Characterising Dark Matter Substructure in Gravitational Lens Galaxies with Deep Learning

Owen Scutt[*]

Project advisor: Simon Dye[†]

Abstract.

We investigate the novel application of two sequential convolutional neural networks (CNNs) for the characterisation of dark matter substructure in lensing galaxies from galaxy-galaxy strong gravitational lensing images. In our configuration, an initial CNN predicts the number of substructures from a gravitationally lensed image and then this number, along with the same image, is input to a second CNN which predicts the power-law slope of the substructure mass distribution function. We have trained and tested the CNNs on simulated images created by lensing a galaxy-like light distribution with a foreground galaxy mass. We find that training and testing the CNNs on images created with a fixed lens geometry allows the number of substructures and the mass function power-law slope to be retrieved well. We then explore the effect of reducing the resolution of images such that the image pixel scale is halved finding that the accuracy of the number of predicted substructures decreases by only 7% while the accuracy of the predicted mass function slope decreases by 25%. When we allow variation in lens geometry between images in the test set, to mimic more physically motivated lens samples, we observe a decrease in accuracy of the number of predicted substructures and the mass function slope of 57% and 81% respectively. We attribute this significant degradation in predicting the mass function power-law slope to the degradation in the performance of the number-predicting CNN by comparing with predictions of the slope that are made when the CNN is given the true number of substructures. We discuss future possible improvements and the impact of the computing hardware available for this work.

1. Introduction. Strong gravitational lensing is an astronomical phenomenon by which several highly distorted images of the same distant source are observed. Lensing occurs as a result of the curvature of space-time caused by the gravitating mass of a foreground lensing object that lies between the source and observer. Multiple distorted images of the source are created as light is deflected about the lensing object along multiple trajectories that converge at a focal point at the observer.

In addition to gaining information about the lensed source, analysis of the images created by gravitational lensing can provide insight into the composition and distribution of the mass of the lensing object [19]. Strong lensing has contributed to progress in the fields of galaxy formation and galaxy evolution through extensive mass distribution analysis [29, 32] and detailed imaging and spectroscopy [26]. Due to its magnification effect, strong lensing has also been used in the discovery of galaxies at much higher redshifts than can be observed with current telescopes [1].

Since dark matter is thought to account for 85% of matter in the universe and constitutes the primary reason why galaxies remain gravitationally bound rather than being thrown apart [31], construction of a theory of dark matter has been of paramount scientific interest for almost fifty years. However, its apparent lack of interaction with the electromagnetic force [11] makes direct observations of dark matter challenging, if not impossible, and conjectures about its properties differ and are thus far inconclusive. One particular challenge is that computer

---

[*]Department of Physics and Astronomy, University of Nottingham (owen.scutt@outlook.com).

[†]Department of Physics and Astronomy, University of Nottingham (simon.dye@nottingham.ac.uk)

simulations of the large-scale structure of the Universe suggest that a Milky Way-sized galaxy should have of the order of 500 satellite galaxies (or 'subhaloes'), whereas only 11 have been actually observed [21]. This so-called 'Missing Satellite Problem' has led to the understanding that while the current cold dark matter model (so-called Λ-CDM) very accurately describes the distribution of matter over large scales in the Universe (i.e., galaxy clusters and larger), it does not offer the same accuracy at small scales.

Although dark matter can not be detected from its electromagnetic radiation, indirect observation is possible through the analysis of gravitational lensing effects. The magnification power of a strong gravitational lens, and a study of its light deflection, can provide crucial information on the distribution of visible and dark matter, offering a unique opportunity to probe the small-scale nature of dark matter.

A limiting factor at present is that strong lensing occurrences are rare due to the precise alignment required to produce lensed arcs or rings, and whilst discoveries have been made with several large telescope surveys, such as the Sloan Lens ACS (SLACS) survey [4] and the Dark Energy Survey (DES) [27], the number of lensing systems is still limited to just a few hundred low-redshift galaxies, too small a sample to make strong conclusions about the missing satellite problem. The small sample size has also meant that there hasn't yet been a strong requirement for highly efficient analysis and modelling of data.

Recently, excitement has begun to grow in the field of gravitational lensing as the sample of known strong lenses is expected to grow by several orders of magnitude [33, 6] in the near future due to the introduction of the next generation of telescopes. The two main facilities responsible for this surge in data volume are the ground-based Vera Rubin Observatory and the European Space Agency's Euclid Satellite, both becoming operational imminently [12, 30]. These large anticipated datasets will challenge the current method of analysis with a new problem of scalability. Previously, quantifying the light distortions produced by galaxy lenses has used maximum likelihood modelling of the observations (for recent examples of maximum likelihood techniques, see Maresca et al. [20], Dye et al. [9], Berta et al. [2], and references contained therein). Many of these studies use a single deflector for their mass model which typically involves the optimisation of anywhere from five to nine non-linear parameters. Some studies use multiple deflectors which further increases the complexity of parameter space, to the point that finding the global best-fit becomes highly unreliable and requires manual intervention to release optimisers from becoming stuck in local minima. This is generally a time-expensive procedure, requiring expert knowledge and typically several weeks of manual effort. With anticipated lens samples comprising tens of thousands of lenses, clearly such an approach is unviable.

Conveniently, this large increase in data is paralleled by increasingly sophisticated computational methods, enabling a new age of automated data processing. The application of convolutional neural networks (CNNs) has already proven more than sufficient to meet the demand for rapid analysis of the petabyte-scale datasets produced by the new telescopes [18]. Specifically in the field of gravitational lensing, the effectiveness of CNNs has been demonstrated by their use in successfully finding previously missed strong gravitational lens images from existing surveys [13, 17]. Building on this success, attention has shifted to preparing CNN-based image analysis methods for use on the forthcoming datasets from next-generation surveys. In preparation for the analysis of the new data, multiple thousand-strong datasets

of simulated lens images are created using physical parameters (chosen to match those from the anticipated surveys) and used to train CNNs.

Hezaveh et al. [10] were the first to use CNNs to estimate the mass model parameters of a galaxy lens. They used four different CNNs to predict the mass profile, showing that lens modelling can be increased in speed by six to seven orders of magnitude and removing the requirement for expert knowledge. This was later extended by Rivero et al. [7] and Pearson et al. [28] who showed that the CNN methodology continues to work when applied to images without first subtracting the light from the lensing galaxy. In addition, Morningstar et al. [22] showed how neural networks allow reconstruction of an unlensed image of the source. Only recently have CNNs been used to infer the properties of subhaloes in lensing galaxies [7].

This study aims to investigate the capability of CNN-based machine learning architectures to infer the properties of the distribution of subhalo masses in strong gravitational lenses. Whereas previous attempts have used the indirect process of simulation-based inference (see, [5], for example), here, an alternative, simpler strategy is investigated, using two regression-based CNNs applied in series to directly detect the number of subhaloes present in any given lens galaxy and measure their distribution in mass.

This report is arranged as follows. Section 2 details the CNN architecture and the methodology used for simulating the gravitational lens images. Section 3 presents the results obtained from training and testing the CNN with different image resolutions and lens parameter distributions. In Section 4 we consider the implications of these results along with the strengths and weaknesses of the investigation and some suggestions for related investigations in the future.

2. Methodology. To train the CNNs, a simulator was produced to create realistic synthetic strongly lensed images labelled by the parameters of the mass distribution of each lens. For background information regarding the theory of gravitational lensing, we refer the reader to [25]. Both the CNN architectures and simulator are described in this section.

2.1. Image Generation. The simulator assumes a Singular Isothermal Ellipsoid (SIE) lens mass distribution for the main lens with Singular Isothermal Sphere (SIS) substructure masses distributed within it. The deflection angles of the SIE and SIS profiles, required for creation of the lensed images by mapping from the source to the image, are given in [15]. We used an identical light source for all images created, defined by a circularly-symmetric 2D Gaussian distribution of standard deviation $\sigma = 0.9$ arcsec centred at the origin of the source plane (i.e. directly on the line of sight).

The parameters of the SIE model are the axis ratio, q, the orientation of the semi-major axis, ϕ, the Einstein radius, θ_E and the lens centroid co-ordinates (x, y). The Einstein radius is effectively a measure of the mass of the main lens which relates to its velocity dispersion (see equation 2.2 below). For the training images, we randomised the lens mass parameters. ϕ and (x, y) were drawn from the following uniform distributions: $0 \leq \phi < \pi$, -0.1 arcsec $\leq x \leq 0.1$ arcsec and -0.1 arcsec $\leq y \leq 0.1$ arcsec. θ_E and q were drawn from realistic normal distributions of mean 1.5 arcsec, width 0.4 arcsec and mean 0.78, width 0.12 respectively, as presented by Pearson et al. [28]. These distributions are shown in Figure 1. In addition, we randomised the redshift of the source according to the distribution of Collett et al.[6] (see Figure 1). Changing source redshift also changes the Einstein radius. We refer the reader to
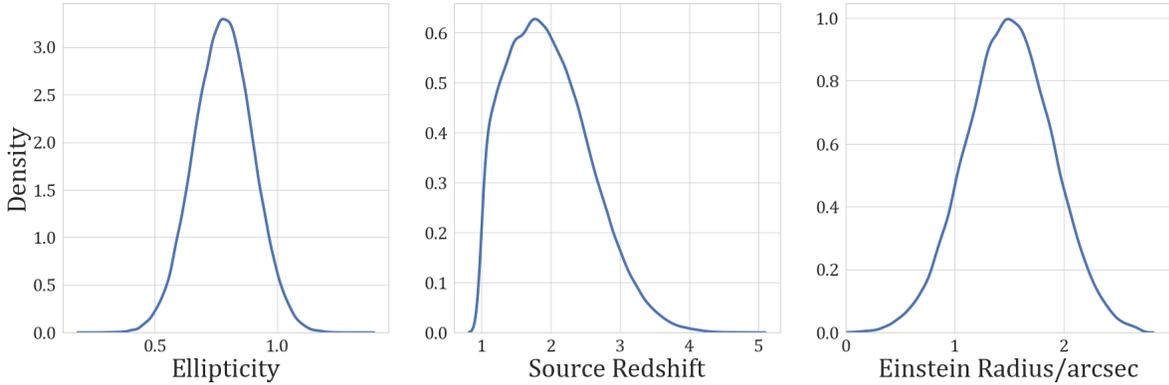
Figure 1. *Number density distributions of lens geometry parameters for a typical training set of 100,000 images. Lens centre position and lens orientation are sampled from the uniform distributions discussed.*

[25] for more details.

The majority of our investigation assumes that lensed images have a pixel scale of 0.1 arcsec to match the resolution of the near infra-red camera on the Euclid Space Observatory. In Section 3.2, we also investigate the impact of changing this resolution to 0.2 arcsec.

Once the source plane redshift and the geometry of the simulated lens has been determined, we generate dark matter substructures to populate the lens plane. The mass of each of these, M, is randomised according to the substructure mass distribution function given by

$$(2.1) \qquad \frac{\mathrm{d}N}{\mathrm{d}M} \propto M^{-\alpha},$$

where N is the number of substructures and α is the power-law slope of the function [34]. We randomise α when generating CNN training images by drawing from a uniform distribution over $-2.5 \leq \alpha \leq -1.5$. For a given simulated lens (and thus a specific value of α), we continue drawing substructures until their total mass accounts for 10% of the total lens mass [34]. This requires calculation of the main lens mass which we obtain in the following way. First, we calculate the lens velocity dispersion, σ_{lens}, using

$$(2.2) \qquad \sigma_{lens} = \frac{c\sqrt{\theta_E}}{4\pi} \frac{D_s}{D_{ds}},$$

where c is the speed of light and D_s and D_{ds} are the angular diameter distances to the source and between the lens and source respectively. The mass of the main lens (in solar units), M_l, is then calculated using

$$(2.3) \qquad M_l = (\frac{A_{1D}}{\sigma_{lens}})^\beta 10^{15},$$

where A_{1D} is the one-dimensional velocity scaling parameter dependent on halo concentration, taken here to be $1200 \mathrm{km\,s}^{-1}$ and we take $\beta = 1/3$ [23]. Some example distributions of subhalo masses for varying values of α are shown in Figure 2.

Substructures are spatially distributed with a uniform surface density across the lens within a radius of $2\theta_E$ from the lens centre [5]. With the lens mass characteristics and source
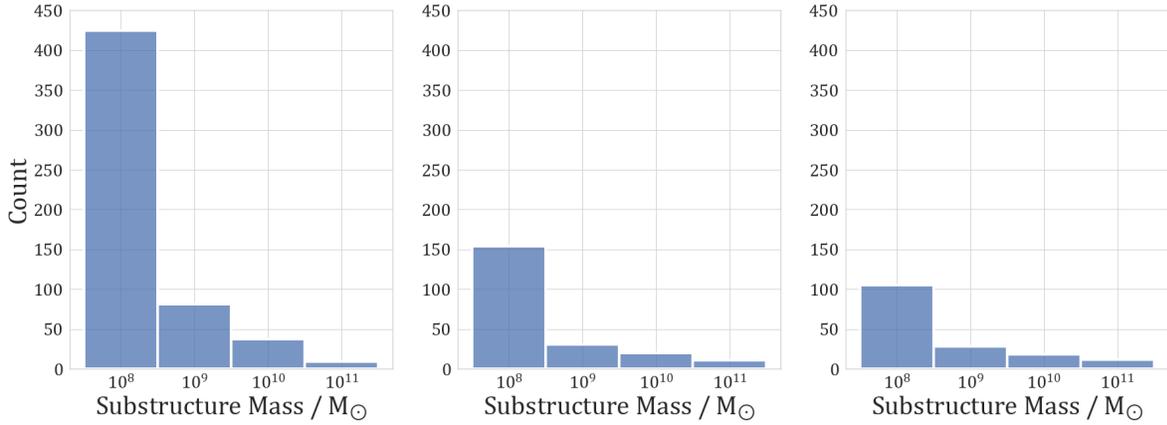
Figure 2. *Populations of substructure masses for lenses with different power law α. From left to right: $\alpha = -2.4$, $\alpha = -2.0$, $\alpha = -1.6$.*
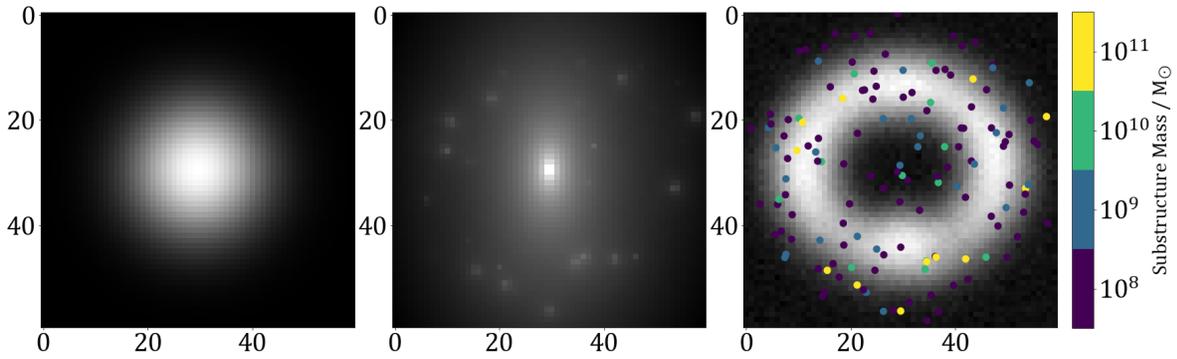

Figure 3. *Example simulated lens image. From left to right: Gaussian source, simulated lens surface mass density and resulting lensed image with substructures plotted. Image resolution is 0.1 arcsec. Lens parameter values are $\theta_E = 1.5\,arcsec$, $q = 0.78$, $(x, y) = (0, 0)$, $\phi = 0$, $z = 1.77$.*

intensity profile fully defined, we created the lensed image using the publicly-available python package Lenstronomy [3].

To increase the realism of the image, we blurred the images with a point spread function (PSF) and added noise. The PSF was set as a 2D Gaussian with standard deviation, σ_{PSF}, varied according to the image resolution: $\sigma_{PSF} = 0.14\,arcsec$ for the $0.2\,arcsec$ resolution images and $\sigma_{PSF} = 0.07\,arcsec$ for the $0.1\,arcsec$ images. We also added normally-distributed noise, obeying a fixed signal-to-noise ratio of 300. An example of a simulated lensed image is shown in Figure 3.

2.2. CNN Architecture. For training our CNNs, we used the mean squared error (MSE) as our loss function, given by

$$(2.4) \qquad MSE = \frac{1}{b}\sum_{i=1}^{b}(y_i - y_{predicted_i})^2,$$

where y and $y_{predicted}$ are the true and predicted subhalo properties (either α or N; see below) respectively, and the sum acts over a batch of b training images.

During preliminary testing, we logged the progression of the MSE as a function of training epoch and altered the CNN structure and parameters accordingly. For example, no decrease in validation loss indicates underfitting, remedied by increasing CNN complexity with added layers. Conversely, sudden increases in validation loss indicate overfitting - fixed by reducing CNN complexity, or decreasing the number of training epochs. This process was repeated for multiple permutations of both single and sequential CNNs with different optimisers (Adam [16], Nadam [8]), learning rates (10^{-4}, 10^{-3}, 10^{-2}), batch sizes (8, 16, 32), number of convolutional layers (1-6), kernel sizes ((3×3), (5×5), (7×7), (9×9), (11×11)), convolution filters (8, 16, 32, 64, 128), and epochs (10-100). Preliminary testing used image sets with fixed lens parameters of resolution 0.1 arcsec, with both true N and true α unknown to the CNNs prior to prediction (as presented in section 3.1.1).

Preliminary testing readily showed that the best performance for fixed lens parameter images was obtained with two sequential CNNs. The first, which we will refer to as the 'N-CNN' hereafter, predicts N from an input image. The second, which we will refer to as the 'α-CNN' hereafter, predicts α from an input image and its estimate of N predicted by the N-CNN. The two CNNs have different architectures and so needed to be trained separately. We trained the N-CNN by labelling images with the true N contained in the lens and the α-CNN by labelling images with both the true α and true N.

In terms of their architecture, for the N-CNN, we used three convolutional layers with a kernel size of (5×5) and 16, 32 and 64 filters respectively from input to output. For the α-CNN, we found that only one convolutional layer was required with 8 filters and a kernel size of (3×3). Following each convolutional layer in both networks, we applied batch normalisation, max-pooling, and dropout. After all convolutions, both networks contained a flatten layer and two fully connected dense layers. Before the final dense layer in the α-CNN, N values were introduced via a concatenate layer. Details of layer order and dimensions are shown in Figure 4.

Each layer except for the flatten layer in both networks used an activation function for forward propagation, and we found the Rectified Linear Unit (ReLU) [24] was most effective in all cases. However, the final fully connected dense layer requires a linear activation function for the CNN to perform regression analysis with output dimensionality of 1. Across all convolutional layers, we chose to use unity stride width and 'same' padding to pad image arrays with zeros until the dimensionality of the input was met. Following each convolution, the extracted feature map was then fed into a max-pooling layer with kernel of dimensions 2×2 and step width of 2, halving its resolution and drastically reducing sample size and computational load while maintaining prediction success. Image pixel values were normalised to a minimum of 0 and a maximum of 1 during image pre-processing to speed training and increase loss stability (although after convolution and pooling, outputs from the first layer extend beyond this range). Batch normalisation following each convolution ensured that feature map pixel values remained normalised throughout training to maintain loss profile stability.

Dropout in a network acts to randomly disable a set fraction of layer outputs before information is passed to the next layer. This promotes exploratory learning by making the network try to pass feature maps along pathways with lower weights that may have been
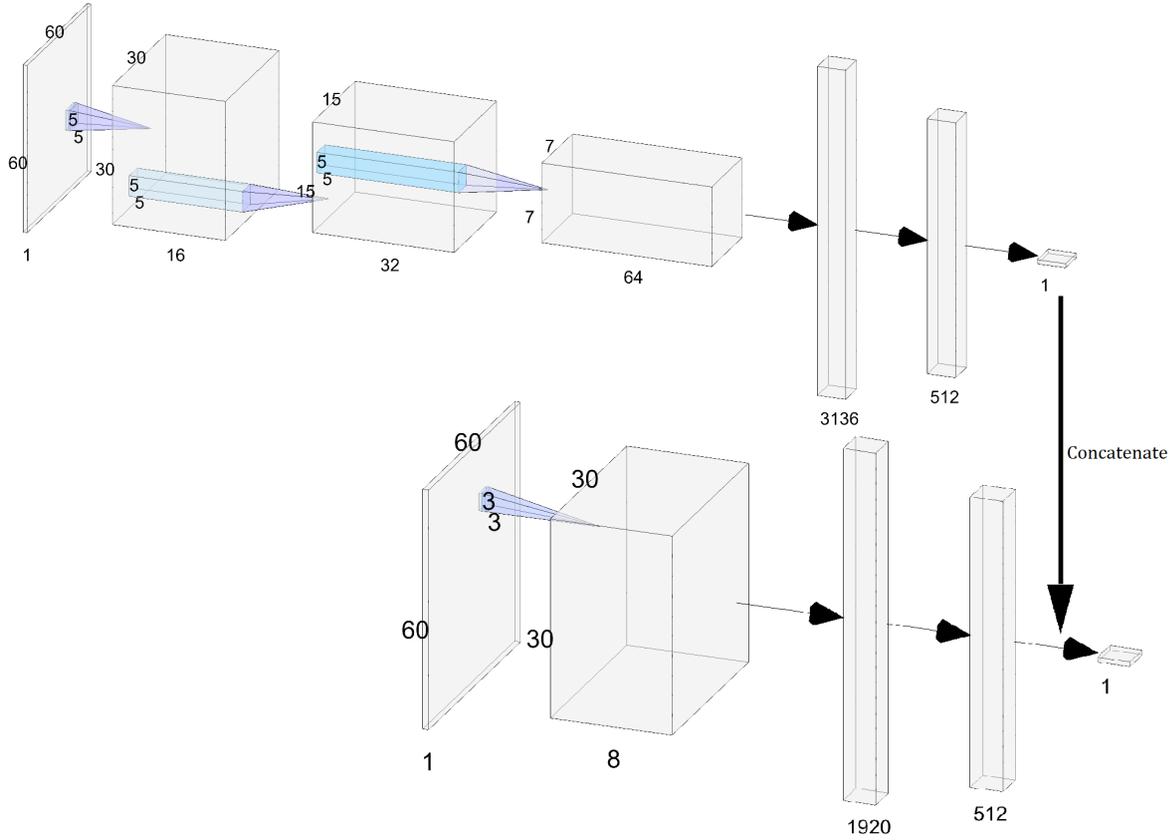
Figure 4. *Architecture of the CNNs used, with N-CNN (top, 1,671,393 trainable parameters) prediction results fed into α-CNN (bottom, 3,687,522 trainable parameters) via concatenate layer during testing.*

previously neglected due to being randomly assigned a low weight during an earlier epoch. We determined that a uniform dropout rate of 0.25 worked well to allow the network to escape local minima in the loss function and reduce over-fitting.

The final dense layers were preceded by a flatten layer to reduce the two-dimensional output of the final convolutional layer into the one-dimensional array format accepted by fully connected layers, before a prediction is output from the final dense layer.

The Adam optimiser with an initial learning rate of 10^{-4} was not only found to be computationally efficient, but also utilised local gradient based learning rate tuning to alter the learning rate during training to reduce cases of local minima trapping in the loss function. Running training on a machine with a Graphics Processing Unit (GPU) reduced training time massively and allowed us more choice in the batch size of input images and the number of training epochs. For all training runs, we found a batch size of 16 to be most time-efficient while retaining learning stability. The optimal number of training epochs was determined by first training a network for 100 epochs, reducing if over-fitting had occurred (signified by a steep increase in validation loss). Using the hyperparameters above, our networks could carry out 100 training epochs with a training set of $100,000$ 60×60 images in under two hours.

We tested CNN performance using test sets of newly generated images, unseen in training,

to avoid false prediction success due to over-fitting and from memorising correct predictions of already seen images. We used a train-to-test set ratio of $5:1$ to allow faster learning without over-fitting.

3. Results. Here we present the prediction performance of the N-CNN and the α-CNN. Both were trained on training sets of $100,000$ images before being tested on test sets of $20,000$ images, in two different scenarios. In the first scenario, we trained and tested the CNNs on image sets generated with a fixed set of lens model parameters and an identical source. In these training and test sets, only the substructure was varied by randomising subhalo positions and masses and the value of α from which subhaloes were drawn. We also tested both 0.1 arcsec and 0.2 arcsec image resolutions. In the second scenario, we trained and tested the CNNs with images generated with the additional variation that the lens model parameters, θ_E, q, ϕ and (x,y) and the source redshift were randomly drawn from the distributions discussed in section 2.1.

In all cases, we ran two separate diagnostics. The CNNs were first tested with sets of $20,000$ images with continuously varied power law. Performance was tested by measuring the slope of the scatter plot of true versus predicted α and N and the root mean square (RMS) of the differences between true and predicted α and N. For the second diagnostic, we tested the CNNs on multiple test sets of $5,000$ images, holding α fixed in each set. For each set, we have plotted the distribution of differences between true and predicted subhalo properties and we have quoted the 68% confidence interval as equal to one standard deviation.

3.1. Fixed Lens Parameters. In this section, we show the performance of both the N-CNN and the α-CNN for images generated with a fixed lens geometry and an identical Gaussian source. The lens model parameters were fixed at the mean values of the physical distributions chosen, namely, $\theta_E = 1.5$ arcsec, $q = 0.78$, $(x,y) = (0,0)$, $\phi = 0$, $z = 1.77$. The only variation in images tested in this section arises from the variation in subhalo masses drawn from the distribution function of slope α, which is itself drawn uniformly between $-2.5 < \alpha < -1.5$, and the randomised positions of substructures within twice the Einstein radius of the lens.

3.1.1. Image resolution: 0.1 arcsec. The top two panels of Figure 5 show the performance of the N-CNN for the case of fixed lens model parameters for 60×60 pixel images of resolution 0.1 arcsec. Fitting a straight line to the scatter plot on the left reveals a gradient of 0.993, indicating that bias in prediction of N is very low over the whole $20,000$ image set. The RMS of the differences between true and predicted values of N (i.e. the scatter about the line $y = x$, labelled 'ideal gradient' in the plot) is 49.0.

The panel in the top right of Figure 5 shows the distribution of differences between true and predicted values of N when tested on five training sets each containing $5,000$ images and each with a fixed value of α as indicated. The brown line in this plot shows the distribution calculated from the $20,000$ images plotted in the top-left panel with randomised values of α for comparison. There are two notable trends present here. First, the error in predicted values of N, as measured by the standard deviation of the distributions, becomes larger with decreasing (i.e. more negative) α. Quantifying this, the range of standard deviations between the lowest and highest values of α tested is $37.9 \leq \sigma \leq 50.4$. Secondly, the number of predicted

substructures monotonically moves from being over-predicted at $\alpha = -1.6$ to under-predicted at $\alpha = -2.4$. We can approximately tie a given value of α to a value of N by noting that, because of the power-law substructure mass function and the fact that we draw subhaloes from it until a fixed fraction of 10% of the total lens mass is reached, a value of $\alpha = -1.7$ results in an average of $N \simeq 200$ subhaloes being drawn whereas at $\alpha = -2.5$, the average is $N \simeq 700$.

The bottom-left and bottom-right panels of Figure 5 are the equivalent of the top two panels but instead show the performance of the α-CNN. For each image tested, its number of subhaloes as predicted by the N-CNN is input to the α-CNN. The best-fitting straight line to the scatter plot in the lower-left panel has a gradient of 0.658, showing significant bias. The scatter of predicted α about the true value is quantified with a root mean square error ($RMSE$) of 0.294. The bottom-right panel shows the same trend of increasing standard deviation with decreasing α as the N-CNN, with a range in standard deviation of $0.212 \leq \sigma \leq 0.310$. Conversely, the trend in bias is reversed compared to that of the N-CNN, such that the predicted α is over-predicted at more negative values of α. Quantitative results for both N-CNN and α-CNN across varied and fixed α test sets are presented in Table 1.

Table 1

Prediction scatter gradient with σ and RMS of the differences between true and predicted values for N-CNN (left) and α-CNN (right) trained and tested on 60×60 images, fixed lens geometry parameters. Fixed α test sets contained $5,000$ images, the varied α test set contained $20,000$ images.

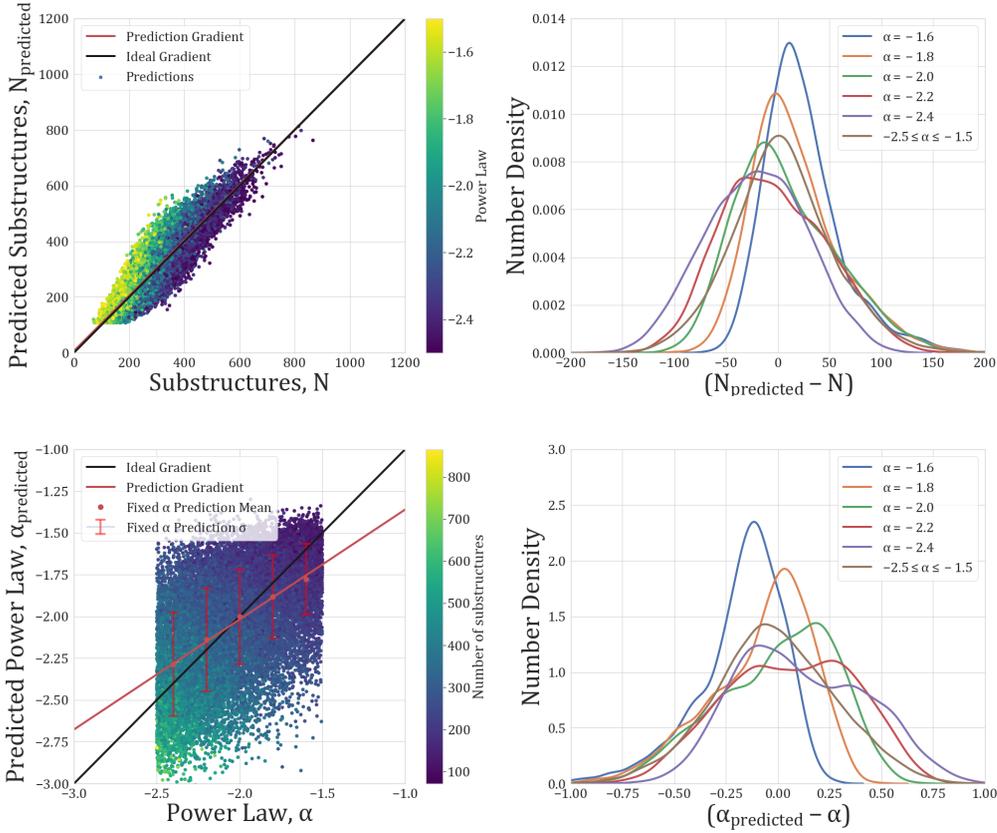|	N-CNN			α-CNN		
α	Gradient	σ	RMSE	Gradient	σ	RMSE
-1.6	-	37.9	44.9	-	0.212	0.276
-1.8	-	43.2	46.5	-	0.248	0.261
-2.0	-	49.1	49.6	-	0.284	0.284
-2.2	-	50.4	50.5	-	0.309	0.315
-2.4	-	48.3	52.6	-	0.310	0.331
Varied	0.993	48.8	49.0	0.658	0.441	0.294

Figure 5. *Prediction results for* 60×60 *images, fixed lens geometry parameters. Top: Prediction scatter (left) and prediction error distributions (right) for N prediction. Bottom: Prediction scatter (left) and prediction error distributions (right) for α prediction.*

3.1.2. Image resolution: 0.2 arcsec.
Here we investigate the effects of image resolution on CNN performance by repeating the analysis of section 3.1.1 but training and testing with 30×30 pixel images with a coarser resolution of 0.2 arcsec.

Figure 6 and Table 2 show our results. Considering the N-CNN first and the top-left panel of the figure, the decrease in image resolution gives rise to slightly more bias in the predicted number of subhaloes overall, as measured by a decrease in the scatter gradient of 7% to 0.923. We also find that the $RMSE$ increases by 26% to 61.9. The network performs similarly to the 0.1 arcsec resolution images for $N < 400$ ($\approx \alpha > -2.2$), however, at values of N above this threshold, the network begins to significantly over-predict N. This is readily apparent in the plot in the top-right of Figure 6 where the trend of increasing bias with decreasing α is much stronger. Despite this, the range of standard deviation of these distributions is $37.8 \leq \sigma \leq 50.6$ which is comparable with the case of 0.1 arcsec images

The lower two panels of Figure 6 show the performance of the α-CNN on the 0.2 arcsec resolution images when fed images with N predicted by the N-CNN. The scatter plot in the lower-left panel has a straight line fit with gradient 0.495, a 25% decrease from the 0.1 arcsec resolution case. At a value of 0.294, the global $RMSE$ of this scatter plot has barely changed in

testing with the lower resolution images, however, at the lower range of α values tested, there is a growing number of predictions of α that are very discrepant with their true values, by up to 70%. The tail of the distribution of differences in true and predicted values of α has become extended compared to the higher resolution case, yet, the spread of these distributions has become slightly smaller, with a range of standard deviation that now spans $0.181 \leq \sigma \leq 0.213$. This is a surprising result but the net effect is that, because the bias is larger, the significance of the bias is higher.

Table 2

Prediction scatter gradient with σ and RMS of the differences between true and predicted values for N-CNN (left) and α-CNN (right) trained and tested on 30×30 images, fixed lens geometry parameters. Fixed α test sets contained $5,000$ images, the varied α test set contained $20,000$ images.

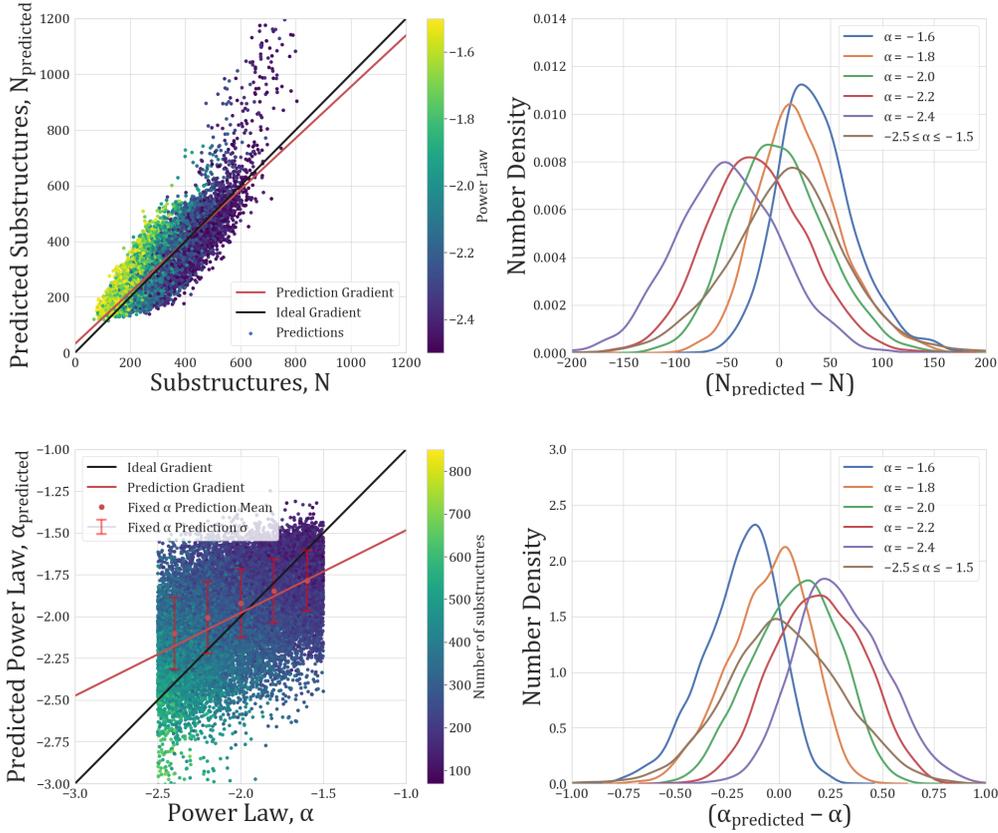| α	N-CNN			α-CNN		
	Gradient	σ	RMSE	Gradient	σ	RMSE
-1.6	-	37.8	52.1	-	0.181	0.259
-1.8	-	40.9	45.6	-	0.191	0.197
-2.0	-	44.9	44.9	-	0.206	0.221
-2.2	-	48.0	52.5	-	0.215	0.290
-2.4	-	50.6	70.7	-	0.213	0.367
Varied	0.923	61.0	61.9	0.495	0.407	0.290

Figure 6. *Prediction results for 30×30 images, fixed lens geometry parameters. Top: Prediction scatter (left) and prediction error distributions (right) for N prediction. Bottom: Prediction scatter (left) and prediction error distributions (right) for α prediction.*

3.2. Varied Lens Parameters. In this section, we investigate the performance of the N-CNN and α-CNN on 0.1 arcsec images measuring 60×60 pixels as in section 3.1.1, but now where the lens model parameters are randomised. Specifically, in addition to randomising subhalo positions and masses, the slope of the mass distribution function and the source redshift, we now draw the lens model parameters, θ_E, q, ϕ and (x, y) randomly according to the distribution functions presented in section 2.1 in both the training and test data sets.

Following the same format as in the previous sections, our results are presented in Figure 7 and Table 3. Considering the top row of the figure which shows the performance of the N-CNN, it is immediately clear that given more varied lens model parameters and therefore more varied lensed image characteristics, the N-CNN struggles to make a reliable prediction of N. The line of best fit now has a gradient of 0.383 as the network more frequently defaults to the mean number of subhaloes encountered during training. Similarly, the global $RMSE$ in this plot has now risen by 216% to 155. This reduction in performance is also seen in the panel in the top-right where the distributions of differences between the true and predicted values of N for various fixed values of α now have standard deviations that are at least double those of the case where lens model parameters were fixed. The previous trend observed that

the predicted N is biased towards higher values at higher α is still present but now much stronger; at $\alpha = -1.6$, the N-CNN over-predicts the true number by an average of 115, an over-estimate of approximately 70%.

Performance of the α-CNN is shown in the bottom two panels of Figure 7. The ability of the network to predict α has also degraded significantly. The gradient of the straight line of best fit has fallen by 81% to 0.123, indicating that the network is almost yielding random values of α. The global $RMSE$ has risen from the fixed lens model 0.1 arcsec case by 29% to 0.380. The bottom-right plot again shows much stronger bias in α, over-predicting towards lower values of α and vice versa. The range of standard deviations of these plots has also increased significantly, with a range of $0.215 < \sigma < 0.259$.

Table 3

Prediction scatter gradient with σ and RMS of the differences between true and predicted values for N-CNN (left) and α-CNN (right) trained and tested on 60×60 images, varied lens geometry parameters. Fixed α test sets contained $5,000$ images, the varied α test set contained $20,000$ images.

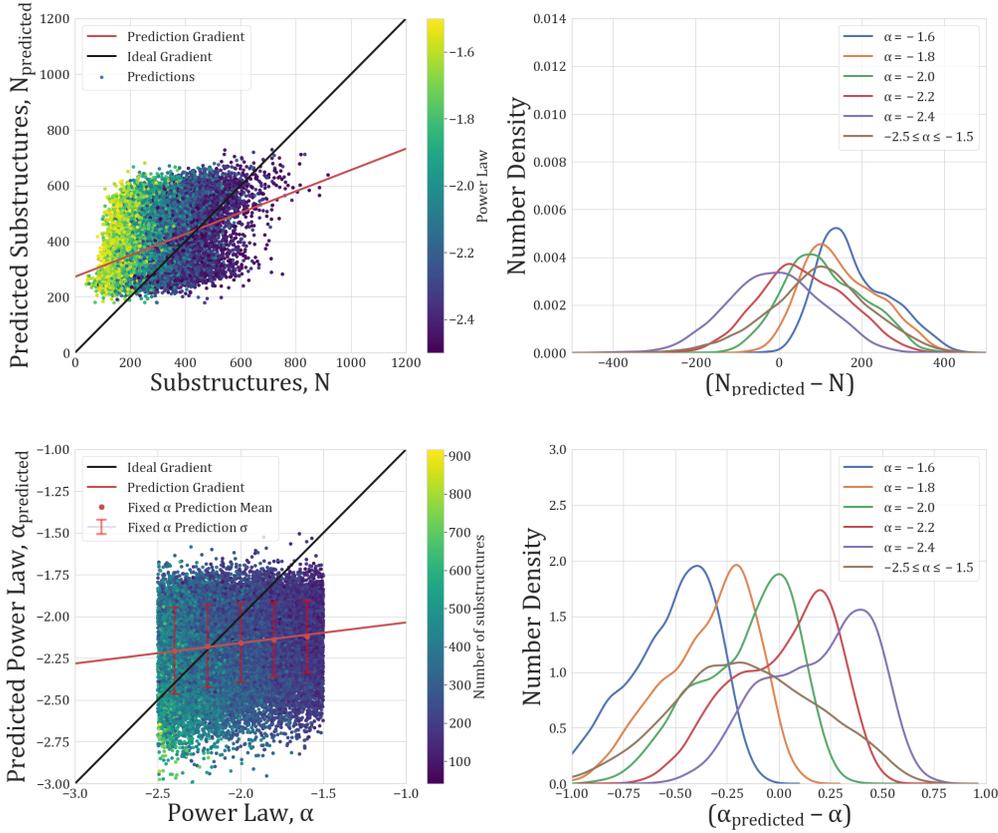|	N-CNN			α-CNN		
α	Gradient	σ	RMSE	Gradient	σ	RMSE
-1.6	-	86.6	202	-	0.215	0.563
-1.8	-	92.0	177	-	0.223	0.405
-2.0	-	97.0	147	-	0.236	0.283
-2.2	-	107	123	-	0.247	0.248
-2.4	-	114	114	-	0.259	0.324
Varied	0.383	121	155	0.123	0.375	0.380

Figure 7. *Prediction results for* 60×60 *images, varied lens geometry parameters. Top: Prediction scatter (left) and prediction error distributions (right) for N prediction. Bottom: Prediction scatter (left) and prediction error distributions (right) for* α *prediction.*

3.2.1. Predicting α with the true N.

To test the degree to which the performance of the α-CNN is degraded by the much reduced ability of the N-CNN to reliably predict N, we ran the exact same analysis as in section 3.2, but instead of providing the α-CNN the predicted N during testing, we gave it the true N.

The results of this test are shown in Figure 8 and Table 4. When using the true N instead of the predicted N, the gradient of the scatter in the panel on the left has increased to 0.669. The global $RMSE$ has also dramatically decreased to 0.223 which is an improvement even on the results obtained in the case of 0.1 arcsec resolution images with fixed lens model parameters. Considering the plot on the right of Figure 8, we find that again, performance is greatly improved, even beyond the 0.1 arcsec resolution fixed lens model case, with smaller bias and a smaller range of standard deviations of $0.141 \leq \sigma \leq 0.239$.

Table 4

Prediction scatter gradient with σ and RMS of the differences between true and predicted values for the α-CNN with true N supplied during testing, 60×60 images, varied lens geometry parameters. Fixed α test sets contained $5{,}000$ images, the varied α test set contained $20{,}000$ images.

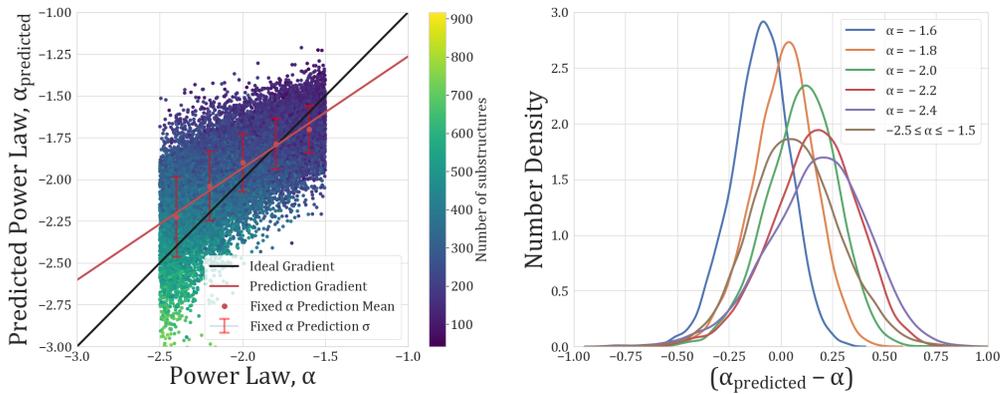| α	Gradient	σ	RMSE
-1.6	-	0.141	0.172
-1.8	-	0.151	0.152
-2.0	-	0.173	0.199
-2.2	-	0.206	0.261
-2.4	-	0.239	0.296
Varied	0.669	0.396	0.223

Figure 8. *Prediction scatter (left) and prediction error distributions (right) for the α-CNN with true N supplied during testing, 60×60 images, varied lens geometry parameters.*

4. Discussion and Conclusions. We have assessed the effectiveness of using deep learning to determine the properties of substructure in gravitational lensing galaxies using the background source images they produce. We trained and tested two different CNN architectures; the first predicts the number of subhaloes, N, in the lens, and the other predicts the slope, α, of the mass distribution power-law function from which they were randomly drawn.

We quantified the performance of both CNNs in two different ways. First, we tested each CNN with a set of images where each image was produced using a value of α drawn from a uniform distribution. We measured the gradient and $RMSE$ of the plot of true values of N and α against their CNN-predicted values; we refer to this as the 'prediction scatter gradient' hereafter. Second, we tested the CNNs' performance by holding α fixed at one of five different values and determining the width of the distribution of differences between true and predicted values of N and α. The former test gives a measure of the global performance of the CNNs, while the second offers measurement of precision as a function of α.

In general, we conclude that prediction scatter gradient offers the most useful single statistic for prediction success. For example, the distribution of the error in predicted α when the CNN is fed $N_{predicted}$ (see section 3.2) is comparable to that obtained when the CNN is fed

the true N, but when comparing the prediction scatter gradients it is clear that the latter outperforms the former. This is also reflected in the probability distributions in Figures 7 and 8, considering the means of the distributions rather than simply their widths.

The results for fixed lens model parameters presented in section 3.1.1 provide a solid benchmark for comparison with other results. The N-CNN results show the best example of ideal network performance with a prediction gradient of 0.993. In comparison the α-CNN prediction scatter gradient of 0.658 fares less well, although it still shows promise; despite not being as close to the ideal unity gradient as the N-CNN, it is clear that significant learning has occurred during training and the CNN is still able to produce reasonable estimations of α. The distributions of errors in this case show a bias such that α is under-predicted and N is over-predicted at low values of α, but the widths of the distributions indicate that these biases are not very significant.

Section 3.1.2 considered CNN prediction success with images of resolution 0.2 arcsec. Despite a 75% reduction in the total number of pixels in training and testing images, the N-CNN only suffers a 7% decrease in predictor accuracy and a 25% increase in predictor uncertainty. However, the N-CNN begins to over-predict significantly for $N > 400$, in some cases by more than 400 substructures. This is propagated through to the α-CNN, where predictions drop to $\alpha_{predicted} = -4.0$, below the minimum value of $\alpha = -2.5$ seen during training. This results in a scatter gradient of 0.495 for predicted α, a decrease of 25% in comparison with the 0.1 arcsec images.

The most likely explanation for this observed degradation in performance is due to substructure over-crowding, i.e. the number of subhaloes that lie within an image pixel. The CNN is only sensitive to pixels containing significant lensed image flux, and therefore performance is better for larger Einstein radii. With a resolution of 0.1 arcsec and an Einstein radius of 1.5 arcsec, ~ 2800 pixels contain significant lens image flux.

Our images rarely have $N > 900$, but even at this maximum limit, due to their random positioning only one in ten pixels will contain multiple subhaloes. At an image resolution of 0.2 arcsec, however, only ~ 700 pixels contain significant image flux, which falls below the maximum N of our images. Given that CNNs can only operate on a pixel-by-pixel basis, this overcrowding of subhaloes causes difficulties at high N. This is further exacerbated for smaller Einstein radii. It would be desirable to repeat the tests presented here with greater resolution than 0.1 arcsec using more powerful hardware (see below).

Section 3.2 reports CNN prediction results when the lens model parameters were varied. For the N-CNN the scatter gradient drops by 81% compared to the tests without lens model variation. Similarly, the standard deviation of the error distributions increases by 150% on average, and predictions cluster heavily within $200 < N_{predicted} < 700$. The resulting α-CNN scatter gradient is 0.123. We found that the performance of the α-CNN was actually better than this without being provided any estimate of N. Our conclusion is therefore that the poor performance of the N-CNN had catastrophic effects on the performance of the α-CNN. Nevertheless, the scatter gradient was not zero, which indicates that the CNN has still learned to a degree.

This degradation in performance of the α-CNN due to poor estimates of N is further evidenced by training and testing the CNN with the true N. In this scenario, the α-CNN achieved the highest accuracy of any of our α-CNN tests; with the same test set as that

discussed immediately above, accuracy increased by 444%.

The variation in our results highlights the potential of using CNN inputs beyond image data alone. We have demonstrated that α prediction can improve substantially if the input $N_{predicted}$ is accurate. This could be improved further with predictions of other lens model parameters. For example, combining predicted θ_E and N values in a simple multi-layer perceptron could weight the 'trustworthiness' of $N_{predicted}$ before it is input to the α-CNN. In principle the inclusion of other lens model parameters could improve performance even further; indeed, the prediction of lens model parameters has already been successfully demonstrated by Pearson et al. [28].

In this study we have considered simplistic lensed images. A logical next step would be to enhance image realism – for example, by introducing variation in the simulated source as well as the lens. For instance, the source could be replaced with galaxy images taken from existing surveys and then lensed when making simulated images. Furthermore, our simulations assume that lens plane redshift is fixed, and lens light has been removed entirely. This neglects the detrimental effects that remaining traces of lens light might have on image clarity.

Also worthy of further investigation is the potential for including images of the same system taken in different wavebands. Such additional information would be likely to increase CNN prediction success, but would also increase computational load and training time. To compensate, more computing power would be required. To put this into context, the computing hardware used in this study took upwards of 4 hours to generate 100,000 60×60 pixel images for training. This was largely due to hard disk access; solid state drive (SSD) support would reduce this time significantly. Storage availability also limited the maximum training set size, as well as the potential to use CNN optimisation packages like Auto-Keras [14]. Limits on training set data footprint were further compounded by available random access memory (RAM). Increased resolution or dataset size requires a decrease in batch size to avoid exceeding RAM limits and crashing during training, which increases training time. RAM limitations also restricted maximum network complexity; with each CNN layer comes more weights, biases and hyperparameters, in turn requiring more RAM during training. While not affecting the credibility of the results presented here, these limitations should be recognised as areas for improvement in any future investigations.

REFERENCES

[1] H. Atek, J. Richard, J.-P. Kneib, M. Jauzac, D. Schaerer, B. Clement, M. Limousin, E. Jullo, P. Natarajan, E. Egami, and et al., *New constraints on the faint end of the uv luminosity function at z~ 7-8 using the gravitational lensing of the hubble frontier fields cluster a2744*, The Astrophysical Journal, 800 (2015), p. 18, https://doi.org/10.1088/0004-637x/800/1/18, http://dx.doi.org/10.1088/0004-637X/800/1/18.

[2] Berta, S., Young, A. J., Cox, P., Neri, R., Jones, B. M., Baker, A. J., Omont, A., Dunne, L., Carnero Rosell, A., Marchetti, L., Negrello, M., Yang, C., Riechers, D. A., Dannerbauer, H., Perez-Fournon, I., van der Werf, P., Bakx, T., Ivison, R. J., Beelen, A., Buat, V., Cooray, A., Cortzen, I., Dye, S., Eales, S., Gavazzi, R., Harris, A. I., Herrera, C. N., Hughes, D., Jin, S., Krips, M., Lagache, G., Lehnert, M., Messias, H., Serjeant, S., Stanley, F., Urquhart, S., Vlahakis, C., and Weiss, A., *Close-up view of a luminous star-forming galaxy at z = 2.95*, A&A, 646 (2021), p. A122, https://doi.org/10.1051/0004-6361/202039743, https://doi.org/10.1051/0004-6361/202039743.

[3] S. Birrer and A. Amara, *lenstronomy: Multi-purpose gravitational lens modelling software package*, Physics of the Dark Universe, 22 (2018), p. 189–201, https://doi.org/10.1016/j.dark.2018.11.002, http://dx.doi.org/10.1016/j.dark.2018.11.002.

[4] A. S. Bolton, T. Treu, L. V. E. Koopmans, R. Gavazzi, L. A. Moustakas, S. Burles, D. J. Schlegel, and R. Wayth, *The sloan lens ACS survey. VII. elliptical galaxy scaling laws from direct observational mass measurements*, The Astrophysical Journal, 684 (2008), pp. 248–259, https://doi.org/10.1086/589989, https://doi.org/10.1086/589989.

[5] J. Brehmer, S. Mishra-Sharma, J. Hermans, G. Louppe, and K. Cranmer, *Mining for dark matter substructure: Inferring subhalo population properties from strong lenses with machine learning*, The Astrophysical Journal, 886 (2019), p. 49, https://doi.org/10.3847/1538-4357/ab4c41, http://dx.doi.org/10.3847/1538-4357/ab4c41.

[6] T. E. Collett, *The population of galaxy–galaxy strong lenses in forthcoming optical imaging surveys*, The Astrophysical Journal, 811 (2015), p. 20, https://doi.org/10.1088/0004-637x/811/1/20, http://dx.doi.org/10.1088/0004-637X/811/1/20.

[7] A. Diaz Rivero and C. Dvorkin, *Direct detection of dark matter substructure in strong lens images with convolutional neural networks*, Physical Review D, 101 (2020), https://doi.org/10.1103/physrevd.101.023515, http://dx.doi.org/10.1103/PhysRevD.101.023515.

[8] T. Dozat, *Incorporating nesterov momentum into adam*, 2016.

[9] S. Dye, S. A. Eales, H. L. Gomez, G. C. Jones, M. W. L. Smith, E. Borsato, A. Moss, L. Dunne, J. Maresca, A. Amvrosiadis, M. Negrello, L. Marchetti, E. M. Corsini, R. J. Ivison, G. J. Bendo, T. Bakx, A. Cooray, P. Cox, H. Dannerbauer, S. Serjeant, D. Riechers, P. Temi, and C. Vlahakis, *A high-resolution investigation of the multiphase ISM in a galaxy during the first two billion years*, Monthly Notices of the Royal Astronomical Society, 510 (2021), pp. 3734–3757, https://doi.org/10.1093/mnras/stab3569, https://doi.org/10.1093/mnras/stab3569, https://arxiv.org/abs/https://academic.oup.com/mnras/article-pdf/510/3/3734/42164443/stab3569.pdf.

[10] Y. D. Hezaveh, L. P. Levasseur, and P. J. Marshall, *Fast automated analysis of strong gravitational lenses with convolutional neural networks*, Nature, 548 (2017), p. 555–557, https://doi.org/10.1038/nature23463, http://dx.doi.org/10.1038/nature23463.

[11] W. Hu, D. J. Eisenstein, M. Tegmark, and M. White, *Observationally determining the properties of dark matter*, Physical Review D, 59 (1998), https://doi.org/10.1103/physrevd.59.023512, http://dx.doi.org/10.1103/PhysRevD.59.023512.

[12] Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, E. Acosta, R. Allsman, D. Alonso, Y. AlSayyad, S. F. Anderson, J. Andrew, and et al., *LSST: From Science Drivers to Reference Design and Anticipated Data Products*, , 873 (2019), 111, p. 111, https://doi.org/10.3847/1538-4357/ab042c, https://arxiv.org/abs/0805.2366.

[13] C. Jacobs, K. Glazebrook, T. Collett, A. More, and C. McCarthy, *Finding strong lenses in cfhtls using convolutional neural networks*, Monthly Notices of the Royal Astronomical Society, 471 (2017), p. 167–181, https://doi.org/10.1093/mnras/stx1492, http://dx.doi.org/10.1093/mnras/stx1492.

[14] H. Jin, Q. Song, and X. Hu, *Auto-keras: An efficient neural architecture search system*, 2019, https:

//arxiv.org/abs/1806.10282.

[15] C. R. KEETON, *A catalog of mass models for gravitational lensing*, 2002, https://arxiv.org/abs/astro-ph/0102341.

[16] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, 2017, https://arxiv.org/abs/1412.6980.

[17] F. LANUSSE, Q. MA, N. LI, T. E. COLLETT, C.-L. LI, S. RAVANBAKHSH, R. MANDELBAUM, AND B. PÓCZOS, *Cmu deeplens: deep learning for automatic image-based galaxy–galaxy strong lens finding*, Monthly Notices of the Royal Astronomical Society, 473 (2017), p. 3895–3906, https://doi.org/10.1093/mnras/stx1665, http://dx.doi.org/10.1093/mnras/stx1665.

[18] M. LOCHNER AND B. BASSETT, *Astronomaly: Personalised active anomaly detection in astronomical data*, Astronomy and Computing, 36 (2021), p. 100481, https://doi.org/10.1016/j.ascom.2021.100481, http://dx.doi.org/10.1016/j.ascom.2021.100481.

[19] R. MANDELBAUM, U. SELJAK, G. KAUFFMANN, C. M. HIRATA, AND J. BRINKMANN, *Galaxy halo masses and satellite fractions from galaxy-galaxy lensing in the sloan digital sky survey: stellar mass, luminosity, morphology and environment dependencies*, Monthly Notices of the Royal Astronomical Society, 368 (2006), p. 715–731, https://doi.org/10.1111/j.1365-2966.2006.10156.x, http://dx.doi.org/10.1111/j.1365-2966.2006.10156.x.

[20] J. MARESCA, S. DYE, A. AMVROSIADIS, G. BENDO, A. COORAY, G. DE ZOTTI, L. DUNNE, S. EALES, C. FURLANETTO, J. GONZÁLEZ-NUEVO, M. GREENER, R. IVISON, A. LAPI, M. NEGRELLO, D. RIECHERS, S. SERJEANT, M. TERGOLINA, AND J. WARDLOW, *Modelling high-resolution ALMA observations of strongly lensed dusty star-forming galaxies detected by Herschel*, Monthly Notices of the Royal Astronomical Society, 512 (2022), pp. 2426–2438, https://doi.org/10.1093/mnras/stac585, https://doi.org/10.1093/mnras/stac585, https://arxiv.org/abs/https://academic.oup.com/mnras/article-pdf/512/2/2426/43137932/stac585.pdf.

[21] M. MATEO, *Dwarf galaxies of the local group*, Annual Review of Astronomy and Astrophysics, 36 (1998), pp. 435–506, https://doi.org/10.1146/annurev.astro.36.1.435, https://doi.org/10.1146/annurev.astro.36.1.435, https://arxiv.org/abs/https://doi.org/10.1146/annurev.astro.36.1.435.

[22] W. R. MORNINGSTAR, L. P. LEVASSEUR, Y. D. HEZAVEH, R. BLANDFORD, P. MARSHALL, P. PUTZKY, T. D. RUETER, R. WECHSLER, AND M. WELLING, *Data-driven reconstruction of gravitationally lensed galaxies using recurrent inference machines*, The Astrophysical Journal, 883 (2019), p. 14, https://doi.org/10.3847/1538-4357/ab35d7, http://dx.doi.org/10.3847/1538-4357/ab35d7.

[23] E. MUNARI, A. BIVIANO, S. BORGANI, G. MURANTE, AND D. FABJAN, *The relation between velocity dispersion and mass in simulated clusters of galaxies: dependence on the tracer and the baryonic physics*, Monthly Notices of the Royal Astronomical Society, 430 (2013), p. 2638–2649, https://doi.org/10.1093/mnras/stt049, http://dx.doi.org/10.1093/mnras/stt049.

[24] V. NAIR AND G. E. HINTON, *Rectified linear units improve restricted boltzmann machines*, in Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, Madison, WI, USA, 2010, Omnipress, p. 807–814.

[25] R. NARAYAN AND M. BARTELMANN, *Lectures on gravitational lensing*, 1997, https://arxiv.org/abs/astro-ph/9606001.

[26] H. NAYYERI, M. KEELE, A. COORAY, D. A. RIECHERS, R. J. IVISON, A. I. HARRIS, D. T. FRAYER, A. J. BAKER, S. C. CHAPMAN, S. EALES, AND ET AL., *Candidate gravitationally lensed dusty star-forming galaxies in theherschelwide area surveys*, The Astrophysical Journal, 823 (2016), p. 17, https://doi.org/10.3847/0004-637x/823/1/17, http://dx.doi.org/10.3847/0004-637X/823/1/17.

[27] B. NORD, E. BUCKLEY-GEER, H. LIN, N. KUROPATKIN, T. COLLETT, D. L. TUCKER, H. T. DIEHL, A. AGNELLO, A. AMARA, T. M. C. ABBOTT, AND ET AL., *Observation and confirmation of nine strong-lensing systems in dark energy survey year 1 data*, Monthly Notices of the Royal Astronomical Society, 494 (2020), p. 1308–1322, https://doi.org/10.1093/mnras/staa200, http://dx.doi.org/10.1093/mnras/staa200.

[28] J. PEARSON, N. LI, AND S. DYE, *The use of convolutional neural networks for modelling large optically-selected strong galaxy-lens samples*, Monthly Notices of the Royal Astronomical Society, 488 (2019), p. 991–1004, https://doi.org/10.1093/mnras/stz1750, http://dx.doi.org/10.1093/mnras/stz1750.

[29] M. POSTMAN, D. COE, N. BENÍTEZ, L. BRADLEY, T. BROADHURST, M. DONAHUE, H. FORD, O. GRAUR, G. GRAVES, S. JOUVEL, AND ET AL., *The cluster lensing and supernova survey with hubble: An*

overview, The Astrophysical Journal Supplement Series, 199 (2012), p. 25, https://doi.org/10.1088/0067-0049/199/2/25, http://dx.doi.org/10.1088/0067-0049/199/2/25.

[30] G. D. RACCA, R. LAUREIJS, L. STAGNARO, J.-C. SALVIGNOL, J. L. ALVAREZ, G. S. CRIADO, L. G. VENANCIO, A. SHORT, P. STRADA, T. BÖNKE, C. COLOMBO, A. CALVI, E. MAIORANO, O. PIERSANTI, S. PREZELUS, P. ROSATO, J. PINEL, H. ROZEMEIJER, V. LESNA, P. MUSI, M. SIAS, A. ANSELMI, V. CAZAUBIEL, L. VAILLON, Y. MELLIER, J. AMIAUX, M. BERTHÉ, M. SAUVAGE, R. AZZOLLINI, M. CROPPER, S. POTTINGER, K. JAHNKE, A. EALET, T. MACIASZEK, F. PASIAN, A. ZACCHEI, R. SCARAMELLA, J. HOAR, R. KOHLEY, R. VAVREK, A. RUDOLPH, AND M. SCHMIDT, *The Euclid mission design*, in Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave, H. A. MacEwen, G. G. Fazio, M. Lystrup, N. Batalha, N. Siegler, and E. C. Tong, eds., vol. 9904, International Society for Optics and Photonics, SPIE, 2016, pp. 235 – 257, https://doi.org/10.1117/12.2230762, https://doi.org/10.1117/12.2230762.

[31] V. TRIMBLE, *Existence and nature of dark matter in the universe*, Annual Review of Astronomy and Astrophysics, 25 (1987), pp. 425–472, https://doi.org/10.1146/annurev.aa.25.090187.002233, https://doi.org/10.1146/annurev.aa.25.090187.002233, https://arxiv.org/abs/https://doi.org/10.1146/annurev.aa.25.090187.002233.

[32] M. VELANDER, E. VAN UITERT, H. HOEKSTRA, J. COUPON, T. ERBEN, C. HEYMANS, H. HILDEBRANDT, T. D. KITCHING, Y. MELLIER, L. MILLER, AND ET AL., *Cfhtlens: the relation between galaxy dark matter haloes and baryons from weak gravitational lensing*, Monthly Notices of the Royal Astronomical Society, 437 (2013), p. 2111–2136, https://doi.org/10.1093/mnras/stt2013, http://dx.doi.org/10.1093/mnras/stt2013.

[33] A. VERMA, T. COLLETT, G. P. SMITH, S. L. S. COLLABORATION, AND THE DESC STRONG LENSING SCIENCE WORKING GROUP, *Strong lensing considerations for the lsst observing strategy*, 2019, https://arxiv.org/abs/1902.05141.

[34] D. XU, D. SLUSE, L. GAO, J. WANG, C. FRENK, S. MAO, P. SCHNEIDER, AND V. SPRINGEL, *How well can cold dark matter substructures account for the observed radio flux-ratio anomalies*, Monthly Notices of the Royal Astronomical Society, 447 (2015), p. 3189–3206, https://doi.org/10.1093/mnras/stu2673, http://dx.doi.org/10.1093/mnras/stu2673.