Galaxy survey data reduction with deep learning

M.Eriksen[1,2] **and L.Cabayol**[1,2]

[1]Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, 08193 Bellaterra (Barcelona), Spain
[2]Port d'Informació Científica (PIC), Campus UAB, C. Albareda s/n, 08193 Bellaterra (Cerdanyola del Vallès), Spain

E-mail: eriksen@pic.es

Abstract. PAUS is a 40 narrow-band imaging survey using the PAUCam instrument installed at the William Herschel Telescope (WHT). Since the survey started in 2015, this instrument has acquired a unique dataset, performing a relatively deep and wide survey, but with a simultaneous excelled redshift accuracy. The survey is a compromise in performance between deep spectroscopic survey and wide field imaging, showing an order of magnitude better redshift resolution than typical broad band surveys. The survey data reduction was designed based on classical data reduction techniques. For example the redshift template fitting needed a different algorithm to properly handle the PAUS data. While the data reduction and redshift estimation worked, it had room for improvements. In this talk, we detail the different efforts of replacing steps in the PAUS data reduction with deep learning algorithms. First, deep learning techniques obtain a 50% reduction in the photo-z scatter for the fainted galaxies. This is achieved through various techniques, including using transfer learning from simulations to handle a small data set. Furthermore, we have constructed multiple algorithms to improve the data reduction stage. Noise estimation from background estimation from a non-uniform background was handled in BKGNet, the galaxy photometry (light measure) was introduced with Lumus. Recent work includes the effort of directly estimating the galaxy distance from images.

1 Introduction

In 1998 two independent teams of astronomers discovered an anomalous accelerated expansion of the Universe using super novae observations [1] [2]. The distant galaxies were moving faster away from us than expected and this observation could either be explained by adding dark energy, a homogeneously distributed invented energy component or by modifying gravity. The past few decades have seen remarkable advancements in our comprehension of the Universe, largely due to galaxy surveys. Technological enhancements in astronomical cameras and computing power have facilitated the acquisition of vast amounts of high-quality data. The cosmic microwave background (CMB), galaxy clusters and galaxy distributions (see [3] and references therein) have supported extending the previous cosmological model.

However, astronomical images must be processed to generate photometric catalogues, which in turn can be used to derive photometric redshifts. With current galaxy surveys having already observed millions of galaxies, upcoming surveys such as Euclid and LSST will soon increase this number to billions, necessitating efficient and accurate methods for extracting photometry and photometric redshifts. Machine learning techniques have been employed to estimate photometric redshifts directly from images [4] [5]. Such algorithms implicitly include steps like background and flux estimation that normally is done using classical algorithms. In this talk we outline the effort of better understanding these steps, the redshift estimation and creating an end-to-end deep learning pipeline for the PAU survey.

2 The PAU survey

PAUCam is an advanced optical narrow-band camera installed on the William Herschel Telescope as part of the Physics of the Accelerating Universe Survey (PAUS) [6]. The focal plane consists of a mosaic of 18 $2k \times 4k$ Hamamatsu fully-depleted CCDs, with high quantum efficiency up to 1 μm in wavelength. To maximize the detector coverage within the FoV, filters are placed in front of the CCDs inside the camera cryostat made out of carbon fiber, using a challenging movable tray system. The camera utilizes a set of 40 narrow-band filters ranging from approximately 4500 to 8500 Å, complemented with six standard broad-band filters ($ugrizY$).

The PAU Survey plans to cover roughly 100 deg^2 over fields with existing deep photometry and galaxy shapes to obtain accurate photometric redshifts for galaxies down to $i_{\rm AB} \sim 22.5$, while also detecting galaxies down to $i_{\rm AB} \sim 23$ with less precision in redshift. With this data set, the survey will measure intrinsic alignments and galaxy clustering and perform galaxy evolution studies in a new range of densities and redshifts.

In the 2019 paper [7], the authors developed the BCNz2 photometric redshift code and characterised its performance using galaxy spectra in the COSMOS field. It achieved a $\sigma_{68}/(1+z)$ of 0.0037 for the redshift range $0 < z < 1.2$ for the best 50% of sources based on a photometric redshift quality cut. For a bright and high quality selection, driven by the identification of emission lines, a higher photo-z precision ($\sigma_{68}/(1+z) \sim 0.001$) is obtained. These result were consistent with expectations from simulations and showed PAUS could achieve the required precision [8]. The BCNz2 code was specifically designed to produce highly precise redshift estimates for the PAUS survey. To do this, BCNz2 used a template-based approach, interpolating between continuum spectral energy density and incorporating additional emission lines while also fitting for zero-points. The code determined a global zero-point per band and allowed for a free scaling between the broad and narrow bands for each galaxy.

3 Photometric redshifts

Despite its theoretical versatility, extending the BCNz2 template fitting code in different directions became practically and computationally challenging (see [9]). For instance, it was difficult to combine the non-linear minimisation with a model that enables variations in individual emission-line strengths while incorporating correlated priors between the lines. Other challenges included extending statistical fitting to account for photometric outliers and efficiently using priors for different galaxy types during minimisation. While the BCNz2 code were extended in [10], the explicit modelling became increasingly difficult and future efforts focused on machine-learning photo-z codes.

Photometric redshift estimation is limited by the number of galaxies with known redshift. This problem is more critical for PAUS, which aims to determine the photo-z precision with an order of magnitude better precision. Transfer learning is a technique commonly used to address limited training data [11]. Rather than training a model from scratch, one can begin with a model already trained on a different data set, which does not need to resemble the data set of interest [12]. For instance, ImageNet is a curated image dataset with millions of images and corresponding classes, often used as a starting point for training image classifiers [13]. Using ImageNet as a pre-training dataset can improve results and require less training. For training the DeepZ deep-learning code [9], we first train the network on galaxy simulations. The primary simulation in this study uses the Flexible Stellar Population Synthesis (FSPS) code [14] [15] which as been modified to incorporate the PAUS filter transmission.

Fig.1 aims to study how the density of spectroscopic redshifts affects the photo-z scatter. It shows the shows the photo-z scatter as a function of the number of galaxies in the bin for bins of $\Delta z = 0.001$. These bins are only used to illustrate the effect of the density and are not used when training the MDN. With the DeepZ code without pretraining, the photo-z scatter is clearly higher in bins with only few training galaxies. The dotted line shows the BCNz2 result which is much less affected by the number of galaxies per bin, specially for very sparse bins. Pretraining on simulations reduces the photo-z scatter with DeepZ, but there is still a region with fewer galaxies where the template fitting works better. Overall, the DeepZ provides a much better photo-z than the template fitting.

The network architecture in this study incorporated an autoencoder, which is a useful tool for feature extraction and noise reduction. The autoencoder consists of an encoder network that compresses the input to a set of ten features, and a decoder network that attempts to reconstruct the original input. Optimising the difference between the input and reconstructed values is known to reduce noise. We observed a 50-70% decrease in photometric errors, with the most significant impact in the blue bands. Additionally, we demonstrated that the autoencoder can result in correlated errors between bands. Incorporating the compressed input features (latent space) obtained from the autoencoder resulted in a moderate decrease in photo-z scatter. The autoencoder is expected to be more critical for wider fields since this type of network can be trained without requiring spectroscopic redshifts.

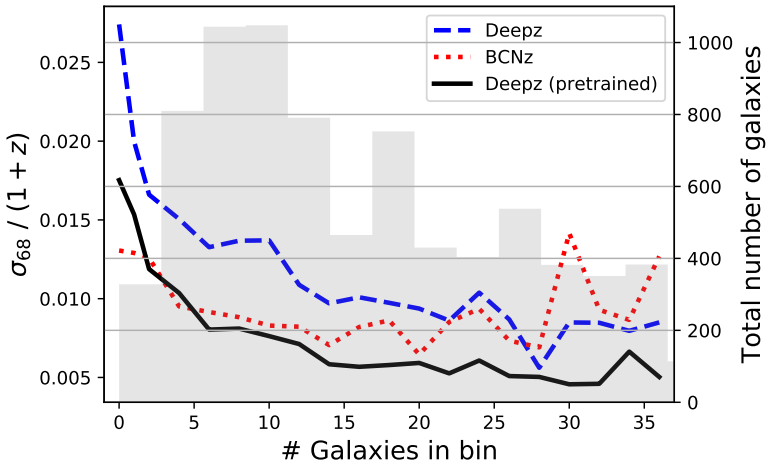

Figure 1: The effect of redshift ranges with a smaller number of galaxies. On the x-axis is the number of galaxies in bins of $\Delta z = 0.001$. The dotted line shows the BCNz2 result, while the continues and dotted lines show the DEEPZ when pretraining or not on simulations. The shaded histogram displays the total number of galaxies for each value on the x-axis.

4 Background estimation

Accurate background-subtraction methods are crucial for obtaining precise source photometry in imaging surveys. The edges of PAUCam images, particularly in the bluer bands, are susceptible to scattered light. Although modifications were made to the camera in 2016 to reduce the amount of scattered light, PAUCam images still contain a notable amount of scattered light. To address the issue of scattered light affecting astronomical images, we have developed a deep-learning algorithm that predicts the background for images captured with PAUCam.

Various methods have been utilised to estimate the sky background in astronomical images. For example SEXTRACTOR [16] use a background map by constructing a mesh and estimating the background at each mesh location. Other techniques aim to be more robust in the presence of nearby sources [17]. For PAUS, we developed the BKGNET deep-neural network to directly predict the background and its corresponding error [18]. The network is trained on image cutouts with only background light (without galaxies). The trained network is more robust to artefacts than classical methods.

5 Photometry

The developed LUMOS photometry code [19] is composed of a CNN followed by a mixture density network (MDN). Traditional photometry algorithms typically provide a single flux value and its corresponding uncertainty. In contrast, LUMOS produces a flux probability distribution by combining five Gaussian distributions. The PDF generated by LUMOS allows for the creation of a co-added flux PDF. The co-added flux PDF captures valuable information about individual flux-exposure distributions that would be missed by combining point estimates.

The LUMOS training uses simulations (Teahupoo) that are constructed combining real PAUS flux measurements and PAUCam background cutouts. However, outliers in either of these can be represented in the Teahupoo simulated images. This can include background images with spurious effects or flux measurements with artificially low or high values. We visually inspect the PAUCam images to filter out poor observations, but we cannot account for local effects in certain regions of the CCD, such as saturated pixels. Therefore, a few outliers may still be present in the Teahupoo catalogue. These simulations will also naturally include closeby located objects, enabling the network to predict the flux of blended objects.

Figure 2 compares photo-z precision of a PAUS photometry sample obtained using two different photometry and photo-z methods. The top panels illustrate the photo-z precision as a function of galaxy properties half-light radius (r_{50}, left) and Sèrsic index (n, right). Both DEEPZ and BCNz2 improves the performance, especially for higher redshifts ($z > 0.9$, bottom panel), while for lower redshifts both

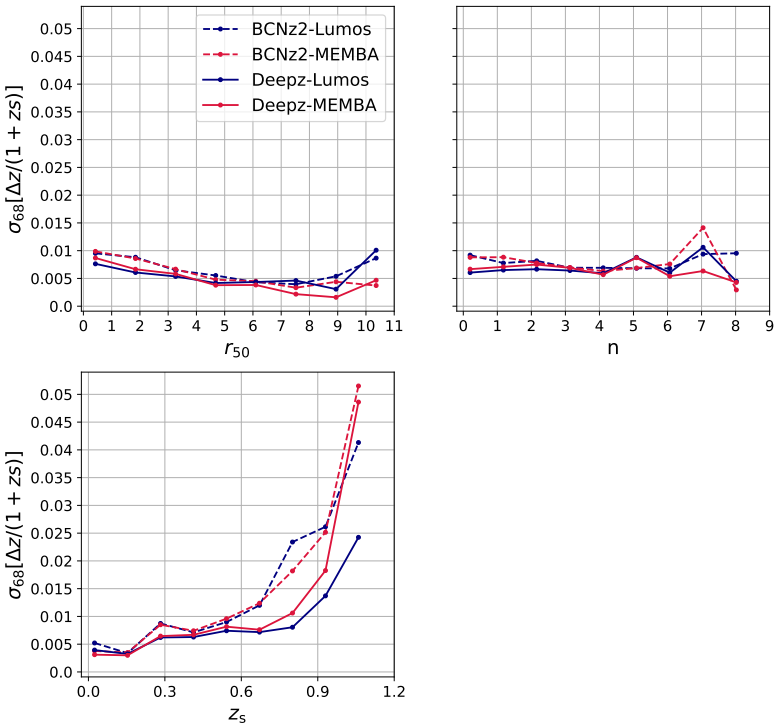

Figure 2: The comparison of the photo-z scatter as a function of different physical properties (r_{50}, n and z) for various combinations of photometric redshift codes (BCNz2, DEEPZ) and photometry (LUMOS, MEMBA).

photometries provide very similar results. This could be due to the higher signal-to-noise ratio of the LUMOS photometry at the faint end. For or smaller galaxies (low r_{50} in the first panel), the combination of Deep-z with LUMOS photometry outperforms the other algorithms, while for larger galaxies, MEMBA photometry is preferred. This discrepancy may be due to LUMOS using a fixed 60×60 cutout, resulting in some of the light from very large galaxies leaking outside the stamp. The bottom panel shows the photo-z performances in spectroscopic redshift bins, revealing clear differences in performance between the two algorithms. With BCNz2, the photo-z precision is better with LUMOS photometry (dashed-blue line) than MEMBA photometry (dashed-red line) at high redshifts ($z > 0.9$).

DEEPZ produces much more precise photo-zs with LUMOS photometry than with MEMBA, especially at the faint end, where the scatter reduces by a factor of approximately two. The fact that DEEPZ benefits more from using LUMOS photometry than BCNz2 could be related with the robustness of the deep-learning photo-z algorithm towards outlier photometries. While DEEPZ can potentially learn to ignore outliers and take advantage of the higher signal-to-noise ratio of the LUMOS photometry, such outliers could be hindering the convergence of the template-fitting BCNz2 algorithm.

6 Combined photo-z and photometry

LUMOS surpasses BKGNET by providing a background-subtracted flux measurement that requires a determination of the background-light contribution. Consequently, LUMOS handles possible correlations between the galaxy flux and the background light, which are not easily addressed analytically. Moreover, we have combined two independent networks, LUMOS and DEEPZ, to obtain the best photometric redshift. This motivates the development of an end-to-end pipeline that supersedes LUMOS by providing galaxy photometry and photometric redshift in a single framework.

ACZIO is a neural network utilizing information from all PAUS narrow bands to measure photometry in each of the bands. This approach enables the network to learn the underlying spectral energy distribution (SED) and use it to improve photometry. On average, ACZIO enhances the signal-to-noise ratio of the LUMOS photometry by a factor of two. For galaxies with $i_{AB} > 21.5$, ACZIO provides 40% more

accurate photo-z estimates than BCNz2, and the precision is comparable to that of DEEPZ. However, the photo-z estimates provided by ACZIO still need improvement, particularly for brighter galaxies where the performance degrades. Work is ongoing to understand and address this trend.

7 Conclusions

This contribution described how the Physics of the Accelerating Universe Survey (PAUS) use deep learning for data analysis. The PAU survey use a special narrow band imaging camera (PAUCam) to observe galaxies in 40 narrow optical bands (100Å wide). Similar to other imaging surveys, one need to transform the images and extract information like the galaxy flux in each band. PAUS has obtained a $\sigma_{68}/(1+z)$ of 0.0037 for the redshift range $0 < z < 1.2$ with a template fitting method for the best 50% of sources based on a photometric redshift quality cut [7]. These precise redshift estimated has multiple applications both in cosmology and galaxy evolution studies.

The DEEPZ [9] network determines PAUS galaxy distances from fluxes, reducing the scatter by 50% at the faint end. To overcome the lack of spectroscopic galaxies for training, the network is trained both on simulated and real data. The network use an auto-encoder that extracts galaxy features, which can also be trained on galaxies without spectroscopic redshifts. Furthermore, the DEEPZ algorithm differs from many other machine-learning codes, it estimates redshift probability distributions (PDFs) using a mixture density network (MDN).

Background and flux estimation are central steps in the image data reduction. To improve and understand their implication for the overall data reduction, we developed two BKGNET [18] and LUMOS [19] networks to perform these tasks. Both networks were trained on simulated data. Comparing to traditional approaches like aperture photometry (MEMBA), we found better results and the networks were more robust to artefacts. The resulting photometry improves the PAUS photometric redshifts. Finally, combining the photometry and photo-z networks we have obtained competitive narrow-band redshift estimates directly from images. A combined network going end-to-end from galaxy images to redshifts is work in progress.

References

[1] Riess A G, Filippenko A V, Challis P, Clocchiatti A, Diercks A, Garnavich P M, Gilliland R L, Hogan C J, Jha S, Kirshner R P, Leibundgut B, Phillips M M, Reiss D, Schmidt B P, Schommer R A, Smith R C, Spyromilio J, Stubbs C, Suntzeff N B and Tonry J 1998 *AJl* **116** 1009–1038 (*Preprint* astro-ph/9805201)

[2] Perlmutter S, Aldering G, Goldhaber G, Knop R A, Nugent P, Castro P G, Deustua S, Fabbro S, Goobar A, Groom D E, Hook I M, Kim A G, Kim M Y, Lee J C, Nunes N J, Pain R, Pennypacker C R, Quimby R, Lidman C, Ellis R S, Irwin M, McMahon R G, Ruiz-Lapuente P, Walton N, Schaefer B, Boyle B J, Filippenko A V, Matheson T, Fruchter A S, Panagia N, Newberg H J M, Couch W J and Project T S C 1999 *AJ* **517** 565–586 (*Preprint* astro-ph/9812133)

[3] Weinberg D H, Mortonson M J, Eisenstein D J, Hirata C, Riess A G and Rozo E 2013 *Physics Reports* **530** 87–255 URL

[4] D'Isanto A and Polsterer K L 2018 **609** A111 (*Preprint* 1706.02467)

[5] Pasquet J, Bertin E, Treyer M, Arnouts S and Fouchez D 2019 *A&A* **621** A26 (*Preprint* 1806.06607)

[6] Padilla C, Ballester O, Cardiel-Sas L, Carretero J, Casas R, Castilla J, Croce M, Delfino M, Eriksen M, Fernández E, Fosalba P, García-Bellido J, Gaztañaga E, Grañena F, Hernández C, Jiménez J, Lopez L, Martí P, Miquel R, Niessner C, Pío C, Ponce R, Sánchez E, Serrano S, Sevilla I, Tonello N and de Vicente J 2016 The PAU camera at the WHT *Ground-based and Airborne Instrumentation for Astronomy VI* (*Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* vol 9908) p 99080Z

[7] Eriksen M, Alarcon A, Gaztanaga E, Amara A, Cabayol L, Carretero J, Castander F J, Crocce M, Delfino M, De Vicente J, Fernand ez E, Fosalba P, Garcia-Bellido J, Hildebrand t H, Hoekstra H, Joachimi B, Norberg P, Miquel R, Padilla C, Refregier A, Sanchez E, Serrano S, Sevilla-Noarbe I, Tallada P, Tonello N and Tortorelli L 2019 *MNRAS* **484** 4200–4215 (*Preprint* 1809.04375)

[8] Martí P, Miquel R, Castander F J, Gaztañaga E, Eriksen M and Sánchez C 2014 **442** 92–109 (*Preprint* 1402.3220)

[9] Eriksen M, Alarcon A, Cabayol L, Carretero J, Casas R, Castander F J, De Vicente J, Fernandez E, Garcia-Bellido J, Gaztanaga E, Hildebrandt H, Hoekstra H, Joachimi B, Miquel R, Padilla C, Sanchez E, Sevilla-Noarbe I and Tallada P 2020 **497** 4565–4579 (*Preprint* 2004.07979)

[10] Alarcon A, Gaztanaga E, Eriksen M, Baugh C M, Cabayol L, Casas R, Carretero J, Castander F J, De Vicente J, Fernandez E, Garcia-Bellido J, Hildebrandt H, Hoekstra H, Joachimi B, Manzoni G, Miquel R, Norberg P, Padilla C, Renard P, Sanchez E, Serrano S, Sevilla-Noarbe I, Siudek M and Tallada-Crespí P 2021 *MNRAS* **501** 6103–6122 (*Preprint* 2007.11132)

[11] Pan S J and Yang Q 2010 *IEEE Transactions on Knowledge and Data Engineering* **22** 1345–1359

[12] Yosinski J, Clune J, Bengio Y and Lipson H 2014 URL https://arxiv.org/abs/1411.1792

[13] Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L 2009 Imagenet: A large-scale hierarchical image database *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp 248–255

[14] Conroy C, Gunn J E and White M 2009 **699** 486–506 (*Preprint* 0809.4261)

[15] Conroy C and Gunn J E 2010 **712** 833–857 (*Preprint* 0911.3151)

[16] Bertin E and Arnouts S 1996 *AASS* **117** 393–404

[17] Popowicz A and Smolka B 2015 *Monthly Notices of the Royal Astronomical Society* **452** 809–823 ISSN 0035-8711 (*Preprint* https://academic.oup.com/mnras/article-pdf/452/1/809/4928371/stv1320.pdf) URL https://doi.org/10.1093/mnras/stv1320

[18] Cabayol-Garcia L, Eriksen M, Alarcón A, Amara A, Carretero J, Casas R, Castander F J, Fernández E, García-Bellido J, Gaztanaga E, Hoekstra H, Miquel R, Neissner C, Padilla C, Sánchez E, Serrano S, Sevilla-Noarbe I, Siudek M, Tallada P and Tortorelli L 2020 **491** 5392–5405 (*Preprint* 1910.02075)

[19] Cabayol L, Eriksen M, Amara A, Carretero J, Casas R, Castander F J, De Vicente J, Fernández E, García-Bellido J, Gaztanaga E, Hildebrandt H, Miquel R, Padilla C, Sánchez E, Serrano S, Sevilla-Noarbe I and Tallada-Crespí P 2021 *MNRAS* **506** 4048–4069 (*Preprint* 2104.02778)