LETTER

doi:10.1038/nature23463

Fast automated analysis of strong gravitational lenses with convolutional neural networks

Yashar D. Hezaveh[1,2]*, Laurence Perreault Levasseur[1,2]* & Philip J. Marshall[1,2]

Quantifying image distortions caused by strong gravitational lensing—the formation of multiple images of distant sources due to the deflection of their light by the gravity of intervening structures—and estimating the corresponding matter distribution of these structures (the 'gravitational lens') has primarily been performed using maximum likelihood modelling of observations. This procedure is typically time- and resource-consuming, requiring sophisticated lensing codes, several data preparation steps, and finding the maximum likelihood model parameters in a computationally expensive process with downhill optimizers[1]. Accurate analysis of a single gravitational lens can take up to a few weeks and requires expert knowledge of the physical processes and methods involved. Tens of thousands of new lenses are expected to be discovered with the upcoming generation of ground and space surveys[2,3]. Here we report the use of deep convolutional neural networks to estimate lensing parameters in an extremely fast and automated way, circumventing the difficulties that are faced by maximum likelihood methods. We also show that the removal of lens light can be made fast and automated using independent component analysis[4] of multi-filter imaging data. Our networks can recover the parameters of the 'singular isothermal ellipsoid' density profile[5], which is commonly used to model strong lensing systems, with an accuracy comparable to the uncertainties of sophisticated models but about ten million times faster: 100 systems in approximately one second on a single graphics processing unit. These networks can provide a way for non-experts to obtain estimates of lensing parameters for large samples of data.

At its core, lens modelling measures the parameters of a highly non-linear image distortion. With recent advances in computer vision and deep learning, convolutional neural networks (Methods) have been shown to excel at many image recognition and classification tasks[6]. This makes them a particularly promising tool for the analysis of gravitational lenses. Recently, these networks have been used to search for gravitational lenses in large volumes of telescope data[7–9] and to simulate weakly lensed galaxy images[10]. Here we show that these networks can also be used for data analysis and parameter estimation.

We train four networks, Inception-v4[11], AlexNet[12], OverFeat[13] and a network of our own design, to analyse strongly lensed systems, by removing their final classification layer and interpreting the outputs of the last fully connected layer as a prediction for lensing parameters, with all weights initialized at random. We train the networks to predict the five parameters of the singular isothermal ellipsoid profile: the Einstein radius, the complex ellipticity and the coordinates of the centre of the lens. We use a squared-difference cost function, averaged over the five parameters. Although in many situations in machine learning collecting sufficiently large training sets is one of the main challenges, here it is possible to simulate the training data extremely fast. We train the networks on half a million simulated strong lensing systems. The lensed background sources are composed of three equal sets of images: the first and second comprise real galaxy images from the Galaxy Zoo[14] machine learning challenge and high-quality images from the GREAT3 training data[15], and the third set is composed of simulated clumpy galaxies with Sérsic and Gaussian clump profiles. The position of the background galaxy in the source plane is chosen randomly for each sample, but limited to regions where strong lensing occurs, that is, inside or on the caustics.

We use a stochastic gradient-descent optimizer to train the networks. At each training step, we select a random sample of simulated data,

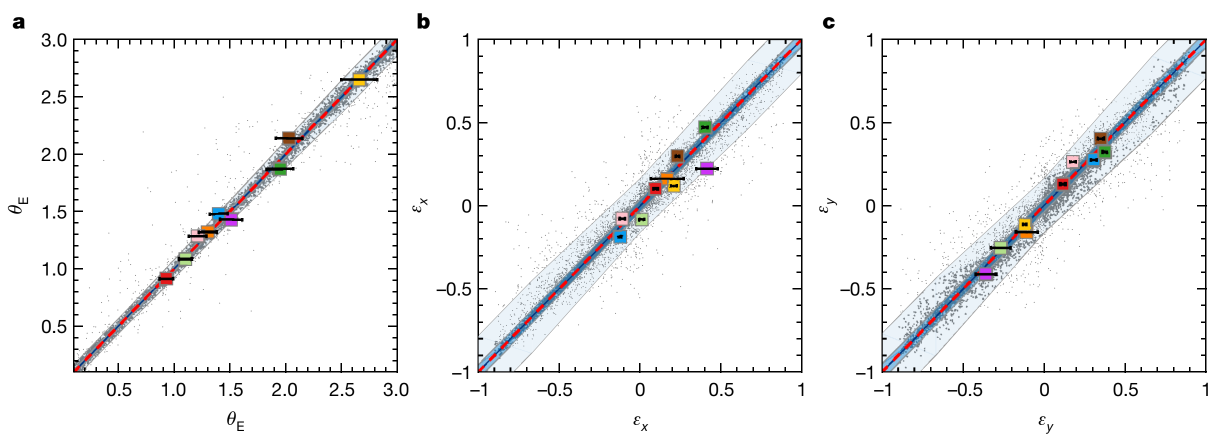


Figure 1 | Comparison of estimated parameters with their true values. The estimated values of the Einstein radius θ_E (**a**) and the x and y components of the complex ellipticity ε_x and ε_y (**b** and **c**) are shown on the y axis; the true values are shown on the x axis. The red dashed line marks the $y = x$ diagonal, on which perfectly recovered parameters should lie. The shaded blue areas represent the 68% and 95% intervals of the parameters recovered from a test set that the network has not been trained on. The small grey dots show the parameters of 10,000 test samples. The coloured data points and their error bars (95% confidence) correspond to real HST images of gravitational lenses, with the true parameters set to previously published values[17].

[1]Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, California, USA. [2]SLAC National Accelerator Laboratory, Menlo Park, California, USA.
*These authors contributed equally to this work.

apply randomly generated, realistic observational effects to each image, and use them to optimize the network weights. These effects include convolution with a point-spread function (PSF), addition of Poisson shot noise, Gaussian random noise with either a white or coloured power spectrum, simulated faint cosmic rays, hot pixels, a zero bias, and a random distribution of circular masks. The parameters of these observational effects, such as noise levels, span a range of realistic values (see Methods for details). Because these effects are randomly generated at each training step, we never encounter two identical realizations of the training data. Combined with the large size of the training set, this substantially mitigates the risk of overfitting. Masks added during training are included to allow for the possibility of masking undesired artefacts in real data that the networks have not been trained on, such as extremely bright cosmic rays and ghosts. Because these masks are allowed to partially cover up to 25% of the flux of the arcs, they also render the networks insensitive to incomplete data. To further increase our accuracy, we combine the predictions in a final trainable layer.

Our validation and test sets are both produced using the same pipeline, but with different random seeds and using background galaxy images that were not used to generate the training set (Extended Data Fig. 1). We quantify the accuracy of our predictions by calculating the interval that contains 68% of the predicted parameters. Our final 68% errors from the combined network on the lensing parameters are 0.02″, 0.04, 0.04, 0.04″ and 0.04″ for the Einstein radius, the x and y components of ellipticity, and the x and y coordinates of the centre of the lens, respectively. These errors are comparable to typical uncertainties on the parameters estimated from lens modelling with maximum likelihood methods for images with similar quality and noise levels[16,17]. In Fig. 1 we show the estimated parameter values of the combined network as a function of their true values. The grey points show the parameters of 10,000 test samples. The blue shaded regions show the 68% and 95% inclusion intervals. Table 1 summarizes the 68% errors of the individual and combined networks.

Table 1 | Errors of the individual and combined networks

Network	θ_E (arcsec)	ε_x	ε_y	x (arcsec)	y (arcsec)
Inception-v4[11]	0.03	0.04	0.05	0.06	0.06
AlexNet[12]	0.03	0.04	0.04	0.05	0.06
OverFeat[13]	0.04	0.05	0.05	0.06	0.06
Our network	0.03	0.05	0.06	0.05	0.05
Combined network	0.02	0.04	0.04	0.04	0.04

The columns present the 68% errors for the Einstein radius (θ_E), the x and y components of complex ellipticity (ε_x and ε_y), and the coordinates of the lensing galaxy (x and y) for each individual network and the combined network. The angular parameters (θ_E, x and y) are given in units of arcseconds.

In addition to the multiply lensed images of background sources, optical data often include light contamination from lensing galaxies. Prior to lens modelling, this light is commonly removed in a preprocessing step by fitting a model, such as Sérsic, to the light distribution of the lens while masking the lensed arcs, which requires an additional supervised optimization procedure[17]. Moreover, lensing galaxies often include complex structures that are not captured by simple parametric models, resulting in substantial residuals.

To fully automate the process of parameter estimation, we use independent component analysis (ICA) to separate the light profiles of the lens and the source arcs using multi-wavelength data. ICA is a method for separating an additive mixture of independent signals into their subcomponents. In this context, the morphologies of the background and foreground galaxies are statistically independent. The colour difference between these galaxies (both intrinsic and due to redshifting) results in different linear combinations of their light in different filters. Therefore, the separation of two components from two filters using ICA can help to remove the lens light from the background arcs. Intrinsic colour variations in the source and lens galaxies and

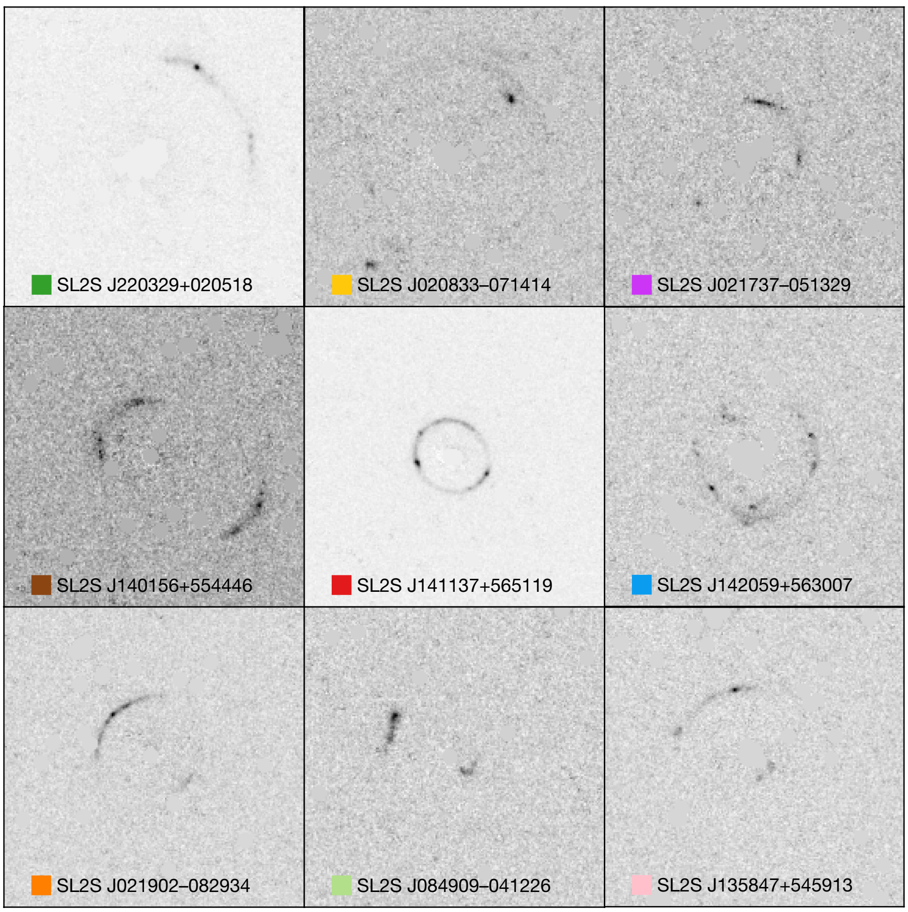


Figure 2 | Hubble Space Telescope images of strongly lensed galaxies from the SL2S survey. These images are used to demonstrate the performance of the network on real data. The light of the lensing galaxies has been removed using independent component analysis of two filters, and circular masks with radii of 0.2″ have been applied to bright cosmic rays and the lens centre. Each panel contains the object name in addition to the data marker used to show its parameters in Fig. 1.

the effects of blurring with different PSFs can result in imperfect ICA separation. However, we find that these issues have only a small effect on the resulting images and do not affect the accuracy of the lensing parameters that are recovered.

We demonstrate the application of this pipeline on real data by analysing Hubble Space Telescope (HST) images of SL2S survey lenses and comparing our estimates with previously published values[17]. From this sample, we select all grade A lenses with at least two Wide Field Camera 3 (WFC3) filters and a minimum signal-to-noise ratio of 10 per pixel. The brightest cosmic rays and the innermost region of the lens in the resulting subsample of nine systems are zeroed with circular masks that were included in the training of the networks. ICA is then applied to the images (Extended Data Fig. 2). In Fig. 2 we show the resulting separated arcs. These images are then fed into our networks. The colour markers in Fig. 1 show the parameters that were obtained for all nine lenses. The values on the x axis and their uncertainties are taken from previously published work[17]. The 95% uncertainties of published values are consistent with our estimated values within the error of the network, an accuracy sufficient for most studies[18].

The only manual preprocessing step in this analysis was the trivial masking of the brightest cosmic rays before the application of ICA. At the time of parameter estimation, our networks do not require the parameters of observational effects, such as the PSF, to be specified. Although only five parameters have been predicted here, estimating these same parameters from maximum likelihood methods requires the inclusion of hundreds to thousands of other nuisance parameters in the model, describing the morphologies of background sources[19,20], which greatly increases the complexity and the computational cost of the parameter search. Moreover, our networks could be modified and trained to predict the morphologies of background sources in addition to a larger number of parameters for more complex density structures, with negligible additional computational cost at evaluation time. Assuming that the typical analysis of a lens with the current density profile would take a few days to complete, our method offers an improvement in speed of about seven orders of magnitude, while also automating parameter estimation. This improvement could be even greater when analysing more complex density structures.

We also trained Inception-v4 to estimate lensing parameters in the presence of lens light, without the application of ICA. The 68% errors of this network are 0.07″, 0.1, 0.1, 0.04″ and 0.04″, higher than those of lens-removed images, but showing substantial promise for further development, especially if colour information is provided.

Although in their current forms our networks predict only global lens parameter solutions, it is in principle possible to use neural networks for parameter uncertainty estimation and posterior mapping using dropout[21] or Bayesian networks[22]. Using these methods, networks similar to those presented here will be able to provide insight into parameter degeneracies, including multimodal posteriors, an important issue for systems that consist of a large number of degenerate lenses, such as clusters[23].

Neural networks provide a fast alternative to the maximum likelihood methods that are commonly used to estimate parameters of interest in astrophysics from imaging data. Their effectiveness for image analysis makes them a powerful tool for applications beyond lensing studies, including stellar-mass estimation from multi-wavelength data and measurements of dust temperature. By streamlining these tasks, these networks can extend the reach of fast parameter estimation to general users.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 May; accepted 3 July 2017.

1. Lefor, A. T., Futamase, T. & Akhlaghi, M. A systematic review of strong gravitational lens modeling software. *New Astron. Rev.* **57,** 1–13 (2013).
2. LSST Science Collaborations. LSST Science Book, Version 2.0. Preprint at https://arxiv.org/abs/0912.0201 (2009).
3. Collett, T. E. The population of galaxy-galaxy strong lenses in forthcoming optical imaging surveys. *Astrophys. J.* **811,** 20 (2015).
4. Hyvärinen, A., Karhunen, J & Oja, E. *Independent Component Analysis* (Wiley, 2001).
5. Kormann, R., Schneider, P. & Bartelmann, M. Isothermal elliptical gravitational lens models. *Astron. Astrophys.* **284,** 285–299 (1994).
6. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vision* **115,** 211–252 (2015).
7. Petrillo, C. E. *et al.* Finding strong gravitational lenses in the Kilo Degree Survey with convolutional neural networks. Preprint at https://arxiv.org/abs/1702.07675 (2017).
8. Jacobs, C., Glazebrook, K., Collett, T., More, A. & McCarthy, C. Finding strong lenses in CFHTLS using convolutional neural networks. *Mon. Not. R. Astron. Soc.* **471,** 167–181 (2017).
9. Lanusse, F. *et al.* CMU DeepLens: deep learning for automatic image-based galaxy-galaxy strong lens finding. Preprint at https://arxiv.org/abs/1703.02642 (2017).
10. Ravanbakhsh, S., Lanusse, F., Mandelbaum, R., Schneider, J & Poczos, B. Enabling dark energy science with deep generative models of galaxy images. In *Proc. 31st AAAI Conf. on Artificial Intelligence* 1488–1494 (AAAI, 2017).
11. Szegedy, C., Ioffe, S., Vanhoucke, V & Alemi, A. A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proc. 31st AAAI Conf. on Artificial Intelligence* 4278–4284 (AAAI, 2017).
12. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (eds Pereira, F. *et al.*) 1097–1105 (Curran Associates, 2012).
13. Sermanet, P. *et al.* OverFeat: integrated recognition, localization and detection using convolutional networks. Preprint at https://arxiv.org/abs/1312.6229 (2014).
14. Willett, K. W. *et al.* Galaxy Zoo: morphological classifications for 120 000 galaxies in HST legacy imaging. *Mon. Not. R. Astron. Soc.* **464,** 4176–4203 (2017).
15. Mandelbaum, R. *et al.* The Third Gravitational Lensing Accuracy Testing (GREAT3) challenge handbook. *Astrophys. J. Suppl. Ser.* **212,** 5 (2014).
16. Bolton, A. S. *et al.* The Sloan Lens ACS Survey. V. The full ACS strong-lens sample. *Astrophys. J.* **682,** 964–984 (2008).
17. Sonnenfeld, A., Gavazzi, R., Suyu, S. H., Treu, T. & Marshall, P. J. The SL2S galaxy-scale lens sample. III. Lens models, surface photometry, and stellar masses for the final sample. *Astrophys. J.* **777,** 97 (2013).
18. Sonnenfeld, A. *et al.* The SL2S galaxy-scale lens sample. V. Dark matter halos and stellar IMF of massive early-type galaxies out to redshift 0.8. *Astrophys. J.* **800,** 94 (2015).
19. Warren, S. J. & Dye, S. Semilinear gravitational lens inversion. *Astrophys. J.* **590,** 673–682 (2003).
20. Suyu, S. H., Marshall, P. J., Hobson, M. P. & Blandford, R. D. A Bayesian analysis of regularized source inversions in gravitational lensing. *Mon. Not. R. Astron. Soc.* **371,** 983–998 (2006).
21. Gal, Y & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *Proc. 33rd Int. Conf. on Machine Learning (ICML-16)* Vol. 48, 1050–1059 (PMLR, 2016).
22. Gal, Y. & Ghahramani, Z. Bayesian convolutional neural networks with Bernoulli approximate variational inference. Preprint at https://arxiv.org/abs/1506.02158 (2016).
23. Johnson, T. L. *et al.* Lens models and magnification maps of the six Hubble Frontier Fields clusters. *Astrophys. J.* **797,** 48 (2014).

Acknowledgements We thank R. Keisler, G. Holder, R. Blandford, R. Wechsler and W. Morningstar for discussions and comments on the manuscript. We also thank G. P. Maher and A. Dwaraknath for comments about neural networks, leading to improved performance. We thank Stanford Research Computing Center and their staff for providing computational resources (Sherlock Cluster) and support. Support for this work was provided by NASA through Hubble Fellowship grant HST-HF2-51358.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS 5-26555. P.J.M. acknowledges support from the US Department of Energy under contract number DE-AC02-76SF00515. Y.D.H. is a Hubble Fellow.

Author Contributions Y.D.H. and L.P.L. contributed equally to all aspects of this project, including design and implementation of the networks and the text of the paper. L.P.L. developed the use of ICA for lens removal. P.J.M. contributed to various aspects of this project, including the choice of tensor libraries and tests on real and simulated data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to Y.D.H. (hezaveh@stanford.edu) and L.P.L. (lplevass@stanford.edu).

Reviewer Information *Nature* thanks Y. Gal, A. Sonnenfeld and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

A feed-forward neural network is a collection of processing units (referred to as neurons) designed to identify underlying relationships in input data. Neurons are organized in layers, with the output of each layer being the input of the next layer. The output of any individual neuron can be written as $f(w\boldsymbol{x})$, where $\boldsymbol{x}$ is a vector input to the neuron (for example, pixel values of an image), w is a matrix of weights (determined through training) and f is a nonlinear function referred to as an activation function. A network processes a given input $\boldsymbol{X}$ and maps it to an output $\boldsymbol{Y} = F_w(\boldsymbol{X})$. A set of training data $\{\boldsymbol{X}_{\rm train}, \boldsymbol{Y}_{\rm train}\}$ is used to determine the values of the weight matrices for this mapping. This is performed by minimizing the deviation between the predictions of the network $F_w(\boldsymbol{X}_{\rm train})$ and the true values $\boldsymbol{Y}_{\rm train}$ for the examples in the training set by optimizing a cost function with respect to the weight matrices. The universal approximation theorem states that, under mild assumptions, a two-layer neural network can approximate any continuous function with arbitrary accuracy with a finite number of neurons[24]. However, in practice, such a simple network would require a large number of neurons and would be difficult to train. More layers can allow networks to identify higher-order complex correlations with fewer parameters (weights) and to be more easily trained. In convolutional neural networks, the weights of a layer are organized in multiple two-dimensional structures, representing a set of filters. The values of neurons are then the result of the convolution of the input to the layer with these filters. This allows networks to retain two- and three-dimensional structures of input data and to extract specific patterns, defined by weight filters, from them. To perform maximum likelihood lens modelling, given a set of lensing parameters $\boldsymbol{p}$, a simulated image $\boldsymbol{M} = L(\boldsymbol{p})$, is produced and compared to real data. Here L is the operation that maps the vector of parameters $\boldsymbol{p}$ to the simulated images $\boldsymbol{M}$, for example, ray-tracing and convolution with a PSF. On the other hand, our networks learn the inverse of this function, mapping each image to a vector of lensing parameters $\boldsymbol{p}$. Given the flexibility of our networks to approximate complex functions, and the two-dimensional structure of images, convolutional neural networks are well suited for estimating lensing parameters from image data.

Simulated datasets. Our datasets use a mixture of images of real and simulated background galaxies. The images of real galaxies comprise 60,000 images from the Galaxy Zoo project and 8,000 images from the GREAT3 challenge data to produce half a million lensed images. The Galaxy Zoo images are averaged over the colour channels to produce a grayscale image. The simulated clumpy galaxies are composed of 1–5 clumps. Their spatial distributions follow a randomly generated correlated Gaussian with a long axis defined as the radius of the galaxy. Each clump follows either a Sérsic profile with index $n = 1$–5, effective radius $R_{\rm eff} = 0.1''$–$0.2''$ and ellipticity $\varepsilon = 0$–0.4 or a Gaussian profile. The apparent unlensed size of the galaxies is allowed to span the range $0.05''$–$0.8''$. The Einstein radius of the lens is sampled from a flat distribution over $0.1''$–$3.0''$. The ellipticity and angle of the lens are also chosen at random from a flat distribution and converted to the x and y components of ellipticity with a maximum ellipticity of 0.9. The coordinates of the centre of the lens cover the central $0.5''$ of the image. To ensure that the position of the sources behind the lens results in strong lensing configurations, we calculate the lensing caustics and place the sources at random inside or close to the edges of the caustics. We also ensure that the lensed images have a minimum total flux magnification of 2.0. The images consist of 192×192 pixels, with a pixel size of $0.04''$, equal to HST WFC3 pixels. Note that for other input pixel sizes, it is possible to interpolate the images on this grid. During training, at each step of a gradient-descent optimization, we add realistic observational effects to the images. We first convolve the images with a Gaussian filter with a randomly chosen root-mean-square (r.m.s., maximum of $0.1''$) to simulate the blurring effect of the PSF. We then convert the normalized intensity of our images to photon counts using a factor of 100–1,000, and generate a Poisson realization map with a mean of these values, effectively adding Poisson noise to the images. We then add random Gaussian noise with either a white power spectrum or a power spectrum measured from blank fields of a sample of HST images. The r.m.s. of the noise is chosen randomly from a flat distribution with a minimum of 1% and a maximum of 10% of the peak signal. To make the networks insensitive to pixel artefacts and low-intensity cosmic rays, we generate 400,000 images that contain simulated cosmic rays and hot pixels. To produce each cosmic ray, we choose a random pixel and populate the neighbouring pixels sequentially, using a friends-of-friends algorithm, until the desired length is achieved. The total number of pixels for each cosmic ray and the total number of such events are also chosen at random for each map. Each artefact map may also include up to 100 hot pixels. These maps are selected at random and added to the lensed images at the training time. The maximum intensity of the resulting images is normalized to 1, with a small deviation (r.m.s. of 0.05), and a zero bias is added to the images (Gaussian, r.m.s. of 0.05). We then generate a number of circular masks with radii of $0.2''$, placed at random throughout the images. Masks are allowed to cover up to 25% of the image flux. The test set is made from 10,000 GREAT3 source images that the networks have not been trained on. For testing, we allow only less than 2% of the lensed flux to be masked. Extended Data Fig. 1 shows a few examples from the test set.

Training. The networks include a first added layer to remove the image intensity biases by filtering the maps with a 4×4 flat filter and subtracting the minimum of the resulting maps from the input images while excluding the masked regions from this calculation. Owing to the different nature of this problem from the classification tasks of ImageNet, we do not initialize our networks from previously trained weights on this dataset. Instead, we use either a Xavier initialization or an initialization from a fixed normal distribution and train entire networks with the Adam optimizer. The network that we designed (in addition to Inception-v4, AlexNet and OverFeat) consists of eight convolutional modules. Each module consists of a convolution with a large filter, followed by a 1×1 convolution with the same depth. The third, fifth and seventh modules use stride 2 convolutions to reduce the image size. The sizes of the primary kernels are 3, 5, 10, 10, 10, 10 and 3, with depths of 32, 32, 32, 32, 64, 64, 128 and 256. The feature maps are then passed to two fully connected layers with 512 and 5 (output) nodes. All layers except the last have a rectified linear activation. We calculate a Euclidean mean-square loss averaged over the five parameters. Because of the parity invariance of the components of ellipticity, we calculate this loss for both $(\varepsilon_x, \varepsilon_y)$ and $(-\varepsilon_x, -\varepsilon_y)$, keeping the minimum of the two loss values. The large size of our training data and the application of randomly generated observational effects for each encounter of the training examples mitigate the risks of overfitting. We do not use dropout layers and notice that even with long training the average cost value for the training data does not exceed that of the validation set. About one day of training on a single graphics processing unit can result in modest accuracy. Improving this accuracy to that quoted in the main text requires an additional few days of training. Neural networks can have a large number of local minima, with some having poorer prediction performance than others. It may sometimes be necessary to restart the training of networks to find a better local minimum than the current solution. However, it has been empirically observed that the highly non-convex cost function of such networks can usually be optimized relatively easily using local updates[25]. To average the results of the four networks, we combine their individual predictions in a final trainable layer. For this, we fix the four networks and train only the parameters of this layer alone.

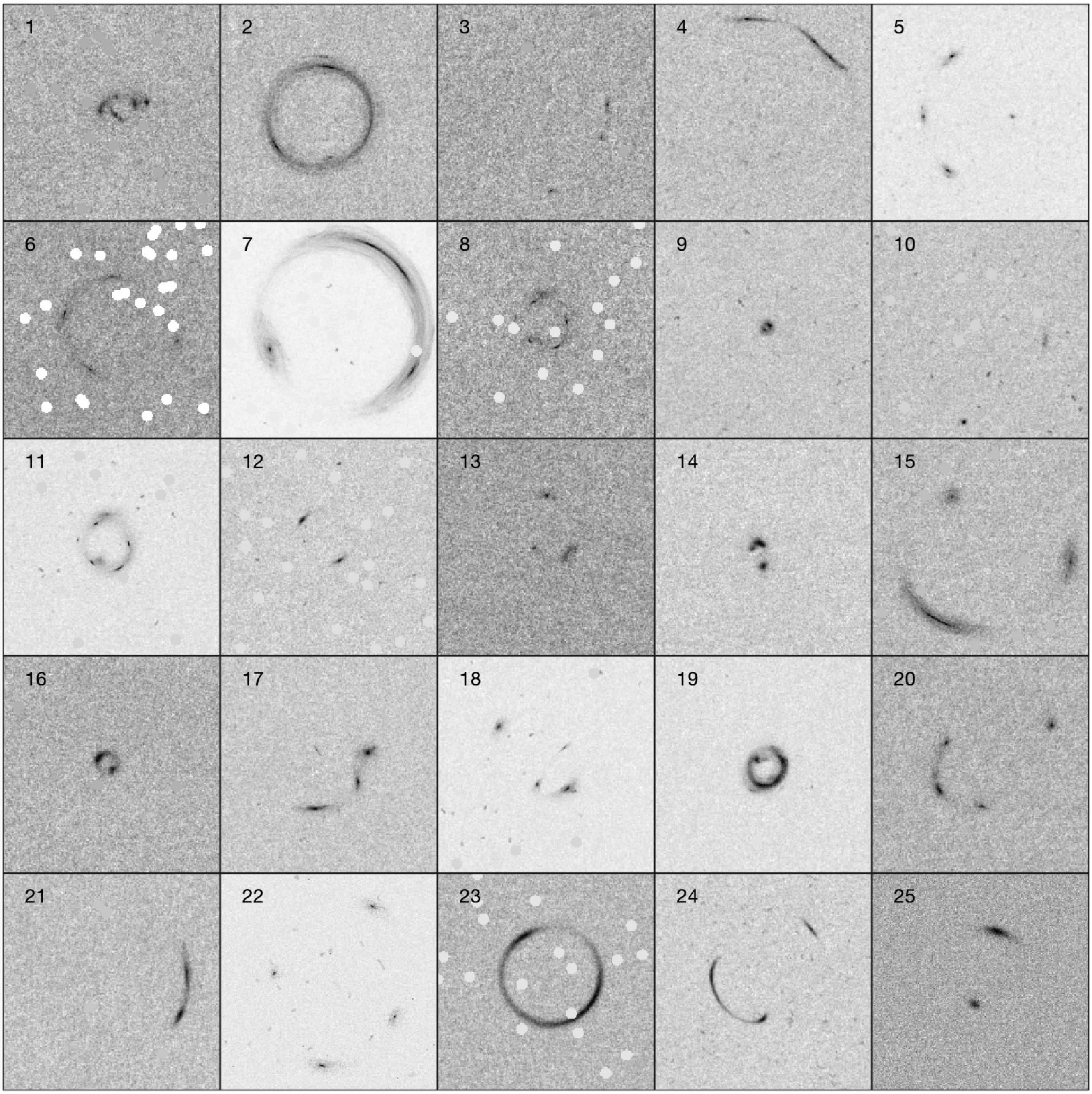
SL2S data and ICA. We obtain the SL2S HST WFC3 data from the Mikulski Archive for Space Telescopes. We select the lenses classified as grade A[17] with a minimum signal-to-noise ratio of 10 per pixel, with at least two available WFC3 filter images. The images are visually cropped to include the lensed arcs near the centre of the images. We apply circular masks of radius $0.2''$ to the bright cosmic rays and to the brightest regions of the lensing galaxies. The images are then passed to a fast ICA code to be unmixed into two subcomponents. Extended Data Fig. 2 shows the resulting separated ICA components from the two filters. The resulting images of the arcs are fed into the networks. True colour variations in each of the galaxies and PSF differences in the two images can result in imperfect component separation. To test if such effects influence the decisions of the networks in determining lensing parameters, we also subtract the light of the lensing galaxies with a Sérsic profile. We do not observe a noticeable difference in the estimated parameters using the two different approaches. To convert the published uncertainties of ellipticity and angle[17] to uncertainties in the x and y components of ellipticity, we assume an uncorrelated normal posterior for the polar parameters. For SL2S J220329+020518, previous analysis excluded a central blue arc, initially thought to be a part of the background source, from lens models, on the basis of an additional spectroscopic analysis[17]. We similarly masked this image before feeding it into our pipeline.

Tests on robustness. We evaluated the performance of the networks for tests convolved with a realistic (non-Gaussian) HST PSF and found that they did not reduce the parameter accuracies substantially. It is expected that the performance of the networks on images with lensing parameters outside the range explored in training data or with a new class of nuisance features that the network has not been trained on (such as a satellite track) will decrease with the strength of these effects. For example, we tested the performance of our networks, trained on data with no external shear, on new examples of lensing systems which included external shear with a maximum shear of 0.3. The r.m.s. errors of the networks for Einstein radius and complex ellipticity increase steadily for increasing shear whereas the lens centre is robustly predicted at all shear values. The r.m.s. errors for ellipticity increase from 0.11 for systems with shear of less than 0.01 to 0.23 for lenses with shear of more than 0.15. However, these errors could be greatly reduced by training the networks on systems that include external shear. Approximate Bayesian neural networks, which produce the posterior of the parameters for each example, can capture the uncertainties of the networks in their predictions for test examples outside the training space[22].

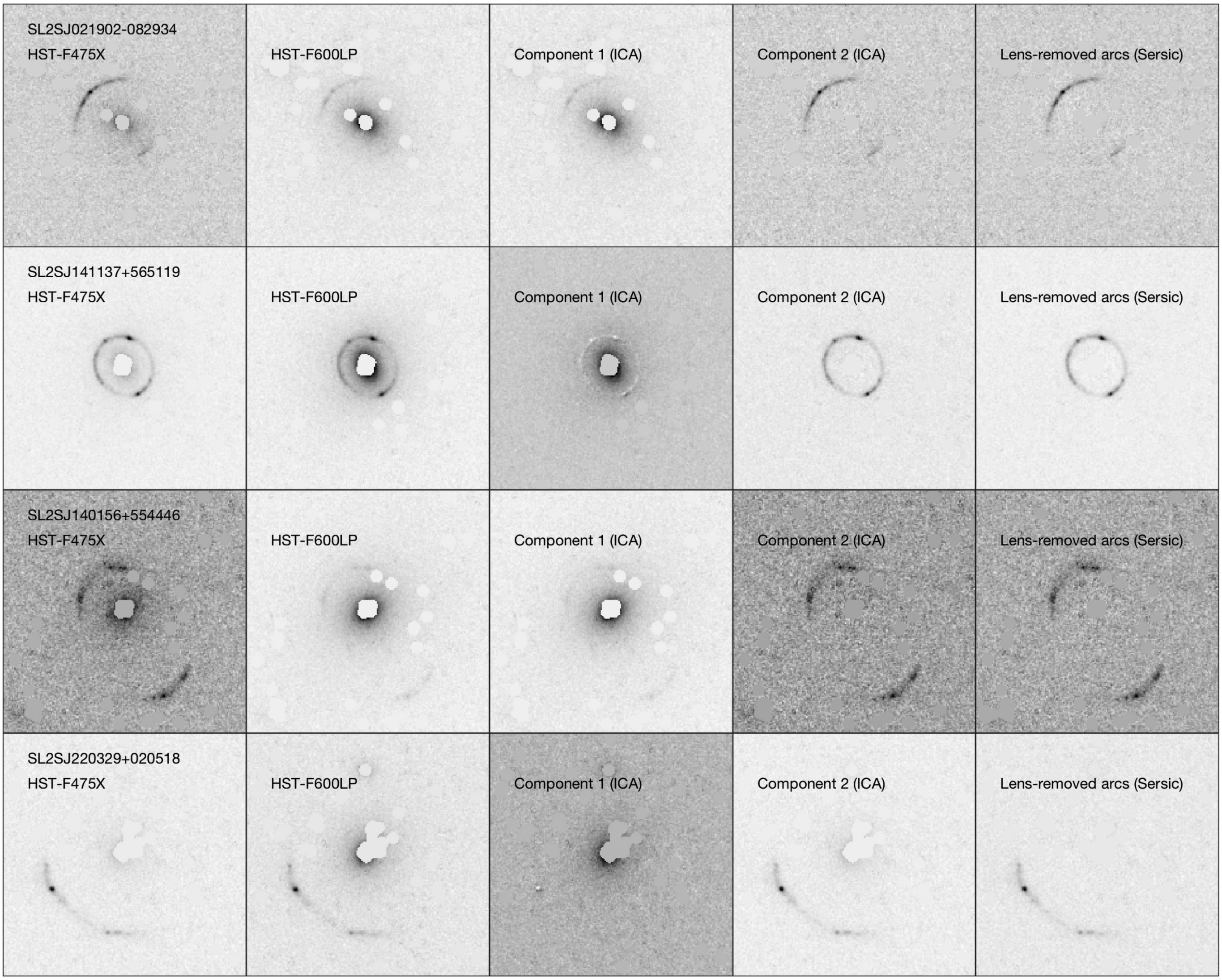
Code availability. The code used to simulate the training, validation and test sets, the code used to estimate the lensing parameters, and the trained network weights are freely available for download at https://github.com/yasharhezaveh/Ensai.

Data availability. The HST data used to test the performance of the networks are publicly available for download from the Mikulski Archive for Space Telescopes (https://archive.stsci.edu/hst/search.php). The unlensed images of background galaxies used to simulate the training, validation and test sets can be obtained from the GREAT3 website (http://great3challenge.info) and the Galaxy Zoo machine learning challenge web page (https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge).

24. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Contr. Signals Syst.* **2,** 303–314 (1989).
25. Baldassi, C. *et al.* Unreasonable effectiveness of learning neural networks: from accessible states and robust ensembles to basic algorithmic schemes. *Proc. Natl Acad. Sci. USA* **113,** E7655–E7662 (2016).



Extended Data Figure 1 | A selection of the test samples used to evaluate the performance of the network. These examples are chosen to illustrate the variations of different effects, including cosmic rays (for example, panels 11 and 12), masks (for example, panels 6 and 23), Einstein radii (for example, panels 7 and 9), noise levels and PSF blurring strengths, and a mixture of lensing image configurations including some unfavourable morphologies (for example, panels 10 and 21).



Extended Data Figure 2 | Examples of the inputs and outputs of the ICA. For each row, the first two panels show the HST images in F475X and F600LP filters. The third and fourth columns show the outputs of the ICA. For comparison, lens-removed arcs using a Sérsic model are shown in the last column. Cosmic rays and the brightest central parts of the lensing galaxies have been masked.