



Article scientifique

Article

2025

Published version

Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

Examining vision foundation models for classification and detection in optical and radio astronomy

Lastufka, Erica Anne; Bait, O.; Drozdova, Mariia; Kinakh, Vitaliy; Piras, Davide; Audard, Marc;
Dessauges, Miroslava; Holotyak, Taras; Schaerer, Daniel Olivier; Voloshynovskyy, Svyatoslav

How to cite

LASTUFKA, Erica Anne et al. Examining vision foundation models for classification and detection in optical and radio astronomy. In: Astronomy & astrophysics, 2025, vol. 703, p. A217. doi: 10.1051/0004-6361/202553691

This publication URL: <https://archive-ouverte.unige.ch/unige:189067>

Publication DOI: [10.1051/0004-6361/202553691](https://doi.org/10.1051/0004-6361/202553691)

Examining vision foundation models for classification and detection in optical and radio astronomy

E. Lastufka^{1,*}, O. Bait³, M. Drozdova¹, V. Kinakh¹, D. Piras¹, M. Audard², M. Dessauges-Zavadsky², T. Holotyak¹, D. Schaerer², and S. Voloshynovskiy¹

¹ Department of Computer Science, University of Geneva, 7 route de Drize, 1227 Carouge, Switzerland

² Department of Astronomy, University of Geneva, 51 Chemin Pegasi, 1290 Versoix, Switzerland

³ SKA Observatory, Jodrell Bank, Lower Withington, Macclesfield SK11 9FT, UK

Received 7 January 2025 / Accepted 8 September 2025

ABSTRACT

Context. 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, and researchers frequently face the challenge of optimizing model performance with limited labeled data of only a few hundred or thousand images.

Aims. This work evaluates the performance of vision foundation models in astrophysics, with a focus on identifying the best practices for adapting these models to domain-specific datasets. We aim to establish a framework for selecting, fine-tuning, and optimizing these models for common tasks in optical and radio astronomy.

Methods. We compared multiple foundation models, including self-supervised, weakly supervised, and distillation-based architectures, across two representative optical and radio datasets. Experiments involved different fine-tuning strategies, projector heads, and data preprocessing techniques, with performance evaluated on classification and detection metrics.

Results. Features extracted by specific foundation models improved classification accuracy for optical galaxy images compared to conventional supervised training. Similarly, these models achieved equivalent or superior performance in object detection tasks with radio images. However, classification performance for radio galaxy images was generally poor, often falling short of traditional supervised approaches.

Conclusions. These findings suggest that selecting suitable vision foundation models for astrophysics applications requires careful consideration of the model characteristics and alignment with the specific requirements of the downstream tasks. This study demonstrates that vision foundation models can be effectively adapted to astrophysical applications, provided practitioners iterate on model selection, training strategies, and data handling. The proposed framework bridges the gap between these advanced models and the unique demands of astronomy, enabling broader adoption of deep learning in the field.

Key words. methods: data analysis – techniques: image processing – radio continuum: general

1. Introduction

Large machine learning models, trained on vast amounts of data encompassing multiple domains, have in recent years enabled significant advancements in both language and image processing. Known as foundation models, these serve as a cornerstone for numerous everyday applications. Foundation models are designed either for representation learning, in which the goal is to capture essential characteristics of the data, or for generative purposes, where the model attempts to create 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 images.

While machine learning is frequently used to manage large amounts of astrophysics image data, vision foundation models are not yet commonly employed. With the exception of ResNets (He et al. 2016) pretrained on ImageNet (Deng et al. 2009), most works in optical and radio astronomy employ models trained from scratch (e.g. Slijepcevic et al. 2024; Lochner & Bassett 2021; Becker et al. 2021). This includes networks trained

for tasks such as image classification and object detection that should be able to benefit directly from the many foundation models available. A more detailed literature review will follow in Sections 3 and 4, but two major themes emerge: firstly, models trained on natural images are not believed to have learned features relevant to astrophysics images, and so are disregarded, and secondly, the desire to employ lightweight models as to not over-fit small labeled datasets precludes the use of modern foundation models, whose smallest architectures tend to have tens of millions of trainable parameters.

Indeed, the proliferation of foundation models in recent years has led to a seeming overabundance of choice to the machine learning practitioner. Each model has its strengths and weaknesses; the diversity and complexity inherent in foundation models can pose substantial hurdles for scientists who wish to apply them for domain-specific inquiries. Therefore, the primary goal of this paper is to experimentally address the initial choice of vision foundation model, given properties of astrophysics data, specific downstream tasks (DSTs), and availability of resources for fine-tuning the foundation model.

Vision foundation models fall broadly into three categories, distinguished by their training methods, architectures, and

* Corresponding author: erica.lastufka@unige.ch

objectives: “self-supervised” models trained on a single modality, “weakly supervised” multimodal models, and “distillation” models that agglomerate learned representations from multiple pretrained models. Supervised models are some of the classic pretrained models available through software libraries such as `huggingface`, `Tensorflow`, and `PyTorch`; convolutional neural networks (CNNs) trained for image classification such as the ResNet family, and more specialized architectures trained for object detection and segmentation such as YOLO (Redmon et al. 2016) and SAM (Kirillov et al. 2023). Self-supervised foundation models do not have a single DST objective, and include masked autoencoders (MAEs, He et al. 2022), masked Siamese networks (MSNs, Assran et al. 2022), DINO, and its improvement DINOv2 (Caron et al. 2021; Oquab et al. 2024). Weakly supervised models enable cross-modal representation learning, first introduced by CLIP (Radford et al. 2021), with its loss function improved upon by SigLIP (Zhai et al. 2023). Knowledge distillation is frequently used to extract smaller models from larger ones, but it can also unify representations of multiple models. AM-RADIO (Ranzinger et al. 2024) distills a student model from multiple teacher models, allowing it to integrate information from several existing foundation models including CLIP and DINOv2.

There are also many types of loss functions used to train foundation models, which impact their effectiveness for specialized DSTs. Self-supervised learning frequently uses a contrastive loss to distinguish between different views of the data, or a reconstruction objective to reconstruct parts of an image. Non-contrastive methods, often employing regularization, are aimed at learning stable, robust features without explicit contrastive pairs. Foundation models also employ various backbone architectures with different numbers of parameters, ranging from CNNs to Vision Transformers (ViTs, Dosovitskiy et al. 2020). Of more concern to scientists is the lack of scientific images in pretraining data. Most foundation models are pretrained on ImageNet or similar large collections of natural images scraped from the web. Certainly the objectives and specific data used for scientific DSTs are neither known nor used during the training of foundation models. Thus, in accordance with the information bottleneck principle (Tishby et al. 2000), 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 this principle suggests, 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; this is one potential explanation for the aforementioned lack of application of pretrained foundation models in the literature. Furthermore, there is likely a misalignment between the training data distribution and the DST data distribution; this problem is known in the field of computer science as “distribution shift”.

In addition to theoretical concerns, practitioners face the challenge of finding optimal fine-tuning techniques. Given that labeled data available for scientific applications is typically limited, identifying the appropriate training objective, data augmentations, model architecture, and training hyperparameters is essential to enhance the model’s performance. In this work, we provide a framework to guide the selection of foundation models for astrophysical imaging, tailored to specific characteristics of the target dataset, DST, and available resources for fine-tuning. Through this empirical approach, our study aims to clarify the strengths and weaknesses of vision foundation models, as well

as to provide practical insights that can inform model selection and adaptation.

The paper is organized as follows. First we discuss the use of foundation models in astrophysics in Section 2, along with our choices of illustrative datasets, various foundation models, and DST methodology. Section 3 examines galaxy morphology classification, and several fine-tuning techniques one can use to improve performance on a challenging dataset. A second DST, object detection, is evaluated in Section 4. Based on these experiments, we generalize our findings into practical guidelines in Section 5, before concluding in Section 6.

2. Use of foundation models in astrophysics

Figure 1 illustrates the differences between natural images that comprise common training datasets such as ImageNet and images from optical and radio astronomy. Unlike natural images, astrophysics images are found nowhere on Earth and tend to have the following properties:

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 methods. 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).

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

Because of these fundamental differences, it is rare to encounter any deep learning applications in astrophysics literature where networks are trained starting from the pretrained weights of a foundation model (the only example we found was Burke et al. 2019). This is especially true for 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 it is possible to use pretrained foundation models in most cases, despite the lack of examples in the literature.

One should be aware that certain tasks, such as rare object discovery, are compromised by the preprocessing conventionally applied for foundation model input. Normalizing high dynamic range images can lead to the loss of critical morphological detail, or the choice of cropping or downscaling size for preparing wide-field images dictates whether compact or extended features are preserved. This trade-off significantly impacts the discovery space for truly new and unexpected structures. Another limitation arises from the fixed input channel requirements of many pretrained foundation models. Models trained on RGB images expect three input channels, so applying these to multiwavelength datasets, such as optical images with five *ugriz* channels, would require either omitting channels, losing valuable information, or replacing the model’s input layer, which can diminish the benefits of pretraining.

It is worth noting that other domains, such as medical or Earth-observing satellite imaging, share some of the aforementioned statistics in their images. To our knowledge, they do not

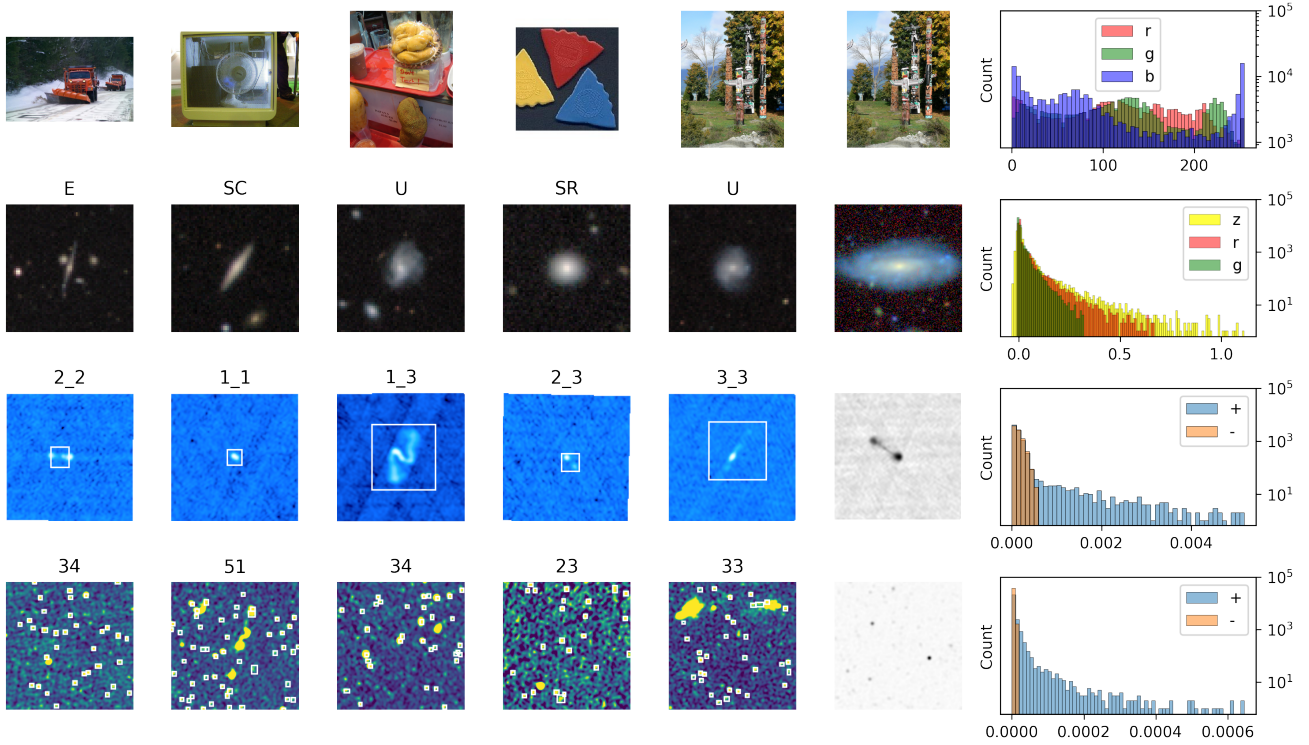


Fig. 1. Random samples from each of the following datasets: ImageNet-1K (top row), GalaxyMNIST (second row), Radio Galaxy Zoo (RGZ, third row), and MeerKAT MGCLS (bottom row). GalaxyMNIST and RGZ are labeled according to morphology class, while MGCLS is labeled by the number of compact sources present. White boxes indicate object bounding boxes used for source detection, when applicable. The rightmost two columns display sample images in their raw data format, unscaled, next to the corresponding histogram. GalaxyMNIST combines data from the Dark Energy Camera’s r , g , and z channels, while radio images are single-channel continuum images reconstructed from Fourier space, which can therefore include negative values as is indicated by the different colors in the histogram.

Table 1. Common tasks in astrophysics and their machine learning analogs.

Astrophysics task	Machine learning task or method	Selected references
Image reconstruction	CNNs, denoising diffusion	(Schmidt et al. 2022; Drozdova et al. 2023)
Source detection	Object detection	(Vafaei Sadr et al. 2019; Jia et al. 2023; Riggi et al. 2023)
Source characterization	Object segmentation	(Farias et al. 2020; Sortino et al. 2023)
Source classification	Image or object classification	(Burke et al. 2019; Riggi et al. 2023; Merz et al. 2023)
Source deblending	Instance segmentation	(Burke et al. 2019; Reiman & Göhre 2019; Hausen & Robertson 2022)
Object/event discovery	Anomaly detection	(Lochner & Bassett 2021; Villar et al. 2021)
RFI detection	CNNs, GANs	fee(Vos et al. 2019; Li et al. 2021)
Background/foreground removal	UNET, denoising diffusion	(Cohen & Lu 2021; Zhou et al. 2023; Chen et al. 2024a)

suffer from all of these effects at once, which is the common case in astrophysics, so we confine the subject of our work and the literature referenced strictly to the astrophysics domain.

2.1. Data

In this work, we chose to evaluate galaxy morphology classification and source detection, two tasks common to astronomy across the electromagnetic spectrum. The scarcity of publicly available, labeled datasets prohibits a comprehensive evaluation of all tasks mentioned in Table 1 in both radio and optical, although in the future we hope that such a study becomes possible. Details of the datasets used to perform galaxy morphology classification and source detection 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, (Walmsley et al.

2022)) 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 (Wong et al. 2024). The observed combinations present in RGZ 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 cutout 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.

Source detection datasets also consist of cutouts from wide-field images; they may have one to six galaxies in the central

Table 2. Datasets used in this study.

Dataset	Number of labels	Image size (pixels)	Data characteristics	Task
GMNIST ¹	10K	64 × 64	Centered on bright optical galaxy	Classification
RGZ ²	7.7K	132 × 132	Noisy, centered on bright radio galaxy	Classification, detection
MGCLS ³	10K	256 × 256	Many radio sources, varying magnitude	Source detection

Notes. ⁽¹⁾https://github.com/mwalmsley/galaxy_mnist ⁽²⁾(Wong et al. 2024) ⁽³⁾<https://doi.org/10.48479/7epd-w356>.

Table 3. Foundation models used in this study.

Model name	Backbone	Number of parameters	Pretraining dataset	EMA
MAE (He et al. 2022)	ViT-Base 16×16	86M	ImageNet-1k	–
DINOv2 (Oquab et al. 2024)	ViT-Base 14×14	86M	LVD-142M (proprietary)	+
DINOv1 (Caron et al. 2021)	ViT-Base 16×16	86M	ImageNet-1k	+
MSN (Assran et al. 2022)	ViT-Base 16×16	86M	ImageNet-1k	+
SigLIP (Zhai et al. 2023)	ViT-Base 16×16	86M	WebLI (English)	–
AM-RADIO (Ranzinger et al. 2024)	ViT-Base 16×16	98.2M	various	–
ResNet 50 (He et al. 2016)	ResNet-50	25.6M	ImageNet-1k	–

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.

The datasets were split into training and test sets with an 80% of samples in the training set and 20% in the test set. Training datasets were further reduced to 10, 30, and 50% of what was available, and data not included in the training sets was used as validation. The learning rate, number of training epochs, and batch size¹ was determined using very minimal hyperparameter tuning on the 30% labeled datasets. The validation set (remaining 70% of the training data) was used to ensure that losses were declining smoothly with the given hyperparameters for both the training and validation dataset. Once this ceased to be the case, training was stopped and the number of training epochs was chosen. When the hyperparameters were chosen, there was no further need of the validation set, so in the case where all available labels were used, these samples were included in the training data. For each dataset and task, the test sets remained the same to ensure the results could be compared to each other, and none of the test images were ever included in either training or validation datasets.

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. Most were pretrained on ImageNet-1k’s 1.2 million training images.

¹ For detection, batch size was limited to maximum 16 by available memory.

Some foundation models are pretrained with a particular DST objective, but even if they are not, they can be used as *backbones* together with particular *projector heads* that enable them to be leveraged for a variety of tasks. This involves removing any specialized head that is present for pretraining – for example, removing the final linear classifier layer of PyTorch’s pretrained ResNet – and attaching the task-specific head.

2.2.1. Categories of foundation models

Earlier, we broadly categorized vision foundation models into three groups: self-supervised models (SSLs), weakly supervised models, and distilled models. Each model type is characterized by different methodologies and architectures aimed at robust representation learning.

The SSL models are designed to extract meaningful representations from large datasets without the need for labels, operating under the assumption that neither the DSTs nor the corresponding labels are available during training. This allows SSL models to focus on learning representations that are generalizable and transferable to many different tasks. Embeddings can be learned through masking regions or patches of the input image, then inferring the missing content as is done by a MSN. Embedding reconstruction models, such as denoising autoencoders and MAEs, utilize an encoder-decoder architecture to compress the input image into a latent representation and reconstruct it. Alternatively, joint embedding models bypass reconstruction by learning representations for augmented views of the same image, with one encoder fixed as a reference and another trainable encoder processing a masked version. However, these models are susceptible to mode collapse in which embeddings lose meaningful variability. Techniques used by models such as SimCLR (Chen et al. 2020), BYOL (Grill et al. 2020), MSN, SwAV (Caron et al. 2020), DINO, Barlow Twins (Zbontar et al. 2021), and VICReg (Bardes et al. 2022) can mitigate this issue. Hybrid models combining the principles of embedding reconstruction and joint embedding also exist, such as CAE (Chen et al. 2024b) and BeIT (Bao et al. 2022), and a recent innovation is joint embedding prediction models such as I-JEPA (Assran et al. 2023) and World Models (Ha & Schmidhuber 2018), which introduce an

additional projector to predict masked representations to enhance training stability.

Weakly supervised learning (WSL) models operate in scenarios where full labels are unavailable but auxiliary data, such as text descriptions or metadata, provides partial supervision. These models have dual-encoder architectures, most often with one processing images and another encoding text to generate semantic embeddings. The objective is to align these embeddings in a shared latent space. Representative models in this category include CLIP, GLIP (Li et al. 2022a), SigLIP, CoCa (Yu et al. 2022), and ALIGN (Jia et al. 2021), which excel in tasks such as zero-shot learning and cross-modal retrieval.

Distilled models rely on a teacher-student paradigm, where a smaller, more efficient model (student) learns to replicate the outputs or distributions generated by a larger, more complex model or models (teacher). The goal is to transfer the teachers' knowledge to the student model; distillation is most often used to create smaller versions of a large foundation model. It can also be used to combine and enhance the outputs of multiple teacher models, such that the student inherits their capabilities across multiple tasks. NVIDIA's AM-RADIO model (Ranzinger et al. 2024) combines multi-teacher distillation with foundation models, distilling representations from various high-performance models in both standard vision and multimodal tasks.

2.2.2. Foundation models used in this study

MAE. The MAE uses masked image modeling (MIM) pre-training, reconstructing masked image patches from only a few visible patches. Unlike natural language processing-inspired approaches such as BeIT, 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.

MSN. Masked Siamese networks combine MIM with Siamese networks to avoid pixel-level and token-level reconstructions. A MSN also uses teacher-student training, with the student network computing the feature representation of a partially masked image.

DINO. DINOv2 improves upon its predecessor, DINOv1 (Caron et al. 2021), by utilizing a larger and more curated dataset, designated LVD-142M. DINOv2 integrates the DINO cross-entropy loss with the MIM objective employed in iBOT (Zhou et al. (2022)). DINOv2 benefits from the Sinkhorn–Knopp batch normalization technique used in SwAV. 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 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.

SigLIP. Sigmoid language image pretraining (SigLIP), is a multimodal method designed to align image and textual representations in a shared embedding space using a modified contrastive learning framework. Pretrained on WebLI's English text-image pairs, SigLIP uses a sigmoid loss that allows the use of extremely large batch sizes in pretraining, when using multiple GPUs.

AM-RADIO. Agglomerative Model–Reduce All Domains Into One (AM-RADIO) merges CLIP, SAM, and DINOv2 into

a unified model via multi-teacher distillation. Merging the concepts of a student learning from an ensemble of teachers with foundation models, AM-RADIO trained student models from CLIP, DINOv2, and SAM, the authors used a cosine distance loss to match a simple student adapter head with the teacher spatial feature vectors and summary vectors (when available). In this way, the student is able to mimic the teacher and perform their targeted DSTs.

In addition to these models, we compare with ResNet, one of the most used networks in machine learning for astrophysics. Residual Networks are a staple of computer vision, introduced by He et al. (2016). They are CNNs that use skip connections to bypass one or more layers, allowing the network to maintain accuracy over very deep networks. ResNet-18 consists of 8 blocks, while ResNet-50 has 16. In this work we use the pretrained weights available through torchvision (Marcel & Rodriguez 2010), 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 (Müller & Hutter 2021), 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

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 distil 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.

As is shown in Figure 2, the UMAP representation of the DINO, MSN, SigLIP, and AM-RADIO feature spaces for GMNIST is quite similar. Classes SR and U, representing SR and unbarred spiral galaxies, are in distinct clusters, although MSN features show more overlap between the two than the rest. The majority of SR and U galaxy images are separated by MAE and ResNet-50 as well, although small clusters of these find themselves in other areas. Galaxies that appear predominantly elliptical – 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 ResNet-50 and AM-RADIO. 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.

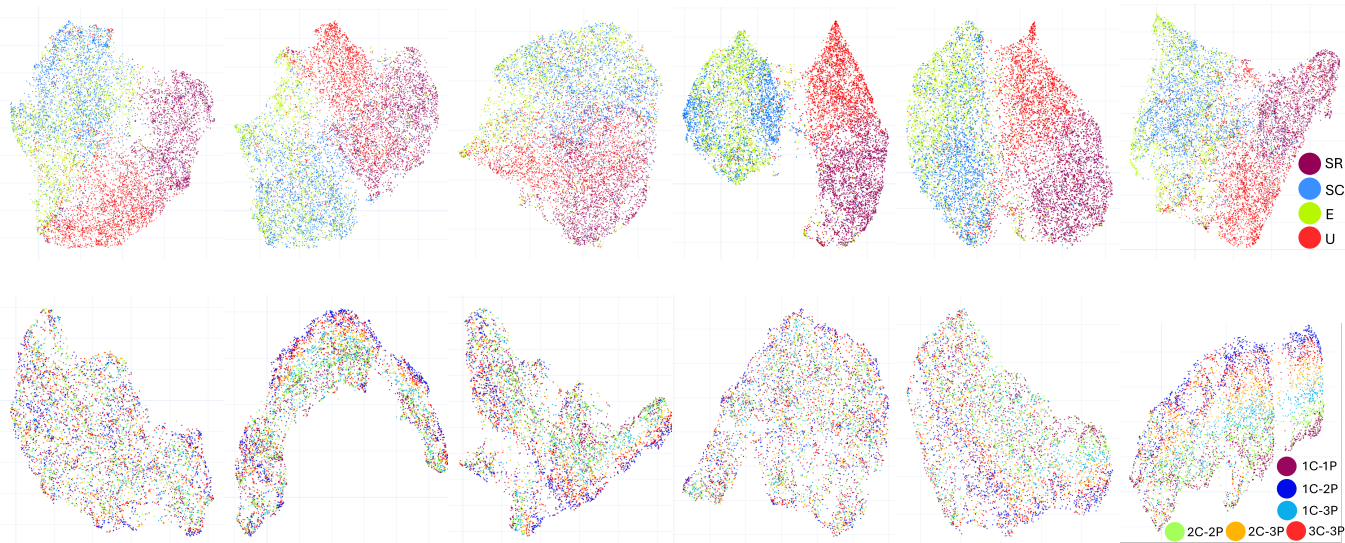


Fig. 2. GMNIST (upper row) and RGZ (lower row) UMAP for features extracted using DINOv2, MSN, MAE, SigLIP, AM-RADIO, and ResNet-50 (from left to right).

2.4. Downstream task study methodology

We examine three approaches to finding appropriate models and datasets for tasks. The first is perhaps the simplest: fine-tuning the model to the given dataset. The second, which is often closely related the other two, is to choose the model based on the DST. The third is to adapt the data to suit the given model.

Fine-tuning a foundation model to perform a task on a particular dataset is commonly known as “transfer learning”. This involves unfreezing parameters of the backbone network and allowing those gradients to update. It can be done layer-by-layer or block-by-block, in order to achieve maximum performance with the minimum fine-tuned parameters, or by unfreezing the entire backbone. In this paradigm, the weights and biases of the model itself are optimized for a given dataset and DST. In this work, we consider fine-tuning for both image classification and source detection.

Related to transfer learning is the idea of adapting the model to the DST. Both this method and full fine-tuning involve fitting a projector head to the foundation model backbone, to enable it to carry out the DST. However, one can choose to only update the parameters in the head rather than the entire head plus backbone configuration. This significantly reduces computational cost. A head architecture that is able to filter task-relevant features from the backbone’s latent space should be able to perform well when large computational resources are not available.

For this work, the projector head used for image classification was a single-layer linear classifier. Source detection was done using Faster-RCNN (Ren et al. 2017), whose projector head is implemented 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 (Li et al. 2022b). The Faster-RCNN implementation using ResNet also uses a FPN, but one that is hierarchical, 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.

The first two approaches are the most familiar to machine learning practitioners; the last will be more intuitive to astrophysicists. Adapting the data to suit the chosen model simply means finding the optimal data preprocessing method, taking into account the characteristics of both the model and the data. In this paper, we closely examine several ways in which this can affect image classification.

3. Galaxy morphology classification

Galaxy morphology classification is a common task in both optical and radio astronomy, and large labeled datasets have been produced thanks to Galaxy Zoo projects (Walmsley et al. 2022; Wong et al. 2024). Previous deep learning approaches in optical classification have relied heavily on CNNs trained from scratch (e.g., Zhu et al. 2019; Maslej-Krešňáková et al. 2021; Pandya et al. 2023; Urechiatu & Frincu 2024). A few works have fine-tuned ImageNet-trained models, such as the DenseNet121 of Hui et al. (2022) and the EfficientNetB5 of Kalvankar et al. (2021). Some works have investigated combining the advantages of ViT’s attention mechanism with CNNs, such as Wei et al. (2024), Dagi (2023) and Cao et al. (2024). However, application of pure ViTs has been limited and resulted in lower performance than CNNs (Lin et al. 2022), although Kumar et al. (2023) reports fine-tuning to be better than training from scratch. Radio galaxy classification work has been more limited due to lack of large datasets and a well-defined labeling scheme, and so far only employs CNNs; Becker et al. (2021) and Ndong’u et al. (2023) provide comprehensive reviews of the topic. Due to the difficult nature of the data, traditional ML techniques, feature engineering, and finding augmentations specific to radio images, are also active fields of study, while some seek to enhance training datasets with synthetic images using generative deep learning.

The datasets we use to demonstrate the galaxy morphology classification task, GMNIST and RGZ, are both cutouts of single optical or radio galaxies labeled courtesy of the Galaxy Zoo project. Twenty percent of RGZ images contain two labeled

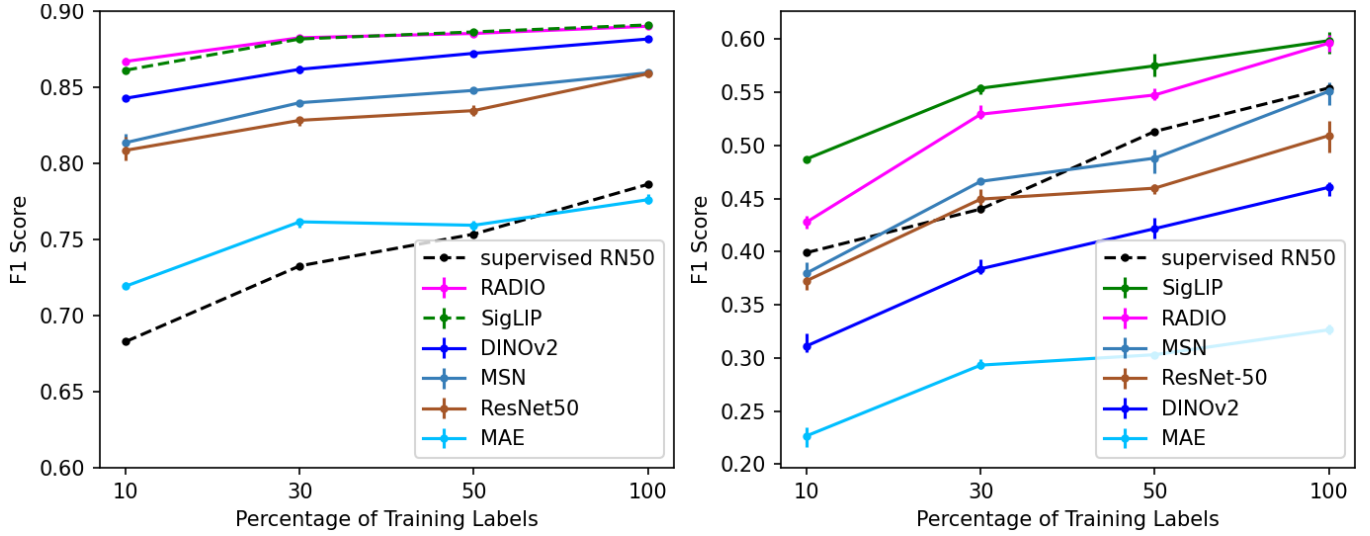


Fig. 3. F1 scores for optical (left) and radio (right) galaxy morphology classification, as a function of percentage of training labels. Error bars show the maximum and minimum scores out of three different runs.

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.

Hyperparameters for the classifier training were similar for both datasets, using a learning rate of 0.0005 and a batch size of 64. Only the classifier head was allowed to train, so the backbone is considered frozen in that the weights and biases of the foundation model backbone stay fixed. 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, as it is an unbalanced dataset. 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 representative models from scratch in a fully supervised manner in which both the backbone and classifier head were trained. The fully supervised models were a ViT-Base model with patch size 16, as well as ResNets with both 50 and 18 layers. The best-performing supervised model was ResNet-50, so its results are shown in Figure 3.

3.1. Classification performance

Figure 3 shows classification F1 score as a function of training label percentage. The F1 score is the harmonic mean of precision and recall, calculated using the number of true and false positives for each class, then averaged. Results for precision and recall are shown in Appendix B; they are not significantly different for these datasets, so the F1 score is a good representation of overall performance.

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 ResNet-18 did not classify as well as ResNet-50.

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. As supervised training takes hours rather than the tens of seconds required to train a linear classifier layer on frozen, there is a clear benefit to using a foundation model for this particular dataset. Even starting from a model pretrained on natural images, one observes large increases in performance and speed.

The distillation model AM-RADIO performed almost exactly as well as SigLIP, with a slight advantage in the low-label regime. Both these models demonstrate a F1 score at least 0.2–0.25 better than the other models. Self-supervised models DINOv2 and MSN both performed better than the ResNets, although ResNet-50 improved the most when trained on 100% of the data. 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. [Balestrieri & LeCun \(2024\)](#) found that the learned representations with the reconstructive power were the least informative for perception. Our findings support that result; it seems that MAE learns features that are irrelevant to classification.

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 is discussed in Section 2.1. Only SigLIP and AM-RADIO consistently and significantly outperformed supervised training.

The poor performance on radio galaxy classification is not consistent with results from previous works such as [Slijepcevic et al. \(2024\)](#) and [Lastufka et al. \(2024\)](#), which have shown much higher F1 scores (up to 0.94 for DINOv2) achieved on the MiraBest dataset ([Porter & Scaife 2023](#)) using pretrained 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.

Challenges with the RGZ dataset may stem from the more subtle characteristics that distinguish the classes. The average F1 score over all classes is strongly determined by that of the simplest and most numerous class, with one radio component

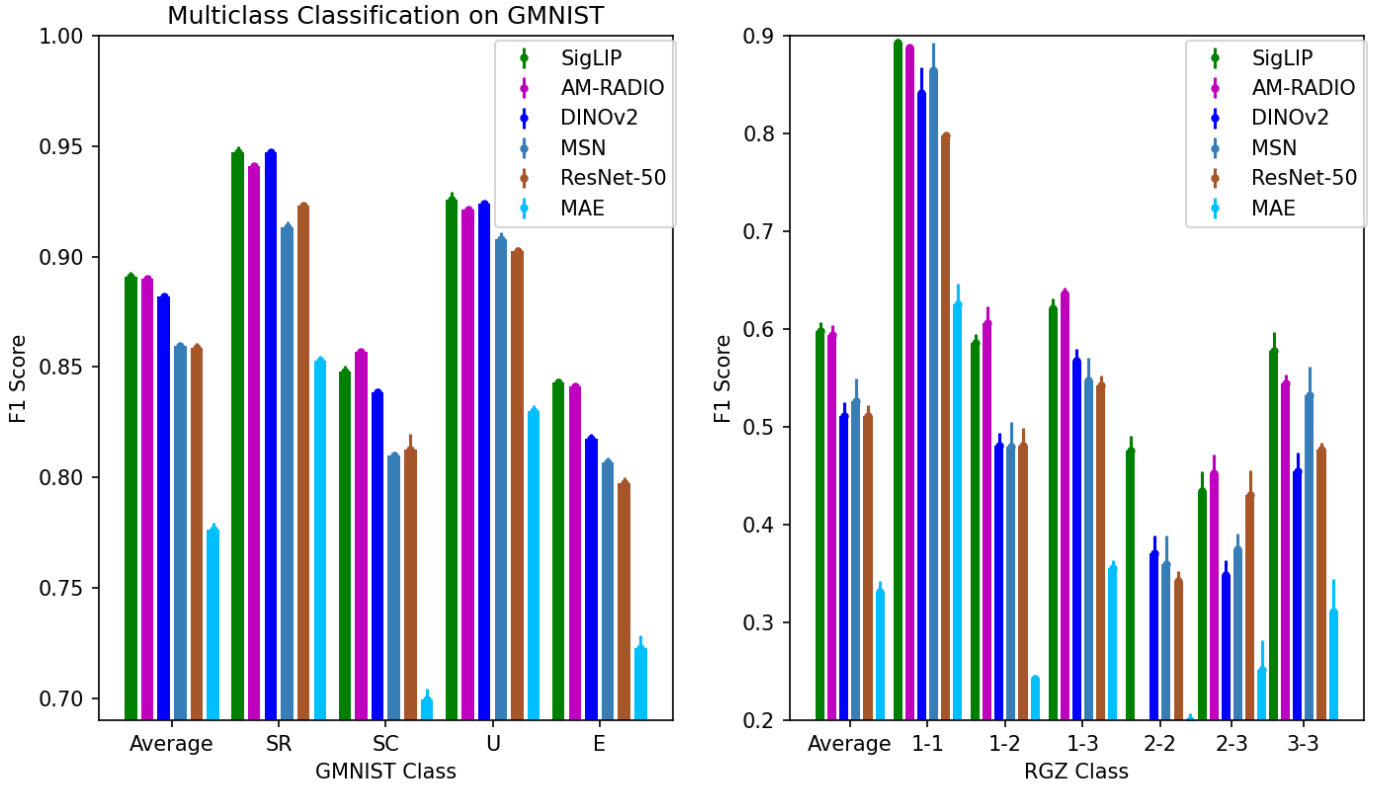


Fig. 4. F1 score per class for classification on GMNIST (left) and RGZ (right). Models can confidently identify single Gaussian sources but struggle with multi-peaked or multi-component sources.

and one bright peak. Use of the weighted cross-entropy loss helps mitigate the problem of class imbalance; however, the fact remains that this particular class is easily distinguishable from the others, and likely to influence the classification even if the dataset were balanced. Figure 4 (right side) also shows that classification for galaxies with two radio components and either two or three bright peaks is especially poor. Relabeling the images by the number of bright peaks resulted in F1 scores of up to 0.74 (seen in Figure 5), with all models but MAE outperforming a supervised baseline. Conversely, when relabeling the images by the number of distinct radio components, the maximum F1 score achieved was 0.63, from MSN with all labels.

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 that 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.

There are two immediately recognizable visual differences between GMNIST, on which classification worked well, and RGZ (see again Figure 1): radio galaxies occupy a much smaller area of the total image than optical ones, and background noise

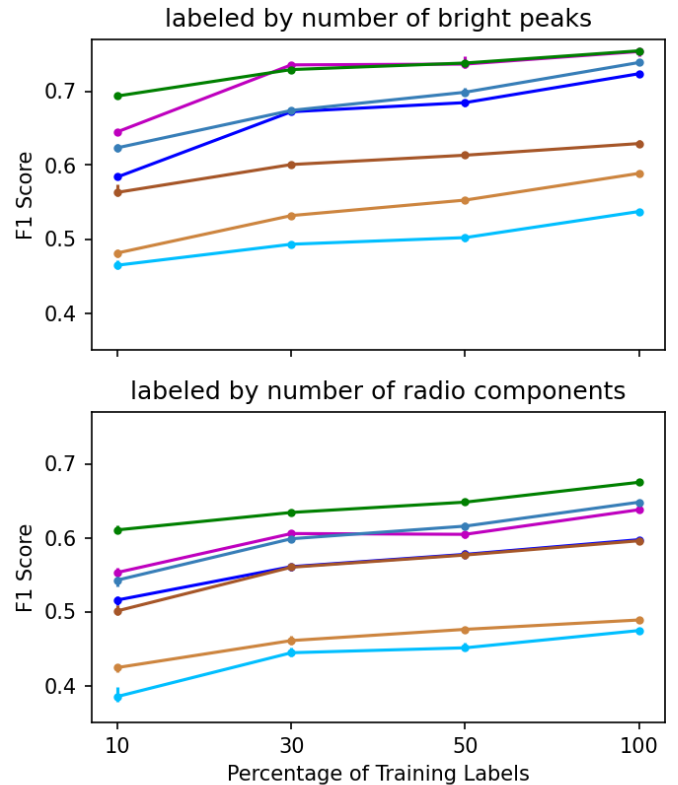


Fig. 5. Label scheme redefined to the number of bright peaks (top) or number of individual radio components (bottom). Higher F1 scores in the top chart show that vision models are better at distinguishing bright peaks in images rather than flux islands defining radio components.

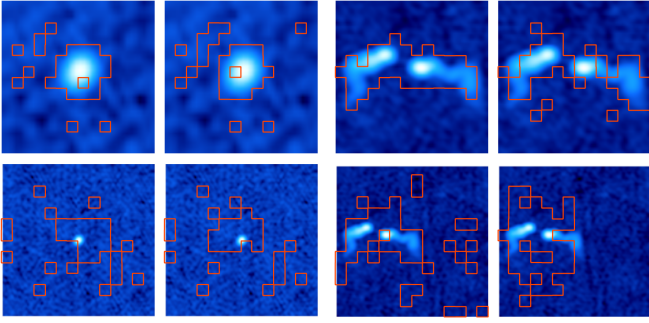


Fig. 6. Attention maps (red contours) overlaid on closely cropped (top) and uncropped (bottom) RGZ images. A typical compact source (labeled as 1C-1P) is shown on the left, with a more complex source with diffuse elements (labeled as 2C-3P) on the right. The attention maps are shown for two different attention heads as contours at the 90th percentile level.

is much higher and patterned. We experimented with bringing the RGZ dataset into better alignment with GMNIST by leveraging the bounding box information to make very close crops around each radio galaxy. This resulted in a F1 score of 0.70 for all labels, using AM-RADIO to extract frozen features.

While this is not yet on par with classification performance on GMNIST, it is clear that the foundation models are able to extract more information from the closely cropped images. Figure 6 shows the output of selected attention heads on both the cropped (top row) and uncropped (bottom row) RGZ images for two different classes of galaxy. Even though the 1C-1P class of the left side image is in general confidently classified, the attention maps show that areas of high interest to the different attention heads correspond mostly to noise in the uncropped images. The central compact source of the left image only occupies about one 16×16 pixel patch in the uncropped image, which has already been upscaled from its original size. Attention maps for sample images from all datasets and all transformer-based backbones can be found in Appendix A. 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, without careful adaptation to suit the chosen foundation model.

3.2. Optimization strategies

We demonstrate various common strategies for improving classification performance on a challenging dataset such as RGZ. These include, roughly in order of the time it takes to implement the changes:

- Use a multilayer perceptron (MLP) for classification instead of a single linear layer.
- Use a bigger model with more parameters.
- Use different preprocessing methods.
- Fine-tune the backbone via Low-Rank Adaptation (LoRA).
- Fine-tune the backbone using a semi-supervised method.

The first two strategies consist of adapting the model, the third adapts the data to suit the model, and the last two are different ways to implement fine-tuning. For purposes of illustration, we report the results using all available training labels and the models AM-RADIO and ResNet-50 in Table 4. With the notable exceptions of employing a close crop and fine-tuning, these techniques improved the classification result of AM-RADIO more than that of ResNet-50. In fact, using either larger input image sizes or larger networks decreased the accuracy of ResNet.

Table 4. Effect of various optimization techniques on RGZ classification.

Technique	ResNet-50 F1	AM-RADIO F1
Whitening layer	0.519 (+0.008)	0.603 (+0.013)
MLP	0.526 (+0.015)	0.623 (+0.034)
RN101/ViT-Large	0.486 (−0.025)	0.625 (+0.036)
RN152/ViT-Huge	0.486 (−0.026)	0.653 (+0.063)
Crop	0.643 (+0.132)	0.705 (+0.115)
Upsample 320	0.481 (−0.031)	0.630 (+0.041)
Upsample 480	0.465 (−0.046)	0.666 (+0.080)
Upsample 512	0.454 (−0.057)	0.667 (+0.078)
LoRA	0.674 (+0.163)	0.692 (+0.102)
Fine-tuning	0.737 (+0.226)	0.721 (+0.132)

The effects of each technique is reported on its own; one can expect that employing multiple techniques is a little less than the sum of each individual improvement. For example, using AM-RADIO and combining a close crop and fine-tuning the backbone gives an F1 of 0.82, rather than 0.84. We briefly describe each technique and its impact on classification score below.

3.2.1. Multilayer perceptron

So far, classification was done using a single linear layer as a classifier. Increasing the number of fully connected layers in the classifier head lends a few advantages. Mainly, the addition of at least one hidden layer enables the classifier to learn nonlinear functions. These dense networks with one or more hidden layers are commonly referred to as MLPs. For a two-layer MLP, we observe a 5.7% increase in the F1 score of AM-RADIO and a 2.9% increase for ResNet-50.

We also investigate the use of a whitening layer (Kalapos & Gyires-Tóth 2024), which scales the frozen features before input to the linear classifier layer. Effectively, this is a two-layer MLP with the weights and biases of the input layer pre-calculated. This leads to a 2.2% increase in the F1 score of AM-RADIO and a 1.5% increase for ResNet-50.

3.2.2. Bigger model

In general, performance increases with the number of parameters in a model (e.g., Zhai et al. 2022). AM-RADIO has pretrained two models larger than 86M parameter ViT-Base: ViT-Large with 307M parameters, and ViT-Huge with 632M parameters. Larger ResNet variants are ResNet-101, with 42.8M parameters, and ResNet-152, with 58.5M parameters, all pretrained for ImageNet classification.

Using features extracted from both these models, we observe an increase in the F1 score of AM-RADIO of 8.1% (ViT-Large) and 10.7% (ViT-Huge). ResNets with more parameters did not lead to an improvement, but rather the opposite; the F1 score decreased by 5% when using features extracted from ResNet-101, and by about the same amount for ResNet-152.

3.2.3. Custom data preprocessing

As is the case for most scientific images, optical and radio astronomy data is not delivered straight from the telescopes. Raw data must be calibrated, to account for instrumental or environmental effects as much as possible, and sometimes further processed

as in the case of imaging the visibility data collected by radio arrays. Much more so than the natural image domain, the data reduction pipeline has a large impact on the capabilities of the resulting dataset.

In the case of RGZ images, a number of steps have taken place before the final images are incorporated into the dataset. First, radio sources were identified from a survey catalog, which is itself the result of a source-finding algorithm. Next, various filters such as the signal-to-noise ratio determined if the image source was kept in the sample or not. Were archival images not available, the initial image would have had to be reconstructed. Experts then had to decide on a cutout size and how to scale each image, which consisted of float values spanning five or more orders of magnitude, before finally converting the image to a PNG. Often, both image reconstruction and scaling favor a certain type of radio emission over another, making a dataset more suited to certain scientific goals or DSTs than others.

The final product of most astrophysics image reduction pipelines are 16- to 32-bit FITS images in physical units, rather than 8-bit PNG scaled between 0 and 255. As only PNG images were available in the GMNIST and RGZ datasets, we can only examine preprocessing methods that operate on those images, which have already lost some detail. One way in which the data has been adapted to fit the model was already mentioned previously; the images had to be re-sized in order to be accepted as input to most models with a ViT-Base backbone. CNNs do not have this input size requirement of 224×224 pixels, as the convolution kernel can operate on images of any size and shape. Similarly, AM-RADIO was designed to take any input image size as long as the dimensions are a multiple of 14.

We resize the RGZ images to 320, 480, and 512 pixels square using bilinear interpolation. Interestingly, increasing the image size does not necessarily increase performance; the optimal size for AM-RADIO is 480×480 , leading to a F1 score increase of 8%. As with increasing the model size, increasing image size has a detrimental effect on the performance of classification via ResNet-50.

Another preprocessing method already mentioned is making a close cutout of the radio source based on the bounding box information. Cropping to the area around the central galaxy leads to a large improvement in F1 score: a 16% increase for AM-RADIO, and a 26% increase for ResNet-50. Similar preprocessing, involving cropping images to less than 50% of their original area, was used since [Dieleman et al. \(2015\)](#) and [Zhu et al. \(2019\)](#) to classify optical galaxy images from SDSS and DeCALs dataset; more recent works such as [Wei et al. \(2024\)](#) and [Urechiatu & Frincu \(2024\)](#) also use this method to achieve F1 scores above 0.95. As is seen from the attention maps in [Figure 6](#), the close crop means that more patches of the ViT contain parts of the object to be classified, allowing the network to use information from more than just one or two patches to perform classification. Custom preprocessing allows the data to complement features the foundation model has already learned, and emphasizes the importance of domain-specific expert knowledge in the development of machine learning pipelines.

3.2.4. Full fine-tuning of the backbone model

Full fine-tuning involves unfreezing parameters the backbone network and allowing them to update. It can be done layer by layer or block by block, but for this experiment we simply unfroze the entire backbone after ten epochs, which was used to allow the newly initialized classifier layer to acclimate.

When fine-tuning the backbone, we observe an increase in the F1 score of AM-RADIO of 22% and 44% for ResNet-50. This is the largest improvement in performance by any of the optimization techniques we investigated.

3.2.5. Fine-tuning the backbone model using LoRA

Low-Rank Adaptation, or LoRA, is a parameter-efficient fine-tuning method. Instead of directly updating model weights, LoRA tracks changes to the weights and then decomposes the large matrix of weight changes into smaller matrices. These low-rank matrices are multiplied to approximate the full matrix of weight changes, allowing for fast and efficient fine-tuning. In practice, it is not much more difficult to implement than full fine-tuning, although there are additional hyperparameters to be optimized.

Using LoRA leads to a 17.3% increase in the F1-score with AM-RADIO as a backbone, and a 31.8% increase when using ResNet-50. These are substantial improvements that result from tuning less than 5% of the model parameters.

4. Source detection

Source detection in both optical and radio images is largely performed analytically. Work using deep learning to replace these time-consuming analytic pipelines has been done by [Wu et al. \(2019\)](#); [Farias et al. \(2020\)](#); [Burke et al. \(2019\)](#); [Riggi et al. \(2023\)](#), mainly through application of Faster- or Mask-RCNN. [Sortino et al. \(2023\)](#) provides an overview of both CNN and ViT-based methods, including YOLO. More recent results based on simulated radio data include [Taran et al. \(2023\)](#) and [Cornu et al. \(2024\)](#), which both surpass the capabilities of analytic source detection.

Here we performed source detection 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 that are not labeled.

As was done for classification, images were upsampled 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 was 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. Otherwise, we kept the default training data augmentations: a randomly applied blur, contrast adjustment, and color jitter.

Our first experiment kept the backbones frozen, allowing only Faster-RCNN’s detection head layers (FPN, RPN, and ROI) to train.

Generally, a learning rate of 0.001 was used, with the exception of training ViT-Base on MGCLS from scratch, and both frozen and fine-tuned MAE on MGCLS, where we used a learning rate of 0.0005 in order to ensure a smooth decay of the training loss. Training was for 100 epochs, with a batch size of 16, performed on 1–2 GPUs depending on availability. Due to training requiring a number of hours to complete, we did not repeat training runs in order to report a range in our chosen metric. We report the mean average precision at an

Table 5. Source detection mAP@50 for the RGZ and MGCLS datasets.

Backbone	RGZ mAP@50		MGCLS mAP@50	
	frozen	Fine-tuned	frozen	Fine-tuned
ViT-Base supervised	–	0.51	–	0.885
MSN	0.70 (+0.19)	0.71 (+0.20)	0.878 (–0.006)	0.891 (+0.007)
DINOv1	0.63 (+0.12)	0.73 (+0.22)	0.886 (+0.001)	0.896 (+0.011)
MAE	0.51 (+0.00)	0.73 (+0.22)	0.867 (–0.018)	0.873 (+0.005)
SigLIP	0.46 (–0.05)	0.679 (+0.17)	0.872 (–0.013)	0.878 (–0.007)
AM-RADIO	0.47 (–0.04)	0.605 (+0.09)	0.867 (–0.018)	0.885
ResNet 50 supervised	–	0.654	–	0.873
ResNet 50	0.627 (–0.027)	0.714 (+0.06)	0.849 (–0.024)	0.873 (+0.00)
ResNet 18 supervised	–	0.641	–	0.511
ResNet 18	0.311 (–0.33)	0.644 (+0.003)	0.020 (–0.490)	0.511 (+0.00)

Notes. Results for frozen and fine-tuned backbones are compared with supervised baselines.

intersection-over-union (IOU) threshold of 50%, known as the mAP@50, in Table 5. This metric considers a prediction correct if the overlap between the predicted and actual object bounding boxes is at least 50%, and is commonly used to evaluate object detection tasks. Depending on the scientific objective, a different IoU threshold might be preferred; for reference, the mean average precision is calculated at different IoU thresholds in Appendix C.

Results showed that MSN and DINO are the best foundation models for this task. They outperformed training from scratch, even when the backbone was frozen, except for MSN on MGCLS. Transformers pretrained using multimodal datasets (SigLIP and AM-RADIO) performed worse than the other ViTs, including MAE, on RGZ, and slightly better on MGCLS, although not as good as DINO and MSN. This is an interesting result, as both SigLIP and AM-RADIO were the best at classification. Representations learned through text-image pairs appear to be less relevant to object detection in images than they are to classification.

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 ResNet-18. 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 1–6 sources in RGZ. Additionally, bounding boxes in RGZ can be quite large, including a lot of noise and making them harder to predict accurately.

These results show that transfer learning, starting with the pretrained weights of a foundation model, is a good practice, as the right choice of model leads to better results even using a frozen backbone than training from scratch, and fine-tuning brings further improvements. 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 DST.

5. Best practice guidelines

We do not claim that the two datasets and two tasks examined here are completely representative of the many applications of deep learning for astrophysics. The lessons learned from

considering various methods in using foundation models in these circumstances may be useful to practitioners, so we gather the results here into a general guideline.

As always, the dataset type – characteristics of the images and type and number of available labels – has the most impact on DST performance. These properties will inform the choice of foundation model, which in turn will determine how the dataset should be customized to complement the model itself. While GMNIST images are complementary enough to ImageNet and similar pretrained backbones, a dataset such as RGZ might benefit from a model with different pretraining. Models pretrained on images with strong background patterns and textures, such as satellite or thermal imagery, or perhaps images from medical scans or microscopy could be good avenues to investigate. The labeling scheme of the dataset is also something to consider; the RGZ labels, while sensible from a scientific point of view, is not well suited to machine learning, as information about the number of radio components is obscured during scaling and conversion of the original images to PNG format.

Aside from the selection of the foundation model for the backbone, there is also the choice of projector head for the DST. Clearly, some projector heads will have to be specifically designed for the task, such as Faster-RCNN for object detection. Other possibilities for detection and segmentation include YOLO, which is one of the best performing architectures on natural images, and can be used with a ResNet backbone. From a practical perspective, the choice of projector might also limit the choice of backbone; YOLO’s latest versions can be adapted for ResNet backbones but not yet for ViTs. For simpler tasks such as classification, the projector head is less important; we found that modifying it with an extra hidden layer or whitening layer only brought improvements of at most 5%.

Fine-tuning the model, once selected, yields the most significant improvements. Model and dataset size determine the necessary time and resources; with many astrophysics datasets having fewer than 10K images, fine-tuning a model of a few million parameters should be feasible. An alternative to full fine-tuning is to only un-freeze the upper layers or blocks of a network. If computational resources are very limited, LoRA is an alternative that will generally be a few percent less accurate than full fine-tuning.

If fine-tuning does not reach the required performance metric, or if one wishes to generally improve the system as much as possible, one should carefully reconsider the data

preprocessing. Together with custom dataset assembly and labeling, this is the most labor-intensive step and requires the most specialized domain knowledge – however it will also have the biggest impact on performance.

For example, the simple task of resizing images to match the input dimensions of the model is often necessary, but can be done optimally. We saw that resizing images can boost performance by up to 8%, but that bigger was not necessarily better and was in fact worse when using ResNet-50. Furthermore, our example of RGZ classification showed that cropping images to focus on regions of interest was the next-best thing to fine-tuning. Considering the earlier discussion of the RGZ labeling scheme in Section 3.1, we also speculate that using different scalings of the original FITS images to include information about not only bright peaks but also flux islands could significantly improve distinction between visually similar classes such as 2C-2P and 2C-3P (see Figure 1 for example). One technique seen in the literature (Aniyān & Thorat 2017, used by Maslej-Krešňáková et al. (2021) for example) is sigma-clipping, where noise is removed by setting all pixels below a certain threshold to zero.

Note that in this work we do not discuss data augmentation, which is normally done during training to help the network learn more robust representations. Several works address augmentation schemes for both optical and radio images; Hayat et al. (2021) appendix D for optical, and Slijepcevic et al. (2024) section 3.9 or Maslej-Krešňáková et al. (2021) section 4.2 for radio.

These general principles can help guide effective application of vision foundation models to astrophysical image data. Experimentation is unlikely to be a linear process; practitioners may need to iteratively revisit choices of foundation models, dataset preparation, and task-specific projector heads in order to achieve the desired result.

6. Conclusions

Our results demonstrate that state-of-the-art vision foundation models can immediately classify optical galaxies and perform source detection of radio galaxies with F1 scores of 0.72–0.88. In the case of optical galaxy classification, all foundation models outperformed supervised training. F1 scores close to 0.87 with only 800 labeled training images (10% of the total dataset) show that foundation models can be used to great effect for early-stage machine learning applications in scientific contexts. Of the models tested, SigLIP and AM-RADIO, pretrained on multimodal datasets, delivered the best performance, with vision transformers such as DINOv2 and MSN also excelling over ResNet-50, which is used or compared with in much of the astrophysics literature. Only MAE, which is optimized for reconstruction loss, failed to encode many task-relevant features for galaxy classification.

Dataset characteristics strongly influenced model performance. For GMNIST, the alignment between the pretraining distribution and the task dataset likely contributed to the strong classification results, as galaxies occupied a significant portion of the image with minimal interference from noise or unrelated objects. Conversely, RGZ images, with small central sources and images dominated by noise and patterns from reconstruction, presented a clear distribution shift that limited the applicability of foundation models. In this case, supervised training outperformed all foundation models except SigLIP and AM-RADIO. While models retained some relevant features, such as bright peaks, we speculated that information about the flux

islands, also relevant to classification, was lost due to the original preprocessing. Adapting dataset-specific preprocessing or label schemes could improve results, though significant domain expertise remains essential for working with radio astronomy datasets.

Source detection results further underscored the importance of aligning data to model architectures and tasks. When radio galaxy images were resized to ensure that transformer patch sizes were slightly smaller than average source dimensions, performance improved markedly. We saw that ViTs did well to detect compact sources in MGCLS, while ResNet was better at finding RGZ galaxies. Models pretrained on multimodal data usually did not perform as well as those pretrained on only images, especially when freezing the backbone. Additionally, fine-tuning improvements were inconsistent and (in the case of MSN on RGZ) much smaller than might be expected. This situation might improve given a more optimal detection head architecture, or additional hyperparameter tuning.

Despite promising results, there remains a significant gap between the performance of foundation models on scientific datasets and their state-of-the-art results on natural image benchmarks. For example, ImageNet-1k classification achieves a top-1 accuracy of 92%, and COCO object detection achieves an average mAP@50 of 0.73 across 90 classes. Closing this performance gap will require techniques that maximize results from limited labeled data and computational resources, a challenge particularly relevant in domains such as astrophysics. Semi-supervised learning frameworks, such as those proposed by Voloshynovskiy et al. (2020), offer exciting possibilities to make use of the abundance of unlabeled data to enhance fine-tuning methods.

In this study, we employed simple fine-tuning strategies, which, while effective, leave room for improvement. Future work will investigate advanced fine-tuning methods that integrate labeled and unlabeled data, potentially enabling a more robust and resource-efficient application of vision foundation models to astrophysical image analysis. This iterative and adaptive approach will be essential for bridging the gap between cutting-edge machine learning methods and the unique demands of astronomical research.

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>.

Acknowledgements. This work has been done in partnership of the Swiss SKA consortium which is funded by the State Secretariat for Education, Research and Innovation (SERI). MGCLS data products were provided by the South African Radio Astronomy Observatory and the MGCLS team and were derived from observations with the MeerKAT radio telescope. The MeerKAT telescope is operated by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation. We would like to thank Philip Denzel for the RGZ object detection dataset. OB and DP were supported by the SNF Sinergia grant CRSII5-193826 “AstroSignals: A New Window on the Universe, with the New Generation of Large Radio-Astronomy Facilities”. MD and VK were supported by the SNF Sinergia project (CRSII5-193716), “Robust deep density models for high-energy particle physics and solar flare analysis (RODEM)”.

References

- Anyan, A. K., & Thorat, K. 2017, *ApJS*, **230**, 20
- Assran, M., Caron, M., Misra, I., et al. 2022, in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI* (Berlin, Heidelberg: Springer-Verlag), 456
- Assran, M., Duval, Q., Misra, I., et al. 2023, in *2023 IEEE/CVF CVPR* (Vancouver, BC, Canada: IEEE), 15619
- Balestriero, R., & LeCun, Y. 2024, in *ICML*
- Bao, H., Dong, L., Piao, S., & Wei, F. 2022, in *ICLR*
- Bardes, A., Ponce, J., & LeCun, Y. 2022, in *ICLR*
- Becker, B., Vaccari, M., Prescott, M., & Grobler, T. 2021, *MNRAS*, **503**, 1828
- Burke, C. J., Aleo, P. D., Chen, Y.-C., et al. 2019, *MNRAS*, **490**, 3952
- Cao, J., Xu, T., Deng, Y., et al. 2024, *A&A*, **683**, A42
- Caron, M., Misra, I., Mairal, J., et al. 2020, in *Advances in Neural Information Processing Systems*, 33 (Curran Associates, Inc.), 9912
- Caron, M., Touvron, H., Misra, I., et al. 2021, in *2021 IEEE/CVF ICCV*, 9630
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. 2020, in *International conference on machine learning (PMLR)*, 1597
- Chen, T., Bianco, M., Tolley, E., et al. 2024a, *MNRAS*, **532**, 2615
- Chen, X., Ding, M., Wang, X., et al. 2024b, *Int. J. Comput. Vis.*, **132**, 208
- Cohen, M., & Lu, W. 2021, *Astron. Comput.*, **37**, 100507
- Cornu, D., Salomé, P., Semelin, B., et al. 2024, *A&A*, **690**, A211
- Dagli, R. 2023, *Astroformer: More Data might not be all you need, Learning to Predict Galaxy Morphologies with Limited Data*
- Deng, J., Dong, W., Socher, R., et al. 2009, in *2009 IEEE Conference on CVPR*, 248
- Dieleman, S., Willett, K. W., & Dambre, J. 2015, *MNRAS*, **450**, 1441
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. 2020, in *Proceedings of the Ninth International Conference on Learning Representations*
- Drozdzova, M., Kinakh, V., Bait, O., et al. 2023, *A&A*, **683**
- Fariás, H., Ortiz, D., Damke, G., Jaque Arancibia, M., & Solar, M. 2020, *Astron. Comput.*, **33**, 100420
- Grill, J.-B., Strub, F., Althé, F., et al. 2020, in *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20* (Red Hook, NY, USA: Curran Associates Inc.), 21271
- Ha, D., & Schmidhuber, J. 2018, in *Advances in Neural Information Processing Systems*, 31 (Curran Associates, Inc.)
- Hausen, R., & Robertson, B. 2022, in *ML4PS Workshop (arXiv)*, [arXiv:2201.04714] [astro-ph]
- Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., & Mustafa, M. 2021, *ApJ*, **911**, L33
- He, K., Zhang, X., Ren, S., & Sun, J. 2016, in *2016 IEEE Conference on CVPR*, 770
- He, K., Chen, X., Xie, S., et al. 2022, in *2022 IEEE/CVF Conference on CVPR*, 15979
- Hui, W., Robert Jia, Z., Li, H., & Wang, Z. 2022, *J. Phys.: Conf. Ser.*, **2402**, 012009
- Jia, C., Yang, Y., Xia, Y., et al. 2021, in *International conference on machine learning (PMLR)*, 4904
- Jia, P., Zheng, Y., Wang, M., & Yang, Z. 2023, *Astron. Comput.*, **42**, 100687
- Kalapos, A., & Gyires-Tóth, B. 2024, in *2024 ICMLA*, 448
- Kalvankar, S., Pandit, H., & Parwate, P. 2021, *Galaxy Morphology Classification using EfficientNet Architectures*, [arXiv:2008.13611] [cs]
- Kirillov, A., Mintun, E., Ravi, N., et al. 2023, in *Proceedings of the IEEE/CVF ICCV*, 4015
- Kumar, R., Sarker, M. K., & Islam, S. R. 2023, in *Deep Learning Theory and Applications*, eds. D. Conte, A. Fred, O. Gusikhin, & C. Sansone (Cham: Springer Nature Switzerland), 115
- Lastufka, E., Bait, O., Taran, O., et al. 2024, *A&A*, **690**, A310
- Li, Z., Yu, C., Xiao, J., Long, M., & Cui, C. 2021, *Astron. Comput.*, **36**, 100482
- Li, L. H., Zhang, P., Zhang, H., et al. 2022a, in *2022 IEEE/CVF Conference on CVPR* (New Orleans, LA, USA: IEEE), 10955
- Li, Y., Mao, H., Girshick, R., & He, K. 2022b, in *European Conference on Computer Vision* (Springer), 280
- Lin, J. Y.-Y., Liao, S.-M., Huang, H.-J., Kuo, W.-T., & Ou, O. H.-M. 2022, in *ML4PS Workshop (arXiv)*, [arXiv:2110.01024]
- Lochner, M., & Bassett, B. A. 2021, *Astron. Comput.*, **36**, 100481
- Marcel, S., & Rodriguez, Y. 2010, in *Proceedings of the 18th ACM international conference on Multimedia, MM'10* (New York, NY, USA: Association for Computing Machinery), 1485
- Maslej-Krešňáková, V., El Boucheffry, K., & Butka, P. 2021, *MNRAS*, **505**, 1464
- Merz, G., Liu, Y., Burke, C. J., et al. 2023, *MNRAS*, **526**, 1122
- Müller, S. G., & Hutter, F. 2021, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 774
- Ndung'u, S., Grobler, T., Wijnholds, S. J., Karastoyanova, D., & Azzopardi, G. 2023, *New Astron. Rev.*, **97**, 101685
- Oquab, M., Darcet, T., Moutakanni, T., et al. 2024, *Trans. Mach. Learn. Res.*
- Pandya, S., Patel, P., O, F., & Blazek, J. 2023, in *ML4PS Workshop (arXiv)*, [arXiv:2311.01500]
- Porter, F. A. M., & Scaife, A. M. M. 2023, *RAS Tech. Instrum.*, **2**, 293
- Radford, A., Kim, J. W., Hallacy, C., et al. 2021, in *International conference on machine learning (PMLR)*, 8748
- Ranzinger, M., Heinrich, G., Kautz, J., & Molchanov, P. 2024, in *Proceedings of the IEEE/CVF Conference on CVPR*, 12490
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. 2016, in *2016 IEEE Conference on CVPR* (Las Vegas, NV, USA: IEEE), 779
- Reiman, D. M., & Göhre, B. E. 2019, *MNRAS*, **485**, 2617
- Ren, S., He, K., Girshick, R., & Sun, J. 2017, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39**, 1137
- Riggi, S., Magro, D., Sortino, R., et al. 2023, *Astron. Comput.*, **42**, 100682
- Schmidt, K., Geyer, F., Fröse, S., et al. 2022, *A&A*, **664**, A134
- Slijepcevic, I. V., Scaife, A. M. M., Walmsley, M., et al. 2024, *RAS Tech. Instrum.*, **3**, 19
- Sortino, R., Magro, D., Fiameni, G., et al. 2023, *Exp. Astron.*, [arXiv:2303.04506] [cs]
- Taran, O., Bait, O., Dessauges-Zavadsky, M., et al. 2023, *A&A*, **674**, A161
- Tishby, N., Pereira, F. C., & Bialek, W. 2000, in *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, (arXiv), 368, [arXiv:physics/0004057]
- Urechiatu, R., & Frincu, M. 2024, *Universe*, **10**, 230
- Vafaei Sadr, A., Vos, E. E., Bassett, B. A., et al. 2019, *MNRAS*, **484**, 2793
- Villar, V. A., Cranmer, M., Berger, E., et al. 2021, *ApJS*, **255**, 24
- Voloshynovskiy, S., Taran, O., Kondah, M., Holotyak, T., & Rezende, D. 2020, *Entropy*, **22**, 943
- Vos, E. E., Francois Luus, P. S., Finlay, C. J., & Bassett, B. A. 2019, in *2019 IEEE 29th International Workshop on MLSP*, 1
- Walmsley, M., Lintott, C., Géron, T., et al. 2022, *MNRAS*, **509**, 3966
- Wei, S., Lu, W., Dai, W., et al. 2024, *AJ*, **167**, 29
- Wong, O. I., Garon, A., Alger, M., et al. 2024, *MNRAS*, **stae2790**
- Wu, C., Wong, O. I., Rudnick, L., et al. 2019, *MNRAS*, **482**, 1211
- Yu, J., Wang, Z., Vasudevan, V., et al. 2022, *Trans. Mach. Learn. Res.*
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. 2021, in *Proceedings of the 38th ICML (PMLR)*, 12310
- Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. 2022, in *Proceedings of the IEEE/CVF CVPR*, 12104
- Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. 2023, in *Proceedings of the IEEE/CVF ICCV*, 11975
- Zhou, J., Wei, C., Wang, H., et al. 2022, in *ICLR*
- Zhu, X.-P., Dai, J.-M., Bian, C.-J., et al. 2019, *Astrophys. Space Sci.*, **364**, 55
- Zhou, X., Gong, Y., Deng, F., et al. 2023, *MNRAS*, **521**, 278

Appendix A: Attention visualization

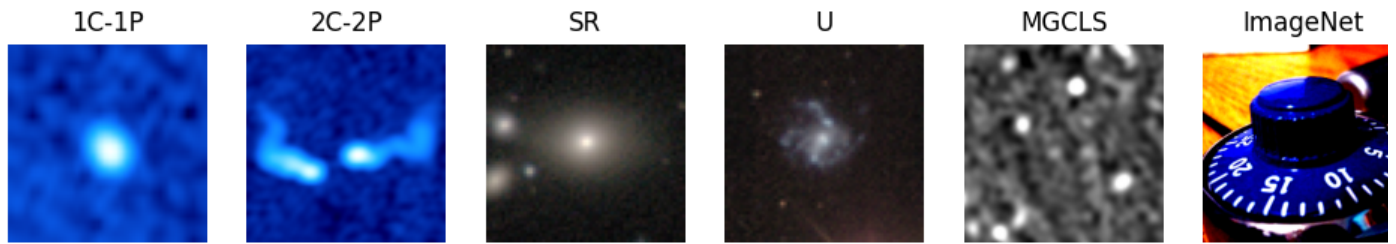


Fig. A.1. Images for the following attention maps

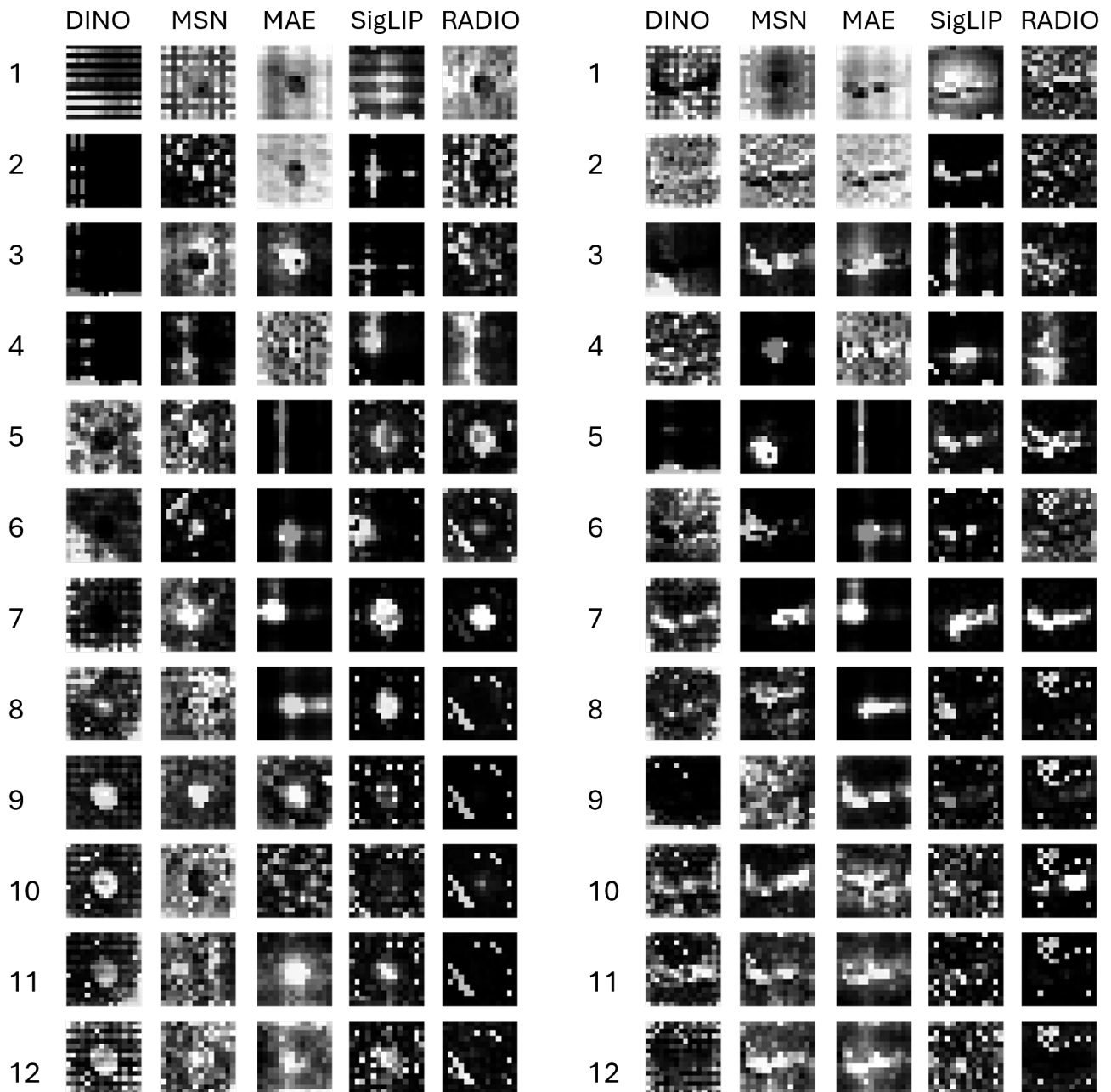


Fig. A.2. Attention maps for all models, RGZ compact source (left, labeled in the dataset as 1C-1P) and extended source (right, labeled in the dataset as 2C-2P), shown for each model and each attention block.

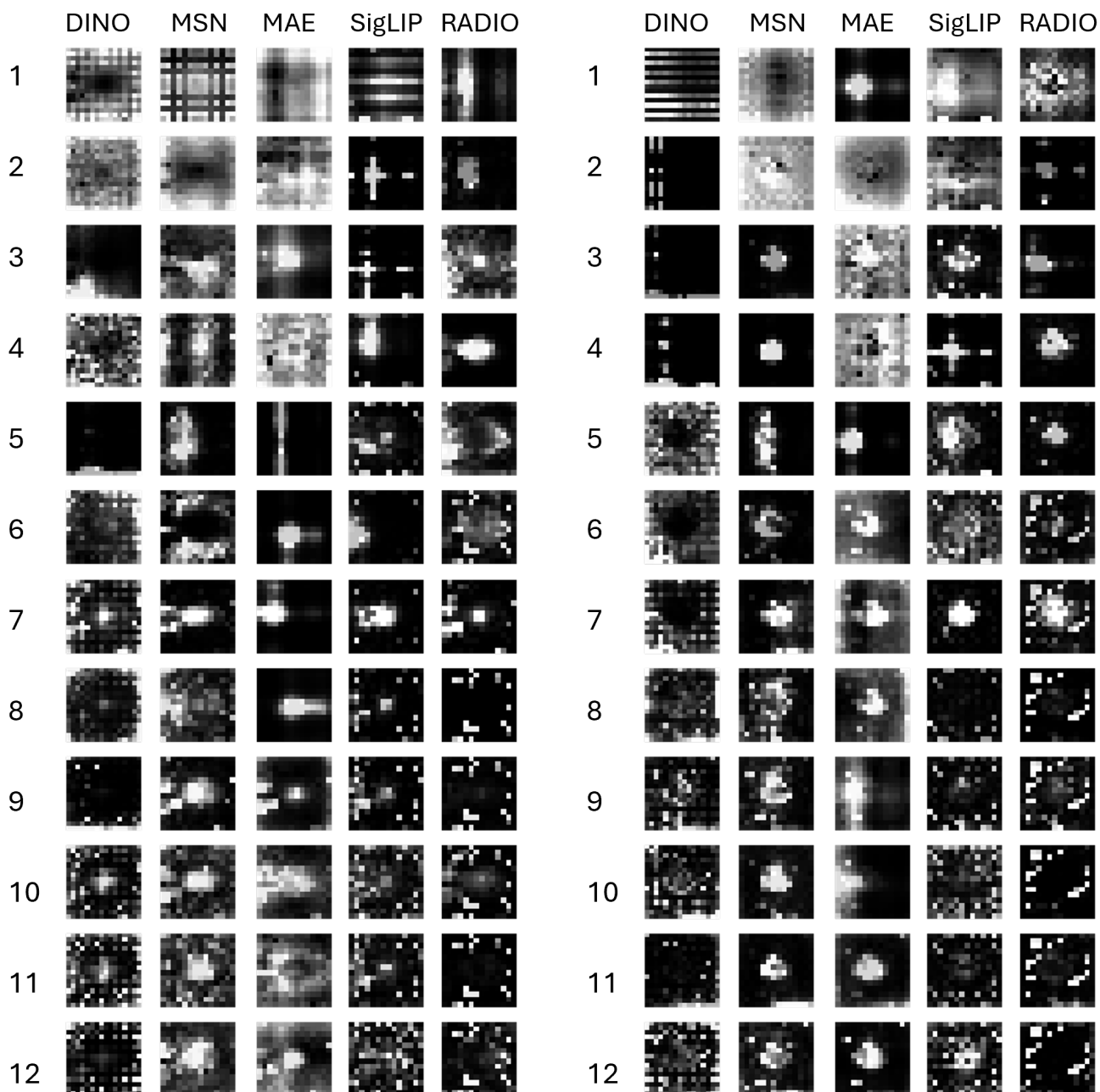


Fig. A.3. Attention maps for all models, GMNIST SR galaxy (left) and unbarred spiral (right), shown for each model and each attention block.

We visualize the attention maps for sample images from each dataset, which are shown in Figure A.1. One map per attention block is shown; ViT-Base, the backbone architecture for all Vision Transformers used in this study, has twelve blocks.

Maps of blocks with higher numbers, which typically capture high-level semantic relationships as opposed to earlier blocks which focus on more low-level, localized features, do not always display meaningful features when calculated for astrophysical images (Figures A.2, A.3 and A.4 left side). This is especially evident with SigLIP and AM-RADIO. Models pretrained on natural images only (DINOv2, MSN, and MAE), however, do display the typical behavior described above in attention maps calculated for an ImageNet example (Figure A.4, right side). Examination of the attention maps therefore illustrates that the semantic relationships, both local and global, learned from natural image pretraining do not always have meaningful counterparts in astrophysical images.

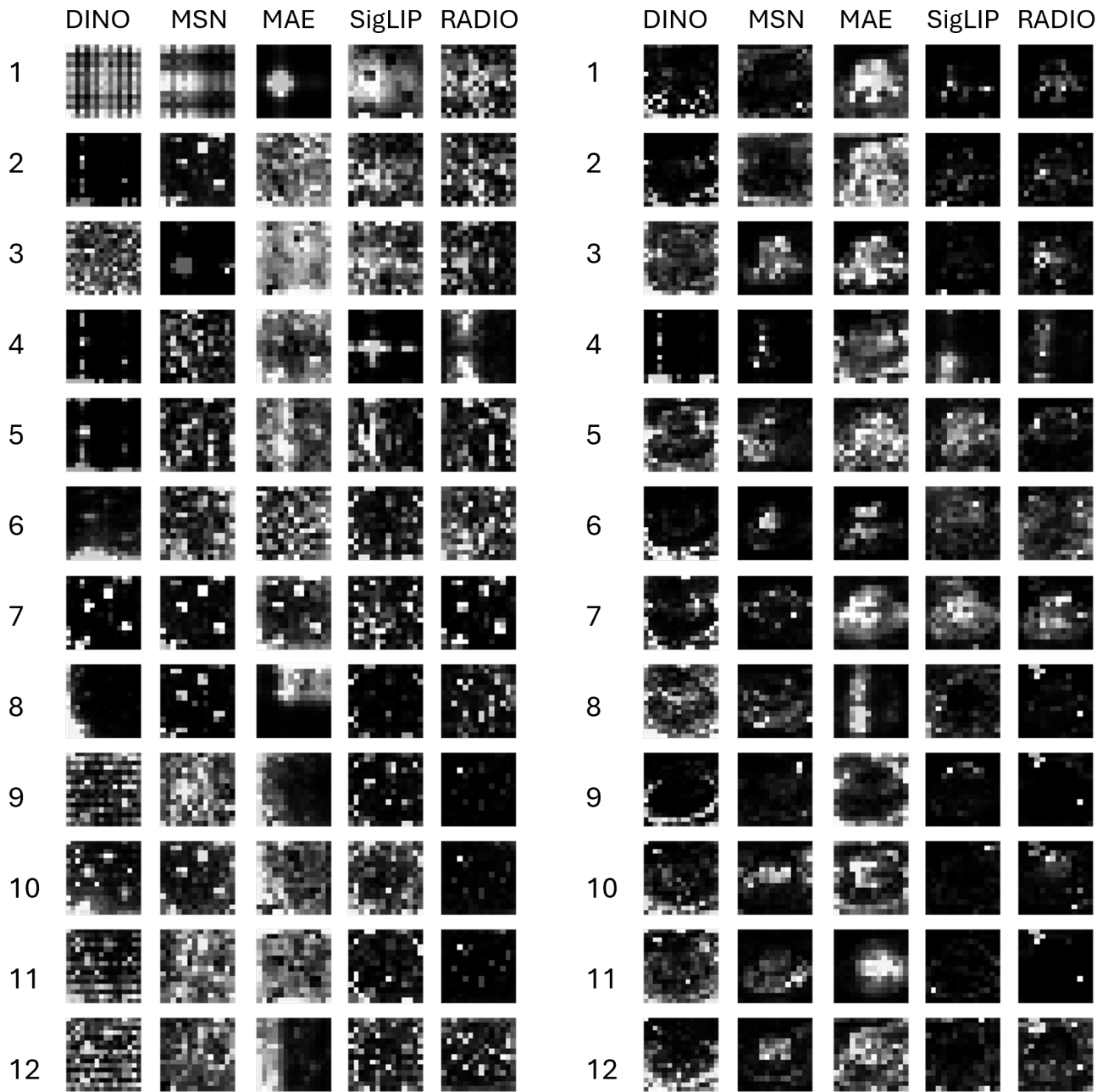


Fig. A.4. Attention maps for all models, MGCLS multi-source crop (left). ImageNet (right), shown for each model and each attention block.

Appendix B: Classification precision and recall

For completeness, Figures B.1 and B.2 show precision and recall for classification on GMNIST and RGZ, respectively.

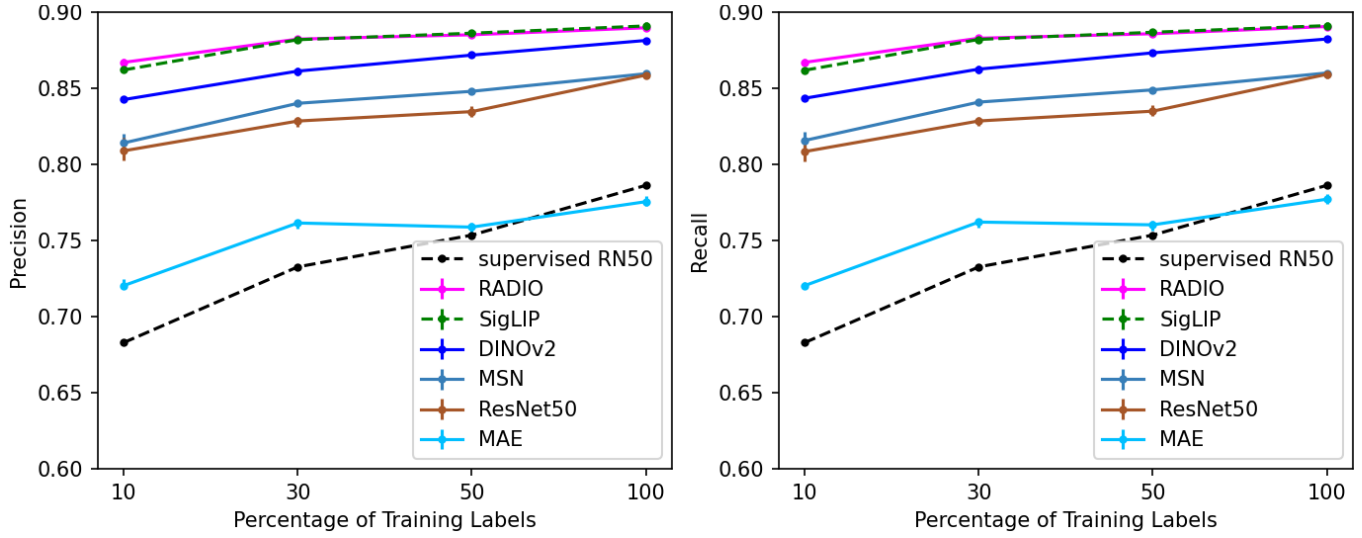


Fig. B.1. Precision (left) and recall (right) for optical galaxy morphology classification using GMNIST, as a function of percentage of training labels. Error bars show the maximum and minimum scores out of three different runs.

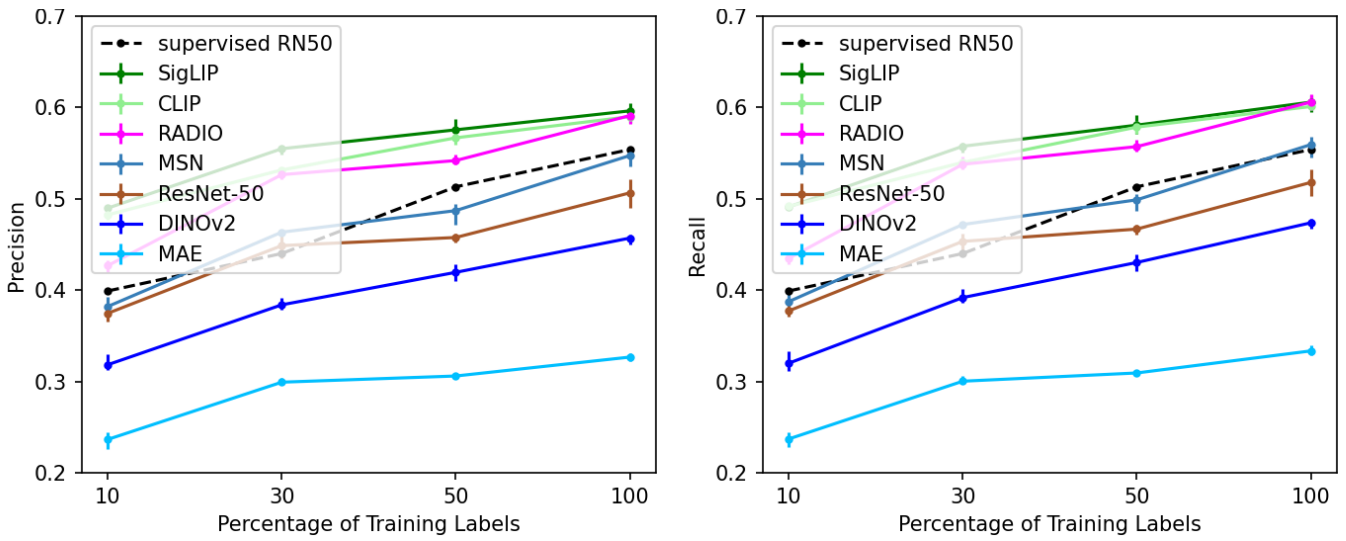


Fig. B.2. Precision (left) and recall (right) for radio galaxy morphology classification using RGZ, as a function of percentage of training labels. Error bars show the maximum and minimum scores out of three different runs.

Appendix C: Detection mAP at different IoU thresholds

Mean average precision is calculated at different IoU thresholds, ranging from 0.1 to 0.9 in steps of 0.1. The results are shown in Figure C.1, relative to the mAP obtained with the base architecture trained from scratch, whose values are tabulated in Table C.1.

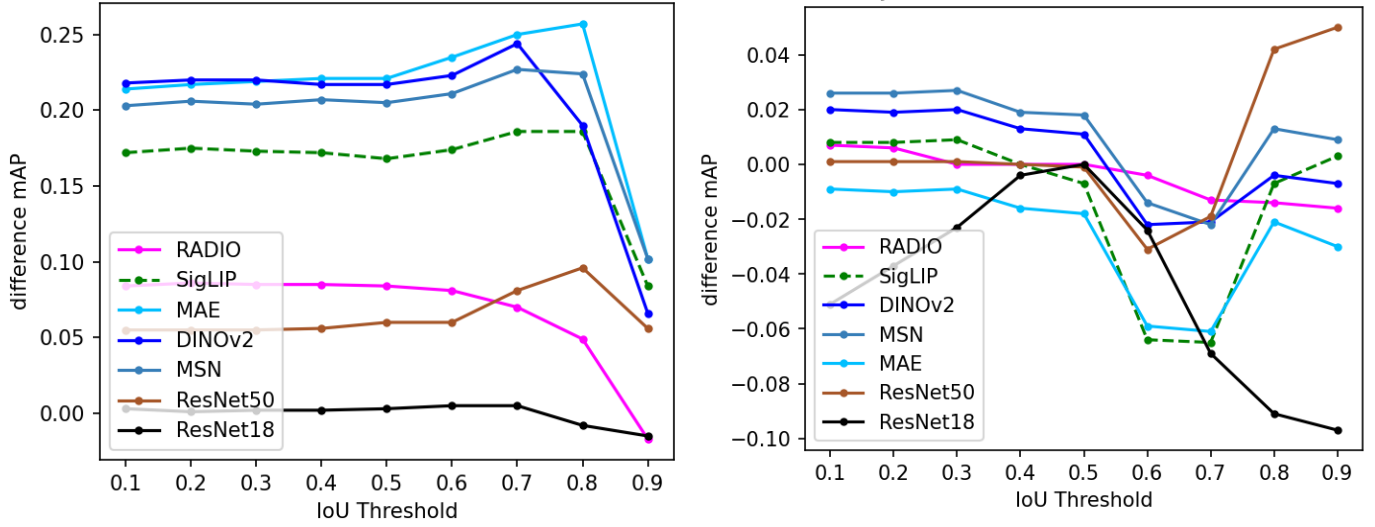


Fig. C.1. The difference in mean average precision (mAP), compared to that obtained with the base architecture trained from scratch, calculated at different intersection-over-union (IoU) thresholds. Results are shown for RGZ object detection (left) and MGCLS object detection (right).

IoU	RGZ mAP	MGCLS mAP
ViT-Base supervised		
0.1	0.522	0.901
0.2	0.519	0.901
0.3	0.517	0.900
0.4	0.514	0.0899
0.5	0.510	0.885
0.6	0.492	0.792
0.7	0.438	0.437
0.8	0.318	0.095
0.9	0.079	0.018
ResNet 50 supervised		
0.1	0.662	0.888
0.2	0.661	0.888
0.3	0.661	0.888
0.4	0.659	0.887
0.5	0.654	0.874
0.6	0.645	0.808
0.7	0.598	0.469
0.8	0.480	0.096
0.9	0.142	0.012
ResNet 18 supervised		
0.1	0.658	0.720
0.2	0.657	0.696
0.3	0.655	0.657
0.4	0.651	0.601
0.5	0.641	0.511
0.6	0.614	0.343
0.7	0.555	0.140
0.8	0.415	0.021
0.9	0.108	0.011

Table C.1. Mean average precision (mAP) calculated at different intersection-over-union (IoU) thresholds for base architecture networks trained from scratch on either RGZ or MGCLS.