Lawrence Berkeley National Laboratory
LBL Publications

Title
Mining for Strong Gravitational Lenses with Self-supervised Learning

Permalink

Journal

ISSN

Authors
Stein, George
Blaum, Jacqueline
Harrington, Peter
et al.

Publication Date

DOI

Peer reviewed

Mining for Strong Gravitational Lenses with Self-supervised Learning

George Stein[1,2] , Jacqueline Blaum[2,3], Peter Harrington[2], Tomislav Medan[2], and Zarija Lukić[2]

[1] Berkeley Center for Cosmological Physics, University of California, Berkeley, Berkeley, CA 94720, USA; gstein@berkeley.edu
[2] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; jrblaum@berkeley.edu, pharrington@lbl.gov, tmedan@lbl.gov, zarija@lbl.gov
[3] Department of Astronomy, University of California, Berkeley, Berkeley, CA 94720, USA

Abstract

We employ self-supervised representation learning to distill information from 76 million galaxy images from the Dark Energy Spectroscopic Instrument Legacy Imaging Surveys' Data Release 9. Targeting the identification of new strong gravitational lens candidates, we first create a rapid similarity search tool to discover new strong lenses given only a single labeled example. We then show how training a simple linear classifier on the self-supervised representations, requiring only a few minutes on a CPU, can automatically classify strong lenses with great efficiency. We present 1192 new strong lens candidates that we identified through a brief visual identification campaign and release an interactive web-based similarity search tool and the top network predictions to facilitate crowd-sourcing rapid discovery of additional strong gravitational lenses and other rare objects: github.com/georgestein/ssl-legacysurvey.

Unified Astronomy Thesaurus concepts: Sky surveys (1464); Strong gravitational lensing (1643); Convolutional neural networks (1938)

1. Introduction

The identification of rare objects is essential to facilitate a number of studies in astronomy. In recent years, the rapid advance of digital sky surveys has increased the size and complexity of data at an ever-increasing pace, giving researchers more opportunities to discover and analyze unique objects. However, this opportunity comes at the cost of an enormous amount of data to search through, such that relying on manual inspection of imagery by experts has long been rendered obsolete. Crowd-sourced classification campaigns such as the Galaxy Zoo projects (Lintott et al. 2008; Willett et al. 2013) allow for the visual inspection of a relatively large number of galaxies ($\mathcal{O}(10^5)$-$\mathcal{O}(10^6)$), but this number still pales in comparison to the full extent of current and incoming surveys. For example, the latest data release of the Dark Energy Spectroscopic Instrument (DESI) Legacy Imaging Surveys (Dey et al. 2019) includes $\mathcal{O}(10^9)$ galaxies, a number far out of reach of even the most ambitious crowd-sourcing effort.

While visual inspection of the full data set to find rare objects is unattainable, the recent success of deep convolutional neural networks (CNNs; LeCun et al. 2015) for a wide variety of supervised classification and regression tasks in computer vision has been shown to be invaluable in conjunction with crowd-sourced astronomical efforts. Given a sufficient sample of human-labeled galaxies, such networks can be trained in a supervised setting to classify images with remarkable accuracy, which has led to their widespread adoption for sky survey science (Huang et al. 2020; Walmsley et al. 2020; Vega-Ferrero et al. 2021). Although powerful, this approach is not without issues, as it directly relies on the quantity and quality of the (manually assigned) labels, which tend to be from "more interesting" galaxies. As such, these labels are biased toward bright and large objects and generally targeted toward more common galaxy morphologies. In particular, supervised CNN classification can prove difficult when the number of known objects of a particular class is very small, as the quality of the supervised network degrades considerably without a sufficient number of labeled samples to learn from (Hayat et al. 2021).

One important class of rare objects is strongly gravitationally lensed galaxy systems, in which light from background sources is lensed by an intervening foreground galaxy or galaxy cluster, magnifying and distorting the appearance of the background source. As the lensing is a function of gravitational potential, its strength is dominated by the dark matter content of the lens and thus can be used to study the dark matter distribution of galaxies (Gilman et al. 2019). Additionally, the magnification of background sources allows for the study of high-redshift galaxy features that would otherwise fall below the detection threshold of astronomical instruments (Caputi et al. 2021). When multiple images of the same background source are observed, the time delay of unique features traversing different path lengths can provide constraints on the expansion of the universe (Wong et al. 2020). We refer the reader to Treu (2010) for a comprehensive review of the subject.

As a result of the scientific value of strong lenses, a large targeted effort has been undertaken to search for them in existing sky survey imaging data sets. The primary challenge in doing so is their scarcity, as we expect only about one in $\mathcal{O}(10^4)$ galaxies to be strongly lensed (Collett 2015). In recent years, the search for strong lenses has been dominated by the use of CNNs trained in a standard supervised learning setup (e.g., Jacobs et al. 2017, 2019a, 2019b; Petrillo et al. 2017, 2019; Lanusse et al. 2018; Cañameras et al. 2020, 2021; He et al. 2020; Huang et al. 2020, 2021; Li et al. 2020; Gentile et al. 2022).[4] After assembling the (small) sample of already known strong lenses (or a set of simulated strong lenses) and selecting a number of galaxy images that are not lensed, the CNN is trained in a binary classification setting on the

---

[4] See the lens-finding challenge (Metcalf et al. 2019) or Stein (2020) for additional examples.

training set until the performance on a holdout validation set is optimized. These studies have succeeded in discovering a number of new strong lens systems, but even after compiling all known lenses to date across all surveys, the number of available training samples is a major limiting factor in achieving accurate predictions with this approach.

The lack of quality training labels coupled with the enormous quantity of data coming from current and near-future sky surveys motivates the need for scalable approaches capable of learning from unlabeled data. Rather than training specialized networks for each individual classification or regression task, learning without labels allows for more generalized models capable of performing multiple tasks in parallel while simultaneously benefiting from access to orders of magnitude more data. A preliminary application of unsupervised methods to lens identification was done by Cheng et al. (2020), who applied a convolutional autoencoder to simulated data and clustered the resulting features to classify lenses. An exciting alternative to this purely unsupervised feature learning is the emerging paradigm of self-supervised learning, which aims to distill or extract useful information from images without requiring supervision labels for each sample in the data set (Zhai et al. 2019). Self-supervised methods can make use of very large unlabeled data sets (e.g., Goyal et al. 2021), which would be cumbersome or otherwise impossible to manually label, and build meaningful embeddings or representations by solving contrived tasks during training that require some high-level understanding of the input features. The representations produced by self-supervised models can then be directly used or fine-tuned for machine learning tasks of interest, and this approach is able to outperform standard supervised machine learning models, particularly when the number of data labels available for supervision is low (Chen et al. 2020a, 2020b, 2020c). Compared to simpler unsupervised feature learners like autoencoders, self-supervised models are known to produce higher-quality representations with far more predictive power (Donahue et al. 2016; Donahue & Simonyan 2019) that generalize well to many different tasks (Caron et al. 2021).

The utility of self-supervised models applied to astronomical imagery was recently showcased by Hayat et al. (2021), who found that self-supervised pretraining on a large unlabeled set of Sloan Digital Sky Survey galaxy images improves performance on tasks like redshift estimation and morphology classification, and that these performance gains are most significant when the number of labels for supervised training is limited. Hayat et al. (2021) also showed that the representation space learned in self-supervised pretraining is semantically meaningful and readily provides a similarity metric that can identify additional examples of a query object, such as an observational error or anomalous galaxy. As automated strong lens detection is a problem inherently limited by the number of labeled examples, self-supervised models are thus an exciting prospect for quickly identifying new candidates given a large set of galaxy imagery.

In this work, we apply self-supervised learning to Data Release 9 (DR9) of the DESI Legacy Surveys (Dey et al. 2019) to conduct a search for strong gravitational lenses. Strong gravitational lensing can result in a number of different visual configurations, including multiple appearances of the same background galaxy or extended arcs curving toward the central source, but determining whether the configuration is a true

gravitational lens by visual inspection can be difficult. Therefore, unless the system is spectroscopically confirmed (or, in the case of faint lens features, observed at higher resolutions using other instruments), we often cannot use images to conclusively determine whether or not a system is truly a gravitational lens. As such, in this work, we consider strong lens candidates: images that contain characteristics of strong gravitational lensing as viewed in the DESI Legacy Survey imaging. In Section 2, we first give an overview of the images and strong lens labels used, followed by a description of our self-supervised model and supervised classifiers in Section 3. In Section 4, we detail our discovery of 1192 new strong gravitational lens candidates and present our publicly available similarity search tool to facilitate the discovery of additional gravitational lenses beyond those presented here. Lastly, we conclude in Section 5.

2. Data Set

2.1. DESI Legacy Survey

The DESI Legacy Imaging Surveys[5] (Dey et al. 2019) cover approximately 19,721 deg^2 of extragalactic sky visible from the northern hemisphere in three optical bands (g, r, and z). The survey footprint is split into two components, north and south. The observations in the north region, $\delta > 32.375$, were undertaken jointly by the Beijing-Arizona Sky Survey (BASS; g and r bands) and the Mayall z-band Legacy Survey (MzLS; z band), while the southern part of the survey, $\delta < 32.375$, uses data from the DECam Legacy Survey (DECaLS). The DECaLS program made use of other DECam data, the most significant of which are from the Dark Energy Survey (DES), which includes a 5000 deg^2 contiguous area in the south Galactic cap. DECaLS explicitly did not reimage the DES area and instead incorporated the DES imaging itself. We use the latest data release, DR9, released in 2021 January. The DR9 does not contain significant new observations but builds on DR8 by improving the reduction techniques and procedures used for the Legacy Surveys. Imaging from the Legacy Surveys is first reduced through the NOIRLab Community Pipeline[6] before being processed using the Tractor.[7] The Tractor identifies unique sources in the imaging survey, measures a large number of photometric features, and fits for a number of quantities, including morphological type.

Our self-supervised approach allows us to learn from the entire data set and does not require restrictive data cuts to narrow down objects to those most likely to show features of strong lensing, such as color or magnitude cuts. Rather, we aim to use every galaxy in the DESI Legacy Imaging Surveys. We keep every source in the sweep catalogs (a subset of the information about each source from the full Tractor catalogs) that was not identified as a star, i.e., did not have a best-fit morphological model of point-spread function (PSF) = stellar. For both the north and south surveys, we compile an initial list of all non-PSF sources and keep those with a z-band magnitude (denoted mag$_z$) less than 20. This cut was introduced to keep the data at a more manageable size and to focus on galaxies that have been included in previous lens searches (Huang et al. 2021). For the south survey, we also include the next 20

---

million sources with $\mathrm{mag}_z > 20$, sorted by decreasing brightness.

For each source in the list, we extracted a 256×256 pixel cutout in each of the (g, r, z) bands at $0\rlap{.}''262$ resolution, centered at the R.A. and decl. of the source. A very small fraction of galaxies that existed in the sweep catalogs resulted in an error when requesting the cutout, so these sources were excluded from our final catalogs. In total, this resulted in 14,174,203 images from the north survey and 62,272,646 images from the south, 42,272,646 with a $\mathrm{mag}_z < 20$, for a total of 76,446,849 images. Sources in, e.g., galaxy clusters can be separated by less than the angular extent of an image, so separate images of nearby sources will include overlapping regions of the sky. The vast majority of galaxies and strong lenses cover angular extents much less than the full 256 pixel cutout, so we choose to train all networks on 96×96 pixel $(25'' \times 25'')$ crops. In order to perform the required set of data augmentations during training, mainly random rotations and jitter, we save 152×152 pixel versions to disk, which results in a data set size of 20 TB. Files are saved in hd5f format and split into arrays of 1 million images (i.e., each file contains an image array of size (1,000,000, 3, 152, 152) and the corresponding sweep catalog information for each source).

2.2. Compilation of Previously Found Lenses

We compiled a list of strong lens candidates from previous lens-finding campaigns using all candidates from the following sources.

1. The Master Lens Database[8] as of late 2021 July.
2. Huang et al. (2020, 2021) with tables available on their project website.[9]
3. A number of previous works with tables published through the VizieR database[10] (Carrasco et al. 2017; Diehl et al. 2017; Jacobs et al. 2017, 2019a, 2019b; Pourrahmani et al. 2018; Cañameras et al. 2020).
4. The Survey of Gravitationally lensed Objects in HSC Imaging Candidate List[11] (Sonnenfeld et al. 2018, 2019, 2020; Wong et al. 2018; Chan et al. 2020; Jaelani et al. 2020).

Some of the above works overlap with the Master Lens Database, but we included their data tables separately to ensure all sources were included. We disregard the "quality" labels indicating the respective team's level of confidence that each candidate is actually a lens, as the classification criteria for these labels are inconsistent throughout the literature. During publication of this paper, after the conclusion of our lens search, Rojas et al. (2021) performed an additional supervised CNN search on the DES region, finding 186 new lens candidates, and Li et al. (2021) presented 97 new high-quality strong lensing candidates. We also became aware of another separate repository of known lenses[12] that were already included in our analysis through the Master Lens Database.

To first determine which of the lens candidates overlap with our image data set, we cross-match the R.A. and decl. from our lens candidate list with the approximately 76 million sources in

our DR9 sample and define a match as any lens that has a DR9 source separated by less than the size of our images (96 pixels). For galaxy clusters with multiple sources near the given lens position, we flag the nearest source as "the lens" regardless of whether it is in the center of the cluster. We additionally flag any DR9 sources that reside near a lens, but are not the closest, by noting their distance from the lens. To prevent any ambiguity during training of our networks in Section 3.3, we make sure not to sample sources that happen to lie near any labeled lenses, as the lens features may still appear in the image.

This procedure results in nearly 6000 lens candidates that overlap with sources in DR9, but by our own lens classification criteria, the majority do not appear to have clear lensing signatures in DR9 imaging.[13] This discrepancy is potentially due to limitations in the angular resolution or noise levels of DR9 or unclear classifications of what constitutes a lens candidate in the variety of sources compiled. As we explain in the following section, we train all models only on south sources with $\mathrm{mag}_z < 20$. In this subset, we visually inspect the lens candidates and keep only the ones that, in our opinion, exhibit distinguishable lensing features in DR9 images. We do not include images with very faint or ambiguous features. We refer to this final catalog of 1615 lenses as DECaLS strong lenses, and we use only this sample for training and validating our supervised networks.

We show the distribution of lenses on the sky in Figure 1. As expected, we see a higher density of lenses in the DES region due to the greater imaging depth and its heightened focus by previous lens searches. In addition, given our highly subjective visual selection of often-ambiguous images (which is based on an already-biased sample of previously discovered lenses), our training set has likely over- or undersampled certain lens classes and configurations, which affects the set of newly uncovered lens candidates. Thus, the lens candidates we present in this work should not be considered as a representative sample of all those remaining to be found in sky survey data sets but rather as a complex combination of the known sample of lenses available for training, the biases introduced by the specific subset of humans who examined them for this work, and the inherent biases of classification with highly nonlinear models in a class-imbalanced problem. Rather than assembling a training set of lenses by visual inspection, one could instead learn from a set of simulated lenses (Jacobs et al. 2017; Li et al. 2020), although this procedure does not come without its own set of biases, as the "true" distribution of lenses is still unknown, and the lensing profiles used in the simulations may not perfectly reflect reality. As we are working from a data-driven perspective and have developed a tool to find lenses given a single example, we chose to opt for the set of human-classified lenses as they appear in our imaging in the hope that they include more outlying examples from those more commonly exhibited in lensing systems. However, we note that a set of simulated lenses would allow for very similar investigations.

3. Methods

In this section, we describe the machine learning methods used to sift through our galaxy catalogs and detect new strong
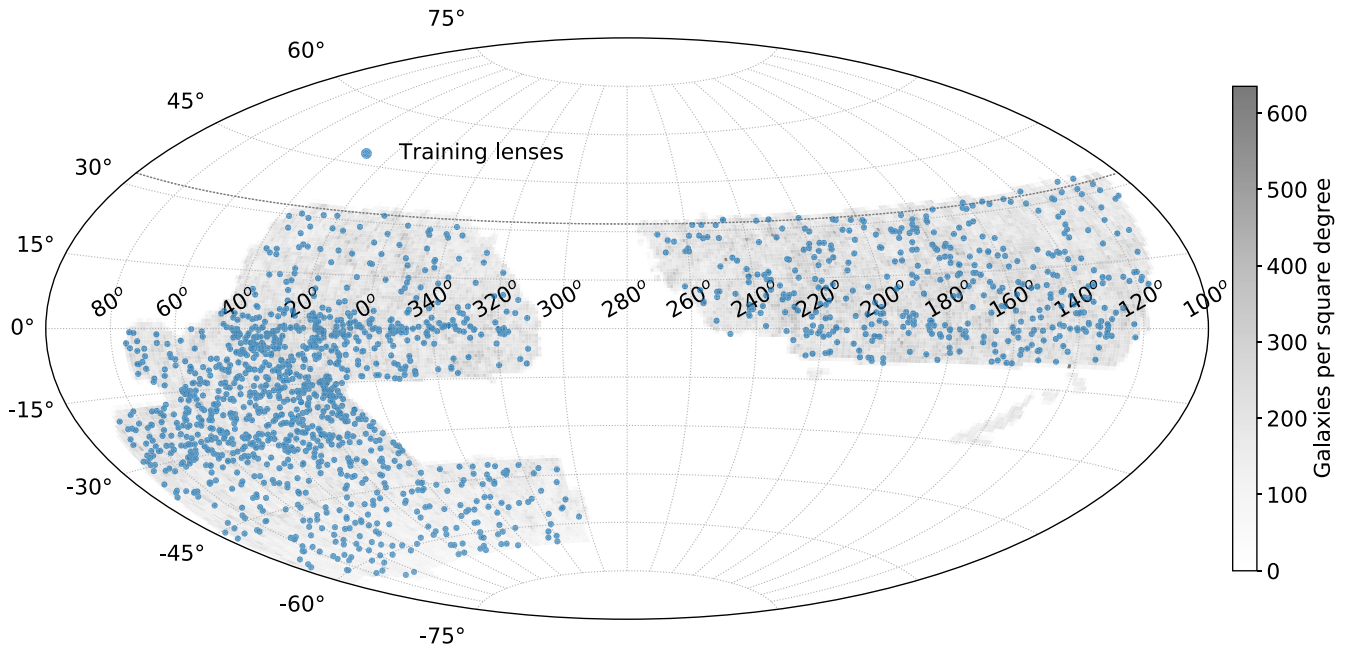

---

[13] We examine images after converting from nanomaggies to RGB values using the same transformation as the legacy survey viewer.

Figure 1. The R.A. and decl. of previously identified DECaLS strong lens candidates, with the background shaded by the number of the 3,500,000 galaxies used to train the self-supervised network in each square-degree region. The gray line at $\delta = 32.375$ denotes the split between the north and south surveys.
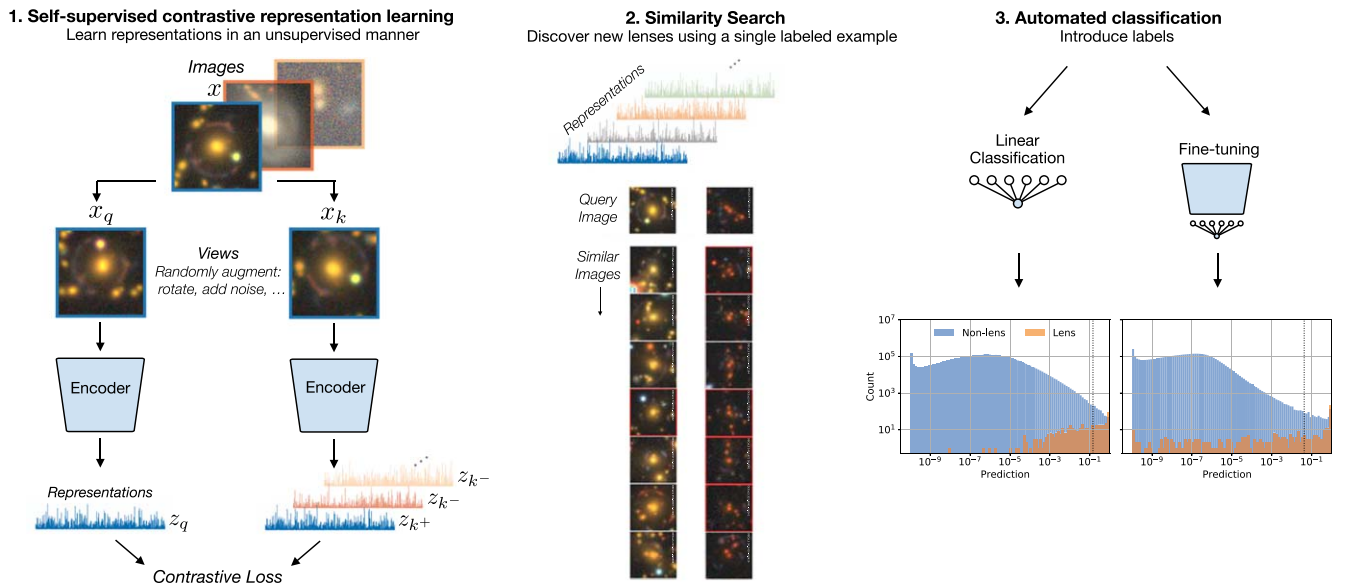
Figure 2. Illustration of self-supervised training (left; see Hayat et al. 2021) and its use cases for discovering objects through similarity search (middle) and automated classification (right).

lens candidates. We provide a schematic of our workflow in Figure 2. The first step in our process is to train a self-supervised model on a large body of unlabeled data, allowing the network to see many types of galaxies and learn from the natural distribution of morphologies and features within the survey sample. Then, we can take the representations learned by the self-supervised model and use them to identify new lenses via similarity searches and automated classification.

3.1. Self-supervised Pretraining on Unlabeled Images

Modern self-supervised learning has been shown to distill highly information-rich representations from standard data sets without utilizing any labeled information, with the classification performance of a simple linear classifier trained on such representations rivaling the performance of fully supervised CNN models (Chen et al. 2020a; Grill et al. 2020; He et al. 2020).[14] Such data sets contain a ground-truth label for every image, allowing for robust measurements of the quality of the learned representations, as better representations will lead to increased performance on downstream supervised tasks using these labels. However, for many scientific data sets, especially sky surveys, the number of ground-truth labels is minimal compared to the number of images, and the labels themselves often have nonnegligible noise or bias. Thus, assessing representation quality via robust downstream supervised evaluations is much more

---

[14] See self-supervised ImageNet linear classification accuracy (https://paperswithcode.com/sota/self-supervised-image-classification-on) by publication date.

challenging, so tuning hyperparameters and choosing data augmentations for the self-supervised model can be difficult.

Rather than supervised evaluation, Reed et al. (2020) found that the evaluation on an additional self-supervised task (e.g., rotating images and predicting the rotation from the representations) was strongly correlated with the supervised evaluations and could be used in place of supervised labels to determine representation quality. Unfortunately, their strongest results were for a rotation evaluation task, which is incompatible with the rotation-invariant representations of galaxies we desire in this study (as there is no preferential orientation of galaxies on the sky). A jigsaw task (Noroozi & Favaro 2016) may prove useful for our data set, but we leave this for future work and instead evaluate the quality of our self-supervised models according to their linear classification performance on our validation set of strong lenses.

To limit the amount of hyperparameter tuning, we closely follow Hayat et al. (2021) in designing the architecture and training procedure for our self-supervised model. In this setting, the backbone of the model is a CNN encoder that takes an image x as input and produces a lower-dimensional representation z. The encoder learns to make meaningful representations by associating augmented views of the same image as similar and views of different images as dissimilar via a contrastive loss function. The representations for different images are maintained in a queue during training, which extends the number of contrasting examples available to the model at each training step beyond just those available in a given minibatch. More details on this approach can be found in Chen et al. (2020c) and Hayat et al. (2021). We use the same ResNet50 network and training hyperparameters as Hayat et al. (2021) but increase the queue length to $K = 262{,}144$ to accommodate our larger training set. As our encoder architecture is a standard ResNet50, the number and size of the convolutional kernels and pooling layers results in a representation vector of 2048 dimensions. This is strictly a result of the encoder architecture used, and while a representation of this size has been shown to perform well on downstream tasks on common curated industry data sets, it is an open question as to whether it is an optimal size for sky survey data sets, which often exhibit a smaller degree of image diversity.

The key to successful self-supervised pretraining is to produce the differing views of each image using a carefully crafted set of image augmentations that reflect symmetries, uncertainties, or noise in the data set. These augmentations must realistically perturb the images, making it nontrivial for the model to associate augmented pairs of the same image while still preserving the important features in the input. We choose the following set of augmentations, applying each of them in succession to images during pretraining in the order listed below.

1. *Galactic extinction.* We first deredden the image according to its tabulated $E(B - V)$ value (Schlegel et al. 1998), simulating a view of the galaxy with no foreground dust. Then, we randomly sample a new $E(B - V)$ value from a lognormal distribution fit to the data set and artificially redden the image with this random $E(B - V)$. Details on the lognormal $E(B - V)$ distribution and photometric conversions between $E(B - V)$ and transmission can be found in Appendix B.

2. *Rotation/orientation.* We randomly flip the image across each axis with 50% probability, then rotate by a random angle sampled from $\mathcal{U}(0, 2\pi)$.

3. *Size scaling.* We randomly resize the image to between 90% and 110% of its original size to simulate views of a galaxy at different distances. We do not account for the change in redshift or resolution corresponding to such changes in distance and simply rescale with bilinear interpolation.

4. *PSF blur.* We model additional PSF smoothing (on top of the existing inherent PSF smoothing in the image) by applying Gaussian blur to each channel. The amount of blurring (i.e., the width σ of the blurring kernel) is parameterized by lognormal fits to the PSF distribution found in the data (see Appendix B for details).

5. *Jitter and crop.* We translate the image center along each respective axis by i, j pixels, where $i, j \sim \mathcal{U}(-7, 7)$, before cropping out the central 96×96 pixels.

6. *Gaussian noise.* We add additional Gaussian noise to the image (on top of the inherent noise already present) by sampling a noise level from lognormal distributions tuned for each filter channel. The noise in each channel is uncorrelated, since images are taken at different times and/or with different telescopes. See Appendix B for details on how each channel's noise distribution is parameterized.

With these augmentations, we perform self-supervised pretraining on our training set, which was constructed with the following considerations.

A significant number of the full 76 million images are faint red galaxies with a spatial size in the image spanning no more than a few pixels. This dominant class of objects would introduce a high degree of degeneracy due to the similarity between samples, which would limit the effectiveness of the contrastive loss used in self-supervised training. Thus, to ensure a more diverse sample of galaxy images, we choose to train on sources with $\mathrm{mag}_z < 20$ and sample our training set equally in mag_z rather than randomly sampling over the full data set. The magnitude cut (aligning with that chosen by Huang et al. 2020) removes the dimmest galaxies, which exhibit the highest degree of visual similarity, and the uniform mag_z sampling helps to ensure that the training set is not dominated by the vastly more numerous dim galaxies. Although the brightest and most nearby galaxies will not constitute strong lenses, we do not use a minimum magnitude cut to remove them from the self-supervised training set; the goal at this stage is only to learn task-agnostic galaxy representations from a diverse set of galaxy images that can later be used for a variety of different downstream tasks. We show the magnitude distribution of the training sample compared to the full sample in Appendix B.

In addition to uniform magnitude sampling, we train only on images from the south region of the survey. The north and south surveys were observed with different telescopes and bandpasses; therefore, images from these regions have different statistical properties. South images have the best signal-to-noise ratio and angular resolution, and lenses and sharp features are easier to distinguish. We posit that training using this higher-quality sample and then simply using the models on the lower-quality north sample will allow the network to learn a more information-rich representation than training on both surveys simultaneously. Careful crafting of region-specific

augmentations could possibly allow for the combination of these two data samples in a self-supervised learning approach, but we leave this for future work.

These efforts result in a curated subset of 3.5 million images, and we are able to train the model in just 8 hr on eight NVIDIA Tesla V100 GPUs. After the self-supervised pretraining phase, our CNN encoder was applied to all 76,446,849 images from both the south (DECaLS) and north (MzLS/BASS) regions of DR9 to distill features from each image into a representation vector z of length 2048. This procedure extrapolates beyond the training data to both objects of lower magnitude (south images with $mag_z > 20$) and images with inherently different statistical properties (north images), yet we will show that the models nevertheless exhibit strong generalization properties outside of the distribution seen during training. We believe this generalizability simply stems from the high degree of image similarity between $mag_z = 20$ objects and those slightly dimmer than the cutoff and from the overall similarity of the north images compared to the south. The north images exhibit a higher level of noise and slightly different colors (specifically in the z band), and they have the same pixel size and the same range of morphologies, features, and shapes as those in the training set from the south. Although subtle selection effects are possibly introduced by applying our models outside of the training set, we find a significant number of lenses in these regions, thus showing that the models have useful generalization abilities for the applications of this paper.

In the following sections, we describe how these representations are used to search the data set for objects similar to known lenses and then demonstrate their ability to directly reveal new lens candidates via automated classification.

3.2. Image Similarity Search in Representation Space

A powerful advantage of self-supervised pretraining is the ability to provide a measure of semantic similarity between any two images. Given the symmetries and noise inherent to the data set, a notion of similarity is difficult to construct in pixel space, as noise levels and image rotations drastically change the individual pixel values while leaving the semantic information in the image unchanged. Rather than operating in the pixel space, the representations we distill from each image in the data set retain the overall information, such as the size and relative orientation of the galaxies, the number of sources, the clustering, and the color, while removing the noise and symmetries described by the data augmentations (for a visualization of how galaxy properties are organized in the representation space of a self-supervised model, see Hayat et al. 2021). While we find that the representations contain information highly relevant to lens identification, it is important to note that the self-supervised pretraining is task agnostic, and not a single lens label has been used during training. We can utilize the similarity search to uncover additional rare objects of any type, not only strong gravitational lenses.

The contrastive loss in self-supervised pretraining encourages semantically similar images to have similar representations, so we can easily define a similarity metric in representation space (which is a lower dimensionality than the input images). We choose to measure similarity between any two images x_i and x_j by the cosine similarity (normalized scalar product) of their representation vectors z_i and z_j:

$$similarity(z_i, z_j) = z_i \cdot z_j / (\|z_i\| \, \|z_j\|). \quad (1)$$

For each of the 42 million images in the south region of DR9 with $mag_z < 20$, we calculate the 999 most similar images and save the resulting "similarity array" (size $\sim$42 million $\times$ 1000) for later reuse. Computations are performed on eight GPUs using Facebook AI Research's Faiss[15] library for efficient similarity search.[16] To avoid unnecessary computations, we did not compute the full $N \times N$ similarity matrix ($N = 42$ million), but for each source, we limited the search to sources within a magnitude range of 0.5 in the r band. We found that this resulted in nearly equivalent results as the full search, as the model naturally learned that brighter galaxies do not have significant similarity to dimmer ones. The precomputed similarity array allows for rapid similarity searches of any image in the data set; given a query image, the most similar images are immediately available to examine (see Stein et al. 2021 for a discussion of additional use cases for our similarity search tool).

3.3. Automated Classification of Strong Lenses

We perform automated classification of strong lens candidates through two different supervised methods, both of which rely on a pretrained self-supervised model. The first approach is to simply perform linear classification directly on the self-supervised representations, and the second is to fine-tune the self-supervised model for classification.

3.3.1. Training and Validation Sets

As described in Section 2.2, our efforts in compiling a list of known lenses from the literature and cross-matching them with our DR9 galaxy images yielded a total of 1615 well-defined lens images in the south region. We use a random selection of 70% (1130) of these lenses for training and save the remaining 30% (485) for validation. We do not train on any images from the north region or on sources with $mag_z > 20$ and only use these for lens searching. Unlike some previous supervised CNN lens identification works, we do not hand-select the nonlens samples to visually resemble different types of galaxy and cluster morphologies. Hand selection is labor intensive and can artificially inflate the occurrence of specific galaxy types in the training set as compared to the full data set that we will be performing the lens search on. Instead, we simply select nonlens samples randomly from the full data set, excluding those with a lensing galaxy near the central galaxy and those flagged as lenses in other data sets but hard to distinguish visually in DR9 imaging. This approach allows for greater coverage of the survey with minimal manual effort and allows us to build training and validation sets that actually reflect the scarcity of lensed galaxies ($\mathcal{O}(10^{-4})$) expected in the data.

Our random selection of nonlenses from the true data distribution introduces a small amount of class leakage, where undiscovered lenses are mislabeled, but the effects of this leakage are marginal due to the rarity of lenses. To confirm this, we experimented with using an initial linear model to predict

---

[15] https://github.com/facebookresearch/faiss
[16] While we calculated the cosine similarity without employing any of the compression techniques available in Faiss, we found that it still outperformed its NumPy equivalent on CPU and benefited from an additional speed-up on GPU.

labels for the training and validation nonlenses and then visually examined the images with the highest predictions to identify and remove this contamination. However, the level of contamination was small (∼40 of the 4 million "nonlenses" were later found to be lens candidates), and retraining the linear classification model with these more accurate labels returned highly similar predictions, so we kept the original training and validation labels.

To construct a realistic validation set reflecting the scarcity of strong lens populations, we selected 10,000 nonlenses for every lens. This differs from many strong lens searches in the literature, which instead sample validation sets using the same lens-to-nonlens ratios (denoted lens:nonlens) as they selected for their training set (e.g., 1:1, Jacobs et al. 2019b; or 1:33, Huang et al. 2020). Using a more equal ratio will significantly inflate the performance metrics calculated across these validation sets compared to the true data (1:10,000) when undertaking the search for new lenses. The optimal model as measured using a 1:1 validation set is not necessarily optimal on a true 1:10,000 population and is likely to flag a significant number of false positives. Thus, we constructed our validation set to instead represent a likely realistic fraction of lenses in the universe. We treat the lens:nonlens ratio in the training set as a hyperparameter and search over different settings to determine the optimal level of class imbalance. We present our findings for this search in Section 4.2.

3.3.2. Model Evaluation

By searching for rare objects, we are performing a classification task in a scenario with several specific characteristics. Namely, our data set is highly class-imbalanced, and we have limited time to manually assess predicted lens candidates for quality, so we opt for slightly different performance metrics than those canonically used in classification. As our goal is to identify the greatest number of lenses in a limited amount of human time, we decided to restrict our time spent visually inspecting the top network predictions to a few hours. After performing an initial test to see how many images three of the authors could classify per unit time, we estimated our "classification rate" to be ∼10,000 images in a few hours. Thus, if our training set was an unbiased sample of the lenses in the universe, an optimal model would maximize the number of known lenses in the top $N = 10^4$ predictions and minimize the number of false positives; for practical purposes, we are unconcerned with the network performance on the rest of the data set. Our human classification procedure does not include a detailed examination of each image but rather a fast scan of 500 images laid out in a 5×100 grid, devoting approximately 0.5–1 s to each individual image. This procedure leads to a lens candidate selection biased toward more distinct lensing characteristics and configurations and likely misses a number of lower-quality candidates, but it allows for a rapid rate of lens finding.

Based on these initial estimates, we use only the top $N = 1139$ predictions on our validation set to evaluate model quality. This particular value of N was chosen by scaling down the total number of images we expect to visually classify (10^4) to adjust for the difference in size between the validation set (4.85 million galaxies) and our full search size (which we initially scoped to contain ∼42.5 million images in the south survey). This procedure is in contrast to the standard approach of choosing a probability cutoff c_p to manage the trade-off

between precision and recall, as our "cutoff" $c_{\text{top } N}$ is simply determined such that the number of samples above the cutoff is $N = 1139$. Then we can easily assess model quality by computing the precision,

$$\text{Prec}_{(\text{top } N)} = \frac{\text{TP}}{\text{FP} + \text{TP}}, \tag{2}$$

where TP and FP are the number of true positives (the number of true lenses) and false positives (the number of nonlenses) in the top N predictions of the model, respectively. We note that the choice of $N = 1139$ is of little consequence to the chosen model's utility; i.e., we found that the model that maximizes the number of lenses in the top 1139 predictions will be the same model that maximizes the number of lenses in the top 2000 predictions. Thus, for consistency, we choose to fix $N = 1139$ throughout the experiments of this paper regardless of the exact final search size or number of lenses in the final validation set.

3.3.3. Linear Classification

As shown in Hayat et al. (2021) for classification of galaxy morphologies, linear classification on self-supervised representations of images achieves superior performance to full supervised training of a CNN when the amount of labeled data is small. This is despite the fact that the self-supervised representations have only been trained to group similar images together and have not been explicitly trained to pick out properties of any specific objects, like strong gravitational lenses. As we will demonstrate using a similarity search, unlabeled images that are predicted as a high probability of being a lens reside near the labeled samples (i.e., known lenses) in self-supervised representation space. Thus, the 2048-dimensional representation space can be partitioned into binary class predictions (nonlens and lens) by a simple linear classifier of the form $\hat{y} = \boldsymbol{W}^T \boldsymbol{z} + b$, where $\hat{y}$ is the predicted lens probability, $\boldsymbol{z}$ is the representation vector, $\boldsymbol{W}$ is a 2048-dimensional vector of weights, and b is a bias. We also experimented with classifying representations using a random forest classifier but found this more complex nonlinear model to be prone to overfitting, and all configurations tested performed worse on the validation set.

We train our linear classifiers in PyTorch (Paszke et al. 2019) using a binary cross-entropy loss function with LBFGS optimization. Our final model is the epoch that achieves the largest number of true positives in the top 1139 predictions on the validation set, as explained in the previous section. The simplicity of the model and the low dimensionality of the representations allows for training to proceed rapidly on a single CPU in a straightforward single-batch gradient descent, and training on the 4.5 million samples of the 1:4000 training set only requires 5–10 minutes on a CPU (45 s per epoch), compared to many hours on multiple GPUs in a supervised CNN setup. This has the potential for a significant impact beyond the strong lens classifier application demonstrated here. Rather than requiring both the access to computational facilities to host the full image data set and perform multi-GPU training and the machine learning expertise to train complicated CNN models in parallel, self-supervised representation learning reduces the problem to one that can be solved with minimal computing resources and computational background, opening the door to crowd-sourced "citizen science" efforts like those

employed for galaxy morphology studies (e.g., Lintott et al. 2008; Willett et al. 2013).

3.3.4. Fine-tuned Self-supervised Classification

While linear classification on the self-supervised representations achieves high-quality classification results, we can instead attach a fully connected layer on the 2048-dimensional output of the self-supervised encoder CNN and fine-tune the network in a supervised setup using the available lens labels. By applying a smaller learning rate to update the CNN encoder parameters and a larger learning rate on the linear classification head, we can train the linear classifier from a random initialization while only marginally modifying the CNN parameters optimized during the self-supervised learning phase. This can improve the classification results beyond those achievable with linear classification on the original self-supervised representations, but the effect can be small in the few-label regime (Hayat et al. 2021) and can overfit with a small number of labels for a highly class-imbalanced task. However, there is evidence that fine-tuning a self-supervised network improves model robustness over purely supervised learning (Hendrycks et al. 2019).

Works focusing on supervised CNN classification of strong lenses utilize minibatch gradient descent due to memory constraints introduced by large CNN models and image dimensionality, with minibatches of size $\sim$128 images. Each minibatch requires instances of both classes to efficiently learn and cannot be completely dominated by the negative class (nonlens). Therefore, to ensure that each minibatch, on average, has a few lenses, either the class imbalance in the training set must be chosen to be much less than the true data (e.g., 1:1, Jacobs et al. 2019b; or 1:33, Huang et al. 2020) or oversampling techniques that sample lenses more often than their nonlens counterparts must be employed. Our linear classifier, trained with a single large batch, does not have these requirements, and we can easily train with a high degree of class imbalance. To use the same training set as the linear classifier, we oversampled lenses during fine-tuning to ensure a nonzero average number of lenses in each batch. This oversampling factor is an additional hyperparameter of the training setup that we explore during training to optimize the model performance on the validation set.

Training was conducted in PyTorch using a binary cross-entropy loss. We augmented images at each training epoch with jitter/crop, random rotations, Gaussian noise, and Gaussian blur; thus, each sampled version of a given lens will have different realizations of the augmentations. We trained using eight NVIDIA Tesla V100 GPUs for 40 epochs with a batch size of 512 using the SGD optimizer. We implemented a learning rate of 0.01 for the classification head and 0.001 for the remaining network parameters as in Hayat et al. (2021). Training requires approximately 1 hr per epoch. Note that this does not take into account the computing time required for self-supervised pretraining, which does not need to be undertaken separately for each supervised hyperparameter setup.

4. Results

4.1. Identification of Lens Candidates via Similarity Search in Representation Space

To assess our self-supervised model quality and search for new strong lens candidates in the DESI Legacy Surveys, we first employed a number of similarity searches using some of the known lenses identified in previous searches. We demonstrate this approach in Figure 3. For an illustrative example, we select three lenses from the training sample, each of different brightnesses and morphologies, and use the interactive similarity search tool we publicly provide to query for the 1000 most similar images. The first query image is a bright single central source with an extended and nearly complete lens arc, the second is a medium-brightness central cluster producing a small blue double-image arc, and the third is a single dim central galaxy producing an arc. For each query, we visually examine the 512 most similar images, requiring only a few minutes, and flag any strong lens candidates. We find a total of 83 strong lens candidates given only three labeled queries—a significant number, considering we have used only three examples, and our training sample of 1615 lenses constitutes all of the high-quality lenses in the south survey discovered to date. Of the 83 candidates found, to our knowledge, 53 of them have not been previously discovered. Compared to the probability of finding one strong lens in 10,000 images expected from a random sample of galaxies, self-supervised similarity search has immediately increased the lens frequency to $\sim$one in 20 (83:512 $\times$ 3 versus 1:10,000), a factor of 500 speed-up. We found that some candidates were picked up by separate similarity searches of known lenses.

The most similar objects returned by the search do not always share the same morphologies and colors as the labeled query but feature "similarity" in a variety of characteristics. This result is most apparent in the first query presented of a bright central galaxy with an extended purple arc, where the most similar lenses returned display a wide variety of arc colors, arc sizes, and central source clustering but overall exhibit hard-to-quantify similar visual characteristics. As such, the 512 most similar objects we chose to visually inspect was an arbitrary cutoff, and additional lens candidates are likely to be uncovered by extending beyond this number.

Similarity search can benefit supervised classification tasks in a number of ways, as a single label is sufficient to search for additional lens candidates, compared to the many hundreds or thousands required at a bare minimum for supervised training. Supervised classification is biased toward predicting the correct answer on the most common type of labels, and the predictive performance on rare occurrences suffers. Conversely, similarity search can be used to help rapidly identify additional members of less common examples but may not guarantee the same degree of completeness as a full automated classification, as the new lens candidates by construction exhibit a high degree of visual similarity to the few labeled images provided. Thus, to maximize completeness, the techniques can be employed in tandem. Importantly, similarity search does not need to focus on finding additional lenses and can instead be used to construct a more robust training sample of nonlenses. For example, ring galaxies, spiral galaxies, or galaxy mergers were commonly flagged as false positives by a number of previous lens-finding studies due to their scarcity and similar appearance to gravitational lenses. While these objects are not as rare as strong lenses, when randomly sampling nonlenses, they occur infrequently in smaller training sets, and the network can mistake them for lenses. With a similarity search, one can rapidly identify additional examples for these objects and include them in the training set to reduce the number of false positives during inference.

With such results, the computational simplicity of a similarity search approach makes it a very attractive option for identifying new lenses, but it is worth highlighting the limitations. Most notably, similarity search does not perform as
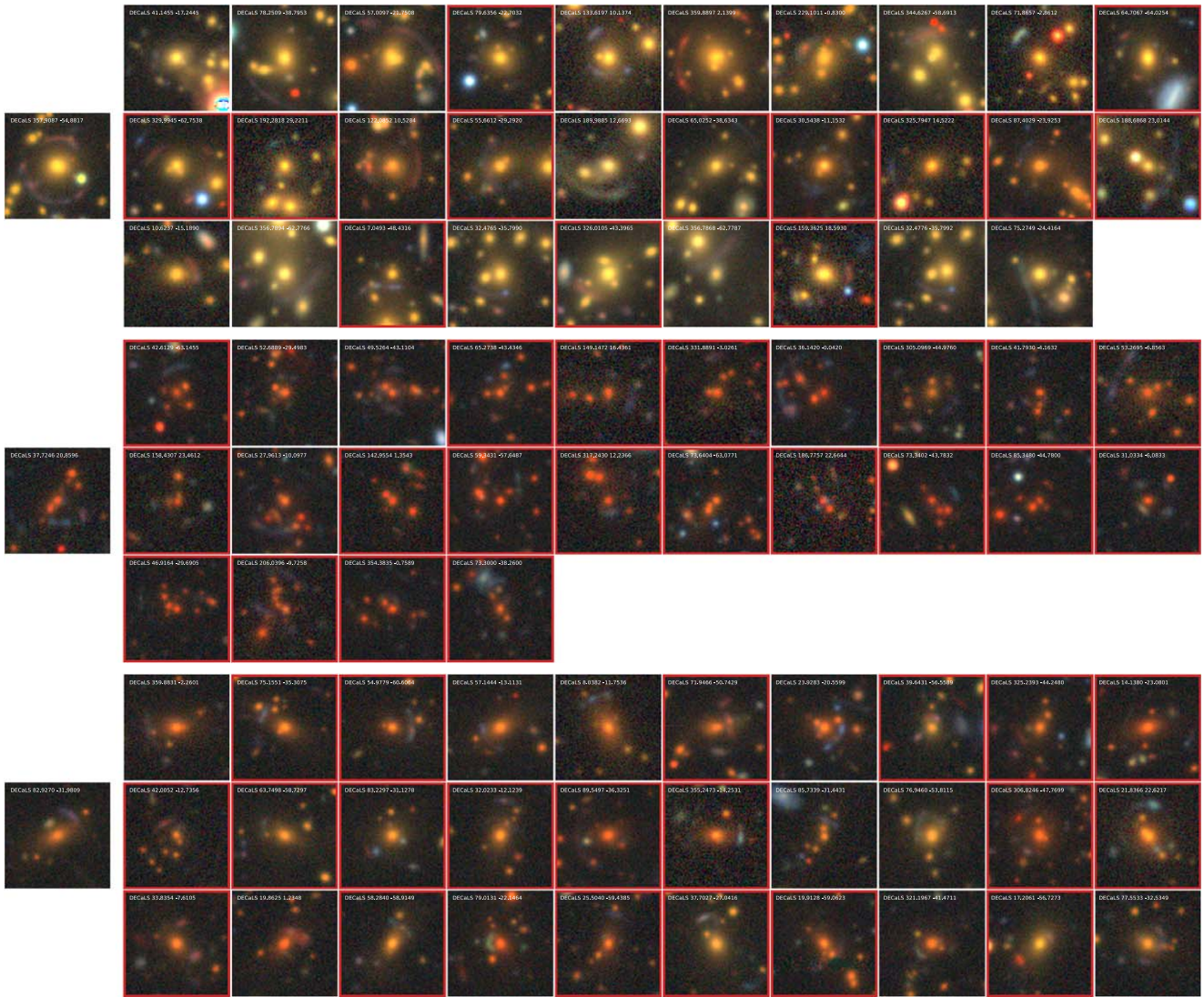
Figure 3. Similarity search for the discovery of new strong lens candidates, achieved without using a single label for training. We queried three images of previously discovered lenses (left), and on the right, we display a sample of new strong lens candidates selected from the top 512 most similar images to each query, ordered by decreasing similarity score. Strong lens candidates with a red border are previously undiscovered. Using only three labeled data samples and ∼10 minutes of visual inspection, we identified 83 strong lens candidates, 53 of them previously undiscovered.

well for images with distinct nonlens features offset from the central source that are unrelated to lensing, such as bright foreground stars or galaxies. For these cases, the most similar images returned are those that also include similar foreground features, and the "similarity focus" has been drawn away from the lensing characteristics we are interested in. This is expected, as our self-supervised model is not given any explicit learning signal on what constitutes lensing features. One potential work-around for this effect is to include such foreground effects as data augmentations in self-supervised pretraining—for example, by randomly adding bright foreground objects to a given view of the original image—which would make representations more invariant to these features. We leave this for future investigation.

We invite readers to explore our publicly available, web-based interactive similarity search tool and try to discover additional strong lenses beyond those presented here.[17]

---
[17] github.com/georgestein/ssl-legacysurvey

4.2. Training Automated Lens Classification Models

Given the large class imbalance between lensed and nonlensed images in the true data, it is critical to consider the balance of lenses and nonlenses in the training data for classification models. A sample with no class imbalance will include only a minimal number of negative examples (due to the small number of positive labels), which will not nearly cover the full distribution of galaxies in the data set and will result in a larger number of false positives on the validation data. Alternatively, training using a significant class imbalance can result in poor convergence, where the model does not predict any images as belonging to the positive class with any high degree of certainty.

Using the 1130 labeled lenses set aside for training, we sample six distinct training sets with increasing class imbalance, testing ratios of 1:1, 1:10, 1:100, 1:1000, 1:4000, and 1:10,000. When increasing the number of nonlenses, we sample additional galaxies while keeping those already included in smaller training sets (e.g., all representations from the 1:1 training set are also included in the 1:10 set). As shown
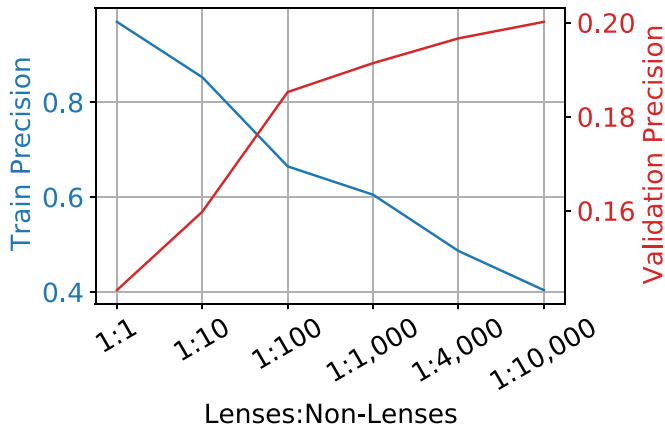
Figure 4. Linear classification precision as a function of the lens:nonlens ratio used for training. Validation is set at 1:10,000 to approximate the class imbalance expected in the full survey. As we increase the lens-to-nonlens ratio, we find lower performance on the training set but much higher performance on the validation data, demonstrating the advantage of training with class imbalances similar to what is expected in the true data.

in Figure 4, we find a clear correlation between the lens-to-nonlens training ratio and the performance of the linear classifier on the validation set. When training with a much smaller degree of class imbalance than exists in the real data (1:1 through 1:100), we find that while the performance on the training set seems impressive, the performance on a realistic validation set suffers, and the model predicts a significant number of false positives. As we increase the lens-to-nonlens ratio, we find lower performance on the training set but much higher performance on the validation data, demonstrating the advantage of training with class imbalances similar to those expected in the true data. An increase of precision (Equation (2)) from 0.14 (1:1) to 0.20 (1:4000) corresponds to a nearly 50% decrease in the manual inspection time required to find a given number of strong lens candidates. Ratios of 1:4000 and 1:10,000 give us our best results, and between the two, performance is comparable. Thus, for our remaining automated lens classification models, we use the 1:4000 ratio, as it is faster to train.

With the training set finalized, linear classification proceeds in a straightforward manner. For fine-tuning, we need to additionally determine the optimal lens oversampling factor. As explained in Section 3.3.4, oversampling is required to ensure that each minibatch of 512 samples includes both positive and negative instances. We varied the oversampling factor to ensure that each minibatch, on average, contains (1%, 2.5%, 5%, 10%) of its samples from the positive class (lens) and found that 5% achieved the best performance on the validation set. This corresponds to an oversampling factor of 40, such that during an epoch, each lens is seen 40 times (with different augmentations applied for each sample), while nonlenses are only seen once per epoch. Linear classification required a few minutes on a CPU, while fine-tuning used 36 hr on eight GPUs.

We show the predicted lens probability over the validation set of the linear classification and fine-tuned models in Figure 5. The results of linear classification and fine-tuning are quite similar, with roughly half the true lenses in the validation set receiving probabilities above the $N = 1139$ cutoff (dotted vertical black lines) determined by the number of images that we have time to visually inspect. For both models, we find a clear correlation between the visual quality of the lens candidates and the predicted lens probability. Images with clear features of strong gravitational

lensing, such as distinctive arcs, are almost always given a high predicted lens probability, while lens candidates with faint features are given low predicted values. We also find great agreement between the predictions of the linear classifier and fine-tuned models for individual image/representation pairs; both models identify similar lenses.

In the right panel of Figure 5, we estimate the number of images that one needs to visually inspect in order to find a given fraction of lenses, assuming that our validation set is representative of the true distribution and there is one lens in every 10^4 galaxies. The initial validation set will contain a number of undiscovered lenses that introduce a certain degree of mislabeling, and we later confirm that a significant number of the "nonlenses" with a high predicted lens probability are in fact previously undiscovered strong gravitational lenses. From the figure, we see that the linear classifier and fine-tuned network perform nearly perfectly on the top 10% of lenses and identify 50% and 60% of the lenses above the manual inspection cutoff, respectively. As we extrapolate to the full DECaLS $mag_z < 20$ sample, the fine-tuned network slightly outperforms the linear classifier if inspecting the top 10^4 predictions (0.02% of the total), but the linear classifier in fact performs better than the fine-tuned if inspecting more than the top 10^5 predictions (0.2% of the total) due to slight overfitting of the fine-tuned model. These estimates do not account for any undiscovered lens configurations that do not have similar examples in the labeled data set.

4.3. Searching for New Strong Lens Candidates

After the linear classification and fine-tuned models were trained using the DECaLS $mag_z < 20$ training sample, we applied them to the remainder of the DECaLS $mag_z < 20$ sources and the full north data set. Although the $mag_z > 20$ sample was dimmer than any data seen during training, we found that both models extrapolated well and resulted in accurate predictions upon visual inspection. The north survey was observed with different instruments and bandpasses and had higher noise levels than the south, thus introducing a train-test discrepancy to such experiments. Nevertheless, as we will demonstrate in the following section, models trained only on south images and applied to the north survey also resulted in the detection of a number of new strong lens candidates.

Using the linear classification and fine-tuned models, a combination of authors J.B., T.M., and G.S. inspected

1. ∼25 similarity searches using lenses selected from the training set,
2. the top ∼7500 predictions over the 43 million DR9 south galaxies with $mag_z < 20$,
3. the top ∼3000 predictions over the 20 million DR9 south galaxies with $mag_z > 20$, and
4. the top ∼5000 predictions over DR9 north.

After an initial test, where all three authors examined an identical set of predictions to ensure consistency between visual classifications, the samples outlined above were split into chunks, and there was little overlap of galaxies inspected by the different authors. The top linear classification and fine-tuned predictions had a significant overlap, and the linear classification model received the most attention. Fine-tuned predictions were inspected after the linear classification to identify any missed lenses in the first pass or any that did not make the linear classification cutoff. We discarded any overlapping lens candidates that resulted from
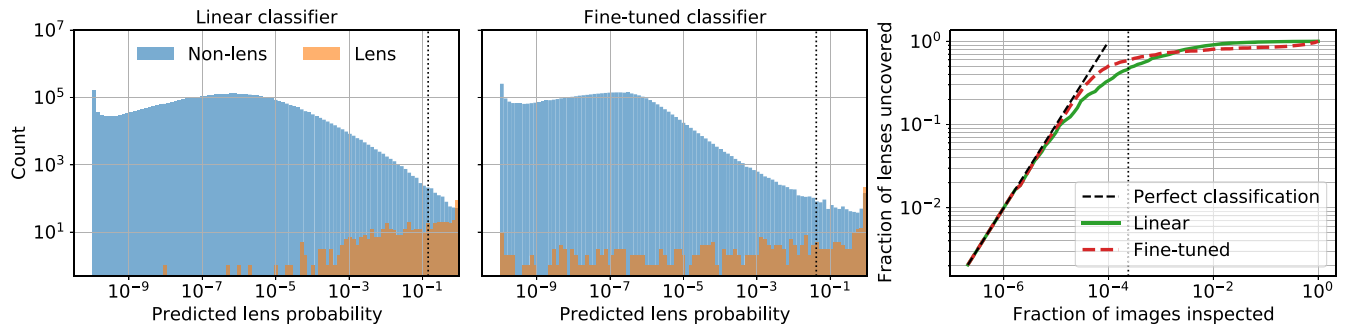
Figure 5. Predicted probability of image containing a lens over the validation set. The vertical dotted black lines designate the predictions over the portion of the survey that we expect to visually inspect in a few hours.

nearby galaxies of the same cluster being independently identified as lenses by different methods.

Extrapolating from Figure 5, we expect that in the south $\mathrm{mag}_z < 20$ sample, we have identified 50%–60% of the lenses, and 30%–40% remain. Approximately 20% of the remaining lenses are likely to be in the top 100,000 predictions, which is an achievable number to manually inspect. This is supported by the fact that we continued to find a number of high-quality lenses after the 7000th highest prediction. The lens candidates remaining to be discovered will likely not be configurations with extremely distinct extended arcs, as we found that those were generally predicted with the highest probabilities in our validation sample. The number of remaining lenses in the south $\mathrm{mag}_z > 20$ or north samples is more difficult to estimate, but we expect it is greater than the fractions quoted above.

We separated our new strong lens candidates into two groups based on their quality. During the lens search outlined above, the authors assigned each lens candidate either a tentative A or B grade, which a single author, G.S., refined into the final lens classifications to achieve consistency. Grade A lens candidates are highly likely to be strong gravitational lenses, as they exhibit distinct lens features such as large extended arcs or clear multiple images. Grade B galaxies exhibit fainter features than grade A and are more ambiguous, with common false positives such as ring galaxies. Nevertheless, grade B lenses still display likely features of gravitational lensing. To include galaxies as either grade A or B strong lens candidates, we erred on the side of caution and attempted to exclude candidates that displayed characteristics of gravitational lensing but were difficult to differentiate from projection or noise effects. As there are no distinct criteria for lens classification, the small number of nonexpert galaxy inspectors likely introduced biases to the type of lenses identified in this work and possibly passed over certain groups of objects. For this reason, we make the network predictions and similarity search tool public and available for anyone to examine.

4.4. Catalog of New Strong Lens Candidates

We compiled the lens candidates discovered through similarity search, linear classification, and fine-tuning into a single catalog consisting of a total of 1192 new strong lens candidates. Table 1 outlines the number of candidates split by lens grade and survey region. The majority of new lens candidates are from the south survey and have $\mathrm{mag}_z < 20$. This is expected, as the south survey, specifically the DES region, has a lower noise level than the north and was also the subset of the data used to train the models. We display the 404 grade A

Table 1
New Strong Gravitational Lens Candidates Split by Survey Region and Lens Grade

Grade	Survey	New Lens Candidates
A	South $\mathrm{mag}_z < 20$	316
	South $\mathrm{mag}_z > 20$	47
	North	41
	All	404
B	South $\mathrm{mag}_z < 20$	583
	South $\mathrm{mag}_z > 20$	131
	North	74
	All	788
Total		1192

lens candidates in Figures 6–8, split by survey region, and show the B candidates in Appendix A.

The most common grade A lens configurations are arcs separated by at least a few arcseconds from the central source(s), more extended arcs separated by a smaller angular distance from a central source, and duplicate background galaxies. We also find numerous Einstein crosses and other rare lens configurations. The majority of the central galaxies have a red or orange appearance, but we also find a number of lensing features around blue central galaxies. The diversity of lens candidates discovered here demonstrates the power of a purely data-driven self-supervised approach capable of learning from every single galaxy in the survey. Our method did not require any cuts on the galaxy or lens sample used; thus, we are able to easily identify lens candidates covering the full variety of appearances. In contrast, the majority of previous studies subsample the data set to include only luminous red galaxies by introducing hand-selected color and magnitude cuts, which, given their redshift distribution and masses, are the most likely type of galaxy to lens background sources (Turner et al. 1984). Targeting only a subsample of the full range of gravitational lenses in the universe will only allow for classification of the given specific lens type and not generalize outside of the limited distribution of data used for training.

Figure 9 illustrates the distribution of the new strong lens candidates on the sky in R.A. and decl., with the background shaded by the number of observations used to construct DR9 images. We find that the DES region (depicted clearly in the bottom left as the dark gray region with decl. $< 5°$) contains the highest density of new lenses. This region was imaged to the greatest depth and thus contains the lowest level of noise, which means that lenses should be more visually distinct and easy to identify. This low noise level also means that the DES region has
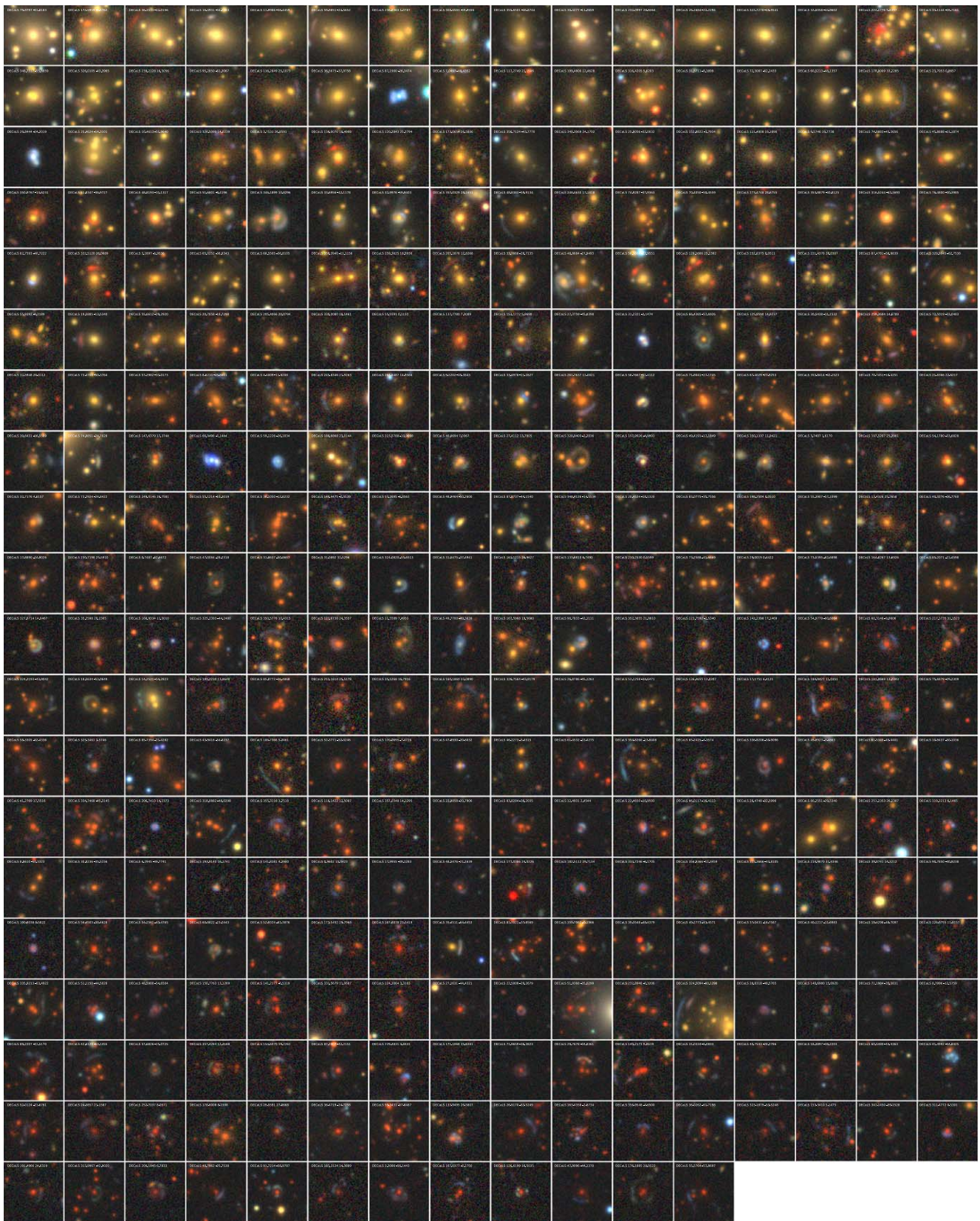
Figure 6. South survey grade A lens candidates with mag$_z$ < 20.

been the focus of the majority of recent previous lens-finding works; thus, many of the lenses have already been found. Nevertheless, in this work, we find a large number of lens candidates in this region that were not identified in any other search. The most significant number of lenses discovered in DESI Legacy Survey data was achieved by Huang et al. (2021),
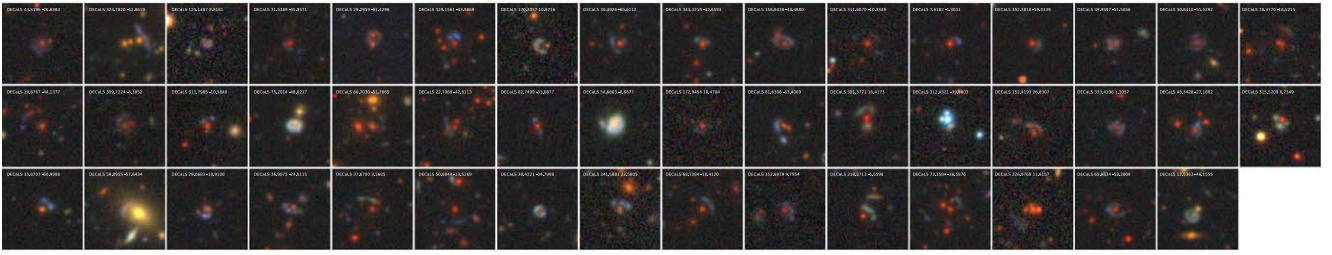
Figure 7. South survey grade A lens candidates with *z*-band magnitude >20.
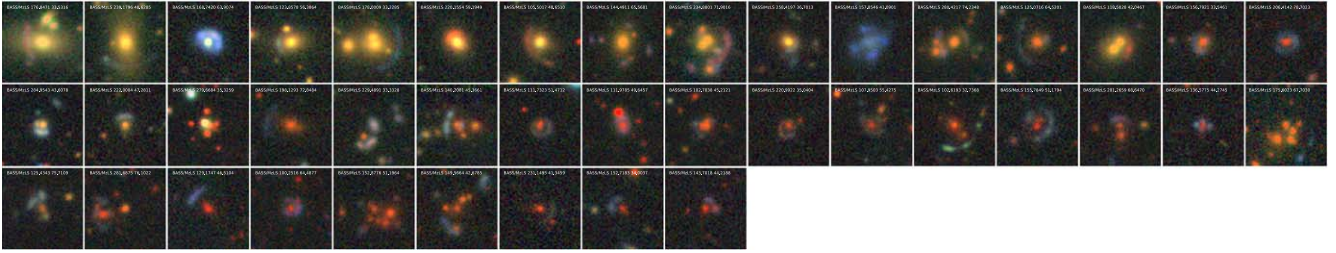
Figure 8. North survey grade A lens candidates.
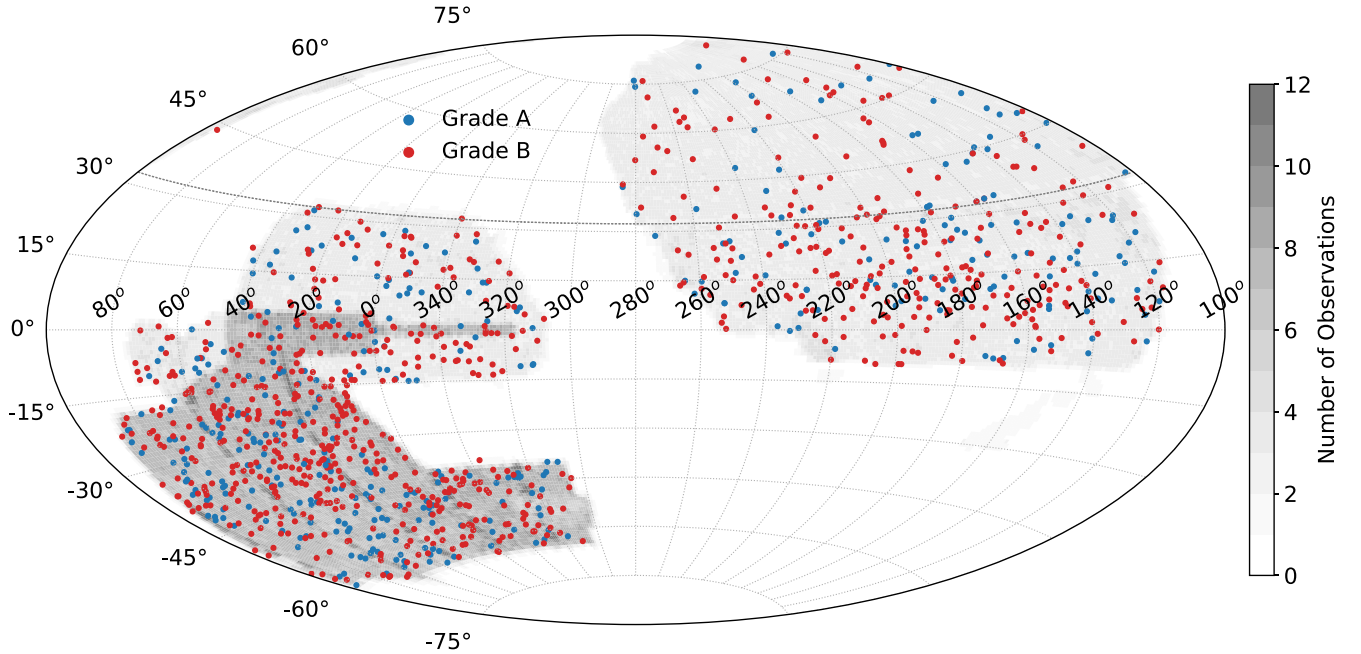
Figure 9. The R.A. and decl. of the new strong lens candidates, with the background shaded by the average number of observations across the three photometric bands in each square-degree region. The gray line at $\delta = 32.375$ denotes the split between the north and south surveys.

who used the previous data release, DR8, which had incomplete coverage below $\delta \approx -32°$. Although in Figure 1 we see a much greater density of training samples above this line than below, we do not find the same with our new lenses. This means we are not simply finding lenses in regions of the sky that have not been searched; rather, we are identifying lenses missed by previous works. It is also clear that the north region is the least densely populated due to both the increased level of noise and the fact that our models were trained on the south images but applied to the north, even though the north images used slightly different photometric bands. We find no distinct correlations between the sky distribution and the lens grade, photometric redshift, or *z*-band magnitude.

We summarize the distribution of magnitudes, photometric redshifts, and the source morphology of the lenses in Figure 10, split by survey region. We find that the newly discovered lenses closely resemble the training set, with the exception of the $mag_z > 20$ south lens candidates, which are, by definition, dimmer than the set used for training. The majority of newly discovered lens candidates are elliptical galaxies (deVaucouleurs = "DEV"), and a number are around spiral galaxies (exponential = "EXP"). Round exponential galaxies with a variable radius ("REX") and Sérsic profiles ("SER") are more rare, but we still find a number with these morphological types.

This new catalog of 1192 strong gravitational lens candidates is summarized in the data tables in Appendix A.
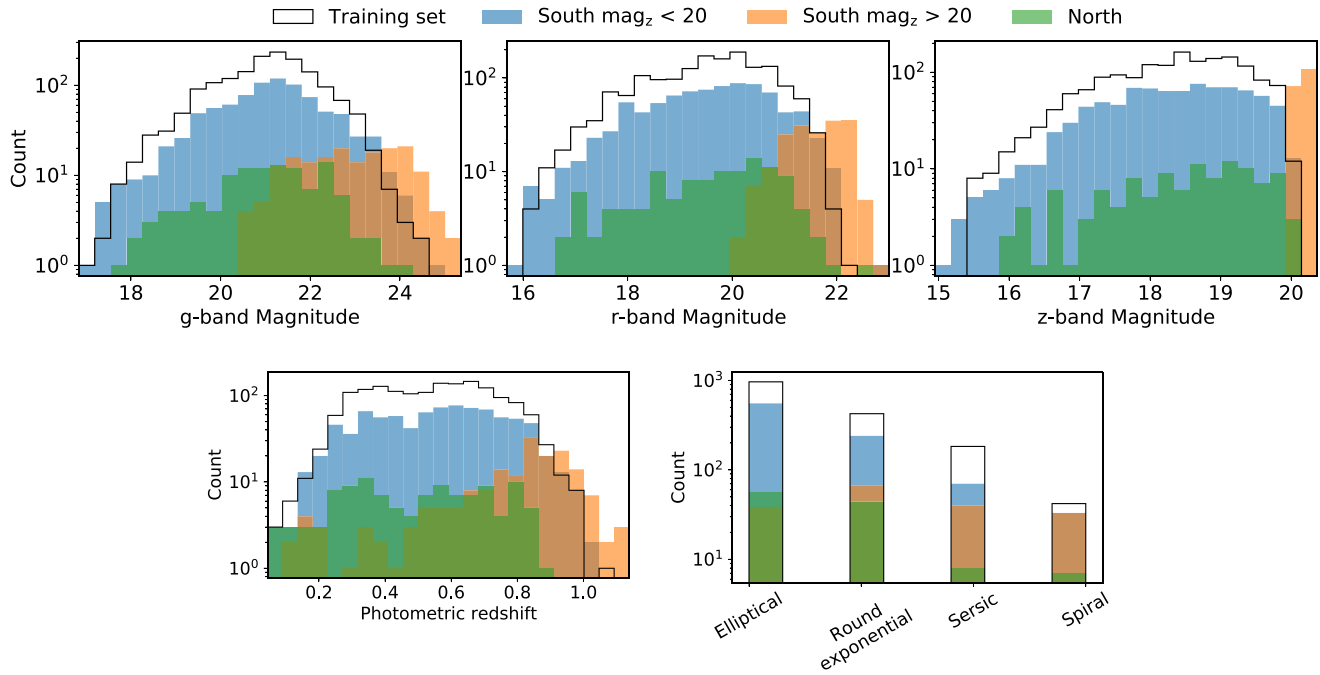
Figure 10. New strong lenses by central source magnitudes, photometric redshifts from Zhou et al. (2021), and the central lens morphology from the Tractor fits.

5. Conclusion

In this work, we demonstrated the first direct application of self-supervised machine learning for finding rare objects in large astronomical data sets, finding 1192 new strong gravitational lens candidates in the DESI Legacy Imaging Surveys. After training the self-supervised model on a large body of unlabeled data, we presented three different methods for the discovery of strong lenses—similarity search, linear classification, and fine-tuning—each with their own distinct advantages over a standard supervised classification setup. The similarity search tool provides an avenue to rapidly identify additional lenses given just a single labeled sample (see Figure 3), which is orders of magnitude less labeled data than would be required to train even the simplest supervised network. Additionally, as the self-supervised model was only trained to distinguish the natural distribution of morphologies and features within the survey sample, it is not limited to identification of a specific class of object but can be used to facilitate the discovery of any object(s) that exist in the data set.

Linear classification and fine-tuning the self-supervised network both proved extremely useful for the automated classification of new strong lens candidates. Although supervised fine-tuning slightly improved the classification results over the fraction of the data set we visually inspected, it is important to consider and contrast the computational resources, the human time investment, and the machine learning expertise required to perform the two types of automated searches and what this means for survey science. Supervised training of the CNN in a distributed GPU setup provides a number of barriers to rapid scientific discovery; it requires a nonnegligible machine learning background, access to a machine capable of storing the full image data on disk (2.5 TB for the training and validation set images and 19 TB for the full DESI Legacy Survey sample used in this work), and access to top-of-the-line GPUs for days just for the final network training (which does not including the time required for hyperparameter searches). Linear classification provides a compelling alternative; it requires only CPU access and minimal or

computational background, and the distilled representations can fit on any modest machine (76 GB for the training and validation sets and 579 GB for the full survey sample).

Using a combination of the three methods, we conducted a fast visual inspection campaign of ∼18,000 DESI Legacy Survey images, from which we identified 1192 new strong lens candidates. This is a significant number when compared to previous supervised CNN studies, despite comparatively little time dedicated to visual identification. For example, using more restrictive data cuts of the DESI Legacy Survey DR8 that prefilter galaxies not likely to be lenses, Huang et al. (2020) found 335 new candidates in ∼50,000 inspections, and Huang et al. (2021) found 948 new candidates in 38,679 inspections of their main ResNet model. We emphasize that direct comparisons between our study and such previous works are difficult due to two competing effects, and as such, these numbers are stated here only as a guideline. On one hand, previous lens searches that utilized previously discovered lenses as their training sample (rather than a simulated sample, e.g., Jacobs et al. 2017, in which lenses are adequately numerous but may lack full realism) generally had a smaller number of labeled lens candidates to train their models, which limits their classification performance. On the other hand, previous searches had a greater number of undiscovered high-quality lenses in the unmined data, and the number of remaining lenses distinguishable at the resolution of the DESI Legacy Surveys has diminished with each successive lens search (Huang et al. 2020, 2021).

We expect that a nontrivial number of lenses in our catalog, on the order of 30%–40% based on the validation set experiment, remain undiscovered. However, they can be discovered by subsequent similarity searches or inspections of network predictions that fell outside of our limited labeling campaign. For this reason, we are releasing the similarity search tool and the network predictions to facilitate the identification of additional strong lenses through crowd-sourcing.[18]

---

[18] Predictions and the location of the similarity search tool can be found at github.com/georgestein/ssl-legacysurvey.

Appendix A
Strong Lens Catalog and Additional Lens Candidates

We provide the 1192 new lens candidates identified in this work in data tables at github.com/georgestein/ssl-legacysurvey and display the grade B lenses in Figures 11–14. Due to the fainter lens features of grade B lenses compared to grade A, it is likely that the grade B catalog contains a larger fraction of false positives. For this reason, we recommend that follow-up observations and
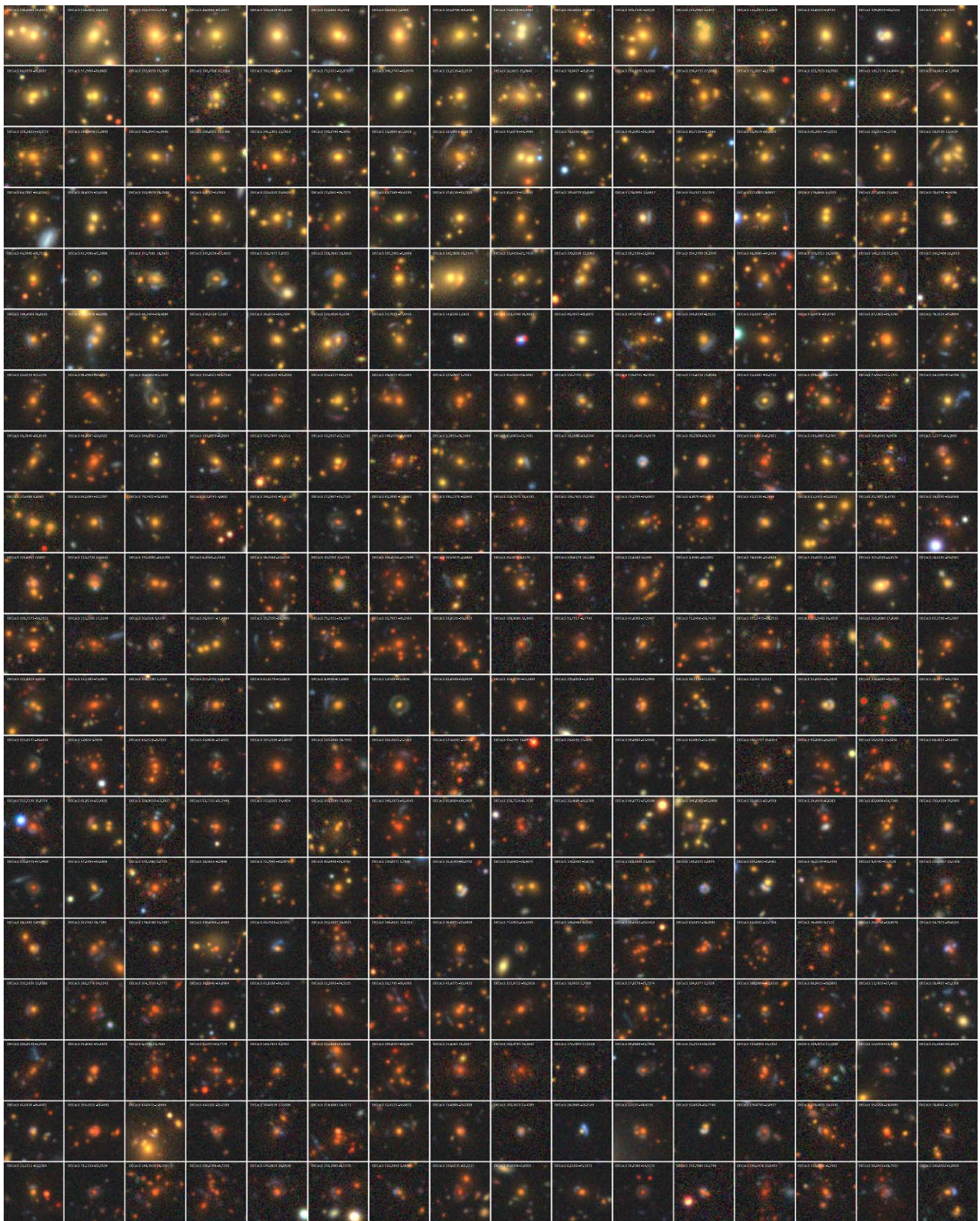
Figure 11. South survey grade B lens candidates with *z*-band magnitude <20.

investigations start with the grade A catalog before moving onto the B. During the writing of this paper, after the conclusion of our lens search, Rojas et al. (2021) performed an additional supervised CNN search on the DES region, finding 186 new lens candidates after examining their top 76,582 network predictions. We have not checked for overlap with our new lens candidates.
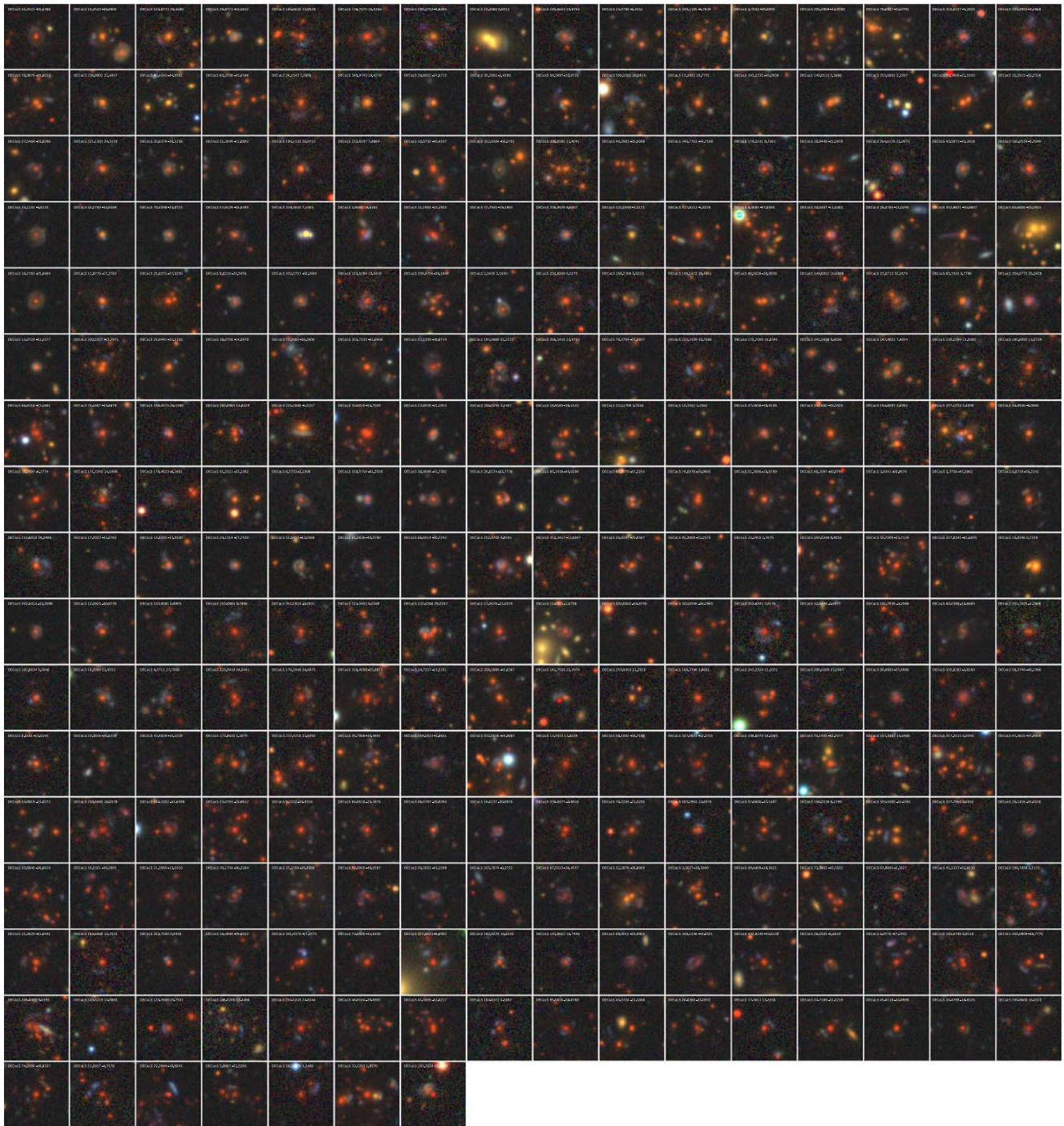
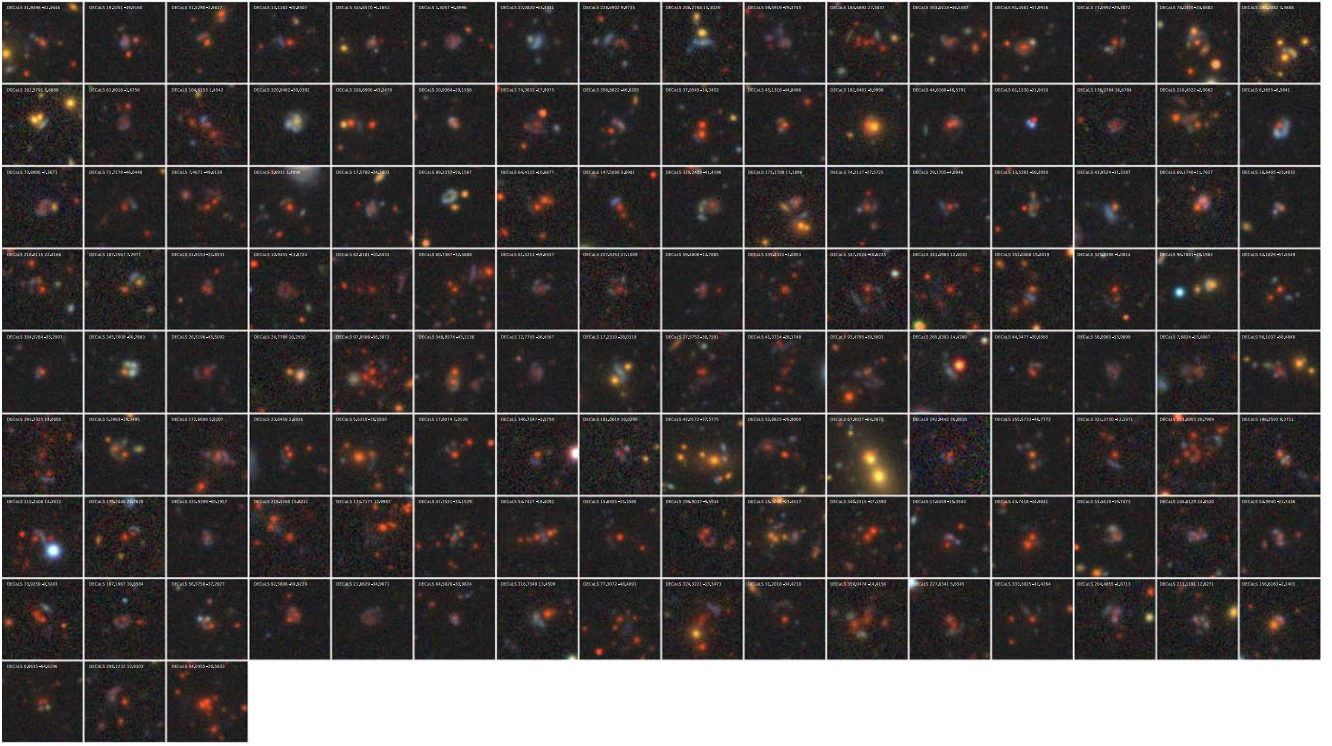**Figure 12.** South survey grade B lens candidates with *z*-band magnitude <20.

Figure 13. South survey grade B lens candidates with *z*-band magnitude >20.
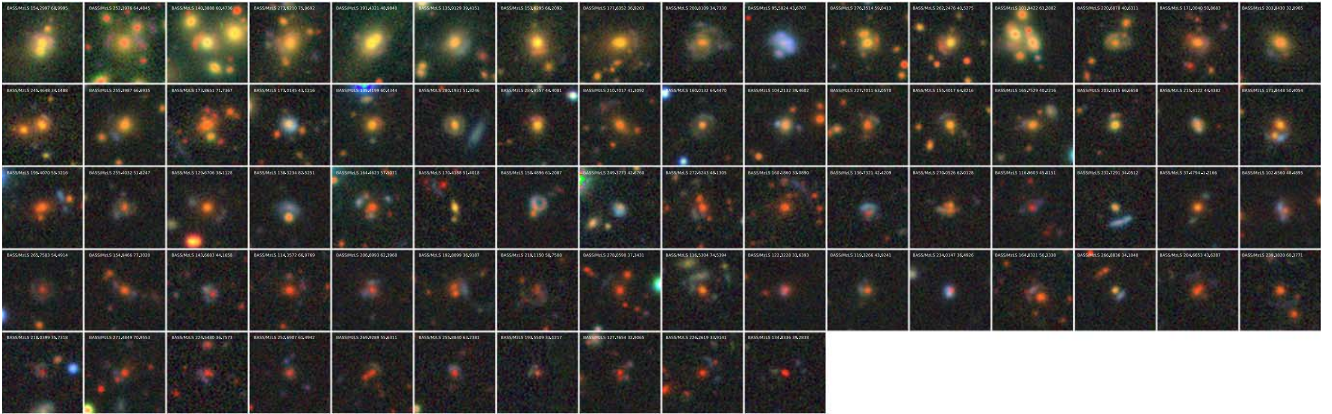

Figure 14. North survey grade B lens candidates.

Appendix B
Data Augmentations

To design the data augmentations, we fit to measured distributions of the Gaussian noise level, point-spread smoothing, and extinction in images from the south survey sample with mag$_z$ < 20, as this was the sample used to train the self-supervised model. Distributions were fit separately in the *g*, *r*, and *z* bands, and the augmentations were considered to be uncorrelated across bands. This is a reasonable approximation given that each image is compiled from a number of different observations taken at different times.

To determine the level of Gaussian noise in the DR9 images, we used the noise equivalent area calculated using the blob-masked version outlined at legacysurvey.org/dr9/nea/,

$$\frac{1}{\sigma_{\mathrm{pix}}^2} = 4\pi \ \mathrm{psfdepth}\left[\frac{\mathrm{psfsize}}{2.3548}\right]^2, \qquad (B1)$$

where psfsize is in units of pixels ($0\rlap{.}''262$), and σ_{pix} is the estimated Gaussian noise level in each pixel. The standard deviation of the PSF was calculated by

$$\sigma_{\mathrm{PSF}} = \frac{\mathrm{psfsize}}{2.3548}, \qquad (B2)$$

where psfsize is again in units of pixels, and the extinction was given in the sweep catalogs. Examining the histogram of the values measured from the 43 million galaxies, we found that all of them are well described by a lognormal distribution separately in each band for the Gaussian noise and PSF. Fitting a lognormal distribution to each with scipy. stats. lognorm.fit, we get the following fit parameters (shape $\equiv \sigma$, loc $\equiv \mu$, scale $\equiv \alpha$), which can be used to get the probability density through

Augmentation	Shape	Loc.	Scale
Gaussian noise	(0.2264926, 0.2431146, 0.1334844)	(−0.0006735, −0.0023663, −0.0143416)	(0.0037602, 0.0067417, 0.0260779)
PSF	(0.2109966, 0.3008485, 0.3471172)	(1.0807153, 1.2394326, 1.1928363)	(1.3153171, 0.9164757, 0.8233702)
Extinction	0.67306	0.001146	0.03338

Note. See Equation (B3).
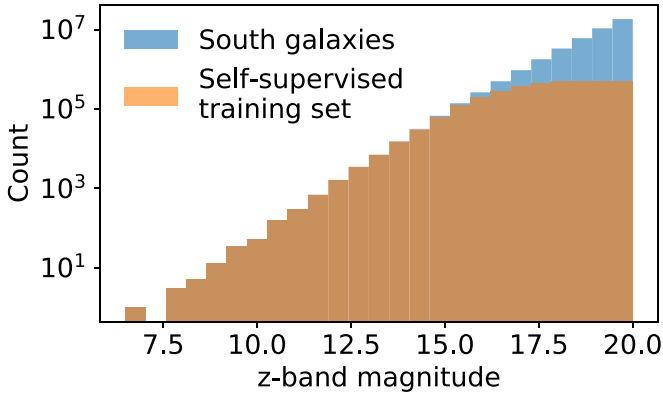
Figure 15. The z-band magnitude of the 3,500,000 galaxy sample used for training the self-supervised model compared to the full distribution. Although the brightest and most nearby galaxies will not constitute strong lenses, we do not use a minimum magnitude cut to remove them from the self-supervised training set; the goal at this stage is only to learn task-agnostic galaxy representations from a diverse set of galaxy images.

$$p(x) = \frac{1}{(x-\mu)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\log((x-\mu)/\alpha)^2}{2\sigma^2}\right), \quad \text{(B3)}$$

and we show the best-fit parameters in Table 2.

Each image in the data set already has a certain level of Gaussian noise and Gaussian blur applied during the observation process, so to add noise on top of what is already in the image, we do not want to directly sample from the fit distributions. Instead, from the power-law distributions, we sample an initial noise level in the image, σ_i, and a desired noise level in the image, σ_f, and apply a differential noise, $\sigma = \sqrt{\sigma_f^2 - \sigma_i^2}$. If σ_f is smaller than σ_i, we do not apply the augmentation for that sample in that band.

Figure 15 displays the z-band magnitude of the 3,500,000 galaxy sample used for training the self-supervised model compared to the full distribution.

ORCID iDs

George Stein ⓘ https://orcid.org/0000-0002-5193-516X

References

Cañameras, R., Schuldt, S., Shu, Y., et al. 2021, A&A, 653, L6
Cañameras, R., Schuldt, S., Suyu, S. H., et al. 2020, A&A, 644, A163
Caputi, K. I., Caminha, G. B., Fujimoto, S., et al. 2021, ApJ, 908, 146
Caron, M., Touvron, H., Misra, I., et al. 2021, arXiv:2104.14294
Carrasco, M., Barrientos, L. F., Anguita, T., et al. 2017, ApJ, 834, 210
Chan, J. H. H., Suyu, S. H., Sonnenfeld, A., et al. 2020, A&A, 636, A87
Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020a, arXiv:2002.05709
Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. 2020b, arXiv:2006.10029
Chen, X., Fan, H., Girshick, R., & He, K. 2020c, arXiv:2003.04297
Cheng, T.-Y., Li, N., Conselice, C. J., et al. 2020, MNRAS, 494, 3750
Collett, T. E. 2015, ApJ, 811, 20
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
Diehl, H. T., Buckley-Geer, E. J., Lindgren, K. A., et al. 2017, ApJS, 232, 15
Donahue, J., Krähenbühl, P., & Darrell, T. 2016, arXiv:1605.09782
Donahue, J., & Simonyan, K. 2019, arXiv:1907.02544
Gentile, F., Tortora, C., Covone, G., et al. 2022, MNRAS, 510, 500
Gilman, D., Birrer, S., Treu, T., Nierenberg, A., & Benson, A. 2019, MNRAS, 487, 5721
Goyal, P., Caron, M., Lefaudeux, B., et al. 2021, arXiv:2103.01988
Grill, J.-B., Strub, F., Altché, F., et al. 2020, arXiv:2006.07733
Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., & Mustafa, M. 2021, ApJL, 911, L33
He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. 2020, in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 9729
He, Z., Er, X., Long, Q., et al. 2020, MNRAS, 497, 556
Hendrycks, D., Mazeika, M., Kadavath, S., & Song, D. 2019, arXiv:1906.12340
Huang, X., Storfer, C., Gu, A., et al. 2021, ApJ, 909, 27
Huang, X., Storfer, C., Ravi, V., et al. 2020, ApJ, 894, 78
Jacobs, C., Collett, T., Glazebrook, K., et al. 2019a, MNRAS, 484, 5330
Jacobs, C., Collett, T., Glazebrook, K., et al. 2019b, ApJS, 243, 17
Jacobs, C., Glazebrook, K., Collett, T., More, A., & McCarthy, C. 2017, MNRAS, 471, 167
Jaelani, A. T., More, A., Oguri, M., et al. 2020, MNRAS, 495, 1291
Lanusse, F., Ma, Q., Li, N., et al. 2018, MNRAS, 473, 3895
LeCun, Y., Bengio, Y., & Hinton, G. 2015, Natur, 521, 436
Li, R., Napolitano, N. R., Spiniello, C., et al. 2021, ApJ, 923, 16
Li, R., Napolitano, N. R., Tortora, C., et al. 2020, ApJ, 899, 30
Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179
Metcalf, R. B., Meneghetti, M., Avestruz, C., et al. 2019, A&A, 625, A119
Noroozi, M., & Favaro, P. 2016, arXiv:1603.09246
Paszke, A., Gross, S., Massa, F., et al. 2019, in Advances in Neural Information Processing Systems 32, ed. H. Wallach et al. (Red Hook, NY: Curran Associates Inc.), 8024
Petrillo, C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129
Petrillo, C. E., Tortora, C., Vernardos, G., et al. 2019, MNRAS, 484, 3879
Pourrahmani, M., Nayyeri, H., & Cooray, A. 2018, ApJ, 856, 68
Reed, C. J., Metzger, S., Srinivas, A., Darrell, T., & Keutzer, K. 2020, arXiv:2009.07724
Rojas, K., Savary, E., Clément, B., et al. 2021, arXiv:2109.00014
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Sonnenfeld, A., Chan, J. H. H., Shu, Y., et al. 2018, PASJ, 70, S29
Sonnenfeld, A., Jaelani, A. T., Chan, J., et al. 2019, A&A, 630, A71
Sonnenfeld, A., Verma, A., More, A., et al. 2020, A&A, 642, A148
Stein, G. 2020, georgestein/ml-in-cosmology: Machine Learning in Cosmology, v1.0, Zenodo, Zenodo, doi:10.5281/zenodo.4024768
Stein, G., Harrington, P., Blaum, J., Medan, T., & Lukic, Z. 2021, arXiv:2110.13151
Treu, T. 2010, ARA&A, 48, 87
Turner, E. L., Ostriker, J. P., & Gott, J. R. I. 1984, ApJ, 284, 1
Vega-Ferrero, J., Domínguez Sánchez, H., Bernardi, M., et al. 2021, MNRAS, 506, 1927
Walmsley, M., Smith, L., Lintott, C., et al. 2020, MNRAS, 491, 1554
Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835
Wong, K. C., Sonnenfeld, A., Chan, J. H. H., et al. 2018, ApJ, 867, 107
Wong, K. C., Suyu, S. H., Chen, G. C. F., et al. 2020, MNRAS, 498, 1420
Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. 2019, in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV) (Piscataway, NJ: IEEE), 1476
Zhou, R., Newman, J. A., Mao, Y.-Y., et al. 2021, MNRAS, 501, 3309