Research Statement

Xingchen Zhou

PhD and Research Assistant at

National Astronomical Observatories, Chinese Academy of Science

Current Research

With the development of observational instruments, the astronomical community has entered the era of big data. The data volume from ongoing and planned observations has reached unprecedented scales, posing significant challenges to traditional data analysis techniques. Machine learning algorithms, particularly deep learning, have emerged as a pivotal approach for data exploration in recent years. These algorithms possess the ability to effectively manage and interpret intricate, high-dimensional data, offering potential for profound insights from vast datasets. My research leverages deep learning to advance the analysis and interpretation of various surveys, aiming to improve efficiency, accuracy, and depth of understanding in critical domains of observational astronomy.

In relatively low redshift, galaxy surveys serve as the primary probes for studying our Universe. The measurements of various quantities for galaxies are fundamental to astronomical and cosmological research. Among these quantities, redshift holds the most basic significance. My research endeavors to develop deep learning models for estimating photometric and spectroscopic redshifts for the China Space Station Telescope (CSST). This telescope has seven photometric bands and three slitless spectroscopic bands, offering observations from near-ultraviolet to near-infrared. Notably, the CSST enables simultaneous photometric and slitless spectroscopic surveys, covering an expansive area of approximately 17,500 square degrees.

For a photometric survey, I develop a pipeline capable of estimating redshift from photometry, galaxy images, and the combination of the two data types. I demonstrate that galaxy images can provide lower outliers for redshift estimations compared to using flux data alone. Furthermore, the combined dataset yields further improvements, as morphological information can naturally be incorporated by convolutional neural networks. Given the significance of uncertainties in various cosmological studies, I applied a Bayesian neural network (BNN) for redshift estimation, generating redshift values along with their corresponding uncertainties. The code for redshift estimation pipeline is publicly available at Github. Additionally, to validate the applicability of BNN in real observations, I utilized it to create a new catalog for 1.8 billion sources in the DESI Legacy Surveys. Specifically, the network is trained using galaxy images from optical and near-infrared bands, along with corresponding spectroscopic redshift provided in the DESI Early Data Release (EDR), leveraging the high accuracy of redshift measured by DESI. We found that the performance varies among the galaxy types, and estimation within individual type provides higher accuracy.

For CSST, spectroscopic surveys are conducted by slitless spectrograph to match the survey speed of photometric surveys. However, due to the low resolution and signal-to-noise ratio of slitless spectra, the redshift estimations are challenging. Therefore, I develop a neural network model to investigate the accuracy of redshift estimations achievable for slitless spectroscopic surveys. The mock spectra were simulated by a slitless spectrum simulation software that accepts spectral energy distribution (SED) and morphological parameters. To enhance the realism of the simulations, BOSS and DESI observational data were employed. From these mock spectra, my network exhibits high accuracy, meeting the requirements of $\sigma_{NMAD} < 0.005$ for BAO and other cosmological studies utilizing CSST spectroscopic surveys.

In higher redshifts nearing the reionization epoch, line intensity mapping (LIM) emerges as the primary probe, as galaxy surveys are infeasible in most cases. LIM measures the integrated emission from specific atom or molecular line transitions across various redshifts. Several lines, such as the 21cm, CII and CO, serve as potential tracers. However, the extraction of a signal is complicated due to foreground emission and interloper lines, thereby limiting the ability of cosmological constraints. I work on foreground removal for CO intensity mapping employing deep learning models. My specific objectives include calibrating the deviated power spectra induced by beam effects during PCA foreground removal and evaluating the performance of different CO luminosity models. The outcomes demonstrate that the deviations in power spectra can be significantly reduced utilizing a straightforward generative model, U-Net.

In deep learning, the volume of training data significantly influences the model's robustness. However, in most cases, real data are insufficient or inexistent to train a robust model. Consequently, we commonly rely on simulations. I am currently working on simulation of galaxies from IllustrisTNG using the SKIRT project for various instruments. SKIRT has the capability to calculate radiative transfer for particles with diverse SEDs. I develop a Python wrapper for comprehensive workflow that encompasses preprocessing and postprocessing stages. Preprocessing involves extracting particles from TNG, assigning SED, and preparing an execution file to run SKIRT. Postprocessing entails calculating galaxy images for various observations, accounting for transmission and instrumental effects, such as PSFs and backgrounds. This wrapper offers remarkable flexibility, enabling users to freely modify configurations, including simulation mode, dust model and other parameters, to tailor galaxy simulation to their specific requirements. The code is publicly available at Github.

With the knowledge in photometric and spectroscopic surveys, I am currently engaged in preliminary target selections for MUltiplexed Spectroscopic Telescope (MUST), a stage-V spectroscopic instrument led by Tsinghua University. I am in charge with selections on Emission Line Galaxies (ELG) and high-redshift Lyman Break Galaxies (LBG) under observational conditions of MUST. Additionally, I calculate the galaxy biases for these targets, which are used to forecast the capability of cosmological constraints. And currently, I am working on simulation of spectra using my knowledge gained from CSST slitless spectra project.

Research Plans

Deep learning holds immense potential to revolutionize the analysis and interpretation of data from next-generation photometric and spectroscopic surveys. These surveys, such as Euclid, Rubin, MUST, WST, will produce vast, high-dimensional datasets that require advanced computational tools for efficient processing. By automating tasks, uncovering hidden patterns, and enhancing the efficiency of data analysis, deep learning will enable us to maximize the scientific return from next-generation surveys. It will accelerate discoveries, refine our understanding of the universe, and tackle challenges and complexity of modern cosmological datasets. I am interested in exploring the applications of deep learning algorithms to problems in cosmological studies. My research plan outlines the following objectives:

Blending and photo-z estimations: Blending effects tends to increase with the depth of observation, necessitating a suitable approach to address this issue. Utilizing galaxy simulation tools from IllustrisTNG by SKIRT mentioned above, blending effects can be comprehensively investigated. Generative models, such as U-Net, GAN, and Diffusion model, are suitable for deblending research. My objective is to construct a generative model and compare it with existing conventional methods. Additionally, it is crucial to analyze the impact of deblending on estimation of photo-z and other quantities.

Spec-z from 2d slitless spectral images: The raw data of slitless spectroscopic observations are two-dimensional spectral images, from which one-dimensional spectra can be extracted. However, spectroscopic redshift (spec-z) measurements are significantly affected by certain issues. One such issue is the calibration of wavelength. Estimating spec-zs from two-dimensional spectral images can be advantageous, as it bypasses the need for wavelength calibrations. Additionally, the blending effects can be straightforwardly simulated using two-dimensional spectral images, allowing for the analysis of their impact on estimations for spec-zs.

Galaxy morphological classifications: The morphological characteristics of galaxies can serve as indicators of their evolutionary history, and this research area has been extensively studied using deep learning techniques or conventional methods. However, morphological classifications at high redshift are seldom investigated, primarily because sources with current labels are insufficient to construct a reliable model. Representation learning employing unsupervised and semi-supervised approaches may offer a potential solution. Furthermore, the current classification framework is subjective and necessitates human intervention for determination. It is intriguing to explore the possibility of classifications that are entirely automated and driven solely by data.

Large-scale structure: Large-scale structure is web-like distribution of matter on cosmological scales, shaped by the gravitational growth of initial density fluctuations in the early universe. This structure is composed of interconnected components, including filaments, walls, and void, which together form the cosmic web. Studying this structure provides crucial insights into the composition, dynamics and evolution of the universe, including the nature of dark mater, dark energy, and the processes that govern galaxy formation and clustering. Observations from wide and deep photometric and spectroscopic surveys, combined with theoretical models and simulations, allow for mapping and analyzing these structures. Advances in deep learning, particularly explainable and interpretable AI, will play a crucial role in constraints on cosmological studies, reconstruction of initial conditions, rapid emulation for large-scale structure, and other challenges that should be addressed in studies on large-scale structure.

Deep learning algorithms: Numerous interesting models have been proposed to address specific challenges in astronomy. Simulation-based inference (SBI) is such a model. Utilizing an emulator and an estimator, this technique offers an additional approach to constrain astronomical and cosmological parameters, yielding comparable or superior performance to conventional methods, such as Markov Chain Monte Carlo (MCMC). Another approach is Graph Neural Network (GNN), which can learn the relationships from structured data, including catalogues, a common data structure in astronomy. While deep learning has achieved remarkable successes, several challenges remain to be addressed. One challenge is domain adaptation from simulated data to real observational data, which serves as the key to applying deep learning models to astronomical and cosmological studies. Another challenge is model interpretation, an active and ongoing topic in deep learning research. By addressing this problem, a deeper understanding and insights of deep learning models can be obtained, thus promoting the development of data analysis methods.