Research in Astronomy and Astrophysics, 24:095012 (11pp), 2024 September https://doi.org/10.1088/1674-4527/ad6fe6

Preparation for CSST: Star-galaxy Classification using a Rotationally Invariant Supervised Machine Learning Method

Shiliang Zhang[1], Guanwen Fang[1], Jie Song[2,3], Ran Li[4], Yizhou Gu[5], Zesen Lin[6], Chichun Zhou[7], Yao Dai[1], and Xu Kong[2,3]
[1] Institute of Astronomy and Astrophysics, Anqing Normal University, Anqing 246133, China; wen@mail.ustc.edu.cn
[2] Deep Space Exploration Laboratory/Department of Astronomy, University of Science and Technology of China, Hefei 230026, China; xkong@ustc.edu.cn
[3] School of Astronomy and Space Science, University of Science and Technology of China, Hefei 230026, China
[4] National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
[5] Tsung-Dao Lee Institute and Key Laboratory for Particle Physics, Astrophysics and Cosmology, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China
[6] Department of Physics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, S.A.R., China
[7] School of Engineering, Dali University, Dali 671003, China
Received 2024 June 24; revised 2024 August 4; accepted 2024 August 14; published 2024 September 18

Abstract

Most existing star-galaxy classifiers depend on the reduced information from catalogs, necessitating careful data processing and feature extraction. In this study, we employ a supervised machine learning method (GoogLeNet) to automatically classify stars and galaxies in the COSMOS field. Unlike traditional machine learning methods, we introduce several preprocessing techniques, including noise reduction and the unwrapping of denoised images in polar coordinates, applied to our carefully selected samples of stars and galaxies. By dividing the selected samples into training and validation sets in an 8:2 ratio, we evaluate the performance of the GoogLeNet model in distinguishing between stars and galaxies. The results indicate that the GoogLeNet model is highly effective, achieving accuracies of 99.6% and 99.9% for stars and galaxies, respectively. Furthermore, by comparing the results with and without preprocessing, we find that preprocessing can significantly improve classification accuracy (by approximately 2.0% to 6.0%) when the images are rotated. In preparation for the future launch of the China Space Station Telescope (CSST), we also evaluate the performance of the GoogLeNet model on the CSST simulation data. These results demonstrate a high level of accuracy (approximately 99.8%), indicating that this model can be effectively utilized for future observations with the CSST.

Key words: methods: data analysis – techniques: image processing – stars: imaging

1. Introduction

In astronomy, precise differentiation between stars and galaxies is paramount due to their representation of distinct astrophysical phenomena. For instance, the systematic contribution resulting from the cross-contamination of star and galaxy samples could significantly impact the field of "precision cosmology" (e.g., Ross et al. 2011; Thomas et al. 2011; Soumagnac et al. 2015; Sevilla-Noarbe et al. 2018). This issue will become increasingly critical in future astronomical research, as the upcoming large-field sky surveys, such as those conducted by the Chinese Space Station Telescope (CSST) (Zhan 2011, 2018), the Euclid Space Telescope (Euclid Collaboration et al. 2022), and the Roman Space Telescope (Spergel et al. 2015), will yield imaging of millions to billions of stars and galaxies. This necessitates the development of methods to accurately and rapidly distinguish between stars and galaxies.

Several methods are currently available to address this issue. The first classification method is morphology-based, involving the determination of an optimal threshold in the space of observable image properties (e.g., MacGillivray et al. 1976; Kron 1980; Leauthaud et al. 2007; Skelton et al. 2014; Sevilla-Noarbe et al. 2018; López-Sanjuan et al. 2019). This method is predicated on the assumption that stars are point-like sources, while galaxies are extended sources. Consequently, stars and galaxies can be differentiated by their distribution in a size-magnitude diagram (e.g., Leauthaud et al. 2007; Skelton et al. 2014). Another approach is color-based, leveraging the distinct spectral shapes of stars and galaxies. This method allows for the differentiation between stars, galaxies, and quasars through color–color diagrams (e.g., Baldry et al. 2010; Saglia et al. 2012). Combining these two methods is also feasible to maximize the utilization of available data (e.g., Kim et al. 2015; Soumagnac et al. 2015; Kim & Brunner 2016).

However, these empirical methods rely on reduced summary information, which could be challenging to obtain. Transforming astronomical images into suitable features necessitates careful engineering and considerable domain expertise. Machine learning, particularly Convolutional Neural Networks (CNNs), offers a potent alternative by facilitating the direct extraction of image features (e.g., Dieleman et al. 2015;

Huertas-Company et al. 2015; Walmsley et al. 2018), circumventing the need for manual feature design. This approach enables a more efficient and accurate classification between stars and galaxies.

The application of machine learning techniques in classifying stars and galaxies was pioneered by Odewahn et al. (1992), and it has since become an integral component of widely used software packages such as SExtractor (Bertin & Arnouts 1996). Subsequently, various successful implementations have emerged to tackle this problem, including decision trees, support vector machines (SVMs), and classifier ensemble strategies (e.g., Weir et al. 1995; Suchkov et al. 2005; Ball et al. 2006; Vasconcellos et al. 2011; Sevilla-Noarbe & Etayo-Sotos 2015; Fadely et al. 2012; Kim & Brunner 2016). However, studies have demonstrated that traditional CNN-based machine learning algorithms often exhibit poor robustness to signal-to-noise ratios (S/N) and image rotations since noise and rotations may break the image features (e.g., Nazaré et al. 2018; Liu et al. 2020; Cheng et al. 2016; Cabrera-Vives et al. 2017; Chen et al. 2018; Cheng et al. 2019). Efforts have been made to address these challenges. For instance, noise reduction has been identified as an effective method for mitigating the impact of image S/N (e.g., Nazaré et al. 2018; Fang et al. 2023). Regarding the influence of image rotation, our previous studies have demonstrated that Adaptive Polar Coordinate Transformation (APCT) can effectively overcome the challenges posed by image rotation (e.g., Fang et al. 2023; Dai et al. 2023; Song et al. 2024). These approaches can significantly enhance the accuracy of machine learning models in recognizing image features.

In this study, we employ the GoogLeNet algorithm (Szegedy et al. 2015) to the Hubble Space Telescope (HST) *I*-band images of the Cosmic Evolution Survey (COSMOS) field to classify stars and galaxies. The GoogLeNet algorithm, previously validated for high-resolution images (e.g., Fang et al. 2023; Dai et al. 2023), is found to effectively distinguish between stars and galaxies after preprocessing the images with noise reduction and APCT. Comparing the results obtained with and without preprocessing, we observe a significant improvement in classification accuracy. Furthermore, in anticipation of the future launch of the CSST, we evaluate the robustness of our framework using the simulated CSST data. The results demonstrate an accuracy approaching 99.8%, indicating the suitability of this framework for future observations with the CSST.

The structure of this paper is as follows. In Section 2, we outline how we select our samples of stars and galaxies. In Section 3, we describe our data preprocessing method and the GoogLeNet algorithm. The classification results are presented in Section 4. In Section 5, we evaluate the performance of our framework on the CSST simulated data. Finally, conclusions are provided in Section 6.

2. Data Set and Sample Selection

2.1. COSMOS Field

The COSMOS field (Scoville et al. 2007) has significantly advanced our understanding of the Universe by providing deep data covering a wide wavelength range from radio to X-rays. In this study, we utilize the high-resolution imaging data acquired from HST with the Advanced Camera for Surveys (ACS) in the F814W band. This data set comprises approximately 590 pointings and encompasses an area of 1.64 square degrees within the COSMOS field, making it the largest contiguous survey conducted by the HST/ACS to date. Its extensive coverage ensures a sufficient number of stars, rendering it suitable for our investigation. The original images were processed by Koekemoer et al. (2007) using the MultiDrizzle package (Koekemoer et al. 2003). The final mosaic images have a pixel scale of $0\farcs03$, and the 5σ depth is 27.2 AB magnitude for point source observations within an aperture diameter of $0\farcs24$. Subsequent analyses in this study are conducted based on these high-resolution mosaic images.

2.2. COSMOS2020 Catalog

The data set utilized in this study is derived from the COSMOS2020 catalog (Weaver et al. 2022), which offers reliable photometric estimations for approximately 1.7 million objects spanning from far-ultraviolet to near-infrared wavelengths. These objects are detected from the "chi-square" $izYJHK_s$ image. The luminosity is estimated using two independent extraction methods: (1) the Classic catalog employing SExtractor, and (2) the Farmer catalog employing parametric modeling via the Tractor package (Lang et al. 2016). Additionally, photometric redshifts and other physical parameters are also estimated through spectral energy distribution (SED) fitting using two different approaches, namely LePhare (Ilbert et al. 2006) and EAZY (Brammer et al. 2008), for both catalogs. Weaver et al. (2022) provided a detailed comparison among these different estimation methods, demonstrating their high consistency. We opt for the Classic catalog with redshifts estimated using LePhare in this study.

Additionally, Weaver et al. (2022) distinguished stars from galaxies by combining morphological and SED criteria. Point-like objects form a tight sequence in the half-light radius versus magnitude space, thus, Weaver et al. (2022) classified all bright sources on this sequence as stars using morphological information obtained from HST/ACS (*I* band) and Subaru/HSC (*i* band) images. They also compared the best-fit χ^2 values obtained using stellar ($\chi^2_{\rm star}$) and galaxy ($\chi^2_{\rm gal}$) templates during SED fitting with LePhare. Objects with $\chi^2_{\rm star}$ smaller than $\chi^2_{\rm gal}$ were classified as stars. In the COSMOS2020 catalog, the parameter "$\rm lp_{type}$" indicates the final classification results between stars and galaxies, with galaxies assigned $\rm lp_{type} = 0$ and stars assigned $\rm lp_{type} = 1$.

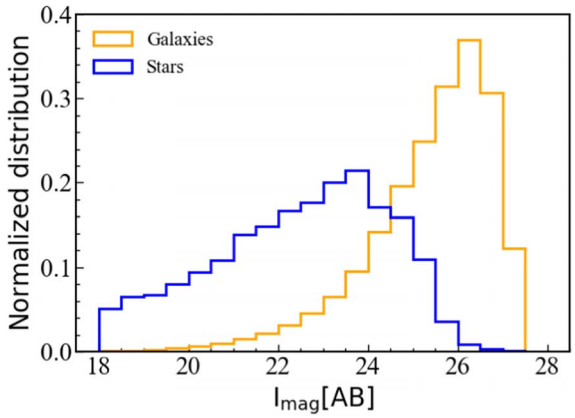

Figure 1. Distributions of I-band magnitudes for stars (blue) and galaxies (yellow), separated by the $\mathrm{lp_{type}}$ parameter, in the COSMOS field. In this work, we only consider objects brighter than 25 mag.

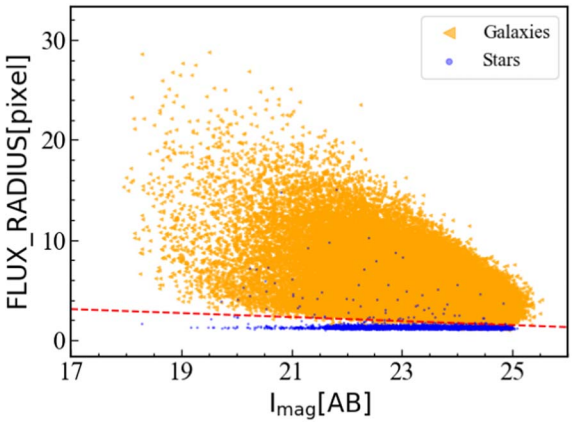

Figure 2. Distributions of FLUX_RADIUS and I_{mag} for our samples, separated by $\mathrm{lp_{type}}$. Galaxies and stars are represented by yellow triangles and blue points, respectively. It is evident from this figure that stars form a tight sequence. We separate the stars and galaxies using a custom criterion, shown as a red line in this figure. All sources above this line are removed from our star sample.

2.3. Selection of Galaxies for Analysis

In this study, we obtain the photometric information from the Classic-version catalog, along with various other physical parameters estimated using `LePhare`. Our galaxy samples are selected based on the following criteria: (1) $\mathrm{lp_{type}} = 0$, indicating the object is classified as a galaxy; (2) $I_{\mathrm{mag}} < 25$, excluding objects that are too faint; (3) $0.2 < z < 1.2$, ensuring that the galaxies are observed in the rest-frame optical wavelength; (4) $\mathrm{FLAG_{COMBINE}} = 0$, indicating that the photometric measurements are not contaminated by bright stars. Additionally, we exclude all sources with bad pixels. Finally, we randomly select 60,000 galaxies from the remaining samples as our galaxy sample. In Figure 1, we present the distribution of I-band magnitudes for stars and galaxies (as separated by the $\mathrm{lp_{type}}$ parameter) in the COSMOS field.

2.4. Selection of Stars for Analysis

To accurately evaluate the performance of the GoogLeNet model in distinguishing between stars and galaxies, it is essential to obtain clean samples of stars. This is particularly important because the number of stars in this field is relatively small compared to the number of galaxies, and contamination could significantly affect the accuracy of star classification. Therefore, in addition to restricting the selection criteria to mirror those used for galaxy identification, which include: (1) $\mathrm{lp_{type}} = 1$, (2) $I_{\mathrm{mag}} < 25$, and (3) $\mathrm{FLAG_{COMBINE}} = 0$, we also take further steps to ensure the reliability of our star sample.

Weaver et al. (2022) had already distinguished stars from galaxies using the half-light radius and magnitude for bright sources. Any bright objects ($I < 23$ mag for HST images and $i < 21.5$ mag for HSC images) falling on the point-like source sequence were classified as stars by Weaver et al. (2022). Considering that our sample includes some fainter sources, we also employed similar methods to obtain a purer sample of stars with I-band magnitudes $I_{\mathrm{mag}} < 25$.

First, Laigle et al. (2016) demonstrated that stars form a tight sequence ($z^{++} - K_{\mathrm{s}} < (B - z^{++}) \times 0.3$–$0.2$) in the BzK_{s} color–color diagram. Following their criteria, we consider only stars that are not too far from this sequence. Additionally, we use the size-magnitude diagram to examine the contamination of the star sample we selected earlier. The results are illustrated in Figure 2, where the distributions of galaxies and stars are represented by yellow triangles and blue points, respectively. Here, we do not attempt to obtain the half-light radius of each object by fitting their light distribution with different models. Instead, we directly use the FLUX_RADIUS provided by `SExtractor`. Some samples classified as stars by $\mathrm{lp_{type}} = 1$ are located in the region of extended sources in this figure, even for objects brighter than 23 mag in I band. Considering the different definitions of the size used by us and Weaver et al. (2022), the presence of these bright extended sources with $\mathrm{lp_{type}} = 1$ is reasonable. We eliminate potential extended sources from our stellar sample with a custom criterion, $\mathrm{FLUX\_RADIUS} > -0.2 \times \mathrm{I_{mag}} + 6.5$, shown as a red dashed line in Figure 2. Any sources located above this line are excluded from the following analysis. A total of 91 stars have been excluded. Moreover, the Classic catalog also provides the class_star parameter, which is estimated by `SExtractor` to distinguish stars from galaxies. Any ambiguous star samples showing class_star values lower than 0.85 are also removed. We have summarized the process of selecting galaxy and star samples in Table 1.

To verify the purity of our selected star sample, we present the distribution of the selected stars and galaxies in the gzK_{s} color–color diagram in Figure 3. Numerous studies have shown that stars follow a tight sequence in this color–color space (e.g., Arcila-Osejo & Sawicki 2013; Weaver et al. 2022). As affirmed in Figure 3, our selected star samples also tightly follow this sequence, demonstrating the reliability of our selection method.

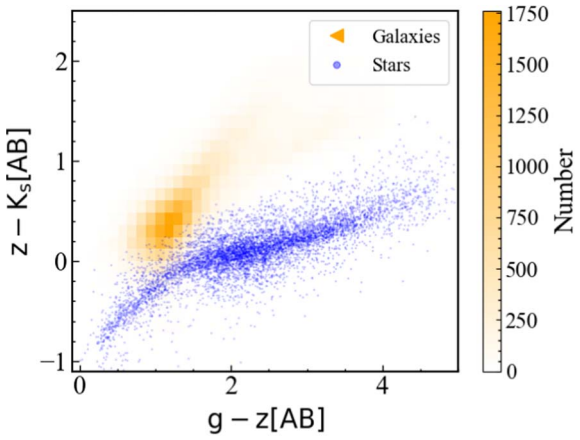

Figure 3. Distributions of our final selected galaxy and star samples in the gzK_s diagram. Galaxies are represented by the yellow 2D histogram with color representing the number in each bin, while stars are represented by blue points. The stars are clearly located on a distinct sequence, indicating that we have obtained clean star samples.

Table 1
Selection of Galaxy and Star Samples

Step	Galaxy Selection	Star Selection
1	$I_{mag} < 25$ $FLAG_{COMBINE} = 0$	$I_{mag} < 25$ $FLAG_{COMBINE} = 0$
2	$lp_{type} = 0$	$lp_{type} = 1$
3	$0.2 < z < 1.2$	$z^{++} - K_s < (B - z^{++}) \times 0.3$–$0.2$
4	⋯	FLUX_RADIUS $< -0.2 \times I_{mag} + 6.5$
5	⋯	class_star < 0.85

Utilizing the aforementioned methods, we ultimately obtain a clean sample comprising 7102 stars and 60,000 galaxies.

3. Preprocessing of Data and SML Models

3.1. Noise Reduction

Several previous studies have shown that noise can disrupt the features extracted from images by machine learning, potentially leading to incorrect classification results (Fang et al. 2023). Noise reduction has been identified as a viable solution to this issue. Masci et al. (2011) demonstrated that image quality can be significantly enhanced by extracting the primary features of images and subsequently reconstructing the images from these extracted features. In this work, we adopt Convolutional Autoencoders (CAEs) for noise reduction, which has been proven effective in many other studies (e.g., Zhou et al. 2022; Dai et al. 2023; Fang et al. 2023; Song et al. 2024). The specific approach we use follows that of Song et al. (2024).

First, we crop our samples into cutouts of 100×100 pixels, centering the objects within the image. We then apply CAE to reduce the noise in the images. The CAE configurations we adopt are identical to those used by Song et al. (2024). For more details, we refer readers to that paper. In Figure 4, we present the results of image denoising for some randomly selected samples. The first column showcases the raw images of four randomly selected objects, while the second column displays the corresponding denoised counterparts. It is evident that the CAE significantly enhances image quality without breaking the main features.

3.2. Adaptive Polar Coordinate Transformation

In the field of astronomy, the classification of stars and galaxies should be independent of image rotations. However, existing supervised machine learning (SML) algorithms, particularly those based on CNNs, may misclassify them when images are rotated. Various methods have been proposed to mitigate this issue, including data augmentation techniques and traditional polar coordinate transformation (e.g., Chen et al. 2018; Liu et al. 2019; Mo & Zhao 2022). However, data augmentation is very computationally intensive and inefficient, making it challenging to apply to future large-scale sky surveys. Polar-coordinate transformation is more efficient, but conventional polar-coordinate transformation cannot perfectly convert the rotations of the raw images into new images due to the integer coordinates of the pixels (as shown in Figure 3 of Fang et al. 2023). To overcome this shortcoming, we have proposed the APCT method in our previous works (e.g., Dai et al. 2023; Fang et al. 2023; Song et al. 2024).

In brief, the APCT method employs rotation-invariant polar axes, making it more robust to image rotation. This involves selecting the pixels with the highest and lowest flux values as the brightest and darkest points. The line connecting the brightest to the darkest point is designated as the polar axis for the polar-coordinate system. This designation ensures the polar axis remains unaffected by rotation. Next, we rotate the axis counterclockwise in increments of 0.05 radians. For each discrete rotation, the axis passes through many pixels of the original image. By stacking pixels along this rotating axis during rotation, a new image is obtained. Considering that CNNs are more sensitive to information in the center of images, we apply a mirroring process to the transformed images. The third column in Figure 4 shows the corresponding results after APCT for our randomly selected samples.

3.3. GoogLeNet Algorithm

In 2014, GoogLeNet (Szegedy et al. 2015) achieved remarkable results in the ImageNet image classification challenge, demonstrating its effectiveness in image classification tasks. The GoogLeNet architecture features nine sequentially stacked inception modules. Each inception module employs parallel applications of convolutions with kernel sizes of 1×1, 3×3, and 5×5. This configuration allows the network to cover a broader area of the input images while

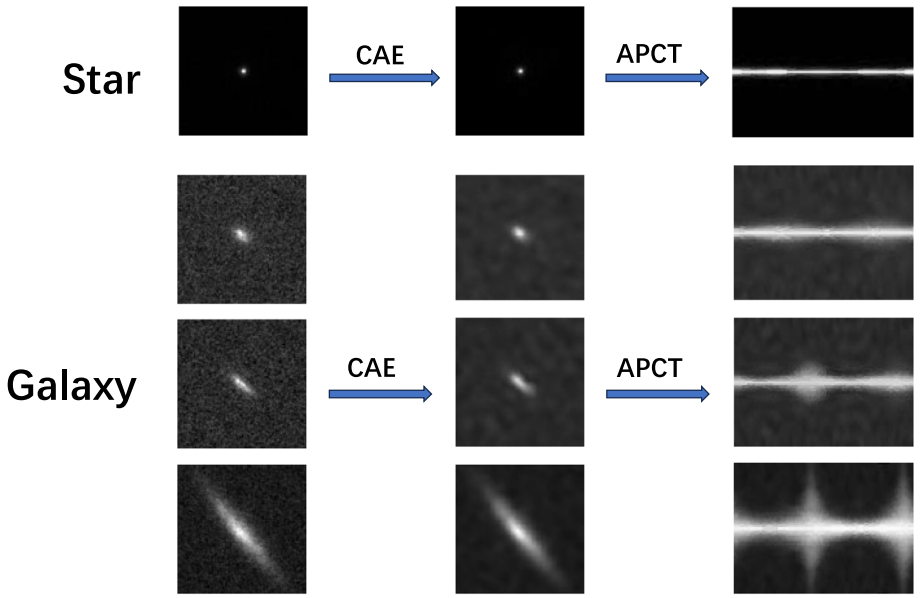

Figure 4. Examples of our preprocessing method. The first column shows images of four randomly selected samples, the second column displays the corresponding denoised images, and the third column presents the results after transformation into polar coordinates.

preserving fine detail in smaller regions. By expanding both the depth and width of the network, it achieves enhanced parameter efficiency while concurrently minimizing the total number of parameters. This advancement significantly contributes to the model's performance and computational efficiency. In our previous work, Fang et al. (2023) tested the performance of three different machine learning frameworks and found that the GoogLeNet model performs best for these deep high-resolution images. Therefore, we also adopt the GoogLeNet model in this work. The algorithmic flow of the GoogLeNet used in our study is illustrated in Figure 5. The model parameters used in this work are the same as those in Fang et al. (2023).

4. Results and Analysis

4.1. Classification Results of COSMOS Data

In Section 2, through meticulous selection, we have obtained reliable samples of stars and galaxies. Using the preprocessed images of these samples, we can evaluate the performance of the GoogLeNet algorithm in distinguishing between stars and galaxies. To avoid overfitting, we randomly divide our sample into training (48,040 galaxies and 5640 stars) and validation (11,960 galaxies and 1462 stars) sets in an 8:2 ratio and estimate the precision and recall rates based on this validation set.

The results of our classification are presented in Figure 6, where the left panel and right panel show the recall and precision rates, respectively. The recall rates for both stars and galaxies exceed 99.0%, while the precision rates are higher than 99.5%. Additionally, when considering the specific number of classification errors, only 6 galaxies and 12 stars are misclassified by the GoogLeNet model, as presented in Table 2. These results demonstrate the GoogLeNet model's robust performance in classifying star and galaxy images, with a low probability of misclassification between them.

Table 2
The Specific Classification Results of the Validation Set

Pred/Real	Galaxies	Stars	Total
Stars	6	1450	1456
Galaxies	11,954	12	11,965

Some previous works have also utilized machine learning for star-galaxy classification. For example, using data from the J-PLUS Early Data Release, López-Sanjuan et al. (2019) implemented a Bayesian classifier for morphological star-galaxy classification based on Probability Density Function analysis. They provided reliable probabilities for statistical analysis of 150,000 stars and 101,000 galaxies, achieving a completeness of approximately 95.0% and a contamination of approximately 5.0% for both stars and galaxies up to $r \sim 20$. Compared to their results, our study demonstrates the superior performance of the GoogLeNet model, with significantly higher precision and recall rates and a minimal misclassification rate, emphasizing the model's robustness and reliability in classifying stars and galaxies.

4.2. The Importance of Preprocessing

In Section 3, we thoroughly explained the necessity of preprocessing. The importance of noise reduction has been carefully discussed in many previous studies. Regarding

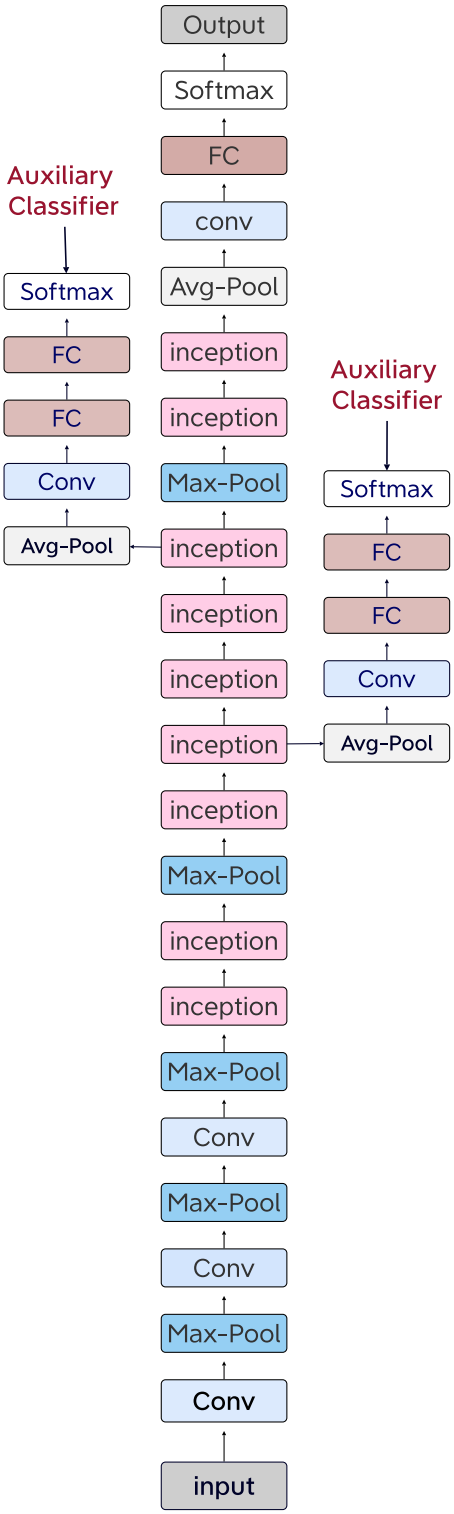

Figure 5. The structure of GoogLeNet used in this work is outlined in this figure: there are 22 layers in total, with no fully connected layers. The operations performed by each layer are represented in the boxes.

APCT, Fang et al. (2023) demonstrated its significance in galaxy morphology classification. When the test set is rotated, the overall accuracy significantly decreases from over 95.0% to approximately 83.0% without APCT. However, after performing APCT, the largest difference is only 1.5% no matter how many degrees the test set is rotated. In this section, we will further investigate the importance of our preprocessing method in the classification of stars and galaxies.

4.2.1. T-SNE Test

The t-Distributed Stochastic Neighbor Embedding (t-SNE) technique provides us with an effective method for mapping high-dimensional data into a lower-dimensional space, thereby transforming distributions into dimensions suitable for examination (e.g., van der Maaten & Hinton 2008; Van Der Maaten 2014; Wattenberg et al. 2016). In Figure 7, we present the t-SNE diagram for our samples before and after preprocessing in panels (a) and (b), respectively. To facilitate visualization, we only present the results of a randomly selected 5.0% subset of our entire sample. We have also examined the distribution of all samples and found that the results remained consistent.

It is generally believed that images with similar morphologies should have similar features, leading to distinct separable boundaries between different classes. Thus, stars and galaxies should be far apart in this diagram. In both panels of Figure 7, stars and galaxies exhibit clear separable boundaries, indicating that we have indeed provided viable prior samples for the GoogLeNet model. Although there is a slight overlap between stars and galaxies in the t-SNE diagram, this can be reasonable given the projection effects. Moreover, compared to panel (a), the boundaries between the two classes appear more distinct in panel (b), indicating the effectiveness of our preprocessing steps.

Since t-SNE can only qualitatively describe the effects of preprocessing, we have also applied an SVM classification to this feature plot to make a quantitative demonstration. The boundary lines between galaxy and star samples are estimated with the SVM technique, and the results are presented as red dashed lines in Figure 7. Since the lower left corner of each panel is predicted as star-like samples, and the upper right corner is predicted as galaxy-like samples by SVM classification, we calculate the accuracy and recall rates for galaxies and stars for this classification. Without preprocessing, the precision and recall rates for galaxies (stars) are 99.7% and 98.6% (98.6% and 99.7%), respectively. After preprocessing, the precision and recall rates improved to 99.9% and 99.4% (99.5% and 99.9%) for galaxies (stars). This indicates that galaxies and stars are indeed more distinctly separated in this t-SNE plot, indicating the effectiveness of our preprocessing process.

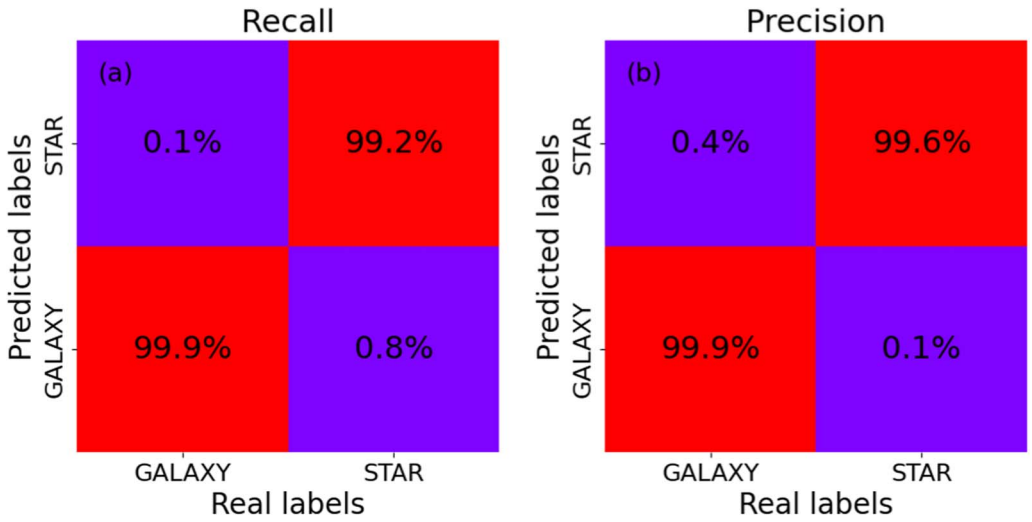

Figure 6. The recall (panel (a)) and precision (panel (b)) rates estimated from the validation set of the GoogLeNet model are depicted. The overall accuracy rates for both stars and galaxies exceed 99.0%, indicating that our framework demonstrates excellent performance in classifying stars and galaxies.

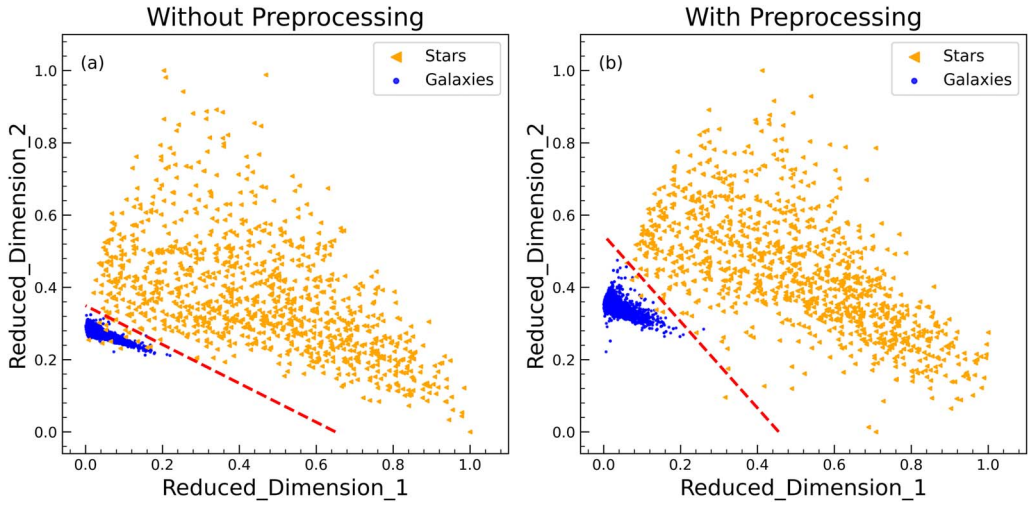

Figure 7. The t-SNE diagram depicts our randomly selected 5.0% samples, with the left panel illustrating the distribution of stars and galaxies without preprocessing, and the right panel showing the results after preprocessing. Using the data after t-SNE dimensionality reduction, an SVM classification on this feature plot is applied and the result is shown as the red dashed line in the figure. It can be seen clearly that after preprocessing, the galaxies and stars are indeed more distinctly separated on this t-SNE plot.

4.2.2. The Effectiveness of Preprocessing

To quantitatively demonstrate the effectiveness of preprocessing, we calculate the overall accuracy of the GoogLeNet model when images in the validation set are rotated by 0°, 90°, 180°, and 270°. In Figure 8, we present the overall accuracy as a function of training steps for the GoogLeNet model applied to the images with and without APCT. The cyan and magenta lines represent the median accuracy rate in each step bin, while the shaded areas enclose the corresponding 16th and 84th percentiles at each step bin. Different panels show the results obtained when the validation set is rotated at different angles.

From this figure, it is evident that regardless of the angle at which the validation set is rotated, preprocessing leads to an overall accuracy approaching 100%. Moreover, after reaching its peak, the accuracy of the validation set remains nearly constant as the number of training steps increases. In contrast, without preprocessing, the results are comparable to those with preprocessing when the validation set images are not rotated. However, when the validation set is rotated, the overall accuracy decreases by approximately 2.0% to 6.0%. Additionally, after reaching its peak, the accuracy of the validation set still exhibits considerable fluctuations as the number of training steps increases. This underscores the effectiveness of our preprocessing

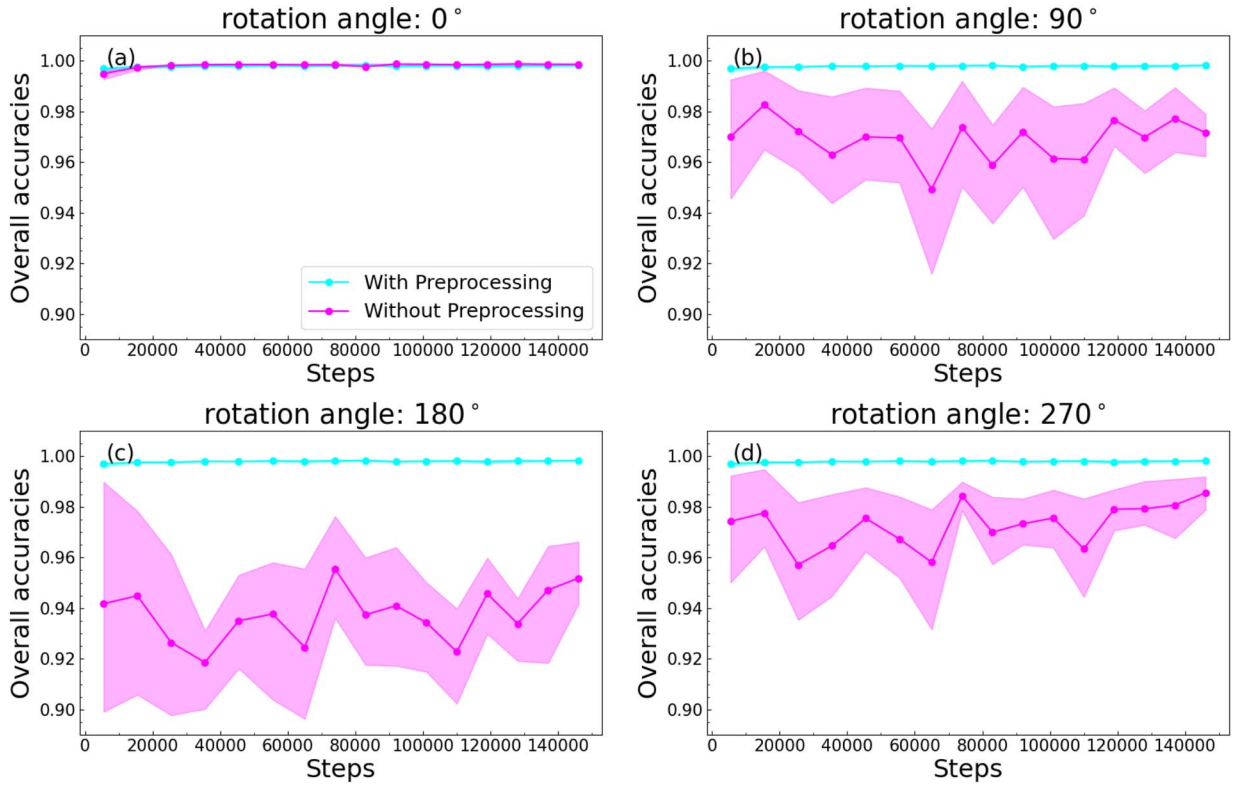

Figure 8. The accuracy as a function of training steps when images in the validation set are rotated with different angles is presented in four panels: panel (a) (0°), panel (b) (90°), panel (c) (180°), and panel (d) (270°). In each panel, the cyan and magenta lines represent the median accuracy rate in each step bin estimated with and without preprocessing, respectively. The shaded regions represent the corresponding 16th and 84th percentiles of accuracy rate at each step bin.

and highlights the poor rotational robustness of the traditional CNN framework.

5. Prepare for CSST

The CSST is a 2 m aperture space telescope that will be positioned in the same orbit as the China Manned Space Station. The CSST's survey camera is outfitted with 18 multi-band imaging detectors, encompassing 7 wavelength bands (*NUV*, *u*, *g*, *r*, *i*, *z*, and *y*). Each detector offers a field of view spanning 11×11 arcmin^2, with a scale of 9k × 9k pixels, and an average pixel size of $0\farcs74$ (Zhan 2021). Scheduled to commence observations in 2026, the CSST is slated for a 10 yr operational span. Over its operational period, the CSST aims to image approximately 15,000 square degrees of the sky at a depth of $r = 26.0$ mag and 400 square degrees at a depth of $r = 27.2$. In anticipation of the CSST's launch, we have evaluated the reliability of our framework using simulated CSST data.

5.1. Simulation Data

In this study, we utilize the data from the CSST main sky survey simulation data[8] to evaluate the performance of GoogLeNet in discriminating between stars and galaxies. The CSST simulation data used in this work are from version C6.2 of the CSST Main Sky Survey Simulation Data. The software used is the CSST Main Sky Survey Simulation Software 2.0.0. For comprehensive details regarding the simulation data, please consult the pertinent website. Briefly, as outlined by Cao et al. (2018) and Fu et al. (2023), the simulations are generated through the following steps: (1) creation of an input catalog containing sufficient physical properties (e.g., magnitude, redshift, source type, source shape, and source size) of all objects utilized in generating CSST observations, and (2) simulation of various physical and instrumental effects during observations, encompassing cosmic rays, sky background, nonlinearity, distortion, dark current, flat field, bias, charge diffusion effect, failed image elements/columns, CCD saturation overflow, instrument platform jitter, gain, readout noise, and others. The input galaxy catalog is derived from the "JiuTian" cosmological simulation catalog (e.g., Luo et al. 2016; Wei et al. 2018; Qiu & Kang 2022), while the input star catalogs are sourced from Gaia DR3 (e.g., Gaia Collaboration et al. 2021; Lindegren et al. 2021; Gaia Collaboration et al. 2023) and one simulated catalog obtained with `Galaxia` (Sharma et al. 2011). The final simulation comprises 137 multi-band exposures centered at R. A. = 244.9727 deg, decl. = 39.8959 deg, covering an area of

[8] https://csst-tb.bao.ac.cn/code/csst-sims/csst_msc_sim.git

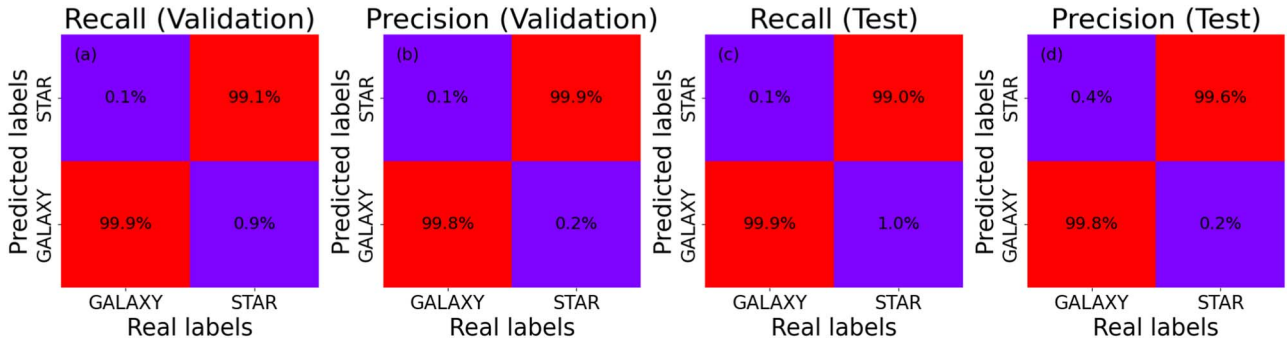

Figure 9. The results obtained from the CSST simulation data are depicted in panels (a) and (b), illustrating the recall and precision rates derived from the validation set, respectively. Similarly, panels (c) and (d) present the results obtained from the test set. Across all panels, the overall accuracy rates surpass 99.0%, indicating the suitability of our framework for future CSST missions.

approximately 1.53 square degrees. In the subsequent analyses, we focus on the data obtained in the *i* band, which is analogous to the *I* band utilized in our previous analysis.

5.2. Classification Results of CSST Simulation Data

Similar to Section 2, we obtain simulated images of stars and galaxies along with their corresponding input catalog. Given the known types of input sources, no additional steps are necessary to ensure a clean sample. We restrict our analysis to objects brighter than 24 mag in the *i* band when considering that the CSST depth is shallower compared to that of the COSMOS field. For galaxies, we further constrain our sample within the redshift range of $0.2 < z < 1.2$. After excluding images contaminated with cosmic rays, we obtain data for 180,135 galaxies and 34,788 stars. Taking into account the CSST pixel size of $0.''074$, we crop these simulated samples into cutouts with a size of 42×42 pixels. Then we apply preprocessing methods, including noise reduction and APCT, to these images.

To replicate the workflow after the observation on CSST, we employ data from the initial 45 exposures to train the GoogLeNet network. Subsequently, we utilize the trained network to classify the types of samples acquired from the remaining 92 exposures (referred to as the test set). During the network training phase, we partition the samples into training and validation sets in an 8:2 ratio to avoid overfitting. The detailed breakdown of stars and galaxies in the various data sets is presented in Table 3. Our classification outcomes are illustrated in Figure 9. Panels (a) and (b) depict the recall and precision rates of the validation set, while panels (c) and (d) exhibit the corresponding results of the test set. Notably, from panels (a) and (b), it is evident that the performance of the GoogLeNet model on the validation set mirrors the outcomes obtained with the COSMOS data, with overall accuracy rates surpassing 99.0%. Furthermore, when assessing the model's performance on a smaller training data set and a larger test data set, the results remain largely consistent. The recall (precision) rate for galaxies on the test set is 99.9% (99.8%), while for stars, the corresponding result is 99.0% (99.6%). These findings suggest that our network can be trained effectively with a limited data set and subsequently deployed with high reliability for the forthcoming CSST surveys.

Table 3
The Numbers of Stars and Galaxies in the Training, Validation, and Test Sets

Type	Training Set	Validation Set	Test Set	Total
Star	9276	2319	23,193	34,788
Galaxy	48,036	12,009	120,090	180,135

5.3. Discussion for Applications on Real CSST Images

In this study, we have demonstrated the effectiveness of our algorithm in classifying galaxies and stars using data from the HST/COSMOS field and CSST simulation data. However, when applying it to future real CSST data, further considerations may be necessary. First, in this study, we only considered relatively bright samples. For the fainter samples, due to their lower S/N, the classification results may not be as reliable. To address this issue, we need more samples of faint sources with reliable labels. For example, we can obtain reliable labels for more faint samples through simulated data and semi-analytic models. Additionally, CSST will provide 150 square degrees of deep field data, which will also aid in our study of faint sources. In addition, we used only single-band data for our samples. By incorporating multiband images of galaxies, we believe that we can achieve even more reliable results. We are also continuously iterating on our algorithm to make it more effective in extracting characteristic information from faint sources.

In addition, for real images, we do not have an input catalog to provide us labels for training samples. However, fortunately, many excellent instruments (e.g., HST, Euclid, Spitzer, and JWST) have already provided us with many high-quality data. By combining CSST's high-resolution images with multiband

photometric data and spectral data from other instruments, we can construct our galaxy and star samples by simultaneously considering the morphology and SED shapes of the targets (just like in Section 2.4 of our study and Section 5.1 of Weaver et al. 2022).

Moreover, due to CSST's high-efficiency survey capability, we will acquire a substantial amount of data over a short period. This imposes high demands on the efficiency of our data processing, such as how to cut out the stamp images of objects more efficiently given their large number. We have currently significantly improved the efficiency of cutting out these images by using multithreading techniques. However, the current method requires us to perform source detection in advance to obtain the coordinates of the targets. In the future, we are considering using some deep learning techniques (e.g., the YOLO algorithm) to automatically detect sources and extract the target galaxies (He et al. 2021). This will greatly enhance the efficiency of our entire workflow, making it more suitable for the large amount of data in the future.

6. Summary

In this study, we assess the efficacy of the GoogLeNet algorithm in distinguishing between stars and galaxies using data sourced from the COSMOS field, with data preprocessing applied. Through meticulous selection, we procure clean samples of stars and galaxies from the COSMOS2020 catalog. Subsequently, we employ the noise reduction technique with CAE and adopt APCT to transform the images into a polar-coordinate system. Partitioning the samples into training and validation sets in an 8:2 ratio, we observe remarkably high accuracy of the GoogLeNet model on the validation set, surpassing an overall accuracy of 99.0%. Notably, when the validation set is subjected to rotation, the accuracy attained without preprocessing exhibits a decrease of approximately 2.0% to 6.0% compared to the accuracy achieved with preprocessing. This underscores the superiority of our preprocessing approach in mitigating the SML method's poor robustness to image rotation. Moreover, in anticipation of the forthcoming CSST missions, we extend our framework to the CSST simulation data. Remarkably, the GoogLeNet model showcases exceptionally high accuracy, affirming the suitability of our framework for future CSST data analysis.

Acknowledgments

This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences (grant No. XDB 41000000), the National Natural Science Foundation of China (NSFC, Grant Nos. 12233008 and 11973038), the China Manned Space Project (No. CMS-CSST-2021-A07) and the Cyrus Chun Ying Tang Foundations. Z.S.L. acknowledges the support from Hong Kong Innovation and Technology Fund through the Research Talent Hub program (GSP028).

References

Arcila-Osejo, L., & Sawicki, M. 2013, MNRAS, 435, 845
Baldry, I. K., Robotham, A. S. G., Hill, D. T., et al. 2010, MNRAS, 404, 86
Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. 2006, ApJ, 650, 497
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Cabrera-Vives, G., Reyes, I., Förster, F., Estévez, P. A., & Maureira, J.-C. 2017, ApJ, 836, 97
Cao, Y., Gong, Y., Meng, X.-M., et al. 2018, MNRAS, 480, 2178
Chen, Y., Lyu, Z. X., Kang, X., & Wang, Z. J. 2018, 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2111
Cheng, G., Han, J., Zhou, P., & Xu, D. 2019, ITIP, 28, 265
Cheng, G., Zhou, P., & Han, J. 2016, ITGRS, 54, 7405
Dai, Y., Xu, J., Song, J., et al. 2023, ApJS, 268, 34
Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
Euclid Collaboration, Scaramella, R., Amiaux, J., et al. 2022, A&A, 662, A112
Fadely, R., Hogg, D. W., & Willman, B. 2012, ApJ, 760, 15
Fang, G., Ba, S., Gu, Y., et al. 2023, AJ, 165, 35
Fu, Z.-S., Qi, Z.-X., Liao, S.-L., et al. 2023, FrASS, 10, 1146603
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, A&A, 649, A1
Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2023, A&A, 674, A1
He, Z., Qiu, B., Luo, A.-L., et al. 2021, MNRAS, 508, 2039
Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841
Kim, E. J., & Brunner, R. J. 2016, MNRAS, 464, 4463
Kim, E. J., Brunner, R. J., & Carrasco Kind, M. 2015, MNRAS, 453, 507
Koekemoer, A. M., Fruchter, A. S., Hook, R. N., & Hack, W. 2003, in HST Calibration Workshop: Hubble after the Installation of the ACS and the NICMOS Cooling System, ed. S Arribas, A Koekemoer, & B Whitmore, 337
Koekemoer, A. M., Aussel, H., Calzetti, D., et al. 2007, ApJS, 172, 196
Kron, R. G. 1980, ApJS, 43, 305
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Lang, D., Hogg, D. W., & Mykytyn, D., 2016 The Tractor: Probabilistic astronomical source detection and measurement, Astrophysics Source Code Library, ascl:1604.008
Leauthaud, A., Massey, R., Kneib, J.-P., et al. 2007, ApJS, 172, 219
Lindegren, L., Klioner, S. A., Hernández, J., et al. 2021, AA, 649, A2
Liu, R., Zhang, Y., Zheng, Y., et al. 2019, CEVT, 10, 590
Liu, Y., Tu, H.-L., Zhou, C.-C., Liu, Y., & Zhang, F.-L. 2020, arXiv:2005.11679
Luo, Y., Kang, X., Kauffmann, G., & Fu, J. 2016, MNRAS, 458, 366
López-Sanjuan, C., Vázquez Ramió, H., Varela, J., et al. 2019, A&A, 622, A177
MacGillivray, H. T., Martin, R., Pratt, N. M., et al. 1976, MNRAS, 176, 265
Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J. 2011, in Artificial Neural Networks and Machine Learning – ICANN 2011, ed. T Honkela et al. (Berlin: Springer), 52
Mo, H., & Zhao, G. 2022, arXiv:2211.11812
Nazaré, T. S., da Costa, G. B. P., Contato, W. A., & Ponti, M. 2018, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 22nd Iberoamerican Congress, CIARP 2017, ed. M Mendoza & S Velastín, (Berlin: Springer), 416
Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., & Zumach, W. A. 1992, AJ, 103, 318
Qiu, Y., & Kang, X. 2022, ApJ, 930, 66
Ross, A. J., Ho, S., Cuesta, A. J., et al. 2011, MNRAS, 417, 1350
Saglia, R. P., Tonry, J. L., Bender, R., et al. 2012, ApJ, 746, 128
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Sevilla-Noarbe, I., & Etayo-Sotos, P. 2015, A&C, 11, 64
Sevilla-Noarbe, I., Hoyle, B., Marchã, M. J., et al. 2018, MNRAS, 481, 5451
Sharma, S., Bland-Hawthorn, J., Johnston, K. V., & Binney, J. 2011, ApJ, 730, 3
Skelton, R. E., Whitaker, K. E., Momcheva, I. G., et al. 2014, ApJS, 214, 24
Song, J., Fang, G., Ba, S., et al. 2024, ApJS, 272, 42
Soumagnac, M. T., Abdalla, F. B., Lahav, O., et al. 2015, MNRAS, 450, 666
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv:1503.03757
Suchkov, A. A., Hanisch, R. J., & Margon, B. 2005, AJ, 130, 2439

Szegedy, C., Liu, W., Jia, Y., et al. 2015, Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ: IEEE), 1
Thomas, S. A., Abdalla, F. B., & Lahav, O. 2011, PhRvL, 106, 241301
Van Der Maaten, L. 2014, JMLR, 15, 3221
van der Maaten, L., & Hinton, G. 2008, JMLR, 9, 2579
Vasconcellos, E. C., de Carvalho, R. R., Gal, R. R., et al. 2011, AJ, 141, 189
Walmsley, M., Ferguson, A. M. N., Mann, R. G., & Lintott, C. J. 2018, MNRAS, 483, 2968
Wattenberg, M., Viégas, F., & Johnson, I. 2016, Distill, 1, e2
Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, ApJS, 258, 11
Wei, C., Li, G., Kang, X., et al. 2018, ApJ, 853, 25
Weir, N., Fayyad, U. M., & Djorgovski, S. 1995, AJ, 109, 2401
Zhan, H. 2011, SSPMA, 41, 1441
Zhan, H. 2018, in 42nd COSPAR Scientific Assembly, E1.16–4–18
Zhan, H. 2021, ChSBu, 66, 1290
Zhou, C., Gu, Y., Fang, G., & Lin, Z. 2022, AJ, 163, 86