Astronomy & Astrophysics manuscript no. output
October 18, 2023

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], H. Domínguez Sánchez[133]

(Affiliations can be found after the references)

Received xxx; accepted yyy

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 the 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 implement a classical CNN architecture, an inception network and a residual network. We train and test our networks on different subsamples of a data set of forty thousand mock images, having characteristics similar to those expected in the wide survey planned with the ESA mission *Euclid*, gradually including larger fractions of faint lenses. We also evaluate the importance of adding information about the color difference between the lens and source galaxies by repeating the same training on single-band and multi-band images. Our models find samples of clear lenses with $\gtrsim 90\%$ precision and completeness. Nevertheless, when including lenses with fainter arcs in the training set, the three models' performance deteriorates with accuracy values of ~ 0.87 to ~ 0.75 depending on the model. Specifically, the classical CNN and the inception network have similar performances 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 investigate what the optimal training strategy for this specific survey is to exploit the scientific potential of the upcoming observations fully. We suggest that training the networks separately on lenses with different morphology might be needed for identifying the faint arcs. We also test how relevant the color information is for the detection of these systems, and we find that it does not yield a significant improvement, with the accuracy ranging from ~ 0.89 to ~ 0.78 for the different models. This result might be due to the lower resolution of the *Euclid* telescope in the infrared bands, with respect to that of the the images in the visual band.

Key words. 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.

Such 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 lens' gravitational potential (e.g., stellar kinematics), one can disentangle the dark and baryonic mass distributions, thus studying 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, which 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 to stress-test the CDM paradigm (see also Meneghetti et al. 2022; Ragagnin et al. 2022). Modeling such 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).

Less than one thousand 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 submm wavelengths (Negrello et al. 2010, 2017), and sources with unusual shapes (Myers et al. 2003). Some arc and ring finders have been developed to analyse optical images, and they typically look 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 due to their rarity, but it is expected that this will change in the next decade, thanks to upcoming imaging surveys. In fact, 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 et al. 2009) performed with the Vera C. Rubin Observatory will observe more than one hundred thousand strong lenses (Collett 2015), thus significantly increasing the number of known systems. Producing such large and homogeneous catalogs of GGSL systems 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 one, 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 do not have access to, to ensure that the learning is not leading to overfitting. The validation happens at the same time as 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 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 for estimating the photometric redshifts of luminous sources (e.g., Pasquet et al. 2019; Shuntov et al. 2020; Li et al. 2022), for performing 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), for constraining the cosmological parameters (e.g., Merten et al. 2019; Fluri et al. 2019; Pan et al. 2020), for identifying cluster members (e.g., Angora et al. 2020), for finding galaxy-scale strong lenses in galaxy clusters (e.g., Angora et al. 2023), for quantifying galaxy metallicities (e.g., Wu & Boada 2019; Liew-Cain et al. 2021), and for estimating galaxy cluster dynamical masses (e.g., Ho et al. 2019; Gupta & Reichardt 2020). Recently, O'Riordan et al. (2023) also tested the use of CNNs for detecting subhalos in simulated *Euclid*-like galaxy-scale strong lenses.

Several CNN architectures were recently used also 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; Napolitano et al. 2020; Li et al. 2021), 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; The Dark Energy Survey Collaboration 2005; Jacobs et al. 2019b,a; 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 not. The first challenge results, presented in Metcalf et al. (2019), show 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 in the identification of GGSL systems. We test them on different subsamples of a data set of *Euclid*-like mock observations. In particular, we evaluate how including faint lenses in the training set affects 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; in Sect. 4, we describe our experiments and we present and discuss our results. In Sect. 5, we summarise our conclusions.

⋆ e-mail: laura.leuzzi3@unibo.it

2. Convolutional neural networks

Artificial neural networks (ANNs; e.g., McCulloch & Pitts 1943; Goodfellow et al. 2016) are an ML algorithm inspired by the biological functioning of the human brain. They consist of artificial neurons, or nodes, that are organised 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 k-th layer $\boldsymbol{h}^k$ depends on the output of the previous layer $\boldsymbol{h}^{k-1}$ (Bengio 2009)

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

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

CNNs are a special class of ANNs that use the convolution operation. Thanks to 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 of typical dimension 3×3 to 7×7 and act as the weights of a generic ANN. They are convolved with the layer's input to produce the feature maps. The feature maps are passed through an activation function, that introduces non-linearity in the network and then they are 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 of typical dimension 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 1-D vector that is processed by fully-connected layers and is then passed to the output layer which 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 $P(Y = i \,|\, \boldsymbol{x})$, where Y is the class associated with the input $\boldsymbol{x}$, among all the possible classes i.

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 initialised, are readjusted so that the network's output predictions are correct for the largest number of possible examples. This step is crucial since the weights are not modified afterward when the final model is applied to other data. The training aims to minimise 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 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 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 generalises to other images. Preventing the model from overfitting (i.e., memorising 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's performance is assessed on the validation set, a small part of the data set (usually $5-10\%$) 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 also a technique used to mitigate overfitting. This method consists in randomly dropping units from the network during training, i.e. temporarely removing incoming and outcoming connections from a given node. Once the training is completed, the performance of the final model is evaluated on the test set, a part of the data set (about $20-25\%$) excluded from the other subsets. Afterward, the CNN can be applied to new images.

CNNs handle large data sets conveniently 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, thanks to graphics processing units (GPUs). Moreover, the feature extraction process during the training is 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 implement 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 apply to the images is the result of several trials in which we have tested different hyperparameters for the optimization (such as the learning rate) and general architectures (such as 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 since 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., of dimension 5×5 or 7×7). Since the concatenation of multiple kernels of

size 3×3 has the same resulting receptive field of 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 and five max pooling layers alternating. Let us 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 in the renormalization of the layer inputs (Ioffe & Szegedy 2015) and is employed to accelerate and stabilise 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 multi-band 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. Since they have a smaller size 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 being 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. In Appendix A, Fig. A.1 shows the VGG-like network configuration we tested on the VIS images (panel a) and on the multi-band 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 of the number of parameters of the model, favoring overfitting and increasing the requirements of computational resources. Szegedy et al. (2015) suggest 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, 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 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 being 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 two million.

The configuration used to analyze the multi-band images has a secondary branch with one initial convolutional layer and seven inception modules. This branch is characterised by approximately one million parameters, thus leading to a total of around three million parameters. In Appendix A, Fig. A.2 shows the IncNet configuration we tested on the VIS images (panel a) and the multi-band 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 $\boldsymbol{x}$ 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 $\mathcal{F}(\boldsymbol{x})$ that the block is expected to infer can be written as

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

where $\mathcal{H}(\boldsymbol{x})$ 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 $\mathcal{F}(\boldsymbol{x}) + \boldsymbol{x}$.

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

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

and serves as the residual function in Eq. (2). Here $\mathcal{T}_i(\boldsymbol{x})$ 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 having only one initial convolutional layer.

The parameters of the model are circa one million in the VIS configuration and about two million in the multi-band 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 networks' architecture and this specific setup outperformed the others, including those that had a higher number of weights. In Appendix A, Fig. A.3 shows the ResNet configuration we applied to the VIS images (panel a) and on the multi-band images (panel b).

3. The data set

Training CNNs requires thousands of labeled examples. Since 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, since we do not have real images observed with the *Euclid* telescope yet. The realism of the simulations is essential to ensure that the evaluation of the model's performance is indicative of the results we may expect from real observations.

The image simulations are used to produce all the images in the data set, i.e. both the lenses and non-lenses. We generate all the images and then divide them into the two classes according to the criteria that will be introduced later. The simulations use 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).

We construct the images using the following procedure.

- We randomly select a trial lens galaxy from the light cone subject to a magnitude cut of 23 in the VIS band from the *Euclid* telescope, i.e. the $I_{\rm E}$ band. After this, we randomly select a background source from a catalog of Hubble Ultra Deep Field (UDF; Coe et al. 2006) sources with known redshift. We decompose 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 which is usually not visible above the noise. We investigate the potential impact of these effects on the results of this paper in Sec. 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 use 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 $0\farcs05$, i.e. twice the final resolution of the VIS instrument. We use these rays to compute the deflection angles that will trace the path of the light back to the sources.
- The code detects any caustics in the field and does some further refinement to characterise 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 arcsec^2 and smaller than 20 arcsec^2, 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 take the parameters for the Sérsic profiles from the Flagship catalog with some randomization. While we place the lens galaxy at the center of the cut-out, the positions of the other galaxies are 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 will have a different angular position.
- We place the background source galaxy at a random point on the source plane within a circle surrounding the caustic. The radius of the circle is 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 arcsecs and then downsamples to 0.1 arcsecs for VIS and 0.3 arcsecs for the infrared bands. The VIS PSF was derived from modeling the instrument (Euclid collaboration et al., in prep.). For the infrared bands, a simple Gaussian model with a width of 0.3 arcsecs is used. The noise is simulated with a Gaussian random field, to reproduce the noise level expected by the Euclid Wide Survey (Euclid Collaboration: Scaramella et al. 2022).
- To avoid repeating a particular lens and to increase the number of images at a low computational cost, we randomise 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's radius is set to 30 arcsecs at the distance of the lens. In addition, the galaxies outside this sphere, but within the field of view, are independently rotated about the primary in the plane of the sky. The mass associated with each galaxy is moved with the galaxy's image. The position angles of each galaxy are also randomly re-sampled.
- A final step is to classify the images as lenses. Some of the images will have low signal-to-noise in some lensed images or not be distorted enough to be recognizable lenses.

This procedure is similar to the one used for the Lens Finding Challenges and 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 non-linearity, charge transfer inefficiency, and a more intricate PSF model, that are included in other studies (e.g., Pires et al. 2020).

The result of these simulations are one hundred thousand *Euclid*-like mock images simulated in the $I_{\rm E}$ band of the VIS instrument and $H_{\rm E}$, $Y_{\rm E}$ and $J_{\rm E}$ bands of the NISP instrument (Euclid Collaboration: Schirmer et al. 2022). 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.

Table 1. Main characteristics of the *Euclid* VIS and NISP (Euclid Collaboration: Schirmer et al. 2022) instruments.

Instrument	Capability	λ range (nm)	Pixel size (arcsec)
VIS	Visual imaging	$I_{\rm E}$ (530–920)	0.1
NISP	NIR imaging	$Y_{\rm E}$ (949.6–1212.3),	0.3
	photometry	$J_{\rm E}$ (1167.6–1567.0),	0.3
		$H_{\rm E}$ (1521.5–2021.4)	0.3

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

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

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:

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

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 unlensed image's pixels 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 non-lenses, 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 low signal-to-noise images might vary, while the clearest ones should be immediately assigned to the correct category.

By using these conditions, we divide the images we simulated into 19 591 lenses and 79 816 non-lenses, 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 use all the lenses for the training, and we randomly select only a subsample of 20 000 non-lenses. As will be discussed in Sec. 4.1, these numbers are increased by data augmentation. We refer to the non-lenses 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 non-lenses (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 sources' redshift distribution, in the top-right panel, is also similar for the two subsets. On the other hand, the simulated sources (middle-right panel) in the non-lenses class are on average fainter than that of the sources in the lenses. This is intuitive, since sources with lower magnitudes (i.e. brighter) 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 will be the distortion of the source's images, hence the lensing system.

[2] http://metcalf1.difa.unibo.it/blf-portal/gg_challenge.html

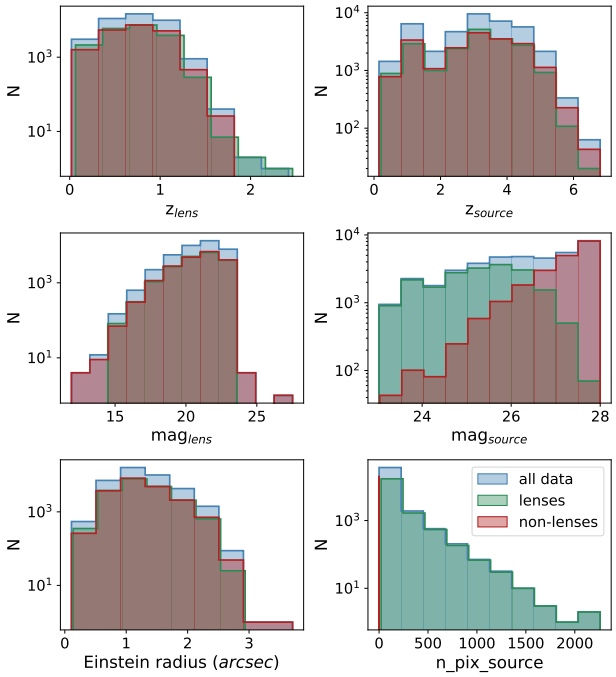

Fig. 1. Distribution of several properties of the simulated images in the data set (blue histograms) selected for training, that consists of 40 000 mocks in total. The distributions of the same properties in the separate subsets of lenses and non-lenses are given by the green and red histograms, respectively. In the upper- and middle-row panels, 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 where the source brightness exceeds 1σ above the background noise level.

4. Results and discussion

4.1. Data preprocessing

The data preparation consists of a sequence of several steps. We divide 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 are randomly assigned to one of these subsets, but we checked that all subsets (training, validation, and testing) are representative of the entire data set. We do 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.

Once the data set is split, we randomly select 20% of the images in the training set for augmentation. We perform five augmentations: we rotate these images by 90°, 180°, and 270° and flip them with respect to the horizontal and vertical axes. After performing these operations the size of the training set is doubled. Neither the test set nor the validation set are augmented.

Afterward, we proceed with the normalization of the images in the data set. We subtract the mean and divide them 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 if 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 implement, train and test 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 use the Adaptive moment estimation (Adam; Kingma & Ba 2017; Reddi et al. 2019) optimizer with initial learning rate of 10^{-4}. We employ the binary cross-entropy $\mathcal{L}$ to estimate the loss at the end of each epoch:

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

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 i-th example has label 1 as predicted by the network so that $1 - y_p$ is the probability that the i-th example has 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 twenty 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 use the best models, i.e. those that have the minimum value of the loss function on the validation set, for our tests.

4.3. Performance evaluation

We assess 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 take into consideration four statistical metrics that are 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, i.e. the True Positives (TP) and the True Negatives (TN), while the off-diagonal terms show the number of misclassified objects, i.e. the False Positives (FP) and the 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

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{6}$$

  which measures the level of purity of the retrieved catalog.
- The recall (R) can be computed as

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{7}$$

  which measures the level of completeness of the retrieved catalog.
- The F1-score (F1) is the harmonic average of P and R,

$$\mathrm{F1} = 2\,\frac{P\,R}{P + R}. \tag{8}$$

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

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{9}$$

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 compute 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 is used to discriminate whether an image contains a lens or not. The area under the ROC curve (AUC) summarises 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 aim at evaluating the ability of CNNs to detect the less clear lenses and at assessing their performance on a diversified data set.

We do this by training the three networks we have presented on four selections of images, named from S1 to S4, which gradually include a greater fraction of objects that present challenging visual identification, as we will discuss shortly for non-lenses and lenses separately. These samples consist of approximately two thousand, ten thousand, twenty thousand images, and forty thousand images, respectively. They are built to have an approximately equal number of lenses and non-lenses (see Table 2). The criteria we adopt to progressively broaden our selections take into account the features that might be employed by the networks to classify the objects as members of the correct category.

In the case of the non-lenses, the lack of a background source, or the absence of its images, makes the classification more likely to be correct. Therefore, we initially consider a sample of the approximately ten thousand non-lenses without a background source. Specifically, we select one thousand of them in S1, five thousand in S2, and ten thousand in S3. In S4, we broaden our sample by including the images where a background source has been added but does not correspond to a visible image, extending our selection to the other objects that are classified as non-lenses 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, as well as more challenging. In fact, the mere presence of an image of the source does not guarantee a straightforward classification of the system, since several factors contribute to the actual clarity of the observable features. Among them are the magnitude of the source and the extension of the image produced by the lensing effect. After several tests involving these parameters and others (such as the Einstein area and the magnification of the sources), we deem `n_pix_source` to be an appropriate parameter to discriminate between clear and faint lenses. The complete sample of lenses is characterised by the minimum value `n_pix_source` > 20. From S4 to S1, we increase this threshold to different levels, which depend on the number of images we

[3] https://keras.io/
[4] https://www.tensorflow.org/
[5] https://keras.io/api/callbacks/early_stopping/

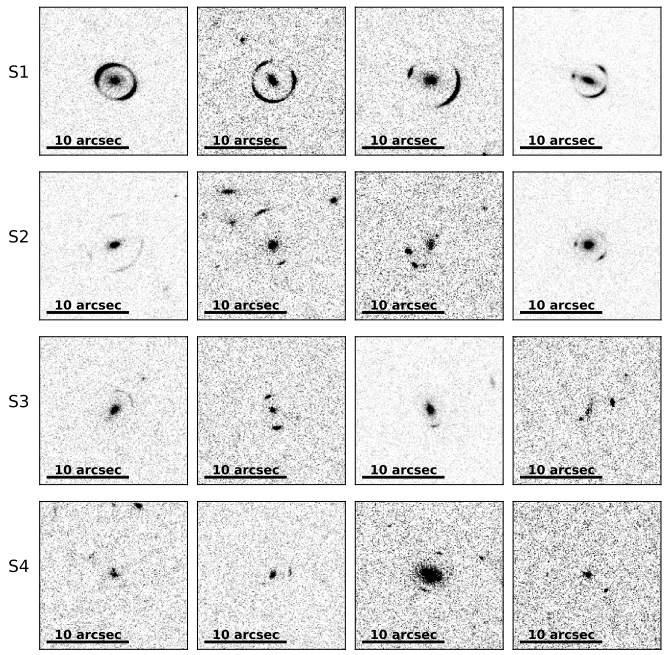

Fig. 2. From top to bottom row, we show four random lenses extracted respectively from the data sets S1, S2, S3, and S4, as simulated in the I_E band. We see from these images that the effect of using lower thresholds of the parameter `n_pix_source` is to select fainter lensing systems. While in the data set S1, most of the lenses are characterised by clear rings and distortions, when we move to S4 we find many examples of fainter, barely visible arcs.

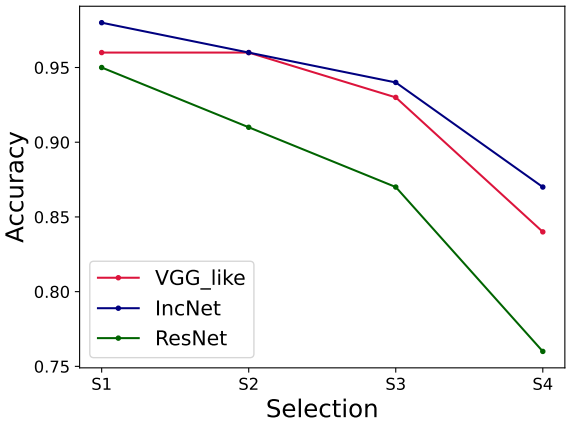

Fig. 3. Trend of the classification accuracy of the single-branch versions of the VGG-like network (red), the IncNet (blue) and the ResNet (green) tested on the four selections of data.

seek to isolate: the higher the value employed, the smaller will be the number of images selected 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 give a summary of the criteria 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 by considering different thresholds for `n_pix_source` in the definition of the training sets.

We train and test on these selections of the data set the three architectures, previously discussed: 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 conduct twenty-four training sessions in total, since we train each architecture on each selection of data. Twelve of them use the VIS images, the other twelve use the NISP bands in addition to the VIS one. Every training was carried out for 100 epochs, since 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 are conducted using the I_E band images, are summarised 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 multi-band images is 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 low signal-to-noise lenses increases. All the results described in the paper are found 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, the ResNet is the worst-performing architecture, with an accuracy of ~ 0.75 on the complete data set.

The precision, recall, and F1-score also have similar global trends to that of the accuracy. They are shown in the top, middle, and bottom panels of Fig. 4, respectively. These metrics are evaluated separately on the non-lenses (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 it actually 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, with the ResNet being 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, with the AUC decreasing by $\sim 10\%$ from S1 to S4. It should, however, be pointed out that the IncNet performs systematically better than the ResNet: while the former has an AUC of 0.92 on S1 and 0.81 on S4, the latter has AUC that 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 carefully checking the predictions of this network on the different selections, we think this is due to a significant difference in the number of objects predicted in the two classes when applying a high threshold to the output probabilities.

Let us focus on the selection S4, i.e. on the performance of our models on the complete data set. We can see in Fig. 6 nine misclassified non-lenses and in Fig. 7 nine misclassified lenses. The images reported in these figures are selected among those misclassified by all three models, therefore they should be char-

Table 2. Summary of the criteria adopted to choose the images included in the different selections of lenses and non-lenses for our experiments. 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 non-lenses is primarily based on the possible presence and visibility of a background source.

Selection	**Lenses**		**Non-lenses**		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

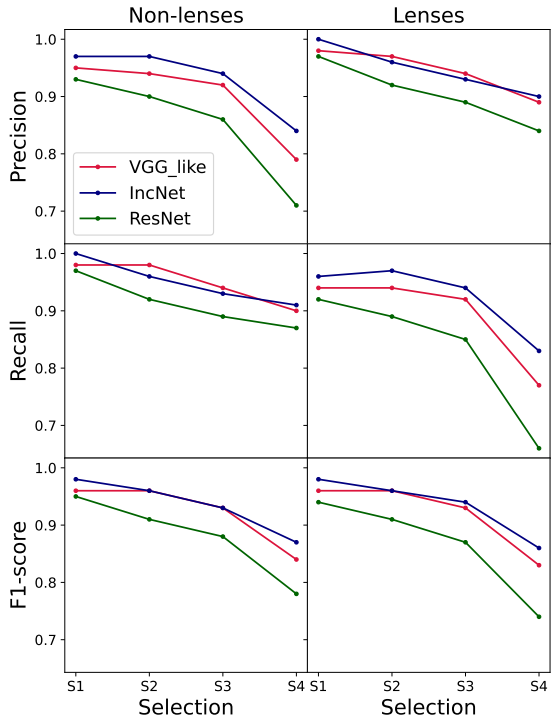

Fig. 4. Trend of the precision (first row), recall (second row), and F1-score (third row) in the classification of the non-lenses (left column) and of the lenses (right column) in the different selections. Different colored lines refer to different networks, as labeled, in the single-branch configuration.

acterised by the features that the networks generally find harder to attribute to the correct class.

The false positives in Fig. 6 are mostly characterised 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. In fact, 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 non-lensed galaxies, such as spiral arms, or isolated, but elongated galaxies. One possible way to mitigate the misclassification of non-lenses with a background source could be to train the networks on multiband images, to benefit from the color information. We will investigate this possibility in Sec. 4.8.

The false negatives in Fig. 7 are partly not even recognizable as lenses by visual inspection. Although being classified as lenses according to the criteria in Eq. (4), many of these objects do not show evident lensing features. Therefore, if the classification was to be carried out on unlabeled observations, we would not expect the models to be able to identify them as lenses. An approach to solve the issue of having non-detectable 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 as lenses those systems 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, signaling 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 actually identify clear lenses, we test them on the images in S2 (test S4/S2). Before doing so, we make sure to remove from S2 the images that the networks trained on S4 have analyzed during training and validation. We do this because otherwise the network performance would be biased to better performance than can be achieved on unseen data. We compare the result of this test to those obtained from training and testing the

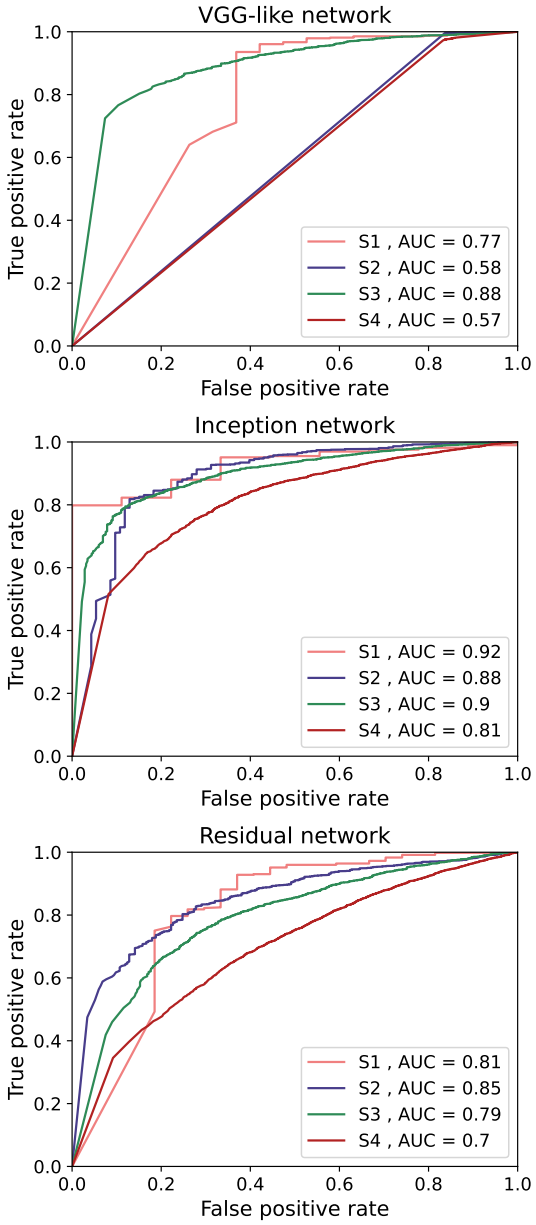

Fig. 5. From top to bottom, each panel of this image shows the ROC curves obtained from the application of the single-branch versions of the VGG-like network, the IncNet, and the ResNet to the test sets of the different selections S1 (pink line), S2 (blue line), S3 (green line) and S4 (red line) of the data set.

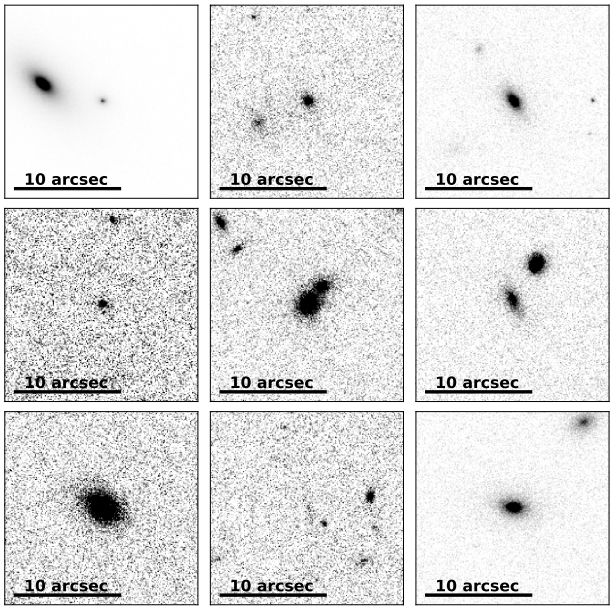

Fig. 6. Example of false positives produced by the three networks in the single branch configuration, when applied to the selection S4, here pictured in the I_E band.

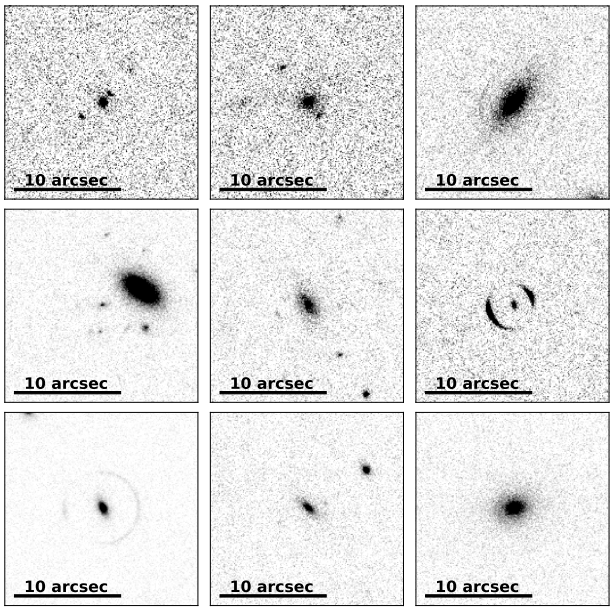

Fig. 7. Example of false negatives produced by the three networks in the single branch configuration, when applied to the selection S4, here pictured in the I_E band.

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 the identification of the lenses in S2 is generally worse than that of the models trained on S2, even though the images that are part of S2 will also inevitably be part of S4 since S4 consists of the complete data set. One reason for this is that the networks used in the test S2/S2 are specifically trained to identify the lenses in S2, while the networks trained on the larger data set S4 have been exposed to a larger variety of systems and are not as specialized on the S2-lenses, but let us look at the results in Table B.2. 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 since 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 since a significant fraction of this selection is composed of non-obvious lenses, that are intrinsically harder to classify. Moreover, there is a deterioration in the ability of the models to recognise the clearest GGSL events in the data set, that are also present in S2.

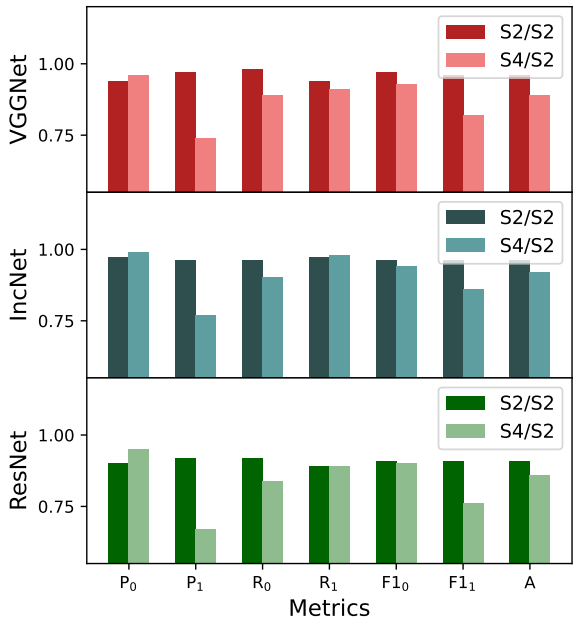

Fig. 8. Comparison of the tests S2/S2 and S4/S2 (darker and lighter histograms, respectively) run with the VGG-like network (top), the IncNet (center), and the ResNet (bottom). In each panel we show the results for the different metrics: from left to right we show the precision on the class of the non-lenses (P_0) and lenses (P_1), the recall on the class of the non-lenses (R_0) and lenses (R_1), the F1-score on the class of the non-lenses ($F1_0$) and lenses ($F1_1$), and the overall accuracy (A).

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 reflects in the fact that the networks might not learn how to properly distinguish them. Wide arcs and rings will be recognizable only in a moderate number of images, thus not being as significant as they are in S2 for the classification of the lenses. Second, the most recurrent features in the training set will be the ones that occur in the low signal-to-noise images, thus contributing 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 classified as non-lenses 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 is extended to include a significant number of fainter arcs, as 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 comparing the metrics in Figs. 3 and 4. The ResNet, on the other hand, performs significantly worse than the others, especially on S4.

4.6. Additional tests

We now test 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 are also included in the training set of S2. This test has the purpose of assessing whether the networks trained on clear examples are flexible enough to detect fainter systems. Also a deterioration of the performance from S2/S3 to S2/S4 is expected, since CNNs mostly generalise to the images that are similar to those in the data set they have been trained with. Consequently, they might perform the same task poorly when dealing with images characterised by features they have never seen 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 $\sim$ 0.85 in the case S2/S3 to $\sim$ 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 non-lenses and lenses, as well as in the accuracy. We report the results of these tests in Table B.3.

The purity of the non-lenses decreases when broader selections are used as test sets: the precision reaches the value of $\sim$ 0.64 with S4. On the other hand, the recall is approximately constant at values of $\sim$ 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, we find a roughly opposite trend. The precision of the classification is roughly constant at $\sim$ 0.94, while the recall decreases drastically from $\sim$ 0.7 in S3 to $\sim$ 0.38 in S4: these values suggest that the networks trained on the S2 sample do not manage to recognise a large fraction of the lenses in the complete data set.

These trends can be interpreted by considering the impact of the inclusion of the fainter features in the test sets. In particular, the training set of S2 mostly includes clear lenses and images of isolated non-lenses, 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 non-lenses. 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 $\sim$ 20% in the recall of the lenses from S2/S2 to S2/S3 and of $\sim$ 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 use the galaxies observed in the UDF as background sources. For the purpose of denoising them, we decompose the galaxies with a shapelet-based approach. The shapelet technique is a very powerful mathematical tool to describe astrophysical objects, and its limitations have been investigated in some works (see e.g., Melchior et al. 2007, 2010). In this Section, we investigate what is the impact of these limitations on the performance of our networks.

We assess 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 non-lensed galaxies of the UDF. The purpose of this test is not to evaluate the performance of our networks on a realistic sample, which would require including a larger number of non-lenses in the test set. We want 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 point to the simulations not being descriptive enough of the characteristics of real galaxies.

We use the networks trained on S2 to carry out this test. We preprocess all the images by normalizing them with a similar procedure to the one we apply to the simulations, described in Sec. 4.1. In the case of the galaxies of the UDF, we also reshape the images to the size expected by the networks.

The results of this test are the following. We recover 129 of the lenses with the IncNet, and 126 of them with the VGG-like Network and with the ResNet. In the case of the non-lensed UDF galaxies, all the three networks correctly classify 296 of them. Given these recovery rates, we reckon that the shapelet decomposition does not introduce significant limitations in our simulations.

4.8. Training with multi-band images

The correct identification of GGSL events may benefit significantly from color information emerging from the analysis of multi-band data. Indeed, lenses and sources typically have different colors, due to their different spectral energy distributions (and redshifts). For example, the most common sources are star-forming galaxies that appear bluer than the lenses, which on the contrary 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 those systems that do not present evident morphological distortions.

For example, Gentile et al. (2022) find that training CNNs on multi-band images results in an improved classification of systems that have small Einstein radii while training on single-band images is more efficient for finding lenses with large radii. Also Metcalf et al. (2019) find that using multi-band images for training substantially improves the performance of the classifiers, when dealing with mock ground-based data, even though the color information comes from observations with poorer spatial resolution.

We evaluate the importance of color information for the identification of the low `n_pix_source` lenses in *Euclid*-like data by repeating the same training discussed so far, this time including the NIR images, 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 change the architecture of our models to take into account the different sizes of the VIS and NISP images, as explained in Sect. 2 and represented in the panels (b) of Figs. A.1, A.2 and A.3, but otherwise keep the same setup as in our previous experiments. We report the results of these tests in Table B.4.

By comparing these values to those of the VIS training (see Table B.1), we do not observe a significant improvement in the models' performance when training with multi-band data. This is expected for the smaller selections, limited to the clearest lenses, whose correct identification through their morphology is relatively easy. Thus, in these cases, color information is expected to be less relevant. However, when looking at the broader selections, in which the morphology of the lenses is less clear, we might expect to see some improvement in the classification performance when feeding the models with color information. Surprisingly, we do not notice 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, i.e. the blue wavelengths of the optical spectrum. Secondly, the images in the NIR bands are characterised by lower resolution than those in the I_E band (also see Table 1), thus the morphological information is degraded in these channels. This also suggests that morphological information is more important than color when identifying lenses, at least in this wavelength range.

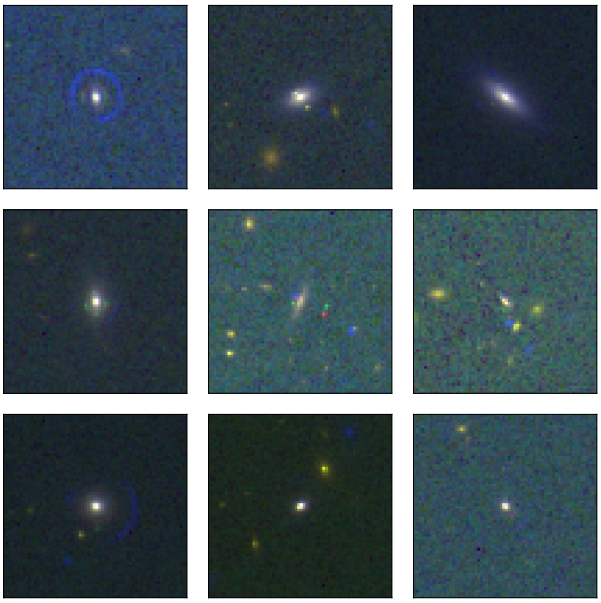

Fig. 9. Example of randomly chosen lenses in the configuration used for multi-band training. For visualization pruposes, the images simulated in the I_E band were downgraded to the resolution of the NISP bands in these examples.

4.9. Finding lenses in unbalanced data sets

As we discussed in Sec. 3, training on a balanced training set is important for the networks to learn how to assign the images to the correct class, but having a balanced test set is not a requirement. In fact, while in all the previous tests we used a balanced test set, with a ratio of around 1:1 between lenses and non-lenses, this is very different from reality, where we reasonably expect to observe less than one lens for 1000 non-lenses (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 will be required to find definite samples of strong lenses. While training on simulations instead of real observations plays a role in this since it is possible that the images will 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 run an additional test with realistic proportions in the number of images of the two subsamples. We focus on the networks trained on S1, that globally have the best performances (Figs. 3 and 4). We apply the networks trained on this selection on a test set that has the same lenses as in the original test set of S1, i.e. 240 lenses, and use the ~ 80 000 non-lenses that we had 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 combine 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 has indeed 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 on 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 having 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, as the systems that are classified with smaller probability will be 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 could improve the classification performance with real data (Domínguez Sánchez et al. 2019; Ćiprijanović et al. 2022). These techniques would require re-training networks that were trained on simulations on a small sample (a few hundreds) of observed lenses and might lead to a significant improvement of the networks performance.

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 essential also for efficient follow-up observations. For example, the 4-meter 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, given the small number of galaxy-galaxy lenses known at the moment. As the data is accumulated, more sophisticated simulations will be done, where the lenses are real galaxies observed by *Euclid*. The networks will be re-trained 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 comes with the challenge of measuring them with good accuracy. Having a large enough separation between the lens galaxy and the source or using efficient de-blending techniques would be decisive in this context.

The greatest advantage of searching for lenses with *Euclid* is that it will resolve faint Einstein rings with small radii ($\sim 0.5''$), mostly lensed by bulges of spiral galaxies, in addition to lenses with larger angular scale. These systems are usually unresolved by ground-based facilities, but will be found thanks to the high resolution of *Euclid*. Moreover, they will be the most common according to forecasts (Collett 2015). *Euclid* observations could also be combined with and complemented by those of other surveys. 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 that will have high resolution will be produced by Square Kilometer Array (Dewdney et al. 2009). They are complementary to the others since 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 that the strong lenses are a biased subsample of. Moreover, they showed how to use the information about the number of non-detections to constrain models of galaxy structure further. More recently, Sonnenfeld et al. (2023) investigated the difference between lens galaxies and lensed sources from their parent population, i.e. the strong lensing bias. Given that *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 did this by using a data set of forty thousand images simulated by the Bologna Lens Factory to mimic the data quality expected by the *Euclid* space mission. The classification was primarily based on the morphology of the systems since we mainly conducted our experiments with the images simulated in the $I_{\rm E}$ band. Still, we evaluated the importance of color information using multi-band images. We trained and tested our CNNs on four data set selections that gradually include a greater fraction of objects characterised by faint lensing features and will be more difficult to recognise. We evaluated the outcome of the classification by estimating the precision, recall, and F1 score of the catalogs of obtained lenses.

We found that the morphological characteristics of the lenses included in the training set influence in a critical way the ability of our CNNs to identify the lenses in a separate test set, 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 $\sim 10\%$, 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 since they become under-represented in the training set.

These results emphasise the importance of building realistic training sets for DL models. This is particularly relevant for the first searches since 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 up to now for training the networks. In particular, they suggest that identifying lenses with

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

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 tackled as a multiclass classification problem, distinguishing the clear and probable lenses from the probable and evident non-lenses. In this last case, however, the distinction between obvious and non-obvious objects should be further investigated and quantified.

We also retrain our models on the same selections of the data set, including a separate channel for processing the NIR images in addition to those in the I_E band, thus assessing how relevant the color information is for identifying the low signal-to-noise lenses. We do not find a significant improvement in the performance of any of our networks. We suggest that this might depend on a combination of two factors: firstly, the images in the I_E band have higher resolution than those in the NIR bands. Secondly, the I_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 applied to the same selections of images. The only exception is the ResNet, whose accuracy on the full data set is $\sim 10\%$ worse than the others. Because of the higher precision the IncNet has on the test with an unbalanced number of images, we would conclude this is the best-performing network among those we tested. The results of this test are, indeed, the closest to what we might expect from real data, hence particularly relevant for the evaluation of 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 characterise better the kind of systems that are most likely to be found or missed.

Acknowledgements. The authors acknowledge the Euclid Consortium, the European Space Agency, and a number of agencies and institutes that have supported the development of *Euclid*, in particular the Academy of Finland, the Agenzia Spaziale Italiana, the Belgian Science Policy, the Canadian Euclid Consortium, the French Centre National d'Etudes Spatiales, the Deutsches Zentrum für Luft- und Raumfahrt, the Danish Space Research Institute, the Fundação para a Ciência e a Tecnologia, the Ministerio de Ciencia e Innovación, the National Aeronautics and Space Administration, the National Astronomical Observatory of Japan, the Netherlandse Onderzoekschool Voor Astronomie, the Norwegian Space Agency, the Romanian Space Agency, the State Secretariat for Education, Research and Innovation (SERI) at the Swiss Space Office (SSO), and the United Kingdom Space Agency. A complete and detailed list is available on the *Euclid* web site (`http://www.euclid-ec.org`). We acknowledge support from the grants PRIN-MIUR 2017 WSCC32, PRIN-MIUR 2020 SKSTHZ and ASI n.2018-23-HH.0. MM was supported by INAF Grant "The Big-Data era of cluster lensing". This work has made use of CosmoHub. CosmoHub has been developed by the Port d'Informació Científica (PIC), maintained through a collaboration of the Institut de Física d'Altes Energies (IFAE) and the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT) and the Institute of Space Sciences (CSIC & IEEC), and was partially funded by the "Plan Estatal de Investigación Científica y Técnica y de Innovación" program of the Spanish government.

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, arXiv:2303.00769
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, vol. 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
Cañameras, R., Schuldt, S., Shu, Y., et al. 2021, A&A, 653, L6
Cañameras, R., Schuldt, S., Suyu, S. H., et al. 2020, A&A, 644, A163
Cabanac, R. A., Alard, C., Dantel-Fort, M., et al. 2007, A&A, 461, 813
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
Cañameras, R., Schuldt, S., Shu, Y., et al. 2023, arXiv e-prints, arXiv:2306.03136
Chollet, F. 2015, keras, https://github.com/fchollet/keras
Ćiprijanović, A., Kafkes, D., Snyder, G., et al. 2022, Machine Learning: Science and Technology, 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, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8442, Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave, ed. M. C. Clampin, G. G. Fazio, H. A. MacEwen, & J. Oschmann, Jacobus M., 84420V
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 Proceedings, 97, 1482
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2019, MNRAS, 484, 93
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, MNRAS, 476, 3661
Euclid Collaboration: Scaramella, R., Amiaux, J., Mellier, Y., et al. 2022, A&A, 662, A112
Euclid Collaboration: Schirmer, M., Jahnke, K., Seidel, G., et al. 2022, A&A, 662, A92
Fluri, J., Kacprzak, T., Lucchi, A., et al. 2019, Phys. Rev. D, 100, 063514
Gavazzi, R., Marshall, P. J., Treu, T., & Sonnenfeld, A. 2014, ApJ, 785, 144
Gavazzi, R., Treu, T., Marshall, P. J., Brault, F., & Ruff, A. 2012, ApJ, 761, 170
Gavazzi, R., Treu, T., Rhodes, J. D., et al. 2007, ApJ, 667, 176
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 Conference on Computer Vision and Pattern Recognition (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, Vol. 37, Proceedings of the 32nd International Conference on Machine Learning, ed. F. Bach & D. Blei (Lille, France: PMLR), 448
Jacobs, C., Collett, T., Glazebrook, K., et al. 2019a, ApJS, 243, 17
Jacobs, C., Collett, T., Glazebrook, K., et al. 2019b, MNRAS, 484, 5330
Jacobs, C., Glazebrook, K., Collett, T., More, A., & McCarthy, C. 2017, MNRAS, 471, 167
Jauzac, M., Klein, B., Kneib, J.-P., et al. 2021, MNRAS, 508, 1206
Kingma, D. P. & Ba, J. 2017, arXiv:1412.6980
Koopmans, L. V. E., Browne, I. W. A., & Jackson, N. J. 2004, New A Rev., 48, 1085
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
LeCun, Y., Boser, B., Denker, J. S., et al. 1989, Neural Computation, 1, 541
Li, R., Napolitano, N. R., Feng, H., et al. 2022, A&A, 666, A85
Li, R., Napolitano, N. R., Spiniello, C., et al. 2021, ApJ, 923, 16
Li, R., Napolitano, N. R., Tortora, C., et al. 2020, ApJ, 899, 30
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:1312.4400
LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201

Maciaszek, T., Ealet, A., Gillard, W., et al. 2022, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 12180, Space Telescopes and Instrumentation 2022: Optical, Infrared, and Millimeter Wave, ed. L. E. Coyle, S. Matsuura, & M. D. Perrin, arXiv:2210.10112
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, Bulletin of Mathematical Biophysics, 5, 115
Melchior, P., Böhnert, A., Lombardi, M., & Bartelmann, M. 2010, A&A, 510, A75
Melchior, P., Meneghetti, M., & Bartelmann, M. 2007, A&A, 463, 1215
Meneghetti, M., Davoli, G., Bergamini, P., et al. 2020, Science, 369, 1347
Meneghetti, M., Melchior, P., Grazian, A., et al. 2008, A&A, 482, 403
Meneghetti, M., Ragagnin, A., Borgani, S., et al. 2022, A&A, 668, A188
Meneghetti, M., Rasia, E., Merten, J., et al. 2010, A&A, 514, A93
Merten, J., Giocoli, C., Baldi, M., et al. 2019, MNRAS, 487, 104
Metcalf, R. B., Meneghetti, M., Avestruz, C., et al. 2019, A&A, 625, A119
Metcalf, R. B. & Petkova, M. 2014, MNRAS, 445, 1942
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., Amber, S., Amvrosiadis, A., et al. 2017, MNRAS, 465, 3558
Negrello, M., Hopwood, R., De Zotti, G., et al. 2010, Science, 330, 800
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, Science China Physics, Mechanics, and Astronomy, 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, On the Convergence of Adam and Beyond, 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, arXiv:2301.13230
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014, Journal of Machine Learning Research, 15, 1929–1958
Stacey, H. R., McKean, J. P., Robertson, N. C., et al. 2018, MNRAS, 476, 5075
Stehman, S. V. 1997, Remote Sensing of Environment, 62, 77
Suyu, S. H., Hensel, S. W., McKean, J. P., et al. 2012, ApJ, 750, 10
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. 2016, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2818
Szegedy, C., Wei Liu, Yangqing Jia, et al. 2015, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1
Tallada, P., Carretero, J., Casals, J., et al. 2020, Astronomy and Computing, 32, 100391
Taufik Andika, I., Suyu, S. H., Cañameras, R., et al. 2023, arXiv e-prints, arXiv:2307.01090
The Dark Energy Survey Collaboration. 2005, arXiv:0510346
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:1505.00853
Zhou, Y.-T. & Chellappa, R. 1988, in IEEE 1988 International Conference on Neural Networks, Vol. 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
[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 Cédex 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, PT1749-016 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, P.O. 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, CH-1211 Genève 4, Switzerland
[58] Department of Physics, P.O. 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, 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, Bat. 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, DK-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, Nova Scotia, 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, FI-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, 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, SE-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 have implemented. In particular, Fig.A.1 shows the VGG-like network, Fig. A.2 shows the IncNet and Fig. A.3 shows the ResNet.

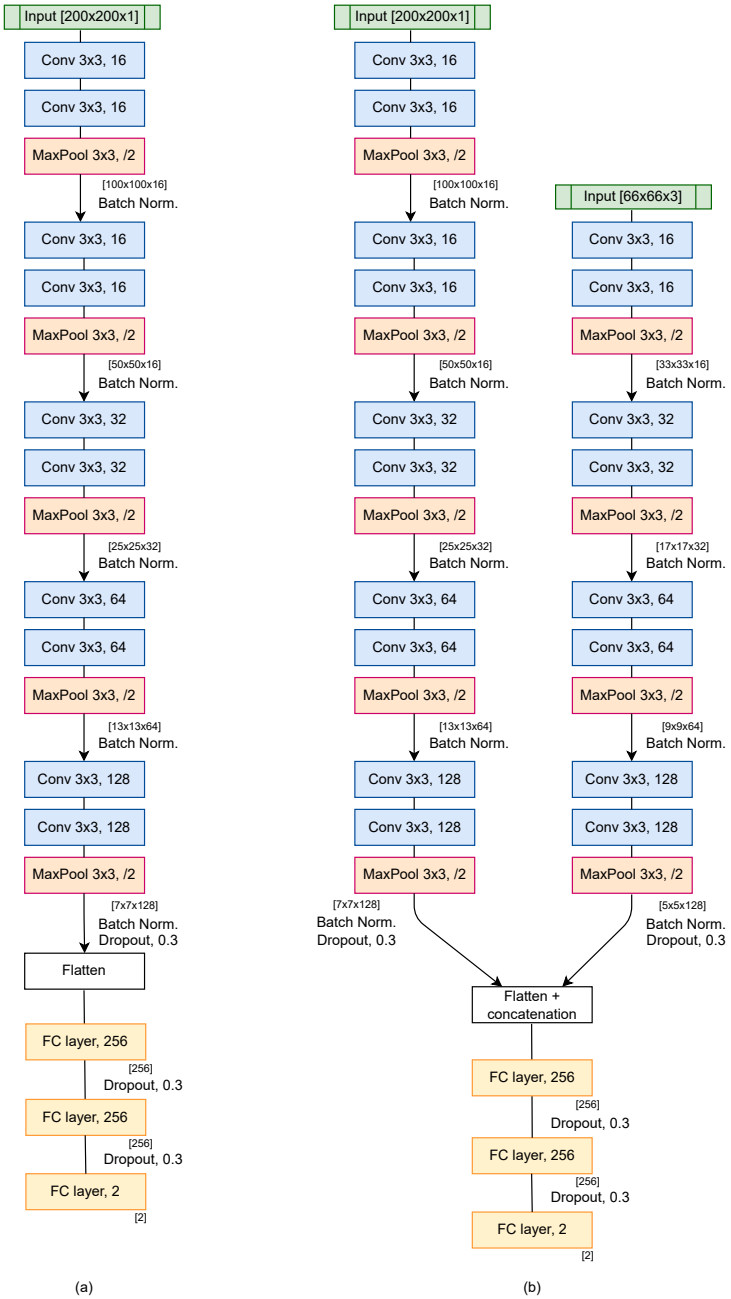

Fig. A.1. VGG-like network configurations tested on (a) VIS images and (b) multi-band images. We report the dimension (D) and number (F) of the filters used in the convolutional layers in the format $D\times D$, F. We also indicate the pooling region (R) and the strides (S) in the pooling layers in the format $R\times R$, $/S$. The numbers in square brackets indicate the dimension and number of the feature maps obtained as the output of the layers in the format [$D\times D\times F$] in the case of the convolutional layers, and the number of nodes in the format [N] in the case of the fully-connected layers.

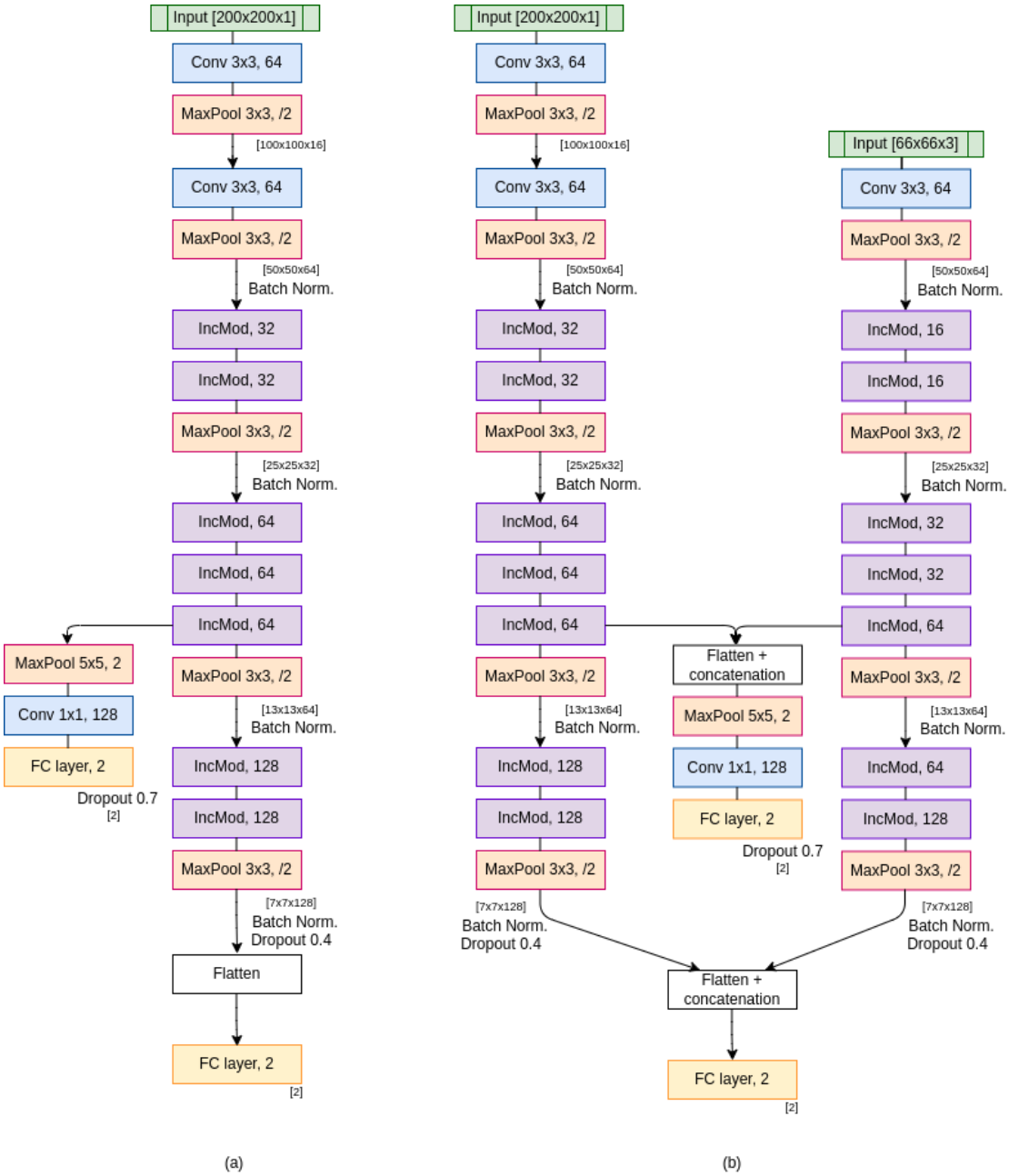

Fig. A.2. Inception Network configurations tested on (a) VIS images and (b) multi-band images. These diagrams use the same notation as those in Fig. A.1. Every inception module (IncMod) is built as described in subsection 2.1.2.

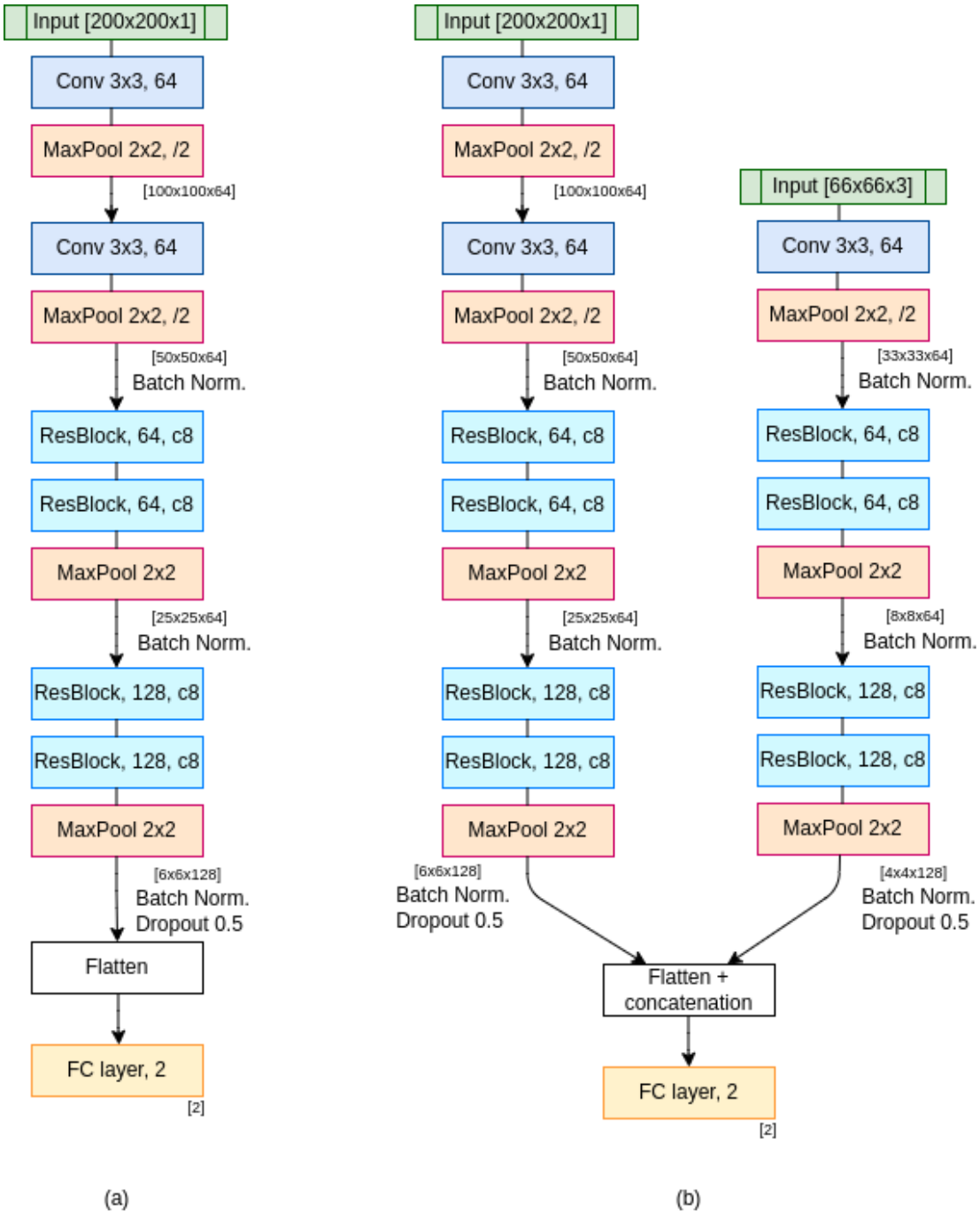

Fig. A.3. Residual Network configurations tested on (a) VIS images and (b) multi-band images. These diagrams use the same notation as those in Fig. A.1. Every residual block (ResBlock) is built as described in subsection 2.1.3, so c8 refers to the cardinality of the block, that we set to be equal to 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 multi-band images; in Table B.5 we present the results of a test with realistic proportions between lenses and non-lenses.

Table B.1. Summary of the performance of the VGG-like network, the IncNet, and the ResNet in the classification of the objects of the four selections of images in the I_E band. The precision, recall, and F1-score are evaluated on the class of the non-lenses (0) and of the lenses (1) separately, while accuracy and AUC are global quantities.

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	

Table B.2. Comparison between the metrics of tests on the selection S2 with the models trained on S2 (top) and on S4 (bottom). Class 0 refers to the non-lenses, while class 1 refers to the lenses.

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	

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 the classification of the objects that are part of the selections S3 and S4. The precision, recall and F1-score are evaluated on the class of the non-lenses (0) and of the lenses (1) separately.

	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	

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

	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	