Prepared for submission to University of Cambridge

Reproducing AstroCLIP: Executive Summary

Adnan Siddiquei

University of Cambridge

E-mail: as3438@cam.ac.uk

[1] Word Count: 999 (not incl. figure captions, bibliography); 1213 (incl. figure captions).

Contents

1 Background and Motivation

Scientific datasets, especially in astronomy, have been expanding rapidly for decades. Spectroscopic surveys like the Sloan Digital Sky Survey (SDSS)[1] (York et al., 2000) and the Dark Energy Spectroscopic Instrument (DESI)[2] have been collating millions of galaxy spectra. Similarly, photometric surveys like the DESI Legacy Survey (Dey et al., 2019) have imaged vast areas of the sky, extracting millions of sources. A spectrum is a plot of light intensity versus wavelength, and a photometric image combines light from various filters, as illustrated in Figures (1) and (2).

Both photometric and spectroscopic surveys are essential tools in modern astronomy and these datasets are used for a variety of scientific purposes, from understanding the large scale structure of the universe; estimating galaxy properties such as redshift, stellar mass, and star formation rate; to identifying rare objects such as quasars and supernovae; and many more. However, the growing data set size and diversity makes much of this difficult and traditional methods are often limited by the quality of the data and its associated labels. One such example is morphological classification, where we desire to classify galaxies into different types based on their shape and structure, such as spiral or elliptical. A decade ago we had crowdsourced campaigns such as Galaxy Zoo 2 (Willett et al., 2013) which classified approximately 300,000 galaxies, we now have tools such as Tractor[3] which can probabilistically identify sources from photometric surveys and infer properties such as morphological classification. More recently, given the unavailability of high quality labels, unsupervised and self-supervised learning methods have been gaining popularity to tackle these sorts of tasks. For example:

- Liang et al. (2023) train a 1D convolutional spectrum autoencoder on spectral data for the purposes of outlier detection;
- Stein et al. (2021) train a 2D convolutional image embedder using a self-supervised technique on galaxy images for the purposes of similarity search;
- Hayat et al. (2021) also use a self-supervised technique to train a 2D convolutional model to estimate distances to galaxies from their photometric images, and further

[1] https://www.sdss.org/

[2] https://www.desi.lbl.gov/

[3] https://github.com/dstndstn/tractor

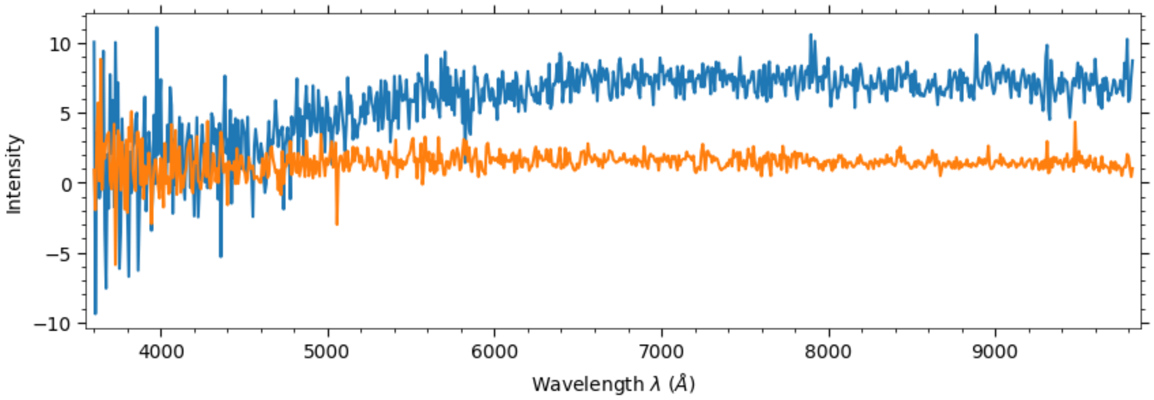


Figure 1: An example of two galaxy spectra.

demonstrate that the learned embeddings can be fine-tuned very effectively for redshift estimation;

Hayat et al. (2021) also show that significantly better performance can be acquired through fine-tuning the self-supervised pre-trained model compared to simply training a supervised model from scratch. This conclusion is not uncommon in the field of machine learning outside of astronomy and demonstrates the power of transfer learning and the importance of foundation models for astronomical datasets.

However, as of yet most of these self-supervised learning methods have only been applied to a single modality despite promising results in cross-modal contrastive learning outside of astronomy, such as contrastive language-image pre-training CLIP (Radford et al., 2021). Parker et al. (2024) pioneer on this front by proposing a multi-modal contrastive learning approach to embed galaxy spectra and galaxy images into a shared low-dimensional latent space, and in this paper we reproduce their results. Given the multi-modal nature of astronomical datasets, a useful astronomical foundation model should be able to embed the varying views of the same object effectively into a shared latent space, allowing for in-modal and cross-modal downstream application through zero-shot or few-shot learning.

2 A Review of AstroCLIP

Parker et al. (2024) present AstroCLIP, a cross-modal foundation model for galaxies. The approach involves two components. Initially, they pre-trained a transformer-based spectrum embedder on spectroscopic data from the DESI Early Data Release (Collaboration et al., 2023) using a self-supervised technique. They also pre-trained a vision transformer for images from the DESI Legacy Survey (Dey et al., 2019) using the DINO v2 framework (Oquab et al., 2024). Secondly, they fine-tuned both embedders under the InfoNCE loss (van den Oord et al., 2019) to align embeddings in a 512-dimensional latent space. The InfoNCE loss promotes proximity in embeddings of the same object and distance for different objects, where proximity is measured through the cosine similarity (or equivalently, the normalised dot product) of the embeddings.

They demonstrate that the embeddings are well-aligned by using the embeddings to perform a variety of downstream zero-shot and few-shot tasks, including in-modal and cross-modal similarity search, redshift prediction, and galaxy property prediction. The model was

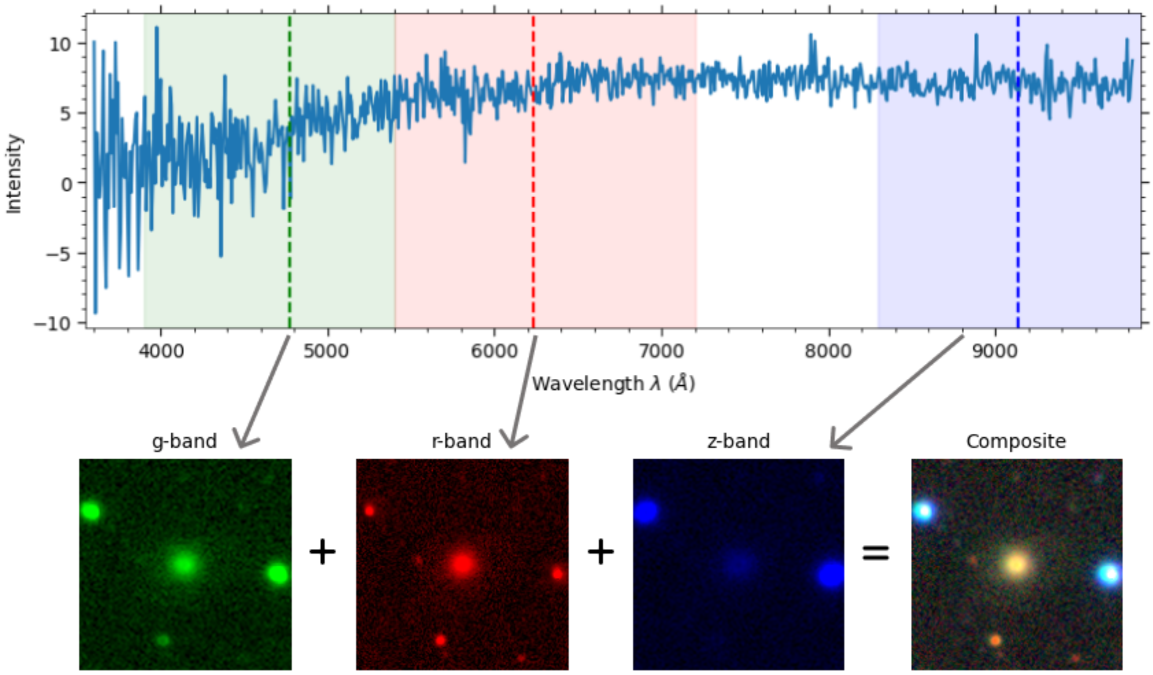


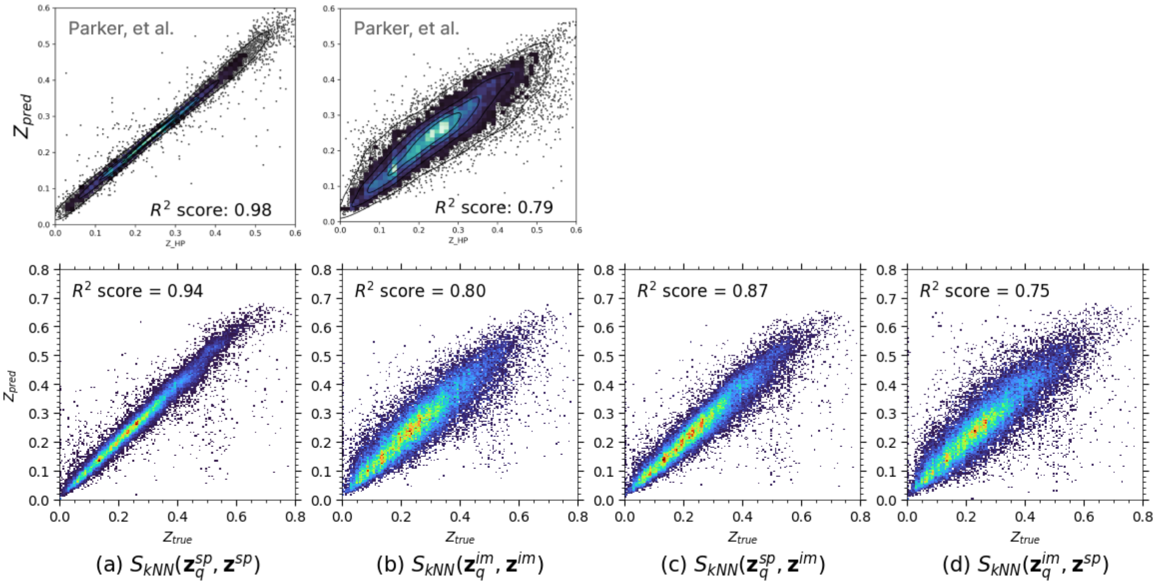
Figure 2: Bandpasses for various filters, illustrating how they combine to form a photometric image. The dotted line indicates peak transmission wavelengths; the shaded areas show the transmission range. This Figure is a conceptual illustration, and not exact.

never explicitly trained to embed these properties into the latent space, but these properties inherently exist in the data and as such, well-aligned embeddings should be able to make effective zero-shot and few-shot predictions on these properties.

3 Reproduction

We reproduce the AstroCLIP model, however we utilise the pre-trained convolutional spectrum embedder by Liang et al. (2023) and the pre-trained convolutional image embedder by Stein et al. (2021) rather than pre-training our own. We use the dataset as provided by Parker et al. (2024), and a variety of general and astronomy specific data augmentations to increase the variety of the dataset. In addition to the 512-dimensional model, we also train AstroCLIP variations to embed the spectra and images into a variety of lower-dimensional latent spaces: $[8, 16, 32, 64, 128, 256]$. We train each of these 7 models for 75 epochs, and choose the model with the lowest validation loss for analysis.

We assess the performance of our models using a subset of the downstream tasks presented in the original paper, specifically zero-shot k-NN redshift estimation and retrieval by cosine similarity, these results are displayed in Figures (3), (4), (5), and (6). Generally, we find equivalent performance in our reproduction. As shown in Figures (3), we outperform their 512-dimensional model in the photometric redshift prediction task (across a larger redshift range) with our 128-dimensional model, achieving an R^2 value of 0.80 compared to their 0.79; but fall short in the spectroscopic redshift prediction task with an R^2 value of 0.94 compared to their 0.98. We also show that model performance stays generally stable with



(a) $S_{kNN}(\mathbf{z}_q^{sp}, \mathbf{z}^{sp})$ (b) $S_{kNN}(\mathbf{z}_q^{im}, \mathbf{z}^{im})$ (c) $S_{kNN}(\mathbf{z}_q^{sp}, \mathbf{z}^{im})$ (d) $S_{kNN}(\mathbf{z}_q^{im}, \mathbf{z}^{sp})$

Figure 3: k-NN regression for zero-shot redshift prediction, showing predicted vs. true redshift. Top row is results from Parker et al. (2024) for their 512-dim embedding model, bottom row is results from our 128-dim embedding model. $S_{kNN}(\mathbf{z}_q^{sp}, \mathbf{z}^{im})$ indicates the cross-modal prediction type where a spectrums redshift was predicted using its 16 closest embeddings derived from galaxy images.

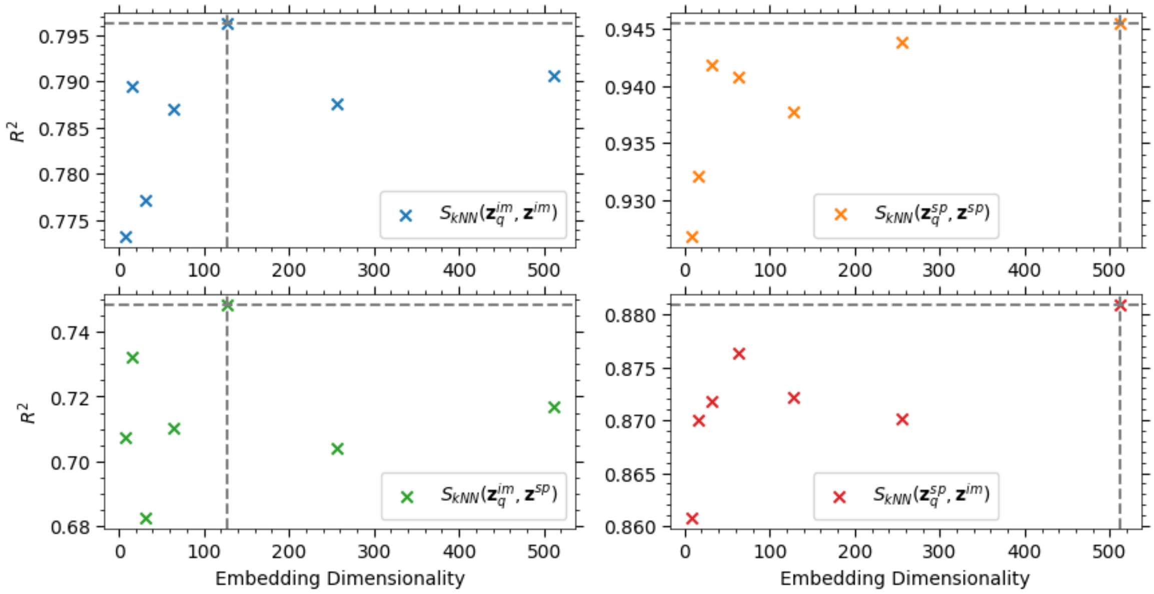
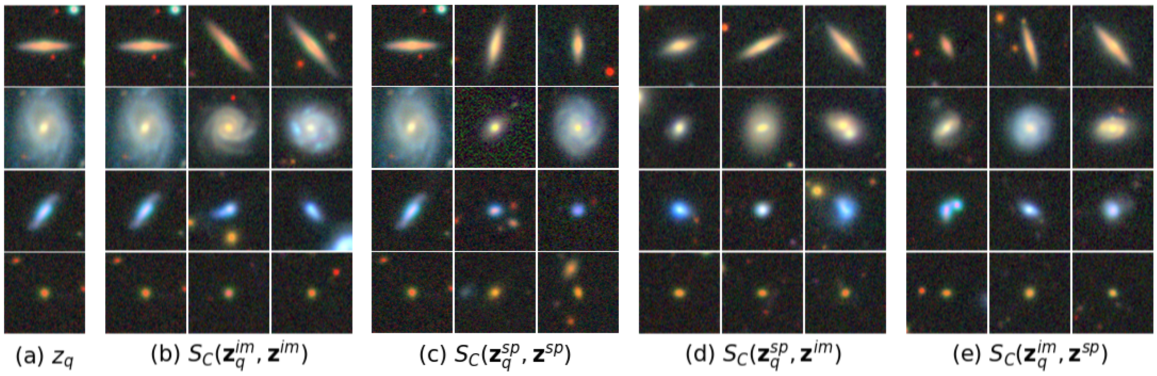


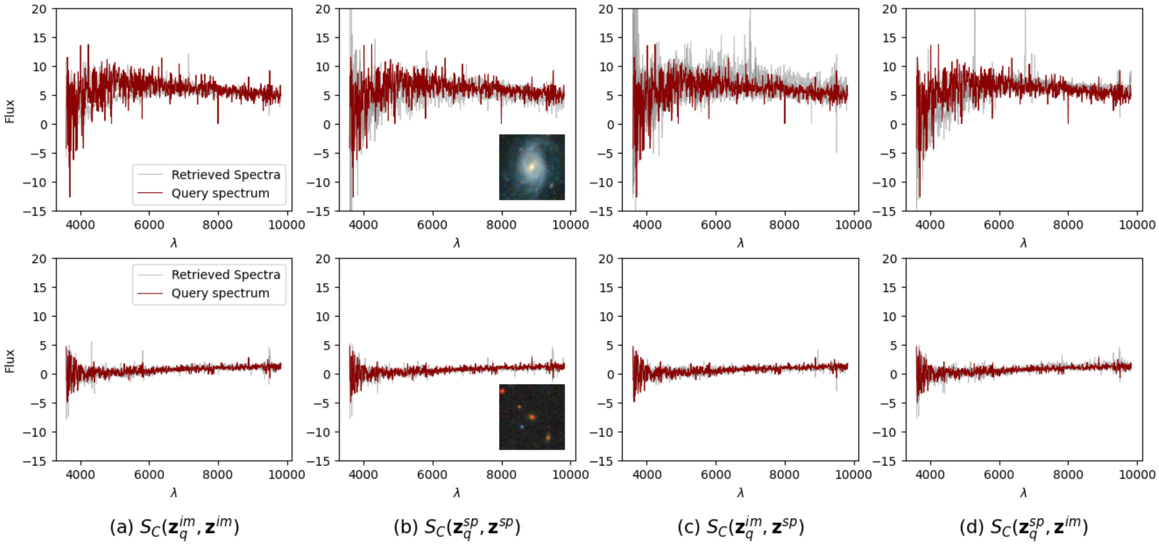
Figure 4: Consolidated results for all embedding dimensions and prediction types, displaying R^2 values vs. embedding dimensionality. These are the R^2 values derived if Figure (3) were plotted for all embedding dimensionalities.

embedding dimensionality as low as 8, as shown in Figure (4). Cosine retrieval results are also consistent with the original paper, as shown in Figures (5) and (6).



(a) z_q (b) $S_C(\mathbf{z}_q^{im}, \mathbf{z}^{im})$ (c) $S_C(\mathbf{z}_q^{sp}, \mathbf{z}^{sp})$ (d) $S_C(\mathbf{z}_q^{sp}, \mathbf{z}^{im})$ (e) $S_C(\mathbf{z}_q^{im}, \mathbf{z}^{sp})$

Figure 5: Cosine similarity search results using the 128-dimensional model. (a) 4 Query galaxies; (b, c, d, e) Top 3 similar galaxies across the 4 prediction types. By construction, the most similar in-modal galaxy to any given galaxy is itself, hence, the first column of images in (b) and (c) are identical to the query image in (a).



(a) $S_C(\mathbf{z}_q^{im}, \mathbf{z}^{im})$ (b) $S_C(\mathbf{z}_q^{sp}, \mathbf{z}^{sp})$ (c) $S_C(\mathbf{z}_q^{im}, \mathbf{z}^{sp})$ (d) $S_C(\mathbf{z}_q^{sp}, \mathbf{z}^{im})$

Figure 6: Cosine similarity search for spectra using the 128-dimensional model. Each row shows a query galaxy (galaxy imaged in (b)), each subfigure shows the 3 most similar spectra to the query galaxy for the given prediction type.

4 Significance of Results

Our results, alongside those of Parker et al. (2024), demonstrate that it is possible to achieve high quality foundation models for astronomical data using cross-modal contrastive pre-training, and that the learned embeddings can be used for a variety of downstream tasks with strong performance. This has a variety of impacts on the field of astronomy, such as enabling cross-modal similarity searches for rare or interesting objects, and enabling the use of pre-trained foundation models for transfer learning on smaller datasets for specific tasks, thereby reducing the requirement for large amounts of high quality labels.

Bibliography

Collaboration, D., Adame, A. G., Aguilar, J., et al. (2023). The early data release of the dark energy spectroscopic instrument. *arXiv:2306.06308.*

Dey, A., Schlegel, D. J., Lang, D., et al. (2019). Overview of the desi legacy imaging surveys. *arXiv:1804.08657.*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929.*

Hayat, K. A., Harrington, P., Stein, G., Lukić, Z., and Mustafa, M. (2021). Estimating galactic distances from images using self-supervised representation learning. *arXiv:2101.04293.*

Liang, Y., Melchior, P., Hahn, C., Shen, J., Goulding, A., and Ward, C. (2023). Outlier detection in the desi bright galaxy survey. *arXiv:2307.07664.*

Oquab, M., Darcet, T., Moutakanni, T., et al. (2024). Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193.*

Parker, L., Lanusse, F., Golkar, S., Cranmer, M., et al. (2024). Astroclip: Cross-modal pre-training for astronomical foundation models. *arXiv:2310.03024.*

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., et al. (2021). Learning transferable visual models from natural language supervision. *arXiv:2103.00020.*

Stein, G., Harrington, P., Blaum, J., Medan, T., Lukic, Z., et al. (2021). Self-supervised similarity search for large scientific datasets. *arxiv:2110.13151.*

van den Oord, A., Li, Y., and Vinyals, O. (2019). Representation learning with contrastive predictive coding. *arXiv:1807.03748.*

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., et al. (2013). Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *https://doi.org/10.1093/mnras/stt1458.*

York, D., Collaboration, T. S., et al. (2000). The sloan digital sky survey: Technical summary. *arXiv:astro-ph/0006396.*