# Population-level Dark Energy Constraints from Strong Gravitational Lensing using Simulation-Based Inference

**Sreevani Jarugula** [1] **Brian Nord** [1 2] **Abhijith Gandrakota** [1] **Aleksandra Ćiprijanović** [1 2]

## Abstract

In this work, we present a scalable approach for inferring the dark energy equation-of-state parameter ($w$) from a population of strong gravitational lens images using Simulation-Based Inference (SBI). Strong gravitational lensing offers crucial insights into cosmology, but traditional Monte Carlo methods for cosmological inference are computationally prohibitive and inadequate for processing the thousands of lenses anticipated from future cosmic surveys. New tools for inference, such as SBI using Neural Ratio Estimation (NRE), address this challenge effectively. By training a machine learning model on simulated data of strong lenses, we can learn the likelihood-to-evidence ratio for robust inference. Our scalable approach enables more constrained population-level inference of $w$ compared to individual lens analysis, constraining $w$ to within $1\sigma$. Our model can be used to provide cosmological constraints from forthcoming strong lens surveys, such as the 4MOST Strong Lensing Spectroscopic Legacy Survey (4SLSLS), which is expected to observe 10,000 strong lenses.

## 1. Introduction

Dark energy, which comprises approximately 70% of the energy density of the universe, plays a pivotal role in driving the accelerated expansion of the universe. Yet, the nature of dark energy remains a prominent puzzle in physics. Central to understanding its fundamental nature is the dark energy equation-of-state parameter $w = p/\rho$, the ratio between the dark energy pressure $p$ and the energy density $\rho$. High-precision constraints on $w$ are crucial for understanding the fate of the expansion of the universe (e.g., Caldwell et al., 2003; Sahni, 2002). Constraints on dark energy have been provided through observations of the Cosmic Microwave Background (CMB; Planck Collaboration, 2020), supernova (Riess et al., 2016), and Baryon Acoustic Oscillations (BAO; Alam et al., 2017) which are in broad agreement with the constant value of $w = -1$ according to the $\Lambda$ cold dark matter ($\Lambda$CDM) cosmological model (Escamilla et al., 2023). However, $w$ is degenerate with other cosmology parameters such as the Hubble constant $H_0$ and the dark matter density $\Omega_m$, necessitating careful statistical analysis combining multiple observational probes. Given the existing tensions in cosmology parameter estimates using probes from the early and late universe observations (e.g., Leizerovich et al., 2023), improving constraints on $w$ using independent cosmological probes like strong gravitational lensing is crucial.

The dark energy equation-of-state parameter can be constrained through the distance ratio from static galaxy-galaxy strong lens systems by combining lensing and stellar dynamical measurements. The constraints on $w$ have been obtained from a population of $\mathcal{O}(100)$ strong lens systems (Cao et al., 2015; Jie et al., 2016) using Markov Chain Monte Carlo (MCMC) with analytic likelihoods (Lewis & Bridle, 2002). However, traditional MCMC methods are limited by the need for accurate and efficient lens modeling with numerous parameters, making it computationally prohibitive to analyze large datasets. More efficient techniques to model strong lensing systems are required in the upcoming era of Legacy Survey of Space and Time (LSST; Ivezic et al., 2008), Euclid Wide Survey (Euclid Collaboration, 2022), and Roman Space Telescope High Latitude Wide Area Survey (Spergel et al., 2015) where $\mathcal{O}(10^5)$ strong lens systems will be discovered (Holloway et al., 2023). Spectroscopic information from surveys such as the 4MOST Strong Lensing Spectroscopic Legacy Survey (4SLSLS; Collett et al., 2023), expected to observe about 10,000 strong lenses, will provide better constraints on $w$ (Li et al., 2024). However, the analyses so far are limited by the complexity of lens modeling in calculating the analytical likelihood, highlighting the need for faster and scalable inference algorithms.

AI-based methods such as Simulation-Based Inference (SBI) using Neural Posterior Estimation (NPE) and Neural Ratio

[*]Equal contribution [1]Fermi National Accelerator Laboratory [2]Kavli Institute for Cosmological Physics & Department of Astronomy and Astrophysics, The University of Chicago. Correspondence to: Sreevani Jarugula <jarugula@fnal.gov>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Estimation (NRE) have emerged as powerful methods for posterior inference without explicitly calculating the likelihood (Cranmer et al., 2020). These methods have been successfully applied in various astrophysical and cosmology studies (Brehmer et al., 2019; Legin et al., 2021; Gerardi et al., 2021; Wagner-Carena et al., 2021; Zhang et al., 2022; Khullar et al., 2022; Poh et al., 2022; Mishra-Sharma & Cranmer, 2022; Wagner-Carena et al., 2023; Moser et al., 2024; Lemos et al., 2024). The NRE method involves training a neural network on simulations of data to estimate the likelihood-to-evidence ratio (Hermans et al., 2020) which can then be used to perform approximate Bayesian inference. This approach allows for efficient posterior inference without the need for explicit likelihood evaluations, making it particularly useful for complex simulations with intractable likelihoods. Moreover, NRE facilitates population-level inference, allowing for inference on parameters common across a population while marginalizing over variables included in the simulation but not of inferential interest (nuisance parameters). A key advantage of these SBI methods is their ability to amortize the computational cost of the inference procedure where after an upfront cost of simulation and neural network model training, efficient inference can be performed on a large number of observations.

This analysis presents the first application of NRE for population-level inference, constraining the dark energy equation-of-state parameter $w$ from strong gravitational lens images. We estimate the posterior distribution of $w$ from the likelihood-to-evidence ratio from NRE using MCMC sampling and analytical calculations. After introducing strong gravitational lensing and the simulation setup (Section 2), we delve into the background of NRE and details of posterior calculation methods (Section 3). Details on the training and test data used for the NRE model are then provided in Section 4. Section 5 focuses on the analysis and results. We evaluate the performance of the trained NRE model by comparing its predicted values of $w$ to the true values. Furthermore, we present the population-level inference results, showcasing the ability of NRE to constrain $w$ simultaneously from multiple lens systems. Finally, in Section 6, the key findings are summarized, accompanied by an outlook on potential future analyses.

## 2. Strong Gravitational Lensing

Strong gravitational lensing is a phenomenon in which the mass distribution of a galaxy (the lens) distorts the light from a background source, producing distorted and magnified images of the source in the observed image plane. This phenomenon arises due to the multiple solutions to the lens equation. In the case of a point-like background source, such as a quasi-stellar object (QSO) or a supernova, the lensing effect can lead to the formation of multiple images of the source. In the case of an extended background galaxy, lensing results in an arc-like structure in the image plane, which we refer to as galaxy-galaxy lensing. Future surveys are expected to discover thousands of galaxy-galaxy strong lenses making this increasingly relevant for high-redshift and cosmology studies. Notably, the strong lensing observations can constrain the dark energy equation-of-state parameter $w$, which plays a crucial role in our understanding of the expansion of the universe. The constraint on $w$ can be derived from the analysis of the distance ratio, defined as the ratio of the angular diameter distance between the lens and the source, and the distance between the source and the observer. In this section, we present some of the equations that are used in the simulator.

The lens equation relates the source position $\beta$ to the lensed source position $\theta$ through the deflection angle $\alpha(\theta)$ as

$$\beta = \theta - \alpha(\theta). \tag{1}$$

The deflection field is described by the projected surface mass density $\kappa(\theta) = \Sigma(\theta)/\Sigma_{cr}$. The critical lensing surface density $\Sigma_{cr}$ is given by

$$\Sigma_{cr} = \frac{1}{4\pi\, G}\frac{D_s}{D_{ls}D_l}, \tag{2}$$

where $D_l$, $D_s$, and $D_{ls}$ are the angular diameter distances between the observer and the lens, the observer and the source, and the lens and the source, respectively. In $w$CDM cosmology, the angular diameter distance depends on the astrophysical (local) parameters of redshift for the source and the lens and on cosmology (population-level) parameters — Hubble constant $H_0$, matter density parameter $\Omega_m$, dark energy density parameter $\Omega_{de} = 1 - \Omega_m$, and the dark energy equation-of-state parameter $w$:

$$D(z, H_0, \Omega_m, w) = \frac{1}{1+z}\frac{c}{H_0}\int_0^z \frac{dz'}{h(z', \Omega_m, w)}, \tag{3}$$

where the Hubble parameter is

$$h^2(z, \Omega_m, w) = \Omega_m(1+z)^3 + (1-\Omega_m)(1+z)^{3(1+w)}. \tag{4}$$

The Einstein radius is given by (Treu, 2010):

$$\theta_E = 4\pi\left(\frac{\sigma_v}{c}\right)^2\frac{D_{ls}}{D_s}. \tag{5}$$

Through observations of the Einstein ring from strong lensing and stellar velocity dispersion from dynamical studies, the cosmology $(\Omega_m, w)$ can be constrained through the distance ratio $\frac{D_{ls}}{D_s}$.

We model the mass of the lens using a Singular Isothermal Ellipsoid (SIE) profile (Kormann et al., 1994). The convergence of the lens is described by $\theta_E$, lens position $(x_l, y_l)$,

and the ellipticity of the lens ($l_{e1}, l_{e2}$). We model the light profile of the source using a Sersic profile (Sersic, 1968) which is described by the magnitude of the source ($m_s$) at the effective radius $R$, the source position ($x_s$, $y_s$), Sersic index $n$, and the source ellipticity ($s_{e1}, s_{e2}$).

## 3. Simulation-based Inference

Given a population of strong lens images, our goal is to infer the dark energy equation-of-state parameter $w$. In the context of population-level inference, we implement an approach akin to hierarchical Bayesian modeling, where we combine information from a sample of observations to obtain tighter constraints on the parameters that are common across the population. In Bayesian inference, the posterior is given by

$$p(w|x) = \frac{p(x|w)}{p(x)} p(w), \tag{6}$$

where $p(w)$ is the prior on $w$, $p(x|w)$ is the likelihood of the data $x$ given $w$, and $p(x)$ is the evidence. The likelihood depends on nuisance parameters ($\nu$) that we integrate over:

$$p(x|w) = \int p(x, \nu|w) d\nu, \tag{7}$$

where $p(x, \nu|w)$ is the combined likelihood of observed variables and nuisance parameters. In many scenarios in modern cosmology, the likelihood is intractable because of the large number of nuisance parameters, which require a high-dimensional integral. In this work, we employ NRE to obtain the posterior $p(w|x)$ (Cranmer et al., 2015; Baldi et al., 2016). With NRE, a classification neural network is trained to model the likelihood ratio between two hypotheses. Each hypothesis represents a probability distribution from which the samples $(x, w)$ are drawn. If one of the hypotheses is that the samples are drawn from $p(x|w)$ and the other is $p_{\rm ref}(x|w)$, the likelihood ratio $r(x|w) = \frac{p(x|w)}{p_{\rm ref}(x|w)}$. Using a Binary Cross Entropy Loss (BCE) function, the optimal classifier that differentiates the samples drawn from these two hypotheses is given by

$$d^*(x, w) = \frac{p(x|w)}{p(x|w) + p_{\rm ref}(x|w)}. \tag{8}$$

The ratio $r(x|w)$ is related to the classifier $d^*(x, w)$ as

$$r(x|w) = \frac{d^*(x, w)}{1 - d^*(x, w)}. \tag{9}$$

The classification neural network learns to map the $(x, w)$ samples to the class probabilities through its final layer. The layer preceding the final activation outputs the *logit*, which is the logarithm of the likelihood ratio $\log r(x|w)$.

In our analysis, we use a classifier to distinguish between the sample-parameter pairs: $(x, w) \sim p(x, w)$ drawn from the joint distribution with class label $y = 1$ and $(x, w) \sim p(x)p(w)$ drawn from the marginal distribution with class label $y = 0$. If we consider the two hypotheses to be the joint and the marginal distribution, the network learns the likelihood-to-evidence ratio $r(x|w) = \frac{p(x,w)}{p(x)p(w)} = \frac{p(x|w)}{p(x)}$ (Hermans et al., 2020).

The joint likelihood-to-evidence ratio from a population of lenses $\{x\}$ can be obtained by combining the likelihood ratio from individual lenses under the assumption that the observations are independent and identically distributed. This joint likelihood ratio can be written as

$$r(\{x\}|w) = \prod_i r(x_i|w) = \sum_i log\, r(x_i|w). \tag{10}$$

An important advantage of estimating the likelihood-to-evidence ratio is its ability to combine the information from a population of observations that have the same underlying parameter of interest. This provides more robust constraints on the population-level parameters (Brehmer et al., 2019; Zhang et al., 2022).

### 3.1. Neural Ratio Estimation Model Architecture

To learn the likelihood ratios, we use a ResNet architecture (He et al., 2015). The model architecture is shown in Table 2 in Appendix A.

We implement the model in *TensorFlow* (Abadi et al., 2016). The inputs to the model are the images and the corresponding parameter of interest ($w$) that generated the images. As a pre-processing step, we normalize the images in the training data. The samples from the marginal distribution are generated in each batch, where $w$ is randomly selected for each image. Performing the randomization at the batch level gives the network more random samples to differentiate between the two distributions. Batch normalization is applied to the data followed by a convolutional layer with batch normalization and *ReLU* activation function, serving as the entry block. Residual connections are introduced to preserve essential information across subsequent convolutional blocks of increasing depth and complexity. Each convolutional block consists of two convolutional layers with batch normalization and *ReLU* activations, followed by max-pooling for spatial downsampling. We perform global average pooling to reduce the size of the images. The residual from the previous convolution block is added to the output from the current convolution layer and this is the input to the next convolution block. The architecture further includes a convolutional layer and global average pooling to output a one-dimensional tensor. The $w$ values are embedded with a dense layer and concatenated with the image representations for joint analysis. Regularization techniques,

such as dropout and L2 regularization, are applied to the dense layers to prevent overfitting. The architecture outputs *logits* which is $\log\left(\frac{p(x|w)}{p(x)}\right)$. We use the output from the trained model to estimate the posterior $p(w|x)$.

We train the network by minimizing the BCE loss function by dynamically adjusting the learning rate when the validation loss is plateaued by a decay factor of 0.1 starting from $1e^{-2}$ to $1e^{-6}$, if the validation loss does not improve over five epochs. The model is trained over 100 epochs with a batch size of 1024 using `Adam` optimizer. We include an option for early stopping if the validation loss does not improve over 20 epochs. The model is trained on NVIDIA A100 GPU for 71 epochs with a typical training time of 25 minutes.

### 3.2. Population-level Posterior Inference

Given the likelihood-to-evidence ratio, we calculate the posterior using two methods. The first is sampling from the joint likelihood ratio using MCMC, and the second is calculating the posterior analytically for each $w$ value drawn from the prior without the need for sampling. Using both MCMC and analytical calculation of the posterior from the likelihood-to-evidence ratio serves as complementary methods to validate the accuracy and consistency of the results.

#### 3.2.1. Posterior from MCMC

We infer the posterior distributions of $w$ by carrying out MCMC sampling using the Metropolis-Hastings algorithm (Hastings, 1970) from the log likelihood-to-evidence ratio obtained from NRE. We implemented MCMC using the `emcee` package (Foreman-Mackey et al., 2013). The prior range to draw $w$ values is the same as the training prior. We initiate five walkers randomly around a starting value of $w = -1.0$. We implement a series of two warm-up periods, with 100 and 500 steps respectively, where the walkers converge to the target distribution. The position of the walkers is reset to the value that was converged on after each warm-up. After the two warm-ups, we run the sampling for 1000 steps. The log likelihood-to-evidence ratio $\log r(\{x\}|w)$ for a population of images from Equation 10 is estimated for each $w$ value drawn from the prior using the trained NRE model followed by summation of the ratios.

#### 3.2.2. Analytical Posterior

The population-level posterior can be analytically calculated from the individual likelihoods when there are few enough latent parameters — i.e., when the evidence contains a tractable integral.

From Equations 6 and 10, the posterior $p(w|\{x\})$ from a population of observations $\{x\}$ is given by

$$p(w|\{x\}) = \frac{p(w)\ \prod_i r(x_i|w)}{\int dw'\ p(w')\ \prod_i r(x_i|w')}, \\ = p(w)\ \left(\int dw'\ p(w')\ \prod_i \frac{r(x_i|w')}{r(x_i|w)}\right)^{-1}. \tag{11}$$

We generate a list of $w$ values in the prior range and analytically calculate the above equation for a population of strong lens images at each $w$.

## 4. Data

We generate images of strong lenses using `deeplenstronomy` (Morgan et al., 2021) which is built on `lenstronomy` (Birrer et al., 2015; Birrer & Amara, 2018). The main components of simulating galaxy-galaxy strong lenses are the lens's mass profile, the source's light profile, and the angular distance diameters between the observer, source, and lens.

We assume that the astrophysics parameters are independent of the cosmology and also independent of each other for simplicity of modeling. We also assume that the lens light is perfectly subtracted from the images. The lens light is often subtracted before lens modeling and there has been work done on automating this process (Hezaveh et al., 2017). We set the lens and source positions to $(0, 0)$, lens redshift $z_l = 0.1$, source redshift $z_s = 2.0$, stellar velocity dispersion = 200 $km/s$, $H_0 = 70\ (km/s)/Mpc$ and $\Omega_m = 0.3$.

In our model, we have:

- observable $x$: strong lens image;
- parameter of interest $w$: dark energy equation-of-state parameter;
- nuisance parameters $\nu$: lens ellipticity ($l_{e1}$, $l_{e2}$), magnitude of the source $m_s$, source effective radius $R$, source Sersic index $n$, and source ellipticity ($s_{e1}$, $s_{e2}$). While these nuisance parameters are incorporated into the simulation process to generate the images, our objective does not involve inferring their values.

The strong lens images are generated by sampling from the uniform prior distribution of these parameters as shown in Table 1. To generate images representative of the survey conditions, we use DES Data Release 1 (DES Collaboration, 2018) with a pixel scale of 0.263 arcsec/pixel, CCD detector gain of 6.083 $e^-$/count, and read noise of 7.0 $e^-$. We simulate $g$-band images with a magnitude zero point of 30.0. We include the sky brightness, seeing, and the effective number of exposures based on the survey's empirical values. The

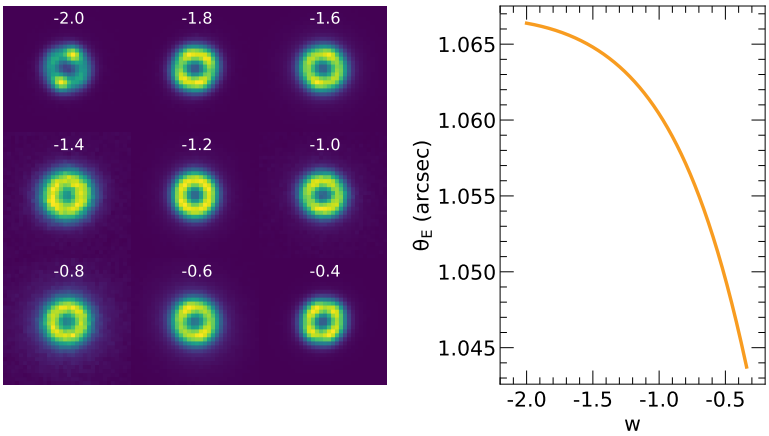


*Figure 1.* Left: Sample of training data with the $w$ value used to generate the image. Right: the Einstein radius $\theta_E$ as a function of the dark energy equation-of-state parameter $w$. The variation in $\theta_E$ is larger at higher values of $w$ than at the lower values.

images have a size of 32 x 32 pixels. The training and validation dataset includes 640,000 images and 160,000 images, respectively. To assess the model's performance, we use a test dataset of 2,000 images generated from the same prior range as the training dataset. For the population-level inference, we create three distinct datasets, each consisting of 3,000 images with a common $w$ value across all images within that dataset. Specifically, these three datasets are generated with $w_{\rm true}$ fixed to $-1.2$, $-1.0$, and $-0.8$, respectively. The purpose of these datasets is to assess the capacity of the model to constrain the common value of $w$ simultaneously from multiple observations within a population. A sample of the training data is shown in Figure 1 (left). In Figure 1, we show the correlation between the Einstein radius $\theta_E$ and $w$ in the training data: the variation in $\theta_E$ is greater at high $w$ values compared to low $w$. The code and the dataset used in this paper can be found in our github repository [1].

## 5. Results

With the trained NRE model, we infer the cosmological parameter $w$ on a dataset of 2,000 individual strong lenses generated from the same distribution as the training data. We also perform population-level inference on a dataset of 3,000 strong lens images that share a common $w$ value. This analysis is done to assess the improvement in constraining $w$ by combining information from multiple observations within the population.

We assess the performance of the trained NRE model using the Receiver Operating Characteristic (ROC) curve for different thresholds as shown in Figure 2 (left). The area under the ROC curve (AUC) is $\sim$ 0.92, which indicates that the model has high discriminating power. We also check the model's stability by training it on the same training dataset but with three different weight initializations—set by three different random seeds. The models from the three seeds have almost identical ROC curves indicating that the model training converges to similar optimal classifiers.

We calculate the posteriors for the 2,000 images using the analytical method described in Section 3.2.2. The analytical posterior approach is computationally less expensive for individual images, especially for large datasets, compared to the MCMC sampling method outlined in Section 3.2.1. We present the mean equation-of-state parameter $w_{\rm pred}$ and associated $1\sigma$ uncertainties obtained from the posterior of each image in Figure 2 (middle, top). The figure indicates that $w_{\rm true}$ most often lies within the predicted $1\sigma$ uncertainty for most data points. However, for observations with $w_{\rm pred} \lesssim -1.4$, the posteriors are less constrained compared to those with $w_{\rm pred} \gtrsim -1.4$. We suggest that this is due to the smaller variation in $\theta_E$ at low $w_{\rm true}$ than at high $w_{\rm true}$ as seen in Figure 1, which limits the model's capacity to predict features as effectively as it does for high $w_{\rm true}$ values. Figure 2 (middle, bottom) illustrates the bias between the true and predicted values relative to the true value, with error bars scaled by the true value accordingly. The low $w_{\rm true}$ values exhibit more bias compared to high $w_{\rm true}$ consistent with the explanation from the model performance in the low $w_{\rm true}$ range. In future studies, we will examine the training set in more detail and attempt to design priors that mitigate biases in predictions.

Figure 2 (right) presents the posterior coverage for three different random seeds for the network weight initialization. The posterior coverage is a measure of how well-calibrated the predicted uncertainties of the posteriors are. In a well-calibrated model, the fraction of observations within a pos-

[1] https://github.com/deepskies/DeepSLIDE/

*Table 1.* Parameter distributions used to generate training and test sets. For the population-level inference, we generate datasets with a fixed $w$ value that is common across all the images. Uniform distributions are denoted by $\mathcal{U}$.

| Parameter | Name | Training Priors | Test Priors and Fixed values |
|---|---|---|---|
| | Cosmology | | |
| $w$ | Dark Energy Equation of State | $\mathcal{U}(-2.0, -0.34)$ | $\mathcal{U}(-2.0, -0.34)$, $-1.2$, $-1.0$, $-0.8$ |
| $H_0$ [km/s/Mpc] | Hubble constant | 70.0 | 70.0 |
| $\Omega_{de}$ | Dark Matter density | 0.7 | 0.7 |
| $\Omega_m$ | Dark Matter density | 0.3 | 0.3 |
| | Lens Parameters | | |
| $l_{e_1}$ | lens ellipticity | $\mathcal{U}(-0.1, 0.1)$ | $\mathcal{U}(-0.1, 0.1)$ |
| $l_{e_2}$ | lens ellipticity | $\mathcal{U}(-0.1, 0.1)$ | $\mathcal{U}(-0.1, 0.1)$ |
| | Source Light Parameters | | |
| $m_s$ | magnitude | $\mathcal{U}(19, 24)$ | $\mathcal{U}(19, 24)$ |
| $R$ [arcsec] | half-light radius | $\mathcal{U}(0.1, 3.0)$ | $\mathcal{U}(0.1, 3.0)$ |
| $n$ | sersic index | $\mathcal{U}(0.5, 8.0)$ | $\mathcal{U}(0.5, 8.0)$ |
| $s_{e_1}$ | source ellipticity | $\mathcal{U}(-0.1, 0.1)$ | $\mathcal{U}(-0.1, 0.1)$ |
| $s_{e_2}$ | source ellipticity | $\mathcal{U}(-0.1, 0.1)$ | $\mathcal{U}(-0.1, 0.1)$ |

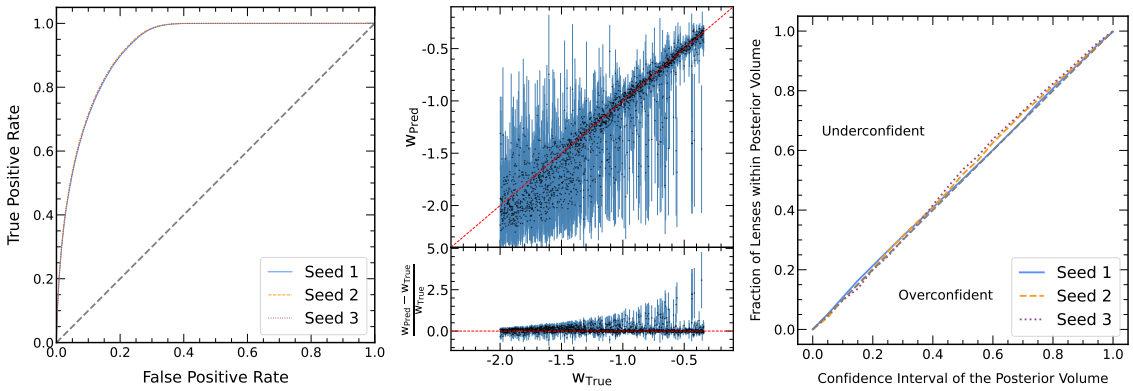


*Figure 2.* Left: The model performance on 2,000 test images is shown by the Receiver Operating Characteristic (ROC) curve for three models with different seed initialization. Middle Top: The parity plot showing true Vs predicted $w$ values with the mean and $1\sigma$ error obtained from the analytical posterior calculation (Section 3.2.2). Middle Bottom: The bias plot with the scaled error. The expected value is shown by the red dashed line. Right: The posterior coverage plot for different weight initialization demonstrating that the model is well-calibrated.

terior volume should match the confidence interval of that volume. For example, a posterior with 68% ($1\sigma$ equivalent) confidence interval should contain the true value within that confidence interval 68%of the time. Figure 2 (right) shows that the model uncertainty is well-calibrated for all random seeds initialization — i.e., all curves follow the diagonal dashed line.

The analytical posteriors for five of the 2,000 images from Figure 2 are shown in Figure 3. The posteriors converge to a similar distribution for different random seeds, indicating the model's stability.

For the population-level analysis, we estimate the posteriors from the joint likelihood-to-evidence ratio using two complementary methods: MCMC sampling and analytical calculation. While MCMC sampling is widely used and particularly efficient in high-dimensional parameter space, it can be computationally expensive. In contrast, the analytical calculation of the posterior is a computationally efficient alternative that is suitable in low dimensions. Comparing the posteriors obtained from both methods will validate the accuracy and consistency of the population-level inference.

We perform population-level inference analysis on three datasets with $w = -0.8, -1.0, -1.2$ as described in Section 4. Figure 4 shows the MCMC posterior densities (top panel) and the analytical posterior (bottom panel) for the three $w$ values, each inferred from the joint likelihood-to-evidence ratio of 500, 1000, 2000, and 3000 strong lens observations. The posterior is better constrained around the true value for the inference performed on the larger population.

Figure 5 presents the mean and $1\sigma$ error of $w$ obtained from MCMC and analytical posterior using the joint likelihood-to-evidence ratio of 5, 100, 500, 1000, 2000, and 3000 images. Consistent with the posterior distributions in Figure 4, the posterior widths of $w$ decrease with increasing population size. This behavior is observed for both the MCMC and analytical approaches to estimating the posteriors. This result highlights the potential gain in precision achievable by combining information from a larger population of observations.

## 6. Conclusions and Outlook

This study presents the first exploration of the capability of Neural Ratio Estimation (NRE) to constrain the dark energy equation-of-state parameter $w$ from a population of strong gravitational lenses.

We use NRE to train a neural network on simulated strong lens images and the corresponding $w$ to estimate the likelihood-to-evidence ratio $\frac{p(x|w)}{p(x)}$. We then use this ratio to compute the posterior $p(w|x)$ using two methods—MCMC and analytical calculation. The main results and conclusions from our analysis are:

1. We find that the likelihood-to-evidence ratio classifier has high discriminatory power with an AUC $\sim 0.92$. We also find that it is stable to changes in the neural network weight initialization.

2. From the analytical posterior analysis of 2,000 test images, we find that the true value lies within $1\sigma$ of the posterior for most of the observations. However, the model's performance is less efficient in constraining low values of $w$ compared to higher values, as the variation in the observable Einstein radius is less pronounced at low $w$, limiting the model's ability to learn in that regime. We aim to address this in future work by updating the training distribution to account for the difference in model performance.

3. From the posterior coverage plots, we find that the model is well-calibrated irrespective of the weight initialization.

4. We perform population-level analysis by estimating the posterior using the joint likelihood-to-evidence ratio. Comparing the joint posterior for $w$ values of -0.8, -1.0, and -1.2 obtained through both MCMC and analytical calculation, we observe that the posterior width decreases with an increasing number of observations in the inference population.

From our analysis, we demonstrate that the dark energy equation-of-state parameter is better constrained from a population of strong lenses than from individual images. In this proof-of-concept analysis, we use a simplified dataset designed with sufficient variation for training: fixed lens redshift, source redshift, and stellar velocity dispersion; and varying the lens ellipticity, source magnitude, half-light radius, Sersic index, and source ellipticity. Improvements upon our study could include inference from more complex and realistic images representing different configurations of sources and lenses, testing observational aspects of other surveys, and performing a joint inference of $w$ and $\Omega_m$. With upcoming surveys anticipated to discover thousands of strong lenses, there is a significant opportunity for scientific discovery. To fully utilize the data from these surveys, we must harness scalable population-level inference methods, such as the one demonstrated in this analysis.

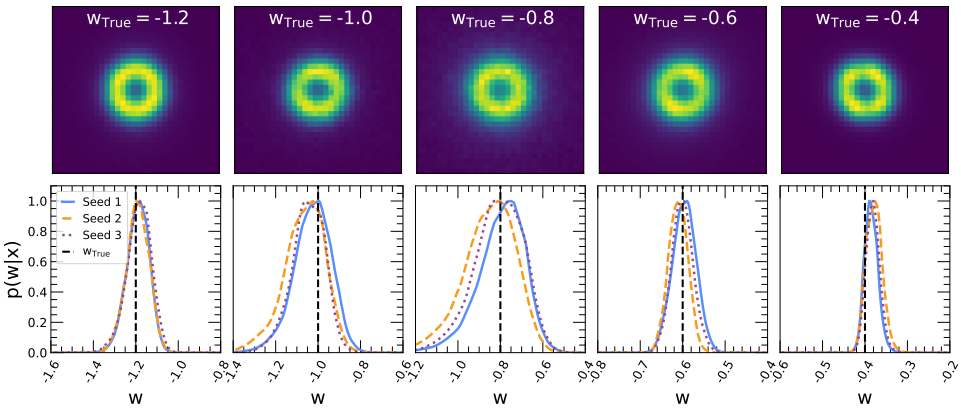


*Figure 3.* Strong lens images (top) and the associated posterior predictions from the analytic method (bottom) with three different random seeds in each case for the neural network weight initialization. The black dashed line shows the true $w$ for each inference.

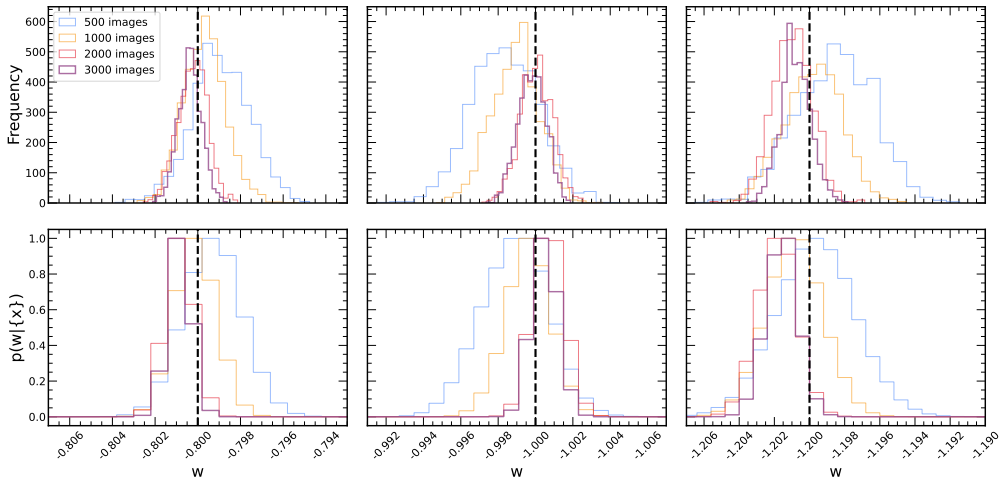


*Figure 4.* The posterior inference $p(w|\{x\})$ from the joint population analysis of 500, 1000, 2000, and 3000 strong lens images is shown. The posteriors for $w = -0.8, -1.0$, and $-1.2$ is shown from left to right. Top: The posterior obtained from MCMC sampling (Section 3.2.1). Bottom: The posterior obtained from analytical calculations (Section 3.2.2). In both the methods, the posterior is more constrained as the number of observations in the population increase.

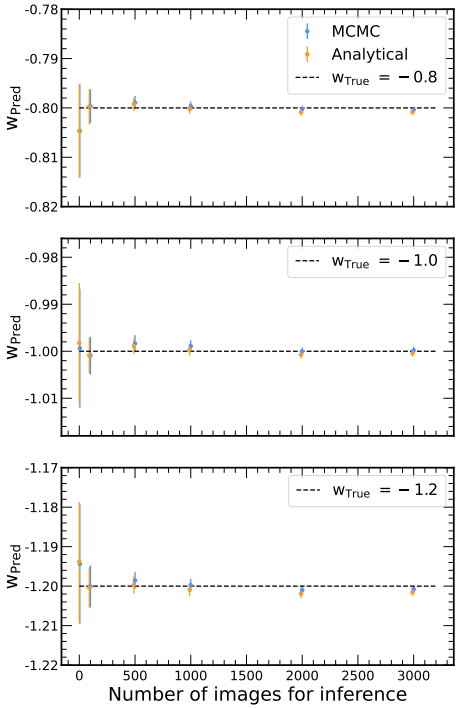


*Figure 5.* Population-level MCMC predictions of the dark energy equation-of-state parameter $w_{\rm pred}$ as a function of the population size (5, 100, 500, 1000, 2000, and 3000 strong lens images) for three different values of $w_{\rm true}$ (-0.8, -1.0, -1.2). The posterior converges to the true value as the number of observations in the population increases.

## References


Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv e-prints*, art. arXiv:1603.04467, March 2016. doi: 10.48550/arXiv.1603.04467.

Alam, S., Ata, M., Bailey, S., Beutler, F., Bizyaev, D., Blazek, J. A., Bolton, A. S., Brownstein, J. R., Burden, A., Chuang, C.-H., Comparat, J., Cuesta, A. J., Dawson, K. S., Eisenstein, D. J., Escoffier, S., Gil-Marín, H., Grieb, J. N., Hand, N., Ho, S., Kinemuchi, K., Kirkby, D., Kitaura, F., Malanushenko, E., Malanushenko, V., Maraston, C., McBride, C. K., Nichol, R. C., Olmstead, M. D., Oravetz, D., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Pellejero-Ibanez, M., Percival, W. J., Petitjean, P., Prada, F., Price-Whelan, A. M., Reid, B. A., Rodríguez-Torres, S. A., Roe, N. A., Ross, A. J., Ross, N. P., Rossi, G., Rubiño-Martín, J. A., Saito, S., Salazar-Albornoz, S., Samushia, L., Sánchez, A. G., Satpathy, S., Schlegel, D. J., Schneider, D. P., Scóccola, C. G., Seo, H.-J., Sheldon, E. S., Simmons, A., Slosar, A., Strauss, M. A., Swanson, M. E. C., Thomas, D., Tinker, J. L., Tojeiro, R., Magaña, M. V., Vazquez, J. A., Verde, L., Wake, D. A., Wang, Y., Weinberg, D. H., White, M., Wood-Vasey, W. M., Yèche, C., Zehavi, I., Zhai, Z., and Zhao, G.-B. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. MNRAS, 470(3): 2617–2652, September 2017. doi: 10.1093/mnras/stx721.

Baldi, P., Cranmer, K., Faucett, T., Sadowski, P., and Whiteson, D. Parameterized neural networks for high-energy physics. *European Physical Journal C*, 76(5):235, May 2016. doi: 10.1140/epjc/s10052-016-4099-4.

Birrer, S. and Amara, A. lenstronomy: Multi-purpose gravitational lens modelling software package. *Physics of the Dark Universe*, 22:189–201, December 2018. doi: 10.1016/j.dark.2018.11.002.

Birrer, S., Amara, A., and Refregier, A. Gravitational Lens Modeling with Basis Sets. ApJ, 813(2):102, November 2015. doi: 10.1088/0004-637X/813/2/102.

Brehmer, J., Mishra-Sharma, S., Hermans, J., Louppe, G., and Cranmer, K. Mining for Dark Matter Substructure: Inferring Subhalo Population Properties from Strong Lenses with Machine Learning. ApJ, 886(1):49, November 2019. doi: 10.3847/1538-4357/ab4c41.

Caldwell, R. R., Kamionkowski, M., and Weinberg, N. N. Phantom Energy: Dark Energy with w<-1 Causes a Cosmic Doomsday. Phys. Rev. Lett., 91(7):071301, August 2003. doi: 10.1103/PhysRevLett.91.071301.

Cao, S., Biesiada, M., Gavazzi, R., Piórkowska, A., and Zhu, Z.-H. Cosmology with Strong-lensing Systems. ApJ, 806 (2):185, June 2015. doi: 10.1088/0004-637X/806/2/185.

Collett, T. E., Sonnenfeld, A., Frohmaier, C., Glazebrook, K., Sluse, D., Motta, V., Verma, A., Anguita, T., Koopmans, L., Tortora, C., Courbin, F., Cabanac, R., Frye, B., Smith, G. P., Diego, J. M., Alteiri, B., Lopez, S., Fassnacht, C., Cooray, A., Goobar, A., Ryczanowski, D., Serjeant, S., Richard, J., Treu, T., Moustakas, L., Li, R., Jacobs, C., Lemon, C., Marchetti, L., Hartley, P., Jullo, E., Lee, C. H., Birrer, S., Fritz, A., Nightingale, J., Napolitano, N., Plazas, A. A., Kruk, S., Spiniello, C., Grillo, C., Suyu, S., Shajib, A., Vernardos, G., Dye, S., Daylan, T., Newman, J., and Schuldt, S. The 4MOST Strong Lensing Spectroscopic Legacy Survey (4SLSLS). *The Messenger*, 190:49–52, March 2023. doi: 10.18727/0722-6691/5313.

Cranmer, K., Pavez, J., and Louppe, G. Approximating Likelihood Ratios with Calibrated Discriminative Classifiers. *arXiv e-prints*, art. arXiv:1506.02169, June 2015. doi: 10.48550/arXiv.1506.02169.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Science*, 117(48):30055–30062, December 2020. doi: 10.1073/pnas.1912789117.

DES Collaboration. The Dark Energy Survey: Data Release 1. ApJS, 239(2):18, December 2018. doi: 10.3847/1538-4365/aae9f0.

Escamilla, L. A., Giarè, W., Di Valentino, E., Nunes, R. C., and Vagnozzi, S. The state of the dark energy equation of state circa 2023. *arXiv e-prints*, art. arXiv:2307.14802, July 2023. doi: 10.48550/arXiv.2307.14802.

Euclid Collaboration. Euclid preparation. I. The Euclid Wide Survey. A&A, 662:A112, June 2022. doi: 10.1051/0004-6361/202141938.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. emcee: The MCMC Hammer. PASP, 125(925): 306, March 2013. doi: 10.1086/670067.

Gerardi, F., Feeney, S. M., and Alsing, J. Unbiased likelihood-free inference of the Hubble constant from light standard sirens. Phys. Rev. D, 104(8):083531, October 2021. doi: 10.1103/PhysRevD.104.083531.

Hastings, W. K. Monte Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika*, 57 (1):97–109, April 1970. doi: 10.1093/biomet/57.1.97.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *arXiv e-prints*, art. arXiv:1512.03385, December 2015. doi: 10.48550/arXiv.1512.03385.

Hermans, J., Begy, V., and Louppe, G. Likelihood-free mcmc with amortized approximate ratio estimators, 2020.

Hezaveh, Y. D., Perreault Levasseur, L., and Marshall, P. J. Fast automated analysis of strong gravitational lenses with convolutional neural networks. Nature, 548(7669): 555–557, August 2017. doi: 10.1038/nature23463.

Holloway, P., Verma, A., Marshall, P. J., More, A., and Tecza, M. On the detectability of strong lensing in near-infrared surveys. MNRAS, 525(2):2341–2354, October 2023. doi: 10.1093/mnras/stad2371.

Ivezic, Z., Axelrod, T., Brandt, W. N., Burke, D. L., Claver, C. F., Connolly, A., Cook, K. H., Gee, P., Gilmore, D. K., Jacoby, S. H., Jones, R. L., Kahn, S. M., Kantor, J. P., Krabbendam, V. V., Lupton, R. H., Monet, D. G., Pinto, P. A., Saha, A., Schalk, T. L., Schneider, D. P., Strauss, M. A., Stubbs, C. W., Sweeney, D., Szalay, A., Thaler, J. J., Tyson, J. A., and LSST Collaboration. Large Synoptic Survey Telescope: From Science Drivers To Reference Design. *Serbian Astronomical Journal*, 176:1–13, June 2008. doi: 10.2298/SAJ0876001I.

Jie, A., Bao-Rong, C., and Li-Xin, X. Cosmic Constraints to the wCDM Model from Strong Gravitational Lensing. *Chinese Physics Letters*, 33(7):079801, July 2016. doi: 10.1088/0256-307X/33/7/079801.

Khullar, G., Nord, B., Ćiprijanović, A., Poh, J., and Xu, F. DIGS: deep inference of galaxy spectra with neural posterior estimation. *Machine Learning: Science and Technology*, 3(4):04LT04, December 2022. doi: 10.1088/2632-2153/ac98f4.

Kormann, R., Schneider, P., and Bartelmann, M. Isothermal elliptical gravitational lens models. A&A, 284:285–299, April 1994.

Legin, R., Hezaveh, Y., Perreault Levasseur, L., and Wandelt, B. Simulation-Based Inference of Strong Gravitational Lensing Parameters. In *Machine Learning and the Physical Sciences Workshop*, pp. 95, January 2021. doi: 10.48550/arXiv.2112.05278.

Leizerovich, M., Landau, S. J., and Scóccola, C. G. Tensions in cosmology: a discussion of statistical tools to determine inconsistencies. *arXiv e-prints*, art. arXiv:2312.08542, December 2023. doi: 10.48550/arXiv.2312.08542.

Lemos, P., Parker, L., Hahn, C., Ho, S., Eickenberg, M., Hou, J., Massara, E., Modi, C., Dizgah, A. M., Blancard, B. R.-S., Spergel, D., and SimBIG Collaboration. Field-level simulation-based inference of galaxy clustering with convolutional neural networks. Phys. Rev. D, 109(8):083536, April 2024. doi: 10.1103/PhysRevD.109.083536.

Lewis, A. and Bridle, S. Cosmological parameters from CMB and other data: A Monte Carlo approach. *Phys. Rev. D*, 66:103511, 2002. doi: 10.1103/PhysRevD.66.103511.

Li, T., Collett, T. E., Krawczyk, C. M., and Enzi, W. Cosmology from large populations of galaxy-galaxy strong gravitational lenses. MNRAS, 527(3):5311–5323, January 2024. doi: 10.1093/mnras/stad3514.

Mishra-Sharma, S. and Cranmer, K. Neural simulation-based inference approach for characterizing the Galactic Center $\gamma$ -ray excess. Phys. Rev. D, 105(6):063017, March 2022. doi: 10.1103/PhysRevD.105.063017.

Morgan, R., Nord, B., Birrer, S., Lin, J., and Poh, J. deeplenstronomy: A dataset simulation package for strong gravitational lensing. *The Journal of Open Source Software*, 6 (58):2854, February 2021. doi: 10.21105/joss.02854.

Moser, B., Kacprzak, T., Fischbacher, S., Refregier, A., Grimm, D., and Tortorelli, L. Simulation-based inference of deep fields: galaxy population model and redshift distributions. J. Cosmology Astropart. Phys., 2024(5): 049, May 2024. doi: 10.1088/1475-7516/2024/05/049.

Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. A&A, 641:A6, September 2020. doi: 10.1051/0004-6361/201833910.

Poh, J., Samudre, A., Ćiprijanović, A., Nord, B., Khullar, G., Tanoglidis, D., and Frieman, J. A. Strong Lensing Parameter Estimation on Ground-Based Imaging Data Using Simulation-Based Inference. *arXiv e-prints*, art. arXiv:2211.05836, November 2022. doi: 10.48550/arXiv.2211.05836.

Riess, A. G., Macri, L. M., Hoffmann, S. L., Scolnic, D., Casertano, S., Filippenko, A. V., Tucker, B. E., Reid, M. J., Jones, D. O., Silverman, J. M., Chornock, R., Challis, P., Yuan, W., Brown, P. J., and Foley, R. J. A 2.4% Determination of the Local Value of the Hubble Constant. ApJ, 826(1):56, July 2016. doi: 10.3847/0004-637X/826/1/56.

Sahni, V. The cosmological constant problem and quintessence. *Classical and Quantum Gravity*, 19(13): 3435–3448, July 2002. doi: 10.1088/0264-9381/19/13/304.

Sersic, J. L. *Atlas de Galaxias Australes*. 1968.

Spergel, D., Gehrels, N., Baltay, C., Bennett, D., Breckinridge, J., Donahue, M., Dressler, A., Gaudi, B. S., Greene, T., Guyon, O., Hirata, C., Kalirai, J., Kasdin, N. J., Macintosh, B., Moos, W., Perlmutter, S., Postman, M., Rauscher, B., Rhodes, J., Wang, Y., Weinberg, D., Benford, D., Hudson, M., Jeong, W. S., Mellier, Y., Traub, W., Yamada, T., Capak, P., Colbert, J., Masters, D., Penny, M., Savransky, D., Stern, D., Zimmerman, N., Barry, R., Bartusek, L., Carpenter, K., Cheng, E., Content, D., Dekens, F., Demers, R., Grady, K., Jackson, C., Kuan, G., Kruk, J., Melton, M., Nemati, B., Parvin, B., Poberezhskiy, I., Peddie, C., Ruffa, J., Wallace, J. K., Whipple, A., Wollack, E., and Zhao, F. Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report. *arXiv e-prints*, art. arXiv:1503.03757, March 2015. doi: 10.48550/arXiv.1503.03757.

Treu, T. Strong Lensing by Galaxies. ARA&A, 48:87–125, September 2010. doi: 10.1146/annurev-astro-081309-130924.

Wagner-Carena, S., Park, J. W., Birrer, S., Marshall, P. J., Roodman, A., Wechsler, R. H., and LSST Dark Energy Science Collaboration. Hierarchical Inference with Bayesian Neural Networks: An Application to Strong Gravitational Lensing. ApJ, 909(2):187, March 2021. doi: 10.3847/1538-4357/abdf59.

Wagner-Carena, S., Aalbers, J., Birrer, S., Nadler, E. O., Darragh-Ford, E., Marshall, P. J., and Wechsler, R. H. From Images to Dark Matter: End-to-end Inference of Substructure from Hundreds of Strong Gravitational Lenses. ApJ, 942(2):75, January 2023. doi: 10.3847/1538-4357/aca525.

Zhang, G., Mishra-Sharma, S., and Dvorkin, C. Inferring subhalo effective density slopes from strong lensing observations with neural likelihood-ratio estimation. MNRAS, 517(3):4317–4326, December 2022. doi: 10.1093/mnras/stac3014.

## A. NRE Network Architecture

The architecture for NRE is shown in Table 2.

| Layer Type | Output Shape | No: of Parameters |
|---|---|---|
| Input Images $x$ | (Batch, 32, 32, 1) | 0 |
| 2D Convolution | (Batch, 16, 16, 8) | 80 $\leftarrow$ |
| 2D Convolution | (Batch, 16, 16, 16) | 1168 |
| 2D Convolution | (Batch, 16, 16, 16) | 2320 |
| 2D MaxPooling | (Batch, 8, 8, 16) | 0 |
| 2D Convolution | (Batch, 8, 8, 16) | 144 |
| Add | (Batch, 8, 8, 16) | 0 $\leftarrow$ |
| 2D Convolution | (Batch, 8, 8, 32) | 4640 |
| 2D Convolution | (Batch, 8, 8, 32) | 9248 |
| 2D MaxPooling | (Batch, 4, 4, 32) | 0 |
| 2D Convolution | (Batch, 4, 4, 32) | 544 |
| Add | (Batch, 4, 4, 32) | 0 $\leftarrow$ |
| 2D Convolution | (Batch, 4, 4, 45) | 13005 |
| 2D Convolution | (Batch, 4, 4, 45) | 18270 |
| 2D MaxPooling | (Batch, 2, 2, 45) | 0 |
| 2D Convolution | (Batch, 2, 2, 45) | 1485 |
| Add | (Batch, 2, 2, 45) | 0 $\leftarrow$ |
| 2D Convolution | (Batch, 2, 2, 64) | 3349 |
| Input $w$ | (Batch, 1) | 0 |
| GlobalAveragePooling | (Batch, 64) | 0 |
| Dense | (Batch, 64) | 128 |
| Dense | (Batch, 128) | 16512 |
| Dropout | (Batch, 128) | 0 |
| Dense | (Batch, 64) | 8256 |
| Dropout | (Batch, 64) | 0 |
| Dense ($\log r$) | (Batch, 1) | 65 |

*Table 2.* Summary of the network architecture with residual skip connections indicated by the teal arrows which are added through the *Add* layer. Batch normalization and activation functions are added to each of the convolution and dense layers.