AstroDINO: Self-Supervised Learning on Astronomical Images

Viraj Manwadkar
Department of Physics, Stanford University
virajvm@stanford.edu

Sagar Kapare
Stanford University
skapare@stanford.edu

1. Introduction

Self-supervised learning is a powerful framework in computer vision where models learn meaningful representations from images alone, without relying on any labeled data. The goal is to extract visual features that can generalize across numerous downstream tasks like image-level classification and pixel-level segmentation.

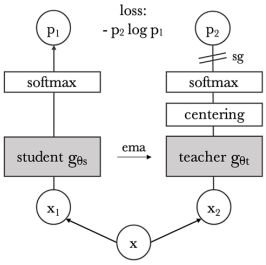
In astronomy, many problems can be framed as computer vision tasks: classifying galaxies, segmenting sources from the background, detecting anomalous images, finding sources of a specific type etc. While supervised approaches have made progress in some of these areas, they are often constrained by the need for extensive, high-quality labeled datasets, which can be significant bottleneck.

Recent work has shown the promise of self-supervised learning in astronomy, using large imaging datasets to learn general-purpose features [7]. These features can then be fine-tuned or adapted for numerous specific tasks like morphological classification, anomaly detection or similarity searches [8].

In this project, we explore the use of DINO (self-**Di**stillation with **No** labels), which is a self-supervised framework that trains vision transformers to learn semantic features without supervision [1, 5]. Furthermore, despite not being trained for segmentation, DINO's attention heads naturally emerge as semantic part detectors, yielding attention maps that can directly serve as unsupervised segmentation masks (see examples in [1, 5]).

This makes DINO especially compelling for astronomy where supervised labels are scarce and interpretability and generalization are important. By applying DINO to astronomical images, we aim to learn representations that not only capture the pixel-level structure of galaxies and sources, but also support a wide range of downstream tasks from anomaly detection to image classification, all within a unified, label-free framework.

In this specific project, we focus on understanding the usability of DINO models for astronomical datasets. The end goal is to have a DINO based model that can do semantic segmentation on astronomical imaging. In Section 4, we motivate and discuss this specific problem statement in more detail.



Algorithm 1 DINO PyTorch pseudocode w/o multi-crop.

```
# gs, gt: student and teacher networks
# C: center (K)
# tps, tpt: student and teacher temperatures
# l, m: network and center momentum rates
gt.params = gs.params
for x in loader: # load a minibatch x with n samples
    x1, x2 = augment(x), augment(x) # random views

    s1, s2 = gs(x1), gs(x2) # student output n-by-K
    t1, t2 = gt(x1), gt(x2) # teacher output n-by-K

    loss = H(t1, s2)/2 + H(t2, s1)/2
    loss.backward() # back-propagate

    # student, teacher and center updates
    update(gs) # SGD
    gt.params = l*gt.params + (1-l)*gs.params
    C = m*C + (1-m)*cat([t1, t2]).mean(dim=0)

def H(t, s):
    t = t.detach() # stop gradient
    s = softmax(s / tps, dim=1)
    t = softmax((t - C) / tpt, dim=1) # center + sharpen
    return - (t * log(s)).sum(dim=1).mean()
```

Figure 1. (Left) A schematic diagram of the DINO framework where x is the input image while x_1 and x_2 are the 2 augmented representations of this input x. (Right) The psuedo-code of the DINO framework (without including the cropping augmentations). Note that both these figures have been taken from [1].

2. The DINO framework

DINO [1] is a self-supervised learning framework that trains vision transformers (ViTs) using a self-distillation approach without requiring any labeled data. The core idea is to have two networks — a student and a teacher — which learn from each other. Unlike traditional supervised or contrastive learning setups, DINO does not rely on labeled pairs or negative/contrastive samples.

2.1. Vision Transformers

Vision Transformers (ViTs) divide an image into small patches and treat each patch as a token, analogous to words in a sentence. These patches are then processed using the transformer architecture, allowing the model to capture long-range dependencies and global context across the image. Unlike convolutional networks like ResNets, ViTs natively support patch-level attention, producing interpretable attention maps that can act as unsupervised segmentation masks, which is our end target.

2.2. Distillation: The Teacher-Student Framework

The left panel in Figure 1 shows a schematic overview of the DINO training process. The training process uses

the distillation framework, where the student network is the only network directly trained via gradient descent and the teacher network is constructed as an exponential moving average (EMA) of the student weights — often referred to as a momentum encoder. This dynamic ensures that the student is always trying to match a slowly changing target, leading to stable training and rich feature learning.

A key component of DINO is the multi-crop augmentation strategy, where multiple views of the same image are generated. Global crops (typically $\geq 50\%$ of the image size) provide context and help the network understand large-scale structures. Local crops (typically $< 50\%$) allow the model to focus on fine details. The teacher processes only global views, while the student sees both global and local views, encouraging consistency in feature representations across different scales and perspectives.

The training objective is to align the student's output distribution with that of the teacher using the cross-entropy loss. Unlike methods like MoCo, DINO does not use contrastive loss with negative pairs. Also note that unlike previous work in knowledge distillations that relies on a pretrained fixed teacher, the DINO teacher is dynamically built during training. This way, knowledge distillation is directly cast as a self-supervised objective instead of being used as a post-processing step.

DINO version 2 (DINOv2; [5]) is an extension and improved version of the DINO framework where it combines two types of training objectives. Just like in the original DINO, the knowledge distillation (KD) framework is still kept where it uses the global and local image crops and the student network learns to match the teacher's output on global views. Additionally, there is a Masked Image Modeling (MIM) part, where DINOv2 has the student network predict features of masked image patches, while the teacher sees the full (unmasked) image.

These two losses, KD and MIM, are weighted and combined in the final training objective. Additionally, DINOv2 adds a regularization term called KoLeo, which encourages the feature representations in each batch to spread out evenly (to avoid collapse and improve diversity). These changes help DINOv2 learn richer and more general features, which lead to better performance on a wide range of tasks like image classification, segmentation, depth estimation etc. [5].

3. Data

In this project, we use data from the DESI Legacy Imaging Surveys Data Release 10 (DR10) [2], which provides uniform, deep optical imaging in four filters (g, r, i, z) over ∼15,000 square degrees of the sky. The Legacy Surveys combine data from multiple telescopes and are processed with a consistent pipeline, yielding a large, homogeneous dataset. We access this dataset via the MultiModalUniverse dataset[1] [9] and separately through direct URL-based queries to retrieve image cutouts at specified sky coordinates. The MultiModalUniverse DR10 dataset contains 160×160 pixel cutouts. Additionally, the dataset includes morphological classifications from Galaxy Zoo[2] for $17,736$ galaxies within the DR10 footprint, allowing for a quantitative evaluation of how well the model learns semantically meaningful representations.

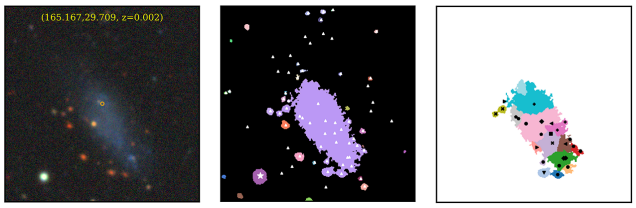


Figure 2. **Example of fragmentation in astronomical images**. The left panel shows the RGB cutout, where a diffuse blue galaxy overlaps with several compact red/orange background galaxies. The center panel displays a traditional segmentation map, which fails to separate the overlapping sources. The right panel shows the primary segment deblended; reconstructing the full extent of the main galaxy from this is non-trivial and requires additional cues from color, spatial context, and morphology—information often lost in binary masks.

4. The Problem Statement: Semantic Segmentation in Astronomical Imaging

A key challenge in astronomical image analysis is the grouping of pixels in physically meaningful ways — a task analogous to semantic segmentation in computer vision. In natural images, semantic segmentation involves separating objects like dogs, trees, or lakes into coherent regions despite overlaps. Similarly, in astronomy, galaxies often appear blended, overlapping in projection or exhibiting complex internal structure, making it non trivial to distinguish their boundaries and separating them.

Traditional source detection methods in astronomy are designed to identify statistically significant peaks in flux relative to the background noise. These peaks, once detected, are modeled as independent sources, often using parametric/non-parametric models [4]. When multiple peaks are spatially close and are fit as multiple sources, they can be flagged as blended sources, however it is non-trivial to turn this into a statement on whether the peaks correspond to distinct galaxies or to substructures within a single extended object.

This ambiguity becomes especially problematic for nearby, irregular, or interacting galaxies, where complex

[1] https://github.com/MultimodalUniverse/MultimodalUniverse/

[2] https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/

morphology and partial resolution of internal structure can cause over-segmentation: a single galaxy may be broken into multiple sources (see Figure 2). While source models can model and explain all the observed flux in an image, they are not designed to assign pixels to semantically meaningful units — e.g., determining that a set of peaks together constitute a single galaxy.

As a result, estimating global properties (e.g., total flux, size, or color) of such objects becomes non-trivial. These properties depend on correctly grouping pixels and sources, a process that involves integrating spatial, morphological, and multi-band color information across the image.

Emerging self-supervised learning methods, such as DINO, offer a promising avenue for addressing this challenge. These models can learn galaxy representations directly from seeing many images, without requiring labeled examples. By developing an internal model of what constitutes a galaxy, including its typical structure, extent, and color gradients, such approaches can help group pixels in a way that aligns with physically meaningful units. This capability could improve the identification of entire galaxies, including irregular and blended systems, and enable more accurate measurements of their global properties.

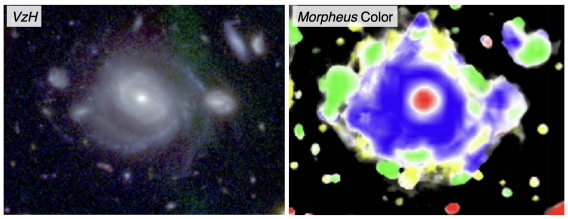


Figure 3. **Example segmentation from the Morpheus framework for morphological classification** [3]. This demonstrates a case of *semantic* image segmentation in astronomy, which goes beyond simply identifying regions of significant flux. Notably, the large spiral galaxy at the center is *fragmented* into multiple segments due to its internal morphological variation. While the Morpheus framework is performing as intended for its classification task, our goal is to develop a complementary and more general approach that instead recognizes and labels such a galaxy as a single cohesive object.

5. Results

5.1. DINO Attention Heads and Naive Segmentation Maps

In this section, we explore the attention heads produced by the DINO model and if they appear to be physically meaningful. We specifically use the pretrained DINO ViT-S/8 model.

To extract attention maps from the pretrained DINO ViT model, we reshape the attention weights. The attention weights have dimensions 6×784 where we have 6 attention heads with 784 attention weights. As the patch size is 8×8, we have a total of $(224/8)^2 = 784$ weights. We reshape each head into 2D maps corresponding to the spatial resolution of the image's patch grid. We computed the feature map size using the known patch size of the model and interpolated these maps back to the original image resolution using nearest-neighbor interpolation. To produce binary segmentation masks, we retained only the most important regions of each head's attention map by selecting pixels that cumulatively account for the top 40% of attention mass (corresponding to a 0.6 threshold). This procedure ensures that each attention head highlights its most focused regions, and the resulting attention maps can be interpreted as unsupervised segmentation maps. Figure 4 shows a few examples of this.

In Section 6, we outline the method of how we hope to produce segmentation maps and discuss potential challenges and considerations associated with this direction.

5.2. Galaxy Zoo morphological classification

To more quantitatively test if the image embeddings produced by the DINOv2 are meaningful, we test how well it does on a simple morphological classification task. Note that there are many other simpler, models that can do morphological classification and can produce similar level of accuracy. The goal here is to specifically test if the DINOv2 embeddings are informative for astronomical images that are not in the ImageNet dataset that DINOv2 has seen.

To classify galaxy images using features extracted by DINOv2, we first preprocessed all images by resizing them to 224×224 pixels, converting them to tensors, and normalizing them using standard ImageNet statistics. We passed the images through a pretrained DINOv2 model to obtain the image embeddings, 1024 for the large model and 384 for the smaller model. These embeddings were saved along with their corresponding Galaxy Zoo 10 labels. We split the dataset into 80% training and 20% test sets, and trained a simple multilayer perceptron (MLP) classifier with one hidden layer (256 units, ReLU activation) and a final output layer matching the number of classes. The model was trained using the Adam optimizer and cross-entropy loss over 30 epochs with a batch size of 64. Table 1 tabulates the model's Top-1 accuracy compared to other models. The accuracies of other models is taken from Table 3 of [9].

It is interesting to see that some labels are well separated in the embedding space, for example, labels 2, 3, 8 and 9. These correspond to Merging Galaxies, Round Smooth Galaxies, Unbarred Loose Spiral Galaxies and Edge-on Galaxies without Bulge, and indeed these morphologies are quite visually distinct. In contrast, labels 6,7 are quite mixed in this space. These are Barred Spiral Galaxies and Unbarred Tight Spiral Galaxies, which have a more subtle visual difference between them.

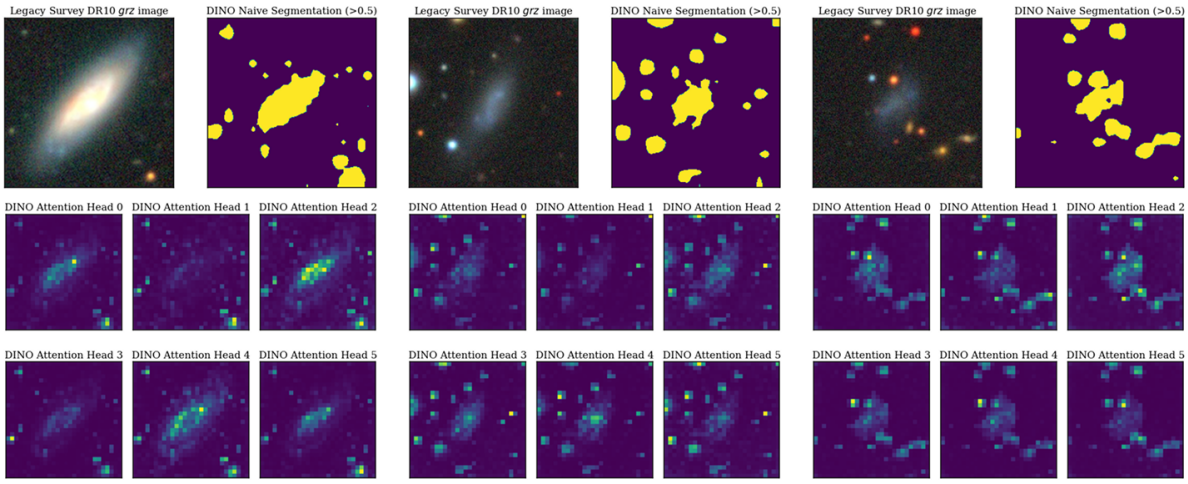


Figure 4. **Examples of attention maps and naive segmentation outputs from the pretrained DINO ViT-S/8 model** for three different galaxy cutouts. In each example, the top-right panel displays the RGB image cutout. While the primary galaxy of interest (the largest, central object) is clearly visible, the images also contain multiple foreground stars and background galaxies. The attention maps highlight how the model is able to pick up on significant sources in the image, however, not yet distinguishing them as separate sources. Retraining the SSL on astronomical images will likely help with this.

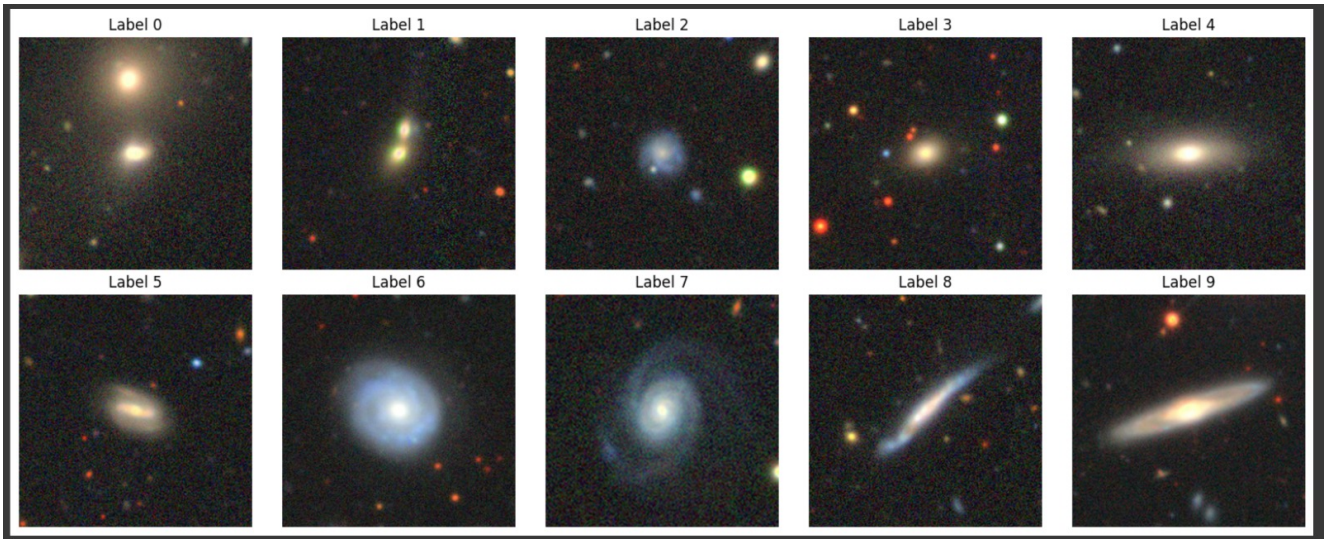


Figure 5. Examples of galaxies from the Galazy Zoo dataset from each of the 10 labels. These morphologies extend from "disturbed galaxies" (label 1) and "merginf galaxies" (label 2) to smooth round galaxies (label 3, 4) and spiral galaxies (labels 6 and beyond). The labels 6 and beyond split spiral galaxies in additional sub classes.

6. Next Steps: Producing Semantic Segmentation Maps

Since DINO is trained in a self-supervised manner and does not include a segmentation head, segmentation maps must be derived from its learned patch embeddings. As detailed in the DINOv2 paper, there are two approaches to produce segmentation maps: a linear head setup and a +ms (multiscale) setup. We summarize these two approaches below and plan out our next steps.

Linear Setup: A linear classifier (e.g., a 1×1 convolution or MLP) is trained to map each DINO patch token to class logits[3]. Since ViT processes images as fixed-size, non-overlapping patches (e.g., 16×16 pixels), an input image of size 224×224 would yields a 8×8 patch grid. This results in a coarse, low-resolution segmentation map, which can be upsampled to full image resolution using bilinear interpolation. As the galaxy images indicate, and depending on the physical resolution of the image, this might be too coarse and so the following method might be better.

[3]Class logits are unnormalized output scores for each class before applying a softmax function

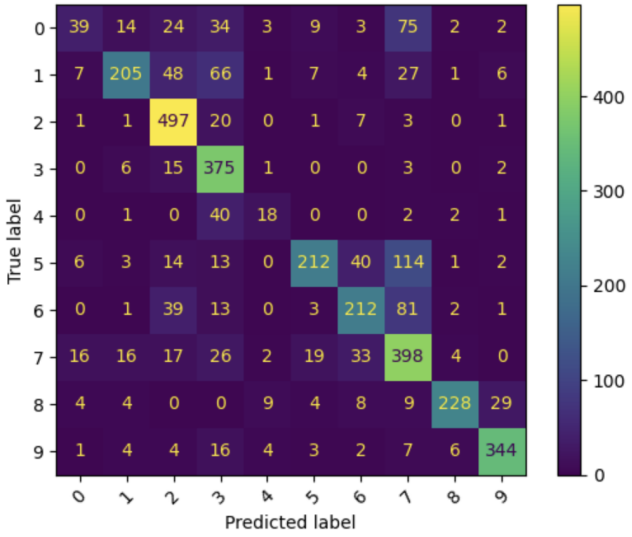


Figure 6. A confusion matrix for the morphological classification task using the Galaxy Zoo dataset using the DINOv2 VIT-L/14 model.

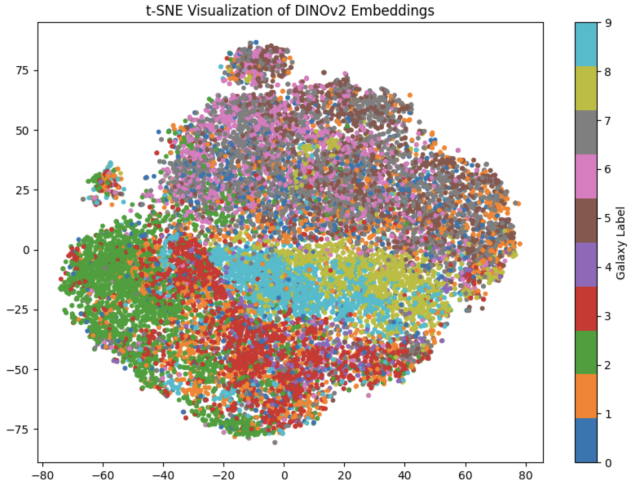


Figure 7. 2D t-SNE projection of galaxy embeddings from DINOv2, colored by galaxy label.

+ms (Multiscale) Setup: This enhanced approach improves segmentation quality by concatenating patch tokens from the last 4 layers of the ViT to improve representation and increasing the input image resolution to produce more patch tokens. Applying multiscale test-time augmentation, where predictions from multiple image scales are averaged. As detailed in [5], this setup achieves performance close to supervised methods using a fully finetuned MAE with an Upernet decoder, demonstrating the strength of DINOv2's learned features.

Both setups are limited by the low spatial resolution of patch-based tokens. To produce finer segmentation maps, several modifications are possible: we could use smaller patch sizes (e.g., 8×8) to get denser token grids. Also, this issue is alleviated further if we use a higher resolution images from telescopes like Hyper Suprime Camera (HSC), Hubble Space Telescope (HST) and James Webb Space Telescope (JWST). The Multi-Modal Universe Dataset includes datasets from these telescopes and we plan on testing our methods there.

Pretraining	Model	Top-1 Accuracy
No pretraining	**DINOv2 ViT-L/14**	**73%**
	DINOv2 ViT-S/14	**67%**
	EfficientNetB0	81%
	ConvNext-nano	76%
	ResNet-18	74%
	DenseNet121	73%
Galaxy Zoo	ConvNext-nano	90%

Table 1. The Top-1 accuracy of the two DINOv2 models we test at morphological classification task on the Galaxy Zoo labelled dataset. We compare the (no pretraining) model performance to other frameworks with no pretraining as well. For the larger DINOv2 model (ViT-L/14), we find comparable accuracy to other models. The accuracies of the other models are taken from [9].

In the very last stages of this project, we came across an implementation of DINOv2 for just the image encoding aspect in the ASTROCLIP model [6]. They use the DINOv2 embeddings to do contrastive learning against the spectral embeddings of the same galaxies. Their goal is very different, but in reading their paper we learned some valuable insights on how we could better process our images. Furthermore, their finetuned DINOv2 model could help us produce better segmentation maps (using methods described above), and we plan on exploring this in the future.

This is how they process their images: Each galaxy image, with spatial resolution $N \times N$ and C channels (e.g., multi-band images), is first divided into fixed-size, non-overlapping patches of size $P \times P$. These patches are flattened into vectors of dimension $P^2 \times C$, resulting in a sequence of $K = N^2/P^2$ patch tokens. Each of these tokens is linearly projected into a latent embedding space of dimension D using a trainable projection matrix. To encode spatial information, they add a learnable one-dimensional positional embeddings to each patch token. They use the ViT-Large (ViT-L) model with a patch size of $P = 12$. For galaxy images with three channels, this results in flattened patch vectors of dimension 432, which are then projected to a 1024-dimensional embedding space. [6] found that such a setup achieved strong performance, and smaller variants were less effective. We therefore intend to use their model and test its use case on our goal of semantic segmentation.

A potential complication of performing segmentation on astronomical images is that, unlike in natural images where objects often have well-defined boundaries, astronomical

objects typically do not. Their brightness peaks at the center and gradually decreases outward, eventually blending smoothly into the background noise. As a result, defining clear segmentation maps becomes challenging, since there is no sharp boundary where the object ends and the background begins. We will need to think carefully about how to best create the small training set of segmented images, but image simulations where the ground truth is better known, is a promising avenue.

7. Conclusions

In this project, we explored the Self-Distillation with No Labels (DINO) framework [1, 5] and evaluated its applicability to astronomical imaging data. Using images from the DESI Legacy Imaging Surveys [2] and morphological labels from the Galaxy Zoo dataset within the MultiModal Universe Dataset [9], we assessed the quality of representations produced by pretrained DINOv2 models.

We find that the DINOv2 model performs well out-of-the-box, achieving a Top-1 accuracy of approximately 73% on morphological classification tasks (see Table 1). Additionally, the image embeddings generated by DINOv2 appear highly informative, indicating that the model—despite not being trained on astronomical images—can extract features that are meaningful for scientific analysis.

Building on these encouraging results, and informed by recent advancements in image embeddings used for contrastive learning [6], we aim to extend this work toward producing semantic segmentation maps of galaxies in future iterations.

References

[1] M. Caron et al. Emerging Properties in Self-Supervised Vision Transformers. *arXiv e-prints*, page arXiv:2104.14294, Apr. 2021. 1, 6

[2] A. Dey et al. Overview of the DESI Legacy Imaging Surveys. , 157(5):168, May 2019. 2, 6

[3] R. Hausen and B. E. Robertson. Morpheus: A Deep Learning Framework for the Pixel-level Analysis of Astronomical Image Data. , 248(1):20, May 2020. 3

[4] P. Melchior et al. SCARLET: Source separation in multi-band images by Constrained Matrix Factorization. *Astronomy and Computing*, 24:129, July 2018. 2

[5] M. Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv e-prints*, page arXiv:2304.07193, Apr. 2023. 1, 2, 5, 6

[6] L. Parker, F. Lanusse, S. Golkar, L. Sarra, M. Cranmer, A. Bietti, M. Eickenberg, G. Krawezik, M. McCabe, R. Morel, R. Ohana, M. Pettee, B. Régaldo-Saint Blancard, K. Cho, S. Ho, and Polymathic AI Collaboration. AstroCLIP: a cross-modal foundation model for galaxies. , 531(4):4990–5011, July 2024. 5, 6

[7] G. Stein et al. Self-supervised similarity search for large scientific datasets. *arXiv e-prints*, page arXiv:2110.13151, Oct. 2021. 1

[8] G. Stein et al. Mining for Strong Gravitational Lenses with Self-supervised Learning. , 932(2):107, June 2022. 1

[9] The Multimodal Universe Collaboration, J. Audenaert, M. Bowles, B. M. Boyd, D. Chemaly, B. Cherinka, I. Ciucă, M. Cranmer, A. Do, M. Grayling, E. E. Hayes, T. Hehir, S. Ho, M. Huertas-Company, K. G. Iyer, M. Jablonska, F. Lanusse, H. W. Leung, K. Mandel, J. R. Martínez-Galarza, P. Melchior, L. Meyer, L. H. Parker, H. Qu, J. Shen, M. J. Smith, C. Stone, M. Walmsley, and J. F. Wu. The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of Astronomical Scientific Data. *arXiv e-prints*, page arXiv:2412.02527, Dec. 2024. 2, 3, 5, 6