UNIVERSITÉ DE GENÈVE

Examining Vision Foundation Models for Optical and Radio Astronomy Applications

E. Lastufka

M. Audard, O. Bait, M. Dessauges-Zavadsky, M. Drozdova, T. Holotyak, V. Kinakh, D. Piras, D. Schaerer, S. Voloshynovskiy

SKACH

What are vision foundation models?

- General-purpose models trained on a large amount (millions to billions) of natural images
- Capable of performing or being fine-tuned to perform multiple diverse tasks

What can VFMs be used for?

Machine Learning task	Astrophysics task	Examples
Image reconstruction	Fourier image reconstruction	Schmidt 2020, Drozdova 2023
Object detection	Source detection	Vafaei Sadr 2019, Jia 2023
Object segmentation	Source characterization	Farias 2020, Sortina 2023
Image or object classification	Source classification	Burke 2019, Riggi 2023
Instance segmentation	Instance segmentation	Hausen & Robertson 2022
Anomaly detection	Object/event discovery	Lochner & Bassett 2021

What can VFMs be used for?

Machine Learning task	Astrophysics task	Examples
Image reconstruction	Fourier image reconstruction	Schm…
Object detection	Source detection	
Object segmentation	So…	…20, Sortina 2023
Image or object cl…		Burke 2019, Riggi 2023
Instance	…segmentation	Hausen & Robertson 2022
Anomaly d…	Object/event discovery	Lochner & Bassett 2021

All these tasks can be facilitated using VFMs!

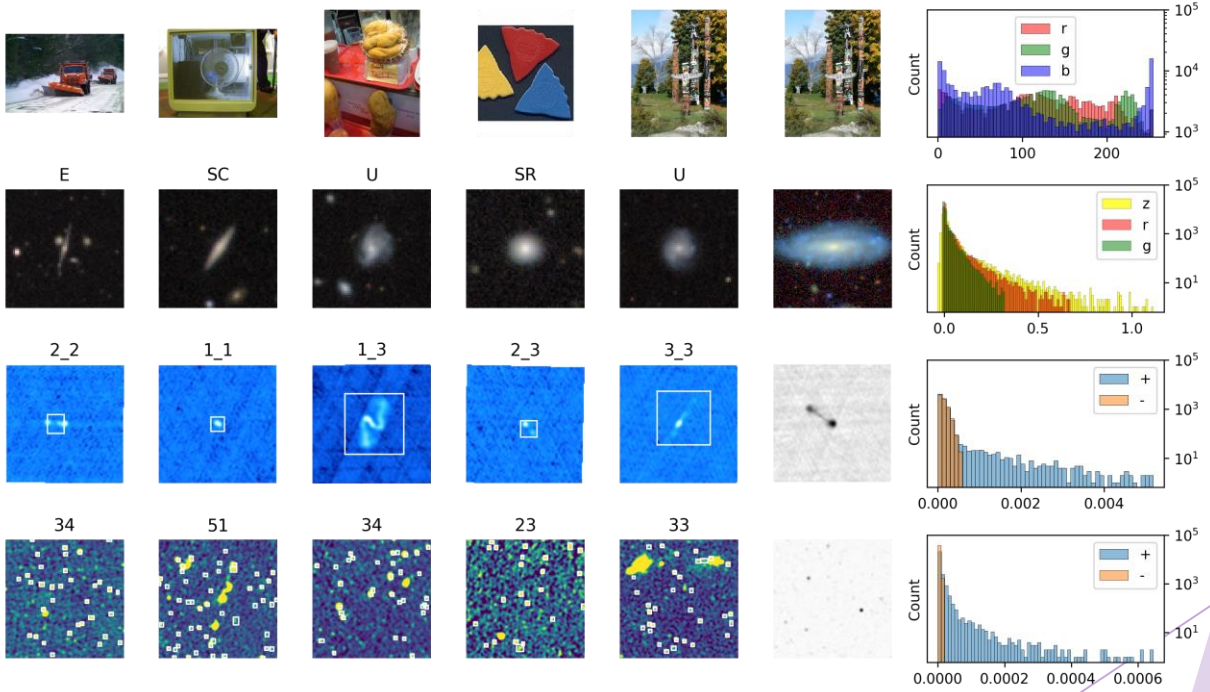
What can VFMs be used for?

Machine Learning task	Astrophysics task	Examples
Image reconstruction	Fourier image reconstruction	Schm…
Object detection	Source detection	
Object segmentation	So…	…20, Sortina 2023
Image or object cl…		Burke 2019, Riggi 2023
Instance	…segmentation	Hausen & Robertson 2022
Anomaly d…	Object/event discovery	Lochner & Bassett 2021

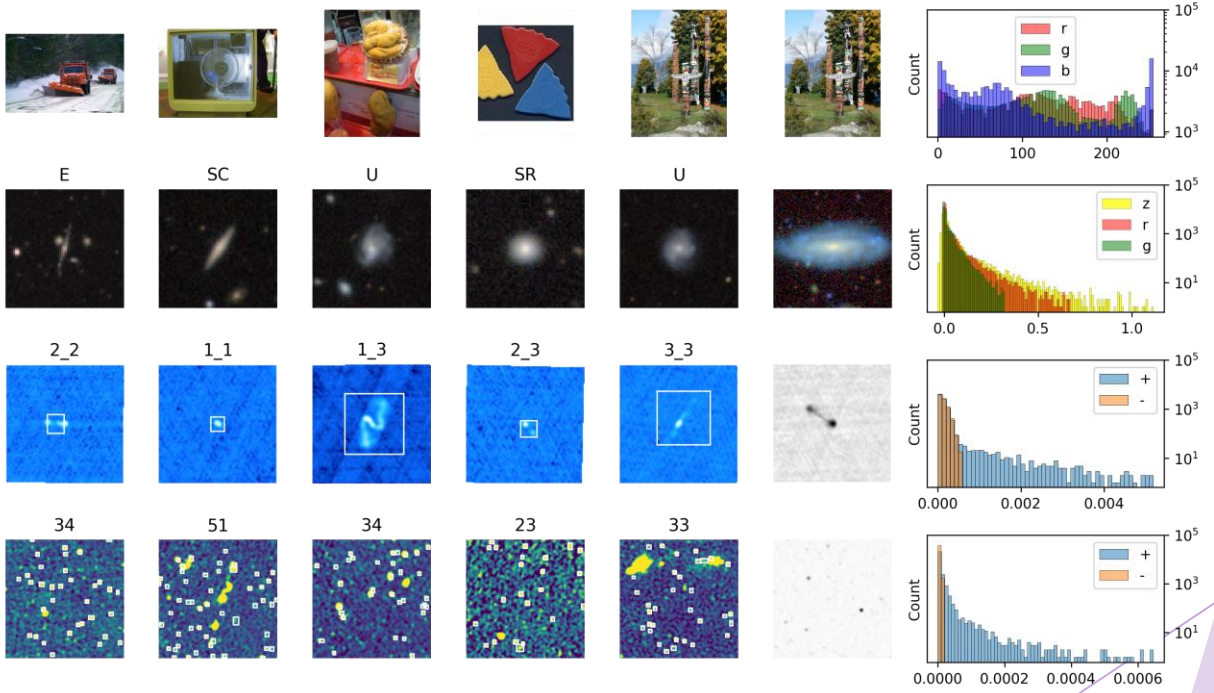
All these tasks can be facilitated using VFMs!

… but hardly anyone does it

Our data is just too different…



Our data is just too different…



… a problem known in computer science as “distribution shift”

Are standard VFMs useful for astrophysics?

Yes*

*short answer

Types of VFMs

- Many factors contribute to differences between VFMs! These include: training data, architecture, training objective

Self-supervised

Trained on a single modality (images)

Different types of losses possible

Weakly-supervised

Trained using multiple modalities (ex: images + text)

Contrastive loss

Distillation

Agglomerate representations from many existing models

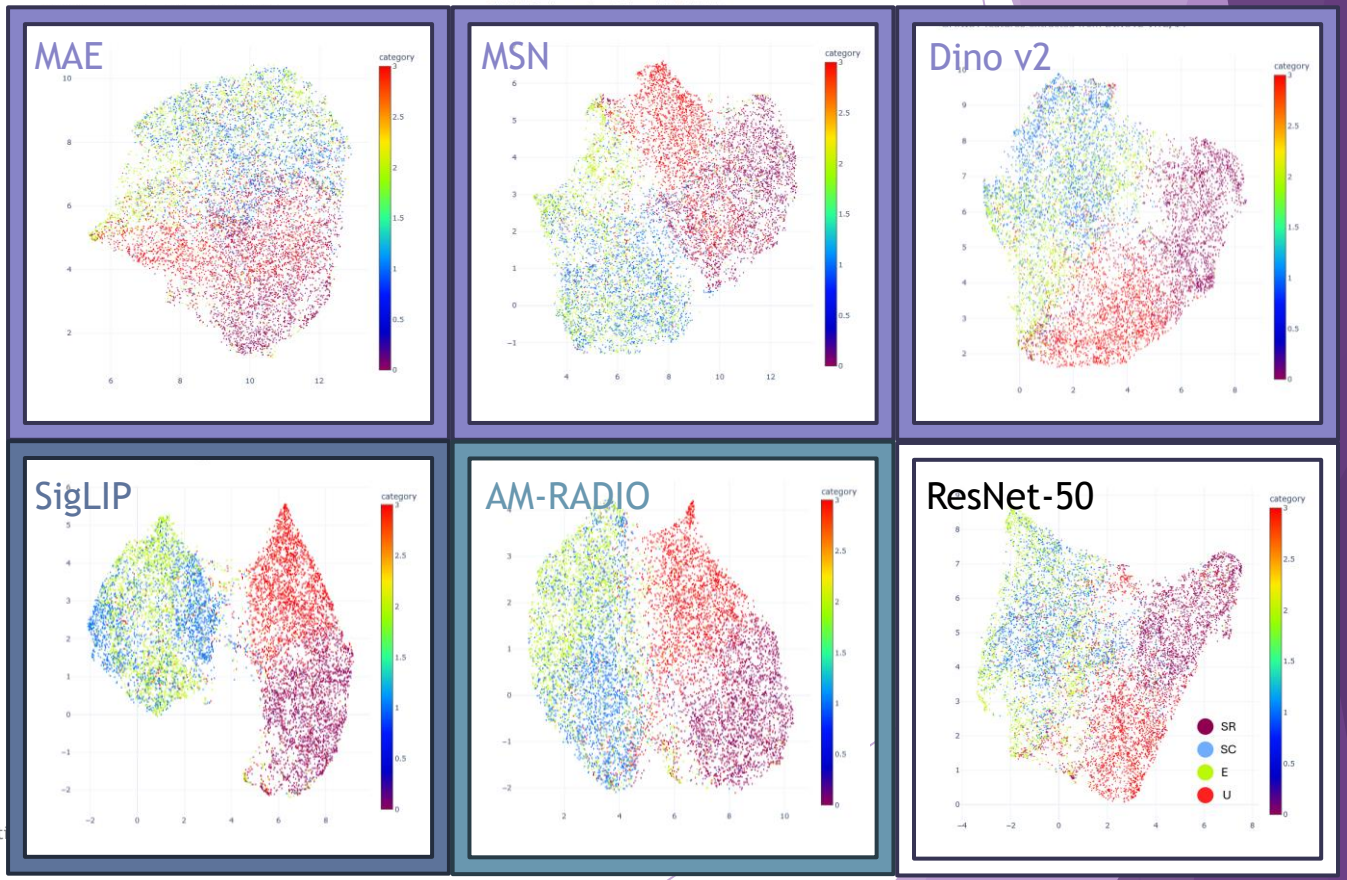
Various weighting methods

VFMs that we studied

Model	Architecture	Parameters	Pre-training Dataset	Training Objective
MAE	ViT-Base, 16x16	86M	ImageNet-1k	Reconstruction
MSN	ViT-Base, 16x16	86M	ImageNet-1k	Predict masked patches
DINOv2	ViT-Base, 14x14	86M	LVD-142M	Local-to-global
ResNet-50	ResNet-50	25.6M	ImageNet-1k	Image classification
ResNet-18	ResNet-18	11.5M	ImageNet-1k	Image classification
SigLIP	ViT-Base, 16x16	86M	Web-LI (English)	Image-text pairs
AM-RADIO	ViT-Base, 16x16	98.2M	various	Knowledge distillation

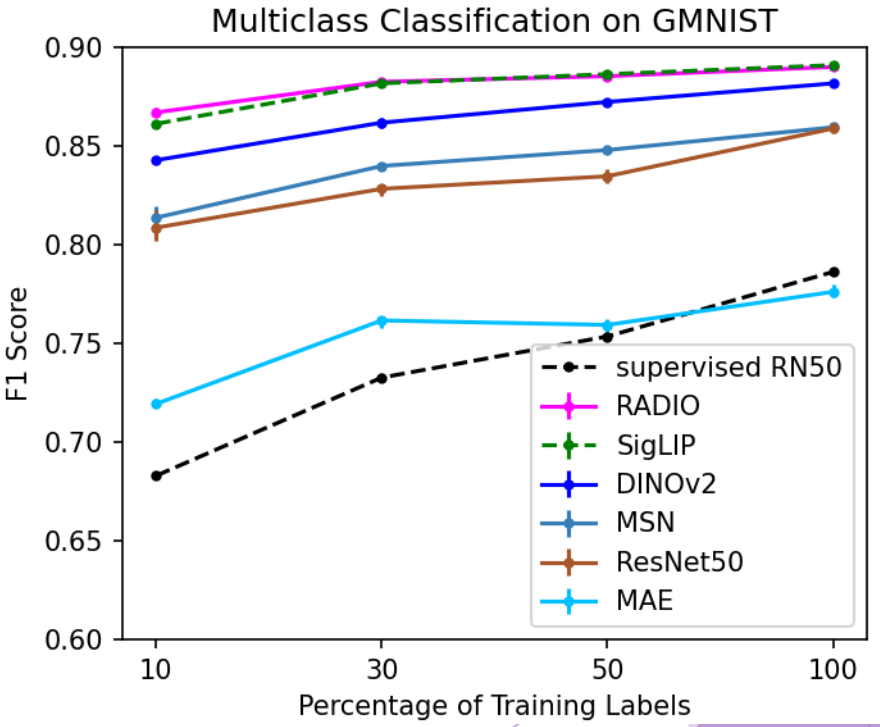
Illustration: VFMs are not all the same!

UMAP latent space representation for GalaxyMNIST dataset (optical images, 4 morphology classes)



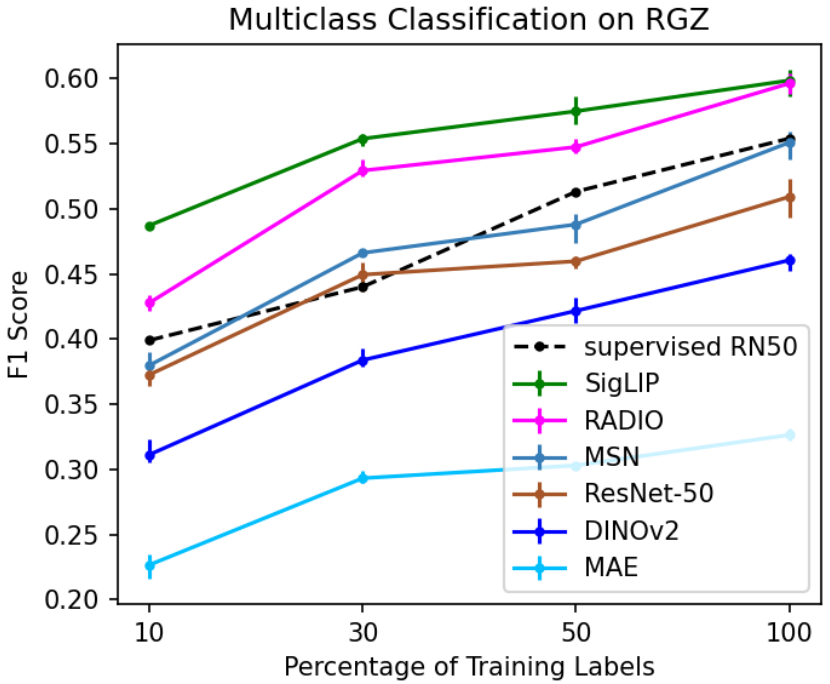
Example: Optical galaxy morphology classification

- F1 score: harmonic mean of precision and recall - higher is better
- Compare against fully supervised training (only show the best supervised result)
- Almost all VFMs out-perform supervised training, for any number of labeled examples!
- MAE's reconstruction objective is not very useful for classification

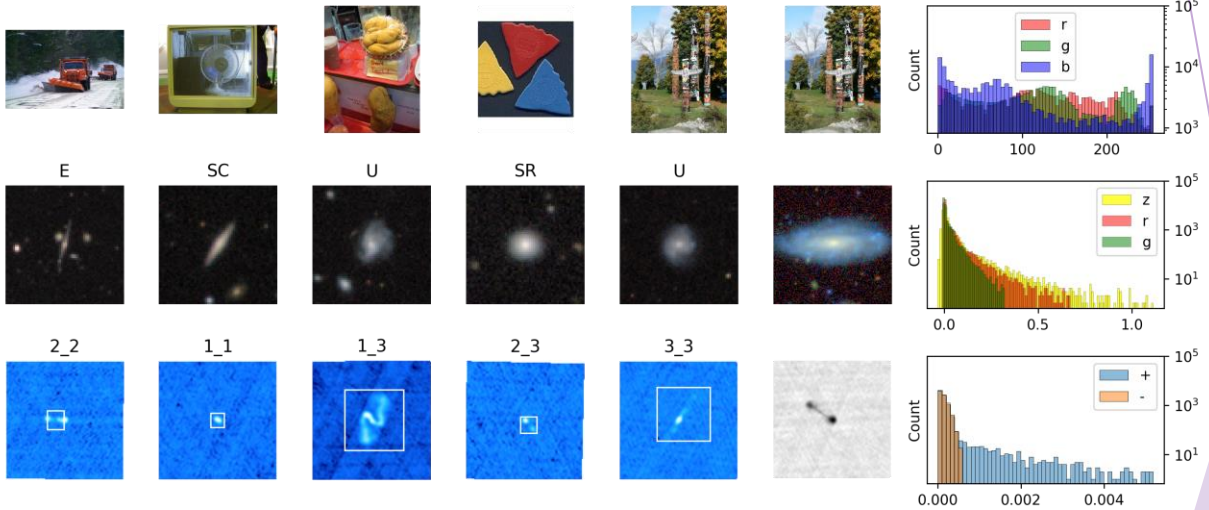


Example: Radio galaxy morphology classification

- F1 scores are much lower than for GMNIST!
- Only SigLIP and AM-RADIO regularly out-perform supervised training



Distribution shift, revisited



- Why are optical galaxies so much easier to classify?

Improving performance by mitigating distribution shift

Align the downstream task data to the training data

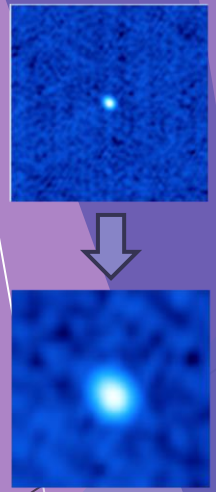
- Crop and re-size so that central galaxy occupies larger percentage of the image

Align the model to the downstream task data

- Add a whitening layer to normalize representations
- Change the model patch size
- More model parameters, either in the backbone or projection head
- Fine-tune the backbone efficiently
- Full fine-tuning of the backbone

Align the downstream task data to the training data

Technique	ResNet-50 F1	AM-RADIO F1
Crop close to center source	**0.643 (+1.32)**	**0.705 (+0.115)**
Resize image to 320x320	0.481 (-0.031)	0.630 (+0.041)
Resize image to 480x480	0.465 (-0.046)	0.666 (+0.080)
Resize image to 512x512	0.454 (-0.057)	0.667 (+0.078)

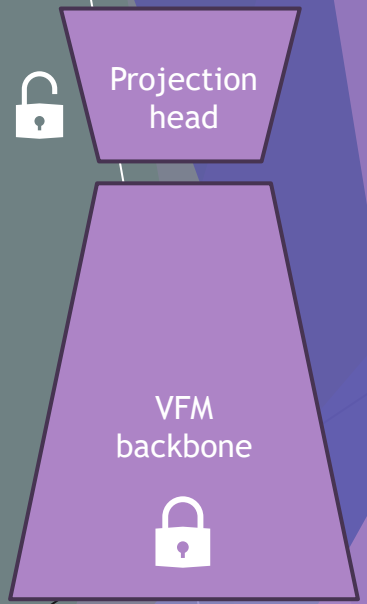


Align the model to the downstream task data

Technique	ResNet-50 F1	AM-RADIO F1
Whitening Layer	0.519 (+0.008)	0.603 (+0.013)
MLP(larger projection head)	0.526 (+0.015)	0.623 (+0.034)
ResNet-101/ViT-Large (larger backbone model)	0.486 (-0.025)	0.625 (+0.036)
ResNet-152/ViT-Huge (much larger backbone model)	0.486 (-0.026)	0.653 (+0.063)
Efficient fine-tuning (LoRA)	0.674 (+0.163)	0.692 (+0.102)
Full fine-tuning	**0.737 (+0.226)**	**0.721 (+0.132)**

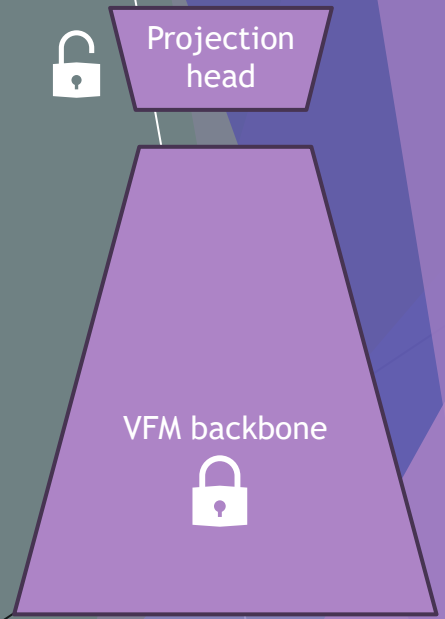
Align the model to the downstream task data

Technique	ResNet-50 F1	AM-RADIO F1
Whitening Layer	0.519 (+0.008)	0.603 (+0.013)
MLP(larger projection head)	0.526 (+0.015)	0.623 (+0.034)
ResNet-101/ViT-Large (larger backbone model)	0.486 (-0.025)	0.625 (+0.036)
ResNet-152/ViT-Huge (much larger backbone model)	0.486 (-0.026)	0.653 (+0.063)
Efficient fine-tuning (LoRA)	0.674 (+0.163)	0.692 (+0.102)
Full fine-tuning	**0.737 (+0.226)**	**0.721 (+0.132)**



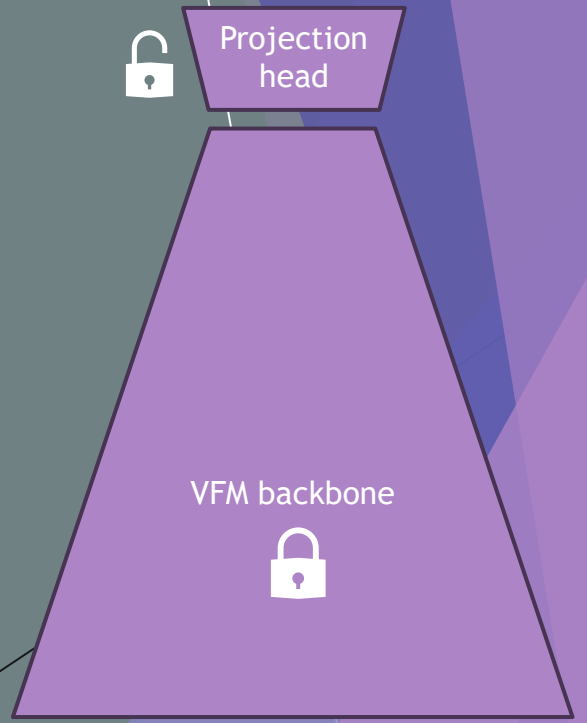
Align the model to the downstream task data

Technique	ResNet-50 F1	AM-RADIO F1
Whitening Layer	0.519 (+0.008)	0.603 (+0.013)
MLP(larger projection head)	0.526 (+0.015)	0.623 (+0.034)
ResNet-101/ViT-Large (larger backbone model)	0.486 (-0.025)	0.625 (+0.036)
ResNet-152/ViT-Huge (much larger backbone model)	0.486 (-0.026)	0.653 (+0.063)
Efficient fine-tuning (LoRA)	0.674 (+0.163)	0.692 (+0.102)
Full fine-tuning	**0.737 (+0.226)**	**0.721 (+0.132)**



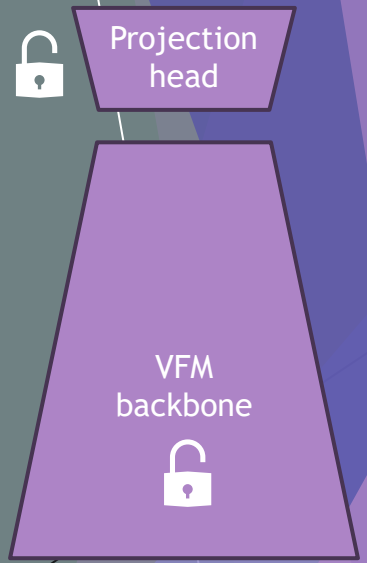
Align the model to the downstream task data

Technique	ResNet-50 F1	AM-RADIO F1
Whitening Layer	0.519 (+0.008)	0.603 (+0.013)
MLP(larger projection head)	0.526 (+0.015)	0.623 (+0.034)
ResNet-101/ViT-Large (larger backbone model)	0.486 (-0.025)	0.625 (+0.036)
ResNet-152/ViT-Huge (much larger backbone model)	0.486 (-0.026)	0.653 (+0.063)
Efficient fine-tuning (LoRA)	0.674 (+0.163)	0.692 (+0.102)
Full fine-tuning	**0.737 (+0.226)**	**0.721 (+0.132)**



Align the model to the downstream task data

Technique	ResNet-50 F1	AM-RADIO F1
Whitening Layer	0.519 (+0.008)	0.603 (+0.013)
MLP(larger projection head)	0.526 (+0.015)	0.623 (+0.034)
ResNet-101/ViT-Large (larger backbone model)	0.486 (-0.025)	0.625 (+0.036)
ResNet-152/ViT-Huge (much larger backbone model)	0.486 (-0.026)	0.653 (+0.063)
Efficient fine-tuning (LoRA)	0.674 (+0.163)	0.692 (+0.102)
Full fine-tuning	**0.737 (+0.226)**	**0.721 (+0.132)**



Improving performance by mitigating distribution shift

Technique	ResNet-50 F1	AM-RADIO F1
Whitening Layer	0.519 (+0.008)	0.603 (+0.013)
MLP(larger projection head)	0.526 (+0.015)	0.623 (+0.034)
ResNet-101/ViT-Large (larger backbone model)	0.486 (-0.025)	0.625 (+0.036)
ResNet-152/ViT-Huge (much larger backbone model)	0.486 (-0.026)	0.653 (+0.063)
Resize image to 320x320	0.481 (-0.031)	0.630 (+0.041)
Resize image to 480x480	0.465 (-0.046)	0.666 (+0.080)
Resize image to 512x512	0.454 (-0.057)	0.667 (+0.078)
Crop close to center source	**0.643 (+1.32)**	**0.705 (+0.115)**
Efficient fine-tuning (LoRA)	0.674 (+0.163)	0.692 (+0.102)
Full fine-tuning	**0.737 (+0.226)**	**0.721 (+0.132)**

Are standard VFMs useful for astrophysics?

Yes

There are so many pre-trained VFMs available, that there is no reason not to start with one! However, some models are more suited to particular tasks and datasets than others.

Questions to consider:

- How much distribution shift?
- Which downstream task?
- What resources (labeled data, compute, personnel) are available?
- Is it more effective to align your data to the training data, or the model to your data?

Full results – paper & code

https://arxiv.org/abs/2409.11175

https://github.com/elastufka/fm4astro