Prepared for submission to University of Cambridge

Reproducing AstroCLIP: A Cross-Modal Foundation Model for Galaxies

Adnan Siddiquei

University of Cambridge

E-mail: as3438@cam.ac.uk

Abstract. We present a reproduction of the AstroCLIP model by Parker et al. (2024), a cross-modal foundation model for galaxies. AstroCLIP is a single cross-modal model that can embed galaxy spectra and images into a shared low-dimensional embedding space. AstroCLIP discovers a meaningful representation of the data by training the model using contrastive training under the InfoNCE loss on the Dark Energy Spectroscopic Instrument (DESI) spectra and images from its corresponding DESI Legacy Imaging Survey. These cross-modal embeddings can be used for a variety of downstream tasks and we reproduce a subset of the downstream tasks from the original paper to assess the performance of our AstroCLIP model, including (1) in-modal and cross-modal similarity search and (2) zero-shot in-modal and cross-modal redshift prediction. Our approach to this reproduction differs from the original authors in that we utilise pre-trained convolutional image (Stein et al., 2021) and spectrum (Liang et al., 2023) embedders as opposed to the transformer architectures used by Parker et al. (2024). We also explore the performance of the model in the downstream tasks as the embedding dimensionality is varied through $[8, 16, 32, 64, 128, 256, 512]$ dimensions. We find generally equivalent performance to the original paper, and show that even low-dimensional embeddings are able to perform well in the downstream tasks. We outperform the original authors in the in-modal photometric redshift prediction task with a 128-dimensional embedding compared to their 512-dimensional embedding, but their transformer architecture outperforms our convolutional architecture in the spectroscopic redshift prediction task.

[1]Word Count: 3491 (text); 365 (figure captions).

Contents

1 Introduction

The size of scientific datasets, particularly in the field of astronomy, has been growing at an ever-increasing rate over the last couple of decades. Spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS) (York et al., 2000) and more recently the Dark Energy Spectroscopic Instrument (DESI)[1] have been collating millions of galaxy spectra, while photometric surveys such as the DESI Legacy Survey (Dey et al., 2019) has been imaging large portions of the sky extracting millions of sources. 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, 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[2] which can probabilistically identify sources from photometric surveys and infer properties such as morphological classification.

[1] https://www.desi.lbl.gov/
[2] https://github.com/dstndstn/tractor

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 from the DESI Early Data Release (Collaboration et al., 2023) Bright Galaxy Survey for the purposes of outlier detection. Similarly, Stein et al. (2021) train a 2D convolutional image embedder using a self-supervised technique on galaxy images from the DESI Legacy Survey for the purposes of similarity search. They use the Moco v2 self-supervised learning framework (He et al., 2020; Chen et al., 2020), which is a technique to learn image embeddings by maximising the similarity of the embedding between augmented views of the same image, and minimising the similarity between the embedding of different images. Hayat et al. (2021) also use this technique to train a 2D convolutional model to estimate distances to galaxies from their photometric images, and further demonstrate that the learned embeddings can be fine-tuned very effectively for redshift estimation. They also show two important conclusions, (1) self-supervised pre-training followed by supervised fine-tuning can achieve the same performance as supervised training from scratch, whilst requiring significantly fewer labels; and (2) fine-tuning the self-supervised pre-trained model on the entire Main Galaxy Sample of the SDSS outperforms state-of-the-art supervised learning methods. These conclusions are not uncommon in the field of machine learning outside of astronomy (NLP: Devlin et al. (2019); Radford et al. (2018); Medical Imaging: Shin et al. (2016)) 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 aim to 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

For a detailed description of AstroCLIP's original implementation, we refer the reader to Parker et al. (2024), but given that this is a reproduction of their work, we provide a brief overview of their method and key results here. Their approach consisted of two main components, (1) self-supervised pre-training of a novel transformer-based spectrum embedder and a vision transformer image embedder to generate high quality single-modal embeddings followed by (2) a contrastive learning approach to align the two embedders into a shared latent space.

Given that galaxy spectra and galaxy images are two different views on the same underlying physical object, the contrastive learning approach relies on the assumption that there is enough mutual information between the two views to align them into a shared latent space. The embedders are aligned by maximising the similarity of the image and spectra embedding that refer to the same object, and minimising the similarity of the embeddings of different objects. This is done under the InfoNCE loss (van den Oord et al., 2019), as per Equation (2.1)

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{K}\sum_{i=1}^{K} \log \frac{\exp(S_C(\mathbf{x}_i, \mathbf{y}_i)/\tau)}{\sum_j^K \exp(S_C(\mathbf{x}_i, \mathbf{y}_j)/\tau)}. \tag{2.1}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are an embedded batch of galaxy spectra and images respectively, K is the batch size, S_C is the cosine similarity function, and τ is a smoothing parameter (often called the temperature). Indices i correspond to positive pairs (embedded image-spectra pairs which correspond to the same galaxy) and j corresponds to negative pairs (embedded image-spectra pairs which correspond to different galaxies). Formally, S_C, as defined in Equation (2.2)

$$S_C(\mathbf{x}_i, \mathbf{y}_j) = \frac{(\mathbf{x}_i)^T \mathbf{y}_j}{||\mathbf{x}_i||_2 \, ||\mathbf{y}_j||_2} = \cos(\theta_{i,j}) \tag{2.2}$$

is the normalised dot product between any two embeddings, which is equivalent to the cosine of the angle $\theta_{i,j}$ between the two embeddings. The similarity measure S_C is bounded in the range [-1, 1].

Parker et al. (2024) use a novel transformer based architecture for the spectrum embedder as opposed to the more traditional 1D convolutional architecture which is typically used for spectral data, and they pre-train it through a self-supervised mask filling task. The image embedder is a standard vision transformer ViT (Dosovitskiy et al., 2021) architecture which is pre-trained using the DINO v2 self-supervised learning framework (Oquab et al., 2024). We refer the reader to the original paper for more details on the pre-training and see Section (3) for details on the dataset used (as we use the same dataset for our reproduction). Following the training, the two embedders are aligned under the InfoNCE loss, using a large batch size of 1024 to ensure that there is a significant number of negative samples per batch. The final shared latent dimensionality is $\mathbf{z} \in \mathbb{R}^{512}$.

2.1 Key Results

To demonstrate the effectiveness of their AstroCLIP embeddings, Parker et al. (2024) evaluate their embeddings on a variety of downstream tasks: (1) in-modal and cross-modal retrieval of galaxy spectra and images through cosine similarity; (2) zero-shot and few-shot in-modal redshift predictions; (3) zero-shot and few-shot in-modal galaxy property (stellar mass, metallicity, age and star formation rate) predictions.

The similarity search successfully extracts visually similar spectra and images. The zero-shot redshift estimations were performed by predicting the redshift of a galaxy using the redshift of its nearest 16 embedded neighbours of the same modality. Their zero-shot image redshift predictions gave an R^2 metric (coefficient of determination) of 0.78 (with predicted redshift plotted against true redshift), and their zero-shot spectrum redshift predictions gave an R^2 metric of 0.98; these are shown on the top row of Figure (4) - this Figure is discussed in detail in Section (6). Their image results outperformed the supervised baseline models (ResNet18 and MLP) as well as the state-of-the-art self-supervised model for galaxy images (Stein et al., 2021). In this paper we attempt to reproduce the metrics listed above, but we refer the reader to the original paper for a more detailed analysis of the results.

3 Data

We use the dataset as provided exactly by Parker et al. (2024), with minor adjustments. The dataset contains 197,976 galaxy image-spectra pairs, along with their redshift measurements (split 80/20 train/validation with 158,377 train samples and 39,599 validation samples). The galaxy spectra were taken from the Dark Energy Spectroscopic Instrument (DESI) Early Data Release (Collaboration et al., 2023) and the galaxy images were taken from the DESI

Legacy Survey (Dey et al., 2019). We summarise the key pre-processing steps relating to the data below.

3.1 DESI Legacy Survey Images

The galaxy image dataset was curated by Stein et al. (2022) from the DESI Legacy Survey Data Release 9[3], we refer the reader to Stein et al. (2022); Parker et al. (2024) for a more comprehensive overview of the dataset and its curation, but the key points are summarised here. These images were taken by 3 different telescopes, with each telescope focusing on a different combination of sky area and wavelength range, creating a survey of the sky with a sky coverage of 14,000 deg^2, at a pixel resolution of 0.262 arcseconds, across the g, r, and z wavelength bands (see Dey et al. (2019) for the exact transmittance by wavelength detail of each of these filters). The Tractor is used to probabilistically identify and infer properties such as morphological classification of sources within the survey. This creates a Tractor catalogue of each identified source, and a sweep catalogue is a subset of this information. Stein et al. (2022) then filter the dataset as follows: they drop any source that were identified as a star in the sweep catalogues (this is where Tractor identifies a best-fit morphological model of point-spread function, which indicates star); and drop any sources with a z-band magnitude (mag_z) larger than 20. Each remaining source is extracted through a 256x256 pixel cut-out centred on the source, in each of the (g, r, z) bands, yielding a total of 76,446,849 images. Parker et al. (2024) cross-match this dataset for their corresponding spectra from the DESI Early Data Release using the targetIDs associated with the sources, to yield a final dataset of 197,976 galaxy spectra and image pairs.

We perform a variety of augmentations on the images including random horizontal and vertical shifts, flips and rotations, as well as adding Gaussian noise. The 256x256 pixel images all cover more sky than the source being analysed, so we follow Stein et al. (2022) in center cropping the images to the central 96x96 pixels, adverse to the 144x144 cut-out performed by Parker et al. (2024). The choice of the 96x96 cut-out as opposed to a 144x144 cut-out is primarily due to the pre-training of our image embedder as explained in Section (4.1). The images are then standardised, ensuring that each channel (g, r, z) of each image had a mean of 0 and a standard deviation of 1. The Gaussian noise $\mathcal{G}(0, 0.03)$ was added on top of this standardised data such that each channel was noised proportionately.

3.2 DESI Early Data Release Spectra

The spectra were extracted from the DESI Early Data Release (Collaboration et al., 2023) and only those that were successfully cross-matched with the galaxy images were kept. The spectra ranged the wavelength range of [3600Å, 9824Å], with a resolution of 7781 bins per spectrum. As with the image data, the spectra were also standardised individually, to mean 0 and standard deviation 1, and then noise $\epsilon_{sp}(\lambda)$ was added in the form of scaled Gaussian noise as given in Equation (3.1).

$$\epsilon_{sp}(\lambda) = \gamma \cdot \sigma_{sp}(\lambda) \cdot \mathcal{G}(0, 1) \tag{3.1}$$

where $\epsilon_{sp}(\lambda)$ is the noise added to the spectrum at wavelength λ, γ is a constant scaling factor, $\sigma_{sp}(\lambda)$ is the standard deviation of the spectrum at wavelength λ, and $\mathcal{G}(0, 1)$ is a standard Gaussian random variable. $\sigma_{sp}(\lambda)$ was computed for the training set such that $\sigma_{sp}(i)$ ($i \in [0, 7781]$) gave the standard deviation of all 158,377 train spectra measurements

[3]https://www.legacysurvey.org/dr9/description/

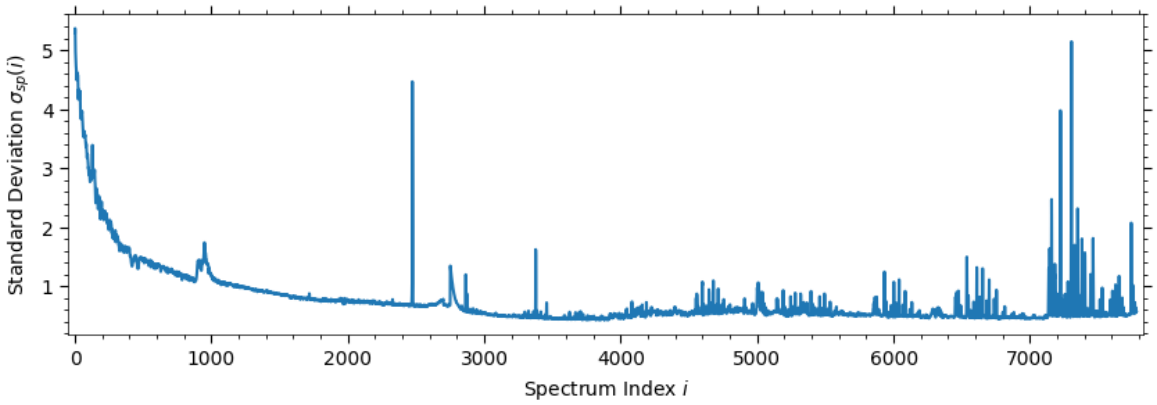


Figure 1: $\sigma_{sp}(\lambda)$ for the training set. Each spectrum had 7781 bins, the x-axis represents the bin index, and the y-axis represents the standard deviation of the standardised bin value across all 158,377 training spectra, for that given bin.

at the i^{th} bin. This computed $\sigma_{sp}(\lambda)$ is shown in Figure (1). This was useful as it meant that the noise added to the spectra at any given wavelength was proportional to the variance of the spectra at the given wavelength, which complemented the fact that some wavelengths naturally had more variance than others, as shown by Figure (1). γ was set to 0.3.

3.3 Redshift measurements

The redshift measurements utilised were the catalog-reported measurements and were compiled and provided directly with the dataset by Parker et al. (2024).

3.4 Further Pre-processing

The data was pre-processed further to remove outliers and ensure data was sensible to improve training dynamics. We dropped any galaxies with a redshift outside the range $[0, 0.8]$ and also dropped any galaxies which had every element in their spectra equal to 0 (given that they were clearly erroneous). This resulted in removing a further 1,508 samples from the training dataset and 380 samples from the validation set. Our redshift range of $[0, 0.8]$ was larger than the range of $[0, 0.6]$ used by Parker et al. (2024) in their final results.

4 AstroCLIP Implementation

Our AstroCLIP model was implemented using a pretrained 2D convolutional image embedder courtesy of Stein et al. (2021) and a pretrained 1D spectrum embedder courtesy of Liang et al. (2023). These were placed into our unified AstroCLIP model and then fine-tuned (with all the weights set as trainable) using contrastive learning under the InfoNCE loss as per Equation (2.1), to learn the shared low-dimensional embedding space. Embeddings were projected into a variety of dimensions: $[8, 16, 32, 64, 128, 256, 512]$ to explore the effect of the dimensionality of the embedding space on the performance of the model, with Parker et al. (2024) using a 512 dimensional embedding for their model. Below we give a brief overview of the pretraining strategies for the image and spectrum embedders, but for a comprehensive report we refer the reader to the original papers.

4.1 Image Embedder

Stein et al. (2021) trained a galaxy image embedder using a self-supervised method for the purposes of image based similarity search. The image embedder was trained via the MoCo v2 framework (He et al., 2020; Chen et al., 2020) on a sample of 42 million galaxies from the DESI Legacy Survey Data Release 9 (Dey et al., 2019). The encoder architecture used was a ResNet-50 architecture. The images underwent a variety of augmentations (galactic extinction, random rotation, blurring, noising) before being cropped to the central 96x96 pixels. The loss function used was the same InfoNCE loss we used in our AstroCLIP model as given in Equation (2.1), and the similarity measure used was also the cosine similarity function as given in Equation (2.2). Their model yielded very promising results, and we refer the reader to the original paper for a thorough discussion on the trained model. The final fully-connected layers of this model contained 3 layers with $[2048, 2048, 128]$ units respectively. The final 128 unit layer was adjusted as required to produce the output dimensionality we desired, but otherwise all weights and biases were initialised as per the pretrained model. This yielded between 27.7M and 28.8M trainable parameters depending on the output dimensionality of the model.

4.2 Spectrum Embedder

Liang et al. (2023) trained a galaxy spectrum autoencoder using galaxy spectra from the DESI Early Data Release (Collaboration et al., 2023), focusing primarily on the Bright Galaxy Survey (BGS) (for precise details on the spectral dataset used, see the original paper). The autoencoder architecture, named SPENDER, was proposed by Melchior et al. (2022) and encoded the spectra into a 6-dimensional embedding space. For our AstroCLIP model, we used the encoder part of the SPENDER model which consisted of 3–1D convolutional blocks followed by 5 fully-connected layers to bring the dimensionality down to 6 (units per layer: $[256, 128, 64, 32, 6]$). We reduced the fully-connected layers to 3 layers with $[256, 128, D_{out}]$ units respectively, for any desired output dimensionality $D_{out} \leq 128$, and $[256, 256, D_{out}]$ for $D_{out} > 128$. All weights and biases were retained for the convolutional layers and all weights and biases that could be retained for the fully-connected layers were retained, the rest were reinitialised from a normal distribution. This yielded between 3.1M to 3.3M trainable parameters, depending on the output dimensionality of the model.

5 Model Training

Our AstroCLIP model unifies these two embedders into a single model, and then fine-tunes the embeddings using the InfoNCE loss (Equation (2.1)). The embedding dimensionality was varied through the range $[8, 16, 32, 64, 128, 256, 512]$ and each model was trained for 75 epochs which took approximately 3.5 hours (per model) on a single NVIDIA A100-SXM4-80GB GPU.

Through some exploratory analysis, we fixed a few hyperparameters: batch size of 512; learning rate of $5e^{-4}$; a weight decay of $1e^{-4}$; and a temperature parameter in the InfoNCE loss of 0.07; and only varied the embedding dimensionality. One small aspect of note is that the batch size was not consistent throughout training, the pre-processing mentioned in Section (3.4) was performed on the fly on each batch of data as and when randomly sampled by the DataLoader during training. This meant that there were small fluctuations in the batch size, and for completeness we show the batch size for each batch throughout the training of the AstroCLIP models in Figure (2). As can be seen on the figure, the batch size fluctuates very little and generally remains between 502 and 510. This fluctuating batch size was intentionally

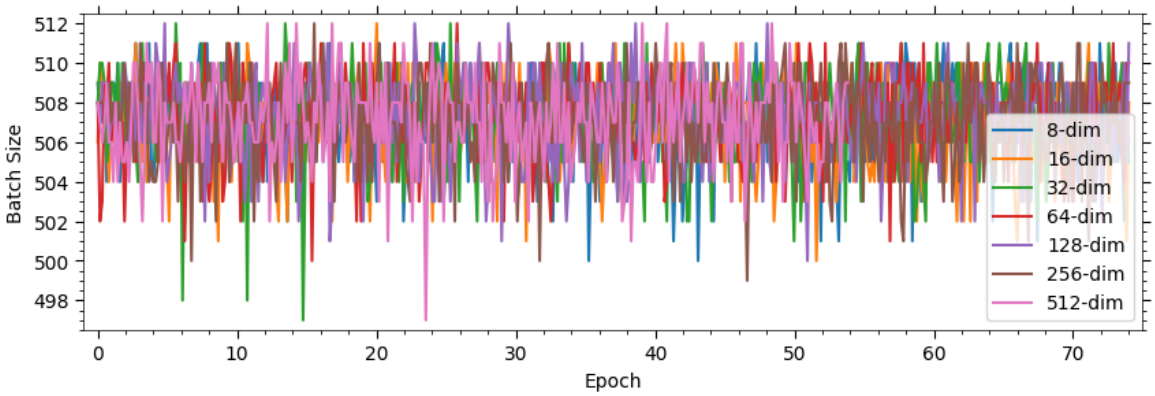


Figure 2: The batch size for each batch after the pre-processing steps were applied to the batch on the fly. Each line represents the batch size throughout the training of each AstroCLIP model with everything identical except the embedding dimensionality. The individual lines are not meant to be discernable from one another, but rather this figure should show the general trend.

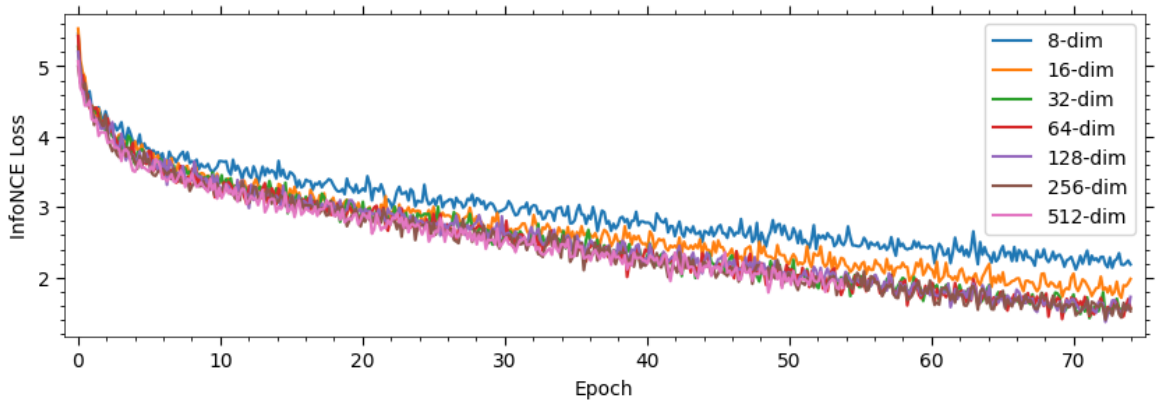
left in the training process as it's effects were negligible and would likely only have a positive effect through the stochasticity it introduces. Figure (3) shows the training and validation loss for each model throughout the 75 epochs, and Figure (3b) details the lowest validation loss achieved for each model and the corresponding epoch. Whilst there is a clear pattern in the training loss where the more flexible model appears to be able to achieve the lowest loss, the same was not true for the validation loss with no clear pattern emerging. However, it is important to note that a thorough sweep of hyperparameters was not performed, and so it is possible that a different set of hyperparameters (such as stronger regularisation on the more flexible models, to prevent over-fitting) could have yielded different results (lower validation loss for the more flexible models).

6 Results

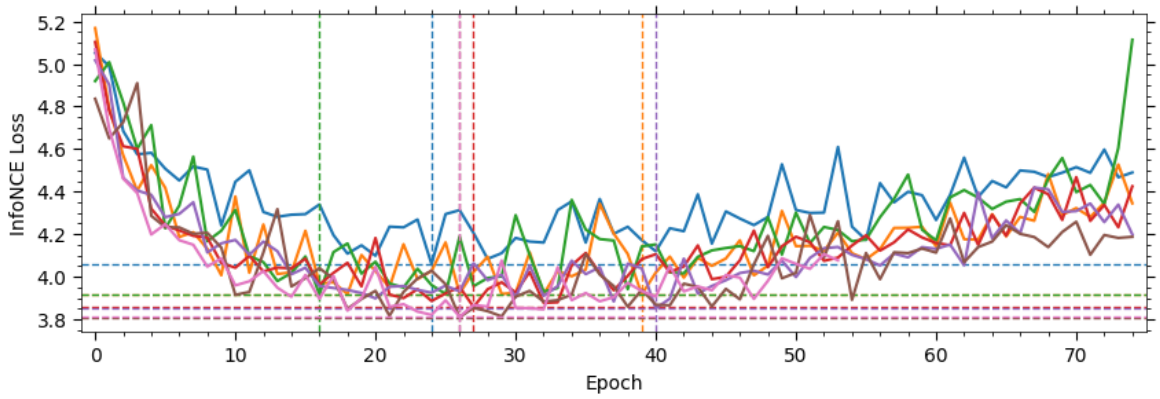
To assess and demonstrate the performance of our AstroCLIP model, we reproduce a subset of the downstream tasks from the original implementation by Parker et al. (2024). The models with the lowest validation loss for each embedding dimensionality (as shown in Figure (3b)) were selected for evaluation. The held-out validation set (composing of 39,599 of the 197,976 image-spectra pairs) was used to evaluate the performance by assessing the accuracy of zero-shot k-NN redshift estimations and qualitative similarity retrieval tasks. Notation used in this section intentionally follows that of the original paper for the reader's convenience.

6.1 Zero-shot k-NN Redshift Estimation

Figure (4) shows the zero-shot redshift predictions using k-NN regression on the learned embeddings for the best 128-dimensional embedding model, and compares this to the 512-dimensional results from Parker et al. (2024). Our 128-dim image embedder outperforms the 512-dim image embedder from Parker et al. (2024) in the photometric redshift prediction task (across a larger range in redshift values) with an R^2 value of 0.80 compared to their 0.79; but falls short in the spectroscopic redshift prediction task with an R^2 value of 0.94



(a) Training loss.



(b) Validation loss. The vertical lines correspond to the epoch at which the lowest validation loss was achieved and the horizontal line corresponds to the loss value at that epoch. The (embedding dimensionality, lowest validation loss, epoch) values are as follows: (8, 4.06, 24); (16, 3.92, 39); (32, 3.92, 16); (64, 3.86, 27); (128, 3.85, 40); (256, 3.81, 26); (512, 3.81, 26).

Figure 3: Train and validation loss for AstroCLIP models with varying embedding dimensionality. The legend on Figure (3a) applies to both plots.

compared to their 0.98. Figure (5) then shows this same data in a more succinct manner for all embedding dimensionalities. For a like-by-like comparison to Parker et al. (2024), our 512-dim embedding performs marginally worse in the photometric redshift prediction task with an R^2 value of 0.79, and marginally better in the spectroscopic redshift prediction task with an R^2 value of 0.95. These results demonstrate strong evidence that our AstroCLIP model is able to learn a meaningful representation of the data, as the redshift predictions are highly correlated with the true redshift values. More interestingly, our results show that even low-dimensional embeddings are able to perform well in the zero-shot redshift prediction task, with the 8-dim and 16-dim embeddings generally performing at a similar level to the higher dimensional embeddings across all prediction types.

Architecturally, these results indicates that the vision transformer architecture used by Parker et al. (2024) gave no significant advantage over the 2D convolutional architecture used in this work for the image embedder. However, the transformer architecture used for the spectrum embedder by Parker et al. (2024) outperformed the 1D convolutional architecture

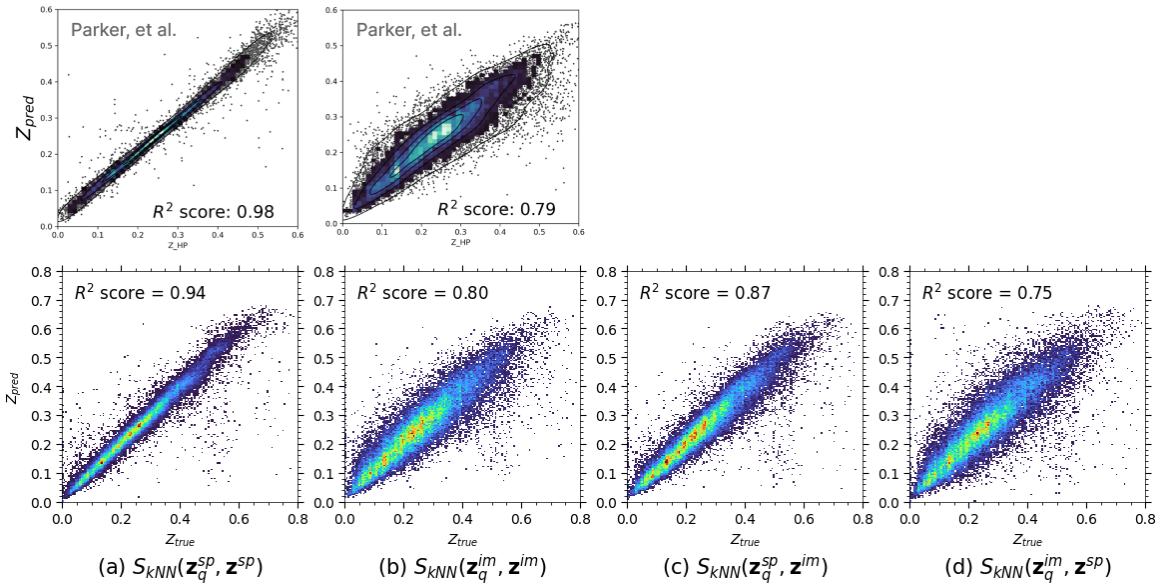


Figure 4: k-NN regression for zero-shot redshift prediction, showing predicted vs. true redshift. Top row is results from Parker et al. (2024) for their 512-dim embedding model, bottom row is results from our 128-dim embedding model. $S_{kNN}(\mathbf{z}_q^{sp}, \mathbf{z}^{im})$ indicates the cross-modal prediction type where a spectrums redshift was predicted using its 16 closest embeddings derived from galaxy images.

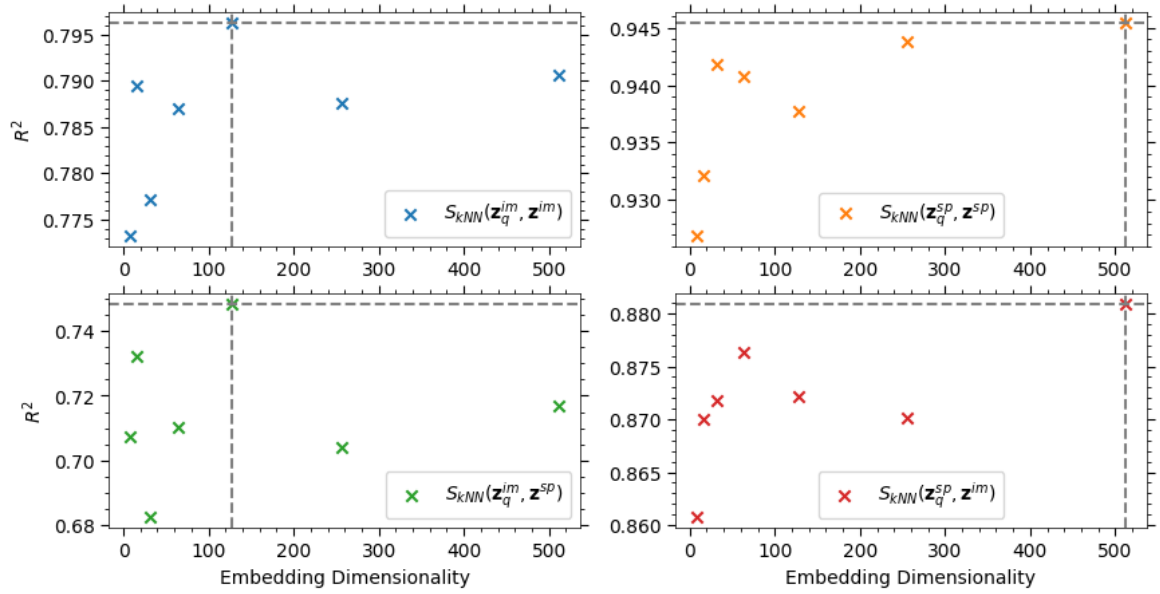


Figure 5: Consolidated results for all embedding dimensions and prediction types, displaying R^2 values vs. embedding dimensionality. These are the R^2 values derived if Figure (4) were plotted for all embedding dimensionalities.

used in this work for the spectrum embedder, and adds promise to the potential of transformer architectures for spectral data.

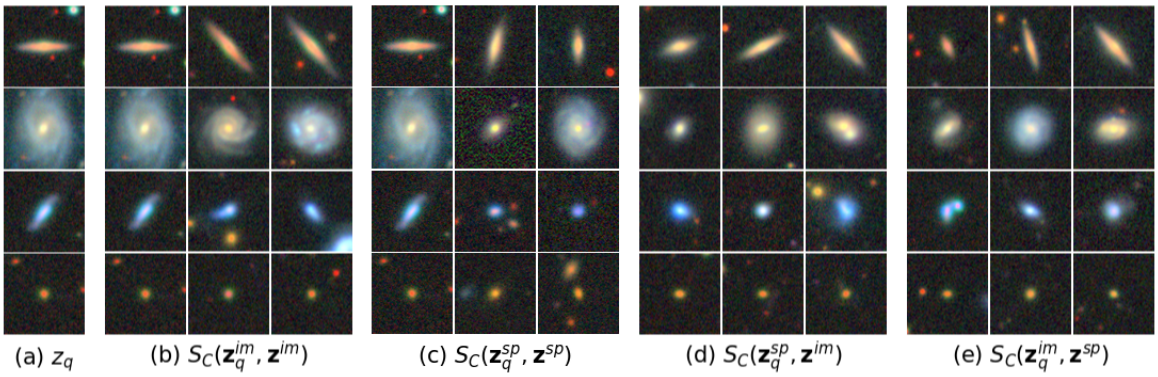


Figure 6: Cosine similarity search results using the 128-dimensional model. (a) 4 Query galaxies; (b, c, d, e) Top 3 similar galaxies across the 4 prediction types. By construction, the most similar in-modal galaxy to any given galaxy is itself, hence, the first column of images in (b) and (c) are identical to the query image in (a).

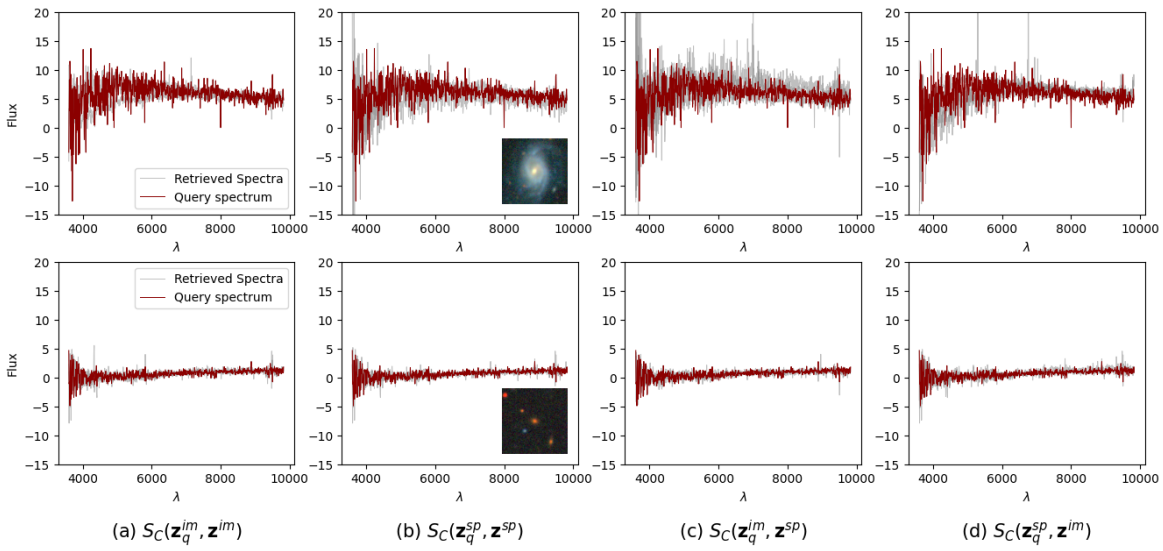


Figure 7: Cosine similarity search for spectra using the 128-dimensional model. Each row shows a query galaxy (galaxy imaged in (b)), each subfigure shows the 3 most similar spectra to the query galaxy for the given prediction type.

6.2 Retrieval by Cosine Similarity

The best 128-dimensional embedding model was used to perform similarity searches using cosine similarity (Equation (2.2)) on the learned embeddings. Unlike Stein et al. (2021), our model is not constrained to only performing in-modality image searches and Figures (6) and (7) show the results of both in-modal and cross-modal similarity searches for images and spectra respectively. The results show that the model is able to retrieve similar images and spectra to the query image and spectrum, respectively, even across modalities. As described by Stein et al. (2021), this ability is critical when searching for rare astronomical sources and the cross-modal ability of the model adds significantly to this.

7 Conclusion

In this report, we present a successful reproduction of the AstroCLIP model by Parker et al. (2024). We utilise different architectures for the image and spectrum embedders, but ultimately achieve similar results and show that the model is able to learn a meaningful representation of the data. We also explore the effect of different embedding dimensionalities on the performance of the model and show that even low-dimensional embeddings are able to capture a meaningful amount of mutual information between the two modalities, given their strong performance in the zero-shot redshift prediction task. Our results do bring into question the optimal architecture for the image embedder and optimal embedding dimensionality, as our 128-dimensional model marginally outperforms the 512-dimensional model by Parker et al. (2024) in the in-modal image to image redshift prediction task. Nonetheless, 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

Chen, X., Fan, H., Girshick, R., and He, K. (2020). Improved baselines with momentum contrastive learning. *arXiv:2003.04297*.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

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*.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *arXiv:1911.05722*.

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*.

Melchior, P., Liang, Y., Hahn, C., and Goulding, A. (2022). Autoencoding galaxy spectra i: Architecture. *arXiv:2211.07890*.

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*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., et al. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *EEE Transactions on Medical Imaging*, 35(5):1285–1298.

Stein, G., Blaum, J., Harrington, P., Medan, T., and Lukic, Z. (2022). Mining for strong gravitational lenses with self-supervised learning. *arxiv:2110.00023*.

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*.