Vision foundation models: can they be applied to astrophysics data?

Erica Lastufka
Department of Computer Science
University of Geneva
Geneva, Switzerland 1211
erica.lastufka@unige.ch

Mariia Drozdova
University of Geneva
Geneva, Switzerland 1211
mariia.drozdova@unige.ch

Vitaliy Kinakh
University of Geneva
Geneva, Switzerland 1211
vitaliy.kinakh@unige.ch

Davide Piras
University of Geneva
Geneva, Switzerland 1211
davide.piras@unige.ch

Svyatoslav Voloshynovskiy*
University of Geneva
Geneva, Switzerland 1211
Svyatoslav.Voloshynovskyy@unige.ch

Abstract

Vision foundation models, which have demonstrated significant potential in many multimedia applications, are often underutilized in the natural sciences. This is primarily due to mismatches between the nature of domain-specific scientific data and the typical training data used for foundation models, leading to distribution shifts. Scientific data often differ substantially in structure and characteristics; researchers frequently face the challenge of optimizing model performance with limited labeled data of only a few hundred or thousand images. To adapt foundation models effectively requires customized approaches in preprocessing, data augmentation, and training techniques. Additionally, each vision foundation model exhibits unique strengths and limitations, influenced by differences in architecture, training procedures, and the datasets used for training. In this work, we evaluate the application of various vision foundation models to astrophysics data, specifically images from optical and radio astronomy. Our results show that using features extracted by specific foundation models improves the classification accuracy of optical galaxy images compared to conventional supervised training. Similarly, these models achieve equivalent or better performance in object detection tasks with radio images. However, their performance in classifying radio galaxy images is generally poor and often inferior to traditional supervised training results. These findings suggest that selecting suitable vision foundation models for astro-

*corresponding author

Preprint. Under review.

physics applications requires careful consideration of the model characteristics and alignment with the specific requirements of the downstream tasks.

1 Introduction

In recent years, foundation models have enabled significant advancements in both language and image processing. These large models, trained on vast amounts of data encompassing multiple domains, have shown remarkable capabilities and serve as a cornerstone for numerous applications in science and technology. Foundation models are designed either for representation learning, where the goal is to capture essential characteristics of the data, or for generative purposes, where the model attempts to generate new data samples similar to the training distribution. In this work, we focus on the potential of the learned representations from vision foundation models applied to astrophysical data.

A myriad of architectural innovations and training methodologies have been employed to capture complex patterns in vast datasets of natural images, some scraped from the internet and others more restricted like ImageNet. For many computer vision applications, this provides a robust starting point for further adaptation via fine-tuning for specialized downstream tasks (DSTs). However, the diversity and complexity inherent in foundation models pose substantial hurdles for scientists, particularly those in the fields of physics, astronomy, and biology, who wish to apply these tools for domain-specific inquiries.

The objectives and data used for scientific DSTs are not known or used during the training of foundation models. The foundation models are typically trained in a self-supervised or weakly supervised way. Thus, in accordance to the information bottleneck principle [TPB00], it is not obvious whether these models are able to retain features related to the DST, namely sufficient statistics, in their learned representations. If, as suggested, only DST-related information should be kept in the model's representations (and the rest filtered out), it is difficult to determine what, if any, relevant information a foundation model trained on ImageNet might have for astronomy. Furthermore, since there might be a misalignment between the training data distribution and the DST data distribution, this can additionally result in the problem of *distribution shift*.

In addition to theoretical concerns, practitioners face the challenge of finding *optimal fine-tuning techniques*. Given the limited labeled data typically available for scientific applications, identifying the appropriate training objective, data augmentations, model architecture, and training hyperparameters is essential to enhance the model's performance. Firstly, there is the challenge of selecting an appropriate model, which is both critical and non-trivial. It requires a deep understanding of the model's compatibility with the DST data distribution. Secondly, the way performance varies across different tasks and datasets can be perplexing. Some models excel with minimal labeled data, while others might require substantial label-rich datasets to perform effectively. This performance disparity is often not linear and depends significantly on the nature and volume of the training examples provided for the DST. Thirdly, the feasibility of deploying these models is constrained by available computational capabilities, which varies dramatically based on the model's scale —- from small to giant models. Finally, all of this investigation requires time and substantial effort from domain experts in both the topic of interest and in machine learning best practices.

The goal of our paper is to investigate how popular vision foundation models can be applied to astronomical data and typical downstream tasks. Our experiments used images from both the optical and radio astronomy domains, and we examined classification and detection tasks. By applying various foundation models to these datasets and downstream tasks, we illustrate the differences between foundation models and their implications for obtaining optimal task-specific performance. In doing so, this paper seeks to bridge the gap between advanced machine learning techniques and practical scientific research.

2 Use of Foundation Models in Astrophysics

Figure 1 illustrates the differences between natural images that comprise common training datasets like ImageNet and images from optical and radio astronomy. Unlike natural images, astrophysics images tend to have the following properties:

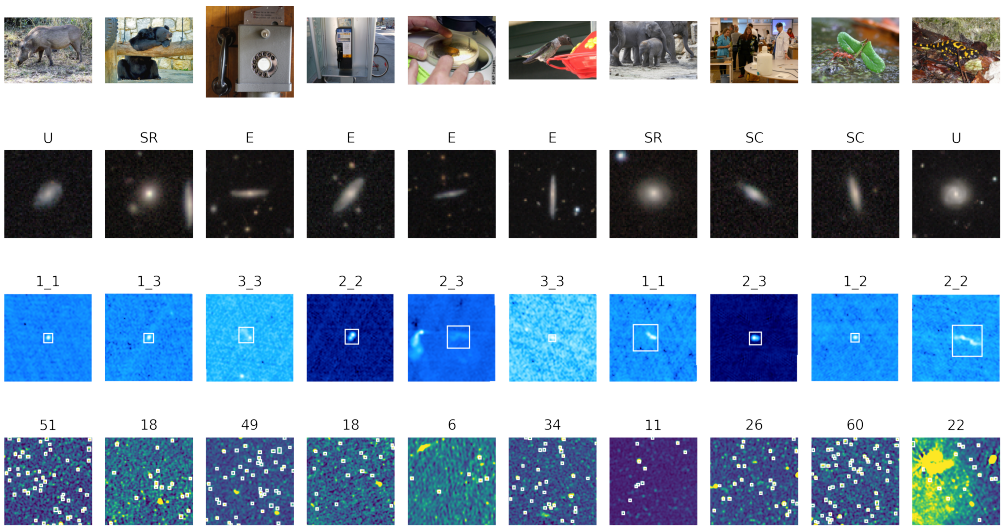


Figure 1: Ten random samples from each of the following datasets: ImageNet-1K (top row), GalaxyMNIST (second row), Radio Galaxy Zoo (RGZ, third row), MeerKAT MGCLS (bottom row). GalaxyMNIST and RGZ are labeled according to morphology class, while MGCLS is labeled by the number of compact sources present. White boxes indicate object bounding boxes used for source detection, when applicable.

Sparseness: most of the images consist of several objects that occupy a small fraction of the total image size.

Noise: systematic noise is present in the images (this is especially notable for the radio images).

High dynamic range: the brightness of the objects in the image can span several orders of magnitude, not easily captured by traditional normalization. Research goals most often define the ideal combination of filters or various weighting schemes for displaying the image (for example, to emphasize compact bright sources vs diffuse extended emission).

Artefacts: instrumental effects or residuals from image reconstruction can form structures of different scales in the images.

Because of these fundamental differences, it is commonly assumed in the academic community that vision models trained on natural images might be inadequate for performing tasks in astronomy, especially radio astronomy where images are mathematically reconstructed from sparse samples measured in the Fourier plane.

While there are many extremely specialized DSTs in both optical and radio astronomy, often taking advantage of the many wavelength channels available, some more common tasks involving a single or small number of channels are listed in Table 1. Machine learning can offer a number of applicable techniques, and pretrained foundation models can be of use in most cases.

Astrophysics Task	Machine Learning Task or Method	References
Image reconstruction	CNNs, de-noising diffusion	[SGF+22, DKB+23]
Source detection	Object detection	[VVB+19, JZWY23, RMS+23]
Source characterization	Object segmentation	[FOD+20, SMF+23]
Source classification	Image or object classification	[BAC+19, RMS+23, MLB+23]
Source deblending	Instance segmentation	[BAC+19, RG19, HR22]
Object/event discovery	Anomaly detection	[LB21, VCB+21]
RFI detection	CNNs, GANs	[VFLFB19, LYX+21]
Background/foreground removal	UNET, de-noising diffusion	[CL21, ZGD+23, CBT+24]

Table 1: Common tasks in astrophysics and their machine learning analogues.

2.1 Data

In this work we evaluated galaxy morphology classification and source detection, two tasks common to both optical and radio astronomy. Details of the datasets used to perform these tasks are in Table 2, and sample images are shown in Figure 1.

Classification datasets are images with a single galaxy centered in the cutout. GalaxyMNIST (GMNIST, [WLG$^+$22]) is a balanced dataset of four categories: smooth and round (SR), smooth and cigar-shaped (SC), edge-on-disk (E), and unbarred spiral (U). Radio Galaxy Zoo (RGZ) is unbalanced, with labels determined according to number of distinct radio components (C) and number of intensity peaks (P) in each source [WGA$^+$24]. Possible combinations are: 1C 1P, 1C 2P, 1C 3P, 2C 2P, 2C 3P, and 3C 3P, a total of 6 classes. Labeling was done by citizen scientists who performed inspection by varying the relative intensity of continuum radio emission and infrared observations. It is not always visually obvious from the normalized PNG images if there is a difference between a cutouts containing the same number of bright peaks – for example, one containing a single galaxy with two peaks, and another two different components with one peak each.

Dataset	Number of labels	Image size (pixels)	Data characteristics	Task
GMNIST[2]	10K	64×64	centered on bright optical galaxy	classification
RGZ[3]	7.7K	132×132	noisy, centered on bright radio galaxy	classification, detection
MGCLS[4]	10K	256×256	many radio sources, varying magnitude	source detection

Table 2: The datasets used in this study.

Source detection datasets also consist of cutouts from wide-field images; they may have one to six galaxies in the central area, as in RGZ, or have tens of galaxies in a single cutout, as in MGCLS. Examples of bounding boxes around sources are shown in Figure 1; MGCLS labels are consistent in that the boxes only designate compact sources, and not other examples of extended emission or larger sources that might also be present in the images. Because RGZ's labels also contain extended sources, bounding boxes can much larger and filled with a large amount of noise.

2.2 Foundation Models

Foundation models exhibit significant variability in terms of their architectures, training, and performance characteristics. These differences arise from factors such as the size and nature of the training datasets, the number of parameters they incorporate, and their underlying architectural frameworks. Moreover, these models leverage a carefully curated set of data augmentation techniques and diverse pretraining strategies, each tailored to minimize a specific loss function. Table 3 lists the foundation models investigated in this study. With the exception of DINOv2, all were pretrained on ImageNet-1k's 1.2 million training images.

Model Name	Backbone	Number of parameters	Pre-training Dataset	EMA
MAE [HCX$^+$21]	ViT-Base 16x16	86M	ImageNet-1k	-
DINOv2 [ODM$^+$23]	ViT-Base 14x14	86M	LVD-142M (proprietary)	+
DINOv1 [CTM$^+$21]	ViT-Base 16x16	86M	ImageNet-1k	+
MSN [ACM$^+$22]	ViT-Base 16x16	86M	ImageNet-1k	+
ResNet 50 [HZRS16]	ResNet-50	25.6M	ImageNet-1k	-
ResNet 18 [HZRS16]	ResNet-18	11.5M	ImageNet-1k	-

Table 3: The foundation models used in this study.

MAE. The Masked AutoEncoder (MAE, [HCX$^+$21]) uses masked image modeling (MIM) pretraining, reconstructing masked image patches from only a few visible patches. Unlike NLP inspired

[2] `https://github.com/mwalmsley/galaxy_mnist`

[3] [WGA$^+$24]

[4] `https://doi.org/10.48479/7epd-w356`

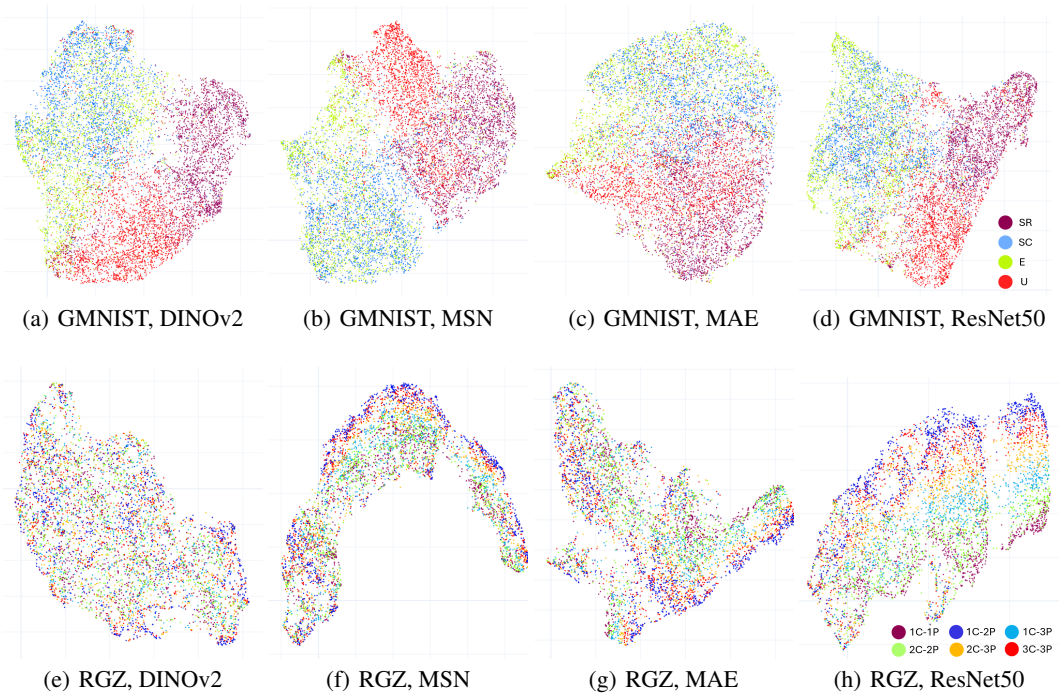
approaches like BeIT [BDPW21], which discretize visual tokens through an autoencoder, MAE directly processes the visible image patches through an encoder. The resulting output is then combined with mask tokens to reconstruct the original image using a decoder. Reconstruction error is used as the objective training function.

DINO. DINOv2 [ODM+23] improves upon its predecessor, DINOv1 [CTM+21], by utilizing a larger and more curated dataset, LVD-142M. DINOv2 integrates the DINO cross-entropy loss with the MIM objective employed in iBOT [ZWW+22]. DINOv2 benefits from the Sinkhorn-Knopp batch normalization technique used in SwAV [CMM+20]. The model employs a teacher-student training paradigm, where a teacher network computes the feature representation of global views of an image, while a student network computes the feature representation of a local views, a series of smaller crops. The model is optimized to train the student network to replicate the teacher network's output. The teacher network is periodically updated using an exponential moving average (EMA) of the student network's parameters.

MSN. Masked Siamese Networks (MSN, [ACM+22]) combine MIM with Siamese networks to avoid pixel-level and token-level reconstructions. MSN also uses teacher-student training, with the student network computing the feature representation of a partially masked image.

ResNet. Residual Networks are a staple of computer vision, introduced by in [HZRS16]. They are convolutional neural networks (CNNs) that use skip connections to bypass one or more layers, allowing the network to maintain accuracy over very deep networks. ResNet18 consists of 8 blocks, while ResNet50 has 16. In this work we use the pre-trained weights available through torchvision [mc16], which were obtained by training in a supervised fashion for classification of ImageNet-1k using cross-entropy loss. A series of very specific data augmentations, including TrivialAugment [MH21], random erasing, mixup and cutmix, helped increase top-1 accuracy relative to the original augmentation scheme of random resized crops and horizontal flips.

2.3 UMAP Illustration



(a) GMNIST, DINOv2 (b) GMNIST, MSN (c) GMNIST, MAE (d) GMNIST, ResNet50

(e) RGZ, DINOv2 (f) RGZ, MSN (g) RGZ, MAE (h) RGZ, ResNet50

Figure 2: GMNIST (upper row) and RGZ (lower row) UMAP for features extracted using DINOv2, MSN, MAE, and Resnet50.

We provide empirical evidence that foundation models learn different representations of the same data. Extracting features from each foundation model and dataset, performing PCA and dimensionality

reduction using the first 10 components with UMAP illustrates the latent space (Figure 2). We chose UMAP over T-SNE as UMAP better preserves global structure, which is important when local structure in the images might be dominated by noise. For the unbalanced RGZ dataset, a random sample with balanced classes is displayed.

Although UMAP attempts to distill the relationships between thousand-dimensional vectors into two dimensions, it can still reveal clusters where images with similar embeddings reside. This can be deceptive because the model may have not learned relevant embeddings; visualizing the associated class labels can offer insight.

The UMAP representation of the DINO and MSN feature spaces for GMNIST is quite similar. Classes SR and U, representing smooth-and-round and unbarred spiral galaxies, are in distinct clusters, although MSN features show more overlap between the two than DINO's. The majority of SR and U galaxy images are separated by ResNet50 as well, although small clusters of these find themselves in other areas. Galaxies that appear predominantly elliptical – smooth-and-cigar-shaped (SC) and edge-on-disk (E) – share similar latent space embeddings; clustering is most distinct for DINOv2.

Latent space grouping according to class is far less evident with the RGZ data. The best visible separation is seen in ResNet50. The ResNet model we used was trained with the goal of image classification, as opposed to DINO and MSN, which seek to reconcile local and global image characteristics, and MAE, whose objective is image reconstruction.

Through visualization of the latent space, we see that our chosen foundation models retain more representations that are class-relevant for GMNIST than RGZ. This is not to say that the learned representations are completely unrelated to the information contained in RGZ, simply that the most important features (according to PCA) are not strongly correlated with image class.

2.4 Downstream Task Study Methodology

For *image classification*, we used the models 'out-of-the-box' by attaching a single-layer linear classifier head to the pre-trained foundation model backbone. Only the classifier was allowed to train, so the backbone is considered frozen in that the weights and biases of the backbone stay fixed. We compare results with fully supervised classifiers; in this case, both the backbone network used as well as the classifier head were allowed to train.

Source detection was done in a similar fashion, using Faster-RCNN [RHGS17]. The implementation of Faster-RCNN is slightly different for Vision Transformers and CNNs. For a Vision Transformer backbone, a simple feature pyramid network (FPN) based on only the output of the last large-stride feature map of the backbone is used [LMGH22]. The Faster-RCNN implementation using ResNet also uses a FPN, but one that is heirarchical, acting on feature maps from different convolutional blocks in the ResNet rather than just the last one.

After the backbone and FPN, object detection is done by applying a sliding window of the region proposal network (RPN) that predicts whether an object is present or not. These proposals are pooled (RoI pooling) and finally bounding boxes and classes are predicted. We considered the FPN, RPN and RoI layers to comprise the detection head in this situation, as these components are what is attached on top of the backbone network.

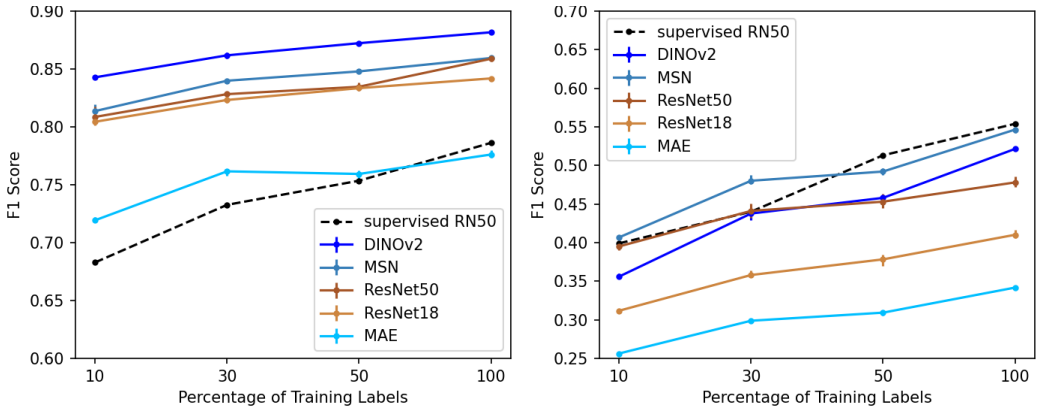
We again considered the case of a frozen backbone and trainable detection head. For this source detection task, we also sometimes allowed the parameters of the backbone network to train; this is referred to as fine-tuning.

3 Galaxy Morphology Classification

We demonstrate the galaxy morphology classification problem with both optical and radio datasets labeled courtesy of the Galaxy Zoo project. Twenty percent of RGZ images contain two labeled galaxies, and four percent contain more than two; these images were all excluded from the classification dataset, reducing the total number of samples to 4692 training images and 1189 test images. Therefore the training data available for RGZ was just over half that available for GMNIST, which was split with 80% of its 10K images in the training set. The images in both datasets were already

normalized, so the only data transformation we performed was to scale the images up to the same size (224 $\times$ 224 pixels) from their native resolutions (GMNIST: 64 $\times$ 64, RGZ: 132 $\times$ 132).

Hyperparameters for the classifier training were similar for both datasets, using a learning rate of 0.0005 and a batch size of 16. Classifiers were trained for 100 epochs on GMNIST and 200 on RGZ. Class weights inversely proportional to the number of samples present in the training dataset were used in the cross-entropy loss function used to classify RGZ. To illustrate the performance differences in low- and high-label regimes, the number of samples used for training was varied at 10%, 30%, 50% and 100%, while the test set was kept the same. For comparison, we trained a ViT-Base model with patch size 16 from scratch in a fully-supervised manner, as well as ResNets with both 50 and 18 layers. The best-performing supervised model in both cases was ResNet50.



(a) GMNIST (b) RGZ

Figure 3: F1 scores for optical (left) and radio (right) galaxy morphology classification, as a function of percentage of training labels. Error bars show the maximum and minimum scores out of three different runs.

Figure 3 shows classification F1 score as a function of training label percentage. While results differ according to dataset, some trends are in common. Performance generally improved when the number of training labels increased, MAE was the worst performer overall, and the smaller ResNet18 did not classify as well as ResNet50.

On GMNIST, it was notable that the use of a foundation model improved classification relative to fully supervised training by up to 15%. In fact, the only configuration where starting from a foundation model was outperformed by supervised training is MAE with all available training labels. Supervised training with all available training labels took 4-6 hours on a single GPU depending on the backbone architecture, while training a linear classifier layer on frozen features took around one minute on a laptop GPU. Therefore, the benefits of using a foundation model for this particular dataset, even one pretrained on natural images, are large increases in performance and speed.

DINOv2 and MSN both performed better than the ResNets, although ResNet50 improved the most when trained on 100% of the data. One might explain the superior performance of DINO and MSN by the number of parameters of their ViT-Base architectures, were it not for MAE which is also ViT-Base. Unlike DINO and MSN, which use student-teacher paradigms to reconcile partially masked or locally cropped views of images, MAE's objective is to reconstruct an entire image from a few patches. We speculate that MAE retains too much information that is irrelevant to classification; because large portions of GMNIST images consist of empty space, this might overly contribute to the learned representations.

Classification was more difficult with Radio Galaxy Zoo, with F1 scores of less than 0.6 even when using all available training labels. This could be because more of the image is dominated by noise, and less by the galaxy to be classified. The labeling schematic could be another reason for low scores, as discussed in Section 2.1. Only MSN out-performed supervised training, and only for a low percentage of training labels.

The poor performance on radio galaxy classification is not consistent with results from previous works such as [SSW+24] and [LBT+24], which have shown much higher F1 scores (up to 0.94 for DINOv2) achieved on the MiraBest dataset [PS23] using pre-trained foundation models. MiraBest is designed for binary classification in the Fanarhoff-Riley scheme, which could be roughly described as a single Gaussian component with one peak versus a single component with two peaks.

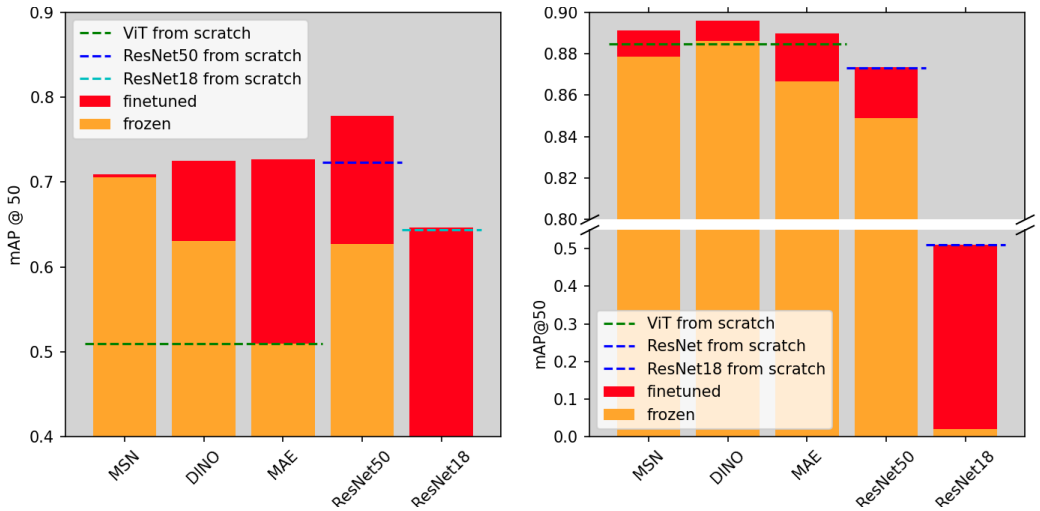
The challenges with the RGZ dataset may stem from the more subtle characteristics that distinguish the classes. In fact, re-labeling the images by the number of bright peaks resulted in F1 scores of up to 0.74, with all models but MAE out-performing a supervised baseline. Conversely, when re-labeling the images by the number of distinct radio components, the maximum F1 score achieved was 0.63, from MSN with all labels – the only time use of a foundation model beat supervised training.

It seems that foundation models struggle to identify distinct radio sources, especially when they involve multiple emission peaks. Radio flux islands, containing one or more peaks, are usually identified through analytic source-finding algorithms via connected-component labeling, starting by selecting regions of pixels above a certain flux threshold. This threshold is usually up to 10 times the background RMS, and the brightest pixels it contains can be orders of magnitude greater. This information of relative flux is lost after compression and normalization of the images.

Unlike GMNIST, RGZ images are reconstructions which contain noise with strong inter-pixel correlations on the scale of the synthesized beam. These statistical differences between ImageNet and radio images suggest that a well-performing model may require knowledge of a very specific set of features; a true case of distribution shift. While the out-of-the-box application of foundation models to GMNIST images performed well, it is not sufficient for classification on RGZ where it offers little to no advantage over supervised training.

4 Source detection

Source detection was performed on radio continuum datasets RGZ, which mostly contains single galaxies labeled by morphology, and MGCLS, where a single image contains from 10-100 individual compact sources. MGCLS images may also contain larger sources which are not labeled.



(a) RGZ (b) MGCLS

Figure 4: Source detection results for all backbones, on RGZ (left) and MGCLS (right).

As was done for classification, images were upscaled from their native resolution. For RGZ, scaling to 224×224 meant that the average single object size became 55×55 pixels, while the 25th percentile were 24×24 pixels. MGCLS crops were doubled in size from 256×256. This was done so that the average compact source occupied 16×16 pixels, and the 25th percentile 14×14 pixels. The size of the sources relative to the transformer patch size is important; sources unable to occupy most of a

patch were difficult to detect. This was proven empirically – performance on detection with MGCLS increased by more than 10% when the image size doubled.

We chose to use Faster-RCNN to compare performance on object detection. Our first experiment kept the backbones frozen, allowing only the detection head layers (FPN, RPN and ROI) to train. Due to differences in architecture this resulted in a different number of trainable parameters: 21M for ViT-Det, 17.8M for Resnet50 with FPN, and 11M for Resnet18. Because of the structure of the FPN for ViT-Det, for this task the DINO backbone used was DINOv1, which had available pre-trained weights for the architecture ViT-Base with 16x16 patch size (DINOv2 uses 14x14). In the second experiment we allowed the weights and biases of the backbone networks to update as well (fine-tuning).

We kept the default training data augmentations: a randomly applied blur, contrast adjustment, and color jitter. Different learning rates, varied in order to ensure a smooth decay of the training loss, were required depending on the dataset and network. Training was for 100 epochs, with a batch size of 16, performed on 1-2 GPUs depending on availability. We report the mean average precision at an intersection-over-union (IOU) threshold of 50%, known as the mAP@50.

Results showed that MSN and DINO are good foundation models for this task. They outperformed training from scratch, even when the backbone was frozen, except for MSN on MGCLS. Although the ResNet models did not outperform training from scratch when used as a frozen backbone, they equaled or exceeded that performance when fine-tuned. Detection on galaxies in the RGZ dataset appears to be a task particularly suited to ResNets, while Vision Transformers do very well at detecting compact sources in MGCLS. Generally, detection was very good on MGCLS for networks larger than ResNet18. Possible reasons for the higher mAP@50 include the single category – compact source – and the tens of sources present in the images compared to the low numbers of sources in RGZ. Additionally, bounding boxes in RGZ can be quite large, including a lot of noise; we saw from classification that networks had trouble identifying distinct source components, so a tendency to predict smaller bounding boxes around individual peaks rather than one large box around an extended single component could result in lower scores.

From these results, it seems that there is nothing to lose by starting from a foundation model, as performance will not be worse than training from scratch and may in fact be better. It is important to ensure that properties of the architecture, such as patch size in the case of ViTs, are complementary to the data being used for the downstream task.

5 Conclusions

Our results demonstrate that state-of-the-art vision foundation models can immediately perform classification of optical galaxies and source detection of radio galaxies with 72 - 88% precision. In the case of optical galaxy classification, all foundation models out-performed supervised training. F1 scores of close to 0.85 for 800 labeled training images, 10% of the available total, show that foundation models can be used to great effect for initial investigations into using machine learning for scientific tasks. Of the models tested, Vision Transformers DINOv2 and MSN performed the best, along with CNN ResNet50. MAE, despite its success on natural images, is optimized for reconstruction loss and failed to encode characteristics which are highly relevant for galaxy classification.

Shared characteristics of the model training distribution and the images of GMNIST might have contributed to good classification performance. Even though GMNIST images contain a lot of empty space, a good portion of the image is occupied by the galaxy of interest. While stars and other objects may be present in the background, they are small relative to the main galaxy, even if the galaxy is seen edge-on. On the other hand, classification of radio galaxies was challenging, as even the main radio source was small and the images were dominated by noise and patterns from reconstruction. In this case, supervised training outperformed all foundation models except MSN.

For a dataset like RGZ with a clear distribution shift from training to down-stream task data, applicability of foundation models is limited. The models did retain some relevant information about the data – in this case, the information about bright peaks – and designing the data augmentations or label scheme to take advantage of that could improve results.

The need for compatibility between the dataset, down-stream task, and foundation model appears in the source detection example as well. Once images were scaled so that the Vision Transformer patch

size was smaller than the average radio source, source detection results were encouraging. High mAP@50 scores showed that foundation models larger than ResNet 18 could identify radio sources, especially when the backbone was allowed to fine-tune. However, improvements due to fine-tuning were inconsistent and (in the case of MSN) much smaller than might be expected. This situation might improve given more optimal methods of fine-tuning.

Even in cases where results are good, such as source detection, there remains a large gap in performance between tasks done on scientific images and tasks done on natural images. State-of-the-art classification on thousand-class ImageNet-1k has a top-1 accuracy of 92%, while the best object detection on the standard COCO dataset has a mAP@50 of 0.73, which may seem low in comparison to our results but is the average metric across all 90 classes. Closing the gap in performance is necessary to achieve scientifically meaningful results, but it will have to be done with far less computing and human resources than are available to large companies. Techniques which can operate with a low number of labeled training examples and with training a small number of network parameters will be especially valuable in extending the capabilities of vision foundation models to scientific images. In this respect, the semi-supervised learning proposed in [VTK$^+$20] seems to be a very interesting option for further investigation.

In this study, we considered very simple way of fine-tuning. Perhaps a more powerful fine-tuning method, based on the simultaneous use of labeled and unlabeled data, might enhance the situation, as suggested in [VTK$^+$20]. This would be particularly exciting for astrophysics, where unlabeled data is plentiful. In future work, we will examine the potential of such semi-supervised learning framework for fine-tuning.

Data availability

GalaxyMNIST (`https://github.com/mwalmsley/galaxy_mnist`) and MGCLS (`https://doi.org/10.48479/7epd-w356`) are public datasets. Radio Galaxy Zoo (`https://radio.galaxyzoo.org/`) will soon have its first data release, and the dataset used here is available upon reasonable request. The code and instructions to reproduce these experiments is available at `https://github.com/elastufka/fm4astro`.

References

[ACM+22] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked Siamese Networks for Label-Efficient Learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 456–473, Berlin, Heidelberg, October 2022. Springer-Verlag.

[BAC+19] Colin J Burke, Patrick D Aleo, Yu-Ching Chen, Xin Liu, John R Peterson, Glenn H Sembroski, and Joshua Yao-Yu Lin. Deblending and classifying astronomical sources with Mask R-CNN deep learning. *Monthly Notices of the Royal Astronomical Society*, 490(3):3952–3965, December 2019.

[BDPW21] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[CBT+24] T Chen, M Bianco, E Tolley, M Spinelli, D Forero-Sanchez, and J P Kneib. The stability of deep learning for 21cm foreground removal across various sky models and frequency-dependent systematics. *Monthly Notices of the Royal Astronomical Society*, 532(2):2615–2634, August 2024.

[CL21] M. Cohen and W. Lu. A diffusion-based method for removing background stars from astronomical images. *Astronomy and Computing*, 37:100507, October 2021.

[CMM+20] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[CTM+21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, October 2021. ISSN: 2380-7504.

[DKB+23] M. Drozdova, V. Kinakh, O. Bait, O. Taran, E. Lastufka, M. Dessauges-Zavadsky, T. Holotyak, D. Schaerer, and S. Voloshynovskiy. Radio-astronomical image reconstruction with a conditional denoising diffusion model. *Astronomy & Astrophysics*, December 2023.

[FOD+20] H. Farias, D. Ortiz, G. Damke, M. Jaque Arancibia, and M. Solar. Mask galaxy: Morphological segmentation of galaxies. *Astronomy and Computing*, 33:100420, October 2020.

[HCX+21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. arXiv:2111.06377 [cs].

[HR22] Ryan Hausen and Brant Robertson. Partial-Attribution Instance Segmentation for Astronomical Source Detection and Deblending, January 2022. arXiv:2201.04714 [astro-ph].

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. ISSN: 1063-6919.

[JZWY23] P. Jia, Y. Zheng, M. Wang, and Z. Yang. A deep learning based astronomical target detection framework for multi-colour photometry sky survey projects. *Astronomy and Computing*, 42:100687, January 2023.

[LB21] Michelle Lochner and Bruce A. Bassett. Astronomaly: Personalised Active Anomaly Detection in Astronomical Data. *Astronomy and Computing*, 36:100481, July 2021. arXiv:2010.11202 [astro-ph].

[LBT+24] Erica Lastufka, Omkar Bait, Olga Taran, Mariia Drozdova, Vitaliy Kinakh, Davide Piras, Marc Audard, Miroslava Dessauges-Zavadsky, Taras Holotyak, Daniel Schaerer, and Svyatoslav Voloshynovskiy. Self-Supervised Learning on MeerKAT Wide-Field Continuum Images, August 2024. arXiv:2408.06147 [astro-ph].

[LMGH22] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring Plain Vision Transformer Backbones for Object Detection, June 2022. arXiv:2203.16527 [cs].

[LYX+21] Z. Li, C. Yu, J. Xiao, M. Long, and C. Cui. Detection of radio frequency interference using an improved generative adversarial network. *Astronomy and Computing*, 36:100482, July 2021.

[mc16] TorchVision maintainers and contributors. TorchVision: PyTorch's Computer Vision library, 2016. Publication Title: GitHub repository.

[MH21] Samuel G. Müller and Frank Hutter. TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation, March 2021.

[MLB+23] Grant Merz, Yichen Liu, Colin J Burke, Patrick D Aleo, Xin Liu, Matias Carrasco Kind, Volodymyr Kindratenko, and Yufeng Liu. Detection, instance segmentation, and classification for astronomical surveys with deep learning (deepdisc): detectron2 implementation and demonstration with Hyper Suprime-Cam data. *Monthly Notices of the Royal Astronomical Society*, 526(1):1122–1137, November 2023.

[ODM+23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, April 2023. arXiv:2304.07193 [cs].

[PS23] Fiona A M Porter and Anna M M Scaife. MiraBest: a data set of morphologically classified radio galaxies for machine learning. *RAS Techniques and Instruments*, 2(1):293–306, January 2023.

[RG19] David M Reiman and Brett E Göhre. Deblending galaxy superpositions with branched generative adversarial networks. *Monthly Notices of the Royal Astronomical Society*, 485(2):2617–2627, May 2019.

[RHGS17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017.

[RMS+23] S. Riggi, D. Magro, R. Sortino, A. De Marco, C. Bordiu, T. Cecconello, A. M. Hopkins, J. Marvil, G. Umana, E. Sciacca, F. Vitello, F. Bufano, A. Ingallinera, G. Fiameni, C. Spampinato, and K. Zarb Adami. Astronomical source detection in radio continuum maps with deep neural networks. *Astronomy and Computing*, 42:100682, January 2023.

[SGF+22] K. Schmidt, F. Geyer, S. Fröse, P.-S. Blomenkamp, M. Brüggen, F. de Gasperin, D. Elsässer, and W. Rhode. Deep learning-based imaging in radio interferometry. *Astronomy & Astrophysics*, 664:A134, August 2022. Publisher: EDP Sciences.

[SMF+23] Renato Sortino, Daniel Magro, Giuseppe Fiameni, Eva Sciacca, Simone Riggi, Andrea DeMarco, Concetto Spampinato, Andrew M. Hopkins, Filomena Bufano, Francesco Schillirò, Cristobal Bordiu, and Carmelo Pino. Radio astronomical images object detection and segmentation: A benchmark on deep learning methods. *Experimental Astronomy*, May 2023. arXiv:2303.04506 [cs].

[SSW+24] Inigo V Slijepcevic, Anna M M Scaife, Mike Walmsley, Micah Bowles, O Ivy Wong, Stanislav S Shabala, and Sarah V White. Radio galaxy zoo: towards building the first multipurpose foundation model for radio astronomy with self-supervised learning. *RAS Techniques and Instruments*, 3(1):19–32, January 2024.

[TPB00] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[VCB+21] V. Ashley Villar, Miles Cranmer, Edo Berger, Gabriella Contardo, Shirley Ho, Griffin Hosseinzadeh, and Joshua Yao-Yu Lin. A Deep-learning Approach for Live Anomaly Detection of Extragalactic Transients. *The Astrophysical Journal Supplement Series*, 255(2):24, August 2021. Publisher: The American Astronomical Society.

[VFLFB19] Etienne E. Vos, P. S. Francois Luus, Chris J. Finlay, and Bruce A. Bassett. A Generative Machine Learning Approach to RFI Mitigation for Radio Astronomy. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, October 2019. ISSN: 1551-2541.

[VTK+20] Slava Voloshynovskiy, Olga Taran, Mouad Kondah, Taras Holotyak, and Danilo Rezende. Variational Information Bottleneck for Semi-Supervised Classification. *Entropy*, 22(9):943, September 2020. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[VVB+19] A Vafaei Sadr, Etienne E Vos, Bruce A Bassett, Zafiirah Hosenie, N Oozeer, and Michelle Lochner. DeepSource: point source detection using deep learning. *Monthly Notices of the Royal Astronomical Society*, 484(2):2793–2806, April 2019.

[WGA+24] O Ivy Wong, A F Garon, M J Alger, L Rudnick, S S Shabala, K W Willett, J K Banfield, H Andernach, R P Norris, J Swan, M J Hardcastle, C J Lintott, S V White, N Seymour, A D Kapinska, H Tang, B D Simmons, and K Schawinski. Radio Galaxy Zoo Data Release 1: 100,185 radio source classifications from the FIRST and ATLAS surveys. *in prep*, 2024.

[WLG+22] Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W Willett, Steven Bamford, Lee S Kelvin, Lucy Fortson, Yarin Gal, William Keel, Karen L Masters, Vihang Mehta, Brooke D Simmons, Rebecca Smethurst, Lewis Smith, Elisabeth M Baeten, and Christine Macmillan. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, January 2022.

[ZGD+23] Xingchen Zhou, Yan Gong, Furen Deng, Meng Zhang, Bin Yue, and Xuelei Chen. Foreground removal of CO intensity mapping using deep learning. *Monthly Notices of the Royal Astronomical Society*, 521(1):278–288, May 2023.

[ZWW+22] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer, January 2022. arXiv:2111.07832 [cs].