amU Aix Marseille Université	Doctoral School 352 Physics and Science of Matter	Centrale Méditerranée

Laboratory:

Laboratoire d’Astrophysique de Marseille

Thesis supervisor:

Arnouts Stéphane (Directeur de Recherche, LAM, UMR 7326)

Thesis co-supervisors:

Pasquet Jérôme (Maître de Conférence, UMR TETIS Montpellier)
Treyer Marie (Chargée de Recherche, LAM, UMR 7326)

Title: Galaxy Evolution and Cosmic Web with Deep Learning: Addressing CNN redshift biases for large photometric surveys.

Summary:

This thesis will develop deep learning (DL) methods to estimate the photometric redshifts (“photo-z”) of galaxies (i.e. their distances) and their physical properties by exploiting multi-band imaging surveys. As our team demonstrated, the convolutional neural network (CNN) technique significantly improves the accuracy of photo-z compared to all previous methods (Pasquet et al., 2019, Ait-Ouahmed et al., 2024). However, one major challenge is to deal with unbalanced and incomplete representativity of the training set and variable multi-band imaging conditions across large photometric surveys. To this end, new DL methodologies will be developed, based on self supervised learning and few shot learning to better consider poorly represented objects in the training set and secure the homogeneity of the CNN redshift’s accuracy across a large area of the sky. The expected gain in photo-z accuracy (dz<0.02) and its angular stability will enable us, for the first time, to reconstruct the cosmic web (CW) in thin redshift slices to high redshift. The science goal will be twofold: 1- to study the link between galaxy properties and their large scale environment; 2- to measure the connectivity of the filamentary structures (mean number of filaments) around CW peaks at different epochs. This measure, which traces the growth rate of the cosmic structures, depends on dark energy and will therefore be a constraint. The PhD thesis is part of an existing collaboration between Marseille, Montpellier and Paris, providing a dynamic and rich environment for a PhD student.

Description of the thesis subject :

The cosmic web (CW) is a complex network of voids, walls, filaments and knots in which galaxies form and evolve. The exchanges of gas and energy (infalls/outflows) between the galaxies and their environment play a crucial role in shaping galaxy properties. Theoretical predictions have emphasized the close link between the large scale cosmic flows and the spin orientation of galaxies, recently observed in the local universe (Tempel et al., 2013). Dependencies between the stellar mass and star formation activity of galaxies and their distances to CW filaments have also been detected (Kraljic et al., 2018 at z<0.3; Malavasi et al., 2017 at z~0.8). These results offer a new way to understand galaxy evolution in a cosmological context with the forthcoming large spectroscopic surveys. Beyond the spectroscopic 3D mapping, the CW can also be reconstructed in thin 2D redshift slices with robust photo-z's, as initially shown by Laigle et al. (2018) with the COSMOS 30 bands survey. This technique offers a great alternative to investigate this topic with gigantic imaging surveys, such as LSST (covering more than a half hemisphere).

To reach this goal, the challenge is to get very accurate photo-z measurements (dz<0.02) with a limited set of filters. One limiting factor of current photo-z techniques has been the extraction of photometric quantities (fluxes and colors) used as input, which capture only a fraction of the information present in the images. Taking advantage of the latest DL techniques, of the GPU acceleration and of the large size of spectroscopic training samples in the local universe (SDSS), photo-z derived with DL algorithms (e.g. dealing directly at the pixel level; Hoyle et al., 2016; d'Isanto, 2017; Pasquet et al., 2019) outperform traditional approaches. In particular, we (Pasquet et al.; 2019) found that DL photo-zs show no bias with redshift, galaxy inclination, …, compared to other machine learning methods and provide robust probability distribution functions, crucial for cosmological analyses. In this thesis, we will exploit the HSC+CLAUDS survey (covering 25 deg2 with multi-band UgrizY imaging down to r~27). It is the best existing survey for this purpose, with a unique combination of depth and area, rich ancillary data and a large spectroscopic training set (>100,000 redshifts).

In our recent works (Ait Ouahmed et al. 2024; Treyer et al. 2024) we have shown the presence of strong biases in large surveys for faint magnitude sources. This issue is a direct consequence of a possible mismatch caused by several factors, such as the spatial distribution, the signal-to-noise ratio, etc., between the training data and the test data, leading to a significant loss of performance. First attempts to overcome these biases (Ait Ouahmed, PhD thesis; Treyer et al., in prep) are to use adversarial losses to force the latent space to be invariant to certain criteria (such as the targeted sky region). This approach helps to limit the bias between the training and test datasets. However, it significantly reduces performance compared to a network that would have also learned from data coming from the test set.

The objective of this thesis is to reduce the performance gap between a training setup where the bias is corrected and one where the training and test datasets are mixed and then re-split into train and test folds. To achieve this, we propose to explore the foundational models, particularly extending self-supervised methods based on contrastive learning approaches, such as SimCLR (Chen et al. 2021) or BYOL, which are capable of extracting relevant information without the need for labels. Recent studies (Hayat et al. 2021) demonstrate that these approaches are effective and can create latent spaces in the case of astronomical surveys, rivaling supervised approaches.

During the thesis, we propose to improve these types of algorithms in order to extract the maximum amount of relevant information from the test dataset. To improve performance and reduce biases, we propose working on the constraints of the latent space. One way to achieve this is by focusing on the uniformity (Fang et al. 2021) of representations across the different layers of the network. To this end, the coherence of distances related to certain image parameters that may introduce biases, such as signal-to-noise ratio, color, and sky region, can be leveraged.

In the first part of the thesis, an unsupervised model constrained by physical parameters will be proposed. In the second part, the thesis will address the issue of limited image

availability in the training set, particularly in the context of fine-tuning models using data from a distribution similar to that of the test set. We will explore algorithms from meta-learning for few-shot learning (Chen et al. 2021), as well as methods such as ProtoNets (Snell et al. 2017). The goal will be to extend these techniques, originally developed for classification tasks, to regression problems, particularly in situations where not all possible values are represented in the training set.
During this thesis, in order to verify and validate all the methods, the image dataset from COSMOS 30 and HSC Deep Imaging Surveys will be regenerated by simulating variable signal-to-noise ratio and image qualities in different sky regions. This will allow mimicking expected behaviors that will impact LSST observations. In the last year (or two) of the thesis, a large number of spectroscopic redshifts from PFS spectroscopic survey will become available to test the performance of the proposed methods during this thesis with the HSC image dataset. First LSST observations will also become available, the proposed model developed in the PhD thesis will be used to make initial predictions of CNN photo-z and the galaxy physical parameters for cosmic web investigations.

During this PhD thesis, the student will interact with researchers at LAM (M. Treyer, D. Vibert, S de La Torre, O. Ilbert), Paris (S. Codis, E. Bertin) and Strasbourg (K. Kraljic) for the astronomical part. For the Deep learning part, in addition to the supervision by Jerome Pasquet at TETIS (Montpellier), s/he will have the opportunity to interact with the LIRMM namely the ICAR research team and Marc Chaumont (Montpellier) to benefit from their advice and expertise in mismatch problem.

Références:

Tempel, Stoica, Saar, 2013 MNRAS 428, 1827
Kraljic, Arnouts, Pichon et al., 2018, MNRAS 474, 547
Malavasi, Arnouts et al., 2017, MNRAS 465, 3817
Laigle, Pichon, Arnouts et al., 2018, MNRAS 474, 5437
Hoyle, 2016, A&C 16, 34 d'Isanto, 2017, IAUS 325, 209
Pasquet, Bringay, Chaumont, EUSIPCO 2014
Pasquet, Bertin, Treyer, Arnouts, Fouchez, 2019, A&A 621, 26
Pasquet, Pasquet, Chaumont, Fouchez, 2019, A&A 627, 15
Snell, Swersky, Zemel, 2017, NeuRIPS
Chen, Liu, Xu, Darrell, Wang, 2021, ICCV
Hayat, Stein, Harrington, Lukić, Mustafa, 2021, ApJ 911, 33
Chen, Kornblith, Norouzi, Hinton, 2021, ICML
Fang, Li, Sun, Wang , 2024, ICLR
Ait Ouahmed, Arnouts, Pasquet, Treyer, Bertin, 2024, A&A 683, 26
Treyer, Ait-Ouahmed, Pasquet, Arnouts, Bertin, Fouchez, 2024, MNRAS 527, 651