Comparative Analysis of the Performance of a U-Net With and Without Attention for Dark Matter Subhalo Detection in Strongly Lensed Images

School of Physics and Astronomy
May 26, 2024
Aleyna Adamson and Benjamin Holmes

Contents

Abstract

Detecting low-mass dark matter subhalos in strongly lensed images enables the exploration of the true nature of dark matter. Current methods to detect subhalos both directly and statistically are complicated and take a long time to analyse each image. Therefore, we implement a machine learning solution to automatically detect the presence of subhalos in strongly lensed images and retrieve their mass. We present a comparison between two convolutional neural networks (CNNs). In one of the networks, we implement a novel technique - an attention gate - which enhances the ability of the network to identify and understand the most important features of an image. The networks are trained on images containing a source, main lens, and subhalos. We find that the network with attention can resolve masses as low as $M \gtrsim 10^{6.5} M_{\odot}$ compared to $M > 10^{7} M_{\odot}$ without. It also predicts up to 80% less false positives when trained on images containing multiple subhalos which proves advantageous when subhalos are overlapping in an image. However, when noise is added to the images the the addition of attention causes the network to match or be outperformed by the baseline network.

1 Introduction

The standard Lambda cold dark matter (ΛCDM) model provides an excellent description of the growth and structure of the observable universe at larger scales from explaining the formation of galaxy clusters (Crocce et al., 2016) to observations of temperature fluctuations in the Cosmic Microwave Background (Planck Collaboration et al., 2020). However, on smaller, sub-galactic scales, the behaviour of dark matter depends on the model. One of the fundamental predictions of the ΛCDM paradigm is that small dark matter structures form first before eventually merging into more massive dark matter halos. The Aquarius Project N-body simulations (Springel et al., 2008) found that a power law function can be used to describe the relative subhalo abundances:

$$\frac{dN}{dM} \propto M^{-\alpha}, \tag{1}$$

where N is the number of subhalos, M is the subhalo mass, and α is the gradient of the overall subhalo mass function (SHMF). This relation is consistent across N-body ΛCDM simulations and results in a relatively high abundance of low-mass subhalos ($M \lesssim 10^8 M_\odot$).

Alternative dark matter models can affect the structure of subhalos and their number for a given mass. For example, warm dark matter candidates (eg. Bode et al., 2001) are non-relativistic but have a non-negligible free-streaming velocity at early times. This would delay structure formation to a time when the Universe is less dense meaning subhalos of a given mass are less abundant. Therefore, searching for lower mass dark matter subhalos can act as a test for the ΛCDM paradigm.

Unfortunately, low-mass subhalos are difficult to detect. As the subhalo mass decreases, star formation becomes increasingly inefficient resulting in halos that are not very luminous as they are largely devoid of stars. This makes detecting low mass subhalos even in the Local Group challenging (Read et al., 2017; Fitts et al., 2017). Gravitational methods are therefore useful as they do not rely on baryonic activity to detect.

Searches for subhalos within the Local Group consist of analysing the effect of subhalos moving through tidal stellar streams (Bonaca et al., 2019; Banik et al., 2021) and the analysis of stellar motion in the Milky Way (Feldmann & Spolyar, 2015). Outside of the Local Group, the only method to detect low-mass subhalos is using images of strong gravitational lenses as these do not rely on the presence and distribution of baryons. Ideally, both the lens and the source will consist of galaxies as this will increase the effective area for perturbations to be observed. In this case, the shape of the lens and most of its mass come from the galaxy's dark matter halo. Subhalos that lie close to the lensed arcs are expected to cause perturbations with the amount of perturbation depending on the mass and location of the subhalo.

There have been many methods developed to detect subhalos from strongly lensed images. Metcalf & Amara (2012) used radio flux anomalies in strongly lensed quasars to measure and constrain the number, position, and mass of subhalos in the lens galaxy. Koopmans (2005) and Vegetti & Koopmans (2009) developed a technique to directly measure the subhalo mass and location based on perturbations to the surface brightness distribution of the lensed image. Daylan et al. (2018) developed a Bayesian data analysis framework to apply probabilistic cataloguing to a set of strongly lensed images using Markov chain Monte Carlo (MCMC) methods. To date, several potential dark matter subhalo candidates have been identified (Vegetti et al., 2010, 2012, 2014; Hezaveh et al., 2016; Nightingale et al., 2022)

Current methods to directly or statistically detect subhalos are not ideal as they require accurate lens modelling and arduous calculations. Inaccuracies in modelling can lead to false positives (Ritondale et al., 2019) and different models can result in different outcomes. For example, Vegetti et al. (2014) detected subhalos that were initially modelled with a Pseudo-Jaffe profile (Dalal & Kochanek, 2002) and were predicted to have significantly lower masses compared to when they modelled the same subhalos with a Navarro–Frenk–White (NFW) profile.

In the coming years, hundreds of thousands of strongly lensed systems are expected to be found by experiments such as Large Scale Synoptic Survey (LSST Science Collaboration et al., 2009), *Euclid*

(Refregier et al., 2010), and the Hubble Space Telescope (Pawase et al., 2014). Using these traditional methods to analyse these images individually would take far too long. Therefore, developing faster methods to find subhalos would increase the ability to use gravitational lenses to probe the nature of dark matter.

This can be achieved by applying machine learning techniques to the problem. Multiple studies focus on utilising machine learning to speed up the lens modelling process. Pearson et al. (2019) used CNNs to estimate the strong gravitational lens mass model parameters and found that errors in parameters increased exponentially with increasing signal-to-noise ratios (SNRs) in the images. Then, in Pearson et al. (2021), an approximate Bayesian CNN was trained to predict mass profile parameters and associated uncertainties.

Direct subhalo detection skips lens modelling and, in turn, drastically decreases computational complexity and cost. Diaz Rivero & Dvorkin (2020) focused on developing a binary classifier that identifies whether an image contains subhalos. The purpose of this network is to be the beginning of a pipeline where more advanced neural networks could investigate further, though it only achieved high accuracies for subhalos $\geq 5 \times 10^9 M_\odot$. Similarly, Alexander et al. (2020b) made a CNN to be used as part of a pipeline but aimed to distinguish different types of dark matter substructure. Their follow-up work developed the CNN to use unsupervised learning when inferring the presence of subhalos such that it is theory agnostic and does not rely on a current dark matter model being correct (Alexander et al., 2020a). Individually inferring subhalos can also be replaced by population-level statistics. Brehmer et al. (2019) built a neural network that assumes subhalos are present and then uses population inference to extract subhalo mass abundances and the slope of the SHMF.

Beyond detection is classification where individual subhalos are found and characterised. (Ostdiek et al., 2022) used a U-Net - a type of encoder-decoder network - to semantically segment images to identify the physical boundaries of the main lens and subhalos. The architecture used was originally developed for bio-medical image segmentation (Ronneberger et al., 2015) and appears to be one of the best architectures for image segmentation. The major pitfall was that the CNN performed poorly for images with high SNRs and light profiles more complex than a single Sérsic profile.

Amongst various machine learning methods to search for dark matter subhalos, similar problems arise. Increasing the SNR and the complexity of the light profiles in the images reduced the accuracy of the networks. Also, the sheer quantity of data and computational energy needed to train these networks requires powerful hardware which can be expensive.

A novel machine learning technique, attention mechanisms (Vaswani et al., 2017) may provide a solution to these problems. Attention can enhance the ability of the network to identify and understand the most important features of an image. This improves the detection capabilities of the network which would be useful when using noisy images or complex lens models. It could also improve the network's ability to detect perturbations caused by low-mass subhalos as many networks struggle to detect subhalos smaller than $M \lesssim 10^{8.5} - 10^9 M_\odot$ (Ostdiek et al., 2022; Diaz Rivero & Dvorkin, 2020). Therefore, with the addition of attention, the network could potentially detect subhalos with masses closer to wear different dark matter models diverge in the SHMF. Attention also offers the ability to focus on key parts of an image which reduces the number of training images needed compared to standard approaches.

In this paper, we present a comparison of two networks used to detect low-mass dark matter subhalos in strongly lensed images. Both have the same architecture with one of them having attention mechanisms implemented. These networks' method of object detection revolves around semantic image segmentation where the output is a prediction mask with the same dimensions as the input and each pixel is labelled as belonging to one of the classes. Each pixel is predicted to belong to one of nine classes: a subhalo from one of eight pre-established mass bins, or none of them (which will be referred to as the background). The specific architecture chosen was the U-Net (Ronneberger et al., 2015) which can learn features at multiple scales through down- and up-sampling features between convolutional layers. The advantage of semantic segmentation is that the pixel predictions can then be translated to physical subhalo predictions containing information about its location and mass. Training each network requires thousands of simulated images with associated ground truth masks that are created

using a separate algorithm. Following this, the networks can detect subhalos in unseen images.

This paper is organised as follows. Section 2 presents our methods which include generating the strongly lensed images and their labels (Section 2.1), the structure and training of our networks (Section 2.2), and the process of counting the subhalos in each image (Section 2.3). Section 3 details how our networks perform when various image parameters are changed. Finally, our conclusions are presented in Sections 4 and 5.

2 Methods

2.1 Data Generation

The aim of this project is automatic subhalo detection and classification using a convolutional neural network (CNN). The network can do this through supervised learning where each image inputted requires an accompanying target label. The training images are strongly lensed images generated using the publicly available software package, LENSTRONOMY (Birrer & Amara, 2018). Each image is 80x80 pixels with a resolution of 0.05" per pixel and contains a source light and gravitational lens (consisting of a smooth lens and subhalos). In Section 3 each image contains the same gravitational lens and source placed in the centre of the image which is shown in Figure 2. Images in Section 3.2.2 also contain noise simulated using Hubble Space Telescope (HST) parameters. The associated mask for each image contains a label for each pixel corresponding to the target objects in the image. The image and mask generation process is detailed below.

2.1.1 Generating Strongly-Lensed Images

Gravitational lensing is a phenomenon observed when an unperturbed ray passes a mass, M, at impact parameter, b, and is deflected by the angle $\hat{\alpha}$ where

$$\hat{\alpha} = \frac{4GM}{c^2 b}. \tag{2}$$

The geometry of a gravitational lens system is illustrated in Figure 1 (Narayan & Bartelmann, 1996a). The ray from source S is deflected by the angle $\hat{\alpha}$ at the lens and then reaches the observer O. The angle between the position of the source and the optical axis is $\boldsymbol{\beta}$ and the position of the image I and the optical axis is $\boldsymbol{\theta}$. The angular diameter distances between the lens and source and the observer and source are D_{ds}, and D_s, respectively. From Figure 1, the lens equation can be derived:

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{D_{ds}}{D_s}\hat{\boldsymbol{\alpha}}(\boldsymbol{r}). \tag{3}$$

For the full derivation see Schneider et al. (1992).

These equations are used to simulate the interactions between the light source, the main lens, and subhalos places in the main lens. The individual components required to simulate strong gravitationally lensed images are discussed below.

Light Source: Extended light sources such as galaxies provide more opportunities to detect substructure compared to point sources such as quasars (Hezaveh et al., 2013). A single elliptical Sérsic profile has been used in this paper to focus on the subhalo detections themselves since additional source complexity would affect results. The parameters of the light source across all images are identical.

A single elliptical Sérsic profile (Cardone, 2004) can be described by

$$I(R) = I_0 \exp[-b_{n_{ser}}(\frac{R}{R_{ser}})^{\frac{1}{n_{ser}}}], \tag{4}$$

where R is the distance from the centre of the ellipse, I_0 is the intensity at $R = 0$, R_{ser} is the radius of the light source, n_{ser}, is the Sérsic index, and $b_{n_{ser}}$ is described by

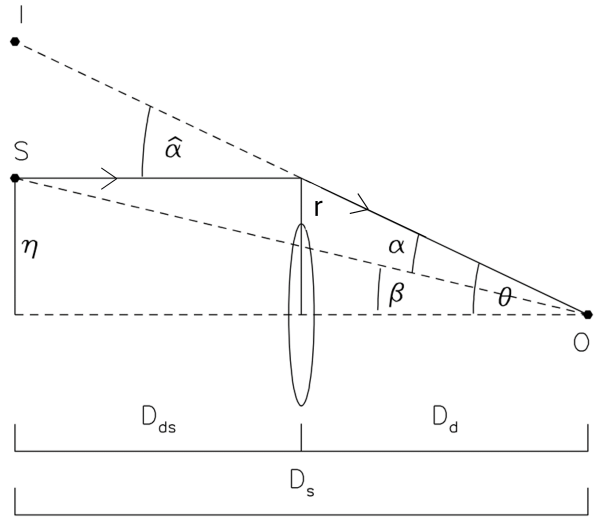

Figure 1: Gravitational lens diagram. The ray from source S is deflected by the angle $\vec{\hat{\alpha}}$ at the lens and then reaches the observer O. The angel between the position of the source and the optical axis is β and the position of the image I and the optical axis is θ. The angular diameter distances between the lens and source and the observer and source are D_{ds}, and D_s, respectively (Narayan & Bartelmann, 1996a).

$$b_{n_{ser}} \approx 2n_{ser} - \frac{1}{3}. \tag{5}$$

The parameters of the Sérsic profile used to generate the images in this paper are as follows:

$$I_0 = 1 \tag{6}$$

in arbitrary units of surface brightness, the radius of the light source is

$$R_{ser} = 0.1". \tag{7}$$

The x and y components of ellipticity ϵ_x and ϵ_y respectively are

$$\begin{aligned} \epsilon_x &= \frac{1}{3} \\ \epsilon_y &= 0, \end{aligned} \tag{8}$$

and the Sérsic index chosen as

$$n_{ser} = 0.5. \tag{9}$$

Finally, the location of the source is set to the centre of the image at $x = 0, y = 0$.

Main Lens: The distance to the main lens is fixed at a redshift $z = 0.2$ and the cosmology was decided from Planck Collaboration et al. (2016). The lens parameters outlined below result in a main lens with a mass of order $10^{13} M_\odot$. Once again, an ideal scenario is presumed and therefore the main lens is modelled as a singular isothermal sphere (SIS) as outlined in (Kormann et al., 1994). Here, the term isothermal indicates that the mass asymptotically decreases with 1/(distance). The surface mass density of a SIS is defined as

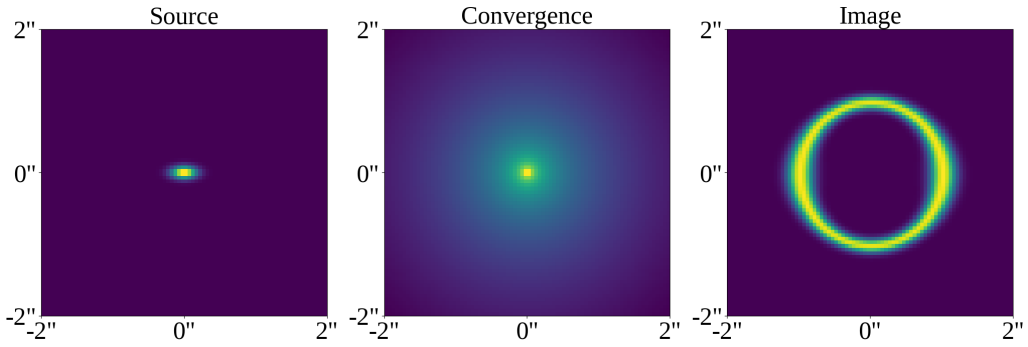

Figure 2: Components of simulating a strong gravitational lens. Light from the source galaxy (left) is deflected according to the convergence map (middle) which represents the mass density distribution of the lensing galaxy. The output from this calculation is the image seen on the right.

$$\Sigma(x, y) = \frac{v^2}{2G} \frac{1}{\sqrt{x^2 + y^2}}, \tag{10}$$

where (x, y) is the projected two-dimensional position on the lens plane, v is the velocity dispersion along the line-of-sight, f is the axis ratio, and G is the gravitational constant.

In LENSTRONOMY, the SIS profile is parameterised by the Einstein radius θ_E with a dimensionless surface mass density

$$\kappa(x, y) = \frac{1}{2}(\frac{\theta_E}{\sqrt{x^2 + y^2}}). \tag{11}$$

The Einstein radius relates to v by

$$\theta_E = 4\pi \frac{v^2}{c^2} \frac{Dds}{D_s} = \hat{\alpha} \frac{Dds}{D_s}, \tag{12}$$

where the angular diameter distances between the lens and source and the observer and source are D_{ds}, and D_s, respectively and $\hat{\alpha}$ is the deflection angle (Narayan & Bartelmann, 1996b).

The Einstein ring has been simulated to be perfectly circular for straightforward benchmarking between network performances. To do so, the Einstein radius is set to be

$$\theta_E = 1". \tag{13}$$

Additionally, the main lens is placed at the centre of the image at $x = 0, y = 0$. It is also implied that the ratio of the minor-to-major axis of the lens, q has to be

$$q = 1, \tag{14}$$

due to the spherical model. Figure 2 illustrates the source, convergence of the lens, and the resulting image.

Subhalos: The subhalos added into the lens are modelled with a truncated NFW profile (Navarro et al., 1997) with a density profile

$$\rho(r) = \frac{r_{trunc}^2}{r^2 + r_{trunc}^2} \frac{\rho_0(\alpha_{R_s})}{\frac{r}{R_s}(1 + \frac{r}{R_s})^2} \tag{15}$$

where R_s is the scale radius, $r_{trunc}^2 = 5R_s$, α is the deflection angle at R_s (Birrer & Amara, 2018).

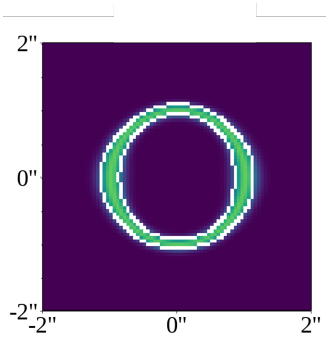

Figure 3: To ensure that subhalos cause detectable perturbations, their locations are randomly chosen from the pixels within the white contours which represent where the pixel brightness is at least 50% of the maximum brightness in the image.

The subhalos are added into the lens with a concentration parameter $c = 15$, defined as

$$c = \frac{R_{200}}{R_s}, \tag{16}$$

where R_{200} is the virial radius.

The subhalos are randomly placed exclusively in bright pixels which are defined as pixels at least 50% as bright as the brightest pixel in the image. This is illustrated in Figure 3. Subhalos are placed in bright pixels to maximise their effects on the Einstein ring enabling the network to learn relevant details from the images.

The masses of the subhalos are selected from eight mass bins uniformly distributed between $10^{6.25}M_{\odot}$-$10^{10.25}M_{\odot}$ with a width of $10^{0.5}M_{\odot}$. In Section 3.2 each image contains one subhalo from the corresponding mass bin for the test, indicated by the x-axis. In Section 3.3 each image contains one subhalo from each mass bin for a total of eight subhalos.

Noise: Within LENSTRONOMY, a SimulationAPI allows realistic simulations of specific telescopes. The infrared WFC3-F160W camera band is chosen and the point spread function (PSF) is modelled using a Tiny Time kernel (Krist et al., 2011). An exposure time of 3300s is used for a total of 4 exposures which is a realistic amount for a small-scale research team applying for HST time (Dye & Warren, 2005). The zero-point magnitude is set to 25.96 with a sky brightness of 22.3 mag/arcsec^2. The noise consists of Gaussian background noise calculated from the factors above, read noise, and shot noise. Since shot noise is dependent on the photon flux, the amplitude parameter (the half-light radius) used to simulate the source light is no longer arbitrary. Section 3.2.2 compares network performance against the source galaxy's apparent magnitude, m, which is used to calculate the amplitude for the simulation:

$$F = 10^{\frac{m+Z}{2.5}}, \tag{17}$$

where F is the flux (equal to double the amplitude parameter) and Z is the zero-point. Examples of the noise simulations for varying light source apparent magnitudes are seen in Figure 4.

2.1.2 Generating the Image Masks

Due to the network using supervised learning, the input data set requires labels as well as the images. These labels aim to provide the network with a target to learn from. For semantic segmentation, the labels are a set of masks that are paired with each generated image. These masks are created as a multidimensional array with a channel for each target. The only targets of concern are the subhalos as modelling lens parameters is not of interest in this project. The final channel is dedicated to the background pixels where the target is any pixel that does not already belong to one of the subhalos.

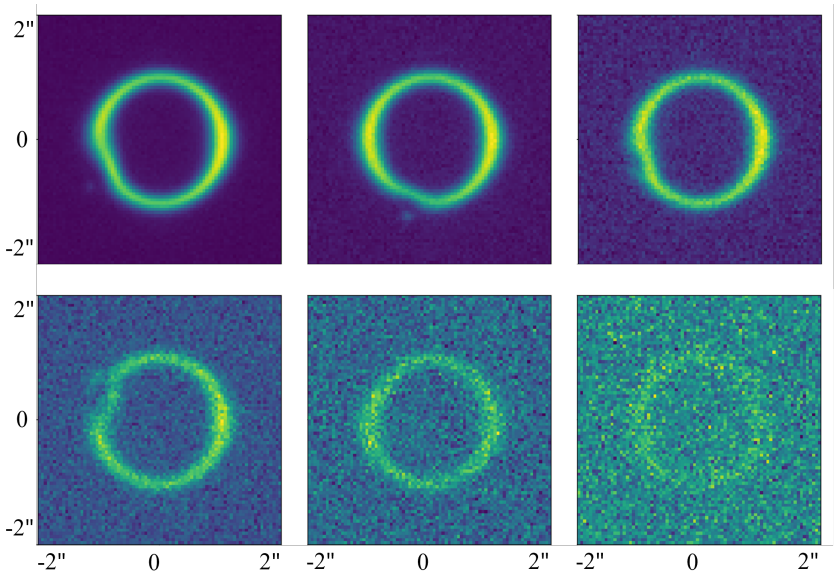

Figure 4: Noise is added based on HST simulations and the apparent magnitude of the source galaxy. From left to right, the apparent magnitudes used in the images are in the range [17, 23]. The subhalo mass used in each image is $10^{10} M_{\odot}$ to highlight the obscuring effect of noise, even for the largest perturbations.

In each of these channels we assign a label to each pixel by changing the value of the pixel. Each channel is a binary problem where any pixel that belonging to the target class is assigned a value of 1 and any other pixel is 0.

For all subhalo targets, the ground truth mask includes a circle of radius 2 pixels on the center of the location of the subhalo where every pixel within this circle belongs to the subhalo mass bin class. The size of the subhalo targets is kept equal across mass bins for more stability when training of the network. If the size of the target changed depending on the mass bin, such as larger circles for larger masses, then there would be an imbalance in the training data where some classes would contain more pixels belonging to the target. Accounting for this would include creating custom loss functions and having different-sized training datasets for different mass bins. An example of how these target masks are made is illustrated in 5 which shows an example of the masks used in Section 3.2.1. The image contains a single subhalo from the $10^{10} M_{\odot}$ mass bin with the target masks for the subhalo and the background.

In Section 3.2 where there is only one subhalo mass bin being tested, the target masks contain two channels: one for the subhalo target and the other for the background. In Section 3.3 where all mass bins are represented, the masks contain nine channels: eight for the subhalo mass bins and one channel for the background target.

2.2 Network Architecture

Image segmentation is a technique for separating features in an image into distinct regions. Semantic segmentation involves assigning a class to each pixel within an image such as, in the context of this project, a subhalo's mass bin or the background. Compared to simple object detection, the primary advantage is the precise object boundaries which can help resolve overlapping structures. Additionally, pixel-level details can be captured which helps to detect perturbations from low-mass subhalos below where the subhalo mass functions of different dark matter models diverge.

Encoder-decoder networks have been used in recent years to achieve semantic segmentation (Long et al., 2014). These networks consist of two parts: an encoder which extracts the features of an image while down-sampling to produce a compact representation before the decoder performs up-sampling to reintroduce spatial information. Unfortunately, these networks can sometimes lose information during this process. One solution is to implement skip connections between the equivalent encoder

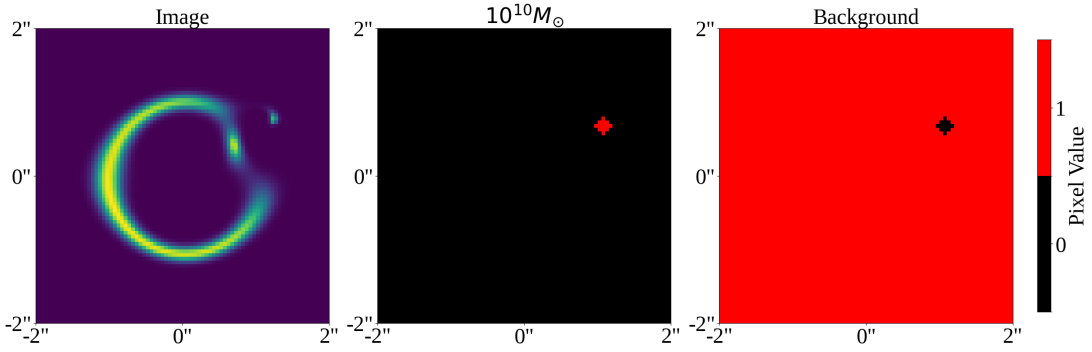

Figure 5: The image on the left contains one $10^{10}M_\odot$ subhalo. For the network to train, the ground truth mask requires a channel for where the subhalo is (middle) and for where there are no subhalos (right). The background mask is effectively the negative space of the combined subhalo masks.

and decoder levels to retain the higher resolution feature representation from earlier in the encoder; this architecture is known as a U-Net (Ronneberger et al., 2015). In this paper, we use the U-Net architecture to semantically classify each pixel in the simulated gravitational lens as belonging to one of the eight investigated mass bins or the background (no subhalo detection).

2.2.1 U-Net

The overall architecture of the U-Net is shown in Figure 6 which we implement using Tensorflow (Abadi et al., 2016). As an initial preprocessing step, we normalise each image by dividing it by the maximum pixel value. This nullifies the effect of surface brightness on the classification process.

Each red arrow represents a sequence of three layers that learn the features of an image. The first layer is a 2D convolution of the image using multiple 3x3 filters that output feature maps. The numbers above the white boxes in the figure indicate the number of feature maps per image at that point in the network. The weights within the convolution filter are iteratively changed throughout the training process to extract key features such that the final output image more accurately depicts each image's ground truth mask. The stride and padding are chosen to keep the image dimensions identical after the convolution. Since the network is trained on batches rather than individual images, batch normalisation (Ioffe & Szegedy, 2015) re-centres the mean and standard deviation across all feature channels (the outputs from the convolutional layer). This puts all values into a similar range with a mean of zero and unit standard deviation, helping to speed up training and preventing overfitting. The final operation in the sequence is a rectified linear unit (ReLU) that sets any negative value to zero while letting all other values through unchanged:

$$\mathrm{ReLU}(z) = \begin{cases} 0, z < 0 \\ z, z >= 0 \end{cases} \tag{18}$$

where z is the pixel value. This function adds non-linearity to the network which facilitates more complex learning.

In the encoder, green arrows indicate max pooling layers that halve the image dimensions by scanning 2x2 areas of the image with a stride of two and only taking the maximum value. Conversely, the blue arrows in the decoder indicate transpose convolutions that double the image dimensions. Due to consistently multiplying or dividing by two, each part of the encoder has an equivalent part of the decoder where the feature maps have the same dimensions. The skip connection (black arrows) between these two network parts concatenates these feature maps to combine contextual and spatial information from multiple scales.

The final convolution of the decoder outputs nine 80x80 prediction masks representing the different classes stacked into the channels of a single image. A Softmax activation function (Bridle, 1989) is applied channel-wise to convert the pixel values into probabilities using the equation:

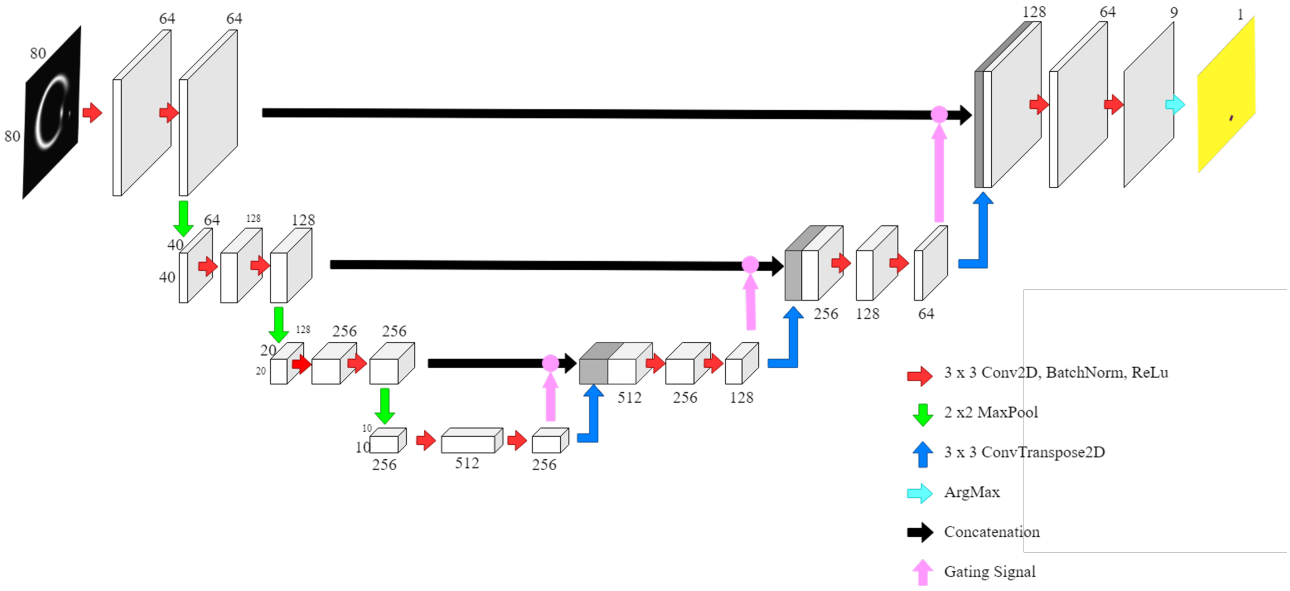

Figure 6: U-Net architecture including attention gates. The input 80x80 image is passed through the network to output an equally-sized prediction mask where each pixel corresponds to the predicted class index.

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}, \tag{19}$$

where z_i is the pixel prediction in channel i, and K is the overall number of output classes. Summing a given pixel's probabilities across all channels equals unity meaning each pixel in a channel represents the network's prediction of it belonging to that class. The final step is to apply an argmax function (cyan arrow) over all channels to collapse the nine channels into one where a pixel's value corresponds to the channel index it had the highest probability.

2.2.2 Attention

While the U-Net architecture has been successful in image segmentation applications, they have some intrinsic problems. An important part of the U-Net is the large number of feature channels which aid in propagating contextual information to higher resolutions (Ronneberger et al., 2015), but some channels may be redundant for the segmentation task and can harm the final output. Even with these additional feature channels, they can fail to capture enough contextual information between different feature scales to properly segment small or overlapping objects (Gu et al., 2021) leading to more false positives predictions. Finally, the non-linear nature of the network limits its application to dark matter subhalo detection as the results would be hard to interpret and confirm.

Attention mechanisms can address each of these issues. Many computer vision techniques for machine learning are based on human biology with attention being inspired by how the brain focuses on key sensory information while filtering out irrelevant stimuli (Lindsay, 2020). By mimicking this, neural networks can increase their accuracy without excessive computational resources.

Introduced by Oktay et al. (2018), attention gates highlight salient features by using the feature maps passed through the skip connection from the encoder and contextual information from the gating signal (the layer before up-sampling occurs in the decoder). Figure 7 illustrates the architecture of the attention gate. The attention gate has two inputs: the feature maps g in the decoder and the feature maps $\hat{x}^l$ in the encoder. The feature maps are down-sampled to the dimensions of the gating signal to minimise the complexity and number of trainable parameters before they are summed together element-wise where aligned weights grow larger while unaligned weights become smaller. The two inputs are merged via a convolution (W_g, W_x) and batch normalization (b_g, b_x), respectively. Following this, a ReLU activation σ_1 is applied to the result to set negative weights (irrelevant regions

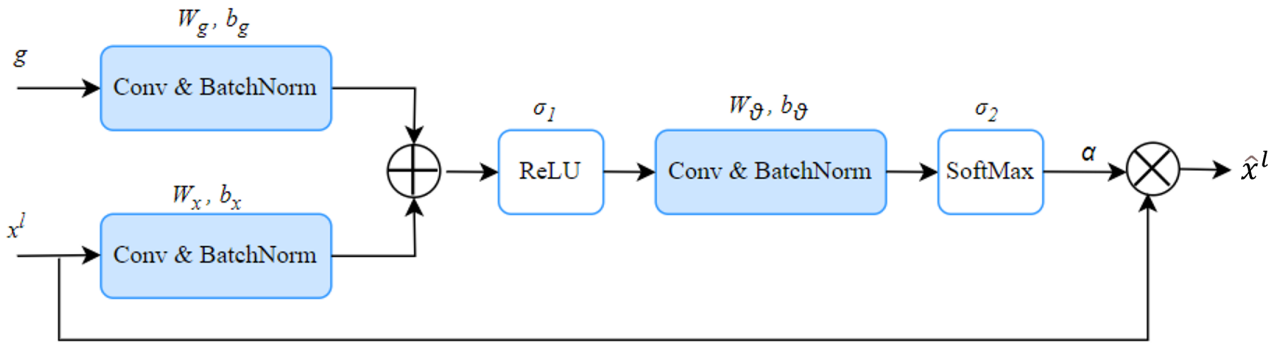

Figure 7: Overview of an attention gate. The attention weights are calculated by considering the activations from the input feature maps x^l as well as the broad contextual information from the gating signal g. x^l is then scaled by the attention weights before continuing through the skip connection (Oktay et al., 2018).

of the image) to zero such that they will be ignored while the remaining positive weights pass through unchanged. Another convolution is applied to the previous result, W_θ, to get a single filter where the values are the attention weights. Subsequently, a Softmax σ_2 activation turns these into probabilities and gives the attention coefficient α_i before up-sampling them to the original dimensions of the feature maps that were passed through the skip connections. Finally, each feature map is multiplied by the α_i so that relevant features are highlighted and continue through the skip connection to where they are concatenated. The feature mapping process in the attention gate can be expressed by

$$F = \sigma_1(W_x^T x^l + W_g^T g_i + b_g) \tag{20}$$

$$\alpha_i = \sigma_2(W_\theta^T F + b_\theta) \tag{21}$$

$$\hat{x}^l = x^l \alpha_i \tag{22}$$

In Figure 2.2.1 the application of attention gates at each skip connection are shown by the pink arrows. This ensures that information from multiple feature scales is aggregated to increase the amount of contextual information. The attention filter can also be visualised in the form of a heatmap overlayed onto the ground truth mask to assist in interpreting the results of the network.

2.2.3 Training

Each dataset generated consists of 10^4 images where 80% are designated for training and the remaining 20% is split equally between validation and testing. For Section 3.2, each image consists of one subhalo, the source light, and a smooth lens. For Section 3.3 each image consists of one subhalo from each mass bin along with the source light and smooth lens.

Network performance is tracked during training to monitor network stability. As this is a multi-class simulation, a categorical-cross-entropy loss function is implemented to quantify the difference between the predicted outputs of the network and the actual target values. The categorical-cross-entropy per pixel is given by

$$L = \frac{-1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{K} y_k^{(i,j)} \log(\hat{p}_k^{(i,j)}), \tag{23}$$

where n is the number of images, p is the number of pixels in an image, K is the number of classes, and $y_k^{(i,j)}$ represents the true probability of pixel j in image i to belonging to class k. Each pixel either belongs or does not belong to a class therefore $y_k^{(i,j)}$ is always a 1 or a 0. Finally, $\hat{p}_k^{(i,j)}$ is the

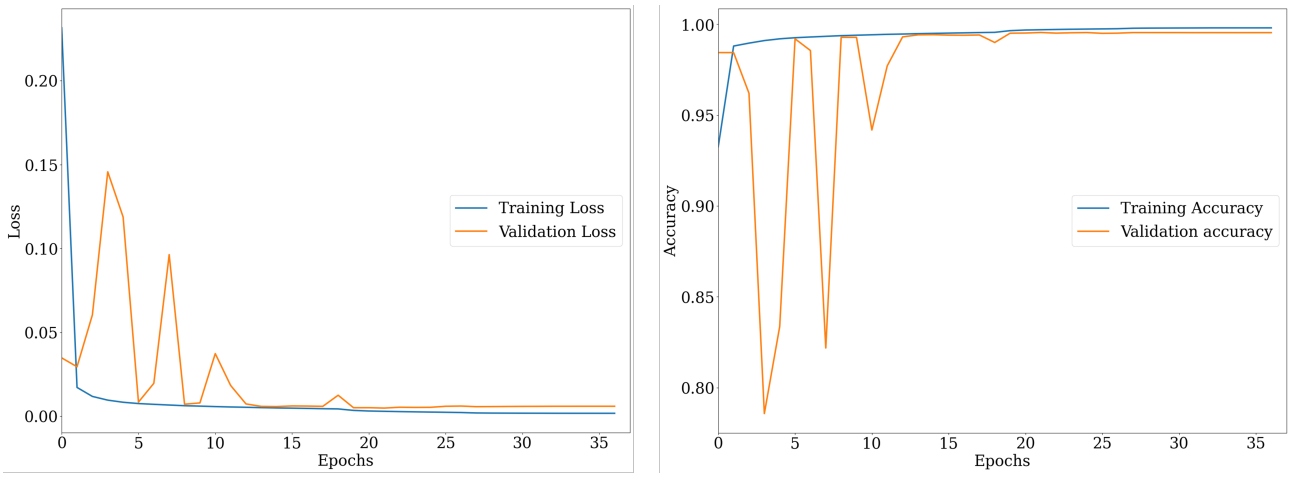

Figure 8: Example of training metrics. The left plot shows the categorical cross-entropy loss as a function of epoch during the training regime. The validation loss does not improve after epoch 22 resulting in the training being stopped early at epoch 37. The right plot shows the combined pixel accuracy for all classes. The dips in accuracy are a result of the Adam optimiser receiving unideal training batches for optimisation.

probability predicted by the model.

During training, the loss is minimised using the Adam optimiser (Kingma & Ba, 2014) with a learning rate of 10^{-3} and the default moving average value. If the validation loss does not decrease in five epochs, the learning rate is reduced by a factor of 10 down to a minimum learning rate of 10^{-6}. The training ends if the validation loss does not decrease in a total of 15 epochs. A batch size of 32 was used for the batch normalisation.

Another metric used to track the network is categorical pixel accuracy which is defined as

$$\text{Pixel Accuracy} = \frac{\text{Number of correct pixels in prediction mask}}{\text{Total number of pixels}}. \tag{24}$$

To determine whether a pixel is correct, a channel-wise Argmax function is applied to both the prediction and ground truth masks. This results in a one-channel image where each pixel value corresponds to the class index where it had the highest probability ($\max \hat{p}_k^{(i,j)}$). If a given pixel in the prediction mask has the same value in the ground truth mask, it is identified as a correct pixel where both the mass bin and location were correctly predicted.

The networks are trained using one NVIDIA L4 GPU accessed through Google Colab. An example of how the accuracy and loss change over the training regime is shown in Figure 8.

2.3 Counting Subhalos

While the pixel accuracy is a valid measure of performance for the network itself, the actual number of subhalos can not be gauged from this metric. Furthermore, the locations and masses of subhalo predictions are equally as important for scientific investigations of, for example, the subhalo mass function (Ostdiek et al., 2022) or galaxy formation and evolution (Diemand et al., 2007).

The solution implemented is based on the pixel accuracy equation from Section 2.2.3 which already considers location and class. Groups of predicted pixels that are connected orthogonally or diagonally are each given a unique label representing that group using connected component analysis. Specifically, a two-pass algorithm (Hoshen & Kopelman, 1976) is applied using a 3x3 structuring element containing only ones. This requires a binary image which has been achieved by simply rounding the probabilities of the prediction masks. Pixels with a probability ≥ 0.5 are rounded to have a value of 1 and those with probabilities < 0.5 are rounded to a pixel value of 0.

Looping through each distinct group, the number of overlapping pixels between the group and the ground mask label is counted. Since there is at least one subhalo in each image, there are no true

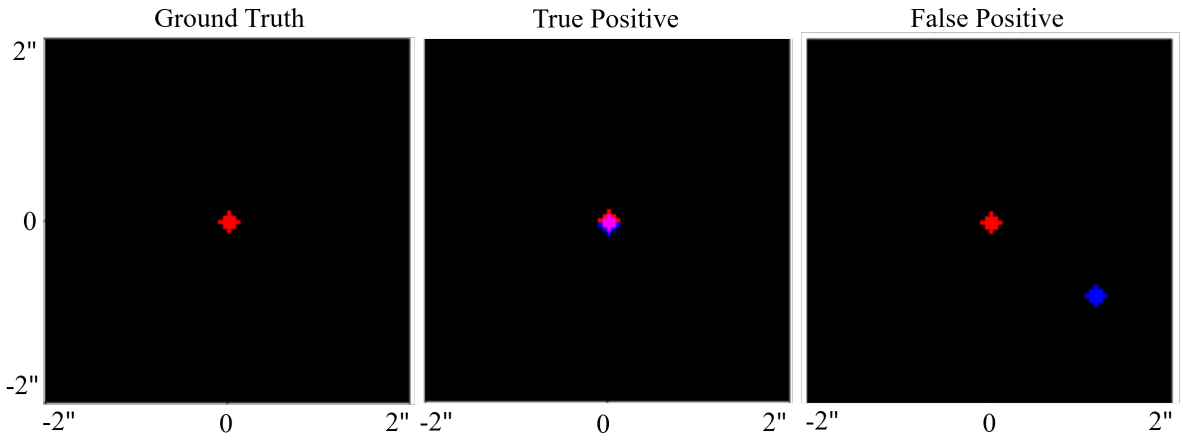

Figure 9: Examples for different classification cases. The left image is the ground truth subhalo mask (red). The middle image shows an example of a TP prediction as there is more than a 30% overlap between the prediction mask (blue) and the ground truth. The image on the right shows a FP prediction where the network completely misses the target. Since there is also no TP in this case, the image would be classified as containing both a FP and a FN.

negative (TN) predictions leaving three classifications a group of pixels can be given:

- **True Positive (TP)** - if more than 30% of pixels in the prediction group overlap, it is determined to be a TP. This threshold was decided to give the classification more leniency and could be changed for more robust investigations. If a second, disjointed group of pixels is also found within the ground truth subhalo mask, it is not counted.
- **False Positive (FP)** - if there is less than 30% overlap - meaning it is almost a true positive or it is in a completely incorrect position - it is counted as a FP.
- **False Negative (FN)** - if there are zero predicted subhalos after rounding, or if none of the predictions were TPs, then the lack of a correct prediction implies a FN.

Figure 9 demonstrates situations where each of these classifications could occur.

Iterating over the test dataset, the number of each of these classifications for each class are accumulated. These values can then be used to assess the quality of the classifications. Accuracy measures how often the model correctly predicts the outcome:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}}. \tag{25}$$

Accuracy does not handle class imbalances well, however, which makes it a misleading metric on its own if there are more subhalos of a given mass such as low mass subhalos in the ΛCDM model Springel et al. (2008).

Precision considers class imbalance by measuring how often positive predictions are correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{26}$$

This metric is important to reduce bias in the results and ensure follow-up validation studies investigate actual subhalos. The downside is that it does not consider FNs when a subhalo is missed.

Finally, recall completes the picture of network classification by measuring the fraction of correctly identified TPs:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{27}$$

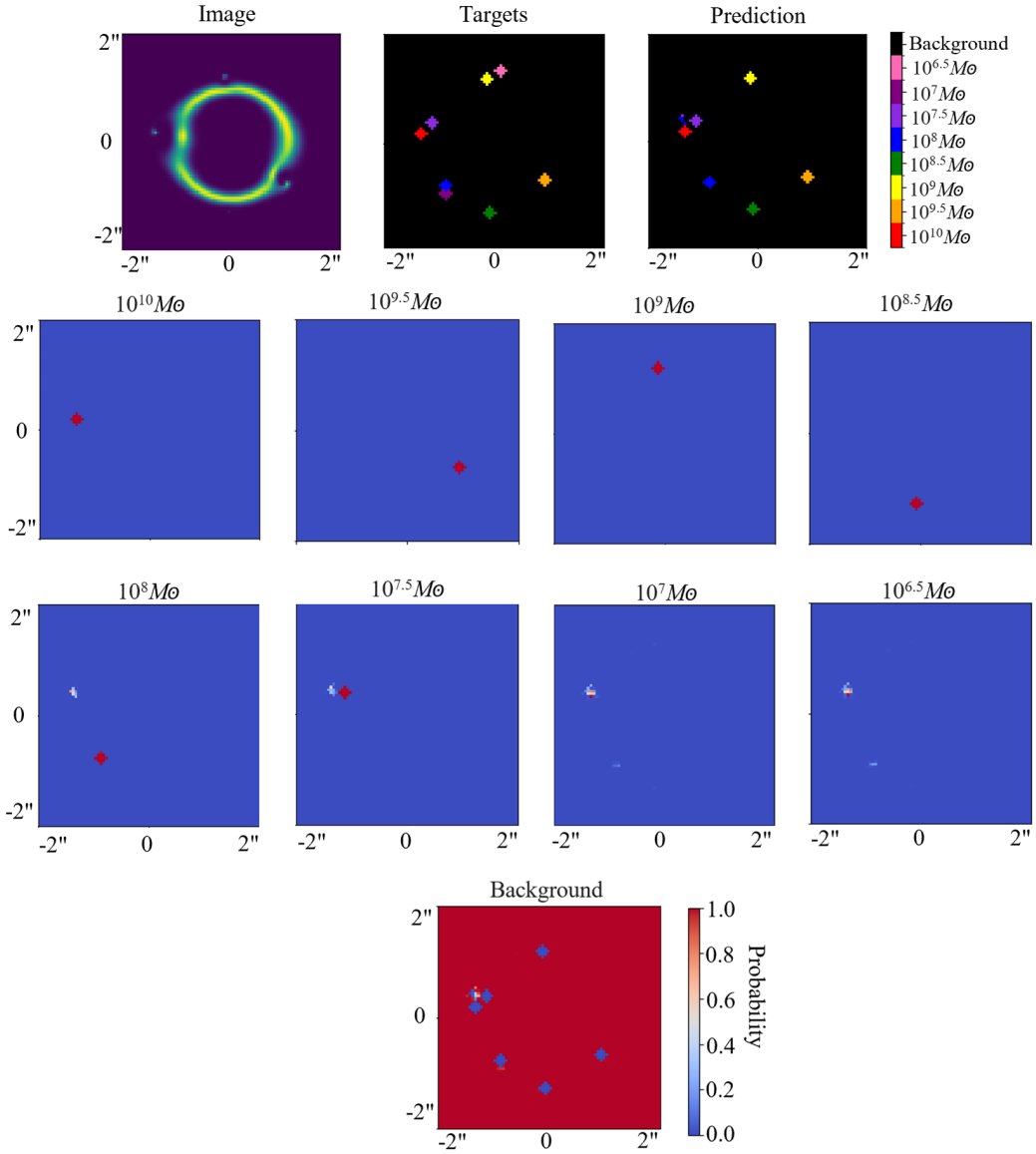

Figure 10: (From left to right) The top row shows an image of a gravitational lens with multiple subhalos added, which is input to the network with attention, the target labels for each pixel, and the prediction from the network. The network prediction for each pixel is determined by the class with the maximum probability for that pixel. The bottom three rows show the probability assigned to each class for each pixel where a red is high probability, the blue is low probability, and the white is around 0.5. The white pixels show where the network is unsure whether the pixel does or does not belong to that class. The classes include eight mass bins and a background class.

It is effectively a measure of network sensitivity that also considers class imbalance and indicates the proportion of missed subhalos. Opposite to precision, it does not consider FPs but all three of these metrics together provide the full scope of the network's capabilities for binary classification within each class.

3 Results

We train the networks on a different data set in each section of the results. Then, the trained networks (the U-Net with and without attention) are tested on unseen images in the test datasets. This section first displays example outputs of the network (Section 3.1) to visualise the prediction

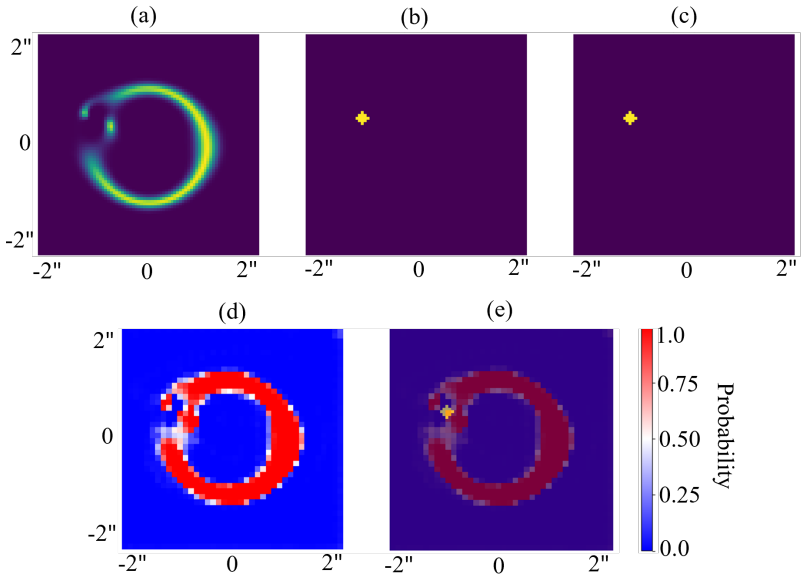

Figure 11: (a) Simulated strong gravitational lens containing a single $10^{10}M_{\odot}$ subhalo. (b) Ground truth mask. (c) Prediction mask outputted from the network. (d) Attention map visualised in the form of a heatmap where red pixels are large attention weights and blue is small. (e) The prediction mask and heatmap overlayed to interpret how the attention U-Net predicted the subhalo's location.

channels and attention heatmaps. Next, the single subhalo tests with variable mass (Section 3.2.1) and noise (Section 3.2.2) demonstrate each network's adaptability to different parameters. Lastly, the test involving multiple subhalos in a single image (Section 3.3) will show the effect that overlapping structures can have on network performance.

3.1 Example Results

Figure 10 shows an example of the output from the network with attention. The first image on the top row shows the image that is input to the network, the second shows the ground truth target labels, and the third shows the prediction of the network. Each pixel in the prediction is assigned to the class that had the highest probability for that pixel. The remaining images show the individual class probabilities where red pixels represent higher confidence and blue pixels represent lower confidence. The white pixels show intermediate probabilities around 0.5 where the network is unsure whether the pixel does or does not belong to that class.

Looking closer at the first row of Figure 10, six out of the eight subhalos have been correctly identified. From a glance, the network appears to be very confident of the position and mass of the most massive subhalos in the second row ($10^{10}M_{\odot}, 10^{9.5}M_{\odot}, 10^{9}M_{\odot}$, and $10^{8.5}M_{\odot}$) as there are no visible white pixels. For classes $10^{7.5}M_{\odot}$ and $10^{8}M_{\odot}$, the network is confident that there is at least one subhalo belonging to that class but the presence of white pixels suggests that it is unsure whether there may be another. Some of these white pixels appear on the prediction mask as additional FP predictions, once again where the largest perturbation in the image is. Nevertheless, the network was able to detect the $10^{7.5}M_{\odot}$ subhalo despite its proximity to the largest mass subhalo.

The two false negative predictions are expectedly the two least massive subhalo mass bins ($10^{7}M_{\odot}$ and $10^{6.5}M_{\odot}$). In both instances, they are close to or overlapping with a more massive subhalo. Since they don't have a distinct, isolated perturbation, the network seems to presume that their effects were incorporated into the largest perturbation in the image caused by the combined effect of the $10^{10}M_{\odot}$ and $10^{7.5}M_{\odot}$ subhalos. This observation is supported by the white pixels showing a lack of confidence in the prediction.

To interpret what attention is contributing to the classification, a map of the attention weights is shown in Figure 11. For clarity, the sample was taken from an image containing a single $10^{10}M_{\odot}$ sub-

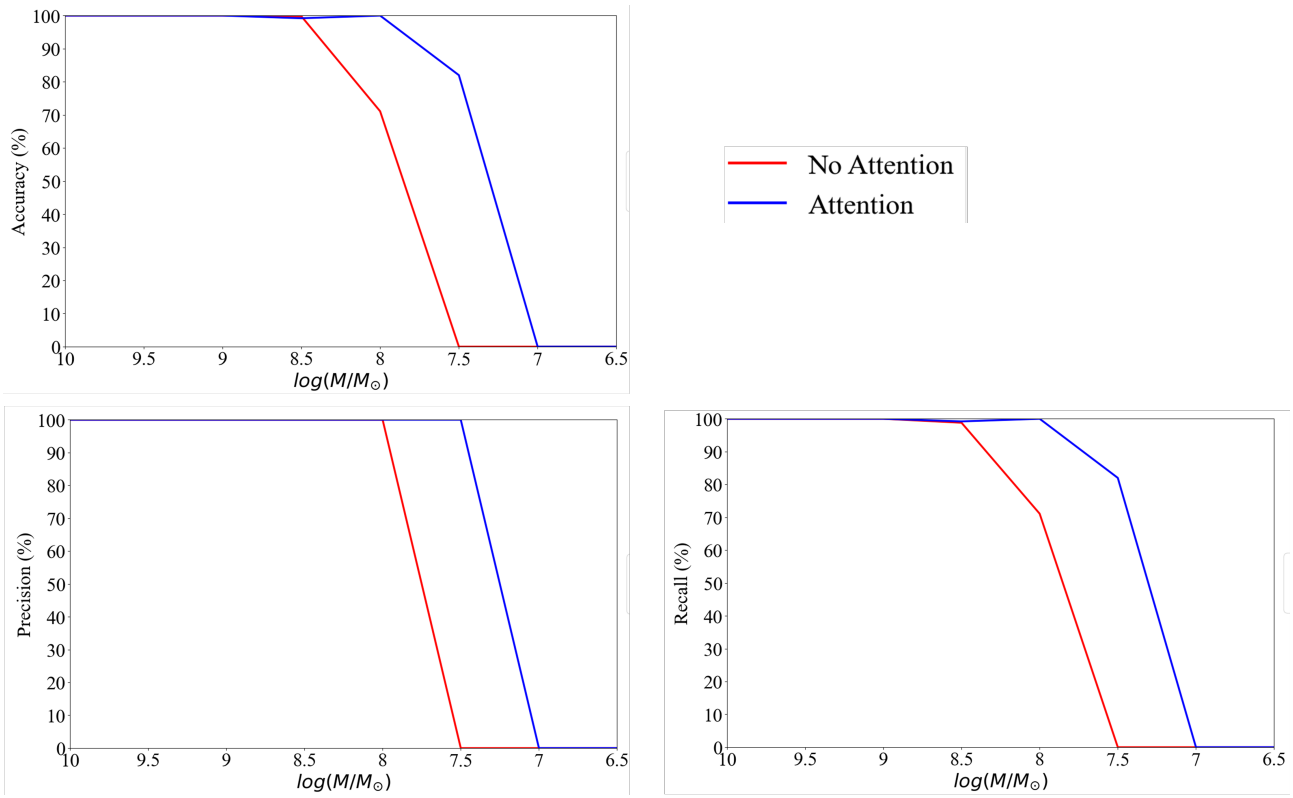

Figure 12: The performance of each network as the mass of the subhalos in the images are decreased. The blue and red lines in each plot represent predictions made by the network with and without attention, respectively. The top left plot shows the accuracy for overall correct subhalo predictions for each mass bin. The bottom plots show the precision (left) and recall (right) of the networks.

halo. Surprisingly, the attention weights almost exactly recreate the shape of the perturbed Einstein ring rather than the subhalo prediction. This is most likely due to a perturbation inherently being a feature of the Einstein ring rather than having distinct qualities itself. The attention network is most likely indirectly predicting the subhalo from the centre of the isolated region of low attention weights using contextual information from the shape of the Einstein ring.

3.2 Single Subhalo Tests

With network predictions now understood, we investigate how the accuracy of our network changes for varying simulation parameters. Network performance for variable subhalo masses can indicate a low-mass detection cutoff point, and the ability to detect subhalos with various magnitudes of noise can show how sensitive the network is to mass when perturbations are obscured.

Each network is trained using images containing a single subhalo. A different set of 10^4 images is used to train the networks for each mass bin and apparent magnitude. After training, the network is tested on an unseen 1000 images where the number of subhalos predicted is counted using the methods outlined in Section 2.3.

3.2.1 Mass Sensitivity

Low-mass subhalos, while difficult to observe due to the suppression of star formation, offer the most information about the dark matter present. It is therefore paramount that future observations are able to accurately detect and characterise these subhalos. This investigation aims to compare how attention can improve the accuracy and possibly lower the mass sensitivity cut-off when compared to the standard U-Net.

In Figure 12, the accuracy, precision, and recall for both networks are found. The blue and red

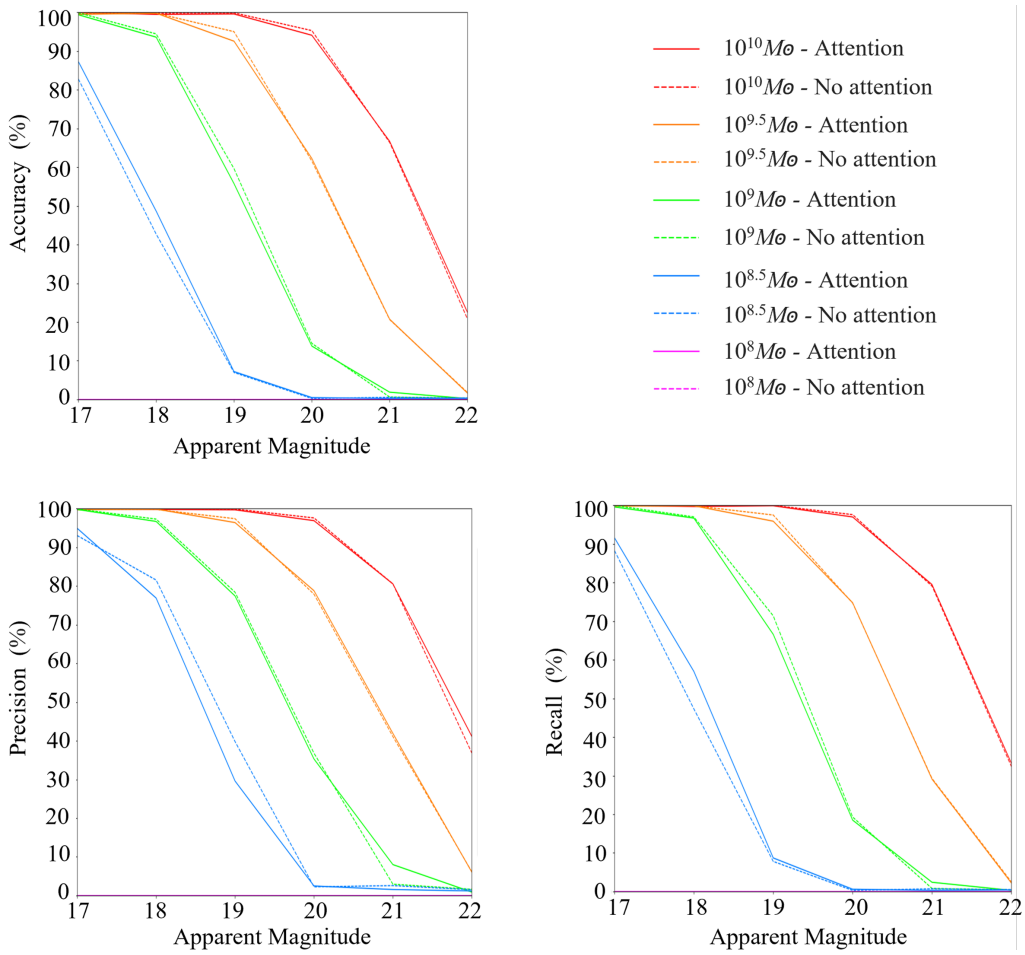

Figure 13: The performance of each network as the SNR is increased. The solid lines in each plot represent the network with attention and the dashed lines represent the network without attention with the mass bin indicated by colour. The top left plot shows the accuracy for overall correct subhalo predictions. The bottom plots show the precision (left) and recall (right) of the networks.

lines in each plot represent predictions made by the network with and without attention, respectively. The x-axis in each plot denotes the subhalo mass bin that the network was trained and tested on.

The top left plot shows the accuracy for overall correct subhalo predictions for each mass bin. Both networks maintain a high degree of accuracy for mass bins down to $10^{8.5}M_\odot$ before the standard U-Net starts declining. While the attention U-Net has does not see this dip for another mass bin, there is a small lapse in accuracy at $10^{8.5}M_\odot$ due to a small number of FNs though it is unclear why this occurred. The overall shape of the plots is roughly the same for both networks with attention allowing detection for an extra subhalo mass bin. Similarly, the precision (bottom left) and recall (bottom right) plots show the same trend. The only other unique feature is that neither network predicts any FPs shown by the precision dropping from 100% to 0% where the mass sensitivity is at its limit. This shows attention does not have a significant effect on how often positive predictions are correct.

It is clear that, in general, the addition of attention improves the performance of the network. When attention is added, the network can detect subhalos masses $M \lesssim 10^7 M_\odot$ compared to $M \lesssim 10^{7.5} M_\odot$ without.

3.2.2 Noise Sensitivity

The aim of classifying network performance on increasing SNRs is to identify the limit where the network begins to fail and confuse noise for perturbations caused by a subhalo. Since the size of a perturbation is partially dependent on the subhalo mass, it will take less noise for low-mass subhalo

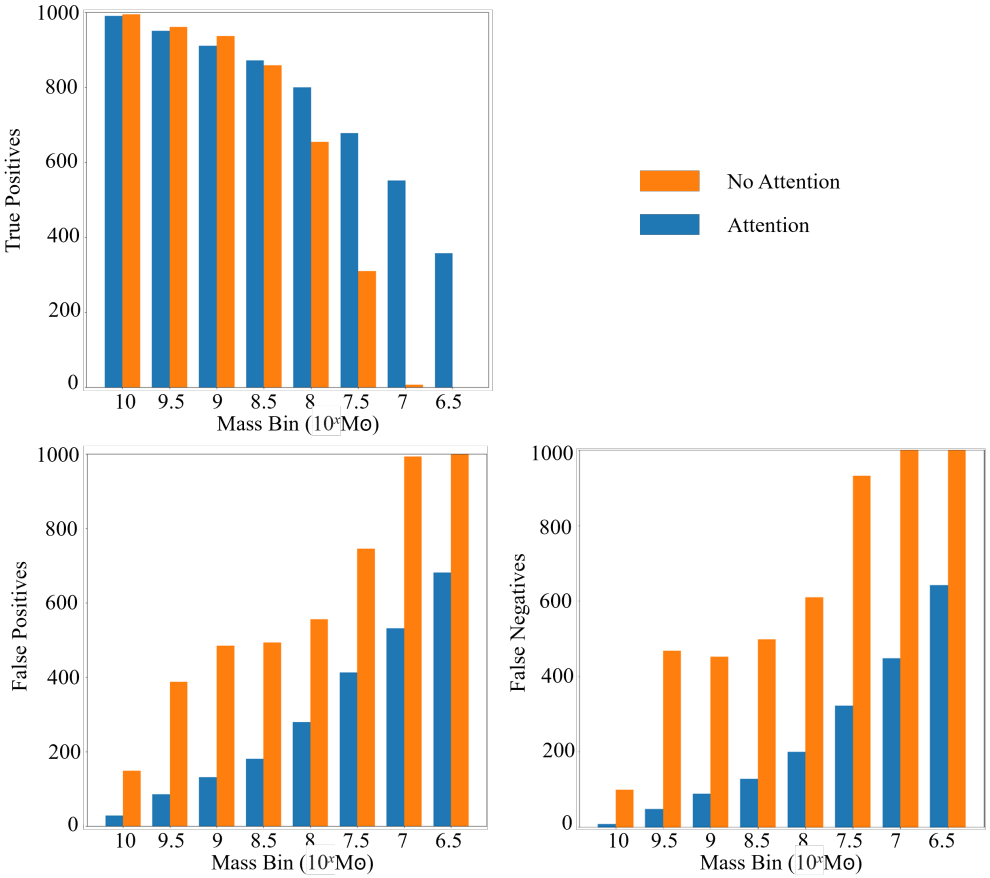

Figure 14: Bar charts plotting the number of TPs, FPs, or FNs per mass bin for both networks. The total value of these three values combined for a given mass bin may add up to more than 1000 (the number of subhalos of that mass in the test dataset).

detections to be missed compared to high-mass subhalos.

In Figure 13, the solid lines in each plot represent the network with attention and the dashed lines represent the network without attention. The mass bin and type of network used are indicated in the legend. The accuracy is shown in the top left, precision in the bottom left, and recall in the bottom right.

Reminiscent of Section 3.2.1, the lines are approximately the same shape in each plot with apparent magnitudes causing a translation effect. In contrast, however, attention changes the results very little. The largest differences are the 5-10% improved precision that attention provides for the $10^{8.5} M_{\odot}$ mass bin up to apparent magnitude 19 light sources and, conversely, the approximately 5-10% decreased recall and accuracy for the same simulation parameters. This implies that attention reduces the rate of FPs but has a higher rate of FNs for this subset of parameters. In general, attention seems to provide some marginal improvements for medium to high-mass subhalos while the standard U-Net slightly outperforms when classifying images with high apparent magnitude light sources.

3.3 Multiple Subhalo Test

While testing singular subhalo detections against varying simulation parameters is important, realistic systems will include many subhalos that are potentially overlapping. To compare network performance in these more complex situations, we simulate each image to contain one subhalo from each mass bin (8 subhalos total). The procedure for training and testing is the same as in Section 3.2.

Figure 14 shows how the amount of TP, FP, and FN predictions vary between each model for the mass bins. The blue and orange bars represent the predictions of the network with and without attention, respectively. The top left plot shows the number of TPs the networks predict. The values

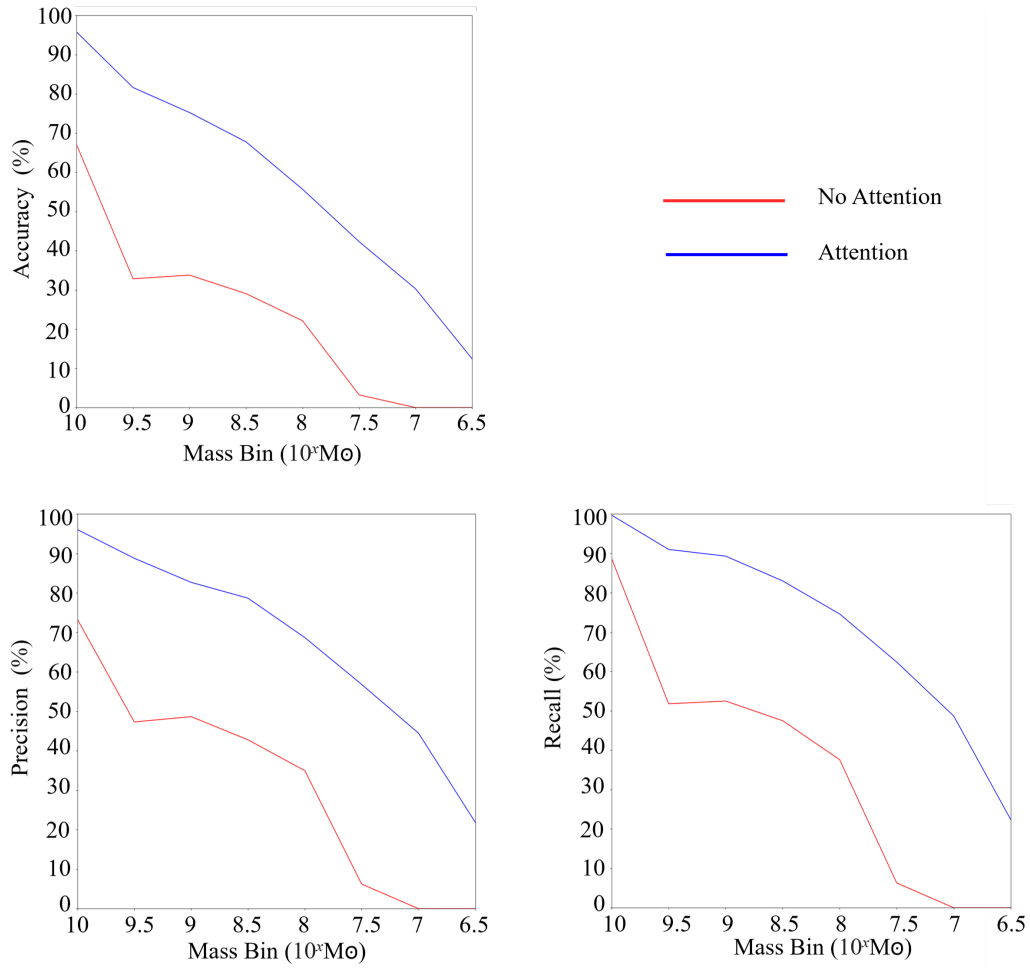

Figure 15: The performance of each network when one subhalo from each mass bin is added to each image (8 total subhalos). The blue and red lines in each plot represent predictions made by the network with and without attention, respectively. The top left plot shows the accuracy for overall correct subhalo predictions for each mass bin. The bottom plots show the precision (left) and recall (right) of the networks

are found by treating each class channel separately and applying the algorithm in Section 2.3 individually. Even with overlapping substructures, the addition of attention improves the performance of the network by correctly detecting more subhalos down to lower mass bins. The number of TPs detected by the standard U-Net also decreases more rapidly as subhalo mass decreases showing that it does not differentiate between overlapping perturbations as well as the attention U-Net.

The bottom two plots show the amount of FP and FN the networks predict. The rate of FP detection with attention is considerably lower than without showing attention is less likely to predict a subhalo where there is not one present. The rate of FN prediction with attention is also considerably lower. This means that with the addition of attention, the network is less likely to predict the absence of a subhalo when in fact there is one present or at least gets a TP prediction more often.

Figure 15 shows the overall performance of the networks. The blue lines in each plot represent predictions made by the network without attention and the red lines represent the predictions made by the network with attention. The top left plot shows the accuracy with which the networks predicted the subhalos correctly for each mass bin. The bottom left plot shows the precision of predictions made by the networks across all mass bins and the bottom right plot shows the recall of the predictions from the networks across all mass bins.

The addition of attention is explicitly shown to improve the performance of the network. When attention is added, the network is not only able to detect more subhalos from each mass bin but can also detect subhalos from lower mass bins. With attention, the network can detect subhalos from

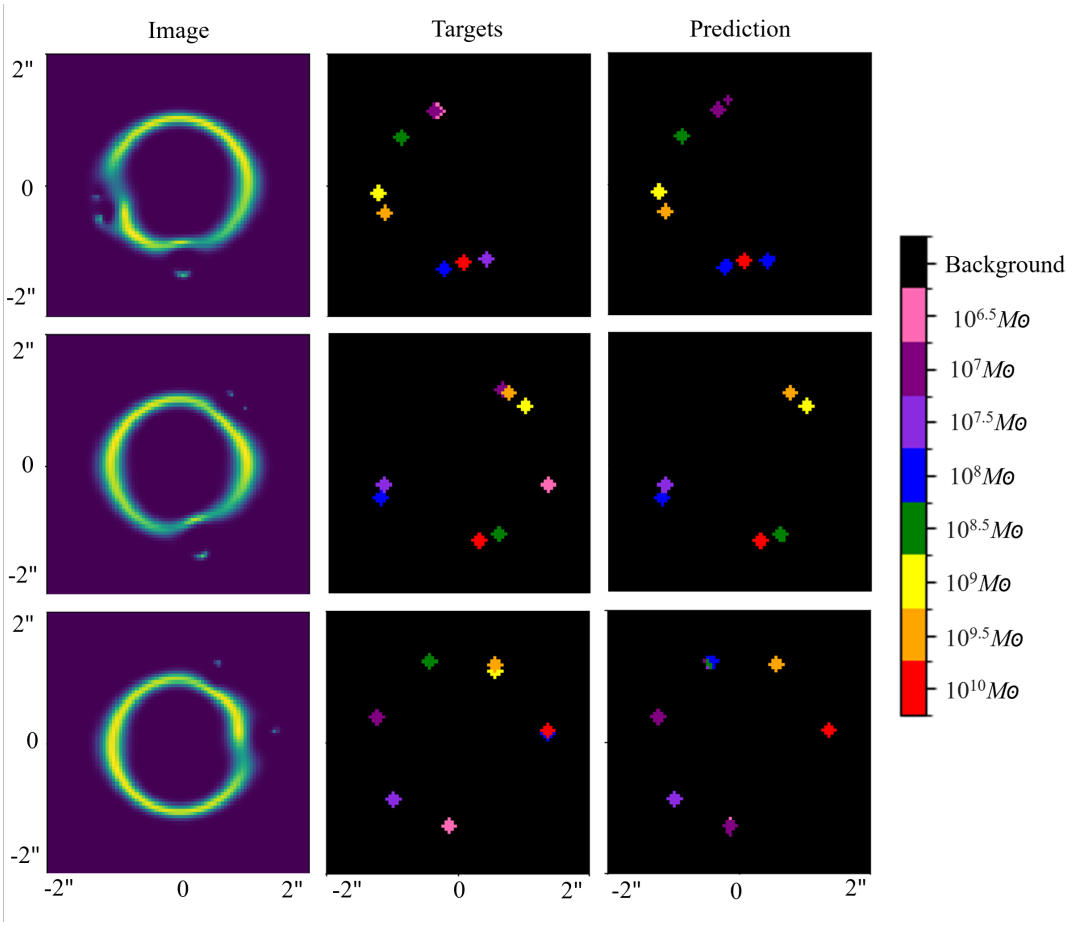

Figure 16: Example outputs of our network with attention. The left column shows the images that are input into the network. The second column shows the target labels of each pixel in the image. The third column shows the prediction of the network. These images show how the network can react when there are multiple subhalos in the images including overlapping subhalos.

mass bins at least as small as $10^{6.5} M_\odot$ whereas it fails at $M \lesssim 10^7 M_\odot$ without attention.

Figure 16 shows three example outputs from the network with attention. The first column shows the images input into the network, the second column shows the targets which the network did not see before making its prediction, and the third column shows the networks prediction. Across all predictions it struggles to find the lowest mass subhalos in the $10^{6.5} M_\odot$ mass bin which we expected. It also struggles when subhalos overlap in an image. In each of the images there is a significant overlap between at least two of the subhalos which the network predicts to be just one. Each time, the networks seems to predict the more massive subhalo of the overlapping pair. It is also evident from Figure 16 that the network can detect the locations of subhalos but struggles more with predicting which mass bin they belong to.

4 Discussion

We have developed a neural network capable of detecting and classifying subhalos in simulated strong gravitational lenses. The network architecture used is a U-Net that outputs a semantically segmented image where each pixel corresponds to one of the subhalo mass bins or none of them. Attention gates were incorporated into a secondary version of the network to compare whether it increases performance. Both networks trained on images containing at least one subhalo placed in a bright pixel to ensure a perturbation is caused. Each network was then tested on an independent dataset and evaluated on classification accuracy, precision, and recall.

For a TP detection, a subhalo prediction needs to overlap with the ground truth subhalo mask

by at least 30%; for our image resolution and the size of the subhalo masks, this corresponds to an accuracy of 0.15". This is a mildly generous threshold while still providing a high degree of certainty that further studies would be observing the correct region of space. More rigorous investigations have the freedom to change the overlap threshold to increase the required accuracy for a TP prediction. The sources of FPs include extraneous subhalo predictions within a given mass bin and predictions in the absence of a TP. Since each image contains a consistent number of subhalos for a given mass bin, there are very few instances of FNs in the absence of a FP.

Results indicate that attention increases the U-Net's sensitivity to subhalo mass in the absence of noise. When a simulated image contains one subhalo, the U-Net with attention can detect subhalos one mass bin below the standard U-Net with better classification metrics for the higher mass bins as well. This remains true when each image contains one subhalo from each mass bin as overlapping subhalo perturbations result in much fewer FNs and FPs for the attention-based U-Net. These results are promising for observing more subhalos $\lesssim 10^8 M_\odot$ where baryonic activity becomes suppressed (Read et al., 2017; Brehmer et al., 2019) which is around where the high-mass WDM and ΛCDM models diverge.

Once noise is introduced, however, both networks perform similarly. The only notable improvement attention makes is increased precision for the $10^{8.5} M_\odot$ mass bin when the light source has an apparent magnitude less than 20. However, the accuracy and recall fall in this same range of parameters. The attention U-Net generally sees marginal improvements for medium to high-mass subhalos while being outperformed to an extent by the standard U-Net when classifying images with high apparent magnitude light sources. This may be due to the noise causing attention weight misalignment leading to FN predictions and more FPs for higher apparent magnitude light sources. The results show that attention would only be suitable for images from deep-sky surveys (e.g. Planck Collaboration et al., 2020) where the SNR is high even for fainter source galaxies.

4.1 The Networks

The U-Net architecture was chosen for its ability to perform semantic segmentation, providing fine-grained details about the physical boundaries of subhalo predictions. Analysis of the physical interpretation is then possible which is what the counting subhalos algorithm described in Section 2.3 accomplishes. While accuracy, precision and recall are excellent measures of network performance, other metrics could provide a more concise description. The F_β score is a weighted mean of the precision and recall measures meaning it is also robust when class imbalances are present (Christen et al., 2023). Depending on the priorities of the investigation, the β value can be altered such that precision or recall influences the score more heavily. An extension to the metrics would be to consider leakage between mass bins. This would give further context to the results and indicate weaknesses in the model that will be able to inform actions to reduce them, such as post-processing techniques or including additional features in the architecture (Bishop & Nasrabadi, 2007).

From the investigation with multiple subhalos, overlaps have been shown to both aid and hinder the detection of low-mass subhalos. In Figure 10, the overlaps resulted in the two smallest mass bins being missed while results in Sections 3.2 and 3.3 reveal that the overlap may benefit low-mass subhalo detection. The type of effect caused by the overlap is most likely determined by the difference between the subhalo masses where nearby large subhalos may completely erase the effect of smaller subhalos while two similar-sized subhalos will produce a combined noticeable effect that can be detected and decomposed into the component contributions. To properly investigate the effect, it would be ideal to run a separate test where each combination of subhalo masses is represented in the training dataset with varying degrees of overlap. Concerning separating the overlapping perturbations, a deeper U-Net architecture with more down and up-sampling could be developed which would be able to learn features on even smaller scales. Alternatively, other neural network architectures could be tested. Rettenberger et al. (2023) found that Mask R-CNNs outperform U-Net architectures with overlapping structures. Additionally, it uses instance segmentation rather than semantic segmentation meaning unique subhalo instances within a class would be automatically counted without the algorithm described in Section

2.3.

We present a stark lack of uncertainties in our data throughout this paper. The results we present have all been obtained after a single round of training and testing due to time constraints. The next logical step in evaluating the performance of our networks is to calculate the uncertainty in its outputs. It is unlikely that every time a network is trained or tested it will yield the same result. Therefore, the consistency of the network and the spread of results can be obtained from repeatedly testing and training the networks.

Attention has been shown to provide increased network capabilities and interpretability in the project. Because of how attention mechanisms focus on relevant parts of the image, the network more efficiently learns features which can compensate for smaller datasets (Schlemper et al., 2019). In this paper, only 10^4 images were used for training which was enough to demonstrate a significant advantage over the conventional U-Net architecture. An additional investigation could find out how few images are needed to obtain a usable result which could reduce the computational power required for future works. Furthermore, other attention mechanisms could be implemented to boost performance further. Squeeze-and-Excitation blocks (Hu et al., 2017) recalibrate relevant feature maps by modelling the relationships between them leading to significant performance improvements. The attention implemented by Qin et al. (2018) uses contextual information to adapt the size of the effective receptive field to the optimal scale for feature representation. Each of these adds little extra computational costs making them an efficient way to improve network performance.

4.2 The Data

All images were generated using the same parameters. Each image contained the same lens which was modelled as a singular isothermal sphere (SIS) - the same source light modelled as a Sérsic ellipse - and at least one subhalo modelled with a truncated NFW profile. We are confident in choosing to model the subhalos with a truncated NFW profile because it reflects the expectations of the mass profiles of dark matter structures that form via hierarchical accretion (eg. Okabe et al., 2013), though we recognise that these models are incredibly simplistic and do not represent expected real-life observations. Therefore, the next logical step to further explore the capabilities of our networks would be to vary the parameters of the lens. The parameters chosen result in a perfect Einstein ring which is unlikely to be present in observations. Varying the ellipticity and Einstein radius of the lens would produce images with a variety of lensed arcs rather than perfect circles. Even modelling the source light with a profile more complex than a single Sérsic profile or modelling the lens as a singular isothermal ellipse (Kormann et al., 1994) with the surface mass density described by

would better represent real lensed systems.

An additional factor in the image simulations is the resolution of the image. The U-Net's down-sampling required a pixel resolution that was divisible by two, though this could be increased to enable more precise network predictions. Doing this would improve the angular resolution set at 0.05" which will be possible with next generation telescopes. For example, the Extremely Large Telescope (ELT) contains the instrument HARMONI which boasts an angular resolution up to 0.004" (Thatte et al., 2022). Perturbations from low-mass subhalos would then be more visible allowing the mass function to be pushed down further.

One aspect of a strongly lensed system we have neglected in this paper is lens light. In reality, it is expected that there may be objects between the lens and the observer. The lens light may not cause additional perturbation to the lens but would obstruct the line of sight which could cause potential confusion in the network. Due to the distances involved in gravitational lensing line-of-sight (LOS), halos that are not associated with the main lens will cause additional perturbations in the lens (Çağan Şengül et al., 2020) which would also confuse the networks.

Due to the simplification and absence of modelling explained above, the results are heavily biased. We predict that with the addition of lens light, varied lens parameters, and more complex light profiles, 10^4 images will not be enough to adequately train the network. In the images expected from future deep sky surveys, there may also be cosmic rays or dead pixels which have not been simulated. Based

on this, we expect that the trained network would perform poorly if given a set of realistic data to test and the comparison between each network's performances would be different.

4.3 Hardware Limitations

We stored the images on Google Drive and executed the code on Google Colab for access to their GPUs. An NVIDIA L4 GPU was used due to its ability to handle large datasets and fast training capabilities. Despite this, datasets containing more than 10,000 images were overloading the RAM meaning the training potential was limited and datasets were not as comprehensive as we would have liked. Other studies on subhalo detection using machine learning use between $10^5 - 10^6$ training images containing thousands of images for each mass bin (Ostdiek et al., 2022) or a mixture of simulated and real strong lensing data (Hezaveh et al., 2017). Lacking the capability for this, we focused on exploring subhalo detection by omitting simulations with an absence of subhalos though this does introduce some bias as the networks learn to expect at least one subhalo in a given mass bin. Continuations of this work should utilise hardware with the ability to generate a complete dataset that will not introduce epistemic uncertainties. Alternatively, one could apply Bayesian statistics to the networks to calculate the epistemic uncertainties resulting from an incomplete dataset (Perreault Levasseur et al., 2017).

5 Conclusion

We present a comparison of the performance between a network with and without attention mechanisms implemented for dark matter subhalo detection in strongly lensed images. The addition of attention has improved the network sensitivity to lower mass subhalos, helped break down overlapping subhalos, and marginally improved sensitivity in the presence of low-level amounts of noise. However, it struggles to distinguish perturbations from heavy noise in images. More investigation into the attention gate is required to see how it performs when the images become more complex, such as when varying the lens parameters and source light profile.

Overall, attention proves to be an exciting technique that has the potential to vastly improve the detection of dark matter subhalos. This is incredibly useful for anticipating the tens of thousands of strong gravitational lens images expected from future deep sky surveys.

References

Abadi M., et al., 2016, arXiv e-prints, p. arXiv:1603.04467

Alexander S., Gleyzer S., Parul H., Reddy P., Toomey M. W., Usai E., Von Klar R., 2020a, arXiv e-prints, p. arXiv:2008.12731

Alexander S., Gleyzer S., McDonough E., Toomey M. W., Usai E., 2020b, ApJ, 893, 15

Banik N., Bovy J., Bertone G., Erkal D., de Boer T. J. L., 2021, J.Cosmology Astropart. Phys., 2021, 043

Birrer S., Amara A., 2018, Physics of the Dark Universe, 22, 189

Bishop C. M., Nasrabadi N. M., 2007, Journal of Electronic Imaging, 16, 049901

Bode P., Ostriker J. P., Turok N., 2001, ApJ, 556, 93

Bonaca A., Hogg D. W., Price-Whelan A. M., Conroy C., 2019, ApJ, 880, 38

Brehmer J., Mishra-Sharma S., Hermans J., Louppe G., Cranmer K., 2019, ApJ, 886, 49

Bridle J. S., 1989, in Proceedings of the 2nd International Conference on Neural Information Processing Systems. NIPS'89. MIT Press, Cambridge, MA, USA, p. 211–217

Cardone V. F., 2004, A&A, 415, 839

Christen P., Hand D. J., Kirielle N., 2023, ACM Comput. Surv., 56

Crocce M., et al., 2016, MNRAS, 455, 4301

Dalal N., Kochanek C. S., 2002, ApJ, 572, 25

Daylan T., Cyr-Racine F.-Y., Diaz Rivero A., Dvorkin C., Finkbeiner D. P., 2018, ApJ, 854, 141

Diaz Rivero A., Dvorkin C., 2020, Phys. Rev. D, 101, 023515

Diemand J., Kuhlen M., Madau P., 2007, ApJ, 667, 859

Dye S., Warren S. J., 2005, ApJ, 623, 31

Feldmann R., Spolyar D., 2015, MNRAS, 446, 1000

Fitts A., et al., 2017, MNRAS, 471, 3547

Gu R., et al., 2021, IEEE Transactions on Medical Imaging, 40, 699–711

Hezaveh Y., Dalal N., Holder G., Kuhlen M., Marrone D., Murray N., Vieira J., 2013, ApJ, 767, 9

Hezaveh Y. D., et al., 2016, ApJ, 823, 37

Hezaveh Y. D., Perreault Levasseur L., Marshall P. J., 2017, Nature, 548, 555

Hoshen J., Kopelman R., 1976, Phys. Rev. B, 14, 3438

Hu J., Shen L., Albanie S., Sun G., Wu E., 2017, arXiv e-prints, p. arXiv:1709.01507

Ioffe S., Szegedy C., 2015, arXiv e-prints, p. arXiv:1502.03167

Kingma D. P., Ba J., 2014, arXiv e-prints, p. arXiv:1412.6980

Koopmans L. V. E., 2005, MNRAS, 363, 1136

Kormann R., Schneider P., Bartelmann M., 1994, "A&A", 284, 285

Krist J. E., Hook R. N., Stoehr F., 2011, in Optical Modeling and Performance Predictions V. p. 81270J, doi:10.1117/12.892762

LSST Science Collaboration et al., 2009, arXiv e-prints, p. arXiv:0912.0201

Lindsay G. W., 2020, Frontiers in Computational Neuroscience, 14

Long J., Shelhamer E., Darrell T., 2014, arXiv e-prints, p. arXiv:1411.4038

Metcalf R. B., Amara A., 2012, MNRAS, 419, 3414

Narayan R., Bartelmann M., 1996a, arXiv e-prints, pp astro–ph/9606001

Narayan R., Bartelmann M., 1996b, arXiv e-prints, pp astro–ph/9606001

Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493

Nightingale J. W., et al., 2022, arXiv e-prints, p. arXiv:2209.10566

Okabe N., Smith G. P., Umetsu K., Takada M., Futamase T., 2013, ApJ, 769, L35

Oktay O., et al., 2018, arXiv e-prints, p. arXiv:1804.03999

Ostdiek B., Diaz Rivero A., Dvorkin C., 2022, ApJ, 927, 83

Pawase R. S., Courbin F., Faure C., Kokotanekova R., Meylan G., 2014, MNRAS, 439, 3392

Pearson J., Li N., Dye S., 2019, MNRAS, 488, 991

Pearson J., Maresca J., Li N., Dye S., 2021, MNRAS, 505, 4362

Perreault Levasseur L., Hezaveh Y. D., Wechsler R. H., 2017, ApJ, 850, L7

Planck Collaboration et al., 2016, A&A, 594, A13

Planck Collaboration et al., 2020, A&A, 641, A1

Qin Y., Kamnitsas K., Ancha S., Nanavati J., Cottrell G., Criminisi A., Nori A., 2018, arXiv e-prints, p. arXiv:1805.08403

Read J. I., Iorio G., Agertz O., Fraternali F., 2017, MNRAS, 467, 2019

Refregier A., Amara A., Kitching T. D., Rassat A., Scaramella R., Weller J., 2010, arXiv e-prints, p. arXiv:1001.0061

Rettenberger L., Münke F., Bruch R., Reischl M., 2023, Current Directions in Biomedical Engineering, 9, 335

Ritondale E., Vegetti S., Despali G., Auger M. W., Koopmans L. V. E., McKean J. P., 2019, MNRAS, 485, 2179

Ronneberger O., Fischer P., Brox T., 2015, arXiv e-prints, p. arXiv:1505.04597

Schlemper J., Oktay O., Schaap M., Heinrich M., Kainz B., Glocker B., Rueckert D., 2019, Medical Image Analysis, 53, 197

Schneider P., Ehlers J., Falco E. E., 1992, Gravitational Lenses, doi:10.1007/978-3-662-03758-4.

Springel V., et al., 2008, MNRAS, 391, 1685–1711

Thatte N. A., et al., 2022, in Ground-based and Airborne Instrumentation for Astronomy IX. p. 1218420, doi:10.1117/12.2628834

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., 2017, arXiv e-prints, p. arXiv:1706.03762

Vegetti S., Koopmans L. V. E., 2009, MNRAS, 392, 945

Vegetti S., Koopmans L. V. E., Bolton A., Treu T., Gavazzi R., 2010, MNRAS, 408, 1969

Vegetti S., Lagattuta D. J., McKean J. P., Auger M. W., Fassnacht C. D., Koopmans L. V. E., 2012, Nature, 481, 341

Vegetti S., Koopmans L. V. E., Auger M. W., Treu T., Bolton A. S., 2014, MNRAS, 442, 2017

Çağan Şengül A., Tsang A., Diaz Rivero A., Dvorkin C., Zhu H.-M., Seljak U., 2020, Phys. Rev.˜D, 102, 063502