NEWASTROCLIP: CROSS-MODAL PRE-TRAINING FOR ASTRONOMICAL FOUNDATION MODELS

Jingjing Gao

Supervisor: Prof. Miles Cranmer

Department of Physics
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy

St.Edmunds's College June 2024

Abstract

Motivated by Astroclip, I employ contrastive learning with InfoNCE loss to align spectrum and image representations. Drawing from Spender, I develop a spectrum encoder to capture spectrum features, while inspired by MocoV2, I use an image encoder to map image features into a shared space. I introduce newAstroCLIP, a methodology aimed at constructing foundational astronomical models that unify diverse observational modalities. My approach demonstrates that employing cross-modal contrastive learning between galaxy images and optical spectra produces highly informative embeddings for both modalities. Specifically, I apply this method to multi-band images and optical spectra obtained from the Dark Energy Spectroscopic Instrument (DESI) and find that: (1) these embeddings effectively align across modalities, facilitating accurate cross-modal searches, and (2) these embeddings encode crucial physical characteristics of galaxies, such as redshift, enabling competitive zero-shot predictions without additional fine-tuning efforts.

Acknowledgements

I gratefully acknowledge the invaluable support, discussion, and supervision provided by Prof. Miles Cranmer. Additionally, I am thankful for the suggestions and assistance from Mohammed Hadi Sotoudeh, Adnan Siddiquei, and Andreas Vrikkis, who are collaborating on the same data analysis project. I extend my appreciation to all members of the 2024 DIS programme. Special thanks to Prof. James Robert Fergusson for the opportunity to participate in the Data Intensive Science program at the University of Cambridge.

Table of contents

1 Introduction **1**

1.1 The Challenge in Astronomical Surveys 1

1.2 AstroCLIP: Cross-Modal Pre-Training for Astronomical Foundation Models 1

1.3 Inspired Methodology: Integrating Spender's Spectrum Encoder and MoCo v2 2

2 Literature Review **3**

2.1 Contrastive Training 3

2.2 AstroCLIP: Contrastive Learning in Galaxies 4

2.3 Self-Supervised Training for Astronomical Images 6

2.4 Representation Learning for Galaxy Spectra 6

3 Methodology **9**

3.1 Training Objective 9

3.1.1 Introduction to Contrastive Learning 9

3.1.2 InfoNCE Loss and Its Importance 10

3.2 Implementation 11

3.2.1 Pretrained Spectrum Embedder 11

3.2.2 Pre-training Strategy of Spectra Encoder 13

3.2.3 Pretrained Image Embedder 16

3.2.4 Contrastive Training: NewAstroCLIP Architecture 21

4 Experiments **23**

4.1 Dataset 23

4.1.1 Dataset Description and Preprocessing 23

4.1.2 Data Augmentation 24

4.2 Training Process Analysis 24

4.2.1 Spectrum Encoder with MLP 24

4.2.2 Spectrum Encoder without MLP 25

4.2.3 Analysis of Validation Loss Behavior 25
4.3 4.3 Downstream Tasks and Results Analysis 26
4.3.1 4.2.1 Similarity Search 26
4.3.2 4.2.2 Zero-shot Regression of Redshift 29

5 Conclusion **35**
5.1 Conclusion 35

References **37**

Chapter 1

Introduction

1.1 The Challenge in Astronomical Surveys

Foundation models in machine learning have revolutionized the ability to create interconnected systems that transcend traditional scientific boundaries. However, in many physical domains, the diversity in measurement methods results in disjoint data representations. Machine learning approaches often specialize in specific measurement types, limiting their comprehensive information extraction from diverse observations. This creates a pressing need for methodologies capable of integrating information from varied observational modalities into a unified framework.

A pivotal area where this integration is crucial is in large-scale astronomical galaxy surveys. These surveys encompass tens of millions to billions of galaxies, observed through different telescopes and instruments. Imaging surveys provide color images of galaxies, while spectroscopic surveys measure optical spectra—quantitative measurements of light across different wavelengths emitted by galaxies. Both types of observations offer complementary insights into the nature of galaxies. However, analyzing these vast datasets remains challenging due to the absence of standardized labels and representations.

1.2 AstroCLIP: Cross-Modal Pre-Training for Astronomical Foundation Models

In October 2023, Polymathic AI introduced AstroCLIP [12] (Cross-Modal Pre-Training for Astronomical Foundation Models). This innovative framework draws inspiration from CLIP [14] (Contrastive Language-Image Pretraining) and aims to integrate galaxy image and spectra data into a unified latent space. AstroCLIP aligns spectroscopic and imaging data

based on shared semantic information, embedding information from both modalities into a shared representation. The underlying concept of AstroCLIP is that different observational perspectives represent filtered views of the same underlying physical processes, implying an inherent shared latent space.

AstroCLIP employs a sophisticated pre-training strategy to maximize mutual information between imaging and spectroscopic modalities, effectively aligning their representations. It utilizes a transformer-based spectrum encoder similar to the GPT-2 [15] architecture, trained to infer masked segments of galaxy spectra in a self-supervised manner. Concurrently, a convolutional image model is adapted to galaxy images, incorporating physical augmentations to enhance robustness. Fine-tuning under a contrastive objective ensures that image-spectra pairs from the same galaxy are closely aligned in the embedding space, while minimizing similarities between different galaxies.

1.3 Inspired Methodology: Integrating Spender's Spectrum Encoder and MoCo v2

Inspired by AstroCLIP, my project integrates Spender's [11] spectrum encoder, renowned for its ability to capture essential spectral features using convolutional layers and attention mechanisms. This encoder is pivotal in extracting meaningful information from galaxy spectra. For the image encoder, I use MoCo v2 [2], which is adept at handling image data with its robust representation learning capabilities.

By aligning these embeddings through contrastive learning, the project demonstrates their utility in downstream tasks. One such task is Principal Component Analysis (PCA), which reveals the embedded representations' ability to encode critical physical properties like galaxy redshift. These embeddings encapsulate valuable high-level information, enabling simple techniques like k-nearest neighbor prediction to meaningfully predict physical properties of galaxies, such as their redshift, from the embedded representations.

Chapter 2

Literature Review

2.1 Contrastive Training

In recent years, contrastive training has emerged as an effective paradigm for learning meaningful representations of data in cross-modal domains. By bringing similar data closer and pushing dissimilar data apart in the embedding space, the model learns robust representations of the underlying data.

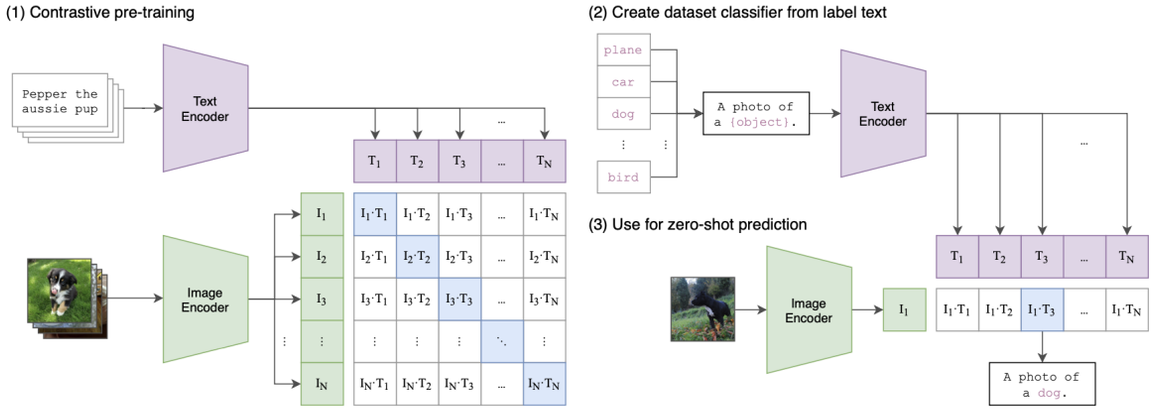

Fig. 2.1 CLIP pre-trains an image encoder and a text encoder to predict which images are paired with which texts. This behavior is then leveraged to turn CLIP into a zero-shot classifier. The approach involves converting all of a dataset's classes into captions, such as "a photo of a dog," and predicting the class of the caption that CLIP estimates best pairs with a given image.

OpenAI released CLIP [14] (Contrastive Language-Image Pretraining), where contrastive learning connects language and image representations of the same objects. They demonstrate that the simple pre-training task of predicting which caption goes with which image is an

efficient and scalable way to learn image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones), enabling zero-shot transfer of the model to downstream tasks. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset-specific training. The summary of their approach is in Fig. 1. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time, the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

2.2 AstroCLIP: Contrastive Learning in Galaxies

Polymathic AI released AstroCLIP [12] - the first cross-modality pretraining model for galaxies. The approach takes information about galaxies from two separate modalities - images and optical spectra - and embeds and aligns both into a shared embedding space, which can be used as a foundation for downstream tasks. The illustration of the AstroCLIP cross-modal training strategy is in Fig. 2. AstroCLIP embeds both the optical spectra and the images of galaxies into a shared embedding space and aligns spectra and images of the same galaxy through self-supervised contrastive learning.

AstroCLIP develops the first transformer-based model for galaxy spectra, along with an effective pre-training strategy for this model. As their architecture for the spectrum embedder, they adopt a transformer model structured similarly to GPT-2. They pretrain this transformer only on spectra first, using a self-supervised learning paradigm. They randomly replace 6 contiguous segments of length 30 (equivalent to length 600 in the original spectra representation) with zeros and train the model to minimize the Mean Square Error loss between the predictions and the ground truth on the replaced segments of the sequence. Once this mask-filling model has been trained, they freeze its weights and use a single cross-attention block (cross-attention layer with 4 heads and embedding dimension of 128 followed by an MLP) to extract a short embedding vector. They use the output of the final transformer block of the mask-filling model as the key and value and use a learnable sequence of size 1×128 as the query vector. The output of this procedure is a single vector of length 128. The weights of this cross-attention block and the 128 parameters of the query vector amount to 362k total parameters which are then trained via the contrastive training procedure.

They demonstrated that their cross-modal embeddings are well-aligned and can be used for accurate cross-modal searches. They demonstrate that their embeddings encode valuable

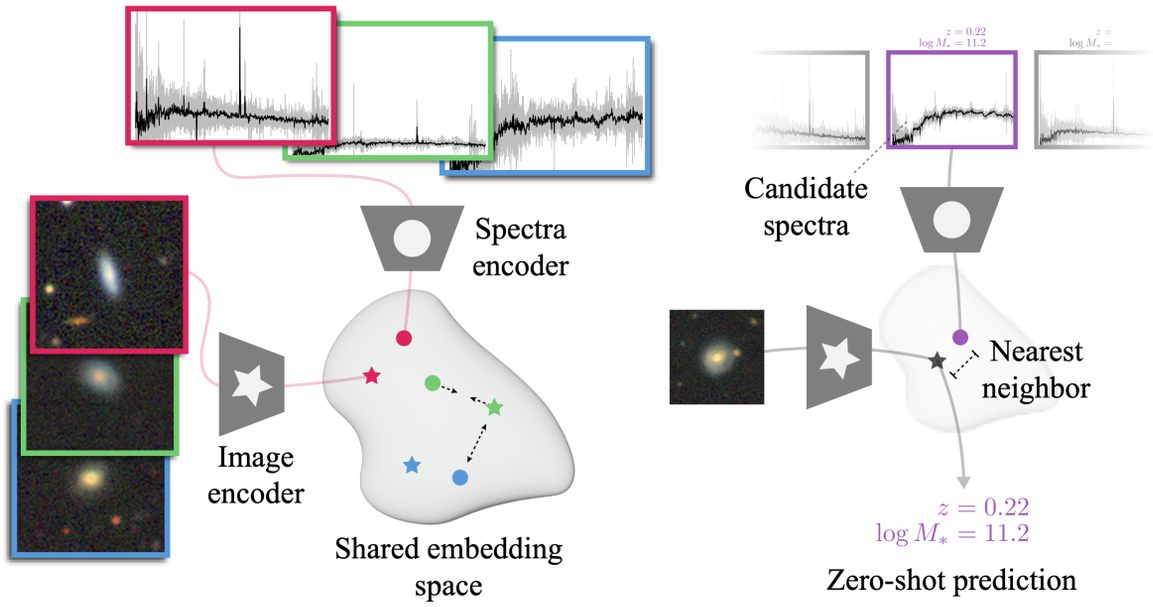

Fig. 2.2 Illustration of the AstroCLIP cross-modal training strategy. This approach embeds the images and optical spectra of galaxies into a shared embedding space, utilizing cross-modal contrastive learning to align these embeddings based on shared semantic information. These embeddings capture physically meaningful high-level features, which can subsequently be used for various downstream tasks, such as k-nearest neighbor regression and zero-shot prediction for redshift and stellar mass.

high-level information that can be used for downstream tasks, as even simple techniques like k-nearest neighbor prediction can meaningfully predict physical properties of galaxies such as their redshift from the embedded representations. The core idea of AstroCLIP is to use a self-supervised image encoder to get image representation and use a spectra encoder to get spectra representation. This includes self-supervised training for astronomical images and representation learning for galaxy spectra.

2.3 Self-Supervised Training for Astronomical Images

Reliable tools to extract patterns from high-dimensionality spaces are becoming more necessary as astronomical datasets increase both in volume and complexity. Contrastive Learning is a self-supervised machine learning algorithm that extracts informative measurements from multi-dimensional datasets. To do so, it maximizes the agreement between the information extracted from augmented versions of the same input data, making the final representation invariant to the applied transformations. Contrastive Learning is particularly useful in astronomy for removing known instrumental effects and for performing supervised classifications and regressions with a limited number of available labels, showing a promising avenue towards Foundation Models.

One of the earliest works in this direction is the application of large-scale, self-supervised contrastive learning to galaxy images [6] [16]), exploiting a MoCo V2 approach [2] in their later work. The embeddings generated from this approach have been shown to encode valuable astrophysical information, which can be used to predict galaxy properties (such as redshift in Hayat et al. [6]) as well as to perform similarity searches (for instance, to identify rare but scientifically interesting events such as strong gravitational lenses [16]). Another prominent example in this field is the application of a similar BYOL self-supervised training strategy [5] for pretraining networks that can further be easily fine-tuned, even in the low data regime, for the task of classifying galaxies according to their morphologies [19].

2.4 Representation Learning for Galaxy Spectra

Finding and accurately extracting the structure in spectra without prior astrophysical knowledge has been the topic of significant attention. While historically, traditional techniques like Principal Component Analysis (PCA) have proven widely successful, a new line of inquiry using unsupervised machine learning techniques has recently emerged. Portillo et al. [13] use a variational auto-encoder (VAE) to reduce the dimensionality of galaxy spectra to a small latent space and demonstrate that the VAE embeddings of the spectra can be easily

used for downstream tasks like outlier detection, interpolation, and galaxy class classification. Teimoorinia et al. [17] improve upon the existing VAE by introducing convolutional elements into the AutoEncoder to extract correlated features from the spectra. Melchior et al. [11] further advance this approach with a specifically designed architecture that combines an attentive convolutional encoder with an explicit redshift transformation after the decoder. Their embedding is then similarly useful for downstream tasks such as anomaly detection (Liang et al., 2023b [9]).

Chapter 3

Methodology

3.1 Training Objective

The central aim of our work is to exploit the shared physical latent space that exists between different observational modalities such as images and spectra. These modalities can be thought of as filtered views of the same underlying physical phenomena. Therefore, our goal is to construct embeddings of these modalities that maximize mutual information about the underlying object, aligning the representations from different modalities around shared semantics.

Formally, let $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^M$ be examples from two different modalities. We aim to develop models $f_\theta : \mathbb{R}^N \to \mathbb{R}^d$ and $g_\theta : \mathbb{R}^M \to \mathbb{R}^d$ that map these examples to a shared d-dimensional space, maximizing the mutual information $I(f_\theta(x), g_\theta(y))$. Specifically, x represents a galaxy image and y represents a spectrum. The encoders $f_\theta(x)$ and $g_\theta(y)$ compress these inputs into a 128-dimensional space.

To achieve this, we use contrastive training under the InfoNCE loss.

3.1.1 Introduction to Contrastive Learning

Contrastive learning is a powerful approach in unsupervised and self-supervised learning where the objective is to learn representations by contrasting positive and negative pairs of data points. The main idea is to bring representations of similar (positive) pairs closer together in the embedding space while pushing representations of dissimilar (negative) pairs farther apart. This approach helps in learning discriminative features that capture the underlying structure of the data without the need for explicit labels.

Contrastive learning has been successfully applied in various domains such as computer vision, natural language processing, and more, showing impressive performance in tasks

like image classification, object detection, and sentence embedding. The learned representations often exhibit robustness and generalization capabilities, making them valuable for downstream tasks.

3.1.2 InfoNCE Loss and Its Importance

InfoNCE [18] (Information Noise-Contrastive Estimation) loss is a popular loss function used in contrastive learning to maximize mutual information between different views or modalities of the same underlying data. It operates by bringing representations of positive pairs (different views of the same object) closer together while pushing apart representations of negative pairs (views of different objects). This loss function is defined as follows:

$$L(x,y,\tau) = -\sum_i \log \frac{\exp(f_\theta(x_i)^T g_\theta(y_i)/\tau)}{\exp(f_\theta(x_i)^T g_\theta(y_i)/\tau) + \sum_{j\neq i} \exp(f_\theta(x_i)^T g_\theta(y_j)/\tau)}$$

where τ is a smoothing parameter (temperature) and j denotes indices of negative examples. We consider the spectrum and image of the same object as a positive pair and all other combinations as negative samples.

The reasons for using InfoNCE loss in our framework are:

1. **Mutual Information Maximization:** InfoNCE is designed to maximize the mutual information between different representations, which aligns with our goal of capturing the shared underlying physical information between images and spectra.

2. **Effective Contrastive Learning:** InfoNCE has been empirically proven to be effective in contrastive learning scenarios, leading to representations that are robust and generalizable.

3. **Scalability:** The formulation of InfoNCE allows for efficient computation even with large datasets, making it suitable for our use case involving millions of galaxy images and spectra.

The aim is for the embeddings from both modalities to capture the shared physical information about the galaxy images and spectra. This shared space should also facilitate highly structured information about the underlying objects, which we validate by achieving good zero-shot prediction of redshift.

To this end, we developed the contrastive learning architecture NewASTROCLIP, which uses different networks for feature extraction from spectra and images of galaxies. The spectra encoder uses a convolutional architecture optimized for up to 256 spectral features,

while the image encoder uses a ResNet50 backbone pretrained with the MoCo V2 [11] framework. Pretraining is done using a subset of the DESI Legacy Survey [4], ensuring robust initial representations. The final embeddings are then refined using contrastive learning to align the representations from both modalities.

Evaluating the new AstroCLIP model through zero-shot prediction is crucial because it tests the generalization capability of the learned embeddings without requiring further training on the specific downstream task. Zero-shot prediction involves using the learned representations directly to make predictions on new, unseen data. This evaluation method highlights the effectiveness of the embeddings in capturing relevant features and their ability to generalize across different tasks.

By performing zero-shot prediction, we can quantitatively assess the quality and utility of the learned embeddings. In our case, we use simple k-Nearest Neighbour (k-NN) regression on the embedded images and spectra to infer the redshift of galaxies. The performance of the model in zero-shot prediction tasks serves as a strong indicator of the robustness and versatility of the learned representations, confirming whether the contrastive learning approach has successfully aligned the different modalities around shared semantics.

In addition, we also evaluated the cross-modal similarity $SC(z_{sp}, z_{im})$ between a query spectrum $z_{sp} = g_\theta(x_{sp})$ and an image $z_{im} = f_\theta(x_{im})$, assessing how well the model aligns different modalities in practice.

3.2 Implementation

Since our representations come from different modalities, we employ two separate models to perform the embeddings, starting with pretrained image and spectrum embedders. These models embed the galaxy image and spectrum into a shared 128-dimensional space, which we then fine-tune using the contrastive InfoNCE loss.

3.2.1 Pretrained Spectrum Embedder

The architecture of our spectra encoder is based on the Spender [11] architecture developed by Melchior et al. The architecture of Spender is shown in figure 3.1. This encoder uses a convolutional architecture to compress 256 spectral features into a low-dimensional latent space and incorporates attention mechanisms to handle varying spectral feature locations due to redshift.

Encoder Details:

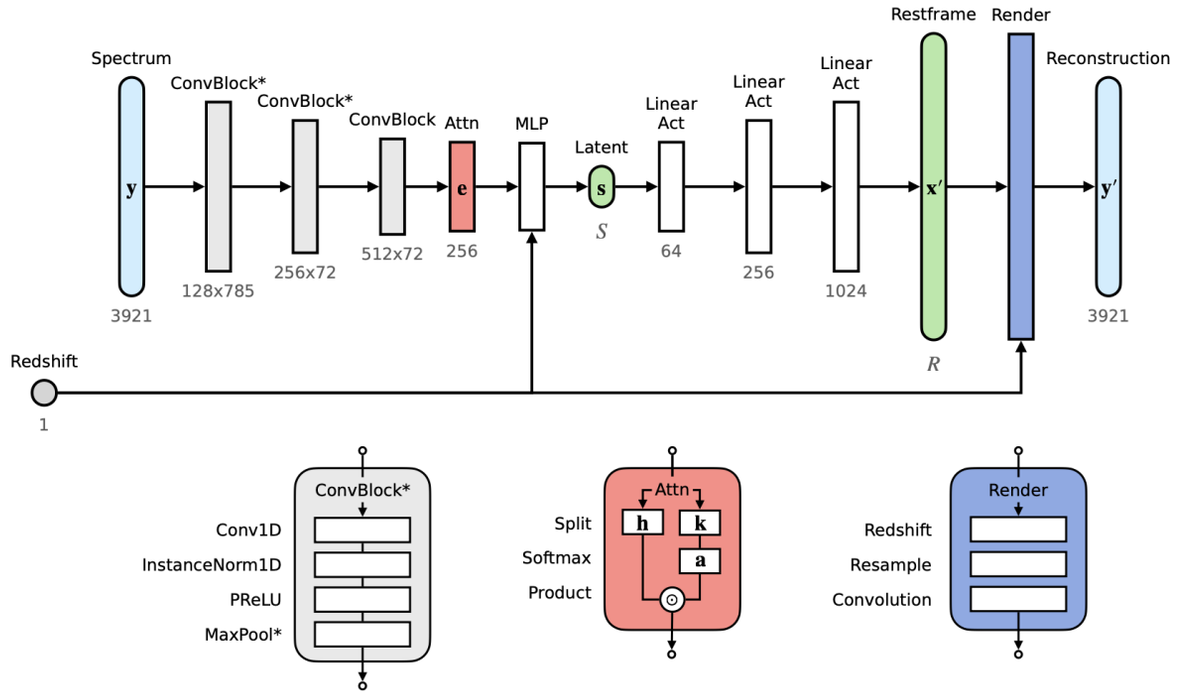

Fig. 3.1 The autoencoder architecture of spender with an attentive convolutional en- coder and an explicit redshifting, resampling, and convolution transformation after the decoder.

- **Convolutional Layers:** The encoder begins with three convolutional layers with kernel sizes of 5, 11, and 21, using trainable PReLU activations and max-pooling. This translates the spectral elements into 512 channels for 72 wavelength segments.
- **Attention Mechanism:** Attention is applied in the wavelength direction, splitting the channels into two parts, h and k, and combining them with softmax to produce attention weights a. This mechanism helps account for the apparent shift of spectral features due to redshift.
- **MLP Compression:** An extra MLP on the encoder side further compresses the attended features and the redshift into latent variables. This allows the encoder to learn the relationship between spectral features and redshift, which is crucial for generalizing across different redshifts.

Decoder Details:

- **MLP Decoder:** The decoder consists of a 3-layer MLP with hidden dimensions (64, 256, 1024), generating an internal rest-frame representation of the spectrum. This representation is then redshifted and resampled to match the observed spectrum.

The Spender architecture ensures a clear assignment of responsibilities across the autoencoder modules, providing a robust framework for encoding galaxy spectra. This architecture, combined with the contrastive learning approach, facilitates the extraction of meaningful representations from both galaxy images and spectra, enhancing our ability to perform zero-shot predictions of redshift and other downstream tasks.

3.2.2 Pre-training Strategy of Spectra Encoder

To effectively leverage the pretrained spectra encoder, I explored two different pretraining strategies for incorporating the spectra features into the shared latent space used in the contrastive learning framework. The goal is to identify the most effective way to compress and represent the spectral features in the absence of explicit redshift information.

Cut-off After MLP

In the first strategy, I followed an approach where the spectra features are compressed into the latent space using an extra Multi-Layer Perceptron (MLP). This differs from the Spender architecture as it does not include redshift information in the MLP structure.

Process:

1. **Convolution and Attention:** The spectra are processed through three convolutional blocks followed by an attention mechanism applied in the wavelength direction.

2. **MLP Compression:** The attended features are then compressed into the latent space using an MLP. This step aims to utilize the broad convolution kernels to encapsulate continuum features, forming a highly informative latent representation.

The rationale behind this strategy is that, in the contrastive learning task, I only have (spectrum, image) pairs without the explicit redshift information. Hence, the MLP is used purely to compress the attended spectral features into the latent space, independent of redshift.

Architecture

The architecture of the "Cut-off After MLP" strategy is described as follows:

- **Convolutional Layers:** The encoder starts with three convolutional layers. Each layer uses a different filter size to capture various spectral features and applies padding to maintain the spectral dimensions.

  - **Filters:** The filters used are of sizes 5, 11, and 21.

  - **Channels:** The number of output channels for the convolutional layers are 128, 256, and 512 respectively.

- **Pooling Layers:** After the convolutional layers, max pooling is applied to reduce the dimensionality and computational load. Two max pooling layers are used for the first two convolutional layers.
- **Attention Mechanism:** The output from the convolutional layers is split into two parts, where one part is used to generate attention weights. The attention weights are applied to the other part to focus on the relevant spectral features.
- **MLP for Latent Space Compression:** An MLP is used to further compress the attended features into a lower-dimensional latent space. The MLP consists of several hidden layers with activation functions and a final output layer that maps the features to the latent space.
- **Softmax for Attention Weights:** A softmax function is applied to the attention weights to normalize them and ensure they sum to one.
- **Output Latent Space:** The final output is a compressed latent representation of the spectrum, which can be used in the contrastive learning framework to align with image embeddings.

This architecture allows the spectrum encoder to learn meaningful representations of the spectra, which are then used in the contrastive learning framework to align with the image embeddings. The use of convolutional layers, attention mechanisms, and MLP ensures that the encoder captures a wide range of spectral features and compresses them effectively into a latent space suitable for contrastive learning.

Cut-off Before MLP

In the second strategy, I designed an alternative architecture where the spectrum encoder is cut off before the MLP stage. Instead of using the MLP to compress the features, a linear layer is used to map the attended feature directly into the shared 128-dimensional space. This strategy utilizes approximately 3.1 million total parameters, which are then trained via the contrastive training procedure described below.

Process:

1. **Convolution and Attention:** The spectra undergo initial processing through three convolutional blocks, followed by an attention mechanism to focus on relevant spectral features.

2. **Linear Layer Mapping:** The attended features are mapped directly to the 128-dimensional shared space using a linear layer.

The rationale behind this strategy is to investigate the necessity of the MLP and to streamline the compression process by using a simpler linear transformation.

Architecture

The architecture of the "Cut-off Before MLP" strategy is described as follows:

- **Convolutional Layers:** The encoder starts with three convolutional layers, each with different filter sizes to capture a variety of spectral features. These layers apply padding to maintain the spectral dimensions.
  - **Filters:** The filters used have sizes 5, 11, and 21.
  - **Channels:** The number of output channels for these convolutional layers are 128, 256, and 512, respectively.
- **Pooling Layers:** After the convolutional layers, two max pooling layers are applied to reduce the dimensionality and computational load for the first two convolutional layers.
- **Attention Mechanism:** The output from the convolutional layers is split into two parts: one part generates attention weights, and the other part, when multiplied by these weights, focuses on the relevant spectral features.
- **Linear Layer for Dimensionality Reduction:** Instead of using an MLP for compression, a single linear layer maps the attended features directly to the 128-dimensional latent space. This linear layer transforms the 256-dimensional attended feature vector into the 128-dimensional shared space used in the contrastive learning framework.
- **Softmax for Attention Weights:** A softmax function is applied to the attention weights to normalize them, ensuring they sum to one.
- **Output Latent Space:** The final output is a compressed latent representation of the spectrum, which can be used in the contrastive learning framework to align with image embeddings.

This architecture allows the spectrum encoder to efficiently learn meaningful representations of the spectra with fewer parameters and a simpler transformation process, facilitating alignment with image embeddings in the shared latent space. The use of convolutional layers and an attention mechanism, followed by a linear layer for compression, ensures the encoder

captures essential spectral features and reduces them to a compact latent representation suitable for contrastive learning.

3.2.3 Pretrained Image Embedder

Self-Supervised Learning in Astronomical Images

The identification of rare astronomical objects has significantly advanced due to digital sky surveys and deep learning technologies. The sheer volume of data from these surveys has made traditional methods of manual image inspection impractical. While crowd-sourced classification campaigns like Galaxy Zoo[10] have inspected many galaxies, they cover only a small fraction of the data generated by modern surveys. For instance, the DESI[4] Legacy Imaging Surveys include around one billion galaxies, far exceeding manual or crowd-sourced capabilities.

Deep convolutional neural networks (CNNs) have shown remarkable promise in automating the classification of astronomical images, achieving high accuracy with sufficient human-labeled data. However, these models rely heavily on labeled data, which tends to be biased towards more prominent galaxies and often neglects rare objects. The scarcity of labeled data and the massive scale of modern sky surveys necessitate scalable, label-efficient approaches.

Self-supervised learning has emerged as a powerful alternative, enabling the extraction of meaningful features from unlabeled data. This method involves solving contrived tasks that require a high-level understanding of the input data, producing robust representations that can be fine-tuned for specific tasks. Recent studies have demonstrated the effectiveness of self-supervised models in astronomy. For instance, it was shown that self-supervised pretraining on large sets of unlabeled galaxy images from the Sloan Digital Sky Survey (SDSS)[1] significantly improved performance on various tasks, especially when labeled data was scarce. This approach also provided a useful similarity metric for identifying additional examples of rare objects.

In the work by Stein et al[16]., self-supervised learning was applied to the DESI Legacy Survey's Data Release 9 to search for strong gravitational lenses. By leveraging the strengths of self-supervised models, they discovered 1,192 new strong gravitational lens candidates. This highlights the potential of self-supervised learning to advance the search for rare astronomical objects, providing a scalable solution for handling the vast data volumes of modern sky surveys. They also made their similarity search tool publicly available to aid in the ongoing discovery of gravitational lenses, facilitating broader engagement and exploration within the astronomical community.

Contrastive Self-Supervised Pretraining Framework of Stein et al.

Modern self-supervised learning techniques have demonstrated the ability to extract highly informative representations from standard machine learning datasets without labeled information. The performance of simple linear classifiers trained on these representations can rival that of fully supervised CNN models. These datasets typically contain ground-truth labels for every image, facilitating robust measurements of representation quality, as better representations lead to improved performance on downstream supervised tasks using these labels. However, for many scientific datasets, particularly sky surveys, ground-truth labels are sparse compared to the number of images, and the labels often contain significant noise or bias. This scarcity makes robust downstream supervised evaluations challenging, complicating hyperparameter tuning and the selection of data augmentations for the self-supervised learning (SSL) model.

To minimize hyperparameter tuning, Stein et al. closely followed previous work in designing the architecture and training procedure for their self-supervised learning model. The Illustration of self-supervised training of Stein's model is in figure 3.2. The backbone of the model is a CNN encoder that processes an image x and produces a lower-dimensional representation z. The encoder learns to create meaningful representations by associating augmented views of the same image as similar and views of different images as dissimilar through a contrastive loss function. Representations of different images are maintained in a queue during training, increasing the number of contrasting examples available to the model at each training step beyond those in each minibatch. Detailed descriptions of this approach can be found in prior works. Stein et al. used the same ResNet-50[8] network and training hyperparameters as previous studies, but increased the queue length to $K = 262,144$ to accommodate their larger training set. Their encoder architecture, a standard ResNet-50, produces a 2048-dimensional representation vector.

Successful self-supervised learning pretraining requires producing differing views of each image through carefully crafted augmentations that reflect symmetries, uncertainties, or noise in the dataset. These augmentations should perturb images realistically, making it challenging for the model to associate augmented pairs of the same image while preserving essential features. Stein et al. applied the following augmentations in succession during pretraining:

- **Galactic extinction:** They deredden the image based on its tabulated SFD $E(B-V)$ value, simulating a view without foreground dust. They then randomly sampled a new $E(B-V)$ value from a lognormal distribution fit to the dataset and artificially reddened the image.

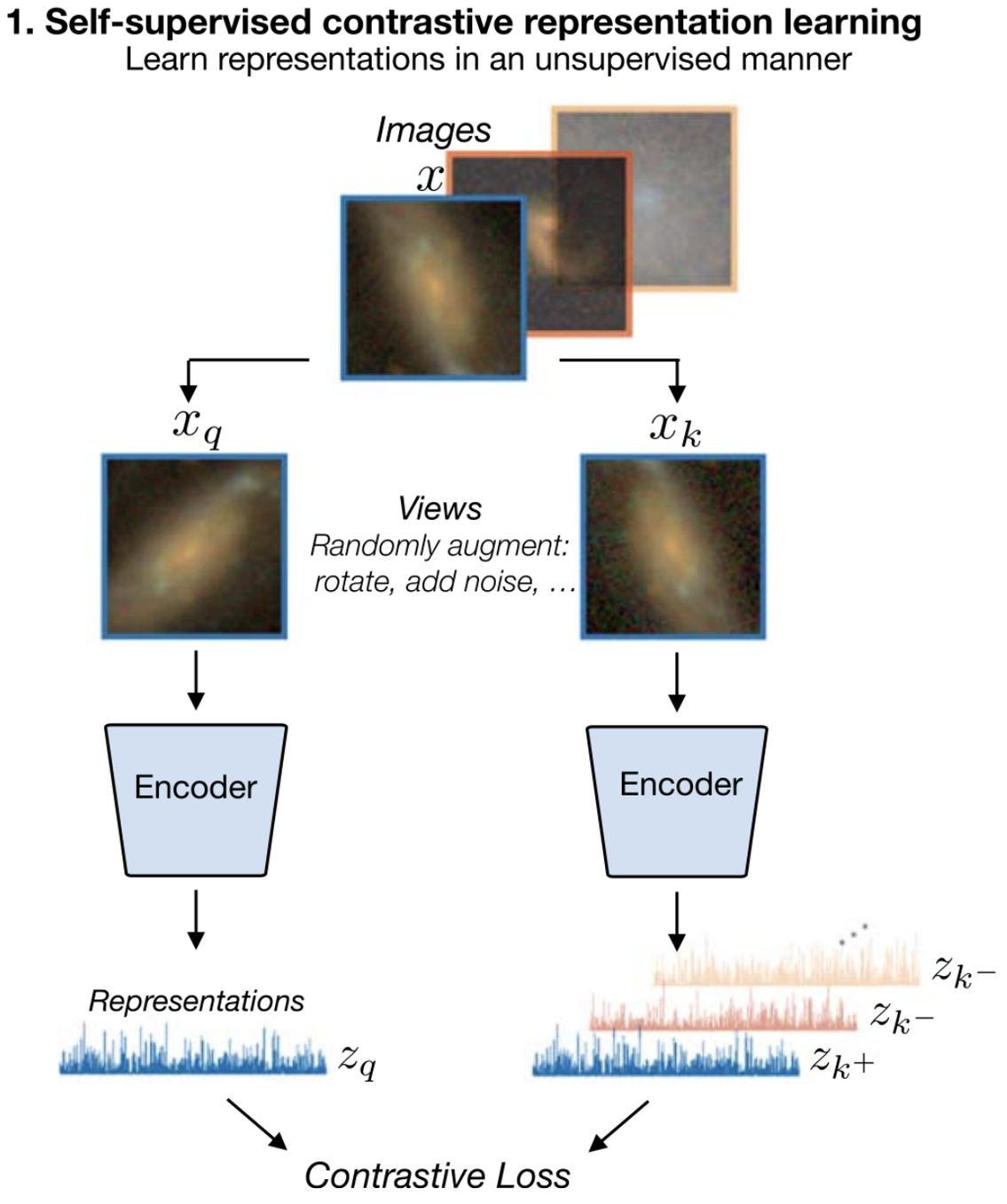

Fig. 3.2 Schematic of the contrastive self-supervised framework of galaxies images.

- **Rotation/Orientation:** They randomly flipped the image across each axis with 50% probability and rotated it by a random angle sampled from $U(0, 2\pi)$.
- **Size Scaling:** They resized the image to 90-110% of its original size to simulate different galaxy distances, rescaling with bilinear interpolation without accounting for changes in redshift or resolution.
- **Point Spread Function (PSF) blur:** They applied additional Gaussian blur to each channel, with the blurring parameterized by lognormal fits to the PSF distribution found in the data.
- **Jitter and crop:** They translated the image center by i, j pixels, where $i, j \sim U(-7, 7)$, before cropping out the central 96×96 pixels.
- **Gaussian noise:** They added Gaussian noise, sampling noise levels from lognormal distributions tuned for each filter channel. The noise in each channel is uncorrelated since images are taken at different times and/or with different telescopes.

The relative importance of these augmentations for producing good representations depends on both the dataset and the implementation of each augmentation. Applying these augmentations to samples x, Stein et al. obtained a pair of views that were denoted "positive" (x_q, x_k^+) when the two came from different transformations of the same image and "negative" (x_q, x_k^-) otherwise. For each of the views, an encoder network extracted a 2048-dimensional representation z = encoder(x) and was trained to make positive pairs have similar representations while making negative pairs have dissimilar representations via the InfoNCE contrastive loss function:

$$L_{(q,k^+,k^-)} = -\log\left(\frac{\exp(\text{sim}(z^q, z_{k^+}))}{\exp(\text{sim}(z^q, z_{k^+})) + \sum_k \exp(\text{sim}(z^q, z_{k^-}))}\right) \tag{1}$$

where $\text{sim}(a, b) = \frac{a \cdot b}{\tau \|a\| \|b\|}$ is the cosine similarity measure between vectors a and b, normalized by a tunable "temperature" hyperparameter τ. This loss (InfoNCE) is minimized when positive pairs have high similarity while negative pairs have low similarity. They closely followed previous work in their self-supervised learning setup.

Self-Supervised Framework

Encoder: They followed prior studies in using a ResNet-50 architecture as their encoder. Specifically, they used the implementation of the TorchVision library, which is part of the PyTorch project. The standard ResNet-50, however, is designed to work on wider images

than their 64×64 ones. To maintain reasonably wide representations by the end of the 50 layers, they changed the first `Conv2d` layer to have a stride of 1 instead of the default stride of 2, and they also removed the first `MaxPool2d` layer. This gives them 4× wider activations throughout the network than what they would get with the defaults of ResNet-50. The output of the `AdaptiveAvgPool2d` layer of the network is their representation, a 2048-dimensional vector.

Following previous studies, they did not use the representation z directly in the loss in Equation (1); instead, they used a two-layer MLP projection head that maps the representations to a space where the contrastive loss is applied. This has been shown to improve the learned representations. The output of the projection head is a 128-dimensional vector. The head is discarded after the self-supervised training process is completed.

Momentum Encoder: In contrastive learning setups, to make the task of identifying positive examples nontrivial, it is crucial to have a large set of negative examples. For this, they used the momentum encoder [7] idea; They maintain a queue of size 62k representations (nearly 5% of the training dataset size) that is continuously being updated during the training process. The representations in the queue are encoded using a momentum encoder, a second encoder with the same architecture as the main encoder but with parameters that are a moving average of the main encoder parameters. This is done by updating the momentum encoder parameters, θ, as a moving average of the main encoder parameters, θ:

$$\theta' \leftarrow m\theta' + (1-m)\theta \tag{2}$$

where m is a momentum parameter that they set to 0.999 as in previous work. This prevents the representations in the queue from changing too quickly during training.

Image Encoder Architecture

To minimize fine-tuning parameters, I used the pretrained galaxy image encoder from previous work [6] [16]. This model, based on MoCo V2 [7] with a ResNet-50 [8] backbone, was pretrained in a self-supervised regime on 3.5 million galaxies from the DESI Legacy Survey. The model has 28 million parameters, with 4.5 million trainable parameters during fine-tuning. The image encoder architecture is based on a ResNet-50 backbone, modified for feature extraction:

Initial Layers:

- **Convolution and Pooling:** 7×7 convolutional layer, batch normalization, ReLU activation, followed by a max pooling layer.

Residual Blocks:

- **Layer 1:** Three bottleneck blocks with 1×1 convolutions, batch normalization, ReLU activation, and skip connections.
- **Layer 2:** Similar to Layer 1, but with larger output channels and downsampling in the first block.
- **Layer 3 and Layer 4:** Continue the pattern of bottleneck blocks, increasing channels and downsampling as necessary.

Global Pooling and Fully Connected Layers:

- **Global Average Pooling:** Reduces spatial dimensions to 1×1 while maintaining depth.
- **Fully Connected Layers:** Two linear layers reduce the feature vector to a compact 128-dimensional representation.

The image encoder architecture leverages a ResNet-50 backbone to extract hierarchical features from input images. Using bottleneck blocks, global pooling, and fully connected layers, the architecture balances model complexity, computational efficiency, and feature representation quality, making it suitable for various downstream tasks like image classification, retrieval, or clustering in astronomical applications.

3.2.4 Contrastive Training: NewAstroCLIP Architecture

In this study, I introduce `newAstroCLIP`, an innovative deep learning architecture tailored for analyzing astronomical data using both image and spectrum inputs. The model comprises two core components: an `image_encoder` and a `spectrum_encoder`. To optimize the learning process, I freeze all layers of the `image_encoder` except for the final fully connected layer, enabling targeted fine-tuning without compromising the pre-trained features in the earlier layers. Meanwhile, the `spectrum_encoder` remains entirely unfrozen to preserve the integrity of its fine-tuned parameters.

The `image_encoder` utilizes a MoCoV2 model with a ResNet backbone, as described in Section 3.2.3. The `spectrum_encoder` employs two different architectures, as discussed in Section 3.2.2: one includes an MLP, while the other uses a linear layer to project the weighted feature into a shared 128-dimensional space with the image.

Both pre-trained models feed into the unified `newAstroCLIP` model. Training utilizes the InfoNCE objective (Equation 1), where embedded representations are considered positive pairs if they pertain to the same galaxy and negative examples if they pertain to different galaxies. I set the queue length to $K = 512$ image-spectrum pairs and apply basic data

augmentation with random vertical and horizontal flips, as well as random rotations on the images. The model is trained for 50 epochs, which takes approximately 18 hours on a single A100 GPU.

A unique feature of my architecture is the implementation of a fixed logit scale, set to the logarithm of 15.5, integrated into the custom `CLIPLoss` function. This design choice facilitates efficient contrastive learning by aligning the embeddings generated from both the image and spectrum encoders.

The forward pass of the model is designed for flexibility, allowing for the embedding of either images or spectra based on the input data. During training and validation, I calculate losses with and without logit scaling, enabling comprehensive monitoring and optimization of the model's performance. These losses are accumulated and logged at the end of each epoch, providing clear metrics for model evaluation.

I employ the AdamW optimizer combined with a cosine annealing learning rate scheduler to ensure efficient learning and adaptation throughout the training process. The use of PyTorch Lightning further enhances the workflow, providing a streamlined framework for model training and management. Overall, `newAstroCLIP` represents a robust and versatile tool for extracting meaningful insights from complex astronomical datasets, demonstrating significant potential in advancing the field of astronomical data analysis.

Chapter 4

Experiments

4.1 Dataset

4.1.1 Dataset Description and Preprocessing

The dataset utilized in this study is derived from the DESI Legacy Survey 3 Data Release 9, as documented by Dey et al. (2019) [4] and prepared by Lanusse et al [12]. Initially, the dataset comprises 41 million images captured in the g, r, z bands, each with a resolution of 152×152 pixels. To focus on relevant regions and ensure consistency, these images are center-cropped to 96×96 pixels. Furthermore, this imaging data is cross-matched with galaxy spectra from the DESI Early Data Release (Collaboration et al. [3]), resulting in a refined subset of 197,976 image-spectrum pairs.

To enhance the dataset's utility for extracting physical properties, these pairs are further cross-matched with the PRObabilistic Value-Added Bright Galaxy Survey (PROVABGS) Catalog (Hahn et al., 2023) [**?**]), which provides additional attributes such as redshift. The final dataset is organized as a dictionary containing keys for 'image', 'spectrum', 'redshift', and 'targetid', ensuring comprehensive cross-referencing and alignment of all data points.

For training and evaluation purposes, the dataset is split into a training set of 159,377 image-spectrum pairs and a test set of 38,599 pairs. Data loading is optimized using PyTorch's `DataLoader`, configured with a batch size of 512 and 10 workers to ensure efficient data handling. These configurations are in strict adherence to the guidelines established by Lanusse et al. (2023) [12].

4.1.2 Data Augmentation

To improve model robustness and prevent overfitting, data augmentation techniques are applied to the images during the training process. The augmentation includes random vertical and horizontal flips to introduce variability and enhance the model's generalization capabilities. Additionally, a center crop is employed to maintain a consistent image size of 96×96 pixels. This augmentation strategy aims to preserve the dataset's integrity while introducing necessary variability, thereby supporting the development of a more resilient and generalized model.

4.2 Training Process Analysis

4.2.1 Spectrum Encoder with MLP

Fig. 4.1 Schematic of the contrastive self-supervised framework of galaxies images.

Figure 4.1 illustrates the training and validation loss trends for the spectrum encoder with MLP.

- **Training Loss Trend:** The training loss consistently decreases over the epochs, indicating that the model is effectively learning from the training data.

- **Validation Loss Trend:** However, the validation loss shows significant oscillations without a clear downward trend.

4.2.2 Spectrum Encoder without MLP

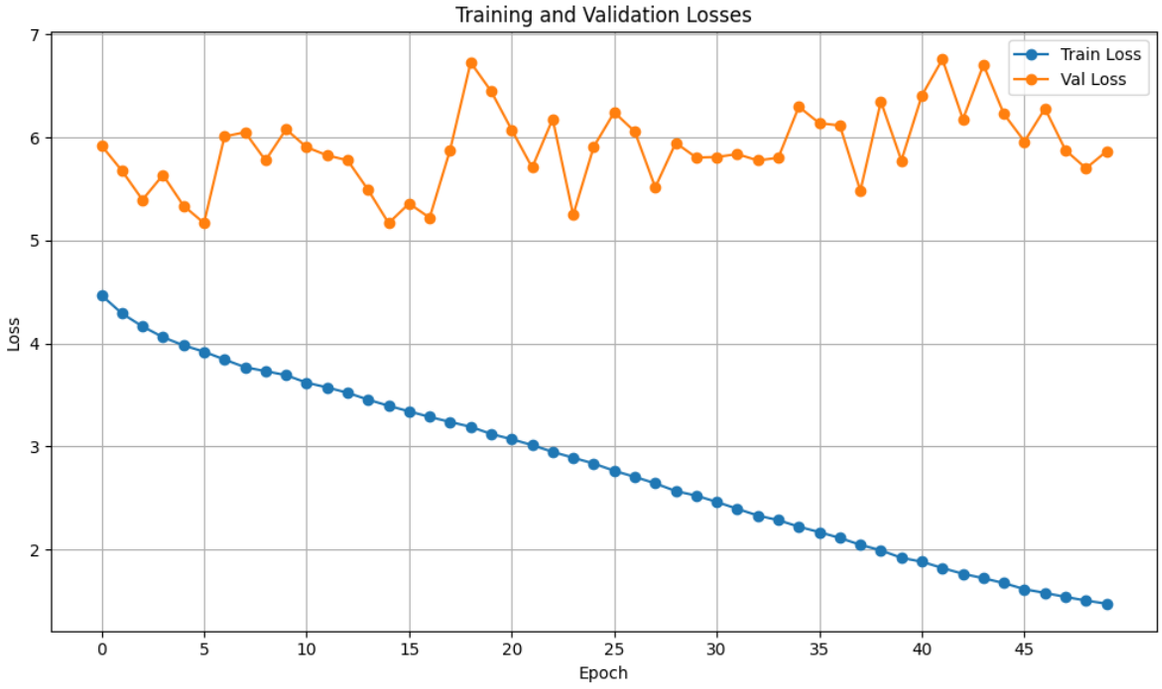

Fig. 4.2 Schematic of the contrastive self-supervised framework of galaxies images.

Figure 4.2 shows the training and validation loss for the spectrum encoder without MLP.

- **Training Loss Trend:** Similar to the MLP variant, the training loss decreases steadily throughout the epochs.
- **Validation Loss Trend:** The validation loss also exhibits oscillations and does not show a consistent decrease.

4.2.3 Analysis of Validation Loss Behavior

From the observed figures (Figure 4.1 and Figure 4.2), several factors could contribute to the oscillatory behavior and lack of significant validation loss reduction:

1. **Model Complexity and Overfitting:** Despite the training loss decreasing, the oscillations in validation loss suggest that the model might be overfitting to the training data.

Overfitting occurs when the model learns specific details and noise in the training data that do not generalize well to unseen validation data. This can be exacerbated by the complexity of the model architecture or insufficient regularization.

2. **Data Quality and Distribution:** Disparities in data quality or distribution between the training and validation sets can also contribute to validation loss oscillations. Ensuring that the validation set is representative of the data distribution encountered during deployment is crucial for accurate model evaluation.

4.3 4.3 Downstream Tasks and Results Analysis

4.3.1 4.2.1 Similarity Search

To visualize the effectiveness of my embedding scheme in aligning representations of galaxies, I perform similarity searches using a cosine similarity metric. Specifically, I query galaxies and find their nearest neighbors in the embedding space. The cross-modal similarity $S_C(z_i^{sp}, z_j^{im})$ between a query spectrum $z_i^{sp} = g_\theta(x_i^{sp})$ and an image $z_j^{im} = g_\theta(x_j^{im})$ is computed as:

$$S_C(z_i^{sp}, z_j^{im}) = \frac{z_i^{sp} \cdot z_j^{im}}{\|z_i^{sp}\| \|z_j^{im}\|}$$

This similarity search is performed for both in-modality similarity, where I determine neighbors based on the similarity between embeddings from the same modality (e.g., image-image or spectrum-spectrum), and cross-modality similarity, where the similarity between embeddings from different modalities (e.g., image-spectrum or spectrum-image) is considered.

As illustrated previously, I have two architectures for the spectrum encoder:

- **A. Spectrum encoder with MLP**
- **B. Spectrum encoder without MLP**

The similarity search results for **A. spectrum encoder with MLP** are shown in Figures 4.3 through 4.6:

- Figure 4.3: Spectrum-Spectrum similarity search.
- Figure 4.4: Image-Image similarity search.
- Figure 4.5: Image-Spectrum similarity search.

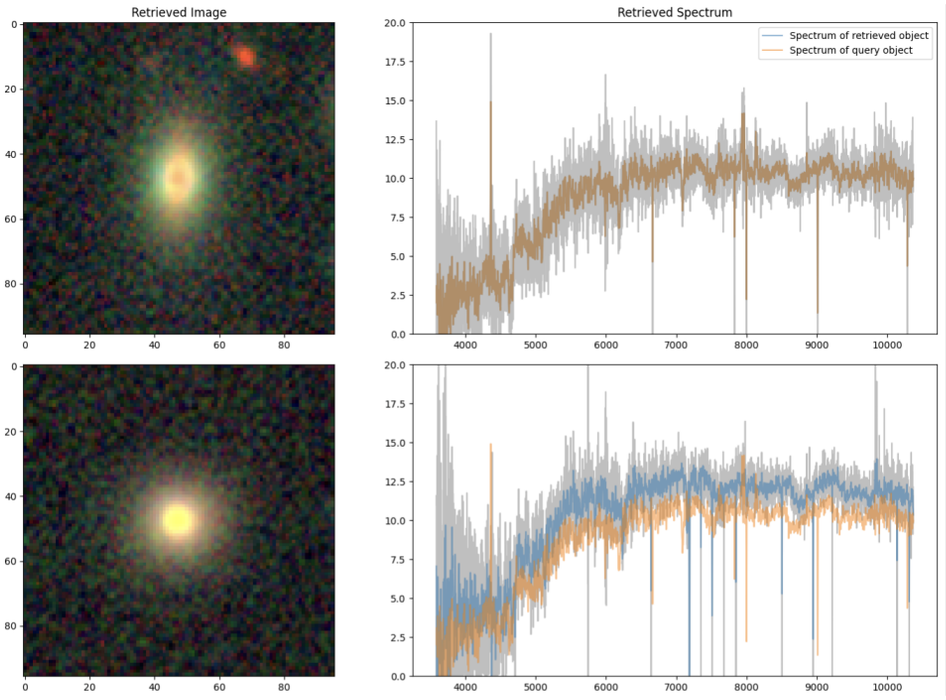

Fig. 4.3 Spectrum-Spectrum similarity search for the spectrum encoder with MLP

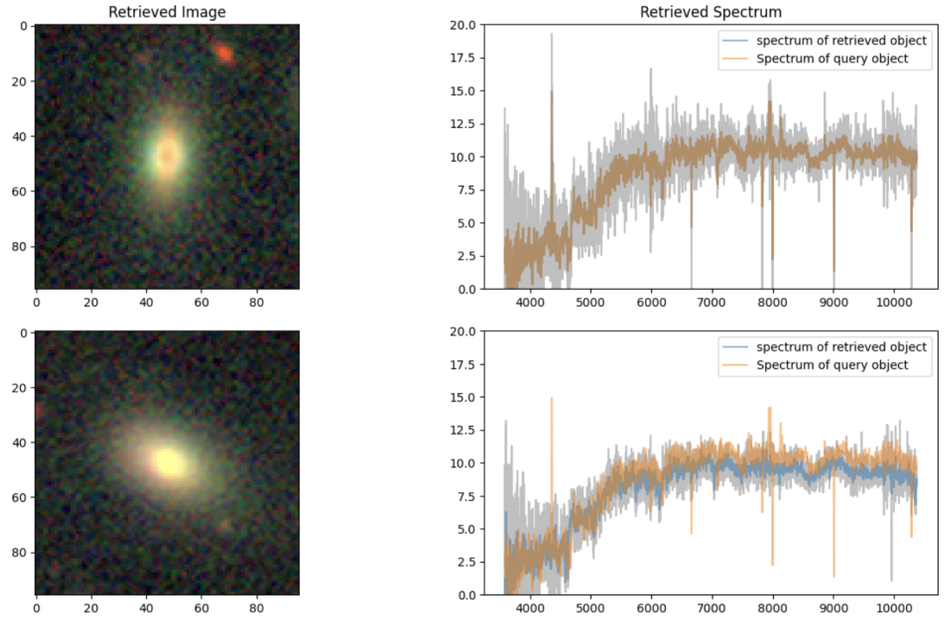

Fig. 4.4 Image-Image similarity search for the spectrum encoder with MLP

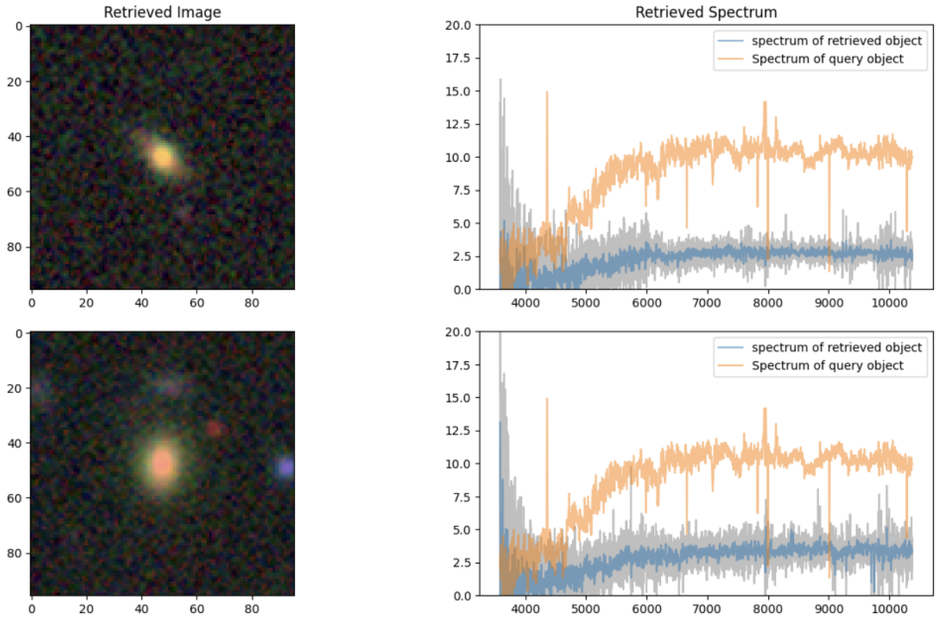

Fig. 4.5 Image-Spectrum similarity search for the spectrum encoder with MLP

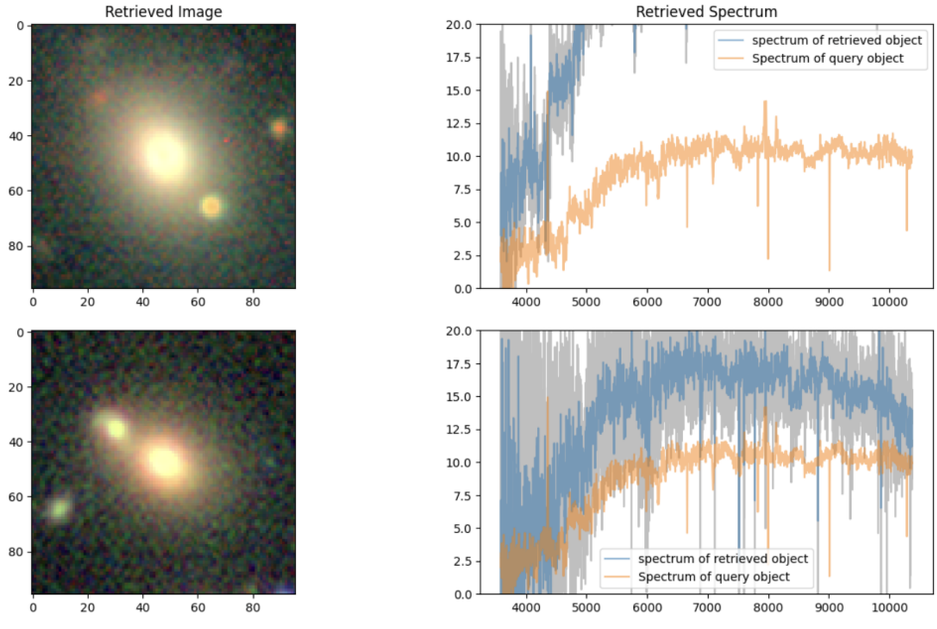

Fig. 4.6 Spectrum-Image similarity search for the spectrum encoder with MLP

- Figure 4.6: Spectrum-Image similarity search.

The similarity search results for **B. spectrum encoder without MLP** are shown in Figures 4.7 through 4.10:

- Figure 4.7: Spectrum-Spectrum similarity search.
- Figure 4.8: Image-Image similarity search.
- Figure 4.9: Image-Spectrum similarity search.
- Figure 4.10: Spectrum-Image similarity search.

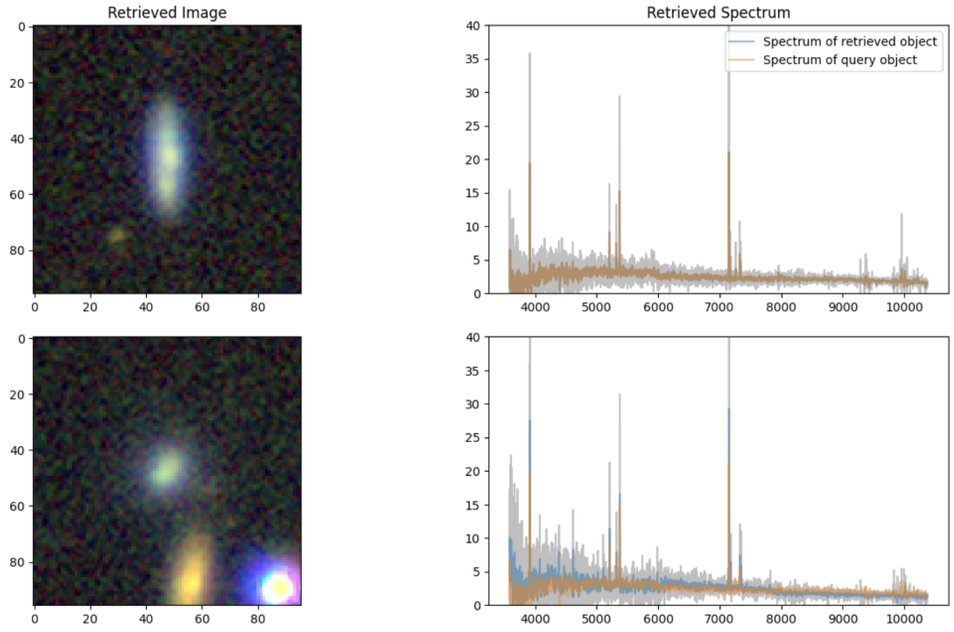

Fig. 4.7 Spectrum-Spectrum similarity search for the spectrum encoder without MLP

These figures present the results for all four possible embedding pairs for a randomly selected query example. Ultimately, these examples demonstrate that the model can consistently represent the same types of objects similarly, regardless of the original modality in which the object is represented. Notably, by design, the closest match in an in-modality similarity search is the object itself.

4.3.2 4.2.2 Zero-shot Regression of Redshift

To make more quantitative statements about the performance of newAstroCLIP pretraining, I assess the model's ability to perform zero-shot prediction on downstream tasks using the

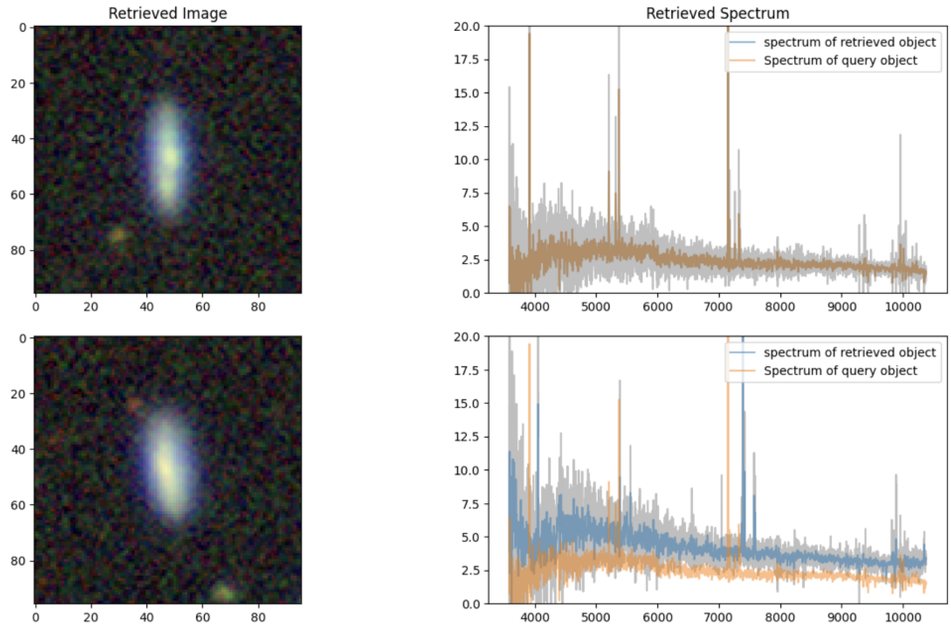

Fig. 4.8 Image-Image similarity search for the spectrum encoder without MLP

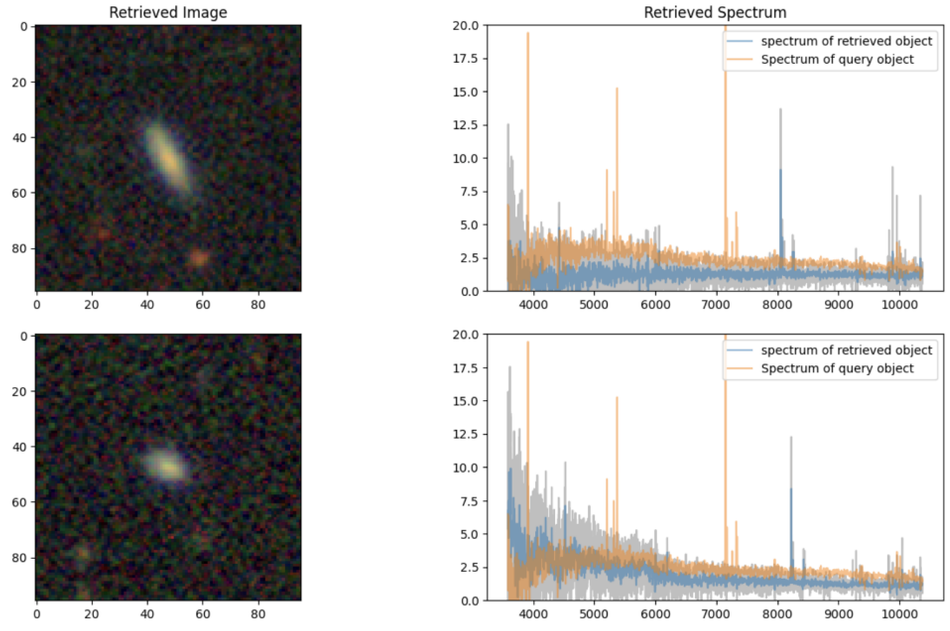

Fig. 4.9 Image-Spectrum similarity search for the spectrum encoder without MLP

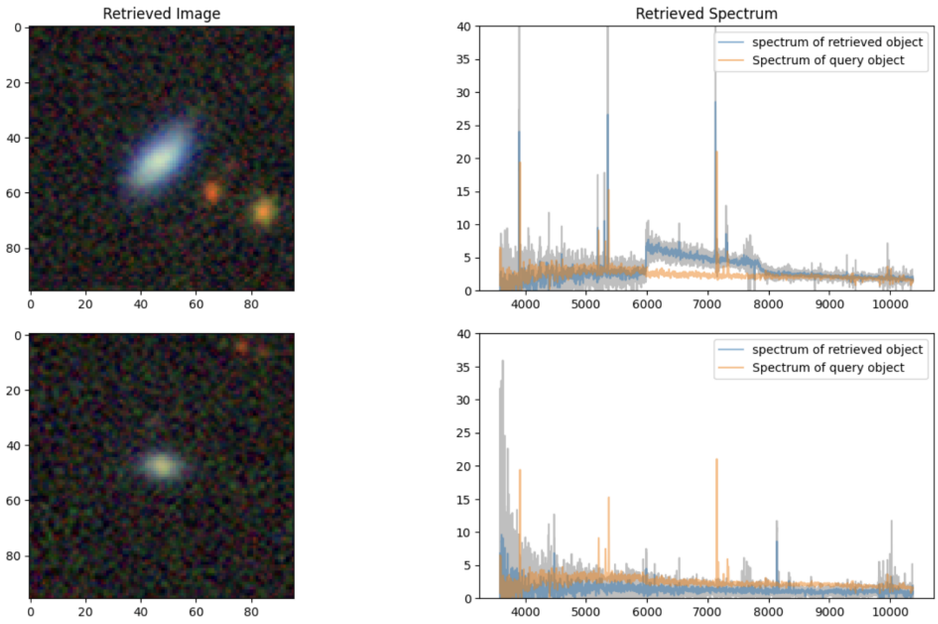

Fig. 4.10 Spectrum-Image similarity search for the spectrum encoder without MLP

embedded galaxy samples. Specifically, I employ simple k-Nearest Neighbour (k-NN) regression on the embedded images and spectra to infer the redshift of galaxies. This regression is conducted on the autocorrelated newAstroCLIP image and spectrum embeddings.

Figures 4.11,12,13,14 illustrate a comparison of the performance of k-NN regression from the newAstroCLIP embeddings for both in-modality. Figure 4.11 and 4.12 show results for the spectrum encoder with MLP, while Figure 4.13 and 4.14 presents results for the spectrum encoder without MLP.

Several important observations can be made:

1. **Similarity in Embedded Space:** Neighbors in the embedded space share similar physical properties, as evidenced by the ability of the k-NN regressor to make accurate predictions. This indicates that the model successfully organizes galaxy samples according to high-level, physically meaningful features.

2. **Performance of Spectrum Encoder with and without MLP:** Interestingly, the spectrum encoder without MLP performs slightly better than the one with MLP. This improvement can be attributed to the use of a simple linear layer, which directly projects the weighted attention features of the spectrum into the shared space. The linear projection preserves the information provided by the convolution and attention blocks of the spectrum encoder, unlike MLP which compresses these features. Consequently, the spectrum encoder without MLP retains more detailed information, leading to better performance.

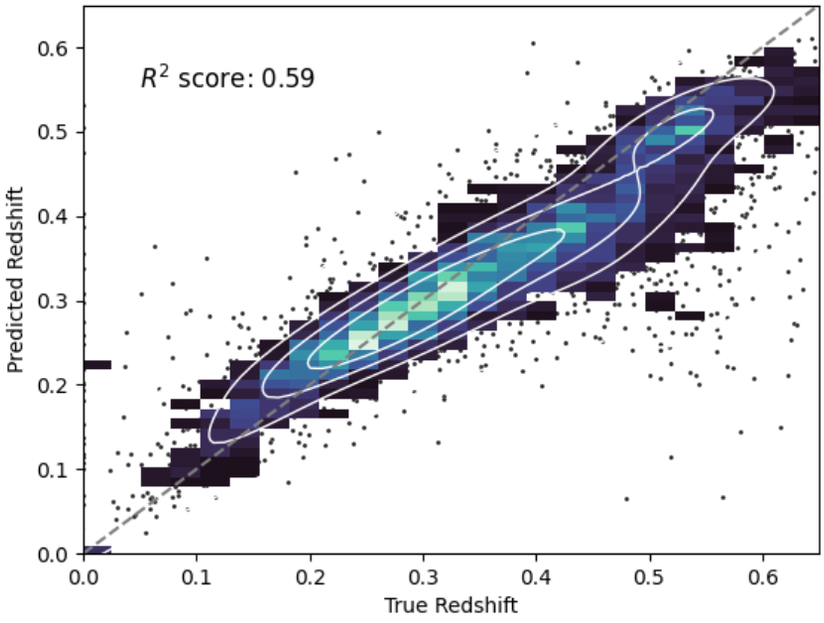

Fig. 4.11 Redshift regression by k-NN using image embeddings for the spectrum encoder with MLP

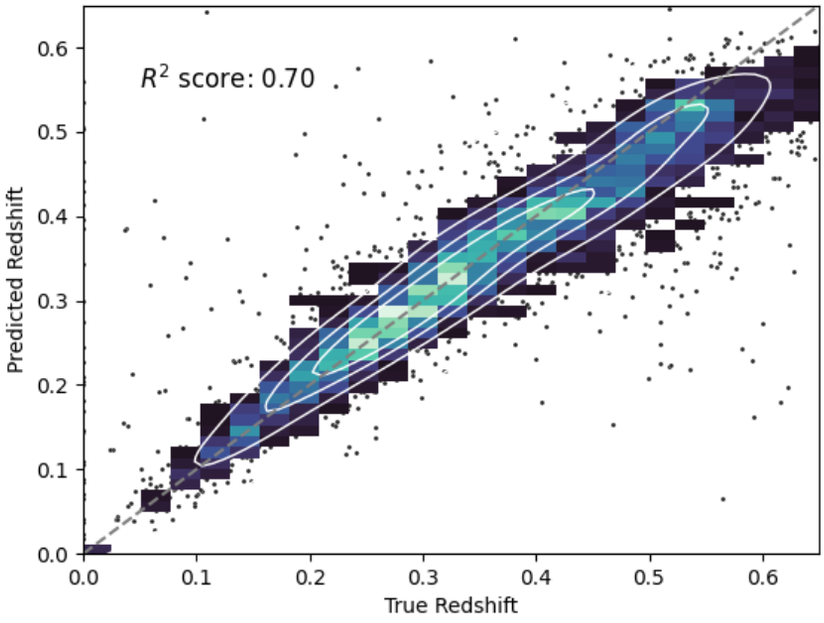

Fig. 4.12 Redshift regression by k-NN using spectral embeddings for the spectrum encoder with MLP

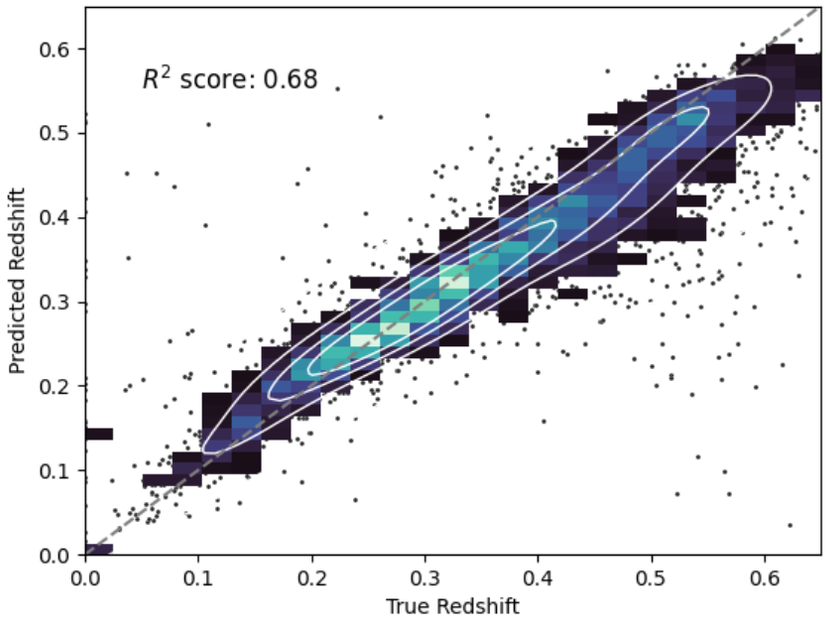

Fig. 4.13 Redshift regression by k-NN using image embeddings for the spectrum encoder without MLP

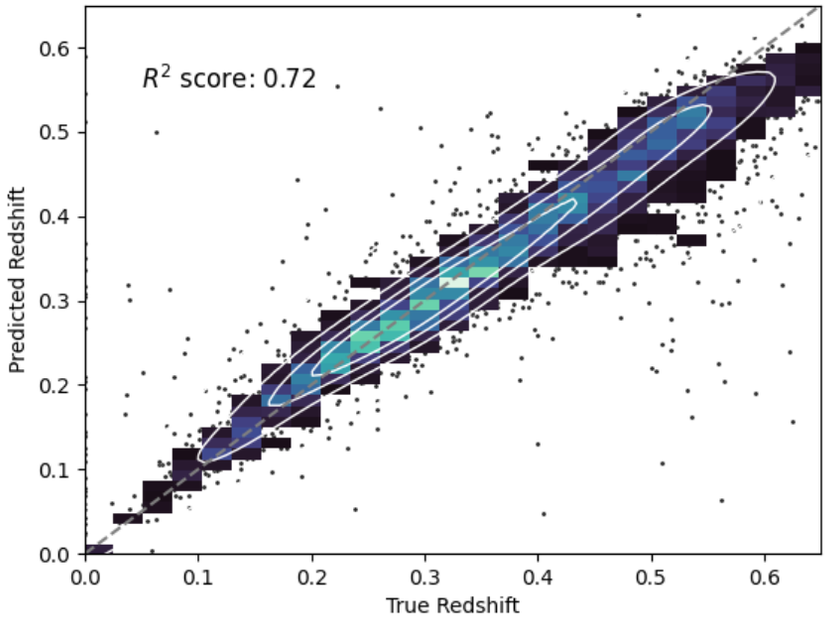

Fig. 4.14 Redshift regression by k-NN using spectral embeddings for the spectrum encoder without MLP

These findings highlight the effectiveness of newAstroCLIP embeddings for zero-shot redshift regression and provide insights into the structuring of the embedding space by the model.

Chapter 5

Conclusion

5.1 Conclusion

My findings underscore the potential of cross-modal contrastive pre-training in establishing high-quality foundational models for astronomical data. These models prove effective in various downstream tasks without requiring fine-tuning. This property is pivotal as it enables the community to develop sophisticated compositional models that leverage pre-trained frozen embeddings, akin to the versatility seen with frozen CLIP embeddings across diverse applications.

Furthermore, my results provide compelling evidence that even when diverse modalities do not perfectly inform each other, the contrastive learning task facilitates the discovery of relevant physical patterns within each modality's embedding. For instance, my spectral embeddings demonstrate an emergent capability to encode information about redshift that surpasses the information contained in the images alone. This discovery paves the way for future research to explore the creation of informative embeddings from an extensive range of data modalities, even in scenarios where direct inter-modality connections are not robustly established.

These insights reinforce the promise of cross-modal contrastive pre-training as a transformative approach in astronomical data analysis, offering avenues for deeper understanding and broader application across the field

References

[1] Almeida, A., Anderson, S. F., Argudo-Fernández, M., Badenes, C., Barger, K., Barrera-Ballesteros, J. K., Bender, C. F., Benitez, E., Besser, F., Bird, J. C., Bizyaev, D., Blanton, M. R., Bochanski, J., Bovy, J., Brandt, W. N., Brownstein, J. R., Buchner, J., Bulbul, E., Burchett, J. N., Díaz, M. C., Carlberg, J. K., Casey, A. R., Chandra, V., Cherinka, B., Chiappini, C., Coker, A. A., Comparat, J., Conroy, C., Contardo, G., Cortes, A., Covey, K., Crane, J. D., Cunha, K., Dabbieri, C., Davidson, J. W., Davis, M. C., de Andrade Queiroz, A. B., De Lee, N., Méndez Delgado, J. E., Demasi, S., Di Mille, F., Donor, J., Dow, P., Dwelly, T., Eracleous, M., Eriksen, J., Fan, X., Farr, E., Frederick, S., Fries, L., Frinchaboy, P., Gänsicke, B. T., Ge, J., González Ávila, C., Grabowski, K., Grier, C., Guiglion, G., Gupta, P., Hall, P., Hawkins, K., Hayes, C. R., Hermes, J. J., Hernández-García, L., Hogg, D. W., Holtzman, J. A., Ibarra-Medel, H. J., Ji, A., Jofre, P., Johnson, J. A., Jones, A. M., Kinemuchi, K., Kluge, M., Koekemoer, A., Kollmeier, J. A., Kounkel, M., Krishnarao, D., Krumpe, M., Lacerna, I., Lago, P. J. A., Laporte, C., Liu, C., Liu, A., Liu, X., Lopes, A. R., Macktoobian, M., Majewski, S. R., Malanushenko, V., Maoz, D., Masseron, T., Masters, K. L., Matijevic, G., McBride, A., Medan, I., Merloni, A., Morrison, S., Myers, N., Mészáros, S., Negrete, C. A., Nidever, D. L., Nitschelm, C., Oravetz, D., Oravetz, A., Pan, K., Peng, Y., Pinsonneault, M. H., Pogge, R., Qiu, D., Ramirez, S. V., Rix, H.-W., Rosso, D. F., Runnoe, J., Salvato, M., Sanchez, S. F., Santana, F. A., Saydjari, A., Sayres, C., Schlaufman, K. C., Schneider, D. P., Schwope, A., Serna, J., Shen, Y., Sobeck, J., Song, Y.-Y., Souto, D., Spoo, T., Stassun, K. G., Steinmetz, M., Straumit, I., Stringfellow, G., Sánchez-Gallego, J., Taghizadeh-Popp, M., Tayar, J., Thakar, A., Tissera, P. B., Tkachenko, A., Toledo, H. H., Trakhtenbrot, B., Fernández-Trincado, J. G., Troup, N., Trump, J. R., Tuttle, S., Ulloa, N., Vazquez-Mata, J. A., Alfaro, P. V., Villanova, S., Wachter, S., Weijmans, A.-M., Wheeler, A., Wilson, J., Wojno, L., Wolf, J., Xue, X.-X., Ybarra, J. E., Zari, E., and Zasowski, G. (2023). The eighteenth data release of the sloan digital sky surveys: Targeting and first spectra from sdss-v. *The Astrophysical Journal Supplement Series*, 267(2):44.

[2] Chen, X., Fan, H., Girshick, R., and He, K. (2020). Improved baselines with momentum contrastive learning.

[3] DESI Collaboration Et Al (2023). The early data release of the dark energy spectroscopic instrument.

[4] Dey, A., Schlegel, D. J., Lang, D., Blum, R., Burleigh, K., Fan, X., Findlay, J. R., Finkbeiner, D., Herrera, D., Juneau, S., Landriau, M., Levi, M., McGreer, I., Meisner, A., Myers, A. D., Moustakas, J., Nugent, P., Patej, A., Schlafly, E. F., Walker, A. R., Valdes, F., Weaver, B. A., Yèche, C., Zou, H., Zhou, X., Abareshi, B., Abbott, T. M. C., Abolfathi, B., Aguilera, C., Alam, S., Allen, L., Alvarez, A., Annis, J., Ansarinejad, B., Aubert, M.,

Beechert, J., Bell, E. F., BenZvi, S. Y., Beutler, F., Bielby, R. M., Bolton, A. S., Briceño, C., Buckley-Geer, E. J., Butler, K., Calamida, A., Carlberg, R. G., Carter, P., Casas, R., Castander, F. J., Choi, Y., Comparat, J., Cukanovaite, E., Delubac, T., DeVries, K., Dey, S., Dhungana, G., Dickinson, M., Ding, Z., Donaldson, J. B., Duan, Y., Duckworth, C. J., Eftekharzadeh, S., Eisenstein, D. J., Etourneau, T., Fagrelius, P. A., Farihi, J., Fitzpatrick, M., Font-Ribera, A., Fulmer, L., Gänsicke, B. T., Gaztanaga, E., George, K., Gerdes, D. W., A Gontcho, S. G., Gorgoni, C., Green, G., Guy, J., Harmer, D., Hernandez, M., Honscheid, K., Huang, L. W., James, D. J., Jannuzi, B. T., Jiang, L., Joyce, R., Karcher, A., Karkar, S., Kehoe, R., Kneib, J.-P., Kueter-Young, A., Lan, T.-W., Lauer, T. R., Guillou, L. L., Van Suu, A. L., Lee, J. H., Lesser, M., Levasseur, L. P., Li, T. S., Mann, J. L., Marshall, R., Martínez-Vázquez, C. E., Martini, P., du Mas des Bourboux, H., McManus, S., Meier, T. G., Ménard, B., Metcalfe, N., Muñoz-Gutiérrez, A., Najita, J., Napier, K., Narayan, G., Newman, J. A., Nie, J., Nord, B., Norman, D. J., Olsen, K. A. G., Paat, A., Palanque-Delabrouille, N., Peng, X., Poppett, C. L., Poremba, M. R., Prakash, A., Rabinowitz, D., Raichoor, A., Rezaie, M., Robertson, A. N., Roe, N. A., Ross, A. J., Ross, N. P., Rudnick, G., Gaines, S., Saha, A., Sánchez, F. J., Savary, E., Schweiker, H., Scott, A., Seo, H.-J., Shan, H., Silva, D. R., Slepian, Z., Soto, C., Sprayberry, D., Staten, R., Stillman, C. M., Stupak, R. J., Summers, D. L., Tie, S. S., Tirado, H., Vargas-Magaña, M., Vivas, A. K., Wechsler, R. H., Williams, D., Yang, J., Yang, Q., Yapici, T., Zaritsky, D., Zenteno, A., Zhang, K., Zhang, T., Zhou, R., and Zhou, Z. (2019). Overview of the desi legacy imaging surveys. *The Astronomical Journal*, 157(5):168.

[5] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning.

[6] Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., and Mustafa, M. (2021). Self-supervised representation learning for astronomical images. *The Astrophysical Journal Letters*, 911(2):L33.

[7] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning.

[8] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

[9] Liang, Y., Melchior, P., Lu, S., Goulding, A., and Ward, C. (2023). Autoencoding galaxy spectra. ii. redshift invariance and outlier detection. *The Astronomical Journal*, 166(2):75.

[10] Masters, K. L. (2019). Twelve years of galaxy zoo. *Proceedings of the International Astronomical Union*, 14(S353):205–212.

[11] Melchior, P., Liang, Y., Hahn, C., and Goulding, A. (2023). Autoencoding galaxy spectra. i. architecture. *The Astronomical Journal*, 166(2):74.

[12] Parker, L., Lanusse, F., Golkar, S., Sarra, L., Cranmer, M., Bietti, A., Eickenberg, M., Krawezik, G., McCabe, M., Morel, R., Ohana, R., Pettee, M., Régaldo-Saint Blancard, B., Cho, K., and Ho, S. (2024). Astroclip: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011.

[13] Portillo, S. K. N., Parejko, J. K., Vergara, J. R., and Connolly, A. J. (2020). Dimensionality reduction of sdss spectra with variational autoencoders. *The Astronomical Journal*, 160(1):45.

[14] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

[15] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[16] Stein, G., Blaum, J., Harrington, P., Medan, T., and Lukić, Z. (2022). Mining for strong gravitational lenses with self-supervised learning. *The Astrophysical Journal*, 932(2):107.

[17] Teimoorinia, H., Archinuk, F., Woo, J., Shishehchi, S., and Bluck, A. F. L. (2022). Mapping the diversity of galaxy spectra with deep unsupervised machine learning. *The Astronomical Journal*, 163(2):71.

[18] van den Oord, A., Li, Y., and Vinyals, O. (2019). Representation learning with contrastive predictive coding.

[19] Walmsley, M., Slijepcevic, I. V., Bowles, M., and Scaife, A. M. M. (2022). Towards galaxy foundation models with hybrid contrastive learning.