifications to the model architectures, our study suggests that models established in the CV community with large number of parameters including pre-trained weights can be used if data adaption and fine-tuning is performed carefully. Using an ensemble of models can also further improve performance to over 90% top-1 accuracy on par with top-performing models in the literature with higher reliability. With these data adaption methods, in combination with additional fine-tuning techniques such as cropping closer to the radio source, image rescaling to the original image size of the pre-training dataset, using larger projection heads, or more efficient pretraining strategies, we expect that pretrained models could yield superior performance compared to models trained from scratch with considerably less effort. A comparison to domain-specific pre-

trained models such as those by Ceccarello et al. (2024) could lead to more insights in this regard and eventually lead to a complementary approach.

In the future, the results of this study can be transferred to surveys such as GLEAM-X (GaLactic and Extragalactic All-sky MWA Extended Survey (Wayth et al. 2015; Hurley-Walker et al. 2017), based on observations with the Murchison Wide-field Array, or to surveys from the upcoming SKAO.

REFERENCES

- Alger, M. J., Banfield, J. K., Ong, C. S., et al. 2018, MNRAS, 478, 5547, doi: 10.1093/mnras/sty1308
- Alhassan, W., Taylor, A. R., & Vaccari, M. 2018, MNRAS, 480, 2085, doi: 10.1093/mnras/sty2038
- Alrubai, S. A. H., & Hameed, A. H. 2018, JUBPAS, 26, 122, doi: 10.29196/jubpas.v26i8.1677
- Aniyan, A. K., & Thorat, K. 2017, ApJ Supplement Series, 230, 20, doi: 10.3847/1538-4365/aa7333
- Astropy Collaboration. 2022, The Astrophysical Journal, 935, 167, doi: 10.3847/1538-4357/ac7c74
- Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. 2006, ApJ, 650, 497, doi: 10.1086/507440
- Banfield, J. K., Wong, O. I., Willett, K. W., et al. 2015, MNRAS, 453, 2326, doi: 10.1093/mnras/stv1688
- Becker, B., & Grobler, T. 2019, in 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), 1–8, doi: 10.1109/IMITEC45504.2019.9015881
- Becker, B., Vaccari, M., Prescott, M., & Grobler, T. 2021, MNRAS, 503, 1828, doi: 10.1093/mnras/stab325
- Becker, R. H., White, R. L., & Helfand, D. J. 1995, Astrophysical Journal, 450, 559, doi: 10.1086/176166
- Berry, D. 2015, Astronomy and Computing, 10, 22, doi: 10.1016/j.ascom.2014.11.004
- Braun, R., Bourke, T. L., Green, J. A., Keane, E., & Wagg, J. 2015, in PoS(AASKA14), doi: 10.22323/1.215.0174
- Burke, C. J., Aleo, P. D., Chen, Y.-C., et al. 2019, MNRAS, 490, 3952, doi: 10.1093/mnras/stz2845
- Carion, N., Massa, F., Synnaeve, G., et al. 2020, CoRR, doi: 10.48550/arXiv.2005.12872
- Ceccarello, T., Riggi, S., Becciani, U., et al. 2024, CoRR, doi: 10.48550/arXiv.2411.14078
- Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. 2020, MNRAS, 493, 4209, doi: 10.1093/mnras/staa501
- Conselice, C. J. 2014, Annual Review of A&A, 52, 291, doi: 10.1146/annurev-astro-081913-040037
- Deng, J., Dong, W., Socher, R., et al. 2009, in 2009 IEEE CVPR, doi: 10.1109/cvpr.2009.5206848
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. 2020, CoRR, doi: 10.48550/arXiv.2010.11929
- Dumitrescu, F., Ceachi, B., Truică, C.-O., Trăscău, M., & Florea, A. M. 2022, Aerospace, 9, doi: 10.3390/aerospace9090520
- Ellien, A., Slezak, E., Martinet, N., et al. 2021, A&A, 649, A38, doi: 10.1051/0004-6361/202038419
- Fanaroff, B. L., & Riley, J. M. 1974, MNRAS, 167, 31P, doi: 10.1093/mnras/167.1.31P
- Fayyad, U. M., Weir, N., & Djorgovski, S. G. 1993, in Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27–29, 1993, ed. P. E. Utgoff (Morgan Kaufmann), 112–119, doi: 10.1016/B978-1-55860-307-3.50021-6
- Franzen, T. M. O., Banfield, J. K., Hales, C. A., et al. 2015, MNRAS, 453, 4020, doi: 10.1093/mnras/stv1866
- Galvin, T. J., Huynh, M. T., Norris, R. P., et al. 2020, MNRAS, 497, 2730, doi: 10.1093/mnras/staa1890
- Greisen, E. W. 2012, AIPS FITS file format, <https://fits.gsfc.nasa.gov/>
- Hancock, P. J., Trott, C. M., & Hurley-Walker, N. 2018, PASA, 35, e011, doi: 10.1017/pasa.2018.3
- He, K., Gkioxari, G., Dollar, P., & Girshick, R. 2017, in 2017 IEEE ICCV, 2961–2969, doi: 10.1109/iccv.2017.322
- He, K., Zhang, X., Ren, S., & Sun, J. 2015, CoRR, doi: 10.48550/arXiv.1512.03385
- Hopkins, A. M., Miller, C. J., Connolly, A. J., et al. 2002, The Astronomical Journal, 123, 1086, doi: 10.1086/338316
- Hosenie, Z. B. 2018, Source classification in deep radio surveys using machine learning techniques, <https://api.semanticscholar.org/CorpusID:215896011>
- Hossain, M. S., Roy, S., Asad, K. M. B., et al. 2023, Procedia Computer Science, 222, 601, doi: 10.1016/j.procs.2023.08.198
- Hurley-Walker, N., Callingham, J. R., Hancock, P. J., et al. 2017, MNRAS, 464, 1146, doi: 10.1093/mnras/stw2337
- Jocher, G., Chaurasia, A., & Qiu, J. 2023, Ultralytics YOLOv8, 8.0.0, <https://github.com/ultralytics/ultralytics>
- Jocher, G., Stoken, A., et al. 2022, ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, <https://zenodo.org/record/7347926>, Zenodo, doi: 10.5281/ZENODO.7347926
- Johnston, K. G., Beuther, H., Linz, H., et al. 2014, A&A, 568, A56, doi: 10.1051/0004-6361/201423943
- Koopmans, L., Pritchard, J., Mellema, G., et al. 2015, in PoS(AASKA14), doi: 10.22323/1.215.0001
- Lastufka, E., Bait, O., Drozdova, M., et al. 2024, CoRR, doi: 10.48550/arXiv.2409.11175
- Li, Y., Yuan, G., Wen, Y., et al. 2022, CoRR, doi: 10.48550/arXiv.2206.01191
- Lin, T.-Y., Maire, M., Belongie, S., et al. 2014, in Computer Vision – ECCV 2014, ed. D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Springer International Publishing), 740–755, doi: 10.1007/978-3-319-10602-1_48
- Lintott, C., Schwamb, K., Bamford, S., et al. 2010, MNRAS, 410, 166, doi: 10.1111/j.1365-2966.2010.17432.x
- Lintott, C. J., Schwamb, K., Slosar, A., et al. 2008, MNRAS, 389, 1179, doi: 10.1111/j.1365-2966.2008.13689.x
- Liu, Z., Lin, Y., Cao, Y., et al. 2021, in 2021 IEEE/CVF ICCV, 9992–10002, doi: 10.1109/iccv48922.2021.00986
- Loshchilov, I., & Hutter, F. 2017, CoRR, doi: 10.48550/arXiv.1711.05101
- Lukic, V., Brüggen, M., Banfield, J. K., et al. 2018, MNRAS, 476, 246, doi: 10.1093/mnras/sty163
- Maslej-Krešníková, V., El Boucheffy, K., & Butka, P. 2021, MNRAS, 505, 1464, doi: 10.1093/mnras/stab1400
- Mohan, D., Scaife, A. M. M., Porter, F., Walmsley, M., & Bowles, M. 2022, MNRAS, 511, 3722, doi: 10.1093/mnras/stac223
- Mostert, R. I. J., Duncan, K. J., Alegre, L., et al. 2022, A&A, 668, A28, doi: 10.1051/0004-6361/202243478
- Mouhcine, M., Ferguson, H. C., Rich, R. M., Brown, T. M., & Smith, T. E. 2005, ApJ, 633, 810, doi: 10.1086/468177
- Oquab, M., Daretz, T., Moutakanni, T., et al. 2023, CoRR, doi: 10.48550/arXiv.2304.07193
- Padilla, R., Netto, S. L., & da Silva, E. A. B. 2020, in 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), 237–242, doi: 10.1109/IWSSIP48289.2020.9145130
- Pesce, D. W., Braatz, J. A., Reid, M. J., et al. 2020, ApJ Letters, 891, L1, doi: 10.3847/2041-8213/ab75f0
- Polsterer, K. L., Gieseke, F., & Igel, C. 2015, Astronomical data analysis software and systems XXIV (ADASS XXIV), 495, 81
- Pritchard, J., Ichiki, K., Mesinger, A., et al. 2015, in PoS(AASKA14), 012, doi: 10.22323/1.215.0012
- Rahaman, R., & Thiery, A. H. 2020, CoRR, doi: 10.48550/arXiv.2007.08792
- Ranzinger, M., Heinrich, G., Kautz, J., & Molchanov, P. 2024, in 2024 IEEE/CVF CVPR, 12490–12500, doi: 10.1109/cvpr52733.2024.01187
- Reddick, R. M., Wechsler, R. H., Tinker, J. L., & Behroozi, P. S. 2013, ApJ, 771, 30, doi: 10.1088/0004-637x/771/1/30
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. 2016, in 2016 IEEE CVPR, 779–788, doi: 10.1109/cvpr.2016.91
- Redmon, J., & Farhadi, A. 2018, CoRR, doi: 10.48550/arXiv.1804.02767
- Ren, S., He, K., Girshick, R., & Sun, J. 2017, IEEE TPAMI, 39, 1137, doi: 10.1109/tpami.2016.2577031
- Riggi, S., Magro, D., Sortino, R., et al. 2022, CoRR, doi: 10.48550/arXiv.2212.02538
- Riggi, S., Bordiu, C., Magro, D., et al. 2023, CoRR, doi: 10.48550/arXiv.2301.02804
- Ruder, S. 2016, CoRR, doi: 10.48550/arXiv.1609.04747a
- Sadeghi, M., Javaherian, M., & Miraghaei, H. 2021, AJ, 161, 94, doi: 10.3847/1538-3881/abd314
- Santos, M., Alonso, D., Bull, P., Silva, M. B., & Yahya, S. 2015, in PoS(AASKA14), 021, doi: 10.22323/1.215.0021
- Scaramella, R., Amiaux, J., Mellier, Y., et al. 2022, A&A, 662, A112, doi: 10.1051/0004-6361/202141938
- Sharma, V. 2022, IJGASR, 1, 23, doi: 10.55938/ijgasr.v1i1.4
- Shimabukuro, H., Hasegawa, K., Kuchinomachi, A., Yajima, H., & Yoshiura, S. 2022, PASJ, 75, S1, doi: 10.1093/pasj/psac042
- Slijepcevic, I. V., Scaife, A. M. M., Walmsley, M., et al. 2022, MNRAS, 514, 2599, doi: 10.1093/mnras/stac1135
- . 2023, CoRR, doi: 10.48550/arXiv.2305.16127
- Sortino, R., Magro, D., Fiameni, G., et al. 2023, Experimental Astronomy, 56, 293, doi: 10.1007/s10686-023-09893-w
- Stahl, N., Falkman, G., Karlsson, A., & Mathiason, G. 2020, in Communications in Computer and Information Science (Springer), 556–568, doi: 10.1007/978-3-030-50146-4_41

- Tan, M., & Le, Q. V. 2019, CoRR,
doi: 10.48550/arXiv.1905.11946
- . 2021, CoRR, doi: 10.48550/arXiv.2104.00298
- Tan, M., Pang, R., & Le, Q. V. 2020, in 2020 IEEE/CVF CVPR, nil, doi: 10.1109/cvpr42600.2020.01079
- Tang, H., Scaife, A. M. M., & Leahy, J. P. 2019, MNRAS, 488, 3358, doi: 10.1093/mnras/stz1883
- Tang, H., Scaife, A. M. M., Wong, O. I., & Shabala, S. S. 2022, MNRAS, 510, 4504, doi: 10.1093/mnras/stab3553
- Taylor, L., & Nitschke, G. 2018, in 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 1542–1547, doi: 10.1109/SSCI.2018.8628742
- Tieleman, T., Hinton, G., et al. 2012, Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf
- Tingay, S. J., Goeke, R., Bowman, J. D., et al. 2013, PASA, 30, e007, doi: 10.1017/pasa.2012.007
- Tody, D. 1986, in Instrumentation in Astronomy VI, ed. D. L. Crawford, Vol. 0627, International Society for Optics and Photonics (SPIE), 733 – 748, doi: 10.1117/12.968154
- Trott, C. M., Jordan, C. H., Midgley, S., et al. 2020, MNRAS, 493, 4711, doi: 10.1093/mnras/staa414
- Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, CoRR, doi: 10.48550/arXiv.1706.03762
- Wang, X., Wei, J., Liu, Y., et al. 2021, Universe, 7, doi: 10.3390/universe7070211
- Wayth, R. B., Lenc, E., Bell, M. E., et al. 2015, PASA, 32, doi: 10.1017/pasa.2015.26
- White, R. L., Helfand, D. J., Becker, R. H., Glikman, E., & de Vries, W. 2007, ApJ, 654, 99, doi: 10.1086/507700
- Wong, O. I., Garon, A. F., Alger, M. J., et al. 2024, MNRAS, 536, 3488, doi: 10.1093/mnras/stae2790
- Wu, C., Wong, O. I., Rudnick, L., et al. 2019a, MNRAS, 482, 1211, doi: 10.1093/mnras/sty2646
- Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. 2019b, Detectron2, <https://github.com/facebookresearch/detectron2>
- Y., Z., & Y., Z. 2004, A and A, 422, 1113, doi: 10.1051/0004-6361:20040141
- Zhang, H., Li, F., Liu, S., et al. 2022a, CoRR, doi: 10.48550/arXiv.2203.03605
- Zhang, Z., Jiang, B., & Zhang, Y. 2022b, PASP, 134, 064503, doi: 10.1088/1538-3873/ac67b1

This paper was built using the Open Journal of Astrophysics L^AT_EX template. The OJA is a journal which provides fast and easy peer review for new papers in the

astro-ph section of the arXiv, making the reviewing process simpler for authors and referees alike. Learn more at <http://astro.theoj.org>.