Euclid preparation

XXXIII. Characterization of convolutional neural networks for the identification of galaxy-galaxy strong-lensing events

Euclid Collaboration: L. Leuzzi[1,2], M. Meneghetti[2,3], G. Angora[4,5], R. B. Metcalf[1], L. Moscardini[1,2,3], P. Rosati[4,2], P. Bergamini[6,2], F. Calura[2], B. Clément[7], R. Gavazzi[8,9], F. Gentile[10,2], M. Lochner[11,12], C. Grillo[6,13], G. Vernardos[14], N. Aghanim[15], A. Amara[16], L. Amendola[17], N. Auricchio[2], C. Bodendorf[18], D. Bonino[19], E. Branchini[20,21], M. Brescia[22,5], J. Brinchmann[23], S. Camera[24,25,19], V. Capobianco[19], C. Carbone[13], J. Carretero[26,27], M. Castellano[28], S. Cavuoti[5,29], A. Cimatti[30], R. Cledassou[31,32], G. Congedo[33], C. J. Conselice[34], L. Conversi[35,36], Y. Copin[37], L. Corcione[19], F. Courbin[7], M. Cropper[38], A. Da Silva[39,40], H. Degaudenzi[41], J. Dinis[40,39], F. Dubath[41], X. Dupac[36], S. Dusini[42], S. Farrens[43], S. Ferriol[37], M. Frailis[44], E. Franceschi[2], M. Fumana[13], S. Galeotta[44], B. Gillis[33], C. Giocoli[2,3], A. Grazian[45], F. Grupp[18,46], L. Guzzo[6,47,48], S. V. H. Haugan[49], W. Holmes[50], F. Hormuth[51], A. Hornstrup[52,53], P. Hudelot[9], K. Jahnke[54], M. Kümmel[55], S. Kermiche[56], A. Kiessling[50], T. Kitching[38], M. Kunz[57], H. Kurki-Suonio[58,59], P. B. Lilje[49], I. Lloro[60], E. Maiorano[2], O. Mansutti[44], O. Marggraf[61], K. Markovic[50], F. Marulli[1,2,3], R. Massey[62], E. Medinaceli[2], S. Mei[63], M. Melchior[64], Y. Mellier[65,9,66], E. Merlin[28], G. Meylan[7], M. Moresco[1,2], E. Munari[44], S.-M. Niemi[67], J. W. Nightingale[62], T. Nutma[68,69], C. Padilla[26], S. Paltani[41], F. Pasian[44], K. Pedersen[70], V. Pettorino[71], S. Pires[43], G. Polenta[72], M. Poncet[31], F. Raison[18], A. Renzi[73,42], J. Rhodes[50], G. Riccio[5], E. Romelli[44], M. Roncarelli[2], E. Rossetti[10], R. Saglia[55,18], D. Sapone[74], B. Sartoris[55,44], P. Schneider[61], A. Secroun[56], G. Seidel[54], S. Serrano[75,76], C. Sirignano[73,42], G. Sirri[3], L. Stanco[42], P. Tallada-Crespí[77,27], A. N. Taylor[33], I. Tereno[39,78], R. Toledo-Moreo[79], F. Torradeflot[27,77], I. Tutusaus[80], L. Valenziano[2,81], T. Vassallo[44], Y. Wang[82], J. Weller[55,18], G. Zamorani[2], J. Zoubian[56], S. Andreon[47], S. Bardelli[2], A. Boucaud[63], E. Bozzo[41], C. Colodro-Conde[83], D. Di Ferdinando[3], M. Farina[84], R. Farinelli[2], J. Graciá-Carpio[18], E. Keihänen[85], V. Lindholm[58,59], D. Maino[6,13,48], N. Mauri[30,3], C. Neissner[26,27], M. Schirmer[54], V. Scottez[65,86], M. Tenti[81], A. Tramacere[41], A. Veropalumbo[47], E. Zucca[2], Y. Akrami[87,88,89,90,91], V. Allevato[5,92], C. Baccigalupi[93,94,44,95], M. Ballardini[4,96,2], F. Bernardeau[97,9], A. Biviano[44,94], S. Borgani[44,98,95,94], A. S. Borlaff[99,100], H. Bretonnière[101], C. Burigana[102,81], R. Cabanac[80], A. Cappi[2,103], C. S. Carvalho[78], S. Casas[104], G. Castignani[1,2], T. Castro[44,95,94], K. C. Chambers[105], A. R. Cooray[106], J. Coupon[41], H. M. Courtois[107], S. Davini[21], S. de la Torre[8], G. De Lucia[44], G. Desprez[108], S. Di Domizio[109], H. Dole[15], J. A. Escartin Vigo[18], S. Escoffier[56], I. Ferrero[49], L. Gabarra[73,42], K. Ganga[63], J. Garcia-Bellido[87], E. Gaztanaga[110,75,16], K. George[46], G. Gozaliasl[58,111], H. Hildebrandt[112], I. Hook[113], M. Huertas-Company[114,83,115,116], B. Joachimi[117], J. J. E. Kajava[118], V. Kansal[119], C. C. Kirkpatrick[85], L. Legrand[57], A. Loureiro[120,33,91], M. Magliocchetti[84], G. Mainetti[121], R. Maoli[122,28], M. Martinelli[28,123], N. Martinet[8], C. J. A. P. Martins[124,23], S. Matthew[33], L. Maurin[15], P. Monaco[98,44,95,94], G. Morgante[2], S. Nadathur[16], A. A. Nucita[125,126,127], L. Patrizii[3], V. Popa[128], C. Porciani[61], D. Potter[129], M. Pöntinen[58], P. Reimberg[65], A. G. Sánchez[18], Z. Sakr[130,80,131], A. Schneider[129], M. Sereno[2,3], P. Simon[61], A. Spurio Mancini[38], J. Stadel[129], J. Steinwagner[18], R. Teyssier[132], J. Valiviita[58,59], M. Viel[93,94,44,95], I. A. Zinchenko[55], and H. Domínguez Sánchez[133]

(Affiliations can be found after the references)

ABSTRACT

Forthcoming imaging surveys will increase the number of known galaxy-scale strong lenses by several orders of magnitude. For this to happen, images of billions of galaxies will have to be inspected to identify potential candidates. In this context, deep-learning techniques are particularly suitable for finding patterns in large data sets, and convolutional neural networks (CNNs) in particular can efficiently process large volumes of images. We assess and compare the performance of three network architectures in the classification of

strong-lensing systems on the basis of their morphological characteristics. In particular, we implemented a classical CNN architecture, an inception network, and a residual network. We trained and tested our networks on different subsamples of a data set of 40 000 mock images whose characteristics were similar to those expected in the wide survey planned with the ESA mission *Euclid*, gradually including larger fractions of faint lenses. We also evaluated the importance of adding information about the color difference between the lens and source galaxies by repeating the same training on single- and multiband images. Our models find samples of clear lenses with ≳90% precision and completeness. Nevertheless, when lenses with fainter arcs are included in the training set, the performance of the three models deteriorates with accuracy values of ∼0.87 to ∼0.75, depending on the model. Specifically, the classical CNN and the inception network perform similarly in most of our tests, while the residual network generally produces worse results. Our analysis focuses on the application of CNNs to high-resolution space-like images, such as those that the *Euclid* telescope will deliver. Moreover, we investigated the optimal training strategy for this specific survey to fully exploit the scientific potential of the upcoming observations. We suggest that training the networks separately on lenses with different morphology might be needed to identify the faint arcs. We also tested the relevance of the color information for the detection of these systems, and we find that it does not yield a significant improvement. The accuracy ranges from ∼0.89 to ∼0.78 for the different models. The reason might be that the resolution of the *Euclid* telescope in the infrared bands is lower than that of the images in the visual band.

Keywords: gravitational lensing: strong – methods: statistical – methods: data analysis – surveys

1. Introduction

Galaxy-galaxy strong-lensing (GGSL) events occur when a foreground galaxy substantially deflects the light emitted by a background galaxy. When the observer, the lens, and the source are nearly aligned and their mutual distances are favorable, the background galaxy appears as a set of multiple images surrounding the lens. These images often have the form of extended arcs or rings.

These events have multiple astrophysical and cosmological applications. For example, GGSL enables us to probe the total mass of the lens galaxies within the so-called Einstein radius (e.g., Treu & Koopmans 2004; Gavazzi et al. 2012; Nightingale et al. 2019). By independently measuring the stellar mass and combining lensing with other probes of the gravitational potential of the lens (e.g., stellar kinematics), we can distinguish the contributions from dark and baryonic mass and thus study the interplay between these two mass components (e.g., Barnabè et al. 2011; Suyu et al. 2012; Schuldt et al. 2019). Accurately measuring the dark matter mass profiles and the substructure content of galaxies also enables us to test the predictions of the standard cold dark matter (CDM) model of structure formation and to shed light on the nature of dark matter (e.g., Grillo 2012; Oguri et al. 2014; Vegetti et al. 2018; Minor et al. 2021). Finally, the lensing magnification makes it possible to study very faint and high-redshift sources that would be not observable in the absence of the lensing effects (e.g., Impellizzeri et al. 2008; Allison et al. 2017; Stacey et al. 2018).

The high-mass density in the central regions of galaxy clusters boosts the strong-lensing cross section of individual galaxies (Desprez et al. 2018; Angora et al. 2020). Thus, the probability for GGSL is particularly high in cluster fields. Meneghetti et al. (2020) suggested that the frequency of GGSL events is a powerful tool for a stress-test of the CDM paradigm (see also Meneghetti et al. 2022; Ragagnin et al. 2022). Modeling these lensing events helps constraining the cluster mass distribution on the scale of cluster galaxies (e.g., Tu et al. 2008; Grillo et al. 2014; Jauzac et al. 2021; Bergamini et al. 2021).

Fewer than 1000 galaxy-scale lenses have been confirmed so far. They have been discovered, along with more candidates, by employing a variety of methods, including searches for unexpected emission lines in the spectra of elliptical galaxies (Bolton et al. 2006), sources with anomalously high fluxes at submillimeter wavelengths (Negrello et al. 2010, 2017), and sources with unusual shapes (Myers et al. 2003). Some arc and ring finders have been developed to analyze optical images, and they typically search for blue features around red galaxies (e.g.,

Cabanac et al. 2007; Seidel & Bartelmann 2007; Gavazzi et al. 2014; Maturi et al. 2014; Sonnenfeld et al. 2018). Assembling extensive catalogs of GGSL systems is arduous because these systems are rare, but this is expected to change in the next decade through upcoming imaging surveys. It has been estimated that the ESA *Euclid* space telescope (Laureijs et al. 2011) and the Legacy Survey of Space and Time (LSST; LSST Science Collaboration 2009) performed with the *Vera C. Rubin* Observatory will observe more than 100 000 strong lenses (Collett 2015), which will significantly increase the number of known systems. Producing large and homogeneous catalogs of GGSL systems like this will be possible because of the significant improvements in spatial resolution, area, and seeing of these surveys compared to previous observations.

Identifying potential candidates will require the examination of hundreds of millions of galaxies; thus, developing reliable methods for analyzing large volumes of data is of fundamental importance. Over the past few years, machine-learning (ML), and specifically, deep-learning (DL), techniques have proven extremely promising in this context. We focus on supervised ML techniques. These automated methods learn to perform a given task in three steps. In the first step, the training, they analyze many labeled examples and extract relevant features from the data. In the second step, the validation, the networks are validated on labeled data whose labels they cannot access to ensure that the learning does not lead to overfitting. The validation occurs at the same time as the training and is used to guide it. In the third step, the architectures are tested on more labeled data that were not used in the previous phases, whose labels are unknown to the models, but that are used to evaluate their performance.

In particular, convolutional neural networks (CNNs; e.g., LeCun et al. 1989) are a DL algorithm that has been successfully applied to several astrophysical problems and is expected to play a key role in the future of astronomical data analysis. Among the many different applications, they have been employed to estimate the photometric redshifts of luminous sources (e.g., Pasquet et al. 2019; Shuntov et al. 2020; Li et al. 2022), to perform the morphological classification of galaxies (e.g., Huertas-Company et al. 2015; Domínguez Sánchez et al. 2018; Zhu et al. 2019; Ghosh et al. 2020), to constrain the cosmological parameters (e.g., Merten et al. 2019; Fluri et al. 2019; Pan et al. 2020), to identify cluster members (e.g., Angora et al. 2020), to find galaxy-scale strong lenses in galaxy clusters (e.g., Angora et al. 2023), to quantify galaxy metallicities (e.g., Wu & Boada 2019; Liew-Cain et al. 2021), and to estimate the dynamical masses of galaxy clusters (e.g., Ho et al. 2019; Gupta & Reichardt 2020).

Recently, O'Riordan et al. (2023) also tested whether CNNs can be used to detect subhalos in simulated *Euclid*-like galaxy-scale strong lenses.

Several CNN architectures were also used recently to identify strong lenses in ground-based wide-field surveys such as the Kilo Degree Survey (KiDS; de Jong et al. 2015; Petrillo et al. 2017, 2019; He et al. 2020; Li et al. 2020, 2021; Napolitano et al. 2020), the Canada–France–Hawaii Telescope Legacy Survey (CFHTLS; Gwyn 2012; Jacobs et al. 2017), the Canada France Imaging Survey (CFIS; Savary et al. 2022), the Hyper Suprime-Cam Subaru Strategic Program Survey (HSC; Aihara et al. 2018; Cañameras et al. 2021; Wong et al. 2022), and the Dark Energy Survey (DES; Dark Energy Survey Collaboration 2005; Jacobs et al. 2019a,b; Rojas et al. 2022). Most of them were also employed in two challenges aimed at comparing and quantifying the performance of several methods to find lenses, either based on artificial intelligence or working without it. The first challenge results, presented in Metcalf et al. (2019), showed that DL methods are particularly promising with respect to other traditional techniques, such as visual inspection and classical arcfinders.

In this work, we investigate the ability of three different network architectures to identify GGSL systems. We test them on different subsamples of a data set of *Euclid*-like mock observations. In particular, we evaluate the effect of including faint lenses in the training set on the classification.

This paper is organized as follows: in Sect. 2 we explain how CNNs are implemented and trained to be applied to image-recognition problems, in Sect. 3 we introduce the data set of simulated images used for training and testing our networks, and in Sect. 4 we describe our experiments and present and discuss our results. In Sect. 5, we summarize our conclusions.

2. Convolutional neural networks

Artificial neural networks (ANNs; e.g., McCulloch & Pitts 1943; Goodfellow et al. 2016) are a ML algorithm inspired by the biological functioning of the human brain. They consist of artificial neurons, or nodes, that are organized in consecutive layers and linked together through weighted connections. The weights define the sensitivity among individual nodes (Hebb 1949) and are adapted to enable the network to carry out a specific task.

The output of the kth layer $\boldsymbol{h}^k$ depends on the output of the previous layer $\boldsymbol{h}^{k-1}$ (Bengio 2009)

$$\boldsymbol{h}^k = f(\boldsymbol{b}^k + \mathsf{W}^k \boldsymbol{h}^{k-1}). \tag{1}$$

Here, $\boldsymbol{b}^k$ is the vector of offsets (biases), and W^k is the weight matrix associated with the layer. The dimension of $\boldsymbol{b}^k$ and W^k corresponds to the number of nodes within the layer, and the symbol f represents the activation function, which introduces nonlinearity in the network that would otherwise only be characterized by linear operations.

The CNNs are a special class of ANNs that use the convolution operation. Through this property, they perform particularly well on pattern recognition tasks. The basic structure of a CNN can be described as a sequence of convolutional and pooling layers, followed by fully connected layers. Convolutional layers consist of a series of filters, also called kernels, which are matrices of weights with a typical dimension of 3×3 to 7×7 and act as the weights of a generic ANN. They are convolved with the layer input to produce the feature maps. The feature maps are passed through an activation function that introduces nonlinearity in the network, and they are then fed as input to the subsequent layer.

In our networks, we use the leaky rectified linear unit (Leaky ReLU; Xu et al. 2015) as the activation function. The organization of the filters in multiple layers ensures that the CNN can infer complex mappings between the inputs and outputs by dividing them into simpler functions, each extracting relevant features from the images. The pooling operation downsamples each feature map by dividing it into quadrants with a typical dimension of 2×2 or 3×3 and substituting them with a summary statistic, such as the maximum (Zhou & Chellappa 1988). This operation has the twofold purpose of reducing the size of the feature maps and therefore the number of parameters of the model, and making the architecture invariant to small modifications of the input (Goodfellow et al. 2016).

After these layers, the feature maps are flattened into a 1D vector that is processed by fully connected layers and is then passed to the output layer that predicts the output. In classification problems, the activation function used for the output layer is often the softmax, providing an output in the range $[0, 1]$ that can be interpreted (Bengio 2009) as an indicator of $P(Y = i \mid \boldsymbol{x})$, where Y is the class associated with the input $\boldsymbol{x}$ of all the possible classes i.

The CNNs master the execution of a given task due to a supervised learning process, called training, in which they analyze thousands of known input-output pairs. The weights of the network, which are randomly initialized, are readjusted so that the output predictions of the network are correct for the largest number of possible examples. This step is crucial because the weights are not modified afterward when the final model is applied to other data. The training aims to minimize a loss (or cost) function that estimates the difference between the outputs predicted by the network and the true labels. To do this, the images are passed to the network several times, and at the end of each pass, called epoch, the gradient of the cost function is computed with respect to the weights and is backpropagated (Rumelhart et al. 1986) from the output to the input layer so that the kernels can be adapted accordingly. The magnitude of the variation of the weights is regulated through the learning rate, a hyperparameter that is to be defined at the beginning of the training, whose specific value is fine-tuned by testing different values to find the one that minimizes the loss function.

In addition to showing good performance on the training set, it is essential that the network generalizes to other images. Preventing the model from overfitting (i.e., memorizing peculiar characteristics of the images in the training set that cannot be used to make correct predictions on other data sets) is possible by monitoring the training with a validation step. At the end of each epoch, the network performance is assessed on the validation set, which is a small part of the data set (usually 5–10%) that was excluded from the training set. If the loss function evaluated on these images does not improve for several consecutive epochs, the training should be interrupted or the learning rate reduced. Dropout (Srivastava et al. 2014) is another technique that is used to mitigate overfitting. This method consists of randomly dropping units from the network during training, that is, temporarily removing incoming and outcoming connections from a given node. When the training is completed, the performance of the final model is evaluated on the test set, which is a part of the data set (about 20–25%) that was excluded from the other subsets. The CNN can then be applied to new images.

The CNNs conveniently handle large data sets for several reasons. While the training can take up to a few days to be completed, processing a single image afterward requires a fraction of a second through graphics processing units (GPUs). Moreover, the feature-extraction process during the training is

Table 1. Main characteristics of the *Euclid* VIS and NISP (Euclid Collaboration 2022a) instruments.

Instrument	Capability	λ range (nm)	Pixel size (arcsec)
VIS	Visual imaging	I_E (530–920)	0.1
NISP	NIR imaging	Y_E (949.6–1212.3),	0.3
	photometry	J_E (1167.6–1567.0),	0.3
		H_E (1521.5–2021.4)	0.3

completely automated. The algorithm selects the most significant characteristics for achieving the best results without any previous knowledge of the data. The following subsections provide more information about the specific architectures we test in this work and technical details about our training.

2.1. Network architectures

We implemented three CNN architectures: a visual geometry group-like network (VGG-like network; Simonyan & Zisserman 2015), an inception network (IncNet; Szegedy et al. 2015, 2016), and a residual network (ResNet; He et al. 2016; Xie et al. 2017). The definition of the final configuration of the networks that we applied to the images is the result of several trials in which we tested different hyperparameters for the optimization (e.g. the learning rate) and general architectures (e.g., the number of layers and kernels) to find the most suitable arrangement for our classification problem.

2.1.1. VGG-like network

The visual geometry group network (VGGNet) was first presented by Simonyan & Zisserman (2015). The most significant innovation introduced with this architecture is the application of small convolutional filters with a receptive field of 3×3, which means that the portion of the image that the filter processes at any given moment is 3×3 pixels wide. This allowed the construction of deeper models because the introduction of small filters keeps the number of trainable parameters in the CNN smaller than that of networks that use larger filters (e.g., with a dimension of 5×5 or 7×7). Because the concatenation of multiple kernels with sizes of 3×3 has the same resulting receptive field as larger filters (Szegedy et al. 2016), it is possible to analyze features of larger scales while building deeper architectures.

Our implementation of the VGGNet comprises ten convolutional layers that alternate with five max pooling layers. We define a convolutional-pooling block as two convolutional layers followed by a pooling layer. At the end of each convolutional-pooling block, we perform the batch normalization of the output of the block. Batch normalization consists of the renormalization of the layer inputs (Ioffe & Szegedy 2015) and is employed to accelerate and stabilize the training of deep networks. After five convolutional-pooling blocks, two fully connected layers of 256 nodes each alternate with dropout layers, and finally, a softmax layer as the output layer. The number of parameters for this architecture is about two million.

When training on multiband observations, we add a second branch to process the *Euclid* Near Infrared Spectrometer and Photometer (NISP; Maciaszek et al. 2022) images, passing them to the network through a second input channel. Because they are smaller than the Visual Imager (VIS; Cropper et al. 2012) images (see Table 1), this branch of the network is only

four convolutional-pooling blocks deep. The outputs of the two branches are flattened and concatenated before they are passed to the output layer. Like in the single-branch version of this architecture, we have two fully connected layers with 256 nodes each, and finally, the output layer. In this configuration, our network uses about three million parameters. Figure A.1 shows the VGG-like network configuration we tested on the VIS images (panel a) and on the multiband images (panel b).

2.1.2. Inception network

The reasons for the IncNet architecture were outlined by Szegedy et al. (2015), who applied the ideas of Lin et al. (2013) to CNNs. Trying to improve the performance of a CNN by enlarging its depth and width leads to a massive increase in the number of parameters of the model, favoring overfitting and increasing the requirements of computational resources. Szegedy et al. (2015) suggested applying filters with different sizes to the same input, making the model extract features on different scales in the same feature maps. This is implemented through the inception module. In the simplest configuration, each module applies filters of several sizes (1×1, 3×3, and 5×5) and a pooling function to the same input and concatenates their outputs, passing the result of this operation as input to the following layer. However, this implementation can be improved by applying 1×1 filters before the 3×3 and 5×5 filters. Introducing 1×1 filters has the main purpose of reducing the dimensionality of the feature maps, and thus the computational cost of convolutions, while keeping their spatial information. This is possible by reducing the number of channels of the feature maps. An IncNet is a series of such modules stacked upon each other. A further improvement of the original inception module design is presented in Szegedy et al. (2016): The 5×5 filters are replaced by two 3×3 filters stacked together in order to decrease the number of parameters required by the model. This version of the inception module is used in our network implementation.

Before they are fed to the inception modules, the images are processed through two convolutional layers alternating with two max pooling layers. The network is composed of seven modules, the fifth of which is connected to an additional classifier. The outputs of the two classifiers are taken into account when computing the loss function by computing the individual losses and then taking a weighted sum of them. The intermediate output layer is weighted with weight 0.3, while the final one is weighted with weight 1.0. Dropout is performed before both output layers, while batch normalization is performed on the output of each max pooling layer. The output layers are both softmax layers. The total number of parameters that compose the model is approximately 2 million.

The configuration used to analyze the multiband images has a secondary branch with one initial convolutional layer and seven inception modules. This branch is characterized by approximately one million parameters, thus leading to a total of around three million parameters. Figure A.2 shows the IncNet configuration we tested on the VIS images (panel a) and the multiband images (panel b).

2.1.3. Residual network

He et al. (2016) introduced residual learning to make the training of deep networks more efficient. The basic idea behind the ResNets is that it is easier for a certain layer (or a few stacked layers) to infer a residual function with respect to the input rather than the complete, and more complicated, full mapping.

In practice, this is implemented using residual blocks with shortcut connections. Let x be the input of a given residual block. The input is simultaneously propagated through the layers within the block and stored without being changed, through the shortcut connection. The residual function $\mathcal{F}(x)$ that the block is expected to infer can be written as

$$\mathcal{F}(x) := \mathcal{H}(x) - x, \tag{2}$$

where $\mathcal{H}(x)$ is the function that a convolutional layer would have to learn in the absence of shortcut connections. Thus, the original function can be computed as $\mathcal{F}(x) + x$.

This architecture was later improved by Xie et al. (2017), who presented the ResNeXt architecture. The main modification introduced in this work is the ResNeXt block, which aggregates a set of transformations, and can be presented as

$$\mathcal{F}(x) = \sum_{i=1}^{C} \mathcal{T}_i(x) \tag{3}$$

and serves as the residual function in Eq. (2). Here, $\mathcal{T}_i(x)$ is an arbitrary function, and C is a hyperparameter called cardinality, which represents the size of the set of transformations to be aggregated.

In our implementation of the ResNet, we use this last ResNeXt block as the fundamental block, with the cardinality set to eight. In particular, the input is initially processed by two convolutional layers alternated with two pooling layers. The resulting feature maps are passed to four residual blocks alternated with two max pooling layers. There follows a dropout layer and finally a softmax layer. Moreover, batch normalization is performed after every max pooling layer. The NISP images are processed by a similar branch, which differs from this one in that it has only one initial convolutional layer.

The parameters of the model are circa one million in the VIS configuration and about two million in the multiband configuration, so they are significantly fewer than those of our implementations of the VGG-like network and of the IncNet. However, we tested different configurations of the ResNet when designing the network architectures, and this specific setup outperformed the others, including those that had a higher number of weights. Figure A.3 shows the ResNet configuration we applied to the VIS images (panel a) and the multiband images (panel b).

3. The data set

Training CNNs requires thousands of labeled examples. Because not enough observed galaxy-scale lenses are known to date, simulating the events is necessary for training a classifier to identify them. In some cases, it is possible to include real observations in the training set, but in our case, it is inevitable to adopt a fully simulated data set because no real images have been observed with the *Euclid* telescope yet. The realism of the simulations is essential to ensure that the evaluation of the model performance is indicative of the results we may expect from real observations.

The image simulations were used to produce all the images in the data set, that is, both the lenses and nonlenses. We generated all the images and then divided them into the two classes according to the criteria that we introduced below. The simulations used the galaxy and halo catalogs provided by the Flagship simulation (v1.10.11; Castander et al., in prep.) through the CosmoHub portal[1] (Carretero et al. 2017; Tallada et al. 2020).

---
[1] https://cosmohub.pic.es/home

We constructed the images using the following procedure. We randomly selected a trial lens galaxy from the light cone subject to a magnitude cut of 23 in the VIS band from the *Euclid* telescope, that is, the I_{E} band. After this, we randomly selected a background source from a catalog of *Hubble* Ultra Deep Field (UDF; Coe et al. 2006) sources with known redshift. We decomposed these sources into shapelets for denoising, following the procedure described in Meneghetti et al. (2008, 2010). This procedure has its limitations because in regions of high magnification, the finite resolution of the shapelets can be apparent and there can be low surface brightness ringing that is usually not visible above the noise. We investigate the potential impact of these effects on the results of this paper in Sect. 4.7. The mass of the lens is represented by a truncated singular isothermal ellipsoid (TSIE) and a Navarro, Frenk & White (NFW; Navarro et al. 1996) halo. The SIE model has been shown to fit existing GGSLs well (Gavazzi et al. 2007).

We used the GLAMER lensing code (Metcalf & Petkova 2014; Petkova et al. 2014) to perform the ray-tracing. Light rays coming from the position of the observer are shot within a $20'' \times 20''$ square centered on the lens object, with an initial resolution of $0\rlap{.}''05$, that is, twice the final resolution of the VIS instrument. We used these rays to compute the deflection angles that trace the path of the light back to the sources. The code detects any caustics in the field and provides some further refinement to characterize them. Specifically, more rays are shot in a region surrounding the caustics to constrain their position with higher resolution. If the area within the largest critical curve is larger than 0.2 arcsec² and smaller than 20 arcsec², the object is accepted as a lens of the appropriate size range.

The lensed image is constructed using the shapelet source and Sérsic profiles for the lens galaxy and any other galaxy that appears within the field. We took the parameters for the Sérsic profiles from the Flagship catalog with some randomization. While we placed the lens galaxy at the center of the cutout, the positions of the other galaxies were determined following the Flagship catalogs as well, with some randomization. In this way, the density of galaxies along the line of sight is the same as that of the Flagship simulations, but the sources have a different angular position. We placed the background source galaxy at a random point on the source plane within a circle surrounding the caustic. The radius of the circle was set to one-half of the largest separation between points in the caustic times 2.5.

A model for the point spread function (PSF) is applied to the image which initially has a resolution of 0.025 arcsec and then downsamples to 0.1 arcsec for VIS and 0.3 arcsec for the infrared bands. The VIS PSF was derived from modeling the instrument (Euclid Collaboration, in prep.). For the infrared bands, a simple Gaussian model with a width of 0.3 arcsec was used. The noise was simulated with a Gaussian random field to reproduce the noise level expected by the Euclid Wide Survey (Euclid Collaboration 2022b).

To avoid repeating a particular lens and to increase the number of images at a low computational cost, we randomized each lens. In this step, all the galaxies within a sphere centered on the primary lens are rotated randomly in three dimensions about the primary lens. The sphere radius was set to 30 arcsec at the distance of the lens. In addition, the galaxies outside this sphere but within the field of view were independently rotated about the primary in the plane of the sky. The mass associated with each galaxy is moved with the galaxy image. The position angles of each galaxy were also randomly resampled.

The final step is the classification of the images as lenses. Some of the images will have low signal-to-noise ratios in some

lensed images or are not distorted enough to be recognizable lenses.

This procedure is similar to the one used for the lens-finding challenges that was described in more detail in Metcalf et al. (2019). These simulations are currently being improved to provide more realistic representations of lens and source galaxies. This is important both for training the CNNs and for statistical studies (see Sect. 4.10). A possible improvement that would be relevant in the context of GGSL searches is a better characterization of the blending between the lens and source galaxies in the definition of `n_pix_source` by taking into consideration the fraction of light from lens and source in each pixel. Moreover, the simulations miss some instrumental effects, such as nonlinearity, charge transfer inefficiency, and a more intricate PSF model, which are included in other studies (e.g., Pires et al. 2020).

The result of these simulations are 100 000 *Euclid*-like mock images simulated in the I_E band of the VIS instrument and H_E, Y_E and J_E bands of the NISP instrument (Euclid Collaboration 2022a). The dimensions of the VIS and NISP images are 200×200 and 66×66 pixels, respectively. Given the resolution of the instruments, reported in Table 1, these correspond to $20'' \times 20''$ images.

When preparing the images for the training, we clean the data set by removing the images with sources at $z > 7$, thus leaving a catalog of 99 409 objects. We do this because there are just a few hundreds of such objects in the simulated data set and their number would not be sufficient to grant generalization after training. Moreover the sources at such high redshift are not as reliable as the others used in the simulations. The images in the data set are considered lenses if they meet the following criteria simultaneously:

$$
\begin{cases}
\texttt{n\_source\_im} > 0; \\
\texttt{mag\_eff} > 1.6; \\
\texttt{n\_pix\_source} > 20.
\end{cases} \tag{4}
$$

Here, `n_source_im` represents the number of images of the background source, `mag_eff` is the effective magnification of the source, and `n_pix_source` is the number of pixels in which the surface brightness of the source is 1σ above the background noise level. For every image, the magnification is computed as the ratio of the sum of all the pixels with a flux above the noise level in the lensed images on the image plane and the pixels of the unlensed image on the source plane. The most discriminatory parameters seems to be `n_pix_source`. The same criteria were adopted in the lens-finding challenge 2.0[2] (Metcalf et al., in prep.).

In many cases, one or more background sources are present in the nonlenses, but they are too faint or too weakly magnified to be classified as a lens, or both. For this reason, the parameters `n_pix_source` and `mag_eff` are also considered in the classification criteria (Eq. (4)). Depending on the sensitivity of the model, the classification of the images with a low signal-to-noise ratio might vary, while the clearest images should be immediately assigned to the correct category.

By using these conditions, we divided the images we simulated into 19 591 lenses and 79 816 nonlenses, thus obtaining two very unbalanced classes out of the complete data set. It is well known that unbalanced classes result in biased classification (Buda et al. 2018). For this reason, we used all the lenses for the training, and we randomly selected only a subsample of 20 000 nonlenses. As we discuss in Sect. 4.1, these numbers were

_____
[2] http://metcalf1.difa.unibo.it/blf-portal/gg_challenge.html
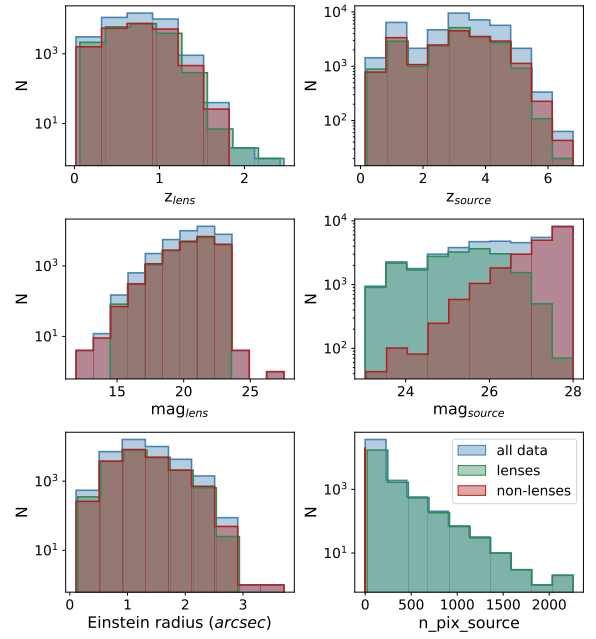
Fig. 1. Distribution of several properties of the simulated images in the data set (blue histograms) selected for training, which consisted of 40 000 mocks in total. The distributions of the same properties in the separate subsets of lenses and nonlenses are given by the green and red histograms, respectively. In the panels in the upper and middle rows, we show the distributions of lens and source redshifts and I_E band magnitudes (in the case of the sources, we refer to the intrinsic magnitude). The bottom panels show the distributions of Einstein radii of the lenses and of the number of pixels for which the source brightness exceeds 1σ above the background noise level.

increased by data augmentation. We refer to the nonlenses as class 0 and to the lenses as class 1. More strategies would be possible to deal with the unbalanced data set, such as using different weights for the two classes in the loss function or optimizing our classifiers with respect to purity, but we did not test them.

In Fig. 1, we report the distribution of some properties of the images in the data set. From top left to bottom right, we show the distribution of the redshifts of the galaxy lenses and sources, of the magnitudes of the galaxy lenses and sources, of the Einstein radii of the largest critical curve in the lensing system, and of `n_pix_source`. The histograms in each panel refer to the lenses (green) and nonlenses (red) separately and to the complete data set (blue). The galaxy lenses in the two classes share similar distributions of redshift, magnitude, and Einstein radius (top, middle, and bottom left panels, respectively). The redshift distribution of the sources in the top right panel is also similar for the two subsets. On the other hand, the simulated sources (middle right panel) in the nonlenses class are fainter on average than that of the sources in the lenses. This is intuitive because sources with lower magnitudes (i.e., brighter sources) will be more evident in the images, and it will be more likely that they produce a clear lensing event. A similar argument can be made about `n_pix_source` (bottom right panel): the higher the value of this parameter, the clearer the distortion of the source images, hence the lensing system.

4. Results and discussion

4.1. Data preprocessing

The data preparation consists of a sequence of several steps. We divided the entire data set into three subsets: the training

set (70%), the validation set (5%), and the test set (25%). The images in the data set were randomly assigned to one of these subsets, but we checked that all subsets (training, validation, and testing) were representative of the entire data set. We did this by inspecting the distributions of several parameters that define the characteristics of the lenses and sources in the data set, such as their redshift, magnitude, and Einstein radius.

After the data set was split, we randomly selected 20% of the images in the training set for augmentation. We performed five augmentations: We rotated these images by 90°, 180°, and 270° and flipped them with respect to the horizontal and vertical axes. After performing these operations, we doubled the size of the training set. Neither the test set nor the validation set were augmented.

Afterward, we proceeded with the normalization of the images in the data set. We subtracted the mean and divided it by the standard deviation of the mean image of the training set. The mean image of the training set is the image that has for every pixel i, j the mean value of the pixel i, j of all images in the training set. The reason for this type of normalization is that the computation of the gradients in the training stage of the networks is easier when the features in the training set are in a similar range. Moreover, scaling the inputs in this way makes the parameter sharing more efficient (Goodfellow et al. 2016).

4.2. Training procedure

We implemented, trained, and tested our networks using the library Keras[3] (Chollet 2015) 2.4.3 with the TensorFlow[4] (Abadi et al. 2016) 2.2.0 backend on an NVIDIA Titan Xp GPU.

We used the adaptive moment estimation (Adam; Kingma & Ba 2017; Reddi et al. 2019) optimizer with an initial learning rate of 10^{-4}. We employed the binary cross-entropy $\mathcal{L}$ to estimate the loss at the end of each epoch,

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} y(\mathbf{x}_i) \ \ln[y_p(\mathbf{x}_i)] + [1 - y(\mathbf{x}_i)] \ \ln[1 - y_p(\mathbf{x}_i)], \quad (5)$$

where N is the number of training examples, $\mathbf{x}_i$ is the batch of images used to compute the loss, y is the ground truth, and y_p is the probability that the ith example has the label 1, as predicted by the network, so that $1 - y_p$ is the probability that the ith example has the label 0.

The performance of the network on the validation set is estimated at the end of every epoch and is used to monitor the training process. If the loss function evaluated on this independent subset does not decrease for 20 consecutive epochs, the training will be stopped with the EarlyStopping[5] class from Keras. This step is particularly useful to avoid overfitting. At the end of training, we used the best models, that is, those with the lowest value of the loss function on the validation set, for our tests.

4.3. Performance evaluation

We assessed the performance of our trained networks by examining the properties of the catalogs produced by the classification of the images in the test set. In particular, we considered four statistical metrics that were immediately derived from the

---

confusion matrix (Stehman 1997). A generic element of the confusion matrix C_{ij} is given by the number of images belonging to the class i and classified as members of the class j. In a binary classification problem like the one considered here, the diagonal elements indicate the number of correctly classified objects, that is, the number of true positives (TP) and the number of true negatives (TN), while the off-diagonal terms show the number of misclassified objects, that is, the number of false positives (FP) and the number of false negatives (FN).

Considering the class of Positives, the combination of these quantities leads to the definition of the following metrics:

– The precision (P) can be computed as

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (6)$$

which measures the level of purity of the retrieved catalog.

– The recall (R) can be computed as

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (7)$$

which measures the level of completeness of the retrieved catalog.

– The F1-score (F1) is the harmonic average of P and R,

$$\text{F1} = 2 \, \frac{P \, R}{P + R}. \quad (8)$$

– The accuracy (A) is the ratio of the number of correctly classified objects and the total number of objects,

$$A = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (9)$$

The first three indicators can be similarly computed for the class of the Negatives, while the accuracy is a global indicator of the performance.

In addition, we computed the receiver operating characteristic (ROC; Hanley 1982) curve, which visually represents the variation of the true-positive rate (TPR) and false-positive rate (FPR) with the detection threshold $t \in (0, 1)$, which was used to discriminate whether an image contains a lens. The area under the ROC curve (AUC) summarizes the information conveyed by the ROC: while 1.0 would be the score of a perfect classifier, 0.5 indicates that the classification is equivalent to a random choice and hence worthless.

4.4. Experiment setup

The identification of GGSL events is primarily based on their distinctive morphological characteristics, namely on the distortion of the images of the background source into arcs and rings, as well as on the color difference between the foreground and background galaxies. However, real lenses can show complex configurations and might not be so easily recognizable. Our experiments aimed at evaluating the ability of CNNs to detect the less clear lenses and at assessing their performance on a diversified data set.

We did this by training the three networks we presented on four selections of images, labeled S1 to S4, which gradually include a greater fraction of objects that present challenging visual identification, as we discuss below for nonlenses and lenses separately. These samples consist of approximately 2000, 10 000, 20 000, and 40 000 images, respectively. They were built

Table 2. Summary of the criteria we adopted to choose the images included in the different selections of lenses and nonlenses for our experiments.

Selection	Lenses		nonlenses		Total
	Criterion	Number of images	Criterion	Number of images	
S1	`n_pix_source >430`	1001	Randomly selected objects with `n_sources = 0`	1000	2001
S2	`n_pix_source >140`	5083	Randomly selected objects with `n_sources = 0`	5000	10 083
S3	`n_pix_source >70`	9709	Randomly selected objects with `n_sources = 0`	10 000	19 709
S4	`n_pix_source >20`	19 591	Randomly selected objects with `n_source_im = 0`	20 000	39 591

Notes. While the identification of the lenses is solely based on the variation of a threshold value for the parameter `n_pix_source`, the identification of the nonlenses is primarily based on the possible presence and visibility of a background source.

to have an approximately equal number of lenses and nonlenses (see Table 2). The criteria we adopted to progressively broaden our selections took the features into account that might be employed by the networks to classify the objects as members of the correct category.

In the case of the nonlenses, the lack of a background source, or the absence of its images, makes the classification more likely to be correct. Therefore, we initially considered a sample of the approximately 10 000 nonlenses without a background source. Specifically, we selected 1000 of them in S1, 5000 in S2, and 10 000 in S3. In S4, we broadened our sample by including the images to which a background source was added, but that do not correspond to a visible image, extending our selection to the other objects that are classified as nonlenses according to the criteria in Eq. (4).

In the case of the lenses, the definition of an effective criterion to identify the clearest examples in the data set is more important and also more challenging. The mere presence of an image of the source does not guarantee a straightforward classification of the system because several factors contribute to the actual clarity of the observable features. They include the magnitude of the source and the extension of the image produced by the lensing effect. After several tests involving these parameters and others (e.g., the Einstein area and the magnification of the sources), we selected `n_pix_source` as an appropriate parameter to distinguish between clear and faint lenses. The complete sample of lenses is characterized by the minimum value `n_pix_source > 20`. From S4 to S1, we increased this threshold to different levels, which depended on the number of images we sought to isolate: the higher the value, the smaller the number of selected images and the clearer the lenses. The thresholds established for the creation of the selections described so far also take into account the necessity to have a comparable number of images of each class, so that the examples passed to the networks in the training phase are balanced. In Table 2, we summarize the criteria we used to identify the images to include in each selection. We also show in Fig. 2 some randomly chosen examples of lenses that are characteristic of each selection to better illustrate which kind of selection we introduce
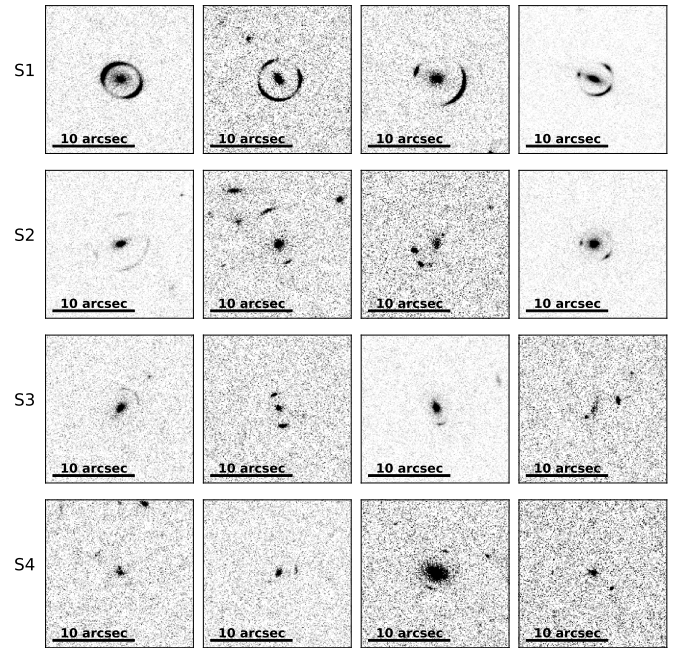

Fig. 2. Examples of the kind of lenses included in all the selections used for training. From top to bottom row, we show four random lenses that were extracted from data sets S1, S2, S3, and S4, as simulated in the $I_{\mathrm E}$ band.

by considering different thresholds for `n_pix_source` in the definition of the training sets.

We trained and tested on these selections of the data set the three architectures we discussed above: a VGG-like network (Simonyan & Zisserman 2015), an IncNet (Szegedy et al. 2015, 2016), and a ResNet (He et al. 2016; Xie et al. 2017). We conducted 24 training sessions in total because we trained each architecture on each selection of data. Twelve of them used the VIS images, and the other 12 used the NISP bands in addition to the VIS one. Every training was carried out for
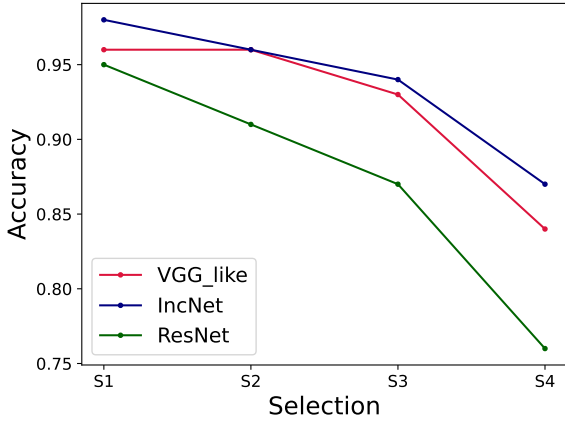
Fig. 3. Trend of the classification accuracy of the single-branch versions of the VGG-like network (red), the IncNet (blue) and the ResNet (green) tested on the four data selections.

Fig. 4. Trend of the precision (first row), recall (second row), and F1-score (third row) in the classification of the nonlenses (left column) and of the lenses (right column) in the different selections. Differently colored lines refer to different networks, as labeled, in the single-branch configuration.

100 epochs because the EarlyStopping method we had set up to prevent overfitting did not interrupt any of them. The best results of each architecture and each classification experiment, which were conducted using the $I_{\rm E}$ band images, are summarized in Table B.1, where the precision, recall, F1-score, accuracy, and AUC obtained from the application of our models are reported. An anologous summary for the training on the multiband images is provided in Table B.4.

4.5. Discussion

By studying how the metrics depend on the selections, we find that the ability of our networks to correctly classify the images tends to deteriorate as the fraction of included lenses with a low signal-to-noise ratio increases. All the results described in the paper were found by considering a classification threshold of 0.5. The trend of the accuracy is shown in Fig. 3. Our three models succeed in the classification of the objects in the selections S1 and S2, where the accuracy is in the range ~0.9 to ~0.96. The IncNet and VGG-like network also perform similarly on S3, while they reach an accuracy level of ~0.87 on S4. On the other hand, ResNet performs worst, with an accuracy of ~0.75 on the complete data set.

The global trends of precision, recall, and F1-score are also similar to that of the accuracy. They are shown in the top, middle, and bottom panels of Fig. 4, respectively. These metrics were evaluated separately on the nonlenses (left panels) and on the lenses (right panels), but the same consideration applies to both classes. This suggests that the degradation of the performance does not only affect the identification of the lenses, but affects the classification of the two categories. In particular, the F1-score, which depends on precision and completeness, peaks at ~0.96 on S1 and decreases to ~0.87 on S4, and ResNet is again the worst-performing network.

In each panel of Fig. 5, we show the ROC curves of one of our networks, evaluated on the test sets of the selections S1, S2, S3, and S4. Their trends for the IncNet (middle panel) and the ResNet (bottom panel) are similar, and the AUC decreases by ~10% from S1 to S4. It should, however, be pointed out that IncNet performs systematically better than ResNet: while the AUC of the former is 0.92 on S1 and 0.81 on S4, the AUC of the latter ranges from 0.81 on S1 to 0.7 on S4. On the other hand, the ROC of the VGG-like network on S2 and S4 has a lower AUC, of ~0.57, compared to the other models, and higher AUC values
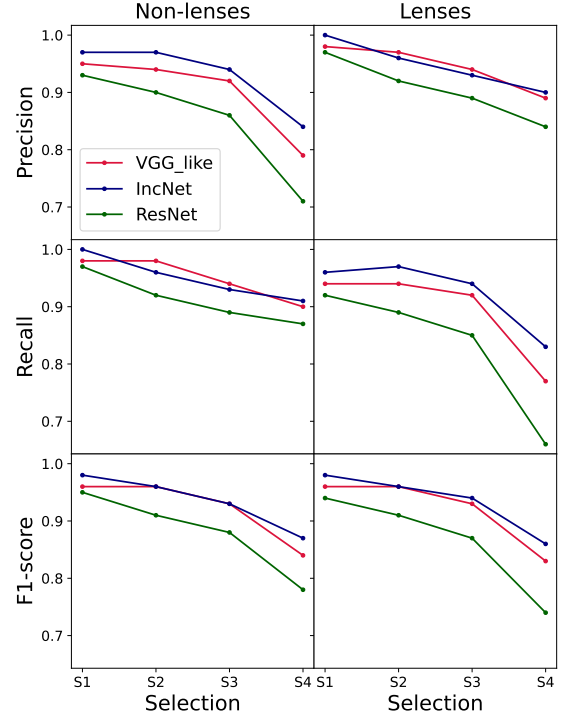
only for the selections S1 and S3. After studying the predictions of this network on the different selections, we think that this is due to a significant difference in the number of objects that is predicted in the two classes when a high threshold is applied to the output probabilities.

We focus on the selection S4, that is, on the performance of our models on the complete data set. Figure 6 shows nine misclassified nonlenses, and Fig. 7 shows nine misclassified lenses. The images reported in these figures were selected from those that were misclassified by all three models, and therefore, they should be characterized by the features that the networks generally find harder to attribute to the correct class.

The false positives in Fig. 6 are mostly characterized by the coexistence of more than one source in addition to the lens galaxy, which might be mistaken for multiple images of the same source. The misinterpretation of these objects might be exacerbated by the inclusion of several low `n_pix_source` lenses in the training set. Many of the lenses in the labeled examples do not present clear arcs or rings, and the faint distortions encountered in the feature-extraction process are likely to resemble specific morphological features of nonlensed galaxies, such as spiral arms, or isolated, but elongated galaxies. One possible way to mitigate the misclassification of nonlenses with a background source could be to train the networks on multiband images to benefit from the color information. We investigate this possibility in Sect. 4.8.
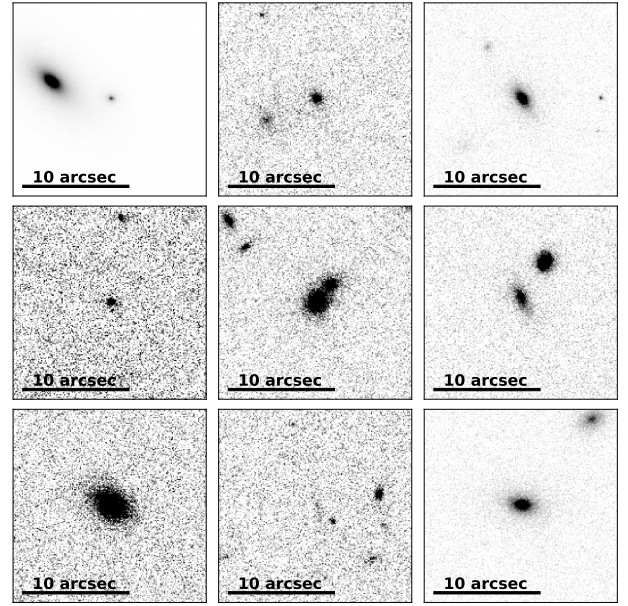
The false negatives in Fig. 7 are partly not even recognizable as lenses by visual inspection. Although they were classified as lenses according to the criteria in Eq. (4), many of these objects do not show evident lensing features. Therefore, if the classification were to be carried out on unlabeled observations, we would not expect the models to be able to identify them as lenses.
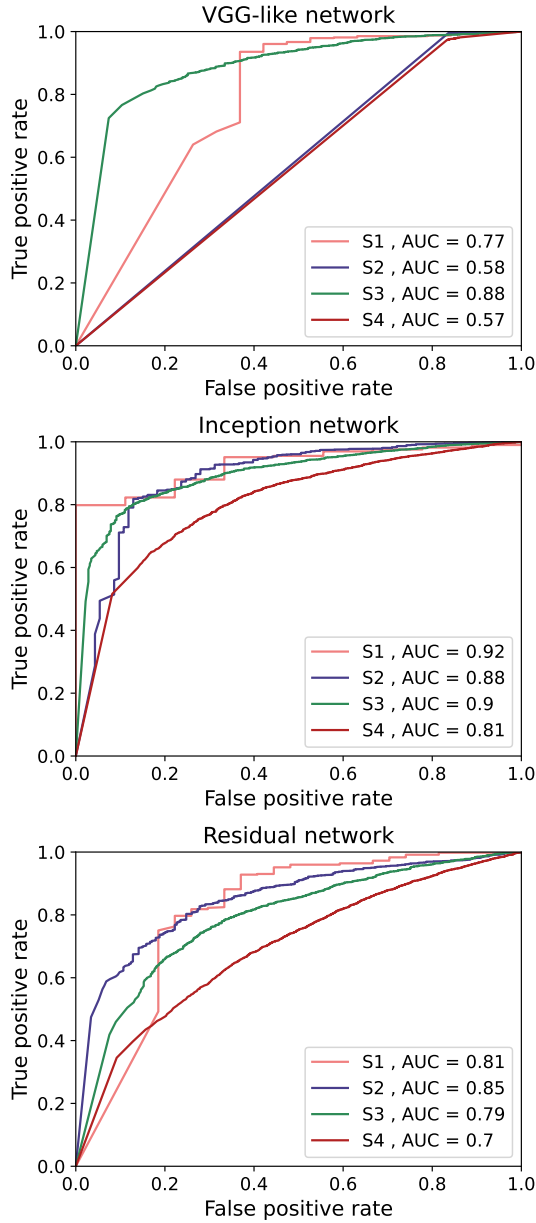
Fig. 5. ROC curves as obtained from the tests of the single-branch versions of our architectures. From top to bottom, each panel of this image shows the ROC curves of the VGG-like network, the IncNet, and the ResNet to the test sets of the different selections S1 (pink line), S2 (blue line), S3 (green line) and S4 (red line) of the data set.


Fig. 6. Example of false positives produced by the three networks in the single-branch configuration when applied to the selection S4, here pictured in the I_E band.
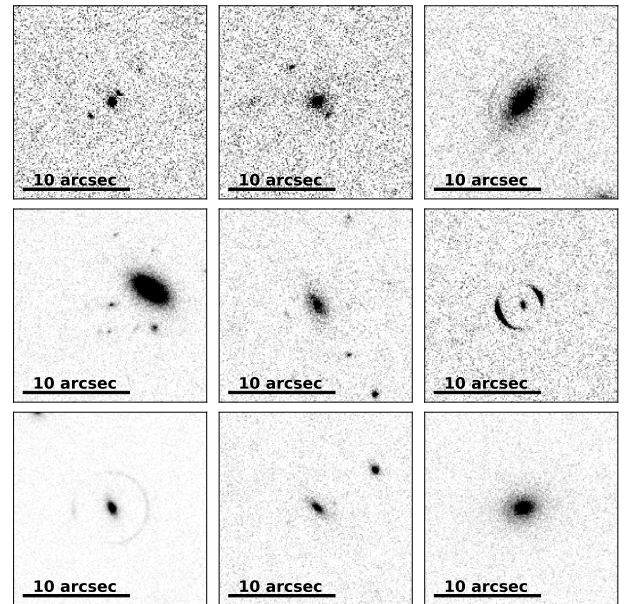
Fig. 7. Example of false negatives produced by the three networks in the single-branch configuration when applied to the selection S4, here pictured in the I_E band.

An approach to solving the issue of nondetectable lenses might be to complement the use of the aforementioned criteria with the visual inspection of the images in the training set. In addition to this, we might include an additional criterion to ensure that the arc is detectable with respect to the other sources in the image. In this case, we would only accept systems as lenses in which the flux of the brightest pixel of the background source is greater than the flux of the other objects along the line of sight at the same pixel (see Shu et al. 2022; Cañameras et al. 2023). However, in some of the images, the arc-shaped and ring-shaped sources are evident. Nevertheless, their classification is incorrect, which signals that some clear lenses might also be missed by our classifiers.

In order to further investigate the ability of the networks trained on S4 to identify clear lenses, we tested them on the

images in S2 (test S4/S2). The networks trained on S4 have analyzed during training and validation some of the images that are part of S2. We removed these images from our test set S4/S2, because otherwise the network performance would be biased to a better performance than can be achieved on unseen data. We compared the result of this test with results obtained from training and testing the networks on S2 (test S2/S2). The results of this comparison are shown in Fig. 8, and more details can be found in Table B.2.

The performance of the models trained on S4 in identifying the lenses in S2 is generally worse than that of the models trained on S2, even though the images that are part of S2 are
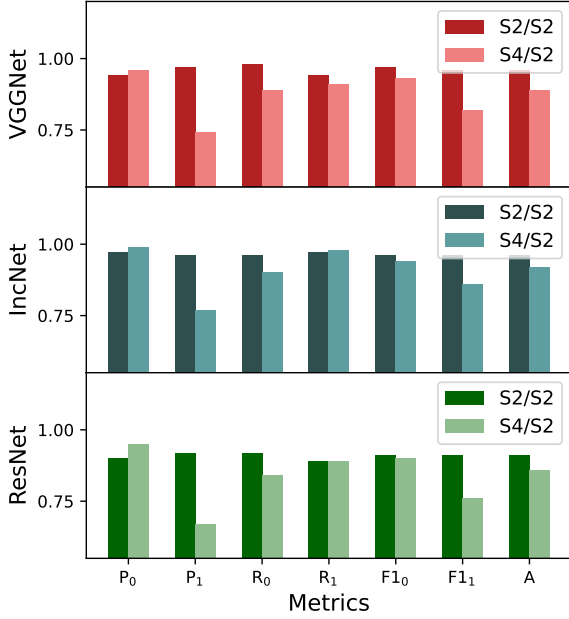
Fig. 8. Comparison of the tests S2/S2 and S4/S2 (darker and lighter histograms) run with the VGG-like network (top), the IncNet (center), and the ResNet (bottom). In each panel, we show the results for the different metrics. From left to right, we show the precision on the class of the nonlenses (P_0) and lenses (P_1), the recall on the class of the nonlenses (R_0) and lenses (R_1), the F1-score on the class of the nonlenses ($F1_0$) and lenses ($F1_1$), and the overall accuracy (A).

also inevitably part of S4 because S4 consists of the complete data set. One reason for this is that the networks we used in the test S2/S2 were specifically trained to identify the lenses in S2, while the networks trained on the larger data set S4 were exposed to a larger variety of systems and are not as specialized on the S2 lenses. We examine the results in Table B.2, however. While the completeness of the retrieved catalog of lenses is constant in the two tests, the precision decreases by ∼20%, passing from ∼0.95 in the test S2/S2 to ∼0.73 in S4/S2, with only minor differences between the different architectures. Even though the magnitude of the overall deterioration is not large per se (the accuracy decreases by ∼5% for the three networks), this is problematic because it is also due to the misclassification of clear lenses, which are also the most useful for scientific purposes.

This result suggests that the performance of the models trained on S4 is worse in general because a significant fraction of this selection is composed of nonobvious lenses that are intrinsically harder to classify. Moreover, the ability of the models to recognize the clearest GGSL events in the data set that are also present in S2 deteriorates.

This effect might result from a combination of two complementary factors regarding the characteristics of the images in the data set. First, the fraction of clear images in the training set of S4 is smaller than in the other selections because of the relevant fraction of low `n_pix_source` lenses included. This is reflected in the fact that the networks might not learn how to properly distinguish them. Wide arcs and rings are recognizable only in a moderate number of images, and they are therefore not as significant as they are in S2 for the classification of the lenses. Second, the most frequently recurring features in the training set are those that occur in images with a low signal-to-noise ratio, and they thus contribute to explaining the misinterpretation of some of the images that present evident lensing features.

As shown in Fig. 7, a large fraction of the lenses that were classified as nonlenses by the networks trained on S4 do not present clear lensing features. However, a non-negligible fraction of evident lenses might also be missed if the training set were extended to include a significant number of fainter arcs because the evident systems might become under-represented. In addition to this, the architecture of the network appears to be influential in the outcome of the classification only to a certain degree. In particular, when trained and tested on the same selections, the IncNet and VGG-like networks generally perform similarly when the metrics in Figs. 3 and 4 are compared. The ResNet, on the other hand, performs significantly worse than the others, especially on S4.

4.6. Additional tests

We tested the models trained on S2 on the wider selections S3 and S4 (tests S2/S3 and S2/S4, respectively) after removing the parts of these samples that were also included in the training set of S2. This test had the purpose of assessing whether the networks trained on clear examples are flexible enough to detect fainter systems. A lower performance from S2/S3 to S2/S4 was also expected because CNNs mostly generalize to the images that are similar to those in the data set they were trained with. Consequently, they might perform the same task poorly for images that are characterized by features they never saw before. In the present case, most images in the training set of S2 show clear lensing features, while the test sets progressively include a greater fraction of images with new features.

The general performance of the networks trained on S2 deteriorates on the other broader selections. The accuracy of the classification varies from ∼0.85 in the case S2/S3 to ∼0.7 in the case S2/S4. By comparing these results with those of the test S4/S4 in Figs. 3 and 4, we observe several differences in the precision, recall, and F1-score, computed separately for the nonlenses and lenses, as well as in the accuracy. We report the results of these tests in Table B.3.

The purity of the nonlenses decreases when broader selections are used as test sets. The precision reaches ∼0.64 with S4. On the other hand, the recall is approximately constant at values of ∼0.96 independently of the considered selection, meaning that the largest fraction of the objects in this class is correctly identified. In the case of the lenses, the trend is roughly reversed. The precision of the classification is roughly constant at ∼0.94, while the recall decreases drastically from ∼0.7 in S3 to ∼0.38 in S4. These values suggest that the networks trained on the S2 sample cannot recognize a large fraction of the lenses in the complete data set.

These trends can be interpreted by considering the impact of including the fainter features in the test sets. In particular, the training set of S2 mostly includes clear lenses and images of isolated nonlenses that are not surrounded by other sources. When processing the images in S3 and S4, the absence of clear arcs and rings, and more generally the faintness of the lensing features induce a growing fraction of lenses to be classified as nonlenses. Our results highlight the inability of our models to recover a considerable fraction of lenses that are not similar to those in S2, leading to a decrease of more than ∼20% in the recall of the lenses from S2/S2 to S2/S3 and of ∼30% from S2/S3 to S2/S4 (see Table B.3 for more details).

4.7. The impact of the shapelet decomposition

In the simulation of the images in our data set, we used the galaxies observed in the UDF as background sources. For the

purpose of denoising them, we decomposed the galaxies with a shapelet-based approach. The shapelet technique is a very powerful mathematical tool for describing astrophysical objects, and its limitations have been investigated in some works (see e.g., Melchior et al. 2007, 2010). In this section, we investigate the impact of these limitations on the performance of our networks.

We assessed this by testing our networks on a sample of 134 real lenses mainly found in the Sloan Lens ACS Survey (SLACS; Bolton et al. 2006) and in the BOSS Emission-Line Lens Survey (BELLS; Brownstein et al. 2012) and on 300 nonlensed galaxies of the UDF. The purpose of this test was not to evaluate the performance of our networks on a realistic sample, which would require including a larger number of nonlenses in the test set. We wished to estimate whether the shapelet decomposition prevents the networks from being applied to real observations. The failure of the networks to identify the observed lenses as lenses would indicate that the simulations are not descriptive enough for the characteristics of real galaxies.

We used the networks trained on S2 to carry out this test. We preprocessed all the images by normalizing them with a procedure similar to the one we applied to the simulations as described in Sect. 4.1. In the case of the galaxies of the UDF, we also reshaped the images to the size expected by the networks.

The results of this test are that we recovered 129 of the lenses with the IncNet and 126 lenses with the VGG-like network and with the ResNet. In the case of the nonlensed UDF galaxies, all the three networks correctly classified 296 of them. Based on these recovery rates, the shapelet decomposition does not introduce significant limitations in our simulations.

4.8. Training with multiband images

The correct identification of GGSL events may benefit significantly from color information emerging from the analysis of multiband data. Lenses and sources typically have different colors because their spectral energy distributions (and redshifts) are different. For example, the most common sources are starforming galaxies that appear bluer than the lenses, which in contrast are often early-type passive galaxies. Moreover, the color similarity of multiple images of the same source can be leveraged to identify strongly lensed sources. This is particularly useful in systems that do not present evident morphological distortions.

For example, Gentile et al. (2022) reported that training CNNs on multiband images resulted in an improved classification of systems with small Einstein radii, while training on single-band images was more efficient for finding lenses with large radii. Metcalf et al. (2019) also found that using multiband images for the training substantially improved the performance of the classifiers for mock ground-based data, even though the color information came from observations with poorer spatial resolution.

We evaluated the importance of color information for the identification of the low `n_pix_source` lenses in *Euclid*-like data by repeating the same training as before, but this time included the NIR images that are also available from the simulations. We show in Fig. 9 some randomly chosen examples of lenses obtained by combining the VIS and NIR bands. We changed the architecture of our models to take the different sizes of the VIS and NISP images into account, as explained in Sect. 2 and represented in panels b of Figs. A.1–A.3, but otherwise, we used the same setup as in our previous experiments. We report the results of these tests in Table B.4.
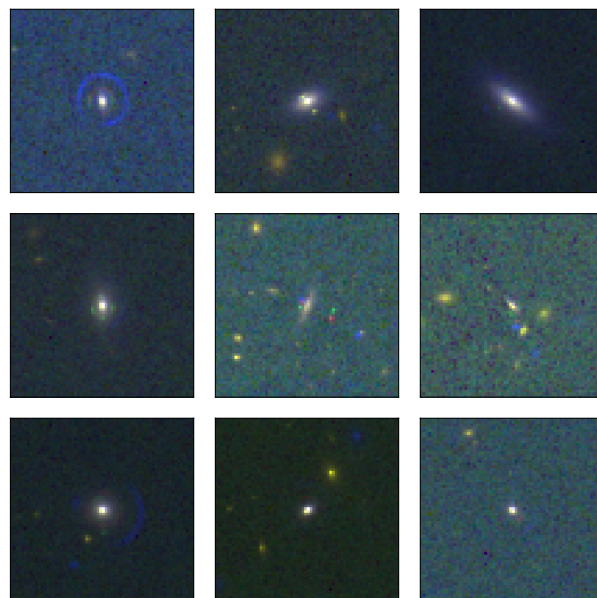
Fig. 9. Example of randomly chosen lenses in the configuration used for multiband training. For visualization pruposes, the images simulated in the I_E band were downgraded to the resolution of the NISP bands in these examples.

By comparing these values to those of the VIS training (see Table B.1), we do not observe a significant improvement in the model performances for a training with multiband data. This is expected for the smaller selections, which are limited to the clearest lenses, whose correct identification through their morphology is relatively easy. In these cases, the color information is therefore expected to be less relevant. However, for broader selections, in which the morphology of the lenses is less clear, we might expect to see some improvement in the classification performance when the models are fed with color information. Surprisingly, we do not note any significant variations in the metrics that quantify the model performance.

We interpret this result as follows. First, the wavelength range covered by the VIS instrument (see Table 1) does not include the wavelengths at which the color difference between the background and foreground galaxies is particularly evident, that is, the blue wavelengths of the optical spectrum. Second, the images in the NIR bands are characterized by lower resolution than those in the I_E band (also see Table 1), which means that the morphological information is degraded in these channels. This also suggests that morphological information is more important than color for identifying lenses, at least in this wavelength range.

4.9. Finding lenses in unbalanced data sets

As we discussed in Sect. 3, training on a balanced training set is important for the networks to learn how to assign the images to the correct class, but a balanced test set is not a requirement. While in all the previous tests we used a balanced test set, with a ratio of about 1:1 between lenses and nonlenses, this is very different from reality, where we reasonably expect to observe less than one lens for 1000 nonlenses (Marshall et al. 2009). In this scenario, even very efficient classifiers will produce a large number of false positives (Savary et al. 2022; Jacobs et al. 2019a,b), and the visual inspection of thousands of candidates is required to find definite samples of strong lenses. While training on simulations instead of real observations plays a role in this

because it is possible that the images present irregular features or shapes that were not included in the training, the high imbalance between the two populations is a major factor to consider.

For this reason, we ran an additional test with realistic proportions in the number of images of the two subsamples. We focused on the networks trained on S1, which globally have the best performances (Figs. 3 and 4). We applied the networks trained on this selection on a test set that has the same lenses as in the original test set of S1, that is, 240 lenses, and used the ~80 000 nonlenses that were excluded from the training (as discussed in Sect. 3). While most of the metrics have similar values to those we found in the test with balanced classes, the precision drops to ~0.15 for the VGG-like network, to ~0.45 for the IncNet and to ~0.13 for the ResNet. This is expected and due to the larger number of false positives predicted by the networks. To reduce the occurrence of false positives, we combined the results of the three networks by averaging their predictions, as this has shown to benefit the rate of correct predictions (e.g., Taufik Andika et al. 2023). We find that the ensemble prediction indeed has a higher precision (with a precision of ~0.46) than those of the VGG-like network and of the ResNet, while it is comparable to that of the IncNet. More details for this test are given in Table B.5.

Even though it is difficult to design a method that will produce a highly pure and complete sample of strong lenses, different strategies are possible to mitigate the issue of many false positives. A common way to reduce their number is to use a high threshold for the classification of the lenses (Petrillo et al. 2019; Gentile et al. 2022) and perform a visual inspection of the candidates that are most likely to be lenses to further refine the selection. The drawback of this method is that the completeness of the sample decreases because the systems that are classified with a lower probability are missed. Another possible strategy is increasing the number of images with misleading features in the negative class of the training set (Cañameras et al. 2020). This should make the networks more familiar with these objects and thus more efficient in recognizing them when applied to real data. Moreover, methods such as transfer learning and domain adaptation might improve the classification performance with real data (Domínguez Sánchez et al. 2019; Ćiprijanović et al. 2022). These techniques would require retraining networks that were trained on simulations on a small sample (a few hundred) of observed lenses and might lead to a significant improvement of the network performances.

4.10. Finding lenses in Euclid

Future *Euclid* observations will offer the opportunity to increase the number of known GGSL events by orders of magnitude as long as potential candidates are efficiently identified. The optimization of the lens-finding strategy, especially in the first year after the launch, is also essential for efficient follow-up observations. For example, the 4-m Multi-Object Spectroscopic Telescope (4MOST; de Jong et al. 2019) Strong Lens Spectroscopic Legacy Survey[6] will observe about 10 000 lens candidates observed by *Euclid* and LSST, providing spectroscopic redshifts for them.

The strategy currently planned for finding lenses in the survey relies both on fully simulated images and data-driven simulations. Training CNNs on simulated images is inevitable in the initial phase of the *Euclid* observations because only so few

galaxy-galaxy lenses are known at the moment. As the data accumulate, more sophisticated simulations will be made, in which the lenses are real galaxies observed by *Euclid*. The networks will be retrained with images that include realistic properties of both lenses and sources, thus improving the performance of the classifiers in the next step of the data analysis. The addition of information about photometric redshifts of the sources might also yield some improvement, but this comes with the challenge of measuring them with good accuracy. A large enough separation between the lens galaxy and the source or efficient deblending techniques are decisive in this context.

The greatest advantage of searching for lenses with *Euclid* is that it will resolve faint Einstein rings with small radii (~0.5″), mostly lensed by bulges of spiral galaxies, in addition to lenses on a larger angular scale. These systems are usually unresolved by ground-based facilities, but will be found with the high resolution of *Euclid*. Moreover, they will be most common according to forecasts (Collett 2015). *Euclid* observations could also be combined with and complemented by those of other surveys. The LSST, for instance, will observe a comparable number of lenses that will likely be skewed to larger radii because of the lower resolution of ground-based observations. A complementary data set of lenses in the radio band with high resolution will be produced by the Square Kilometer Array (Dewdney et al. 2009). They are complementary to the others because the parent population of the systems observed in radio is different from that of the systems observed in optical and infrared bands (Koopmans et al. 2004).

The fully simulated data sets are also critical for studying the selection functions of the algorithms that will be used for finding lenses in the survey. An accurate characterization of the selection function is necessary for the scientific exploitation of the GGSLs found by *Euclid*. For example, Sonnenfeld (2022) discussed the importance of characterizing the selection function for inferring the properties of the population of galaxies of which the strong lenses are a biased subsample. Moreover, they showed how the information about the number of nondetections can be used to further constrain models of galaxy structure. More recently, Sonnenfeld et al. (2023) investigated the difference between lens galaxies and lensed sources from their parent population, that is, the strong-lensing bias. Because *Euclid* will provide the largest sample of homogeneously discovered strong lenses ever gathered, this type of study will be more significant than in the past.

5. Conclusions

In this work, we have presented a detailed analysis of the performance of three CNN architectures in identifying GGSL events. We used a data set of 40 000 images simulated by the Bologna Lens Factory to mimic the data quality expected from the *Euclid* space mission. The classification was primarily based on the morphology of the systems because we mainly conducted our experiments with the images simulated in the $I_{\rm E}$ band. Still, we evaluated the importance of color information using multiband images. We trained and tested our CNNs on four data-set selections that gradually included a greater fraction of objects characterized by faint lensing features and that will be more difficult to recognize. We evaluated the outcome of the classification by estimating the precision, recall, and $F1$ score of the lens catalogs we obtained.

We found that the morphological characteristics of the lenses included in the training set influence the ability of our CNNs to

---

[6] https://www.4most.eu/cms/science/extragalactic-community-surveys/

identify the lenses in a separate test set in a critical way, whether they show clear or faint lensing features. We found that the inclusion of a large fraction of images deteriorates the performance of our models, causing a decrease in the overall accuracy of ∼10%, from ∼0.95 to ∼0.85 for the IncNet and VGG-like network, and an even greater decrease for the ResNet, which reaches an accuracy of ∼0.74. Moreover, we also found that it impacts the ability of our models to identify the most evident lenses because they become under-represented in the training set.

These results emphasize the importance of building realistic training sets for DL models. This is particularly relevant for the first searches because we will not have real lensing systems at our disposal, and the simulations of large data sets will be the only option for training. In this phase, the inclusion of the real galaxies observed by *Euclid* in the simulation will make the mocks more realistic than those used so far to train the networks. In particular, they suggest that identifying lenses with different morphologies might require specific training focused on the type of lenses of interest for a certain purpose. Alternatively, the classification of the lenses might be considered and solved a multiclass classification problem, distinguishing the clear and probable lenses from the probable and evident nonlenses. In this last case, however, the distinction between obvious and nonobvious objects should be further investigated and quantified.

We also retrained our models on the same selections of the data set, including a separate channel for processing the near-IR images in addition to those in the $I_{\rm E}$ band, thus assessing how relevant the color information is for identifying lenses with a low signal-to-noise ratio. We found no significant improvement in the performance of any of our networks. We suggest that this might depend on a combination of two factors. First, the images in the $I_{\rm E}$ band have a higher resolution than those in the near-IR bands. Second, the $I_{\rm E}$ band covers a wavelength range in which the color difference between lens and source galaxies might not be important (see Table 1).

Finally, we highlight that the three architectures retrieve catalogs with similar characteristics in terms of completeness and precision when they are applied to the same selections of images. The only exception is ResNet, whose accuracy on the full data set is lower by ∼10% than the others. Because of the higher precision of IncNet in the test with an unbalanced number of images, we would conclude that this is the best-performing network of those we tested. The results of this test are indeed the closest to what we might expect from real data, and they are therefore particularly relevant to evaluate the performance of our models.

In the future, we could improve our selection method by testing a combination of physical parameters to differentiate between faint and clear lenses instead of using `n_pix_source`, which we have as a result of our simulations, but is not a physical property of the galaxies. It would also be useful to study whether there is a bias in the properties of the lenses found by our models to characterize the type of systems better that are most likely to be found or missed.

References

Abadi, M., Barham, P., Chen, J., et al. 2016, 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 265
Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, PASJ, 70, S4
Allison, J. R., Moss, V. A., Macquart, J. P., et al. 2017, MNRAS, 465, 4450
Angora, G., Rosati, P., Brescia, M., et al. 2020, A&A, 643, A177
Angora, G., Rosati, P., Meneghetti, M., et al. 2023, A&A, 676, A40
Barnabè, M., Czoske, O., Koopmans, L. V. E., Treu, T., & Bolton, A. S. 2011, MNRAS, 415, 2215
Bengio, Y. 2009, Foundation and Trends in Machine Learning, 2, 1
Bergamini, P., Rosati, P., Vanzella, E., et al. 2021, A&A, 645, A140
Bolton, A. S., Burles, S., Koopmans, L. V. E., Treu, T., & Moustakas, L. A. 2006, ApJ, 638, 703
Brownstein, J. R., Bolton, A. S., Schlegel, D. J., et al. 2012, ApJ, 744, 41
Buda, M., Maki, A., & Mazurowski, M. A. 2018, Neural Networks, 106, 249
Cabanac, R. A., Alard, C., Dantel-Fort, M., et al. 2007, A&A, 461, 813
Cañameras, R., Schuldt, S., Suyu, S. H., et al. 2020, A&A, 644, A163
Cañameras, R., Schuldt, S., Shu, Y., et al. 2021, A&A, 653, A6
Cañameras, R., Schuldt, S., Shu, Y., et al. 2023, A&A, submitted https://doi.org/10.1051/0004-6361/202347072
Carretero, J., Tallada, P., Casals, J., et al. 2017, in Proceedings of the European Physical Society Conference on High Energy Physics, 5–12 July, 488
Chollet, F. 2015, https://github.com/fchollet/keras
Ćiprijanović, A., Kafkes, D., Snyder, G., et al. 2022, Mach. Learn. Sci. Technol., 3, 035007
Coe, D., Benítez, N., Sánchez, S. F., et al. 2006, AJ, 132, 926
Collett, T. E. 2015, ApJ, 811, 20
Cropper, M., Cole, R., James, A., et al. 2012, SPIE Conf. Ser., 8442, 84420V
Dark Energy Survey Collaboration (Abbott, T., et al.) 2016, MNRAS, 460, 1270
de Jong, J. T. A., Verdoes Kleijn, G. A., Boxhoorn, D. R., et al. 2015, A&A, 582, A62
de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, The Messenger, 175, 3
Desprez, G., Richard, J., Jauzac, M., et al. 2018, MNRAS, 479, 2630
Dewdney, P. E., Hall, P. J., Schilizzi, R. T., & Lazio, T. J. L. W. 2009, IEEE Proc., 97, 1482
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, MNRAS, 476, 3661
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2019, MNRAS, 484, 93
Euclid Collaboration (Schirmer, M., et al.) 2022a, A&A, 662, A92
Euclid Collaboration (Scaramella, R., et al.) 2022b, A&A, 662, A112
Fluri, J., Kacprzak, T., Lucchi, A., et al. 2019, Phys. Rev. D, 100, 063514
Gavazzi, R., Treu, T., Rhodes, J. D., et al. 2007, ApJ, 667, 176
Gavazzi, R., Treu, T., Marshall, P. J., Brault, F., & Ruff, A. 2012, ApJ, 761, 170
Gavazzi, R., Marshall, P. J., Treu, T., & Sonnenfeld, A. 2014, ApJ, 785, 144
Gentile, F., Tortora, C., Covone, G., et al. 2022, MNRAS, 510, 500
Ghosh, A., Urry, C. M., Wang, Z., et al. 2020, ApJ, 895, 112
Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (The MIT Press)
Grillo, C. 2012, ApJ, 747, L15
Grillo, C., Gobat, R., Presotto, V., et al. 2014, ApJ, 786, 11
Gupta, N., & Reichardt, C. L. 2020, ApJ, 900, 110
Gwyn, S. D. J. 2012, AJ, 143, 38
Hanley, J. V., & McNeil, B. 1982, Radiology, 143, 29
He, K., Zhang, X., Ren, S., & Sun, J. 2016, 2016 IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR), 770
He, Z., Er, X., Long, Q., et al. 2020, MNRAS, 497, 556
Hebb, D. O. 1949, The Organization of Behavior: A Neuropsychological Theory (Wiley)
Ho, M., Rau, M. M., Ntampaka, M., et al. 2019, ApJ, 887, 25
Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, ApJS, 221, 8
Impellizzeri, C. M. V., McKean, J. P., Castangia, P., et al. 2008, Nature, 456, 927
Ioffe, S., & Szegedy, C. 2015, in Proceedings of Machine Learning Research, Proceedings of the 32nd International Conference on Machine Learning, eds. F. Bach & D. Blei (Lille: PMLR), 37, 448

Jacobs, C., Glazebrook, K., Collett, T., More, A., & McCarthy, C. 2017, MNRAS, 471, 167

Jacobs, C., Collett, T., Glazebrook, K., et al. 2019a, ApJS, 243, 17

Jacobs, C., Collett, T., Glazebrook, K., et al. 2019b, MNRAS, 484, 5330

Jauzac, M., Klein, B., Kneib, J.-P., et al. 2021, MNRAS, 508, 1206

Kingma, D. P., & Ba, J. 2017, arXiv e-prints [arXiv:1412.6980]

Koopmans, L. V. E., Browne, I. W. A., & Jackson, N. J. 2004, New Astron. Rev., 48, 1085

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints [arXiv:1110.3193]

LeCun, Y., Boser, B., Denker, J. S., et al. 1989, Neural Comput., 1, 541

Li, R., Napolitano, N. R., Tortora, C., et al. 2020, ApJ, 899, 30

Li, R., Napolitano, N. R., Spiniello, C., et al. 2021, ApJ, 923, 16

Li, R., Napolitano, N. R., Feng, H., et al. 2022, A&A, 666, A85

Liew-Cain, C. L., Kawata, D., Sánchez-Blázquez, P., Ferreras, I., & Symeonidis, M. 2021, MNRAS, 502, 1355

Lin, M., Chen, Q., & Yan, S. 2013, arXiv e-prints [arXiv:1312.4400]

LSST Science Collaboration (Abell, P. A., et al.) 2009, ArXiv e-prints [arXiv:0912.0201]

Maciaszek, T., Ealet, A., Gillard, W., et al. 2022, SPIE Conf. Ser., 12180, 121801K

Marshall, P. J., Hogg, D. W., Moustakas, L. A., et al. 2009, ApJ, 694, 924

Maturi, M., Mizera, S., & Seidel, G. 2014, A&A, 567, A111

McCulloch, W., & Pitts, W. 1943, Bull. Math. Biophys., 5, 115

Melchior, P., Meneghetti, M., & Bartelmann, M. 2007, A&A, 463, 1215

Melchior, P., Böhnert, A., Lombardi, M., & Bartelmann, M. 2010, A&A, 510, A75

Meneghetti, M., Melchior, P., Grazian, A., et al. 2008, A&A, 482, 403

Meneghetti, M., Rasia, E., Merten, J., et al. 2010, A&A, 514, A93

Meneghetti, M., Davoli, G., Bergamini, P., et al. 2020, Science, 369, 1347

Meneghetti, M., Ragagnin, A., Borgani, S., et al. 2022, A&A, 668, A188

Merten, J., Giocoli, C., Baldi, M., et al. 2019, MNRAS, 487, 104

Metcalf, R. B., & Petkova, M. 2014, MNRAS, 445, 1942

Metcalf, R. B., Meneghetti, M., Avestruz, C., et al. 2019, A&A, 625, A119

Minor, Q., Gad-Nasr, S., Kaplinghat, M., & Vegetti, S. 2021, MNRAS, 507, 1662

Myers, S. T., Jackson, N. J., Browne, I. W. A., et al. 2003, MNRAS, 341, 1

Napolitano, N. R., Li, R., Spiniello, C., et al. 2020, ApJ, 904, L31

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Negrello, M., Hopwood, R., De Zotti, G., et al. 2010, Science, 330, 800

Negrello, M., Amber, S., Amvrosiadis, A., et al. 2017, MNRAS, 465, 3558

Nightingale, J. W., Massey, R. J., Harvey, D. R., et al. 2019, MNRAS, 489, 2049

Oguri, M., Rusu, C. E., & Falco, E. E. 2014, MNRAS, 439, 2494

O'Riordan, C. M., Despali, G., Vegetti, S., Lovell, M. R., & Moliné, Á. 2023, MNRAS, 521, 2342

Pan, S., Liu, M., Forero-Romero, J., et al. 2020, Sci. China Phys. Mech. Astron., 63, 110412

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26

Petkova, M., Metcalf, R. B., & Giocoli, C. 2014, MNRAS, 445, 1954

Petrillo, C. E., Tortora, C., Chatterjee, S., et al. 2017, MNRAS, 472, 1129

Petrillo, C. E., Tortora, C., Vernardos, G., et al. 2019, MNRAS, 484, 3879

Pires, S., Vandenbussche, V., Kansal, V., et al. 2020, A&A, 638, A141

Ragagnin, A., Meneghetti, M., Bassini, L., et al. 2022, A&A, 665, A16

Reddi, S. J., Kale, S., & S., K. 2019, arXiv e-prints [arXiv:1904.09237]

Rojas, K., Savary, E., Clément, B., et al. 2022, A&A, 668, A73

Rumelhart, D., Hinton, G. E., & Williams, R. J. 1986, Nature, 323, 533

Savary, E., Rojas, K., Maus, M., et al. 2022, A&A, 666, A1

Schuldt, S., Chirivì, G., Suyu, S. H., et al. 2019, A&A, 631, A40

Seidel, G., & Bartelmann, M. 2007, A&A, 472, 341

Shu, Y., Cañameras, R., Schuldt, S., et al. 2022, A&A, 662, A4

Shuntov, M., Pasquet, J., Arnouts, S., et al. 2020, A&A, 636, A90

Simonyan, K., & Zisserman, A. 2015, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings

Sonnenfeld, A. 2022, A&A, 659, A132

Sonnenfeld, A., Chan, J. H. H., Shu, Y., et al. 2018, PASJ, 70, S29

Sonnenfeld, A., Li, S.-S., Despali, G., Shajib, A. J., & Taylor, E. N. 2023, A&A, 678, A4

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014, J. Mach. Learn. Res., 15, 1929

Stacey, H. R., McKean, J. P., Robertson, N. C., et al. 2018, MNRAS, 476, 5075

Stehman, S. V. 1997, Rem. Sens. Environ., 62, 77

Suyu, S. H., Hensel, S. W., McKean, J. P., et al. 2012, ApJ, 750, 10

Szegedy, C., Wei Liu, Yangqing Jia, et al. 2015, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. 2016, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2818

Tallada, P., Carretero, J., Casals, J., et al. 2020, Astron. Comput., 32, 100391

Taufik Andika, I., Suyu, S. H., Cañameras, R., et al. 2023, A&A, 678, A103

Treu, T., & Koopmans, L. V. E. 2004, ApJ, 611, 739

Tu, H., Limousin, M., Fort, B., et al. 2008, MNRAS, 386, 1169

Vegetti, S., Despali, G., Lovell, M. R., & Enzi, W. 2018, MNRAS, 481, 3661

Wong, K. C., Chan, J. H. H., Chao, D. C. Y., et al. 2022, PASJ, 74, 1209

Wu, J. F., & Boada, S. 2019, MNRAS, 484, 4683

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. 2017, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5987

Xu, B., Wang, N., Chen, T., & Li, M. 2015, arXiv e-prints [arXiv:1505.00853]

Zhou, Y.-T., & Chellappa, R. 1988, in IEEE 1988 International Conference on Neural Networks, 2, 71

Zhu, X.-P., Dai, J.-M., Bian, C.-J., et al. 2019, Ap&SS, 364, 55

[1] Dipartimento di Fisica e Astronomia "Augusto Righi" – Alma Mater Studiorum Universitá di Bologna, via Piero Gobetti 93/2, 40129 Bologna, Italy
e-mail: laura.leuzzi3@unibo.it

[2] INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy

[3] INFN-Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy

[4] Dipartimento di Fisica e Scienze della Terra, Universitá degli Studi di Ferrara, Via Giuseppe Saragat 1, 44122 Ferrara, Italy

[5] INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy

[6] Dipartimento di Fisica "Aldo Pontremoli", Universitá degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy

[7] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland

[8] Aix-Marseille Université, CNRS, CNES, LAM, Marseille, France

[9] Institut d'Astrophysique de Paris, UMR 7095, CNRS, and Sorbonne Université, 98 bis boulevard Arago, 75014 Paris, France

[10] Dipartimento di Fisica e Astronomia, Universitá di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy

[11] Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town 7535, South Africa

[12] South African Radio Astronomy Observatory, 2 Fir Street, Black River Park, Observatory 7925, South Africa

[13] INAF-IASF Milano, Via Alfonso Corti 12, 20133 Milano, Italy

[14] Observatoire de Sauverny, Ecole Polytechnique Fédérale de Lausanne, 1290 Versoix, Switzerland

[15] Université Paris-Saclay, CNRS, Institut d'astrophysique spatiale, 91405 Orsay, France

[16] Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK

[17] Institut für Theoretische Physik, University of Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany

[18] Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, 85748 Garching, Germany

[19] INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese (TO), Italy

[20] Dipartimento di Fisica, Universitá di Genova, Via Dodecaneso 33, 16146, Genova, Italy

[21] INFN-Sezione di Genova, Via Dodecaneso 33, 16146, Genova, Italy

[22] Department of Physics "E. Pancini", University Federico II, Via Cinthia 6, 80126, Napoli, Italy

[23] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal

[24] Dipartimento di Fisica, Universitá degli Studi di Torino, Via P. Giuria 1, 10125 Torino, Italy

[25] INFN-Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy

[26] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona), Spain

[27] Port d'Informació Científica, Campus UAB, C. Albareda s/n, 08193 Bellaterra (Barcelona), Spain

[28] INAF-Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monteporzio Catone, Italy

[29] INFN section of Naples, Via Cinthia 6, 80126 Napoli, Italy

[30] Dipartimento di Fisica e Astronomia "Augusto Righi" – Alma Mater Studiorum Universitá di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy

[31] Centre National d'Etudes Spatiales – Centre spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France

[32] Institut national de physique nucléaire et de physique des particules, 3 rue Michel-Ange, 75794 Paris Cedex 16, France

[33] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK

[34] Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK

[35] European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Roma, Italy

[36] ESAC/ESA, Camino Bajo del Castillo s/n, Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain

[37] University of Lyon, Univ Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, 69622 Villeurbanne, France

[38] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK

[39] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, 1749-016 Lisboa, Portugal

[40] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal

[41] Department of Astronomy, University of Geneva, ch. d'Ecogia 16, 1290 Versoix, Switzerland

[42] INFN-Padova, Via Marzolo 8, 35131 Padova, Italy

[43] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, 91191 Gif-sur-Yvette, France

[44] INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, 34143 Trieste, Italy

[45] INAF-Osservatorio Astronomico di Padova, Via dell'Osservatorio 5, 35122 Padova, Italy

[46] University Observatory, Faculty of Physics, Ludwig-Maximilians-Universität, Scheinerstr. 1, 81679 Munich, Germany

[47] INAF-Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, Italy

[48] INFN-Sezione di Milano, Via Celoria 16, 20133 Milano, Italy

[49] Institute of Theoretical Astrophysics, University of Oslo, PO Box 1029 Blindern, 0315 Oslo, Norway

[50] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA, 91109, USA

[51] von Hoerner & Sulger GmbH, SchloßPlatz 8, 68723 Schwetzingen, Germany

[52] Technical University of Denmark, Elektrovej 327, 2800 Kgs. Lyngby, Denmark

[53] Cosmic Dawn Center (DAWN), Denmark

[54] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

[55] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstrasse 1, 81679 München, Germany

[56] Aix-Marseille Université, CNRS/IN2P3, CPPM, Marseille, France

[57] Université de Genève, Département de Physique Théorique and Centre for Astroparticle Physics, 24 quai Ernest-Ansermet, 1211 Genève 4, Switzerland

[58] Department of Physics, PO Box 64, 00014 University of Helsinki, Finland

[59] Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland

[60] NOVA optical infrared instrumentation group at ASTRON, Oude Hoogeveensedijk 4, 7991PD Dwingeloo, The Netherlands

[61] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany

[62] Department of Physics, Institute for Computational Cosmology, Durham University, South Road, Durham DH1 3LE, UK

[63] Université Paris-Cité, CNRS, Astroparticule et Cosmologie, 75013 Paris, France

[64] University of Applied Sciences and Arts of Northwestern Switzerland, School of Engineering, 5210 Windisch, Switzerland

[65] Institut d'Astrophysique de Paris, 98bis Boulevard Arago, 75014 Paris, France

[66] CEA Saclay, DFR/IRFU, Service d'Astrophysique, Bât. 709, 91191 Gif-sur-Yvette, France

[67] European Space Agency/ESTEC, Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands

[68] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands

[69] Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands

[70] Department of Physics and Astronomy, University of Aarhus, Ny Munkegade 120, 8000 Aarhus C, Denmark

[71] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, Astrophysique, Instrumentation et Modélisation Paris-Saclay, 91191 Gif-sur-Yvette, France

[72] Space Science Data Center, Italian Space Agency, via del Politecnico snc, 00133 Roma, Italy

[73] Dipartimento di Fisica e Astronomia "G. Galilei", Universitá di Padova, Via Marzolo 8, 35131 Padova, Italy

[74] Departamento de Física, FCFM, Universidad de Chile, Blanco Encalada 2008, Santiago, Chile

[75] Institut d'Estudis Espacials de Catalunya (IEEC), Carrer Gran Capitá 2–4, 08034 Barcelona, Spain

[76] Institut de Ciencies de l'Espai (IEEC-CSIC), Campus UAB, Carrer de Can Magrans, s/n Cerdanyola del Vallés, 08193 Barcelona, Spain

[77] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain

[78] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda, 1349-018 Lisboa, Portugal

[79] Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, Plaza del Hospital 1, 30202 Cartagena, Spain

[80] Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, 31400 Toulouse, France

[81] INFN-Bologna, Via Irnerio 46, 40126 Bologna, Italy

[82] Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA

[83] Instituto de Astrofísica de Canarias, Calle Vía Láctea s/n, 38204, San Cristóbal de La Laguna, Tenerife, Spain

[84] INAF-Istituto di Astrofisica e Planetologia Spaziali, via del Fosso del Cavaliere 100, 00100 Roma, Italy

[85] Department of Physics and Helsinki Institute of Physics, Gustaf Hällströmin katu 2, 00014 University of Helsinki, Finland

[86] Junia, EPA department, 41 Bd Vauban, 59800 Lille, France

[87] Instituto de Física Teórica UAM-CSIC, Campus de Cantoblanco, 28049 Madrid, Spain

[88] CERCA/ISO, Department of Physics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA

[89] Laboratoire de Physique de l'École Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, 75005 Paris, France

[90] Observatoire de Paris, Université PSL, Sorbonne Université, LERMA, 750 Paris, France

[91] Astrophysics Group, Blackett Laboratory, Imperial College London, London SW7 2AZ, UK

[92] Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy

[93] SISSA, International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, TS, Italy

[94] IFPU, Institute for Fundamental Physics of the Universe, via Beirut 2, 34151 Trieste, Italy

[95] INFN, Sezione di Trieste, Via Valerio 2, 34127 Trieste, TS, Italy

[96] Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara, Via Giuseppe Saragat 1, 44122 Ferrara, Italy

[97] Institut de Physique Théorique, CEA, CNRS, Université Paris-Saclay, 91191 Gif-sur-Yvette Cedex, France

[98] Dipartimento di Fisica – Sezione di Astronomia, Universitá di Trieste, Via Tiepolo 11, 34131 Trieste, Italy

[99] NASA Ames Research Center, Moffett Field, CA 94035, USA

[100] Kavli Institute for Particle Astrophysics & Cosmology (KIPAC), Stanford University, Stanford, CA 94305, USA

[101] Department of Astronomy and Astrophysics, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA

[102] INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, 40129 Bologna, Italy

[103] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France

[104] Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University, 52056 Aachen, Germany

[105] Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822, USA

[106] Department of Physics & Astronomy, University of California Irvine, Irvine, CA 92697, USA

[107] UCB Lyon 1, CNRS/IN2P3, IUF, IP2I Lyon, 4 rue Enrico Fermi, 69622 Villeurbanne, France

[108] Department of Astronomy & Physics and Institute for Computational Astrophysics, Saint Mary's University, 923 Robie Street, Halifax, NS B3H 3C3, Canada

[109] Dipartimento di Fisica, Universitá degli studi di Genova, and INFN-Sezione di Genova, via Dodecaneso 33, 16146 Genova, Italy

[110] Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain

[111] Department of Computer Science, Aalto University, PO Box 15400, Espoo 00 076, Finland

[112] Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing (GCCL), 44780 Bochum, Germany

[113] Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK

[114] Instituto de Astrofísica de Canarias (IAC), Departamento de Astrofísica, Universidad de La Laguna (ULL), 38200 La Laguna, Tenerife, Spain

[115] Université Paris-Cité, 5 rue Thomas Mann, 75013 Paris, France

[116] Université PSL, Observatoire de Paris, Sorbonne Université, CNRS, LERMA, 75014 Paris, France

[117] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

[118] Department of Physics and Astronomy, Vesilinnantie 5, 20014 University of Turku, Turku, Finland

[119] AIM, CEA, CNRS, Université Paris-Saclay, Université de Paris, 91191 Gif-sur-Yvette, France

[120] Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, Stockholm 106 91, Sweden

[121] Centre de Calcul de l'IN2P3/CNRS, 21 avenue Pierre de Coubertin, 69627 Villeurbanne Cedex, France

[122] Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 2, 00185 Roma, Italy

[123] INFN-Sezione di Roma, Piazzale Aldo Moro 2, c/o Dipartimento di Fisica, Edificio G. Marconi, 00185 Roma, Italy

[124] Centro de Astrofísica da Universidade do Porto, Rua das Estrelas, 4150-762 Porto, Portugal

[125] Department of Mathematics and Physics E. De Giorgi, University of Salento, Via per Arnesano, CP-I93, 73100 Lecce, Italy

[126] INAF-Sezione di Lecce, c/o Dipartimento Matematica e Fisica, Via per Arnesano, 73100 Lecce, Italy

[127] INFN, Sezione di Lecce, Via per Arnesano, CP-193, 73100 Lecce, Italy

[128] Institute of Space Science, Str. Atomistilor, nr. 409 Măgurele, Ilfov 077125, Romania

[129] Institute for Computational Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[130] Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 226, 69120 Heidelberg, Germany

[131] Université St Joseph; Faculty of Sciences, Beirut, Lebanon

[132] Department of Astrophysical Sciences, Peyton Hall, Princeton University, Princeton, NJ 08544, USA

[133] Centro de Estudios de Física del Cosmos de Aragón (CEFCA), Plaza San Juan, 1, planta 2, 44001 Teruel, Spain

Appendix A: Network architectures

The three figures in this appendix show the architectures of the networks we implemented. In particular, Fig.A.1 shows the VGG-like network, Fig. A.2 shows the IncNet, and Fig. A.3 shows the ResNet.
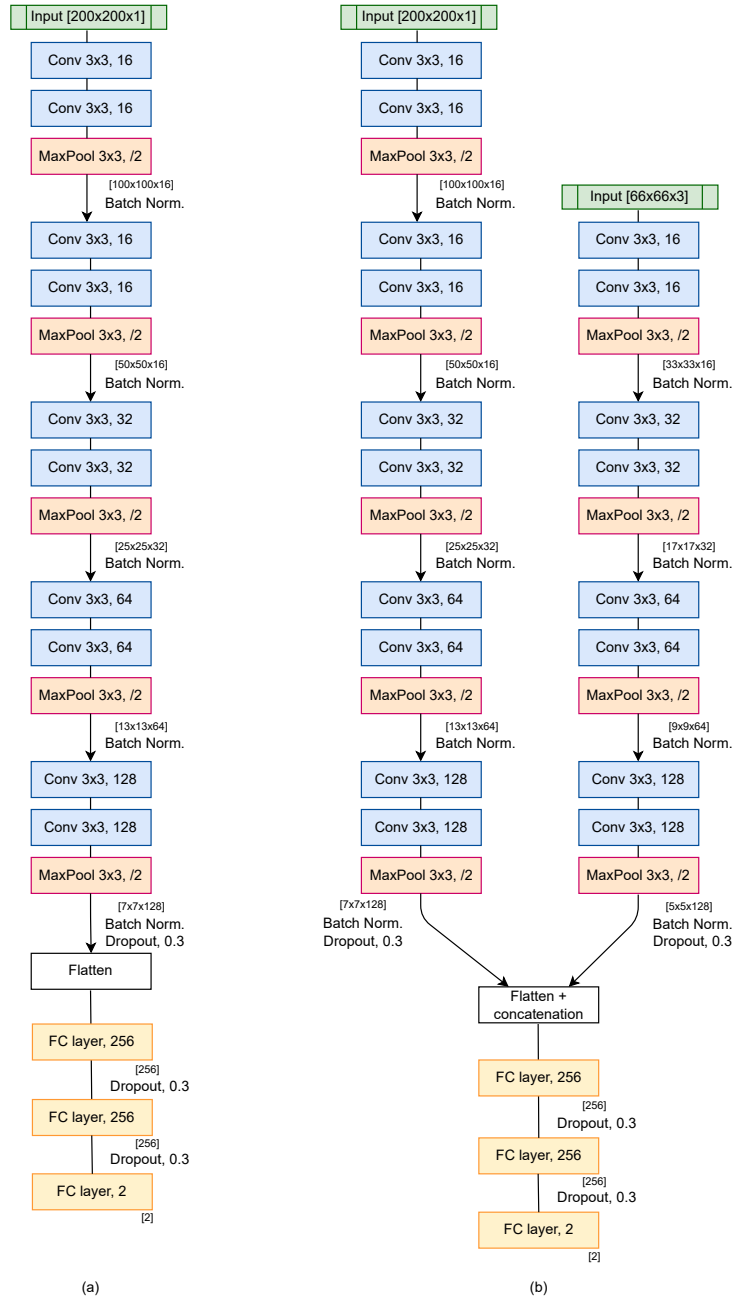
Fig. A.1. VGG-like network configurations tested on (a) VIS images and (b) multiband images. We report the dimension (*D*) and number (*F*) of the filters used in the convolutional layers in the format *D*×*D*, *F*. We also indicate the pooling region (*R*) and the strides (*S*) in the pooling layers in the format *R*×*R*, /*S*. The numbers in square brackets indicate the dimension and number of the feature maps obtained as the output of the layers in the format [*D*×*D*×*F*] in the case of the convolutional layers, and the number of nodes in the format [N] in the case of the fully connected layers.
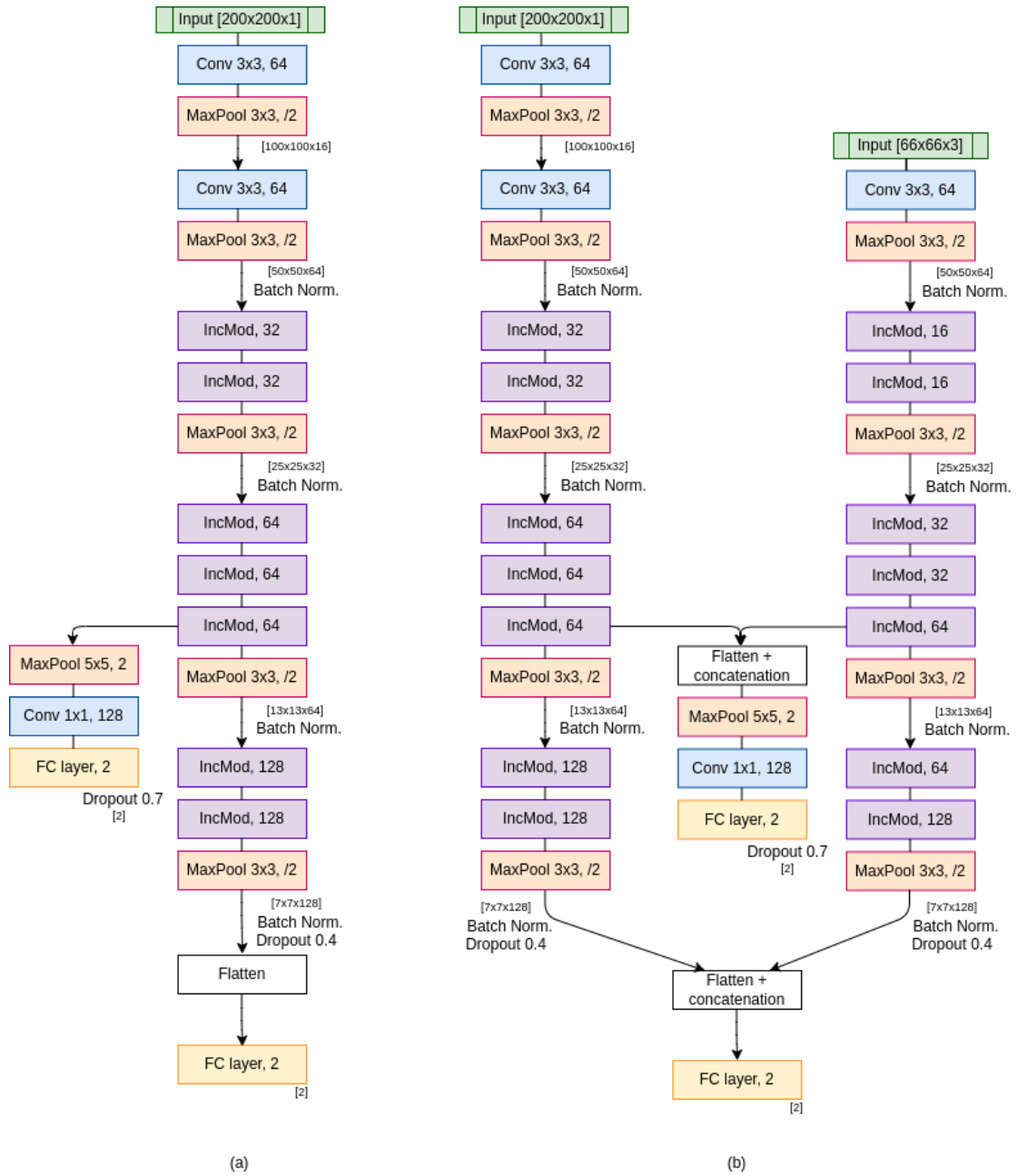
Fig. A.2. Inception network configurations tested on (a) VIS images and (b) multiband images. These diagrams use the same notation as in Fig. A.1. Every inception module (IncMod) is built as described in subsection 2.1.2.
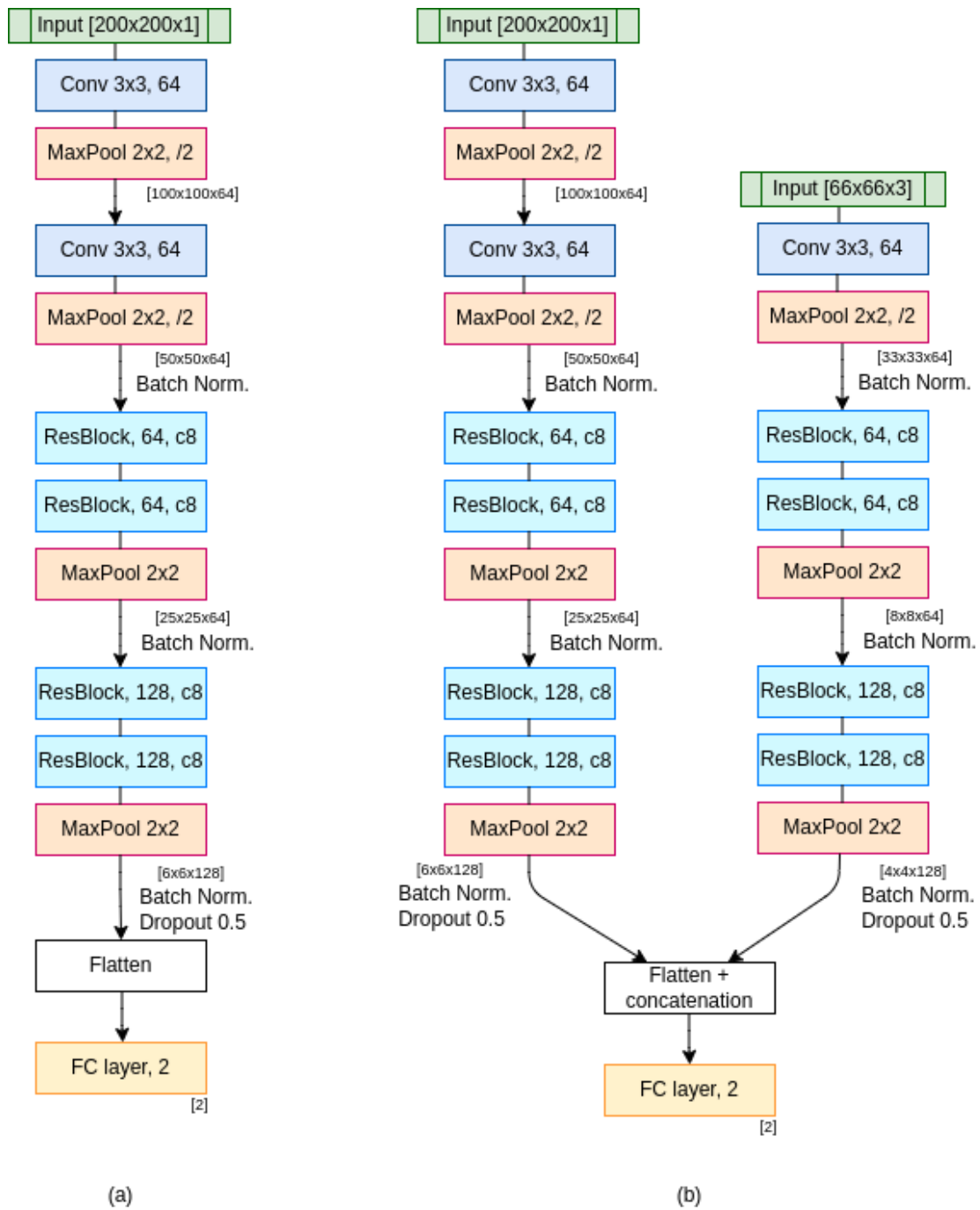
Fig. A.3. Residual network configurations tested on (a) VIS images and (b) multiband images. These diagrams use the same notation as in Fig. A.1. Every residual block (ResBlock) is built as described in subsection 2.1.3: c8 refers to the cardinality of the block, which we set to be equal to eight.

Appendix B: Tables

In this appendix, we summarize the main results of our tests. In Table B.1 we show the results of training our models on VIS images, in Table B.2 we compare the results of applying our models trained on S2 and on S4 to the test set S4, in Table B.3 we show the results of two additional tests, S2/S3, and S2/S4, in Table B.4 we show the results of training our models on multiband images, and in Table B.5 we present the results of a test with realistic proportions between lenses and nonlenses.

Table B.1. Summary of the performance of the VGG-like network, the IncNet, and the ResNet in classifying the objects of the four selections of images in the I_{E} band.

	VGG-like network							
	S1		S2		S3		S4	
Class	0	1	0	1	0	1	0	1
Precision	0.95	0.98	0.94	0.97	0.92	0.94	0.79	0.89
Recall	0.98	0.94	0.98	0.94	0.94	0.92	0.90	0.77
F1-score	0.96	0.96	0.96	0.96	0.93	0.93	0.84	0.83
Accuracy	0.96		0.96		0.93		0.84	
AUC	0.77		0.58		0.88		0.57	
	Inception Network							
	S1		S2		S3		S4	
Class	0	1	0	1	0	1	0	1
Precision	0.97	1.0	0.97	0.96	0.94	0.93	0.84	0.90
Recall	1.0	0.96	0.96	0.97	0.93	0.94	0.91	0.83
F1-score	0.98	0.98	0.96	0.96	0.93	0.94	0.87	0.86
Accuracy	0.98		0.96		0.94		0.87	
AUC	0.92		0.88		0.90		0.81	
	Residual Network							
	S1		S2		S3		S4	
Class	0	1	0	1	0	1	0	1
Precision	0.93	0.97	0.90	0.92	0.86	0.89	0.71	0.84
Recall	0.97	0.92	0.92	0.89	0.89	0.85	0.87	0.66
F1-score	0.95	0.94	0.91	0.91	0.88	0.87	0.78	0.74
Accuracy	0.95		0.91		0.87		0.76	
AUC	0.81		0.85		0.79		0.70	

Notes. The precision, recall, and F1-score are evaluated on the class of the nonlenses (0) and of the lenses (1) separately, while accuracy and AUC are global quantities.

Table B.2. Comparison between the metrics of tests on the selection S2 with the models trained on S2 (top) and on S4 (bottom).

	S2/S2					
	VGG-like network		**Inception Network**		**Residual Network**	
Class	0	1	0	1	0	1
Precision	0.94	0.97	0.97	0.96	0.90	0.92
Recall	0.98	0.94	0.96	0.97	0.92	0.89
F1-score	0.96	0.96	0.96	0.96	0.91	0.91
Accuracy	0.96		0.96		0.91	
AUC	0.58		0.88		0.85	
	S4/S2					
	VGG-like network		**Inception Network**		**Residual Network**	
Class	0	1	0	1	0	1
Precision	0.96	0.74	0.99	0.77	0.95	0.67
Recall	0.89	0.91	0.90	0.98	0.85	0.89
F1-score	0.93	0.82	0.94	0.86	0.90	0.76
Accuracy	0.89		0.92		0.86	
AUC	0.51		0.88		0.75	

Notes. Class 0 refers to the nonlenses, while class 1 refers to the lenses.

Table B.3. Summary of the performance of the VGG-like network, the inception network, and the residual network, trained on the selection S2, in classifying the objects that are part of the selections S3 and S4.

	VGG-like network				**Inception Network**				**Residual Network**			
	S2/S3		S2/S4		S2/S3		S2/S4		S2/S3		S2/S4	
Class	0	1	0	1	0	1	0	1	0	1	0	1
Precision	0.77	0.97	0.62	0.95	0.82	0.96	0.65	0.93	0.75	0.88	0.64	0.85
Recall	0.98	0.68	0.98	0.33	0.97	0.76	0.97	0.42	0.92	0.67	0.94	0.40
F1-score	0.86	0.80	0.76	0.48	0.89	0.85	0.78	0.58	0.83	0.76	0.76	0.55
Accuracy	0.83		0.68		0.87		0.71		0.80		0.69	
AUC	0.57		0.52		0.81		0.7		0.78		0.65	

Notes. The precision, recall, and F1-score are evaluated on the class of the nonlenses (0) and of the lenses (1) separately.

Table B.4. Same as in Table B.1, but using images in the VIS and NISP bands.

	VGG-like network							
	S1		S2		S3		S4	
Class	0	1	0	1	0	1	0	1
Precision	0.99	0.97	0.98	0.97	0.91	0.96	0.81	0.91
Recall	0.97	0.99	0.97	0.98	0.96	0.91	0.92	0.79
F1-score	0.98	0.98	0.98	0.98	0.94	0.93	0.86	0.84
Accuracy	0.98		0.98		0.93		0.85	
AUC	0.65		0.87		0.67		0.62	
	Inception Network							
	S1		S2		S3		S4	
Class	0	1	0	1	0	1	0	1
Precision	0.98	0.96	0.97	0.98	0.96	0.96	0.87	0.91
Recall	0.96	0.98	0.98	0.96	0.96	0.96	0.91	0.87
F1-score	0.97	0.97	0.97	0.97	0.96	0.96	0.89	0.89
Accuracy	0.97		0.97		0.96		0.89	
AUC	0.77		0.9		0.92		0.84	
	Residual Network							
	S1		S2		S3		S4	
Class	0	1	0	1	0	1	0	1
Precision	0.96	0.95	0.92	0.94	0.86	0.92	0.74	0.85
Recall	0.94	0.96	0.94	0.92	0.92	0.87	0.87	0.71
F1-score	0.95	0.95	0.93	0.93	0.90	0.89	0.80	0.77
Accuracy	0.95		0.93		0.90		0.78	
AUC	0.81		0.88		0.81		0.72	

Table B.5. Results of testing our best-performing networks, trained on S1, on a test set with 200 lenses and 80 000 nonlenses.

	VGG-like network		Inception Network		Residual Network		Ensemble Network	
Class	0	1	0	1	0	1	0	1
Precision	1.0	0.15	1.0	0.45	1.0	0.13	1.0	0.46
Recall	0.98	0.94	0.99	0.96	0.98	0.92	1.0	0.97
F1-score	0.99	0.26	0.99	0.61	0.99	0.23	1.0	0.63
Accuracy	0.98		0.99		0.98		1.0	
AUC	0.76		0.83		0.81		0.99	

Notes. Class 0 refers to the nonlenses, while class 1 refers to the lenses. Ensemble network refers to the combination of the predictions of the three networks.