RAS Techniques and Instruments

RASTAI **3,** 174–197 (2024) https://doi.org/10.1093/rasti/rzae013
Advance Access publication 2024 April 3

Transfer learning for galaxy feature detection: Finding giant star-forming clumps in low-redshift galaxies using Faster Region-based Convolutional Neural Network

Jürgen J. Popp [1]★ Hugh Dickinson [1] Stephen Serjeant [1] Mike Walmsley [2] Dominic Adams [3] Lucy Fortson [3] Kameswara Mantha,[3] Vihang Mehta [4] James M. Dawson,[5] Sandor Kruk[6] and Brooke Simmons[7]

[1]*School of Physical Sciences, The Open University, Milton Keynes MK7 6AA, UK*
[2]*Jodrell Bank Centre for Astrophysics, Department of Physics & Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK*
[3]*School of Physics and Astronomy, University of Minnesota, 116 Church Street SE, Minneapolis, MN 55455, USA*
[4]*IPAC, California Institute of Technology, Mail Code 314-6, 1200 E. California Blvd., Pasadena, CA 91125, USA*
[5]*Centre for Radio Astronomy Techniques & Technologies, Rhodes University, Artillery Road, Grahamstown 6140, South Africa*
[6]*ESAC/ESA, Camino Bajo del Castillo, s/n. Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain*
[7]*Physics Department, Lancaster University, Lancaster LA1 4YB, UK*

Accepted 2024 April 1. Received 2024 March 8; in original form 2023 August 16

ABSTRACT

Giant star-forming clumps (GSFCs) are areas of intensive star-formation that are commonly observed in high-redshift ($z \gtrsim 1$) galaxies but their formation and role in galaxy evolution remain unclear. Observations of low-redshift clumpy galaxy analogues are rare but the availability of wide-field galaxy survey data makes the detection of large clumpy galaxy samples much more feasible. Deep Learning (DL), and in particular Convolutional Neural Networks (CNNs), have been successfully applied to image classification tasks in astrophysical data analysis. However, one application of DL that remains relatively unexplored is that of automatically identifying and localizing specific objects or features in astrophysical imaging data. In this paper, we demonstrate the use of DL-based object detection models to localize GSFCs in astrophysical imaging data. We apply the Faster Region-based Convolutional Neural Network object detection framework (FRCNN) to identify GSFCs in low-redshift ($z \lesssim 0.3$) galaxies. Unlike other studies, we train different FRCNN models on observational data that was collected by the Sloan Digital Sky Survey and labelled by volunteers from the citizen science project 'Galaxy Zoo: Clump Scout'. The FRCNN model relies on a CNN component as a 'backbone' feature extractor. We show that CNNs, that have been pre-trained for image classification using astrophysical images, outperform those that have been pre-trained on terrestrial images. In particular, we compare a domain-specific CNN – '*Zoobot*' – with a generic classification backbone and find that *Zoobot* achieves higher detection performance. Our final model is capable of producing GSFC detections with a completeness and purity of $\geq$0.8 while only being trained on $\sim$5000 galaxy images.

Key words: Machine Learning – Deep Learning – Data Methods – Object Detection – Transfer Learning – Galaxies: Structure.

1. INTRODUCTION

Deep field observations of high-redshift star-forming galaxies with the *Hubble Space Telescope* (*HST*) showed galaxy morphologies which differ from low-redshift galaxies. The dominating spiral and elliptical shapes in the local Universe are replaced by more irregular and chaotic morphologies at higher redshifts (Cowie, Hu & Songaila 1995; van den Bergh et al. 1996; Elmegreen et al. 2005, 2007, 2009; Förster Schreiber et al. 2009, 2011; Guo et al. 2015, 2018). While these early *HST*-based studies suggest that the formation of galaxies with disc morphologies happened late in the cosmological timeline, recent studies using data from the *JWST* Early Release observations (Ferreira et al. 2022) and the *JWST* CEERS observations (Ferreira et al. 2023) find a high number of regular disc galaxies already at early times. With the longer wavelength filters and higher spatial resolution from *JWST* more faint morphological features of galaxies could be resolved revealing different morphologies for previously peculiar galaxy types.

Hα line emission and rest ultraviolet (UV)/optical continuum emissions show that most galaxies at $z > 1$ are dominated by several giant star-forming knots or 'clumps' (GSFCs, or clumps for short) which appear much more luminous and larger in extent than H II regions of local galaxies. Unlensed observations report clump sizes of $\sim$1 kpc (Elmegreen et al. 2007; Förster Schreiber et al. 2011) and stellar masses ranging 10^7–$10^9\,M_\odot$ (Elmegreen et al. 2007; Guo et al. 2012, 2018; Zanella et al. 2019; Mehta et al. 2021). For

★ E-mail: jurgen.popp@open.ac.uk

these extended regions of star-formation, highly elevated specific star-formation rates (sSFRs) have been observed (Guo et al. 2012, 2018; Fisher et al. 2016).

However, clumps that are observed in high-redshift galaxies are likely to be unresolved. Observations by *HST* from lensed galaxies found clump sizes of ∼30–100 pc (Livermore et al. 2012; Adamo et al. 2013; Cava et al. 2017) and from recent *JWST* Early Release observations of lensed galaxies at $z = 1$–8.5 clumps with sizes of <10 to 100s of pc have been detected (Claeyssens et al. 2023). These *JWST* observations also revealed clump masses as low as $10^5\,M_\odot$. Other reports indicate that kpc-scale clumps observed at redshifts of $z \sim 0.1$–0.3 consist of smaller coalesced clumps which are not resolved with existing instruments (Overzier et al. 2009; Fisher et al. 2014; Messa et al. 2019).

The formation and evolution of GSFCs are still debated in the literature. There are believed to be two principal modes of GSFC formation: (1) formation by gravitational instabilities in a gas-rich disc (Elmegreen & Elmegreen 2005; Bournaud, Elmegreen & Elmegreen 2007; Bournaud et al. 2013; Mandelker et al. 2014; Romeo & Agertz 2014; Fisher et al. 2016) and (2) formation due to galaxy interactions and mergers (Conselice, Yang & Bluck 2009; Mandelker et al. 2016; Zanella et al. 2019). However, the ways in which clumps do contribute to the evolution of the host galaxy towards modern elliptical and spiral types is not yet fully understood.

The fraction of clumpy galaxies appears to peak at ∼ 55–65 per cent around $z \sim 2$ (Guo et al. 2015; Shibuya et al. 2016) but this fraction decreases with decreasing redshift (e.g. Adams et al. 2022). Due to the scarcity of clumpy galaxies in the local Universe most surveys of clumpy galaxies have focused on intermediate- and high-redshift galaxies. Therefore, their evolution and properties have not been fully studied for a continuous redshift range between $0 \leq z \lesssim 0.3$. Comparable studies for local galaxies are faced with the challenge of identifying enough clumps in galaxies to base population statistics on a reasonable sample size. Extensive surveys like the Sloan Digital Sky Survey (SDSS, York et al. 2000), the Dark Energy Camera Legacy Survey (DECaLS, Dey et al. 2019), and the Hyper Suprime-Cam Subaru Strategic Program (HSC SSP, Aihara et al. 2018) are providing wide field imaging data that make systematic searches for large numbers of low-redshift clumpy galaxies possible but are limited by the resolution constraints of ground-based telescopes.

With forthcoming instruments like the *Euclid* space telescope and wide-field surveys like the Vera Rubin Observatory Legacy Survey of Space and Time, vast amounts of high-resolution imaging data of local galaxies will become available.

Such huge data volumes require automatic analysis. Deep Learning (DL), and in particular Convolutional Neural Networks (CNNs, e.g. LeCun, Bengio & Hinton 2015), have been successfully applied to image classification tasks in astrophysical data analysis (for an overview see Huertas-Company & Lanusse 2023). However, one application of DL, that of automatically identifying and localizing specific objects or features in astrophysical imaging data either through object detection (e.g. Huertas-Company et al. 2020) or image segmentation (e.g. Aragon-Calvo 2019; Burke et al. 2019; Merz et al. 2023; Zavagno et al. 2023), has only been recently used in astrophysical data analysis.

Modern object detection algorithms like the Faster Region-based Convolutional Neural Network framework (Faster R-CNN or FRCNN for short, Ren et al. 2015) are widely used in 'terrestrial' applications, e.g. self-driving cars or face-recognition software. Those networks usually incorporate a pre-trained classification backbone which can be used for transfer learning and so the whole object detection model only needs fine-tuning for a specific use case to be able to produce a high detection performance.

To quantify the benefits of transfer learning for astrophysical image data this paper tests '*Zoobot*' (Walmsley et al. 2023) as a feature extraction backbone for object detection in galaxy images. *Zoobot* is a classification-CNN which has been already pre-trained on morphological features from >1000 000 galaxies and will be benchmarked against FRCNN models with feature extraction backbones that have been trained using terrestrial (i.e. not related to galaxies) imaging data.

Even fine-tuning a classifier or an object detection model still requires *some* training data. A major challenge for astronomy is the lack of sufficiently large labelled data sets to train supervised DL models. Previous studies have used simulated training images with known labels (e.g. Burke et al. 2019; Huertas-Company et al. 2020; Ginzburg et al. 2021) or classifications from publicly available catalogues (e.g. Chan & Stott 2019). In contrast, the object detection models that we describe in this paper were trained using observational data labelled by volunteers from the citizen science project 'Galaxy Zoo: Clump Scout' (GZCS, Adams et al. 2022; Dickinson et al. 2022).

To assess the sample size required to obtain a scientifically useful GSFC detection performance, we train the different FRCNN models using training data sets with different sizes. The results of these tests can be used to estimate the required effort if labels are needed to fine-tune the FRCNN model for a new data set.

This paper is organized as follows. Section 2 provides a brief introduction to the techniques of object detection with DL, followed by a section describing the data sources and the necessary pre-processing steps (Section 3). Section 4 explains the details of our model design and training process together with an evaluation of the achieved detection performance. We describe applications of the object detection models on different sets of imaging data in Section 5 and discuss the implications of our findings in Section 6. The paper concludes with a summary of our results in Section 7.

2. DL FOR OBJECT DETECTION

Object detection is one of many technologies used in computer vision and image processing. Its main application is in detecting and recognizing instances of *semantic* objects, i.e. objects of meaningful physical origin, in images or videos (e.g. Dasiopoulou et al. 2005). Object detection algorithms generally make use of Machine Learning or DL to produce automatic detections, localizations, and classifications on large data sets like video-feeds or image catalogues. It is commonly used in computer vision tasks like face recognition or traffic sign recognition in driver assistance systems (see e.g. Erhan et al. 2014; Pavel, Tan & Abdullah 2022).

Among computer vision tasks, object detection can be seen as a combination of image classification and object localization (e.g. Szegedy, Toshev & Erhan 2013). Whereas image classification attempts to assign a label or class to an entire image, object localization tries to locate a single instance of a specific object in an image and marks it with a tightly cropped bounding box centred on the instance. Object detection not only tries to locate all instances of multiple objects in an image but also assigns a label to each instance found. Dealing with a variable number of objects and instances of different sizes are the main challenges for object detection algorithms.

In this paper, we train a version of the Faster R-CNN architecture proposed by Ren et al. (2015). We chose an object detection algorithm over object instance segmentation algorithms, like Mask R-CNN (He et al. 2017), as we are mainly interested in the object localization and

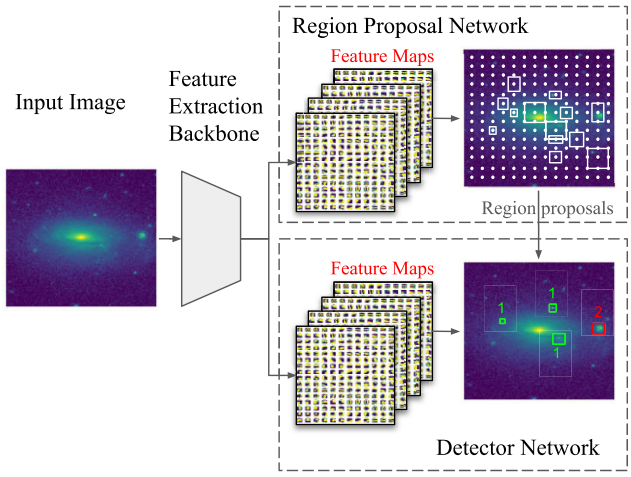


Figure 1. Schematic view of the Faster R-CNN architecture. The input image is fed into the backbone feature extractor and the resulting feature maps are used as input for the RPN and the detector network. Sample anchor points (dots in the upper right galaxy image) and anchor boxes (rectangles in the upper right galaxy image) are shown for the RPN. The detector network then uses the region proposals from the RPN and the feature maps from the feature extraction backbone to output the final, classified detections of class 1 objects (rectangles with label 1 in the lower right galaxy image) and class 2 objects (rectangles with label 2 in the lower right galaxy image), for example.

expect that the resolution of our imaging data is too low to extract useful segmentation masks of detected objects.

Briefly, FRCNN models comprise three components. First, a CNN is used as a 'backbone' to extract spatial hierarchies of patterns or features from an input image. These features are then used as input to two separate sub-networks (Fig. 1).

The first is called the Region Proposal Network (RPN) and identifies (or proposes) regions in the image that are likely to contain objects. It sets anchor points at every pixel location of the output feature map of the feature extracting backbone and places at each anchor point position a set of k anchor boxes with default sizes and aspect ratios. The RPN optimizes these initial anchor boxes depending on the overlap with the ground-truth object boxes from the training set and generates a twofold output. The prediction scores ('objectness') for the two generic classes, 'object' and 'background', and for each of the k anchor boxes are one output. The other output are regression coefficients for each of the four attributes: centre coordinates x, y, the width w, and the height h, of the k anchor boxes.

The second sub-network, the detector network, is then used to classify the contents of the proposed regions of class 'object' into one of the n final object categories using the corresponding features for those parts of the image that were extracted by the backbone CNN. It also further refines the predicted bounding boxes.

The final output of the FRCNN model is a collection of rectangular bounding boxes identifying groups of pixels in the image that contain objects and a classification identifying the type of object that each box contains.

3. DATA

In this section, we describe the criteria used to select the galaxy images that we use to train our models and the methods used to label them. Table 1 lists the number of galaxy images that remain after each stage of our image selection and labelling pipeline.

Table 1. Reductions applied for the final galaxy sample.

Selection	Galaxy count
GZ2	304 122
With spectroscopic redshift	243 500
With $0.02 \leq z \leq 0.25$	225 085
With $f_{\text{featured}} > 0.5$ (GZCS)	53 613
After consensus aggregation	20 683
With bulge markings removed	20 646
After padding	18 772

Our starting point in this paper is the set of galaxy images that were used for the GZCS citizen science project (Adams et al. 2022, see also Appendix A1 for a brief description of the image creation process). GZCS ran on the Zooniverse platform (www.zooniverse.org/) from the 2019 September 19 to the 2021 February 11. For the GZCS project, the participating volunteers were asked to annotate visible clumps on image cutouts of 53 613 SDSS galaxies. These were selected from over 300 000 galaxies that were classified by volunteers who contributed to the 'Galaxy Zoo 2' citizen science project (GZ2, Willett et al. 2013). For GZCS, galaxies were selected for which the majority of GZ2-volunteers answered with 'No' to the question: 'Is the galaxy simply smooth and rounded, with no sign of a disc?', since it seemed unlikely that galaxies containing prominent GSFCs would match this description. The sample was further reduced to only contain galaxies with a documented spectroscopic redshift between $0.02 \leq z \leq 0.25$. The redshift constraint was applied to ensure that most of the clumps, which were anticipated to be of ∼kpc size, appear as point-like sources throughout all sample images (see also Appendix A1).

Volunteers who participated in the GZCS project were asked to identify the locations of the clumps within the selected galaxies. The annotation process is described in detail by Adams et al. (2022), but a brief summary is given here. First the volunteers were asked to mark the central bulge of the galaxy to help them recognize that this should not be interpreted as a clump even though it has a similar appearance. The volunteers were also equipped with a 'normal clump marker' and an 'unusual clump marker'. The latter allowed volunteers to mark foreground stars which might overlap with the centre galaxy's spatial extent and could look similar to clumps in terms of colours and being a comparable point source in the SDSS images. We used the volunteers' markings for normal and unusual clumps as the classification label in our training set and further refer to them as 'normal' and 'odd' clumps, respectively.

Each galaxy image was inspected and annotated by at least 20 independent volunteers and their markings were then aggregated to derive consensus clump locations. Dickinson et al. (2022) developed a framework to aggregate two-dimensional image annotations into a consensus label which further reduced the sample to 20 683 clumpy galaxies (Table 1).

Fig. 2 shows that the vast majority of the aggregated clumps (∼ 99 per cent) fall into a central square of half the original side length of each image with the target galaxy at its centre. To reduce the computing time needed to train the network and to make the algorithm focus on the central galaxy only, the outer area was later cropped to the size of the central square during the image augmentation step for model training. Furthermore, any remaining bulge markings and clumps located within a 10 per cent pixel margin from the borders of each image were removed. This resulted in a final set of 18 772 galaxies containing 39 745 aggregated clump annotations (Table 1).

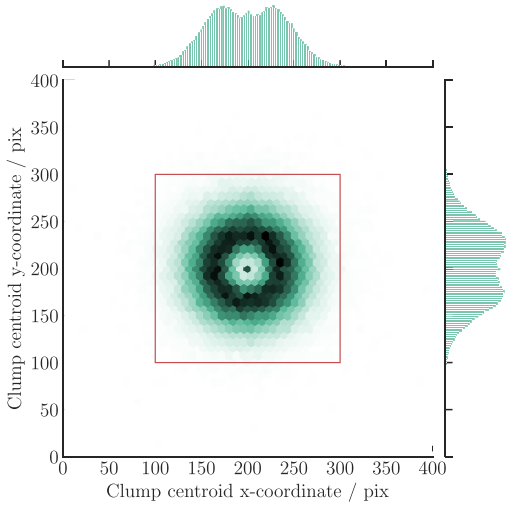


Figure 2. Spatial distribution of the clump centroids within the image dimensions from the final set of 18 772 galaxies containing 39 745 annotated clumps, before central bulge markings and clumps too close to the cropped image dimensions have been removed. The cropped imaged dimensions are marked by the square and contain ∼ 99 per cent of the annotated clumps. After the image creation process, the median corresponding value for 1 pixel in the RGB-composite images is ∼0.2 arcsec (see Appendix A1).

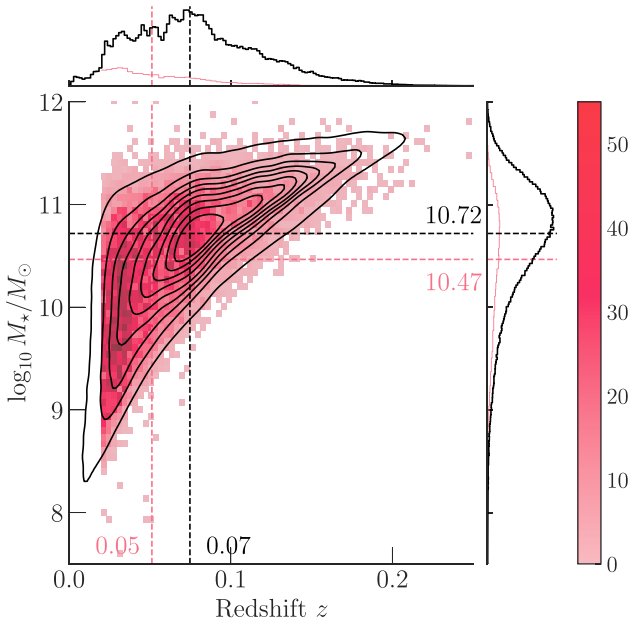


Figure 3. The galaxy stellar mass as a function of redshift for the galaxies used for developing the object detection models. Overlaid with contours are the galaxies from the parent GZ2 sample with spectroscopic redshifts. The dashed lines mark the median of each distribution.

Fig. 3 illustrates the mass-redshift distribution of our final sample of galaxies with at least one off-centre clump we used for developing the object detection models. We did not apply further limits to our selection of host galaxies as these are used primarily to train the object detection models. Stellar mass estimates for galaxies in our final sample were taken from the SDSS DR7 MPA-JHU value-added catalogue (Kauffmann et al. 2003; Brinchmann et al. 2004) and range in order of $10^7\,\mathrm{M}_\odot \gtrsim M_\star \gtrsim 10^{12}\,\mathrm{M}_\odot$.

4. DEVELOPING THE OBJECT DETECTION MODEL

In the following sections, we describe the specific FRCNN-implementation with the different feature extraction backbones that we compare with each other (Section 4.1). The training set-up and execution is covered in Section 4.2 and in Section 4.3 we explain the post-processing steps that we perform. Finally, in Section 4.4 the detection performance is evaluated.

4.1 Feature extraction backbone

The CNN as a feature extraction backbone plays an important role in the FRCNN object detection framework. With its ability to extract features from the input images it provides crucial inputs for the RPN and the region classifier or detector network. In this paper, we investigate the performance of five backbone CNNs that use different initial weight configurations and training strategies (Table 2).

The *Zoobot* model is a CNN developed to classify galaxies based on their morphological features by Walmsley et al. (2023). We use a version of *Zoobot* based on the ResNet50 architecture (He et al. 2016), which has been trained to morphologically classify galaxies in SDSS, *HST*, and DECaLS imaging data. In total, our *Zoobot* version has been trained using more than 1000 000 classified galaxy images, and training on more images is still on-going.

The domain-specificity of the *Zoobot* model, combined with the diversity of instruments that provided its training data, suggests that:

(1) Its weights may extract features that are very well suited for the task of identifying clumps in galaxies and,
(2) Using it as an FRCNN feature extraction backbone may allow the FRCNN model to be more easily adapted to novel imaging data sets using fewer labelled training examples.

However, neither of these hypotheses are necessarily true. CNNs used in computer vision applications have been very often pre-trained on massive 'terrestrial' data sets. The high-level abstraction of such a CNN might have reached a high enough generalization level after being pre-trained on a sample like the ImageNet data set, which consists of 1.2 million images belonging to more than 1000 classes and has become a standard challenge for benchmarking DL models in computer vision (ImageNet Large Scale Visual Recognition Challenge, Russakovsky et al. 2015).

To test whether an FRCNN model with *Zoobot* as a feature extracting backbone (henceforth named *Zoobot-backbone*) does indeed provide better performance and flexibility, we tested the performance versus a ResNet50 feature extraction backbone that has been pre-trained using the ImageNet data set. We refer to the backbone trained using these terrestrial images as *Imagenet-backbone*.

In addition to the unmodified *Zoobot* model, we also tested a version of *Zoobot* that we specifically train as a classifier to distinguish clumpy and non-clumpy galaxies (*Zoobot-clumps-backbone*). With this approach we address the possibility that specific objects might have been underrepresented in the data set used for training *Zoobot* and the feature extraction backbone has not learned to extract features resembling GSFCs in galaxies well enough. We outline the fine-tuning process in Appendix B.

We kept the weights of the ResNet50 architecture with 48 convolutional layers in four blocks (see He et al. 2016 for details of the ResNet architecture) fixed for all three models described above and only allowed the additional layers of the RPN and the detector network to adjust during training. In this mode, we tested the transfer learning ability of the three backbone feature extractors.

Table 2. Backbone classifiers used during the Faster R-CNN model development.

	Model name	Feature extractor architecture	Weight initialization	Learning mode	Trainable blocks
1	*Imagenet-backbone*	ResNet50	ImageNet	Transfer learning	–
2	*Imagenet-backbone-finetuned*	ResNet50	ImageNet	Fine-tuning	2, 3, and 4
3	*Zoobot-clumps-backbone*	ResNet50	*Zoobot Clumps*	Transfer learning	–
4	*Zoobot-backbone*	ResNet50	*Zoobot*	Transfer learning	–
5	*Zoobot-backbone-finetuned*	ResNet50	*Zoobot*	Fine-tuning	2, 3, and 4

For testing a fine-tuning approach, we added two more variants of the *Imagenet-backbone* and *Zoobot-backbone* models, in which some weights were allowed to vary during the FRCNN model training. Specifically, we allowed the upper ResNet50 blocks (2, 3, and 4) to vary their weights. We refer to these partially trainable backbones as *Imagenet-backbone-finetuned* and *Zoobot-backbone-finetuned*.

All five FRCNN models (Table 2) were parametrized in the same way. The models expect the same RGB-composite images used for GZCS (see Appendix A1) as input together with a list of bounding box corner pixel-coordinates and class-labels as derived from the consensus aggregation process (Dickinson et al. 2022). As the galaxy cutouts were scaled to a size of 400 × 400 pixels for GZCS, they vary in pixel scale between 0.1 and 1.3 arcsec pixel^{-1} with a median of 0.2 arcsec pixel^{-1}, depending on the angular size of the central galaxy. The RPN was initialized with default anchor box sizes of 32 × 32, 64 × 64, 128 × 128, 256 × 256, and 512 × 512 pixels and aspect ratios of 0.5, 1.0, and 2.0.

4.2 Model training

We trained the models over 20 runs each with increasing training sample sizes, which we divided into random train/validation/test splits of size 70 per cent/20 per cent/10 per cent (see Appendix C for details). The images were augmented by random horizontal and vertical flips and cropped to a size of 200 × 200 pixels, keeping the central galaxy but removing the parts where very few clumps have been marked (see Section 3 and Fig. 2). This helped to improve training time and made the FRCNN model focus on the central galaxy while removing the parts which would only produce unnecessary anchor boxes not containing any clumps.

For all models and run-groups we used the 'adaptive moment estimation' optimiser (*Adam*, Kingma & Ba 2014) with an initial learning rate of 10^{-4}. We trained every configuration over 120 epochs each using PyTorch's 'distributed data parallel' configuration on a multi-GPU environment of eight NVIDIA A100 GPUs and used batch sizes of 32.

We monitored the training and validation loss at each epoch. Fig. 4 shows examples for run 2 (423 training samples), run 10 (1949 training samples), and run 20 (13 140 training samples). Training and validation loss converge at the beginning of the training runs for all models. We observed overfitting for the models *Imagenet-backbone*, *Imagenet-backbone-finetuned*, and *Zoobot-clumps-backbone* where the validation loss diverges from the training loss again. Overfitting reduces with increasing size of the training data set for the models *Imagenet-backbone* and *Zoobot-clumps-backbone* (e.g. run 2 versus run 20, Fig. 4), but remains strong for the FRCNN model *Imagenet-backbone-finetuned*. In contrast, both unchanged Zoobot models, *Zoobot-backbone* and *Zoobot-backbone-finetuned* are stable at all sample sizes tested and express a robust behaviour over all 120 epochs.

4.3 Post-processing

4.3.1 Non-maximum suppression

The output of an object detection model consists of many overlapping bounding boxes with different objectness scores attached. We applied a process called non-maximum suppression (NMS) which uses the Jaccard distance $J(A, B)$ (Jaccard 1912) to determine the Intersection over Union (IoU) of the areas A and B:

$$J(A, B) = \frac{A \cap B}{A \cup B} \tag{1}$$

and keeps only the bounding box with the highest score from the overlapping bounding boxes.

Fig. 5 illustrates the result of NMS after being applied to a sample galaxy image. In most cases the raw model proposals consisted of multiple small bounding boxes which were fully contained within larger boxes with a lower objectness score. A threshold of IoU $\geq$ 0.2, which we applied in this paper, proved to be suitable to discard most of those larger bounding box proposals while keeping partially overlapping, adjacent clump proposals.

The output of this step is a set of clump candidates with either the 'normal' or 'odd' classification.

4.3.2 Spatial exclusion and aggregation of clump candidates

With the seeing of SDSS, clumps are assumed to appear as point-like sources (see also Section 3) with a light profile equal to the instrumental point spread function (PSF). Therefore, we merged adjacent clump candidates not further apart than one r-band PSF-full width at half-maximum (FWHM) in each subject image into one single detection. We measured the distance between the clump centroids, i.e. the midpoint of the surrounding bounding boxes, and set the new location of the merged clump to the midpoint between the clump centroids. A new label was assigned so, that if at least one of the clumps is classified as an 'odd' clump, that label is assigned to the new aggregated clump (Fig. 6).

Next, we removed all clump candidates located outside the extent of the host galaxy. For each galaxy, we smoothed the r-band image to generate a segmentation map with the photutils Python package (Bradley et al. 2023), that outlines the extent of the host galaxy. The r-band images were smoothed using a Gaussian kernel with the size of the corresponding r-band PSF-FHWM. Accounting for the low surface brightness of the galaxy outskirts, we applied a threshold of 1σ per pixel above the background noise to outline the central galaxy. Clumps located outside the galaxy outline were discarded (Fig. 7).

4.4 Detection performance

We used a two-step approach to evaluate detection performance. In the first step, during the model training phase, performance was evaluated on the validation sample set of each run-group (20 per cent, see Section 4.2) after each epoch. Based on the general detection

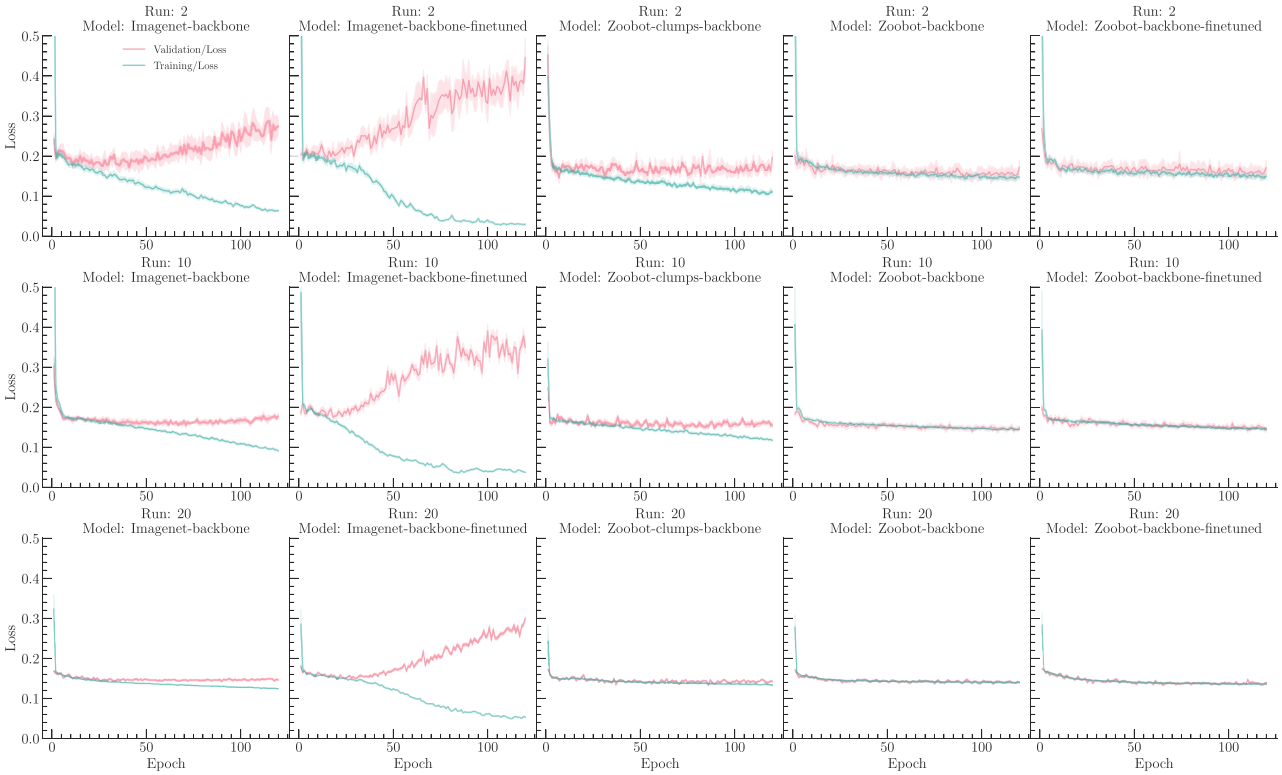
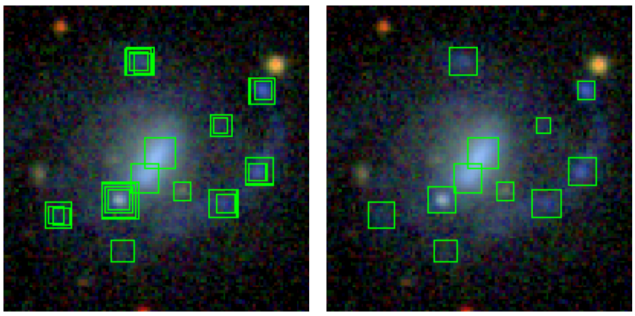


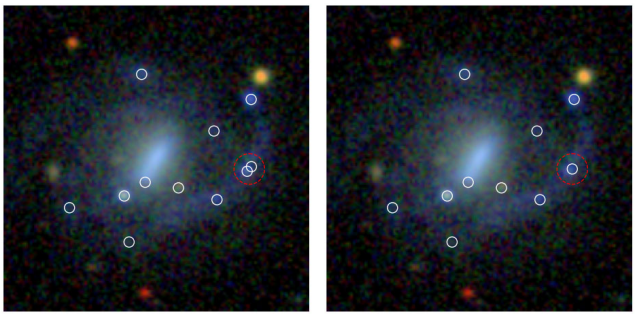
Figure 4. The mean training and validation loss of the Faster R-CNN for run 2 (423 samples), run 10 (1949 samples), and run 20 (13 140 samples), where the shaded areas show the corresponding 1σ standard error of the loss.



(a) Before non-maximum suppression.

(b) After non-maximum suppression.

Figure 5. Comparison of bounding boxes for a sample galaxy before and after NMS.



(a) Before merging.

(b) After merging.

Figure 6. If the distance between clump centroids is less than the *r*-band PSF-FWHM, the clumps are merged into one with a new location at the midpoint between the clump centroids. If at least one of the clumps is classified as an 'odd' clump, that label is assigned to the new aggregated clump.

performance and taking into account when the model first showed signs of overfitting (e.g. Fig. 4), we determined a best model version for each run. In the second step, we compared the trained model versions in an astrophysical context using the test sample set of each run-group (10 per cent, Section 4.2).

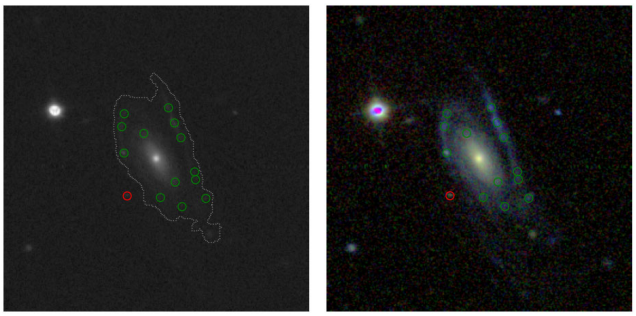
4.4.1 Determining the best model using the COCO metrics

We evaluated the detection performance simultaneously on the validation set during the training process for all models using the metrics from the COCO Object Detection Challenge (Lin et al. 2014). A short description of the metrics can be found in Appendix D.

We exported a current model version after each epoch during the training process and calculated average precision (AP) and average recall (AR) and $F1$-scores for detection score thresholds ranging from 0.1 to 0.9 after post-processing (see Section 4.3) the results. Based on the $F1$-score the best models were chosen from this pool of model versions, separately for each training run. The model versions we chose for training run 20, which used all of the 18 772 galaxy images for training, validation, and testing, are listed in Table 3.

4.4.2 Completeness and purity for model detections

In astrophysical applications, domain-specific post-processing steps are very often necessary and detections need to be reassessed with the help of additional morphological and physical parameters. Moreover, the appropriate object detection score threshold (or objectness thresh-



(a) r-band FITS with outline of the galaxy segmentation map. (b) RGB image with clumps marked.

Figure 7. Applying the galaxy segmentation map to exclude detections located outside the host galaxy.

Table 3. Best model versions chosen for training run 20, based on a prediction score threshold of ≥0.3 and IoU threshold of ≥0.5.

Model name	Epoch	AP	AR	f_1
Imagenet-backbone	20	0.40	0.39	0.40
Imagenet-backbone-finetuned	20	0.42	0.82	0.56
Zoobot-clumps-backbone	60	0.07	0.73	0.13
Zoobot-backbone	120	0.44	0.68	0.54
Zoobot-backbone-finetuned	120	0.44	0.67	0.53

old, see Section 2) very often depends on the scientific questions asked. A complete sample can be a more appropriate outcome than a pure sample and *vice versa*.

As our clumps are based on assumed point-like sources not resolved by SDSS, we applied a method which is different to the IoU approach used by the COCO metrics (see Appendix D) to compute the overlap between the ground-truth (i.e. the volunteers' labels from GZCS) and our model detections. A successful clump candidate is counted, if the distance between the centroid of a predicted clump candidate and the centroid of the ground-truth clump is less than 0.75 of the image-specific PSF-FWHM.

Fig. 8 shows completeness against purity for the final models from training run 20 and with an increasing detection score threshold c_n from 0.0 to 0.9. The models using a feature extraction backbone initialized with the unmodified *Zoobot* weights generally show a better detection performance compared with the *Imagenet-backbone* model. Only if we allow the ImageNet-based model to adjust its weights for the last convolutional blocks of the backbone feature extractor (*Imagenet-backbone-finetuned*), completeness and purity do reach similar levels. The object detection model which uses a backbone feature extractor pre-trained on detecting clumps (*Zoobot-clumps-backbone*) appears to not benefit from the even more domain-specific initialization.

We observe a similar performance ranking of the different FRCNN models if completeness is plotted against purity for all training runs (Fig. 9, as an example for a score threshold of ≥0.3).

4.4.3 *Performance per relative clump flux bin*

Guo et al. (2015) suggested a GSFC definition based on the ratio of clump to host galaxy UV-luminosity. A clump needs to exceed a ratio of 8 per cent to be classified as a GSFC and not as a normal star-forming region. We used this definition to determine the completeness and purity of our FRCNN models in terms of astrophysically

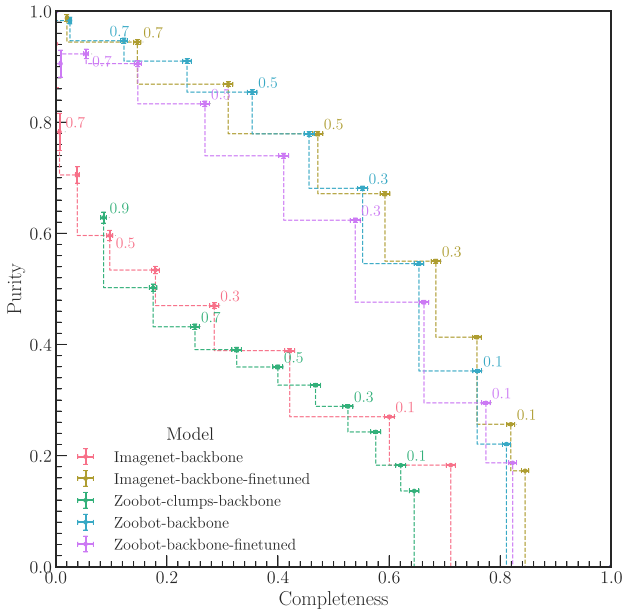


Figure 8. Purity and completeness for all five models. The detection score threshold c_n is increasing from 0.0 (right) to 0.9 (left) as indicated by the annotations. A clump candidate is considered to be a True Positive (TP), if the distance between the centroid of a predicted clump candidate and the centroid of the ground-truth clump is less than 0.75 of the image-specific PSF-FWHM. All models have been trained on the full sample size (run 20).

relevant GSFC-detections and also applied a 3 per cent flux ratio threshold, similar to Adams et al. (2022).

We measured the u-band flux from SDSS for each clump candidate (for details see Appendix E), as it is closest to UV, and retrieved the host galaxy u-band flux from the SDSS DR15 PhotoPrimary table (Aguado et al. 2019). The log of the ratio of clump to host galaxy flux was grouped into equally spaced bins of size 0.5.

We then compared the completeness of the discovered clumps against the log of the relative flux for each model and each training run. As an example, Fig. 10a plots completeness for the models trained on 5061 galaxy images (run 15), which is considerably smaller than the full set of training samples. The figure shows the highest completeness for the model *Zoobot-backbone* over most of the flux ratio range. Only for the faintest clumps does the completeness drop to 0.4 and below the completeness values of the model *Imagenet-backbone-finetuned*. *Zoobot-backbone-finetuned* reaches similar completeness levels within the error margins but the remaining two models, *Zoobot-clumps-backbone* and *Imagenet-backbone*, are significantly lower in completeness.

Note, that the completeness is already high ($\gtrsim$ 0.8) in the flux ratio ranges of ≥ 3 per cent and ≥ 8 per cent for the FRCNN model *Zoobot-backbone* after we trained the model on only 5061 galaxy images. For comparison, Fig. 10b shows the completeness versus relative flux ratio for run 20, where we trained the models on the full set of training samples. If trained on the full set of 13 140 galaxy images, *Imagenet-backbone-finetuned* now reaches the highest completeness values, although the completeness of the other models, apart from *Imagenet-backbone*, are similar in the clump-specific flux ratio ranges.

The completeness shown by the two FRCNN models using a version of the unmodified *Zoobot* as their feature extraction backbone is similar after either being trained on a reduced train set (run 15) or

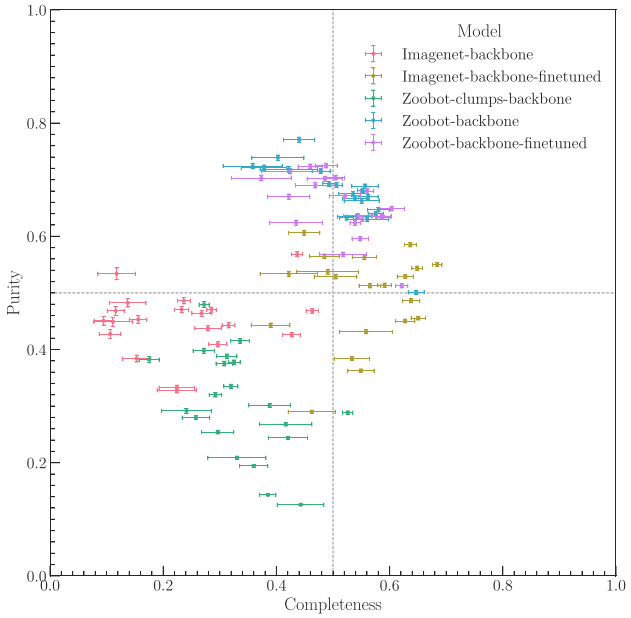


Figure 9. Model completeness and purity for a score threshold ≥0.3. Error bars show the 95 per cent confidence interval. The different points for each model represent different training runs with different sample sizes but are not labelled for better visibility.

on the full train set (run 20). This indicates that a reasonable detection performance can already be achieved through a relatively small labelled training sample. Figs 10a and b also show, that *Imagenet-backbone-finetuned* and *Zoobot-clumps-backbone* require a larger training set to achieve comparable completeness performance. Fine-tuning a terrestrial feature extraction backbone, in our case *Imagenet-backbone-finetuned*, does result in a high completeness performance but only through using a large data set. In contrast, applying the same model in transfer learning mode (*Imagenet-backbone*) does not result in comparable completeness levels as seen from the FRCNN models with domain-specific feature extracting backbones.

We also calculated purity for the same flux ratio bins for all models and training runs. Figs 11a and b show the two resulting plots after training run 15 and 20. The achieved purity levels are highest for relatively bright clumps (flux ratios > 3 per cent). The purity of the model detections tend to be closer together for all models except the *Zoobot-clumps-backbone* model. Unlike before, where we compared the completeness, there does not appear to be a clear difference between models with terrestrial and domain-specific feature extraction backbones.

To further illustrate the differences between the five feature extraction backbones and the increase in purity and completeness, if the clump candidates are limited to relevant flux ratios, we show completeness against purity plots for clump candidates with a u-band flux ratio of ≥ 3 per cent in Fig. 12a and for a ratio ≥ 8 per cent in Fig. 12b.

5. DETECTION RESULTS

We applied the five different FRCNN models to two different data sets, (1) GZCS (SDSS images) and (2) Hyper Suprime-Cam (HSC) images.

In Section 5.1, we compare the detections from each model on SDSS imaging data to the ground-truth data set, which consists of the GZCS galaxies with annotated clumps from the volunteers. In Section 5.2, we describe a first test of transferability of the object detection models after we applied the models 'out-of-the-box', i.e. without any additional training, on the HSC imaging data.

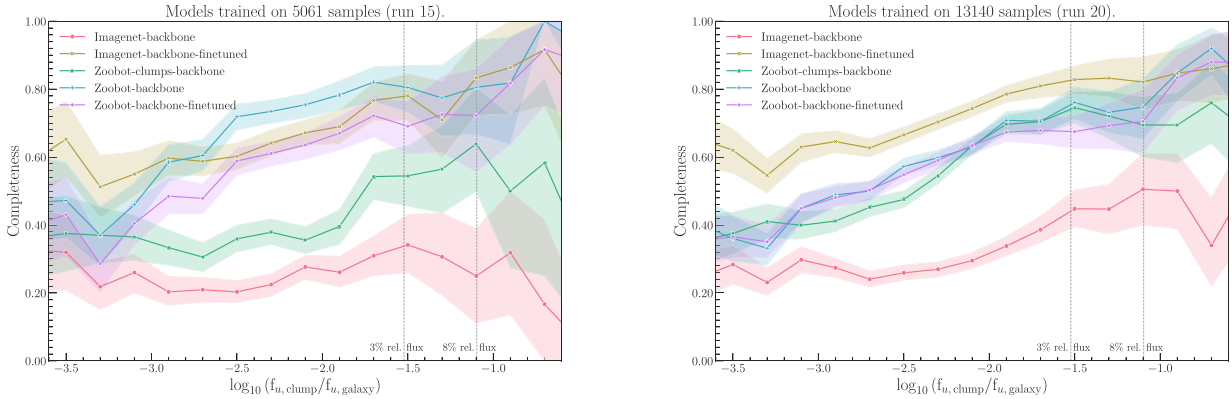
Unless otherwise indicated, all clump candidate detections were made with models developed on the full training set (run 20) of the GZCS sample and for detection scores ≥0.3.

5.1 Clump candidate detection – GZCS images

5.1.1 Visual comparison

We first compared the detections by the five FRCNN models visually with each other and with the GZCS volunteers' labels. Fig. 13 shows the differences in detection performance (Section 4.4) for one example galaxy.

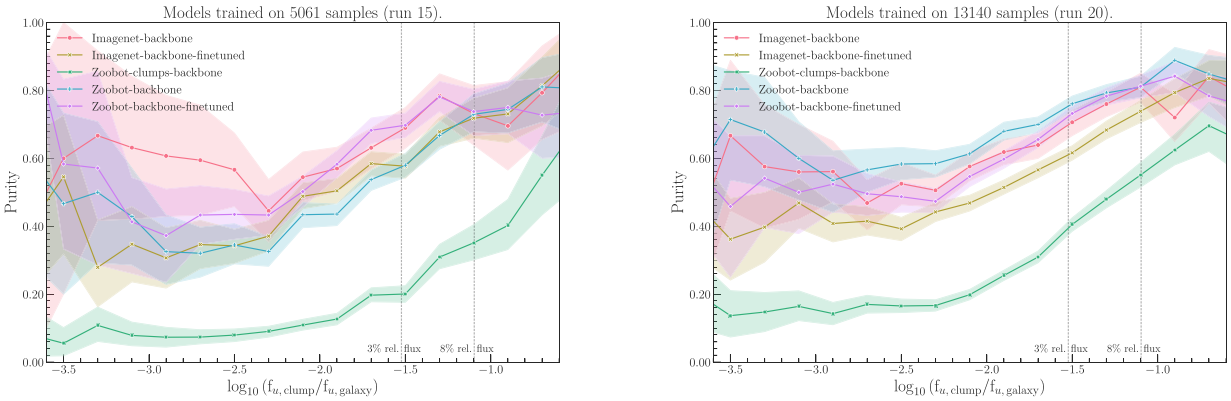
This example illustrates the general observations we made while the resulting detections went through a visual vetting process. *Zoobot-clumps-backbone* tends to produce far bigger bounding boxes whereas the model *Imagenet-backbone* generally detects fewer



(a) Completeness for the models trained on 5,061 galaxy images (run 15).

(b) Completeness for the models trained on 13,140 galaxy images (run 20).

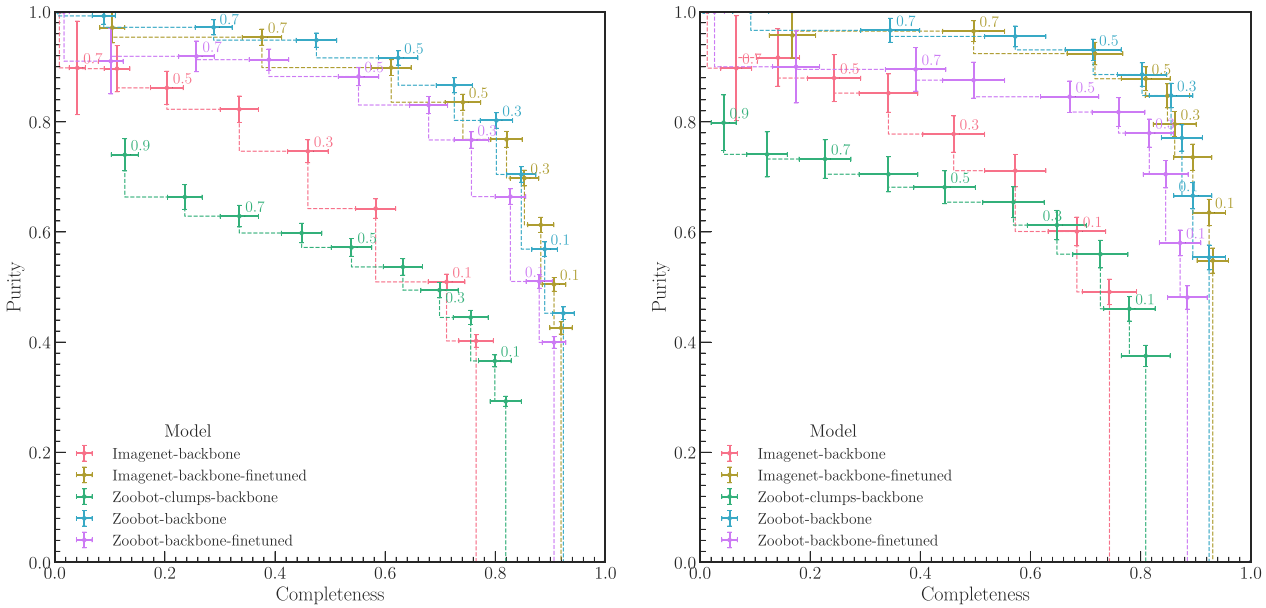
Figure 10. Model completeness with respect to the volunteers' labels from GZCS per relative clump flux for training run 15 and run 20 (training sample size of 5061 and 13 140, respectively, and for a score threshold of ≥0.3). Shaded areas showing the 95 per cent confidence interval. The 3 per cent and 8 per cent threshold for the flux ratio are indicated with vertical dashed lines.



(a) Purity for the models trained on 5,061 galaxy images (run 15).

(b) Purity for the models trained on 13,140 galaxy images (run 20).

Figure 11. Model purity with respect to the volunteers' labels from GZCS per relative clump flux for training run 15 and run 20 (training sample size of 5061 and 13 140, respectively, and for a score threshold of ≥0.3). Shaded areas showing the 95 per cent confidence interval. The 3 per cent and 8 per cent threshold for the flux ratio are indicated with vertical dashed lines.



(a) Purity and completeness for u-band flux ratio ⩾ 3%.

(b) Purity and completeness for u-band flux ratio ⩾ 8%.

Figure 12. Purity and completeness for all five models for clumps and clump candidates with a measured u-band flux ratio compared with the host galaxy of ≥ 3 per cent and ≥ 8 per cent. The detection score threshold c_n is increasing from 0.0 (right) to 0.9 (left) as indicated by the annotations. Error bars show the 95 per cent confidence interval. All models have been trained on the full sample size (run 20).

clump instances than the other models, which explains the relatively low completeness for *Imagenet-backbone*.

We observed that the models often predict additional clump instances which were not marked by the volunteers. These can be non-blue objects like foreground stars but in many cases these detections show typical visual clump characteristics and were not marked by a high enough number of volunteers for a consensus label. This can be seen from the second image in in Fig. 13, which shows the original GZCS volunteers' annotations before the aggregation process (see Section 3 and Dickinson et al. 2022).

Figs G1, G2, and G3 in the Appendix provide additional comparisons of SDSS galaxy examples.

5.1.2 Comparison to the GZCS sample

We determined the number of detected normal and odd clumps predicted by the FRCNN models and the clump per galaxy ratios (Table 4).

The clump predictions vary considerably between the models. For example, the models *Zoobot-clumps-backbone* and *Imagenet-backbone-finetuned* predict far more clumps than the other models and what has been originally annotated by the volunteers. These models detect clumps in almost all galaxies from the GZCS set, whereas the models *Zoobot-backbone* and *Zoobot-backbone-finetuned* both detect ∼ 15 per cent fewer clumpy galaxies. The model *Imagenet-backbone* predicts the lowest number of clumps in only ∼ 60 per cent of the GZCS galaxies.

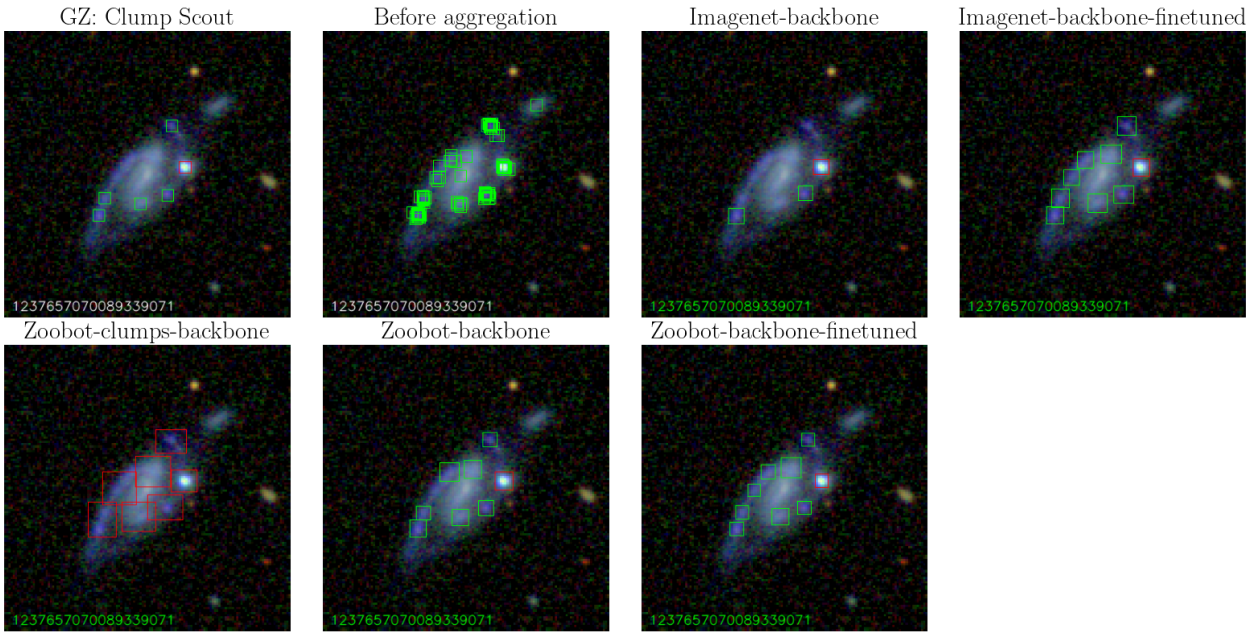


Figure 13. Clump candidates for a SDSS galaxy for all models. From left to right, top row: GZCS volunteers' labels, original GZCS volunteers' labels before the aggregation process, and *Imagenet-backbone* and *Imagenet-backbone-finetuned*. From left to right, bottom row: *Zoobot-clumps-backbone*, *Zoobot-backbone*, and *Zoobot-backbone-finetuned*. Normal clumps are marked with green boxes, odd or unusual clumps with red boxes. The images are labelled with the SDSS-DR7 object number of the host galaxy. Detection score threshold is ≥0.3.

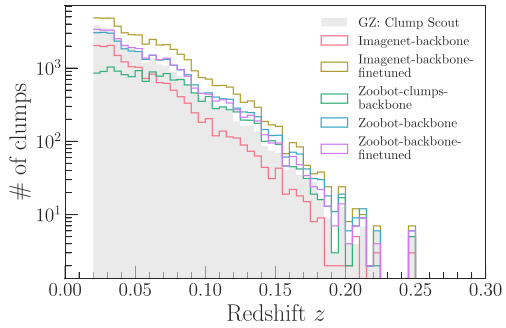
Table 4. Clumpy galaxies, clumps, and clumps per galaxy ratios for the final models per *u*-band clump/galaxy flux ratio for a detection threshold of ≥0.3.

Model	*u*-Band flux ratio $f_{u,\,clump}/f_{u,\,galaxy}$	Clumpy galaxies Count	Clumps, all Count	Clumps, all Average	Clumps, normal Count	Clumps, normal Average	Clumps, odd Count	Clumps, odd Average
GZCS	All	18 772	39 745	2.12	29 619	1.89	10 126	1.20
	≥0.03	7329	10 106	1.38	7464	1.34	2642	1.07
	≥0.08	2810	3406	1.21	2007	1.22	1399	1.05
Imagenet-backbone	All	11 191	17 804	1.59	16 223	1.57	1581	1.05
	≥0.03	4427	5359	1.21	4382	1.22	977	1.03
	≥0.08	1796	2015	1.12	1277	1.14	738	1.03
Imagenet-backbone-finetuned	All	17 836	48 659	2.73	46 271	2.66	2388	1.06
	≥0.03	8300	12 698	1.53	10 995	1.54	1703	1.04
	≥0.08	3085	3882	1.26	2592	1.33	1290	1.04
Zoobot-clumps-backbone	All	17 491	69 966	4.00	14 625	1.61	55 341	3.95
	≥0.03	8472	15 149	1.79	2802	1.23	12 347	1.76
	≥0.08	2951	4027	1.36	605	1.20	3422	1.32
Zoobot-backbone	All	15 858	31 199	1.97	29 018	1.93	2181	1.05
	≥0.03	7210	9712	1.35	7913	1.35	1799	1.05
	≥0.08	2752	3272	1.19	1879	1.23	1393	1.04
Zoobot-backbone-finetuned	All	15 937	32 923	2.07	30 301	2.03	2622	1.07
	≥0.03	7122	9684	1.36	7800	1.36	1884	1.06
	≥0.08	2821	3330	1.18	1871	1.22	1459	1.05

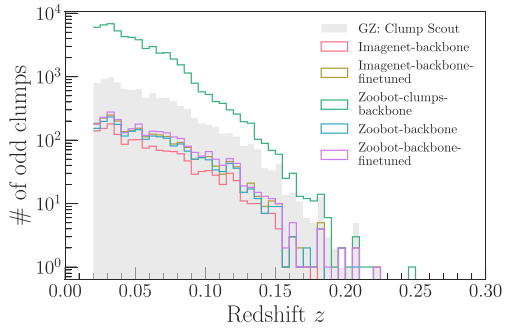
These differences result not only from normal clump candidates but also from the number of detected odd clump candidates. While both models based on the unmodified *Zoobot* feature extractor, detect a similar number of normal clumps, they detect fewer odd clumps compared with the number marked by the GZCS volunteers. We also observed, that *Imagenet-backbone-finetuned* detects ∼ 60 per cent more normal clumps, whereas a model like *Zoobot-clumps-backbone* predicts more than five times the number of odd clumps annotated in the GZCS sample.

Comparing the number of clumps per galaxy (Table 4), we noted that the models *Zoobot-backbone* and *Zoobot-backbone-finetuned* are closest to the distribution of the GZCS sample but tend to find slightly more normal clumps per galaxy.

Considering the redshift of the host galaxies, the normal clump candidates predicted by *Zoobot-backbone* and *Zoobot-backbone-finetuned* are close to the GZCS-distribution of the normal clumps. Although these models do detect fewer clumps for host galaxies at redshift 0.02–0.04, the other models differ substantially from



(a) Normal clump candidates.



(b) Odd clump candidates.

Figure 14. Histograms of detected clump candidates per redshift bin by the five different models, separately for normal and odd clumps. Detection score threshold is ≥0.3. The underlying shaded distributions are from GZCS volunteers' labels.

the GZCS distribution (Fig. 14a). Specifically, the model *Imagenet-backbone-finetuned* predicts far more clump candidates on the full redshift range of our sample of host galaxies, whereas the model *Imagenet-backbone* detects fewer clumps compared with the number of clumps resulting from the GZCS project. We also observe that the number of predicted normal clumps from the model *Zoobot-clumps-backbone* is far lower for redshifts ≲0.08 but much higher for predicted odd clumps at all redshifts (Fig. 14b).

We also applied the same thresholds for the clump to host galaxy u-band flux ratio as we did for assessing the detection performance of the models (see Section 4.4.3). The number of detected clumps with a u-band flux ratio of ≥ 3 per cent (≥ 8 per cent) reduces to $\sim 20-30$ per cent ($\sim 6-11$ per cent). Again, the normal clump candidates predicted by the models *Zoobot-backbone* and *Zoobot-backbone-finetuned* are close to the GZCS-distribution of the normal clumps if either of the flux ratio thresholds are applied, but now the number of detected odd clumps is also similar to the number of odd clumps identified from the GZCS project (Table 4).

Plotted on a $(g-r)/(r-i)$ colour–colour diagram, the normal clump candidates predicted by most of the models (Fig. 15) tend to be bluer than the clumps marked by the volunteers from GZCS. Only the normal clumps detected by *Zoobot-clumps-backbone* are redder in comparison to the GZCS sample (Fig. 15c).

The predictions from this model also differ notably for odd clumps (Fig. 16). *Zoobot-clumps-backbone* not only detects far more odd clumps, but they also tend to be bluer and spread over a wider $(g-r)$ and $(r-i)$ range than the GZCS-distribution. This is in contrast to the odd clump candidate detections made by the other models which are closely resembling the $(g-r)$ and $(r-i)$ colour distribution from the GZCS sample and are following a tight locus of the colour–colour space, indicative of foreground stars.

To see whether the host galaxy has an effect on our model predictions, we further compared the host galaxies' sSFRs, stellar masses, and redshifts for which our models did detect normal clumps to the host galaxies for which the volunteers from GZCS have marked normal clumps. We obtained stellar masses $M_\star$ and sSFR for the host galaxies from the SDSS DR7 MPA-JHU value-added catalogue (Kauffmann et al. 2003; Brinchmann et al. 2004). We caution the reader that these figures are only intended to compare the performance characteristics of our models with the performance characteristics of the GZCS volunteers. We make no claims about the true distribution of clumpy galaxies as a function of redshift, stellar mass, or sSFR.

Apart from *Zoobot-clumps-backbone*, the distributions of the sSFR of the host galaxies containing predicted clumps are similar to those of the GZCS galaxies (Fig. 17). The majority of the normal clump candidates are found in star-forming [$\log_{10}(\mathrm{sSFR}) \gtrsim -11.2$] galaxies. At the higher end of the redshift range of our sample galaxies ($z > 0.05$), the distributions of the predicted clumps still resemble that of the GZCS-distribution for the models *Zoobot-backbone* and *Zoobot-backbone-finetuned* (Figs 17d and e). However, the model predictions from *Imagenet-backbone* and *Imagenet-backbone-finetuned* are notably different compared with the GZCS-galaxies for $\log_{10}(\mathrm{sSFR}) \gtrsim -11$ (Figs 17 a and b).

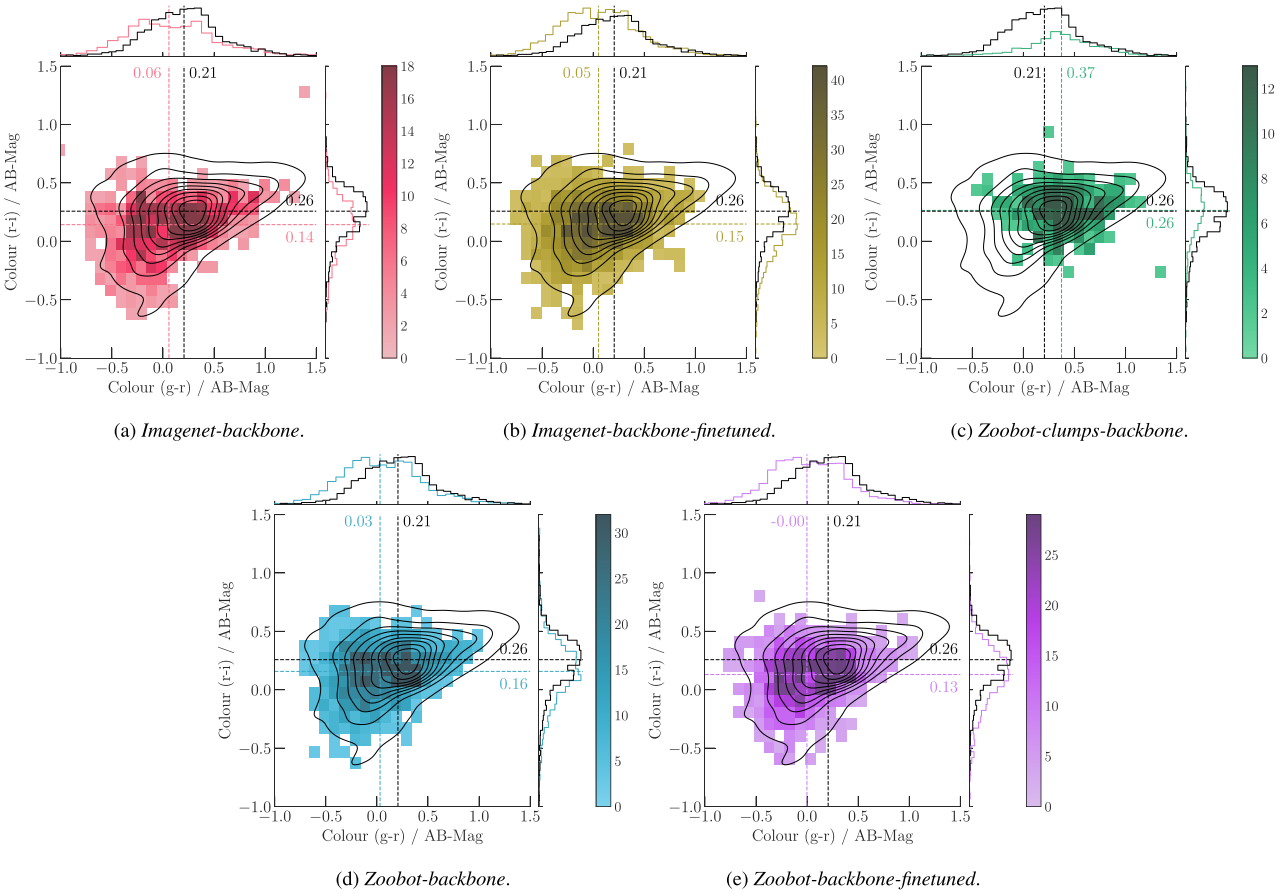
We also observed, that the model *Imagenet-backbone* tends to predict fewer normal clumps in more massive galaxies [$\log_{10}(M_\star) \gtrsim 9.4$] for redshifts $z < 0.05$ (Fig. 18a), whereas *Imagenet-backbone-finetuned* detects more normal clumps in comparison to the GZCS-distribution for galaxies with stellar masses of $\log_{10}(M_\star) \gtrsim 10.2$ (Fig. 18b). *Zoobot-clumps-backbone*, on the other hand, produces normal clump candidates which are predominantly located in higher redshift galaxies with $\log_{10}(M_\star) \gtrsim 10.2$ (Fig. 18c). The *Zoobot-backbone* and *Zoobot-backbone-finetuned* models closely resemble the host galaxy distribution with clumps labelled by the GZCS volunteers (Figs 18d and e).

5.1.3 *Clump catalogue release and use*

We applied the model *Zoobot-backbone* on the 53 613 galaxies from the original GZCS set (Table 1) and released a catalogue containing the detected clump candidates and their estimated properties along with this paper. The catalogue contains normal and odd clump candidates for a detection score ≥0.3 together with measured fluxes and magnitudes for each of the *ugriz*-filter bands from SDSS. Table 5 describes the columns in compact form.

When using the catalogue, we recommend excluding entries with the label 'odd clumps' as these are very likely non-clump candidates (i.e. foreground stars, background galaxies, or other point-like sources). Currently, our model framework does not include a classification score for the labels 'normal clump' and 'odd clump' that would allow for a more specific selection. Purity and completeness can be varied by filtering on the detection score, where a higher score threshold results in higher purity and *vice versa* (see Fig. 8).

We also caution the reader that we did not apply any survey completeness limits to this catalogue. Such completeness limits need to be applied for further scientific analysis of our sample. For an example, where galaxy stellar mass and redshift limits were applied, see Adams et al. (2022).



(a) *Imagenet-backbone*. (b) *Imagenet-backbone-finetuned*. (c) *Zoobot-clumps-backbone*.

(d) *Zoobot-backbone*. (e) *Zoobot-backbone-finetuned*.

Figure 15. Colour–colour diagrams for normal clump candidates with a u-band clump/galaxy flux ratio of $\geq$0.08 for each of the five models. The colour bars indicate the specific counts of normal clumps for each $(g-r)/(r-i)$ bin (with bin size of 0.1). The small histograms on the top and right side of the plots are showing the distributions of $(g-r)$ and $(r-i)$, separately. The colour–colour distribution of the clumps annotated by the GZCS volunteers are overlaid with contours and small histograms. Vertical and horizontal lines mark the median colour determined for the normal clumps, which are annotated with the corresponding values.

5.2 Clump candidate detection – HSC images

The HSC SSP (Aihara et al. 2018) partly overlaps with the SDSS footprint, but the wider aperture of the *Subaru Telescope* allows for much deeper imaging with the HSC and exposes many more morphological details.

We did not train the models specifically for the HSC imaging data set or reproduce exactly the same image pre-processing as for the SDSS galaxy images (see Appendix A). We note, however, that the *Zoobot*-based feature extraction backbones had some existing 'memory' learned from galaxy images with a higher spatial resolution than SDSS as such images have already been used in their pre-training as classifiers.

A cross-match between the GZCS-catalogue and the HSC-catalogue found 2424 objects surveyed by SDSS and the HSC SSP for the GZCS sample, which we used for a direct visual comparison of the clump candidates on images with different spatial resolution. Fig. 19 shows such a comparison of two sample galaxies from both catalogues with the clump candidates overlaid. The higher resolution of HSC is immediately apparent and the *Zoobot-backbone* is capable of detecting multiple clump candidates with varying sizes of the bounding boxes on galaxy images from this source.

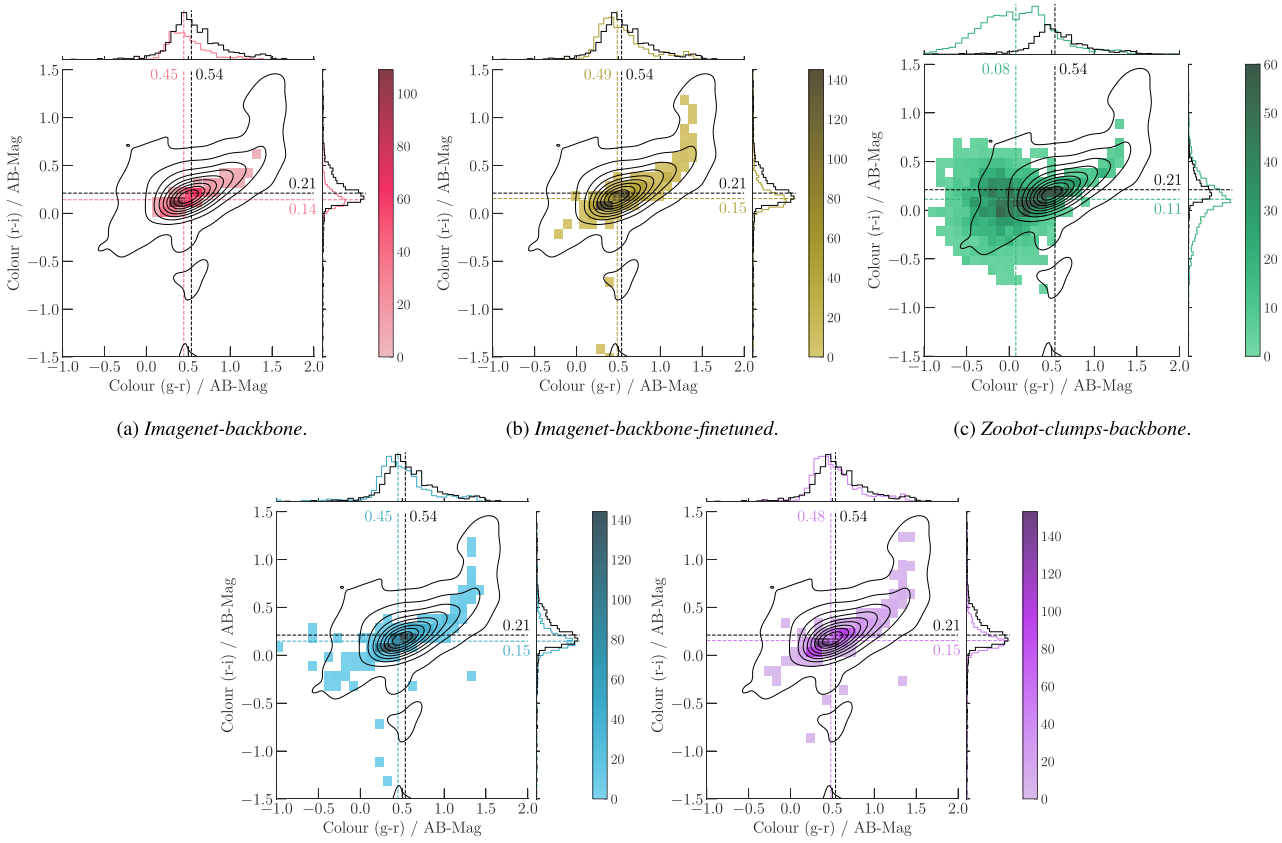
From visual inspection, the *Zoobot-backbone* model appears to produce the most reasonable detections. A comparison with the model *Imagenet-backbone* can be seen from the additional examples in Figs G4, G6, and G6 in the Appendix. *Imagenet-backbone* appears to find far fewer clumps and some bounding boxes for detected objects are too large for clump-like objects.

6. DISCUSSION

We have used a set of galaxy images annotated with clump markings from the GZCS citizen science project (Adams et al. 2022) to train Faster R-CNN models for object detection which are capable of producing plausible predictions of clump locations within the host galaxies.

This supervised DL approach faces several challenges. The first challenge concerns the training (or ground-truth) data from which the models learn. Observational data have the advantage over simulated data as it does not require prior assumptions of the photometric and physical properties of the objects to be detected. On the other hand, even if the completeness of the sample has been corrected against expert labels and probabilistic algorithms applied to aggregate the volunteers' annotations into a consensus location and label (Dickinson et al. 2022), it is likely that the training set contains misidentifications and is missing genuine clump instances.

Another challenge lies in how the imaging data are presented to the neural network. In this study, we have used three channels per image as input for our models (e.g. *gri* bands converted into RGB for



(a) *Imagenet-backbone*. (b) *Imagenet-backbone-finetuned*. (c) *Zoobot-clumps-backbone*.

(d) *Zoobot-backbone*. (e) *Zoobot-backbone-finetuned*.

Figure 16. Colour–colour diagrams for odd or unusual clump candidates with a u-band clump/galaxy flux ratio of $\geq$0.08 for each of the five models. The colour bars indicate the specific counts of odd clumps for each $(g-r)/(r-i)$ bin (with bin size of 0.1). The small histograms on the top and right side of the plots are showing the distributions of $(g-r)$ and $(r-i)$, separately. The colour–colour distribution of the odd clumps annotated by the GZCS volunteers are overlaid with contours and small histograms. Vertical and horizontal lines mark the median colour determined for the odd clumps, which are annotated with the corresponding values.

the SDSS and HSC images). We expect this approach to be superior to using only single-channel input data (e.g. Huertas-Company et al. 2020), but this still needs to be validated.

In comparing the performance of *Zoobot* as a feature extraction backbone against CNNs trained on 'terrestrial' data sets, we have also limited ourselves to these three channels. From the SDSS imaging data, two additional channels are available, namely the u and z band. To overcome these limitations, we are considering using the feature extraction backbones in an ensemble configuration in future works. Possible configurations could consist of:

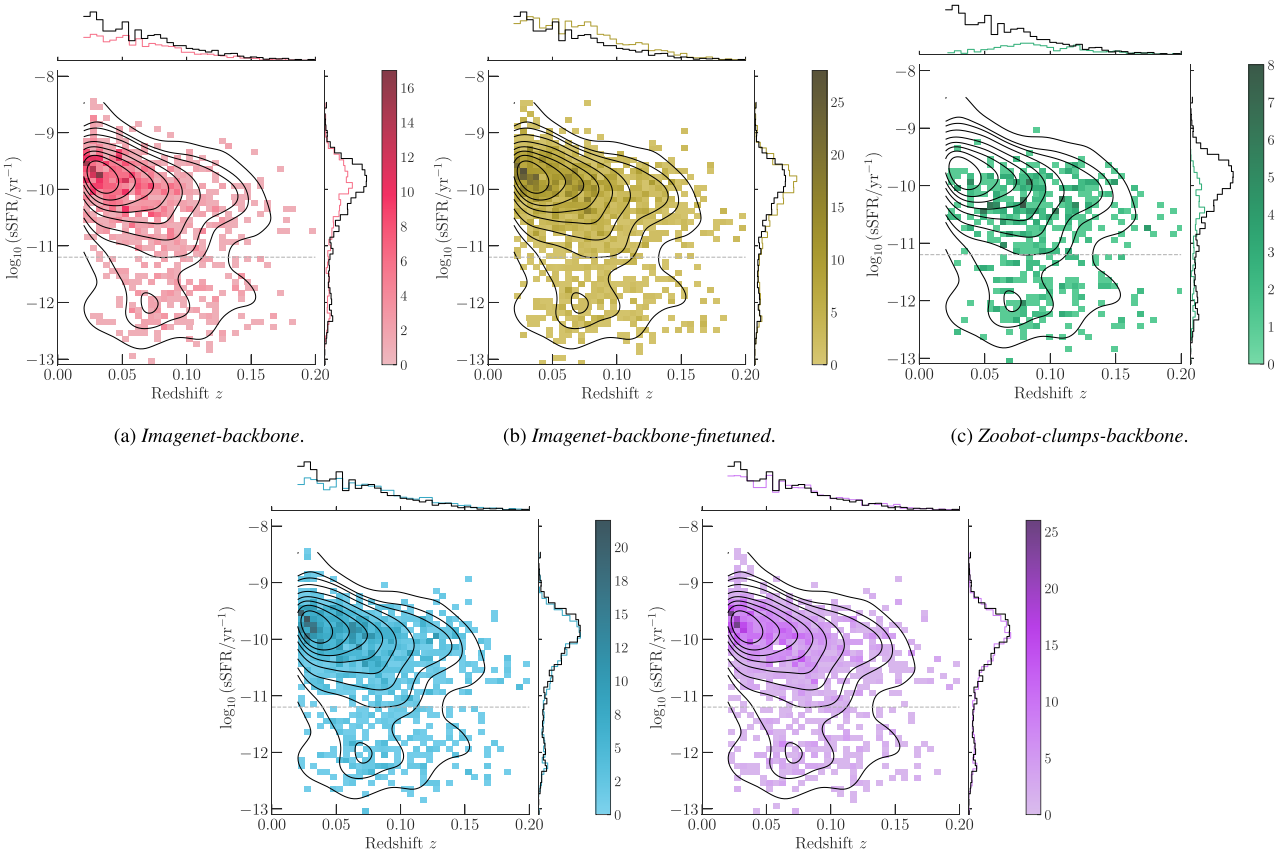
(1) Two ResNet-based feature extraction backbones, where the sixth channel will be left unused,
(2) An ensemble of five backbones, one backbone for each of the *ugriz* bands.

Furthermore, fine-tuning a classification network before being used as a feature extraction backbone, like we tried with the *Zoobot-clumps-backbone* model (see Appendix B), leads to a worse detection performance. This is possibly caused by the feature extraction backbone 'forgetting' previously learned features and now relying on different features for classifying clumpy galaxies. We expected the *Zoobot-clumps-backbone* model to more closely resemble the GZCS population but instead, the model produces much bigger bounding boxes (Fig. 13), many more odd clump candidates (Fig. 14b) and the normal clump candidates are mainly detected in more massive galaxies compared with the GZCS sample (Fig. 18c).

From an astrophysical point of view, we favour a more complete over a more pure set of detections. As Dickinson et al. (2022) pointed out, the volunteers' from GZCS tend to mark more faint features in galaxies as clumps than experts, especially when those features appear blue in colour. The authors also noted, that volunteers are more likely to mark a clump as 'normal', despite being labelled by experts as 'unusual' or 'odd'.

This disagreement makes the GZCS sample less suitable for acting as a benchmark to determine the purity of the model outputs. Instead, we argue that a higher completeness is better suited to scientific analyses. We find support for the robustness of our model completeness after we tested them on galaxy images with simulated clumps from Adams et al. (2022, see Appendix F), where they were able to produce similar performance results. We therefore suggest that a final clump selection process needs to include a refinement of the model output by considering directly observable clump characteristics or derived photometric or physical properties (e.g. colour, flux, and stellar mass estimates of the clumps).

Taking into account the detection performances of the five FRCNN models as described in Section 4.4 and the resulting clump population statistics in comparison to the GZCS data set, we selected the FRCNN model *Zoobot-backbone* as the best and most robust detection model. We argue that:



(a) *Imagenet-backbone*. (b) *Imagenet-backbone-finetuned*. (c) *Zoobot-clumps-backbone*.

(d) *Zoobot-backbone*. (e) *Zoobot-backbone-finetuned*.

Figure 17. sSFR versus redshift z of the host galaxy for normal clump candidates with a u-band clump/galaxy flux ratio of $\geq$0.08. The colour bars indicate the specific counts of clumps for each $\log_{10}$(sSFR) and z bin (with bin sizes of 0.1 and 0.005, respectively). The small histograms on the top and right side of the plots are showing the distributions of $\log_{10}$(sSFR) and z, separately. The distribution of the normal clumps annotated by the GZCS volunteers are overlaid using contours and small histograms. The horizontal dashed line marks the separation between star-forming and quiescent galaxies at $\log_{10}(\mathrm{sSFR}) = -11.2$.

(1) The detected clumps are closest to the GZCS population in terms of total number, clumps per galaxy, and identified galaxies with at least one off-centre clump.

(2) Completeness and purity are among the highest, especially for the clump-galaxy flux-ratio of $\geq$ 3 per cent.

(3) The model performs well after only being trained on relatively small data sets and is robust against overfitting.

(4) It uses the unmodified *Zoobot* CNN as the backbone feature extractor, for which no problem-specific fine-tuning is necessary.

(5) This FRCNN model will likely benefit from the continued training of *Zoobot* for galaxy classifications on more and more imaging data sets.

Utilizing the 'feature knowledge' *Zoobot* has already acquired can help to facilitate the extension of DL-based object detection to a wide range of morphological features observed in galaxies, e.g. bars, rings, or spiral arms. Furthermore, extending Faster R-CNN from object detection to object instance segmentation will add a valuable alternative to 'traditional' segmentation methods using contrast detection techniques with SExtractor (Bertin & Arnouts 1996), for example. A framework like Mask R-CNN (He et al. 2017), which uses a similar architecture to Faster R-CNN, will likely benefit from a pre-trained feature extraction backbone like *Zoobot*.

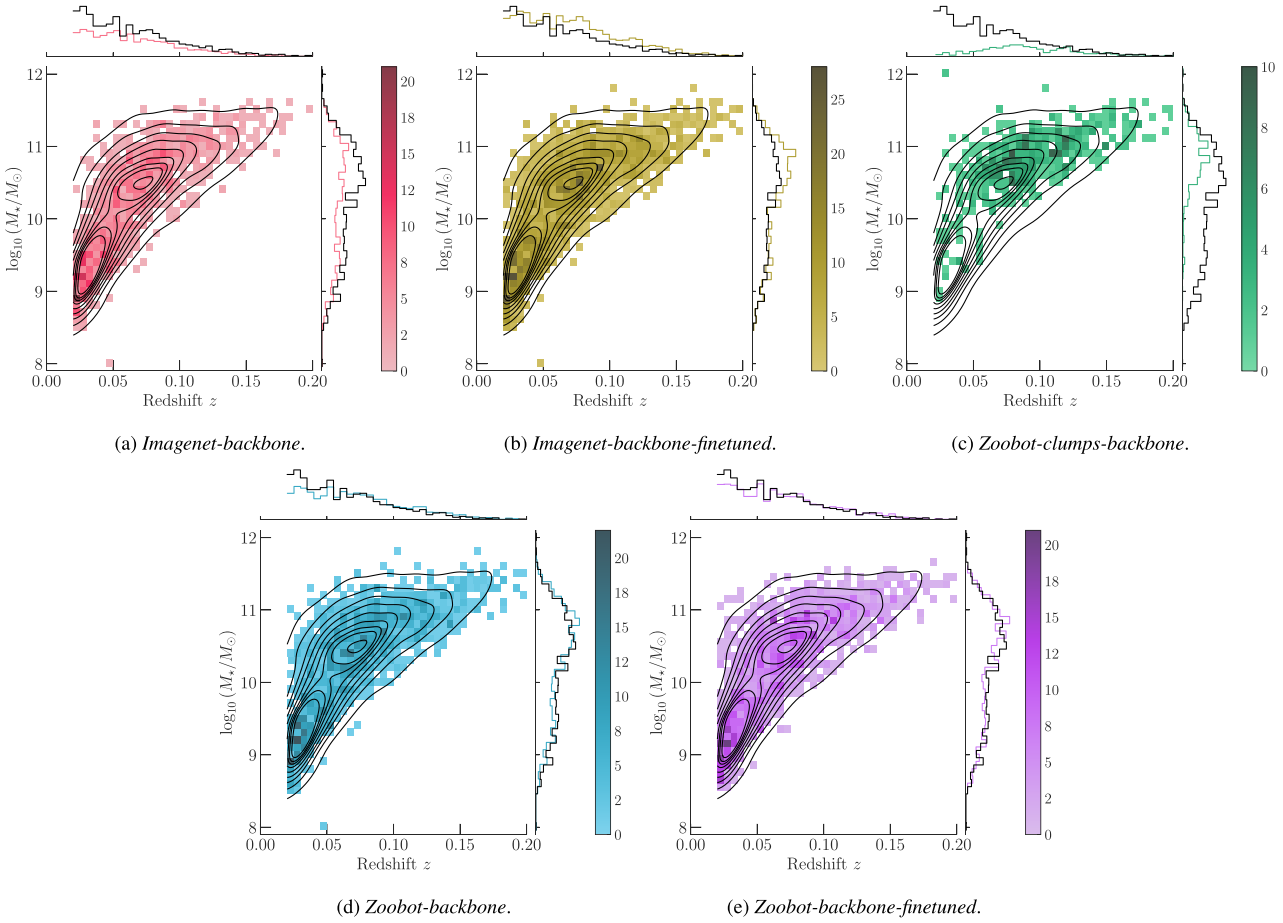
The ability of our models not only to detect clumps but also to classify the detections into normal and odd clumps can greatly facilitate the removal of foreground star contamination in massive sets of galaxy images. We present in this paper first promising results but will continue to train and evaluate our models with new training data which will be specifically labelled for this purpose.

7. CONCLUSIONS

This paper has presented a DL model for detecting GSFCs in low-redshift galaxies. We developed object detection models using the Faster R-CNN framework and trained the models on real observations from the GZCS project. These 18 772 low-redshift galaxy images, taken from the SDSS, were annotated with clump markings by non-expert volunteers and aggregated to 39 745 potential clump locations using a probabilistic aggregation algorithm.

We tested five Faster R-CNN models with different feature extraction backbones, all of which are based on the ResNet50 architecture but initialized with either terrestrial or domain-specific pre-trained weights in different training modes. We trained each model on 20 different training sets with varying sample sizes, ranging from 500 to 18 772 galaxy images with marked clump locations. For all 100 training runs, we evaluated the detection performance using the standard COCO metrics and also determined completeness and purity within the GSFC-specific context after applying necessary post-processing steps.

The key results are summarized in the following points:



(a) *Imagenet-backbone*. (b) *Imagenet-backbone-finetuned*. (c) *Zoobot-clumps-backbone*.

(d) *Zoobot-backbone*. (e) *Zoobot-backbone-finetuned*.

Figure 18. Stellar mass $M_\star$ versus redshift z of the host galaxy for normal clump candidates with a u-band clump/galaxy flux ratio of $\geq$0.08. The colour bars indicate the specific counts of clumps for each $\log_{10}(M_\star)$ and z bin (with bin sizes of 0.1 and 0.005, respectively). The small histograms on the top and right side of the plots are showing the distributions of $\log_{10}(M_\star)$ and z, separately. The distribution of the normal clumps annotated by the GZCS volunteers are overlaid using contours and small histograms.

Table 5. Description of the clump catalogue for the GZCS galaxies.

Columns	Property	Units	Source
1–2	Object IDs		SDSS
3–8	Clump candidate detection score, label and coordinates		
9–18	Clump *ugriz*-fluxes and errors	Jy	
19–33	Clump *ugriz*-magnitudes and corrections	AB mag	
34–37	Clump colours		
38–40	Est. clump/galaxy near-UV flux ratio (u-band)		
41–45	Host galaxy coordinates, redshift and axis ratio		SDSS
46	Host galaxy stellar mass (log)	$\log_{10}(M_\odot)$	SDSS
47	Host galaxy sSFR (log)	$\log_{10}(yr^{-1})$	SDSS
48–52	Host galaxy *ugriz*-fluxes	Jy	SDSS
53–67	Host galaxy *ugriz*-magnitudes and corrections	AB mag	SDSS

(1) DL-based object detection models have been trained on large samples of real observational data instead of simulated data. This paper has shown that Faster R-CNN can be successfully applied for detecting clumps in galaxy images. The models we present in this paper are capable of producing clump candidate detections with a completeness and purity of $\geq$0.8 on SDSS imaging data.

(2) The same models can be used without additional training ('out-of-the-box') on imaging data from the HSC SSP, which have increased spatial resolution compared with the data, that has been used for training the models.

(3) Using *Zoobot* as a feature extraction backbone for the FRCNN model has shown the effectiveness of transfer learning within an astrophysical context. The models using the unmodified *Zoobot* feature extractor are robust against overfitting and produce the best results for detecting GSFCs.

(4) This paper has also shown how domain adaptation made it possible to apply FRCNN models to problem sets which are generally too small for a 'training from scratch' approach. The final model, *Zoobot-backbone*, achieved a high detection performance while only being trained on ∼5000 samples or 39 per cent of the full training

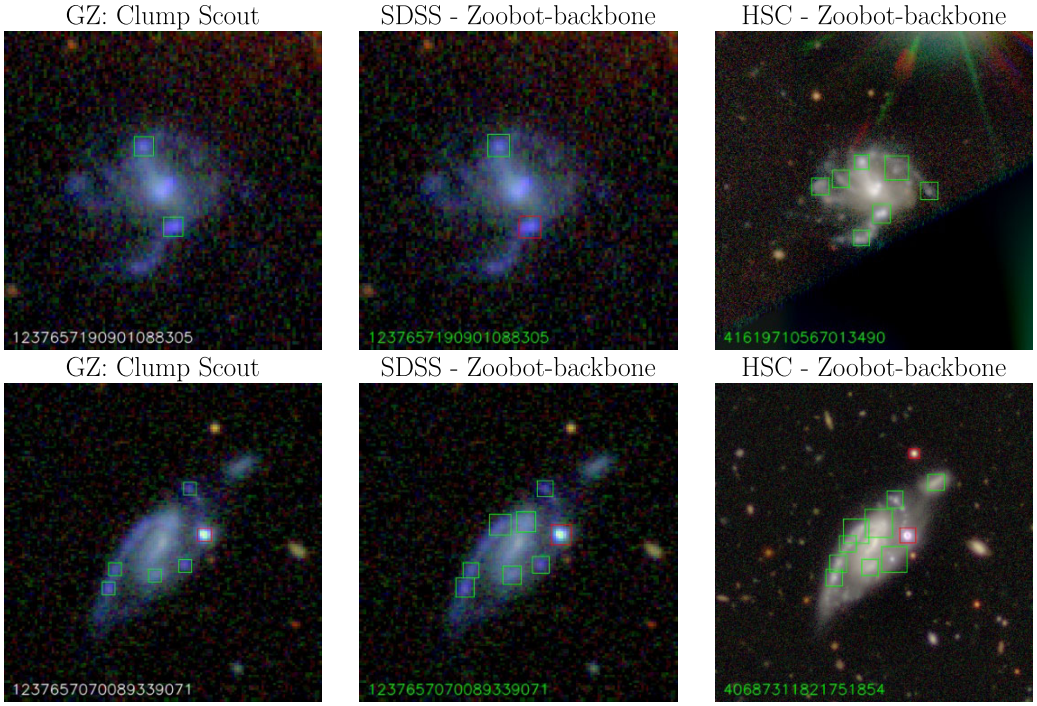


Figure 19. Clump candidates on two example HSC images. Column 1: SDSS image with volunteers' labels from GZCS, column 2: SDSS image with detections by *Zoobot-backbone*, and column 3: HSC image with detections by *Zoobot-backbone*. Normal clumps are marked with green boxes, odd or unusual clumps with red boxes. The images are labelled with their SDSS-DR7 and HSC object numbers, respectively. Detection score threshold is ≥0.3.

set. To achieve good prediction results, the effort of creating labelled data sets can be reduced as training the model does not require large data sets and will be less computationally expensive.

(5) The model output can be made suitable for further scientific analysis after necessary post-processing steps and corrections for sample completeness have been applied.

DATA AVAILABILITY

The data underlying this paper were used in Adams et al. (2022) and can be obtained as a machine-readable table by downloading the associated article data from https://doi.org/10.3847/1538-4357/ac6512. The final models and code are made publicly available at: https://github.com/ou-astrophysics/Faster-R-CNN-for-Galaxy-Zoo-Clump-Scout. A detailed catalogue of giant star-forming clumps, detected for the full set of Galaxy Zoo: Clump Scout galaxies observed by SDSS, can be downloaded from https://doi.org/10.5281/zenodo.8228890.

ACKNOWLEDGEMENTS

We thank Miguel Aragon and the second anonymous referee for their useful and constructive comments that led to improvements in this manuscript.

JP acknowledges funding from the Science and Technology Facilities Council (STFC) Grant code ST/X508640/1. MW acknowledges funding from the Science and Technology Facilities Council (STFC) Grant code ST/R505006/1. DA, LFF, and KBM acknowledge partial funding from the U.S. National Science Foundation Award IIS 2006894 and NASA Award 80NSSC20M0057.

This research made use of the open-source Python scientific computing ecosystem, including NumPy (Harris et al. 2020), Matplotlib (Hunter 2007), seaborn (Waskom 2021), and Pandas (McKinney 2010). This research made use of Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration 2022) and the associated Python libraries APLpy (Robitaille 2019) and photutils Python package (Bradley et al. 2023).

For the DL model development the Python frameworks TensorFlow Object Detection API (Huang et al. 2017) and PyTorch/Torchvision (Paszke et al. 2019) were used.

This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

The authors acknowledge the Minnesota Supercomputing Institute (MSI, https://www.msi.umn.edu/) at The University of Minnesota for providing high-performance computing (HPC) resources that have contributed to the research results reported within this paper.

REFERENCES

Adamo A., Östlin G., Bastian N., Zackrisson E., Livermore R. C., Guaita L., 2013, ApJ, 766, 105
Adams D., Mehta V., Dickinson H., Scarlata C., Fortson L., Kruk S., Simmons B., Lintott C., 2022, ApJ, 931, 16
Aguado D. S. et al., 2019, ApJS, 240, 23
Aihara H. et al., 2018, PASJ, 70, S8
Aragon-Calvo M. A., 2019, MNRAS, 484, 5771
Astropy Collaboration, 2022, ApJ, 935, 167
Bertin E., Arnouts S., 1996, A&A, 117, 393
Bournaud F., Elmegreen B. G., Elmegreen D. M., 2007, ApJ, 670, 237
Bournaud F. et al., 2013, ApJ, 780, 57
Bradley L. et al., 2023, astropy/photutils: 1.7.0. Zenodo. Available at: https://doi.org/10.5281/zenodo.7804137

Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Burke C. J., Aleo P. D., Chen Y.-C., Liu X., Peterson J. R., Sembroski G. H., Lin J. Y.-Y., 2019, MNRAS, 490, 3952
Cava A., Schaerer D., Richard J., Pérez-González P. G., Dessauges-Zavadsky M., Mayer L., Tamburello V., 2017, Nature Astron., 2, 76
Chan M. C., Stott J. P., 2019, MNRAS, 490, 5770
Claeyssens A., Adamo A., Richard J., Mahler G., Messa M., Dessauges-Zavadsky M., 2023, MNRAS, 520, 2180
Conselice C. J., Yang C., Bluck A. F. L., 2009, MNRAS, 394, 1956
Cowie L. L., Hu E. M., Songaila A., 1995, AJ, 110, 1576
Dasiopoulou S., Mezaris V., Kompatsiaris I., Papastathis V.-K., Strintzis M., 2005, IEEE Trans. Circuits Syst. Video Technol., 15, 1210
Dey A. et al., 2019, AJ, 157, 168
Dickinson H. et al., 2022, MNRAS, 517, 5882
Elmegreen B. G., Elmegreen D. M., 2005, ApJ, 627, 632
Elmegreen D. M., Elmegreen B. G., Rubin D. S., Schaffer M. A., 2005, ApJ, 631, 85
Elmegreen D. M., Elmegreen B. G., Ravindranath S., Coe D. A., 2007, ApJ, 658, 763
Elmegreen D. M., Elmegreen B. G., Marcus M. T., Shahinyan K., Yau A., Petersen M., 2009, ApJ, 701, 306
Erhan D., Szegedy C., Toshev A., Anguelov D., 2014, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Scalable Object Detection Using Deep Neural Networks. IEEE, Columbus, USA, p. 2155
Ferreira L. et al., 2022, ApJ, 938, L2
Ferreira L. et al., 2023, ApJ, 955, 94
Fisher D. B. et al., 2014, ApJ, 790, L30
Fisher D. B. et al., 2016, MNRAS, 464, 491
Förster Schreiber N. M. et al., 2009, ApJ, 706, 1364
Förster Schreiber N. M., Shapley A. E., Erb D. K., Genzel R., Steidel C. C., Bouché N., Cresci G., Davies R., 2011, ApJ, 731, 65
Ginzburg O., Huertas-Company M., Dekel A., Mandelker N., Snyder G., Ceverino D., Primack J., 2021, MNRAS, 501, 730
Guo Y., Giavalisco M., Ferguson H. C., Cassata P., Koekemoer A. M., 2012, ApJ, 757, 120
Guo Y. et al., 2015, ApJ, 800, 39
Guo Y. et al., 2018, ApJ, 853, 108
Harris C. R. et al., 2020, Nature, 585, 357
He K., Zhang X., Ren S., Sun J., 2016, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Deep Residual Learning for Image Recognition. IEEE, CA, USA, p. 770
He K., Gkioxari G., Dollár P., Girshick R., 2017, preprint (arXiv:1703.06870)
Huang J. et al., 2017, Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. IEEE, CA, USA, p. 3296
Huertas-Company M., Lanusse F., 2023, PASA, 40, e001
Huertas-Company M. et al., 2020, MNRAS, 499, 814
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Jaccard P., 1912, New Phytologist, 11, 37
Kauffmann G. et al., 2003, MNRAS, 341, 33
Kingma D. P., Ba J., 2014, preprint (arXiv:1412.6980)
LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436
Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L., 2014, in Fleet D., Pajdla T., Schiele B., Tuytelaars T., eds, Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science. Springer, Berlin, p. 740
Livermore R. C. et al., 2012, MNRAS, 427, 688
Lupton R., Blanton M. R., Fekete G., Hogg D. W., O'Mullane W., Szalay A., Wherry N., 2004, PASP, 116, 133
McKinney W., 2010, Proc. Python in Science Conference (SciPy 2010). SciPy, Austin, TX
Mandelker N., Dekel A., Ceverino D., Tweed D., Moody C. E., Primack J., 2014, MNRAS, 443, 3675
Mandelker N., Dekel A., Ceverino D., DeGraf C., Guo Y., Primack J., 2016, MNRAS, 464, 635
Mehta V. et al., 2021, ApJ, 912, 49
Merz G., Liu Y., Burke C. J., Aleo P. D., Liu X., Carrasco Kind M., Kindratenko V., Liu Y., 2023, MNRAS, 526, 1122
Messa M., Adamo A., Östlin G., Melinder J., Hayes M., Bridge J. S., Cannon J., 2019, MNRAS, 487, 4238
Oke J. B., Gunn J. E., 1983, ApJ, 266, 713
Overzier R. A. et al., 2009, ApJ, 706, 203
Paszke A. et al., 2019, Advances in Neural Information Processing Systems, Vol. 32. Curran Associates, Inc., Vancouver, BC, p. 8026
Pavel M. I., Tan S. Y., Abdullah A., 2022, Appl. Sci., 12, 6831
Ren S., He K., Girshick R., Sun J., 2015, in Cortes C., Lawrence N., Lee D., Sugiyama M., Garnett R., eds, Advances in Neural Information Processing Systems, Vol. 28. Curran Associates, Inc., Barcelona, Spain
Robitaille T., 2019, APLpy v2.0: The Astronomical Plotting Library in Python (2.0). Zenodo. Available at: https://doi.org/10.5281/zenodo.2567476
Romeo A. B., Agertz O., 2014, MNRAS, 442, 1230
Russakovsky O. et al., 2015, Int. J. Comput. Vis., 115, 211
Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Shibuya T., Ouchi M., Kubo M., Harikane Y., 2016, ApJ, 821, 72
Stoughton C. et al., 2002, AJ, 123, 485
Szegedy C., Toshev A., Erhan D., 2013, in Burges C., Bottou L., Welling M., Ghahramani Z., Weinberger K., eds, Advances in Neural Information Processing Systems, Vol. 26. Curran Associates, Inc.
Tan M., Le Q. V., 2019, preprint (arXiv:1905.11946)
van den Bergh S., Abraham R. G., Ellis R. S., Tanvir N. R., Santiago B. X., Glazebrook K. G., 1996, AJ, 112, 359
Walmsley M. et al., 2023, J. Open Source Softw., 8, 5312
Waskom M. L., 2021, J. Open Source Softw., 6, 3021
Willett K. W. et al., 2013, MNRAS, 435, 2835
York D. G. et al., 2000, AJ, 120, 1579
Zanella A. et al., 2019, MNRAS, 489, 2792
Zavagno A. et al., 2023, A&A, 669, A120

APPENDIX A: CREATING THE GALAXY IMAGES

A1. SDSS galaxy images

The SDSS galaxy images were created as cutouts from the SDSS DR15 Legacy survey data. To ensure a comparable visual size of the target galaxies, we scaled the cutouts to six times the 90-percent *r*-band Petrosian radius while keeping a scale of 0.396 arcsec per pixel to match the native SDSS resolution. We then created RGB-composite images from the three single *g*-, *r*-, and *i*-band FITS, where the *g*, *r*, and *i* map to red, green, and blue channels using 'Lupton'-scaling (Lupton et al. 2004):

$$I'_x = \frac{1}{Q}\operatorname{asinh}\left[Q \cdot \frac{\left(\frac{I_x}{\beta_x} - m\right)}{\alpha}\right], \tag{A1}$$

where I_x is the input pixel intensity in band x and I'_x is the scaled pixel intensity. Table A1 lists the parameters used for the GZCS-cutouts.

In a final step, we resized the RGB-composite images to 400 × 400 pixels so that the visual sizes of each central galaxy is similar but with varying resolutions.

Table A1. Parameters Lupton-scaling for *g*, *r*, and *i* bands.

Parameter	Band *g*	Band *r*	Band *i*
Q	7	7	7
Stretch α	0.2	0.2	0.2
Minimum m	0	0	0
Channel scales β	0.7	1.17	1.818

The SDSS DR15 PhotoPrimary table (Aguado et al. 2019) lists a redshift range of $0.02 \leq z \leq 0.25$ with a median value of $z_{\rm median} = 0.05$ for the clumpy galaxies from GZCS. The *r*-band PSF-FWHM from SDSS for the individual observations varies from 0.60 arcsec $\leq {\rm PSF}_{{\rm FWHM},\,r} \leq 2.08$ arcsec with a median value of ${\rm PSF}_{\rm median} = 1.12$ arcsec. For these redshift ranges, cosmological angular size–redshift relations can be simplified giving a median size of objects seen at an angle of one PSF-FWHM of 1.14 kpc. This corresponds to the initial assumption that clumps with ∼1 kpc physical sizes are unresolved with SDSS. We note, however, that clumps marked by the volunteers in galaxies at the higher end of our sample redshift range are unresolved up to a size of ∼5 kpc and might not necessarily represent genuine GSFCs.

A2. HSC galaxy images

We obtained the HSC galaxy cutouts using the command-line tools to access the PDR3 Data Access Service from HSC SSP. We applied a field of view (FOV) of 60 arcsec which matches the median FOV from the GZCS cutouts. The *g*-, *r*-, and *i*-filter bands were used to create a RGB-composite image using an asinh stretch to the images:

$$I'_x = \frac{\mathrm{asinh}(e^{10}\, I_x)/\mathrm{asinh}(e^{10}) + 0.05}{0.72}, \tag{A2}$$

where I_x is the input pixel intensity in band x and I'_x is the scaled pixel intensity. Finally, the RGB-composites were resized to 400 × 400 pixels.

APPENDIX B: FINE-TUNING *ZOOBOT* FOR CLASSIFYING CLUMPY GALAXIES

We used the set of the GZCS imaging data after it has been reduced by the aggregation algorithm from Dickinson et al. (2022) to develop a classification CNN specifically for separating clumpy and non-clumpy galaxies (*Zoobot Clumps*). The 45 643 images were split into a training set (36 514 or 80 per cent), validation set (4564 or 10 per cent), and a test set (4565 or 10 per cent) with the same distribution of clumpy and non-clumpy galaxies as the full data set (Table B1). A galaxy containing at least one off-centre clump, either a 'normal' or 'odd' clump (see Section 3), after the consensus label aggregation from Dickinson et al. (2022), was given the class 'clumpy' and 'w/o clumps' otherwise.

During training, the model was presented not only with the original set of input galaxy images but also with variations of it. This image augmentation helps to improve the generalization ability of the model

Table B1. Train, validation, and test data sets for the *Zoobot Clumps* classification model.

Label	ID	Train	Validation	Test	Total
W/o clumps	0	20 503 (56.15 per cent)	2541 (55.67 per cent)	2612 (57.22 per cent)	25 656 (56.21 per cent)
Clumpy	1	16 011 (43.85 per cent)	2023 (44.33 per cent)	1953 (42.78 per cent)	19 987 (43.79 per cent)
Total	–	36 514 (100.00 per cent)	4564 (100.00 per cent)	4565 (100.00 per cent)	45 643 (100.00 per cent)

Table B2. Parameters used for training the *Zoobot Clumps* classification model.

	ResNet50	EfficientNetB0
Python framework	PyTorch	PyTorch, Tensorflow
Infrastructure	NVIDIA A100 SXM4-40GB GPU	NVIDIA A100 SXM4-40GB GPU
Batch size	32	32
Optimiser	Adam	Adam
Initial learning rate	10^{-4}	10^{-4}
Epochs	100	100
Accuracy	0.8291 ± 0.0109 (at epoch = 62)	0.8484 ± 0.0104 (at epoch = 74)

and increases the subset of the learning data. The variations were mostly randomly applied and consisted of the following techniques:

(1) Random resizing, keeping the aspect ratio between 0.9 and 1.1,
(2) Random cropping to 224 × 224 pixels within a 10 per cent margin of the original image,
(3) Random horizontal flip,
(4) Random rotation of 90°, and
(5) Normalization of the pixel values in each channel.

We train two *Zoobot Clumps* versions, based on the ResNet50 (He et al. 2016) and EfficientNetB0 (Tan & Le 2019) architecture, over 100 epochs without extensive hyperparameter tuning. Learning rate and optimiser settings were varied around values gained from expert knowledge (Dickinson and Walmsley, private communication), but resulting model performance did not change significantly. We chose the *Adam* optimiser (Kingma & Ba 2014) over the standard stochastic gradient descent optimiser as it introduces an adaptive learning rate and increases computation speed. We show the parameters used for developing *Zoobot Clumps* in Table B2.

The models with the best classification performance achieved an accuracy of 0.8291 ± 0.0109 and 0.8484 ± 0.0104 (within a 2σ confidence interval) based on the ResNet50 and EfficientNetB0 architecture, respectively.

APPENDIX C: TRAINING RUN DETAILS

For the several training runs, we randomly assigned the remaining 18 772 galaxy images, after we have applied all exclusions (see Table 1), into 20 run-groups where the group size increased exponentially from 500 to 18 772. Each run-group was further split into a training (70 per cent), validation (20 per cent), and test set (10 per cent). The split was also done randomly but using a stratification based on the ratio of odd (or unusual) to normal clumps in each galaxy to maintain a comparable distribution of both clump classes in all groups. For the whole data set the ratio (odd clumps/normal clumps) = 0.293 ± 0.004 and the ratio of (clumps/galaxy) = 2.12 ± 0.009. Table C1 lists the sample sizes of the various run-groups and sets.

Note, that only galaxies with at least one off-centre clump were used for training. True negative samples are obtained from areas in the image where no instances of the objects to be detected are located. Additional images containing no instances of the objects are usually not required as the number of negative and positive anchor boxes need to be balanced for the RPN as otherwise a bias towards negative samples will occur (Ren et al. 2015).

Table C1. Number of galaxy images and annotated clumps (in brackets) for each set per run-group.

Run group	Training	Validation	Test	Total
1	349 (752)	100 (223)	51 (104)	500 (1079)
2	424 (893)	120 (244)	61 (128)	605 (1265)
3	512 (1056)	146 (292)	74 (151)	732 (1499)
4	620 (1298)	177 (366)	89 (194)	886 (1858)
5	751 (1671)	214 (456)	108 (224)	1073 (2351)
6	908 (1923)	260 (557)	130 (257)	1298 (2737)
7	1,099 (2361)	314 (685)	158 (316)	1571 (3362)
8	1330 (2788)	380 (822)	191 (429)	1901 (4039)
9	1610 (3426)	460 (953)	231 (482)	2301 (4861)
10	1949 (4145)	557 (1190)	279 (589)	2785 (5924)
11	2358 (5027)	674 (1426)	338 (690)	3370 (7143)
12	2855 (6082)	816 (1733)	408 (849)	4079 (8664)
13	3455 (7323)	987 (2076)	494 (1095)	4936 (10 494)
14	4181 (8808)	1195 (2526)	598 (1289)	5974 (12 623)
15	5061 (10 712)	1443 (3025)	724 (1510)	7230 (15 247)
16	6124 (12 782)	1750 (3730)	876 (1759)	8750 (18 271)
17	7412 (15 851)	2118 (4429)	1060 (2272)	10 590 (22 552)
18	8971 (18 986)	2563 (5415)	1282 (2680)	12 816 (27 081)
19	10 857 (22 959)	3102 (6633)	1552 (3259)	15 511 (32 851)
20	13 140 (27 825)	3754 (7941)	1878 (3979)	18 772 (39 745)

APPENDIX D: COCO METRICS

For object detection models performance is typically evaluated by AP metrics. A common set of metrics is used for the COCO Object Detection Challenge (COCO metrics, Lin et al. 2014).

Given the prediction score (objectness) c_i as the probability whether an anchor box i contains an object or not, a threshold value $c_0 \in [0, 1]$ is set so that if $c_i \geq c_0$, box i is defined to contain an object and not, if $c_i < c_0$. We calculated the IoU (see Section 4.3) for each bounding box containing objects with respect to the ground-truth, in this case the annotations from the GZCS-volunteers. If the IoU is $\geq$0.5, e.g. the detection is a TP, otherwise a False Positive (FP). On the other hand, if a bounding box has high enough overlap with a ground-truth object (IoU is $\geq$0.5) but the objectness score is below the threshold, so $c_i < c_0$, then this is called a False Negative (FN). The IoU-threshold can be varied depending on the specific task.

Precision p can then be defined as

$$p(c_0) = \frac{\#\mathrm{TP}(c_0)}{\#\mathrm{TP}(c_0) + \#\mathrm{FP}(c_0)} \tag{D1}$$

and recall r as

$$r(c_0) = \frac{\#\mathrm{TP}(c_0)}{\#\mathrm{TP}(c_0) + \#\mathrm{FN}(c_0)}, \tag{D2}$$

where $\#\mathrm{TP}(c_0)$, $\#\mathrm{FP}(c_0)$, and $\#\mathrm{FN}(c_0)$ are the number of TPs, FPs, and FNs, respectively, depending on the prediction score threshold c_0.

Precision and recall are equivalent to purity and completeness, respectively, which are more commonly used in an astrophysical context. Completeness is the number of objects in a data set that are detected over the number that exists. Purity is the number of true detections over the number of all detections. High values for both metrics show that a detector is returning accurate results ('high precision') as well as returning a majority of all true results ('high completeness').

We observed a trade-off between precision and recall which is typically for object detection problems. High precision can be achieved with low-score thresholds c_0 which results in a lower recall and *vice versa* (see Fig. 8 for an example where purity and completeness are used instead of precision and recall). Plotting precision against recall for a discrete set of score thresholds, e.g. $c_n \in [0.0, 0.1, \ldots, 1.0]$, the area under this curve is used as a model comparison metric, called AP:

$$\mathrm{AP} = \sum_n (r_n - r_{n-1}) p_n, \tag{D3}$$

where r_n and p_n are the corresponding values for recall and precision specific to the score threshold c_n. In other words, the AP summarizes such a r–p plot as the weighted mean of precision achieved at each threshold, using the step increase in recall from the previous threshold as the weight. A value close to 1 represents both high recall and high precision.

For multiclass detection problems the mean average precision ($\overline{\mathrm{AP}}$) is defined by

$$\overline{\mathrm{AP}} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{AP}_i, \tag{D4}$$

for the $k > 1$ classes.

Instead of iterating through discrete values of the detection score c_n, different thresholds for the IoU are used for calculating the AR. From the recall–IoU curve, where IoU $\in$ [0.5, 1.0], the area under the curve, multiplied by two, is used as the value for AR:

$$\mathrm{AR} = 2 \int_{0.5}^{1.0} r(\mathrm{IoU})\, \mathrm{d}(\mathrm{IoU}). \tag{D5}$$

This is then averaged over all classes k to define the mean average recall ($\overline{\mathrm{AR}}$):

$$\overline{\mathrm{AR}} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{AR}_i. \tag{D6}$$

$\overline{\mathrm{AP}}$ and $\overline{\mathrm{AR}}$ can be combined into a single performance metric. This metric is called $F1$-score f_1 and is defined as

$$f_1 = 2 \times \frac{\overline{\mathrm{AP}}\,\overline{\mathrm{AR}}}{\overline{\mathrm{AP}} + \overline{\mathrm{AR}}}. \tag{D7}$$

The F1-score can be used to compare different models, especially if they vary in precision and recall and if both metrics are equally important for evaluating detection performance.

APPENDIX E: CLUMP PHOTOMETRY

The flux of each clump was measured from each of the *ugriz*-bands FITS using the PHOTUTILS Python package (Bradley et al. 2023). Fig. E1 shows the distribution of the sizes of the bounding boxes described by the radius of a circle from the centre of the box to the closest side of each box. Allowing for some margin, we chose an aperture with a radius of 1.125 × the band-specific PSF-FWHM centred on the clump midpoint. We acknowledge, that this fixed aperture is not suitable for the wider range of bounding box sizes output by the model *Zoobot-clumps-backbone*. As this model tends to produce bigger bounding boxes for similar object detections (see Section 5.1, Figs G1, G2, and G3 in the Appendix), we kept the same aperture size for all models for better comparison.

Annuli spanning three to five PSF-FWHM where used to compute the median background flux around the aperture. We multiplied this per-pixel value with the aperture area and then subtracted the result from the flux measured in the clump aperture.

Here, background refers to the diffuse light of the host galaxy in which the clumps are embedded. There are several factors that can impact the background flux estimate for each clump. For clumps close to the rims of the galaxy the background estimate will be affected by the area outside the galaxy extent and not only resulting from the diffuse light of the host galaxy. Also, adjacent clumps might fall into the area of the annulus and will obscure the background estimate. To mitigate these effects, we took photometry measurements after masking the area outside, and all other identified clumps within, the host galaxy (Fig. E2).

We also corrected the background-subtracted fluxes for the flux loss due to the small aperture sizes as it was assumed that the clumps are unresolved at this scale. Using a Gaussian profile for the PSFs, the fluxes were multiplied by a factor of ∼1.03 for the aperture correction.

Flux values as observed by SDSS are reported in nanomaggies (Stoughton et al. 2002) and were converted into Jansky using the factor 3.631×10^{-6} Jy per nMgy. Also, we converted the flux values f into AB magnitudes $m_{\rm AB}$ (Oke & Gunn 1983) using

$$m_{\rm AB} = 22.5 - 2.5 \log_{10}(f). \quad \text{(E1)}$$

Further corrections to the AB magnitudes were applied for galactic extinction. For each clump location the reddening $E(B - V)$ is retrieved from the Schlegel, Finkbeiner & Davis (1998) dust maps and converted into an extinction A_λ applied to each *ugriz* AB magnitude using the tabulated factors from Schlafly & Finkbeiner (2011). We determined the clump colours $(u - g)$, $(g - r)$, $(r - i)$, and $(i - z)$ as differences between the background-subtracted clump magnitudes.

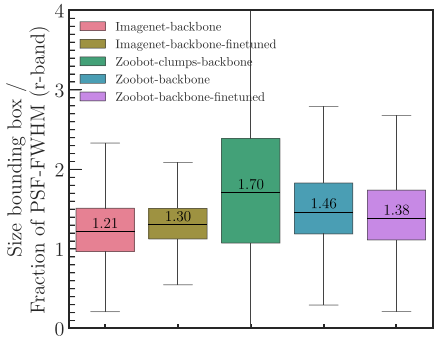
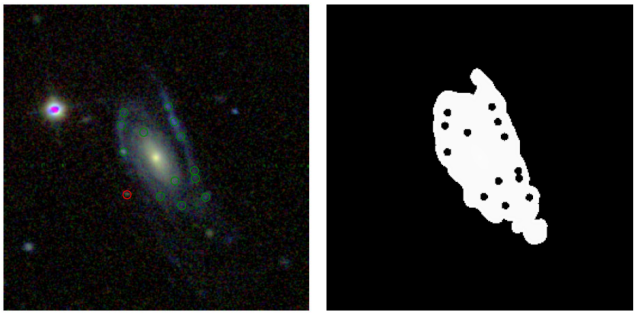


Figure E1. Distribution of bounding box sizes for all five models as fractions of the *r*-band PSF-FWHM. The size of a bounding box is described by the radius of a circle from the centre of the box to the closest side of the box. The annotations show the median values of the sizes.



(a) RGB image with clumps inside (green) and outside (red) the galaxy's spatial extent marked.

(b) *r*-band FITS with background and clump mask.

Figure E2. For clump photometry measurements the host galaxy background and the other clumps are masked for correct background subtraction.

APPENDIX F: GALAXY IMAGES WITH SIMULATED CLUMPS

Besides the real observational data, galaxy images containing simulated clumps were made available from Adams et al. (2022). This data set consists of 84 565 clumps placed in 26 736 galaxies with comparable characteristics with the main GZCS sample. The simulated clumps were placed randomly within the central galaxy's spatial extent but with a higher probability for inner-galactic area ($p = 0.75$) compared with the outer, low surface brightness area ($p = 0.25$).

Luminosity, mass, and spectral properties of the artificial clumps were carefully modelled to span a wide property range allowing for a robust completeness and purity assessment of the object detection models. The detailed process for generating these images with simulated clumps can be found in Adams et al. (2022).

We evaluated our models on the set of 26 736 galaxies with simulated clumps and measured lower overall completeness and purity compared with the levels reached by the models after we applied them to the real clumps. The simulated clumpy galaxies created by Adams et al. (2022) contain many more faint clumps

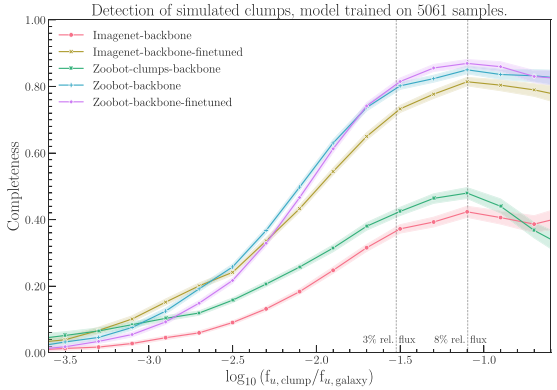


Figure F1. Model completeness per relative clump flux for training run 15 with a training sample size of 5061 at a score threshold of ≥0.3 (simulated clumps). Shaded areas showing the 95 per cent confidence interval.

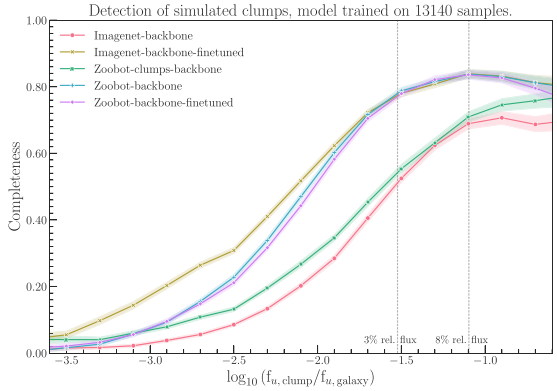


Figure F2. Model completeness per relative clump flux for training run 20 at a score threshold of ≥0.3 (simulated clumps). Shaded areas showing the 95 per cent confidence interval.

which were used for probing the volunteers' recovery capability and completeness during the GZCS project. As faint clumps have rarely been labelled by the volunteers in the data sets with real observations used to train the FRCNN models, we expected detection performance to be lower for faint clumps and, hence, in overall detection completeness. This can be seen from Figs F1 and F2, where completeness drops for a clump-galaxy flux ratio below 3 per cent for all models.

With focus on the clump-specific flux ratio ranges the ranking of the models in terms of completeness performance is similar to what we observed from the detections on the real imaging data. Both FRCNN models using a version of the unmodified *Zoobot* as their feature extraction backbone are reaching the highest completeness levels regardless of how many samples were used for training. *Imagenet-backbone-finetuned* is capable of reaching similar completeness levels but only after intensive training. The completeness curve for this model is rising after being trained on 5061 samples (Fig. F1) to 13 140 samples in the training set (Fig. F2). Again, *Zoobot-clumps-backbone* and *Imagenet-backbone* are much worse compared with the other models.

APPENDIX G: ADDITIONAL VISUAL EXAMPLES OF DETECTED CLUMPS IN SDSS AND HSC GALAXY IMAGES

The following images show nine different examples of galaxies annotated by the volunteers from GZCS, where we compare the detections from all five FRCNN models with the volunteers' markings. All models were trained on the full training set (13 140 galaxies, Table 1) and for all detections we applied a detection score threshold of ≥0.3.

Similar to Fig. 13, we compare the clump candidates for SDSS galaxies in Figs G1, G2, and G3. The galaxy images were chosen to cover many of the challenges the models can face during the object detection inference process, e.g. prominent foreground stars and image artefacts (e.g. Fig. G3).

In addition, Figs G4, G5, and G6 compare the clump markings and detections on the same nine galaxies with detections made on HSC galaxies, for which we found a cross-match with our GZCS sample. We compare only the detection results from the models *Imagenet-backbone* and *Zoobot-backbone* for clarity and to highlight the differences between a terrestrial and a domain-specific/astrophysical feature extraction backbone.

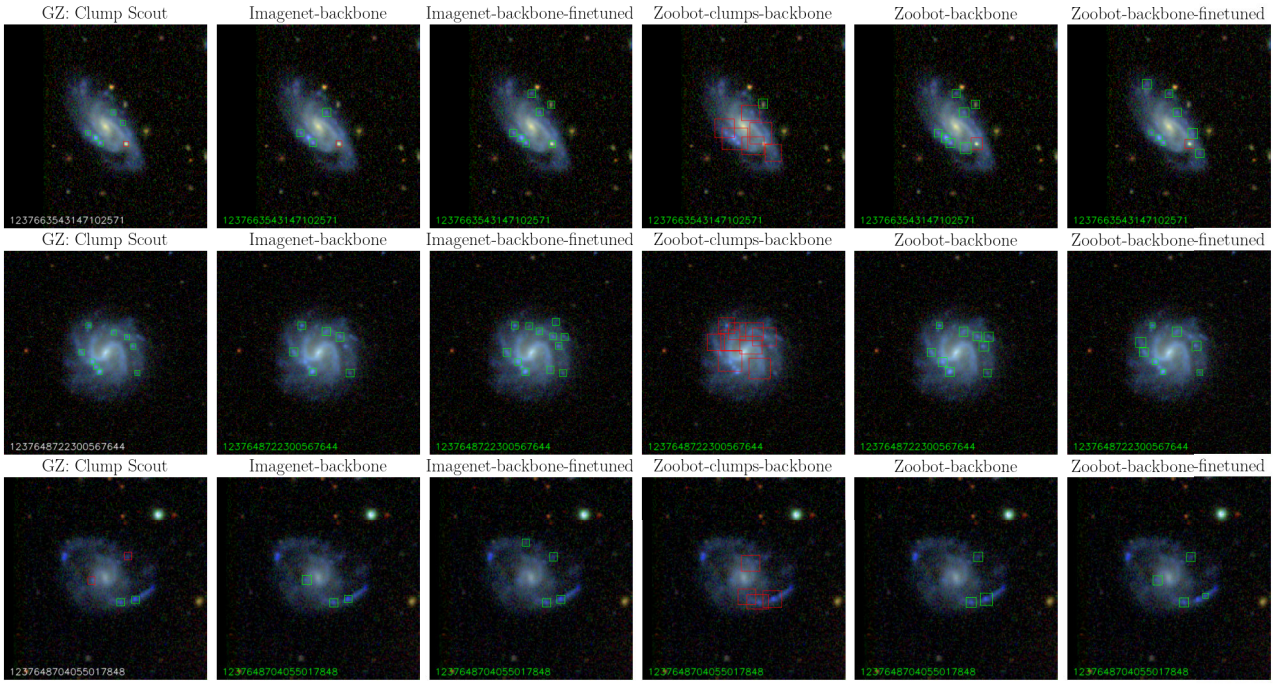


Figure G1. Comparison of detections on SDSS images from all models. Column 1: SDSS image with volunteers' labels from GZCS, column 2: with object detections by *Imagenet-backbone*, column 3: with object detections by *Imagenet-backbone-finetuned*, column 4: with detections by *Zoobot-clumps-backbone*, column 5: with detections by *Zoobot-backbone*, and column 6: with object detections by *Zoobot-backbone-finetuned*. Normal clumps are marked with green boxes, odd or unusual clumps with red boxes. The images are labelled with their SDSS-DR7 object numbers. Detection score threshold is ≥0.3.

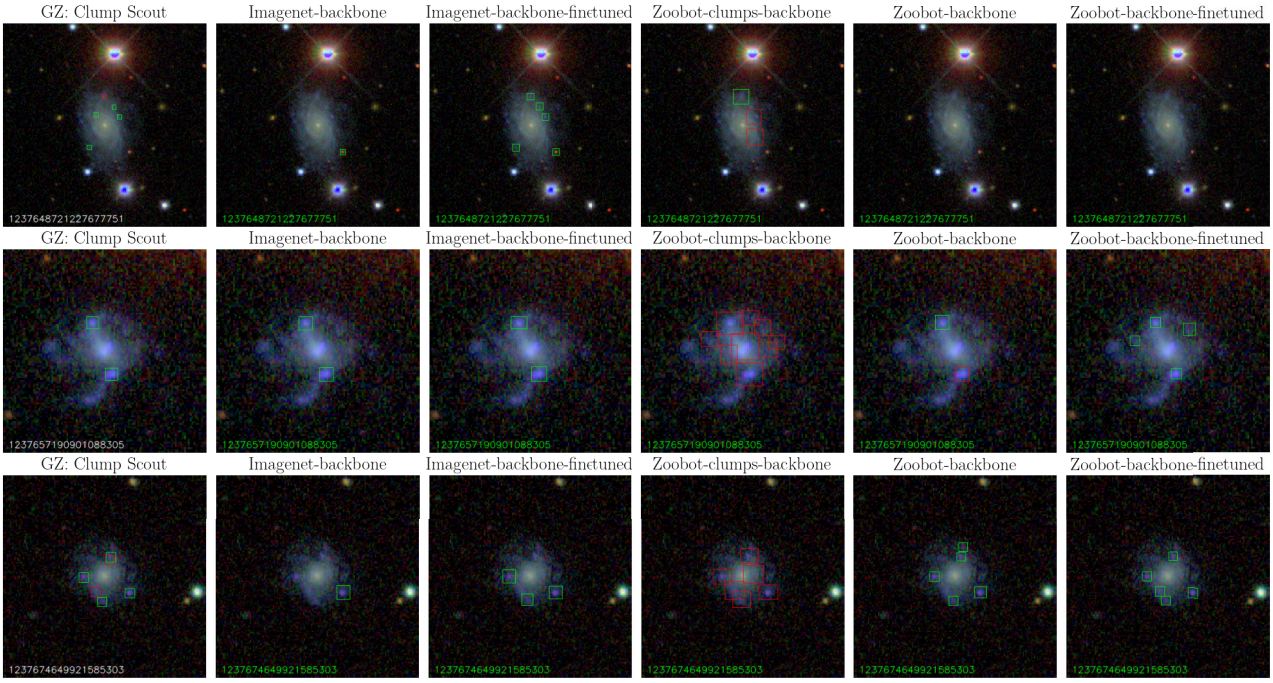


Figure G2. Comparison of detections on SDSS images from all models. Images and detections as described in the previous figure.

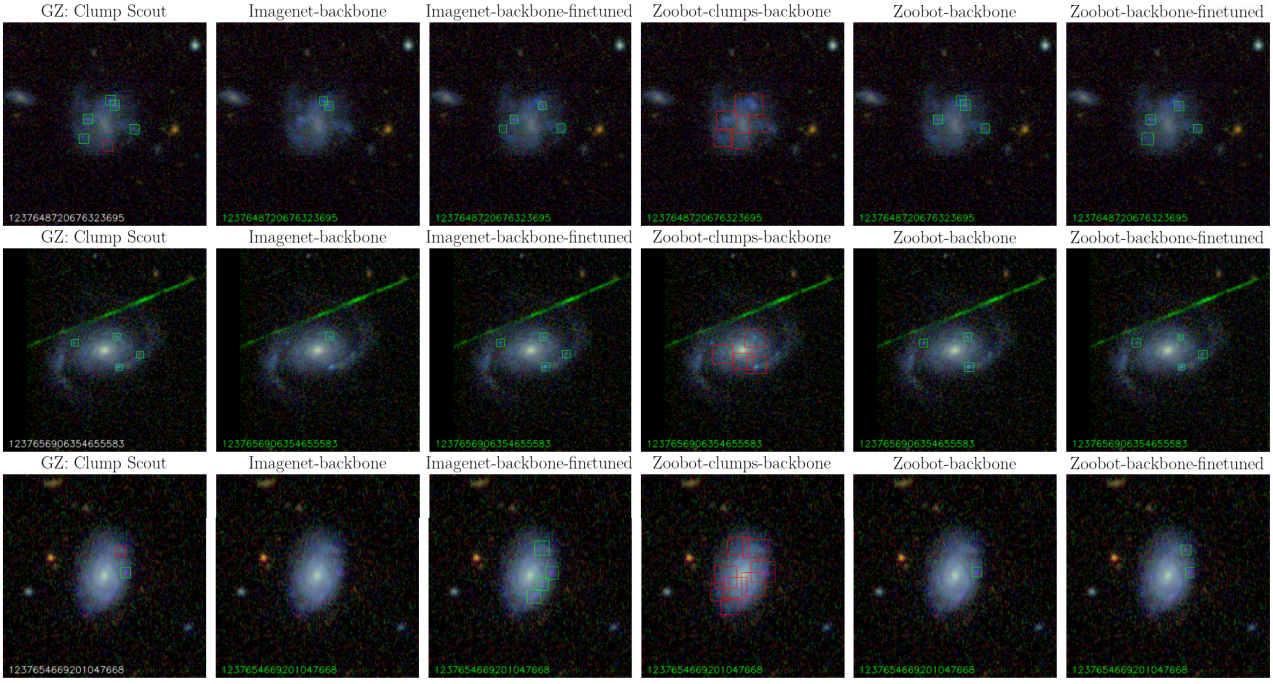


Figure G3. Comparison of detections on SDSS images from all models. Images and detections as described in the previous figure.

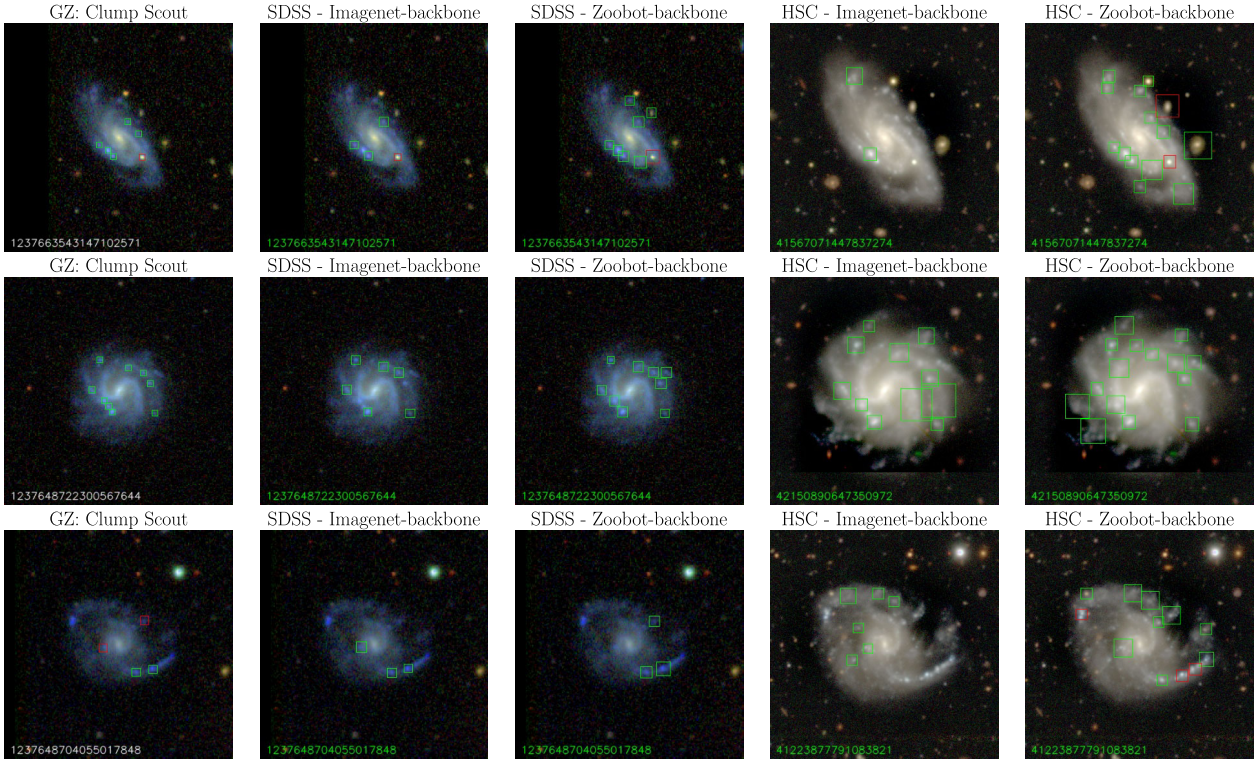


Figure G4. Comparison of detections on SDSS and HSC images. Column 1: SDSS image with volunteers' labels from GZCS, column 2: SDSS image with object detections by *Imagenet-backbone*, column 3: SDSS image with detections by *Zoobot-backbone*, column 4: HSC image with object detections by *Imagenet-backbone*, and column 5: HSC image with detections by *Zoobot-backbone*. Normal clumps are marked with green boxes, odd or unusual clumps with red boxes. The images are labelled with their SDSS-DR7 and HSC object numbers, respectively. Detection score threshold is $\geq$0.3.

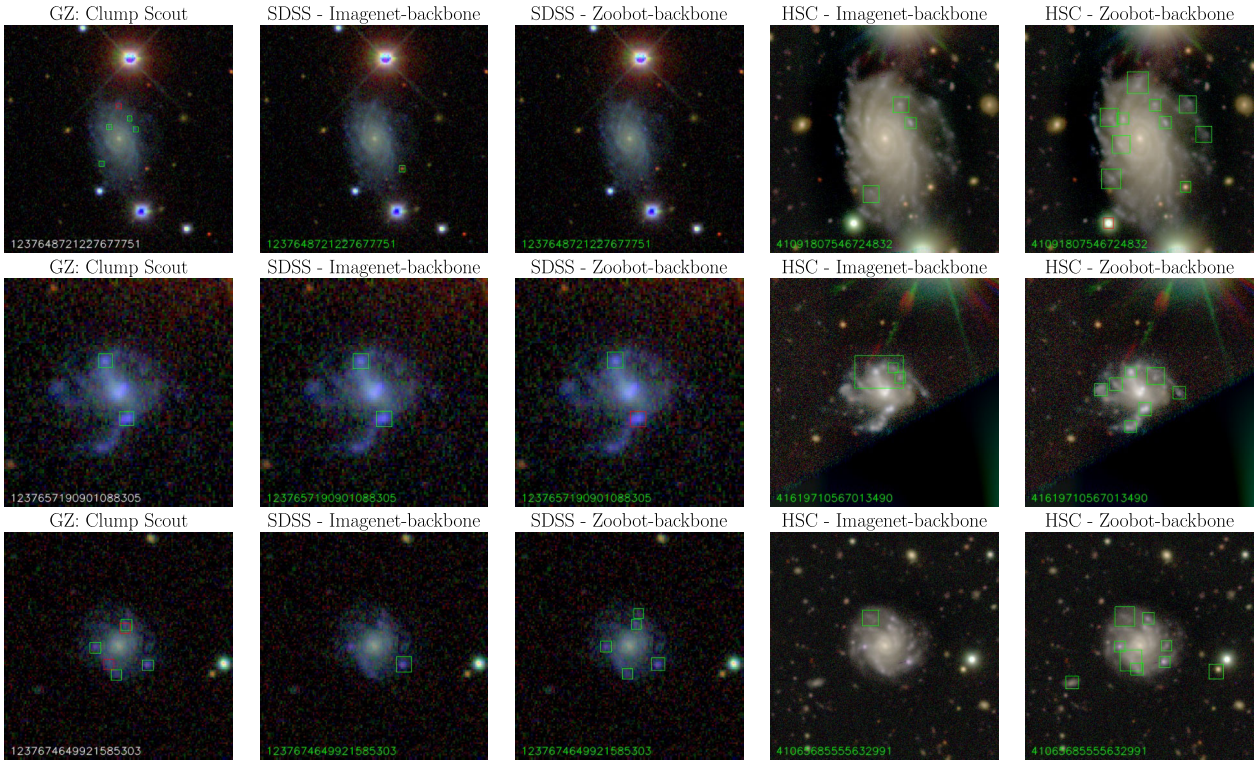


Figure G5. Comparison of detections on SDSS and HSC images. Images and detections as described in the previous figure.

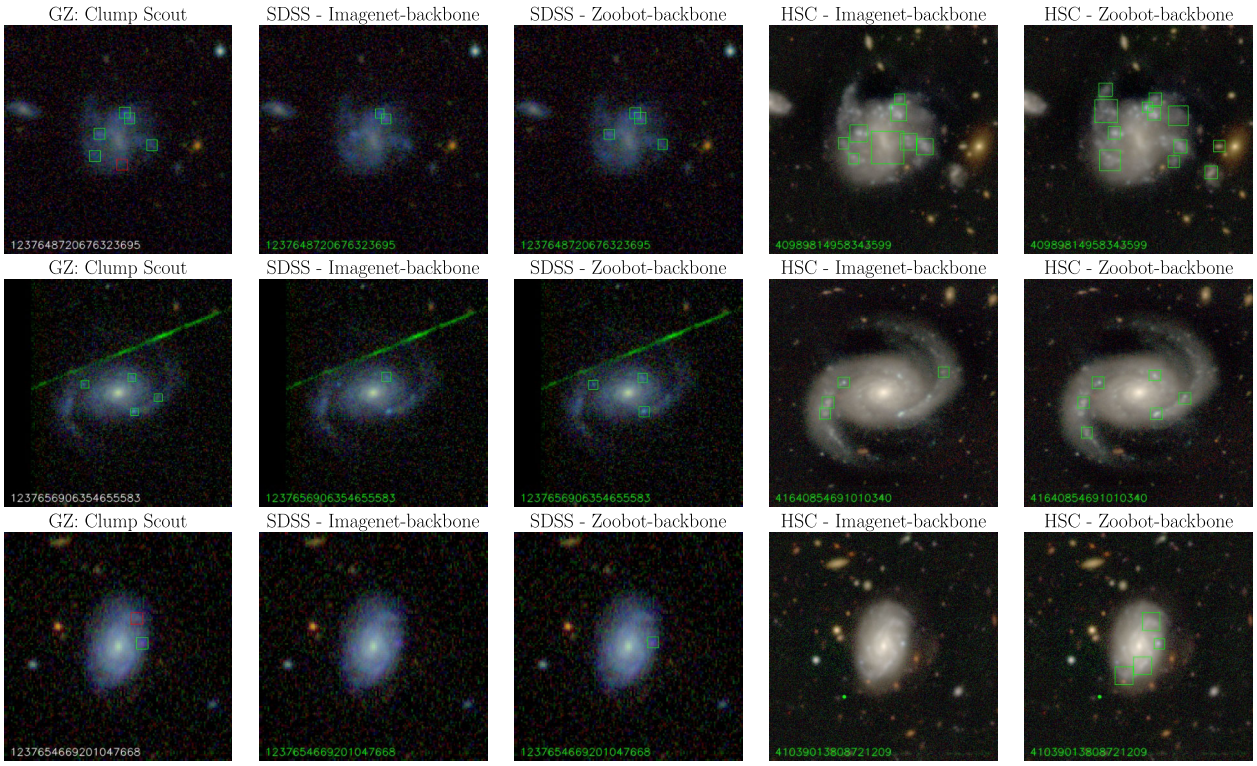


Figure G6. Comparison of detections on SDSS and HSC images. Images and detections as described in the previous figure.

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