





Developing a victorious strategy to the second strong gravitational lensing data challenge

C. R. Bom ^{1,2★} B. M. O. Fraga ¹ L. O. Dias,¹ P. Schubert,^{1†} M. Blanco Valentin,^{1‡} C. Furlanetto ³,
M. Makler ^{1,4} K. Teles,¹ M. Portes de Albuquerque¹ and R. Benton Metcalf^{5,6}

¹Centro Brasileiro de Pesquisas Físicas, Rua Dr Xavier Sigaud 150, CEP 22290-180 Rio de Janeiro, RJ, Brazil

²Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Rodovia Mário Covas, lote J2, quadra J, CEP 23810-000 Itaguaí, RJ, Brazil

³Departamento de Física, Universidade Federal do Rio Grande do Sul, CEP 91501-970 Porto Alegre, RS, Brazil

⁴International Center for Advanced Studies & ICIFI, ECyT-UNSAM & CONICET, 25 de Mayo y Francia, C.P.: 1650, San Martín, Buenos Aires, Argentina

⁵Dipartimento di Fisica e Astronomia, Università di Bologna, Via Gobetti 93/2, I-40129 Bologna, Italy

⁶INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Gobetti 93/3, I-40129 Bologna, Italy

Accepted 2022 July 18. Received 2022 June 11; in original form 2022 March 11

ABSTRACT

Strong lensing is a powerful probe of the matter distribution in galaxies and clusters and a relevant tool for cosmography. Analyses of strong gravitational lenses with deep learning have become a popular approach due to these astronomical objects' rarity and image complexity. Next-generation surveys will provide more opportunities to derive science from these objects and an increasing data volume to be analysed. However, finding strong lenses is challenging, as their number densities are orders of magnitude below those of galaxies. Therefore, specific strong lensing search algorithms are required to discover the highest number of systems possible with high purity and low false alarm rate. The need for better algorithms has prompted the development of an open community data science competition named strong gravitational lensing challenge (SGLC). This work presents the deep learning strategies and methodology used to design the highest scoring algorithm in the second SGLC (II SGLC). We discuss the approach used for this data set, the choice of a suitable architecture, particularly the use of a network with two branches to work with images in different resolutions, and its optimization. We also discuss the detectability limit, the lessons learned, and prospects for defining a tailor-made architecture in a survey in contrast to a general one. Finally, we release the models and discuss the best choice to easily adapt the model to a data set representing a survey with a different instrument. This work helps to take a step towards efficient, adaptable, and accurate analyses of strong lenses with deep learning frameworks.

Key words: gravitational lensing; strong – methods: numerical – techniques: image processing.

1 INTRODUCTION

The strong gravitational lensing (SL) effect is a phenomenon produced by massive objects along the line of sight, typically matter haloes in cluster or galaxy scales, that deflect light from sources farther away. As a result, those systems commonly present magnified and multiple images of sources, which can be highly distorted in the form of rings or arcs.

Gravitationally lensed systems can be used as unique probes in many astrophysical and cosmological studies. For instance, the light deflection produces magnified images acting as a ‘gravitational telescope’, enabling the assessment of distant source objects or features that would be beyond the magnitude limit or resolution of a given survey if not lensed (e.g. Marshall et al. 2007; Poindexter, Morgan & Kochanek 2008; Jones et al. 2010; Richard et al. 2011;

Ebeling et al. 2018; Akhshik et al. 2020; Man et al. 2021). Lensing systems can also be used as non-dynamical probes of the mass distribution of galaxies (e.g. Treu & Koopmans 2002a,b; Koopmans et al. 2006), and galaxy clusters (e.g. Kovner 1989; Abdelsalam, Saha & Williams 1998; Natarajan, De Lucia & Springel 2007; Carrasco et al. 2010; Coe et al. 2010; Zackrisson & Riehm 2010), providing a relevant observational probe to dark matter (see e.g. Meneghetti et al. 2004).

Because of the cosmological distances involved, strong lensing has also been used to derive cosmological constraints on the cosmic expansion, dark energy, and dark matter (see e.g. Bartelmann et al. 1998; Cooray 1999; Yamamoto et al. 2001; Treu & Koopmans 2002b; Meneghetti et al. 2004; Jullo et al. 2010; Schwab, Bolton & Rappaport 2010; Enander & Mörtzell 2013; Pizzuti et al. 2017). In particular, accurate time-delay distance measurements of multiply imaged lensed quasi-stellar objects (QSO) systems have enabled precise measurements of the Universe’s cosmic expansion (Oguri 2007; Suyu et al. 2010; Wong et al. 2020). More recently, time-delay cosmography was also obtained with the strongly lensed supernova ‘Refsdal’ (Grillo et al. 2020). Furthermore, strong lensing can be used to constrain dark matter models (Vegetti et al. 2012; Hezaveh

* E-mail: debom@cbpf.br

† In memoriam.

‡ Present address: Electrical and Computer Engineering Department, McCormick School, Northwestern University, 633 Clark St, Evanston, IL 60208, USA.

et al. 2016; Bayer et al. 2018; Gilman et al. 2018), as well as to assess dark matter substructures along the line of sight (McCully et al. 2017; Despali et al. 2018).

Those systems' many applications and studies motivated an increasing number of searches for strong lensing systems. Earlier searches have been carried out on high-quality space-based data from the *Hubble Space Telescope* (*HST*), for instance, the *Hubble Deep Field* (HDF; Hogg et al. 1996), the Great Observatories Origins Deep Survey (GOODS; Fassnacht et al. 2004), the *HST* Medium Deep Survey (Ratnatunga, Griffiths & Ostrander 1999), the *HST* Archive Galaxy-scale Gravitational Lens Survey (Pawase et al. 2014) just to name a few.

However, the abundance of data in ground-based experiments, in particular wide-field surveys, fomented the exploration and identification (by visual inspection or automated search on images) of most of the known high-quality strong lensing candidates and, therefore, the majority of confirmed ones. Many candidates were identified in the Red-Sequence Cluster Survey (RCS; Gladders et al. 2003), in the Sloan Digital Sky Survey (SDSS; Estrada et al. 2007; Belokurov et al. 2009; Kubo et al. 2010; Wen, Han & Jiang 2011; Bayliss 2012), the Canada–France–Hawaii Telescope Legacy Survey (CFHTLS; Cabanac et al. 2007; More et al. 2012, 2016; Gavazzi et al. 2014; Maturi, Mizera & Seidel 2014; Paraficz et al. 2016), the Deep Lens Survey (DLS; Kubo & Dell'Antonio 2008), the Dark Energy Survey (DES; e.g. Nord et al. 2016; Diehl et al. 2017), and the Kilo-Degree Survey (KiDS; Petrillo et al. 2017). The future 2020s and 2030s observatories, like the Vera C. Rubin (Ivezic et al. 2019), *Euclid* (Laureijs et al. 2011), and *Nancy Grace Roman* (Green et al. 2012) telescopes, are expected to increase the number of strong lensing candidates by a few orders of magnitude than what is currently known (see e.g. Collett 2015).

Many of the earlier catalogues of strong lensing systems were found through visual searches only. Nevertheless, the current large data sets from wide-field surveys triggered the development of fully or human-assisted automated search methods to find and classify (Gavazzi et al. 2014; Joseph et al. 2014; de Bom et al. 2015; Avestruz et al. 2019; Cheng et al. 2020a), and more recently inferring the properties (Hezaveh, Levasseur & Marshall 2017; Bom et al. 2019; Legin et al. 2021; Pearson et al. 2021; Schuldt et al. 2021) of lens candidates.

Most of those techniques are based on image processing and/or neural networks. In particular, several works have established that both traditional neural networks (Estrada et al. 2007; Bom et al. 2017) and deep neural networks (Glazebrook et al. 2017; Lanusse et al. 2018; Jacobs et al. 2019; Metcalf et al. 2019; Petrillo et al. 2019a,b; Morgan et al. 2022) can be used to identify lenses from non-lenses, with minimal human intervention. Although there is a certain intuition on how those techniques work, except when the methods explicitly take morphological features or colours rather than the raw image, it is not completely clear which features are used by deep learning algorithms in strong lensing identification. Even so, several searches with deep learning techniques have successfully found a large number of candidates (see e.g. Glazebrook et al. 2017; Jacobs et al. 2019; Petrillo et al. 2019b). The interest in the field led to a community effort to develop the best solutions in the two strong gravitational lensing challenges (henceforth SGLC; Metcalf et al. 2019; Metcalf et al., in preparation). Both were data challenges where the participants developed different methods to find strong lensing from different sets. The first challenge was performed in 100 000 simulated images mimicking KiDS-quality survey data, with a 48-h time limit. The second challenge used 100 000 images with *Euclid*-like conditions, including four bands, one of them with a

different resolution in terms of pixel scale. The participants used this set of images, the tagged set, to search for the best possible model since their respective labels were also provided. For methods that require a training phase, this tagged set was split into three groups: training/validation/test sets. The final results of the challenges used a different blind set of 100 000 images. In this work, we present the path that ultimately led our team to develop the winning solution of the second SGLC (II SGLC). We discuss the lessons learned and how to work on the data. Most of our pipeline is generic enough and thus valuable for other potential applications in deep learning image processing problems in astronomy. We present how to pre-process the data, and what is the deep learning architecture choice, including how to combine images with different resolutions in the same network. Additionally, we use a technique to identify what region on the image the algorithm is using to make a classification, and therefore, we may assess the decision-making process. We further discuss the adaptability of our model to other data sets and the detection limits in terms of lensed pixels above the background. We make our pipeline and model weights publicly available.¹

This paper is organized as follows. In Section 2, we briefly describe the data and present the initial processing and data exploration. Later, in Section 3, we introduce the deep learning models, the architecture definition, and the choice of parameters used in this work. Then, in Section 4.1, we describe the training process, convergence, and overfitting. Following that, in Section 4.2, we describe the model's performance in the test set. In Section 4.3, we use a technique to infer which features are relevant to the deep learning classification. In Section 5, we evaluate the sensitivity of the current method and define the detection limits. Later, in Section 6, we present a prescription on how to adapt the method for a different data set. Finally, in Section 7, we make a discussion and present the conclusions of this work and an outlook for the future.

2 DATA EXPLORATION

2.1 Catalogue and available data

The data for the II SGLC are composed of 100 000 simulated objects in four different bands: VIS (visible) and *H*, *J*, and *Y* (infrared), for a total of 400 000 images (the training set mentioned in Section 1). The VIS images have a 200×200 pixels resolution with a pixel scale of 0.1 arcsec, while the other bands have ~ 3 times lower resolution, with 66×66 pixels images with 0.3 arcsec of pixel scale. The images are an update from the first SGLC (I SGLC; for a detailed description of the data sets, see Metcalf et al. 2019; Metcalf et al., in preparation).

The simulation of the lenses started with catalogues derived from the Millennium Observatory project (Overzier et al. 2013) that describes both the galaxies and dark matter haloes within a simulated light-cone. The galaxies were matched to dark matter haloes using the semi-analytic model (SAM) of Guo et al. (2011). The halo catalogue lists the mass, half-mass radius, maximum circular velocity, redshift, and angular position of each halo, including subhaloes. The galaxy catalogue includes parameters for both a disc and spheroidal component for each galaxy. The effective radius, position angle, and photometry in 10 bands along with other information are given.

A mass model is constructed for each dark matter halo consisting of a spherically symmetric Navarro–Frenk–White (NFW) profile with an external shear to account for some asymmetry. In addition, for each galaxy a truncated elliptical singular isothermal sphere (TESIS) is

¹https://github.com/cdebom/cast_lensfinder

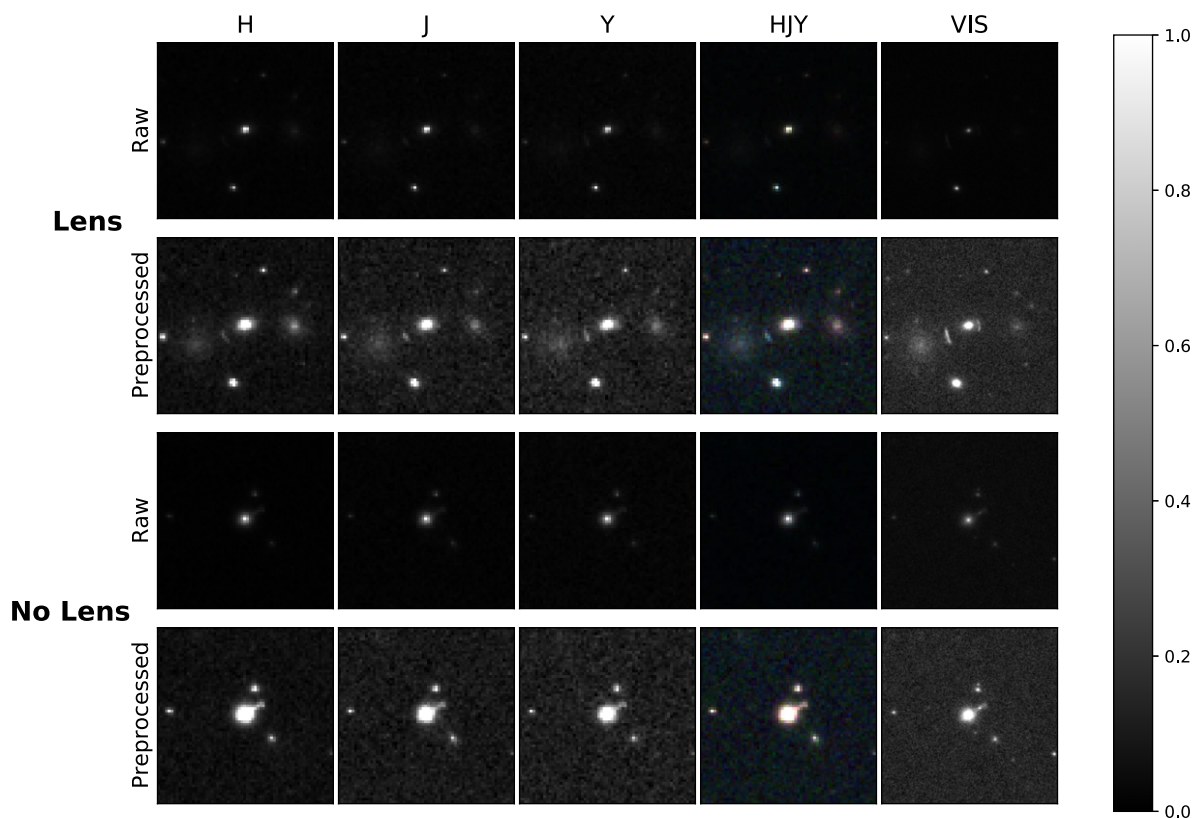


Figure 1. *H*, *J*, *Y*, *VIS*, and *HJY* band images combined from two objects on data set, a lensed and a non-lensed one, with and without any pre-processing. No features can be seen without pre-processing, and the emission from the galaxy itself is hardly visible. The images were normalized individually.

constructed that matches the velocity dispersion of the galaxy from the catalogue. The mass in these TESISs is subtracted from the host dark matter halo.

The image of each lens galaxy consists of a disc and spheroidal component. The disc has an inclined exponential profile and the spheroidal component has a Sérsic profile with an index near 4 with a random 10 per cent fluctuation. The discs have analytic spiral surface brightness fluctuations to mock spiral arms. The position angle of the mass matches that of the lens light.

The sources that are placed behind the lenses were taken from the *Hubble Ultra Deep Field* (UDF). Their UDF images were decomposed into shapelet functions to remove noise by Meneghetti et al. (2010). These images have more realistic shapes than simple analytic profiles. The redshifts and colours are taken from the UDF catalogue. The sources redshifts in the II SGLC were between 1.27 and 11.08, with a mean of 2.76.

The mass modelling and the subsequent ray shooting are done within the GLAMER lensing code (Metcalf & Petkova 2014; Petkova, Metcalf & Giocoli 2014). A particular lens was constructed by picking a galaxy at random from among the bright galaxies and a source from the source catalogue. A grid of rays with twice the resolution and the same range as the final image was shot backward in time to the redshift of the sources and critical curves and associated caustic curves were found. If the critical curve was within pre-set bounds, the field was accepted. Then it was randomly decided if a source should be added. The fraction of cases without sources were low because a large fraction of the cases with sources still did not meet the observational criterion for being a lens because of low source magnification or brightness.

If a source was added, it was placed within a circle with a radius that is 1.5 times greater than a circle that would circumscribe the largest tangential caustic in the field. The neighbouring objects within 0.5 arcmin are included. Some randomization of the position angle, inclination, neighbour positions, and other parameters are done so that the same object could be used multiple times without producing very similar images.

Properties of the lens and lensed images were calculated and logged. The image was then blurred, reduced in resolution, and noise was added according to what is expected for the *Euclid* observations. The objects had the same surface brightness profiles in the different bands although their relative brightness was changed according to the input colours. The spherical and disc components of galaxies generally had different colours.

During the competition’s duration, the organizers provided no details on how the simulations were generated.

Together with the images, a catalogue of properties was also provided, instead of a simple truth table. The catalogue contains, among others, the coordinates of the centre of the critical curves, the redshifts of source and lens objects, source effective magnification, and the number of source pixels with intensity 1σ above the background level. Thus, we may have simulated images where just 1 pixel from the lensed source was visible. For the challenge, a detectable strong lens system was defined by the following criteria:

- (i) number of separated groups of source pixels larger than zero;
- (ii) source’s effective magnification in all bands larger than 1.6;
- (iii) number of source pixels with intensity above 1σ of the background level larger than 20.

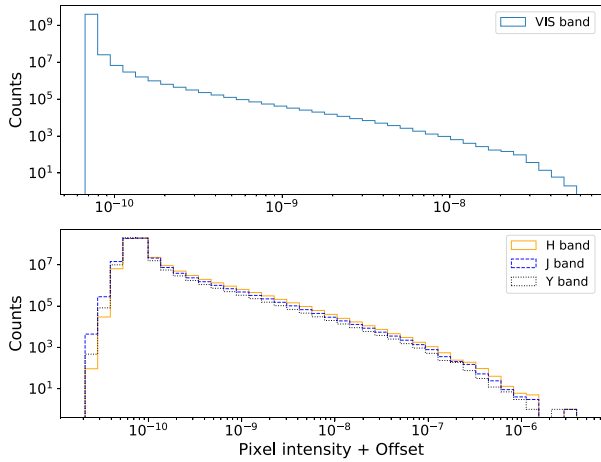


Figure 2. Pixel histogram from H , J , Y , and VIS image bands. The pixel values are offset by a fixed amount to remove negative values and allow the use of a logarithmic scale.

49 213 images in the data set followed these criteria and were therefore classified as a lens, making it reasonably balanced. Unlike a simple truth table, a catalogue like this allows us to explore how changing the detectable strong lens definition (e.g. the number of source pixels above the background level) can affect the classification.

Fig. 1 shows two example objects from the catalogue, one a lens (top two rows) and a non-lens object (bottom two rows), in each of the four bands, and joining H , J , and Y together. For the sake of comparison, we show also the same images after applying our pre-processing scheme (discussed in Section 2.2). The images without any pre-processing show almost no visible features and the galaxy itself is hard to see. After the pre-processing, not only the central galaxy appears (together with some background galaxies) but also the lens’ arc is now visible.

The pixel intensity distributions for all four bands of all images are shown in Fig. 2. They are all extremely skewed towards higher values, with the minimum and the median pixel values comparable. At the same time, the maximum is three to four orders of magnitude (depending on the band) larger than the median. Hence, to present the distribution, we use a logarithmic scale so that features are more easily seen. We also apply a constant offset to make all pixel values positive. These extremely skewed distributions help explain the lack of visual features when displaying the images: the brightest pixels are orders of magnitude brighter than the other ones, almost completely dominating the entire image and making the lensing features difficult to identify by visual inspection without any image processing. Furthermore, this extreme difference could make a neural network look only at the high-intensity pixels, almost completely disregarding the rest.

2.2 Pre-processing

The deep learning methods are sensitive to visual features by construction, so we pre-process the data to enhance the visualization. For completeness, we also evaluate the performance of our deep learning pipeline without any pre-processing. A simple normalization would not change the skewness of the pixel intensity distribution. Therefore, we first make a contrast adjustment by clipping the image histograms, i.e. we choose a lower and an upper bound and set every pixel above or below those values to be equal to the nearest bound. We define two sets of contrast adjustments. The first one was used for the II SGLC: since the images present negative pixel values, we chose the lower

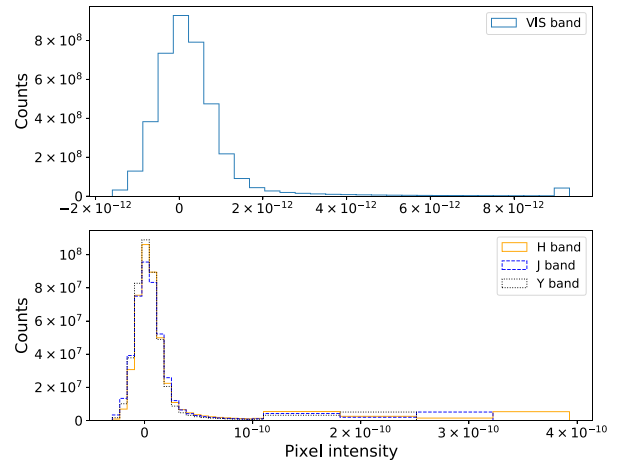


Figure 3. Pixel intensity histogram for VIS , H , J , and Y image bands with clipped images.

bound by first inverting the image, taking the 99.9 percentile, and multiplying it by -1 ; the upper bound was set as the 98 percentile, both using the full-pixel intensity distribution of each band. However, upon further visual inspection, we found that the H , J , and Y images were saturated in several cases, making the small features of objects or lenses disappear. After a search around the bounds used in the II SGLC and a visual inspection of some example images after each clipping, we chose a second contrast adjustment to the 0.1 and 99 percentile as the lower and upper bound, respectively. Fig. 3 shows the pixel intensity distribution after this clipping; the extremeness of the original distribution can now be appreciated: the 99 percentile for all bands is three to four orders of magnitude lower than the maximum. With the clipping, now the median the maximum are comparable.

The effect in the images can be seen in Fig. 4, which shows five examples of clipped lensed images in the VIS band and combined HJY band. Several features are now visible, such as the lenses and other galaxies in the background. However, while some of the lenses are easily noticeable in both VIS and infrared bands, some can only be readily seen in the VIS band, namely the ones in the second and last columns. This suggests that, at least in some cases, resolution might be more important than colour when visually searching for gravitational lenses. While this new pre-processing helped the lensing features visualization and thus motivates its usage, there is no guarantee that it will improve the results before the final deep learning performance evaluation.

3 DEEP LEARNING MODELS

3.1 EfficientNets

We made use of a family of convolutional neural network (CNN) models known as EfficientNet (Tan & Le 2019), which were built to be high performing, i.e. state of the art, in benchmarking image classification data sets such as ImageNet, while also being easily scalable. In order to improve a CNN’s performance, one can scale up the net in three ways: increasing the number of layers (depth), the number of channels (width), or the input image resolution. Increasing the depth could make the network learn more of the complex features but also makes it more prone to suffer from the vanishing gradients problem, where the derivative of the loss with respect to the weights approaches zero, and the learning stalls (Goodfellow, Bengio & Courville 2016). Furthermore, deeper networks tend to saturate on accuracy (He et al. 2016). On the other hand, while wide and shallow networks would

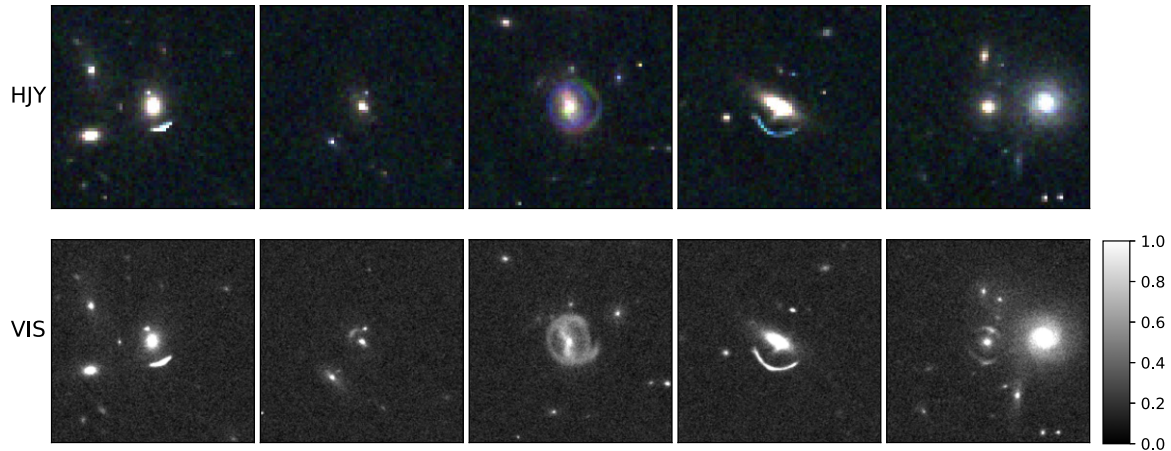


Figure 4. VIS and *HJY* band images combined from five lensed objects after pre-processing. The galaxies and lenses are now visible, and the difference in resolution can be appreciated: the lens cannot be seen in some of the *HJY* images, but it is visible in all VIS images.

avoid this problem, they would fail to learn more complex features. Therefore, Tan & Le (2019) introduced the idea of compound scaling, where depth, width, and image resolution are scaled up at the same time while maintaining a balance between them:

$$\begin{aligned} \text{depth: } d &= \alpha^\phi, \\ \text{width: } w &= \beta^\phi, \\ \text{resolution: } r &= \gamma^\phi, \end{aligned} \quad (1)$$

where ϕ is an integer named compound coefficient and α , β , and γ are constants. floating-point operations per second (FLOPS) on a convolutional operation scale as $d w^2 r^2$; thus, for it to scale as 2^ϕ for any new compound coefficient, α , β , and γ are subject to the following constraints:

$$\begin{aligned} \alpha \beta^2 \gamma^2 &\approx 2, \\ \alpha \geq 1, \beta \geq 1, \gamma \geq 1. \end{aligned} \quad (2)$$

Before scaling up, a good base model is needed; Tan & Le (2019) use a multi-objective architecture search, optimizing for accuracy and FLOPS, using the same parameter space as Tan et al. (2019, MnasNet). The resulting network is similar to MnasNet but with more parameters, called EfficientNet-B0, having a mobile inverted bottleneck with a squeeze and excitation connections as its building block (see Fig. 5b). A small grid search is done to obtain the best values for α , β , and γ for B0, which are then left constant. A family of EfficientNets (B1–B7 originally) can be then built by increasing the compound coefficient ϕ . EfficientNet-B7 achieved the best results for top-1 accuracy in ImageNet, outperforming more complex architectures in terms of the number of parameters. However, among the EfficientNet model family, many of them already have a high performance, around 80 per cent, in top-1 accuracy, i.e. the accuracy in multiclass problems considering as a correct prediction only the highest probability class.

3.2 Proposed architecture

EfficientNet models defined by Tan & Le (2019) are built to have as inputs three channels. Since *H*, *J*, and *Y* images have the same resolution, it is natural to combine them. However, the same cannot be done for the VIS band. Therefore, we define a second EfficientNet model to operate with the VIS images. We use two simple methods to create two extra channels in the VIS model:

- (i) propagate the images to the other channels, effectively creating three equal ones, which we call VIS (repeated);
- (ii) fill the other channels with null arrays, which we call VIS (zeros).

While this approach is suitable if we want to train either the *HJY* or VIS bands alone, it does not help if we want to use them together. In order not to scale up or down the images, keeping them in their original form, we chose to create a net with two EfficientNet branches, one for each input. The final model has the last two dense layers of the original architecture removed and concatenates both outputs before passing them through a fully connected layer with softmax activation (see Fig. 5a). We made an initial test of performance with the EfficientNet models and did not find any significant improvement in performance beyond the B2 model, considering cross-validation uncertainties, while the more complex B3–B7 models have more parameters and can take considerably longer to train. Thus, we use in our branches the EfficientNet-B2 architecture for our main results, which ended up being the model that achieved the highest score on the II SGLC. Furthermore, it is worth noticing that EfficientNet-B2 was also successfully implemented in image classifications of astrophysical sources for galaxy morphology catalogues as described in Bom et al. (2021). Walmsley et al. (2022) used an EfficientNet-B0 model for the same kind of problem.

Hence, we test the following architectures: one-branched EfficientNet, using either VIS or *HJY*, and two-branched with all bands. We tested the two versions for all cases with VIS band, VIS (repeated) or VIS (zeros). After some visual assessment in the near-infrared bands we found that many of the objects were visually very similar, so we also added a simpler model where we chose one of the near-infrared bands, *Y*, and VIS, filling the other two channels with null arrays. Therefore, we evaluate a total of six models presented in Table 1, each of them with the two pre-processing schemes described in the previous sections and without any pre-processing.

4 RESULTS

4.1 Training

The previously described models were trained using one of the state-of-the-art optimizers: the Rectified Adam (RADAM; Liu et al. 2020), with a binary cross-entropy loss. We initialize the networks with pre-trained weights from the ImageNet data set in order to improve the computing time needed for convergence, the overall performance,

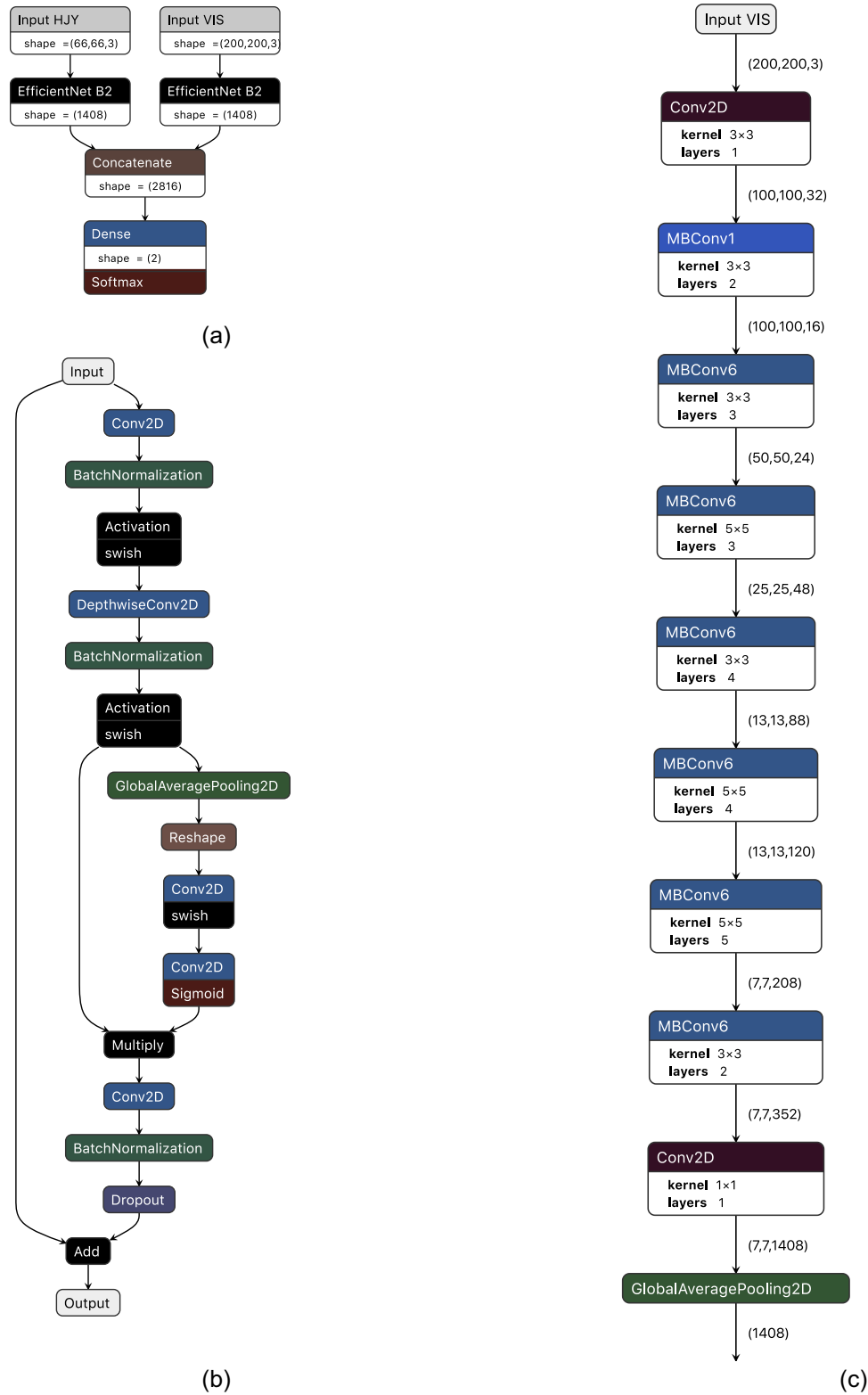


Figure 5. The deep learning architecture used in this work. The model was based on the EfficientNet-B2 model. Panel (a): our two-branched model used to train both images with *HJY* and *VIS* bands at the same time. Panel (b): mobile inverted bottleneck block, with squeeze and excitation phases. Panel (c): full architecture of the EfficientNet-B2 model. MBCConv blocks are the Mobile Inverted Bottleneck mentioned previously.

Table 1. Area under curve (AUC) for receiver operating characteristic (ROC) and precision–recall (PR) curves, and the F_β score for different bands and different pre-processing, on a reduced sample (20 000 images), using a cross-validation procedure. The scores are calculated on a blind test sample comprising 10 per cent of this reduced sample. As per the challenge, we use $\beta^2 = 0.001$. The combination submitted to the challenge is highlighted in red, and the best one in blue.

| | | No pre-processing | II SGLC pre-processing | Alternative pre-processing |
|--------------------------------|-----------|---------------------|----------------------------|----------------------------|
| <i>HJY</i> | ROC | 0.4994 ± 0.00135 | 0.5389 ± 0.01891 | 0.5350 ± 0.02414 |
| | PR | 0.4950 ± 0.0155 | 0.5535 ± 0.0210 | 0.5342 ± 0.0303 |
| | F_β | 0.620002 ± 0.102507 | 0.918411 ± 0.069823 | 0.800559 ± 0.098771 |
| VIS (repeated) | ROC | 0.5040 ± 0.0108 | 0.6589 ± 0.0423 | 0.6266 ± 0.0445 |
| | PR | 0.5129 ± 0.0779 | 0.6936 ± 0.0366 | 0.6634 ± 0.0434 |
| | F_β | 0.526797 ± 0.066389 | 0.982306 ± 0.007477 | 0.979158 ± 0.011016 |
| VIS (zeros) | ROC | 0.5001 ± 0.0101 | 0.7002 ± 0.0306 | 0.6534 ± 0.0414 |
| | PR | 0.4884 ± 0.0384 | 0.7326 ± 0.0256 | 0.6864 ± 0.0387 |
| | F_β | 0.548153 ± 0.066831 | 0.986763 ± 0.005173 | 0.979588 ± 0.010142 |
| <i>HJY</i> + VIS (zeros) | ROC | 0.5017 ± 0.0069 | 0.8018 ± 0.0173 | 0.8147 ± 0.0161 |
| | PR | 0.4916 ± 0.0074 | 0.8239 ± 0.0128 | 0.8359 ± 0.0132 |
| | F_β | 0.613814 ± 0.094273 | 0.990140 ± 0.003437 | 0.991610 ± 0.003873 |
| <i>HJY</i> + VIS (repeated) | ROC | 0.4932 ± 0.0102 | 0.8016 ± 0.0186 | 0.8295 ± 0.0088 |
| | PR | 0.4921 ± 0.0175 | 0.8230 ± 0.0149 | 0.8469 ± 0.0081 |
| | F_β | 0.619433 ± 0.106441 | 0.990402 ± 0.004087 | 0.992070 ± 0.003427 |
| <i>Y</i> + VIS (zeros) | ROC | 0.5011 ± 0.0174 | 0.8015 ± 0.0186 | 0.8149 ± 0.0144 |
| | PR | 0.4939 ± 0.0124 | 0.8243 ± 0.0135 | 0.8369 ± 0.0120 |
| | F_β | 0.613662 ± 0.125263 | 0.989978 ± 0.004941 | 0.990235 ± 0.004237 |

and to make the training more stable as in Bom et al. (2021). All images were normalized in the $[0, 1]$ range before being fed to the net. A 10-fold cross-validation was performed: the sample is divided in 10 parts; at each iteration, one of these parts is used as a validation sample, while the rest is used for training. This helps to avoid biases that could arise from the selection of specific training/validation sets. For each fold, we trained the net for 50 epochs with a batch size of 64.

Furthermore, we employ data augmentation strategies consisting of random rotations, mirroring in both axes, and zooming in or out the images between 0.8 and 1.2 times. We trained the models in a multi-GPU server with 8 RTX 3090 with 24 GB of GPU memory each. Since cross-validation is a high resource-consuming procedure, we selected a random sample of 20 000 images (four bands each) from the full data set for training/validation. The models and training were implemented in TensorFlow 2 (Abadi et al. 2015). The remaining 80 000 images are used as an independent test set for network performance assessment.

The training curves with no pre-processing show no or a negligible decrease in the training and validation losses, with the latter having large oscillations in some epochs indicating an underfit. The one-branched models containing *HJY* or *VIS* have the training loss decrease rapidly in the first few epochs, and more slowly later on; however, there is a strong overfit (validation loss much larger than training loss) already in the first 3–4 epochs, with the validation loss increasing and having again large oscillations. Training a two-branched network with *HJY* + *VIS* or *Y* + *VIS* simultaneously removes this large oscillation, and both losses decrease at approximately the same pace for 15–20 epochs when we start to see some overfitting.

To avoid being contaminated by overfitting at each fold, we use the model with the lowest validation loss to make our predictions on the test set. In Fig. 6, we present the difference between the initial and the lowest validation loss for each combination of bands and pre-processing. The error bars are standard deviations, as measured in all 10 trainings for each configuration. This quantity is related to how well the model learned the problem and its ability to generalize. Smaller values mean that the network did not learn a more general solution, as the validation loss did not decrease. This happens in all models with no pre-processing, and in models with only one

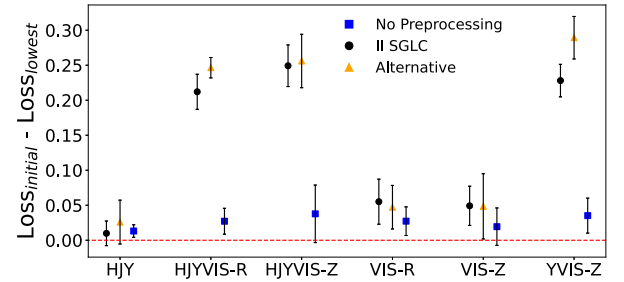


Figure 6. Mean validation loss difference between the initial and the best epoch colour coded for the different pre-processing methods used. Blue: no pre-processing; black: pre-processing as submitted to the challenge; and orange: new pre-processing. Larger differences indicate that the model learned more. Here R stands for repeated and Z for zeros.

branch (*HJY* or *VIS*). Furthermore, the II SGLC pre-processing and the current alternative one present similar results considering the errors.

Overfitting occurs when the difference between training and validation loss starts to diverge. In Fig. 7, we present the difference between the mean validation loss and the mean training loss at the last epoch and at the epoch with the lowest validation loss for every combination of bands and pre-processing. Smaller values suggest that overfitting did not occur or it is negligible. Nevertheless, the overfitting results should also consider the underfitting results from Fig. 6. In a situation of underfitting, one might also expect that the losses stay nearly constant for all epochs and thus small differences between training and validation loss. Confirming what was seen before, the models with no pre-processing did not learn, as the training and validation loss changed negligibly during training. Furthermore, one-branched models have a volatile training phase, with high variance in the loss at the last epoch and a little overfitting even in the best epoch. Training a two-branched network presents, as before, the best results, with almost no variance between the losses at each fold and little to no overfitting when considering the best epoch. However, it can be seen that for most models, there is a tendency

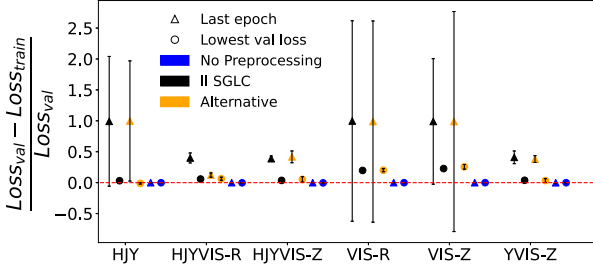


Figure 7. Difference between training and validation mean losses at the last and best epoch, colour coded for the different pre-processing methods used. Blue: no pre-processing; black: pre-processing as submitted to the challenge; and orange: new pre-processing. Triangles show the difference in the last epoch, while circles show it at epoch with the lowest validation loss. Lower differences indicate less overfitting or no learning at all. Here R stands for repeated and Z for zeros.

to overfit at later epochs. Again, the II SGLC and the alternative pre-processing are compatible within the errors except in the case of *HJY* and *VIS* (repeated), where training with the alternative pre-processing shows negligible overfit at all even in the last epoch.

4.2 Performance evaluation

To assess the models’ performance we consider several metrics, such as precision, recall, and the false alarm rate. The precision can be defined as

$$P = \frac{|\{\text{SL}\} \cap \{\text{Systems classified as SL}\}|}{|\{\text{Systems classified as SL}\}|}. \quad (3)$$

The recall can be defined as

$$R = \frac{|\{\text{SL}\} \cap \{\text{Systems classified as SL}\}|}{|\{\text{SL}\}|}. \quad (4)$$

The false alarm rate is

$$F = \frac{|\{\text{Not SL}\} \cap \{\text{Systems classified as SL}\}|}{|\{\text{Not SL}\}|}. \quad (5)$$

The precision is a measurement of how pure the sample is, i.e. the percentage of the elements classified as a given class that are correct. The recall represents how complete is the classified sample or how many elements of a class were correctly identified. The false alarm rate is the percentage of fake detections. The output of the neural network pipeline is a number associated with the probability of a given object being a strong lensing system. Therefore, to obtain the aforementioned metrics is then necessary to define a probability threshold t : varying this threshold in the range $[0, 1]$, one can obtain a curve of P and R and other for R and F , with the latter known as receiver operating characteristic (ROC) curve. This is a typical process to assess the quality of a given classification algorithm (see e.g. Metcalf et al. 2019; Cheng et al. 2020a; Bom et al. 2021; Fraga et al. 2021; Magro et al. 2021), and can also be used to define the best threshold balancing the precision, recall, or false alarm rate. Additionally, the area under curve (AUC) of the ROC is also an intuitive quantity to evaluate the classification performance: a perfect classifier would have $\text{AUC} = 1$. In contrast, for a random choice classifier, we would expect $\text{AUC} = 0.5$. Analogously, the area under the P and R curve can also be used as a quality metric.

We summarize our results in Table 1, obtained in a validation sample at each iteration of the cross-validation. There we show the mean AUC for the ROC and precision–recall (PR) curves with the error corresponding to one standard deviation when considering all

trainings. We also show the mean and error for the F_β score, defined as

$$F_\beta = (1 + \beta^2) \frac{P \times R}{\beta^2 P + R}. \quad (6)$$

A lower β , ~ 0 will favour the precision P , while recall R dominates higher values > 1 , and $\beta = 1$ represents the harmonic mean between precision and recall. For ranking purposes in the II SGLC, $\beta^2 = 0.001$. Since they both depend on the probability threshold t chosen to separate between the classes, we take the maximum value of F_β at each fold,

$$F_\beta = \max_t F_\beta(t). \quad (7)$$

The results in Table 1 confirm that the net is unable to learn from inputs without any pre-processing, and it is only slightly better using the *HJY* bands alone. The results improve when using only the *VIS* band, with some slightly numerical advantage for *VIS* with zeros. The comparison between *HJY* and *VIS* alone suggests that increasing the resolution of the images might give better results than using colours in the current configuration of the II SGLC data set, as was previously seen in Fig. 4. The best performance was using all bands as the input, combining the higher resolution *VIS* images and colour information from the infrared bands. Interestingly, the results using only the *Y* band are similar to the ones using *H*, *J*, and *Y* together. Although visually more appealing, the alternative pre-processing did not always translate into better results.

Moreover, both ways of inputting the *VIS* band also give similar results when using it alone or when combining it with other bands. In fact, all results using a two-branched EfficientNet are compatible considering the errors, except for *HJY* + *VIS* (repeated) with the alternative pre-processing, which is only marginally better than the two-band configuration with the II SGLC pre-processing when considering 1σ errors. We highlight in red the configuration used in the submission to the II SGLC, and in blue the one with the best nominal F_β score.

We also present the results for the independent test set with the resulting 80 000 images neither used in training nor validation in Fig. 8. We used the models with the lowest and fifth lowest validation loss from the 10 multiple trainings for inference in this sample. The results agree with the ones reported in the table, apart from for the fold with best model using only *VIS*. Interestingly, we found that the best model in *VIS* with zeros has a better performance considering one standard deviation in the blind test set compared to the validation set results. However, the fifth best model is in agreement with Table 1, and the AUCs for all folds confirm that the best model using *VIS* alone is an outlier.

4.3 Strong lensing classification interpretation

Deep learning predictions are often hard to interpret and obtain intuition. Moreover, there is no straightforward standard procedure to justify the algorithm choices. One of the reasons for that is the complexity of those models, including the number of free parameters and sparsity. Therefore, several techniques were proposed to infer and interpret the outputs from a given network; among them, we chose a popular approach named local interpretable model-agnostic explanations (LIME; Ribeiro, Singh & Guestrin 2016). This is a framework that can be applied to several machine learning methods and has been applied in several machine learning pipelines (see e.g. Mishra, Sturm & Dixon 2017; Haunschmid, Manilow & Widmer 2020; Hassan et al. 2022). The method makes perturbations in small parts of the input and evaluates how the predictions change. It treats

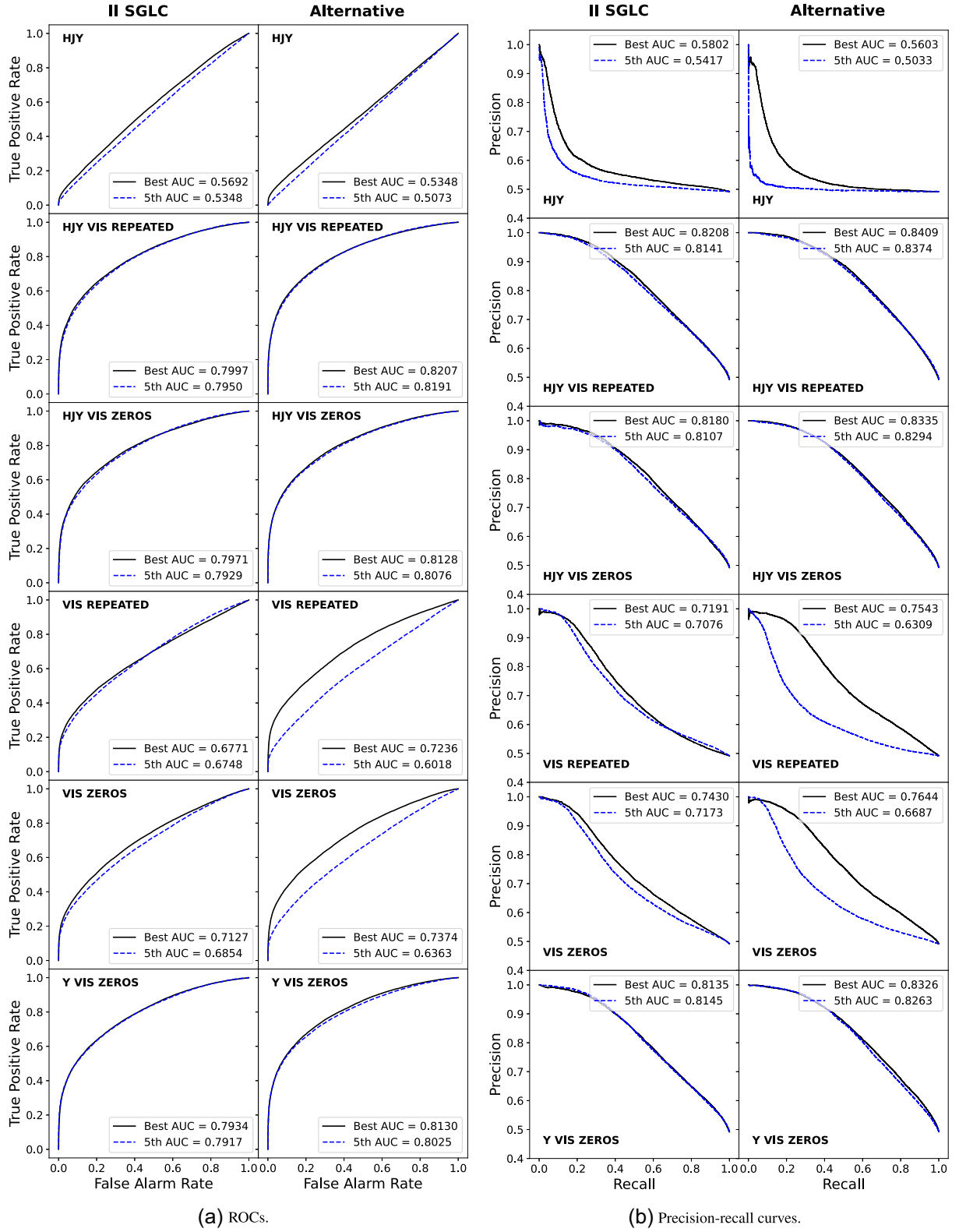


Figure 8. Results of our models on the independent test group with 80 000 objects. The best performing fold (lowest validation loss) is shown as a black solid line, and the fifth best one as a blue dashed one.

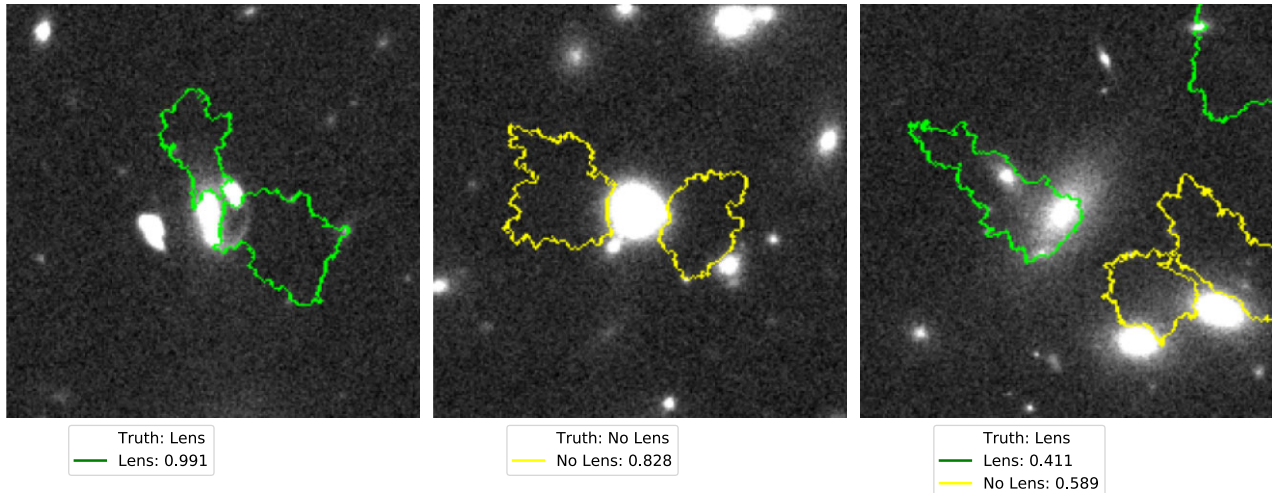


Figure 9. LIME analysis for three images in the data set. Highlighted regions correspond to the two most important superpixels when determining the classification: green for lenses and yellow for non-lenses. The predictions of the network and the ground truth are also shown.

the model as a complete black box (hence the name *model agnostic*), approximating it locally by a linear model, which is simpler to analyse compared to the global model.

For image classification, LIME does this by running the model several times, perturbing different image regions, and checking the predictions. The superpixels (group of adjacent pixels) defined by the LIME technique are then classified according to their importance to the classification.

We applied LIME to our sample images and made a visual assessment of what are the superpixels relevant for its classification. In Fig. 9, we present three example images from our data set with their respective predictions and ground truth, with the two most important superpixels highlighted. In the left-hand and middle panels, the probabilities are fairly high for one of the classes, so we show only the superpixels important to that class, while on the right-hand panel, we show the regions important for both classes since the probabilities are close. It is worth noticing that the model searched around the central galaxy for the lens and gave high probabilities for being a lens or not based on finding it, which agrees with the intuition of how a human classifier would analyse the image. However, when the field is crowded with bright objects, it searches around background galaxies for lenses, finding none. This suggests that deep learning algorithms might be misguided in crowded fields. Nevertheless, since it also searched around the central galaxy for the lens, both probabilities are comparable.

5 DETECTION LIMITS

Conceptually, a strong lensing system is defined when the light of a given source is strongly deflected by the lens. However, the strict definition of whether the system is detectable or not is not clearly defined, depending on the survey definitions, observational conditions, and instrument configurations. For instance, the simulated images can have just a few source pixels above the background level so that even if there is a signal, it might be undetectable. The II SGLC organizers gave their own definition of detectable strong lensing, which we reproduce in order to obtain a truth table of lenses and non-lenses, as presented in Section 2.1. The experience from I SGLC showed that human visual inspection is less sensitive than automated deep learning algorithms in simulations. Thus, we investigate the

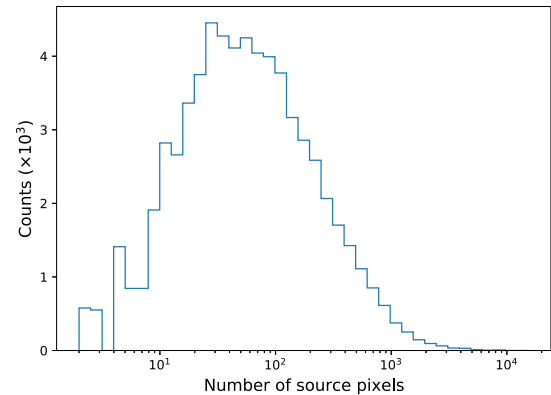


Figure 10. Distribution of number of source pixels above the background for images satisfying both other criteria for being considered a lens (effective magnification and number of groups of sources).

detection limits of our method, in particular, how the number of source pixels above 1σ of the background level, N_{pix} , which was used as a criterion to define lenses in the II SGLC, can affect our deep learning model's performance. Fig. 10 shows the distribution of N_{pix} for images that fulfil the other two criteria for lenses, all with $N_{\text{pix}} > 0$ (since, in principle, any image with at least one source pixel above the background level could be considered a strong lensing system). Even though the challenge set a minimum of 20, the median of the distribution is approximately 50, with still a relevant number of objects with $N_{\text{pix}} > 100$.

In order to assess the performance of our model in more adverse conditions, we test our trained *HJY* + VIS (repeated) models in test sets composed of lensed systems in a given N_{pix} range, fulfilling the other criteria mentioned in Section 2. This set is complemented with the non-lenses category defined, for the purposes of this test only, by an equal number of objects with N_{pix} less than the minimum value for that given range. All these objects are taken from the 80 000 objects that we used as an independent test set (see Section 4.2). We start with $N_{\text{pix}} = 10$ and go to 150 in steps of 10; for example, in the first range the lensed systems have $10 \leq N_{\text{pix}} < 20$ and non-lensed have $N_{\text{pix}} < 10$. We use the results of our 10 trained models (one for each fold in the cross-validation scheme) for each range to obtain the

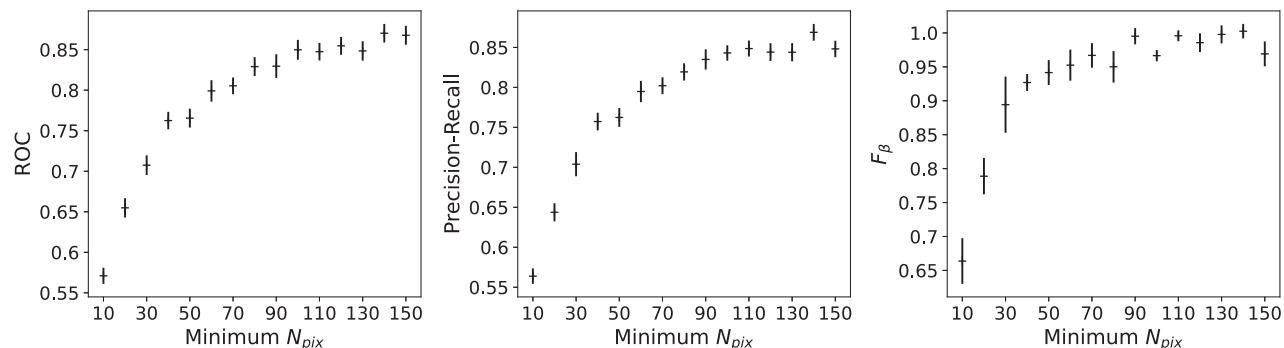


Figure 11. Metrics for different minimum values of N_{pix} . The points correspond to the mean, and the error bars to one standard deviation considering all models.

error bars. Fig. 11 shows the mean and one standard deviation ROC AUC, precision–recall AUC, and F_β considering the results from all 10 models. For $N_{\text{pix}} < 30$, the net is not much better from random guessing, while for $N_{\text{pix}} \geq 70$ we start seeing results comparable to the inference on the full sample.

6 ADAPTABILITY

Most of the deep learning algorithms are tailor-made for a specific data set. This scenario is reasonable in a context of a data competition. However, the pursuance of an adaptable algorithm can save lots of development in new data sets making the efforts to analyse new data more efficient and accessible, not requiring a great amount of time from deep learning experts, increasing the model usage capability, and relevance to the community. Deep learning models are usually data hungry, relying upon massive simulated data sets in specific surveys, which are not always available or convenient, as simulations will focus on a range of parameters and models. Additionally, adapting the method for a different scenario also validates the methodology.

The data set suitable to evaluate whether the deep learning is adaptable should be considered uniform in a given survey condition that is different from the data set used in the initial training. It also needs to be abundant enough to determine the algorithm’s performance and how many training samples are required to make a fine-tuning. Therefore, we evaluate the use of a different data set, namely, the one used in I SGLC in the current trained algorithm. Differently from the II SGLC data, based in *Euclid*, space-based conditions, the main multiband I SGLC data set was built to represent ground-based images using the ESO’s Very Large Telescope (VLT) Survey Telescope in KiDS-like data. This included the observational conditions of the KiDS survey, the level of noise, and the bands. This data set, the same used as training in I SGLC, presents a total of 20 000 systems with a truth table among lenses and non-lenses in four bands u , g , r , and i . Because of the nature of the I SGLC simulations, the data set is relevant to highlight performance in different surveys/observational conditions. However, they are simulated with the same kind of strong lensing algorithms, making it not suitable to determine the limitations of specific strong lensing modelling methods.

We start our investigation by applying the trained model in II SGLC data using $HJY + VIS$ bands in the new KiDS-like data. After some initial tests, we found that the best way to split the four bands of this data to fit our $3 + 1$ scheme was to leave the r band alone, combining u , g , and i . The images were pre-processed and

adjusted using the same kind of procedure employed in the II SGLC. However, the images had a different size in pixels to the *Euclid*-like data set used to train the models, so we resize them using a standard routine from the OpenCV library (Bradski 2000). The performance of the model when trying to make inference directly in this new data set was consistent with a random guess, i.e. AUC of ROC ~ 0.5 . In order to improve this result, we select a small number of samples in the new data and use them for training the model, leaving every weight in the net free. The cross-validation procedure was the following: first, we split the initial 20 000 systems into 10 groups of 2000. Each group will be used as an independent test. For each test group, we select a given number of images (N_{img}) from the 18 000 remaining ones for training and the same number for validation.

In Fig. 12, we present our metrics for this test using different numbers of training images. In black, we report the results using the trained model in the II SGLC; for comparison, we also show in blue the same metrics in the same number of images, now using a network trained from scratch, i.e. initialized with random weights. We see that the trained model already has an AUC of ROC above the randomness level with 40 images even though the model trained from scratch is still guessing. With 200 images, we see that the results are comparable to the ones obtained with the II SGLC data set. It can also be seen that in all cases, retraining the model gives better and more consistent results.

7 DISCUSSION AND CONCLUDING REMARKS

7.1 Summary

In this contribution, we present a pipeline and the strategy employed to obtain the highest performance result in terms of F_β score of the II SGLC. We discuss the choices in the code, architecture definition, how to work with images with different resolutions, the importance of a pre-processing evaluating it by the code performance, and visual assessment. We use the LIME technique to infer the relevant features used in the classification. We also present a prescription on how to adapt the pipeline for a different data set and the detectability limit in terms of lensed pixels. Finally, we made a public release of the code, including neural network weights. In the next paragraphs, we summarize the lessons learned.

7.2 Data visualization, pre-processing, and normalization

The sample’s visual inspection shows that just normalizing the images without a proper contrast adjustment, for instance, clipping

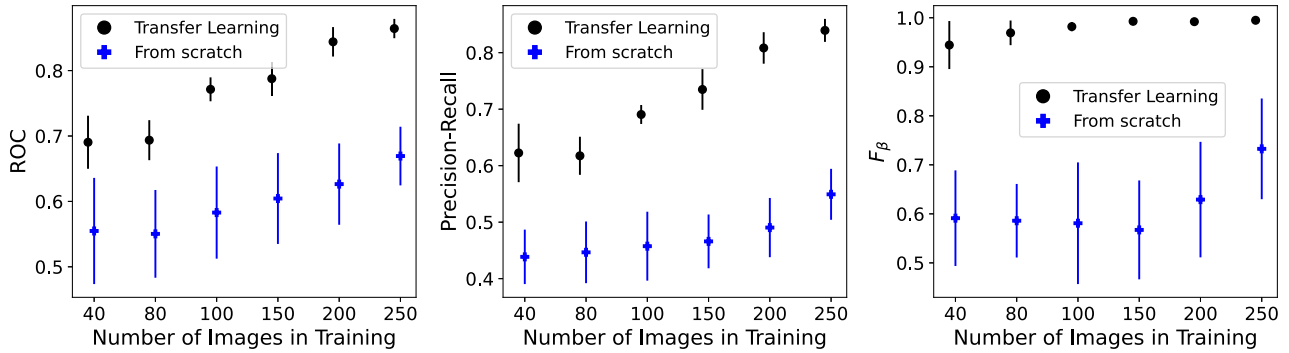


Figure 12. Metrics for our model tested on I SGLC data, retraining one of our models (transfer learning, black circles) and using random weights (from scratch, blue crosses). We use an increasing number of images for training, and report one standard deviation errors for the cross-validation. See text for details.

the histogram, makes us see no lensing feature at all. This was later confirmed also by the deep learning pipeline. However, by comparing the two pre-processing we use, even if we perceive a visual gain, this does not necessarily reflect into better performance, considering the 1σ errors. Therefore, we conclude that pre-processing the images in order to make them visually convincing as lenses is an important step. However, fine-tuning might be more critical to visual assessments than to deep learning algorithms.

7.3 Multiple bands and resolution relevance

There are multiple sources of spurious detections in strong lensing analysis, and some of them are just image artefacts. None the less, others are images that might look like a lens, as in the case of edge-on galaxies and some spiral galaxies. One popular approach to reduce this issue is to use colours and look for red galaxies with a blue object close. The colour information is considered an important feature to find lenses (see e.g. Ostrovski et al. 2018; Spiniello et al. 2018). The use of colour queries is considered at least an interesting way to make a pre-selection and improve the final deep learning result purity. In fact, even in the I SGLC results the single-band scenario found lower performance in terms of the considered metric, the AUC of ROC. The II SGLC in a *Euclid*-like scenario allowed to make deep learning classifications using images with colour information in multiple bands but also higher resolution images. Interestingly the networks using only *HJY* bands did not find a competitive fit. On the other hand, using the *VIS* band only with a higher resolution found better results than *HJY*, which is close to a random guess. Still, the *VIS* band only has inferior results if compared to the runs using all information. This result suggests that the use of high-resolution images might play an important role and was in fact, more relevant than the multiple bands in the cases tested in this contribution. It is worth mentioning that the threshold defined by the maximum F_β , with $\beta = 0.001$, privileged a pure sample instead of a complete one, obtaining in our best network, *HJY* + *VIS* with alternative pre-processing, around ~ 99 per cent purity with completeness of around 45 per cent. This choice is justified for the same reason a pre-selection of targets is implemented by colour queries and other methods to reduce the number of non-lenses in a survey where we have billions of non-lenses for thousands of lenses.

7.4 Underfitting and overfitting

During the network definition and training process, we found, on several occasions, underfitting results, i.e. training losses were not

minimized, and also overfitting, where the discrepancy between training and validation losses increases with the epochs. The underfitting was found mainly when using no pre-processing, just a simple data normalization; this result is presented in Fig. 7. It also agrees with the fact, already mentioned, that without the pre-processing no features were visible to human eyes.

None the less, we also find overfitting in many of our tests. Even the best configurations presented a tendency to overfit if we just let the training proceed unbounded. Rather than choose an early stopping method that might miss a big picture of training by quickly interrupting it to avoid overfitting, we deal with that by just choosing the epoch where the validation loss was the lowest to perform our evaluation of the model. This was an important strategy to the competition, enabling us to choose the network weights that could better generalize to the validation sample. Interestingly, when we delivered our submission to the II SGLC, the higher performance network was not the one that best performed in terms of F_β in the testing sample, but precisely the ones that generalized better in the validation sample. This suggests that, in order to adapt for this blind sample, the best strategy was to choose the one with the lowest validation loss. A more detailed discussion on the network entries is a topic to be presented in the II SGLC results paper.

Additionally to the observed tendency to overfitting, we also found unexpected overfitting in the very early epochs, i.e. from epoch 2 by using the TensorFlow native implementation of EfficientNets. This was an overfitting so early that happened before the validation loss dropped a level where we could find a classification better than random guess. After some investigation, we find that there are some differences between the TensorFlow implementation and ours, based on the original EfficientNet GitHub,² where the former incorporates into the architecture some pre-processing layers. Since our data were already fully prepared, these extra steps caused the issues mentioned.

7.5 Neural network complexity

From the computer vision challenges (Russakovsky et al. 2015) experience, the use of deeper networks with similar architectures in terms of layers and structure are usually high performing if we are able to train successfully, with no strong overfitting. This scenario changes when some innovation on the architecture is presented, like the ResNet (He et al. 2016). This effect was also presented in the EfficientNet paper. The scaled networks, with a bigger number of parameters, presented a better performance compared to the ones with

²<https://github.com/qubvel/efficientnet>

lower parameters, this effect was more evident with the initial B0–B4 than the later up to B7 where the performance gain was smaller (for further details, see fig. 1 of Tan & Le 2019). Nevertheless, these deeper networks require more computation time, high-performing hardware and the gain in performance is not always sufficient to justify this choice depending on the specific problem, particularly in a cross-validation scheme where we evaluate the variability of the results for a given train/validation sets. In our experiments with strong lensing we found this scenario with the EfficientNets with a great number of free parameters, and we choose the B2 architecture as a trade-off between computation time efficiency and performance.

7.6 Deep learning strong lensing finding decision making

There is not a general interpretation theory for deep learning decision making. In fact this is an active research topic (see e.g. Cheng et al. 2020b). Thus, we made use of LIME technique to infer what is more important to the deep learning decision-making process. This is not the only possible choice, for instance, in a recent paper by Wilde et al. (2022), the authors used a different set of techniques to perform assessment of a strong lensing finder and found that in high confidence lenses their CNN model highlights the arc or ring shapes. Our results reveal that the important regions for classification are the edges of the central galaxy, which agrees with the visual inspection intuition. We noticed that crowded regions were influenced by the neighbouring objects. This might be an important issue for lenses with multiple images but no pronounced arcs and regions close to the centre of clusters. One possibility is to work with smaller images. In principle, it is worth noticing that one could work with the same trained network but using some stamps with smaller sizes by resizing the images and with a fine-tuning. This approach was implemented successfully when we adapted the network for the I SGLC sample, where the original images had 101×101 pixels and the r band was resized to fit the VIS network branch where we had 200×200 pixels images.

7.7 Generic strong lensing finders based on deep learning models are possible?

The deep learning algorithms are usually tailor-made for a specific survey. Commonly this means enormous efforts to define a suitable high-performing architecture and relies in massive simulations to train the models, which are ideally made to represent the expected population of strong lenses. Here, we presented a simple prescription to adapt the pipeline to different survey conditions, where one can use a small amount of data. This could be used to fine-tuning by training in real data, which is usually very limited and avoid the need to produce new strong lensing finders from scratch. Depending on the specific goal of the strong lensing search, this approach has to be used carefully, as the known sample of real lenses is highly inhomogeneous and subject to complex selection functions (see e.g. Guy et al. 2022). However, this might be of particular interest to specialize the algorithm to a specific population of rare systems. The two-branched architecture presented here to deal with images of different resolutions also offers a possible path to integrate multiple data from current and future multiband surveys.

ACKNOWLEDGEMENTS

The authors would like to thank to the Bologna Lens Factory team for organizing the II SGLC and producing the full data sets used in this work. CRB acknowledges the financial support from CNPq (316072/2021-4) and FAPERJ (201.456/2022).

CF acknowledges the financial support from CNPq (processes 433615/2018-4 and 314672/2020-6). MM acknowledges financial support from CNPq and FAPERJ. The authors acknowledge the LITCOMP/COTEC/CBPF multi-GPU development team for all the support in the artificial intelligence infrastructure and Sci-Mind's High-Performance multi-GPU system.

DATA AVAILABILITY

We make the deep learning algorithms publicly available at https://github.com/cdebom/cast_lensfinder. The trained deep learning models are also public and can be downloaded from <https://doi.org/10.5281/zenodo.6344064>. The image data sets developed in the context of the I and II SGLC are available at http://metcalf1.difa.unibo.it/blf-portal/gg_challenge.html

REFERENCES

- Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Available at <https://www.tensorflow.org/>
- Abdelsalam H. M., Saha P., Williams L. L. R., 1998, *MNRAS*, 294, 734
- Akhshik M. et al., 2020, *ApJ*, 900, 184
- Avestruz C., Li N., Zhu H., Lightman M., Collett T. E., Luo W., 2019, *ApJ*, 877, 58
- Bartelmann M., Huss A., Colberg J. M., Jenkins A., Pearce F. R., 1998, *A&A*, 330, 1
- Bayer D., Chatterjee S., Koopmans L. V. E., Vegetti S., McKean J. P., Treu T., Fassnacht C. D., 2018, preprint ([arXiv:1803.05952](https://arxiv.org/abs/1803.05952))
- Bayliss M. B., 2012, *ApJ*, 744, 156
- Belokurov V., Evans N. W., Hewett P. C., Moiseev A., McMahon R. G., Sanchez S. F., King L. J., 2009, *MNRAS*, 392, 104
- Bom C. R., Makler M., Albuquerque M. P., Brandt C. H., 2017, *A&A*, 597, A135
- Bom C., Poh J., Nord B., Blanco-Valentin M., Dias L., 2019, preprint ([arXiv:1911.06341](https://arxiv.org/abs/1911.06341))
- Bom C. R. et al., 2021, *MNRAS*, 507, 1937
- Bradski G., 2000, Dr. Dobb's J. Softw. Tools, 120, 122 (The OpenCV Library)
- Cabanac R. A. et al., 2007, *A&A*, 461, 813
- Carrasco E. R. et al., 2010, *ApJ*, 715, L160
- Cheng T.-Y., Li N., Conselice C. J., Aragón-Salamanca A., Dye S., Metcalf R. B., 2020a, *MNRAS*, 494, 3750
- Cheng K., Wang N., Shi W., Zhan Y., 2020b, *J. Comput. Res. Development*, 57, 1208
- Coe D., Benítez N., Broadhurst T., Moustakas L. A., 2010, *ApJ*, 723, 1678
- Collett T. E., 2015, *ApJ*, 811, 20
- Cooray A. R., 1999, *A&A*, 341, 653
- de Bom C. R., Furlanetto C., More A., Brandt C., Makler M., Santiago B., 2015, in Rosquist K., Jantzen R. T., Ruffini R., eds, Thirteenth Marcel Grossmann Meeting: On Recent Developments in Theoretical and Experimental General Relativity, Astrophysics and Relativistic Field Theories. World Scientific Press, Singapore, p. 2088
- Despali G., Vegetti S., White S. D. M., Giocoli C., van den Bosch F. C., 2018, *MNRAS*, 475, 5424
- Diehl H. T. et al., 2017, *ApJS*, 232, 15
- Ebeling H., Stockmann M., Richard J., Zabl J., Brammer G., Toft S., Man A., 2018, *ApJ*, 852, L7
- Enander J., Mörtzell E., 2013, *J. High Energy Phys.*, 2013, 31
- Estrada J. et al., 2007, *ApJ*, 660, 1176
- Fassnacht C. D., Moustakas L. A., Casertano S., Ferguson H. C., Lucas R. A., Park Y., 2004, *ApJ*, 600, L155
- Fraga B. M. O., Barres de Almeida U., Bom C. R., Brandt C. H., Giommi P., Schubert P., de Albuquerque M. P., 2021, *MNRAS*, 505, 1268
- Gavazzi R., Marshall P. J., Treu T., Sonnenfeld A., 2014, *ApJ*, 785, 144
- Gilman D., Birrer S., Treu T., Keeton C. R., Nierenberg A., 2018, *MNRAS*, 481, 819

- Gladders M. D., Hoekstra H., Yee H. K. C., Hall P. B., Barrientos L. F., 2003, *ApJ*, 593, 48
- Glazebrook K., Jacobs C., Collett T., More A., McCarthy C., 2017, *MNRAS*, 471, 167
- Goodfellow I., Bengio Y., Courville A., 2016, *Deep Learning*. MIT Press, Cambridge, MA
- Green J. et al., 2012, preprint ([arXiv:1208.4012](https://arxiv.org/abs/1208.4012))
- Grillo C., Rosati P., Suyu S. H., Caminha G. B., Mercurio A., Halkola A., 2020, *ApJ*, 898, 87
- Guo Q. et al., 2011, *MNRAS*, 413, 101
- Guy L. P. et al., 2022, preprint ([arXiv:2201.03862](https://arxiv.org/abs/2201.03862))
- Hassan M. R., Islam M. F., Uddin M. Z., Ghoshal G., Hassan M. M., Huda S., Fortino G., 2022, *Future Generation Comput. Syst.*, 127, 462
- Haunschmid V., Manilow E., Widmer G., 2020, preprint ([arXiv:2009.02051](https://arxiv.org/abs/2009.02051))
- He K., Zhang X., Ren S., Sun J., 2016, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Piscataway, NJ, p. 770
- Hezaveh Y., Dalal N., Holder G., Kisner T., Kuhlen M., Perreault Levasseur L., 2016, *J. Cosmol. Astropart. Phys.*, 2016, 048
- Hezaveh Y. D., Levasseur L. P., Marshall P. J., 2017, *Nature*, 548, 555
- Hogg D. W., Blandford R., Kundic T., Fassnacht C. D., Malhotra S., 1996, *ApJ*, 467, L73
- Ivezić Ž. et al., 2019, *ApJ*, 873, 111
- Jacobs C. et al., 2019, *MNRAS*, 484, 5330
- Jones T. A., Swinbank A. M., Ellis R. S., Richard J., Stark D. P., 2010, *MNRAS*, 404, 1247
- Joseph R. et al., 2014, *A&A*, 566, A63
- Jullo E., Natarajan P., Kneib J.-P., D'Aloisio A., Limousin M., Richard J., Schmid C., 2010, *Science*, 329, 924
- Koopmans L. V. E., Treu T., Bolton A. S., Burles S., Moustakas L. A., 2006, *ApJ*, 649, 599
- Kovner I., 1989, *ApJ*, 337, 621
- Kubo J. M., Dell'Antonio I. P., 2008, *MNRAS*, 385, 918
- Kubo J. M. et al., 2010, *ApJ*, 724, L137
- Lanusse F., Ma Q., Li N., Collett T. E., Li C.-L., Ravanbakhsh S., Mandelbaum R., Póczos B., 2018, *MNRAS*, 473, 3895
- Laureijs R. et al., 2011, preprint ([arXiv:1110.3193](https://arxiv.org/abs/1110.3193))
- Legin R., Hezaveh Y., Perreault Levasseur L., Wandelt B., 2021, preprint ([arXiv:2112.05278](https://arxiv.org/abs/2112.05278))
- Liu L., Jiang H., He P., Chen W., Liu X., Gao J., Han J., 2020, in *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*. Addis Ababa, Ethiopia
- McCully C., Keeton C. R., Wong K. C., Zabludoff A. I., 2017, *ApJ*, 836, 141
- Magro D., Zarb Adami K., DeMarco A., Riggi S., Sciacca E., 2021, *MNRAS*, 505, 6155
- Man A. W. S. et al., 2021, *ApJ*, 919, 20
- Marshall P. J. et al., 2007, *ApJ*, 671, 1196
- Maturi M., Mizera S., Seidel G., 2014, *A&A*, 567, A111
- Meneghetti M., Dolag K., Tormen G., Bartelmann M., Moscardini L., Perrotta F., Baccigalupi C., 2004, *Mod. Phys. Lett. A*, 19, 1083
- Meneghetti M., Rasia E., Merten J., Bellagamba F., Ettori S., Mazzotta P., Dolag K., Marri S., 2010, *A&A*, 514, A93
- Metcalfe R. B., Petkova M., 2014, *MNRAS*, 445, 1942
- Metcalfe R. B. et al., 2019, *A&A*, 625, A119
- Mishra S., Sturm B. L., Dixon S., 2017, in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*. ISMIR, Suzhou, China, p. 537
- More A., Cabanac R., More S., Alard C., Limousin M., Kneib J.-P., Gavazzi R., Motta V., 2012, *ApJ*, 749, 38
- More A. et al., 2016, *MNRAS*, 455, 1191
- Morgan R. et al., 2022, *ApJ*, 927, 109
- Natarajan P., De Lucia G., Springel V., 2007, *MNRAS*, 376, 180
- Nord B. et al., 2016, *ApJ*, 827, 51
- Oguri M., 2007, *ApJ*, 660, 1
- Ostrovski F. et al., 2018, *MNRAS*, 473, L116
- Overzier R., Lemson G., Angulo R. E., Bertin E., Blaizot J., Henriques B. M. B., Marleau G. D., White S. D. M., 2013, *MNRAS*, 428, 778
- Paraficz D. et al., 2016, *A&A*, 592, A75
- Pawase R. S., Courbin F., Faure C., Kokotanekova R., Meylan G., 2014, *MNRAS*, 439, 3392
- Pearson J., Maresca J., Li N., Dye S., 2021, *MNRAS*, 505, 4362
- Petkova M., Metcalf R. B., Giocoli C., 2014, *MNRAS*, 445, 1954
- Petrillo C. E. et al., 2017, *MNRAS*, 472, 1129
- Petrillo C. E. et al., 2019a, *MNRAS*, 482, 807
- Petrillo C. E. et al., 2019b, *MNRAS*, 484, 3879
- Pizzuti L. et al., 2017, *J. Cosmol. Astropart. Phys.*, 2017, 023
- Poindexter S., Morgan N., Kochanek C. S., 2008, *ApJ*, 673, 34
- Ratnatunga K. U., Griffiths R. E., Ostrander E. J., 1999, *AJ*, 117, 2010
- Ribeiro M. T., Singh S., Guestrin C., 2016, in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery (ACM), New York, p. 1135
- Richard J., Jones T., Ellis R., Stark D. P., Livermore R., Swinbank M., 2011, *MNRAS*, 413, 643
- Russakovsky O. et al., 2015, *Int. J. Comput. Vision*, 115, 211
- Schuld S., Suyu S., Meinhardt T., Leal-Taixé L., Cañameras R., Taubenberger S., Halkola A., 2021, *A&A*, 646, A126
- Schwab J., Bolton A. S., Rappaport S. A., 2010, *ApJ*, 708, 750
- Spiniello C. et al., 2018, *MNRAS*, 480, 1163
- Suyu S. H., Marshall P. J., Auger M. W., Hilbert S., Blandford R. D., Koopmans L. V. E., Fassnacht C. D., Treu T., 2010, *ApJ*, 711, 201
- Tan M., Le Q., 2019, in *Chaudhuri K., Salakhutdinov R., eds, Proceedings of the 36th International Conference on Machine Learning*. PMLR, New York, NY, p. 6105
- Tan M., Chen B., Pang R., Vasudevan V., Sandler M., Howard A., Le Q. V., 2019, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Piscataway, NJ, p. 2820
- Treu T., Koopmans L. V. E., 2002a, *ApJ*, 575, 87
- Treu T., Koopmans L. V. E., 2002b, *MNRAS*, 337, L6
- Vegetti S., Lagattuta D. J., McKean J. P., Auger M. W., Fassnacht C. D., Koopmans L. V. E., 2012, *Nature*, 481, 341
- Walmsley M. et al., 2022, *MNRAS*, 509, 3966
- Wen Z.-L., Han J.-L., Jiang Y.-Y., 2011, *Res. Astron. Astrophys.*, 11, 1185
- Wilde J., Serjeant S., Bromley J. M., Dickinson H., Koopmans L. V. E., Metcalf R. B., 2022, *MNRAS*, 512, 3464
- Wong K. C. et al., 2020, *MNRAS*, 498, 1420
- Yamamoto K., Kadoya Y., Murata T., Futamase T., 2001, *Progress Theor. Phys.*, 106, 917
- Zackrisson E., Riehm T., 2010, *Adv. Astron.*, 2010, 478910

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.