Mem. S.A.It. Vol. 94, 111
© SAIt 2023

Memorie della

Reconstructing blended galaxies with Machine Learning

L. Nemani[1], A. Fontana[1], E. Merlin[1], and F. Caro[1]

Istituto Nazionale di Astrofisica – Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monte Porzio Catone RM, Italy e-mail: `lavanya.nemani@inaf.it`

Received: 28-11-2022; Accepted: 28-02-2023

Abstract. Galaxy blending is a confusion effect created by the projection of photons from galaxies on the same line of sight, to the 2D plane (Dawson & Schneider 2014). The upcoming deep extragalactic surveys like LSST and Euclid expect to see a blending fraction of up to 50% in the densest regions (Reiman & Göhre 2019). For standard aperture photometry and for more complex techniques such as PSF-fitting and template-fitting algorithms, deblending, the process of reconstructing the individual light profiles from blended sources, becomes crucial.
The current standard deblending algorithms like SExtractor (Bertin & Arnouts 1996) are based on threshold methods that simply assign each pixel to a single object, often failing to correctly take into account the real properties of the blended galaxies. With the advent of Machine Learning (ML) and Computer Vision in Astronomy we want to explore an unbiased and more accurate method of reconstructing individual light profiles using generative models.

Key words. Machine Learning, Computer Vision, Deblending

1. Deblending with ML

ML techniques have been previously employed to provide methods for deblending that use generative adversarial networks (GANs) (Reiman & Göhre 2019) which show promising results but GANs don't have properly defined metrics for evaluation and can be computationally heavy to train. We use a flavor of auto-encoders (AEs) called variational auto-encoders (VAEs) for deblending in the Euclid framework. AEs (Rumelhart & Williams 1986) were first introduced as neural networks that are trained to reconstruct their inputs. They are generative models that consists of two parts: an encoder and a decoder. The encoder part of the model learns to reduce the high-dimensional input to an encoded representation and the decoder part learns how to reconstruct the input from the lower dimensional representation which is called the latent space or the bottleneck layer. VAEs (Kingma & Welling 2015), are a flavor of AEs that use probablistic distribution for data generation, usually a normal distribution, such that instead of mapping the high-dimensional input to a fixed encoded representation, it is mapped to a distribution, giving us more control over how we want to model our latent distribution (Higgins 2016).

Two distinct VAE networks are needed to deblend galaxies: One which learns how to reconstruct galaxy light profiles in isolation,

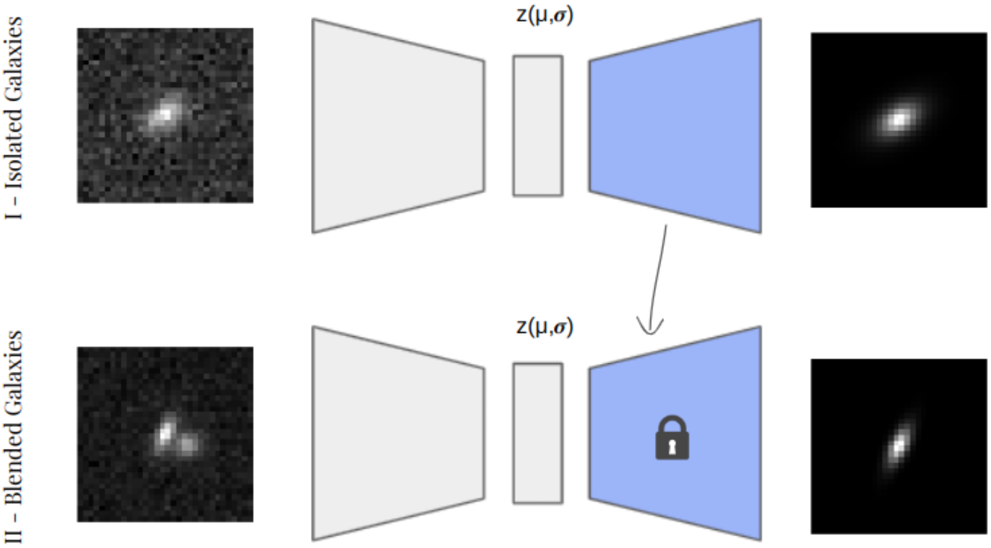


Fig. 1: The first VAE works on isolated galaxies and learns to reconstruct them. Using the decoder of the first VAE and fixing its weights the second VAE works on blended galaxies and learns to map central galaxy to z(μ, σ) s.t. reconstruction ~ original isolated.

and another one which uses the trained decoder (with fixed weights) of the first network to actually deblend overlapping galaxies by reconstructing their individual light profiles (Arcelin 2020). The second network's encoder, that works on blended galaxy images, learns to map the high-dimensional input to a point in the latent space as if it was isolated over the training period. The latent space point can then be reconstructed to a full galaxy stamp using the first network's decoder which is already trained to reconstruct galaxy stamps in isolation as shown in Fig. 1

2. Data Generation

To test the results of using this ML technique for deblending we simulate stamps of galaxy images for the Euclid survey in the VIS band. The galaxy images are built starting from a mock input catalogue created using the Empirical Galaxy Generator (EGG) (Schreiber 2017), a code that can generate mock galaxy catalogs with realistic positions, morphologies and fluxes. The catalog feeds the image simulation toolkit GalSim (Rowe 2014) to generate analytical double-Sersic profiles. We also use more realistic simulations of galaxy images for the Euclid VIS that have been created as part of the Euclid morphology challenge (Bretonnière 2022; Merlin et al. 2022).

For the second VAE we created artificially blended galaxy images by superimposing one galaxy in an annulus around another and adding the pixel values. We only use two galaxies per blended galaxy image to study the algorithm for a simple test case. For both simulations, analytic and realistic, gaussian noise is added to the noiseless simulated galaxy stamps. The gaussian noise corresponds to the limiting magnitude of the Euclid VIS band and is $m_{\text{lim}} = 27.1$. For both simulations a total of 150,000 images are used which are split into two parts for training the first VAE on isolated galaxies and the second VAE on blended galaxies.

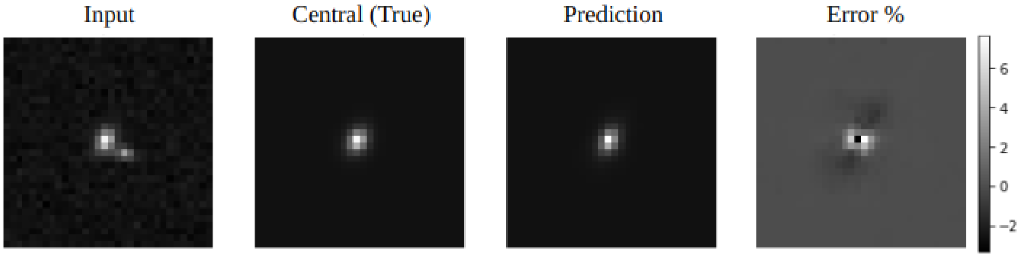


Fig. 2: Example of deblending by the VAE for the analytic dataset

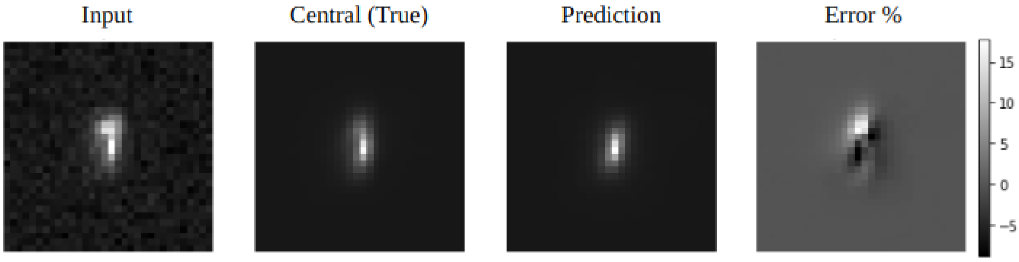


Fig. 3: Example of deblending by the VAE for the realistic dataset

3. Results

Lets look at some examples of deblending for the two different datasets described in Sec. 2. Fig. 2 and Fig. 3 show an example of deblending for the analytic dataset and the realistic dataset respectively. The input to the second VAE as shown in Fig. 1 is seen in the first panel. The second panel shows the true central galaxy in a blended pair. The third panel shows the prediction of the VAE and it can been seen that for both the datasets the VAE is able to reproduce the shape, size and orientation of the galaxy. The last panel shows the % error between the original central galaxy and the prediction by the VAE. The realistic galaxies have more complex morphology and hence it's harder for the algorithm to deblend.

The results for flux estimates from the analytic simulations are shown in Fig. 4 and from the realistic simulations are shown in Fig. 5. For the isolated galaxies as seen in Fig. 4a and Fig. 5a on the X-axis is the true magnitude and on the Y-axis is the normalised residuals of the estimated flux (by summing the pixels of the image predicted by the VAE). For the blended galaxies as seen in Fig. 4b and Fig. 5b on the X-axis is the true magnitude of the central galaxy in the blended pair and on the Y-axis is the normalised residuals of the estimated flux of the central galaxy (that is deblended by the VAE and estimated by summing the pixels of the predicted image).

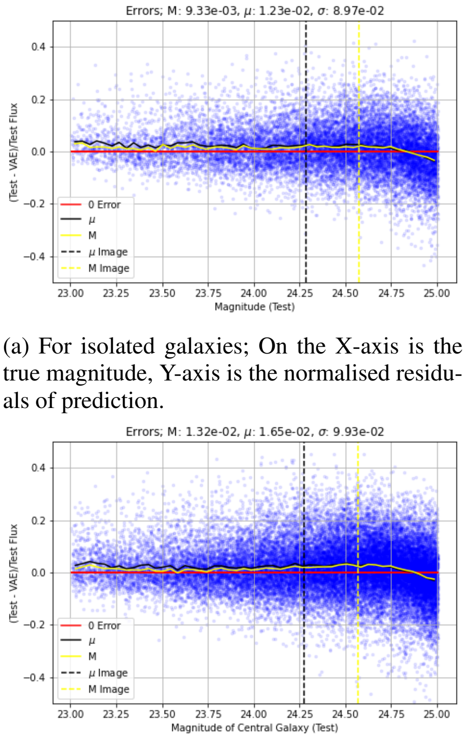
In the case of realistic simulations for bright central galaxies the deblended flux is underestimated by ~10% and this can be due to an imbalance in the dataset which reflects the true distribution of galaxies, i.e. there are more fainter galaxies than brighter ones which can be improved by artificially homogenizing the dataset. On the other hand standard methods like SExtractor tend to underestimate flux by ~12% because of their conservative apertures for photometry (Boucaud et al. 2019).

4. Conclusions

The main focus of our work is to obtain accurate flux and morphology estimates for blended objects and clean light profiles to be used as priors for template fitting codes like T-PHOT (Merlin 2015). The summary of the analysis

Table 1: Summary of standard deviation of normalised residuals

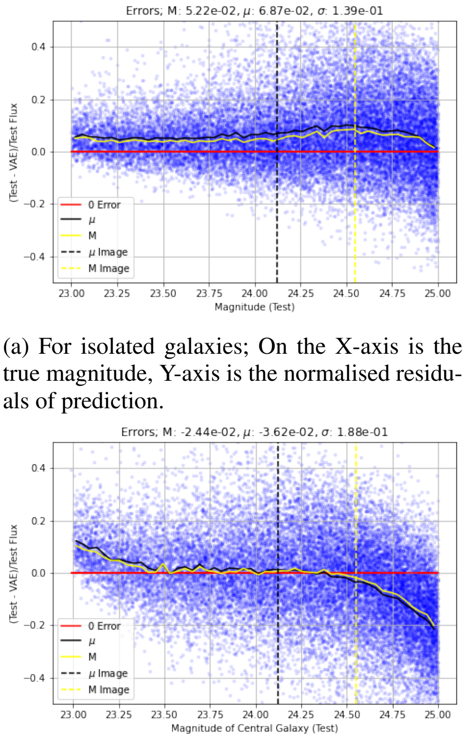
Dataset	Isolated	Blended
Analytic	8.97%	9.93%
Realistic	13.9%	18.8%



(a) For isolated galaxies; On the X-axis is the true magnitude, Y-axis is the normalised residuals of prediction.

(b) For blended galaxies; On the X-axis is the true magnitude of the central galaxy, Y-axis is the normalised residuals of prediction of the central galaxy.

Fig. 4: Results with the analytic simulations. The running mean and median are shown in black and yellow lines and the mean and median of the test dataset is marked with black and yellow dashed lines. The points above the 0 error line (red) are cases where flux is underestimated.



(a) For isolated galaxies; On the X-axis is the true magnitude, Y-axis is the normalised residuals of prediction.

(b) For blended galaxies; On the X-axis is the true magnitude of the central galaxy, Y-axis is the normalised residuals of prediction of the central galaxy.

Fig. 5: Results with the realistic simulations. The running mean and median are shown in black and yellow lines and the mean and median of the test dataset is marked with black and yellow dashed lines. The points above the 0 error line (red) are cases where flux is underestimated.

can be seen in Tab. 1 that shows that the deblending algorithm using VAEs is able to recover the flux of the central galaxy in blended pair with ∼10% accuracy for the analytic simulations and with ∼19% accuracy for the realistic simulations. In the future, we want to optimize the VAE's use for deblending by implementing hyper-parameter tuning the model architecture (for eg. the optimizing the size of the latent space) and additionally including data augmentation techniques (such as rotating, flipping the isolated galaxies and create multiple realizations of blending with the same set of galaxies) for better performances.

References

Arcelin, B., D. C. e. a. 2020, MNRAS, 500, 531

Bertin, E. & Arnouts, S. 1996, A&A supplement series, 117, 393

Boucaud, A., Huertas-Company, M., Heneka, C., et al. 2019, Monthly Notices of the Royal Astronomical Society, 491, 2481

Bretonnière, H. e. a. 2022, A&A, 657, A90

Dawson, W. & Schneider, M. 2014

Higgins, I., M. L. e. a. 2016, arXiv

Kingma, D. & Welling, M. 2015, arXiv

Merlin, E., Castellano, M., Bretonnière, H., et al. 2022, arXiv preprint arXiv:2209.12906

Merlin, E. e. a. 2015, A&A, 582, A15

Reiman, D. & Göhre, B. 2019, MNRAS, 485, 2617

Reiman, D. M. & Göhre, B. E. 2019, Monthly Notices of the Royal Astronomical Society, 485, 2617

Rowe, B., J. M. e. a. 2014, Astronomy and Computing, 10, 121

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

Schreiber, C. e. a. 2017, A&A, 602, A96