THE UNIVERSITY OF CHICAGO

SHEDDING LIGHT ON THE LOW-SURFACE-BRIGHTNESS UNIVERSE WITH GALAXY SURVEYS AND MACHINE LEARNING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ASTRONOMY AND ASTROPHYSICS

BY
DIMITRIOS TANOGLIDIS

CHICAGO, ILLINOIS
AUGUST 2022

To my teachers and mentors

The young poet Evmenes

complained to Theokritos one day;

"I've been writing for two years now

and only one idyll I've composed.

This is my only accomplished work.

Alas, it's tall, this I can see,

it's very tall Poetry's ladder;

it pains me that from here, where I stand

on this first step, I won't climb any higher".

"Words like that", Theokritos replied,

"are inappropriate and blasphemous.

And if you stand on this first step

you must be proud and happy.

Arriving this far is not a little thing.

So much you've done, a great glory.

Even this step, the first one,

is quite distanced from the ordinary world.

In order to be standing on this step

you must by right be a citizen in the ideas' city.

It's hard and rare to be accepted in that city.

You will find there Legislators

that no charlatan can fool.

Arriving this far is not a little thing,

So much you've done, a great glory".

C.P. Cavafy

TABLE OF CONTENTS

LIST OF FIGURES . . . viii

LIST OF TABLES . . . xiv

ACKNOWLEDGMENTS . . . xv

ABSTRACT . . . xvii

1 INTRODUCTION . . . 1
- 1.1 Prologue: A veil waiting to be lifted . . . 1
  - 1.1.1 On the shores of Crete . . . 1
  - 1.1.2 A hidden Universe . . . 2
- 1.2 Galaxies in the Universe . . . 3
- 1.3 Low-Surface-Brightness Galaxies . . . 5
  - 1.3.1 Significance of LSBG studies . . . 5
  - 1.3.2 History and recent observations . . . 7
- 1.4 Galaxy Surveys . . . 9
  - 1.4.1 The Dark Energy Survey (DES) . . . 11
  - 1.4.2 The Legacy Survey of Space and Time (LSST) . . . 12
- 1.5 Thesis Outline . . . 12

2 MACHINE LEARNING IN ASTROPHYSICS . . . 14
- 2.1 Introduction . . . 14
- 2.2 Machine Learning Taxonomy . . . 15
- 2.3 ML Algorithms used in this work . . . 16
  - 2.3.1 Support Vector Machines . . . 17
  - 2.3.2 Random Forests . . . 17
  - 2.3.3 Artificial Neural Networks (ANNs) . . . 18
  - 2.3.4 Convolutional Neural Networks (CNNs) . . . 20
- 2.4 Machine Learning workflow . . . 21

3 SHADOWS IN THE DARK: LOW-SURFACE-BRIGHTNESS GALAXIES DISCOVERED IN THE DARK ENERGY SURVEY . . . 24
- 3.1 Introduction . . . 24
- 3.2 DES Data . . . 27
- 3.3 LSBG Catalog . . . 29
  - 3.3.1 Initial sample selection . . . 29
  - 3.3.2 Machine Learning Classification . . . 30
  - 3.3.3 Visual Inspection . . . 34
  - 3.3.4 Sérsic Model Fitting . . . 36
  - 3.3.5 Extinction Correction and Final Cuts . . . 38
- 3.4 Detection Efficiency around the Fornax Cluster . . . 39

3.5 LSBG Properties 42
3.6 Clustering of LSBGs 47
3.6.1 Comparison to other galaxy samples 51
3.6.2 Cross-correlation between galaxy samples 54
3.7 Associations with Galaxy Clusters and Groups 55
3.7.1 Radial Profiles 61
3.7.2 Size–Luminosity Relation 62
3.8 Summary and Conclusions 65
3.9 Surface-Brightness Limits 68
3.10 Selection Criteria 68
3.11 Magnitude Distributions 70

4 DEEPSHADOWS: SEPARATING LOW SURFACE BRIGHTNESS GALAXIES FROM ARTIFACTS USING DEEP LEARNING 73
4.1 Introduction 73
4.2 Data 77
4.2.1 The Dark Energy Survey 77
4.2.2 LSBGs and artifacts in DES 77
4.2.3 Generation of datasets 79
4.2.4 HSC SSP dataset 81
4.3 Methods 82
4.3.1 Machine Learning 82
4.3.2 Deep Learning 83
4.4 Classification Results 86
4.4.1 Classification Metrics 86
4.4.2 Results 87
4.4.3 Interpretation of Results 91
4.5 Transfer Learning 94
4.6 Uncertainty Quantification 97
4.6.1 Bootstrap resampling of the test set 99
4.6.2 DES dataset assignment 100
4.6.3 Impact of mislabeling 101
4.7 Discussion and Conclusions 103
4.8 Three-class classification 106
4.9 Grad-CAM details 109

5 DEEPGHOSTBUSTERS: USING MASK R-CNN TO DETECT AND MASK GHOSTING AND SCATTERED-LIGHT ARTIFACTS FROM OPTICAL SURVEY IMAGES 111
5.1 Introduction 111
5.2 Data 114
5.2.1 Dark Energy Survey Data 114
5.2.2 Annotation process 116
5.3 Methods 120

5.4 Results 124
5.4.1 CCD-based metrics 126
5.4.2 Comparison with the Ray-Tracing algorithm 130
5.4.3 Standard object detection evaluation metrics 132
5.4.4 Using Mask R-CNN to classify ghost-containing vs. ghost-free images 134
5.5 Summary and Conclusions 137
5.6 Human annotator agreement 139
5.7 Training History 141
5.8 Masking CCDs 141
5.9 False Positive and False Negative examples 143

6 INFERRING STRUCTURAL PARAMETERS OF LOW-SURFACE-BRIGHTNESS GALAXIES WITH UNCERTAINTY QUANTIFICATION USING BAYESIAN NEURAL NETWORKS 147
6.1 Introduction 147
6.2 Simulated Data 149
6.3 Bayesian Neural Networks 151
6.3.1 Variational Inference 152
6.3.2 BNN training 153
6.4 Implementation and training 156
6.5 Results 157
6.5.1 Parameter posteriors 157
6.5.2 Calibration 158
6.5.3 Comparison of point estimates 160
6.6 Discussion and Future Work 162
6.7 Bayes by Backprop details 163

7 CONCLUSIONS AND SUMMARY 165

REFERENCES 167

LIST OF FIGURES

1.1 DES footprint (observing area) in relation to the Milky Way plane and the large and small Magellanic clouds. Credit: J. Prat. 10

3.1 The distribution of the objects visually classified as LSBGs in the seven 4° × 4° regions used to create the labeled set for classification and validation. The Fornax galaxy cluster is located at (RA, DEC) $\sim (55^\circ, -35^\circ)$. 31

3.2 The confusion matrix of our final SVM classifier evaluated on the validation set. The quoted numbers correspond to the number of the validation instances (objects) based on their true and predicted label. The false-negative rate is $\sim 9\%$. 33

3.3 $30'' \times 30''$ cutouts of 20 candidates, positively classified by our machine learning algorithm (Section 3.3.2). Candidates 2, 3, 8, 11, 12, 13, 14, 15, and 18 are visually classified as LSBGs, while the other candidates are rejected as false positives and/or duplicates. 35

3.4 The dwarf galaxies present in the NGFS catalog (in blue) and the matches from our DES LSBG catalog (red). The NGFS catalog is separated into nucleated (denoted by an 'X') and non-nucleated (circles) galaxies. We plot DES LSBGs that were not matched to NGFS objects in light red (these are generally located outside the NGFS area). The black cross denotes the nominal center of the Fornax cluster. 39

3.5 Color–color diagram of our LSBG sample, using (a) `SourceExtractor` `MAG_AUTO` parameters and (b) magnitudes derived by fitting with `galfitm`. In both cases, we observe a bimodality in the $g - i$ and $g - r$ color distributions. We separate the total sample into red and blue galaxies, based on their $g - i$ color value: we fit the $g - i$ distribution with a Gaussian mixture model with two Gaussians (gray dashed lines in the top panels) and find the intersection point. This is at $g - i = 0.66$ and $g - i = 0.60$ for the `SourceExtractor` and `galfitm` cases, respectively (black vertical dashed lines). We use the intersection point derived from the `galfitm` distribution to define red and blue LSBG samples. 43

3.6 Examples of (a) blue and (b) red LSBGs in our sample. We randomly selected red galaxies with $g - i$ above the median for the red population ($g - i > 0.76$) and blue galaxies below the median of the blue population ($g - i < 0.40$) to make the color difference more prominent. Each cutout is $30'' \times 30''$ in size. 45

3.7 (a) Joint distribution of the red and blue LSBGs in the space of effective radius, $R_{\rm eff}$, and mean surface brightness (within the effective radius), $\bar{\mu}_{\rm eff}$, both in the g band. The two populations are defined according to the $g-i$ color criterion described in Section 3.5. The dashed horizontal and vertical lines correspond to the limits of the selection criteria $r_{1/2} > 2.5''$ and $\bar{\mu}_{\rm eff}(g) > 24.2\,{\rm mag\,arcsec^{-2}}$, respectively. Note that although surface brightness is independent of distance, and thus the scatter shown here reflects the intrinsic properties of our sample, much of the scatter in the angular effective radius comes from the fact that the LSBGs lie at different distances. (b) Sérsic index, n, versus central surface brightness, $\mu_0(g)$ [e.g., Graham and Driver, 2005], for the galaxies in our red and blue subsamples. The black dashed line corresponds to our selection criterion, $\bar{\mu}_{\rm eff}(g) = 24.2\,{\rm mag\,arcsec^{-2}}$. 46

3.8 Sky positions of (a) blue LSBGs ($g-i < 0.60$; 7,671 galaxies) and (b) red LSBGs ($g-i \geq 0.60$; 16,119 galaxies) within the DES footprint. The distribution of the red LSBGs is more strongly clustered than that of the blue LSBGs. 48

3.9 The angular autocorrelation function of the total LSBG sample (dark gray line), and the red and blue LSBG subsamples (red and blue lines, accordingly). The errors were calculated using the jackknife method. The correlation function of the red LSBGs has a higher amplitude than that of the blue LSBGs across all angular scales. 51

3.10 The angular autocorrelation function of all LSBGs (gray line), the HSBG sample extracted from the DES data (blue line) and the 2MPZ sample (red line). We see that the LSBG exhibits a turnover at lower angular scales that is not observed either at the HSBG or 2MPZ samples. 52

3.11 (a) The cross-correlation function, $\xi(\theta)$, between (i) the DES LSBG and HSBG samples (orange line), (ii) the LSBG and 2MPZ samples (blue line), and (iii) the DES HSBG and 2MPZ samples (green line). (b) The square of the cross-correlation coefficient between the same samples as in panel (a), in order to cancel out the contribution of the different galaxy biases and compare the different cross-correlation levels. In both panels, the shaded regions correspond to the errors in the estimated cross-correlations. 54

3.12 KDE map of the distribution of our LSBG sample. Blue regions denote areas of low density, while regions of high density are indicated in yellow/red. Open red circles indicate the positions of the 82 prominent density peaks identified as described in Section 3.7. We have labeled the 10 most prominent peaks, which are summarized in Table 3.2. 57

3.13 Normalized radial profiles of the distribution of LSB galaxies (blue) and galaxies from the 2MPZ catalog (red) around the associations of the most prominent LSBG overdensity peaks, presented in Table 3.2. We have assumed that all galaxies that are within a radius that corresponds to a physical scale of 1.5 Mpc at the distance of the association belong to that association. The normalization constant corresponds to the mean number density of galaxies within the 1.5 Mpc radius. 60

3.14 Size–luminosity relation for LSBGs around the associations of the most prominent overdensity peaks, presented in Table 3.2. We have assumed that all LSBGs within an angle corresponding to a physical radius of 0.5 Mpc at the distance of the association belong to it. With the dashed horizontal lines, we show the physical scale corresponding to the radius cut $r_{1/2}(g) > 2.5''$ at the distance of the cluster. We also show (dashed, diagonal gray lines) the lines of constant mean surface-brightness. 63

3.15 Size–luminosity relation of LSBGs around the nine most prominent overdensities (red points) in the i band. The sample consists of 555 galaxies. For comparison, we overplot the dwarf galaxies found around Fornax in the NGFS survey [Eigenthaler et al., 2018, Ordenes-Briceño et al., 2018]. 41 galaxies in our sample have effective radii exceeding 1.5 kpc in the g band (black circles) and central surface brightness $\mu_0(g) > 24.0\,\mathrm{mag\,arcsec^{-2}}$, which is a conventional definition for ultra-diffuse galaxies [UDG; van Dokkum et al., 2015b]. 64

3.16 Surface-brightness limits at 3σ estimated from the surface-brightness contrast in $10'' \times 10''$ regions over the DES coadd tiles in the g band (top), r band (middle), and i band (bottom). 69

3.17 Normalized distribution of the g-, r-, and i-band magnitudes of our LSBG sample. 70

3.18 Joint distributions of the red and blue LSBGs in the space of g-band magnitude vs (a) effective radius, R_{eff}, and (b) Sérsic index, n, both in g-band. 71

3.19 g-band magnitude distributions of the LSBG sample and the DES catalog matches on the 2MPZ sample. 72

4.1 Example images of (a) LSBGs and (b) artifacts in our dataset. Each cutout corresponds to a $30'' \times 30''$ angular region on the sky. We remind the reader that our artifact class consists of any object that passes the low surface brightness selection criteria but is not an LSBG (see example categories in Sec. 4.1). 78

4.2 The g-band (a) magnitude and (b) surface brightness distributions of the LSBGs (red) and artifacts (blue) in our sample. 80

4.3 Examples of (a) LSBGs and (b) artifacts from the HSC SSP survey. As in Fig. 4.1, each cutout corresponds to a $30'' \times 30''$ angular region on the sky. 81

4.4 Schematic overview of the *DeepShadows* CNN architecture. There are three convolutional layers (yellow), each followed by a max pooling layer (red). The number of filters are 16, 32 and 64 for each layer, respectively. Two dense layers (purple) follow the last pooling layer after flattening. The output is a probability score that the image contains an LSBG. Figure was created using the PlotNeuralNet code [Iqbal, 2018]. 83

4.5 Training and validation accuracy/loss as a function of the training epoch for the *DeepShadows* model. Training was performed on 30,000 images and validation on 5,000. The model reaches a training accuracy of $\sim 92\%$ after 100 epochs. . . 88

4.6 (a) ROC curves for the *DeepShadows* CNN model (orange, solid line), the SVM model (blue dashed line), and the random forest model (red dashed-dotted line) evaluated on the test set. The diagonal dashed line corresponds to the performance of a random classifier. We also show the 95% confidence intervals on the vertical direction (true positive rate). These were estimated using the bootstrap method on the test set of images (see Sec. 4.6). (b) Confusion matrix of the *DeepShadows* model predictions on the test set. The values in parentheses correspond to the normalized version of the matrix, obtained by dividing the number of objects in each case by the total number of objects in each category (true label). 89

4.7 Balanced accuracy score of the *DeepShadows* predictions on the test set, for the objects in three different surface brightness bins. 91

4.8 Output scores from the *DeepShadows* CNN for the LSBGs and artifacts in the test set. 92

4.9 Examples of objects classified with high output score in their respective class and corresponding Grad-CAM visualization maps for the same objects. Clockwise: True Negatives, False Negatives, True Positives, False Positives (same arrangement as the confusion matrix). 93

4.10 Transfer learning results on the HSC SSP test set. (a) Confusion matrix (raw and normalized values) before fine tuning. (b) Confusion matrix (raw and normalized values) after fine tuning. (c) ROC curves and AUC scores, before (orange, dashed line) and after (red, solid line) fine tuning. The bands correspond to the 95% confidence level intervals. 95

4.11 Accuracy, evaluated on the test set, for the *DeepShadows* model trained on progressively larger datasets. 97

4.12 (a) ROC curves of the performance on the test set using the baseline split (orange line) between training-validation-test sets and using splits with different random seeds (blue lines). (a) ROC curves of the performance on the test set using different levels of label noise in the training set, starting from no noise (orange line) to 33% random mislabeling. 98

4.13 Ratio of the ROC curves with different label noise levels to the one without noise. More specifically, we plot the ratio of the true positive rates, as a function of the false negative rate. The horizontal, orange, dashed, line corresponds to ratio = 1. Notice that the range in the false positive rate plotted is 0–0.5 since for higher values all curves converge to a value equal to one. 101

4.14 Randomly selected examples of type 1 artifacts (see text). 107

4.15 (a) Confusion matrix for the three-class classification problem, where two artifact classes are considered. The numbers in parentheses correspond to the normalized values. (b) “reduced” confusion matrix, where the two artifact classes have been combined. 107

5.1 Examples of full-focal-plane DECam images containing ghosts and scattered-light artifacts. The corresponding "ground truth" masks (right) were manually annotated. There are three categories of ghosting artifacts: image (a) contains a scattered-light artifact classified as 'Rays'; image (b) shows the masks for the 'Rays' in red; image (c) contains both 'Bright' and 'Faint' ghosts, and the corresponding masks in blue and yellow, respectively, are shown in image (d). 116

5.2 Histograms of the distribution in size (area) of the three artifact types presented in this work. The areas are quoted as multiples of the area of a single CCD. . . 118

5.3 Masks created by the eight different annotators (overlaid on top of each other) for the same two images presented in Fig. 5.1. The colors indicate the number of annotators that have labeled a given pixel as containing a ghost, from dark purple (one annotator) to light yellow (all the eight annotators). 119

5.4 High-level schematic overview of the Mask R-CNN model. Figure adapted from Weng [2017]. 121

5.5 Total loss of the Mask R-CNN model as function of the training epoch. The training is performed using a progressively smaller learning rate, α. 122

5.6 Predicted masks on four example images that contain the three distinct artifact types — scattered-light 'Rays' (red), 'Bright' ghosts (blue), and 'Faint' ghosts (yellow). The top panels correspond to the images presented in Fig. 5.1. 125

5.7 CCD-based (a) precision (blue) and recall (orange), and (b) $F1$ score as a function of the CCD area threshold (see main text) from the Mask R-CNN model and for the three ghosting artifact categories ('Rays', 'Bright', and 'Faint'). 128

5.8 CCD-based (a) precision and (b) recall of the Mask R-CNN model (blue lines) and the Ray-Tracing algorithm (orange lines). We consider both the combination of all types of artifacts (solid lines) and the combination of 'Rays'+'Bright' (dashed lines). 128

5.9 CCD-based $F1$ scores for the same models and ghost type combinations as in Fig. 5.8. 129

5.10 Precision-Recall curves and Average Precision scores at different IoU threshold values in the range $0.50 - 0.90$. We show these metrics for the different ghost types in this work ('Rays'-'Bright'-'Faint'), and for all ghost types, combined. . 133

5.11 (a) Confusion matrix of predictions of the Mask R-CNN model on a dataset containing an even number of ghost containing and clean images. An image is predicted to 'have ghost' if even a single ghost is detected in that image by the Mask R-CNN model. (b) Confusion matrix of the predictions of the combined CNN + Mask R-CNN model (CNN model from Chang et al. [2021]). An image is said to 'have ghost' if and only if both the CNN and the Mask R-CNN models agree on that (otherwise the prediction is 'clean'). 135

5.12 Distribution of the number of votes each pixel in our dataset has received as containing a ghost, from the eight annotators. We include only pixels that have received at least one vote. We present the distributions for each ghost type separately (panels (a)-(c)) and without distinguishing between the different types (panel (d)). 140

5.13 Classification loss as a function of the training epoch. 142
5.14 Bounding box loss as a function of the training epoch. 142
5.15 Mask loss as a function of the training epoch. 143
5.16 CCDs masked as ghost-containing (in blue) when even a single pixel of the predicted ghost mask lies within the CCD (0% threshold, panels (a) and (c)), and when at least half of the CCD area has to be covered by the CCD (50% threshold, panels (b) and (d)). The yellow contours correspond to the mask predictions of the Mask R-CNN model (without distinguishing between the different types of artifacts). 144
5.17 (a) Example images classified as ghost-containing (false positives) and the corresponding predicted masks (lower panel, (b)). 145
5.18 Examples of false negatives, i.e. images that were classified as 'clean' by the Mask R-CNN model (no objects detected). In practice, the artifacts present in these images are very small and faint, and often go undetected by human annotators. 146

6.1 Examples of simulated LSBG images. The inset text refers to the effective radius, r_e, (top) and the surface brightness, I_e (bottom). 150
6.2 Schematic representation of the BNN architecture used in this work. 155
6.3 Training and validaton (a) loss, and (b) mean absolute error of the BNN model, as a function of the training epoch. 157
6.4 Predicted posterior distributions of structural parameters for a simulated LSBG at the bright end (I_e = 24.4 mag/arcsec^2, panel (a)), and one at the faint end (I_e = 25.3 mag/arcsec^2, panel (b)) of the surface brightness range we consider in this work. The dashed lines indicate the true input parameter values for the simulated. 158
6.5 Calibration curves of the marginalized BNN posteriors, for the five Sérsic model parameters considered in this work. 159
6.6 Comparison of the parameter predictions from the BNN model (blue dots) and from `pyImfit` (red dots). We also show the coefficients of determination. 161

LIST OF TABLES

3.1 Detection efficiency around the Fornax Cluster 39
3.2 Characteristics of the 10 most prominent density peaks and their associations . 56

4.1 Architecture of the *DeepShadows* CNN. 84
4.2 Comparison of the classification metrics for the three machine learning models presented in Sec. 4.4.2. 90
4.3 95% Confidence intervals on the classification metrics from Bootstrap resampling of the test set. 99
4.4 95% Confidence Intervals on the classification metrics from Bootstrap for the transfer learning task, after fine-tuning, evaluated on the HSC SSP test set. . . . 100
4.5 Classification metrics for the baseline split of training-validation-test sets and six alternative splits using different random seeds. Because of the small number of runs we present each case individually and not as an interval like in Table 4.3. . 100
4.6 Classification metrics using the baseline split and different levels of random label noise (mislabeling) in the training set. 102

5.1 CCD-based evaluation metrics (precision, recall, $F1$ score) for the Mask R-CNN and Ray-Tracing algorithms, at 0% CCD area threshold. 132

ACKNOWLEDGMENTS

Completing a PhD is not a trivial task and requires grit and determination. It is also a long journey; without the help of many people, the successful completion of this journey would not be possible –and certainly it would be less fun.

First, I would like to thank my advisor, Alex Drlica-Wagner, for believing in me and giving me the guidance and the freedom to pursue the projects I was interested in, in the intersection of astrophysics and machine learning. I also thank my thesis committee members, Chihway Chang, Andrey Kravtsov, and Brian Nord for their guidance and feedback. I was lucky to work with Chihway on one of my first grad school projects; have Andrey as a teacher and also be his TA, and be part of the DeepSkies community, co-led by Brian Nord. I'm also glad that, during my first couple of years in grad school, I had the opportunity to work with Scott Dodelson, Dan Hooper, and Joshua Frieman. I was lucky to interact with many excellent scientists: Ting Li, Burcin Mutlu-Pakdil, and Aleksandra Ćiprijanović were great collaborators at different stages of this thesis. I would like to especially thank Aleksandra (Alex) for being such a fantastic collaborator and supporter in my machine learning ventures!

My Chicago adventure would had never started without the mentorship of Vaso Pavlidou, back in Crete. Vaso, I will always be grateful for that! The late Theodore Tomaras, also a professor in Crete, was an excellent mentor, too, and helped in very hard times. I wish I could express my gratitude to him when that was possible. I am also indebted to the people of my small village, Vafes, who –in times of extreme financial hardship– supported me and helped to get my college degree. I thank them all!

Beyond growing as a scientist, my University of Chicago years, were also an opportunity of personal growth. I have to admit that I would had been much more productive in my own narrow research topic if I was in a less interesting campus –in a campus where it would not be possible to attend lectures from world-class scholars in different fields and mingle with wonderful, interesting, people. But that would had been so boring actually. I would like to

thank all the people that I met while at UChicago, but let me mention just a few by name, that hold a special position in my heart. Irene Farah was the first person I met outside of the (astro)physics community and a catalyst for personal growth. Thank you so much for showing me a whole new world! Rodrigo Valdés Ortiz, for being such a great friend. I will always remember our long walks and discussions during the pandemic. Thanks for teaching me about networks and for flying from California for my defense! Jessy Morgan, thank you for caring about me all these years and for all planning all my defense celebrations! Without all of you, these six long years would not be the same.

I also want to thank Ying Cai, Sam Passaglia, Georgios Zacharegkas, Melissa Cheng, Alfonso Castillo, Fei Xu, Rayne Liu, for being there at different stages of my long PhD path. The International House at the University of Chicago, of which I had the honor to be a non-residential fellow, offered me great moments and a place to meet and discuss with students from all over the world.

Finally, I just want to express an untargeted gratitude, for what I have done and experienced is something that I could had never imagined a decade ago. Thanks!

ABSTRACT

Low-surface-brightness galaxies (LSBGs), conventionally defined as galaxies with central surface brightness at least one magnitude less than the surface brightness of the ambient dark sky, have remain largely elusive in past wide-field surveys, since the majority of them lie below the surface brightness thresholds of those surveys. At the same time, observational and theoretical arguments point towards an LSBG-dominated galaxy population, especially in the dwarf galaxy regime. Recent observations of radially extended LSBGs, called ultra diffuse galaxies (UDGs) have posed challenges in our understanding of the galaxy formation process.

New, large wide-field and deep galaxies surveys are going to shed light to this elusive low-surface-brightness regime. To do this, and due to the large amount of data they are going to generate, new analysis techniques should be developed, with machine learning providing a promising path for automating and expediting the discovery of LSBGs.

In this work, I start by presenting the discovery and description of a large catalog of LSBGs (21,370 galaxies) from the first three years of the Dark Energy Survey, the largest catalog of LSBGs from a wide-field survey to date. Then I use this catalog to train a deep neural network able to accurately distinguish LSBGs from artifacts in a list of LSBG candidates. Afterward, I show how a computer vision model can be used to detect and remove spurious light reflections in astronomical images, another potential source of noise in LSBG searches. Finally, I show how neural networks can be used to automatically infer structural parameters (radius, surface brightness etc) of galaxies, with simultaneous uncertainty quantification, faster and with similar accuracy as traditional light-profile fitting methods. I conclude with an overview of this work, discussing potential future directions and uses of the results presented in this thesis.

CHAPTER 1

INTRODUCTION

1.1 Prologue: A veil waiting to be lifted

1.1.1 On the shores of Crete

This thesis is about a veil, a veil of light, waiting to be lifted. It is also a meeting point: a meeting point of the most ancient of the sciences, Astronomy, with the newest of the sciences, Computer and Data Science. We will discuss galaxies and artificial intelligence (AI); we will adapt a method developed to locate faces and objects in photographs posted in social media to process images from one of the largest astronomical surveys; we will develop algorithms that can distinguish if an image contains a galaxy or not – much faster than any human astronomer; and we will set the stage for future discovery in a regime that astronomers have just started to be probing.

Our story begins, not so many years ago, in a beautiful remote beach on the island of Crete, around dusk. A young couple with their very little son was enjoying the majestic red and pink color of the horizon. Our nearest star, the Sun, was about to set, and had adopted a deep yellow-red color; half of it was already below the horizon. “Look, look, the Sun is about to say goodbye and go to sleep!”, said the mother. “I don’t want the Sun to disappear and sleep!” replied, almost crying, the boy. “But then you’ll be able to see the moon and the stars!”, continued the mother. “Why can’t I see them now? Why does the Sun has to go to sleep to see them?”.

We know the answer to this question. During the day, when the Sun is up, the incoming sunlight is scattered by the atmospheric particles, spreads and brightens the whole sky. The light from the stars is too dim; we receive many more photons from the scattered sunlight than from the stars. The sunlight blocks our view of the rest of the Universe.

1.1.2 A hidden Universe

What about during the night? It turns out that even the darkest sky, like those in the idyllic beach described above, is not totally dark. There is a limit in how dark the sky can get, even in the absence of any artificial light sources, the Sun and the Moon. There are three main sources of irreducible natural sky brightness: the zodiacal light (a very faint and diffuse glow), the night airglow (emission from the upper parts of the atmosphere), and the scattered starlight.

This thesis investigates objects of very low surface brightness, so let us formally define this quantity. For extended astronomical objects, the surface brightness quantifies the apparent brightness per angular area. It is usually measured in units of magnitudes per square arcsecond (mag/arcsec^2); for a source with total apparent magnitude[1] m and area A, the surface brightness, S, can be calculated as:

$$S = m + 2.5 \log_{10}(A). \tag{1.2}$$

Due to the contributions mentioned above, the surface brightness of the dark night sky is ~ 22 mag/arcsec^2 [e.g., Crumey, 2014, Neilsen et al., 2016] (the exact number depends on the obsered band). The existence of this veil of light poses a question: is our picture of the composition of the Universe, especially of the galaxies that inhabit it, biased and incomplete, since surveys are systematically biased against objects that are intrinsically fainter than the dark sky [Zwicky, 1957, Bothun et al., 1997]? If this is indeed the case, what are the implication for our understanding of the galaxy formation and evolution process, and the dark matter physics (we discuss the connection between galaxies and the dark matter halos

[1] We remind the reader that the apparent magnitude of an object, relative to that of a reference astronomical object (Vega is a common reference choice, for which a magnitude of $m_{\rm ref} = 0$ is assumed) is given by:

$$m - m_{\rm ref} = -2.5 \log_{10}\left(\frac{F_1}{F_{\rm ref}}\right), \tag{1.1}$$

where F is the flux.

the inhabit in the following section)? How can we best utilize the vast amounts of data from current and upcoming galaxy surveys, that produce detailed maps of the sky, to learn more about the properties of the low-surface-brightness galaxies that inhabit our Universe and complete our picture of the galaxy population and its properties?

Although we cannot give an answer to all these questions in the present thesis, we will briefly summarize the recent advances to the growing field of the low-surface-brightness science, and we will develop machine learning and data analysis tools that can be used to discover and analyze the properties of low-surface-brightness galaxies (LSBGs, conventionally defined as galaxies at least one magnitude fainter than the ambient dark sky, Bothun et al. [1997]) in galaxy surveys.

1.2 Galaxies in the Universe

Galaxies are gravitationally bound systems consisted of stars, stellar remnants, interstellar matter (both gas and dust), and (predominantly - depending on the distance from their center) dark matter. Today, with images of beautiful distant galaxies being ubiquitous, it is hard to believe that they were conclusively proven to be stellar systems outside our own Galaxy, the Milky Way, less than 100 years ago. Galaxies (from the Greek work for milk, $\gamma\alpha\lambda\alpha$, "gala" due to the appearance of them on the sky), were initially simply called nebulae and their nature was unknown. Because of their distance, it was impossible for early astronomers to resolve individual stars in them, and they were ever considered nuisance for the interesting astronomical observations of the time. Indeed, when C. Messier published in 1774 his famous catalog of nebulae (many of them found later to be galaxies, like M31 - the Andromeda galaxy) he did so as to provide a catalog of objects astronomers interested in discovering comets should not confuse as such.

After long debates on the nature of those nebulae (whether or not they lie within our own Galaxy), Edwin Hubble, using the 100-inch Mt. Wilson Observatory telescope discovered variable stars (Cepheids) in the arms of spiral nebuae (M31 and M33) [Hubble, 1925]. That was in turn allowed him to estimate the distances to those objects, proving that they lie far from the Milky Way, and thus are separate stellar systems.

Galaxies come in a variety of masses (halo masses from $\sim 10^8 M_\odot$ for dwarf galaxies to $\sim 10^{14} M_\odot$ for giant galaxies), radii (1 kpc–100 kpc), luminosities, and morphologies. Morphologically, galaxies are broadly divided into two categories: *ellipticals*, that lack any sign of structure or disk and have an overall smooth elliptical shape, and *spirals* that are disc-shaped with two or more arms stemming from their center. There are also sub-categories of these, and also irregular galaxies that do not present a structure or a smooth elliptical shape.

Morphology is also related to the stellar and gas content of a galaxy, and thus to its color. Spiral galaxies tend to contain both young and old stellar populations (known as Pop I and Pop II stars, respectively); these galaxies contain clouds of gas and dust and thus they have high ongoing star formation. The population of younger, short-lived, blue stars make them appear bluer, too. Ellipticals, on the other hand, contain older stars, almost absent ongoing star formation and thus they appear redder.

Galaxy formation is a complex process, involving the physics of dark matter, baryonic matter, star formation and feedback loops emerging from it, the effects of the environment where the galaxies are formed, interactions between galaxies etc.

In the standard model of cosmology galaxies are formed inside dark matter halos. A halo is a gravitationally bound region of (dark) matter, decoupled from the Hubble expansion of the universe, and collapsed [e.g., Wechsler and Tinker, 2018]. Galaxies form in a hierarchical way, with small halos (and galaxies) forming first and then merging into larger galaxies and groups of galaxies.

As we said above, galaxy formation is a very complex process, but schematically it is a competition between the gravitational collapse of baryonic matter (gas) that happens within dark matter haloes (and their gravitational potential) and the pressure of gaseous matter that heats up as it collapses [Cimatti et al., 2019]. Secondary processes, such as gas cooling, due to the emission of radiation (and thus energy) are very important in further allowing the process of galaxy formation.

The connection between the galaxies and the dark matter halos they reside in, and the prediction of the galaxy properties from the halo properties is known as the galaxy-halo connection. One basic assumption is that exactly one galaxy is formed within each one dark matter (sub)halo. To connect the distribution of dark matter halos, as obtained from cosmological simulations, one makes the assumption that that the most massive galaxies live in the most massive halos etc [Wechsler and Tinker, 2018]. This is known as abundance matching. With this assumption, the relationship between the typical stellar mass of a galaxy and for a given halo mass, for example, can be inferred. Studies of the mass-halo connection have revealed, for example, that star formation is most efficient in halo masses around $10^{12} M_{\odot}$ (these are usually called L_* galaxies). However, this relationship is much less constrained for dwarf, low-mass galaxies, a regime that, as we will see in the following section, is expected to be dominated by low-surface-brightness galaxies.

1.3 Low-Surface-Brightness Galaxies

1.3.1 Significance of LSBG studies

Our understanding of galaxies, their formation and evolution, and the connection between the galaxies and the dark matter halos they inhabit, as described in the previous section, is constrained by the surface brightness limits of past wide-field galaxy surveys. We note that untargeted, wide-field, galaxy surveys and searches for LSBGs are fundamental if we want to understand the statistical properties of the LSBG galaxy population and role of the environment (inside galaxy clusters vs field galaxies) in their formation history.

As we have said, past wide-field galaxy surveys were lacking depth and thus were highly incomplete in the low-surface-brightness regime. One of the largest, completed, galaxy surveys is the Sloan Digital Sky Survey[2] (SDSS; e.g., York and et al [2000], Abazajian [2009]) has covered $\sim 35\%$ of the sky with photometric observations of nearly a billion objects, and has greatly contributed to our understanding of extragalactic astrophysics and cosmology. SDSS also produced the first large ($\sim$ 12300 galaxies) catalog of LSBGs [Zhong et al., 2008]; however most of them lie in the bright-end limit of the LSB regime, with a median value of central surface brightness of $\mu_0(B) = 22.42$ mag/arcsec^2. Indeed, SDSS starts getting incomplete for surface brightnesses $\mu \sim 23$ mag/arcsec^2, and the completeness drops to $\sim 10\%$ at surface brightnesses $\mu \sim 24$ mag/arcsec^2 [Kniazev et al., 2004], see also the discussion in Jackson et al. [2021]. Despite these caveats, the sample from SDSS was used to study the statistical distribution and the effect of the environment in the formation of LSBGs.

While the LSB universe, especially for surface brightness values > 24.5 mag/arcsec^2, has remained elusive to wide-field surveys, both targeted deep observations and theoretical work suggest that most galaxies in the the dwarf regime (low stellar masses) are LSBGs. For example, Dalcanton et al. [1997] analyzing data from a 17.5 deg^2 survey, down to a surface brightness of ~ 25 mag/arcsec^2, concluded that the number density of LSBGs is comparable or greater than the number density of normal galaxies. Martin et al. [2019], on the theory side, using high-resolution cosmological hydrodynamical simulations, showed that LSBGs contribute $\sim 50\%$ of the local number density of galaxies for stellar masses $M_* > 10^8\, M_\odot$ (the fraction increases to $\sim 85\%$ if we go down to masses $M_* > 10^7 M_\odot$).

Thus, a significant fraction (and potentially the majority) of the galaxy population is absent from the catalogs produced by past wide-field surveys. Since our understanding of galaxy formation and evolution depends on the available galaxy samples, the fact that the high-surface-brightness galaxies (HSBGs, with $\mu \lesssim 23$ mag arcsec^{-2}) that dominate those

[2] https://www.sdss.org/

samples are only a fraction of the underlying galaxy population, most probably renders our picture of galaxy formation and evolution incomplete. Thus, the systematic discovery of a complete galaxy sample in progressively lower surface brightness limits is important.

1.3.2 History and recent observations

Despite the quite recent renewed interest in LSBGs, which will discuss shortly, the existence of "hidden" galaxies with surface brightness much lower to that of the ambient dark sky has been known to the astronomers for over than forty years now. Disney [1976] first theorized the existence of LSBGs, described how the brightness of the night sky imposes strong selection effects on the observations of galaxies, and that the known galaxies represent only the "tip of the iceberg" of the true underlying galaxy population. Sandage and Binggeli [1984] presented examples of large diameter, low-central-surface-brightness dwarf galaxies in the Virgo cluster, while Bothun et al. [1987] reported the discovery of the first verified giant LSB galaxy (Malin 1), a spiral galaxy, with central surface brightness $\mu_V \cong 25.5$ mag arcsec^{-2}. Soon after this first confirmed LSBG discovery, more LSBGs were discovered in the Virgo [Impey et al., 1988] and Fornax [Bothun et al., 1991] clusters. Despite these, and some other examples of early discoveries [e.g., Bothun et al., 1990, Sprayberry et al., 1993, Turner et al., 1993, McGaugh et al., 1995b, Schwartzenberg et al., 1995] LSBGs (especially in large populations) remained largely elusive; see [Bothun et al., 1997, Impey and Bothun, 1997] for a review of the first 20 years of LSBG searches. Indeed, one of the first large-scale, wide-field survey searches for LSBGs came with SDSS that, as we saw in the previous section, is heavily incomplete at surface brightness levels $\mu < 24$ mag/arcsec^2.

The renewed interest in LSBGs was sparked by the discovery by van Dokkum et al. [2015b] of a number (47) of extreme objects in the Coma cluster, with large effective radii ($R_{\text{eff}} = 1.5 - 4.5$ kpc) and low central surface brightnesses ($\mu_0(g) = 24 - 26$ mag/arcsec^2). Those galaxies, appropriately called "ultra-diffuse galaxies" (UDGs), have been found to be occupy the two extremes in terms of dark matter content: van Dokkum et al. [2016, 2019b], for example, using velocity dispersion measurements, showed that one of those UDGs (named Dragonfly 44), has a total mass of $\sim 10^{12} M_{\odot}$ (within the half-light radius of $r_{1/2} = 4.6$ kpc), similar to that of the Milky Way, but only $\sim 1\%$ of its luminosity, suggesting a dark matter fraction of $\sim 98\%$. On the other hand, using similar methods of velocity dispersion measurements, the same team found galaxies [van Dokkum et al., 2018, 2019a] with a halo mass to stellar mass ratio close to unity, suggesting that they lack dark matter. Those observations pose a significant challenge to our understanding of the details of theories of galaxy formation and, as we are going to see below, a number of formation mechanisms have been proposed to explain them. Note that these discoveries were made possible by the introduction of specialized telescopes, optimized for deep observations at the LSB regime, such as the Dragonfly Telephoto Array [Abraham and van Dokkum, 2014].

Targeted, deep, observations have revealed a large number of LSBGs and UDGs around galaxy clusters [see e.g., Koda et al., 2015, Mihos et al., 2015, Muñoz et al., 2015, Martínez-Delgado et al., 2016, van der Burg et al., 2016, Venhola et al., 2017]. Populations of LSBGs have also been discovered in galaxy groups [e.g., Smith Castelli et al., 2016, Müller et al., 2017, Román and Trujillo, 2017b], and in the field [e.g., Zhong et al., 2008, Javanmardi et al., 2016, Papastergis et al., 2017] (see also the discussion in Jackson et al. [2021]). However the relative abundances and the statistical properties of populations across different environments is much less studied, due to the lack of such populations from past wide-field surveys. Current, deeper, surveys have just started producing large LSBG catalogs [Greco et al., 2018, Tanoglidis et al., 2021b, Zaritsky et al., 2022], that will bridge that gap in our understanding of the LSBG population.

1.4 Galaxy Surveys

Astronomical surveys are untargeted observations that map large parts of the sky. Not focusing on a particular object or region, their goal is to provide large samples of objects for statistical analyses. Since their main goal is usually to create catalogs of galaxies, they are also simply called galaxy surveys (notice, however, that the analysis of survey data can lead to other interesting discoveries, such as the comet described in Bernardinelli et al. [2021]). Although in this thesis we focus on uncovering LSBGs from survey data, in order to better understand the galaxy population and provide samples that will help the community to better understand the galaxy formation process, we note that the main goal of the surveys mentioned below is to constrain cosmological models by studying the statistical properties of the galaxy distribution and comparing with theory predictions [e.g., Zhan and Tyson, 2018, Abbott et al., 2022].

Galaxy surveys are broadly divided in two main categories: imaging (also known as photometric) and spectroscopic surveys. In photometric surveys, the light emitted from an object in different wavelength bands is collected using digital cameras consisted of charged-coupled devices (CCDs). The light can then be analyzed to derive physical properties of the objects (brightness, morphological characteristics for extended objects like galaxies, etc), and even estimate distances in a method known as photometric redshift estimation [e.g., Sánchez et al., 2014]. However, the distances obtained using this methods are less accurate than those obtained through spectroscopy [see, e.g., Tanoglidis et al., 2020, for a discussion]. Spectroscopic surveys, on the other hand, are able to provide accurate distance (redshift) measurements; obtaining spectra, however, is expensive, and spectroscopic surveys produce smaller galaxy samples.

In this thesis we focus on data coming from photometric surveys; this means that for the LSBGs we will not be able to have accurate distance estimates (in practice, for nearby galaxies, like the discovered LSBGs discussed in chapter 3, these estimates are totally un-

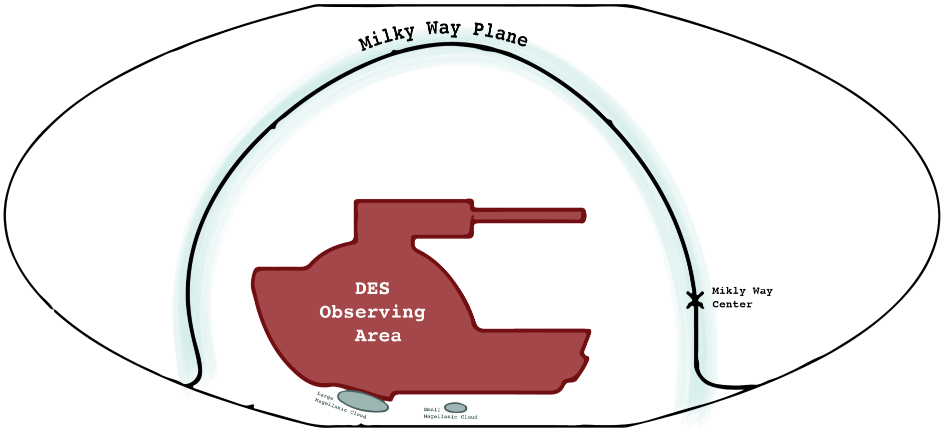


Figure 1.1: DES footprint (observing area) in relation to the Milky Way plane and the large and small Magellanic clouds. Credit: J. Prat.

reliable). Developing techniques for the measurement of distances to LSBGs is actually an active area of current research [e.g., Greco et al., 2021].

We are going to describe in more detail two important galaxy surveys, the Dark Energy Survey (DES), a recently completed survey, that is the source of most of the data used in this thesis; and the Legacy Survey of Space and Time, an upcoming survey that is going to produce large amounts of data and where many of the techniques developed in this thesis can find potential applications. Notice that these are, by no means, the only galaxy surveys that have contributed or are going to contribute to the study of the low-surface-brightness universe. For example, we have seen discussed the SDSS survey; the Hyper Suprime-Cam SSP Survey [Aihara et al., 2018] was used by Greco et al. [2018] to assemble a catalog of 781 LSBGs; and DESI Legacy Imaging Surveys [Dey et al., 2019] were used to produce a catalog of 275 large UDG candidates [Zaritsky et al., 2022]. In terms of future surveys, Euclid, a space-based telescope, will be ideally posed for studies of the low surface brightness Universe [Euclid Collaboration et al., 2022].

1.4.1 The Dark Energy Survey (DES)

The Dark Energy Survey is an optical and near-infrared imaging survey, covering ~ 5000 deg^2 of the southern Galactic cap ($\sim 12\%$ of the full sky area). A schematic representation of the DES footprint (observed area) can be seen in Fig. 1.1. DES is a photometric survey, conducting observations in five broad wavelength bands ($grizY$), using the 570 Megapixel (3 deg^2 field of view) Dark Energy Camera [DECam Flaugher et al., 2015], mounted on the 4-m Blanco telescope at the Cerro Tololo Inter-American Observatory (CTIO) in Chile.

Over its 6 years of observations (758 nights between 2013–2019) it scanned its footprint ten times. The first three years (Y3) of observations (that are used in this work) include ~ 400M distinct astronomical objects, out of which ~ 310M galaxies and ~ 80M stars [DES Collaboration et al., 2018b]. The second DES public data release, stemming from 6 years of operations (Y6), increased that sample to ~ 691M individual objects, out of each ~ 543M galaxies and ~ 145M stars [Abbott et al., 2021].

While the main target of DES is to probe the nature of dark energy and put constraints on cosmological parameters by analyzing the statistical properties of the spatial distribution of galaxies, its depth ($g = 24.33$ in Y3, and $g = 24.7$ in Y6; $i = 23.44$ in Y3 and $i = 23.8$ in Y6) makes it also ideal for studies of faint and low-surface-brightness objects [Dark Energy Survey Collaboration et al., 2016]. Indeed, DES observations have revealed ultra-faint galaxies [Bechtol et al., 2015, Drlica-Wagner et al., 2015], faint stellar systems [e.g., Luque et al., 2017, Cerny et al., 2021], faint stellar overdensities [e.g., Li et al., 2016], stellar streams [Shipp et al., 2018], intracluster light [Zhang et al., 2019], and LSBGs [Tanoglidis et al., 2021b], as we are going to discuss in Chapter 3. We give more details about the DES data we use in corresponding sections within the chapters that follow.

1.4.2 The Legacy Survey of Space and Time (LSST)

The Legacy Survey of Space and Time (LSST) on the Vera C. Rubin Observatory, is an upcoming (expected to begin science operations in 2024) photometric survey, that is going to cover $\sim 18,000$ deg^2 of the southern sky. It will observe in six filters ($ugrizy$) from the ultraviolet to the near infrared, using a powerful 3.2 Gigapixel camera consisted of 189 4k×4k science CCDs (field of view 9.6 deg^2), mounted on the 8.4m Simonyi Survey Telescope in Cerro Pachón, Chile. It is expected to reach a point-source depth of $g = 27.4$ and $i = 26.8$,

Over 10 years of observations, it is expected to produce a catalog of $\sim$ 20B galaxies and $\sim$ 17B resolved stars. It will produce 20 TB per night, for a final database size of 15 PB. While the primary science goals of LSST include probing the nature of the dark matter and the dark energy, due to its unprecedented depth for a wide-field survey, it is expected to play a key role in exploring the low-surface-brightness universe [e.g, Brough et al., 2020, Martin et al., 2022]. LSB targets include and LSBGs and dwarf galaxies, and LSB structures such as merger-induced tidal features and intra-cluster light.

Although we do not focus on LSST in this work, many of the machine learning techniques developed in later chapters can find applications in future analyses of LSST data; notice that the vast amount of data that will be generated by LSST will require fast and automated methods of analysis, such as those we are going to develop in this thesis.

1.5 Thesis Outline

This thesis focuses on the discovery of LSBGs from the Dark Energy Survey data, as well as the development of tools (algorithms) that will further enable the discovery and analysis of LSBGs in future surveys. As we have discussed, there has been significant interest and recent discoveries of LSBGs, which observational and theoretical work indicated that constitute the majority of the galaxy population; however they have remained largely elusive in past wide-field galaxy surveys.

Obtaining large samples of LSBGs across different environments (clusters, groups, field) is important for statistical analyses and for understanding the role of the environment in their formation and evolution. In this thesis we present such a large catalog, stemming from the analysis of the first thee years (Y3) of DES data.

As we enter the era of big data in astrophysics, with galaxy surveys producing vast amount of data, as we saw in the previous section, new analysis techniques would be required to process those data in a fast and efficient way. Especially in the low-surface-brightness regime, where a large number of artifacts can interfere with the astronomical objects of interest, it is imperative to develop algorithms that will allow us to separate the wheat from the chaff automatically, quickly, and accurately.

In Chapter 2 we present a short introduction to the basic concepts and algorithms of machine learning (ML) and a brief overview of ML applications in astrophysics. In Chapter 3 we present the discovery and analysis of a catalog of 23,790 LSBGs from DES Y3 data [Tanoglidis et al., 2021b]. Then, in Chapter 4 we use images of LSBGs and artifacts that we manually labeled during the creation of that catalog to show that, once trained, a convolutional neural network can accurately distinguish between the two image classes, potentially saving significant human effort and time in future LSBG searches [Tanoglidis et al., 2021b]. In Chapter 5 we use a deep learning-based object detection and segmentation model (Mask R-CNN) to identify and remove spurious reflection artifacts from full-focal-plane survey images, artifacts that –due to their low surface brightness– can be a source of noise in future LSBG searches [Tanoglidis et al., 2022]. In Chapter 6 we introduce a method that enables the automatic and fast inference of LSBG structural parameters (such as radius, surface brightness, etc), with uncertainty quantification, that can accelerate the analysis of large LSBG samples from upcoming surveys. Finally in Chapter 7 we summarize and discuss the main results of this thesis, we show examples in the literature where the results of this thesis have already been used, and also some possible future directions of the LSB research.

CHAPTER 2

MACHINE LEARNING IN ASTROPHYSICS

2.1 Introduction

Astrophysics has entered the era of big data. As we saw in the previous chapter, past surveys have produced large astronomical datasets, and future surveys are only going to increase the size of the produced datasets. Past analysis methods are going to be either extremely human-time consuming (e.g. visual inspection of all potential candidates for a specific class/object of interest, like LSBGs or strong lensing events), or will potentially fail to utilize important information (like the traditional 2-point statistical analysis of the distribution of galaxies, which is ubiquitous in cosmology), present in the data. Another limitation of requiring human intervention and inspection of data is that it makes it very difficult to characterize the efficiency and completeness of the process, for example when trying to assemble catalogs of astronomical objects.

This era of data-driven discovery in astrophysics requires the development of new tools and techniques able to tackle those problems. The past decade has seen an explosive development and widespread application of machine learning (ML) algorithms, across different human endeavors, from industrial applications to scientific discovery. Indeed, the number of works in astrophysics and cosmology that use ML has increased over the years (for a review and discussion, see Baron [2019], Dvorkin et al. [2022]). This machine learning revolution of the past decade was made possible mainly because of three factors: development of more efficient algorithms (especially on the deep learning side, that we are going to discuss below), the availability of larger and larger quantities of training data, and improvements in the hardware used to process them (e.g. the introduction of Graphics Processing Units –GPUs– in training deep learning models).

Machine learning algorithms work by identifying patterns in data, and, once trained, can be used to perform tasks common in the astrophysics workflow (like classifying images or inferring the values of parameters of a theoretical model that describes some observed quantity) without the need of human intervention. In this chapter we start by describing the taxonomy of the various machine learning algorithms. Then we briefly present some popular machine learning algorithms that we are going to be used later in this thesis. Finally, we give an (incomplete) overview of some applications of ML in astrophysics, beyond the applications on the LSB regime presented in this thesis.

2.2 Machine Learning Taxonomy

All machine learning algorithms generally try to find (generalizable) patterns in the data, without being explicitly programmed on what patterns to detect (by setting a set of rules, for example). By generalizable, we mean that patterns extracted from a dataset used to train the algorithm, hold when the algorithm is presented with new, unseen data.

Although this is a very general definition, machine learning algorithms can be further categorized based on whether or not they require human supervision (in terms of providing labels to the training dataset) and the function they perform (e.g. classification, as we will soon see). In most cases, the algorithms try to perform the task at hand by minimizing a specific cost (or loss) function, e.g. the number of misclassifications in a classification problem.

In **supervised learning** we provide the algorithm with a number of training data and their corresponding labels. Common supervised learning tasks are classification and regression. In *classification* the data labels belong to a number of distinct categories and the algorithm learns to distinguish between them. Another common task is regression, where the target values are continuous.

In **unsupervised learning**, on the other hand, we have unlabeled data (e.g. categories or target values are not provided). Common unsupervised learning tasks include *clustering* (where one wants to assign each one of the data points to one of a number of different clusters without a specific label associated with them), *dimensionality reduction* (where one wants to project the parameter space of data into a lower-dimensional space), and anomaly detection.

Beyond these two very commonly used ML categories, we also have semi-supervised learning algorithms (where a small amount of labeled data is combined with a larger amount of unlabeled data durning training), and reinforcement learning (where an agent is learning by taking actions, interacting with its a environment and getting feedback/reward). In this thesis we use supervised learning algorithms (mainly for classification, but also for a regression problems), and we will describe a few of them in the section that follows.

Finally let us discuss the difference between **machine learning** and **deep learning**. Deep learning is in practice a subset of machine learning; deep learning algorithms are based on artificial neural networks, that are inspired by the way the brain works. The difference that matters to us is that deep learning algorithms usually work better on *unstructured* data, such as images. While classical machine learning algorithms perform well on structured data (tables) composed of human-selected features, deep learning algorithms are able to extract the features that are important for the task at hand from the data themselves. Usually, training a deep learning/ neural network model requires more time and more data than classical machine learning algorithms, however the results are almost always guaranteed to be superior.

2.3 ML Algorithms used in this work

In what follows, when describing the algorithms, we consider a number of examples, N, each described by a feature vector $\mathbf{x}_i$, $i = 1, \ldots, N$ and with labels y^i.

2.3.1 Support Vector Machines

The SVM classifier [Cortes and Vapnik, 1995] seeks to find a separating hyperplane of the form $\mathbf{w} \cdot \mathbf{x} - b = 0$, with the weights $\mathbf{w}$ and the bias b selected to maximize the margin (distance) between this hyperplane and the training samples that are closest to it (support vectors). In other words, the problem can be characterized as trying to minimize the function $\frac{1}{2}||\mathbf{w}||^2$ such as $y^i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$ for every i. This formulation does not allow for misclassification (hard-margin) though, and thus is prone to overfitting. In practice, we leave some room for misclassification by using a soft-margin SVM, by trying to minimize the following function:

$$\frac{1}{2}||\mathbf{x}||^2 + C\left(\sum_{i}^{N} \xi^i\right), \tag{2.1}$$

subject to $y^i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi^i \ \forall i$. The ξ^i are called slack variables and allow for misclassifications. The variable C controls the "softness" (how much error is allowed) of the classfier and is one of the tunable parameters (*hyperparameters*) of the model.

The above description assumes that the different classes are linearly separable by a hyperplane, though this is not always true. In such cases a kernel function is introduced to perform a non-linear transformation that maps the data to a space where they are linearly separable: $(\mathbf{x}, \mathbf{x'}) \rightarrow K(\mathbf{x}, \mathbf{x'})$. A popular kernel, and the one used in this work, is the Gaussian or Radial Basis Function (RBF) kernel [Orr, 1995]: $K(\mathbf{x}, \mathbf{x'}) = \exp\left(-\gamma||\mathbf{x} - \mathbf{x'}||^2\right)$, where $\gamma = 1/2\sigma^2$ is another tunable hyperparameter.

2.3.2 Random Forests

Another popular and powerful classifier, which we are also going to consider in this thesis is the random forest classifier. Random Forests [Tin Kam Ho, 1995, Breiman, 2001] is an ensemble classifier, in the sense that use a collection of simple classifiers, namely decision trees (DTs), and the output is the majority class derived from them.

A DT divides the dataset into subsets by setting splitting criteria on the features, trying to make these subsets as homogeneous (with respect to the labels of the examples in them) as possible. A DT starts by splitting the data on the feature that results in maximum information gain (imagine asking the most informative question when trying to make a decision) and continues until pure final subsets (nodes) are produced.

The information gain, for a split of a parent node into two (left and right) child nodes, is defined as:

$$IG(D_p, f) = I(D_p) - \frac{N_{\text{left}}}{N_p} I(D_{\text{left}}) - \frac{N_{\text{right}}}{N_p} I(D_{\text{right}}), \tag{2.2}$$

where f is the feature to perform the split, $D_p, D_{\text{left}}, D_{\text{right}}$ refer to the parent, left, and right node datasets, respectively, and $N_p, N_{\text{left}}, N_{\text{right}}$ their corresponding sizes. I is the information gain measure; a popular one is called the Gini impurity, defined as:

$$I = \sum_{i=1}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2, \tag{2.3}$$

with $p(i|t)$ being the proportion of samples that belongs to the class i for a particular node t. This impurity measure is maximized when the classes are perfectly mixed in a node, which minimized the information gain.

The random forests classifier considers a set of n (hyperparameter) DTs and for each one uses a random selection of $\sqrt{m}$ features (m is the total number of features) to construct a DT and train it on a randomly selected subset of the training examples. As mentioned above, the final result is the majority vote (class output) from these n trees.

2.3.3 Artificial Neural Networks (ANNs)

The building blocks of Artificial Neural Networks (ANNs) are the neurons. A neuron takes as input a vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ with weights $\mathbf{w} = (w_1, w_2, \ldots, w_m)$ and a bias parameter,

b, and produces an output:

$$y = g(\mathbf{w} \cdot \mathbf{x} + b), \tag{2.4}$$

where g is called the activation function (or non-linearity), and its purpose is to introduce non-linearities that allow the approximation of arbitrary complex functions. A popular such function, and the one used when deep neural nets are used in this work, is the rectified linear unit (ReLU), $g(x) = \max(0, x)$.

A (feed-forward) deep neural network consists of several layers of neurons, with the outputs of the neurons of the previous layer being the inputs of the neurons of the following one. Let us denote as $\mathbf{x}_{n-1}$ the input vector to the n-th layer, $\mathbf{W}_n$ is a matrix containing the weights of all the neurons of that layer and $\mathbf{b}_n$ a vector of biases. Then, in analogy to Eq. (2.4), the output of the n-th layer is given by:

$$\mathbf{x}_n = g(\mathbf{W}_n \mathbf{x}_{n-1} + \mathbf{b}_n). \tag{2.5}$$

The final layer is called the output layer (the intermediate ones are called hidden) and for a binary classification problem can be just a single neuron. The activation function of the output layer has a sigmoid form, thus the output is a real number between zero and one (thus interpreted as probability) for each one of the examples j, that we denote as $y_j \in \{0, 1\}$. The goal of training a deep learning model is to find a set of weights such as the output vector $\mathbf{y}$ is close to the target output $\hat{\mathbf{y}}$, with each target value being binary, $\hat{y}_j \in [0, 1]$. This is formalized by introducing the loss (or cost) function $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$ and demanding its minimization. To achieve this, the weights and biases are updated at each step via gradient descent:

$$\mathbf{W}_n \leftarrow \mathbf{W}_n - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}_n} \tag{2.6}$$

$$\mathbf{b}_n \leftarrow \mathbf{b}_n - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{b}_n}, \tag{2.7}$$

where η (> 0) is known as the *learning rate* that controls the magnitude of the update at each step. The derivatives can be computed using the backpropagation algorithm (chain rule). For a binary classification problem that outputs probabilities, the binary cross-entropy is commonly used as the loss function:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N} \sum_{j=1}^{N} \hat{y}_j \log_2 y_j + (1 - \hat{y}_j) \log_2 (1 - y_j), \tag{2.8}$$

where the sum is over the training samples. In practice though at each step only a small sample (mini-batch) is used to update the weights at each step, as a way to speed up the computations.

2.3.4 Convolutional Neural Networks (CNNs)

The architecture ($\equiv$ arrangement of neurons and connections between them) we described in the previous subsection corresponds to a fully connected neural network, in which each neuron of a given layer gets is connected to all neurons of the previous one. Such an arrangement has a very large number of trainable parameters (weights) and also it does not preserve spatial information, so it is not optimal for computer vision tasks.

Convolutional Neural Networks (CNNs) were designed to overcome these limitations and were inspired by the way the visual cortex works. The main difference between a fully connected and a convolutional layer is that in the latter the connections happen within a given receptive field.

Each neuron in a convolutional layer is connected only to neurons within a small rectangle in the previous layer, usually of $3-5$ pixels in size, in each direction. The output of such a layer is a predefined number of *feature maps*, generated by convolving the feature maps of each previous layer with different *filters* (or *kernels*), whose trainable weights can capture more and more abstract visual features as we progress on the convolutional network.

If we have $k = 1, \ldots, K$ input feature maps and $\ell = 1, \ldots, L$ output feature maps, in analogy to Eq. (4.1) we write the ℓ-th output map of the n-th layer as:

$$\mathbf{x}_n^{\ell} = g\left(\sum_{k=1} \mathbf{W}_n^{k,\ell} * \mathbf{x}_{n-1}^{k} + b_n^{\ell}\right), \tag{2.9}$$

where $*$ represents the convolution operation, while the matrix $\mathbf{W}_n^{k,\ell}$ contains the filters (weights) corresponding to layer n.

Convolutional layers are almost always followed by *pooling* layers; their main purpose is to subsample the output of the convolutional layer, reducing the number of trainable parameters and the required computer memory. Pooling layers have no weights; what they do is to keep the maximum (max pooling) or the mean (average pooling) within a small (usually 2×2 pixels) window sliding over the input feature map. Training a CNN follows the same procedure as with the ANNs described in the previous subsection.

Note that in this section we have described the training of classic, deterministic, neural networks where some (fixed) weights are learned during the training process and which they output a single numerical prediction when presented with a new example. In Ch. 6 we will describe Bayesian Neural Networks in detail. During the training of a Bayesian Neural Network we learn probability distributions around the weights and the output is a random number drawn from the learned distribution, different one each time the network is presented with the same example.

2.4 Machine Learning workflow

In the previous section we described some popular machine learning algorithms, that we are going to use later in this thesis. Using machine learning, requires more than just picking an appropriate algorithm, though.

From data collection and preparation to selecting and tuning the appropriate model, there is a workflow that has to be followed before deploying a machine learning model and applying to new, unseen, data. Here we very briefly describe that workflow, that is used in every case where machine learning is being applied in this work (we note that in Chapter 3 we have skipped the testing step, because of the limited number of training examples) .

1. **Data collection, annotation, and standardization.** The first, and most important, step in any ML project is the collection of a training dataset. For supervised algorithms, that also means that labels have to be provided, too. Obtaining those labels can be a time-consuming process; for example, as we will see in the following chapter, when trying to classify galaxies vs artifacts, we have to visually inspect candidates and label them as belonging to each one of the two classes.

This process can present an important time bottleneck; at the same time, we know the performance of an ML algorithm significantly improves with larger training datasets, so there is always a trade-off between spending more more time generating training data and improving the performance, and deploying the model earlier.

Finally, before using the data, they usually have to be standardized. In classic ML algorithms the data consist of a number of features (e.g. colors, magnitudes, radii etc in astrophysical problems). Since the numerical values of the different features may have very different scales, some of them may spuriously be interepreted as having greater importance. By standardizing the features (subtracting the mean and diving by the standard deviation), we bring all of them at the same scale.

2. **Train-validation-test split.** Before selecting and training ML models, it is also good to *randomly* split our annotated dataset into three separate datasets: training, validation, and test sets. The proportions usually vary, but a good rule of thumb 70% training and 15% for validation and testing. The reason for this split is that an algorithm can learn very well the statistical properties of the training set but fail to generalize when presented with new data (overfitting). By having these separate datasets, we can train an algorithm, tune

its hyperparameters (using the validation set, see next paragraph), and test on new, unseen, data.

3. **Hyperparameter tuning and model selection.** Hyperparameters are any parameters of the ML algorithms that control the learning process. Such can be the number of trees in a random forrest, the number of layers in a deep neural network. The values of these parameters cannot be learned during training, but rather have to be externally set. In practice, we train a model for a grid of different hyperparameter values and evaluate the performance, for each combination, on the validation set. Then, we re-train the model using the best-performing combination of parameters. Another choice is to perform k-fold cross validation, where the training set is split in k parts, with $k-1$ used for training and the other for validation, and then repeating this process k times.

4. **Performance evaluation.** The final step is to evaluate the final model, once trained, on the left-out test set. There are a number of evaluation metrics (we are going to see them in detail in future chapters), but for classification problems common metrics are the accuracy (fraction of predictions the model got right), precision (proportion of positive identifications was actually correct, can also be called purity), and recall (proportion of actual positives that were identified correctly). Combined metrics, such as the $F1$ score, which is the harmonic mean of the precision and recall scores, also exist. The choice of metric to optimize for depends on the problem (e.g. preferring completeness over purity).

CHAPTER 3

SHADOWS IN THE DARK: LOW-SURFACE-BRIGHTNESS GALAXIES DISCOVERED IN THE DARK ENERGY SURVEY

The text of this chapter was published in Tanoglidis et al. ApJS, 252, 18 (2021)

3.1 Introduction

The low-surface-brightness universe is notoriously difficult to characterize due to the significant impact of observational selection effects [e.g., Disney, 1976, McGaugh et al., 1995a]. Low-surface-brightness galaxies (LSBGs) are conventionally defined as galaxies with central surface brightnesses fainter than the night sky [Bothun et al., 1997]. While these faint galaxies are thought to contribute a minority (a few percent) of the local luminosity and stellar mass density [e.g., Bernstein et al., 1995, Driver, 1999, Hayward et al., 2005, Martin et al., 2019], they may account for $\sim 15\%$ of the dynamical mass budget in the present-day universe [e.g., Driver, 1999, O'Neil et al., 2000, Minchin et al., 2004]. However, due to the observational challenges in detecting these faint systems, LSBGs remain difficult to study as an unbiased population.

LSBGs are known to span a wide range of physical sizes and environments, ranging from the ultra-faint satellites of the Milky Way [e.g., McConnachie, 2012, Simon, 2019], to satellites of other nearby galaxies [e.g., Martin et al., 2013, Merritt et al., 2016a, Martin et al., 2016, Danieli et al., 2017, Cohen et al., 2018b], and members of massive galaxy clusters like Virgo [e.g., Sabatini et al., 2005a, Mihos et al., 2015, 2017], Perseus [e.g., Wittmann et al., 2017], Coma [e.g., Adami et al., 2006, van Dokkum et al., 2015b, Koda et al., 2015], Fornax [e.g., Ferguson, 1989, Hilker et al., 1999, Muñoz et al., 2015, Venhola et al., 2017], and other nearby clusters [e.g., van der Burg et al., 2016]. Untargeted searches have also found a large population of LSBGs in the field [e.g., Zhong et al., 2008, Rosenbaum et al., 2009, Galaz et al., 2011, Greco et al., 2018]. Understanding how LSBGs come to populate this wide range of environments may inform models of cosmology and galaxy evolution. Are LSBGs truly outliers relative to the rest of the galaxy population, or are they merely a natural continuation of the galaxy size–luminosity relation?

The standard model of cosmology (ΛCDM) predicts that galaxies form hierarchically, with smaller galaxies forming first and assembling to form larger galaxies, galaxy groups, and galaxy clusters [e.g., Peebles, 1980, Davis et al., 1985, White and Frenk, 1991]. The formation and growth of galaxies over cosmic time is connected to the growth of the dark matter halos in which they reside (the so-called “galaxy–halo connection”; e.g., Wechsler and Tinker 2018). Many attempts have been made to use the properties of dark matter halos to predict the properties of the galaxies that inhabit them [e.g., Behroozi et al., 2013, Moster et al., 2013]. As extremes in the relationship between galaxy size and luminosity, LSBGs provide a litmus test for models that predict galaxy properties from cosmological principles [e.g., Ferrero et al., 2012, Papastergis et al., 2015]. It has been suggested that LSBGs form naturally within the ΛCDM framework, either primordially in halos with high angular velocity [Dalcanton et al., 1997, Amorisco and Loeb, 2016] or through evolution in dense environments [Tremmel et al., 2019, Martin et al., 2019]. On the other hand, observations of LSBGs with anomalously low dark matter content [van Dokkum et al., 2018, 2019a] may necessitate modified models of galaxy formation [e.g., Papastergis et al., 2017, Sales et al., 2019] and/or dark matter physics [e.g., Carleton et al., 2019]. Disentangling the contributions of various mechanisms for LSBG formation has been historically challenging due to the small volume and highly biased observational samples available.

Over the last few decades, the rapid advance of wide-area, homogeneous, digital imaging has greatly increased our sensitivity to LSBGs. The Sloan Digital Sky Survey (SDSS) enabled statistical studies of large samples of LSBGs down to central surface brightnesses of $\mu_0(B) \sim 24\,\mathrm{mag\,arcsec^{-2}}$ [Zhong et al., 2008, Rosenbaum et al., 2009, Galaz et al., 2011].

Smaller telescopes optimized for the low-surface-brightness regime [i.e., the Dragonfly Telephoto Array; Abraham and van Dokkum, 2014] have illuminated the populations of LSBGs in nearby groups [Merritt et al., 2016a, Danieli et al., 2017, Cohen et al., 2018b] and clusters [van Dokkum et al., 2015b, Janssens et al., 2017], extending down to unprecedented central surface brightnesses of $\mu_0(g) > 27\,\mathrm{mag\,arcsec^{-2}}$. Recently, the Hyper Suprime-Cam Subaru Strategic Program (HSC SSP) revealed a large population of LSBGs with $\bar{\mu}_{\mathrm{eff}}(g) > 24.3\,\mathrm{mag\,arcsec^{-2}}$ in an untargeted search of the first $\sim 200\,\mathrm{deg}^2$ from the Wide layer of the HSC SSP [Greco et al., 2018]. However, results from these deep photometric surveys are still limited to relatively small areas of sky, limiting our ability to characterize the faintest galaxies in an unbiased manner.

Untargeted searches for LSBGs are essential to understand the role that environment plays in their formation and evolution. However, such searches are challenging due to the deep imaging and wide area coverage that is required to provide a statistically significant population of LSBGs. Here we use data from the first three years of the Dark Energy Survey (DES) to detect LSBGs with half-light radii $r_{1/2} > 2.5''$ and mean surface brightness $\bar{\mu}_{\mathrm{eff}}(g) > 24.2\,\mathrm{mag\,arcsec^{-2}}$ over $\sim 5000\,\mathrm{deg}^2$ of the southern Galactic cap. Through a combination of classical cut-based selections on measured photometric properties, machine learning (ML) techniques, and visual inspection, we produce a high-purity catalog of 23,790 LSBGs. We present the spatial, morphological, and photometric properties of this sample based on detailed multi-band Sérsic model fits.

This chapter is organized as follows: In Section 3.2 we describe the DES data set and object catalog used for our search. In Section 3.3 we describe our multi-step selection and measurement pipeline, resulting in our catalog of LSBGs. In Section 3.4 we estimate the efficiency of our catalog selection method by comparing against deeper data around the Fornax galaxy cluster. In Section 3.5, we describe the observed properties of this sample, and in Section 3.6 we examine the statistical clustering of LSBGs. In Section 3.7, we examine the properties of LSBGs that are close in projection to nearby galaxy groups and clusters. We summarize the results of this work in Section 3.8.

3.2 DES Data

DES is an optical–near infrared imaging survey covering $\sim 5000\,\mathrm{deg}^2$ of the southern Galactic cap using the Dark Energy Camera [DECam; Flaugher et al., 2015] on the 4-m Blanco Telescope at the Cerro Tololo Inter-American Observatory (CTIO). The DECam focal plane comprises 62 2k×4k CCDs dedicated to science imaging and 12 2k×2k CCDs for guiding, focus, and alignment. The DECam field of view covers $3\,\mathrm{deg}^2$ with a central pixel scale of $0.263''$. DES observes with a dithered exposure pattern to account for gaps between CCDs [Neilsen et al., 2019] and combines the individual exposures into coadded images that are $0.73 \times 0.73\,\mathrm{deg}$ in size [Morganson et al., 2018]. The median sky brightness levels in the DES exposures are $g = 22.01, r = 21.15$, and $i = 19.89\,\mathrm{mag\,arcsec^{-2}}$ [DES Collaboration et al., 2018b].

We use data collected from the first three years of DES observing (DES Y3). This data set shares the same single-image processing, image coaddition, and object detection as the first DES data release [DR1; DES Collaboration et al., 2018b]. In particular, object detection was performed on $r + i + z$ coadded detection images using `SourceExtractor` [Bertin, 2006]. Photometric measurements were performed in each band using `SourceExtractor` in "dual image" mode using the band of interest in combination with the detection image. The depth of the DES Y3 object catalog at signal-to-noise ratio (S/N) = 10 based on the `SourceExtractor` adaptive aperture fit (`MAG_AUTO`) is $g = 23.52$, $r = 23.10$, and $i = 22.51$ [DES Collaboration et al., 2018b]. The DES pipeline was optimized for the detection and measurement of galaxies at cosmological distances, which are generally faint and relatively small in projected size.

Sky background estimation is an important component in the detection of extended

LSBGs. In DES Y3, sky background estimation and subtraction were performed in two phases [Morganson et al., 2018]. First, the background was fit using a principal components analysis (PCA) algorithm applied to the full focal plane binned into 128 × 128 superpixels that are $\sim 1'$ in size [Bernstein et al., 2018]. Next, `SourceExtractor` was used to fit the residual local background on each CCD using a bicubic spline fit to 256 × 256 pixel blocks, which are again $\sim 1'$ in size [Bertin, 2006, Morganson et al., 2018]. For comparison, the half-light radii of the LSBGs in this study range from $2.5''$ to $\sim 20''$ in radius. Background modeling may reduce the efficiency for detecting larger and lower surface-brightness sources, and we leave further background modeling optimization to future work.

We estimated the surface-brightness contrast on $10'' \times 10''$ scales for each DES coadd tile using the `sbcontrast` module from Multi-Resolution Filtering packaged developed for the Dragonfly Telephoto Array [van Dokkum et al., 2020].[1] This procedure bins each coadd image on the desired scale, subtracts a local background from each binned pixel based on the surrounding 8 pixels, and calculates the variation among the binned and background-subtracted pixels [e.g., Gilhuly et al., 2020]. We applied this procedure to each DES coadd tile after masking bad pixels and sources detected by `SourceExtractor`. We find that on $10'' \times 10''$ scales, the median surface brightness limit at 3σ is $g = 28.26^{+0.09}_{-0.13}, r = 27.86^{+0.10}_{-0.15}, i = 27.37^{+0.10}_{-0.13}\,\mathrm{mag\,arcsec^{-2}}$, where the upper and lower bounds represent the 16th and 84th percentiles of the distribution over DES tiles (Appendix 3.9).[2] These values can be directly compared to the 3σ surface-brightness contrast of $g = 28.616, r = 28.936\,\mathrm{mag\,arcsec^{-2}}$ reported for Dragonfly observations of NGC 4565 [Gilhuly et al., 2020]. However, we note that the DES source detection pipeline has not been optimized for the detection of large, low surface-brightness sources, and so the source detection threshold cannot be directly compared to other catalogs optimized to this purpose.

[1] `https://github.com/AstroJacobLi/mrf`

[2] The uncertainty within individual tiles is sharply peaked at a median value of $0.004\,\mathrm{mag\,arcsec^{-2}}$.

3.3 LSBG Catalog

Here we describe the pipeline used to identify and measure LSBGs in the DES Y3 data. Briefly, we start with a generic catalog of `SourceExtractor` detections and use the morphological and photometric properties to identify a subset of LSBG candidates. We train a machine learning algorithm to remove artifacts and visually inspect the resulting candidate list to assemble a high-purity catalog of LSBGs. We then fit a Sérsic profile to each identified LSBG in order to determine photometric properties in a manner that is consistent with previous work [e.g. Greco et al., 2018]. Our full catalog of DES LSBGs is available as supplemental material.[3]

3.3.1 Initial sample selection

We began with the DES Y3 Gold coadd object catalog (v2.2) assembled from `SourceExtractor` detections [Sevilla-Noarbe et al., 2020]. We first removed objects classified as point-like based on the i-band `SourceExtractor` `SPREAD_MODEL` parameter (see Appendix 3.10 and Sevilla-Noarbe et al. 2020 for more details). Following Greco et al. [2018], we defined our initial sample of candidate LSBGs based on angular size and surface brightness. Because these cuts were primarily intended to reject imaging artifacts, no correction for interstellar extinction was applied at this stage. We required that sources have half-light radii in the g band (as estimated by `SourceExtractor` `FLUX_RADIUS`) to be in the range $2.5'' < r_{1/2}(g) < 20''$[4] and mean surface brightness $24.2 < \bar{\mu}_{\rm eff}(g) < 28.8$ mag/arcsec^2.[5] We also restricted our selection

[3] `https://des.ncsa.illinois.edu/releases/other/y3-lsbg`

[4] After assembling our catalog, we inspected all the candidates ($\sim$1,500) satisfying our color and surface brightness cuts and having $r_{1/2}(g) > 20''$. We found 6 LSBGs that were subsequently included in our catalog.

[5] Note that there is a difference in the mean surface brightness selection, compared to Greco et al. [2018] that uses $24.3 < \bar{\mu}_{\rm eff}(g) < 28.8$mag/arcsec^2. Our definition is slightly more inclusive, and the reader should keep this in mind when comparing to the HSC catalog from Greco et al. [2018].

to objects with colors (based on the `SourceExtractor` `MAG_AUTO` magnitudes) in the range:

$$-0.1 < g - i < 1.4 \tag{3.1}$$

$$(g - r) > 0.7 \times (g - i) - 0.4 \tag{3.2}$$

$$(g - r) < 0.7 \times (g - i) + 0.4. \tag{3.3}$$

These color cuts were guided by the HSC SSP analysis of Greco et al. [2018], and were found to produce similar results in DES. Furthermore, we required the objects in our catalog to have ellipticity < 0.7, to eliminate some high-ellipticity spurious artifacts (i.e., diffraction spikes). Our complete selection criteria are presented in Appendix 3.10. After performing the cuts described above, our sample consisted of 419,895 objects from an initial catalog of $\sim$400 million objects.

3.3.2 Machine Learning Classification

Visual inspection of a few thousand candidates passing the cuts described in the previous section revealed that $\lesssim 8\%$ of the objects passing these selections were LSBGs. The most common sources of contamination were:

1. Faint, compact objects blended in the diffuse light from nearby bright stars or giant elliptical galaxies.
2. Bright regions of Galactic cirrus.
3. Knots and star-forming regions in the arms of large spiral galaxies.
4. Tidal ejecta connected to high-surface-brightness host galaxies.

The large size and low purity of our initial candidate list was well suited to the application of conventional ML classification algorithms. Our goal with ML classification was to reject

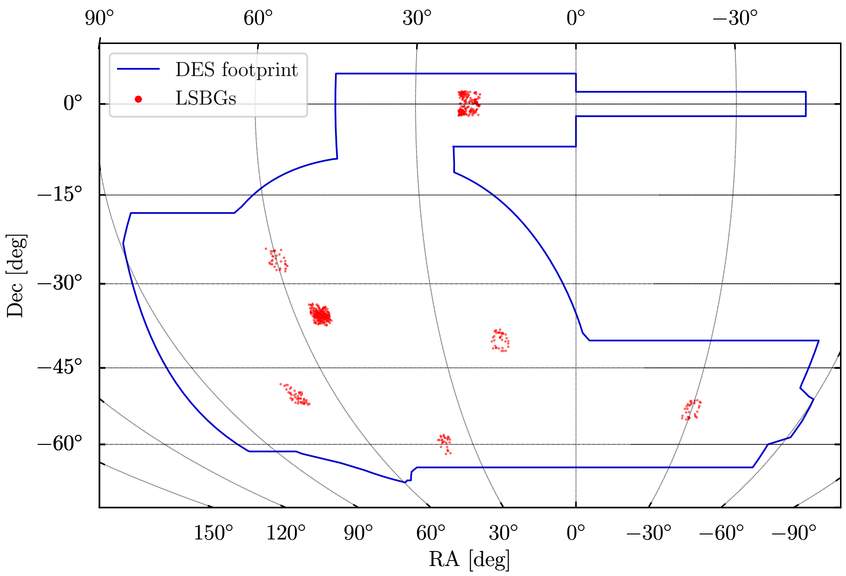


Figure 3.1: The distribution of the objects visually classified as LSBGs in the seven 4° × 4° regions used to create the labeled set for classification and validation. The Fornax galaxy cluster is located at (RA, DEC) $\sim (55°, -35°)$.

a large fraction of false positives while retaining high completeness for true LSBGs.

Training Set

In order to train a supervised ML classification algorithm, we required a sample of objects where the true classification was known. To avoid biases when training the classifier, we seek to assemble a labeled training sample that is representative of the full LSBG candidate sample. We created a labeled sample by visually inspecting all objects that pass the cuts defined in Section 3.3.1 in seven patches spread over the DES footprint, comprising $\sim 100\,\mathrm{deg}^2$ (Figure 3.1). One of these regions was centered on the Fornax galaxy cluster, which is known to contain a high concentration of LSBGs [e.g., Muñoz et al., 2015], while the locations of the other regions were selected at random. Our training set consists of 7760 visually inspected objects, of which 640 were classified as LSBGs.

Features and Classifiers

We split the labeled objects into two sets: 75% of the labeled objects were used as a training set, while the remaining 25% were used as a validation set. We used the validation set to evaluate the performance of different classifiers and tune their hyperparameters. Since the ML classifier was used solely as a precursor to visual inspection, we were not concerned with precisely characterizing its performance. Thus, rather than allocating an independent testing sample, we used our entire labeled data set for training and validation.

In the classification, we used 18 features derived from the `SourceExtractor` measured properties without correcting for interstellar extinction. Specifically, we used:

1. The adaptive aperture magnitudes in the g, r, i bands, `MAG_AUTO`.

2. The colors $(g - r)$, $(g - i)$, and $(i - r)$ derived from the adaptive aperture magnitudes.

3. The size of a circular isophote containing half the flux in the g, r, i bands, `FLUX_RADIUS`.

4. The effective surface brightness in the g, r, i bands, `MU_EFF_MODEL`.

5. The maximum surface brightness measured by `SourceExtractor` in the g, r, i bands, `MU_MAX`.

6. The semi-major and semi-minor axes of the isophotal ellipse containing half the light, `A_IMAGE` and `B_IMAGE`.

7. The isophotal ellipticity, $1 -$ `B_IMAGE/A_IMAGE`.

We tested a number of popular classification algorithms, as implemented in the Python library `scikit-learn` [Pedregosa et al., 2011b].[6]. Specifically, we tested naive Bayes, AdaBoost, nearest neighbor, random forest, linear support vector machines (SVM), and SVM with radial basis function (RBF) kernel classifiers. Due to the relatively small size of our training set (and specifically the small number of positive instances), we did not attempt classification using deep learning techniques.

[6] `https://scikit-learn.org/stable/`

Our goal was to find a classifier that minimized the false-negative rate (FNR)—i.e., true LSBGs classified as false detections—while keeping the true-positive rate (TPR) reasonably high. In other words, we favored completeness over purity in the sample classified as LSBGs. This choice was motivated by our goal to reduce the candidate sample to a tractable size for visual inspection (which would reject the remaining false positives), without losing many real LSBGs in the process.

Note that the samples in our training data were heavily imbalanced: from the 5820 objects (7760 × 0.75) only 480 (640 × 0.75) were true LSBGs. Class imbalance can lead to low accuracy in predicting the label of objects belonging to the less frequent class. We dealt with this by weighting the classes using the `class_weight` parameter. Setting this parameter equal to `"balanced"` assigns each class a weight that is inversely proportional to its frequency, $w_j = n/2n_j$, where w_j is the weight of the $j-$th class and n, n_j are the total number of observations and observations of the $j-$th class, respectively.

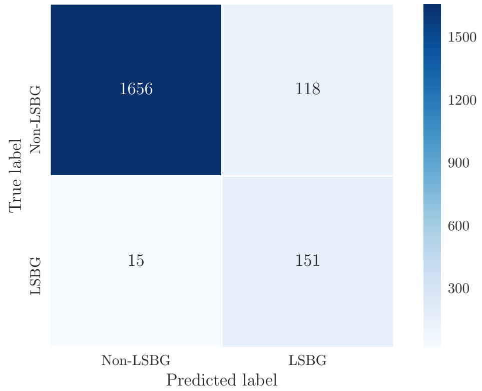


Figure 3.2: The confusion matrix of our final SVM classifier evaluated on the validation set. The quoted numbers correspond to the number of the validation instances (objects) based on their true and predicted label. The false-negative rate is $\sim 9\%$.

We found that the optimal classifier for our specified goal was an SVM classifier with an RBF kernel and parameters $C = 10^4$ and $\gamma = 0.012$ (These parameters are related to the sensitivity to the missclassification rate of training examples vs simplicity of the decision boundary, and the influence of a single training example, respectively. For more details on

SVMs see, e.g., Hastie et al. [2001]). In Figure 3.2, we present the confusion matrix for this classifier, evaluated on the validation set. We see that the FNR, defined as the fraction of true LSBGs classified as non-LSBGs ($\mathrm{FNR} = \mathrm{FN}/(\mathrm{FN} + \mathrm{TP})$), is $\sim 9\%$. We visually inspected the 15 LSBGs rejected by the SVM classifier, as well as examples of LSBGs that were correctly classified. Comparing the two cases, we find that the rejected objects are systematically fainter (about one magnitude in mean surface brightness) than the LSBGs that passed the classification step.

From the same plot, we expect that $\sim 44\%$ of the objects classified as LSBGs are false positives. Subsequent visual inspection (Section 3.3.3) showed that the number of false positives was consistent with the estimate presented here.

Using the optimized classifier, as described in the above section, we classified the 419,895 LSBG candidates that were selected by the cuts defined in Section 3.3.1. The classification returned 44,979 objects classified as LSBGs, thus reducing the sample by about an order of magnitude.

3.3.3 Visual Inspection

The next step in the generation of our LSBG sample was visual inspection of objects that were classified as LSBGs by our ML classifier. We generate $30'' \times 30''$ cutouts centered at the coordinates of each of the candidates, and we inspect candidates in batches of 500. For cutout generation, we use the DESI Legacy Imaging Surveys sky viewer to access the DES DR1 images.[7]

Figure 3.3 shows cutouts around 20 candidates passing our ML classifier. Our visual inspection procedure classified candidates 2, 3, 8, 11, 12, 13, 14, 15, and 18 as LSBGs. Some of these objects are elliptical galaxies while others are spirals. We see that candidates 10 and 11 represent the same object, as do 4, 5, 6, and 7. These duplicates come from

[7] http://legacysurvey.org/

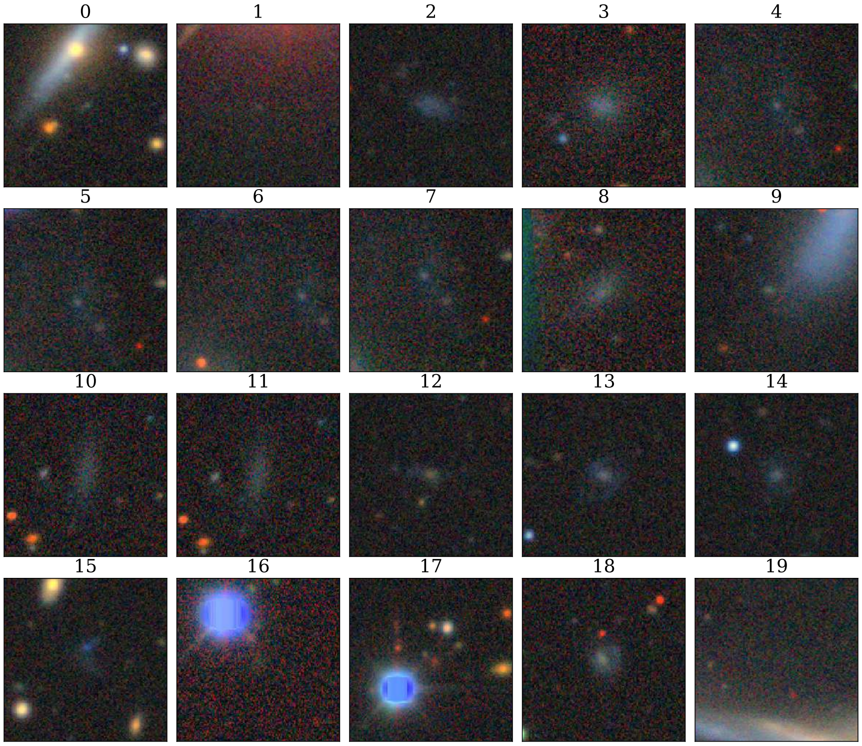


Figure 3.3: $30'' \times 30''$ cutouts of 20 candidates, positively classified by our machine learning algorithm (Section 3.3.2). Candidates 2, 3, 8, 11, 12, 13, 14, 15, and 18 are visually classified as LSBGs, while the other candidates are rejected as false positives and/or duplicates.

`SourceExtractor` shredding larger galaxies into smaller constituents. When we find sources that have been shredded in this way, we make an effort to "stitch" the segmentation maps back together for the `galfitm` (Section 3.3.4). In these cases, we picked the candidate that was best centered on the galaxy; in the example presented here, these are candidates 11 and 4. To avoid further contamination from duplicates in our sample, we also ran an automated spatial cross-match on our final catalog to remove duplicate objects separated by $< 4''$. Candidates 0, 1, 9, 16, 17, and 19 were rejected by visual inspection as false positives. For some candidates (i.e., number 4), it is not immediately clear whether they are isolated LSBGs or tidal debris from larger nearby galaxies. In these cases, we used the DES Sky Viewer[8] to inspect the region surrounding the candidate. The DES Sky Viewer provides flexible

[8] `https://desportal2.cosmology.illinois.edu/sky/`

zooming and scaling, and we ended up rejecting candidate 4, because it is a point-like object blended with the diffuse light of a large galaxy centered outside of the cutout. We note that we make no attempt to distinguish between small, low-luminosity, nearby LSBGs and large, luminous, distant LSBGs.

After visual inspection, our sample contains 21,292 objects. Although we tried to minimize false positives, this sample may still contain a small fraction of low-surface-brightness contaminants such as:

1. Ejecta from large galaxies that reside outside the small angular size of the cutouts.
2. Small background galaxies in the halos of bright stars.
3. Recent mergers with extended halos of stellar debris.

3.3.4 Sérsic Model Fitting

To compare the properties of our LSBG catalog against similar catalogs in the literature [e.g., Greco et al., 2018], we fit each galaxy with a single-component Sérsic light profile. We use `galfitm`, a multi-band implementation of `galfit` developed in the context of the `MegaMorph` project [Peng et al., 2002a, Barden et al., 2012, Häußler et al., 2013], to perform a multi-band fit for each galaxy using the DES coadd images from the g, r, and i bands. We started by creating square cutout images centered on each galaxy. The cutout size was set to be $10 \times the$`FLUX_RADIUS` of each galaxy (rounded up to the nearest 50 pixel step). A minimum cutout size of $201 \times 201\,\mathrm{pix}$ ($\sim 50''$ on a side) was used for small galaxies. We assembled a mask in each band by combining the segmentation map from the DES detection coadd (a combination of the r, i, z images) with the bad pixel mask from each individual band. The `galfitm` "sigma image" was derived from the inverse variance weights plane produced by `SCAMP` [Bertin, 2006] for each of the DES coadded images.

Large LSBGs are sometimes segmented into several catalog objects by `SourceExtractor`. Since we are using the segmentation map as a mask, regions of the image associated with other SourceExtractor sources are excluded from the galfitm analysis by default. These "siblings" of the LSBG often consist of foreground stars, background galaxies, and various stellar overdensities associated with the LSBG itself (e.g., globular clusters, star forming regions, nuclei of recently merged satellites, etc.), as well as spurious shredding of the (mostly) smooth emission of the LSBG. To avoid unnecessary masking, we visually inspect the segmentation maps of each LSBG in our sample. We remove mask regions associated with spurious shredding, while retaining masks associated with compact, high-surface brightness objects. Approximately 5% of our LSBG sample had segmentation maps modified in this way.

The parameters of the Sérsic model fit were initialized based on the values of the SourceExtractor catalog. The centroid was initialized at the position derived by SourceExtractor, and was constrained within 10% of the FLUX_RADIUS. The Sérsic effective radius was similarly initialized based on the FLUX_RADIUS and was constrained to be within a factor of 2 from this initial value. The Sérsic index was initialized at a value of $n = 1.0$ and was constrained to lie within the range $0.2 < n < 5.0$. The galfitm package uses a series of Chebyshev polynomials to parameterize the morphological parameters as a function of wavelength [Häußler et al., 2013].

When performing the fit with galfitm, we tied the centroid position, Sérsic index, ellipticity, and position angle across the three bands. In contrast, the flux normalization of the model was allowed to vary independently in each band according to a quadratic function of wavelength, and the effective radius was fit in each band as a linear function of wavelength. This has the effect of constraining color gradients to vary monotonically with wavelength. We visually inspect the residuals of each fit to identify and correct catastrophic errors. The resulting best-fit Sérsic model parameters are provided as supplemental material.

While the Sérsic model fit provides consistent properties across all objects in our sample and allows comparison to similar catalogs in the literature, it is not a sufficiently complex model to provide a good fit for all LSBGs. In particular, we note that a subset of our objects would be fit better through the inclusion of a nuclear point source, while others show clear indications of irregular, peculiar, or spiral structure. We provide a local estimate of the reduced χ^2 (χ^2 per degree of freedom) of our model in each band calculated within the central region of each LSBG. This information can be used to identify objects that were poorly fit by the simple Sérsic model, and can be followed up with more detailed modeling. The most common modeling issue comes from the existence of compact nuclear sources, which often lead to local $\chi^2 > 3$.

3.3.5 Extinction Correction and Final Cuts

We corrected for the effects of Galactic interstellar extinction on the magnitudes and other derived quantities (color and surface brightness) of our sample. We used the fiducial DES interstellar extinction coefficients [see Section 4.2 of DES Collaboration et al., 2018b]. Briefly, these were derived from the $E(B-V)$ maps of Schlegel et al. [1998] with the normalization adjustment of Schlafly and Finkbeiner [2011] using the reddening law of Fitzpatrick [1999] with $R_V = 3.1$. For the remainder of this paper, we refer only to the extinction-corrected properties of our sample.

As a final step in defining our LSBG sample, we require that galaxies have $R_{\rm eff}(g) > 2.5''$ and $\bar{\mu}_{\rm eff}(g) > 24.2\,{\rm mag\,arcsec}^{-2}$ [9] based on the extinction-corrected Sérsic profile fit. After performing these cuts, our final sample contains 23,790 LSBGs distributed over the $\sim 5000\,{\rm deg}^2$ DES Y3 footprint. Interestingly, the average angular number density of LSBGs in DES Y3 ($\sim 4.5\,{\rm deg}^{-2}$) is similar to that found in the first $\sim 200\,{\rm deg}^2$ of HSC SSP [$\sim 3.9\,{\rm deg}^{-2}$, Greco et al., 2018]

[9] Note that there is no consensus in the literature about the definition of the effective radius of the LSBGs. Some authors use the semi-major axis $R_{\rm eff} = a$ of the ellipse used in the Sérsic model fit, while others use the circularized effective radius, defined as $\overline{R_{\rm eff}} = R_{\rm eff}\sqrt{b/a}$. We use the first option, and then we estimate the mean surface brightness as the total flux contained within the ellipse over its area.

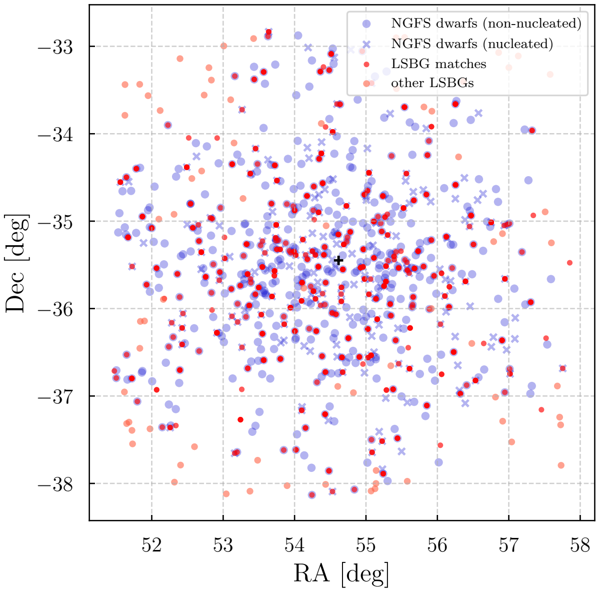


Figure 3.4: The dwarf galaxies present in the NGFS catalog (in blue) and the matches from our DES LSBG catalog (red). The NGFS catalog is separated into nucleated (denoted by an 'X') and non-nucleated (circles) galaxies. We plot DES LSBGs that were not matched to NGFS objects in light red (these are generally located outside the NGFS area). The black cross denotes the nominal center of the Fornax cluster.

3.4 Detection Efficiency around the Fornax Cluster

Table 3.1: Detection efficiency around the Fornax Cluster

Cuts applied	**All galaxies**	**Nucleated**	**Non-nucleated**
No cuts	76.6%	89.5%	71.6%
Surface-brightness cut only	63.1%	58.6%	64.9%
Angular size cut only	56.4%	81.8%	46.4%
Both cuts	43.4%	52.5%	40.3%
Final result (after ML/Vis. inspection)	37.7%	46.9%	34.1%

To estimate the efficiency of our multi-step LSBG selection procedure, we compare our LSBG catalog to similar catalogs produced with deeper data (note that here by deeper we refer to the point-source depth, not the surface brightness). The Fornax galaxy cluster (Abell S373) resides within the DES footprint and is known to host a large population of faint galaxies [e.g., Ferguson, 1989, Hilker et al., 1999, Muñoz et al., 2015, Venhola et al., 2017]. In particular, the Next Generation Fornax Survey [NGFS; Muñoz et al., 2015] has used DECam to image the region around Fornax to an S/N = 5 point-source depth of $g = 26.1$ and $i = 25.3$, which is approximately 2 magnitudes deeper than the DES Y3 imaging in this region of the sky. The NGFS has assembled catalogs of dwarf galaxies covering $\sim 30\,\mathrm{deg}^2$ around the Fornax cluster. The NGFS has reported a total dwarf galaxy population of 643 galaxies, which is split into nucleated (181) and non-nucleated (462) galaxies [Eigenthaler et al., 2018, Ordenes-Briceño et al., 2018].

The NGFS dwarf galaxy catalogs were assembled through visual inspection of the DECam data surrounding Fornax. The NGFS catalog creation process was specifically focused on identifying dwarf galaxies/LSBGs, and it did not apply any cuts similar to those that we imposed on the photometric DES catalog. This makes the NFGS an interesting independent data set to quantitatively evaluate the efficiency of our catalog creation and LSBG sample selection procedures.

We match the NGFS catalogs from Eigenthaler et al. [2018] and Ordenes-Briceño et al. [2018] with the DES Y3 Gold catalog using a matching radius of $3''$ (we find that using a larger matching radius does not significantly increase the number of matches). In Table 3.1, we report the fraction of objects from the NGFS catalog that are matched to objects in the DES Y3 Gold catalog before any cuts, and the resulting change in the matched fraction of galaxies as we apply each of the LSBG selection criteria defined in Section 3.3. This allows us to estimate the efficiency of each cut and the completeness of our final LSBG sample relative to the NGFS sample. We also examine the efficiency of our selection to nucleated and non-nucleated galaxies separately, since the non-nucleated galaxies in the NGFS were found to be fainter and smaller than their nucleated counterparts.

Table 3.1 shows that $\sim 77\%$ of the NGFS galaxies were matched to objects in the DES Y3 Gold catalog generated with `SourceExtractor`. As expected, the recovery fraction is higher for the nucleated LSBGs where the DES detection efficiency reaches $\sim 90\%$. Our surface-brightness cut significantly reduces the number of detected objects, affecting nucleated galaxies more strongly due to their higher central surface brightnesses. The angular size cut, $r_{1/2} > 2.5''$, results in a more significant reduction in the efficiency for recovering non-nucleated galaxies. We expect that this angular size cut will result in an even more severe reduction in the number of distant LSBGs that pass our cuts, since more distant galaxies will be required to have larger physical sizes.

After applying both surface-brightness and size criteria, the detection efficiency drops to 43.4% overall, with a detection efficiency of 52.2% and 40.3% for the nucleated and non-nucleated subsamples, respectively. We further examine the decrease in efficiency from applying our machine learning classification and visual inspection. We find that the drop in efficiency (difference between the last two rows of Table 3.1) corresponds to an absolute drop of ~13% in the number of LSBGs in the field that were not detected. That number is consistent with our expectation that the ML classification has FNR ~ 10% (Figure 3.2). Furthermore, visual inspection of misclassified galaxies showed that most were either extremely faint/hard to distinguish from random background fluctuations or too compact to be included in our LSBG catalog.

Figure 3.4 shows a scatter plot of the NGFS dwarfs, matched LSBGs from our catalog, and unmatched LSGBs in the region around the Fornax cluster. Some of them (~5) are close to an NGFS object and would have been matched with a slightly larger matching radius. This figure also shows the presence of LSBGs detected in our catalog but not present in the NGFS catalog. Most of these galaxies reside outside of the NGFS footprint. Within half the projected virial radius of the Fornax cluster [~ 700 kpc, Drinkwater et al., 2001], we find 11 LSBGs not present in the NGFS catalog.

Overall, our analysis here shows that our pipeline is able to retrieve most of NGFS LSBGs, as we defined them based on the surface-brightness and radius cuts.

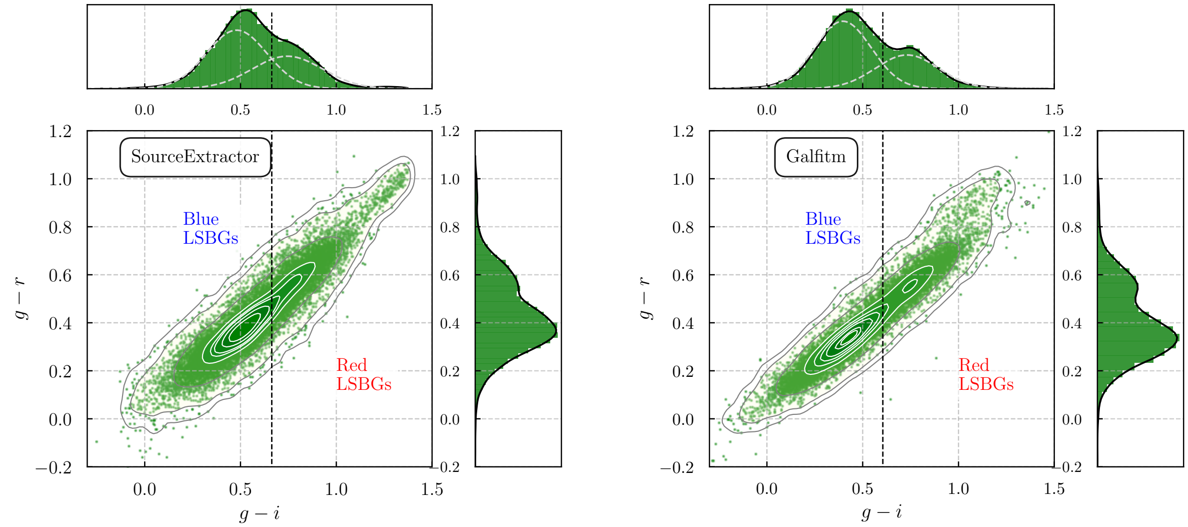
NGFS has the benefit of having been conducted with the same instrument as DES, thus optimal for comparison with our catalog. However, completeness estimates are not provided. The Fornax Deep Survey (FDS) provides a catalog of 564 dwarf galaxies around Fornax, together with completeness estimates from simulations [Venhola et al., 2017, 2018]. This catalog is $\geq 50\%$ complete at a mean surface brightness (in the r band) of $\bar{\mu}_{\text{eff}}(r) = 26.0\,\text{mag}\,\text{arcsec}^{-2}$.

We match our sample with the FDS catalog using a matching radius of $3''$. Before applying any cuts, we find that $\sim$ 92% of the galaxies in FDS are also present in the DES data. We repeat this matching after applying cuts of $\bar{\mu}_{\text{eff}}(r) > 24.2\,\text{mag}\,\text{arcsec}^{-2}$ and $R_{\text{eff}}(r) > 2.5''$ (only r-band data were provided for FDS) to both the DES catalog and the FDS catalog. We find that $\sim$ 66% of the galaxies in the FDS catalog are contained in the DES catalog. A more detailed analysis of efficiency as a function of surface brightness and radius is not very informative given the small number of galaxies that pass the LSBG selection. However, we find that the DES LSBG catalog is $80-90\%$ complete for the lowest- and highest-surface-brightness galaxies.

3.5 LSBG Properties

The large sky area covered by DES ($\sim 5000\,\text{deg}^2$) gives us a unique opportunity to study the statistical properties of the LSBG population. Our search results in a sample of 23,790 LSBGs with effective radii $R_{\text{eff}}(g) > 2.5''$ and extinction-corrected mean effective surface brightnesses $\bar{\mu}_{\text{eff}}(g) > 24.2\,\text{mag}\,\text{arcsec}^{-2}$. This is the largest such catalog of LSBGs to date. In this section, we divide our catalog of LSBGs into red and blue subsamples and compare the properties of these samples to each other and to previous results [i.e., Greco et al., 2018].

The optical colors of galaxies are indicative of their stellar populations. Colors are known to correlate strongly with galaxy morphology and environment. Galaxies are conventionally divided based on color into two well-known sequences of red and blue galaxies [e.g., Strateva et al., 2001, Blanton and Moustakas, 2009]. Less is known about how the colors of LSBGs



(a)

(b)

Figure 3.5: Color–color diagram of our LSBG sample, using (a) `SourceExtractor` `MAG_AUTO` parameters and (b) magnitudes derived by fitting with `galfitm`. In both cases, we observe a bimodality in the $g-i$ and $g-r$ color distributions. We separate the total sample into red and blue galaxies, based on their $g-i$ color value: we fit the $g-i$ distribution with a Gaussian mixture model with two Gaussians (gray dashed lines in the top panels) and find the intersection point. This is at $g-i=0.66$ and $g-i=0.60$ for the `SourceExtractor` and `galfitm` cases, respectively (black vertical dashed lines). We use the intersection point derived from the `galfitm` distribution to define red and blue LSBG samples.

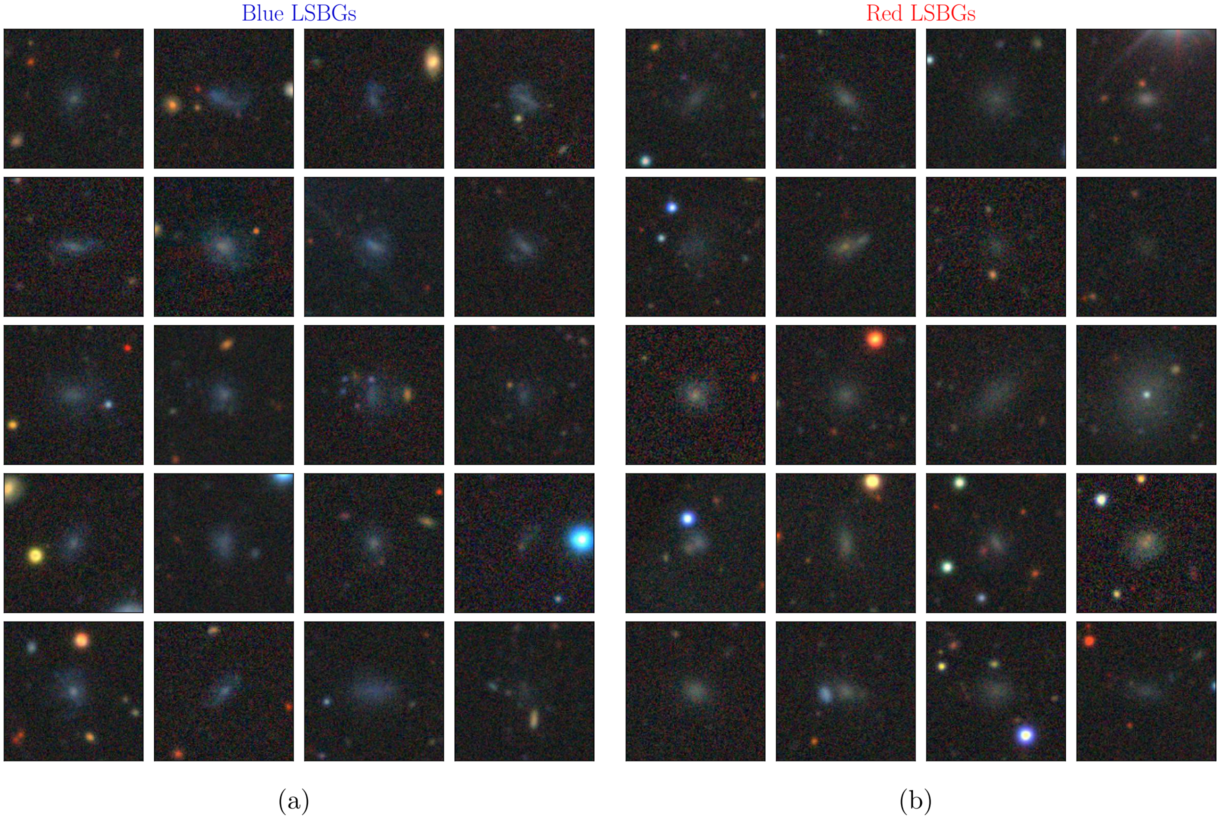
correlate with morphology, star formation history, and environment. For example, O'Neil et al. [1997] found that classical disk LSBGs span a range of blue and red colors. Similar to high-surface-brightness galaxies (HSBGs), blue colors are generally associated with actively star forming spiral or irregular systems, while red colors tend to be indicative of spheroidal or elliptical morphology [e.g., Larson et al., 1980, Strateva et al., 2001, Baldry et al., 2004, Lintott et al., 2011]. Red galaxies are found preferentially in denser environments, where quenching from massive hosts prevents ongoing star formation [Bamford et al., 2009, Geha et al., 2017, Román and Trujillo, 2017b]. Greco et al. [2018] found that LSBGs detected in HSC showed a clear bimodality in color, with two apparently distinct populations separated at $g'-i'=0.64$ (where g' and i' are used to indicate extinction-corrected magnitudes in the HSC filters). They found that blue LSBGs had a brighter mean surface brightness, while galaxies that are large ($R_{\rm eff} > 6''$) and faint ($\bar{\mu}_{\rm eff}(g) > 26\,{\rm mag\,arcsec^{-2}}$) are almost exclusively red.

In Figure 3.5, we present the distribution of our LSBG sample in the $g-i$ vs. $g-r$ color space. We show the color-color diagrams derived from the `SourceExtractor` `MAG_AUTO` quantities (left panel), and the magnitudes derived from the `galfitm` Sérsic model fit (right panel). The color distributions are similar and present signs of bimodality that are slightly more prominent using colors from the Sérsic model fit. Having established the similarity of the color distributions derived from these two fits, in the remainder of this paper, we quote photometric parameters (magnitudes, colors, surface brightness) derived from the `galfitm` model. Thus, photometric and structural parameters (Sérsic index, effective radius) come from the same model fit and can be consistently compared to results in the literature.

We separate the total LSBG sample into red and blue subsamples, according to their $g-i$ color. To do so, we use the following procedure: we fit a two-component Gaussian mixture model (GMM) to the 1D $g-i$ color distribution. The components can be seen in the top panels of Figure 3.5 (dashed gray lines). We find that the two Gaussians intersect at $g-i = 0.60$ (`galfit` case; for comparison using the distribution coming from the `SourceExtractor` quantities the same point is at $g-i = 0.66$). We define a red galaxy sample as galaxies with $g-i \geq 0.60$ (7,671 galaxies), and a blue galaxy sample as galaxies with $g-i < 0.60$ (16,119 galaxies). Note that in the upper-right corner in both panels a "tail" of objects is clearly visible. Inspecting them visually and checking the χ^2 of their `galfit` model fit, we found that most of these are poorly fitted spiral LSBGs.

Our $g-i$ separation threshold is bluer than that of Greco et al. ($g'-i' = 0.64$ in the HSC bandpass).[10] Note that Greco et al. used the median of the distribution to separate the two populations, which was effective as the two populations had similar size. However, the

[10] From a comparison of matched point sources in the HSC SSP Wide and DES Y3 Gold catalogs, we find that the difference between HSC and DES colors is $\Delta(g-i) = 0.013$ for sources with $0.3 < (g'-r') < 0.6$.



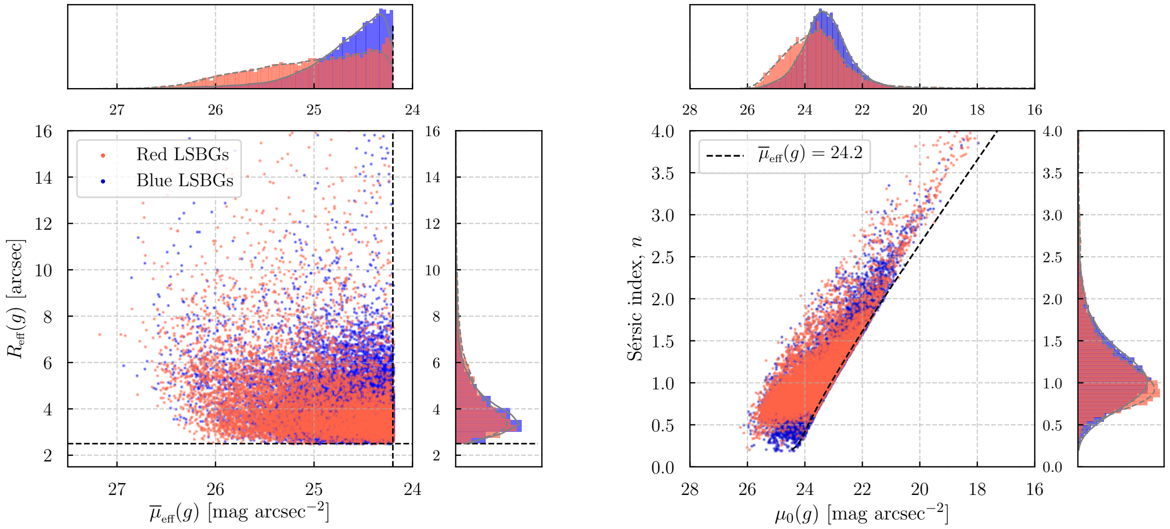
(a)

(b)

Figure 3.6: Examples of (a) blue and (b) red LSBGs in our sample. We randomly selected red galaxies with $g - i$ above the median for the red population ($g - i > 0.76$) and blue galaxies below the median of the blue population ($g - i < 0.40$) to make the color difference more prominent. Each cutout is $30'' \times 30''$ in size.

DES LSBG sample is dominated by blue galaxies, which shifts the median to $(g - i) = 0.60$. The median colors of our red and blue LSBG subsamples are $g - i = 0.76$ and $g - i = 0.40$, respectively.

In Figure 3.6 we show examples of randomly selected blue galaxies with $g - i < 0.40$ (below the median of the blue population) and red galaxies with $g - i > 0.76$ (above the median of the red population). As we can see, the two subsamples show morphological differences. The blue sample is composed primarily of irregular galaxies and galaxies with signs of spiral structure. The red sample consists predominantly of nucleated and non-nucleated spherical and elliptical galaxies.



(a)

(b)

Figure 3.7: (a) Joint distribution of the red and blue LSBGs in the space of effective radius, $R_{\rm eff}$, and mean surface brightness (within the effective radius), $\bar{\mu}_{\rm eff}$, both in the g band. The two populations are defined according to the $g-i$ color criterion described in Section 3.5. The dashed horizontal and vertical lines correspond to the limits of the selection criteria $r_{1/2} > 2.5''$ and $\bar{\mu}_{\rm eff}(g) > 24.2\,{\rm mag\,arcsec^{-2}}$, respectively. Note that although surface brightness is independent of distance, and thus the scatter shown here reflects the intrinsic properties of our sample, much of the scatter in the angular effective radius comes from the fact that the LSBGs lie at different distances. (b) Sérsic index, n, versus central surface brightness, $\mu_0(g)$ [e.g., Graham and Driver, 2005], for the galaxies in our red and blue subsamples. The black dashed line corresponds to our selection criterion, $\bar{\mu}_{\rm eff}(g) = 24.2\,{\rm mag\,arcsec^{-2}}$.

In the left panel of Figure 3.7, we present the joint distribution of our red and blue LSBG samples in the space of effective radius, $R_{\rm eff}(g)$, and mean surface brightness (within the effective radius), $\bar{\mu}_{\rm eff}(g)$. Both populations have sizes ranging from $2.5'' - 16''$. Despite the wide range in angular sizes, most LSBGs in our sample (90%) have radii less than $6''$, with a median of $\sim 4''$. Note that the scatter in angular sizes does not necessarily mean that our galaxies occupy a wide range in physical sizes; much of the scatter comes from the fact that our sample contains galaxies at different distances. For example, in Section 3.7, we show that overdensities in the distribution of LSBGs are associated with galaxy clusters that lie in a range of distances between $\sim 20\,\mathrm{Mpc}$ and $\sim 100\,\mathrm{Mpc}$. For a typical galaxy size of $\sim 1\,\mathrm{kpc}$, that translates into a range of angular sizes between $2''$–$10''$.

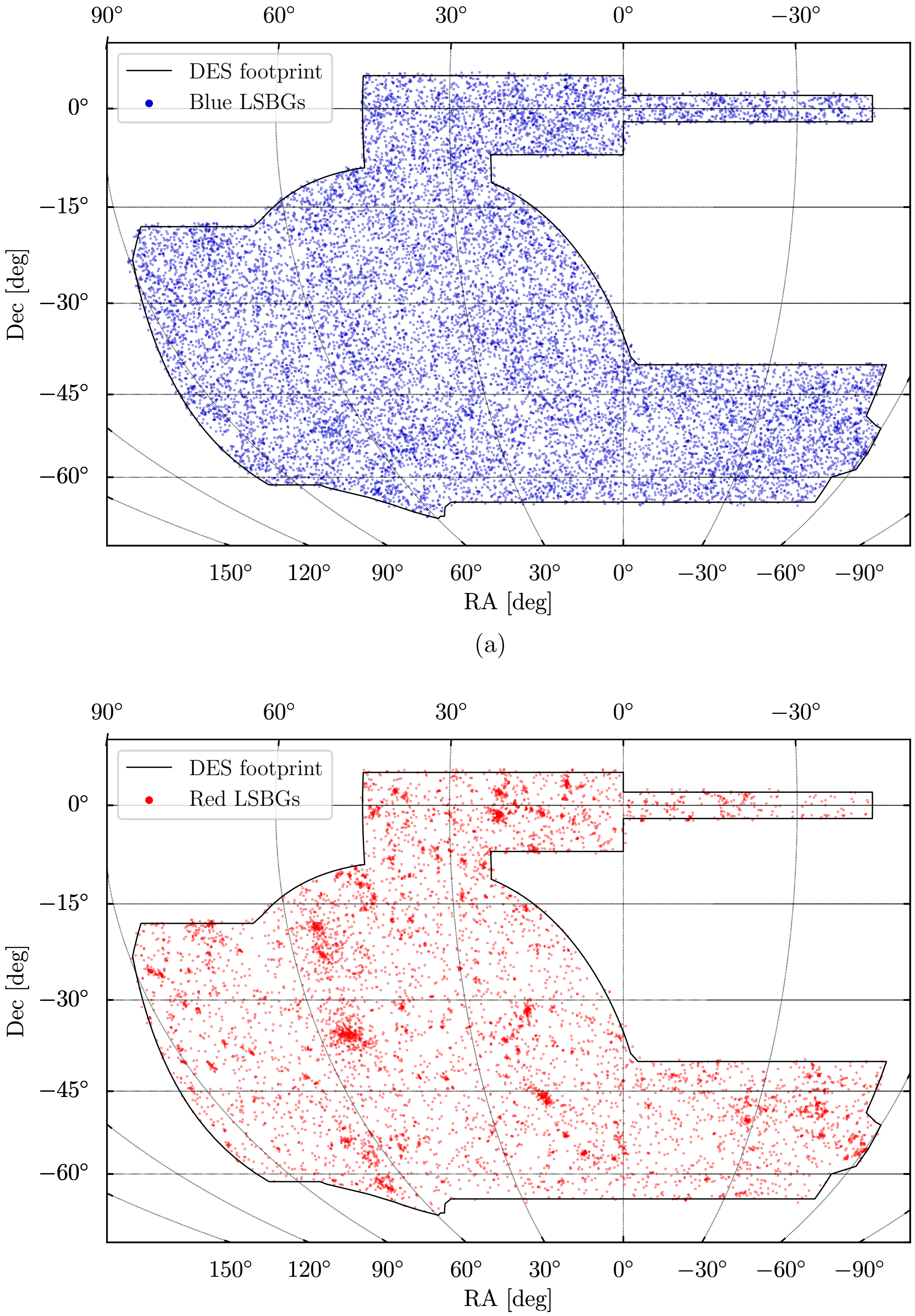
We find that the red galaxy population has a larger tail toward lower surface brightness (larger values of $\bar{\mu}_{\mathrm{eff}}(g)$), while the blue galaxies tend to have higher mean surface brightness. The 50th, 80th, and 90th percentiles in surface brightness are $\bar{\mu}_{\mathrm{eff}}(g) = 24.6, 24.9, 25.2\,\mathrm{mag\,arcsec}^{-2}$ for the red sample and $\bar{\mu}_{\mathrm{eff}}(g) = 24.9, 25.6, 25.9\,\mathrm{mag\,arcsec}^{-2}$ for the blue sample. This result is interesting in the context of early studies that showed no pronounced relationship between color and surface brightness [e.g., Bothun et al., 1997]. However, extrapolating the size–luminosity relationship for red and blue galaxies in SDSS [Shen et al., 2003] suggests that at lower luminosities, red galaxies should be larger than their blue counterparts. A similar result has been shown for the LSBG sample from HSC SSP [Greco et al., 2018].

In the right panel of Figure 3.7 we plot the Sérsic index, n, versus the central surface brightness, $\mu_0(g)$, for our red and blue LSBG samples [e.g., Graham and Driver, 2005]. The distribution in the Sérsic index is similar for two samples, with $0.2 \lesssim n \lesssim 4.0$ and median of $n \sim 1.0$. We do note that the red LSBGs tend to be underrepresented in the regime of small Sérsic index, $n < 0.7$. Unsurprisingly, we find that blue galaxies tend to have higher central surface brightness; however, the difference in central surface brightness between red and blue galaxies is not as striking as the difference in mean surface brightness. The median of the red population is at $\mu_0(g) = 23.6\,\mathrm{mag\,arcsec}^{-2}$, while that of the blue population at $\mu_0(g) = 23.3\,\mathrm{mag\,arcsec}^{-2}$.

3.6 Clustering of LSBGs

Greco et al. tentatively suggested that the spatial distribution of LSBGs in the HSC SSP may be correlated with low-redshift galaxies from the NASA-Sloan Atlas[11]. However, due to the relatively small area covered by their HSC SSP data set ($\sim 200\,\mathrm{deg}^2$), they were unable

[11] `http://nsatlas.org/`



(a)

(b)

Figure 3.8: Sky positions of (a) blue LSBGs ($g-i < 0.60$; 7,671 galaxies) and (b) red LSBGs ($g-i \geq 0.60$; 16,119 galaxies) within the DES footprint. The distribution of the red LSBGs is more strongly clustered than that of the blue LSBGs.

to make any firm statistical statement about possible correlations. Our DES Y3 LSBG catalog covers a contiguous region ~ 25 times larger than that of Greco et al., allowing us to perform a detailed exploration of the spatial distribution of LSBGs. In particular, we are able to *separately* explore the clustering of our red and blue LSBG subsamples (as defined in Section 3.5). In Figure 3.8, we present the spatial distribution of blue and red LSBGs over the DES footprint. We find a stark contrast in the spatial distribution of these two LSBG subpopulations: red LSBGs are highly clustered, while blue galaxies are more uniformly distributed.

To quantify the clustering of our LSBG sample and the red/blue subsamples, we calculate the angular two-point autocorrelation function of LSBGs, $w(\theta)$ [e.g., Peebles, 1980, Connolly et al., 2002]. We use `treecorr` [Jarvis, 2015][12] to calculate $w(\theta)$ using the estimator of Landy and Szalay [1993] with a random sample of points drawn from the DES Y3 Gold footprint mask derived from the DES imaging data using `mangle` [e.g., Swanson et al., 2008]. In Figure 3.9 we plot $w(\theta)$ for the full LSBG sample, as well as the red and blue subsamples (gray, red, and blue curves, respectively). We estimate the errors on $w(\theta)$ using jackknife resampling [e.g., Efron and Gong, 1983]. As expected from Figure 3.8, we find that the amplitude of the autocorrelation function of red LSBGs is more than an order of magnitude larger than that of blue LSBGs at angular scales $\theta \lesssim 3^{\circ}$.

The differences in clustering amplitude between red and blue galaxies has been studied extensively in spectroscopic surveys [e.g., Zehavi et al., 2002, 2005, 2011, Law-Smith and Eisenstein, 2017]. In particular, it has been noted that there is a strong difference in the amplitude and shape of the autocorrelation function of intrinsically faint red galaxies relative to brighter and/or bluer galaxies [e.g., Norberg et al., 2002, Hogg et al., 2003, Zehavi et al., 2005, Swanson et al., 2008, Cresswell and Percival, 2009, Zehavi et al., 2011]. We find the same pronounced difference in the amplitude and shape of $w(\theta)$ for red LSBGs relative to

[12] `https://github.com/rmjarvis/TreeCorr`

the blue LSBG subsample and the power-law behavior observed in higher-surface brightness galaxies, $w(\theta) \propto \theta^{-0.7}$ [e.g., Connolly et al., 2002, Maller et al., 2005, Zehavi et al., 2011, Wang et al., 2013]. The observed shape of the angular autocorrelation function of red LSBGs (which is also manifested in the total LSBG population) can be produced if the LSBG sample has a preferred scale for clustering. We find that we can reproduce the shape of the LSBG $w(\theta)$ by selectively enhancing overdense regions at scales of a few degrees.

Previous theoretical modeling has suggested that the strong clustering of faint red galaxies is the result of these galaxies being dominantly satellites of massive dark matter halos [Berlind et al., 2005, Wang et al., 2009, Zehavi et al., 2011]. Zehavi et al. [2011] note a strong inflection in the clustering of faint red galaxies ($M_r < -19$) at a scale of $\sim 3h^{-1}$ Mpc. By mapping this physical scale to the enhanced clustering observed in the red LSBG sample at angular scales of $\theta \lesssim 3^{\circ}$, we derive an estimated distance of ~ 40 Mpc for the clustered red LSBG sample.

To assess whether the difference in clustering observed between red and blue LSBGs could be attributed solely to a difference in stellar mass, we subdivide our red and blue LSBG samples into samples of faint red galaxies ($21 < g < 22$) and bright blue galaxies ($19.5 < g < 20.5$). Blue galaxies generally have a higher luminosity at a given stellar mass than red galaxies [e.g., Conroy, 2013]. Following Greco et al. [2018], we find that the $(g - i)$ colors of our blue and red LSBGs are well represented by a simple stellar population from Marigo et al. [2017] with $[\mathrm{Fe/H}] = -0.4$ and an age of 1 Gyr and 4 Gyr, respectively. We find that these populations differ in total absolute g-band magnitude by $\Delta(M_g) \sim 1.5$. We also find that the angular autocorrelation functions of the bright red and faint blue samples do not differ significantly from the total red and blue LSBG samples, respectively. This suggests that the difference in clustering shape and amplitude cannot be attributed to a difference in stellar mass alone.

Some authors have argued that observations support a decrease in the number of LSBGs close to the cores of galaxy clusters [e.g., van der Burg et al., 2016, Wittmann et al., 2017]. Such a suppression could reduce the clustering power on small scales, leading to a flattening in the autocorrelation function. However, rigorously testing for a suppression in the abundance of LSBGs in dense regions would require end-to-end simulations with injected LSBGs to characterize the DES detection efficiency as a function of local galaxy density. [e.g., using a tool like `Balrog`; Suchyta et al., 2016, Everett et al., in prep.]. We leave a detailed characterization of the DES selection function for LSBGs to future work.

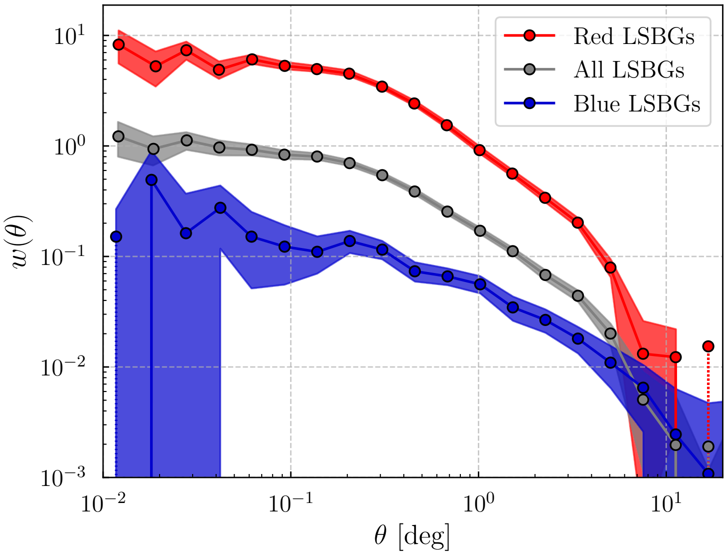


Figure 3.9: The angular autocorrelation function of the total LSBG sample (dark gray line), and the red and blue LSBG subsamples (red and blue lines, accordingly). The errors were calculated using the jackknife method. The correlation function of the red LSBGs has a higher amplitude than that of the blue LSBGs across all angular scales.

3.6.1 Comparison to other galaxy samples

We compare the clustering properties of our LSBG sample to two other galaxy samples: a catalog of HSBGs extracted from the DES Y3 Gold catalog, and an external sample of low-redshift galaxies from the 2MASS Photometric Redshift (2MPZ) catalog. Our goals here are twofold: (1) to compare the clustering of DES galaxies as a function of surface brightness

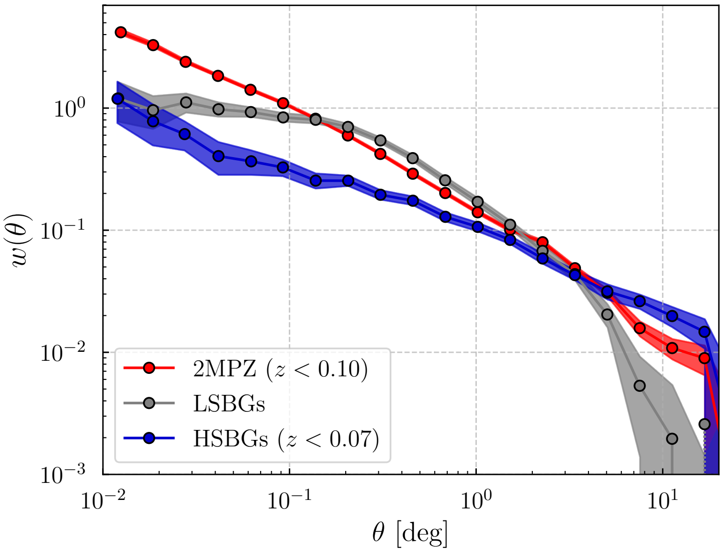


Figure 3.10: The angular autocorrelation function of all LSBGs (gray line), the HSBG sample extracted from the DES data (blue line) and the 2MPZ sample (red line). We see that the LSBG exhibits a turnover at lower angular scales that is not observed either at the HSBG or 2MPZ samples.

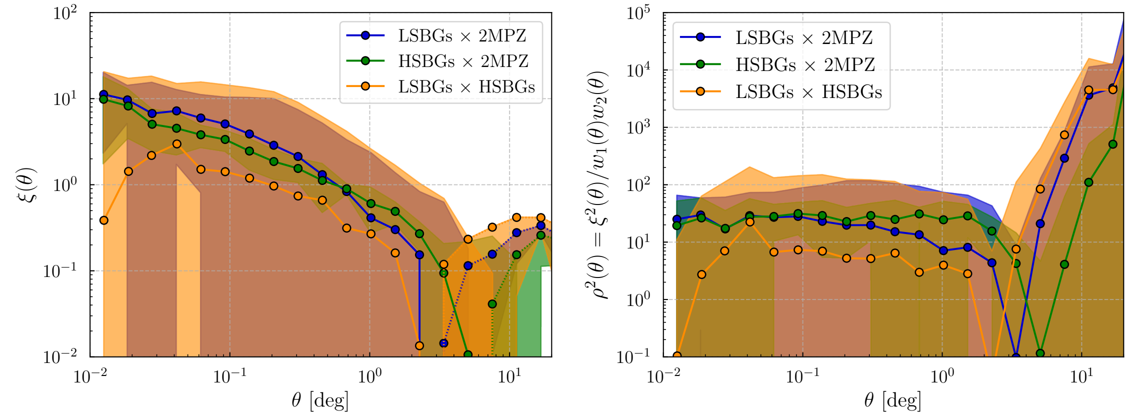
and (2) to use the superior redshifts of the 2MPZ sample to approximately determine the redshift distribution of our LSBGs.

We construct an HSBG sample from the DES Y3 Gold catalog by applying the same star–galaxy separation, color, and ellipticity cuts described in Section 3.3.1 and summarized in Appendix 3.10. We do not apply any angular size restriction on the HSBG sample, but rather we require that the HSBGs have mean surface brightness $20.0 < \bar{\mu}_{\rm eff}(g) < 22.0\,{\rm mag\,arcsec}^{-2}$. Ideally, we would be able to compare the clustering of LSBGs and HSBGs with the same stellar mass and redshift distributions. Since the redshift distribution of the LSBGs is unknown, we scanned over a range of redshifts for the HSBGs using redshifts estimated trough the Directional Neighbourhood Fitting algorithm [DNF; De Vicente et al., 2016] derived from the DES multi-object fitting (MOF) photometry.

For each redshift-selected sample of HSBGs, we select a random subset of galaxies that produces the same distribution in g-band apparent magnitude as our LSBG sample in the range $18 < g < 22$ (see Appendix 3.11). We compare the clustering amplitude of the LSBG and HSBG samples, and find that the best match is achieved for a photometric redshift cut of $z < 0.07$. However, even for this optimal selection, we find less clustering in the HSBG sample than the LSBG sample in the intermediate angular range $\theta \sim 0.1^\circ - -4^\circ$ (Figure 3.10). We note that it is likely that the HSBG sample is contaminated by distant galaxies due to the large photometric redshift uncertainty of DES, which is $\sigma_{68}(z) \sim 0.1$ overall and is known to have a large outlier fraction at low redshift [e.g., Hoyle et al., 2018].

We perform a similar analysis for the 2MPZ catalog [Bilicki et al., 2014], an optical-IR all-sky photometric redshift catalog based on SuperCOSMOS, 2MASS, and WISE extending to $z \sim 0.3$ (peaking at $z \sim 0.07$). We select this catalog due to its uniform sky coverage and accurate photometric redshifts ($\sigma_z = 0.015$). We note that 2MPZ has a very different selection function than DES, as it requires detection in the IR bands. By matching 2MPZ galaxies with galaxies in the DES Y3 Gold catalog, we retrieved information about DES-measured magnitude and surface-brightness distribution of 2MPZ galaxies. We find that the DES-measured mean surface brightness for matched 2MPZ galaxies is significantly brighter ($19.0 < \bar{\mu}_{\rm eff}(g) < 23.0\,{\rm mag\,arcsec^{-2}}$) than the LSBG sample. The g-band magnitude (`MAG_AUTO`) of the 2MPZ sample lies in the range $14.0 < g < 18.5$, while the LSBG sample range is $18 < g < 22$ (see Appendix 3.11, Figure 3.19). We thus expect the 2MPZ sample to consist of brighter, higher stellar mass galaxies compared to the LSBG sample. As before, we identified a redshift cut that resulted in an angular autocorrelation function that is best-matched to that of the LSBGs. In the case of 2MPZ galaxies, we find that this is achieved with a redshift cut of $z < 0.10$.

In Figure 3.10 we plot the angular autocorrelation function, $w(\theta)$, of the LSBGs (gray line), the DES HSBGs with $z < 0.07$ (blue line), and the 2MPZ catalog with $z < 0.10$ (red line). We find that both the DES HSBG and 2MPZ samples have lower clustering amplitude than the LSBG sample at intermediate angular scales ($0.1^\circ \lesssim \theta \lesssim 4^\circ$). Overall, we find that



(a) (b)

Figure 3.11: (a) The cross-correlation function, $\xi(\theta)$, between (i) the DES LSBG and HSBG samples (orange line), (ii) the LSBG and 2MPZ samples (blue line), and (iii) the DES HSBG and 2MPZ samples (green line). (b) The square of the cross-correlation coefficient between the same samples as in panel (a), in order to cancel out the contribution of the different galaxy biases and compare the different cross-correlation levels. In both panels, the shaded regions correspond to the errors in the estimated cross-correlations.

the amplitude of the angular correlation function of LSBGs is better matched by the 2MPZ catalog than the DES HSBG catalog.

3.6.2 Cross-correlation between galaxy samples

The previous autocorrelation analysis compares the clustering properties of the LSBG, HSBG and 2MPZ catalogs individually. However, it does not indicate whether these galaxy samples probe the underlying matter density field in a similar way, i.e., whether the peaks and troughs in their distributions coincide on a statistical basis. Galaxies are known to be biased traces of the underlying matter density field. For large angular scales, the two fields are connected by a (linear) galaxy bias factor, b_g, defined as $\delta_g(z) \equiv b_g(z)\delta_m(z)$, where δ refers to the overdensity field and the subscripts g and m refer to galaxies and matter, respectively. In general, these are functions of redshift, while the bias factor is different for different galaxy samples. The galaxy angular autocorrelation function can be defined as

$w(\theta) = \langle \delta_g(\hat{\mathbf{n}})\delta_g(\hat{\mathbf{n}}+\theta)\rangle = b_g^2 \langle \delta_m(\hat{\mathbf{n}})\delta_m(\hat{\mathbf{n}}+\theta)\rangle$, where $\hat{\mathbf{n}}$ is the direction in the sky.

To address whether the galaxy samples studied in the previous section trace the matter density field in a similar way, we calculate the cross-correlation function, $\xi(\theta)$, between the LSBG and HSBG samples, the LSBG and the 2MPZ samples, and the HSBG and 2MPZ samples (left panel of Figure 3.11). The cross-correlation between two galaxy samples (labeled 1 and 2) is given by $\xi_{12}(\theta) = \langle \delta_{g,1}(\hat{\mathbf{n}})\delta_{g,2}(\hat{\mathbf{n}}+\theta)\rangle = b_{g,1}b_{g,2}\langle \delta_m(\hat{\mathbf{n}})\delta_m(\hat{\mathbf{n}}+\theta)\rangle$. We define the cross-correlation coefficient between the two samples as

$$\rho_{12}(\theta) = \frac{\xi_{12}(\theta)}{\sqrt{w_1(\theta)w_2(\theta)}}, \tag{3.4}$$

where $w_{1,2}(\theta)$ are the autocorrelation functions of the individual samples. In this case, we can cancel the corresponding bias factors present in the different samples, and we can compare the correlations between the matter fields probed by the two samples. We plot the (square of the) cross-correlation coefficient between the same samples as those described above in the right panel of Figure 3.11.

Although the uncertainties are large, we find that the 2MPZ×LSBG sample exhibits a larger cross-correlation signal than the LSBG×HSBG. This likely reflects the better agreement between the redshift distributions of the LSBG and 2MPZ samples, which is expected due to the superior redshift information provided by the 2MPZ. The stronger cross-correlation signal motivates our use of the 2MPZ sample when constructing radial profiles of HSBGs associated with the prominent peaks in the LSBG distribution.

3.7 Associations with Galaxy Clusters and Groups

In the previous section, we described a statistical study of the clustering of LSBGs, which can also be demonstrated visually when plotting the positions of LSBGs (Figure 3.8). In this section, we instead focus on identifying the most prominent spatial overdensities of

LSBGs and associating them with known galaxy clusters, galaxy groups, and individual bright galaxies. Associating peaks in the LSBG distribution to external catalogs provides useful information, such as:

1. Associating a peak in the LSBG distribution with a galaxy system at a known distance allows us to estimate the distances to the LSBGs (assuming a physical association between the LSBGs and reference object). Distances allow us to estimate the intrinsic properties of the LSBGs, such as physical size and luminosity.

2. Defining a sample of likely LSBG cluster members allows us to compare the properties of the LSBGs in cluster environments to those in the field. Such comparisons can be useful for testing models of LSBG formation and evolution. For example, we can compare the radial distributions of LSBG and HSBG cluster members to test for observable signatures of environmental effects that may be responsible for the formation of LSBGs.

3. Peaks in the LSBG density that are not associated to known clusters or groups can be potentially interesting, indicating different clustering patterns for LSBGs and HSBGs.

Table 3.2: Characteristics of the 10 most prominent density peaks and their associations

Peak Number	**(RA,Dec)$_{\text{peak}}$ (deg,deg)**	**Best Association**	**(RA,Dec)$_{\text{assoc}}$ (deg,deg)**	**Redshift** z	**Distance Mpc**	N
1	(21.5012, -1.4286)	Abell 194	(21.4200, -1.4072)	0.018	75.07 ± 5.26	68
2	(54.9388, -18.4712)	RXC J0340.1-1835	(55.0475, -18.5875)	0.0057	23.41 ± 1.64	48
3	(9.8887, 3.1829)	NGC 199	(9.8882, 3.1385)	0.0153	62.81 ± 4.41	46
4	(17.4972, -45.9398)	Abell 2877	(17.6017, -45.9228)	0.0247	106.61 ± 7.45	41
5	(18.4983, -31.7043)	Abell S141	(18.4758, -31.7519)	0.020	84.80 ± 5.94	42
6	(53.9377, -35.3133)	Fornax (Abell S373)	(54.6162, -35.4483)	0.0046	18.97 ± 1.33	32
7	(16.8965, -46.7418)	Abell 2870	(16.9299, -46.9165)	0.0237	102.03 ± 3.89	36
8	(55.3393, -35.5138)	Fornax (Abell S373)	(54.6162, -35.4483)	0.0046	18.97 ± 1.33	28
7	(21.3014, 1.7794)	RXC J0125.5+0145	(21.3746, 1.7627)	0.01739	72.32 ± 5.10	28
10	(9.8888, -55.9649)	Abell 2806	(10.0270, -56.1167)	0.0277	120.23 ± 8.42	32

We use kernel density estimation (KDE) to estimate the projected density of our full LSBGs sample. We apply a Gaussian smoothing kernel with a bandwidth of 0.3°, using the haversine distance metric to account for the cosine dependence on declination [Pedregosa et al., 2011b]. The kernel bandwidth was selected to be similar to the characteristic angular scale of the overdensities present in Figure 3.8. This kernel size is further motivated by the

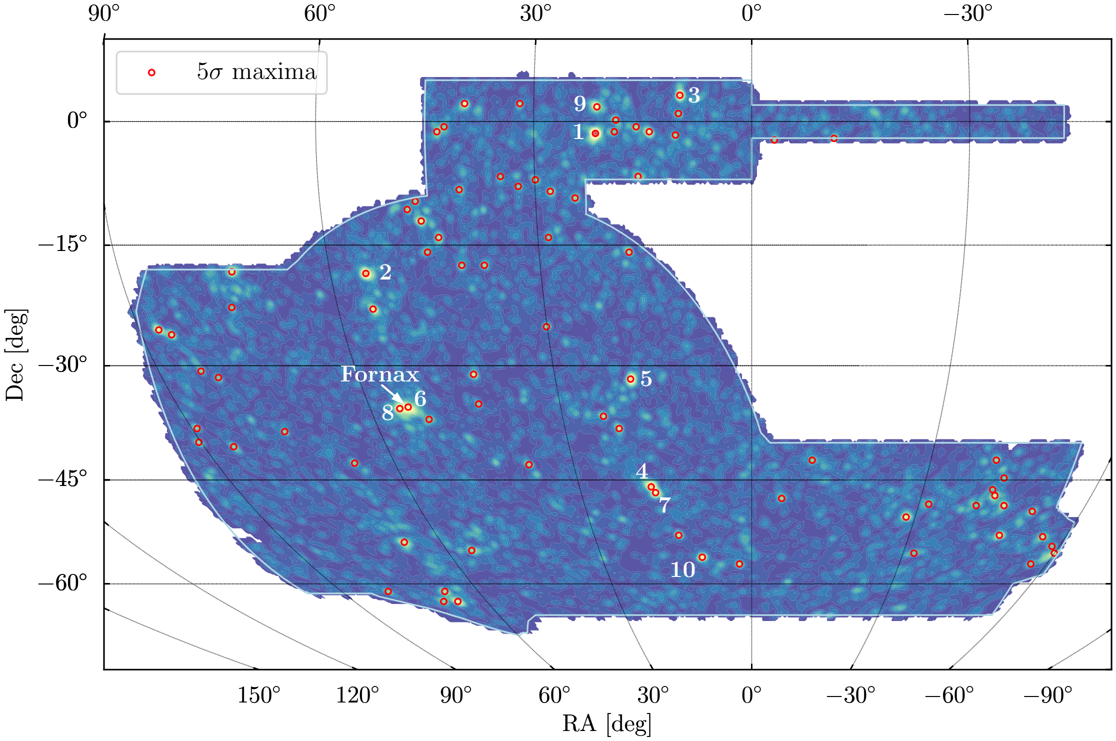


Figure 3.12: KDE map of the distribution of our LSBG sample. Blue regions denote areas of low density, while regions of high density are indicated in yellow/red. Open red circles indicate the positions of the 82 prominent density peaks identified as described in Section 3.7. We have labeled the 10 most prominent peaks, which are summarized in Table 3.2.

radial profiles of LSBGs around peaks (see Figure 3.13), where it is seen that the typical scale of cluster cores is of the order of $\sim 0.5\,\mathrm{Mpc}$. The median distance of clusters associated to our sample is $\sim 80\,\mathrm{Mpc}$, which results into a typical angular size of $\sim 0.35°$. For more distant clusters, that typical angular size is smaller ($\sim 0.28°$ at a distance of $100\,\mathrm{Mpc}$), while for the closest clusters, the typical angular size is significantly larger (e.g., for Fornax at a distance of $\sim 19\,\mathrm{Mpc}$ this scale is $1.5°$). In fact, a bandwidth of $0.3°$ resolves the Fornax cluster into two peaks.

The resulting KDE map is presented in Figure 3.12, with blue regions representing areas of lower density and yellow/red regions representing areas of higher density. To detect outliers in this map, we perform an iterative sigma-clipping procedure where at each step, values that exceed the median by 5σ or more are rejected. We find the local maxima in the regions of the KDE map that are above the 5σ threshold value returned from sigma-clipping. We locate 82 peaks passing our criteria, which are indicated with red open circles in Figure 3.12. We furthermore number the 10 most prominent of them (as defined by their KDE value) and present their coordinates in Table 3.2. In the seventh column of that table, we also present the number of LSBGs within 0.5 degrees from the center of each peak. The complete catalog can be found in a machine-readable form in the supplemental material `https://des.ncsa.illinois.edu/releases/other/y3-lsbg`

Next, we cross-match our list of high-density LSBG peaks with known overdensities in the low-redshift universe. Specifically, we cross-match against:

1. The Abell catalog of rich clusters [southern survey, Abell et al., 1989].

2. The ROSAT-ESO Flux Limited X-ray (REFLEX) Galaxy cluster survey [Böhringer et al., 2004].

3. A catalog of galaxy groups built from the sample of the 2MASS Redshift Survey [Tully, 2015]. We keep only those groups that have more than five members.

4. Bright galaxies from the revised New General Catalogue [Sulentic and Tifft, 1999].

For each peak in the LSBG distribution, we overplotted the distribution of LSBGs and external catalog objects in a region ±0.5° from the nominal center of the peak. To identify associations (if any), we selected the object from the external catalogs that is closest to the center of the LSBG peak, giving priority to objects according to ordering listed above. For example, if an LSBG peak is matched to both an NGC galaxy and an Abell cluster, we select the Abell cluster as the association. From the 82 peaks, we find that 32 are associated with an Abell cluster, 11 with a REFLEX cluster, 10 with a 2MASS group, 16 with an NGC galaxy, while 13 peaks have no association assigned by our criteria. We used the DES Sky Viewer tool to visually inspect the regions around the 13 LSBG peaks that were not associated with objects in our external catalogs. In seven cases, we identified nearby bright galaxies/galaxy clusters that were not included in the external catalogs we used for the matching. Interestingly, in six cases we did not find an obvious nearby galaxy cluster, galaxy group, or bright nearby galaxy. As an interesting case, we mention a peak at (RA,DEC)$\sim (-50.978^\circ, -49.348^\circ)$ with 18 LSBGs in a 0.5° area around it. We leave the more detailed study of these systems for future work.

In Table 3.2 we present the coordinates of the ten most prominent LSBG overdensities and their best associations, along with the coordinates, redshifts, and distances of these associations (retrieved from the NASA Extragalactic Database)[13]. We also report the number of LSBGs within 0.5° from the center of each peak. Note that two peaks are both associated with the Fornax cluster (Abell S373). The full table of associations can be found in the supplemental material, where we provide an additional column characterizing the quality of association: I (very good), II (good), to III (not so good). The quality of the association was determined based on the projected, angular distance of the association from the peak and the presence (or absence) of other potential associations in the vicinity of the peak. Our classification is qualitative, though, and is just a guide for follow-up research. For the cases where we did not find an association using any of the catalogs mentioned above, we visually inspected the region around the peak using the DES Sky Viewer. If there was not any visible high-surface-brightness counterpart around, we indicated quality = I, otherwise (visible clusters of bright galaxies) we indicated quality = III.

By assuming a physical association between these LSBG overdensities and the matched external systems, we can use the known distances of the external systems to estimate the distance to the associated LSBGs. This information is otherwise absent due to our inability to accurately estimate the photometric redshift for these galaxies from the DES data alone. In

[13] https://ned.ipac.caltech.edu/

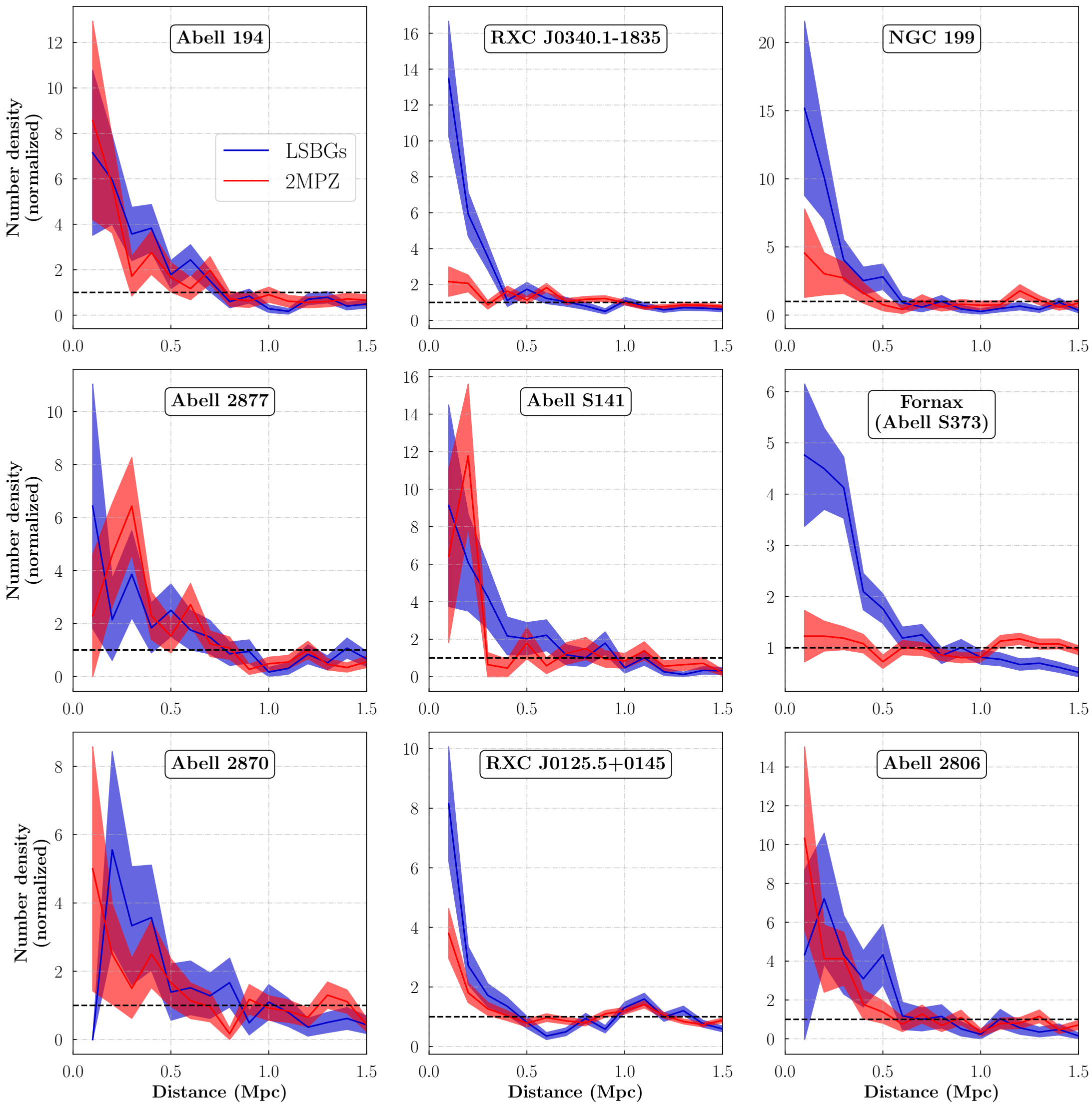


Figure 3.13: Normalized radial profiles of the distribution of LSB galaxies (blue) and galaxies from the 2MPZ catalog (red) around the associations of the most prominent LSBG overdensity peaks, presented in Table 3.2. We have assumed that all galaxies that are within a radius that corresponds to a physical scale of 1.5 Mpc at the distance of the association belong to that association. The normalization constant corresponds to the mean number density of galaxies within the 1.5 Mpc radius.

the remainder of this section, we will use distance information from the nine most prominent associations to (i) study the radial distribution of LSBGs around clusters and (ii) derive the size–luminosity relation for associated LSBGs.

3.7.1 Radial Profiles

Comparing the distribution of LSBGs and HSBGs in dense environments may help illuminate the processes governing the formation and evolution of LSBGs. In Figure 3.13 we plot the number density of LSBGs and 2MPZ galaxies with redshift $z < 0.10$ around the nine most prominent associated systems (clusters and NGC galaxies; Table 3.2). For each of these nine associations, we select all LSBGs and 2MPZ galaxies that reside within an angle corresponding to 1.5 Mpc at the distance of each associated object. We calculate the radial profiles of LSBGs and 2MPZ galaxies in fifteen annuli of width 0.1 Mpc. In order to compare the LSBGs and 2MPZ galaxies on the same scale, we normalize the number densities to the mean number density of galaxies in each sample within the 1.5 Mpc region—i.e., a flat line with unit amplitude indicates a homogeneous distribution of galaxies within the 1.5 Mpc region. We estimate the uncertainty on our radial profile by combining the Poisson uncertainties on the measured number of galaxies per annulus and the total number of galaxies in the 1 Mpc region.

In all cases, we find that the LSBG distribution is peaked within 0.5 Mpc and flattens at distances $\gtrsim 1$ Mpc. We find that the normalized number density of LSBGs peaks at similar amplitudes for most systems, with the most peaked overdensity found around the lenticular galaxy NGC 199. This may be expected given that this association represents the dwarf satellite population of a single central bright galaxy. We find three cases where the normalized radial distributions of the LSBG and 2MPZ samples appear quite different. RXC J0340.1−1835 and Fornax are at significantly lower redshift than the other systems, $z = 0.0057$ and $z = 0.0046$, respectively (the next closest associated system is NGC 1200 at $z = 0.013$.) The 2MPZ catalog includes just a few objects with such low redshifts; there are only 24 objects with $z < 0.005$ and 42 objects with $z < 0.006$. Thus, in these two cases it is likely that the 2MPZ sample consists of background galaxies. The third case where the distribution of 2MPZ and LSBG galaxies differ is around NGC 199. Again, the LSBGs are much more peaked than the 2MPZ sample, suggesting that the observed LSBG overdensity is caused by dwarf galaxies surrounding a single central host. Despite the small sample size, we can say qualitatively that the radial distribution of LSBGs and 2MPZ galaxies appear to largely agree. We use the Kolmogorov–Smirnov test to quantitatively evaluate the similarity of the radial distributions of LSBGs and 2MPZ galaxies surrounding these systems. We calculate the p-values for the null hypothesis that the two galaxy samples are drawn from the same underlying distribution. We find that for RXC J0340.1−1835 and Fornax, $p \ll 0.01$ (thus strongly rejecting the null hypothesis), $p = 0.015$ for NGC 199 (making the null hypothesis unlikely), while for all the other systems $p > 0.1$.

3.7.2 Size–Luminosity Relation

Distance information from our external catalog systems allows us to calculate the physical properties of associated LSBGs. For the nine most prominent peaks in the LSGB distribution, we assume that all LSBGs that reside within a projected distance of 0.5 Mpc are associated to these systems and reside at the same distance. Using this distance, we can estimate the physical effective radii (in pc) and absolute magnitudes of these LSBGs.

In Figure 3.14, we present the size–luminosity relationship for the LSBGs around these nine peaks, based on the physical effective radius, $R_{\rm eff}(g)$, and the absolute magnitude in the g band, M_g. We see that the number of LSBGs associated with each system varies significantly; the smallest number of LSBGs (17) is associated with Abell 2870, while the largest number of LSBGs (175) are associated to Fornax. In Figure 3.14 we also indicate the physical scale corresponding to the angular selection criterion, $R_{\rm eff}(g) > 2.5''$, at the distance of the associated system (dashed black line). Since Fornax is the closest cluster, this angular selection criterion corresponds to the smallest physical size (~ 230 pc), resulting in more faint galaxies passing the selection. Similarly, RXC J0340.1−1835 is also a nearby cluster and has a large number of LSBGs (102). We also show lines of constant mean surface

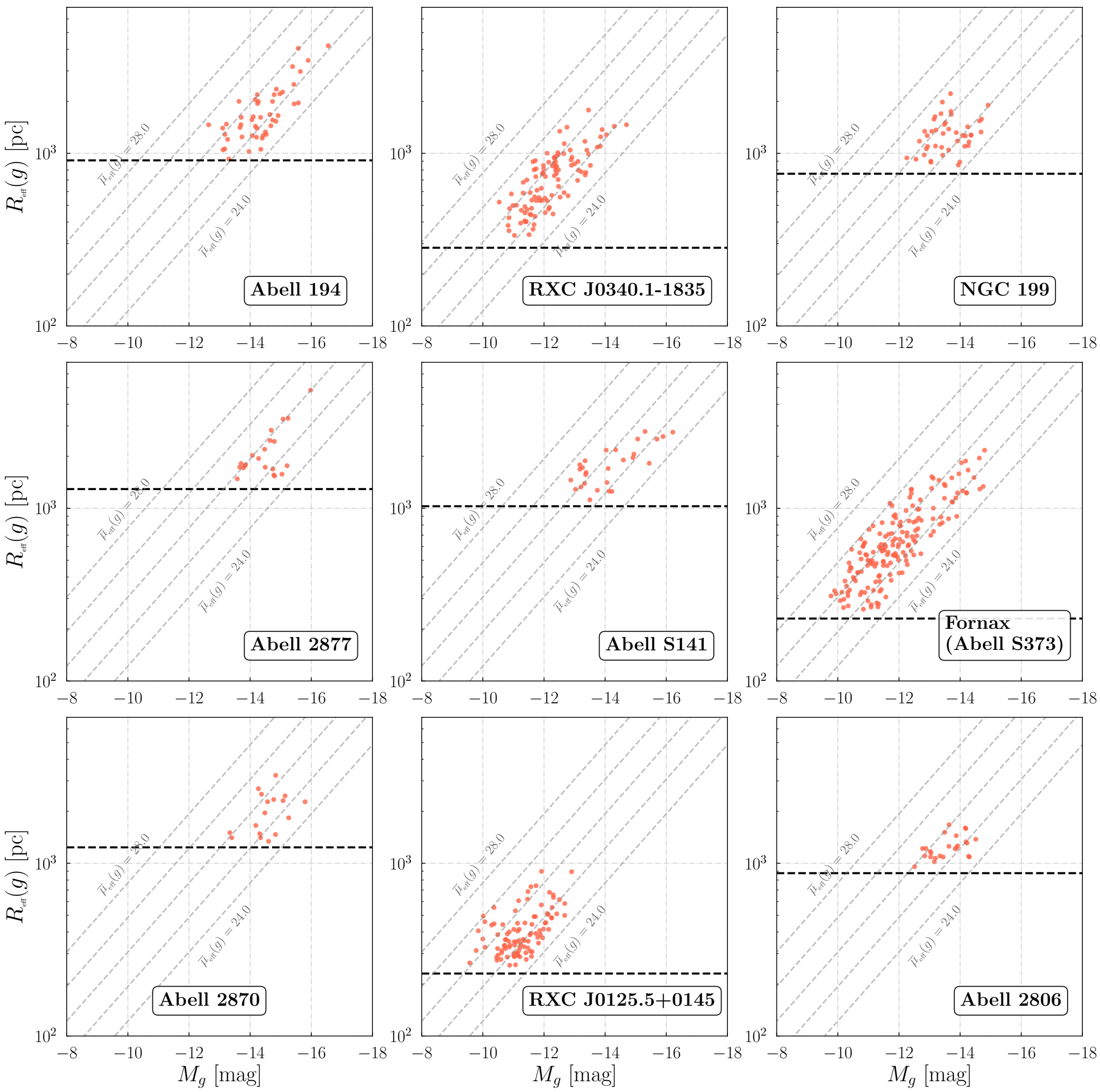


Figure 3.14: Size–luminosity relation for LSBGs around the associations of the most prominent overdensity peaks, presented in Table 3.2. We have assumed that all LSBGs within an angle corresponding to a physical radius of 0.5 Mpc at the distance of the association belong to it. With the dashed horizontal lines, we show the physical scale corresponding to the radius cut $r_{1/2}(g) > 2.5''$ at the distance of the cluster. We also show (dashed, diagonal gray lines) the lines of constant mean surface-brightness.

brightness. The bright-end limit is largely set by the requirement $\bar{\mu}_{\rm eff}(g) > 24.2\,{\rm mag\,arcsec^{-2}}$ used to produce our catalog. Only two associated galaxies have surface brightness $\bar{\mu}_{\rm eff}(g) > 27.0\,{\rm mag\,arcsec^{-2}}$.

In Figure 3.15, we combine the observations of LSBGs from the nine clusters in a single

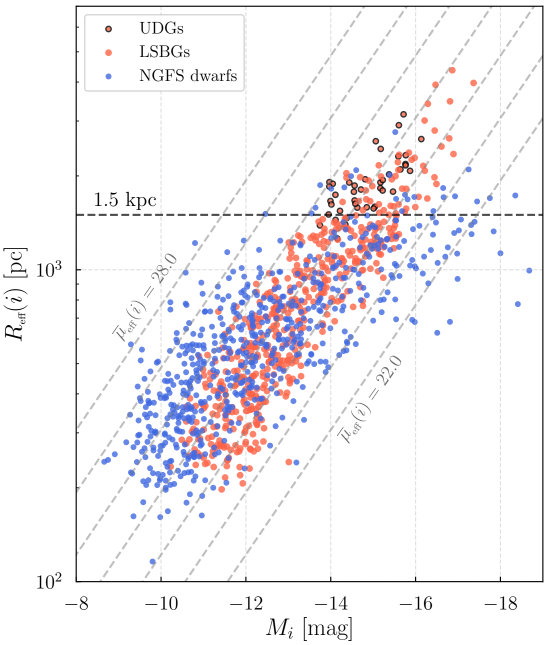


Figure 3.15: Size–luminosity relation of LSBGs around the nine most prominent overdensities (red points) in the i band. The sample consists of 555 galaxies. For comparison, we overplot the dwarf galaxies found around Fornax in the NGFS survey [Eigenthaler et al., 2018, Ordenes-Briceño et al., 2018]. 41 galaxies in our sample have effective radii exceeding 1.5 kpc in the g band (black circles) and central surface brightness $\mu_0(g) > 24.0\,\mathrm{mag\,arcsec}^{-2}$, which is a conventional definition for ultra-diffuse galaxies [UDG; van Dokkum et al., 2015b].

size–luminosity plot. We compare the distribution of our sample to that of the dwarf galaxies discovered in the NGFS survey, described in Section 3.4. Since the NGFS only provides magnitudes and effective radii in the i band [Eigenthaler et al., 2018, Ordenes-Briceño et al., 2018], we choose to plot against the i-band quantities of our sample. We see that the two samples occupy a similar region in the size–luminosity parameter space, with the NGFS sample spanning a larger range of absolute magnitudes. The NGFS extends to fainter absolute magnitudes due to their deeper imaging data, while the lack of an explicit surface-brightness cut extends their sample to brighter magnitudes.

Recently, much attention has been paid to the class of ultra-diffuse galaxies (UDGs), which have been conventionally defined as galaxies with central surface brightness $\mu_0(g) >$ 24.0 and effective radius $R_{\rm eff}(g) > 1.5\,\mathrm{kpc}$ [e.g., van Dokkum et al., 2015b]. The LSBGs in our associated sample span a wide range of physical sizes, from $0.26\,\mathrm{kpc} \lesssim R_{\mathrm{eff}}(g) \lesssim 4.83\,\mathrm{kpc}$, with a median of $R_{\mathrm{eff}}(g) = 0.8\,\mathrm{kpc}$ (the i-band values presented in Figure 3.15 are $0.20\,\mathrm{kpc} \lesssim R_{\mathrm{eff}}(i) \lesssim 4.36\,\mathrm{kpc}$ with a median of $R_{\mathrm{eff}}(i) = 0.75\,\mathrm{kpc}$). The lower limit is largely set by our angular size selection criterion, translated to a physical size for the nearest cluster (Fornax). We find 41 galaxies have size $R_{\mathrm{eff}}(g) > 1.5\,\mathrm{kpc}$ and surface brightness $\mu_0(g) > 24.0\,\mathrm{mag\,arcsec^{-2}}$, thus satisfying the conventional UDG definition. We note again that our angular size selection requires distant galaxies to have larger physical sizes.

The sample covers a wide range of absolute g-band magnitude, $-9.8 \gtrsim M_g \gtrsim -16.5$, with a median of $M_g \sim -12.4$. We see that the galaxies in the sample discussed here span the same range in mean surface brightness ($24.2 \lesssim \bar{\mu}_{\mathrm{eff}}(g) \lesssim 27.0\,\mathrm{mag\,arcsec^{-2}}$), regardless of their sizes: both small and large galaxies populate the range of surface brightnesses. Thus, UDGs seem to be a natural continuation of the LSBG population in the regime of large size and low surface brightness, and not a distinct population that is well separated in the size–luminosity space from other LSBGs [a similar conclusion was drawn by Conselice, 2018].

3.8 Summary and Conclusions

In this paper, we have selected and analyzed 23,790 extended, LSBGs from the first three years of DES imaging data. Our sample selection pipeline consists of the following steps:

1. We selected objects from the DES Y3 Gold catalog based on `SourceExtractor` parameters. The most important selections were based on the half-light radius, $r_{1/2} > 2.5''$ and mean surface brightness, $\bar{\mu}_{\mathrm{eff}}(g) > 24.2\,\mathrm{mag\,arcsec^{-2}}$. The selection criteria are summarized in Appendix 3.10.

2. We applied an SVM classifier tuned to reduce the incidents of false negatives (LSBGs classified as non-LSBGs). This reduced the number of false-positive candidates by an order of magnitude.

3. A visual inspection that eliminated the remaining false positives to produce a high-purity sample of LSBGs.

4. We fit each galaxy with a single-component Sérsic profile, and we made a final selection based on the derived size and surface brightness.

We divided the total LSBG sample into two subsamples according to their $g - i$ color. We study the photometric, structural and spatial clustering properties of the red ($g - i \geq 0.60$) and blue ($g - i < 0.60$) subsamples. Our main findings are the following:

1. The distributions in angular size (effective radius) are similar for the two subsamples with the red population having slightly higher median value ($\sim 3.90''$) compared to the blue population ($\sim 3.76''$).

2. Both samples have a similar median Sérsic index of $n \sim 1.0$.

3. The mean surface-brightness distributions differ noticeably between the two populations: blue galaxies tend to be brighter. We note this behavior is not as prominent as previously observed by Greco et al. [2018]. The distribution in the central surface brightness, $\mu_0(g)$, does not present as large a difference between the two subsamples.

4. The spatial distribution of red LSBGs is much more clustered than that of blue LSBGs, which have an almost homogeneous distribution. This is quantified in the two-point angular correlation function, which is an order of magnitude higher for the red subsample than the blue subsample.

Furthermore, we compared the clustering of the full LSBG sample with a sample of HSBGs selected from the DES and with an external catalog of low-redshift galaxies from the 2MPZ. We find a similar autocorrelation amplitude (and also a high cross-correlation signal) between the LSBG sample and the 2MPZ catalog with a redshift cut of $z < 0.1$ (which is indicative of the low redshift of our LSBG sample). An interesting feature is the lower amplitude of clustering for LSBGs at angular scales less than ~ 0.1 deg.

The spatial distribution of LSBGs contains prominent overdensities. We cross-match the 82 most prominent overdensities with external catalogs of galaxy clusters, galaxy groups, and individual bright galaxies. The association of peaks with objects (clusters, groups, and galaxies) of known distance provides us with distance information for a subset of LSBGs. The distances of associated systems range from $\sim 19\,\mathrm{Mpc}$ (Fornax cluster) to $\sim 354\,\mathrm{Mpc}$ (Abell 2911), with a median distance of $82\,\mathrm{Mpc}$. The mean distance is $106\,\mathrm{Mpc}$ with a standard deviation of $\sim 66\,\mathrm{Mpc}$.

By associating LSBGs with other systems at known distances, we are able to further explore the physical properties of some LSBGs and their host systems. In particular, we present:

1. Projected radial profiles of the distribution of the LSBG and 2MPZ galaxies around the nine most prominent associations. We find that in galaxy clusters, the radial distributions of these two galaxy samples are similar.

2. A physical size–absolute magnitude relationship for LSBGs belonging to the nine most prominent associations. We find that LSBGs in our sample, span a range in physical size (effective radius) from $\sim 0.26\,\mathrm{kpc}$ up to $\sim 4.83\,\mathrm{kpc}$, with a median size of $0.8\,\mathrm{kpc}$. Out of the 555 LSBGs studied, 41 can be classified as UDGs–i.e., have effective radii $R_{\mathrm{eff}}(g) > 1.5\,\mathrm{kpc}$ and central surface brightness $\mu_0(g) > 24.0\,\mathrm{mag\,arcsec}^{-2}$. UDGs appear to be a continuation of the LSBG population.

Our catalog is the largest catalog of LSBGs ($R_{\mathrm{eff}}(g) > 2.5''$ and $\bar{\mu}_{\mathrm{eff}}(g) > 24.2\,\mathrm{mag\,arcsec}^{-2}$) assembled to date. We have presented a general statistical analysis of our catalog, with the hope of enabling more detailed analyses of individual systems and the ensemble population. Future quantitative comparisons can test galaxy formation models in the low-surface-brightness regime, including studies of properties of LSBGs in different environments (clusters/field) and constraints on the mean mass of LSBGs using weak lensing [e.g., Sifón et al., 2018]. Our sample can also be used to better prepare for the next generation galaxy surveys (e.g., with the Vera C. Rubin Observatory). Automated selection procedures result in a large false-positives fraction, necessitating the visual inspection of LSBG candidates. However, visual inspection will become infeasible for the large data sets collected by future surveys. Our LSBG sample can serve as training set for machine and deep learning algorithms, in the hope of fully automating the selection process. The potential of such algorithms will be further explored in upcoming projects. Furthermore, we plan to build upon the know-how we developed constructing the catalog presented in this paper to study LSBGs using the upcoming, deeper data from the total six years of DES observations.

3.9 Surface-Brightness Limits

We estimate the surface-brightness limit of the DES data by applying the `sbcontrast` module from Multi-Resolution Filtering packaged developed for the Dragonfly Telephoto Array [van Dokkum et al., 2020].[14] This procedure bins each coadd image into $10'' \times 10''$ regions, subtracts a local background from each binned pixel based on the surrounding 8 pixels, and calculates the variation among the binned and background-subtracted pixels. We applied this procedure to each DES coadd tile after masking bad pixels and sources detected by `SourceExtractor`. The resulting maps and 1-D distributions of 3σ surface-brightness limits are shown in Figure 3.16. The tail to lower surface-brightness limits comes dominantly from tiles around the survey boarder, which have fewer tilings and less homogenous coverage.

3.10 Selection Criteria

Removal of point sources (star–galaxy separation):

```
(EXTENDED_CLASS_COADD != 0) &
(SPREAD_MODEL_I + 5/3*SPREADERR_MODEL_I > 0.007)
```

[14] https://github.com/AstroJacobLi/mrf

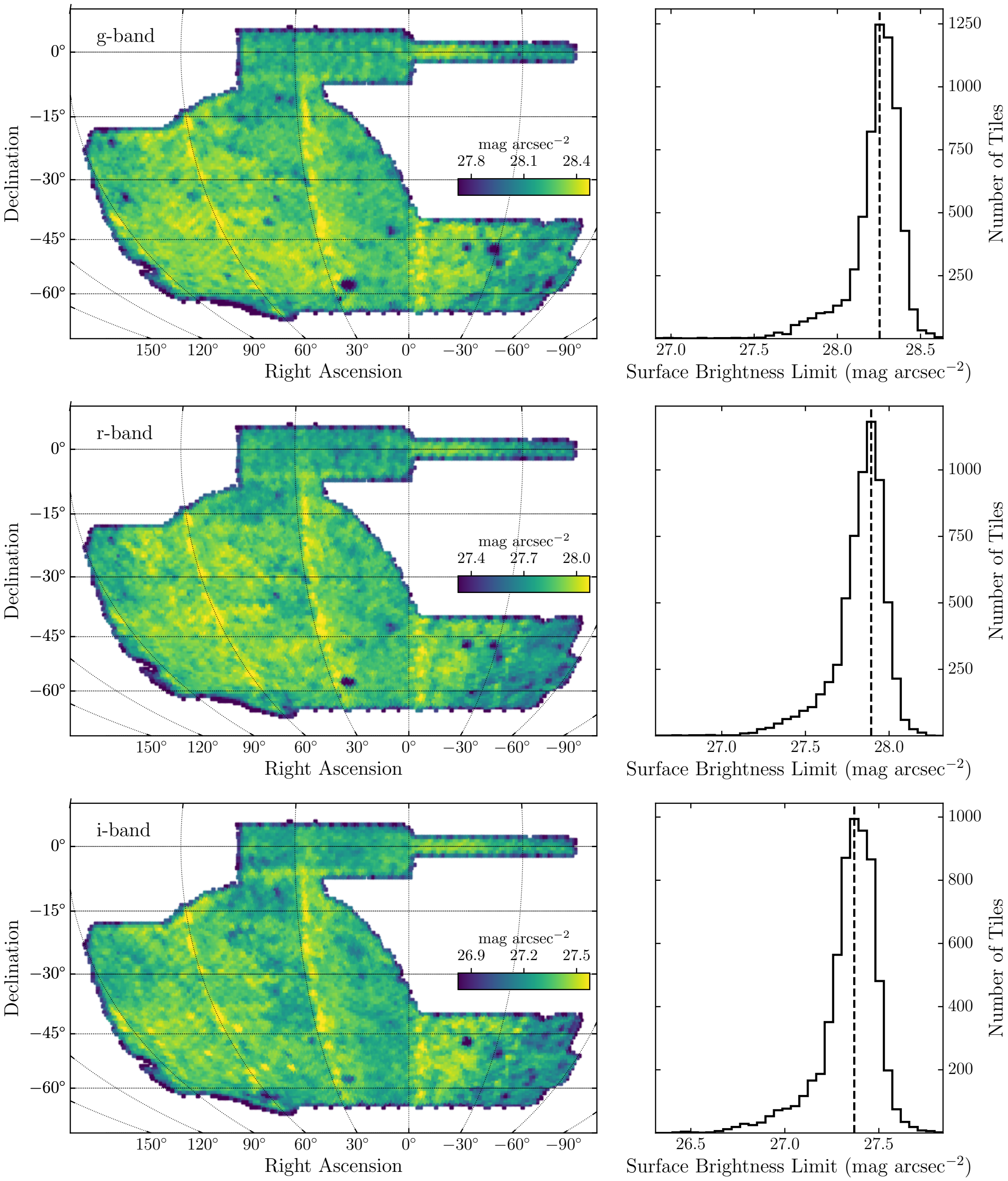


Figure 3.16: Surface-brightness limits at 3σ estimated from the surface-brightness contrast in $10'' \times 10''$ regions over the DES coadd tiles in the g band (top), r band (middle), and i band (bottom).

Selection of LSBG candidates:

- Surface-brightness and radius cuts:

```
(FLUX_RADIUS_G > 2.5) & (FLUX_RADIUS_G < 20)

(MU_MEAN_MODEL_G > 24.2) & (MU_MEAN_MODEL_G < 28.8)
```

- Ellipticity cut:

```
(1 - B_IMAGE/A_IMAGE) < 0.7
```

- Color cuts:

```
-0.1 < (MAG_AUTO_G-MAG_AUTO_I) < 1.4

(MAG_AUTO_G - MAG_AUTO_R) > 0.7*(MAG_AUTO_G - MAG_AUTO_I) - 0.4

(MAG_AUTO_G - MAG_AUTO_R) < 0.7*(MAG_AUTO_G - MAG_AUTO_I) + 0.4
```

3.11 Magnitude Distributions

This appendix presents supplemental plots characterizing the magnitude distribution of our LSBG sample and associated external 2MPZ sample.

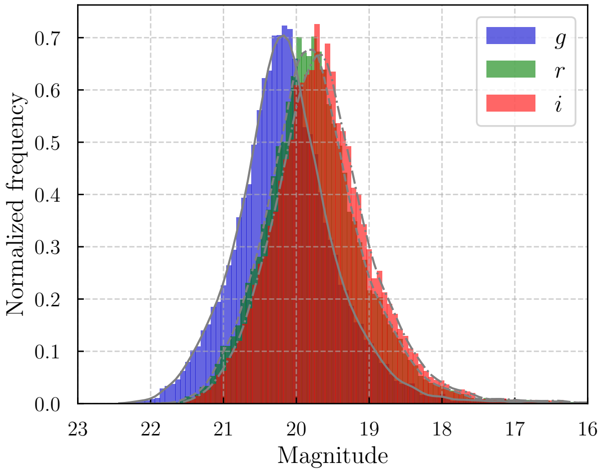
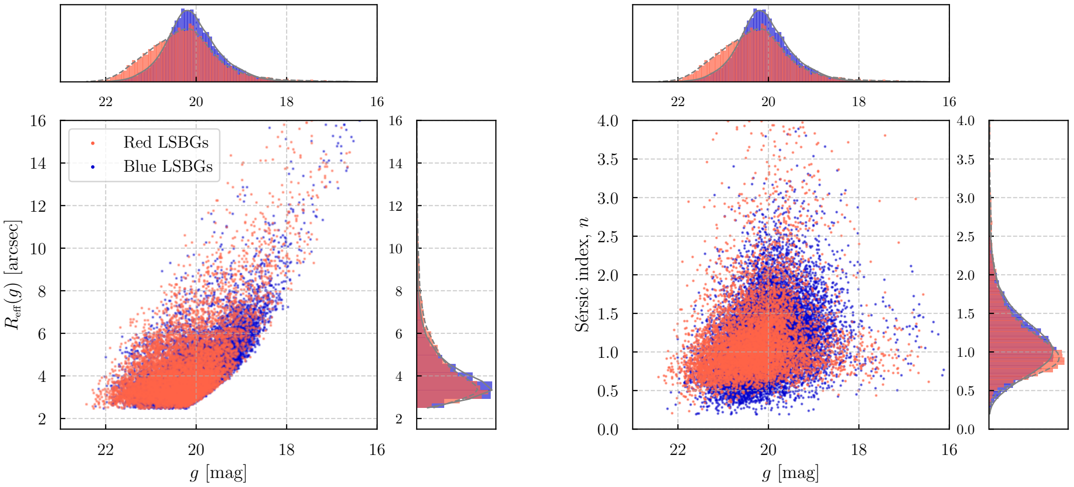


Figure 3.17: Normalized distribution of the g-, r-, and i-band magnitudes of our LSBG sample.

In Figure 3.17 we present the g, r, and i-band magnitude distributions of our LSBG sample. The magnitudes come from the `galfitm` Sérsic model fitting of the sample. The median magnitudes in each band are $g = 20.2$, $r = 19.8$, and $i = 19.7$.



(a)

(b)

Figure 3.18: Joint distributions of the red and blue LSBGs in the space of g-band magnitude vs (a) effective radius, R_{eff}, and (b) Sérsic index, n, both in g-band.

Similar to Figure 3.7, in Figure 3.18 we present joint distributions of the blue and red LSBG subsamples in the space of (a) effective radius, R_{eff}, and (b) Sérsic index vs the g-band magnitude this time. We note that there is no strong color dependence of the g-magnitude distribution.

Finally, in Figure 3.19, we compare the g-band magnitude distributions of the LSBG sample and the 2MPZ galaxy sample that we used in the main text. Because the 2MPZ catalog did not provide such magnitudes, we matched the 2MPZ catalog with the DES Y3 GOLD catalog. The distribution presented here is derived from the `SourceExtractor`'s `MAG_AUTO` magnitudes of these matches. That sample is significantly brighter than the LSBGs, with a median magnitude $g \sim 16.8$.

Note that we do not consider the HSBG sample separately in this section, as by construction it has the same magnitude distributions as the LSBG sample.

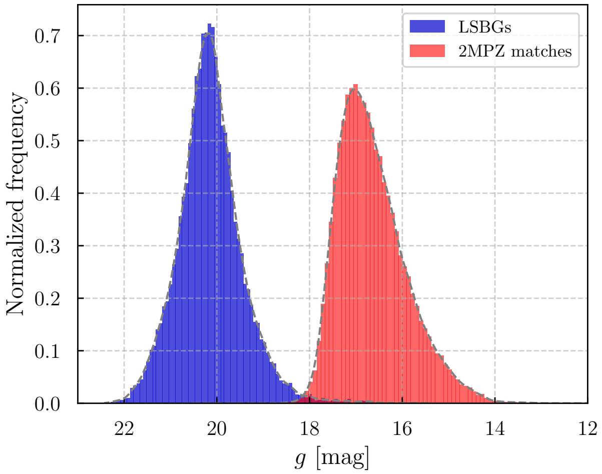


Figure 3.19: g-band magnitude distributions of the LSBG sample and the DES catalog matches on the 2MPZ sample.

CHAPTER 4

DEEPSHADOWS: SEPARATING LOW SURFACE BRIGHTNESS GALAXIES FROM ARTIFACTS USING DEEP LEARNING

The text of this chapter was published as Tanoglidis, Ćiprijanović, & Drlica-Wagner, A&C, 35, 100469 (2021)

4.1 Introduction

Our understanding of galaxy formation, evolution, and the relationship between galaxies and the dark matter halos that they inhabit [e.g., the "galaxy–halo connection"; Wechsler and Tinker, 2018] is constrained by our ability to detect faint galaxies [e.g., Kaviraj, 2020]. Low-surface-brightness galaxies (LSBGs) are conventionally defined as galaxies with a central surface brightness fainter than the night sky ($\mu(g) \gtrsim 22$ mag/arcsec^2). Thus, by definition, they are very difficult to detect and characterize. While contributing only a small fraction to the observed luminosity of the local universe, theoretical [e.g., Martin et al., 2019] and observational [e.g., Dalcanton et al., 1997] arguments suggest that LSBGs account for the majority of galaxies, which thus remains relatively unexplored.

Most of the searches for LSBGs to date have targeted small regions of the sky and have revealed LSBG populations in massive galaxy clusters such as Virgo [e.g., Sabatini et al., 2005b, Mihos et al., 2015, 2017], Coma [e.g., Adami et al., 2006, van Dokkum et al., 2015a] and Fornax [e.g., Hilker et al., 1999, Muñoz et al., 2015, Venhola et al., 2017], as well as faint satellites around the nearby galaxies [e.g., McConnachie, 2012, Martin et al., 2013, Merritt et al., 2016b, Danieli et al., 2017, Cohen et al., 2018a]. To better understand and test galaxy formation models in the low-surface-brightness regime, it is imperative to study LSBGs over a wide sky area and across different environments (inside galaxy clusters vs.

field). Wide-field galaxy surveys have already started to reveal a large number of LSBGs. For example, the Hyper Suprime-Cam Subaru Strategic Program (HSC SSP)[1] discovered 781 radially extended (half-light radius $r_{1/2} > 2.5''$) LSBGs with $\bar{\mu}_{\text{eff}}(g) > 24.3$ mag/arcsec^2 in an analysis of the first $\sim$ 200 deg^2 of its Wide layer [Greco et al., 2018]. More recently, an analysis of the first three years of data from the Dark Energy Survey (DES)[2], covering $\sim 5,000$ deg^2 on the southern sky, brought to light a population of >20,000 LSBGs with similar size and surface brightness limits [Tanoglidis et al., 2021b].

Searches for LSBGs in survey data are plagued by the presence of a large number of low-surface-brightness artifacts in astronomical images. Note that we define the artifact class to consist of any object that passes the selection criteria outlined above but is not an LSBG. This includes imaging artifacts as well as:

- Faint, compact objects blended in the diffuse light from nearby bright stars or giant elliptical galaxies;
- Bright regions of Galactic cirrus;
- Knots and star-forming regions in the arms of large spiral galaxies;
- Tidal ejecta connected to high-surface-brightness host galaxies.

Such objects often dominate the sample of candidate LSBGs. For example, in DES there were 413,000 LSBG candidates, with only ∼5% of them being genuine LSBGs. Even after a feature-based machine learning (ML) classification step that reduced the sample by approximately an order of magnitude, a large number of false-positives remained (∼50% of objects classified as LSBGs). These false-positives had to be manually rejected through visual inspection. Similarly, the authors of the HSC SSP study had to go through a visual

[1] `https://hsc.mtk.nao.ac.jp/ssp/`

[2] `https://www.darkenergysurvey.org/`

inspection step, since their pipeline produced a sample that also had a ∼50% contamination rate from artifacts [Greco et al., 2018].

Visual inspection is time consuming and difficult to perform systematically. Upcoming galaxy surveys, such as the Legacy Survey of Space and Time (LSST)[3] on the Vera C. Rubin Observatory[4] and Euclid[5] are expected to produce massive volumes of data. LSST will observe $\sim$ 20,000 deg^2 of sky, produce 20TB of data per night, and observe ∼10 billion galaxies over its 10 years of observations.[6] With such volumes of data, rejecting artifacts via visual inspection will be impossible. Clearly, the process has to be automated.

Machine learning, and in more recent years deep learning, have started to revolutionize astronomy as the sizes of astronomical datasets grow [for reviews see e.g., Ball and Brunner, 2010, Baron, 2019]. Classification tasks are one of the classical examples where machine learning techniques can be applied. In cases dealing with high-dimensional feature spaces where large training sets are available, deep learning usually outperforms other machine learning algorithms and reaches human-level performance [LeCun et al., 2015].

Convolutional neural networks [CNNs; LeCun et al., 1998] constitute a specific class of deep learning algorithms, inspired by the visual cortex and optimized for computer vision tasks. For that reason they are a promising tool for analyzing astronomical images. Furthermore, working directly at the image level (with the pixels as inputs) eliminates the need for deriving and selecting parameters (sizes, magnitudes, colors etc.) as features to be used for the classification task, which can be subjective and non-optimal.

CNNs were first introduced in astronomy by Dieleman et al. [2015] to perform automatic morphological classification of galaxies and have since found a number of applications. For example, other authors further explored their use in classifying galaxy morphologies [e.g.,

[3] https://www.lsst.org/

[4] https://www.vro.org/

[5] https://www.euclid-ec.org/

[6] https://www.lsst.org/scientists/keynumbers

Dai and Tong, 2018, Domínguez Sánchez et al., 2018, Cheng et al., 2020], separating stars from galaxies [e.g., Kim and Brunner, 2017], identifying strong lenses [e.g., Lanusse et al., 2018, Jacobs et al., 2019, Davies et al., 2019, Bom et al., 2019], eliminating polarimetric artifacts [Paranjpye et al., 2020], evaluating flare statistics in young stars [Feinstein et al., 2020], classifying galaxy mergers [Ćiprijanović et al., 2020b], reconstructing lensing of the Cosmic Microwave Background [Caldeira et al., 2019], setting constraints on the cosmological parameters from weak lensing [e.g., Ribli et al., 2019], and many other applications.

In this paper we present an application of CNNs in classifying LSBGs in astronomical images and we demonstrate that they can significantly help in automating this process. We take advantage of the fact that we have available a large sample of LSBGs and an equally large number of labeled artifacts from visual inspection [Tanoglidis et al., 2021b]. These large labeled training sets are necessary to successfully train CNN models. We compare the performance of our CNN architecture to conventional machine learning models (support vector machines and random forests) trained on features extracted from the same objects. We also study how well the model trained on the DES images can classify images from the HSC SSP, thus demonstrating promise for using a similar technique in upcoming surveys, such as LSST.

This chapter is organized as follows: In Sec. 4.2 we describe the datasets we use for the classification problem. In Sec. 4.3 we briefly summarize the theory and formalism of neural networks and present the specific architecture we use to tackle the problem at hand, which we call *DeepShadows*. In Sec. 4.4 we present the classification results. In Sec. 4.5 we use transfer learning to classify objects from the HSC SSP for which we have labels. In Sec. 4.6 we study the uncertainties present in our study and their impact on the metrics used to assess the classification performance of the model. We discuss our results, propose paths for future investigation, and conclude in Sec. 4.7.

The code and data related to this work are publicly available at the GitHub page of this project: `https://github.com/dtanoglidis/DeepShadows`.

4.2 Data

In this section we describe the datasets used for training and evaluating the performance of the CNN and other machine learning models. We briefly describe the astronomical surveys and selection procedures used to obtain these data.

4.2.1 The Dark Energy Survey

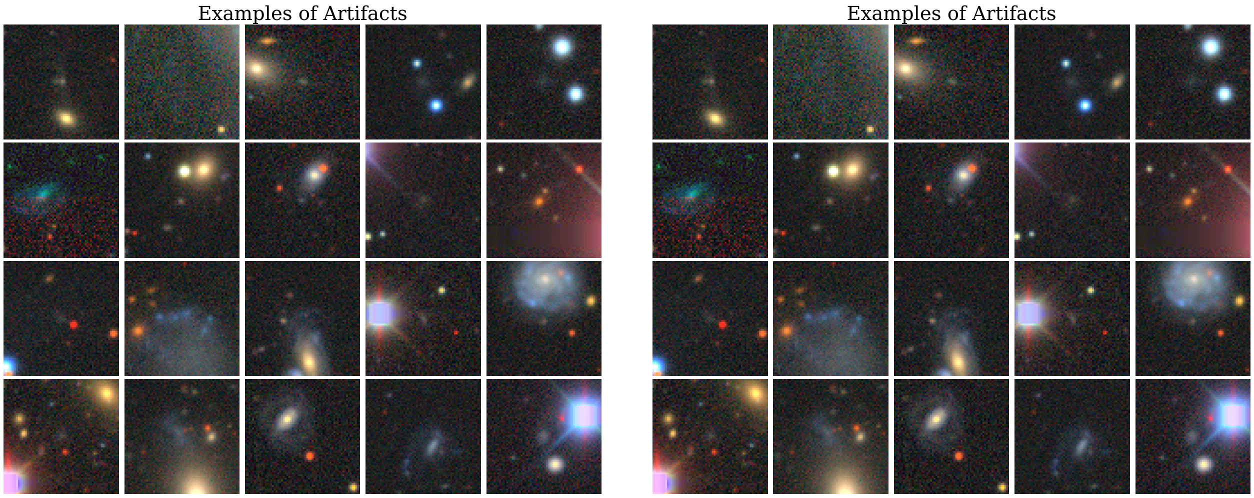
Our primary dataset comes from the first three years (Y3) of observations by DES. DES is an optical/near infrared imaging survey that covers $\sim 5,000$ deg^2 of the southern Galactic cap in five photometric filters, $grizY$, to a depth of $i \sim 24$ over the course of a six-year observational program with the 570-megapixel Dark Energy Camera (DECam) on the 4m Blanco Telescope at the Cerro Tololo Inter-American Observatory (CTIO) in Chile. The DECam field-of-view covers 3 deg^2 with a central pixel scale of $0.263''$ [Flaugher et al., 2015].

Objects were detected in astronomical images using `SourceExtractor` [Bertin and Arnouts, 1996], which provides a catalog of photometric parameters for each object, such as magnitudes, flux radii, mean and central surface brightnesses (adaptive aperture measurements) etc. For a more detailed description of the DES data, see DES Collaboration et al. [2018c].

4.2.2 LSBGs and artifacts in DES

We train, validate and test the performance of our models on LSBGs and artifacts detected by DES, as described in Tanoglidis et al. [2021b]. Here, we briefly outline the main steps followed in that paper for the LSBG catalog construction:

1. Selection cuts were performed using the `SourceExtractor` parameters from the full



(a)

(b)

Figure 4.1: Example images of (a) LSBGs and (b) artifacts in our dataset. Each cutout corresponds to a $30'' \times 30''$ angular region on the sky. We remind the reader that our artifact class consists of any object that passes the low surface brightness selection criteria but is not an LSBG (see example categories in Sec. 4.1).

DES catalog. The most important cuts are on the angular size (half-light radius in the g-band, $r_{1/2} > 2.5''$) and on the mean surface brightness within the effective radius ($\bar{\mu}_{\mathrm{eff}}(g) > 24.3$ mag/arcsec^2).[7] The resulting candidate sample consists of ~ 0.5 million objects.

2. Classification was performed using Support Vector Machines (SVMs) trained on `SourceExtractor` output parameters (features) and a manually annotated set of $\sim$ 8,000 objects, out of which 640 were LSBGs. This step reduces the candidate sample by roughly an order of magnitude. However, the resulting sample still includes $\sim 50\%$ non-LSBG artifacts.

3. Visual inspection was used to reject false positives from more than 40,000 objects positively classified in the previous step.

4. Sérsic model fitting and interstellar extinction correction was applied to objects passing

[7] An updated version of Tanoglidis et al. [2021b] uses a brighter selection, $\bar{\mu}_{\mathrm{eff}}(g) > 24.2$ mag/arcsec^2. However, we keep the older definition, since the LSBG/artifact separation is more challenging in the fainter regime.

the visual inspection, and new selection cuts were performed on the updated parameters.

Note that this dataset contains detection and selection biases that are difficult to estimate due to the use of human visual inspection as described above.

For the current classification study, we randomly select 20,000 LSBGs from those visually verified in Step 3 to be used as the positive class. In the main body of our paper we use as the negative sample 20,000 objects from those visually rejected in Step 3. These are the most challenging artifacts to be separated from LSBGs, since they passed the feature-based classification step in Step 2. We consider a three-class classification problem in 4.8, where we add another class of artifacts, 20,000 randomly selected objects from those rejected in Step 2.

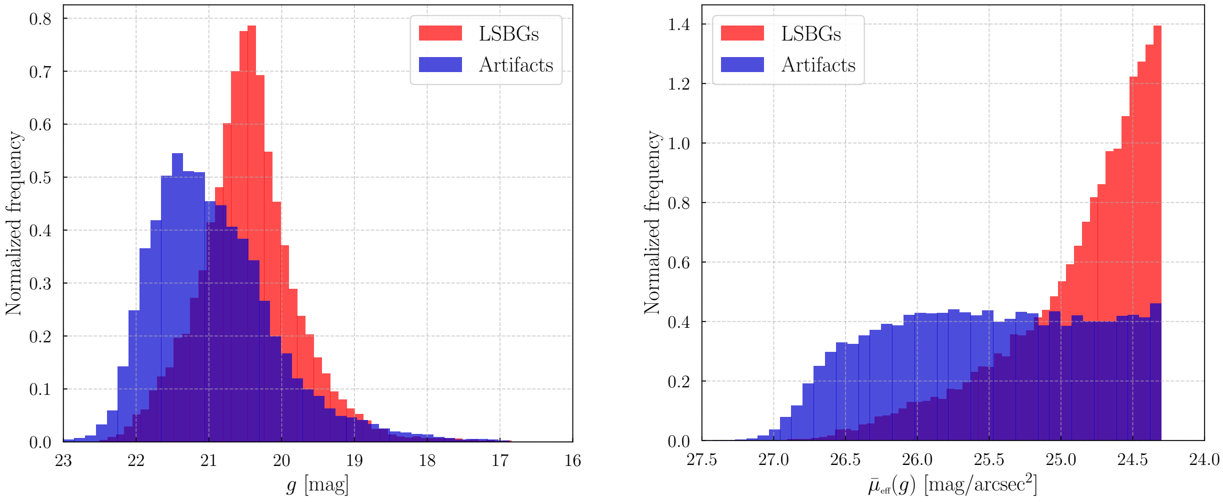
4.2.3 Generation of datasets

For the LSBGs and artifacts we consider two datasets: parameters from `SourceExtractor` to be used as features for the classical machine learning models (SVMs and random forests) and images to be used in our *DeepShadows* CNN model.

For the classical machine learning models, we select the following features:

- Ellipticity of the detected objects.
- `MAG_AUTO` magnitudes in the three bands, g, r, i.
- Colors $g - i$, $g - r$, $r - i$.
- Mean, central and effective surface brightnesses in the three bands, g, r and i.

Each of these properties is derived from the `SourceExtractor` output provided by DES Data Release 1 (DR1, Abbott et al. 2018).



(a)

(b)

Figure 4.2: The g-band (a) magnitude and (b) surface brightness distributions of the LSBGs (red) and artifacts (blue) in our sample.

For the *DeepShadows* CNN model, we generate the image cutouts using the DESI Legacy Imaging Surveys Sky Viewer [Dey et al., 2019][8] to access the DES DR1 images. Each image corresponds to a $30'' \times 30''$ region on the sky and is centered at the coordinates of the candidate object (LSBG or artifact). The initial size of each image is 256×256 pixels that we resize to 64×64 pixels to reduce the dataset size and the memory needs for its processing. The images also have inputs in the three RGB channels (which correspond to g, r, z astronomical bands), so their size is finally $64 \times 64 \times 3$ (we follow a "channels last" format). The code we used to generate the cutouts can be found in the GitHub page of the project. In Fig. 4.1 we show examples of cutouts of LSBGs and artifacts.

Before training, we split our full sample of 40,000 objects into a training set of 30,000 examples, a validation set of 5,000 examples and a test set of 5,000 examples. The selection of objects to be included in each set is random and each set contains an equal number of positive (LSBG) and negative (artifact) examples. We split the datasets of `SourceExtractor` parameters and images in the same way (same objects in each set). In Fig. 4.2 we present

[8] `http://legacysurvey.org/`

the g-band (a) magnitude and (b) surface brightness distributions for the LSBGs (red) and the artifacts (blue) in our (full) sample.

4.2.4 HSC SSP dataset

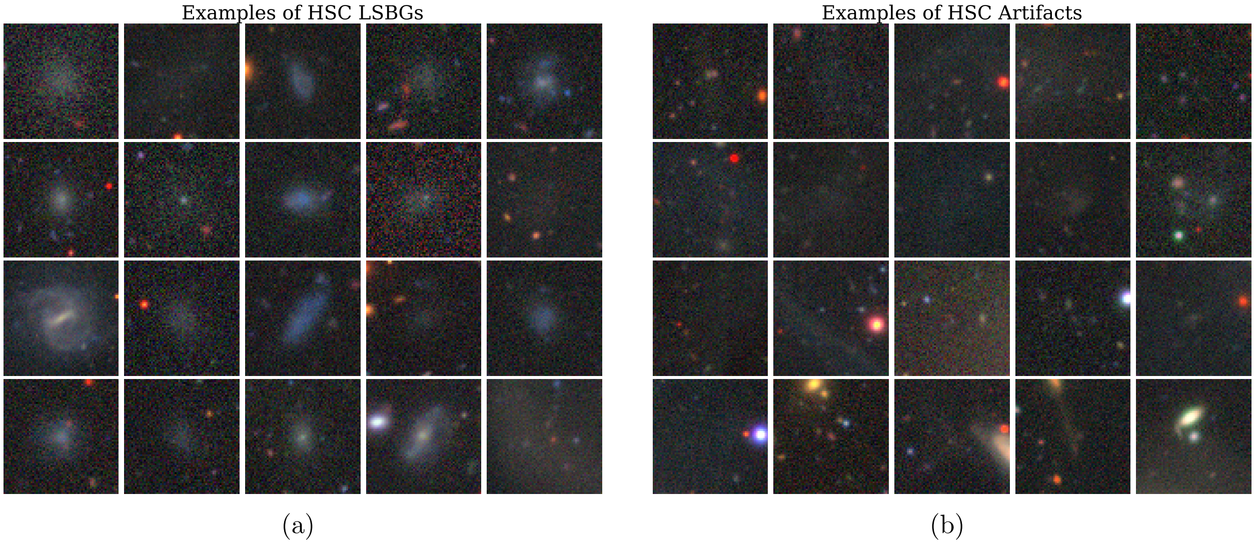


Figure 4.3: Examples of (a) LSBGs and (b) artifacts from the HSC SSP survey. As in Fig. 4.1, each cutout corresponds to a $30'' \times 30''$ angular region on the sky.

We also consider a dataset of 640 LSBGs and 640 artifacts discovered in the HSC SSP as described in Greco et al. [2018]. This is an independent set from a survey with different specifications and different human biases (in the labeling of LSBGs/artifacts) that can be used to test the ability of our model trained on the DES images to classify LSBG candidates in other surveys. We generate image cutouts for HSC SSP, using the Sky Viewer used for the DES images. We split the full dataset of 1,280 objects into a small set of 320 objects to be used for re-training of the classifier (transfer learning) and one of 960 objects for testing the performance.

4.3 Methods

In this section we introduce the notation and briefly describe the machine learning models, the neural networks, and the specific CNN architecture that we use. Our discussion is parsimonious. For a more detailed discussion we suggest the classic book by Hastie et al. [2001] as well as that by Ivezic et al. [2014] that discusses machine learning with a focus on astrophysical applications. The book by Goodfellow et al. [2016] has become a standard reference for deep learning.

4.3.1 Machine Learning

We consider two machine learning classification algorithms that have been proven to be powerful in a number of astrophysical problems (see the review papers mentioned in the introduction and references therein), namely SVMs and random forests. These algorithms perform best with *structured* data (i.e., features/properties of the objects under study) and we apply them to the `SourceExtractor` output properties.

Consider a number of examples, N, each described by a feature vector $\mathbf{x}_i$, $i = 1, \ldots, N$ and with labels y^i (they usually are denoted as $\{1, -1\}$ or $\{1, 0\}$ in two-class problems).

The SVM classifier [Cortes and Vapnik, 1995] seeks to find a hyperplane that separates the two classes, of the form $\mathbf{w} \cdot \mathbf{x} - b = 0$, with the weights $\mathbf{w}$ and the bias b selected to maximize the margin (distance) between this hyperplane and the training samples that are closest to it (support vectors). In practice some misclassification is allowed (soft margin SVM), controlled by a tunable hyperparameter. Furthermore, in many cases, a non-linear transformation is performed on the data that maps them into a space where they are more easily linearly separable.

Random forests is another powerful classification method. Random forests [Tin Kam Ho, 1995, Breiman, 2001] are an ensemble classifier, in the sense that use a collection of simple classifiers, known as decision trees (DTs). Specifically, it considers a set of n DTs and for each one uses a random selection of $\sqrt{m}$ features (m is the total number of features) to construct a DT and train it on a randomly selected subset of the training examples. The output of the random forest is the majority class derived from the DTs.

The values of the hyperparameters are selected (tuned) by exploring a grid of possible values, and obtaining those that give the best results either by evaluating on the validation set, or using $k-$fold cross validation, where the training set is split in k parts, with $k-1$ used for training and the other for validation, and then repeating this process k times.

4.3.2 Deep Learning

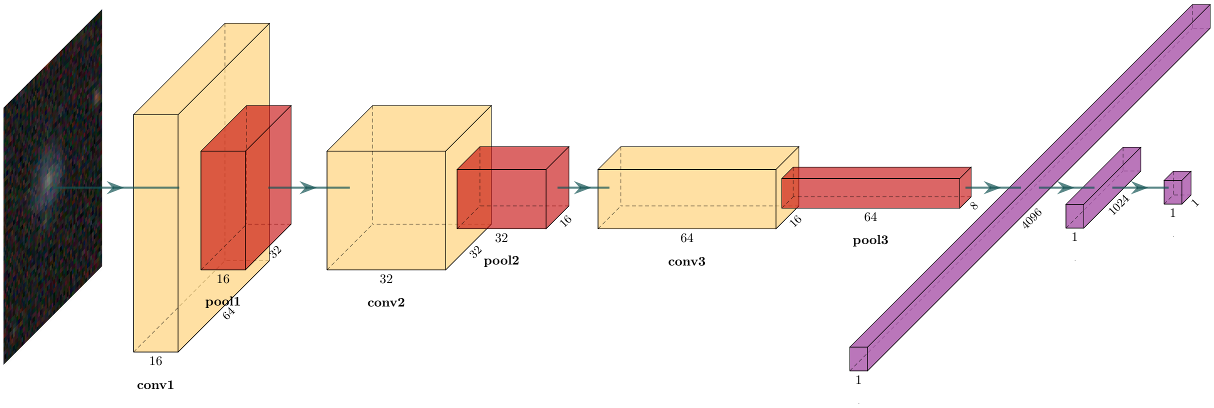


Figure 4.4: Schematic overview of the *DeepShadows* CNN architecture. There are three convolutional layers (yellow), each followed by a max pooling layer (red). The number of filters are 16, 32 and 64 for each layer, respectively. Two dense layers (purple) follow the last pooling layer after flattening. The output is a probability score that the image contains an LSBG. Figure was created using the PlotNeuralNet code [Iqbal, 2018].

The standard deep learning architecture is that of a multi-layer neural network, with the outputs of the previous layer being the inputs to the following one. For example, for the n-th layer:

$$\mathbf{x}_n = g(\mathbf{W}_n \mathbf{x}_{n-1} + \mathbf{b}_n), \tag{4.1}$$

where $\mathbf{W}_n$ a matrix containing the weights of all the neurons of that layer and $\mathbf{b}_n$ a vector

Table 4.1: Architecture of the *DeepShadows* CNN.

Layers	**Properties**	**Stride**	**Padding**	**Output Shape**	**Parameters**
Input	$64 \times 64 \times 3$[a]	-	-	(64, 64, 3)	0
Convolution (2D)	Filters: 16	1×1	Same	(64, 64, 16)	448
	Kernel: 3×3	-	-	-	-
	Activation: ReLU	-	-	-	-
	Reg: L2 (0.13)	-	-	-	-
Batch Normalization	-	-	-	(64, 64, 16)	64
MaxPooling	Kernel: 2×2	2×2	Valid	(32, 32, 16)	0
Dropout	Rate: 0.4	-	-	(32, 32, 16)	0
Convolution (2D)	Filters: 32	1×1	Same	(32, 32, 32)	4640
	Kernel: 3×3	-	-	-	-
	Activation: ReLU	-	-	-	-
	Reg: L2 (0.13)	-	-	-	-
Batch Normalization	-	-	-	(32, 32, 32)	128
MaxPooling	Kernel: 2×2	2×2	Valid	(16, 16, 32)	0
Dropout	Rate: 0.4	-	-	(16, 16, 32)	0
Convolution (2D)	Filters: 64	1×1	Same	(16, 16, 64)	18496
	Kernel: 3×3	-	-	-	-
	Activation: ReLU	-	-	-	-
	Reg: L2 (0.13)	-	-	-	-
Batch Normalization	-	-	-	(16, 16, 64)	256
MaxPooling	Kernel: 2×2	2×2	Valid	(8, 8, 64)	0
Dropout	Rate: 0.4	-	-	(8, 8, 64)	0
Flatten	-	-	-	(4096)	-
Fully connected	Activation: ReLU	-	-	(1024)	4195328
	Reg: L2 (0.12)	-	-	-	-
Fully connected	Activation: Sigmoid	-	-	(1)	1025

[a] We use "channel last" image data format.

of biases. The *activation function*, g, and its purpose is to introduce non-linearities in the network. We use the rectified linear unit (ReLU), $g(x) = \max(0, x)$, activation function except in the final layer, where the activation function takes a sigmoid form (for a binary problem) to allow the output to be interpreted as a score that the example belongs to the positive class.

CNNs are modified versions of the network described in Eq. 4.1, with each neuron in a convolutional layer connected only to neurons within a small rectangle in the previous layer, usually 3×3 to 5×5 pixels in size. The output of such a layer is a predefined number of *feature maps*, generated by convolving the feature maps of each previous layer with different *filters* (or *kernels*), whose trainable weights can capture abstract visual features.

If we have $k = 1, \ldots, K$ input feature maps and $\ell = 1, \ldots, L$ output feature maps, in analogy to Eq. (4.1) we write the ℓ-th output map of the n-th layer as:

$$\mathbf{x}_n^\ell = g\left(\sum_{k=1}^{K} \mathbf{W}_n^{k,\ell} * \mathbf{x}_{n-1}^k + b_n^\ell\right), \tag{4.2}$$

where $*$ represents the convolution operation.

Convolutional layers are almost always followed by *pooling* layers whose purpose is to subsample the output of the convolutional layer, reducing the number of trainable parameters. They have no weights and instead keep the maximum (max pooling) or the mean (average pooling) within a small window (usually 2×2 pixels) sliding over the input feature map. Here we use max pooling.

In Fig. 4.4 we present a schematic overview of the CNN architecture we use for LSBG/artifact classification, that we call *DeepShadows*. It is further described in more detail in Table 4.1. *DeepShadows* is a simple sequential architecture, consisted of three convolutional layers (yellow) alternating with pooling layers (red); after the last pooling layer the array is flattened and followed by two fully-connected layer (purple), the last one being a single neuron that outputs the score (0 to 1) that can be interpreted as the confidence that the input image contains an LSBG. All convolutional layers use kernels of size 3×3, while the pooling layers use kernels of size 2×2. Between each convolutional and pooling layer we perform batch normalization [Ioffe and Szegedy, 2015] to make training faster and more stable.

To tackle overfitting we employ the following methods: first, we use dropout [Srivastava et al., 2014] after each pooling layer. Dropout sets a specific fraction (here we use 0.4) of randomly selected weights equal to zero. We also use L2 (also known as ridge or Tikhonov) regularization [e.g., Hastie, 2020] applied on the weights of the convolutional layers with a penalty term $\lambda = 0.13$ and on the first fully connected layer with penalty $\lambda = 0.12$. We provide more training details for the *DeepShadows* model in Sec. 4.4.2.

4.4 Classification Results

4.4.1 Classification Metrics

To evaluate and compare the performance of the classifiers used in this work we use a number of useful classification metrics, each of which quantifies a different aspect of what a "good" classification is. For binary probabilistic classifiers, we assume (unless otherwise specified) that an example with sigmoid output score (loosely interpreted as probability) $P_{\text{out}} > 0.5$ is classified as an LSBG, while an example with $P_{\text{out}} < 0.5$ is classified as an artifact. We also refer to the LSBGs as the positive class [1] and artifacts as the negative class [0].

True positives (TP) are the correctly classified positive examples, and we analogously define the true negatives (TN), false positives (FP) and false negatives (FN). All the classification metrics, in a binary setting, can be expressed as combinations of these quantities.

The Receiver Operating Characteristic (ROC) curve is a commonly used graphical way to evaluate the performance of a binary classifier; the true positive rate (TPR = TP/(TP+FN)) is plotted versus the false positive rate (FPR = FP/(FP+TN)) at different output thresholds; A derived metric is the Area Under the ROC Curve (AUC). ROC curves are useful for visual inspection of the performance of different classifiers and of their uncertainties.

One of the most widely used evaluation metrics is the *accuracy*, which measures the fraction of the correct predictions among the total sample examined:

$$accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}. \tag{4.3}$$

However, specific problems or applications may have specific requirements that are not fully captured in the overall accuracy. For example, we may be interested in the *completeness* of our classification, in other words, what fraction of the LSBGs were actually classified as such (this metric is also known as "*recall*", "*sensitivity*" or "*True Positive Rate*" in the machine learning literature):

$$completeness = \frac{\text{TP}}{\text{TP+FN}}. \tag{4.4}$$

Another useful quantity is the *purity* of the classification: the fraction of objects classified as LSBGs that are true LSBGs (this quantity is also known as "*precision*" or "*Positive Predictive Value*" in the machine learning literature):

$$purity = \frac{\text{TP}}{\text{TP+FP}}. \tag{4.5}$$

Finally, we also present the confusion matrix, which includes all four TP, TN, FP, FN values. The confusion matrix can be used to construct a number of other classification metrics.

4.4.2 Results

We start by considering the classification results from the two machine learning models (SVMs with an RBF kernel and random forests) described in Sec. 4.3.1. We use these classification algorithms as implemented in the Python library `scikit_learn` [Pedregosa et al., 2011a].[9]

We train these models on the dataset of features derived from `SourceExtractor`, as described in Sec. 4.2.3. The hyperparameters of the models were tuned by searching a grid of values and using five-fold cross validation on the validation set. The best values were found to be $C = 10^4$ (controls how much error is allowed in the soft-margin SVM), $\gamma = 0.001$ (controls the width of the kernel of the non-linear transformation of the data) for the SVM model, while for the random forests model we tune the number of trees in the forest (`n_estimators` = 100) and the number of samples required to split internal nodes (`min_samples_split`=10).

[9] `https://scikit-learn.org/stable/index.html`

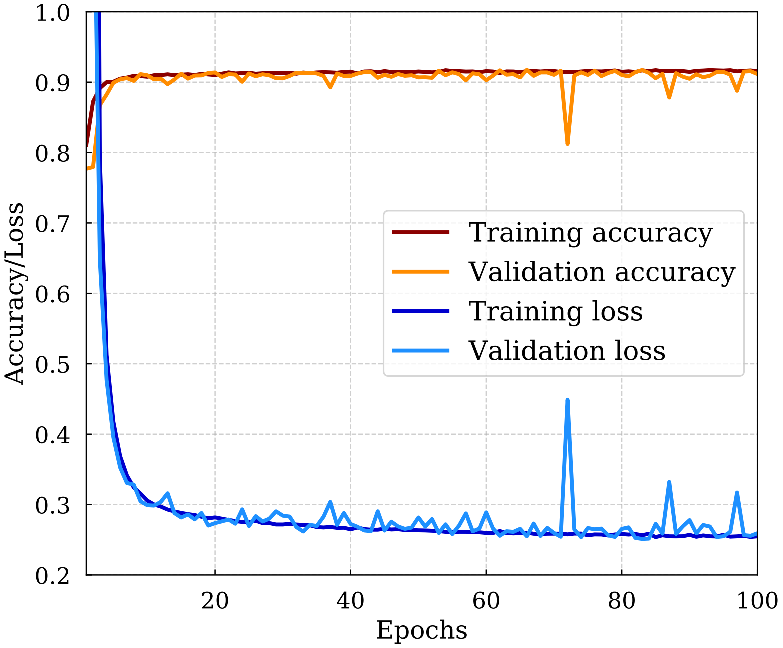


Figure 4.5: Training and validation accuracy/loss as a function of the training epoch for the *DeepShadows* model. Training was performed on 30,000 images and validation on 5,000. The model reaches a training accuracy of $\sim 92\%$ after 100 epochs.

The performance of the models was evaluated on the test set. SVMs reach slightly higher accuracy (81.9% vs 79.7%), higher completeness (or recall, 86.7% vs 80.4%), similar purity (or precision, 79.6% vs 79.7%) and higher AUC (0.894 vs 0.872) compared to the random forest classifier[10]. These results serve as a baseline to compare the performance of our *DeepShadows* CNN model with.

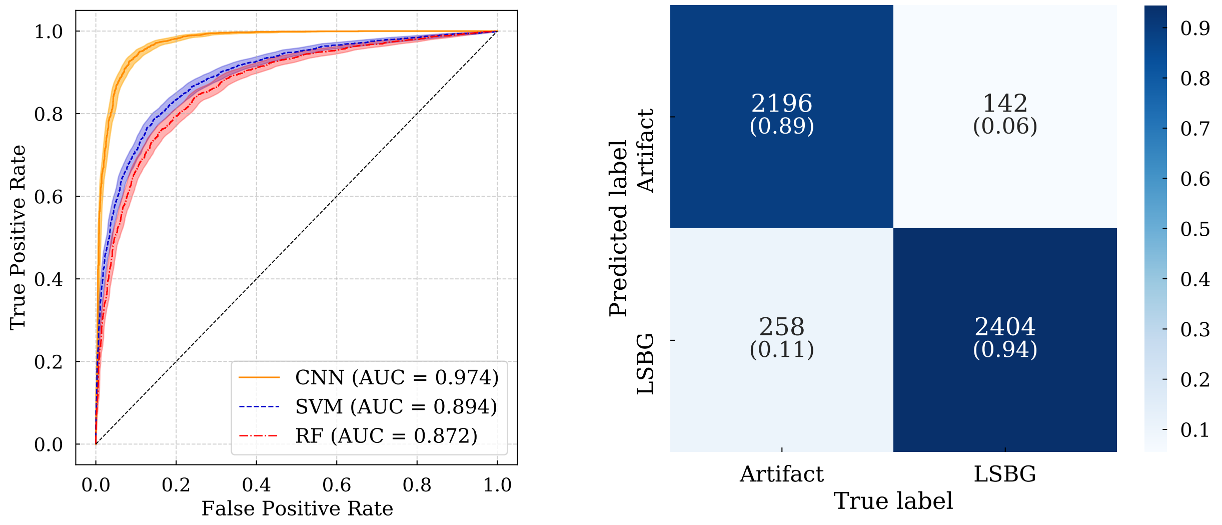
We implement the *DeepShadows* architecture, as described in Sec. 4.3.2 and Table 4.1, using the `Keras`[11] framework on a `TensorFlow`[12] backend. We train the model on the training set of 30,000 images described in Sec. 4.2.3. The weights were updated using Adadelta [Zeiler, 2012], an optimized version of the vanilla stochastic gradient descent algorithm, with

[10] We note again that the SVM model discussed here is different from the one described in the second bullet of Sec. 4.2.2. The SVM presented here is trained on the same annotated dataset of LSBGs and artifacts that was used to train *DeepShadows* and that had confused the original SVM model (Sec. 4.2.2).

[11] `https://keras.io/`

[12] `https://www.tensorflow.org/`

a learning rate of $\eta = 0.1$. The loss function used is the binary cross-entropy. The update is performed iteratively in batches of images; we use a batch size of 64. A training *epoch* occurs once every image in the training set has been used to update the network weights. We train our model for 100 epochs; we do not continue for more epochs since the results do not improve more. We validate the process on the validation set of 5,000 images described in Sec. 4.2.3. In Fig. 4.5 we present the training history of our model (accuracy/loss as a function of training epoch). We can see that our model converges well and that there are no signs of overfitting or underfitting (the training and validation curves closely follow each other).



(a)

(b)

Figure 4.6: (a) ROC curves for the *DeepShadows* CNN model (orange, solid line), the SVM model (blue dashed line), and the random forest model (red dashed-dotted line) evaluated on the test set. The diagonal dashed line corresponds to the performance of a random classifier. We also show the 95% confidence intervals on the vertical direction (true positive rate). These were estimated using the bootstrap method on the test set of images (see Sec. 4.6). (b) Confusion matrix of the *DeepShadows* model predictions on the test set. The values in parentheses correspond to the normalized version of the matrix, obtained by dividing the number of objects in each case by the total number of objects in each category (true label).

The *DeepShadows* CNN classifier reaches an accuracy of 92.0%, completeness (recall) of

Table 4.2: Comparison of the classification metrics for the three machine learning models presented in Sec. 4.4.2.

Metric	**SVM**	**RF**	**CNN**
Accuracy	0.819	0.797	0.920
Completeness	0.867	0.804	0.944
Purity	0.796	0.797	0.903
AUC score	0.894	0.872	0.974

94.4% and purity (precision) of 90.3% and AUC score equal to 0.974, all evaluated on the test set of $5,000$ images. These values are significantly higher than those obtained from the SVM and random forest models (see also Table 4.2 for a direct comparison). Note that these classical machine learning models were trained on a physically motivated set of features that is not guaranteed to be optimal, while *DeepShadows* works directly at the pixel level.

The fact that *DeepShadows* is a more powerful classifier can be visually demonstrated by plotting the ROC curves of the three models (left panel of Fig. 4.6). In the same figure we also show the AUC scores, as well as the 95% confidence intervals on the ROC curves, estimated using the bootstrap method on the test set (see Sec. 4.6).

On the right-hand side of Fig. 4.6 we present the confusion matrix for *DeepShadows* that shows the number of the correctly classified and misclassified objects. Many common classification metrics can be derived from the confusion matrix. In parentheses we present the entries of the normalized confusion matrix, which can be obtained by dividing by the total number of objects in each (true label) category. We see that most misclassification cases occur in artifacts classified as LSBGs, something that is also evident from the lower value of purity compared to completeness.

As can be seen in Fig. 4.2b, the objects in our sample (LSBGs and artifacts) span a wide range in surface brightness ($24.3 < \bar{\mu}_{\rm eff}(g) \lesssim 27.0$ mag/arcsec^2). To investigate whether the performance of our classifier depends on the surface brightness, we split the test sample into three bins according to their g-band surface brightness ($[27-26, 26-25, 25-24.3]$ mag/arcsec^2), and we evaluate the performance of the classifier in each bin independently.

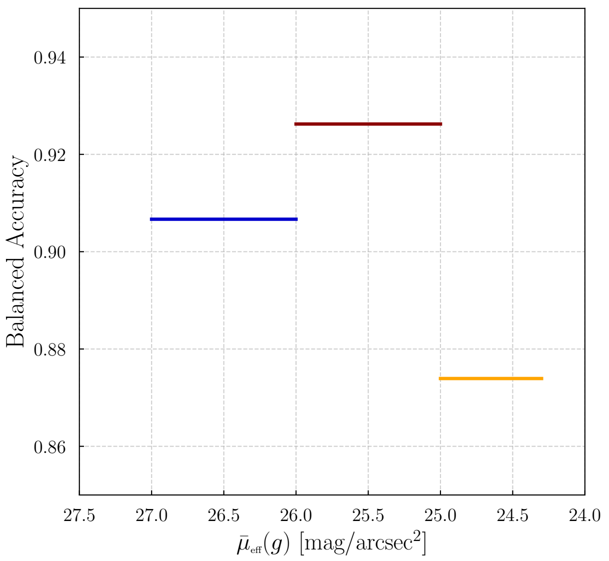


Figure 4.7: Balanced accuracy score of the *DeepShadows* predictions on the test set, for the objects in three different surface brightness bins.

Since the samples are not balanced in these bins (i.e., there are more artifacts than LSBGs in the faintest bin, while the opposite is true in the brightest bin), we calculate the *balanced accuracy* in each bin, which is defined as the average of recall obtained for each class.

We present the results of this study in Fig. 4.7. The balanced accuracies in the three bins (starting from the faintest) are 0.906, 0.926, and 0.874, respectively. This variation in performance is small and comparable to the expected intrinsic uncertainty in the prediction of the classifier (see Sec. 4.6). Interestingly, we see that the worst performance comes from the brightest bin. We find that a larger fraction of artifacts are misclassified as LSBGs in that bin.

4.4.3 Interpretation of Results

The *DeepShadows* CNN is a classifier that outputs a score that can be interpreted as approximately the probability that an example is an LSBG. The classification results presented so far assume a threshold $P_{\rm out} = 0.5$ (any image with higher output score is classified as an LSBG, while any image with lower output score is classified an artifact). We can get more

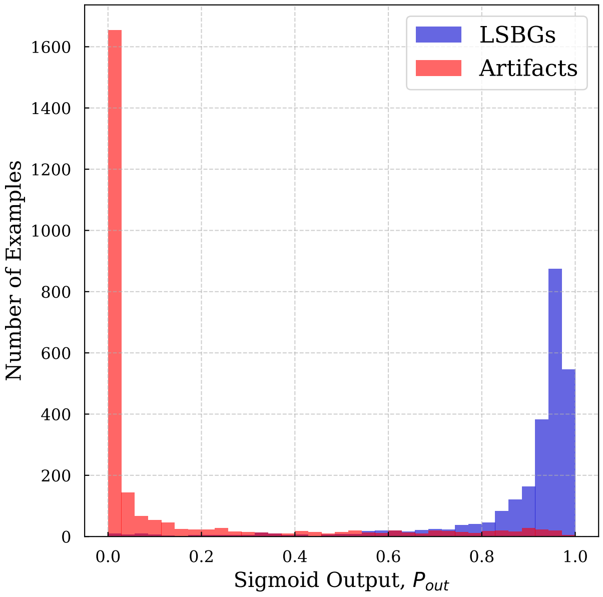


Figure 4.8: Output scores from the *DeepShadows* CNN for the LSBGs and artifacts in the test set.

insight about the classification outcomes, and try to interpret the results, by examining the predicted output scores for the objects in the test set. We plot these sigmoid output scores in Fig. 4.8. Output scores of those objects with true label "artifact" are in red and those with true label "LSBG" are in blue. Artifacts are found to be more concentrated towards the $P_{\rm out} = 0$, implying that most can be easily distinguished from LSBGs. However, there is also a long tail in this distribution with some objects that are labeled (true value) as artifacts but have been assigned a high confidence level of being LSBGs. LSBGs, on the other hand, have a wider distribution in output scores (less concentrated towards $P_{\rm out} = 1$) but a less significant tail to very low scores.

To better understand our results it is useful to inspect some of the objects that: (a) were assigned to the wrong class but with high confidence, or (b) were assigned to the correct class with high confidence. To interpret the classification results we employ a recent technique, called Gradient-weighted Class Activation Mapping [Grad-CAM, Selvaraju et al., 2016]. Grad-CAM allows us to produce "visual explanations" for the classification results,

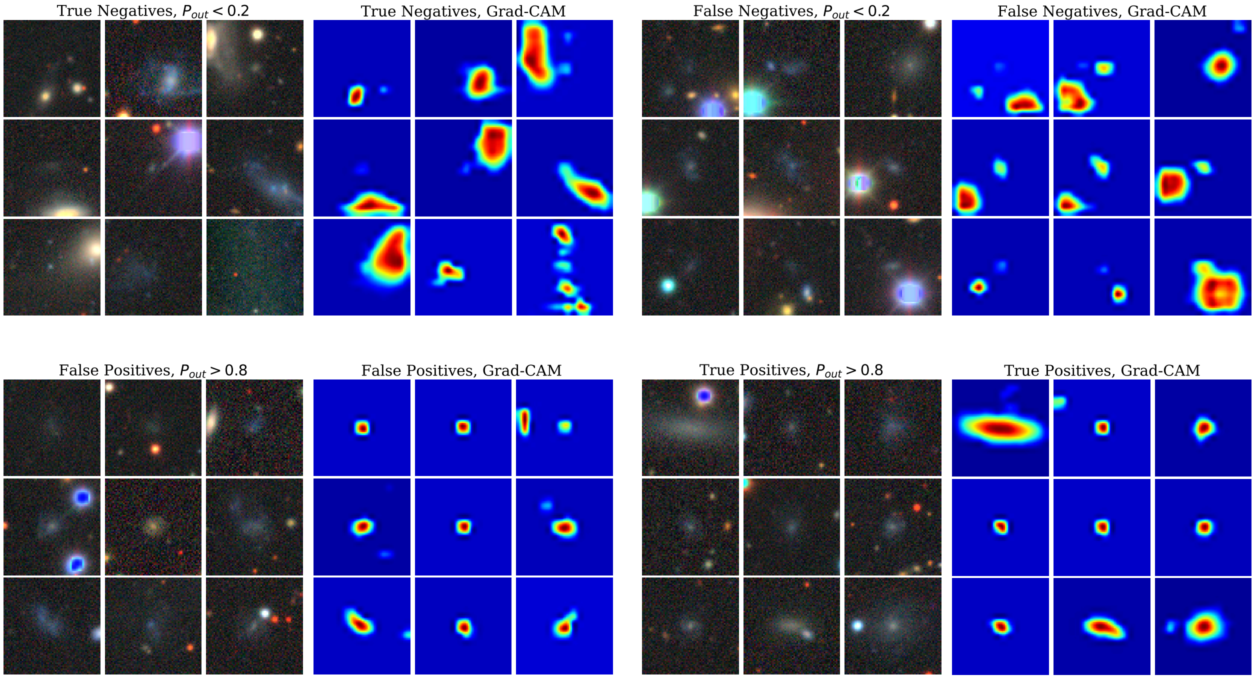

True Negatives, $P_{out} < 0.2$
True Negatives, Grad-CAM
False Negatives, $P_{out} < 0.2$
False Negatives, Grad-CAM
False Positives, $P_{out} > 0.8$
False Positives, Grad-CAM
True Positives, $P_{out} > 0.8$
True Positives, Grad-CAM

Figure 4.9: Examples of objects classified with high output score in their respective class and corresponding Grad-CAM visualization maps for the same objects. Clockwise: True Negatives, False Negatives, True Positives, False Positives (same arrangement as the confusion matrix).

by highlighting the most important regions in the classification procedure. We provide more technical details about Grad-CAM in 4.9; here we just note that these images are produced by calculating the gradients of the feature maps of the last convolutional layer with respect to the output score for each class.

In Fig. 4.9 we present examples and corresponding Grad-CAMs for randomly selected (clockwise): true negatives, false negatives, true positives, false positives. All these examples were classified with high confidence to their assigned categories ($P_{\text{out}} > 0.8$ for those classified as LSBGs and $P_{\text{out}} < 0.2$ for those classified as artifacts).

The images of the objects classified as negatives (artifacts), both true and false, are characterized by the presence of off-centered light sources (such as stars or components of galaxies). The fact that these are the important regions for the classification problems is also confirmed by Grad-CAM maps. Especially in the case of false negatives, we see that the central LSBGs are shadowed by the presence of other nearby objects that contribute to the decision of *DeepShadows* to classify them as artifacts.

On the other hand, the images of those objects classified as LSBGs (positive class) are dominated by the presence of a central object; this can also be seen in the highlighted regions of the Grad-CAM maps. Interestingly, the high-confidence false positives presented here seem to be real galaxies. These objects were likely rejected out of an abundance of caution when visually selecting LSBGs, since these objects were generally more compact or faint compared to other LSBGs. The neural network classifier is able to "correct" the human labeling.

4.5 Transfer Learning

In the previous section, both the training and evaluation of the classifiers used data from the same survey, namely DES. These results demonstrate that a CNN classifier trained on a large training sample can be used to separate LSBGs from artifacts with a high accuracy (and purity/completeness). However, a *large* number of labeled training examples is required. In this section we explore whether we can a classifier trained on one survey and apply it to data from another survey. If such an approach is successful, it can significantly reduce the need to generate large training sets via visual inspection in future surveys.

Transfer learning refers to the process of training a machine learning algorithm to perform a task and then using it to perform another related task or perform the same task on a dataset with different specifications [e.g., Weiss et al., 2016]. Transfer learning has found many applications, including image recognition problems [e.g., Pan and Yang, 2010, Bengio, 2012, Yosinski et al., 2014, Zhuang et al., 2019]. Its power has recently been explored in astronomy [Vilalta, 2018], especially in the context of cross-survey classification, namely in the field of galaxy morphology prediction [Domínguez Sánchez et al., 2019b]. The use of transfer learning has also been investigated for classification of galaxy mergers [Ackermann

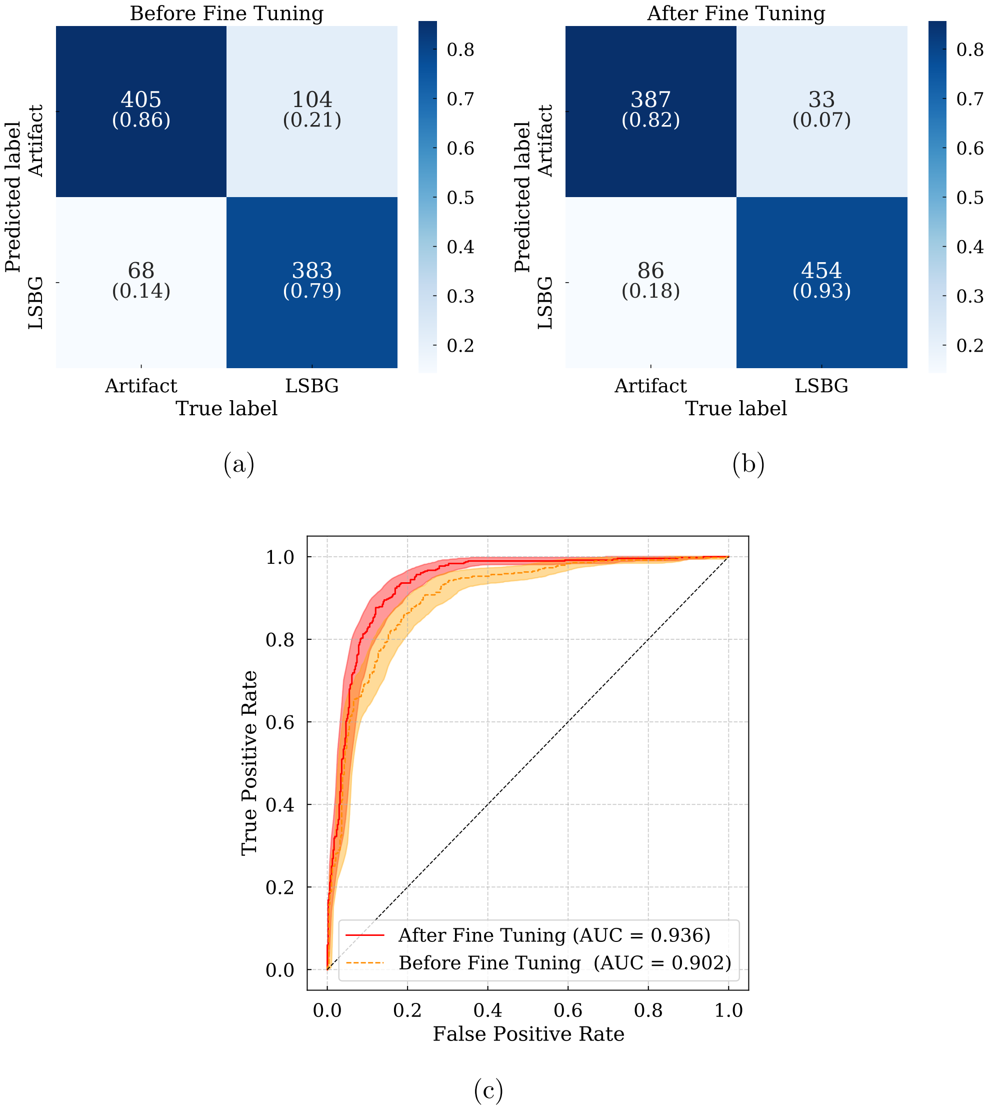


Figure 4.10: Transfer learning results on the HSC SSP test set. (a) Confusion matrix (raw and normalized values) before fine tuning. (b) Confusion matrix (raw and normalized values) after fine tuning. (c) ROC curves and AUC scores, before (orange, dashed line) and after (red, solid line) fine tuning. The bands correspond to the 95% confidence level intervals.

et al., 2018], radio galaxy classification [Tang et al., 2019], star-galaxy classification [Wei et al., 2020], even glitch classification of LIGO events [George et al., 2018].

Here we use data from the HSC SSP, for which we have a small visually classified sample of LSBGs (Sec. 4.2.4), to study how successfully a model trained on one survey can be used to distinguish between LSBGs and artifacts in another survey. Following Domínguez Sánchez et al. [2019b] we consider two cases:

1. Apply the *DeepShadows* model, which was trained on DES data, directly to the HSC SSP data (test set) without any further training.

2. Before predicting on the HSC SSP test set, we use a small set of 320 objects from the HSC SSP to perform a *fine-tuning* step. We re-train the whole model using a much smaller learning rate (in order to keep the change of the weights low and avoid overfitting), $\eta = 0.005$, for 30 epochs, using a batch size of 16 (Note that, alternatively, one can re-train only the final, dense layers. We do not consider this case here).

We present the results (confusion matrices and ROC curves) for the two cases in Fig. 4.10. Before fine tuning, *DeepShadows* has an accuracy of 82.1%, purity (precision) of 84.9% and completeness (recall) of 78.6%. Classification performance significantly improves with fine tuning, as can be seen by inspecting the two ROC curves (and the corresponding AUC scores) in panel (c) of Fig. 4.10. The better performance is driven by the increased number of true positives and correspondingly smaller number of false negatives, as can be seen by comparing the confusion matrices in panels (a) and (b) of Fig. 4.10. The overall accuracy after fine tuning reaches a value of 87.6%, the purity a value of 84.1% and the completeness an impressive 93.2% (almost as good as the application to the one we get from applying to the DES data).

We have demonstrated that we can achieve good performance classifying LSBGs in HSC SSP using transfer learning with a fairly small set for re-training. We would like to quantify the benefit of transfer learning by estimating the size the HSC training set that would be needed to reach comparable accuracy without using transfer learning. Unfortunately, given the small size of the HSC data set, we cannot directly answer this question. However, we can use the DES data to quantify model performance as a function of training set size. We train *DeepShadows* from scratch using progressively larger subsets of the full DES training set. We randomly select 360, 500, 1000, 3000, 5000, and 10000 examples objects in each run;

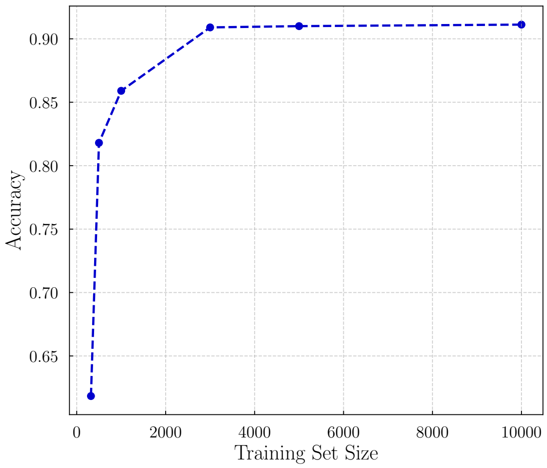


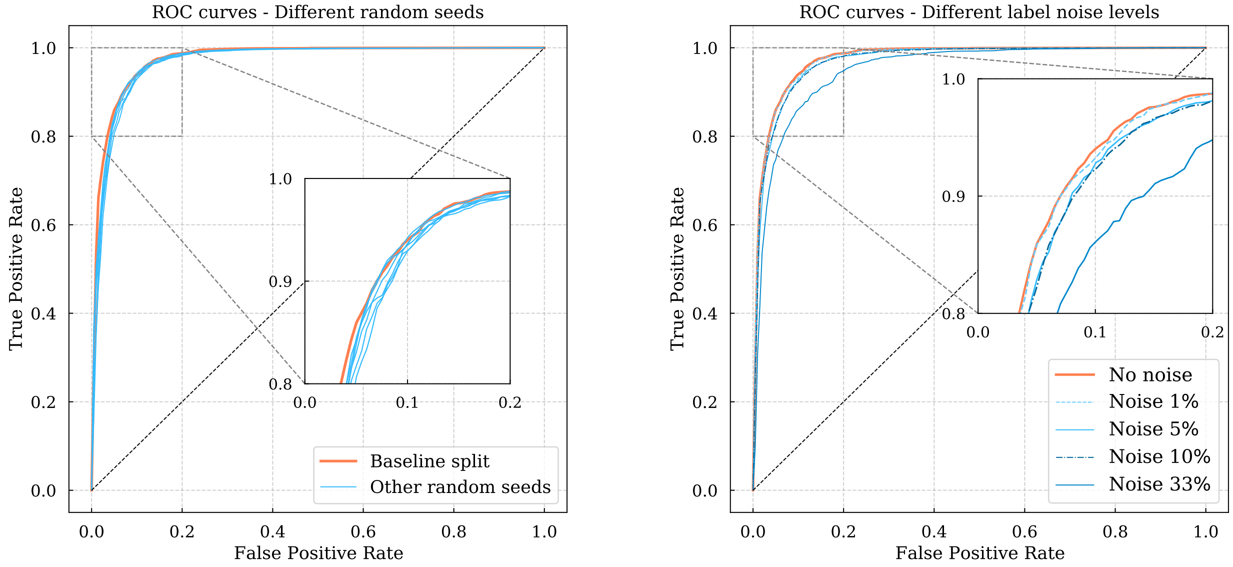
Figure 4.11: Accuracy, evaluated on the test set, for the *DeepShadows* model trained on progressively larger datasets.

we evaluate the performance (accuracy) on the same test set as the one used in Sec. 5.4. To prevent overfitting, which is likely to happen when small training sets are used, we employ early stopping.

The results are presented in Fig. 4.11. The accuracy when the small training set of 360 examples is used is very low, $\sim 61.6\%$; it rapidly increases though reaching $\sim 81.8\%$ when 500 examples are used and $\sim 85.9\%$ when the training set has a size of 1000. After increasing the training size to 3000 the accuracy almost plateaus with the three larger training sets giving accuracies of 90.9%, 91.0%, and 91.2%, respectively. To achieve an accuracy of 87.6% (the one achieved using transfer learning) a training set of ~ 2000 object is required, about one order of magnitude larger than the sample we used for fine tuning. We note that the accuracy reaches high values (comparable to those reached when the full dataset was used) even with relatively small training sets.

4.6 Uncertainty Quantification

We have presented performance metrics for our models without discussing potential uncertainties in these estimates. We now consider three potential sources of uncertainty:



(a)

(b)

Figure 4.12: (a) ROC curves of the performance on the test set using the baseline split (orange line) between training-validation-test sets and using splits with different random seeds (blue lines). (a) ROC curves of the performance on the test set using different levels of label noise in the training set, starting from no noise (orange line) to 33% random mislabeling.

1. Random statistical uncertainties when calculating the evaluation metrics on the test set (subsection 4.6.1).

2. Uncertainties arising from randomly splitting the data into training, validation, and test sets (subsection 4.6.2).

3. Label noise from the presence of mislabeled examples in our training set (subsection 4.6.3).

Our analysis is not intended to be exhaustive. For a more complete discussion of the uncertainties present in deep learning models see e.g., Kendall and Gal [2017], Hüllermeier and Waegeman [2019], Caldeira and Nord [2020], and references therein.

4.6.1 Bootstrap resampling of the test set

We estimate 95% confidence intervals on the classification metrics by bootstrapping the test set 1000 times with replacement and classifying each realization. We present these confidence intervals in Table 4.3; the 95% intervals for the true positive rate are also presented in panel (a) of Fig. 4.6. Note that this approach is not comprehensive. In principle, we could have bootstrapped the training set to evaluate model error in addition to statistical prediction error. However, re-training *DeepShadows* 1000 times was computationally prohibitive.

Table 4.3: 95% Confidence intervals on the classification metrics from Bootstrap resampling of the test set.

Metric	**95% Confidence Interval**
Accuracy	0.912 – 0.928
Completeness	0.935 – 0.953
Purity	0.891 – 0.914
AUC score	0.970 – 0.974

The typical interval for all parameters is (∼1.5%). This is much smaller than the difference between the performance of *DeepShadows* and the other machine learning models, robustly demonstrating that *DeepShadows* performs better in classifying galaxies and artifacts. The confidence intervals of the SVM and random forest models are of similar width.

Another popular method for uncertainty estimation in deep learning models is Monte Carlo dropout [Gal and Ghahramani, 2015]. This method uses dropout at testing time to produce slightly different label predictions on the test set each time we make predictions. We have implemented this method, and we found results comparable to those from bootstrapping.

We have seen in Fig. 4.10 that the confidence intervals in the ROC curves (true positive rates) in the case where we explored transfer learning are wider than those of the main section (training and test on the DES data). In Table 4.4 we present 95% confidence intervals for the transfer learning task, after fine tuning, evaluated on the HSC SSP test set. As we can see the typical range in the evaluation metrics in this case is ∼0.04–0.05, significantly larger

Table 4.4: 95% Confidence Intervals on the classification metrics from Bootstrap for the transfer learning task, after fine-tuning, evaluated on the HSC SSP test set.

Metric	**95% Confidence Interval**
Accuracy	0.856 – 0.896
Completeness	0.910 – 0.954
Purity	0.808 – 0.871
AUC score	0.921 – 0.951

than before.

4.6.2 DES dataset assignment

Table 4.5: Classification metrics for the baseline split of training-validation-test sets and six alternative splits using different random seeds. Because of the small number of runs we present each case individually and not as an interval like in Table 4.3.

Seed #	Accuracy	Completeness	Purity	AUC score
Baseline	0.920	0.944	0.903	0.974
1st run	0.912	0.974	0.868	0.967
2nd run	0.915	0.972	0.870	0.971
3rd run	0.917	0.950	0.889	0.968
4th run	0.914	0.936	0.896	0.968
5th run	0.920	0.917	0.924	0.972
6th run	0.915	0.937	0.898	0.971

We have noticed that the model performance is sensitive (at a level similar to the uncertainties described in the previous section) to the random assignment of examples into the training-validation-test sets. We study variations in model performance stemming from random assignment by splitting the whole dataset into training-validation-test using 6 different random seeds, retraining using each new training set, and evaluating on the new test sets. Note that we do not change the sizes of these sets; these are always 30,000 (training), 5,000 (validation), 5,000 (test).

In Table 4.5 we present the classification metrics for each one of the runs while in the left-hand side of Fig. 4.12 we show the ROC curves for each one of these cases. Due to the limited number of runs (constrained by the cost of re-training the model each time) we do not present summary statistics (like 95% confidence intervals); however we can see that for each metric, the values span a range similar to those presented in Table 4.3. The baseline split, used in Sec. 5.4 was selected as one of the better performing models after several trials. In the GitHub page of this project we make available the on-sky coordinates (right ascension, declination) of the training, validation and test sets under the baseline split.

4.6.3 Impact of mislabeling

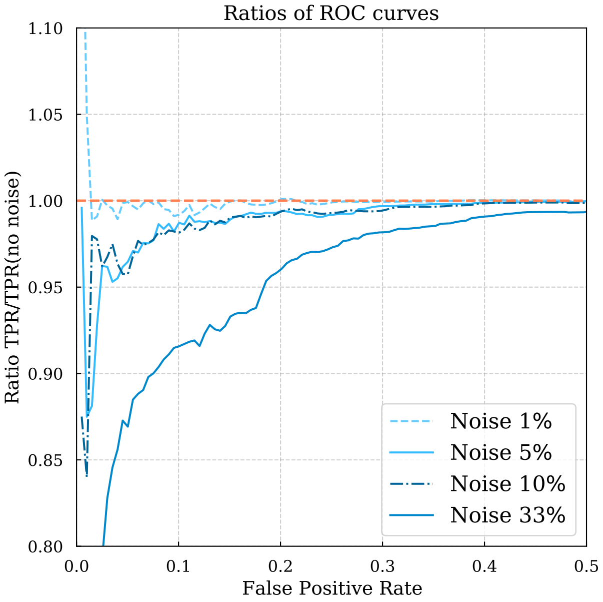


Figure 4.13: Ratio of the ROC curves with different label noise levels to the one without noise. More specifically, we plot the ratio of the true positive rates, as a function of the false negative rate. The horizontal, orange, dashed, line corresponds to ratio = 1. Notice that the range in the false positive rate plotted is 0–0.5 since for higher values all curves converge to a value equal to one.

A final source of uncertainty that we consider is the presence of mislabeled examples in the training set, also known as *label noise*. In Sec. 4.4.3 we showed that our dataset contains some LSBGs that are labeled as artifacts. Generally, we know that we were conservative when performing the visual inspection, meaning that we favored purity over completeness when assembling the LSBG catalog (if we were unsure if an object was a LSBG we preferred

Table 4.6: Classification metrics using the baseline split and different levels of random label noise (mislabeling) in the training set.

Noise level	Accuracy	Completeness	Purity	AUC score
No noise	0.920	0.944	0.903	0.974
Noise 1%	0.918	0.945	0.899	0.974
Noise 5%	0.914	0.935	0.890	0.970
Noise 10%	0.909	0.897	0.923	0.969
Noise 33%	0.883	0.892	0.880	0.950

to flag it as an artifact).

Label noise has been extensively studied in the deep learning literature [for overview papers see e.g., Frenay and Verleysen, 2014, Algan and Ulusoy, 2019, Song et al., 2020]. The general conclusion is that neural networks are relatively robust to label noise, in some cases proving to perform well even in the presence of large noise [Rolnick et al., 2017]. Most of these studies consider well-known annotated datasets (MNIST, CIFAR, etc.) and introduce artificial label noise. However, it is interesting and useful to study potential effects of mislabeling in our dataset.

The best way to do this study would be to identify the mislabeled examples by carefully inspecting and reclassifying them. However, such a detailed study is time-intensive and beyond the scope of this work. Here we consider the impact of label noise in the following way: we randomly select a number of examples (LSBGs and artifacts alike) equal to $1\%, 5\%, 10\%$ and 33% of images from the training set and we flip their label. We retrain the model each time and we evaluate on the test set.

The classification performance metrics for each noise level (and for the baseline case without noise) can be found in Table 4.6. The corresponding ROC curves, that allow for a visual comparison of the performance of the models, can be seen in the right-hand panel of Fig. 4.12. Furthermore, in Fig. 4.13 we plot the ratios of the ROC curves for the models with different noise levels to that without label noise, for better inspection of the differences. We can see that for small ($\lesssim 10\%$) label noise levels the reduction in performance is minimal.

Reduced accuracy only becomes noticeable once the label noise reaches 33%, though even in this case the accuracy reduction is not very large. Our results thus confirm the studies of Rolnick et al. [2017].

So far we have considered a purely random noise, in the sense that we randomly selected an equal number of LSBGs and artifacts and flipped their labels. We repeated the above exercise introducing targeted noise—i.e., we selected only artifacts and changed their label to LSBGs in the first case, and LSBGs that changed their labels to artifacts in the second case. In both cases the results were qualitatively similar to the random noise case. We conclude that the presence of label noise in our sample, either random or biased against one of the two categories, does not significantly change the model performance.

4.7 Discussion and Conclusions

In this work we presented the application of deep learning to the problem of automatic LSBG/artifact classification in astronomical images of LSBG candidates. This study was enabled by the availability of large samples of both LSBGs and artifacts from DES [Tanoglidis et al., 2021b].

We showed that a simple CNN architecture with three convolutional and two dense layers can achieve classification accuracy of 92.0% (completeness 94.4% and purity 90.3%) and significantly improves over conventional machine learning models (SVMs and random forests) trained on `SourceExtractor`-derived features (accuracy 81.9% and 79.7%, respectively). This performance is found to be relatively robust to label noise.

We also demonstrated that knowledge obtained from one survey (training on DES data) can be transferred to another survey (prediction on HSC SSP data, accuracy 82.1%) and the performance of this *transfer learning* can be significantly improved when the model is retrained on a small sample of examples from the new survey (accuracy 87.6%).

These results are promising and impactful for two reasons. First, automating the classification process (or, at least, significantly reducing the need for visual inspection) will be necessary given the data volumes of future surveys such as LSST and Euclid, and even future analyses of current surveys, such as the full 6-year of DES observations and future data releases from HSC SSP. Second, automated classification makes it much easier to characterize, in an unbiased way, the completeness/detection efficiency of future LSBG catalogs. The standard way to characterize detection efficiency is by injecting a large number of mock galaxies with a known parameters (e.g., effective radius, surface brightness, Sérsic index etc.) into the imaging data and then applying the same detection pipeline. The efficiency can be calculated as a function of galaxy parameters by measuring the fraction of mock galaxies that are recovered [e.g., Song et al., 2012, Suchyta et al., 2016, Venhola et al., 2018]. To characterize the detection efficiency over the allowed space of galaxy parameters, it is often necessary to simulate more mock galaxies than are observed in the data itself. This makes unbiased human classification very challenging.

We have presented a study of CNNs for LSBG–artifact separation[13]. Our primary goal was to demonstrate the feasibility of such an approach, and we briefly outline possible further investigations. Specifically, improvements in the CNN model, the training data, the use of domain adaptation techniques for transfer learning, and the systematic study of uncertainties would all improve future models for LSBG classification and discovery.

In particular, CNN architectures have a large number of tunable hyperparameters—e.g., number of layers, filters, kernel sizes, dropout levels, regularization parameters. Finding the optimal combination in a manner similar to that used for the SVM and random forest classifiers is computationally expensive. We have tested that the architecture presented here is robust to small changes – e.g., the model with 3 convolutional layers performs better compared to one with 2 or 4 layers, etc. However, a grid of hyperparameters should be

[13] Note that the reason we consider only these two categories is because of the preprocessing steps described in Sec. 4.2.3; because of them the final candidate sample is forced to contain only those two categories. Without this preprocessing one may consider a multi-class classification, like normal galaxies, stars etc.

explored to ensure an architecture that gives the best results. This would likely require parallelizing the model training and evaluation processes to reduce computational time. Furthermore, more complex types of networks, such the Residual Neural Networks [ResNet He et al., 2015] that allow for very long (large number of epochs) training, should be explored.

The quality of data used for training and testing is also very important. In Sec. 4.6.3 we showed that the presence of label noise does not significantly change performance. However, the selection of objects for which the labels were flipped was totally random; in practice the objects that are more challenging to characterize and thus prone to mislabeling are of a specific type – for example very faint or very compact. It would be useful to reclassify our sample into different confidence categories and check how the performance of the classifier changes when only high-confidence LSBGs and artifacts are used for training. Having a test set without label noise is also important; we have seen that some of the "misclassifications" were actually result of label noise, thus leading to slightly misestimated performance metrics (here we refer to noise introduced by a human labeler, not the artificial label we introduced in Sec. 4.6.3). Furthermore, *data augmentation* is a commonly used technique that could be explored in the future [e.g., Shorten and Khoshgoftaar, 2019]. Data augmentation is a regularization technique used to avoid overfitting, where one increases the number of training examples by adding slightly modified copies of the existing images (rotated, resized etc.).

The topic that likely requires the most detailed further study is that of transfer learning from one survey to another. Here our exploration was minimal, either applying the model trained on DES data directly to HSC SSP data or retraining the whole model on a very small set from HSC SSP. More *domain adaptation* techniques [e.g., Kouw and Loog, 2018, Wang and Deng, 2018] (techniques that allow algorithms trained in one or more "source domains" to be successfully used in a different, but related, "target domain") should be explored before choosing an approach to apply to forthcoming surveys (for domain adaptation applications in astronomy, see e.g., Vilalta et al. 2019, Ćiprijanović et al. 2020a, 2021). These techniques would allow the models to be successfully applied to the new data without the need to retrain the model later and more importantly to manually label new "target" datasets since these techniques often use unlabeled target datasets. This makes the process much faster. Furthermore, the benefit of using larger example sets from the target survey for re-training of the model should also be explored.

Finally, the topic of uncertainty quantification in deep learning is a very active area of research; here a simple error estimation was presented. Future exploration should include bootstrapping the training set and not just the test set (something that would be computationally expensive and should be parallelized), sources of statistical and systematic uncertainties (known as "epistemic" and "aleatoric" in the machine learning community) should be studied separately, as well as potential correlations between the two.

We plan to address some of these questions in future work, but the results presented here are already very promising for the upcoming analysis of the full six years of DES data. However, given the inherent challenges present in classifying very low-surface-brightness objects, the performance of a CNN classifier may never reach human-level accuracy. We argue that as we enter a new, big-data based, era in astronomy, the community should be ready to take a leap of faith in accepting the presence of small and well-characterized classification errors in favor of the great statistical power that comes when assembling large catalogs of objects, using automated deep learning methods, that can illuminate the low-surface-brightness universe.

4.8 Three-class classification

In the main body of this paper we considered a two-class classification: one (positive) category of LSBGs and one (negative) category of artifacts – those visually rejected in the third step of the LSBG catalog generation in Tanoglidis et al. [2021b], also described in Sec. 4.2.2. These artifacts were the hardest to classify, since they had been classified as LSBGs by an

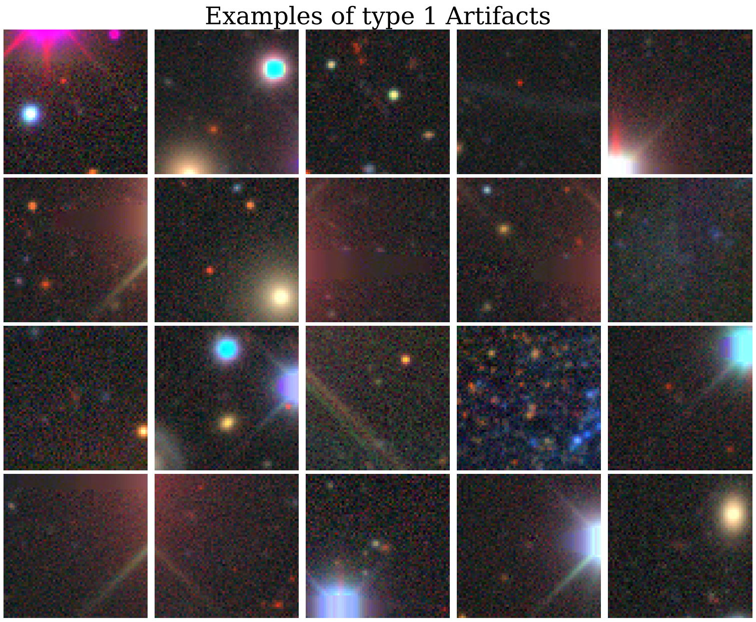
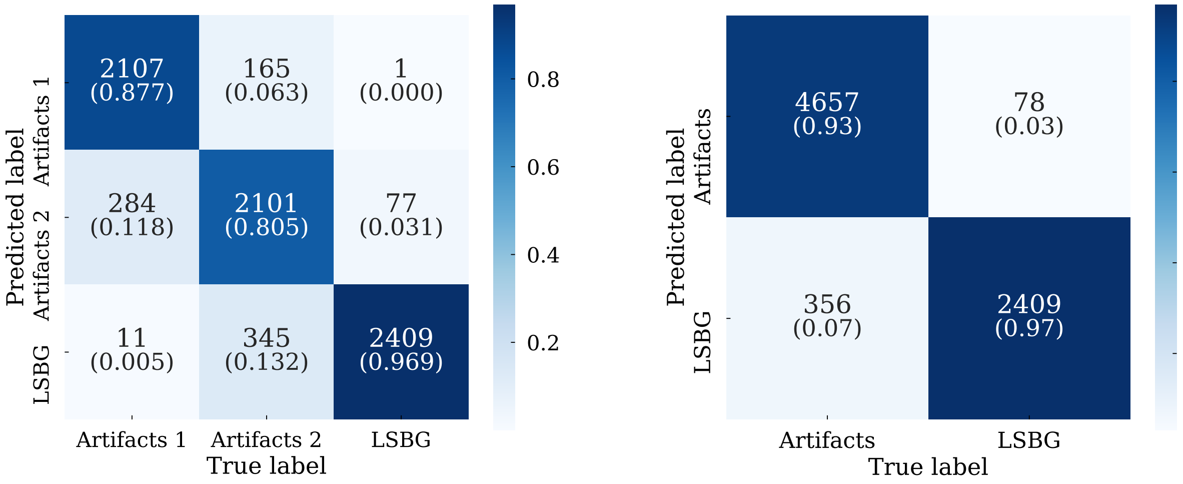


Figure 4.14: Randomly selected examples of type 1 artifacts (see text).



(a)

(b)

Figure 4.15: (a) Confusion matrix for the three-class classification problem, where two artifact classes are considered. The numbers in parentheses correspond to the normalized values. (b) "reduced" confusion matrix, where the two artifact classes have been combined.

SVM classifier trained on `SourceExtractor` features. Before the visual inspection the same classifier had rejected a very large number of artifacts (most of them correctly).

We also investigated the performance of *DeepShadows* on a three-class classification problem where two artifact classes are considered, by adding a sample of 20,000 randomly selected artifacts from those rejected by the SVM classifier. We call this class "Artifacts 1", while the artifacts considered in the main body of the paper are called "Artifacts 2". In Fig. 4.14 we present a randomly selected sample of artifacts of the first kind. We see that these artifacts are dominated by the presence of strong diffraction spikes, and they are less confusing (more easily recognized as artifacts) than those presented in Fig. 4.1.

We keep the same architecture for *DeepShadows*, except that the last dense layer has size of three. We also change the final activation function to softmax and the loss to categorical crossentropy. We train again the model for 100 epochs with a batch size of 64. We split the total dataset of 60,000 object into 45,000 (training), 7,500 (validation), and 7,500 (test) sets.

The resulting three-class confusion matrix of the predictions on the test set can be seen on the left-hand side of Fig. 4.15. We can see that there is very small confusion between the "Artifacts 1" class and "LSBGs" categories, confirming our notion that these are artifacts that can be very easily excluded. Interestingly, the classifier is able to distinguish between the two categories of artifacts, too (with some confusion, of course, accuracy $\sim 90\%$ when calculating for the submatrix between artifacts only).

In the right-hand panel of Fig. 4.15 we combine the two artifact categories, in order to better see the confusion between artifacts and LSBGs. Note that this is now an imbalanced two-class problem, since the artifacts class is twice as large as the LSBGs class. For that reason, accuracy is not a good metric but we can still calculate the completeness and purity, which are more meaningful metrics for the problem at hand. These numbers are 96.8% and 87.1%, respectively and they are comparable (slightly less in purity) to those from the 2-class model discussed in the text.

Our conclusion is that there is not much benefit in using a three-class classification, unless we prefer to eliminate the SVM classification step in future applications.

4.9 Grad-CAM details

We present here some technical details of the Grad-CAM technique for highlighting the most important regions for classification in an image. More details can be found in the original article [Selvaraju et al., 2016]. We also suggest the following blog post[14] for an explanation of the technique and its application using `Keras`.

We define A^k to be the $k-$th ($k = 1, \ldots, K$) feature map of the last convolutional layer, that has dimensions $m \times n = Z$ (Pixel values A^k_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$). Let also y^c the output score (probability) for the class c (obviously, here we have only one positive class, thus only one probability score).

If each convolutional kernel captures a specific visual pattern, then each feature map of the final layer will show where this visual pattern exists in the image. We can thus imagine that the classification output depends on a weighted sum of all the feature maps, with weights depending on the importance each feature has for class c. So, the Grad-CAM maps can be written as: $L^c_{\text{Grad-CAM}} \sim \sum_k \alpha^c_k A^k$.

What are the class-dependent weights a^c_k? The idea is that the gradients of the output score with respect to the (i, j) pixel of the $k-$th feature map, $\partial y^c / \partial A^k_{ij}$, measures the effect of that pixel to the classification score. Grad-CAM then proposes to take the average of all pixels (also known as average pooling) as that the weight for map k and class c is:

$$a^c_k = \frac{1}{Z} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial y^c}{\partial A^k_{ij}}. \tag{4.6}$$

Finally, a ReLU function is applied to the weighted sum, to keep the positive regions, so we get:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k=1}^{K} a^c_k A^k\right), \tag{4.7}$$

[14] `https://fairyonice.github.io/Grad-CAM-with-keras-vis.html`

which produces a localization map that retains the spatial information present in the last convolutional layer.

CHAPTER 5

DEEPGHOSTBUSTERS: USING MASK R-CNN TO DETECT AND MASK GHOSTING AND SCATTERED-LIGHT ARTIFACTS FROM OPTICAL SURVEY IMAGES

The text of this chapter was published as Tanoglidis et al. A&C, 39, 100580 (2022)

5.1 Introduction

Wide-field photometric surveys at optical and near-infrared wavelengths have provided a wealth of astronomical information that has enabled a better understanding of the processes that govern the growth and evolution of the Universe and its contents. Near-future surveys, such as the Vera C. Rubin Observatory's Legacy Survey of Space and Time [LSST; Ivezić et al., 2019][1], will further expand our knowledge of the Universe by extending measurements to unprecedentedly faint astronomical systems. Such surveys will produce terabytes of data each night and measure tens of billions of stars and galaxies.

Images collected by optical/near-infrared surveys often contain imaging artifacts caused by scattered and reflected light (commonly known as "ghosting artifacts" or "ghosts") from bright astronomical sources. These image artifacts are an unavoidable feature of many optical systems. The effective mitigation of ghosts and scattered-light artifacts, and the spurious brightness variations they introduce, is important for the detection and precise measurement of faint astronomical systems. In particular, since many ghosts cover a large image area with relatively low surface brightness [e.g., Slater et al., 2009], they constitute a significant source of contamination in studies of the low-surface-brightness Universe, a major goal of current and upcoming surveys [e.g., Greco et al., 2018, Brough et al., 2020, Kaviraj, 2020, Tanoglidis et al., 2021b].

[1] `https://www.lsst.org/`

Modern wide-field telescopes and instruments greatly reduce the occurrence and intensity of ghosts and scattered-light artifacts by introducing light baffles, and high efficiency anti-reflective coatings on key optical surfaces. Strict requirements on the number and intensity of ghosts and scattered-light artifacts were achieved during the construction of the Dark Energy Camera [DECam; Abbott et al., 2009, Flaugher et al., 2015], which has enabled state-of-the-art cosmological analyses with the Dark Energy Survey [DES; DES Collaboration, 2005, 2016, DES Collaboration et al., 2018a, 2021b].[2], Other smaller surveys have implemented novel optical designs to mitigate the presence of ghosts and scattered-light artifacts [Abraham and van Dokkum, 2014].

Despite these successful efforts, it is often impossible to completely remove ghosts and scattered-light artifacts. For example, the DES 3-year cosmology analyses masked ∼3% of the survey area around the brightest stars, and ∼10% of the survey area around fainter stars [Sevilla-Noarbe et al., 2021]. Additional mitigation steps that go beyond the original survey design requirements are particularly important for studies of low-surface-brightness systems.

The large datasets produced by surveys like DES make the rejection of these residual artifacts by visual inspection infeasible. The situation will become even more intractable in upcoming surveys, like LSST, which will collect ∼ 20TB/night and ∼ 15PB of data over its nominal 10-year survey.[3] Furthermore, the deeper imaging of LSST will place even tighter requirements on low-surface-brightness artifacts [LSST Science Collaboration, 2009, Brough et al., 2020].

To mitigate residual ghosts and scattered-light artifacts, DES uses a predictive Ray-Tracing algorithm as the core of its detection process. This algorithm forward models the physical processes that lead to ghosting/scattered-light events [Kent, 2013], such as the configuration of the telescope and camera optics, and the positions and brightnesses of

[2] `https://www.darkenergysurvey.org/`

[3] `https://www.lsst.org/scientists/keynumbers`

known stars obtained from catalogs external to the survey (for a more detailed description of the Ray-Tracing algorithm, see Kent 2013 and Sec. 2 of Chang et al. 2021). While the Ray-Tracing algorithm is largely successful in predicting the presence and location of artifacts in the images, this algorithm is also limited in predicting the amplitude of the ghost image by the accuracy of the optical model and the external star catalogs used.

Recently, Chang et al. [2021] demonstrated an alternative approach using a convolutional neural network [CNN; **?**] to classify DES images containing ghosts and scattered-light artifacts. CNNs constitute a class of deep neural networks that are inspired by the visual cortex and optimized for computer vision problems. Since their invention, CNNs have found numerous applications in the field of astronomy, including galaxy morphology prediction [e.g., Dieleman et al., 2015, Cheng et al., 2021], star-galaxy separation [e.g., Kim and Brunner, 2017], identification of strongly lensed systems [e.g., Lanusse et al., 2018, Davies et al., 2019, Bom et al., 2019, Huang et al., 2020, 2021], classifying galaxy mergers [e.g., Ćiprijanović et al., 2021], and many other applications. The CNN developed by Chang et al. [2021] was able to predict whether an image contained ghosts or scattered-light artifacts with high-accuracy ($\sim 96\%$ in the training set, $\sim 86\%$ in the test set), but did not identify the specific pixels of the image that were affected by the presence of artifacts. Since ghosts and scattered-light artifacts often affect a subregion of an image, flagging entire images rejects a significant amount of high-quality data.

In contrast to classic CNNs, object detection algorithms are designed to determine the location of objects in an image (e.g., place bounding boxes around objects or mask exact pixels that belong to objects). In this work, we study the use of a deep learning-based object detection algorithm, namely a Mask Region-Based Convolutional Neural Network [Mask R-CNN; He et al., 2017], to predict the location of ghosts and scattered-light artifacts in astronomical survey images. Mask R-CNNs have recently been demonstrated as an accurate tool to detect, classify, and deblend astronomical sources (stars and galaxies) in images

[Burke et al., 2019].

Using 2000 manually annotated images, we train a Mask R-CNN model to identify artifacts in DES images. Comparing the results to those of the Ray-Tracing algorithm on ghost-containing images, we find that Mask R-CNN performs better in masking affected regions — indicated by the value of the $F1$ score (a combination of precision and recall). This demonstrates that deep learning-based object detection algorithms can be effective in helping to address a challenging problem in astronomical surveys without any *a priori* knowledge of the optical system used to generate the images.

This chapter is organized as follows. In Sec. 5.2, we present the dataset, including the annotation process, used in this work. In Sec. 5.3, we describe the Mask R-CNN algorithm, implementation, and the training procedure. In Sec. 5.4 we present results from the Mask R-CNN model, including examples of predicted masks, custom and commonly used evaluation metrics, and we compare its performance to that of a conventional algorithm. We further summarize our results and their applications, and conclude in Sec. 5.5. The code and data related to this work are publicly available at the GitHub page of this project: `https://github.com/dtanoglidis/DeepGhostBusters`.

5.2 Data

In this section, we describe the datasets used for training and evaluating the performance of the Mask R-CNN algorithm for detecting ghosts and scattered-light artifacts. We briefly describe the DES imaging data, our manual annotation procedure, the creation of masks, and the agreement between the human annotators who performed these tasks.

5.2.1 Dark Energy Survey Data

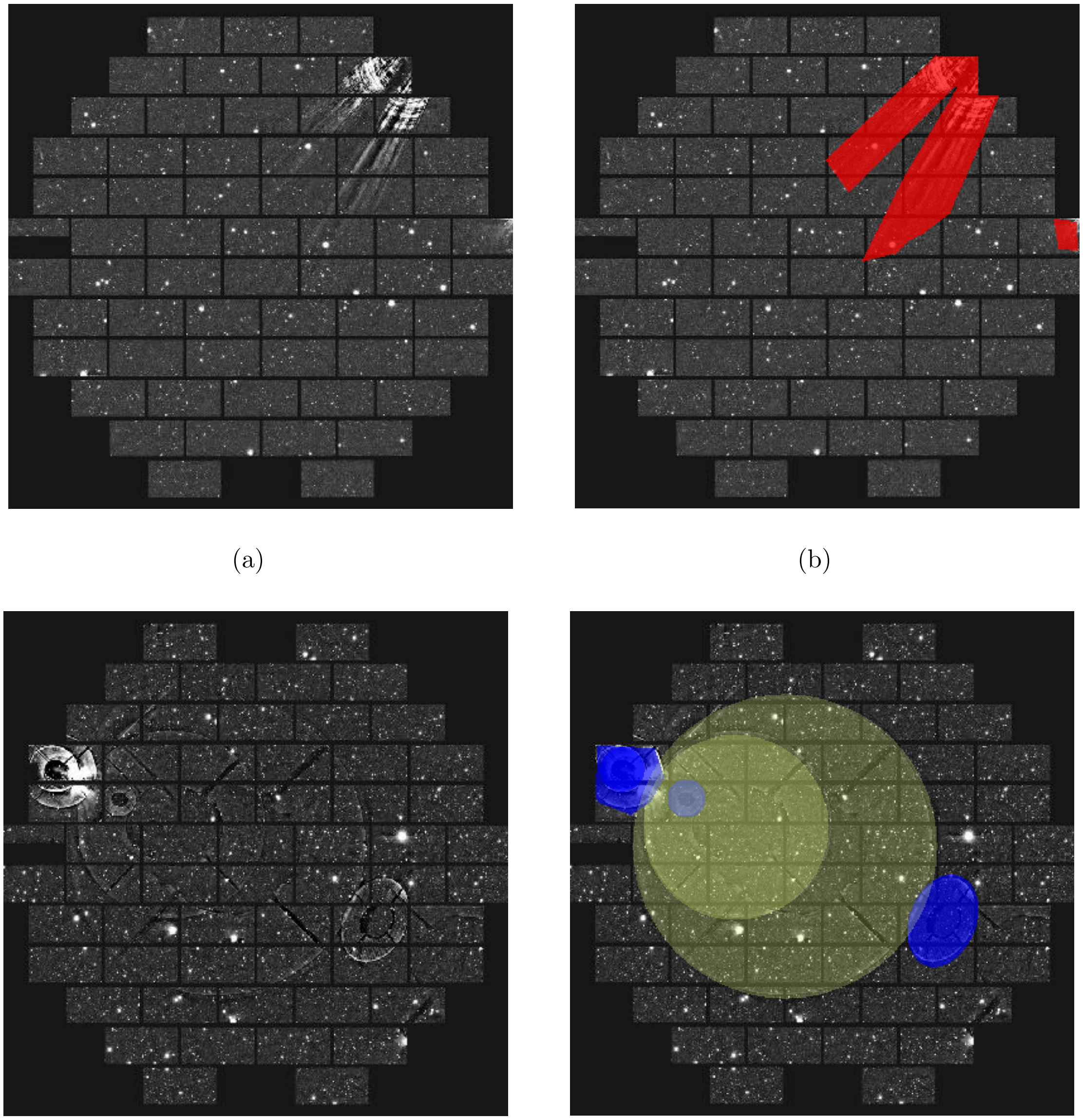
DES is an optical/near-infrared imaging survey that completed six years of observations in January 2019. The DES data cover ~ 5000 deg^2 of the southern Galactic cap in five photometric filters, $grizY$, to a depth of $i \sim 24$ mag [DES Collaboration et al., 2021a]. The observations were obtained with DECam, a 570-megapixel camera mounted on the 4m Blanco Telescope at the Cerro Tololo Inter-American Observatory (CTIO) in Chile [Flaugher et al., 2015]. The focal plane of DECam consists of 62 2048×4096-pixel red-sensitive scientific charge-coupled devices (CCDs), while its field-of-view covers 3 $\deg^2$ with a central pixel scale of $0.263''$.

Our data come from the full six years of DES observations [DES Collaboration et al., 2021a]. For the training, validation, and testing of the Mask R-CNN model, we use 2000 images that cover the full DECam focal plane and are known to contain ghosts and scattered-light artifacts. These are part of the positive sample used in Chang et al. [2021] to train a CNN classifier to distinguish between images with and without ghosts. This dataset was assembled by selecting images that the Ray-Tracing program identified as likely to contain ghosts, and subsequently visually inspecting them to correct for false detections.

As described in Chang et al. [2021], the image data were down-sampled images of the full DECam focal plane. Images were produced with the STIFF program [Bertin, 2012], assuming a power-law intensity transfer curve with index $\gamma = 2.2$. Minimum and maximum intensity values were set to the 0.005 and 0.98 percentiles of the pixel value distribution, respectively. The pixel values in each image were then normalized to a range whose minimum and maximum corresponded, respectively, to the first quartile $Q_1(x)$ and third quartile $Q_3(x)$ of the full distribution in the image, by multiplying each pixel value, x_i, by a factor $s_i = \frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$. Focal plane images were originally derived as 800 × 723-pixel, 8-bit grayscale images in Portable Network Graphics format, which were then downsampled to 400 × 400 pixels for use with the Mask R-CNN. We note that different choices of image scaling could perhaps improve the results [e.g., González et al., 2018] and could be explored in a future work. The data from Chang et al. [2021] are publicly available.[4]

[4] https://des.ncsa.illinois.edu/releases/other/paper-data

5.2.2 Annotation process



(a)

(b)

(c)

(d)

Figure 5.1: Examples of full-focal-plane DECam images containing ghosts and scattered-light artifacts. The corresponding "ground truth" masks (right) were manually annotated. There are three categories of ghosting artifacts: image (a) contains a scattered-light artifact classified as 'Rays'; image (b) shows the masks for the 'Rays' in red; image (c) contains both 'Bright' and 'Faint' ghosts, and the corresponding masks in blue and yellow, respectively, are shown in image (d).

Training the Mask R-CNN algorithm requires both images and ground-truth segmentation masks identifying objects of interest in each image. To create these masks, we used the VGG Image Annotator (VIA; Dutta and Zisserman [2019])[5], a simple manual annotation software for images, audio, and video. We split the 2000 images into batches of 100 images, and we randomly assigned each batch to one of eight authors for annotation.[6]

During manual annotation, we categorized the ghosting and scattered-light artifacts into three distinct morphological categories:

1. 'Rays': These are scattered-light artifacts originating from the light of off-axis stars scattering off of the DECam filter changer [Kent, 2013]. They emanate from one of the edges of the image and span several CCDs. This is the most distinct artifact category and is not commonly confused with either of the other two categories.

2. 'Bright': These are high-surface-brightness ghosting artifacts that come from multiple reflections off the DECam focal plane and the C4 or C5 lenses [Kent, 2013]. They are usually relatively small in size and circular or elliptical in shape. They have more distinct borders and are considerably brighter compared to the following category.

3. 'Faint': These are lower-surface-brightness ghosting artifacts that come from multiple reflections between the focal plane and the C3 lens or filter, or internal reflections off of the faces of the C3, C4, and C5 lenses [Kent, 2013]. They are circular or elliptical in shape and are usually larger in size and significantly fainter than 'Bright' ghosts.

In Fig. 5.1, we present two examples of DECam images that contain ghosts and scattered-light artifacts, along with the annotated ground truth masks. We trained the Mask R-CNN for these three distinct categories due to their significant morphological difference.

In total, our dataset contains 1566 'Rays', 2197 'Bright', and 2949 'Faint' artifact instances. In Fig. 5.2, we present the distribution in size (area) of these three ghost categories.

[5] `https://www.robots.ox.ac.uk/~vgg/software/via/`

[6] Note that not every author annotated the same number of images; six of us annotated 200 images and two of us annotated 400 images.

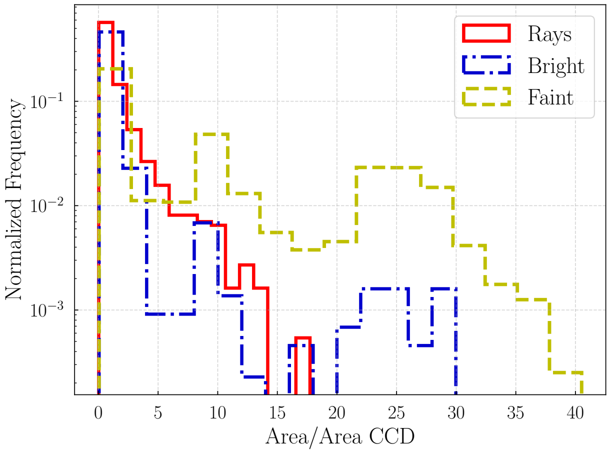
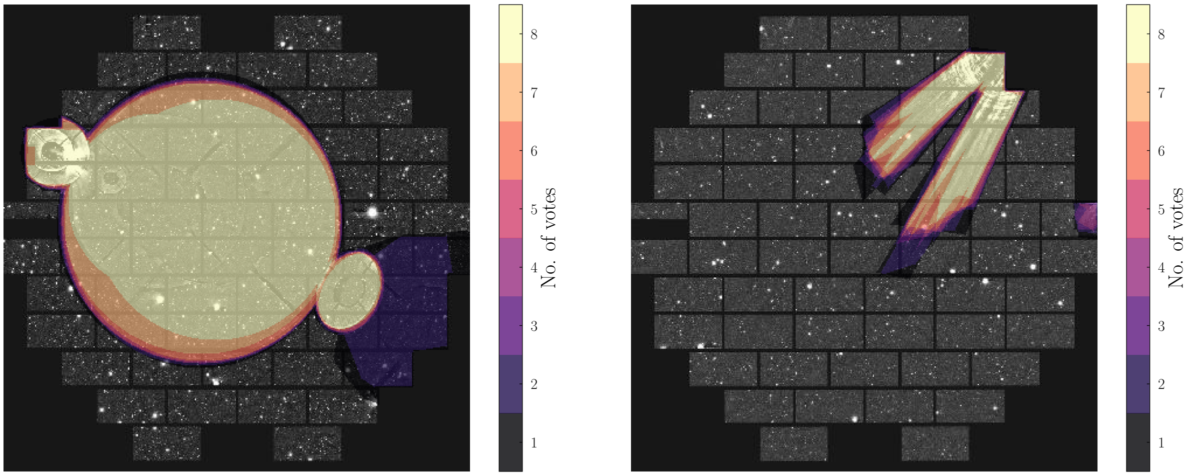


Figure 5.2: Histograms of the distribution in size (area) of the three artifact types presented in this work. The areas are quoted as multiples of the area of a single CCD.

The area of each ghosting artifact is presented as a fraction of the area of a single DECam CCD (area of artifacts in pixel over area of a CCD in pixels). Most 'Rays' have an area that covers fewer than 10 CCDs. 'Bright' ghosts are also relatively small in size, with a few spanning more than a couple of CCDs. On the other hand, 'Faint' ghosts are large in size, with a significant fraction of them covering an area of 20–30 CCDs. Many images contain multiple ghosts or scattered-light artifacts.

By measuring the surface brightness of a small subset of this dataset we find that 'faint' ghost have surface brightnesses that lie in the same regime as those of the DES low-surface-brighness galaxies [24.2–26.5 mag/arcsec^2 Tanoglidis et al., 2021a], consistent with the results of Slater et al. [2009]. Artifacts of the type 'Rays' and 'bright' are typically 1–2 magnitudes brighter.

We note that the ghosting and scattered-light artifacts do not always have clear boundaries (especially those of type 'Rays') and that the distinction between 'Bright' and 'Faint' ghosts is not always well defined. For that reason we expect some disagreement between the human annotators in the extent and shape of the ground truth masks and in the assigned labels.



(a)

(b)

Figure 5.3: Masks created by the eight different annotators (overlaid on top of each other) for the same two images presented in Fig. 5.1. The colors indicate the number of annotators that have labeled a given pixel as containing a ghost, from dark purple (one annotator) to light yellow (all the eight annotators).

In Fig. 5.3, we overlay the masks generated by all eight annotators for the same two DECam images presented in Fig. 5.1. The colors correspond to the number of annotators that have labeled the region as containing an artifact; dark purple corresponds to fewer votes, while light yellow corresponds to more votes. We do not distinguish between the different artifact types in this image.

The right panel of Fig. 5.3 shows a significant variation in the masks created by the different annotators for the 'Rays'. The left panel shows generally good agreement between the different annotators for the most prominent ghosts in the image; however, there is a large area on the right of the image that is labeled by only two annotators. We discuss the agreement between the human annotators in more detail in 5.6. In Section 5.4, we demonstrate that the Mask R-CNN is able to out-perform conventional algorithms even in the presence of the label noise introduced by disagreements in the existence, mask region, and classification of artifacts by individual annotators. Reduction in label noise from more uniform annotation could improve the performance of the algorithm in the future.

5.3 Methods

We use Mask R-CNN [He et al., 2017], a popular, state-of-the art instance segmentation algorithm, to detect and mask ghost and scattered-light artifacts.

Mask R-CNN is a powerful and complex algorithm, a recently-developed, popular model in the rapidly-advancing field of computer vision. It is part of a series of object detection models, collectively known as the R-CNN family. It builds upon many deep learning and computer vision techniques; we refer the reader to Weng [2017] for a detailed description of the R-CNN family.

Instance segmentation [e.g., for a review, Mueed Hafiz and Mohiuddin Bhat, 2020] combines the functions of object detection and image segmentation algorithms. Object detection [e.g., for a review, Zhao et al., 2018] is an active area of research in computer vision, with the goal of developing algorithms that can find the positions of objects within an image. Semantic segmentation [e.g., for a review, Minaee et al., 2020] on the other hand refers to the problem of pixel-level classification of different parts of an image into pre-defined categories. Instance segmentation is used to simultaneously detect objects in an image and to create a segmentation mask for each object.

A schematic description of the Mask R-CNN workflow is presented in Fig. 5.4. In the first stage of the model, the input images are fed into a pre-trained deep CNN — such as VGG [Simonyan and Zisserman, 2014] or ResNet [He et al., 2015] — also called the *backbone* network. The last, fully connected, classification layers of this network have been removed, and thus its output is a feature map. This feature map[7] is passed into the Region Proposal

[7] In practice, most Mask R-CNN implementations – like the one we are using in this work – use a Feature Pyramid Network [FPN; Lin et al., 2016] on top of the backbone. The FPN combines low-level features extracted from the initial stages of the backbone CNN with the high-level feature map output of the last layer. This improves the overall accuracy of the model, since it better represents object at multiple scales.

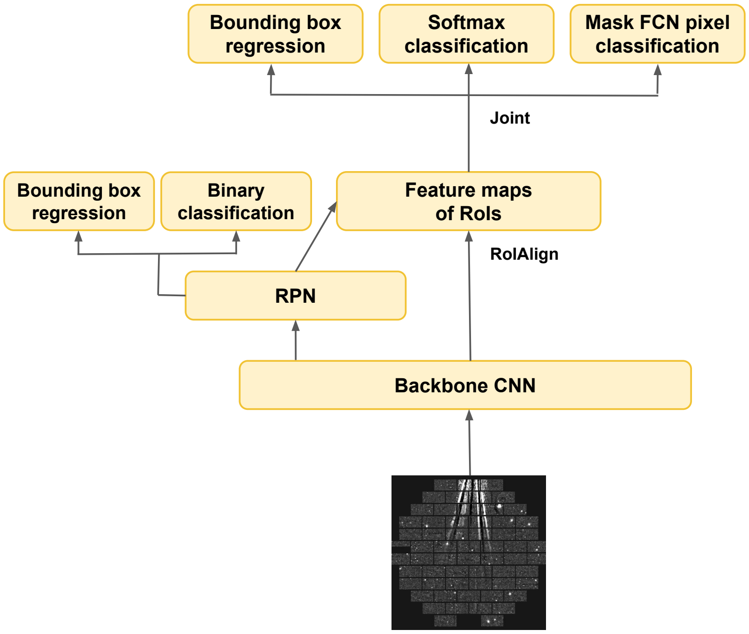


Figure 5.4: High-level schematic overview of the Mask R-CNN model. Figure adapted from Weng [2017].

Network (RPN) to produce a limited number of Regions of Interest (RoIs) to be passed to the main network – i.e., candidate regions that are most likely to contain an object.

The RPN is a simple CNN that uses a sliding window to produce a number of *anchor boxes* – boxes of different scales and aspect ratios – at each position. When training the RPN network, two problems are considered — classification and regression. For classification, the algorithm considers the possibility that there is an object (without considering the particular class) that fits inside an anchor box. For regression, the best anchor box coordinates are predicted. The anchor boxes with the highest object-containing probability scores are passed as RoIs in the next step. The loss of the RPN network is composed of a binary classification loss, $L_{\mathrm{RPN,cls}}$, and a bounding box regression loss, $L_{\mathrm{RNP,bbox}}$, such that $L_{\mathrm{RPN}} = L_{\mathrm{RPN,cls}} + L_{\mathrm{RNP,bbox}}$.

Each of the proposed RoIs has a different size. However, the fully connected networks used for prediction require inputs of the same size. For that reason, the RoIAlign method is used to perform a bilinear interpolation on the feature maps within the area of each RoI and output the interpolated values within a grid of specific size, giving fixed-size feature maps of the candidate regions.

Finally, these reshaped regions are passed to the last part of the Mask R-CNN that performs three tasks in parallel. A softmax classifier learns to predict the class of the object within the RoI; the output is one of the $K+1$ classes, where K are the different possible object types (L_{cls} loss), plus one background class. A regressor learns the best bounding box coordinates (L_{bbox} loss). Finally, the regions pass through a Fully Convolutional Network (FCN) that performs semantic segmentation (L_{mask} loss), i.e. a per-pixel classification, that creates the masks. The total loss of this Mask R-CNN part is thus $L_{\text{tot}} = L_{\text{cls}} + L_{\text{bbox}} + L_{\text{mask}}$.

The *DeepGhostBusters* algorithm is the Mask R-CNN implementation by Abdulla [2017], trained on our manually annotated dataset of ghosting and scattered-light artifacts. This code is written in Python using the high-level `Keras`[8] library using a `TensorFlow`[9] backend. We use the default 101-layer deep residual network (ResNet-101; He et al. 2015) as the backbone convolutional neural network architecture.

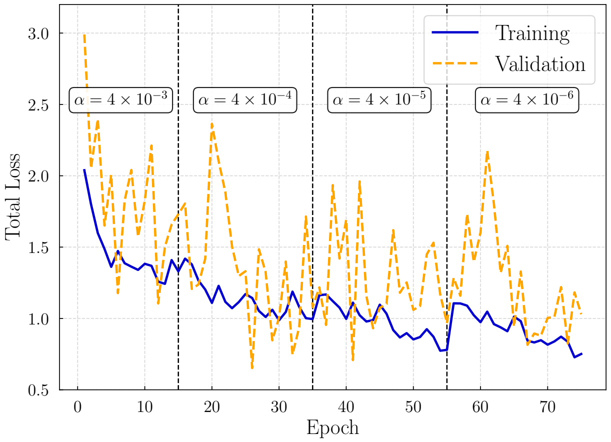


Figure 5.5: Total loss of the Mask R-CNN model as function of the training epoch. The training is performed using a progressively smaller learning rate, α.

[8] `https://keras.io/`

[9] `https://www.tensorflow.org/`

Before training, we randomly split the full dataset of 2000 images into a training set (1400 images), a validation set (300 images), and a test set (300 images). The annotation process was performed before this random split. Such a random split is generally important in machine learning problems for these three sets to be representative of the general population, but it becomes even more important here because different human annotators have different annotation styles. This could create significant systematic differences between the ground truth masks in the datasets if not properly randomized.

In computer vision problems where only a small training set is available, it is common to use *transfer learning* to improve results (for recent reviews, see Wang and Deng 2018 and Zhuang et al. 2019). Transfer learning is a process where the weights of a network that has already been trained for one detection task are used for a different, but related, task, usually with some further training. This speeds up the training process, reduces overfitting, and produces more accurate results. Here, we initialize the learning procedure (i.e., use transfer learning) using the weights learned from training Mask R-CNN on the Microsoft Common Objects in Context (MS COCO) dataset[10] [Lin et al., 2014], which consists of $\sim$ 330k images ($\sim$ 2.5M object instances) of 91 classes of common or everyday objects.

To reduce overfitting, we employ data augmentation [e.g., Shorten and Khoshgoftaar, 2019], by performing geometric transformations on the images and the masks. Specifically, we randomly apply zero to three of the following transformations:

- Rotation of the image and the masks by 270 degrees.
- Left-right mirroring/flip of the images and masks.
- Up-down mirroring/flip of the images and masks.

We re-train our model using stochastic gradient descent to update the model parameters. Similarly to what was proposed in Burke et al. [2019], the training is performed in different

[10] `https://cocodataset.org/#home`

stages with progressively smaller learning rates, α, at each stage. This allows for a deeper learning and finer tuning of the weights, while minimizing the risk of overfitting.

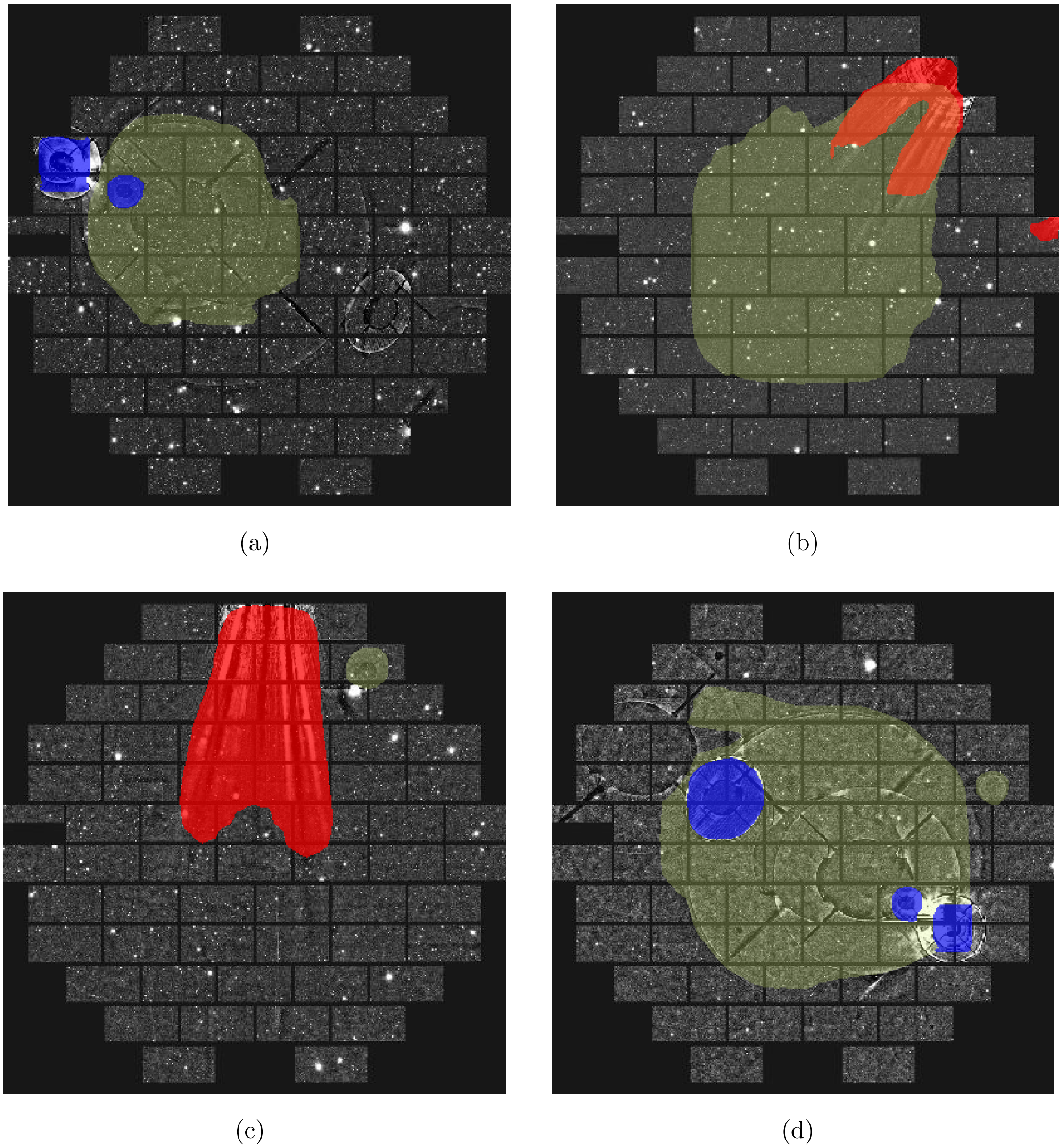
Specifically, in the first stage (15 epochs), we re-train the top layers only and use a learning rate of $\alpha = 4\times10^{-3}$. Then, we train all the layers with decreasing learning rates: 20 epochs at $\alpha = 4\times10^{-4}$, 20 epochs at $\alpha = 4\times10^{-5}$, and 20 epochs at $\alpha = 4\times10^{-6}$. In total, we trained the model for 75 epochs, after which overfitting occurs. In all stages (training, validation, test) we ignore detections with less than 80% confidence (`DETECTION_MIN_CONFIDENCE = 0.8`). We utilized the 25 GB high-RAM Nvidia P100 GPUs available through the Google Colaboratory (Pro version). The training took $\sim$ 4 hours to complete. The inference time is $\sim$ 0.34s per image to predict.

In Fig. 5.5, we present the total loss as a function of the training epoch for both training and validation sets. In 5.7, we show the training history for the individual components of the total loss.

5.4 Results

We use an independent DECam test set to evaluate the performance of the *DeepGhostBusters* Mask R-CNN in detecting and masking ghost and scattered-light artifacts. We use both custom metrics appropriate for the problem at hand and metrics commonly used in the object detection literature. We also compare the performance of *DeepGhostBusters* with the conventional Ray-Tracing algorithm. Finally, we test the classification performance of *DeepGhostBusters* when it is presented with a dataset that also contains images that lack any ghosts or scattered-light artifacts.

We first present the mask and class predictions of the *DeepGhostBusters* Mask R-CNN model on four example images (Fig. 5.6). The two top panels, (a) and (b), correspond to the same images whose ground truth masks were presented in Fig. 5.1. As in Fig. 5.1, the different colors represent the different ghosting artifact types: red for ‘Rays’, blue for


(a)

(b)

(c)

(d)

Figure 5.6: Predicted masks on four example images that contain the three distinct artifact types — scattered-light ‘Rays’ (red), ‘Bright’ ghosts (blue), and ‘Faint’ ghosts (yellow). The top panels correspond to the images presented in Fig. 5.1.

‘Bright’, and yellow for ‘Faint’.

These examples demonstrate both the successes and failures of our model. For example, in panel (a) the model has successfully masked most of the central ‘Faint’ ghost, but it has also missed a significant part of its periphery, as well as the prominent ghost on the right of the image. Furthermore, although it has successfully deblended and separately masked the small ‘Bright’ ghost that is superimposed on the larger ‘Faint’ one, it has only partially masked the one on the left. Panel (b) presents a characteristic example of a false positive detection: predicting a mask for a ‘Faint’ ghost that is not there. The Mask R-CNN has predicted a mask that successfully covers most of the prominent ‘Rays’-type artifact; it is also able to detect the smaller ‘Rays’ on the right. However, it has also erroneously masked a large central region (containing the edges of the rays) as a ‘Faint’ ghost. Panels (c) and (d) present mostly successful detections, although with some false negatives, as the undetected ‘Faint’ ghost on the top-left corner of panel (d). We next formally quantify and evaluate the performance of the Mask R-CNN model and compare it with that of the conventional Ray-Tracing algorithm.

5.4.1 CCD-based metrics

The DECam focal plane consists of 62 science CCDs. The conventional Ray-Tracing algorithm used by DES flags affected focal plane images on a CCD-by-CCD basis — i.e., if a CCD contains a ghost or scattered-light artifact, the entire CCD is removed from processing. To compare the performance of the Mask R-CNN to the conventional algorithm, we develop metrics that are based on whether a CCD contains a ghost or scattered-light artifact.

The resulting metrics depend on the size of individual artifacts. This is important for the problem at hand: for example, we care how well the algorithm can mask a larger ghost compared to a smaller one. At the same time, given the challenges of this problem (e.g., overlapping sources and borders that are not always well defined), assessing the performance at the CCD-level can be more robust than comparisons at the more granular pixel level.

We consider each image as a 1D array of length 62 with entries 0 and 1, where 0 corresponds to CCDs that do not contain a ghost, and 1 corresponds to those that do contain a ghost. For a batch of M images containing $N = 62 \times M$ CCDs, we define the number of true positives (N^{TP}), true negatives (N^{TN}), false positives (N^{FP}), and false negatives (N^{FN}). Then, we define the CCD-based precision (purity) and recall (completeness) as:

$$\text{Precision}_{CCD} = \frac{N^{TP}}{N^{TP} + N^{FP}}, \tag{5.1}$$

$$\text{Recall}_{CCD} = \frac{N^{TP}}{N^{TP} + N^{FN}}. \tag{5.2}$$

Based on the science case of interest, one may want to maximize either the precision or the recall. For example, for systematic studies of low-surface-brightness galaxies, high recall for ghosts and scattered-light artifacts may be preferred at the expense of some loss in precision.

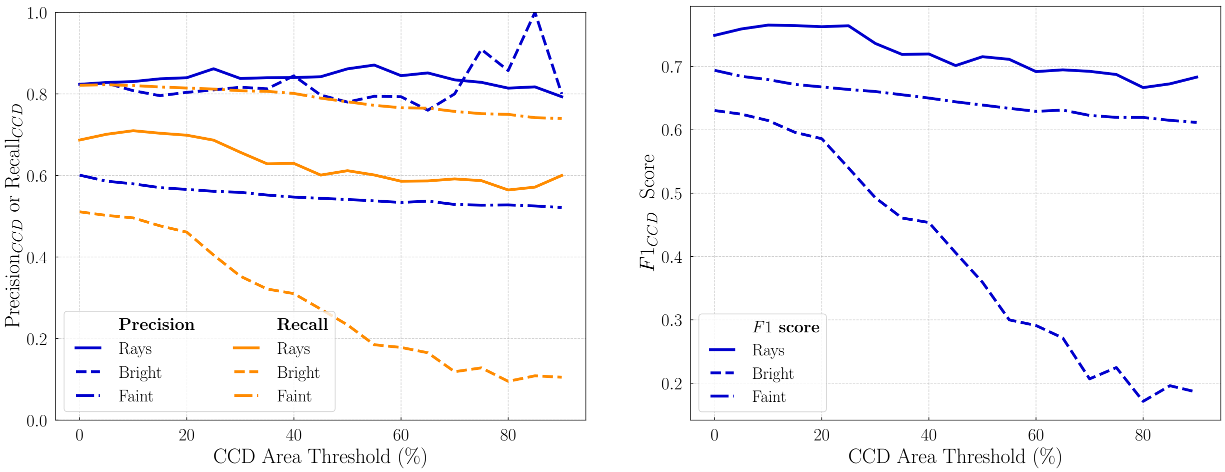
One approach to assessing the trade-off between precision and recall is to define the $F1$ score, which is the harmonic mean of the precision and recall,

$$F1_{CCD} = 2\left(\frac{\text{Precision}_{CCD} \cdot \text{Recall}_{CCD}}{\text{Precision}_{CCD} + \text{Recall}_{CCD}}\right) \tag{5.3}$$

Note that we can use the above definitions for each type of artifact individually or for all artifact types combined.

The above metrics are based on the notion of a binary classification of CCDs as affected by ghosts or scattered-light artifacts. In reality, the ghosts and scattered-light artifacts will only cover some fraction of the CCD area. Thus, we define a threshold for the fraction of the CCD area that must be covered for the CCD to be classified as affected. In 5.8 we present examples of masked CCDs for two different area thresholds. Here, we study how the performance metrics change as a function of that threshold.

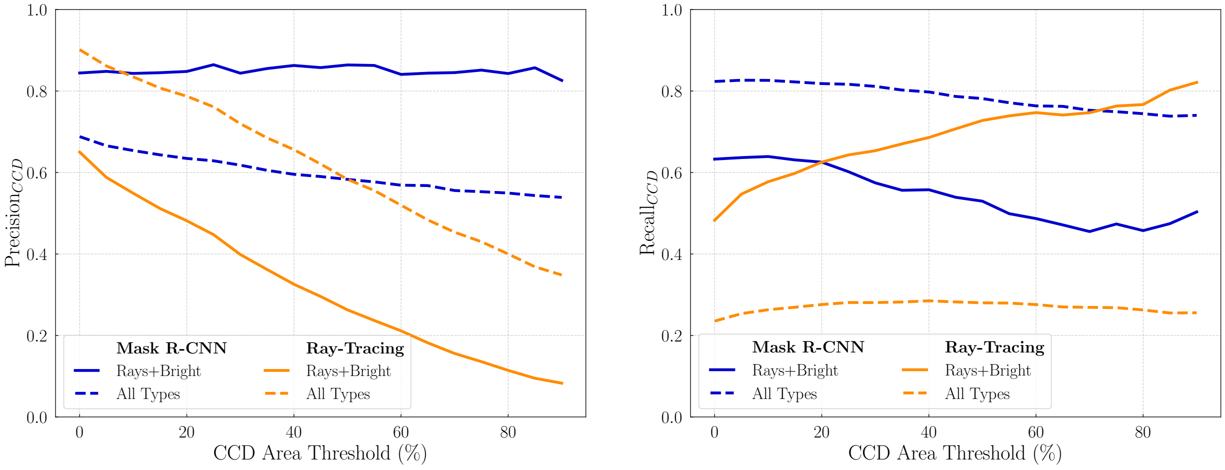
In panel (a) of Fig. 5.7, we present precision and recall as a function of the CCD area threshold for the three artifact categories individually. These metrics are related to the number of CCDs (as opposed to the number of artifacts) that were correctly or incorrectly classified. Therefore, the differences we observe between the artifact types depend on the



(a)

(b)

Figure 5.7: CCD-based (a) precision (blue) and recall (orange), and (b) $F1$ score as a function of the CCD area threshold (see main text) from the Mask R-CNN model and for the three ghosting artifact categories (‘Rays’, ‘Bright’, and ‘Faint’).



(a)

(b)

Figure 5.8: CCD-based (a) precision and (b) recall of the Mask R-CNN model (blue lines) and the Ray-Tracing algorithm (orange lines). We consider both the combination of all types of artifacts (solid lines) and the combination of ‘Rays’+‘Bright’ (dashed lines).

different sizes of the artifacts. For example, as we have seen (Fig. 5.2), ‘Faint’ ghosts tend to cover $\sim 10-30$ CCDs, while ‘Bright’ ghosts are significantly smaller, covering $\sim 1-3$ CCDs. Thus, the classification and masking of a single large ‘Faint’ object has a greater effect on the metrics than the detection of two or three ‘Bright’ ghosts.

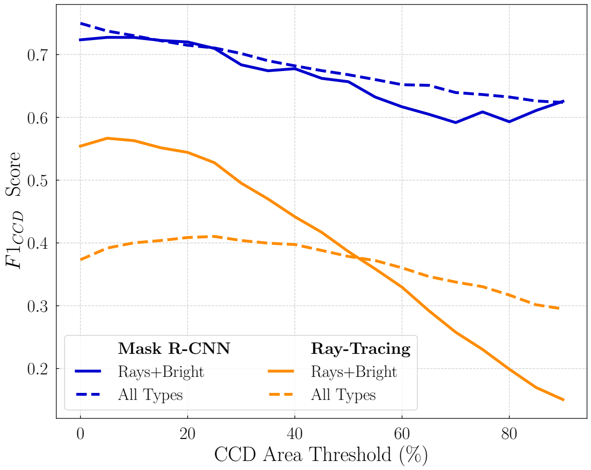


Figure 5.9: CCD-based $F1$ scores for the same models and ghost type combinations as in Fig. 5.8.

There are a few interesting trends to notice in this figure. First, for 'Rays' and 'Bright' ghosts, the precision is higher than the recall and almost constant as the area threshold changes. The high precision score ($\sim 80\%$) for these categories is easy to understand: these are the most distinct and prominent ghosts, and thus it is hard for a CCD with a 'Faint' ghost (or for a CCD without a ghost) to be mistaken as containing either of these types of artifacts.

Second, the recall score for 'Rays' is $\sim 70\%$ and constant as a function of the threshold. The recall score for 'Bright' ghosts greatly degrades with area threshold and it is generally low (less than 50%). 'Bright' ghosts are relatively small, only partially covering the CCDs that contain them; as we increase the area threshold, only a few such ghosts can pass it.

A third interesting point is that 'Faint' ghosts have higher recall than precision, in contrast to the two other categories. 'Faint' ghosts are usually large: even though some may go undetected, the largest cover many CCDs and are usually detected (at least partially), thus pushing the CCD-based recall (completeness) to higher values. On the other hand, some 'Bright' ghosts, especially those with a significant overlap with larger 'Faint' ghosts can be misclassified as 'Faint', leading to a lower precision.

In panel (b) of Fig. 5.7, we present the $F1$ score as a function of the CCD area threshold. The $F1$ score (see Eq. (5.3)) is useful as a way to compare the performance of the classifier for different ghost types using a single metric. As we can see in this figure, the Mask R-CNN performs best in finding CCDs containing 'Rays', while CCDs containing 'Faint' ghosts are identified with higher efficiency than CCDs containing 'Bright' ghosts.

In practice, we are interested in the ability of the *DeepGhostBusters* Mask R-CNN to detect combinations of ghosts and scattered-light artifacts. We present the CCD-based precision and recall as a function of the area threshold in Fig. 5.8 (panels (a) and (b), respectively); we also present the $F1$ score in Fig. 5.9 for the two combinations, 'Rays'+'Bright' (solid blue lines) and 'Rays'+'Bright'+'Faint' (all ghost types, dashed blue line).

We chose this combination for two reasons: first, it allows a fairer comparison with the Ray-Tracing algorithm, which is not tuned for very low-surface-brightness ghosts (see next subsection); second, for a practical application, we may not need to reject CCDs containing very faint ghosts, because these have little influence on the surface brightness of real sources and can be effectively deblended.

5.4.2 Comparison with the Ray-Tracing algorithm

Next, we compare the performance of our Mask R-CNN model in detecting ghost-containing CCDs to that of the Ray-Tracing algorithm. We note a few details of this comparison:

- The test dataset consists only of images known to contain at least one ghost or scattered-light artifact.

- When plotting metrics as a function of the CCD area threshold, this threshold is applied only to the ground-truth masks. This accounts for the fact that we only have predictions from the Ray-Tracing algorithm on a CCD-by-CCD basis.

- The available output from the Ray-Tracing algorithm does not distinguish between the different artifact categories. Furthermore, the Ray-Tracing algorithm applies a threshold to the predicted surface-brightness of artifacts, and thus is not optimized to detect 'Faint' ghosts. For that reason we exclude 'Faint' ghosts when evaluating metrics to compare performance between the Ray-Tracing and Mask R-CNN algorithms.

We plot the CCD-based precision and recall (Fig. 5.8) and $F1$ score (Fig. 5.9) resulting from the Ray-Tracing algorithm (orange lines) and Mask R-CNN (blue lines), as a function of the ground truth threshold area. We consider two categories of artifacts selected based on the ground truth masks: all ghost types combined (solid lines) and the combination of 'Rays'+'Bright' ghosts (dashed line).

We first consider the limit of zero percent CCD area threshold: a single pixel of an artifact has to be in the CCD to be classified as ghost-containing. The Ray-Tracing algorithm achieves a high precision score, which, for the case when the combination of all ghost types is considered, is higher than that from the Mask R-CNN for the same case ($\sim$ 0.9 vs. $\sim$ 0.7). However, for the same case the recall is much lower ($\sim$ 0.8 vs. $\sim$ 0.3). In other words, Ray-Tracing produces results high in purity but low in completeness. When the combination of only 'Rays'+'Bright' ghosts is considered, both the precision and the recall from the *DeepGhostBusters* Mask R-CNN model are significantly higher than those from the Ray-Tracing algorithm.

Fig. 5.8 shows that precision decreases, while recall increases as a function of the CCD area threshold for both artifact combinations. As we increase the threshold, fewer CCDs are labeled as containing artifacts and thus the purity decreases while the completeness increases.

The $F1$ score, which combines precision and recall, demonstrates that the performance of the Mask R-CNN model is significantly higher than that of the Ray-Tracing algorithm for all area threshold values and for both artifact combinations (Fig. 5.9).

To facilitate the numerical comparison of the performance of the algorithms, we present

Table 5.1: CCD-based evaluation metrics (precision, recall, $F1$ score) for the Mask R-CNN and Ray-Tracing algorithms, at 0% CCD area threshold.

Metric	Mask R-CNN		Ray-Tracing	
	Rays+Bright	Rays+Bright+Faint	Rays+Bright	Rays+Bright+Faint
Precision	84.3%	68.7%	64.7%	89.9%
Recall	63.6%	82.5%	48.4%	23.5%
$F1$ score	72.5 %	75.0%	55.4%	37.3%

in Table 5.1 the values of the different metrics for the two models, at a one pixel (> 0%) CCD area threshold, for both algorithms. The results for both artifact category combinations ('Rays'+'Bright' and 'Rays'+'Bright'+'Faint') are presented.

5.4.3 Standard object detection evaluation metrics

We now examine the Average Precision [AP; Everingham et al., 2010], a metric that is commonly used by the computer vision community to assess the performance of object detection algorithms. The AP is defined as the area under the Precision-Recall (PR) curve:

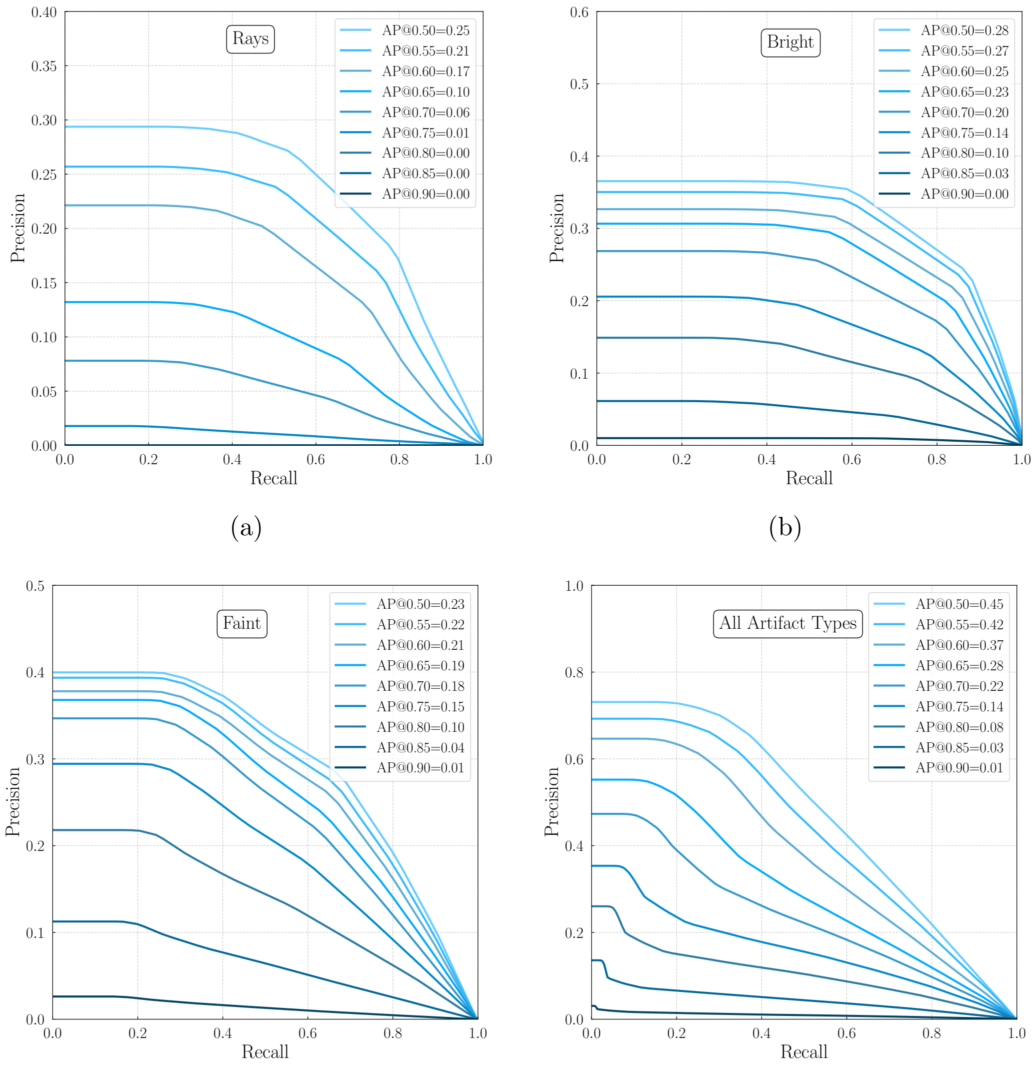
$$\mathrm{AP} = \int_0^1 p(r)dr, \tag{5.4}$$

where $p(r)$ is the precision, p, at recall level r. In practice, an 11-point interpolation method is used, and the AP score is calculated as:

$$\mathrm{AP} = \frac{1}{11} \sum_{r_i \in R} \tilde{p}(r_i), \tag{5.5}$$

where $\tilde{p}$ is the maximum precision at each recall bin and $R = \{0.0, 0.1, \ldots, 1.0\}$. Precision and recall are defined using the common formulae (Eqs. 5.1 and 5.2), but here the number of true positives, true negatives etc. refer to detections of individual artifacts and not single CCDs.

To define the detection of an artifact, we introduce the concept of the Intersection over



(a)

(b)

(c)

(d)

Figure 5.10: Precision-Recall curves and Average Precision scores at different IoU threshold values in the range $0.50 - 0.90$. We show these metrics for the different ghost types in this work (‘Rays’-‘Bright’-‘Faint’), and for all ghost types, combined.

Union (IoU; also known as the Jaccard index; Jaccard 1912), which quantifies the overlap between the masks of the ground truth and the prediction. As the name suggests, it is defined as the ratio of the area of the intersection of the predicted mask (pm) and the ground truth (gt) mask over the area of the union of the predicted and ground truth masks:

$$\text{IoU} = \frac{\text{area of intersection}}{\text{area of union}} = \frac{area(gt \cap pm)}{area(gt \cup pm)}. \tag{5.6}$$

An IoU threshold is then used to determine if a predicted mask is a TP, FP, or FN. It is common to evaluate the AP score at different IoU levels, and we denote the AP at a IoU threshold β as "AP@β".

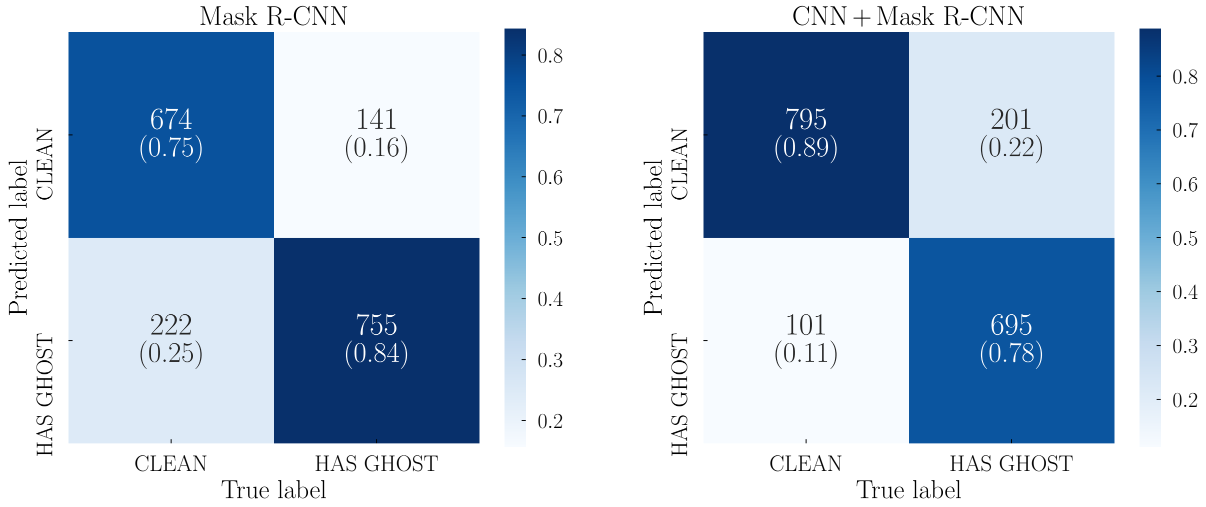
By calculating the PR curves and the AP score at different IoU threshold and for the different artifact categories, we evaluate the performance of the Mask R-CNN model for different artifact categories. Furthermore, by determining how AP varies with increasing IoU, we evaluate the agreement between the true and predicted masks.

In Fig. 5.10, we present the PR curves and the corresponding AP scores for IoU thresholds in the range $0.5 - 0.9$ (with step size 0.05) for the three artifact types in panels (a)-(c), individually, and for all artifact types combined in panel (d). We find that 'Bright' ghosts are most easily detected by the Mask R-CNN, while 'Faint' ghosts are the most challenging to detect — in agreement with our expectations. Furthermore, for 'Rays', the AP decreases rapidly with increasing IoU threshold: the model struggles to accurately reproduce the ground truth masks for these artifacts. This is expected, because these artifacts do not have clear boundaries, as demonstrated by variation in the mask regions defined by the human annotators.

In that section, we have shown that the Mask R-CNN algorithm is superior to the Ray-Tracing in detecting CCDs affected by ghosts or scattered-light artifacts.

5.4.4 Using Mask R-CNN to classify ghost-containing vs. ghost-free images

So far, the images used for training and testing the performance of the Mask R-CNN model were known (by visual inspection) to contain at least one ghost or scattered-light artifact. However, most DECam images do not contain prominent ghost or scattered-light artifacts, and thus they systematically differ from those used to train and test the model. Such differences may result in a large number of false positive detections (e.g., real astronomical sources, especially large and bright objects) or systematically failing to detect ghosts in some



(a)

(b)

Figure 5.11: (a) Confusion matrix of predictions of the Mask R-CNN model on a dataset containing an even number of ghost containing and clean images. An image is predicted to 'have ghost' if even a single ghost is detected in that image by the Mask R-CNN model. (b) Confusion matrix of the predictions of the combined CNN + Mask R-CNN model (CNN model from Chang et al. [2021]). An image is said to 'have ghost' if and only if both the CNN and the Mask R-CNN models agree on that (otherwise the prediction is 'clean').

images — for example, images that contain only very small or very faint ghosts.

To test the performance of the Mask R-CNN on images that do not contain ghosts, we use a set of 1792 images with an equal number of ghost-free and ghost-containing images. This set of images is independent of the 2000 images used to train, validate, and test the Mask R-CNN model. They constitute the test set used in Chang et al. [2021]. For this dataset, the ground truth labels refer to the presence of a ghost in the image — not the number of ghosts or the regions affected by ghosts.

We run Mask R-CNN on this dataset: when the algorithm predicts the existence of even a single ghost or scattered-light artifact in the image, we assign a predicted label 'HAS GHOST' to that image. Otherwise the assigned predicted label 'CLEAN'. The confusion matrix resulting from this process is shown in Fig. 5.11. The accuracy is 79.7%, the precision is 77.3%, and the recall is 84.3%. Both the numbers of false positive and false negative cases are high: false positives occur at $\sim 22.7\%$ of the total number of images classified as positives, and the false negatives occur at $\sim 17.3\%$ of the number of images classified as clean.

However, visual inspection of false positive examples and the predicted masks revealed that most contain objects or exhibit features similar to those found in ghost-containing images. These include bright streaks from artificial Earth-orbiting satellites (mimicking 'Rays'), low-surface-brightness emission from Galactic cirrus, images with poor data quality (due to cloud coverage that diffuses starlight), or large resolved stellar systems (e.g., dwarf galaxies and globular clusters). These are very similar to the cases of false positives returned by the CNN classifier in Chang et al. [2021]. Similarly, most of the false negatives contain very small and faint ghosts (and usually each image contains only one such ghost) that could have been easily missed even by a human annotator.[11] Thus, we conclude that the false positives/negatives are qualitatively different from the true positives/negatives. and that – in practice – the Mask R-CNN is much better in classifying images that contain unusual and/or problematic areas, compared to what one would naively assume from the confusion matrix (Fig. 5.11).

We note that in practical applications of Mask R-CNN, we can reduce the number of false positives by first applying the CNN classifier presented in Chang et al. [2021], and then applying the Mask R-CNN only to those images that are identified as containing ghosts or scattered-light artifacts. The results of this process on the test dataset are presented in panel (b) of Fig. 5.11. We find that we are able to reduce the number of false positives to less that of the Mask R-CNN alone, but at the expense of increasing the number of false negatives. This combined model has an overall accuracy of 83.1%, precision of 87.3%, and recall of 75.6%. Because of this trade-off, the final decision of pre-processing with a CNN depends on the particular problem and whether we are willing to reject otherwise real astronomical objects (false positives) or to have residual ghost and scattered-light artifacts (false negatives).

[11] Examples of false positives and false negatives can be found in 5.9.

5.5 Summary and Conclusions

In this work, we applied a state-of-the art object detection and segmentation algorithm, Mask R-CNN, to the problem of finding and masking ghosts and scattered-light artifacts in astronomical images from DECam. The effective detection and mitigation of these artifacts is especially important for low-surface-brightness science [e.g., Slater et al., 2009], an important target of future surveys.[12] Given the sheer volume of data generated by current and upcoming surveys, automated methods must be used for the identification of these artifacts.

In this paper, we compared the performance of the Mask R-CNN algorithm to two previous approaches, each of which has benefits and limitations. First, the conventional Ray-Tracing algorithm currently used by DES identifies individual CCDs affected by ghosting or scattered-light artifacts. This is a predictive model that does not use the actual imaging data to detect artifacts. Thus, its performance is limited by the accuracy of the optical model and external catalogs of bright stars, and it fails to detect a significant number of artifacts. Second, we compared to a relatively standard CNN [Chang et al., 2021], which does not depend on modeling the optical processes that lead to the generation of artifacts or on external catalogs of bright astronomical objects. Furthermore, it separates “ghost-containing” from “clean” images with high accuracy. However, as a classifier, it does not identify the affected subregion(s) within the image: if used without further investigation, it can lead to the rejection of useful information from non–affected parts of the image.

The Mask R-CNN approach presented in this work has the benefits of a deep learning approach — i.e., it does not depend on physical modeling, except through that training data, themselves — that can predict the locations of ghosts and scattered-light artifacts, which can be used to create CCD- and pixel-level masks of the affected region of an image.

We compare the ability of Mask R-CNN in masking affected CCDs in *ghost-containing*

[12] See, for example, `https://sites.google.com/view/lsstgsc/working-groups/low-surface-brightness-science`

images with that of the Ray-Tracing algorithm. We find that the Mask R-CNN model has superior performance, as measured by the $F1$ score, which is the harmonic mean of the precision (purity) and the recall (completeness). These results hold across different CCD area thresholds and for the two combinations of the morphological classes discussed in this work — 'Bright'+'Rays' and 'Bright'+ 'Rays'+'Faint'. At the threshold of one pixel ($> 0\%$), for example, and for the combination 'Rays+Bright' the $F1$ score of the Mask R-CNN model is 72.5% as opposed to 55.4% of the Ray-Tracing algorithm.

One weakness of our method is that it produces a large number of false positives when presented with images that do not contain ghosts or scattered-light artifacts — although many of these false positives contain other types of artifacts or bright astronomical objects. We show that, to mitigate this problem, a CNN classifier similar to that discussed in Chang et al. [2021] can be used as a pre-processing step before the Mask R-CNN is applied to images that are predicted to contain ghosts or scattered-light artifacts. This process reduces the number of false positives by a factor of two and increases the number of false negatives, improving the overall accuracy. Alternatively, one can add more classes, by generating ground truth masks, for example of some of the most common types of false positives (e.g. satellite trails, or cirrus).

The results presented here highlight the promise of object detection and segmentation methods in tackling the identification of ghosts and scattered-light artifacts. Since deep learning models that are trained on one data set can be adapted to a new data set with many fewer examples through transfer learning, the *DeepGhostBusters* algorithm trained on DECam images can potentially be adapted and retrained to identify such artifacts in future surveys. Indeed, cross-survey transfer learning has already been shown to significantly reduce the need for large annotated datasets in deep learning-based classification cases [e.g., Domínguez Sánchez et al., 2019a, Khan et al., 2019, Tanoglidis et al., 2021a]. Additionally, these results indicate that such techniques are also promising for different, but related, problems, such as the the detection of artifacts from cosmic rays, satellite trails, etc. [e.g., Goldstein et al., 2015, Desai et al., 2016, Melchior et al., 2016, Zhang and Bloom, 2020, Román et al., 2020, Paillassa et al., 2020]. Such automated techniques can facilitate the efficient separation of artifacts from scientifically useful data in upcoming surveys like LSST.

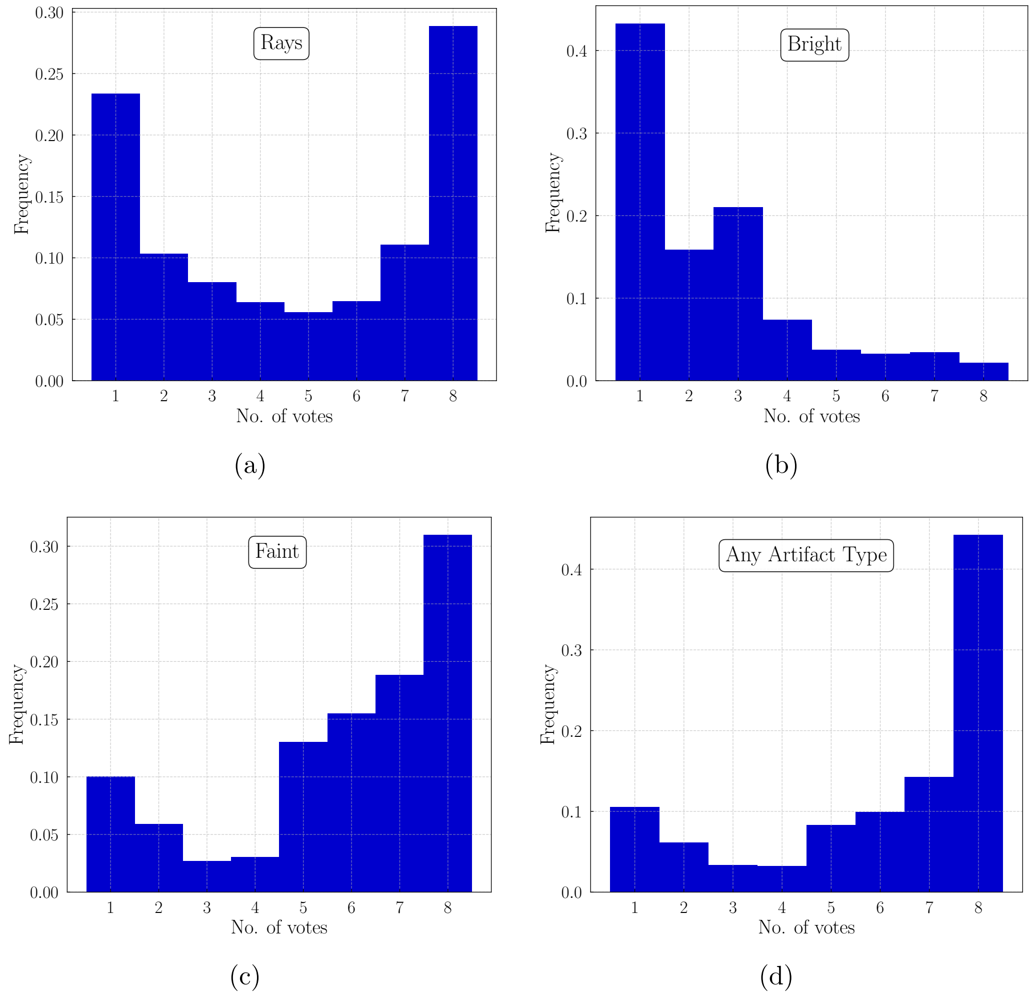
5.6 Human annotator agreement

As mentioned in Sec. 5.2.2, human annotators do not always agree on the mask boundaries and the artifact types. A significant disagreement may affect the performance of the Mask R-CNN, so we study extent of the disagreement in more detail, which may suggest avenues for improvement of the annotation process.

All eight annotators were given a common subset of 50 images that were randomly drawn from the full dataset described in Sec. 5.2.1. When an annotator creates a mask for a specific artifact, they give a 'vote' to the region covered by that mask. A second annotator will create a different mask around the same object. The pixels where there is an overlap between the two masks will receive two votes in total while the non-overlapping parts only one. The same process continues for all the eight annotators. The same region may receive multiple different classifications (e.g., votes for both 'Bright' and 'Faint' ghosts).

In Fig. 5.12, we present histograms of the distribution of the number of votes each pixel in the dataset received during the annotation process. We restrict it to pixels that have received at least one vote. We present the distributions for each artifact category separately in panels (a)-(c), and the case where we do not distinguish between different types in panel (d). A distribution that has a strong peak in the region of ~ 8 votes indicates that there is a very good agreement between the annotators.

The histogram for 'Rays' shows a strong bimodality, with many pixels receiving 8 votes , but also many pixels receive just 1–2 votes. These artifacts are distinct and bright, and hard to confuse with any one of the other two types. However, they do not have very clear



(a)

(b)

(c)

(d)

Figure 5.12: Distribution of the number of votes each pixel in our dataset has received as containing a ghost, from the eight annotators. We include only pixels that have received at least one vote. We present the distributions for each ghost type separately (panels (a)-(c)) and without distinguishing between the different types (panel (d)).

boundaries, so, while annotators agree on the bulk of the pixels affected by a ghost, they do not agree on the extent/edges of the masks they create.

The histogram of votes for ‘Bright’ artifacts, panel (b), presents a peak at the low end (1–3 votes). This can be explained by the fact that there is significant confusion about the class of some large ghosts, which most annotators classify as ‘Faint’, while a few classify as ‘Bright’. Since they are much larger compared to other typical ‘Bright’ ghosts, the distribution is dominated by the pixels belonging to these confusing artifacts.

Generally, there is a good agreement between the annotators when it comes to ‘Faint’ ghosts, with over 30% of the pixels having received the full eight votes. When not distinguishing between the different types of artifacts (panel (d)), we see very good agreement between the annotators in masking ghost-containing pixels, with $\sim 45\%$ of those pixels having received the maximum 8 votes, and an additional $\sim 25\%$ having received seven votes. Only $\sim 10\%$ of the pixels have received only one vote.

From the above discussion, we conclude that there is generally good agreement in the mask-creation process. Some confusion exists between ‘Faint’ and ‘Bright’ ghost types, because the distinction between the two is quite arbitrary. Some potential avenues for improvement are to consider these two categories as one, define more specific criteria for each class, or have multiple persons annotate the same images and assign each artifact to the class that receives the most votes.

5.7 Training History

In Fig. 5.5, we presented the total loss as a function of the training epoch (training history). The total loss, $L_{\rm tot}$, is the sum of the classification, bounding box, and mask loss (see Sec. 5.3). We present the training histories for these losses individually in Figs. 5.13, 5.14, and 5.15, respectively. As described in the main text, we train the model using progressively smaller learning rates for a finer tuning of the parameters. We stopped the process at 75 epochs due to overfitting thereafter.

5.8 Masking CCDs

To help the reader better understand how the imposed area threshold affects the number of CCDs classified as ghost-containing (Sec. 5.4.1), in this Appendix we present the predicted artifact masks and the affected CCDs for two different threshold levels, for the same images

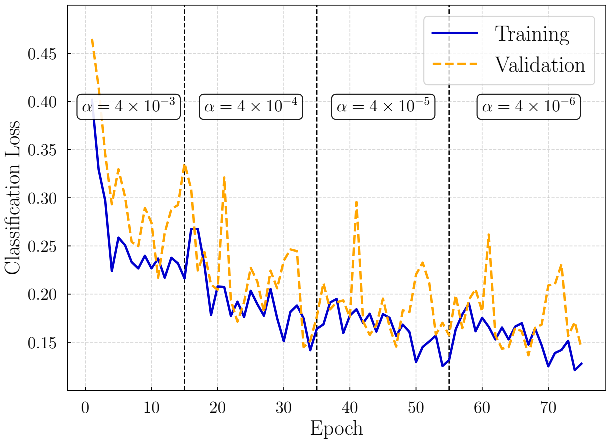


Figure 5.13: Classification loss as a function of the training epoch.

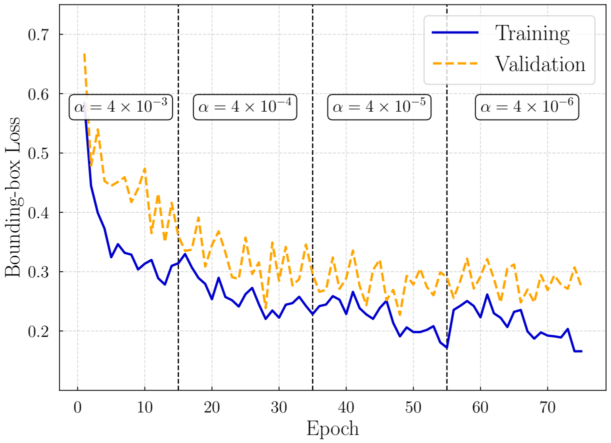


Figure 5.14: Bounding box loss as a function of the training epoch.

presented in the top row of Fig. 5.6.

Specifically, in the panels (a) and (c) of Fig. 5.16 we map (in blue) those CCDs that are classified as ghost-containing when even a single pixel of the predicted artifact mask lies within that CCD ($> 0\%$ threshold). In panels (b) and (d) we show, for the same images, the CCDs masked as ghost-containing when at least half of area of the CCD has to be covered by an artifact to be classified as such (50% threshold). To make the comparison easier, we overlay (yellow contours) the mask predictions of the Mask R-CNN model, without distinguishing between the different ghosting and scattered-light artifact types.

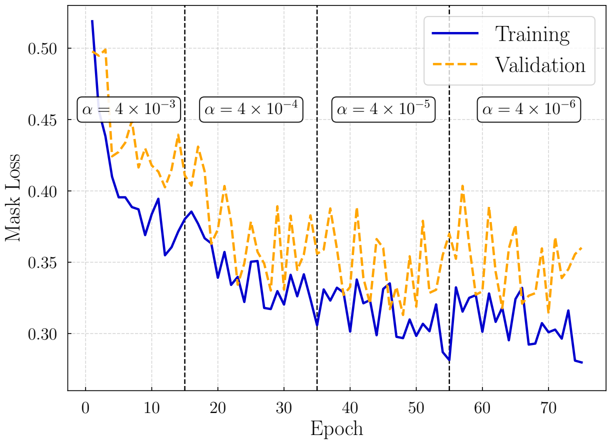


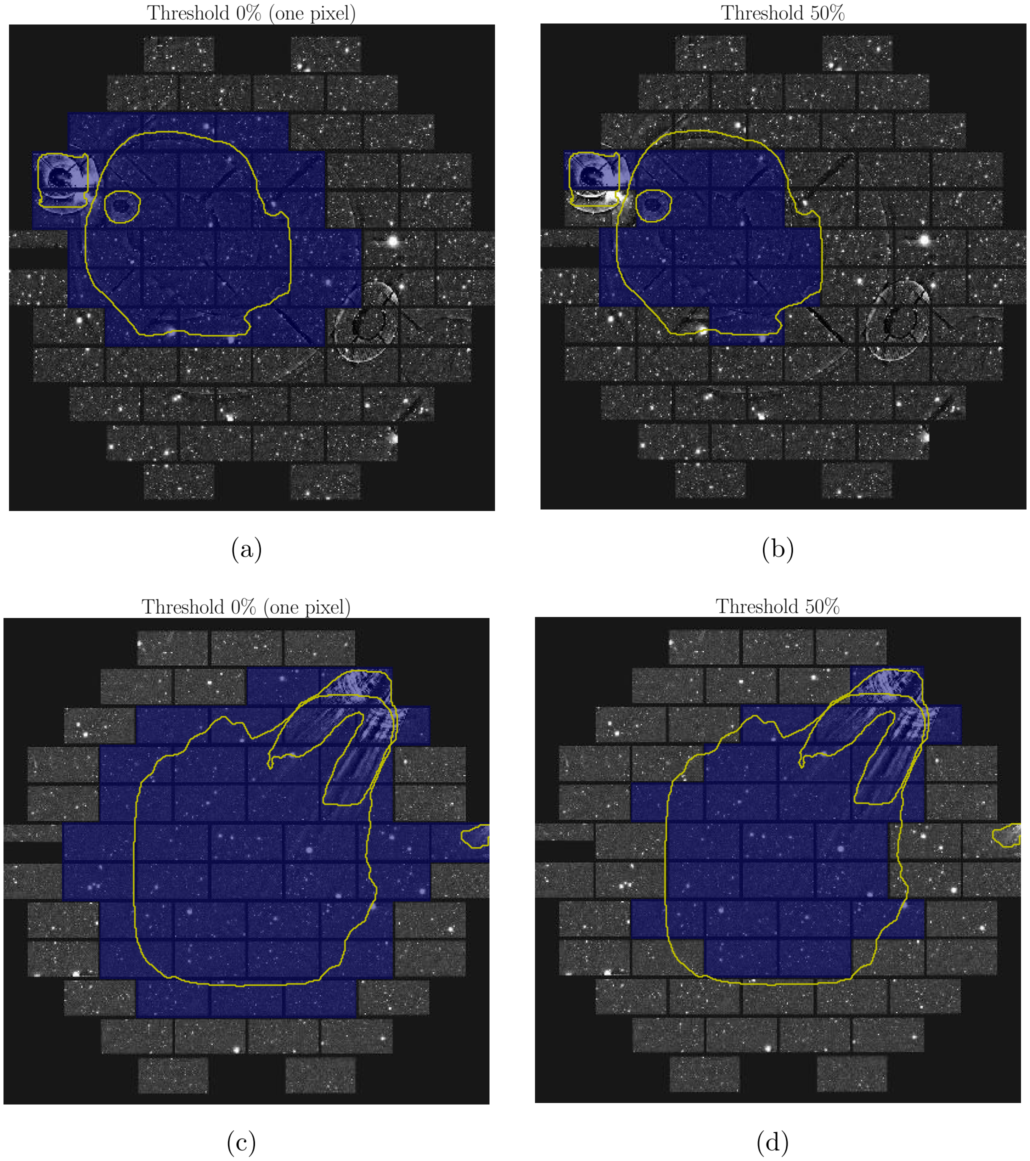
Figure 5.15: Mask loss as a function of the training epoch.

5.9 False Positive and False Negative examples

Here we present examples of false positive and false negative classifications of ghosts and scattered-light artifacts from the Mask R-CNN method outlined in Sec. 5.4.4. Fig. 5.17 presents examples of false positives (panel (a)) and the corresponding mask predictions of the Mask R-CNN model (panel (b)) for the same images. The color scheme of the predicted masks follows that of the main text (see Fig. 5.6).

As discussed in the main text, Sec. 5.4.4, most of those images are qualitatively different from other ghost-free images and contain either other types of artifacts — for example, Earth-orbiting satellites ((2,2), (3,1)), airplane trails (1,4), structured cloud cover ((1,5), (3,2), (3,3)) or large galaxies ((2,2), (2,5)) and resolved stellar systems (4,1), where the tuplets signify rows and columns, respectively.

Fig. 5.18 presents some examples of false negatives. These images contain ghosts (as confirmed by visual inspection), but they are actually very small or faint and hard to distinguish at the resolution presented here. Thus, it is not a surprise that these have been classified as “clean” by the mask R-CNN model, because they are different from the more prominent ghost-containing images that the network was trained on.



Threshold 0% (one pixel)

Threshold 50%

(a)

(b)

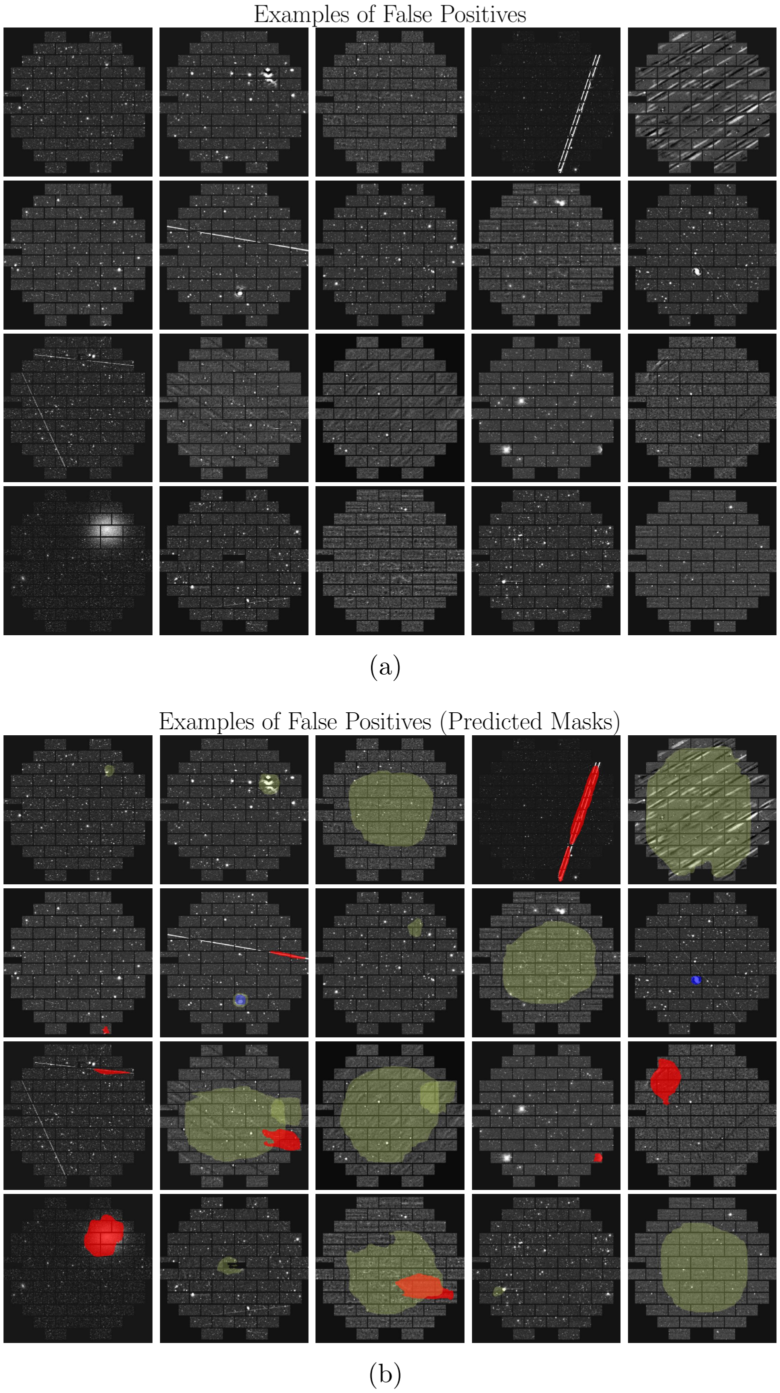
Threshold 0% (one pixel)

Threshold 50%

(c)

(d)

Figure 5.16: CCDs masked as ghost-containing (in blue) when even a single pixel of the predicted ghost mask lies within the CCD (0% threshold, panels (a) and (c)), and when at least half of the CCD area has to be covered by the CCD (50% threshold, panels (b) and (d)). The yellow contours correspond to the mask predictions of the Mask R-CNN model (without distinguishing between the different types of artifacts).


Examples of False Positives
Examples of False Positives (Predicted Masks)

(a)

(b)

Figure 5.17: (a) Example images classified as ghost-containing (false positives) and the corresponding predicted masks (lower panel, (b)).

Examples of False Negatives

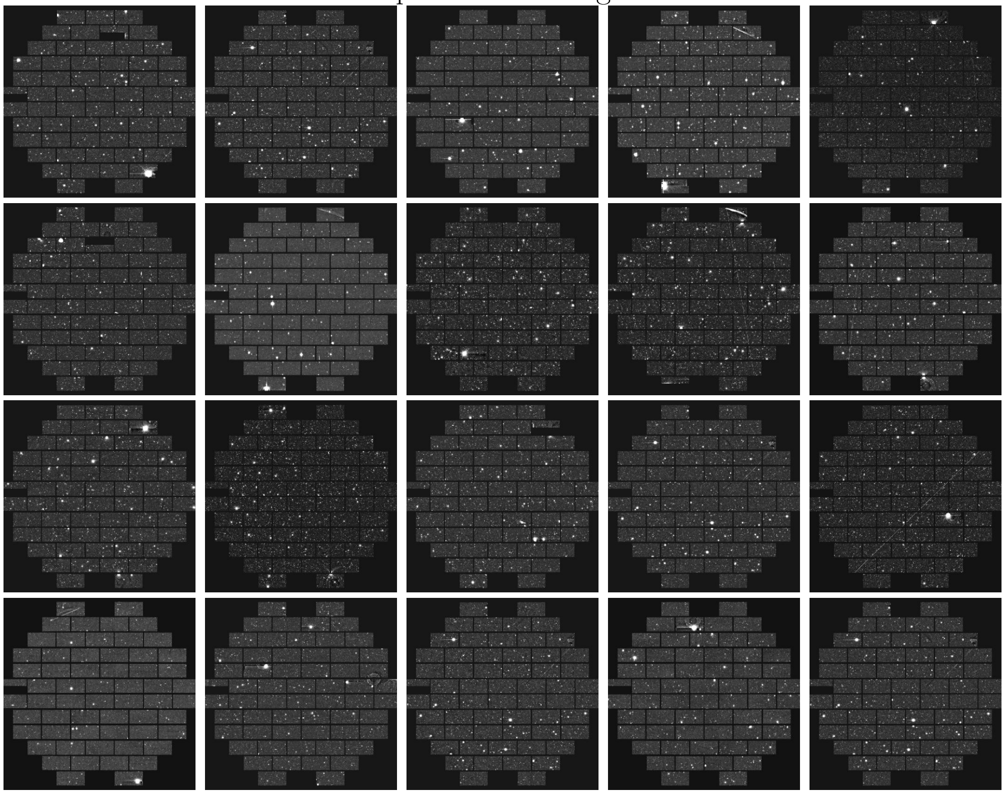


Figure 5.18: Examples of false negatives, i.e. images that were classified as ‘clean’ by the Mask R-CNN model (no objects detected). In practice, the artifacts present in these images are very small and faint, and often go undetected by human annotators.

CHAPTER 6

INFERRING STRUCTURAL PARAMETERS OF LOW-SURFACE-BRIGHTNESS GALAXIES WITH UNCERTAINTY QUANTIFICATION USING BAYESIAN NEURAL NETWORKS

A shorter version of this chapter was presented at the 2022 ICML workshop on Machine Learning and Astrophysics, arXiv: 2207.03471

6.1 Introduction

Despite their morphological diversity and complexity, the light distribution of galaxies can be well-described by analytic fitting functions with a limited number of free parameters, such as their orientation, size (radius), light concentration, total brightness, etc. Measurements of these parameters allow for a quantitative comparison of different galaxy populations and the derivation of empirical scaling relations [e.g., Courteau et al., 2007], which in turns facilitates the testing of galaxy formation models.

Traditionally, these parameters are measured using galaxy profile fitting software (two widely used options being `GALFIT`; Peng et al. [2002b] and `Imfit`; Erwin [2015]) that performs a χ^2 minimization between the chosen analytic model and a given galaxy image, to derive the best-fit parameters. Despite their success, these codes have their limitations, too: they are not optimized to fit a large number of galaxies quickly, and they usually require some manual intervention (for example, the selection of good initial model parameters). The low speed is an even more significant problem if one wants to obtain accurate estimates of the uncertainties associated with those measurements, for example via bootstrap resampling, or by using a Markov-Chain Monte Carlo (MCMC) approach to sample the parameter posterior distribution.

Large galaxy surveys, such as the Dark Energy Survey (DES)[1] and the upcoming Legacy Survey of Space and Time (LSST)[2] on the Vera C. Rubin Observatory, observe hundreds of millions (the former) to tens of billions (the latter) of galaxies. With the advent of these surveys, fast, automated, and reliable methods for measuring the structural parameters of galaxies are needed. Deep learning methods are well-suited to tackle this problem since, once trained, they are able to make predictions on new, unseen, examples very quickly.

Indeed, several works [e.g., Tuccillo et al., 2018, Aragon-Calvo and Carvajal, 2020, Li et al., 2022] have demonstrated that Convolutional Neural Networks (CNNs), trained on simulated galaxy images are significantly faster, and almost as accurate as the standard profile fitting methods in predicting galaxy parameters. However, those works used standard, deterministic, neural networks, that output single-point estimates and thus are unable to quantify the uncertainty associated with their predictions.

A rigorous uncertainty quantification is imperative for studies of challenging, low signal-to-noise objects, such as low-surface-brightness galaxies (LSBGs). LSBGs, defined as galaxies with a central brightness at least a magnitude fainter than that of the ambient dark sky, are challenging to observe and galaxy surveys have only recently started to produce large LSBG catalogs [Greco et al., 2018, Tanoglidis et al., 2021b, Zaritsky et al., 2022], althought they are expected to dominate the galaxy population. LSBGs are a target of future surveys, in the quest of understanding the galaxy formation process in a relatively unexplored regime.

In this chapter, we explore the use of Bayesien Neural Networks [BNNs; e.g., Valentin Jospin et al., 2020] for the problem of LSBG structural parameter estimation with simultaneous uncertainty quantification. BNNs replace signe node weights with distributions over the weights, output posterior probability distributions instead of point estimates for their predictions, and thus they naturally offer a way to quantify the uncertainties associated with

[1] `https://www.darkenergysurvey.org/`

[2] `https://https://www.lsst.org/`

neural network predictions. Specifically, we use a simulated dataset of DES-like LSBGs, to train, validate, and test a convolutional BNN model and compare the speed and accuracy of its predictions with those obtained using `pyImfit`[3], for a single-component Sérsic light-profile model.

This chapter is organized as follows: In Sec. 6.2 we describe the dataset of simulated LSBG images we use to train and test our BNN. In Sec. 6.3 we present the theory behind Bayesian Neural Networks, and how uncertainty in predictions can be quantified. In Sec. 6.4 we describe the architecture and the training of our BNN model. We present the results of our analysis in Sec. 6.5. We discuss our work and possible future directions in Sec. 6.6.

The code related to this work is publicly available on the GitHub page of this project: `https://github.com/dtanoglidis/BayesianNN`.

6.2 Simulated Data

We use `PyImfit` to create a simulated dataset of 170,000 LSBG images. Each image has dimensions 64 × 64 pixels, and it has two components: a uniform background $I(x, y) = I_{\text{sky}}$, and a Sérsic function [Sérsic, 1963] that describes the surface-brightness profile of the galaxies:

$$I_{\text{gal}}(r) = I_e \exp\left\{-b_n\left[\left(\frac{r}{r_e}\right)^{1/n} - 1\right]\right\}, \tag{6.1}$$

where $r = (x^2 + y^2/q^2)^{1/2}$ is the elliptical radius, (x, y) are the coordinates with origin at the center of the image, and q is the axis ratio. The axis ratio is connected to the ellipticity as $q = 1 - \epsilon$. The other free parameters of the model are: the effective half-light radius, r_e, the surface brightness at the effective elliptical half-light radius, I_e, the Sérsic index, n, that controls the shape of the light distribution, and the position angle that defines the

[3] `pyImfit` is a Python wrapper around `Imfit`.

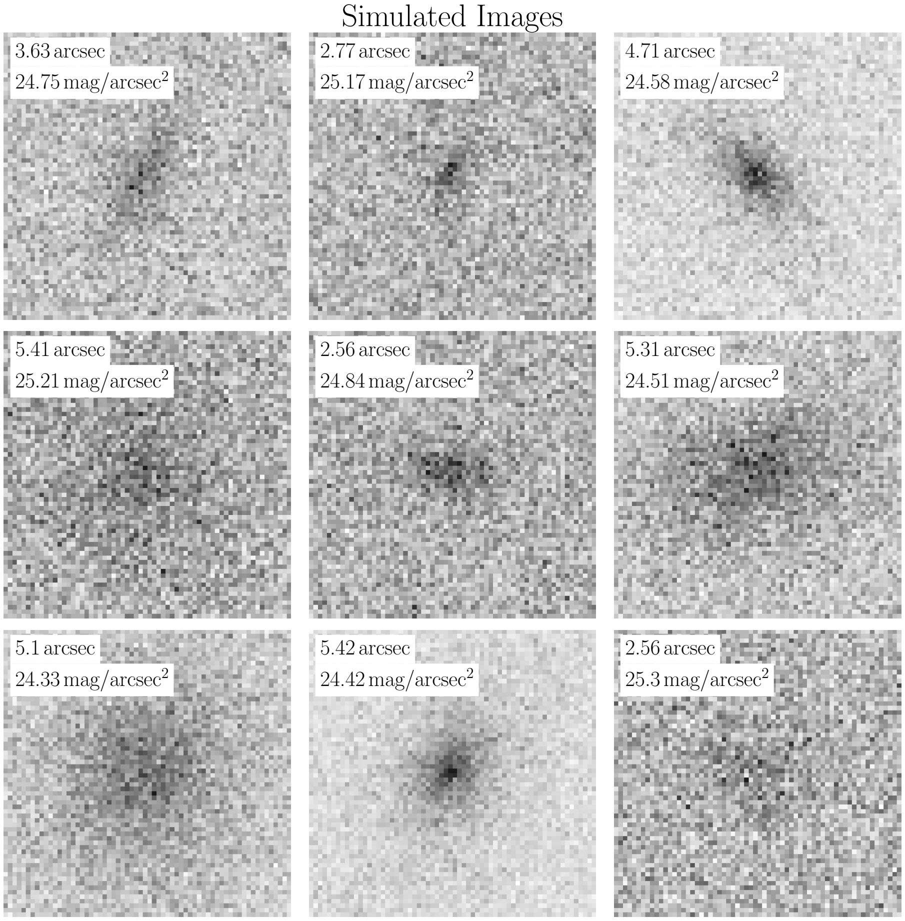


Figure 6.1: Examples of simulated LSBG images. The inset text refers to the effective radius, r_e, (top) and the surface brightness, I_e (bottom).

orientation of the galaxy profile. As for the value of b_n (not a free parameter of the model), is formally given by the solution of the transcendental equation $\Gamma(2n) = 2\gamma(2n, b_n)$, where $\Gamma(\alpha)$ the gamma function and $\gamma(\alpha, x)$ the incomplete gamma function; `pyImfit` uses the polynomial approximation by Ciotti and Bertin [1999].

We want our simulated images to resemble real LSBG images. For that reason we sample parameters uniformly, from a range that roughly corresponds to the bulk of the LSBGs discovered by DES, as described in Tanoglidis et al. [2021b] (Chapter 3):

- Position angle, PA $\in [0, 180]$ degrees,

- Ellipticity, $\epsilon \in [0.05, 0.7]$,
- Sérsic index, $n \in [0.5, 1.5]$,
- Surface brightness, $I_e \in [24.3, 25.5]$ mag/arcsec^2,
- Effective radius, $r_e \in [2.5, 6.0]$ arcsec.

Furthermore, we assume a pixel to angular scale conversion 1 pix = 0.263 arcsec (as it is for DECam), and background sky surface brightness $I_{\text{sky}} = 22.23$ mag/arcsec^2 [Neilsen et al., 2016]. At each pixel we randomly assign a number of photons (counts) drawn from a Poisson distribution with mean predicted from the total surface brightness model $I_{\text{tot}} = I_{\text{gal}} + I_{\text{sky}}$. In Fig. 6.1 we present a small subset of the simulated galaxy images.

6.3 Bayesian Neural Networks

Standard neural networks —with deterministic weights— once trained, output a single point estimate prediction for each example being presented to the network. Furthermore, the prediction is the same every time the same example is presented to the network. Thus, deterministic neural networks do not provide uncertainties in their predictions.

In Bayesian Neural Networks, the single weights are being replaced by appropriate probability distributions, that can be subsequently used to provide a measure of how (un)certain a model is in its predictions.

There are two types of uncertainty in the predictions of a BNN: *aleatoric* and *epistemic*. Aleatoric uncertainty is related to the intrinsic randomness of the data-generating or measurement process and it is not reducible (for example by collecting more data). Epistemic uncertainty, on the other hand, is the uncertainty related to the model and can be reduced by using a more appropriate model (e.g., a different neural network architecture) or by collecting more training data. As we are going to describe in more detail below, we capture aleatoric uncertainty by allowing the output of the neural network model to be a probability distribution, while the epistemic uncertainty is modelled by placing a probability distribution around the weights of the network.

6.3.1 Variational Inference

To understand the approach taken by BNNs, let us set up the problem by introducing a training set of observations, $\mathcal{D} = \{(x_1, y_1), \ldots, (x_D, y_D)\}$, and a new data point, x^*, whose label, y^*, we want to predict.

As we have mentioned, we can capture the epistemic uncertainty by replacing the weights of a neural network, w, with a distribution over the weights. This (posterior) distribution of the weights, given the data, $p(w|\mathcal{D})$, can be obtained using Bayes' theorem:

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}, \tag{6.2}$$

where $p(w)$ is the prior over the weights, $p(\mathcal{D}|w)$ the likelihood of the observed data given the model with weights w, and $p(\mathcal{D}) = \int p(\mathcal{D}, w)dw = \int p(\mathcal{D}|w)p(w)dw$ the evidence.

However, calculating the posterior (6.2) is usually a computationally intractable problem, because of presence of the evidence factor (which has to be integrated over all values of weights in a multi-dimensional space). For that reason, approximate methods should be employed.

In the variational inference (VI) approach, we approximate the true posterior, $p(w|\mathcal{D})$, with a distribution $q(w|\theta)$ of a known form (most commonly a multivariate Gaussian), with free parameters θ to be learned. The idea behind VI is to find those parameters θ (in the case of a Gaussian form, the mean μ and the covariance matrix) such as the variational posterior q matches the true posterior, p, as closely as possible.

A measure of the similarity between the two distributions is the Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(q(w|\theta)||p(w|\mathcal{D})) \equiv \int q(w|\theta) \log \frac{q(w|\theta)}{p(w|\mathcal{D})} dw. \tag{6.3}$$

From its definition, KL$\geq$ 0 with its value being smaller for more similar distributions. So, now the problem of finding the most appropriate posterior parameters can be rephrased as a minimization problem:

$$q(w|\hat{\theta}) = \mathrm{argmin}_{\theta} \mathrm{KL}(q(w|\theta)||p(w|\mathcal{D})). \tag{6.4}$$

By inserting the posterior, as defined in Eq. (6.2), one can show that this is equal to minimizing the following loss function:

$$\mathcal{L}(\mathcal{D}, \theta) = \mathbb{E}_{q(w|\theta)} \left[\log q(w|\theta) - \log p(w) p(\mathcal{D}|w)\right] \tag{6.5}$$

This is commonly referred as the *variational free energy* or the negative *evidence lower bound* (ELBO). For more details, see Appendix 6.7.

6.3.2 BNN training

Training a neural network requires to find those parameters that minimize the loss function for a given training dataset. This is usually achieved by employing an iterative parameter update method, called *gradient descent*, that requires taking derivatives of the loss with respect to those parameters.

The loss function, Eq. (6.5), can be evaluated by sampling weights from the variational distribution, $w \sim q(w|\theta)$. However, now one cannot take derivatives of the loss, exactly because these weights are now stochastic Monte Carlo samples. In order for the backpropagation to work it has been proposed to use a *reparametrization* trick (RT). Instead of sampling the weights directly from the variational distribution, we sample some noise, ϵ, from a simple distribution, $p(\epsilon)$. The weights are then related to that noise through a deterministic function, $w = g(\epsilon, \theta)$, that is differentiable with respect to the parameters, θ, of the initial distribution.

In the case of Gaussian variables, the noise can be drawn from the standard Normal distribution, $\epsilon \sim p(\epsilon) = \mathcal{N}(0, 1)$. Then, instead of sampling the weights from the multidimensional Gaussian, $w \sim q(w|\theta) = \mathcal{N}(\mu, \sigma)$, we can write:

$$w = \mu + \sigma \odot \epsilon. \tag{6.6}$$

Now, to perform backpropagation, the equations can be modified to take the derivative with respect to, and update, the parameters μ, σ. We present the mathematical details of this approach (called *Bayes by Backprop*) in Appendix 6.7.

Note that by having to learn the parameters μ, σ the total number of trainable parameters is doubled, which adds to the computational costs.

Another problem because it is computationally prohibitive to sample a unique noise variable ϵ for each example in a mini-batch, in the RT implementation, the sample weights are the same for all the examples within a batch. This causes the gradients between different samples in the same batch to be correlated, thus preventing the reduction of variance during training.

Flipout has been proposed as a solution to this problem ; it solves the correlation problem by randomly multiplying each perturbation/error by ±1 (pseudo random sign matrix), so in this way we reparametrize the weights as $w = \mu \pm \sigma \odot \epsilon$. The ± sign is randomly chosen. This can be easily vectorized and give us pseudo-independent weight perturbations in a computationally efficient way.

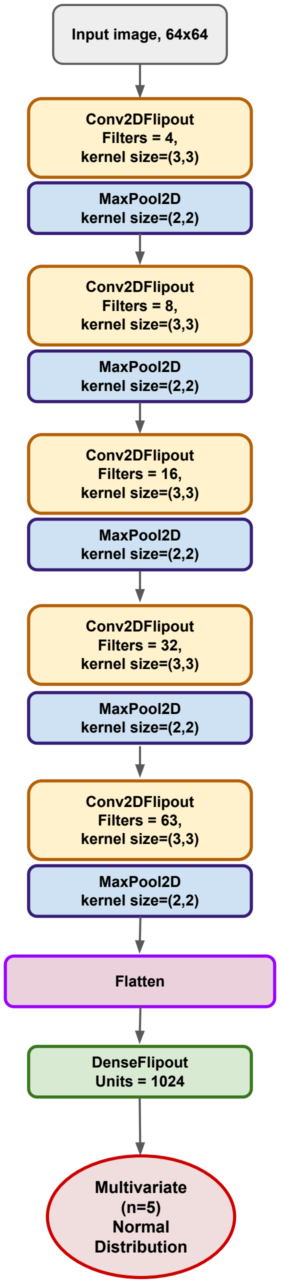


Figure 6.2: Schematic representation of the BNN architecture used in this work.

6.4 Implementation and training

In Fig. 6.2 we present a schematic overview of the BNN architecture used in this work for the problem of LSBG structural parameter regression. Our architecture consists of five (probabilistic) 2D convolutional layers. The number of filters and the kernel size used in each layer can be seen in the figure. Each convolutional layer is followed by a Max Pooling layer, with a kernel size (2,2). After flattening we have a dense layer. The output of the model is a multidimensional normal distribution (the five parameters our model tries to learn), with full covariance that allows to capture the correlations between the parameters.

We implemented the model architecture using the `Keras`[4] library on a `TensorFlow`[5] backend, and the `Tensorflow Probability`[6] extension of it, for the probabilistic layers. For the probabilistic convolutional and dense layers we use the `Convolution2DFlipout` and `DelseFlipout`, respectively, from `Tensorflow Probability` that use the Flipout estimator, as described in Sec. 6.3.2.

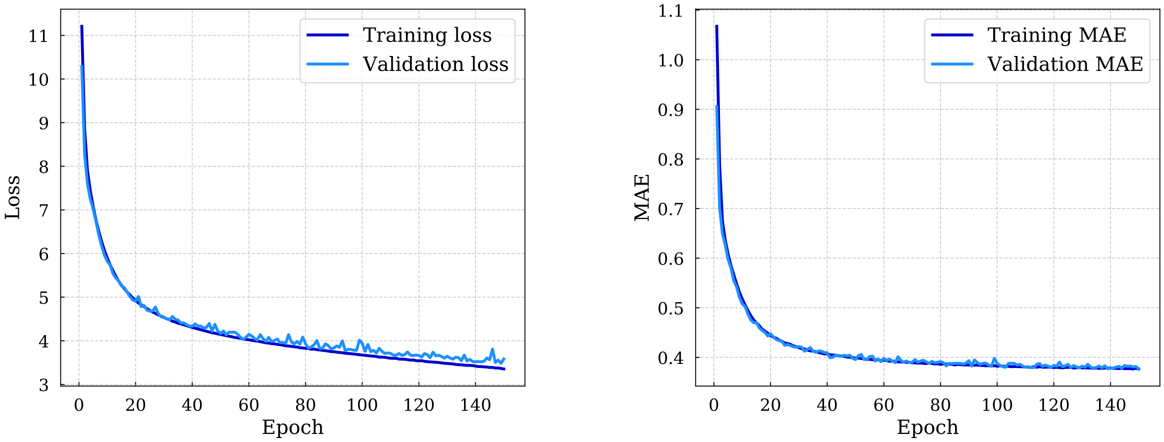
Before training, we randomly split the full simulated dataset to a training (150k), a validation (10k), and a test (10k) set. We perform training with a learning rate $\eta = 0.2$ (`Adadelta` optimizer), for 150 epochs, using a batch size of 64. During training we observed no signs of overfitting. We utilized the 25 GB high-RAM Nvidia P100 GPUs available through the Google Colab Pro. The training took approximately three hours to complete.

In Fig. 6.3 we plot the training history (loss and mean absolute error, MAE, as a function of the training epoch), for the training and validation sets. We do not see signs of overfitting, since the training and validation set curves follow each other. Furthermore, we see that the the MAE plateaus, meaning that we would not have significant gains from a longer training.

[4] `https://keras.io/`

[5] `https://www.tensorflow.org/`

[6] `https://www.tensorflow.org/probability`



(a)

(b)

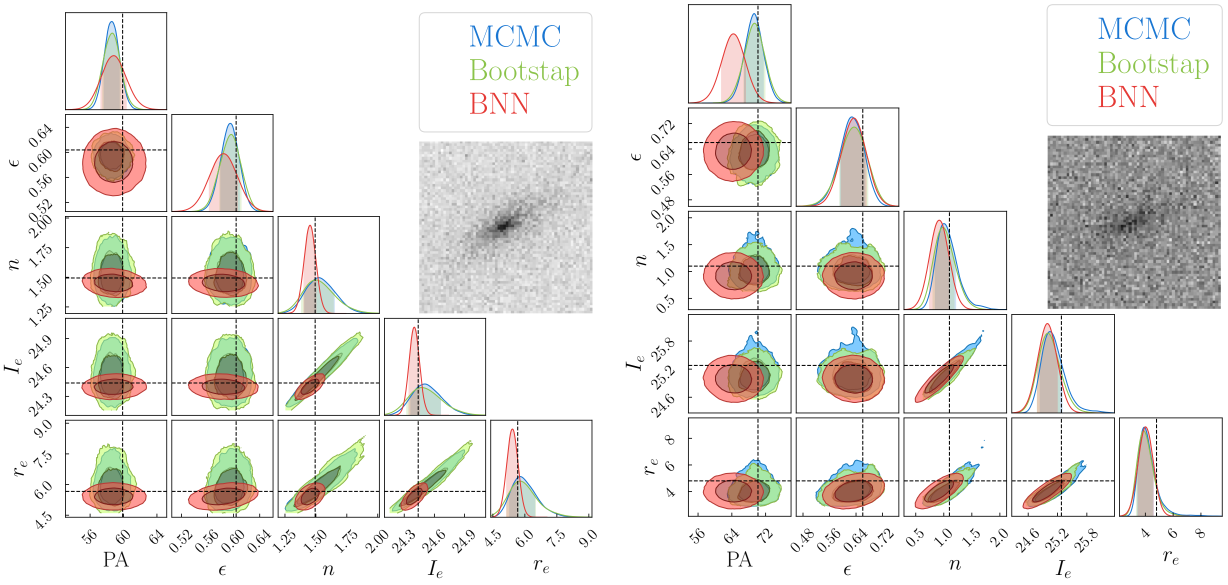
Figure 6.3: Training and validaton (a) loss, and (b) mean absolute error of the BNN model, as a function of the training epoch.

6.5 Results

6.5.1 Parameter posteriors

In Fig. 6.4 we present the predicted posterior distributions for the five parameters of the Sérsic model, described in Sec. 6.2, for a simulated galaxy at the bright end of the surface brightness distribution ($I_e = 24.4$ mag/arcsec^2, panel (a)), and one at the faint end ($I_e = 25.3$ mag/arcsec^2, panel (b)). We present the predictions of the BNN model (red contours) and those from the `pyImfit` model, using two different estimation methods: bootstrap resampling (green contours), and Markov chain Monte Carlo (MCMC; blue contours).

To get the BNN posteriors, we stack together the output distributions from 400 forward passes (predictions) of the model, for each one of the LSBG images. We see that the constraints on the parameters are tighter (as in the case of the brighter LSBG) or comparable (as in the case of the fainter LSBG) to those obtained using `pyImfit`. Although we present only two examples here, we have confirmed that this is true for a larger number of randomly selected LSBG images (see also the section that follows).



(a)

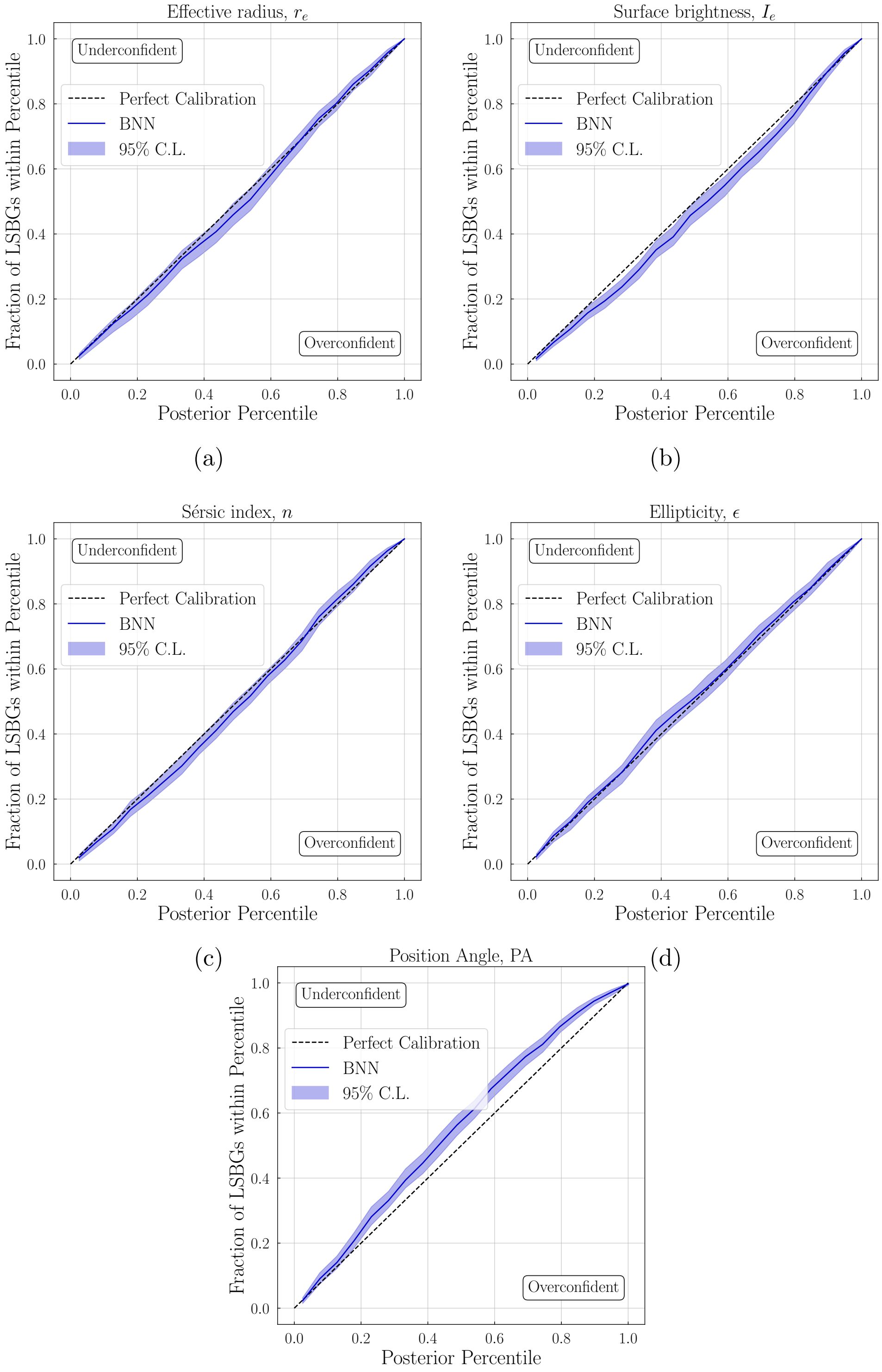
(b)

Figure 6.4: Predicted posterior distributions of structural parameters for a simulated LSBG at the bright end (I_e = 24.4 mag/arcsec^2, panel (a)), and one at the faint end (I_e = 25.3 mag/arcsec^2, panel (b)) of the surface brightness range we consider in this work. The dashed lines indicate the true input parameter values for the simulated.

In terms of speed, obtaining the posterior distribution for a single LSBG example using BNNs was significantly faster than running MCMC ($\sim$ 1 minute vs $\sim$ 6 minutes), and comparable in time to the bootstrap method. The real gain in time comes when one wants to process a large number of galaxy images simultaneously; for example, obtaining full posterior estimates for 1000 LSBGs using BNNs takes $\sim$ 7 minutes on our machine, while processing the same number of images using `pyImfit` and bootstrap resampling would had taken $\sim$ 16 hours.

6.5.2 Calibration

We have seen that BNNs fit Sérsic model parameters with tighter or similar uncertainties to those produced by `pyImfit`. However, in order to be interpreted as confidence intervals, we have to demonstrate that a x% interval contains the true value x% of the time – in other



(a)

(b)

(c)

(d)

(e)

Figure 6.5: Calibration curves of the marginalized BNN posteriors, for the five Sérsic model parameters considered in this work.

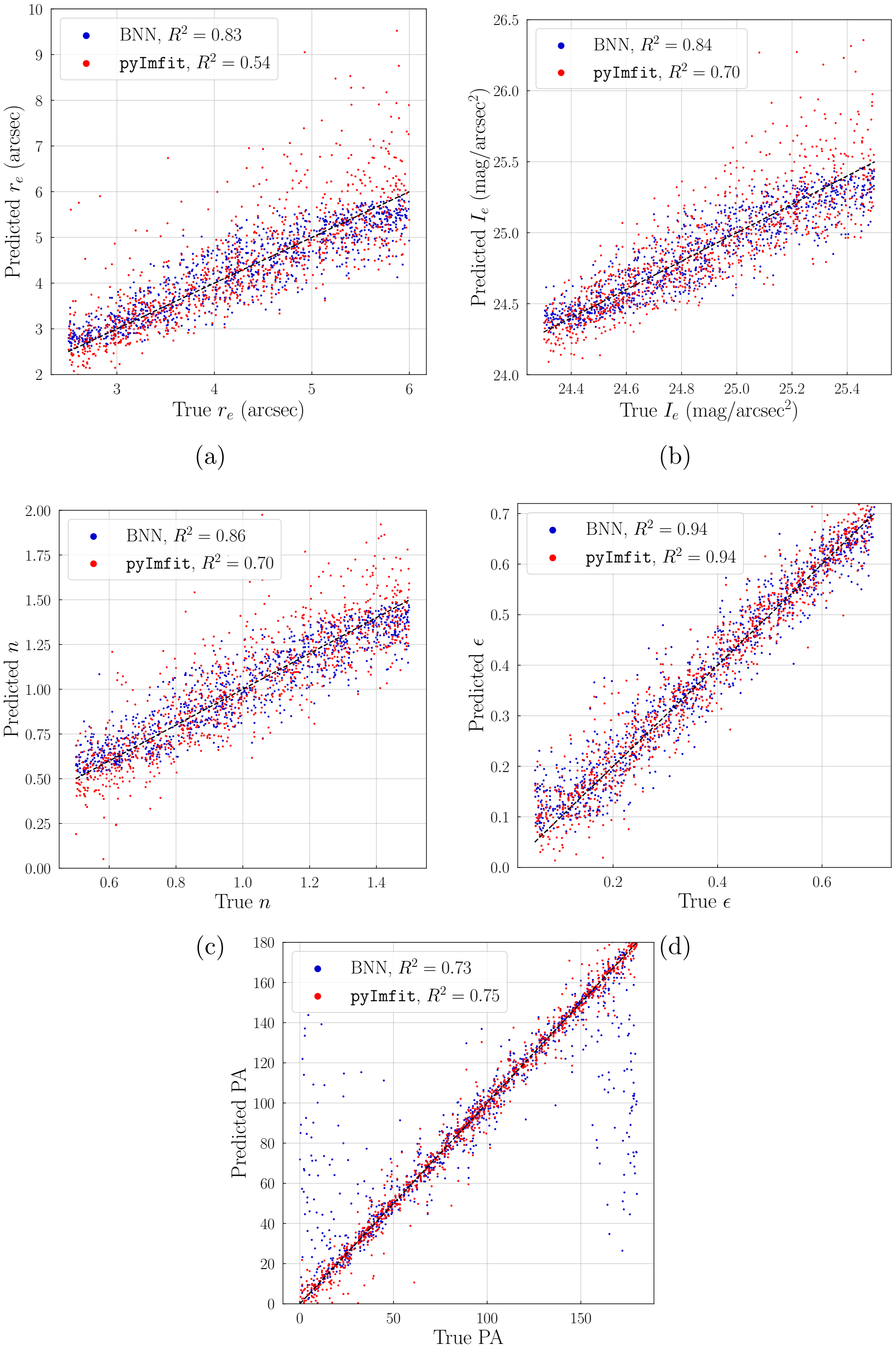
words, that the posterior is well-calibrated.

To investigate that, we consider the parameter posterior predictions on 1000 simulated LSBGs drawn from the test set. Following Wagner-Carena et al. [2021], Park et al. [2021], we calculate, at different posterior percentile levels, the fraction of LSBGs with true values within the limits of that percentile interval. For well-calibrated posteriors, those two quantities (percentile and fraction) should be equal. On the other hand, when the predictions of the BNN are overconfident, a smaller fraction of LSBGs are within the limits of a given percentile, while when the predictions are underconfident a larger fraction of LSBGs are within the same limits.

Here we consider the (marginalized) posterior of each parameter, separately (note that the above references consider the volume of the full posterior in a multidimensional space), in order to investigate if some of the parameter posteriors are better calibrated than others. We present the calibration curves for the five parameter posteriors in Fig. 6.5. The shaded areas correspond to 95% confidence intervals. They were obtained by running bootstrap resampling –with replacement– on the 1000 simulated LSBGs mentioned above, recalculating and stacking together the resulting calibration curves. As we can see, with the exception that for the position angle (PA), the posteriors predicted by the BNN are within the statistical uncertainly limits of being perfectly calibrated. The predictions (confidence intervals) for the position angle seem to be slightly underconfident; this is also consistent with the fact that the BNN confidence intervals presented in the previous section were slightly larger than those coming from `pyImfit` for this parameter (using either the Bootstrap or the MCMC methods). This behavior of the PA comes the discontinuity at PA=0° to PA=180°

6.5.3 Comparison of point estimates

We have demonstrated that the BNN model outputs well-calibrated uncertainties, and we have seen examples where we compared the output parameter posteriors from the BNN and



(a)

(b)

(c)

(d)

(e)

Figure 6.6: Comparison of the parameter predictions from the BNN model (blue dots) and from `pyImfit` (red dots). We also show the coefficients of determination.

the pyImfit algorithm. We now compare the point estimate (mean) prediction from the BNN with the best fit parameter output from pyImfit.

In Fig. 6.6 we plot the true value of the five Sérsic model parameters (for the same 1000 simulated LSBGs as in the previous section) vs the predicted ones, using both methods. The point estimates produced by the BNN method tend to be closer to the true values, as indicated by the higher coefficients of determination, R^2 (for all parameters except the position angle). We also see that BNN performs significantly better for LSBGs with larger effective radius, surface brightness, and Sérsic index values.

6.6 Discussion and Future Work

We have used a Bayesian Neural network model to predict structural parameters of LSBGs in simulated galaxy images. We compare the posterior parameter predictions from the BNN method with those from a profile-fitting algorithm (pyImfit) for simulated LSBGs and we show that the BNN gives comparable or even tighter parameter constraints.

We furthermore show that the uncertainties estimated using the BNN method are well calibrated, and that, for a sample of simulated LSBGs, the BNN gives better point-estimate parameter predictions (higher coefficient of determinations) compared to those from pyImfit.

A significant strength of our BNN method is its speed. For example, it can predict the full posterior distribution of the five-parameter Sérsic model for 1000 LSBGs images within ~ 7 minutes (on the machine used here); using pyImfit and the bootstrap resampling methods to get parameter constraints for the same number of images, would require ~ 16 hours.

An important next step, that we plan to address in future work, is to test the performance of our method on real LSBG images, and investigate ways to improve it if necessary, for example by re-training on real data, as in Tuccillo et al. [2018], or by adopting domain adaptation techniques [e.g. Ćiprijanović et al., 2020a]. Other areas of future investigation include testing different BNN architectures, testing the performance of the model on data outside of the training range, and a more rigorous comparison of the performance (parameter constraints and speed) between BNNs and `pyImfit`.

6.7 Bayes by Backprop details

In this appendix we present the mathematical details of the Bayes by Backprop [Blundell et al., 2015] method used to train Bayesian Neural Networks using the backpropagation methods, that is presented in Sec. 6.3.

First, let us prove that minimizing the KL divergence between the variational and true posterior distributions is equivalent to minimizing the loss function given by Eq. (6.5).

$$\mathrm{KL}(q(w|\theta)||p(w|\mathcal{D})) = \int q(w|\theta) \log \frac{q(w|\theta)p(\mathcal{D})}{p(\mathcal{D}|w)p(w)} dw = \tag{6.7}$$

$$= \int q(w|\theta) \log \frac{q(w|\theta)}{p(w)} dw - \int q(w|\theta) \log p(\mathcal{D}|w) dw + \log p(\mathcal{D}) \int q(w|\theta) dw = \tag{6.8}$$

$$= \mathrm{KL}(q(w|\theta)||p(w)) - \mathbb{E}_{q(w|\theta)} \log p(\mathcal{D}|w) + \log p(\mathcal{D}). \tag{6.9}$$

Since the evidence does not depend on the parameters θ, minimizing the KL divergence is equal to minimizing the loss as given by Eq. (6.5).

Let a random variable ϵ with distribution $q(\epsilon)$. If we can write $w = t(\theta, \epsilon)$, where $t(\theta, \epsilon)$ a deterministic function. Then for a function $f(w, \theta)$:

$$\frac{\partial}{\partial\theta} \mathbb{E}_{q(w|\theta)}[f(w,\theta)] = \frac{\partial}{\partial\theta} \int f(w,\theta) q(w|\theta) dw = \tag{6.10}$$

$$= \frac{\partial}{\partial\theta} \int f(w,\theta) q(\epsilon) d\epsilon = \tag{6.11}$$

$$= \int \frac{\partial}{\partial\theta} = f(w,\theta) q(\epsilon) d\epsilon = \mathbb{E}_{q(\epsilon)} \frac{\partial}{\partial\theta} [f(w,\theta)] = \tag{6.12}$$

$$= \mathbb{E}_{q(\epsilon)} \left[\frac{\partial f(w,\theta)}{\partial w} \frac{\partial w}{\partial\theta} + \frac{\partial f(w,\theta)}{\partial\theta} \right] \tag{6.13}$$

We see now that the loss function can be written as:

$$\mathcal{L}(\mathcal{D}, \theta) = \mathbb{E}_{q(w|\theta)} \left[f(w, \theta) \right] \tag{6.14}$$

with: $f(w, \theta) = \log q(w|\theta) - \log p(w)p(\mathcal{D}|w)$.

Let the posterior parameters be $\theta = (\mu, \sigma)$. We have also said that $w = \mu \pm \sigma \odot \epsilon$.

From the proposition we have proven above, we have:

$$\frac{\partial}{\partial \mu} \mathcal{L}(\mathcal{D}, \theta) \equiv \nabla_{\mu} \mathcal{L} = \mathbb{E}_{q(\epsilon)} \left[\frac{\partial f(w, \theta)}{\partial w} \frac{\partial w}{\partial \mu} + \frac{\partial f(w, \theta)}{\partial \mu} \right] \tag{6.15}$$

and similarly for the derivative with respect to σ. Thus, we can write the backpropagation update equation:

$$\mu = \mu - \alpha \nabla_{\sigma} \mathcal{L} \tag{6.16}$$

$$\sigma = \sigma - \alpha \nabla_{\sigma} \mathcal{L} \tag{6.17}$$

CHAPTER 7

CONCLUSIONS AND SUMMARY

The advent of deep and wide galaxy surveys, recent discoveries, and the development of new analysis techniques make our time ideal for the exploration of the low-surface-brightness universe. Low-surface-brightness galaxies (LSBGs) in particular, that constituted the main objects of interest in this thesis, are defined as galaxies with surface brightness at least a magnitude less than that of the background dark sky and below the surface brightness limits of past surveys.

This work provided a large catalog (23,790) of LSBGs discovered in the first three years of observation of the Dark Energy Survey. The catalog has been made publicly available at the DES Data Managment website, `https://des.ncsa.illinois.edu/releases/other/y3-lsbg`, and has been well-received by the community. For example, Kado-Fong et al. [2021] used this catalog to infer the intrinsic shapes of LSBGs and constrain their formation mechanism. [Müller and Schnider, 2021] used the same catalog to train a LSBG morphology classifier. Furthermore, the lessons learned while assembling this catalog are currently being used for the discovery of LSBGs from the deeper data coming from the full five years of DES observations.

The large datasets produced by current and upcoming surveys (see, e.g., `https://www.lsst.org/scientists/keynumbers`) require the development of new analysis techniques, especially for the discovery of the challenging Low-Surface-Brightness systems, such as the galaxies described in this work. For example, while producing the DES Y3 LSBG catalog, we realized that a large number of artifacts (almost ~ 0.5M) were passing the LSBG selection criteria, without being genuine galaxies. Machine Learning (ML) algorithms are well-suited to tackle problems like this, since one can train a ML model to distinguish between the two categories.

We followed this approach in our original work, and we expanded it (Chapter 4) using modern neural network techniques, achieving classification accuracy that is comparable with that from expert human annotators.

Another problem we tackled in this thesis included the detection and removal of artifacts generated by spurious light reflections off the telescope optical surfaces. For that problem we adapted a deep-learning based object detection model, developed and used for detecting everyday objects, to work with survey images and ghosting artifacts. These ML-based solutions described above required the manual generation of annotations and labels, in order to perform the training of the (supervised) machine learning models. This is a significant (time) bottleneck, common in many problems, since, while there is an abundance of data in astrophysics, the number of labeled data is not so abundant.

Finally, we demonstrated that a Bayesian Neural Network, trained and tested on simulated LSBG images, is able to infer structural galaxy parameters (parameters of a single-component Sérsic model) accurately, with similar uncertainties to those inferred using light profile-fitting software, and significantly faster.

As we have said in multiple occasions, low-surface-brightness science is a key goal of upcoming galaxy surveys, like the Legacy Survey of Space and Time on the Vera C. Rubin observatory (see for example `https://sites.google.com/view/lsstgsc/working-groups/low-surface-brightness-science`); the techniques developed and discussed in this thesis will hopefully contribute to the advancement of the efforts to further explore that elusive domain.

REFERENCES

Daniel B. Abazajian, Kevork N.and et al. The Seventh Data Release of the Sloan Digital Sky Survey. *ApJS*, 182(2):543–558, June 2009. doi: 10.1088/0067-0049/182/2/543.

T. M. C. Abbott, J. Annis, D. L. DePoy, B. Flaugher, S. M. Kent, H. Lin, and W. Merritt. Dark energy camera specifications and technical requirements, 2009. URL `https://www.noao.edu/meetings/decam/media/DECam_Technical_specifications.pdf`.

T. M. C. Abbott, F. B. Abdalla, S. Allam, A. Amara, et al. The Dark Energy Survey: Data Release 1. *ApJS*, 239(2):18, December 2018. doi: 10.3847/1538-4365/aae9f0.

T. M. C. Abbott, M. Adamów, M. Aguena, S. Allam, A. Amon, J. Annis, S. Avila, D. Bacon, M. Banerji, K. Bechtol, M. R. Becker, G. M. Bernstein, E. Bertin, S. Bhargava, S. L. Bridle, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, A. Choi, C. Conselice, M. Costanzi, M. Crocce, L. N. da Costa, T. M. Davis, J. De Vicente, J. DeRose, S. Desai, H. T. Diehl, J. P. Dietrich, A. Drlica-Wagner, K. Eckert, J. Elvin-Poole, S. Everett, A. E. Evrard, I. Ferrero, A. Ferté, B. Flaugher, P. Fosalba, D. Friedel, J. Frieman, J. García-Bellido, E. Gaztanaga, L. Gelman, D. W. Gerdes, T. Giannantonio, M. S. S. Gill, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, S. R. Hinton, D. L. Hollowood, K. Honscheid, D. Huterer, D. J. James, T. Jeltema, M. D. Johnson, S. Kent, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, T. S. Li, C. Lidman, H. Lin, N. MacCrann, M. A. G. Maia, T. A. Manning, J. D. Maloney, M. March, J. L. Marshall, P. Martini, P. Melchior, F. Menanteau, R. Miquel, R. Morgan, J. Myles, E. Neilsen, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón, D. Petravick, A. Pieres, A. A. Plazas, C. Pond, M. Rodriguez-Monroy, A. K. Romer, A. Roodman, E. S. Rykoff, M. Sako, E. Sanchez, B. Santiago, V. Scarpine, S. Serrano, I. Sevilla-Noarbe, J. Allyn Smith, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, C. To, P. E. Tremblay, M. A. Troxel, D. L. Tucker, D. J. Turner, T. N. Varga, A. R. Walker, R. H. Wechsler, J. Weller, W. Wester, R. D. Wilkinson, B. Yanny, Y. Zhang, R. Nikutta, M. Fitzpatrick, A. Jacques, A. Scott, K. Olsen, L. Huang, D. Herrera, S. Juneau, D. Nidever, B. A. Weaver, C. Adean, V. Correia, M. de Freitas, F. N. Freitas, C. Singulani, G. Vila-Verde, and Linea Science Server. The Dark Energy Survey Data Release 2. *ApJS*, 255(2):20, August 2021. doi: 10.3847/1538-4365/ac00b3.

T. M. C. Abbott, M. Aguena, A. Alarcon, S. Allam, O. Alves, A. Amon, F. Andrade-Oliveira, J. Annis, S. Avila, D. Bacon, E. Baxter, K. Bechtol, M. R. Becker, G. M. Bernstein, S. Bhargava, S. Birrer, J. Blazek, A. Brandao-Souza, S. L. Bridle, D. Brooks, E. Buckley-Geer, D. L. Burke, H. Camacho, A. Campos, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, A. Chen, R. Chen, A. Choi, C. Conselice, J. Cordero, M. Costanzi, M. Crocce, L. N. da Costa, M. E. da Silva Pereira, C. Davis, T. M. Davis, J. De Vicente, J. DeRose, S. Desai, E. Di Valentino, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, C. Doux, A. Drlica-Wagner, K. Eckert, T. F. Eifler, F. Elsner, J. Elvin-Poole, S. Everett, A. E. Evrard, X. Fang, A. Farahi, E. Fernandez, I. Ferrero, A. Ferté, P. Fosalba, O. Friedrich, J. Frieman, J. García-Bellido,

M. Gatti, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, G. Giannini, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, I. Harrison, W. G. Hartley, K. Herner, S. R. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, E. M. Huff, D. Huterer, B. Jain, D. J. James, M. Jarvis, N. Jeffrey, T. Jeltema, A. Kovacs, E. Krause, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, P. F. Leget, P. Lemos, A. R. Liddle, C. Lidman, M. Lima, H. Lin, N. MacCrann, M. A. G. Maia, J. L. Marshall, P. Martini, J. McCullough, P. Melchior, J. Mena-Fernández, F. Menanteau, R. Miquel, J. J. Mohr, R. Morgan, J. Muir, J. Myles, S. Nadathur, A. Navarro-Alsina, R. C. Nichol, R. L. C. Ogando, Y. Omori, A. Palmese, S. Pandey, Y. Park, F. Paz-Chinchón, D. Petravick, A. Pieres, A. A. Plazas Malagón, A. Porredon, J. Prat, M. Raveri, M. Rodriguez-Monroy, R. P. Rollins, A. K. Romer, A. Roodman, R. Rosenfeld, A. J. Ross, E. S. Rykoff, S. Samuroff, C. Sánchez, E. Sanchez, J. Sanchez, D. Sanchez Cid, V. Scarpine, M. Schubnell, D. Scolnic, L. F. Secco, S. Serrano, I. Sevilla-Noarbe, E. Sheldon, T. Shin, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, M. Tabbutt, G. Tarle, D. Thomas, C. To, A. Troja, M. A. Troxel, D. L. Tucker, I. Tutusaus, T. N. Varga, A. R. Walker, N. Weaverdyck, R. Wechsler, J. Weller, B. Yanny, B. Yin, Y. Zhang, J. Zuntz, and DES Collaboration. Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing. , 105(2):023520, January 2022. doi: 10.1103/PhysRevD.105.023520.

Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. `https://github.com/matterport/Mask_RCNN`, 2017.

G. O. Abell, H. G. Corwin, Jr., and R. P. Olowin. A catalog of rich clusters of galaxies. *ApJS*, 70:1–138, May 1989. doi: 10.1086/191333.

R. G. Abraham and P. G. van Dokkum. Ultra-Low Surface Brightness Imaging with the Dragonfly Telephoto Array. , 126:55, January 2014. doi: 10.1086/674875.

Sandro Ackermann, Kevin Schawinski, Ce Zhang, and et al. Using transfer learning to detect galaxy mergers. , 479(1):415–425, September 2018. doi: 10.1093/mnras/sty1398.

C. Adami, R. Scheidegger, M. Ulmer, F. Durret, A. Mazure, M. J. West, C. J. Conselice, M. Gregg, S. Kasun, R. Pelló, and J. P. Picat. A deep wide survey of faint low surface brightness galaxies in the direction of the Coma cluster of galaxies. , 459(3):679–692, Dec 2006. doi: 10.1051/0004-6361:20053758.

Hiroaki Aihara, Nobuo Arimoto, Robert Armstrong, Stéphane Arnouts, Neta A. Bahcall, Steven Bickerton, James Bosch, Kevin Bundy, Peter L. Capak, James H. H. Chan, Masashi Chiba, Jean Coupon, Eiichi Egami, Motohiro Enoki, Francois Finet, Hiroki Fujimori, Seiji Fujimoto, Hisanori Furusawa, Junko Furusawa, Tomotsugu Goto, Andy Goulding, Johnny P. Greco, Jenny E. Greene, James E. Gunn, Takashi Hamana, Yuichi Harikane, Yasuhiro Hashimoto, Takashi Hattori, Masao Hayashi, Yusuke Hayashi, Krzysztof G. Hełminiak, Ryo Higuchi, Chiaki Hikage, Paul T. P. Ho, Bau-Ching Hsieh, Kuiyun Huang, Song Huang, Hiroyuki Ikeda, Masatoshi Imanishi, Akio K. Inoue, Kazushi Iwasawa, Ikuru Iwata, Anton T. Jaelani, Hung-Yu Jian, Yukiko Kamata, Hiroshi Karoji, Nobunari

Kashikawa, Nobuhiko Katayama, Satoshi Kawanomoto, Issha Kayo, Jin Koda, Michitaro Koike, Takashi Kojima, Yutaka Komiyama, Akira Konno, Shintaro Koshida, Yusei Koyama, Haruka Kusakabe, Alexie Leauthaud, Chien-Hsiu Lee, Lihwai Lin, Yen-Ting Lin, Robert H. Lupton, Rachel Mandelbaum, Yoshiki Matsuoka, Elinor Medezinski, Sogo Mineo, Shoken Miyama, Hironao Miyatake, Satoshi Miyazaki, Rieko Momose, Anupreeta More, Surhud More, Yuki Moritani, Takashi J. Moriya, Tomoki Morokuma, Shiro Mukae, Ryoma Murata, Hitoshi Murayama, Tohru Nagao, Fumiaki Nakata, Mana Niida, Hiroko Niikura, Atsushi J. Nishizawa, Yoshiyuki Obuchi, Masamune Oguri, Yukie Oishi, Nobuhiro Okabe, Sakurako Okamoto, Yuki Okura, Yoshiaki Ono, Masato Onodera, Masafusa Onoue, Ken Osato, Masami Ouchi, Paul A. Price, Tae-Soo Pyo, Masao Sako, Marcin Sawicki, Takatoshi Shibuya, Kazuhiro Shimasaku, Atsushi Shimono, Masato Shirasaki, John D. Silverman, Melanie Simet, Joshua Speagle, David N. Spergel, Michael A. Strauss, Yuma Sugahara, Naoshi Sugiyama, Yasushi Suto, Sherry H. Suyu, Nao Suzuki, Philip J. Tait, Masahiro Takada, Tadafumi Takata, Naoyuki Tamura, Manobu M. Tanaka, Masaomi Tanaka, Masayuki Tanaka, Yoko Tanaka, Tsuyoshi Terai, Yuichi Terashima, Yoshiki Toba, Nozomu Tominaga, Jun Toshikawa, Edwin L. Turner, Tomohisa Uchida, Hisakazu Uchiyama, Keiichi Umetsu, Fumihiro Uraguchi, Yuji Urata, Tomonori Usuda, Yousuke Utsumi, Shiang-Yu Wang, Wei-Hao Wang, Kenneth C. Wong, Kiyoto Yabe, Yoshihiko Yamada, Hitomi Yamanoi, Naoki Yasuda, Sherry Yeh, Atsunori Yonehara, and Suraphong Yuma. The Hyper Suprime-Cam SSP Survey: Overview and survey design. , 70:S4, January 2018. doi: 10.1093/pasj/psx066.

Görkem Algan and Ilkay Ulusoy. Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey. *arXiv e-prints*, art. arXiv:1912.05170, December 2019.

N. C. Amorisco and A. Loeb. Ultradiffuse galaxies: the high-spin tail of the abundant dwarf galaxy population. , 459(1):L51–L55, Jun 2016. doi: 10.1093/mnrasl/slw055.

M. A. Aragon-Calvo and J. C. Carvajal. Self-supervised learning with physics-aware neural networks - I. Galaxy model fitting. , 498(3):3713–3719, November 2020. doi: 10.1093/mnras/staa2228.

Astropy Collaboration, Thomas P. Robitaille, Erik J. Tollerud, Perry Greenfield, Michael Droettboom, Erik Bray, Tom Aldcroft, Matt Davis, Adam Ginsburg, Adrian M. Price-Whelan, Wolfgang E. Kerzendorf, Alexander Conley, Neil Crighton, Kyle Barbary, Demitri Muna, Henry Ferguson, Frédéric Grollier, Madhura M. Parikh, Prasanth H. Nair, Hans M. Unther, Christoph Deil, Julien Woillez, Simon Conseil, Roban Kramer, James E. H. Turner, Leo Singer, Ryan Fox, Benjamin A. Weaver, Victor Zabalza, Zachary I. Edwards, K. Azalee Bostroem, D. J. Burke, Andrew R. Casey, Steven M. Crawford, Nadia Dencheva, Justin Ely, Tim Jenness, Kathleen Labrie, Pey Lian Lim, Francesco Pierfederici, Andrew Pontzen, Andy Ptak, Brian Refsdal, Mathieu Servillat, and Ole Streicher. Astropy: A community Python package for astronomy. , 558:A33, Oct 2013. doi: 10.1051/0004-6361/201322068.

Ivan K. Baldry, Karl Glazebrook, Jon Brinkmann, Željko Ivezić, Robert H. Lupton, Robert C. Nichol, and Alexander S. Szalay. Quantifying the Bimodal Color-Magnitude Distribution of Galaxies. , 600(2):681–694, Jan 2004. doi: 10.1086/380092.

Nicholas M. Ball and Robert J. Brunner. Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D*, 19(7):1049–1106, January 2010. doi: 10.1142/S0218271810017160.

Steven P. Bamford, Robert C. Nichol, Ivan K. Baldry, Kate Land, Chris J. Lintott, Kevin Schawinski, Anže Slosar, Alexander S. Szalay, Daniel Thomas, Mehri Torki, Dan Andreescu, Edward M. Edmondson, Christopher J. Miller, Phil Murray, M. Jordan Raddick, and Jan Vandenberg. Galaxy Zoo: the dependence of morphology and colour on environment*. , 393(4):1324–1352, Mar 2009. doi: 10.1111/j.1365-2966.2008.14252.x.

Marco Barden, Boris Häußler, Chien Y. Peng, Daniel H. McIntosh, and Yicheng Guo. GALAPAGOS: from pixels to parameters. , 422(1):449–468, May 2012. doi: 10.1111/j.1365-2966.2012.20619.x.

Dalya Baron. Machine Learning in Astronomy: a practical overview. *arXiv e-prints*, art. arXiv:1904.07248, April 2019.

K. Bechtol, A. Drlica-Wagner, E. Balbinot, A. Pieres, J. D. Simon, B. Yanny, B. Santiago, R. H. Wechsler, J. Frieman, A. R. Walker, P. Williams, E. Rozo, E. S. Rykoff, A. Queiroz, E. Luque, A. Benoit-Lévy, D. Tucker, I. Sevilla, R. A. Gruendl, L. N. da Costa, A. Fausti Neto, M. A. G. Maia, T. Abbott, S. Allam, R. Armstrong, A. H. Bauer, G. M. Bernstein, R. A. Bernstein, E. Bertin, D. Brooks, E. Buckley-Geer, D. L. Burke, A. Carnero Rosell, F. J. Castander, R. Covarrubias, C. B. D'Andrea, D. L. DePoy, S. Desai, H. T. Diehl, T. F. Eifler, J. Estrada, A. E. Evrard, E. Fernandez, D. A. Finley, B. Flaugher, E. Gaztanaga, D. Gerdes, L. Girardi, M. Gladders, D. Gruen, G. Gutierrez, J. Hao, K. Honscheid, B. Jain, D. James, S. Kent, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, T. S. Li, H. Lin, M. Makler, M. March, J. Marshall, P. Martini, K. W. Merritt, C. Miller, R. Miquel, J. Mohr, E. Neilsen, R. Nichol, B. Nord, R. Ogando, J. Peoples, D. Petravick, A. A. Plazas, A. K. Romer, A. Roodman, M. Sako, E. Sanchez, V. Scarpine, M. Schubnell, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, J. Thaler, D. Thomas, W. Wester, J. Zuntz, and DES Collaboration. Eight New Milky Way Companions Discovered in First-year Dark Energy Survey Data. , 807(1):50, July 2015. doi: 10.1088/0004-637X/807/1/50.

Peter S. Behroozi, Risa H. Wechsler, and Charlie Conroy. The Average Star Formation Histories of Galaxies in Dark Matter Halos from z = 0-8. , 770(1):57, Jun 2013. doi: 10.1088/0004-637X/770/1/57.

Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. volume 27 of *Proceedings of Machine Learning Research*, pages 17–36, Bellevue, Washington, USA, 02 Jul 2012. JMLR Workshop and Conference Proceedings. URL `http://proceedings.mlr.press/v27/bengio12a.html`.

Andreas A. Berlind, Michael R. Blanton, David W. Hogg, David H. Weinberg, Romeel Davé, Daniel J. Eisenstein, and Neal Katz. Interpreting the Relationship between Galaxy Luminosity, Color, and Environment. , 629(2):625–632, August 2005. doi: 10.1086/431658.

Pedro H. Bernardinelli, Gary M. Bernstein, Benjamin T. Montet, Robert Weryk, Richard Wainscoat, M. Aguena, S. Allam, F. Andrade-Oliveira, J. Annis, S. Avila, E. Bertin, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, R. Cawthon, C. Conselice, M. Costanzi, L. N. da Costa, M. E. S. Pereira, J. De Vicente, H. T. Diehl, S. Everett, I. Ferrero, B. Flaugher, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, S. R. Hinton, D. L. Hollowood, K. Honscheid, D. J. James, K. Kuehn, N. Kuropatkin, O. Lahav, M. A. G. Maia, J. L. Marshall, F. Menanteau, R. Miquel, R. Morgan, R. L. C. Ogando, F. Paz-Chinchón, A. Pieres, A. A. Plazas Malagón, M. Rodriguez-Monroy, A. K. Romer, A. Roodman, E. Sanchez, M. Schubnell, S. Serrano, I. Sevilla-Noarbe, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, G. Tarle, C. To, M. A. Troxel, T. N. Varga, A. R. Walker, Y. Zhang, and DES Collaboration. C/2014 UN$_{271}$ (Bernardinelli-Bernstein): The Nearly Spherical Cow of Comets. , 921(2):L37, November 2021. doi: 10.3847/2041-8213/ac32d3.

G. M. Bernstein, R. C. Nichol, J. A. Tyson, M. P. Ulmer, and D. Wittman. The Luminosity Function of the Coma Cluster Core for -25&lt;M/R&lt;-9.4. , 110:1507, Oct 1995. doi: 10.1086/117624.

G. M. Bernstein, T. M. C. Abbott, R. Armstrong, D. L. Burke, H. T. Diehl, R. A. Gruendl, M. D. Johnson, T. S. Li, E. S. Rykoff, A. R. Walker, W. Wester, and B. Yanny. Photometric Characterization of the Dark Energy Camera. , 130(987):054501, May 2018. doi: 10.1088/1538-3873/aaa753.

E. Bertin. Automatic Astrometric and Photometric Calibration with SCAMP. In C. Gabriel, C. Arviset, D. Ponz, and S. Enrique, editors, *Astronomical Data Analysis Software and Systems XV*, volume 351 of *Astronomical Society of the Pacific Conference Series*, page 112, Jul 2006.

E. Bertin. Displaying Digital Deep Sky Images. In P. Ballester, D. Egret, and N. P. F. Lorente, editors, *Astronomical Data Analysis Software and Systems XXI*, volume 461 of *Astronomical Society of the Pacific Conference Series*, page 263, September 2012.

E. Bertin and S. Arnouts. SExtractor: Software for source extraction. , 117:393–404, June 1996. doi: 10.1051/aas:1996164.

Maciej Bilicki, Thomas H. Jarrett, John A. Peacock, Michelle E. Cluver, and Louise Steward. Two Micron All Sky Survey Photometric Redshift Catalog: A Comprehensive Three-dimensional Census of the Whole Sky. *ApJS*, 210(1):9, Jan 2014. doi: 10.1088/0067-0049/210/1/9.

John P. Blakeslee, Andrés Jordán, Simona Mei, Patrick Côté, Laura Ferrarese, Leopoldo Infante, Eric W. Peng, John L. Tonry, and Michael J. West. The ACS Fornax Cluster Survey. V. Measurement and Recalibration of Surface Brightness Fluctuations and a

Precise Value of the Fornax-Virgo Relative Distance. , 694(1):556–572, Mar 2009. doi: 10.1088/0004-637X/694/1/556.

Michael R. Blanton and John Moustakas. Physical Properties and Environments of Nearby Galaxies. , 47(1):159–210, Sep 2009. doi: 10.1146/annurev-astro-082708-101734.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. *arXiv e-prints*, art. arXiv:1505.05424, May 2015.

H. Böhringer, P. Schuecker, L. Guzzo, C. A. Collins, W. Voges, R. G. Cruddace, A. Ortiz-Gil, G. Chincarini, S. De Grandi, A. C. Edge, H. T. MacGillivray, D. M. Neumann, S. Schindler, and P. Shaver. The ROSAT-ESO Flux Limited X-ray (REFLEX) Galaxy cluster survey. V. The cluster catalogue. , 425:367–383, October 2004. doi: 10.1051/0004-6361:20034484.

Clecio Bom, Jason Poh, Brian Nord, Manuel Blanco-Valentin, and Luciana Dias. Deep Learning in Wide-field Surveys: Fast Analysis of Strong Lenses in Ground-based Cosmic Experiments. *arXiv e-prints*, art. arXiv:1911.06341, November 2019.

G. Bothun, C. Impey, and S. McGaugh. Low-Surface-Brightness Galaxies: Hidden Galaxies Revealed. , 109:745–758, Jul 1997. doi: 10.1086/133941.

Gregory D. Bothun, Christopher D. Impey, David F. Malin, and Jeremy R. Mould. Discovery of a Huge Low-Surface-Brightness Galaxy: A Proto-Disk Galaxy at Low Redshift? , 94:23, July 1987. doi: 10.1086/114443.

Gregory D. Bothun, James M. Schombert, Christopher D. Impey, and Stephen E. Schneider. Discovery of a Second Giant Low Surface Brightness Galaxy: Further Confirmation of Slowly Evolving Disk Galaxies. , 360:427, September 1990. doi: 10.1086/169133.

Gregory D. Bothun, Christopher D. Impey, and David F. Malin. Extremely Low Surface Brightness Galaxies in the Fornax Cluster: Properties, Stability, and Luminosity Fluctuations. , 376:404, August 1991. doi: 10.1086/170290.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL `http://dx.doi.org/10.1023/A%3A1010933404324`.

Sarah Brough, Chris Collins, Ricardo Demarco, Henry C. Ferguson, Gaspar Galaz, Benne Holwerda, Cristina Martinez-Lombilla, Chris Mihos, and Mireia Montes. The Vera Rubin Observatory Legacy Survey of Space and Time and the Low Surface Brightness Universe. *arXiv e-prints*, art. arXiv:2001.11067, January 2020.

Colin J. Burke, Patrick D. Aleo, Yu-Ching Chen, and et al. Deblending and classifying astronomical sources with Mask R-CNN deep learning. , 490(3):3952–3965, December 2019. doi: 10.1093/mnras/stz2845.

J. Caldeira, W. L. K. Wu, B. Nord, C. Avestruz, S. Trivedi, and K. T. Story. DeepCMB: Lensing reconstruction of the cosmic microwave background with deep neural networks. *Astronomy and Computing*, 28:100307, July 2019. doi: 10.1016/j.ascom.2019.100307.

João Caldeira and Brian Nord. Deeply Uncertain: Comparing Methods of Uncertainty Quantification in Deep Learning Algorithms. *arXiv e-prints*, art. arXiv:2004.10710, April 2020.

John M. Cannon, Riccardo Giovanelli, Martha P. Haynes, Steven Janowiecki, Angela Parker, John J. Salzer, Elizabeth A. K. Adams, Eric Engstrom, Shan Huang, Kristen B. W. McQuinn, Jürgen Ott, Amélie Saintonge, Evan D. Skillman, John Allan, Grace Erny, Palmer Fliss, and AnnaLeigh Smith. The Survey of H I in Extremely Low-mass Dwarfs (SHIELD). , 739(1):L22, September 2011. doi: 10.1088/2041-8205/739/1/L22.

Timothy Carleton, Raphaël Errani, Michael Cooper, Manoj Kaplinghat, Jorge Peñarrubia, and Yicheng Guo. The formation of ultra-diffuse galaxies in cored dark matter haloes through tidal stripping and heating. , 485(1):382–395, May 2019. doi: 10.1093/mnras/stz383.

W. Cerny, A. B. Pace, A. Drlica-Wagner, P. S. Ferguson, S. Mau, M. Adamów, J. L. Carlin, Y. Choi, D. Erkal, L. C. Johnson, T. S. Li, C. E. Martínez-Vázquez, B. Mutlu-Pakdil, D. L. Nidever, K. A. G. Olsen, A. Pieres, E. J. Tollerud, J. D. Simon, A. K. Vivas, D. J. James, N. Kuropatkin, S. Majewski, D. Martínez-Delgado, P. Massana, A. E. Miller, E. H. Neilsen, N. E. D. Noël, A. H. Riley, D. J. Sand, L. Santana-Silva, G. S. Stringfellow, D. L. Tucker, and Delve Collaboration. Discovery of an Ultra-faint Stellar System near the Magellanic Clouds with the DECam Local Volume Exploration Survey. , 910(1):18, March 2021. doi: 10.3847/1538-4357/abe1af.

Chihway Chang, Alex Drlica-Wagner, Stephen M. Kent, Brian Nord, Donah Michelle Wang, and Michael H. L. S. Wang. A Machine Learning Approach to the Detection of Ghosting and Scattered Light Artifacts in Dark Energy Survey Images. *arXiv e-prints*, art. arXiv:2105.10524, May 2021.

Ting-Yun Cheng, Christopher J. Conselice, Alfonso Aragón-Salamanca, and et al. Optimizing automatic morphological classification of galaxies with machine learning and deep learning using Dark Energy Survey imaging. , 493(3):4209–4228, April 2020. doi: 10.1093/mnras/staa501.

Ting-Yun Cheng, Christopher J. Conselice, Alfonso Aragón-Salamanca, and et al. Galaxy Morphological Classification Catalogue of the Dark Energy Survey Year 3 data with Convolutional Neural Networks. *arXiv e-prints*, art. arXiv:2107.10210, July 2021.

Andrea Cimatti, Filippo Fraternali, and Carlo Nipoti. Introduction to Galaxy Formation and Evolution. From Primordial Gas to Present-Day Galaxies. *arXiv e-prints*, art. arXiv:1912.06216, December 2019.

L. Ciotti and G. Bertin. Analytical properties of the R$^{1/m}$ law. , 352:447–451, December 1999.

A. Ćiprijanović, D. Kafkes, S. Jenkins, and et al. Domain adaptation techniques for improved cross-domain study of galaxy mergers. *arXiv e-prints*, art. arXiv:2011.03591, November 2020a.

A. Ćiprijanović, G. F. Snyder, B. Nord, and J. E. G. Peek. DeepMerge: Classifying high-redshift merging galaxies with deep neural networks. *Astronomy and Computing*, 32: 100390, July 2020b. doi: 10.1016/j.ascom.2020.100390.

A. Ćiprijanović, D. Kafkes, K. Downey, S. Jenkins, G. N. Perdue, S. Madireddy, T. Johnston, G. F. Snyder, and B. Nord. DeepMerge II: Building Robust Deep Learning Algorithms for Merging Galaxy Identification Across Domains. *arXiv e-prints*, art. arXiv:2103.01373, March 2021.

Yotam Cohen, Pieter van Dokkum, Shany Danieli, and et al. The Dragonfly Nearby Galaxies Survey. V. HST/ACS Observations of 23 Low Surface Brightness Objects in the Fields of NGC 1052, NGC 1084, M96, and NGC 4258. , 868(2):96, December 2018a. doi: 10.3847/1538-4357/aae7c8.

Yotam Cohen, Pieter van Dokkum, Shany Danieli, Aaron J. Romanowsky, Roberto Abraham, Allison Merritt, Jielai Zhang, Lamiya Mowla, J. M. Diederik Kruijssen, Charlie Conroy, and Asher Wasserman. The Dragonfly Nearby Galaxies Survey. V. HST/ACS Observations of 23 Low Surface Brightness Objects in the Fields of NGC 1052, NGC 1084, M96, and NGC 4258. , 868(2):96, December 2018b. doi: 10.3847/1538-4357/aae7c8.

Andrew J. Connolly, Ryan Scranton, David Johnston, Scott Dodelson, Daniel J. Eisenstein, Joshua A. Frieman, James E. Gunn, Lam Hui, Bhuvnesh Jain, Stephen Kent, Jon Loveday, Robert C. Nichol, Liam O'Connell, Marc Postman, Roman Scoccimarro, Ravi K. Sheth, Albert Stebbins, Michael A. Strauss, Alexander S. Szalay, István Szapudi, Max Tegmark, Michael S. Vogeley, Idit Zehavi, James Annis, Neta Bahcall, J. Brinkmann, István Csabai, Mamoru Doi, Masataka Fukugita, G. S. Hennessy, Robert Hindsley, Takashi Ichikawa, Željko Ivezić, Rita S. J. Kim, Gillian R. Knapp, Peter Kunszt, D. Q. Lamb, Brian C. Lee, Robert H. Lupton, Timothy A. McKay, Jeff Munn, John Peoples, Jeff Pier, Constance Rockosi, David Schlegel, Christopher Stoughton, Douglas L. Tucker, Brian Yanny, and Donald G. York. The Angular Correlation Function of Galaxies from Early Sloan Digital Sky Survey Data. , 579(1):42–47, November 2002. doi: 10.1086/342787.

Charlie Conroy. Modeling the Panchromatic Spectral Energy Distributions of Galaxies. , 51 (1):393–455, August 2013. doi: 10.1146/annurev-astro-082812-141017.

Christopher J. Conselice. Ultra-diffuse Galaxies Are a Subset of Cluster Dwarf Elliptical/Spheroidal Galaxies. *Research Notes of the American Astronomical Society*, 2(1):43, March 2018. doi: 10.3847/2515-5172/aab7f6.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

Stéphane Courteau, Aaron A. Dutton, Frank C. van den Bosch, Lauren A. MacArthur, Avishai Dekel, Daniel H. McIntosh, and Daniel A. Dale. Scaling Relations of Spiral Galaxies. , 671(1):203–225, December 2007. doi: 10.1086/522193.

James G. Cresswell and Will J. Percival. Scale-dependent galaxy bias in the Sloan Digital Sky Survey as a function of luminosity and colour. , 392(2):682–690, January 2009. doi: 10.1111/j.1365-2966.2008.14082.x.

Martín Crocce, Anna Cabré, and Enrique Gaztañaga. Modelling the angular correlation function and its full covariance in photometric galaxy surveys. , 414(1):329–349, June 2011. doi: 10.1111/j.1365-2966.2011.18393.x.

Andrew Crumey. Human contrast threshold and astronomical visibility. , 442(3):2600–2619, August 2014. doi: 10.1093/mnras/stu992.

Jia-Ming Dai and Jizhou Tong. Galaxy Morphology Classification with Deep Convolutional Neural Networks. *arXiv e-prints*, art. arXiv:1807.10406, July 2018.

Julianne J. Dalcanton, David N. Spergel, James E. Gunn, Maarten Schmidt, and Donald P. Schneider. The Number Density of Low-Surface Brightness Galaxies with 23 &lt; mu_0 &lt; 25 V Mag/arcsec$\hat{2}$. , 114:635–654, Aug 1997. doi: 10.1086/118499.

Shany Danieli, Pieter van Dokkum, Allison Merritt, Roberto Abraham, Jielai Zhang, I. D. Karachentsev, and L. N. Makarova. The Dragonfly Nearby Galaxies Survey. III. The Luminosity Function of the M101 Group. , 837(2):136, Mar 2017. doi: 10.3847/1538-4357/aa615b.

Dark Energy Survey Collaboration, T. Abbott, F. B. Abdalla, J. Aleksić, S. Allam, A. Amara, D. Bacon, E. Balbinot, M. Banerji, K. Bechtol, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, J. Blazek, C. Bonnett, S. Bridle, D. Brooks, R. J. Brunner, E. Buckley-Geer, D. L. Burke, G. B. Caminha, D. Capozzi, J. Carlsen, A. Carnero-Rosell, M. Carollo, M. Carrasco-Kind, J. Carretero, F. J. Castander, L. Clerkin, T. Collett, C. Conselice, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, T. M. Davis, S. Desai, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, A. Drlica-Wagner, J. Estrada, J. Etherington, A. E. Evrard, J. Fabbri, D. A. Finley, B. Flaugher, R. J. Foley, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, D. A. Goldstein, D. Gruen, R. A. Gruendl, P. Guarnieri, G. Gutierrez, W. Hartley, K. Honscheid, B. Jain, D. J. James, T. Jeltema, S. Jouvel, R. Kessler, A. King, D. Kirk, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, T. S. Li, M. Lima, H. Lin, M. A. G. Maia, M. Makler, M. Manera, C. Maraston, J. L. Marshall, P. Martini, R. G. McMahon, P. Melchior, A. Merson, C. J. Miller, R. Miquel, J. J. Mohr, X. Morice-Atkinson, K. Naidoo, E. Neilsen, R. C. Nichol, B. Nord, R. Ogando, F. Ostrovski, A. Palmese, A. Papadopoulos, H. V. Peiris, J. Peoples, W. J. Percival, A. A. Plazas, S. L. Reed, A. Refregier, A. K. Romer, A. Roodman, A. Ross, E. Rozo, E. S. Rykoff, I. Sadeh, M. Sako, C. Sánchez, E. Sanchez, B. Santiago, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, E. Sheldon, M. Smith, R. C. Smith, M. Soares-Santos, F. Sobreira, M. Soumagnac, E. Suchyta, M. Sullivan, M. Swanson, G. Tarle, J. Thaler,

D. Thomas, R. C. Thomas, D. Tucker, J. D. Vieira, V. Vikram, A. R. Walker, R. H. Wechsler, J. Weller, W. Wester, L. Whiteway, H. Wilcox, B. Yanny, Y. Zhang, and J. Zuntz. The Dark Energy Survey: more than dark energy - an overview. , 460(2):1270–1299, August 2016. doi: 10.1093/mnras/stw641.

Andrew Davies, Stephen Serjeant, and Jane M. Bromley. Using convolutional neural networks to identify gravitational lenses in astronomical images. , 487(4):5263–5271, August 2019. doi: 10.1093/mnras/stz1288.

M. Davis, G. Efstathiou, C. S. Frenk, and S. D. M. White. The evolution of large-scale structure in a universe dominated by cold dark matter. , 292:371–394, May 1985. doi: 10.1086/163168.

J. De Vicente, E. Sánchez, and I. Sevilla-Noarbe. DNF - Galaxy photometric redshift by Directional Neighbourhood Fitting. , 459(3):3078–3088, July 2016. doi: 10.1093/mnras/stw857.

DES Collaboration. The Dark Energy Survey. *arXiv e-prints*, art. astro-ph/0510346, October 2005.

DES Collaboration. The Dark Energy Survey: more than dark energy - an overview. , 460 (2):1270–1299, August 2016. doi: 10.1093/mnras/stw641.

DES Collaboration, T. M. C. Abbott, F. B. Abdalla, A. Alarcon, and et al. Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing. , 98(4):043526, August 2018a. doi: 10.1103/PhysRevD.98.043526.

DES Collaboration, T. M. C. Abbott, F. B. Abdalla, S. Allam, A. Amara, J. Annis, J. Asorey, S. Avila, O. Ballester, M. Banerji, W. Barkhouse, L. Baruah, M. Baumer, K. Bechtol, M. R. Becker, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, J. Blazek, S. Bocquet, D. Brooks, D. Brout, E. Buckley-Geer, D. L. Burke, V. Busti, R. Campisano, L. Cardiel-Sas, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, X. Chen, C. Conselice, G. Costa, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, R. Das, G. Daues, T. M. Davis, C. Davis, J. De Vicente, D. L. DePoy, J. DeRose, S. Desai, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, A. Drlica-Wagner, T. F. Eifler, A. E. Elliott, A. E. Evrard, A. Farahi, A. Fausti Neto, E. Fernandez, D. A. Finley, B. Flaugher, R. J. Foley, P. Fosalba, D. N. Friedel, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, M. S. S. Gill, K. Glazebrook, D. A. Goldstein, M. Gower, D. Gruen, R. A. Gruendl, J. Gschwend, R. R. Gupta, G. Gutierrez, S. Hamilton, W. G. Hartley, S. R. Hinton, J. M. Hislop, D. Hollowood, K. Honscheid, B. Hoyle, D. Huterer, B. Jain, D. J. James, T. Jeltema, M. W. G. Johnson, M. D. Johnson, T. Kacprzak, S. Kent, G. Khullar, M. Klein, A. Kovacs, A. M. G. Koziol, E. Krause, A. Kremin, R. Kron, K. Kuehn, S. Kuhlmann, N. Kuropatkin, O. Lahav, J. Lasker, T. S. Li, R. T. Li, A. R. Liddle, M. Lima, H. Lin, P. López-Reyes, N. MacCrann, M. A. G. Maia, J. D. Maloney, M. Manera, M. March, J. Marriner, J. L. Marshall, P. Martini, T. McClintock, T. McKay, R. G. McMahon, P. Melchior, F. Menanteau, C. J. Miller, R. Miquel, J. J.

Mohr, E. Morganson, J. Mould, E. Neilsen, R. C. Nichol, F. Nogueira, B. Nord, P. Nugent, L. Nunes, R. L. C. Ogand o, L. Old, A. B. Pace, A. Palmese, F. Paz-Chinchón, H. V. Peiris, W. J. Percival, D. Petravick, A. A. Plazas, J. Poh, C. Pond, A. Porredon, A. Pujol, A. Refregier, K. Reil, P. M. Ricker, R. P. Rollins, A. K. Romer, A. Roodman, P. Rooney, A. J. Ross, E. S. Rykoff, M. Sako, M. L. Sanchez, E. Sanchez, B. Santiago, A. Saro, V. Scarpine, D. Scolnic, S. Serrano, I. Sevilla-Noarbe, E. Sheldon, N. Shipp, M. L. Silveira, M. Smith, R. C. Smith, J. A. Smith, M. Soares-Santos, F. Sobreira, J. Song, A. Stebbins, E. Suchyta, M. Sullivan, M. E. C. Swanson, G. Tarle, J. Thaler, D. Thomas, R. C. Thomas, M. A. Troxel, D. L. Tucker, V. Vikram, A. K. Vivas, A. R. Walker, R. H. Wechsler, J. Weller, W. Wester, R. C. Wolf, H. Wu, B. Yanny, A. Zenteno, Y. Zhang, J. Zuntz, S. Juneau, M. Fitzpatrick, R. Nikutta, D. Nidever, K. Olsen, and A. Scott. The Dark Energy Survey: Data Release 1. *ApJS*, 239(2):18, Dec 2018b. doi: 10.3847/1538-4365/aae9f0.

DES Collaboration, T. M. C. Abbott, F. B. Abdalla, S. Allam, A. Amara, et al. The Dark Energy Survey: Data Release 1. *ApJS*, 239(2):18, December 2018c. doi: 10.3847/1538-4365/aae9f0.

DES Collaboration, T. M. C. Abbott, M. Adamow, M. Aguena, and et al. The Dark Energy Survey Data Release 2. *arXiv e-prints*, art. arXiv:2101.05765, January 2021a.

DES Collaboration, T. M. C. Abbott, M. Aguena, A. Alarcon, and et al. Dark Energy Survey Year 3 Results: Cosmological Constraints from Galaxy Clustering and Weak Lensing. *arXiv e-prints*, art. arXiv:2105.13549, May 2021b.

S. Desai, J. J. Mohr, E. Bertin, M. Kümmel, and M. Wetzstein. Detection and removal of artifacts in astronomical images. *Astronomy and Computing*, 16:67–78, July 2016. doi: 10.1016/j.ascom.2016.04.002.

Arjun Dey, David J. Schlegel, Dustin Lang, et al. Overview of the DESI Legacy Imaging Surveys. , 157(5):168, May 2019. doi: 10.3847/1538-3881/ab089d.

Sander Dieleman, Kyle W. Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. , 450(2):1441–1459, June 2015. doi: 10.1093/mnras/stv632.

M. J. Disney. Visibility of galaxies. , 263(5578):573–575, Oct 1976. doi: 10.1038/263573a0.

H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D. Tuccillo, and J. L. Fischer. Improving galaxy morphologies for SDSS with Deep Learning. , 476(3):3661–3676, February 2018. doi: 10.1093/mnras/sty338.

H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, and et al. Transfer learning for galaxy morphology from one survey to another. , 484(1):93–100, March 2019a. doi: 10.1093/mnras/sty3497.

H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, and et al. Transfer learning for galaxy morphology from one survey to another. , 484(1):93–100, March 2019b. doi: 10.1093/mnras/sty3497.

Michael J. Drinkwater, Michael D. Gregg, and Matthew Colless. Substructure and Dynamics of the Fornax Cluster. , 548(2):L139–L142, February 2001. doi: 10.1086/319113.

S. P. Driver. The Contribution of Normal, Dim, and Dwarf Galaxies to the Local Luminosity Density. , 526(2):L69–L72, Dec 1999. doi: 10.1086/312379.

A. Drlica-Wagner, K. Bechtol, E. S. Rykoff, E. Luque, A. Queiroz, Y. Y. Mao, R. H. Wechsler, J. D. Simon, B. Santiago, B. Yanny, E. Balbinot, S. Dodelson, A. Fausti Neto, D. J. James, T. S. Li, M. A. G. Maia, J. L. Marshall, A. Pieres, K. Stringer, A. R. Walker, T. M. C. Abbott, F. B. Abdalla, S. Allam, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, D. Brooks, E. Buckley-Geer, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, M. Crocce, L. N. da Costa, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, T. F. Eifler, A. E. Evrard, D. A. Finley, B. Flaugher, P. Fosalba, J. Frieman, E. Gaztanaga, D. W. Gerdes, D. Gruen, R. A. Gruendl, G. Gutierrez, K. Honscheid, K. Kuehn, N. Kuropatkin, O. Lahav, P. Martini, R. Miquel, B. Nord, R. Ogando, A. A. Plazas, K. Reil, A. Roodman, M. Sako, E. Sanchez, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Tucker, V. Vikram, W. Wester, Y. Zhang, J. Zuntz, and DES Collaboration. Eight Ultra-faint Galaxy Candidates Discovered in Year Two of the Dark Energy Survey. , 813(2):109, November 2015. doi: 10.1088/0004-637X/813/2/109.

A. Drlica-Wagner, K. Bechtol, S. Mau, and et al. Milky Way Satellite Census. I. The Observational Selection Function for Milky Way Satellites in DES Y3 and Pan-STARRS DR1. , 893(1):47, April 2020. doi: 10.3847/1538-4357/ab7eb9.

Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6889-6/19/10. doi: 10.1145/3343031.3350535. URL https://doi.org/10.1145/3343031.3350535.

Cora Dvorkin, Siddharth Mishra-Sharma, Brian Nord, V. Ashley Villar, Camille Avestruz, Keith Bechtol, Aleksandra Ćiprijanović, Andrew J. Connolly, Lehman H. Garrison, Gautham Narayan, and Francisco Villaescusa-Navarro. Machine Learning and Cosmology. *arXiv e-prints*, art. arXiv:2203.08056, March 2022.

Bradley Efron and Gail Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983. ISSN 00031305. URL http://www.jstor.org/stable/2685844.

Paul Eigenthaler, Thomas H. Puzia, Matthew A. Taylor, Yasna Ordenes-Briceño, Roberto P. Muñoz, Karen X. Ribbeck, Karla A. Alamo-Martínez, Hongxin Zhang, Simón Ángel, Massimo Capaccioli, Patrick Côté, Laura Ferrarese, Gaspar Galaz, Eva K. Grebel, Maren Hempel, Michael Hilker, Ariane Lançon, Steffen Mieske, Bryan Miller, Maurizio Paolillo, Mathieu Powalka, Tom Richtler, Joel Roediger, Yu Rong, Ruben Sánchez-Janssen, and Chelsea Spengler. The Next Generation Fornax Survey (NGFS). II. The Central Dwarf Galaxy Population. , 855(2):142, Mar 2018. doi: 10.3847/1538-4357/aaab60.

Peter Erwin. IMFIT: A Fast, Flexible New Program for Astronomical Image Fitting. , 799 (2):226, February 2015. doi: 10.1088/0004-637X/799/2/226.

Euclid Collaboration, A. S. Borlaff, P. Gómez-Alvarez, B. Altieri, P. M. Marcum, R. Vavrek, R. Laureijs, R. Kohley, F. Buitrago, J. C. Cuillandre, P. A. Duc, L. M. Gaspar Venancio, A. Amara, S. Andreon, N. Auricchio, R. Azzollini, C. Baccigalupi, A. Balaguera-Antolínez, M. Baldi, S. Bardelli, R. Bender, A. Biviano, C. Bodendorf, D. Bonino, E. Bozzo, E. Branchini, M. Brescia, J. Brinchmann, C. Burigana, R. Cabanac, S. Camera, G. P. Candini, V. Capobianco, A. Cappi, C. Carbone, J. Carretero, C. S. Carvalho, S. Casas, F. J. Castander, M. Castellano, G. Castignani, S. Cavuoti, A. Cimatti, R. Cledassou, C. Colodro-Conde, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, L. Corcione, J. Coupon, H. M. Courtois, M. Cropper, A. Da Silva, H. Degaudenzi, D. Di Ferdinando, M. Douspis, F. Dubath, C. A. J. Duncan, X. Dupac, S. Dusini, A. Ealet, M. Fabricius, M. Farina, S. Farrens, P. G. Ferreira, S. Ferriol, F. Finelli, P. Flose-Reimberg, P. Fosalba, M. Frailis, E. Franceschi, M. Fumana, S. Galeotta, K. Ganga, B. Garilli, B. Gillis, C. Giocoli, G. Gozaliasl, J. Graciá-Carpio, A. Grazian, F. Grupp, S. V. H. Haugan, W. Holmes, F. Hormuth, K. Jahnke, E. Keihanen, S. Kermiche, A. Kiessling, M. Kilbinger, C. C. Kirkpatrick, T. Kitching, J. H. Knapen, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, P. Liebing, S. Ligori, P. B. Lilje, V. Lindholm, I. Lloro, G. Mainetti, D. Maino, O. Mansutti, O. Marggraf, K. Markovic, M. Martinelli, N. Martinet, D. Martínez-Delgado, F. Marulli, R. Massey, M. Maturi, S. Maurogordato, E. Medinaceli, S. Mei, M. Meneghetti, E. Merlin, R. B. Metcalf, G. Meylan, M. Moresco, G. Morgante, L. Moscardini, E. Munari, R. Nakajima, C. Neissner, S. M. Niemi, J. W. Nightingale, A. Nucita, C. Padilla, S. Paltani, F. Pasian, L. Patrizii, K. Pedersen, W. J. Percival, V. Pettorino, S. Pires, M. Poncet, L. Popa, D. Potter, L. Pozzetti, F. Raison, R. Rebolo, A. Renzi, J. Rhodes, G. Riccio, E. Romelli, M. Roncarelli, C. Rosset, E. Rossetti, R. Saglia, A. G. Sánchez, D. Sapone, M. Sauvage, P. Schneider, V. Scottez, A. Secroun, G. Seidel, S. Serrano, C. Sirignano, G. Sirri, J. Skottfelt, L. Stanco, J. L. Starck, F. Sureau, P. Tallada-Crespí, A. N. Taylor, M. Tenti, I. Tereno, R. Teyssier, R. Toledo-Moreo, F. Torradeflot, I. Tutusaus, E. A. Valentijn, L. Valenziano, J. Valiviita, T. Vassallo, M. Viel, Y. Wang, J. Weller, L. Whittaker, A. Zacchei, G. Zamorani, and E. Zucca. Euclid preparation. XVI. Exploring the ultra-low surface brightness Universe with Euclid/VIS. , 657:A92, January 2022. doi: 10.1051/0004-6361/202141935.

S. Everett, B. Yanny, et al. Measuring the transfer function of the dark energy survey with balrog. *in prep*, in prep.

Mark Everingham, Luc Van Gool, C. K. I. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge, 2010.

Adina D. Feinstein, Benjamin T. Montet, Megan Ansdell, and et al. Flare Statistics for Young Stars from a Convolutional Neural Network Analysis of *TESS* Data. *arXiv e-prints*, art. arXiv:2005.07710, May 2020.

Henry C. Ferguson. Population Studies in Groups and Clusters of Galaxies. II. A Catalog of Galaxies in the Central 3.5 Degrees of the Fornax Cluster. , 98:367, August 1989. doi: 10.1086/115152.

G. J. Ferland, R. L. Porter, P. A. M. van Hoof, R. J. R. Williams, N. P. Abel, M. L. Lykins, G. Shaw, W. J. Henney, and P. C. Stancil. The 2013 Release of Cloudy. , 49:137–163, Apr 2013.

Ismael Ferrero, Mario G. Abadi, Julio F. Navarro, Laura V. Sales, and Sebastián. Gurovich. The dark matter haloes of dwarf galaxies: a challenge for the Λ cold dark matter paradigm? , 425(4):2817–2823, Oct 2012. doi: 10.1111/j.1365-2966.2012.21623.x.

Edward L. Fitzpatrick. Correcting for the Effects of Interstellar Extinction. , 111(755): 63–75, January 1999. doi: 10.1086/316293.

B. Flaugher, H. T. Diehl, K. Honscheid, and et al. The Dark Energy Camera. , 150(5):150, November 2015. doi: 10.1088/0004-6256/150/5/150.

K. C. Freeman. On the Disks of Spiral and S0 Galaxies. , 160:811, Jun 1970. doi: 10.1086/150474.

B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv e-prints*, art. arXiv:1506.02142, June 2015.

Gaspar Galaz, Rodrigo Herrera-Camus, Diego Garcia-Lambas, and Nelson Padilla. Low Surface Brightness Galaxies in the SDSS: The Link Between Environment, Star-forming Properties, and Active Galactic Nuclei. , 728(2):74, Feb 2011. doi: 10.1088/0004-637X/728/2/74.

Marla Geha, Risa H. Wechsler, Yao-Yuan Mao, Erik J. Tollerud, Benjamin Weiner, Rebecca Bernstein, Ben Hoyle, Sebastian Marchi, Phil J. Marshall, Ricardo Muñoz, and Yu Lu. The SAGA Survey. I. Satellite Galaxy Populations around Eight Milky Way Analogs. , 847(1):4, Sep 2017. doi: 10.3847/1538-4357/aa8626.

Daniel George, Hongyu Shen, and E. A. Huerta. Classification and unsupervised clustering of LIGO data with Deep Transfer Learning. , 97(10):101501, May 2018. doi: 10.1103/PhysRevD.97.101501.

Colleen Gilhuly, David Hendel, Allison Merritt, Roberto Abraham, Shany Danieli, Deborah Lokhorst, Qing Liu, Pieter van Dokkum, Charlie Conroy, and Johnny Greco. The Dragonfly Edge-on Galaxies Survey: Shaping the Outer disk of NGC 4565 via Accretion. , 897 (2):108, July 2020. doi: 10.3847/1538-4357/ab9b25.

D. A. Goldstein, C. B. D'Andrea, J. A. Fischer, and et al. Automated Transient Identification in the Dark Energy Survey. , 150(3):82, September 2015. doi: 10.1088/0004-6256/150/3/82.

R. E. González, R. P. Muñoz, and C. A. Hernández. Galaxy detection and identification using deep learning and data augmentation. *Astronomy and Computing*, 25:103–109, October 2018. doi: 10.1016/j.ascom.2018.09.004.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016. ISBN 0262035618.

K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. , 622:759–771, April 2005. doi: 10.1086/427976.

Alister W. Graham and Simon P. Driver. A Concise Reference to (Projected) Sérsic $R^{1/n}$ Quantities, Including Concentration, Profile Slopes, Petrosian Indices, and Kron Magnitudes. , 22(2):118–127, January 2005. doi: 10.1071/AS05001.

Johnny P. Greco, Jenny E. Greene, Michael A. Strauss, Lauren A. Macarthur, Xzavier Flowers, Andy D. Goulding, Song Huang, Ji Hoon Kim, Yutaka Komiyama, Alexie Leauthaud, Lukas Leisman, Robert H. Lupton, Cristóbal Sifón, and Shiang-Yu Wang. Illuminating Low Surface Brightness Galaxies with the Hyper Suprime-Cam Survey. , 857(2):104, Apr 2018. doi: 10.3847/1538-4357/aab842.

Johnny P. Greco, Pieter van Dokkum, Shany Danieli, Scott G. Carlsten, and Charlie Conroy. Measuring Distances to Low-luminosity Galaxies Using Surface Brightness Fluctuations. , 908(1):24, February 2021. doi: 10.3847/1538-4357/abd030.

Trevor Hastie. Ridge Regularizaton: an Essential Concept in Data Science. *arXiv e-prints*, art. arXiv:2006.00371, May 2020.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Boris Häußler, Steven P. Bamford, Marina Vika, Alex L. Rojas, Marco Barden, Lee S. Kelvin, Mehmet Alpaslan, Aaron S. G. Robotham, Simon P. Driver, I. K. Baldry, Sarah Brough, Andrew M. Hopkins, Jochen Liske, Robert C. Nichol, Cristina C. Popescu, and Richard J. Tuffs. MegaMorph - multiwavelength measurement of galaxy structure: complete Sérsic profile information from modern surveys. , 430(1):330–369, Mar 2013. doi: 10.1093/mnras/sts633.

C. C. Hayward, J. A. Irwin, and J. N. Bregman. The Cosmological Unimportance of Low Surface Brightness Galaxies. , 635(2):827–831, Dec 2005. doi: 10.1086/497565.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv e-prints*, art. arXiv:1512.03385, December 2015.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *arXiv e-prints*, art. arXiv:1703.06870, March 2017.

M. Hilker, M. Kissler-Patig, T. Richtler, L. Infante, and H. Quintana. The central region of the Fornax cluster. I. A catalog and photometric properties of galaxies in selected CCD fields. , 134:59–73, January 1999. doi: 10.1051/aas:1999433.

David W. Hogg, Michael R. Blanton, Daniel J. Eisenstein, James E. Gunn, David J. Schlegel, Idit Zehavi, Neta A. Bahcall, Jon Brinkmann, Istvan Csabai, Donald P. Schneider, David H. Weinberg, and Donald G. York. The Overdensities of Galaxy Environments as a Function of Luminosity and Color. , 585(1):L5–L9, March 2003. doi: 10.1086/374238.

B. Hoyle, D. Gruen, G. M. Bernstein, M. M. Rau, J. De Vicente, W. G. Hartley, E. Gaztanaga, J. DeRose, M. A. Troxel, C. Davis, A. Alarcon, N. MacCrann, J. Prat, C. Sánchez, E. Sheldon, R. H. Wechsler, J. Asorey, M. R. Becker, C. Bonnett, A. Carnero Rosell, D. Carollo, M. Carrasco Kind, F. J. Castander, R. Cawthon, C. Chang, M. Childress, T. M. Davis, A. Drlica-Wagner, M. Gatti, K. Glazebrook, J. Gschwend, S. R. Hinton, J. K. Hoormann, A. G. Kim, A. King, K. Kuehn, G. Lewis, C. Lidman, H. Lin, E. Macaulay, M. A. G. Maia, P. Martini, D. Mudd, A. Möller, R. C. Nichol, R. L. C. Ogando, R. P. Rollins, A. Roodman, A. J. Ross, E. Rozo, E. S. Rykoff, S. Samuroff, I. Sevilla-Noarbe, R. Sharp, N. E. Sommer, B. E. Tucker, S. A. Uddin, T. N. Varga, P. Vielzeuf, F. Yuan, B. Zhang, T. M. C. Abbott, F. B. Abdalla, S. Allam, J. Annis, K. Bechtol, A. Benoit-Lévy, E. Bertin, D. Brooks, E. Buckley-Geer, D. L. Burke, M. T. Busha, D. Capozzi, J. Carretero, M. Crocce, C. B. D'Andrea, L. N. da Costa, D. L. DePoy, S. Desai, H. T. Diehl, P. Doel, T. F. Eifler, J. Estrada, A. E. Evrard, E. Fernandez, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, D. W. Gerdes, T. Giannantonio, D. A. Goldstein, R. A. Gruendl, G. Gutierrez, K. Honscheid, D. J. James, M. Jarvis, T. Jeltema, M. W. G. Johnson, M. D. Johnson, D. Kirk, E. Krause, S. Kuhlmann, N. Kuropatkin, O. Lahav, T. S. Li, M. Lima, M. March, J. L. Marshall, P. Melchior, F. Menanteau, R. Miquel, B. Nord, C. R. O'Neill, A. A. Plazas, A. K. Romer, M. Sako, E. Sanchez, B. Santiago, V. Scarpine, R. Schindler, M. Schubnell, M. Smith, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, D. L. Tucker, V. Vikram, A. R. Walker, J. Weller, W. Wester, R. C. Wolf, B. Yanny, J. Zuntz, and DES Collaboration. Dark Energy Survey Year 1 Results: redshift distributions of the weak-lensing source galaxies. , 478(1):592–610, July 2018. doi: 10.1093/mnras/sty957.

X. Huang, C. Storfer, V. Ravi, A. Pilon, M. Domingo, D. J. Schlegel, S. Bailey, A. Dey, R. R. Gupta, D. Herrera, S. Juneau, M. Landriau, D. Lang, A. Meisner, J. Moustakas, A. D. Myers, E. F. Schlafly, F. Valdes, B. A. Weaver, J. Yang, and C. Yèche. Finding Strong Gravitational Lenses in the DESI DECam Legacy Survey. , 894(1):78, May 2020. doi: 10.3847/1538-4357/ab7ffb.

X. Huang, C. Storfer, A. Gu, and et al. Discovering New Strong Gravitational Lenses in the DESI Legacy Imaging Surveys. , 909(1):27, March 2021. doi: 10.3847/1538-4357/abd62b.

E. P. Hubble. Cepheids in spiral nebulae. *The Observatory*, 48:139–142, May 1925.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *arXiv e-prints*, art. arXiv:1910.09457, October 2019.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Chris Impey and Greg Bothun. Low surface brightness galaxies. *Annual Review of Astronomy and Astrophysics*, 35(1):267–307, 1997. ISSN 0066-4146. doi: 10.1146/annurev.astro.35.1.267.

Chris Impey, Greg Bothun, and David Malin. Virgo Dwarfs: New Light on Faint Galaxies. , 330:634, July 1988. doi: 10.1086/166500.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv e-prints*, art. arXiv:1502.03167, February 2015.

Haris Iqbal. Harisiqbal88/plotneuralnet v1.0.0, December 2018. URL `https://doi.org/10.5281/zenodo.2526396`.

Zeljko Ivezic, Andrew J. Connolly, Jacob T. VanderPlas, and Alexander Gray. *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton University Press, USA, 2014. ISBN 0691151687.

Željko Ivezić, Steven M. Kahn, J. Anthony Tyson, and et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. , 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c.

Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2): 37–50, 1912. doi: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x. URL `https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x`.

R. A. Jackson, G. Martin, S. Kaviraj, M. Ramsøy, J. E. G. Devriendt, T. Sedgwick, C. Laigle, H. Choi, R. S. Beckmann, M. Volonteri, Y. Dubois, C. Pichon, S. K. Yi, A. Slyz, K. Kraljic, T. Kimm, S. Peirani, and I. Baldry. The origin of low-surface-brightness galaxies in the dwarf regime. , 502(3):4262–4276, April 2021. doi: 10.1093/mnras/stab077.

C. Jacobs, T. Collett, K. Glazebrook, et al. Finding high-redshift strong lenses in DES using convolutional neural networks. , 484(4):5330–5349, April 2019. doi: 10.1093/mnras/stz272.

Steven Janssens, Roberto Abraham, Jean Brodie, Duncan Forbes, Aaron J. Romanowsky, and Pieter van Dokkum. Ultra-diffuse and Ultra-compact Galaxies in the Frontier Fields Cluster Abell 2744. , 839(1):L17, April 2017. doi: 10.3847/2041-8213/aa667d.

Mike Jarvis. TreeCorr: Two-point correlation functions, Aug 2015.

B. Javanmardi, D. Martinez-Delgado, P. Kroupa, C. Henkel, K. Crawford, K. Teuwen, R. J. Gabany, M. Hanson, T. S. Chonis, and F. Neyer. DGSAT: Dwarf Galaxy Survey with Amateur Telescopes. I. Discovery of low surface brightness systems around nearby spiral galaxies. , 588:A89, April 2016. doi: 10.1051/0004-6361/201527745.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001. URL `http://www.scipy.org/`.

Erin Kado-Fong, Mihai Petrescu, Majid Mohammad, Johnny Greco, Jenny E. Greene, Elizabeth A. K. Adams, Song Huang, Lukas Leisman, Ferah Munshi, Dimitrios Tanoglidis, and Jordan Van Nest. The Intrinsic Shapes of Low Surface Brightness Galaxies (LSBGs): A Discriminant of LSBG Galaxy Formation Mechanisms. , 920(2):72, October 2021. doi: 10.3847/1538-4357/ac15f0.

Sugata Kaviraj. The low-surface-brightness Universe: a new frontier in the study of galaxy evolution. *arXiv e-prints*, art. arXiv:2001.01728, January 2020.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *arXiv e-prints*, art. arXiv:1703.04977, March 2017.

Stephen M. Kent. Ghost Images in DECam. FERMILAB-SLIDES-20-114-SCD, 2013.

Asad Khan, E.A. Huerta, Sibo Wang, Robert Gruendl, Elise Jennings, and Huihuo Zheng. Deep learning at scale for the construction of galaxy catalogs in the dark energy survey. *Physics Letters B*, 795:248–258, 2019. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2019.06.009. URL `https://www.sciencedirect.com/science/article/pii/S0370269319303879`.

Edward J. Kim and Robert J. Brunner. Star-galaxy classification using deep convolutional neural networks. , 464(4):4463–4475, February 2017. doi: 10.1093/mnras/stw2672.

Alexei Y. Kniazev, Eva K. Grebel, Simon A. Pustilnik, Alexander G. Pramskij, Tamara F. Kniazeva, Francisco Prada, and Daniel Harbeck. Low Surface Brightness Galaxies in the Sloan Digital Sky Survey. I. Search Method and Test Sample. , 127(2):704–727, February 2004. doi: 10.1086/381061.

Jin Koda, Masafumi Yagi, Hitomi Yamanoi, and Yutaka Komiyama. Approximately a Thousand Ultra-diffuse Galaxies in the Coma Cluster. , 807(1):L2, Jul 2015. doi: 10.1088/2041-8205/807/1/L2.

Wouter M. Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv e-prints*, art. arXiv:1812.11806, December 2018.

Stephen D. Landy and Alexander S. Szalay. Bias and Variance of Angular Correlation Functions. , 412:64, July 1993. doi: 10.1086/172900.

François Lanusse, Quanbin Ma, Nan Li, and et al. CMU DeepLens: deep learning for automatic image-based galaxy-galaxy strong lens finding. , 473(3):3895–3906, January 2018. doi: 10.1093/mnras/stx1665.

R. B. Larson, B. M. Tinsley, and C. N. Caldwell. The evolution of disk galaxies and the origin of S0 galaxies. , 237:692–707, May 1980. doi: 10.1086/157917.

Jamie Law-Smith and Daniel J. Eisenstein. The Color and Stellar Mass Dependence of Small-scale Galaxy Clustering in SDSS-III BOSS. , 836(1):87, February 2017. doi: 10.3847/1538-4357/836/1/87.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539. URL `https://doi.org/10.1038/nature14539`.

R. Li, N. R. Napolitano, N. Roy, C. Tortora, F. La Barbera, A. Sonnenfeld, C. Qiu, and S. Liu. Galaxy Light Profile Convolutional Neural Networks (GaLNets). I. Fast and Accurate Structural Parameters for Billion-galaxy Samples. , 929(2):152, April 2022. doi: 10.3847/1538-4357/ac5ea0.

T. S. Li, E. Balbinot, N. Mondrik, J. L. Marshall, B. Yanny, K. Bechtol, A. Drlica-Wagner, D. Oscar, B. Santiago, J. D. Simon, A. K. Vivas, A. R. Walker, M. Y. Wang, T. M. C. Abbott, F. B. Abdalla, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, L. N. da Costa, D. L. DePoy, S. Desai, H. T. Diehl, P. Doel, J. Estrada, D. A. Finley, B. Flaugher, J. Frieman, D. Gruen, R. A. Gruendl, G. Gutierrez, K. Honscheid, D. J. James, K. Kuehn, N. Kuropatkin, O. Lahav, M. A. G. Maia, M. March, P. Martini, R. Ogando, A. A. Plazas, K. Reil, A. K. Romer, A. Roodman, E. Sanchez, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Tucker, Y. Zhang, and DES Collaboration. Discovery of a Stellar Overdensity in Eridanus-Phoenix in the Dark Energy Survey. , 817(2):135, February 2016. doi: 10.3847/0004-637X/817/2/135.

Tsung-Yi Lin, Michael Maire, Serge Belongie, and et al. Microsoft COCO: Common Objects in Context. *arXiv e-prints*, art. arXiv:1405.0312, May 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. *arXiv e-prints*, art. arXiv:1612.03144, December 2016.

Chris Lintott, Kevin Schawinski, Steven Bamford, Anå¾e Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert C. Nichol, M. Jordan Raddick, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. , 410(1):166–178, Jan 2011. doi: 10.1111/j.1365-2966.2010.17432.x.

LSST Science Collaboration. LSST Science Book, Version 2.0. *arXiv e-prints*, art. arXiv:0912.0201, December 2009.

E. Luque, A. Pieres, B. Santiago, B. Yanny, A. K. Vivas, A. Queiroz, A. Drlica-Wagner, E. Morganson, E. Balbinot, J. L. Marshall, T. S. Li, A. Fausti Neto, L. N. da Costa, M. A. G. Maia, K. Bechtol, A. G. Kim, G. M. Bernstein, S. Dodelson, L. Whiteway, H. T. Diehl, D. A. Finley, T. Abbott, F. B. Abdalla, S. Allam, J. Annis, A. Benoit-Lévy, E. Bertin, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, C. E. Cunha, C. B. D'Andrea, S. Desai, P. Doel, A. E. Evrard, B. Flaugher, P. Fosalba, D. W. Gerdes, D. A. Goldstein, D. Gruen, R. A. Gruendl, G. Gutierrez, D. J. James, K. Kuehn, N. Kuropatkin, O. Lahav, P. Martini, R. Miquel, B. Nord, R. Ogando, A. A. Plazas, A. K. Romer, E. Sanchez, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, and A. R. Walker. The Dark Energy Survey view of the Sagittarius stream: discovery of two faint stellar system candidates. , 468(1):97–108, June 2017. doi: 10.1093/mnras/stx405.

Ariyeh H. Maller, Daniel H. McIntosh, Neal Katz, and Martin D. Weinberg. The Galaxy Angular Correlation Functions and Power Spectrum from the Two Micron All Sky Survey. , 619(1):147–160, January 2005. doi: 10.1086/426181.

Paola Marigo, Léo Girardi, Alessandro Bressan, Philip Rosenfield, Bernhard Aringer, Yang Chen, Marco Dussin, Ambra Nanni, Giada Pastorelli, Thaíse S. Rodrigues, Michele Trabucchi, Sara Bladh, Julianne Dalcanton, Martin A. T. Groenewegen, Josefina Montalbán, and Peter R. Wood. A New Generation of PARSEC-COLIBRI Stellar Isochrones Including the TP-AGB Phase. , 835(1):77, January 2017. doi: 10.3847/1538-4357/835/1/77.

G. Martin, S. Kaviraj, C. Laigle, J. E. G. Devriendt, R. A. Jackson, S. Peirani, Y. Dubois, C. Pichon, and A. Slyz. The formation and evolution of low-surface-brightness galaxies. , 485(1):796–818, May 2019. doi: 10.1093/mnras/stz356.

G. Martin, A. E. Bazkiaei, M. Spavone, E. Iodice, J. C. Mihos, M. Montes, J. A. Benavides, S. Brough, J. L. Carlin, C. A. Collins, P. A. Duc, F. A. Gómez, G. Galaz, H. M. Hernández-Toledo, R. A. Jackson, S. Kaviraj, J. H. Knapen, C. Martínez-Lombilla, S. McGee, D. O'Ryan, D. J. Prole, R. M. Rich, J. Román, E. A. Shah, T. K. Starkenburg, A. E. Watkins, D. Zaritsky, C. Pichon, L. Armus, M. Bianconi, F. Buitrago, I. Busá, F. Davis, R. Demarco, A. Desmons, P. García, A. W. Graham, B. Holwerda, D. S. H. Hon, A. Khalid, J. Klehammer, D. Y. Klutse, I. Lazar, P. Nair, E. A. Noakes-Kettel, M. Rutkowski, K. Saha, N. Sahu, E. Sola, J. A. Vázquez-Mata, A. Vera-Casanova, and I. Yoon. Preparing for low surface brightness science with the Vera C. Rubin Observatory: Characterization of tidal features from mock images. , 513(1):1459–1487, June 2022. doi: 10.1093/mnras/stac1003.

Nicolas F. Martin, Rodrigo A. Ibata, Alan W. McConnachie, A. Dougal Mackey, Annette M. N. Ferguson, Michael J. Irwin, Geraint F. Lewis, and Mark A. Fardal. The PAndAS

View of the Andromeda Satellite System. I. A Bayesian Search for Dwarf Galaxies Using Spatial and Color-Magnitude Information. , 776(2):80, October 2013. doi: 10.1088/0004-637X/776/2/80.

Nicolas F. Martin, Rodrigo A. Ibata, Geraint F. Lewis, Alan McConnachie, Arif Babul, Nicholas F. Bate, Edouard Bernard, Scott C. Chapman, Michelle M. L. Collins, Anthony R. Conn, Denija Crnojević, Mark A. Fardal, Annette M. N. Ferguson, Michael Irwin, A. Dougal Mackey, Brendan McMonigal, Julio F. Navarro, and R. Michael Rich. The PAndAS View of the Andromeda Satellite System. II. Detailed Properties of 23 M31 Dwarf Spheroidal Galaxies. , 833(2):167, December 2016. doi: 10.3847/1538-4357/833/2/167.

David Martínez-Delgado, Ronald Läsker, Margarita Sharina, Elisa Toloba, Jürgen Fliri, Rachael Beaton, David Valls-Gabaud, Igor D. Karachentsev, Taylor S. Chonis, Eva K. Grebel, Duncan A. Forbes, Aaron J. Romanowsky, J. Gallego-Laborda, Karel Teuwen, M. A. Gómez-Flechoso, Jie Wang, Puragra Guhathakurta, Serafim Kaisin, and Nhung Ho. Discovery of an Ultra-diffuse Galaxy in the Pisces–Perseus Supercluster. , 151(4):96, April 2016. doi: 10.3847/0004-6256/151/4/96.

Alan W. McConnachie. The Observed Properties of Dwarf Galaxies in and around the Local Group. , 144(1):4, Jul 2012. doi: 10.1088/0004-6256/144/1/4.

Stacy S. McGaugh, Gregory D. Bothun, and James M. Schombert. Galaxy Selection and the Surface Brightness Distribution. , 110:573, Aug 1995a. doi: 10.1086/117543.

Stacy S. McGaugh, James M. Schombert, and Gregory D. Bothun. The Morphology of Low Surface Brightness Disk Galaxies. , 109:2019, May 1995b. doi: 10.1086/117427.

P. Melchior, E. Sheldon, A. Drlica-Wagner, and et al. Crowdsourcing quality control for Dark Energy Survey images. *Astronomy and Computing*, 16:99–108, July 2016. doi: 10.1016/j.ascom.2016.04.003.

Allison Merritt, Pieter van Dokkum, Shany Danieli, Roberto Abraham, Jielai Zhang, I. D. Karachentsev, and L. N. Makarova. The Dragonfly Nearby Galaxies Survey. II. Ultra-Diffuse Galaxies near the Elliptical Galaxy NGC 5485. , 833(2):168, Dec 2016a. doi: 10.3847/1538-4357/833/2/168.

Allison Merritt, Pieter van Dokkum, Shany Danieli, and et al. The Dragonfly Nearby Galaxies Survey. II. Ultra-Diffuse Galaxies near the Elliptical Galaxy NGC 5485. , 833(2):168, Dec 2016b. doi: 10.3847/1538-4357/833/2/168.

J. Christopher Mihos, Patrick R. Durrell, Laura Ferrarese, John J. Feldmeier, Patrick Côté, Eric W. Peng, Paul Harding, Chengze Liu, Stephen Gwyn, and Jean-Charles Cuillandre. Galaxies at the Extremes: Ultra-diffuse Galaxies in the Virgo Cluster. , 809(2):L21, Aug 2015. doi: 10.1088/2041-8205/809/2/L21.

J. Christopher Mihos, Paul Harding, John J. Feldmeier, Craig Rudick, Steven Janowiecki, Heather Morrison, Colin Slater, and Aaron Watkins. The Burrell Schmidt Deep Virgo

Survey: Tidal Debris, Galaxy Halos, and Diffuse Intracluster Light in the Virgo Cluster. , 834(1):16, Jan 2017. doi: 10.3847/1538-4357/834/1/16.

Shervin Minaee, Yuri Boykov, Fatih Porikli, and et al. Image Segmentation Using Deep Learning: A Survey. *arXiv e-prints*, art. arXiv:2001.05566, January 2020.

R. F. Minchin, M. J. Disney, Q. A. Parker, P. J. Boyce, W. J. G. de Blok, G. D. Banks, R. D. Ekers, K. C. Freeman, D. A. Garcia, B. K. Gibson, M. Grossi, R. F. Haynes, P. M. Knezek, R. H. Lang, D. F. Malin, R. M. Price, M. Putman, I. M. Stewart, and A. E. Wright. The cosmological significance of low surface brightness galaxies found in a deep blind neutral hydrogen survey. , 355(4):1303–1314, Dec 2004. doi: 10.1111/j.1365-2966.2004.08409.x.

B. Moore, G. Lake, J. Stadel, and T. Quinn. The fate of Low Surface Brightness galaxies in clusters and the origin of the diffuse intra-cluster light. In J. I. Davies, C. Impey, and S. Phillips, editors, *The Low Surface Brightness Universe*, volume 170 of *Astronomical Society of the Pacific Conference Series*, page 229, Jan 1999.

E. Morganson, R. A. Gruendl, F. Menanteau, M. Carrasco Kind, Y. C. Chen, G. Daues, A. Drlica-Wagner, D. N. Friedel, M. Gower, M. W. G. Johnson, M. D. Johnson, R. Kessler, F. Paz-Chinchón, D. Petravick, C. Pond, B. Yanny, S. Allam, R. Armstrong, W. Barkhouse, K. Bechtol, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, E. Buckley-Geer, R. Covarrubias, S. Desai, H. T. Diehl, D. A. Goldstein, D. Gruen, T. S. Li, H. Lin, J. Marriner, J. J. Mohr, E. Neilsen, C. C. Ngeow, K. Paech, E. S. Rykoff, M. Sako, I. Sevilla-Noarbe, E. Sheldon, F. Sobreira, D. L. Tucker, W. Wester, and DES Collaboration. The Dark Energy Survey Image Processing Pipeline. , 130(989):074501, July 2018. doi: 10.1088/1538-3873/aab4ef.

Benjamin P. Moster, Thorsten Naab, and Simon D. M. White. Galactic star formation and accretion histories from matching galaxies to dark matter haloes. , 428(4):3121–3138, Feb 2013. doi: 10.1093/mnras/sts261.

Roberto P. Muñoz, Paul Eigenthaler, Thomas H. Puzia, Matthew A. Taylor, Yasna Ordenes-Briceño, Karla Alamo-Martínez, Karen X. Ribbeck, Simón Ángel, Massimo Capaccioli, Patrick Côté, Laura Ferrarese, Gaspar Galaz, Maren Hempel, Michael Hilker, Andrés Jordán, Ariane Lançon, Steffen Mieske, Maurizio Paolillo, Tom Richtler, Ruben Sánchez-Janssen, and Hongxin Zhang. Unveiling a Rich System of Faint Dwarf Galaxies in the Next Generation Fornax Survey. , 813(1):L15, Nov 2015. doi: 10.1088/2041-8205/813/1/L15.

Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A Survey on Instance Segmentation: State of the art. *arXiv e-prints*, art. arXiv:2007.00047, June 2020.

Oliver Müller and Eva Schnider. Dwarfs from the Dark (Energy Survey): a machine learning approach to classify dwarf galaxies from multi-band images. *The Open Journal of Astrophysics*, 4(1):3, March 2021. doi: 10.21105/astro.2102.12776.

Oliver Müller, Helmut Jerjen, and Bruno Binggeli. New low surface brightness dwarf galaxies in the Centaurus group. , 597:A7, January 2017. doi: 10.1051/0004-6361/201628921.

Jr. Neilsen, Eric H., Gary Bernstein, Robert Gruendl, and Stephen Kent. Limiting Magnitude, τ, t_{eff}, and Image Quality in DESY Year 1. 3 2016. doi: 10.2172/1250877.

Jr. Neilsen, Eric H., James T. Annis, H. Thomas Diehl, Molly E. C. Swanson, Chris D'Andrea, Stephen Kent, and Alex Drlica-Wagner. Dark Energy Survey's Observation Strategy, Tactics, and Exposure Scheduler. *arXiv e-prints*, art. arXiv:1912.06254, December 2019.

Peder Norberg, Carlton M. Baugh, Ed Hawkins, Steve Maddox, Darren Madgwick, Ofer Lahav, Shaun Cole, Carlos S. Frenk, Ivan Baldry, Joss Bland -Hawthorn, Terry Bridges, Russell Cannon, Matthew Colless, Chris Collins, Warrick Couch, Gavin Dalton, Roberto De Propris, Simon P. Driver, George Efstathiou, Richard S. Ellis, Karl Glazebrook, Carole Jackson, Ian Lewis, Stuart Lumsden, John A. Peacock, Bruce A. Peterson, Will Sutherland, and Keith Taylor. The 2dF Galaxy Redshift Survey: the dependence of galaxy clustering on luminosity and spectral type. , 332(4):827–838, June 2002. doi: 10.1046/j.1365-8711.2002.05348.x.

K. O'Neil, G. D. Bothun, and J. Schombert. Red, Gas-Rich Low Surface Brightness Galaxies and Enigmatic Deviations from the Tully-Fisher Relation. , 119(1):136–152, Jan 2000. doi: 10.1086/301160.

Karen O'Neil, G. D. Bothun, and Mark E. Cornell. A Wide Field CCD Survey for Low Surface Brightness Galaxies:I.Data Acquisition, Description, and Initial Results. , 113: 1212, Apr 1997. doi: 10.1086/118338.

Yasna Ordenes-Briceño, Paul Eigenthaler, Matthew A. Taylor, Thomas H. Puzia, Karla Alamo-Martínez, Karen X. Ribbeck, Roberto P. Muñoz, Hongxin Zhang, Eva K. Grebel, Simón Ángel, and et al. The next generation fornax survey (ngfs). iii. revealing the spatial substructure of the dwarf galaxy population inside half of fornax's virial radius. *The Astrophysical Journal*, 859(1):52, May 2018. ISSN 1538-4357. doi: 10.3847/1538-4357/aaba70. URL `http://dx.doi.org/10.3847/1538-4357/aaba70`.

M. J. L. Orr. Regularization in the selection of radial basis function centers. *Neural Computation*, 7(3):606–623, 1995. doi: 10.1162/neco.1995.7.3.606.

M. Paillassa, E. Bertin, and H. Bouy. MAXIMASK and MAXITRACK: Two new tools for identifying contaminants in astronomical images using convolutional neural networks. , 634:A48, February 2020. doi: 10.1051/0004-6361/201936345.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

E. Papastergis, R. Giovanelli, M. P. Haynes, and F. Shankar. Is there a "too big to fail" problem in the field? , 574:A113, Feb 2015. doi: 10.1051/0004-6361/201424909.

E. Papastergis, E. A. K. Adams, and A. J. Romanowsky. The HI content of isolated ultra-diffuse galaxies: A sign of multiple formation mechanisms? , 601:L10, May 2017. doi: 10.1051/0004-6361/201730795.

D. Paranjpye, A. Mahabal, A. N. Ramaprakash, and et al. Eliminating artefacts in polarimetric images using deep learning. , 491(4):5151–5157, February 2020. doi: 10.1093/mnras/stz3250.

Ji Won Park, Ashley Villar, Yin Li, Yan-Fei Jiang, Shirley Ho, Joshua Yao-Yu Lin, Philip J. Marshall, and Aaron Roodman. Inferring Black Hole Properties from Astronomical Multivariate Time Series with Bayesian Attentive Neural Processes. *arXiv e-prints*, art. arXiv:2106.01450, June 2021.

F. Pedregosa, G. Varoquaux, A. Gramfort, and et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011a.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011b. URL `http://jmlr.org/papers/v12/pedregosa11a.html`.

P. J. E. Peebles. *The large-scale structure of the universe.* Princeton University Press, 1980.

Chien Y. Peng, Luis C. Ho, Chris D. Impey, and Hans-Walter Rix. Detailed Structural Decomposition of Galaxy Images. , 124(1):266–293, Jul 2002a. doi: 10.1086/340952.

Chien Y. Peng, Luis C. Ho, Chris D. Impey, and Hans-Walter Rix. Detailed Structural Decomposition of Galaxy Images. , 124(1):266–293, July 2002b. doi: 10.1086/340952.

Dezső Ribli, Bálint Ármin Pataki, José Manuel Zorrilla Matilla, and et al. Weak lensing cosmology with convolutional neural networks on noisy data. , 490(2):1843–1860, December 2019. doi: 10.1093/mnras/stz2610.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep Learning is Robust to Massive Label Noise. *arXiv e-prints*, art. arXiv:1705.10694, May 2017.

Javier Román and Ignacio Trujillo. Spatial distribution of ultra-diffuse galaxies within large-scale structures. , 468(1):703–716, Jun 2017a. doi: 10.1093/mnras/stx438.

Javier Román and Ignacio Trujillo. Ultra-diffuse galaxies outside clusters: clues to their formation and evolution. , 468(4):4039–4047, Jul 2017b. doi: 10.1093/mnras/stx694.

Javier Román, Ignacio Trujillo, and Mireia Montes. Galactic cirri in deep optical imaging. , 644:A42, December 2020. doi: 10.1051/0004-6361/201936111.

S. D. Rosenbaum, E. Krusch, D. J. Bomans, and R. J. Dettmar. The large-scale environment of low surface brightness galaxies. , 504(3):807–820, Sep 2009. doi: 10.1051/0004-6361/20077462.

S. Sabatini, J. Davies, W. van Driel, M. Baes, S. Roberts, R. Smith, S. Linder, and K. O'Neil. The dwarf low surface brightness galaxy population of the Virgo Cluster - II. Colours and HI line observations. , 357(3):819–833, Mar 2005a. doi: 10.1111/j.1365-2966.2005.08608.x.

S. Sabatini, J. Davies, W. van Driel, and et al. The dwarf low surface brightness galaxy population of the Virgo Cluster - II. Colours and HI line observations. , 357(3):819–833, Mar 2005b. doi: 10.1111/j.1365-2966.2005.08608.x.

Laura V. Sales, Julio F. Navarro, Louis Penafiel, Eric W. Peng, Sungsoon Lim, and Lars Hernquist. The Formation of Ultra-Diffuse Galaxies in Clusters. *arXiv e-prints*, art. arXiv:1909.01347, September 2019.

C. Sánchez, M. Carrasco Kind, H. Lin, R. Miquel, F. B. Abdalla, A. Amara, M. Banerji, C. Bonnett, R. Brunner, D. Capozzi, A. Carnero, F. J. Castander, L. A. N. da Costa, C. Cunha, A. Fausti, D. Gerdes, N. Greisel, J. Gschwend, W. Hartley, S. Jouvel, O. Lahav, M. Lima, M. A. G. Maia, P. Martí, R. L. C. Ogando, F. Ostrovski, P. Pellegrini, M. M. Rau, I. Sadeh, S. Seitz, I. Sevilla-Noarbe, A. Sypniewski, J. de Vicente, T. Abbot, S. S. Allam, D. Atlee, G. Bernstein, J. P. Bernstein, E. Buckley-Geer, D. Burke, M. J. Childress, T. Davis, D. L. DePoy, A. Dey, S. Desai, H. T. Diehl, P. Doel, J. Estrada, A. Evrard, E. Fernández, D. Finley, B. Flaugher, J. Frieman, E. Gaztanaga, K. Glazebrook, K. Honscheid, A. Kim, K. Kuehn, N. Kuropatkin, C. Lidman, M. Makler, J. L. Marshall, R. C. Nichol, A. Roodman, E. Sánchez, B. X. Santiago, M. Sako, R. Scalzo, R. C. Smith, M. E. C. Swanson, G. Tarle, D. Thomas, D. L. Tucker, S. A. Uddin, F. Valdés, A. Walker, F. Yuan, and J. Zuntz. Photometric redshift analysis in the Dark Energy Survey Science Verification data. , 445(2):1482–1506, December 2014. doi: 10.1093/mnras/stu1836.

A. Sandage and B. Binggeli. Studies of the Virgo cluster. III. A classification system and an illustrated Atlas of Virgo cluster dwarf galaxies. , 89:919–931, July 1984. doi: 10.1086/113588.

Edward F. Schlafly and Douglas P. Finkbeiner. Measuring Reddening with Sloan Digital Sky Survey Stellar Spectra and Recalibrating SFD. , 737(2):103, August 2011. doi: 10.1088/0004-637X/737/2/103.

David J. Schlegel, Douglas P. Finkbeiner, and Marc Davis. Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. , 500(2):525–553, June 1998. doi: 10.1086/305772.

J. M. Schwartzenberg, S. Phillipps, R. M. Smith, W. J. Couch, and B. J. Boyle. A deep field survey for low surface brightness galaxies. , 275(1):121–128, July 1995. doi: 10.1093/mnras/275.1.121.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, and et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv e-prints*, art. arXiv:1610.02391, October 2016.

J. L. Sérsic. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina*, 6:41–43, February 1963.

I. Sevilla-Noarbe, K. Bechtol, M. Carrasco Kind, and et al. Dark Energy Survey Year 3 Results: Photometric Data Set for Cosmology. *ApJS*, 254(2):24, June 2021. doi: 10.3847/1538-4365/abeb66.

Ignacio Sevilla-Noarbe, Keith Bechtol, et al. Dark energy survey year 3 results: Photometric data set for cosmology. *in prep*, 2020.

Shiyin Shen, H. J. Mo, Simon D. M. White, Michael R. Blanton, Guinevere Kauffmann, Wolfgang Voges, J. Brinkmann, and Istvan Csabai. The size distribution of galaxies in the Sloan Digital Sky Survey. , 343(3):978–994, August 2003. doi: 10.1046/j.1365-8711.2003.06740.x.

N. Shipp, A. Drlica-Wagner, E. Balbinot, P. Ferguson, D. Erkal, T. S. Li, K. Bechtol, V. Belokurov, B. Buncher, D. Carollo, M. Carrasco Kind, K. Kuehn, J. L. Marshall, A. B. Pace, E. S. Rykoff, I. Sevilla-Noarbe, E. Sheldon, L. Strigari, A. K. Vivas, B. Yanny, A. Zenteno, T. M. C. Abbott, F. B. Abdalla, S. Allam, S. Avila, E. Bertin, D. Brooks, D. L. Burke, J. Carretero, F. J. Castander, R. Cawthon, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, C. Davis, J. De Vicente, S. Desai, H. T. Diehl, P. Doel, A. E. Evrard, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. Hartley, K. Honscheid, B. Hoyle, D. J. James, M. D. Johnson, E. Krause, N. Kuropatkin, O. Lahav, H. Lin, M. A. G. Maia, M. March, P. Martini, F. Menanteau, C. J. Miller, R. Miquel, R. C. Nichol, A. A. Plazas, A. K. Romer, M. Sako, E. Sanchez, B. Santiago, V. Scarpine, R. Schindler, M. Schubnell, M. Smith, R. C. Smith, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, D. L. Tucker, A. R. Walker, R. H. Wechsler, and DES Collaboration. Stellar Streams Discovered in the Dark Energy Survey. , 862(2):114, August 2018. doi: 10.3847/1538-4357/aacdab.

Connor Shorten and T. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48, 2019.

Cristóbal Sifón, Remco F. J. van der Burg, Henk Hoekstra, Adam Muzzin, and Ricardo Herbonnet. A first constraint on the average mass of ultra-diffuse galaxies from weak gravitational lensing. , 473(3):3747–3754, January 2018. doi: 10.1093/mnras/stx2648.

Joshua D. Simon. The Faintest Dwarf Galaxies. , 57:375–415, August 2019. doi: 10.1146/annurev-astro-091918-104453.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*, art. arXiv:1409.1556, September 2014.

Colin T. Slater, Paul Harding, and J. Christopher Mihos. Removing Internal Reflections from Deep Imaging Data Sets. , 121(885):1267, November 2009. doi: 10.1086/648457.

A. V. Smith Castelli, F. R. Faifer, and C. G. Escudero. Stellar systems in the direction of the Hickson Compact Group 44. I. Low surface brightness galaxies. , 596:A23, November 2016. doi: 10.1051/0004-6361/201628969.

Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Learning from Noisy Labels with Deep Neural Networks: A Survey. *arXiv e-prints*, art. arXiv:2007.08199, July 2020.

Jeeseon Song, Joseph J. Mohr, Wayne A. Barkhouse, and et al. A Parameterized Galaxy Catalog Simulator for Testing Cluster Finding, Mass Estimation, and Photometric Redshift Estimation in Optical and Near-infrared Surveys. , 747(1):58, March 2012. doi: 10.1088/0004-637X/747/1/58.

D. Sprayberry, C. D. Impey, M. J. Irwin, R. G. McMahon, and G. D. Bothun. Discovery of a Third Giant Low Surface Brightness Disk Galaxy. , 417:114, November 1993. doi: 10.1086/173296.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

Iskra Strateva, Željko Ivezić, Gillian R. Knapp, Vijay K. Narayanan, Michael A. Strauss, James E. Gunn, Robert H. Lupton, David Schlegel, Neta A. Bahcall, Jon Brinkmann, Robert J. Brunner, Tamás Budavári, István Csabai, Francisco Javier Castander, Mamoru Doi, Masataka Fukugita, Zsuzsanna Győry, Masaru Hamabe, Greg Hennessy, Takashi Ichikawa, Peter Z. Kunszt, Don Q. Lamb, Timothy A. McKay, Sadanori Okamura, Judith Racusin, Maki Sekiguchi, Donald P. Schneider, Kazuhiro Shimasaku, and Donald York. Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data. , 122 (4):1861–1874, Oct 2001. doi: 10.1086/323301.

E. Suchyta, E. M. Huff, J. Aleksić, P. Melchior, S. Jouvel, N. MacCrann, A. J. Ross, M. Crocce, E. Gaztanaga, K. Honscheid, B. Leistedt, H. V. Peiris, E. S. Rykoff, E. Sheldon, T. Abbott, F. B. Abdalla, S. Allam, M. Banerji, A. Benoit-Lévy, E. Bertin, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, D. L. DePoy, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, T. F. Eifler, J. Estrada, A. E. Evrard, B. Flaugher, P. Fosalba, J. Frieman, D. W. Gerdes, D. Gruen, R. A. Gruendl, D. J. James, M. Jarvis, K. Kuehn, N. Kuropatkin, O. Lahav, M. Lima, M. A. G. Maia, M. March, J. L. Marshall, C. J. Miller, R. Miquel, E. Neilsen, R. C. Nichol, B. Nord, R. Ogando, W. J. Percival, K. Reil, A. Roodman, M. Sako, E. Sanchez, V. Scarpine, I. Sevilla-Noarbe, R. C. Smith, M. Soares-Santos, F. Sobreira, M. E. C. Swanson, G. Tarle, J. Thaler, D. Thomas, V. Vikram, A. R. Walker, R. H. Wechsler, Y. Zhang, and DES Collaboration. No galaxy left behind: accurate measurements with the faintest objects in the Dark Energy Survey. , 457(1):786–808, March 2016. doi: 10.1093/mnras/stv2953.

J. W. Sulentic and W. G. Tifft. VizieR Online Data Catalog: Revised New General Catalogue (Sulentic+, 1973). *VizieR Online Data Catalog*, 7001, April 1999.

M. E. C. Swanson, Max Tegmark, Andrew J. S. Hamilton, and J. Colin Hill. Methods for rapidly processing angular masks of next-generation galaxy surveys. , 387(4):1391–1402, July 2008. doi: 10.1111/j.1365-2966.2008.13296.x.

H. Tang, A. M. M. Scaife, and J. P. Leahy. Transfer learning for radio galaxy classification. , 488(3):3358–3375, September 2019. doi: 10.1093/mnras/stz1883.

D. Tanoglidis, A. Ćiprijanović, and A. Drlica-Wagner. DeepShadows: Separating low surface brightness galaxies from artifacts using deep learning. *Astronomy and Computing*, 35: 100469, April 2021a. doi: 10.1016/j.ascom.2021.100469.

D. Tanoglidis, A. Drlica-Wagner, K. Wei, et. al,, and (DES Collaboration). Shadows in the Dark: Low-surface-brightness Galaxies Discovered in the Dark Energy Survey. *ApJS*, 252 (2):18, February 2021b. doi: 10.3847/1538-4365/abca89.

D. Tanoglidis, A. Ćiprijanović, A. Drlica-Wagner, B. Nord, M. H. L. S. Wang, A. Jacob Amsellem, K. Downey, S. Jenkins, D. Kafkes, and Z. Zhang. DeepGhostBusters: Using Mask R-CNN to detect and mask ghosting and scattered-light artifacts from optical survey images. *Astronomy and Computing*, 39:100580, April 2022. doi: 10.1016/j.ascom.2022.100580.

Dimitrios Tanoglidis, Chihway Chang, and Joshua Frieman. Optimizing galaxy samples for clustering measurements in photometric surveys. , 491(3):3535–3552, January 2020. doi: 10.1093/mnras/stz3281.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995. doi: 10.1109/ICDAR.1995.598994.

Michael Tremmel, Anna C. Wright, Alyson M. Brooks, Ferah Munshi, Daisuke Nagai, and Thomas R. Quinn. The Formation of Ultra-Diffuse Galaxies from Passive Evolution in the RomulusC Galaxy Cluster Simulation. *arXiv e-prints*, art. arXiv:1908.05684, August 2019.

D. Tuccillo, M. Huertas-Company, E. Decencière, S. Velasco-Forero, H. Domínguez Sánchez, and P. Dimauro. Deep learning for galaxy surface brightness profile fitting. , 475(1): 894–909, March 2018. doi: 10.1093/mnras/stx3186.

R. Brent Tully. Galaxy Groups: A 2MASS Catalog. , 149(5):171, May 2015. doi: 10.1088/0004-6256/149/5/171.

J. A. Turner, S. Phillipps, J. I. Davies, and M. J. Disney. A deep CCD search for low surface brightness galaxies in A 3574. , 261:39–51, March 1993. doi: 10.1093/mnras/261.1.39.

Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on Bayesian Neural Networks – a Tutorial for Deep Learning Users. *arXiv e-prints*, art. arXiv:2007.06823, July 2020.

Remco F. J. van der Burg, Adam Muzzin, and Henk Hoekstra. The abundance and spatial distribution of ultra-diffuse galaxies in nearby galaxy clusters. , 590:A20, May 2016. doi: 10.1051/0004-6361/201628222.

S. Van Der Walt, S. C. Colbert, and G. Varoquaux. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13:22–30, February 2011. doi: 10.1109/MCSE.2011.37.

Pieter van Dokkum, Roberto Abraham, Jean Brodie, Charlie Conroy, Shany Danieli, Allison Merritt, Lamiya Mowla, Aaron Romanowsky, and Jielai Zhang. A High Stellar Velocity Dispersion and ∼100 Globular Clusters for the Ultra-diffuse Galaxy Dragonfly 44. , 828 (1):L6, September 2016. doi: 10.3847/2041-8205/828/1/L6.

Pieter van Dokkum, Shany Danieli, Yotam Cohen, Allison Merritt, Aaron J. Romanowsky, Roberto Abraham, Jean Brodie, Charlie Conroy, Deborah Lokhorst, Lamiya Mowla, Ewan O'Sullivan, and Jielai Zhang. A galaxy lacking dark matter. , 555(7698):629–632, Mar 2018. doi: 10.1038/nature25767.

Pieter van Dokkum, Shany Danieli, Roberto Abraham, Charlie Conroy, and Aaron J. Romanowsky. A Second Galaxy Missing Dark Matter in the NGC 1052 Group. , 874(1):L5, Mar 2019a. doi: 10.3847/2041-8213/ab0d92.

Pieter van Dokkum, Asher Wasserman, Shany Danieli, Roberto Abraham, Jean Brodie, Charlie Conroy, Duncan A. Forbes, Christopher Martin, Matt Matuszewski, Aaron J. Romanowsky, and Alexa Villaume. Spatially Resolved Stellar Kinematics of the Ultra-diffuse Galaxy Dragonfly 44. I. Observations, Kinematics, and Cold Dark Matter Halo Fits. , 880(2):91, August 2019b. doi: 10.3847/1538-4357/ab2914.

Pieter van Dokkum, Deborah Lokhorst, Shany Danieli, Jiaxuan Li, Allison Merritt, Roberto Abraham, Colleen Gilhuly, Johnny P. Greco, and Qing Liu. Multi-resolution Filtering: An Empirical Method for Isolating Faint, Extended Emission in Dragonfly Data and Other Low Resolution Images. , 132(1013):074503, July 2020. doi: 10.1088/1538-3873/ab9416.

Pieter G. van Dokkum, Roberto Abraham, Allison Merritt, and et al. Forty-seven Milky Way-sized, Extremely Diffuse Galaxies in the Coma Cluster. , 798(2):L45, Jan 2015a. doi: 10.1088/2041-8205/798/2/L45.

Pieter G. van Dokkum, Roberto Abraham, Allison Merritt, Jielai Zhang, Marla Geha, and Charlie Conroy. Forty-seven Milky Way-sized, Extremely Diffuse Galaxies in the Coma Cluster. , 798(2):L45, Jan 2015b. doi: 10.1088/2041-8205/798/2/L45.

Pieter G. van Dokkum, Aaron J. Romanowsky, Roberto Abraham, Jean P. Brodie, Charlie Conroy, Marla Geha, Allison Merritt, Alexa Villaume, and Jielai Zhang. Spectroscopic

Confirmation of the Existence of Large, Diffuse Galaxies in the Coma Cluster. , 804(1): L26, May 2015c. doi: 10.1088/2041-8205/804/1/L26.

Aku Venhola, Reynier Peletier, Eija Laurikainen, Heikki Salo, Thorsten Lisker, Enrichetta Iodice, Massimo Capaccioli, Gijs Verdois Kleijn, Edwin Valentijn, Steffen Mieske, Michael Hilker, Carolin Wittmann, Glenn van de Ven, Aniello Grado, Marilena Spavone, Michele Cantiello, Nicola Napolitano, Maurizio Paolillo, and Jesús Falcón-Barroso. The Fornax Deep Survey with VST. III. Low surface brightness dwarfs and ultra diffuse galaxies in the center of the Fornax cluster. , 608:A142, Dec 2017. doi: 10.1051/0004-6361/201730696.

Aku Venhola, Reynier Peletier, Eija Laurikainen, Heikki Salo, Enrichetta Iodice, Steffen Mieske, Michael Hilker, Carolin Wittmann, Thorsten Lisker, Maurizio Paolillo, Michele Cantiello, Joachim Janz, Marilena Spavone, Raffaele D'Abrusco, Glennvande Ven, Nicola Napolitano, GijsVerdoes Kleijn, Natasha Maddox, Massimo Capaccioli, Aniello Grado, Edwin Valentijn, Jesús Falcón-Barroso, and Luca Limatola. The Fornax Deep Survey with the VST. IV. A size and magnitude limited catalog of dwarf galaxies in the area of the Fornax cluster. , 620:A165, December 2018. doi: 10.1051/0004-6361/201833933.

Ricardo Vilalta. Transfer Learning in Astronomy: A New Machine-Learning Paradigm. In *Journal of Physics Conference Series*, volume 1085 of *Journal of Physics Conference Series*, page 052014, September 2018. doi: 10.1088/1742-6596/1085/5/052014.

Ricardo Vilalta, Kinjal Dhar Gupta, Dainis Boumber, and Mikhail M. Meskhi. A General Approach to Domain Adaptation with Applications in Astronomy. , 131(1004):108008, October 2019. doi: 10.1088/1538-3873/aaf1fc.

Sebastian Wagner-Carena, Ji Won Park, Simon Birrer, Philip J. Marshall, Aaron Roodman, Risa H. Wechsler, and LSST Dark Energy Science Collaboration. Hierarchical Inference with Bayesian Neural Networks: An Application to Strong Gravitational Lensing. , 909 (2):187, March 2021. doi: 10.3847/1538-4357/abdf59.

Mei Wang and Weihong Deng. Deep Visual Domain Adaptation: A Survey. *arXiv e-prints*, art. arXiv:1802.03601, February 2018.

Y. Wang, R. J. Brunner, and J. C. Dolence. The SDSS galaxy angular two-point correlation function. , 432(3):1961–1979, July 2013. doi: 10.1093/mnras/stt450.

Yu Wang, Xiaohu Yang, H. J. Mo, Frank C. van den Bosch, Neal Katz, Anna Pasquali, Daniel H. McIntosh, and Simone M. Weinmann. The Nature of Red Dwarf Galaxies. , 697(1):247–257, May 2009. doi: 10.1088/0004-637X/697/1/247.

Risa H. Wechsler and Jeremy L. Tinker. The Connection Between Galaxies and Their Dark Matter Halos. , 56:435–487, September 2018. doi: 10.1146/annurev-astro-081817-051756.

Wei Wei, E. A. Huerta, Bradley C. Whitmore, and et al. Deep transfer learning for star cluster classification: I. application to the PHANGS-HST survey. , 493(3):3178–3193, February 2020. doi: 10.1093/mnras/staa325.

Karl R. Weiss, T. Khoshgoftaar, and Dingding Wang. A survey of transfer learning. *Journal of Big Data*, 3:1–40, 2016.

Lilian Weng. Object detection for dummies part 3: R-cnn family. *lilianweng.github.io/lil-log*, 2017. URL `http://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html`.

Simon D. M. White and Carlos S. Frenk. Galaxy Formation through Hierarchical Clustering. , 379:52, Sep 1991. doi: 10.1086/170483.

Carolin Wittmann, Thorsten Lisker, Liyualem Ambachew Tilahun, Eva K. Grebel, Christopher J. Conselice, Samantha Penny, Joachim Janz, John S. Gallagher, Ralf Kotulla, and James McCormac. A population of faint low surface brightness galaxies in the Perseus cluster core. , 470(2):1512–1525, Sep 2017. doi: 10.1093/mnras/stx1229.

Donald G. York and SDSS Collaboration et al. The Sloan Digital Sky Survey: Technical Summary. , 120(3):1579–1587, September 2000. doi: 10.1086/301513.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv e-prints*, art. arXiv:1411.1792, November 2014.

Dennis Zaritsky, Richard Donnerstein, Ananthan Karunakaran, C. E. Barbosa, Arjun Dey, Jennifer Kadowaki, Kristine Spekkens, and Huanian Zhang. Systematically Measuring Ultra-Diffuse Galaxies (SMUDGes). III. The Southern SMUDGes Catalog. *arXiv e-prints*, art. arXiv:2205.02193, May 2022.

Idit Zehavi, Michael R. Blanton, Joshua A. Frieman, David H. Weinberg, Houjun J. Mo, Michael A. Strauss, Scott F. Anderson, James Annis, Neta A. Bahcall, Mariangela Bernardi, John W. Briggs, Jon Brinkmann, Scott Burles, Larry Carey, Francisco J. Castander, Andrew J. Connolly, Istvan Csabai, Julianne J. Dalcanton, Scott Dodelson, Mamoru Doi, Daniel Eisenstein, Michael L. Evans, Douglas P. Finkbeiner, Scott Friedman, Masataka Fukugita, James E. Gunn, Greg S. Hennessy, Robert B. Hindsley, Željko Ivezić, Stephen Kent, Gillian R. Knapp, Richard Kron, Peter Kunszt, Donald Q. Lamb, R. French Leger, Daniel C. Long, Jon Loveday, Robert H. Lupton, Timothy McKay, Avery Meiksin, Aronne Merrelli, Jeffrey A. Munn, Vijay Narayanan, Matt Newcomb, Robert C. Nichol, Russell Owen, John Peoples, Adrian Pope, Constance M. Rockosi, David Schlegel, Donald P. Schneider, Roman Scoccimarro, Ravi K. Sheth, Walter Siegmund, Stephen Smee, Yehuda Snir, Albert Stebbins, Christopher Stoughton, Mark SubbaRao, Alexander S. Szalay, Istvan Szapudi, Max Tegmark, Douglas L. Tucker, Alan Uomoto, Dan Vanden Berk, Michael S. Vogeley, Patrick Waddell, Brian Yanny, and Donald G. York. Galaxy Clustering in Early Sloan Digital Sky Survey Redshift Data. , 571(1):172–190, May 2002. doi: 10.1086/339893.

Idit Zehavi, Zheng Zheng, David H. Weinberg, Joshua A. Frieman, Andreas A. Berlind, Michael R. Blanton, Roman Scoccimarro, Ravi K. Sheth, Michael A. Strauss, Issha Kayo, Yasushi Suto, Masataka Fukugita, Osamu Nakamura, Neta A. Bahcall, Jon Brinkmann, James E. Gunn, Greg S. Hennessy, Željko Ivezić, Gillian R. Knapp, Jon Loveday, Avery Meiksin, David J. Schlegel, Donald P. Schneider, Istvan Szapudi, Max Tegmark, Michael S. Vogeley, Donald G. York, and SDSS Collaboration. The Luminosity and Color Dependence of the Galaxy Correlation Function. , 630(1):1–27, September 2005. doi: 10.1086/431891.

Idit Zehavi, Zheng Zheng, David H. Weinberg, Michael R. Blanton, Neta A. Bahcall, Andreas A. Berlind, Jon Brinkmann, Joshua A. Frieman, James E. Gunn, Robert H. Lupton, Robert C. Nichol, Will J. Percival, Donald P. Schneider, Ramin A. Skibba, Michael A. Strauss, Max Tegmark, and Donald G. York. Galaxy Clustering in the Completed SDSS Redshift Survey: The Dependence on Color and Luminosity. , 736(1):59, July 2011. doi: 10.1088/0004-637X/736/1/59.

Matthew D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. *arXiv e-prints*, art. arXiv:1212.5701, December 2012.

Hu Zhan and J. Anthony Tyson. Cosmology with the Large Synoptic Survey Telescope: an overview. *Reports on Progress in Physics*, 81(6):066901, June 2018. doi: 10.1088/1361-6633/aab1bd.

Keming Zhang and Joshua S. Bloom. deepCR: Cosmic Ray Rejection with Deep Learning. , 889(1):24, January 2020. doi: 10.3847/1538-4357/ab3fa6.

Y. Zhang, B. Yanny, A. Palmese, D. Gruen, C. To, E. S. Rykoff, Y. Leung, C. Collins, M. Hilton, T. M. C. Abbott, J. Annis, S. Avila, E. Bertin, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, J. De Vicente, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, A. Drlica-Wagner, T. F. Eifler, A. E. Evrard, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, D. L. Hollowood, K. Honscheid, B. Hoyle, D. J. James, T. Jeltema, K. Kuehn, N. Kuropatkin, T. S. Li, M. Lima, M. A. G. Maia, M. March, J. L. Marshall, P. Melchior, F. Menanteau, C. J. Miller, R. Miquel, J. J. Mohr, R. L. C. Ogando, A. A. Plazas, A. K. Romer, E. Sanchez, V. Scarpine, M. Schubnell, S. Serrano, I. Sevilla-Noarbe, M. Smith, M. Soares-Santos, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, W. Wester, and DES Collaboration. Dark Energy Survey Year 1 Results: Detection of Intracluster Light at Redshift ~ 0.25. , 874(2):165, April 2019. doi: 10.3847/1538-4357/ab0dfd.

Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object Detection with Deep Learning: A Review. *arXiv e-prints*, art. arXiv:1807.05511, July 2018.

G. H. Zhong, Y. C. Liang, F. S. Liu, F. Hammer, J. Y. Hu, X. Y. Chen, L. C. Deng, and B. Zhang. A large sample of low surface brightness disc galaxies from the SDSS - I. The sample and the stellar populations. , 391(2):986–999, Dec 2008. doi: 10.1111/j.1365-2966.2008.13972.x.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, and et al. A Comprehensive Survey on Transfer Learning. *arXiv e-prints*, art. arXiv:1911.02685, November 2019.

F. Zwicky. *Morphological Astronomy.* Springer Verlag, 1957.