A versatile framework for analyzing galaxy image data by incorporating Human-in-the-loop in a large vision model[*]

Ming-Xiang Fu (傅溟翔)[1,3,4#] Yu Song (宋宇)[2#] Jia-Meng Lv (吕佳蒙)[2#] Liang Cao (曹亮)[2] Peng Jia (贾鹏)[2†]
Nan Li (李楠)[1,3,4‡] Xiang-Ru Li (李乡儒)[5] Ji-Feng Liu (刘继峰)[1,4] A-Li Luo (罗阿理)[3,6,7] Bo Qiu (邱波)[8]
Shi-Yin Shen (沈世银)[9] Liang-Ping Tu (屠良平)[15] Li-Li Wang (王丽丽)[10] Shou-Lin Wei (卫守林)[11]
Hai-Feng Yang (杨海峰)[12] Zhen-Ping Yi (衣振萍)[13] Zhi-Qiang Zou (邹志强)[7,14]

[1]National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China
[2]College of Electronic Information and Optical Engineering, Taiyuan University of Technology, Taiyuan 030024, China
[3]School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 101408, China
[4]Key lab of Space Astronomy and Technology, National Astronomical Observatories, Beijing 100101, China
[5]School of Computer Science, South China Normal University, Guangzhou 510631, China
[6]CAS Key Laboratory of Optical Astronomy, National Astronomical Observatories, Beijing 100101, China
[7]University of Chinese Academy of Sciences, Nanjing, Nanjing 211135, China
[8]University of Science and Technology Beijing, Beijing 100083, China
[9]Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China
[10]School of Computer and Information, Dezhou University, Dezhou 253023, China
[11]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
[12]School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 30024, China
[13]School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China
[14]Nanjing University of Posts & Telecommunications, Nanjing 210023, China
[15]School of Science, University of Science and Technology Liaoning, Anshan 114051, China

Abstract: The exponential growth of astronomical datasets provides an unprecedented opportunity for humans to gain insight into the Universe. However, effectively analyzing this vast amount of data poses a significant challenge. In response, astronomers are turning to deep learning techniques, but these methods are limited by their specific training sets, leading to considerable duplicate workloads. To overcome this issue, we built a framework for the general analysis of galaxy images based on a large vision model (LVM) plus downstream tasks (DST), including galaxy morphological classification, image restoration, object detection, parameter extraction, and more. Considering the low signal-to-noise ratios of galaxy images and the imbalanced distribution of galaxy categories, we designed our LVM to incorporate a Human-in-the-loop (HITL) module, which leverages human knowledge to enhance the reliability and interpretability of processing galaxy images interactively. The proposed framework exhibits notable few-shot learning capabilities and versatile adaptability for all the abovementioned tasks on galaxy images in the DESI Legacy Imaging Surveys. In particular, for the object detection task, which was trained using 1000 data points, our DST in the LVM achieved an accuracy of 96.7%, while ResNet50 plus Mask R-CNN reached an accuracy of 93.1%. For morphological classification, to obtain an area under the curve (AUC) of ~0.9, LVM plus DST and HITL only requested 1/50 of the training sets that ResNet18 requested. In addition, multimodal data can be integrated, which creates possibilities for conducting joint analyses with datasets spanning diverse domains in the era of multi-messenger astronomy.

Keywords: artificial intelligence, large vision model, human-in-the-loop, astronomy, galaxies

DOI: 10.1088/1674-1137/ad50ab

Received 13 March 2024; Accepted 24 May 2024; Published online 25 May 2024

[*] The authors acknowledge the support from the National Natural Science Foundation of China (Grant Nos. 12173027, 12303105, 12173062), the National Key R & D Program of China (Grant Nos. 2023YFF0725300, 2022YFF0503402), the Science Research Grants from the Square Kilometre Array (SKA) (2020SKA0110100), the Science Research Grants from the China Manned Space Project (Grant Nos. CMS-CSST-2021-A01, CMS-CSST-2021-A07, CMS-CSST-2021-B05), and the CAS Project for Young Scientists in Basic Research, China (Grant No. YSBR-062). This work was supported by the Young Data Scientist Project of the National Astronomical Data Center and the Program of Science and Education Integration at the School of Astronomy and Space Science, University of Chinese Academy of Sciences, China

[†] E-mail: robinmartin20@gmail.com

[‡] E-mail: nan.li@nao.cas.cn

[#] These authors contributed equally as the first authors

I. INTRODUCTION

A vast expansion of available data is an invaluable resource across various scientific disciplines, notably physics and astronomy, because it offers many opportunities and challenges for human beings to understand the universe. Artificial intelligence (AI) techniques have emerged as a leading approach for comprehending the complexities intrinsic to big data challenges in physics, such as interpreting data collected from large-scale sky surveys, gravitational wave detectors, and colliders. These datasets are more than an order of magnitude larger in size than previous datasets, but require a shorter processing time to promptly respond to transient events [1]. They have also supported significant successes, such as the prediction of multivariate time series data drawn from particle accelerators [2], the execution of many-body variational calculations in nuclear physics [3], and many other accomplishments in experimental and theoretical physics (see Ref. [4] and references therein).

The big data challenge in astronomy and astrophysics is especially important [5] because large-scale sky surveys such as LSST[1], Euclid[2], CSST[3], and SKA[4] continue to gather data, leading astronomy and astrophysics into an exciting new era. However, the vast and intricate nature of astronomical datasets poses a significant challenge to astronomers who want to extract meaningful scientific information. Deep learning techniques have been used to address this difficulty (see [6] and references therein). For example, astronomers have leveraged specific data in supervised learning to teach computers how to solve problems, which has been successful in detecting celestial objects [7], classifying their morphology [8, 9] and identifying their spectra [10, 11]. In addition, unsupervised learning algorithms can explore unlabeled data and have demonstrated their effectiveness in classifying galaxy types [12−15] and in characterizing (or improving) the performance of telescopes [16−19]. Furthermore, reinforcement learning algorithms have succeeded in various applications, such as efficiently managing instruments via developing simulators and enabling interactions with observations [20, 21].

However, for the machine learning-based applications discussed above, certain issues still need to be addressed, including interpretability, data labeling, and universality. Persistent issues that hinder their advancement and utility require preparing separate training sets and constructing distinct models for different tasks. Despite this, various tasks may share a common foundation of prior information about celestial objects. For example, tasks such as detecting strong gravitational lensing systems, identifying different types of nebulae or galaxies, and segmenting galaxies share the same need for multicolor structural features. Therefore, creating a foundational model that provides general information and attaches subprocesses for multiple purposes is sensible. Moreover, effectively training a machine learning algorithm typically requires thousands of data units, further exacerbating matters, as obtaining specific data and labels (e.g., the positions of rare astronomical targets or segmentation labels for galaxies) is complex. Therefore, an interactive technique is ideal for building training sets from scratch and maintaining their development.

To overcome the abovementioned shortcomings of existing applications of deep learning to astronomical vision tasks, especially galaxy image processing tasks, we have developed a comprehensive framework containing a foundational model, multiple machine learning models for downstream tasks, and a human-in-the-loop (HITL) interface. The foundational model is based on the Swin-Transformer model [22], and the galaxy images from the ssl-legacysurvey project [23], which contains 76 million galaxy images extracted from the Dark Energy Spectroscopic Instrument (DESI) Legacy Survey [24] Data Release 9, are selected as pre-training data. Covering 14000 square degrees of extragalactic sky in three optical bands (g, r, z), these data constitute a relatively complete description of galaxies in the nearby universe. Different neural networks are then attached to the trained model for downstream tasks, including classification, image restoration, and outlier detection. The model requires far fewer training samples than the current supervised learning algorithms and is suitable for various purposes. To further enhance the performance of the model, a HITL module based on the FLASK web framework [25] is connected to our framework. This module takes advantage of human knowledge to further decrease the workload of data labeling and to improve the reliability, universality, and interpretability of the framework for different image processing tasks.

II. A FOUNDATIONAL VISION MODEL FOR ASTRONOMY

According to deep learning theory [26], there has been a proliferation of neural networks featuring progressively deeper architectures, from millions to billions of parameters, that encode prior knowledge about specific domains of problems. Two examples of these networks are large language models (LLM) [27−30] and large vision models (LVM) [31, 32]. These so-called

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

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

[3] http://nao.cas.cn/csst/

[4] https://www.skao.int/

large models can be used as the backbone for various tasks, offering proficient few-shot learners capable of handling various data processing challenges. In this study, an LVM is developed as the foundational model on the basis of the Swin-Transformer architecture. The LVM was trained in an unsupervised manner using 76 million stamp images with g, r, z-bands [23, 33] from the DESI Legacy Imaging Surveys. More details on the LVM are presented in the remaining parts of this section.

A. Design of the large vision model

Figure 1 illustrates the architecture of our LVM, which is based on the SUNET framework [34] and possesses a parameter count of approximately 100 million. For demonstration, Fig. 1 only displays four layers of the Swin-Transformer Block (STB). The core structure of the LVM follows an encoder-decoder paradigm, with Swin-Transformers serving as the fundamental building blocks. The use of Swin-Transformers is pivotal in amplifying the interpretability of abstract image features, which is a crucial factor for grasping the fundamental elements and inherent characteristics contained in the data. The Swin-Transformer effectively processes local information through its window attention, gradually expands the receptive field, and integrates global information through shifted window attention. Additionally, compared to traditional Transformers or ViTs, the Swin-Transformer significantly reduces computational complexity and memory requirements without compromising the model's performance. In essence, LVM attempts to reconstruct the original images utilizing the sparse features extracted by the encoder and decoder [35]. This process involves learning a mapping function that translates three channels of two-dimensional image data into semantic features in the latent space.

The LVM encoder is comprised of four layers, each containing eight consecutive Swin-Transformer layers (STL), and the decoder has an identical structure. Moreover, 3×3 convolution kernels are adopted along with the STB. By engaging in feature processing across multiple tiers and assimilating global information, the model attains a deeper understanding of the interconnectedness and dependencies inherent within the images. In turn, this facilitates more effective inference processes. This architectural design, featuring a deep feature pyramid, significantly fortifies the performance of the model across tasks encompassing various scales. Figure 2 shows two STLs. These blocks encompass normalization layers, a window-based self-attention layer (Window MSA), a shift window-based self-attention layer (Shift-Window MSA), and a multilayer perceptron (MLP). These layers enhance the perceptive capabilities of the Swin-Trans-

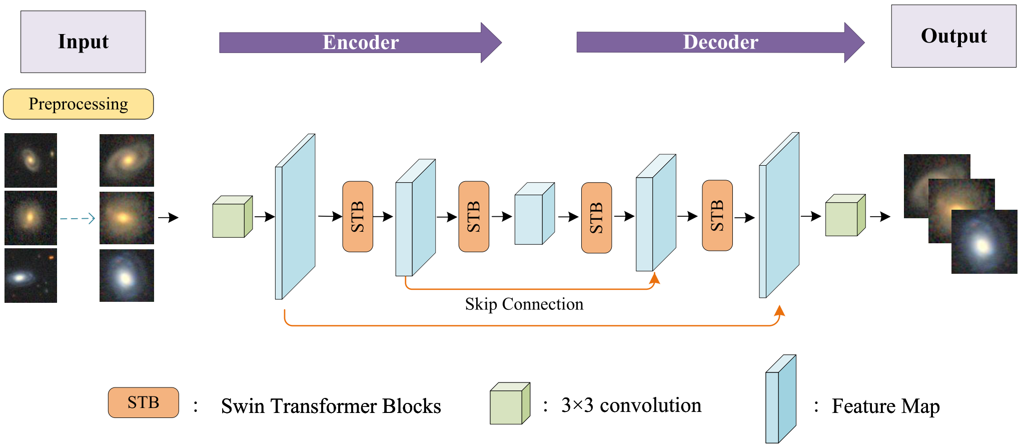


Fig. 1. (color online) Structure of the large vision model (LVM).

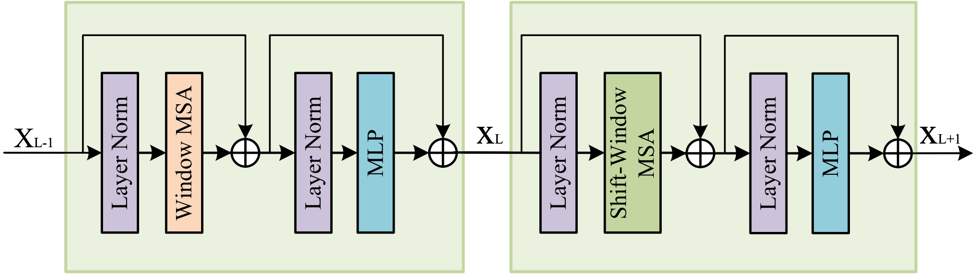


Fig. 2. (color online) Structure of the Swin-Transformer layer (STL) in the LVM.

former to a greater degree than traditional convolutional neural networks. These components are integrated into the U-net structure, effectively increasing the receptive field of the neural network, which substantially amplifies the capacity of the model to represent data samples within the feature space.

B. Pre-training of the large vision model

The LVM undergoes pre-training through a self-supervised method [36, 37] using images of celestial objects from the DESI Legacy Imaging Surveys DR9. Each instance presented to the model is a galaxy image, and an identical galaxy image is generated as its output. These images are comprised of three channels (the $[g, r, z]$-bands) and are resized to $152 \times 152 \times 3$ pixels. The LVM initially compresses galaxy images into feature vectors via the encoder and subsequently reconstructs galaxy images based on these vectors. By utilizing the Mean Squared Error (MSE) loss, the difference between the reconstructed galaxy images and the originals can be measured, which facilitates effective learning of galaxy image representations.

When processing galaxy images, the varying effective sizes of different galaxy images poses a challenge. Leaving this problem unsolved could lead to some galaxies appearing relatively small in the images, making effective analysis and recognition problematic. To overcome this problem, an OpenCV-based algorithm was devised to adaptively crop images [38]. The algorithm calculates the effective area that the galaxies occupy in each stamp according to the grayscale level. Then, it cuts and resizes the original images to create new stamp images with a fixed size of $128 \times 128 \times 3$. This step ensures that each galaxy occupies an appropriate area within the image without losing much information, thus easing subsequent processing and analysis. In addition, the data was augmented by applying flips, rotations, and croppings to generate a more diverse set of training samples from the images in order to enhance the coverage of the training sets in latent space, which enhanced the generality of the model and its overall robustness.

The batch size in the pre-training stage was set to 512, which balanced the efficiency and hardware limits. For each iteration, the MSE is computed for all images in a batch, and then the model parameters are updated using the Adam optimizer. Approximately 196 hours were required for eight NVIDIA A100s with 80 GB of graphic memory to train the LVM. After training, the encoder within the LVM acquires the ability to learn the inherent features of the celestial objects. It is feasible to cut this encoder from the LVM and connect it to the following neural networks for further training. This extended training could involve various downstream tasks and the HITL strategy, which is detailed in the following section.

III. TRAINING THE LARGE VISION MODEL FOR MULTIPLE DOWNSTREAM TASKS

A. Training of multiple downstream tasks

Given that common foundational knowledge is suitable for various downstream tasks, the encoder's proficiency can be enhanced within the LVM by concurrently engaging it in multiple downstream tasks. This approach aims not only to enhance the versatility of the LVM, but also to optimize task-specific performance. In line with this philosophy, three downstream tasks were identified: galaxy classification, image restoration, and image reconstruction. For each task, a task-specific neural network is incorporated alongside the LVM encoder, as illustrated in Fig. 3. During the multitask training stage, the model parameters for the entire multitask framework are updated with these tasks. An active learning strategy is used to dynamically adjust the proportion of training dedicated to different tasks.

The image classification task aims to classify galaxies according to their morphologies. This was achieved by adding two fully connected layers following the LVM encoder. Additionally, a dataset containing images of galaxies was constructed, with four distinct classes: elliptical, edge-on, spiral, and other (including irregular and merging galaxies). These galaxy images were obtained from the DESI Legacy Imaging Surveys, and the labels indicating their morphologies were obtained from the Galaxy Zoo 2 project [1]. The data processing method discussed in [39] was used to obtain high-quality labels. For the multitask training, a training dataset containing 500 galaxy images per category (for a total of 2000 images), was used for model training. Furthermore, a test dataset consisting of 250 images in each category (for a total of 1000 images), was utilized to assess the model's performance.

The image restoration task aims to generate high-quality original images from blurred ones. This was achieved by incorporating a decoder module with convolutional layers following the LVM. The dataset was comprised of two components: 1) reference images, which were high-quality raw galaxy images obtained from the DESI Legacy Imaging Surveys, and 2) blurred images generated by introducing noise and blurred point spread functions (PSFs) using the method outlined in [40]. In the experiment, the Moffat model was employed, assuming that the full widths at half-maximum (FWHMs) of the PSFs were distributed in the 2.0−8.0 pixels range. Additionally, to simulate the blurred data, the noise source was assumed to be a Gaussian function with a standard devi-

[1] https://data.galaxyzoo.org/

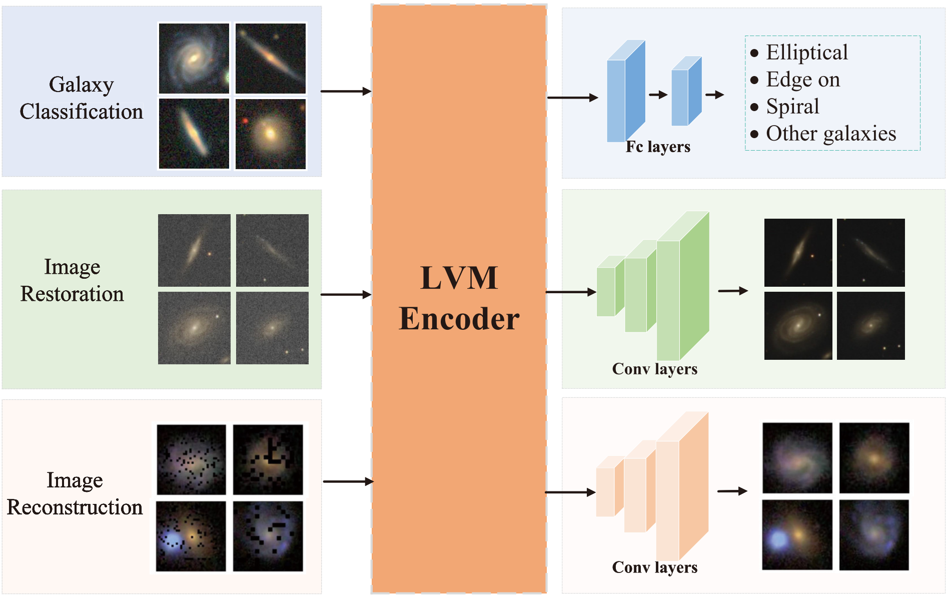


Fig. 3. (color online) Schematic of the multitask training process.

ation uniformly distributed between 1.0 and 15.0. These blurred images simulated the degradation and noise found in real observations. For the purpose of multitask training, a training dataset containing 1000 blurred images and a test dataset consisting of 100 images were utilized to assess the model's performance. Both the training and test datasets were derived from simulated data, ensuring a controlled environment for the training and evaluation of the model.

The image reconstruction task aims to mend the obstructed sections of images, facilitating the segmentation of individual galaxies from several adjacent galaxies. This was achieved by integrating a decoder module comprised of convolutional layers following the LVM. The dataset for this task consisted of two components: 1) reference images, which were original images without bad pixels or other defects obtained from the DESI Legacy Imaging Surveys, and 2) masked images, which were original images masked by varying patch sizes from 0% to 70% with a random scale (this emulated image degradation processes that may occur during observation and acquisition). For multitask training, a training dataset consisting of 1000 data pairs and a test dataset containing 100 pairs were employed to assess the model's performance in image reconstruction.

Various loss functions are applied during the training process for different tasks. Cross-entropy is used as the loss function for galaxy classification tasks, and the MSE is employed for image restoration and reconstruction tasks. For comparative studies, we adopted two training strategies for training downstream task models. The first strategy ($Multi\_uniform$) maintains equal weights for each task during training, meaning that the training proportion for each task is consistent. The second strategy ($Multi\_active$) actively updates the training proportion of each task according to the characteristics and performance exhibited during the training process. This strategy quantifies the proportion of each task allocated during training by evaluating its performance (using, for example, the MSE or F1 score) on the test set. Tasks that demonstrate better performance metrics are allocated a smaller proportion of training data, while tasks with lower performance metrics receive a larger proportion.

B. Performance evaluation

A series of comparative experiments were conducted to evaluate the performance of a model not trained by multitasking (Pre_train model) against two multitasking-trained models (the $Multi\_uniform$ and $Multi\_activate$ models) using the training strategies mentioned earlier across image classification, image reconstruction, and image restoration tasks. The Pre_train model is given by using the frozen training approach (i.e., the weights of the LVM are kept constant and only the weights in the task head are updated).

In the image classification task, our model was trained using a dataset of 1000 images, and its performance was evaluated using a separate set of 500 images. The results are presented in Table 1, which displays the classification accuracy (Acc), precision (Pre), recall (Recall), and F1 scores for the three distinct training strategies. Our analysis indicated that the models trained on multiple tasks outperformed those trained exclusively on the restoration task (Pre_train model) with respect to the classification accuracy and various other metrics. In particular, the actively selected task strategy

Table 1. Classification results for model trained with various training strategies.

Classification Task	Acc	Pre	Recall	F1
Pre_train model	0.784	0.767	0.794	0.771
Multi_uniform model	0.842	0.844	0.850	0.846
Multi_activate model	0.854	0.823	0.834	0.844

Table 2. Image restoration results for different training strategies.

Restoration Task	MSE	PSNR	SSIM
Blurred images	0.00094	31.11	0.48
Pre_train model	0.00084	31.31	0.51
Multi_uniform model	0.00083	31.35	0.54
Multi_activate model	0.00049	33.34	0.56

(*Multi_activate* model) demonstrated a substantial improvement in accuracy compared to the other two. These results suggest that the *Multi_activate* training strategy can augment the model's classification performance, and that actively selecting the task further enhances the accuracy by slightly reducing the precision, recall, and F1 score (see Table 1).

To gauge the effectiveness of our model in terms of image restoration, we employed a variety of metrics, including the PSNR (Peak Signal-to-Noise Ratio), MSE, and SSIM (Structural Similarity Index). These metrics were utilized to evaluate the agreement between the processed images and the original unprocessed images. Higher PSNR, higher SSIM, and lower MSE values indicated better agreement. Our findings, as presented in Table 2, indicated that, while the models trained with the *Multi_uniform* strategy outperformed the *Pre_train* models, the multitasking plus active learning strategy (*Multi_activate*) was the optimal model.

For the image reconstruction task, a dataset consisting of 1000 samples was used for training, and a dataset consisting of 100 samples was used for testing. The performance was also evaluated using the metrics of the PSNR, MSE, and SSIM. The multitasking-trained models (the *Multi_uniform* and *Multi_activate* models) exhibited superior performance in the image reconstruction task, surpassing the non-multitasking-trained model (the *Pre_train* model) in the reconstruction of masked regions, as shown in Table 3. Furthermore, a comprehensive evaluation of the outcomes was conducted by analyzing the image reconstruction performance under varying levels of missing data and patch sizes. The results presented in Table 4 and Fig. 4 demonstrate the remarkable performance of the model, even when processing highly degraded images with a missing content rate of up to 70% and a patch size of 8×8 (for an input data of size 128×128).

Table 3. Image reconstruction results for models trained with different strategies.

Reconstruction Task	MSE	PSNR	SSIM
Masked images	0.0248	15.64	0.36
Pre_train model	0.0089	22.36	0.49
Multi_uniform model	0.0040	26.07	0.61
Multi_activate model	0.0038	26.84	0.64

Table 4. Statistical analysis of image reconstruction performance. Comparison between PSNR, SSIM, and MSE for various patch sizes and masking proportions (using frozen and fine-tuned LVM model parameters).

Patch size	MSE	PSNR	SSIM
4×4 (Masked images)	0.03187	15.92	0.27
Multi_activate model	0.0067	26.54	0.58
8×8 (Masked images)	0.023	17.55	0.36
Multi_activate model	0.0049	24.81	0.55
16×16 (Masked images)	0.03149	17.47	0.42
Multi_activate model	0.0112	23.61	0.48

In summary, multitasking training performed in conjunction with active learning significantly enhanced the performance of the model across different tasks. Compared to the model that did not undergo multitask training, the utilization of multitask training facilitated a more effective acquisition of feature representation and enhanced the generalization ability of the neural network, rendering it suitable for a variety of astronomical image processing tasks.

IV. DEPLOYMENT OF TWO SAMPLE APPLICATIONS

Two astronomical vision tasks were chosen to showcase the capabilities of our LVM model: galaxy morphology classification and strong lens detection in a large field of view. The LVM was utilized as the backbone and two separate downstream models were employed for the two tasks. Detailed information on each of these applications is presented below.

A. Classifying galaxy morphology with few-shot learning based on LVM

The performance of the proposed algorithm was further evaluated according to its ability to classify the morphologies of galaxies in the DESI Legacy Imaging Surveys using the few-shot learning approach. The training and testing sets included image data from the DESI Legacy Imaging Survey and labels from the Galaxy Zoo project. The galaxies were categorized into five types

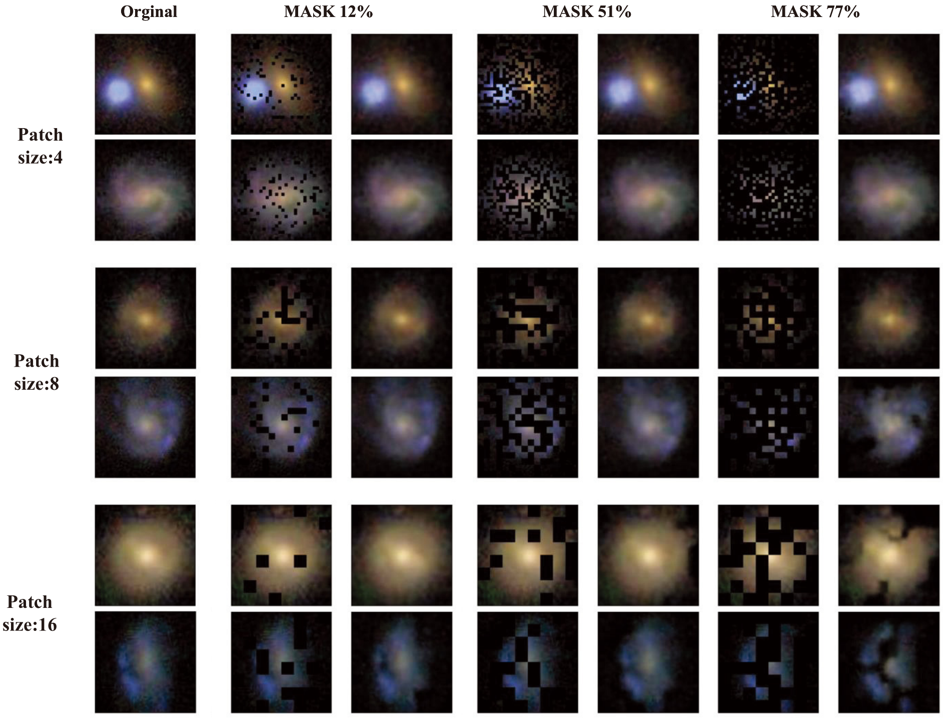


Fig. 4. (color online) Images reconstructed by the model with varying patch block sizes and masking proportions.

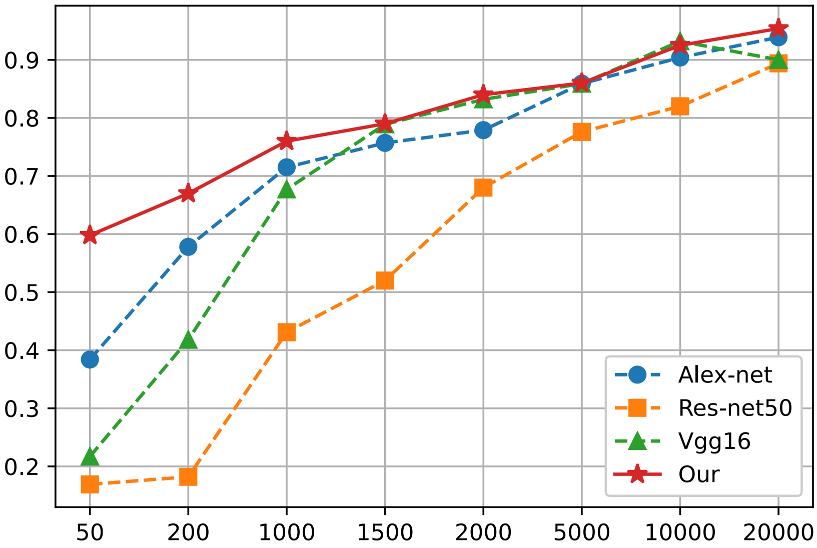


Fig. 5. (color online) Classification accuracy of four different models (our model, AlexNet, VGG16, and ResNet50) for the galaxy classification task as a function of the dataset size.

[41]. After the LVM encoder, a fully connected neural network is employed for the task of galaxy morphological classification and then trained with the above training sets. A comparative analysis was performed by evaluating the results of our model and those of AlexNet [42], VGG16 [43], and ResNet50 [44], which are deep learning architectures that have been proven effective in various image recognition tasks. As Fig. 5 illustrates, the LVM + Downstream Tasks model maintained a higher accuracy, especially in scenarios with minimal data (only 10 images per class). Moreover, as the amount of data increased, the model's performance gradually improved, further confirming its scalability to large datasets. These experimental results not only demonstrated the effectiveness of the LVM + Downstream Tasks model for galaxy morphology classification tasks, but also revealed its stability and generalization ability when handling datasets of different sizes.

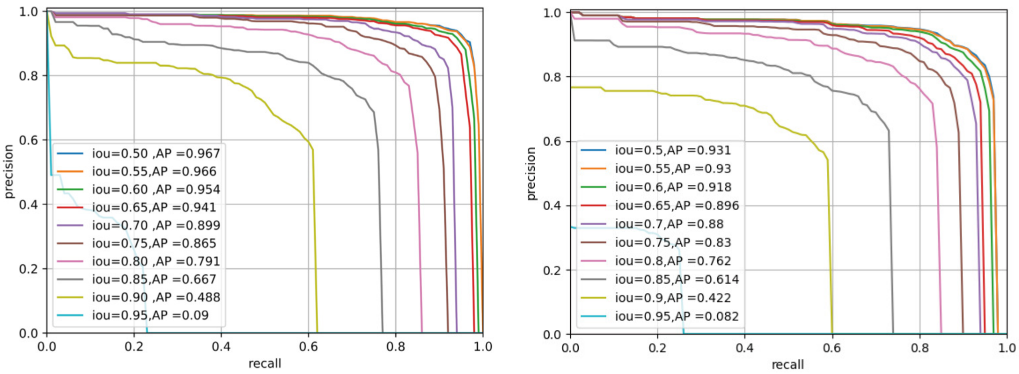


Fig. 6. (color online) Target detection results with different backbones. The left panel shows the target detection results achieved by Mask R-CNN using the LVM as the backbone, and the right panel displays the target detection results obtained by Mask R-CNN using ResNet50 as the backbone. The LVM significantly enhanced the detection capabilities of the neural network.

B. Identifying strong lensing systems with the LVM + Mask R-CNN

To replicate this trend in source detection, a strong lens dataset containing 1000 training images and 1000 testing images was constructed. These images were extracted from the DESI website[1] using the catalog of strong lensing system candidates available in the *NeuraLens Database*[2]. For the downstream task of finding strong lensing systems within a large field of view, Mask R-CNN [45] was chosen as our model. Additionally, ResNet50 was employed as the backbone of Mask R-CNN for comparison. The results, which are presented in Fig. 6, demonstrated that our LVM + Mask R-CNN model achieved an impressive average precision (AP) of 96.7% with 1000 training images. In contrast, the ResNet + Mask R-CNN model achieved a slightly lower AP value of 93.1%. This comparison underscores the effectiveness of our LVM approach in enhancing the performance of Mask R-CNN for strong lens detection.

V. LARGE VISION MODEL WITH THE HUMAN-IN-THE-LOOP MODULE

To interactively integrate human knowledge, we developed a HITL module [46] based on the Flask Web Framework[3] and integrated it into the LVM. Taking the binary classification task as an example, an MLP [47] model with a hidden layer size of 2048 is used to predict the types of galaxies and to introduce the HITL module with an adaptive algorithm in order to find potential objects and boost the model's purity, completeness, and several other metrics. These objects are labeled and included in the training sets in the MLP training procedure. With this module, astronomers can create training sets iteratively from scratch for their specific purposes and direct the model's optimization path as necessary.

To evaluate its feasibility, the HITL was used to distinguish between spiral and elliptical galaxies. It achieved an area under the precision-recall curve (AUPR) of 0.8895 by starting with 10 initial prompts (five positive and five negative) and following one interaction step with 10 recommended examples. Its performance surpassed that obtained by training the LVM with 30 examples (15 positive and 15 negative, AUPR = 0.8683) and training ResNet18 with 100 examples (50 positive and 50 negative, AUPR = 0.8561), and comparable to that achieved by training the LVM with 1000 random examples (AUPR = 0.8921). Figure 7 presents the results of testing the HITL on specific target identification tasks with few-shot learning, such as finding galaxies with bars, strong lensing systems, and galaxy mergers. The results further proved the capacity of this module and demonstrated its broad application potential for various tasks. More details are discussed below.

A. Design of the human-in-the-loop module

Figure 8 shows the overall design of the HITL module and its relationship to the foundational LVM model and downstream network. The HITL module contains a frontend and a backend. Because the frontend was constructed using HTML and CSS, training the AI model to label the images can be performed by clicking on the images (left panel in Fig. 9). These actions are then passed to the backend, which was constructed using the Flask framework and which communicates between the HITL module and LVM. In addition, a user interface based on the Jupyter Notebook constructed for this purpose is available for those who do not run web applications (right panel in Fig. 9).

[1] https://www.legacysurvey.org/

[2] https://sites.google.com/usfca.edu/neuralens/

[3] https://flask.palletsprojects.com/

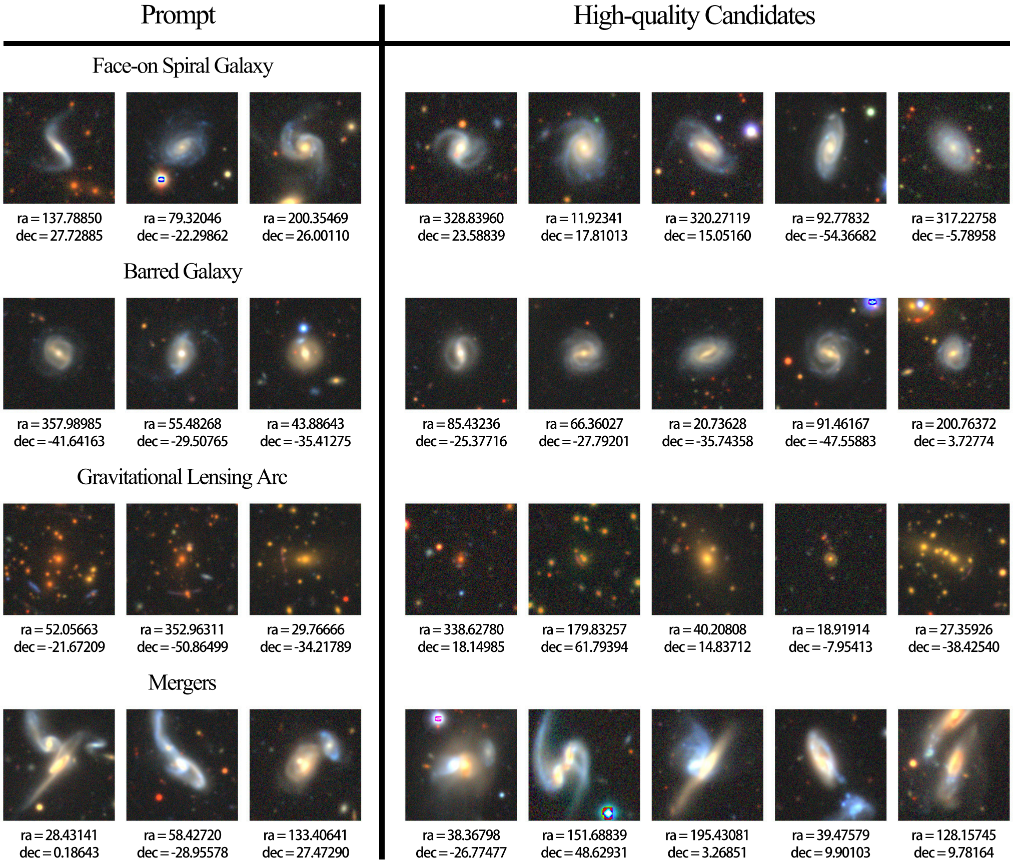

Prompt
High-quality Candidates
Face-on Spiral Galaxy
ra = 137.78850 dec = 27.72885
ra = 79.32046 dec = -22.29862
ra = 200.35469 dec = 26.00110
ra = 328.83960 dec = 23.58839
ra = 11.92341 dec = 17.81013
ra = 320.27119 dec = 15.05160
ra = 92.77832 dec = -54.36682
ra = 317.22758 dec = -5.78958
Barred Galaxy
ra = 357.98985 dec = -41.64163
ra = 55.48268 dec = -29.50765
ra = 43.88643 dec = -35.41275
ra = 85.43236 dec = -25.37716
ra = 66.36027 dec = -27.79201
ra = 20.73628 dec = -35.74358
ra = 91.46167 dec = -47.55883
ra = 200.76372 dec = 3.72774
Gravitational Lensing Arc
ra = 52.05663 dec = -21.67209
ra = 352.96311 dec = -50.86499
ra = 29.76666 dec = -34.21789
ra = 338.62780 dec = 18.14985
ra = 179.83257 dec = 61.79394
ra = 40.20808 dec = 14.83712
ra = 18.91914 dec = -7.95413
ra = 27.35926 dec = -38.42540
Mergers
ra = 28.43141 dec = 0.18643
ra = 58.42720 dec = -28.95578
ra = 133.40641 dec = 27.47290
ra = 38.36798 dec = -26.77477
ra = 151.68839 dec = 48.62931
ra = 195.43081 dec = 3.26851
ra = 39.47579 dec = 9.90103
ra = 128.15745 dec = 9.78164

Fig. 7. (color online) Input prompts (left columns) and objects recommended (right columns) by the HITL after several rounds of interaction. For simple tasks such as identifying face-on spirals, one round was enough, and for more complex tasks such as identifying mergers, no more than 10 rounds were used.

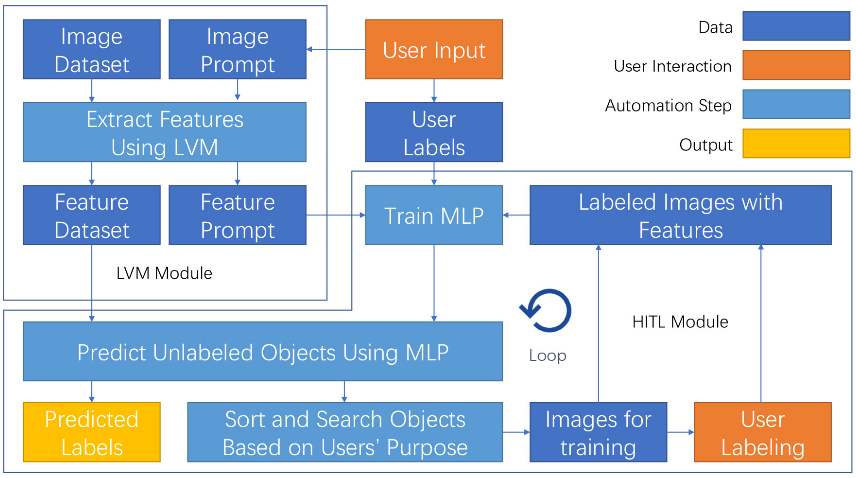
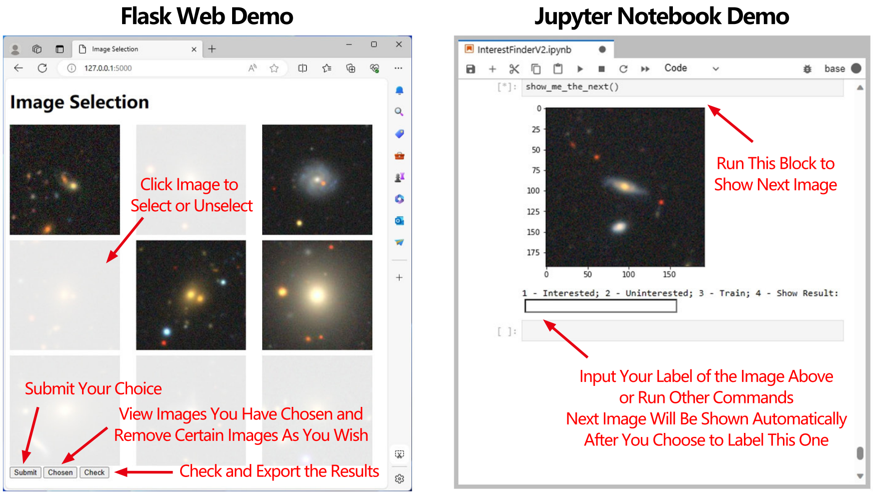


Fig. 8. (color online) Overall design of the human-in-the-loop (HITL) module.

The downstream network is an MLP with a hidden layer of size 2048 running on fixed features extracted by our LVM. This setup resulted in a significantly lower computational cost compared to training a model for a specific task from scratch, and also proved adequate for optimizing the capabilities of our LVM. To maximize the benefits of this interaction to the model for various purposes, we set a parameter as $0 \leq \alpha \leq 1$, which represents a threshold of the ratio between positives and negatives labeled during the latest interaction loop (P/N). When



```
[*]: show_me_the_next()
1 - Interested; 2 - Uninterested; 3 - Train; 4 - Show Result:
[ ]:
```

Click Image to Select or Unselect

Submit Your Choice

View Images You Have Chosen and Remove Certain Images As You Wish

Check and Export the Results

Run This Block to Show Next Image

Input Your Label of the Image Above or Run Other Commands Next Image Will Be Shown Automatically After You Choose to Label This One

Fig. 9. (color online) Interface for our HITL classification web demo (left panel) and Jupyter notebook demo (right panel). In the web demo, users can classify images by clicking, and can also view and export classification results. In the Jupyter notebook demo, users can perform the same actions by viewing images and executing appropriate commands.

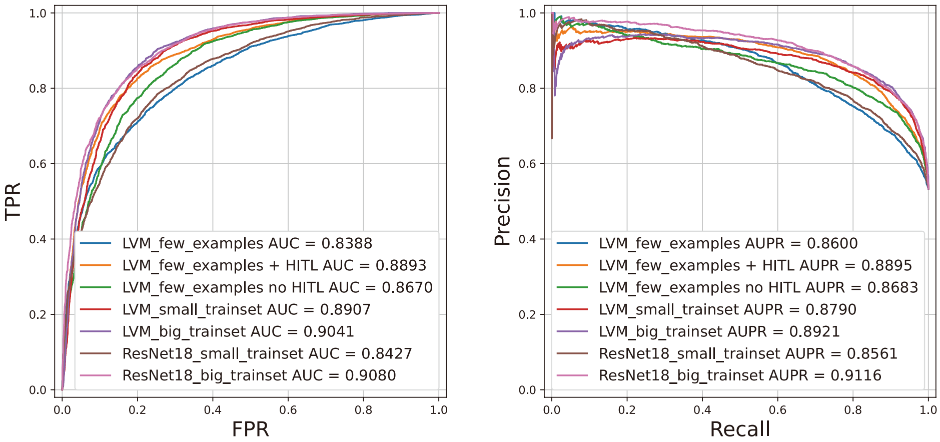


Fig. 10. (color online) Comparison between the performances of ResNet18, our LVM, and our LVM + HITL using different training sets with the same number of images.

$P/N \le \alpha$, the HITL module selects objects with higher scores for user labeling; when $P/N > \alpha$, the HITL module selects objects with lower scores for user labeling. Changing α can guide the downstream model to converge in the required directions. For example, $\alpha = 0.9$ and $\alpha = 0.1$ generate models with high precision and recall rates, respectively, while $\alpha = 0.5$ produces a model with high area under the curve (AUC) and AUPR values.

B. Comparison to conventional models of supervised learning

To evaluate the effectiveness of our HITL module, the classification results for 8522 galaxy images from Galaxy Zoo DECaLS [41] using different approaches (which were not included in the training sets used to train our LVM) were compared with the LVM + HITL and traditional supervised learning approaches. These galaxies were classified as face-on spirals (4546 images) or other galaxies (3976 images, including ellipticals and edge-on galaxies) [48]. We tested the performance of supervised ResNet18[1)], our LVM, and our LVM + HITL with training datasets of different sizes.

Specifically, we first gave five positives and five negatives to the downstream classification network following the LVM, and the outcomes were named *LVM_few_examples* in Fig. 10. Then, we labeled an extra 10 examples recommended by the HITL module and observed a boost in performance ("*LVM_few_examples +HITLHITL*" in Fig. 10). In addition, 15 positive and 15 negative examples fed to our LVM ("*LVM_few_examples* no HITL" in Fig. 10), 50 positive and 50 negative examples fed to our LVM and ResNet18 (*small_*

1) https://pytorch.org/vision/stable/index.html

training_set in Fig. 10), and 1000 randomly chosen examples fed to our LVM and ResNet18 (*big_training_set* in Fig. 10) were evaluated for comparison. As indicated, the performance of the LVM + HITL approach was better than that of the LVM alone and similar to that of the *big_training_set*.

C. Discovering targets in the DESI Bright Galaxy Survey using LVM + HITL

To test the feasibility of the LVM + HITL approach on real observations, we constructed serial tasks for object detection in 201319 galaxy images selected from the DESI Bright Galaxy Survey (BGS) [49], which is a selection of bright galaxies in the DESI Legacy Imaging Surveys (this selection was excluded from the training sets for the LVM). These galaxies have half-light radii between 6.4 and 9.6 arcsec. To maintain wide image margins, the fits files of the galaxies in the g, r, and z channels were cropped into images of size $(H, W, C) =$ (192, 192, 3) with a pixel scale of 0.262 arcsec/pix. This strategy is beneficial for identifying objects such as gravitational lensing arcs and mergers.

First, two characteristic types of galaxies were selected for target-finding experiments: face-on spiral galaxies and barred galaxies (relatively common objects). Starting with only five positives, our LVM + HITL achieved precision rates of 0.91 and 0.75 when identifying face-on spiral galaxies and barred galaxies, respectively, within 10 rounds of interactions. These findings demonstrated that our LVM + HITL method can assist astronomers in identifying their targets for specific scientific goals with a reference sample containing only a few objects.

Moreover, the LVM + HITL model was utilized to identify strong gravitational lensing systems and galaxy mergers to examine its feasibility in searching for rare and complex astronomical objects. Figure 7 shows that our approach can discover these targets successfully. However, the outcomes included many more false positives than the tasks that aimed to identify common objects (e.g., the precision rate of finding galaxy mergers was only 0.15). In principle, this issue can be improved by adopting an appropriate method in the LVM for handling the feedback from the HITL module beyond depending on α alone, which will be a primary focus of our future investigations.

VI. SUMMARY

In this study, we created a framework that utilized a HITL module on top of an LVM for various astronomical vision tasks. The downstream neural networks, combined with the LVM, allowed for versatility without the need for expensive re-training. Furthermore, the HITL module incorporated human knowledge to guide the AI model toward specific objectives, which reduced the workload for composing training sets and enhanced the framework's universality and interpretability. The experiments showed that our framework outperformed traditional supervised machine learning models in classical vision tasks in astronomy, such as object detection, galaxy morphological classification, and observational image reconstruction. Considering that the reliability of AI models in handling scientific data is crucial for valid discoveries [6, 50, 51], we evaluated our framework's reliability through different experiments using labels in the Galaxy Zoo 2 datasets. However, for data in the bands other than g, r, and z and those provided by space-borne telescopes, Galaxy Zoo 2 was insufficient. Therefore, to assess the framework's reliability in a broader context in the future, we are planning on constructing a standard dataset of galaxy images covering a larger feature space from various observations. Using the transfer learning strategy, we plan on extending the framework to encompass various data modalities, including photometry, spectra, and lightcurves. This will lead to a continually evolving AI model that can proficiently handle intricate datasets from a variety of major observing projects (e.g., DESI, LSST, Euclid, and CSST), which is crucial in the age of multi-messenger astronomy.

ACKNOWLEDGMENTS

The authors thank George Stein for publicly sharing the cutouts of galaxy images from DESI Legacy Surveys online.

The DESI Legacy Imaging Surveys consist of three individual and complementary projects: the Dark Energy Camera Legacy Survey (DECaLS), the Beijing-Arizona Sky Survey (BASS), and the Mayall z-band Legacy Survey (MzLS). DECaLS, BASS and MzLS together include data obtained, respectively, at the Blanco telescope, Cerro Tololo Inter-American Observatory, NSF's NOIRLab; the Bok telescope, Steward Observatory, University of Arizona; and the Mayall telescope, Kitt Peak National Observatory, NOIRLab. NOIRLab is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation. Pipeline processing and analyses of the data were supported by NOIRLab and the Lawrence Berkeley National Laboratory (LBNL). Legacy Surveys also uses data products from the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE), a project of the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. Legacy Surveys was supported by: the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility; the U.S. National

Science Foundation, Division of Astronomical Sciences; the National Astronomical Observatories of China, the Chinese Academy of Sciences and the Chinese National Natural Science Foundation. LBNL is managed by the Regents of the University of California under contract to the U.S. Department of Energy. The complete acknowledgments can be found here[1]*.*

References

[1] M. Agarwal, J. Alameda, J. Audenaert *et al.*, (2023), arXiv: 2306.08106

[2] S. Li and A. Adelmann, Phys. Rev. Accel. Beams **26**, 024801 (2023)

[3] A. Boehnlein, M. Diefenthaler, N. Sato *et al.*, Rev. Mod. Phys. **94**, 031003 (2022)

[4] R. Suresh, H. Bishnoi, A. V. Kuklin *et al.*, Frontiers in Physics **12**, 1322162 (2024)

[5] Y. Zhang and Y. Zhao, Data Science Journal **14**, 11 (2015)

[6] M. Huertas-Company and F. Lanusse, Publications of the Astronomical Society of Australia **40**, (2023), arXiv:2210.01813

[7] B. Lao, T. An, A. Wang *et al.*, Science bulletin **66**, 2145 (2021)

[8] M. Banerji, O. Lahav, C. J. Lintott *et al.*, Monthly Notices of the Royal Astronomical Society **406**, 342 (2010)

[9] C. Wu, O. I. Wong, L. Rudnick *et al.*, Monthly Notices of the Royal Astronomical Society **482**, 1211 (2019)

[10] Y. B. Li, A. L. Luo, C. D. Du *et al.*, The Astrophysical Journal Supplement Series **234**, 31 (2018)

[11] J. Xu, Q. Yin, P. Guo *et al.*, Monthly Notices of the Royal Astronomical Society **499**, 1972 (2020)

[12] G. Martin, S. Kaviraj, A. Hocking *et al.*, Monthly Notices of the Royal Astronomical Society **491**, 1408 (2020)

[13] C. Logan and S. Fotopoulou, Astronomy & Astrophysics **633**, A154 (2020), arXiv:1911.05107

[14] Q. Xu, S. Shen, R. S. de Souza *et al.*, Monthly Notices of the Royal Astronomical Society **526**, 6391 (2023)

[15] M. A. Hayat, G. Stein, P. Harrington *et al.*, The Astrophysical Journal Letters **911**, L33 (2021), arXiv:2012.13083

[16] P. Jia, R. Sun, W. Wang *et al.*, Monthly Notices of the Royal Astronomical Society **470**, 1950 (2017)

[17] W. Wang, P. Jia, D. Cai *et al.*, Monthly Notices of the Royal Astronomical Society **478**, 5671 (2018)

[18] S. Ni, Y. Li, L. Y. Gao *et al.*, The Astrophysical Journal **934**, 83 (2022), arXiv:2204.02780

[19] L. Y. Gao, Y. Li, S. Ni *et al.*, Monthly Notices of the Royal Astronomical Society **525**, 5278 (2023), arXiv:2212.08773

[20] P. Jia, Q. Jia, T. Jiang *et al.*, The Astronomical Journal **165**, 233 (2023)

[21] P. Jia, Q. Jia, T. Jiang *et al.*, Astronomy and Computing 100732 (2023)

[22] Z. Liu, Y. Lin, Y. Cao *et al.*, in *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021)

[23] G. Stein, P. Harrington, J. Blaum *et al.*, (2021), arXiv: 2110.13151

[24] A. Dey, D. J. Schlegel, D. Lang *et al.*, The Astronomical Journal **157**, 168 (2019)

[25] M. Grinberg, *Flask web development: developing web applications with python*, (O'Reilly Media, Inc. 2018)

[26] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015)

[27] J. Devlin, M. W. Chang, K. Lee *et al.*, (2018), arXiv: 1810.04805

[28] Z. Dai, Z. Yang, Y. Yang *et al.*, (2019), arXiv: 1901.02860

[29] T. Brown, B. Mann, N. Ryder *et al.*, Advances in neural information processing systems **33**, 1877 (2020)

[30] H. Touvron, T. Lavril, G. Izacard *et al.*, (2023), arXiv: 2302.13971

[31] A. Kirillov, E. Mintun, N. Ravi *et al.*, (2023), arXiv: 2304.02643

[32] F. Lanusse, L. Parker, S. Golkar *et al.*, (2023), arXiv: 2310.03024

[33] G. Stein, J. Blaum, P. Harrington *et al.*, The Astrophysical Journal **932**, 107 (2022), arXiv:2110.00023

[34] C. M. Fan, T. J. Liu, and K. H. Liu, in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2333–2337, (IEEE2022)

[35] K. H. R. Chan, Y. Yu, C. You *et al.*, J. Mach. Learn. Res. 23, 1 (2022)

[36] K. He, X. Chen, S. Xie *et al.*, *Masked autoencoders are scalable vision learners (2021)*, arXiv: 2111.06377

[37] Z. Xie, Z. Zhang, Y. Cao *et al.* (2022), arXiv: 2111.09886

[38] G. Bradski, Dr. Dobb's Journal of Software Tools (2000).

[39] X. P. Zhu, J. M. Dai, C. J. Bian *et al.*, Astrophysics and Space Science **364**, 1 (2019)

[40] K. Schawinski, C. Zhang, H. Zhang *et al.*, Monthly Notices of the Royal Astronomical Society: Letters **467**, L110 (2017)

[41] M. Walmsley, C. Lintott, T. Géron *et al.*, Monthly Notices of the Royal Astronomical Society **509**, 3966 (2022), arXiv:2102.08414

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Communications of the ACM **60**, 84 (2017)

[43] K. Simonyan and A. Zisserman, (2014), arXiv: 1409.1556

[44] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016)

[45] K. He, G. Gkioxari, P. Dollár *et al.*, in *Proceedings of the IEEE international conference on computer vision*, 2961–2969 (2017)

[46] X. Wu, L. Xiao, Y. Sun *et al.*, Future Generation Computer Systems **135**, 364 (2022)

[47] D. E. Rumelhart and J. L. McClelland, *Learning Internal Representations by Error Propagation*, 318–362 (1987)

[48] Z. Zhang, Z. Zou, N. Li *et al.*, Research in Astronomy and Astrophysics **22**, 055002 (2022), arXiv:2202.08172

[49] C. Hahn, M. J. Wilson, O. Ruiz-Macias *et al.*, The Astronomical Journal **165**, 253 (2023), arXiv: 2208.08512

[50] Y. Du, L. Cui, X. Guan *et al.*, Physics **53**, 147 (2024)

[51] P. Martinez-Azcona, A. Kundu, A. del Campo *et al.*, Phys. Rev. Lett. **131**, 160202 (2023)

[1] https://www.legacysurvey.org/acknowledgment/