An Unsupervised Hunt for Gravitational Lenses

Stephen Sheng UC Davis
Keerthi Vasan G.C. UC Davis
Chi Po Choi UC Davis
James Sharpnack Amazon[1]
Tucker Jones UC Davis

Abstract

Strong gravitational lenses allow us to peer into the farthest reaches of space by bending the light from a background object around a massive object in the foreground. Unfortunately, these lenses are extremely rare, and manually finding them in astronomy surveys is difficult and time-consuming. We are thus tasked with finding them in an automated fashion with few if any, known lenses to form positive samples. To assist us with training, we can simulate realistic lenses within our survey images to form positive samples. Naively training a ResNet model with these simulated lenses results in a poor precision for the desired high recall, because the simulations contain artifacts that are learned by the model. In this work, we develop a lens detection method that combines simulation, data augmentation, semi-supervised learning, and GANs to improve this performance by an order of magnitude. We perform ablation studies and examine how performance scales with the number of non-lenses and simulated lenses. These findings allow researchers to go into a survey mostly "blind" and still classify strong gravitational lenses with high precision and recall.

1 Introduction

Massive galaxies can deflect the light from background sources through the effect of gravitational lensing, creating magnified "arcs" and multiple images of background galaxies when they are located directly along the line of sight. Such alignments are rare, and these lensing systems are important to astronomers for a range of studies, such as glimpsing into the farthest regions of space where the light of distant objects is ordinarily too faint to detect. With strong gravitational lenses, this light becomes focused and amplified. Additionally, the lensing information can be used to study the mass distribution in foreground galaxies, notably including the non-baryonic dark matter which comprises most mass in the universe (Metcalf et al., 2019).

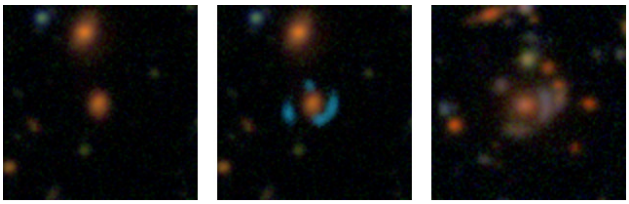

Figure 1: Non-lens (left), simulated lens (middle), real lens (right)

A principal challenge is that strong gravitational lenses are incredibly rare. Across the entire sky only of order a thousand such systems are currently known (Metcalf et al., 2019). Previous efforts to find strong gravitational lenses have largely been done manually by individuals visually inspecting images. This is both impractical and expensive. In recent years, various groups have turned to deep learning methods to search for lens systems (e.g., Jacobs et al., 2017; Sonnenfeld et al., 2018; Pourrahmani et al., 2018; Jacobs et al., 2019; Huang et al., 2020; Li et al., 2020; Cañameras et al., 2020). These early attempts were rather simplistic as they typically only train and evaluate their models in a supervised fashion either on small numbers of known lenses or by making simulated lenses from their own surveys. Nonetheless, deep learning is proving to be a fruitful and efficient approach.

Several surveys are planned for the next decade to observe wide areas of the sky at unprecedented depth and angular resolution (e.g., Rubin Observatory [LSST Science Collaboration et al., 2009], Euclid [Laureijs et al., 2011] and Roman Space Telescope [Spergel et al.,

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

[1]Work done prior to joining Amazon

2015]). These will enable the detection of orders-of-magnitude more strong lenses than with current data (Oguri and Marshall, 2010). The early challenge of analyzing these surveys is that astronomers will not have access to lenses to build their classifiers. In this case, there are two primary options: (1) use lenses found from other surveys and hope the features are effective and transferable, or (2) create simulated lenses based on each survey and train a classifier on those. For (1), the biggest issue is that the transferred performance may vary significantly. This is due to the fact that images from these other surveys are produced with different instruments as well as different preprocessing techniques. Therefore, the samples used for training may be too distributionally dissimilar (i.e. covariate shift) from their target to be useful. One possibility to ameliorate the effects of covariate shifts is to use CycleGANs (Zhu et al., 2017) to transform these images to look like the target data distribution. However, it still doesn't solve the issue of the extremely small number of known lenses with heterogeneous imaging. So (2) is the realistic option for producing consistent performance across surveys by simulating lenses directly on the target set.

Using simulations for training data is quite common in deep learning (Nikolenko, 2019). The problem with option (2) though is that researchers will be creating simulated lenses without a reference point for how they look in their survey. This results in classifiers having good performance when evaluating on held out simulations, but poor performance when classifying real lenses. This is especially problematic for multi-channel images (see Fig. 1) since getting the channel information incorrect can lead to an ineffective classifier. Instead of trying to get all the channel information of the arcs correct, one possibility is to simulate lenses on a single channel and build classifiers to detect lenses in this setting (Cañameras et al., 2020). This sidesteps the issue of getting the channel information correct, but this workaround causes us to lose some contextual information about the "coloring" of the lenses and the surrounding objects, which may actually help the model learn to detect lenses. As a result, we do not explore this option in this paper. We also do not explore using pretrained networks here. Instead, we will focus on a completely self-contained regimen for building classifiers from simulated data. Data augmentation is one way to address this issue of realism without sacrificing this multi-channel information from the image. Secondly, while we can obtain a small sample of non-lensed images to train our classifier, the majority of the survey remains unlabeled, so the use of semi-supervised learning (SSL) algorithms is also a prudent direction to boost the performance of the classifier. By understanding the correct ways to leverage these methods in concert, we can show that you can create highly effective classifiers for detecting lenses even if you only train on potentially "bad" simulated lenses.

2 SSL And Unsupervised Learning

The simplest approach to building a classifier is to use the simulated lenses as our target and train a fully supervised classifier. The limitations of course is that the unlabelled data isn't leveraged and the simulated lens distribution may differ from that of the real lenses.

2.1 Semi-supervised Learning

We find that SSL algorithms are another indispensable tool for classifying lenses. In recent years, the field of deep learning has seen significant progress in the area of semi-supervised learning algorithms (Yang et al., 2021; van Engelen and Hoos, 2019). Instead of covering all of them, we will focus on a narrow collection of state-of-the-art algorithms: Pseudo-label (Lee, 2013), Π-model (Laine and Aila, 2017), Mean Teacher (Tarvainen and Valpola, 2017), VAT (Tarvainen and Valpola, 2017), MixMatch (Berthelot et al., 2019).

For semi-supervised learning algorithms, there are usually two primary goals: consistency regularization and entropy minimization. Some SSL methods (e.g. consistency regularization) considered here require data augmentation (DA), and we summarize the DA methods used in Table 2. These methods are chosen specifically with this application in mind.

Consistency regularization is based on the idea that a classifier should output the same predictions even if the image has been augmented. This is usually carried out by appending a regularizing term to the loss that computes the "distance" between the outputs of the classifier evaluated on two stochastically augmented versions of the same image. Almost all the algorithms we listed above utilize this in some form or another, with the exception of pseudo-label. The set of augmentations is also usually something predefined, which means that the application isn't domain agnostic, and performance will largely depend on the domain-specific augmentations. The exception of course is VAT (Miyato et al., 2019), which generates the augmentations during training instead of being predefined.

Entropy minimization is based on the idea that the decision boundary of the classifier should lie in low-density regions. Worded another way, if two images x_1 and x_2 are close in a high-density region then the predictions y_1 and y_2 should be close as well. Pseudo-label and MixMatch both try to enforce these properties. Pseudo-label does it by assigning pseudo-labels to unlabeled images which are determined by the class

with the highest predicted value. MixMatch does this too but less dramatically by sharpening the predicted values to be used as the label instead of hard thresholding the predictions to produce a pseudo-label.

Typically, the SSL setting assumes that the training and test distributions are the same. However, we will train on simulated images and test on real lenses. A priori it was unclear if the SSL algorithms would improve the metrics in question. Furthermore, there is also the question of whether or not SSL algorithms will even improve over baselines tuned with data augmentations since it has been shown in the past that a classifier's performance can often match the state-of-the-art SSL algorithms by choosing the correct data augmentations (Oliver et al., 2018). Nevertheless, we find that SSL algorithms are an indispensable tool in our arsenal.

2.2 Unsupervised Learning

In unsupervised learning, the typical use case is to learn a data distribution. In deep learning, this is typically done by training a GAN (Goodfellow et al., 2014; Arjovsky et al., 2017; Gulrajani et al., 2017). The outcome of training a GAN is a generator that can produce similar samples from the data distribution it learned from. Those samples are then used for training the classifier. One can think of this as another form of data augmentation. The difference in our case is that the data distribution(i.e. the simulated lenses) we would use to train our GAN does not come from our target distribution(i.e. the real lenses). However, we believe that this can still be helpful because, as we mentioned earlier, we have no a priori notion of what a lensed image would look like coming from the DLS survey. And because GANs do not necessarily faithfully reproduce the data distribution it was trained on, this more exotic form of augmentation should nevertheless be beneficial for improving our classifier's ability to generalize to real lenses.

3 Data And Experimental Setup

The data that will be the focus of our study comes from the Deep Lens Survey (DLS; Wittman et al., 2002). Due to the paucity of known lenses in this survey, we do not allow any training or validation (model tuning) to be done on real lenses. Rather they were reserved for the final comparison of a handful of methods attempted.

3.1 Deep Lens Survey (DLS) and Lens Simulations

The Deep Lens Survey consists of 5 independent fields of 4 $\deg^2$ each, with images taken over $\sim$100 nights using the 4-meter Blanco and Mayall telescopes. The full 20 $\deg^2$ area contains $\sim$5 million cataloged galaxies imaged in 4 different astronomical filters (B,V,R,z) which roughly cover the visible spectrum (i.e. 3000-10000Å). The throughput curve for the filters is published in Schmidt and Thorman (2013) and the data products from the survey are available for public use. For this work, we make use of only the BVR filters as they have the highest SNR. A total of 267,961 galaxies which are likely to act as strong lenses (i.e. which appear to be massive galaxies at moderate cosmological redshifts) were photometrically selected for this analysis by applying a R band magnitude cut ($17.5 < R < 22$). Color images are then constructed for these galaxies using HumVI (Marshall et al., 2015) with the target galaxies centered in the images.

For any given image from the survey, we create a simulated lens counterpart which we use for training. We assume a background galaxy is present behind the central galaxy, and use the `glafic` (Oguri, 2010) lens modeling code to trace the background galaxy's light through the foreground lensing potential. We add the resulting simulated lensed arcs to the DLS survey images using HumVI. The values chosen for the background galaxy and the lensing potential used in the simulations do not rely on any physical property of the foreground and background galaxy. Instead, they randomly probe a range of Einstein radii and redshift values, appropriate for the selected target galaxies, yielding a wide variety of lens configurations. Each simulated lens image has exactly one non-lensed image pair. In other words, we do not use the same non-lensed image to create multiple simulated lensed images in different lens configurations. For this work, we limit our simulations to background sources with relatively blue colors, as these are the most common at high redshifts ($z > 1$) and the most likely to be detectable in DLS data. Finally, we visually inspected the simulated lenses and removed any images where the central galaxy was significantly brighter than the arc. This resulted in 259,489 simulated lenses.

3.2 Training Data

From the non-lenses and the corresponding simulated lenses, we make two training sets: TrainingV1 and TrainingV2. For TrainingV1 we use 266,301 images for non-lenses and 257,874 corresponding simulations as lenses. For TrainingV2 we use the 7,074 human-labeled objects as non-lenses and 6,929 corresponding

Table 1: Summary Of Images In Datasets

Dataset	Non-lenses	Simulated lenses	Real lenses	Unlabeled data	Percentage of lenses
TrainingDataPure	267961	259489	0	0	-
TrainingV1	266301	257874	0	0	-
TrainingV2	7074	6929	0	259248	-
SimTest	786	773	0	0	-
TestV1	874	0	52	0	5.6156
TestV2	874	0	27	0	2.9967

Table 2: Augmentations Used During Training

Name	Description
RGB-shuffle	Randomly perturb the order of the channels in the images
JPEG quality	Randomly apply JPEG compression with quality between 50-100%
Rot90	Randomly rotate the images by a multiple of 90 degrees
Translations	Randomly translate the images by at most 20 pixels in the up, down, left and right directions
Horizontal flips	Randomly flips the images across the x-axis
Color augmentation	Randomly perturb the brightness(-0.1-0.1), saturation(0.9-1.3), hue(0.96-1.00), and gamma(1.23-1.25) of the images

simulations as lenses. The rest of the 259,248 images serve as unlabelled data. During training, we do a 90-10 split whereby 90 percent of this data is used for training the ResNet model and 10 percent is reserved for validation. We also created a holdout set, SimTest, of 786 images for non-lenses and 773 images for lenses to be explicitly used for testing in the simulated setting. We summarize these details in Table 1.

3.3 Initial Lens Discovery and Testing Data

Prior to this work, there were only a few real lenses, with which we might form our test set, known in the entire DLS survey. Since manually searching the entire survey for lenses (which are very rare) is a laborious and time-consuming task, we use a pilot model to perform an initial search of the survey. The pilot model was built with the convolution neural network ResNet (He et al., 2016a,b) of 11 layers depth with polar-transformed images as the input. The motivation behind transforming the images to a polar coordinate system is that at lower Einstein radii and galaxy scales, the arcs are approximately symmetric around the image center and a polar transform captures this symmetry as a straight line. The pilot model was an ensemble of 5 ResNet model instances, and each instance was trained on a randomly selected subset of a pilot training dataset which contained 200,000 simulated lenses and 200,000 non-lenses. The entire sample of unlabeled survey images from the DLS survey (279,149 images in total) was scored by the ResNet ensemble. Around 3000 galaxy images (1% of survey) that had the highest median scores were taken for human inspection by a team of astronomers. 52 were labeled as good lens candidates and form our test set TestV1. 27 out of the 52 were deemed very likely to be strong real lens candidates and form our test set TestV2.

In order to populate our test sets with real non-lenses, the same team of astronomers were asked to label a fraction of images ($\sim 3\%$) from the survey. Since most of the images are expected to be non-lenses, this is an easy task and does not warrant an ML model. A total of 8734 galaxy images were labelled to be non-lenses from the entire survey and 874 (i.e., 10%) were randomly chosen out of those to be included as non-lenses for both the TestV1 and TestV2 test sets. Therefore with the help of the pilot model and human labeling, two test sets: TestV1 (52 Lenses, 874 NonLenses) and TestV2 (27 Lenses, 874 NonLenses) are constructed.

This initial lens discovery was done independently of the experiments in order to prevent data leakage between the real lens test sets and the supervised models trained in the main experiment. In addition, we set the number of candidate images returned by the pilot model to be much larger than the expected number of lenses in the DLS survey. As a result, the number of discovered lenses is reasonable given what has been found in other surveys. With this being the case, we believe that the main source of error in the test set labels is human error in the hand labeling of real lenses.

3.4 Experimental Setup And Implementations

We used a standard ResNet-11 architecture for all experiments. The models we use are broadly broken up into 3 groups: supervised, semi-supervised, and GAN+semi-supervised. There are four supervised models: SupervisedV1, SupervisedV1+DA, Super-

visedV2, and SupervisedV2+DA. SupervisedV1 and SupervisedV2 are trained on TrainingV1 and TrainingV2, respectively, but without data augmentation. SupervisedV1+DA and SupervisedV2+DA are trained on TrainingV1 and TrainingV2, respectively, but with data augmentation.

For the semi-supervised models, we have five: MixMatch, Pseudo-label, Mean Teacher, Π-model, and VAT. These are all trained on TrainingV2 with data augmentations. Lastly, we have the GAN+semi-supervised models. The only difference here is that we expand TrainingV2 with 7000 randomly generated lenses from the WGAN we trained on simulated lenses from TrainingV1. The architecture used for the WGAN is nearly the same as the one used in Gulrajani et al. (2017), with some modifications. The model selection for the WGAN was based on visual inspection rather than measurements like FID because we aren't trying to perfectly recreate the simulated data distribution. Data augmentation is applied to the GAN+semi-supervised models as well.

We trained each model for 100 epochs using the Adam optimizer with a learning rate of 0.001 and a mini-batch size of 32. The only exceptions are the performance scaling experiments where we used 200 and 400 epochs when the number of simulated lenses and non-lenses were 2000 and 1000, respectively. For all models, we used a 90-10 split for training and validation. We trained four candidates for each model which we used to evaluate the test sets to get an average performance. For the pseudo-label models we used $\alpha = 0.1$. For Π-model we used $\lambda_U = 0.1$. For MixMatch we used $T = 0.5$, $K = 2$, $\alpha = 0.75$, $\lambda_U = 0.1$, ema_decay $= 0.999$. For Mean Teacher we used $\lambda_U = 0.1$ and ema_decay $= 0.999$. For VAT we used $\epsilon = 0.001, \epsilon_{\text{adv}} = 0.1$ and $\alpha = 0.01$.

4 Results

4.1 Data Augmentation Significantly Improves Performance

From our results, we see that data augmentations have a significant effect on a classifier's performance. This improvement can be seen from SupervisedV2 and SupervisedV2+DA where we see a 5-10 times performance uplift when recall is below 80% in Table 3 and Table 4. This is in line with expectations since we assumed that simulated lenses would look different from their real counterparts. The good news is that data augmentation can compensate for this, and it substantially closes the gap so that the classifier is usable. Higher recall levels are more challenging however and we see this reflected in those tables as well. For recall above 80%, we see that data augmentation can no longer compensate, which suggests that the simulated data in the training sets were unable to capture the relevant details for those remaining cases.

4.2 SSL Algorithms Lead The Charts

When we look at average precision, we see that the most performant algorithms are all SSL algorithms. Adding GAN-generated lenses to the training set also offers a meaningful boost to performance. From Table 3 and Table 4 we see that GAN+MixMatch is first or second in terms of precision at all recall levels. This model is also a substantial improvement over the baseline supervised model, SupervisedV2+DA, by nearly doubling the performance in precision at all recall levels.

Let us summarize some key findings from Table 3 and Table 4. The first thing to notice is that training with more non-lenses seems to hurt performance. This seemingly counterintuitive result can be seen in the performance gap between SupervisedV1+DA and SupervisedV2+DA. Note that this decrease in performance only really happens when the additional non-lenses are used in a supervised fashion. On the other hand, when we treat the additional non-lenses as unlabeled and use them in the SSL algorithms we see a boost in performance. This is most likely due to the fact that supervised losses tend to "push" the decision boundary for the non-lenses into regions where true lenses reside; whereas the unsupervised losses tend to regulate and refine the preexisting boundary defined by the labeled non-lenses, causing a "pull" or retraction of the decision boundary. Furthermore, this pull effect moreso comes from consistency regularization rather than entropy minimization since we see roughly the same performance between SupervisedV2+DA and Pseudo-label, which does pure entropy minimization. This also suggests that using more simulated or GAN-generated lenses should be beneficial because we want to push the decision boundary into the regions where true lenses reside. Lastly, the improvement in performance from using GAN generated samples also suggests possibly that asymmetry is important between the non-lenses and simulated lenses. By this, we mean that the source images for generating the simulated lenses should not also be used in training as non-lenses. We explain and justify these findings in more detail in the following sections.

4.3 More Non-lenses In The Training Set Eventually Decreases Performance

Recall that TrainingV1 is the larger training set that consists of nearly the whole survey and the correspond-

Table 3: Average Precision (%) For TestV1

Model	Recall 50%	Recall 60%	Recall 70%	Recall 80%	Recall 90%	Recall 100%
SupervisedV1	3.4 ± 0.10	3.99 ± 0.15	4.35 ± 0.10	4.67 ± 0.07	5.14 ± 0.03	5.62 ± 0.01
SupervisedV1+DA	18.32 ± 6.59	12.70 ± 5.39	7.68 ± 1.78	5.99 ± 0.48	5.71 ± 0.24	5.66 ± 0.03
SupervisedV2	17.93 ± 9.00	10.09 ± 7.62	8.39 ± 6.07	6.85 ± 3.13	5.64 ± 0.85	5.65 ± 0.02
SupervisedV2+DA	65.46 ± 15.00	55.54 ± 13.13	46.45 ± 14.49	33.02 ± 10.26	22.60 ± 4.78	8.13 ± 2.49
MixMatch	89.67 ± 6.70	72.81 ± 11.95	54.58 ± 18.8	40.77 ± 11.95	30.33 ± 10.06	12.28 ± 5.09
Pseudo-label	64.17 ± 16.67	54.81 ± 18.02	46.20 ± 23.08	26.56 ± 14.91	10.62 ± 6.13	5.75 ± 0.23
Mean Teacher	81.30 ± 6.08	73.75 ± 4.04	63.01 ± 6.61	43.43 ± 7.15	24.43 ± 2.66	6.61 ± 0.99
Π-Model	75.86 ± 8.35	67.71 ± 9.16	54.54 ± 10.09	43.94 ± 12.35	30.74 ± 3.77	13.41 ± 2.33
VAT	68.17 ± 9.00	50.47 ± 9.08	43.85 ± 7.39	33.69 ± 3.81	20.72 ± 11.50	9.25 ± 5.7
GAN + Supervised	84.95 ± 8.70	79.88 ± 6.30	69.42 ± 4.60	54.35 ± 4.57	34.12 ± 7.47	8.25 ± 2.85
GAN + MixMatch	$\mathbf{96.86 \pm 3.70}$	93.93 ± 4.37	$\mathbf{83.90 \pm 1.92}$	63.90 ± 5.69	$\mathbf{46.56 \pm 9.83}$	14.13 ± 6.53
GAN + Pseudo-label	88.3 ± 7.84	84.00 ± 8.18	78.27 ± 11.24	48.16 ± 7.67	28.53 ± 13.21	6.87 ± 1.33
GAN + Mean Teacher	94.7 ± 4.45	$\mathbf{94.32 \pm 5.07}$	80.83 ± 4.92	67.33 ± 3.22	36.95 ± 6.50	$\mathbf{15.57 \pm 6.50}$
GAN + Π-Model	92.89 ± 1.36	86.14 ± 12.7	79.1 ± 18.02	$\mathbf{69.24 \pm 17.25}$	41.47 ± 12.87	15.2 ± 6.21
GAN + VAT	87.96 ± 10.23	75.15 ± 15.2	60.51 ± 12.40	48.92 ± 13.48	24.62 ± 6.46	11.15 ± 8.77

Table 4: Average Precision (%) For TestV2

Model	Recall 50%	Recall 60%	Recall 70%	Recall 80%	Recall 90%	Recall 100%
SupervisedV1	1.90 ± 0.09	2.20 ± 0.07	2.39 ± 0.07	2.63 ± 0.06	2.82 ± 0.03	3.01 ± 0.02
SupervisedV1+DA	8.55 ± 3.92	3.81 ± 0.99	3.03 ± 0.15	2.97 ± 0.13	2.92 ± 0.04	3.03 ± 0.03
SupervisedV2	8.55 ± 3.75	4.74 ± 3.41	4.23 ± 2.86	3.69 ± 1.60	2.98 ± 0.17	3.06 ± 0.04
SupervisedV2+DA	55.77 ± 17.43	44.09 ± 12.94	37.24 ± 12.85	24.37 ± 10.50	10.99 ± 4.25	5.79 ± 3.93
MixMatch	83.54 ± 11.31	77.98 ± 12.13	51.26 ± 17.7	29.44 ± 13.24	12.06 ± 3.17	6.84 ± 3.00
Pseudo-label	53.06 ± 20.53	47.04 ± 18.00	37.44 ± 23.18	17.99 ± 18	4.18 ± 2.16	3.07 ± 0.13
Mean Teacher	72.35 ± 6.79	65.05 ± 4.41	60.09 ± 4.09	28.96 ± 13.14	5.74 ± 2.81	3.55 ± 0.55
Π-Model	63.33 ± 10.34	55.66 ± 9.53	44.98 ± 10.73	30.48 ± 8.12	15.96 ± 1.09	8.68 ± 1.49
VAT	61.93 ± 10.11	46.85 ± 11.96	40.21 ± 11.91	21.37 ± 4.59	9.43 ± 7.97	7.79 ± 8.68
GAN + Supervised	80.19 ± 17.08	78.03 ± 12.45	68.05 ± 12.96	40.98 ± 17.12	16.33 ± 8.661	6.05 ± 2.69
GAN + MixMatch	$\mathbf{96.88 \pm 6.25}$	$\mathbf{92.84 \pm 5.69}$	78.90 ± 5.65	39.51 ± 9.01	20.54 ± 11.69	7.97 ± 3.93
GAN + Pseudo-label	80.58 ± 10.88	77.75 ± 14.94	70.42 ± 13.68	30.36 ± 3.56	9.49 ± 3.91	3.69 ± 0.74
GAN + Mean Teacher	89.51 ± 8.70	87.99 ± 12.04	$\mathbf{81.83 \pm 14.03}$	46.41 ± 18.12	14.61 ± 5.22	8.87 ± 3.92
GAN + Π-Model	85.62 ± 2.38	74.9 ± 13.99	69.58 ± 17.84	$\mathbf{54.26 \pm 13.05}$	$\mathbf{30.45 \pm 8.72}$	$\mathbf{22.27 \pm 7.71}$
GAN + VAT	78.84 ± 17.21	60.40 ± 4.77	44.46 ± 10.12	27.29 ± 7.02	11.99 ± 5.15	8.63 ± 5.71

ing simulated lenses. TrainingV2 on the other hand is the smaller training set made from human-labeled non-lenses and the corresponding simulated lenses. As we mentioned before, there is a noticeable performance gap between models trained on TrainingV1 versus TrainingV2. This can be seen from the difference in precision between SupervisedV1+DA and SupervisedV2+DA in Table 3 and Table 4. This performance gap isn't due to overfitting since validation and testing accuracy on simulated data was nearly perfect. Instead, the drop in performance seems to be caused by the increasing likelihood of non-lenses, which look like lenses, being present in the data that pushes the non-lens decision boundary further into the regions where real lenses reside. To see this we randomly sampled images of non-lenses(z_{nl}), simulated lenses(z_{sl}) and real lenses(z_{rl}) and interpolated between them(z) to see how the decision boundary changes between these points.

$$z = z_{nl}(1 - x - y) + z_{sl}(y) + z_{rl}(x) \tag{1}$$

Contour plots of the predicted likelihood that the image is a lens are given in Figure 2, where the x-axis and y-axis correspond to the x and y values in the interpolation formula above. Notice how the contour lines are pushed further away from the real lens in SupervisedV1+DA than SupervisedV2+DA. We also notice that the contour lines are bundled in a much tighter neighborhood for SupervisedV1+DA than SupervisedV2+DA. The effect of tighter neighborhoods is poorer generalization because it leads to less granularity for thresholding. This increases the likelihood of false positives being selected since the regions intersect more highly with the regions containing real lenses. If we look at GAN+MixMatch, we see that we can overcome these limitations by adding GAN-generated lenses to the training set. Notice also that the decision boundary gets pushed further into the region where real lenses reside allowing a threshold on the output to reliably distinguish between real lenses and non-lenses.

To further confirm our assertions we ran another set of experiments where we used TrainingV1 but varied the number of non-lenses in the training set. We train a supervised model on these different training sets. The results can be seen in Table 5. Overall, we see that while increasing the number of non-lenses increases precision

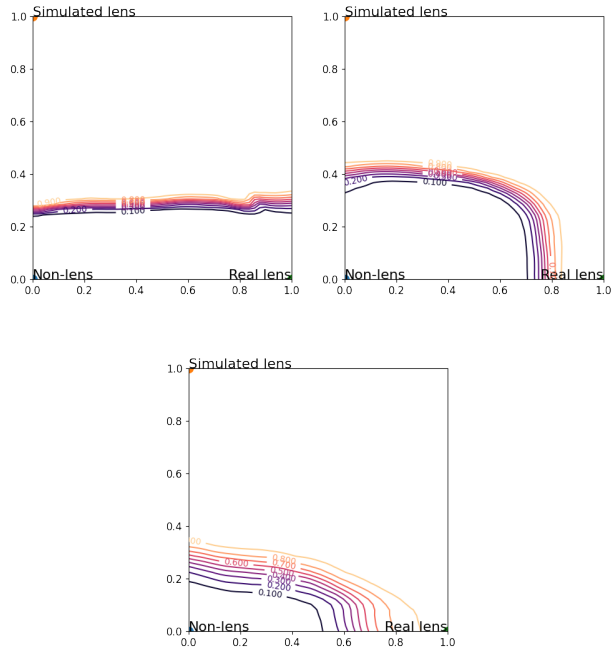

Figure 2: Linear interpolation between randomly sampled non-lens, simulated lens, and real lens for methods: SupervisedV1+DA (top left), SupervisedV2+DA (top right), GAN+MixMatch (bottom).

initially, it will eventually peak and from there it dramatically falls in value until it plateaus.

4.4 Performance Breakdown By Object Type

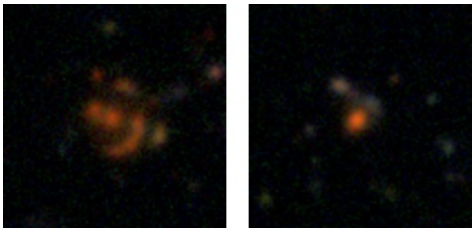

Figure 3: Frequently misclassified lenses

We also looked into whether the performance of the models was consistent based on the properties of the central galaxy in each image. From our analysis, there wasn't any indication that the particular galaxy type affected the model's ability to correctly classify lenses. This is reassuring since the lens simulations used in our training data (see 3.1) are not dependent on any physical property of the central galaxy. The performance of models also remains consistent across different fields of the survey. Instead, the main obstacle seems to just be the coloring of the arc and the brightness of the arc itself. As we can see in Figure 3, orange arcs are misclassified as non-lenses simply because our training set doesn't account for them. Data augmentations by themselves cannot compensate for this because there is no way to turn the red-orange object in the middle and the blue arc to simultaneously become red-orange in color. There are also faint arcs that we can't properly capture because the arcs used in our simulations are comparatively brighter.

4.5 Data Augmentation Ablation Study

The ablation study was performed with GAN+Supervised as our model. As we can see from Table 6 (see Tables 7-12 in the Supplementary section for more recall values), the most impactful augmentations were RGB shuffle, JPEG quality and GAN. Nearly, 40% of the precision we are seeing comes directly from RGB shuffling. The reason RGB shuffle is one of the most effective augmentations is that we don't know a priori how real lenses would look in our survey. As a result, the simulated lens we produce will never perfectly align with the channel intensities of the arcs from real lenses. By shuffling the channels we allow the model to be less sensitive on the coloration of the arc and more on the structural properties of the arc in relation to the surrounding objects. This doesn't mean that it's not important to create realistically colored arcs for the simulations, but rather we can compensate for it when it is significantly off from the coloration of arcs around real lenses. JPEG quality and GAN also were significant contributors to the overall performance of the model.

4.6 Spectroscopic Lens Confirmation

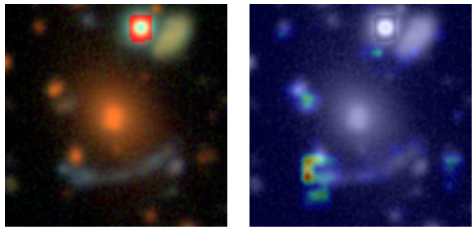

Figure 4: Spectroscopically confirmed lens system DLS212072337(left), Grad-CAM++ overlay from GAN+MixMatch(p-value: 1.00, right), Red-Green-Blue in overlay indicate High-Mid-Low importance of the feature to the model when making the prediction.

To confirm the lensing nature of these systems, spectroscopic redshifts are required to establish that the candidate deflector is indeed a massive galaxy, and that the arc is located at a greater distance (higher redshift) than the deflector. We obtained spectra of a high confidence lens found from this work (DLS212072337;

Table 5: Performance Scaling Of Number Of Non-lenses In TrainingV1 On TestV1 Precision(%)

Number of non-lenses	Recall 50%	Recall 60%	Recall 70%	Recall 80%	Recall 90%	Recall 100%
1000	16.61	12.5	9.38	8.46	6.66	5.76
2000	44.72	23.71	21.29	12.36	5.73	5.62
4000	33.55	33.34	24.79	13.8	8.89	5.7
8000	56.56	52.46	38.96	30.34	13.9	5.96
16000	50	42.39	37.95	34.15	14.35	8.14
32000	49.07	35.56	33.49	26.38	17.4	8.33
64000	11.71	9.69	9.09	6.86	5.75	5.62
128000	13.17	8.14	6.04	5.48	5.34	5.62
256000	17.10	10.52	6.83	5.12	5.33	5.66

Table 6: Ablation performance for 80% Recall

Augmentation removed	TestV1 Precision(%)	TestV2 Precision(%)	TestV1 difference(%)	TestV2 difference(%)
-	54.35 ± 4.57	40.98 ± 17.12	-	-
GAN	33.02 ± 10.26	24.37 ± 10.50	-21.33	-16.61
RGB shuffle	15.34 ± 7.59	11.75 ± 4.90	-39.01	-29.23
JPEG quality	21.33 ± 8.29	15.25 ± 3.41	-33.02	-25.73
Rot90	62.63 ± 12.43	52.2 ± 13.65	+8.28	+11.22
Translations	52.17 ± 12.92	35.09 ± 16.64	-2.18	-5.89
Horizontal flips	60.54 ± 9.78	36.46 ± 9.03	+6.19	-4.52
Color augmentation	39.47 ± 13.24	33.84 ± 10.92	14.88	-7.14

see Figure 4) using the NIRES instrument (Wilson et al., 2004) at W.M. Keck Observatory on the night of 31 March 2021, and measured a secure redshift for the arc of $z = 1.81$ via [O III] $\lambda\lambda$4959,5007 and Hα emission lines (Tran et al., in prep). The central deflector is known to be a massive galaxy at a lower redshift of $z = 0.43$ from BOSS (Dawson et al., 2013), thus confirming this as a true lens system. While this is only a single object, these results demonstrate the feasibility of characterizing a larger sample identified from this work.

5 Discussion

We have shown in this paper that combining data augmentation, SSL algorithms, and GAN-generated lenses significantly improves the lens classifier performance. In terms of data augmentation, RGB-shuffle produced a significant boost to the lens classifier's performance. Of the SSL algorithms, we found that consistency regularization is the dominant factor as to why they outperform supervised baselines. Figure 2 shows us that consistency regularization more definitively separates the non-lenses from the lenses. Our findings indicate that one must take care when balancing the number of non-lenses and simulated lenses in the training data, as increasing one does not necessarily translate to better performance (see Table 5).

In addition to improving the coverage of non-lenses, we also realized that a greater diversity of simulated lenses was needed. Data augmentation is unable to compensate for cases where arcs are not blue or when arcs are relatively faint in real lenses. One direction worth exploring is if there are any beneficial augmentations that can be applied to the arcs themselves. Since simulated lenses consist of superimposing an image of an arc onto a non-lens image these augmentations can also be generated on the fly. **The effect of additional data augmentations specific to the survey (e.g., varying the Point Spread Function) might also be worth exploring.** Another possible improvement would be to create multiple different classes of simulated lenses based on the binned color and brightness of the arcs, turning this binary classification problem into a multiclass classification problem. Breaking up the lens class into multiple subclasses may also offer improvements in performance as well (Hoffmann et al., 2001; Luo, 2008).

An additional approach that we didn't explore in this paper is to take the labeled data and combine it with these various simulated lenses and learn representations through self-supervised learning(Chen et al., 2020; Caron et al., 2020; Grill et al., 2020). Recently, other researchers have started looking into these approaches (Hayat et al., 2021; Stein et al., 2021). The latter preprint is applied to lens discovery in the much larger DESI survey, and simulations are not used in training. Another similar approach by (Cheng et al., 2020) utilizes a convolutional autoencoder to construct an embedding of the simulated strong lens images and fit a Gaussian mixture model to extract the clusters in the embedding, which corresponds to various visual features of the galaxies and lenses. The results are not directly comparable to our own due to differences in the experimental setup.

In this work, we use a two-stage process where a pilot model is used to obtain the test set and then use this to select the best out of a handful of models to obtain better candidate lenses. In practice, one may want to try out several variations of our top-performing methods without the two-stage approach. One way to achieve this is to "recommend" to the astronomer potential lenses in a streaming fashion. We can then associate a reward to each newly discovered lens and use a multi-armed bandit algorithm to perform online model selection.

In summary, we have isolated several key ingredients that are essential to training a lens classifier using only simulations in an astronomical survey. This method is able to improve on existing models by an order of magnitude, and it has already led to the discovery and confirmation of a novel gravitational lens system.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia. PMLR.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 5049–5059. Curran Associates, Inc.

Cañameras, R., Schuldt, S., Suyu, S. H., Taubenberger, S., Meinhardt, T., Leal-Taixé, L., Lemon, C., Rojas, K., and Savary, E. (2020). HOLISMOKES. II. Identifying galaxy-scale strong gravitational lenses in Pan-STARRS using convolutional neural networks. A&A, 644:A163.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Cheng, T.-Y., Li, N., Conselice, C. J., Aragón-Salamanca, A., Dye, S., and Metcalf, R. B. (2020). Identifying strong lenses with unsupervised machine learning using convolutional autoencoder. *Monthly Notices of the Royal Astronomical Society*, 494(3):3750–3765.

Dawson, K. S., Schlegel, D. J., Ahn, C. P., Anderson, S. F., Aubourg, É., Bailey, S., Barkhouser, R. H., Bautista, J. E., Beifiori, A., Berlind, A. A., Bhardwaj, V., Bizyaev, D., Blake, C. H., Blanton, M. R., Blomqvist, M., Bolton, A. S., Borde, A., Bovy, J., Brandt, W. N., Brewington, H., Brinkmann, J., Brown, P. J., Brownstein, J. R., Bundy, K., Busca, N. G., Carithers, W., Carnero, A. R., Carr, M. A., Chen, Y., Comparat, J., Connolly, N., Cope, F., Croft, R. A. C., Cuesta, A. J., da Costa, L. N., Davenport, J. R. A., Delubac, T., de Putter, R., Dhital, S., Ealet, A., Ebelke, G. L., Eisenstein, D. J., Escoffier, S., Fan, X., Filiz Ak, N., Finley, H., Font-Ribera, A., Génova-Santos, R., Gunn, J. E., Guo, H., Haggard, D., Hall, P. B., Hamilton, J.-C., Harris, B., Harris, D. W., Ho, S., Hogg, D. W., Holder, D., Honscheid, K., Huehnerhoff, J., Jordan, B., Jordan, W. P., Kauffmann, G., Kazin, E. A., Kirkby, D., Klaene, M. A., Kneib, J.-P., Le Goff, J.-M., Lee, K.-G., Long, D. C., Loomis, C. P., Lundgren, B., Lupton, R. H., Maia, M. A. G., Makler, M., Malanushenko, E., Malanushenko, V., Mandelbaum, R., Manera, M., Maraston, C., Margala, D., Masters, K. L., McBride, C. K., McDonald, P., McGreer, I. D., McMahon, R. G., Mena, O., Miralda-Escudé, J., Montero-Dorta, A. D., Montesano, F., Muna, D., Myers, A. D., Naugle, T., Nichol, R. C., Noterdaeme, P., Nuza, S. E., Olmstead, M. D., Oravetz, A., Oravetz, D. J., Owen, R., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Parejko, J. K., Pâris, I., Percival, W. J., Pérez-Fournon, I., Pérez-Ràfols, I., Petitjean, P., Pfaffenberger, R., Pforr, J., Pieri, M. M., Prada, F., Price-Whelan, A. M., Raddick, M. J., Rebolo, R., Rich, J., Richards, G. T., Rockosi, C. M., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Rubiño-Martin, J. A., Samushia, L., Sánchez, A. G., Sayres, C., Schmidt, S. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Shelden, A., Sheldon, E., Shen, Y., Shu, Y., Slosar, A., Smee, S. A., Snedden, S. A., Stauffer, F., Steele, O., Strauss, M. A., Streblyanska, A., Suzuki, N., Swanson, M. E. C., Tal, T., Tanaka, M., Thomas, D., Tinker, J. L., Tojeiro, R., Tremonti, C. A., Vargas Magaña, M., Verde, L., Viel, M., Wake, D. A., Watson, M., Weaver, B. A., Weinberg, D. H., Weiner, B. J., West, A. A., White, M., Wood-Vasey, W. M., Yeche, C., Zehavi, I., Zhao, G.-B., and Zheng, Z. (2013). The Baryon Oscillation Spectroscopic Survey of SDSS-III. AJ, 145(1):10.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA. MIT Press.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. (2020). Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., and Mustafa, M. (2021). Self-supervised representation learning for astronomical images. 911(2):L33.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer.

Hoffmann, A., Kwok, R., and Compton, P. (2001). Using subclasses to improve classification learning. In De Raedt, L. and Flach, P., editors, *Machine Learning: ECML 2001*, pages 203–213, Berlin, Heidelberg. Springer Berlin Heidelberg.

Huang, X., Storfer, C., Ravi, V., Pilon, A., Domingo, M., Schlegel, D. J., Bailey, S., Dey, A., Gupta, R. R., Herrera, D., Juneau, S., Landriau, M., Lang, D., Meisner, A., Moustakas, J., Myers, A. D., Schlafly, E. F., Valdes, F., Weaver, B. A., Yang, J., and Yèche, C. (2020). Finding strong gravitational lenses in the DESI DECam legacy survey. *The Astrophysical Journal*, 894(1):78.

Jacobs, C., Collett, T., Glazebrook, K., Buckley-Geer, E., Diehl, H. T., Lin, H., McCarthy, C., Qin, A. K., Odden, C., Caso Escudero, M., Dial, P., Yung, V. J., Gaitsch, S., Pellico, A., Lindgren, K. A., Abbott, T. M. C., Annis, J., Avila, S., Brooks, D., Burke, D. L., Carnero Rosell, A., Carrasco Kind, M., Carretero, J., da Costa, L. N., De Vicente, J., Fosalba, P., Frieman, J., García-Bellido, J., Gaztanaga, E., Goldstein, D. A., Gruen, D., Gruendl, R. A., Gschwend, J., Hollowood, D. L., Honscheid, K., Hoyle, B., James, D. J., Krause, E., Kuropatkin, N., Lahav, O., Lima, M., Maia, M. A. G., Marshall, J. L., Miquel, R., Plazas, A. A., Roodman, A., Sanchez, E., Scarpine, V., Serrano, S., Sevilla-Noarbe, I., Smith, M., Sobreira, F., Suchyta, E., Swanson, M. E. C., Tarle, G., Vikram, V., Walker, A. R., Zhang, Y., and DES Collaboration (2019). An Extended Catalog of Galaxy-Galaxy Strong Gravitational Lenses Discovered in DES Using Convolutional Neural Networks. ApJS, 243(1):17.

Jacobs, C., Glazebrook, K., Collett, T., More, A., and McCarthy, C. (2017). Finding strong lenses in CFHTLS using convolutional neural networks. MNRAS, 471(1):167–181.

Kurakin, A., Li, C.-L., Raffel, C., Berthelot, D., Cubuk, E. D., Zhang, H., Sohn, K., Carlini, N., and Zhang, Z. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*.

Laine, S. and Aila, T. (2017). Temporal ensembling for semi-supervised learning. *ArXiv*, abs/1610.02242.

Laureijs, R., Amiaux, J., Arduini, S., Augueres, J. L., Brinchmann, J., Cole, R., Cropper, M., Dabin, C., Duvet, L., Ealet, A., Garilli, B., Gondoin, P., Guzzo, L., Hoar, J., Hoekstra, H., Holmes, R., Kitching, T., Maciaszek, T., Mellier, Y., Pasian, F., Percival, W., Rhodes, J., Saavedra Criado, G., Sauvage, M., Scaramella, R., Valenziano, L., Warren, S., Bender, R., Castander, F., Cimatti, A., Le Fèvre, O., Kurki-Suonio, H., Levi, M., Lilje, P., Meylan, G., Nichol, R., Pedersen, K., Popa, V., Rebolo Lopez, R., Rix, H. W., Rottgering, H., Zeilinger, W., Grupp, F., Hudelot, P., Massey, R., Meneghetti, M., Miller, L., Paltani, S., Paulin-Henriksson, S., Pires, S., Saxton, C., Schrabback, T., Seidel, G., Walsh, J., Aghanim, N., Amendola, L., Bartlett, J., Baccigalupi, C., Beaulieu, J. P., Benabed, K., Cuby, J. G., Elbaz, D., Fosalba, P., Gavazzi, G., Helmi, A., Hook, I., Irwin, M., Kneib, J. P., Kunz, M., Mannucci, F., Moscardini, L., Tao, C., Teyssier, R., Weller, J., Zamorani, G., Zapatero Osorio, M. R., Boulade, O., Foumond, J. J., Di Giorgio, A., Guttridge, P., James, A., Kemp, M., Martignac, J., Spencer, A., Walton, D., Blümchen, T., Bonoli, C., Bortoletto, F., Cerna, C., Corcione, L., Fabron, C., Jahnke, K., Ligori, S., Madrid, F., Martin, L., Morgante, G., Pamplona, T., Prieto, E., Riva, M., Toledo, R., Trifoglio, M., Zerbi, F., Abdalla, F., Douspis, M., Grenet, C., Borgani, S., Bouwens, R., Courbin, F.,

Delouis, J. M., Dubath, P., Fontana, A., Frailis, M., Grazian, A., Koppenhöfer, J., Mansutti, O., Melchior, M., Mignoli, M., Mohr, J., Neissner, C., Noddle, K., Poncet, M., Scodeggio, M., Serrano, S., Shane, N., Starck, J. L., Surace, C., Taylor, A., Verdoes-Kleijn, G., Vuerli, C., Williams, O. R., Zacchei, A., Altieri, B., Escudero Sanz, I., Kohley, R., Oosterbroek, T., Astier, P., Bacon, D., Bardelli, S., Baugh, C., Bellagamba, F., Benoist, C., Bianchi, D., Biviano, A., Branchini, E., Carbone, C., Cardone, V., Clements, D., Colombi, S., Conselice, C., Cresci, G., Deacon, N., Dunlop, J., Fedeli, C., Fontanot, F., Franzetti, P., Giocoli, C., Garcia-Bellido, J., Gow, J., Heavens, A., Hewett, P., Heymans, C., Holland, A., Huang, Z., Ilbert, O., Joachimi, B., Jennins, E., Kerins, E., Kiessling, A., Kirk, D., Kotak, R., Krause, O., Lahav, O., van Leeuwen, F., Lesgourgues, J., Lombardi, M., Magliocchetti, M., Maguire, K., Majerotto, E., Maoli, R., Marulli, F., Maurogordato, S., McCracken, H., McLure, R., Melchiorri, A., Merson, A., Moresco, M., Nonino, M., Norberg, P., Peacock, J., Pello, R., Penny, M., Pettorino, V., Di Porto, C., Pozzetti, L., Quercellini, C., Radovich, M., Rassat, A., Roche, N., Ronayette, S., Rossetti, E., Sartoris, B., Schneider, P., Semboloni, E., Serjeant, S., Simpson, F., Skordis, C., Smadja, G., Smartt, S., Spano, P., Spiro, S., Sullivan, M., Tilquin, A., Trotta, R., Verde, L., Wang, Y., Williger, G., Zhao, G., Zoubian, J., and Zucca, E. (2011). Euclid Definition Study Report. *arXiv e-prints*, page arXiv:1110.3193.

Lee, D.-H. (2013). Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.

Li, R., Napolitano, N. R., Tortora, C., Spiniello, C., Koopmans, L. V. E., Huang, Z., Roy, N., Vernardos, G., Chatterjee, S., Giblin, B., Getman, F., Radovich, M., Covone, G., and Kuijken, K. (2020). New High-quality Strong Lens Candidates with Deep Learning in the Kilo-Degree Survey. ApJ, 899(1):30.

LSST Science Collaboration, Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., Armus, L., Arnett, D., Asztalos, S. J., Axelrod, T. S., Bailey, S., Ballantyne, D. R., Bankert, J. R., Barkhouse, W. A., Barr, J. D., Barrientos, L. F., Barth, A. J., Bartlett, J. G., Becker, A. C., Becla, J., Beers, T. C., Bernstein, J. P., Biswas, R., Blanton, M. R., Bloom, J. S., Bochanski, J. J., Boeshaar, P., Borne, K. D., Bradac, M., Brandt, W. N., Bridge, C. R., Brown, M. E., Brunner, R. J., Bullock, J. S., Burgasser, A. J., Burge, J. H., Burke, D. L., Cargile, P. A., Chandrasekharan, S., Chartas, G., Chesley, S. R., Chu, Y.-H., Cinabro, D., Claire, M. W., Claver, C. F., Clowe, D., Connolly, A. J., Cook, K. H., Cooke, J., Cooray, A., Covey, K. R., Culliton, C. S., de Jong, R., de Vries, W. H., Debattista, V. P., Delgado, F., Dell'Antonio, I. P., Dhital, S., Di Stefano, R., Dickinson, M., Dilday, B., Djorgovski, S. G., Dobler, G., Donalek, C., Dubois-Felsmann, G., Durech, J., Eliasdottir, A., Eracleous, M., Eyer, L., Falco, E. E., Fan, X., Fassnacht, C. D., Ferguson, H. C., Fernandez, Y. R., Fields, B. D., Finkbeiner, D., Figueroa, E. E., Fox, D. B., Francke, H., Frank, J. S., Frieman, J., Fromenteau, S., Furqan, M., Galaz, G., Gal-Yam, A., Garnavich, P., Gawiser, E., Geary, J., Gee, P., Gibson, R. R., Gilmore, K., Grace, E. A., Green, R. F., Gressler, W. J., Grillmair, C. J., Habib, S., Haggerty, J. S., Hamuy, M., Harris, A. W., Hawley, S. L., Heavens, A. F., Hebb, L., Henry, T. J., Hileman, E., Hilton, E. J., Hoadley, K., Holberg, J. B., Holman, M. J., Howell, S. B., Infante, L., Ivezic, Z., Jacoby, S. H., Jain, B., R, Jedicke, Jee, M. J., Garrett Jernigan, J., Jha, S. W., Johnston, K. V., Jones, R. L., Juric, M., Kaasalainen, M., Styliani, Kafka, Kahn, S. M., Kaib, N. A., Kalirai, J., Kantor, J., Kasliwal, M. M., Keeton, C. R., Kessler, R., Knezevic, Z., Kowalski, A., Krabbendam, V. L., Krughoff, K. S., Kulkarni, S., Kuhlman, S., Lacy, M., Lepine, S., Liang, M., Lien, A., Lira, P., Long, K. S., Lorenz, S., Lotz, J. M., Lupton, R. H., Lutz, J., Macri, L. M., Mahabal, A. A., Mandelbaum, R., Marshall, P., May, M., McGehee, P. M., Meadows, B. T., Meert, A., Milani, A., Miller, C. J., Miller, M., Mills, D., Minniti, D., Monet, D., Mukadam, A. S., Nakar, E., Neill, D. R., Newman, J. A., Nikolaev, S., Nordby, M., O'Connor, P., Oguri, M., Oliver, J., Olivier, S. S., Olsen, J. K., Olsen, K., Olszewski, E. W., Oluseyi, H., Padilla, N. D., Parker, A., Pepper, J., Peterson, J. R., Petry, C., Pinto, P. A., Pizagno, J. L., Popescu, B., Prsa, A., Radcka, V., Raddick, M. J., Rasmussen, A., Rau, A., Rho, J., Rhoads, J. E., Richards, G. T., Ridgway, S. T., Robertson, B. E., Roskar, R., Saha, A., Sarajedini, A., Scannapieco, E., Schalk, T., Schindler, R., Schmidt, S., Schmidt, S., Schneider, D. P., Schumacher, G., Scranton, R., Sebag, J., Seppala, L. G., Shemmer, O., Simon, J. D., Sivertz, M., Smith, H. A., Allyn Smith, J., Smith, N., Spitz, A. H., Stanford, A., Stassun, K. G., Strader, J., Strauss, M. A., Stubbs, C. W., Sweeney, D. W., Szalay, A., Szkody, P., Takada, M., Thorman, P., Trilling, D. E., Trimble, V., Tyson, A., Van Berg, R., Vanden Berk, D., VanderPlas, J., Verde, L., Vrsnak, B., Walkowicz, L. M., Wandelt, B. D., Wang, S., Wang, Y., Warner, M., Wechsler, R. H., West, A. A., Wiecha, O., Williams, B. F., Willman, B., Wittman, D., Wolff, S. C., Wood-Vasey, W. M., Wozniak, P., Young, P., Zent-

ner, A., and Zhan, H. (2009). LSST Science Book, Version 2.0. *arXiv e-prints*, page arXiv:0912.0201.

Luo, Y. (2008). Can subclasses help a multiclass learning problem? In *2008 IEEE Intelligent Vehicles Symposium*, pages 214–219.

Marshall, P., Sandford, C., More, A., and Buddelmeijerr, H. (2015). HumVI: Human Viewable Image creation.

Metcalf, R. B., Meneghetti, M., Avestruz, C., Bellagamba, F., Bom, C. R., Bertin, E., Cabanac, R., Courbin, F., Davies, A., Decencière, E., Flamary, R., Gavazzi, R., Geiger, M., Hartley, P., Huertas-Company, M., Jackson, N., Jacobs, C., Jullo, E., Kneib, J. P., Koopmans, L. V. E., Lanusse, F., Li, C. L., Ma, Q., Makler, M., Li, N., Lightman, M., Petrillo, C. E., Serjeant, S., Schäfer, C., Sonnenfeld, A., Tagore, A., Tortora, C., Tuccillo, D., Valentín, M. B., Velasco-Forero, S., Verdoes Kleijn, G. A., and Vernardos, G. (2019). The strong gravitational lens finding challenge. A&A, 625:A119.

Miyato, T., Maeda, S., Koyama, M., and Ishii, S. (2019). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993.

Nikolenko, S. I. (2019). Synthetic data for deep learning. *Synthetic Data for Deep Learning*.

Oguri, M. (2010). glafic: Software Package for Analyzing Gravitational Lensing.

Oguri, M. and Marshall, P. J. (2010). Gravitationally lensed quasars and supernovae in future wide-field optical imaging surveys. MNRAS, 405(4):2579–2593.

Oliver, A., Odena, A., Raffel, C., Cubuk, E. D., and Goodfellow, I. J. (2018). Realistic evaluation of deep semi-supervised learning algorithms. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3239–3250, Red Hook, NY, USA. Curran Associates Inc.

Pourrahmani, M., Nayyeri, H., and Cooray, A. (2018). LensFlow: A Convolutional Neural Network in Search of Strong Gravitational Lenses. ApJ, 856(1):68.

Schmidt, S. J. and Thorman, P. (2013). Improved photometric redshifts via enhanced estimates of system response, galaxy templates and magnitude priors. MNRAS, 431(3):2766–2777.

Sonnenfeld, A., Chan, J. H. H., Shu, Y., More, A., Oguri, M., Suyu, S. H., Wong, K. C., Lee, C.-H., Coupon, J., Yonehara, A., Bolton, A. S., Jaelani, A. T., Tanaka, M., Miyazaki, S., and Komiyama, Y. (2018). Survey of Gravitationally-lensed Objects in HSC Imaging (SuGOHI). I. Automatic search for galaxy-scale strong lenses. PASJ, 70:S29.

Spergel, D., Gehrels, N., Baltay, C., Bennett, D., Breckinridge, J., Donahue, M., Dressler, A., Gaudi, B. S., Greene, T., Guyon, O., Hirata, C., Kalirai, J., Kasdin, N. J., Macintosh, B., Moos, W., Perlmutter, S., Postman, M., Rauscher, B., Rhodes, J., Wang, Y., Weinberg, D., Benford, D., Hudson, M., Jeong, W. S., Mellier, Y., Traub, W., Yamada, T., Capak, P., Colbert, J., Masters, D., Penny, M., Savransky, D., Stern, D., Zimmerman, N., Barry, R., Bartusek, L., Carpenter, K., Cheng, E., Content, D., Dekens, F., Demers, R., Grady, K., Jackson, C., Kuan, G., Kruk, J., Melton, M., Nemati, B., Parvin, B., Poberezhskiy, I., Peddie, C., Ruffa, J., Wallace, J. K., Whipple, A., Wollack, E., and Zhao, F. (2015). Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report. *arXiv e-prints*, page arXiv:1503.03757.

Stein, G., Blaum, J., Harrington, P., Medan, T., and Lukic, Z. (2021). Mining for strong gravitational lenses with self-supervised learning. *arXiv e-prints*, page arXiv:2110.00023.

Tarvainen, A. and Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 1195–1204. Curran Associates, Inc.

van Engelen, J. E. and Hoos, H. H. (2019). A survey on semi-supervised learning. *Machine Learning*, 109:373–440.

Wilson, J. C., Henderson, C. P., Herter, T. L., Matthews, K., Skrutskie, M. F., Adams, J. D., Moon, D.-S., Smith, R., Gautier, N., Ressler, M., Soifer, B. T., Lin, S., Howard, J., LaMarr, J., Stolberg, T. M., and Zink, J. (2004). Mass producing an efficient NIR spectrograph. In Moorwood, A. F. M. and Iye, M., editors, *Ground-based Instrumentation for Astronomy*, volume 5492 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 1295–1305.

Wittman, D. M., Tyson, J. A., Dell'Antonio, I. P., Becker, A., Margoniner, V., Cohen, J. G., Norman, D., Loomba, D., Squires, G., Wilson, G., Stubbs, C. W., Hennawi, J., Spergel, D. N., Boeshaar, P., Clocchiatti, A., Hamuy, M., Bernstein, G., Gonzalez, A., Guhathakurta, P., Hu, W., Seljak, U., and Zaritsky, D. (2002). Deep lens survey. In Tyson,

J. A. and Wolff, S., editors, *Survey and Other Telescope Technologies and Discoveries*, volume 4836 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 73–82.

Yang, X., Song, Z., King, I., and Xu, Z. (2021). A survey on deep semi-supervised learning. *ArXiv*, abs/2103.00550.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.

Supplementary Material:
An Unsupervised Hunt for Gravitational Lenses

A Additional Ablation Results

Table 7: Ablation performance for 50% Recall

Augmentation removed	TestV1 Precision(%)	TestV2 Precision(%)	TestV1 difference(%)	TestV2 difference(%)
-	84.95 ± 8.70	80.19 ± 17.08	-	-
GAN	65.46 ± 15.00	55.77 ± 17.43	-19.49	-24.42
RGB shuffle	44.8 ± 24.32	35.45 ± 21.75	-40.15	-44.74
JPEG quality	65.30 ± 13.84	56.84 ± 14.06	-19.65	-23.35
Rot90	91.81 ± 4.41	89.96 ± 6.33	+6.86	+9.77
Translations	89.47 ± 10.34	84.59 ± 13.04	+4.52	+4.4
Horizontal flips	84.63 ± 10.85	75.74 ± 13.96	-0.32	-4.45
Color augmentation	71.88 ± 17.35	68.23 ± 15.2	-13.07	-11.96

Table 8: Ablation performance for 60% Recall

Augmentation removed	TestV1 Precision(%)	TestV2 Precision(%)	TestV1 difference(%)	TestV2 difference(%)
-	79.88 ± 6.30	78.03 ± 12.45	-	-
GAN	55.54 ± 13.13	44.09 ± 12.94	- 24.34	-33.94
RGB shuffle	34.25 ± 24.49	26.14 ± 15.9	- 45.63	-51.89
JPEG quality	52.75 ± 24.68	51.69 ± 17.91	- 27.13	-26.34
Rot90	84.94 ± 7.04	78.99 ± 5.23	+5.06	+ 0.96
Translations	74.95 ± 16.11	72.26 ± 24.65	-4.93	-5.77
Horizontal flips	71.15 ± 11.46	60.64 ± 12.27	-8.73	-17.39
Color augmentation	63.24 ± 15.67	50.91 ± 15.53	-16.64	-27.12

Table 9: Ablation performance for 70% Recall

Augmentation removed	TestV1 Precision(%)	TestV2 Precision(%)	TestV1 difference(%)	TestV2 difference(%)
-	69.42 ± 4.60	68.05 ± 12.96	-	-
GAN	46.45 ± 14.49	37.24 ± 12.85	-22.97	-30.81
RGB shuffle	26.68 ± 17.28	18.75 ± 11.78	-42.74	-49.3
JPEG quality	40.38 ± 18.64	40.52 ± 21.74	- 29.04	-27.53
Rot90	79.97 ± 12.15	73.69 ± 7.11	+10.55	+5.64
Translations	67.69 ± 12.67	59.31 ± 22.34	-1.73	-8.74
Horizontal flips	67.36 ± 11.95	55.94 ± 14.71	-2.06	-12.11
Color augmentation	52.79 ± 14.92	45.93 ± 13.97	-16.63	-22.12

Table 10: Ablation performance for 80% Recall

Augmentation removed	TestV1 Precision(%)	TestV2 Precision(%)	TestV1 difference(%)	TestV2 difference(%)
-	54.35 ± 4.57	40.98 ± 17.12	-	-
GAN	33.02 ± 10.26	24.37 ± 10.50	-21.33	-16.61
RGB shuffle	15.34 ± 7.59	11.75 ± 4.90	-39.01	-29.23
JPEG quality	21.33 ± 8.29	15.25 ± 3.41	-33.02	-25.73
Rot90	62.63 ± 12.43	52.2 ± 13.65	+8.28	+11.22
Translations	52.17 ± 12.92	35.09 ± 16.64	-2.18	-5.89
Horizontal flips	60.54 ± 9.78	36.46 ± 9.03	+6.19	-4.52
Color augmentation	39.47 ± 13.24	33.84 ± 10.92	14.88	-7.14

Table 11: Ablation performance for 90% Recall

Augmentation removed	TestV1 Precision(%)	TestV2 Precision(%)	TestV1 difference(%)	TestV2 difference(%)
-	34.12 ± 7.47	16.33 ± 8.661	-	-
GAN	22.60 ± 4.78	10.99 ± 4.25	-11.52	-5.34
RGB shuffle	10.37 ± 4.68	5.88 ± 2.61	-23.75	-10.45
JPEG quality	15.15 ± 6.37	6.18 ± 2.41	-18.97	-10.15
Rot90	40.72 ± 12.74	21.79 ± 3.08	+6.6	+5.46
Translations	32.88 ± 3.35	16.14 ± 6.72	-1.24	-0.19
Horizontal flips	30.81 ± 20.44	14.72 ± 6.59	-3.31	-1.61
Color augmentation	23.86 ± 8.13	14.25 ± 8.45	-10.26	-2.08

Table 12: Ablation performance for 100% Recall

Augmentation removed	TestV1 Precision(%)	TestV2 Precision(%)	TestV1 difference(%)	TestV2 difference(%)
-	8.25 ± 2.85	6.05 ± 2.69	-	-
GAN	8.13 ± 2.49	5.79 ± 3.93	-0.12	-0.26
RGB shuffle	6.43 ± 0.97	3.84 ± 1.02	-1.82	-2.21
JPEG quality	5.88 ± 0.21	4.14 ± 2.09	-2.37	-1.91
Rot90	14.76 ± 8.42	13.00 ± 5.16	+6.51	+6.95
Translations	10.83 ± 2.46	7.37 ± 3.66	+2.58	+1.32
Horizontal flips	15.74 ± 3.06	8.86 ± 1.84	+7.49	+2.81
Color augmentation	11.42 ± 7.25	6.84 ± 4.12	+3.17	+0.79

B Performance Breakdown By Object Type

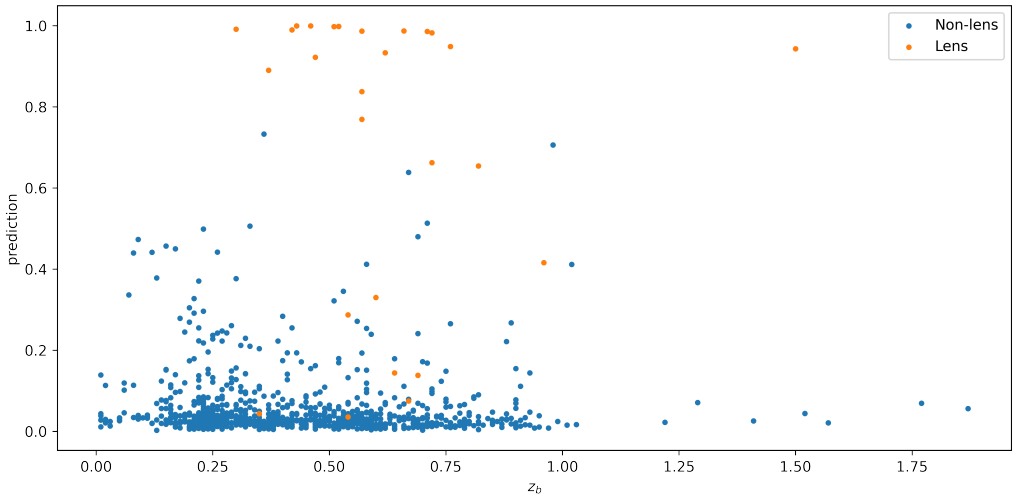

Figure 5: GAN+MixMatch plot of predictions versus the photometric redshift(z_b) for images in TestV2

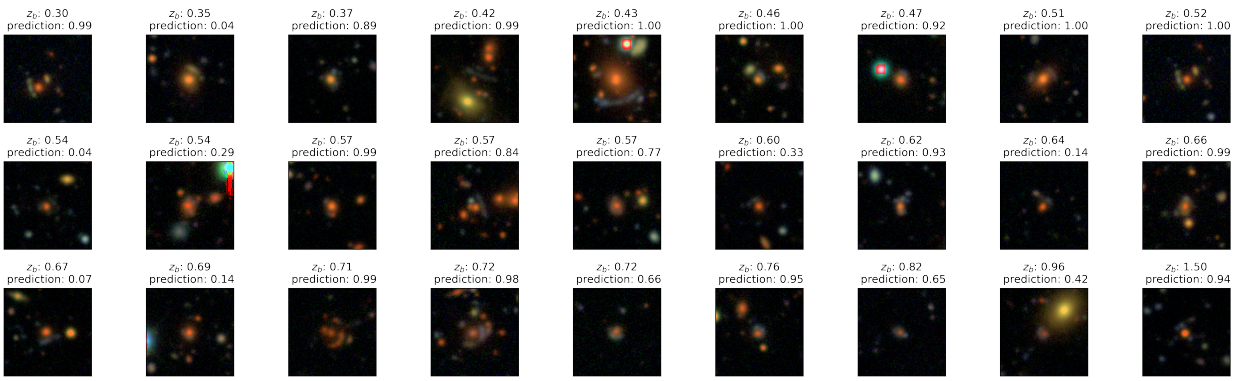

Figure 6: Lenses from TestV2 with prediction from GAN+MixMatch

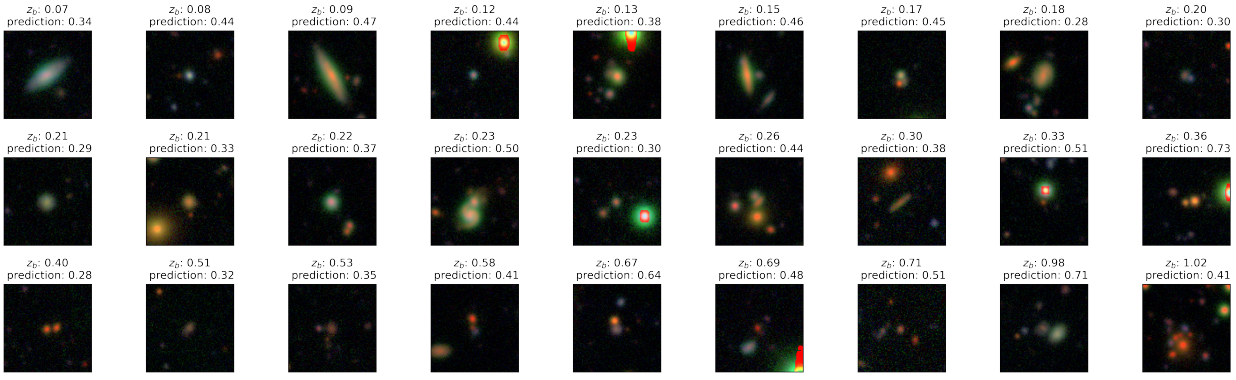

Figure 7: Non-lenses from TestV2 with the highest predictions from GAN+MixMatch

C FixMatch Performance On TestV1 And TestV2

Table 13: Average Precision (%) For TestV1

Model	Recall 50%	Recall 60%	Recall 70%	Recall 80%	Recall 90%	Recall 100%
FixMatch	22.48 ± 2.17	19.83 ± 4.04	17.42 ± 5.26	13.89 ± 5.32	9.29 ± 5.58	6.98 ± 2.63
GAN + FixMatch	65.88 ± 14.34	46.21 ± 17.24	32.57 ± 9.00	20.95 ± 5.68	15.94 ± 6.27	6.47 ± 0.50
GAN + FixMatch + SA	96.49 ± 4.88	90.48 ± 8.61	78.40 ± 10.43	64.39 ± 9.06	45.68 ± 14.84	26.48 ± 9.91

Table 14: Average Precision (%) For TestV2

Model	Recall 50%	Recall 60%	Recall 70%	Recall 80%	Recall 90%	Recall 100%
FixMatch	13.49 ± 3.28	11.40 ± 2.48	10.62 ± 3.76	9.78 ± 3.74	5.89 ± 2.43	4.44 ± 1.11
GAN + FixMatch	46.31 ± 20.24	31.10 ± 13.88	20.23 ± 6.83	10.81 ± 2.43	6.70 ± 2.69	3.47 ± 0.28
GAN + FixMatch + SA	91.50 ± 6.66	86.37 ± 12.11	81.46 ± 13.67	58.92 ± 16.61	33.73 ± 15.91	24.81 ± 13.82

FixMatch(Kurakin et al., 2020) is another SSL algorithm that we considered post-hoc. Although it wasn't available at the time of the original study, we are including additional results here to be thorough. The FixMatch algorithm stipulates the use of only weak augmentations in the supervised loss which means it can't deal with generalizing from "bad" simulated data very well. This can be seen from the results in Tables 13-14. By allowing strong augmentations(SA) such as those in Table 2, we are able to salvage this method and make it perform comparable to the other SSL algorithms we have looked at previously.