Using self-supervised learning to find merging galaxies

Brown AI Winter School 2025
January 16, 2025

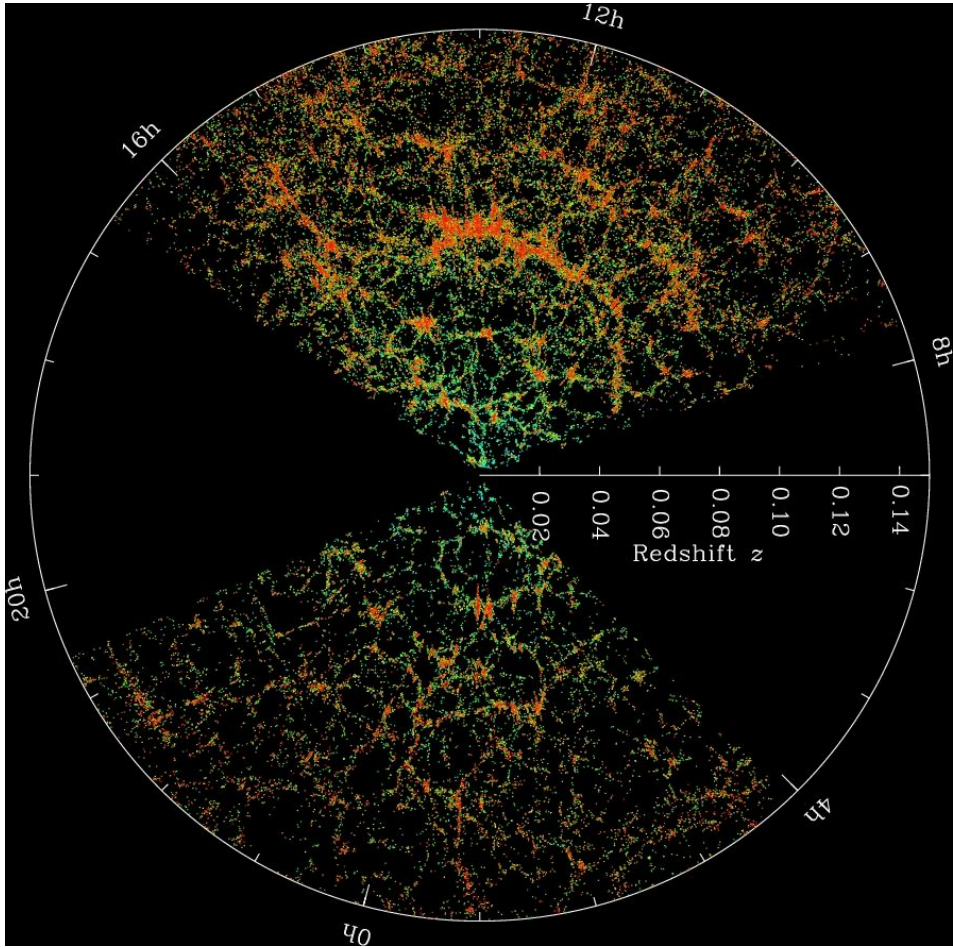
Ian Dell'Antonio (presenter)
Philip LaDuca (co-presenter and module developer!)

Galaxies–the base of cosmic structure

Matter in the Universe is not randomly distributed in the Universe.

The large scale structure is composed of (mostly) distinguishable systems.

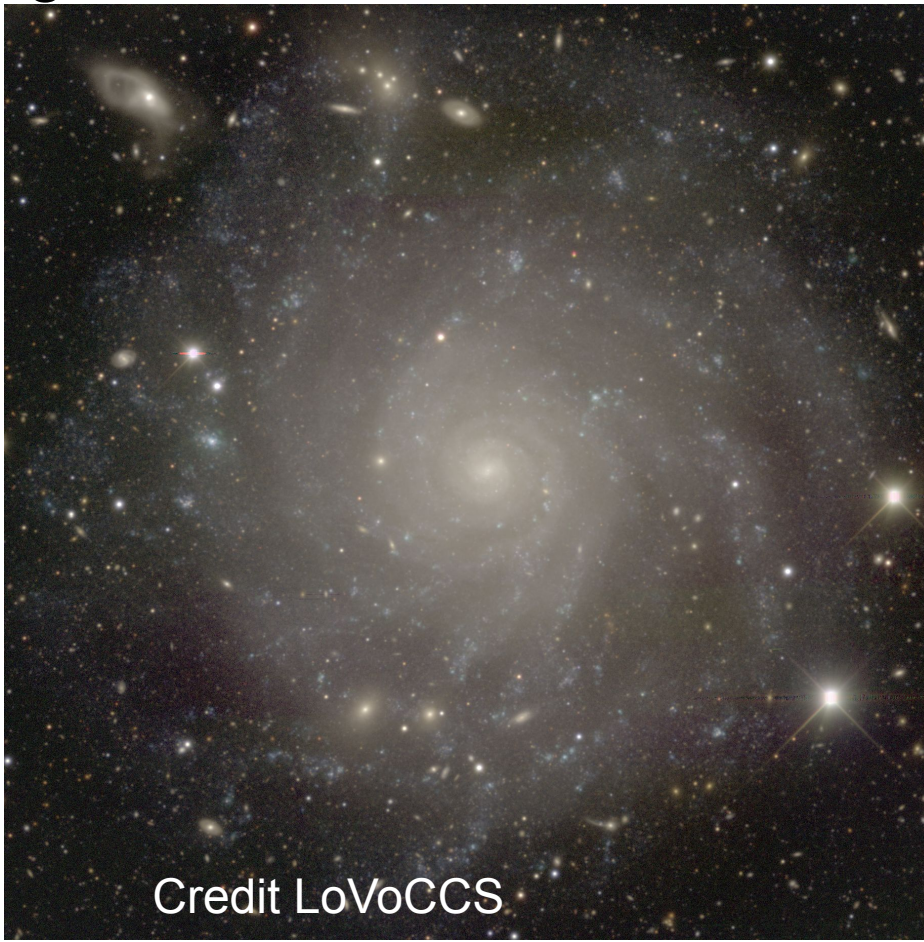
Galaxies!



Galaxies are collections of stars, gas, dust and dark matter

Galaxies come in many shapes and sizes, but they mostly fall into two categories:

Spiral



Elliptical

Galaxies interact more than stars do

Stars inside (and outside of) galaxies mostly don't collide.

This is because the typical size of stars is 10^{8}-10^{10}m, while their typical separations are 10^{16}-10^{17}m –the separation to size ratio is 10-100 million to 1!

Even though the galaxy crossing time for a star is much smaller than the age of the Universe, their relative size makes it unlikely that any collisions happen.

The situation is completely different for galaxies. Galaxy sizes vary a lot, but largish galaxies have sizes ~10^{20}m, but the inter-galaxy separation is around 10^{22}m – a ratio <100 to 1. Galaxies take ~10^{9} years to cross these distances, still smaller than the age of the Universe. Most galaxies have collided (and merged)!

Why galaxy interactions matter

These collisions were even more common in the past (why?--Expansion!)

They were/are responsible for triggering bursts of star formation

They evolve galaxies from one type to another

They replenish the galaxy's gas supply (sometimes) or drain it (sometimes)

They make galaxies grow.

Galaxy interaction timescale

Interactions and their effects last ~10^8 years. This is an awkward time.

1) Much too long for an individual astronomer to see in their lifetime
2) Much shorter than the age of the Universe–so you have to catch galaxies at the right time!

This means that rather than studying one interaction through all its phases (as we can simulate), we must study lots of *different* interactions each at its own phase.

Examples of galaxy interactions 1: tidal tails

These galaxies have undergone a close encounter with a medium-large galaxy, and some of the stars have become stripped through tidal interactions. These interactions remove some stars to float in the potential of the group/cluster (the “ICL”, but also increase the kinetic energy of the stars in the galaxy, puffing it up (and paradoxically “cooling” it).

Examples of interactions 2: mergers/post-merger morphology

These occur when galaxies merge–the stars retain a memory of the tidal interactions–this results in “shells” corresponding to the turnaround of bound galaxies (1/pass)--these shells can be used to determine mass ratios and histories by matching to N-body simulations

Example of interactions 3: jellyfish galaxies

The gas in galaxy clusters is in virial equilibrium with the gravitational potential–it has a temperature of 10-100 million K! As galaxies pass through it, ram pressure stripping drives the cold gas out (trailing behind the motion). This gas shocks and clumps and forms trailers of star formation behind the cluster–the length and color tell you the velocity of the galaxy through the medium!

Why surveys are necessary to find interacting galaxies

Although galaxies interact, mergers “only” last ~10^8 years–thus at any given point in time* fewer than 1% of galaxies are interacting. This means that catching all the phases of interactions requires looking through many galaxies. Because there are many interactions, reconstructing the physics on average requires millions or (to probe many mass ratios) billions of galaxies.

In early surveys, astronomers (cf. Zwicky (1956,1958) and Vorontsov-Vel’yaminov (1959, 1977)) combed through photographs of tens of thousands of galaxies to find <1000 interacting ones)

The problem–scaling from thousands to millions to billions

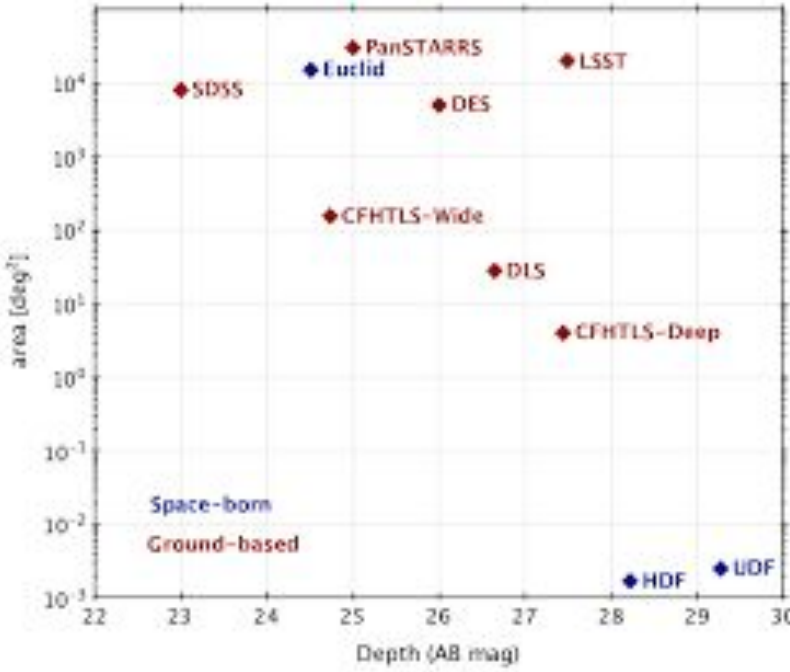
However, as we have taken deeper and deeper images of the sky, we have dramatically increased the number of galaxies.

In the diagram on the right the number of

Galaxies scales as area*$10^{(0.4*depth)}$

The next generation of surveys will have

Billions of galaxies to classify! (200-500x SDSS)

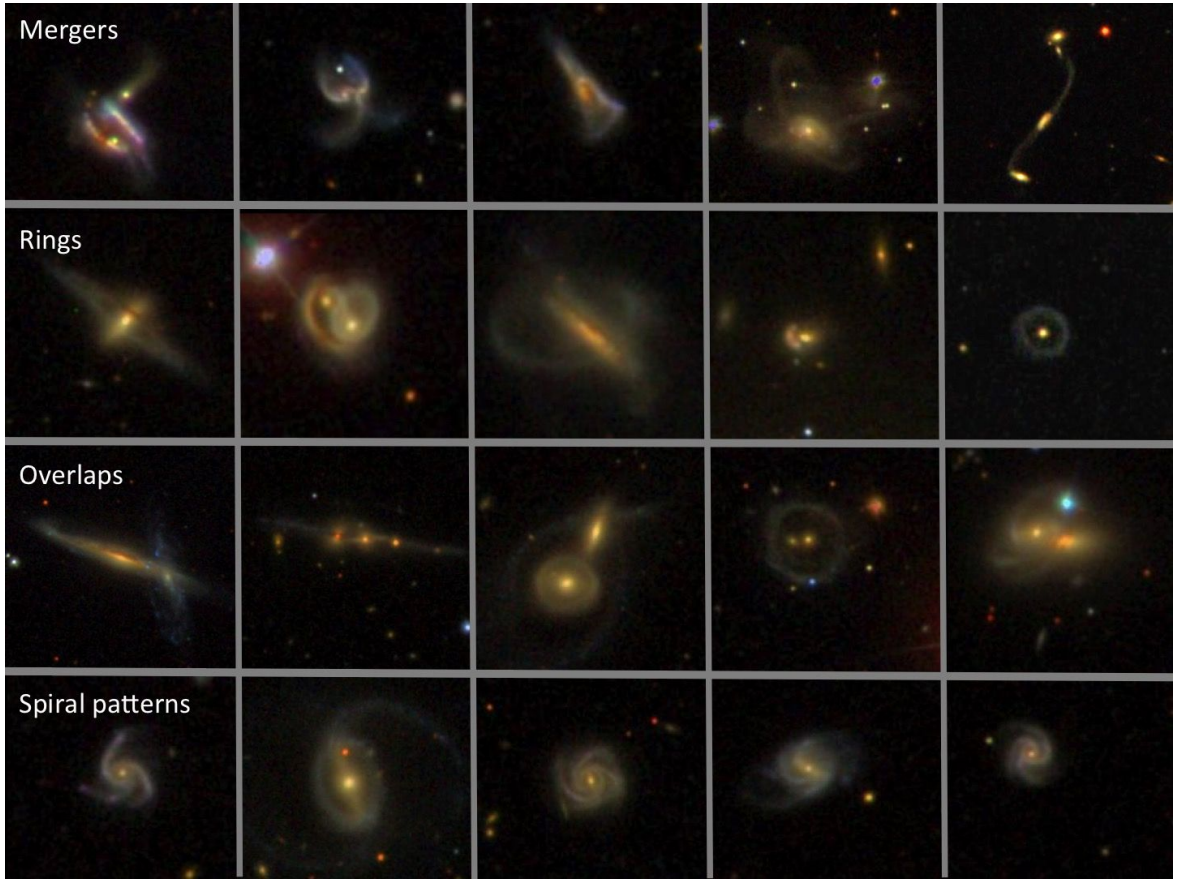


Citizen science as a partial cure

Galaxy Zoo, started in 2007–a project to enlist volunteers from the public to classify images–leveraging “RI”. Thousands of volunteers signed up–about 1 million galaxies have been classified (by multiple people). About 5000 interacting galaxies found in the Sloan Digital Sky Survey and the CANDELS Hubble survey fields…

But the new surveys will exceed this by a factor of 100. To long even for citizen science…

This is why machines should do it for us!

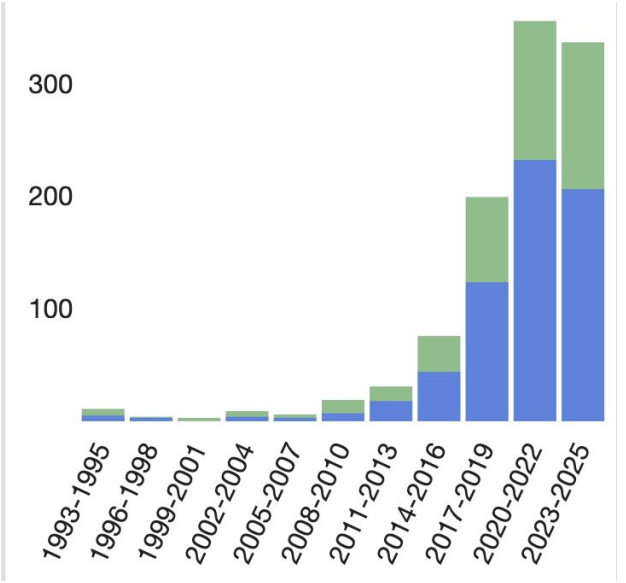


Not an original idea….

Even for galaxy interactions:

See Abraham et al. 2025, Ferreira et al. 2024, Desmons et al. 2024, Margalef-Bentabol et al. 2024…

The implementation we will show you is also not unique. Based on Desmons et al. 2024, but you can and should explore others…



ADS search results for “machine learning galaxy classification”

AI interlude– supervised machine learning

Machine learning has become a exciting and popular approach to exploring patterns and classifying objects in astronomy. Applying machine learning techniques to the classification of galaxies and their tidal features allows us to both leverage the large amounts of data to train robust models and utilize the trained models to pick up classification when citizen science efforts become insufficient.

Supervised learning is the basic form of deep learning. In the case of classifying tidal features, a model is trained using previously classified images allowing it to learn the rules governing the classification of a tidal feature and then apply those rules to new classification.

What about self-supervised learning?

One of the main limitations to supervised learning is that the training data is limited to the amount of labeled data available for training. This means that improving the model requires further manual classification.

Self-supervised models are trained without labeled data, allowing the model to build an understanding of the data by encoding the images into an embedding space. For this model, we use contrastive learning which trains the model to group similar features within the embedding space and separate dissimilar features.

Once the self-supervised model learns representations of the data in the embedding space, we can fine-tune the model by training it with the labeled data to make classifications, leveraging the large amount of unlabeled data to aid the model.

Self-supervised models also have another strength which we'll talk about later…

Set up of the notebook

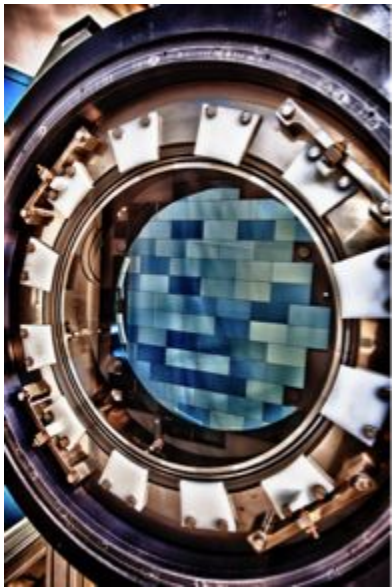
To explore both supervised and self-supervised approaches to tidal feature classification, we will start by getting the colab notebooks ready and mounted to your google drive. We will then:

1. Look at the data augmentation for the models.
2. Finalize the training for a supervised model.
3. Explore the embedding space of a pre-trained self-supervised model.
4. Attempt the classification of individual galaxies using both models.

The data–introducing the LoVoCCS survey

In this session you'll be working with data from the Local Volume Complete Cluster Survey–a survey of 100 regions of the sky around the most massive galaxy clusters in the nearby* Universe (nearby means within 1 billion light years).

The survey covers about 500 square degrees of the sky (200 full Moon's' worth) and measures about 200 million galaxies, about 2,000,000 of which are big enough to unambiguously measure the shapes.



You'll be working with data from just 3 clusters, in order to keep the

Data volume and training times short enough to fit into this timeslot.

The training set will consists of images of galaxies taken from the

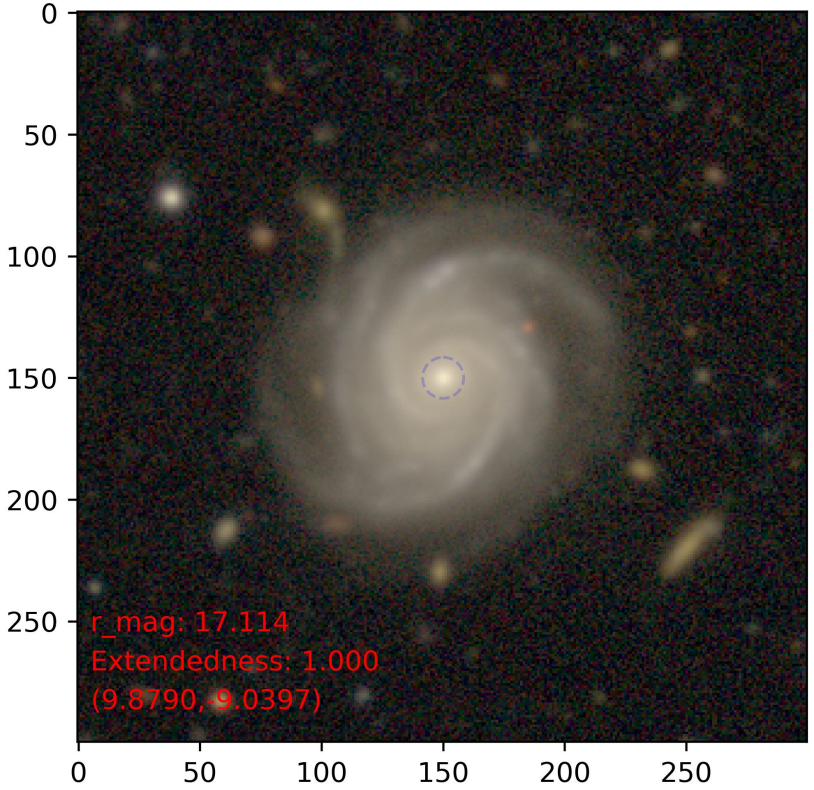
Clusters Abell 2029, Abell 3667, and Abell 85 (no worries if the names

Don't mean anything to you.!)

Data inputs

For each cluster, we will be using “postage stamps” –200x200 pixel images that are extracted automatically–a (non-AI) program selects all bright objects in the image (that don’t meet the criterion for stars), and automatically extracts the same region in the images taken with 5 different filters–the combination of multiple filters are shown at right. The network has the 5 black and white images, rather than the color composite to work with.

Three members of our group have looked through 6000 stamps (tedious!) and classified them. This is important because in Supervised learning you need to give “truth” values to train the network. Today, you’ll only be using 900 of them to keep things fast.



Description of the training set

In order to train the model, a dataset of both galaxies with tidal features and galaxies without tidal features is necessary. To have a balanced dataset, half of the galaxies are labeled 1, containing a tidal feature, or 0, without a tidal feature.

Why might it be important for the dataset to have a balance between the two classifications?

Why the data is augmented (description part 1)

During the training of a model, it is important to augment the data in order to order to get a wider distribution of data and help improve the model's accuracy. By including data augmentations, the model can become more robust to factors such as noise, the orientation of galaxies, and the positioning of a galaxy in the image. This can help improve the model's accuracy when classifying unseen data and by having random augmentations applied to the training data, the model will avoid memorizing specific classifications for images known as overfitting and instead learn more general rules.

One of the key goals in training is to avoid having the network fit for something (like the orientation of a galaxy) that is not important!

How the data is augmented

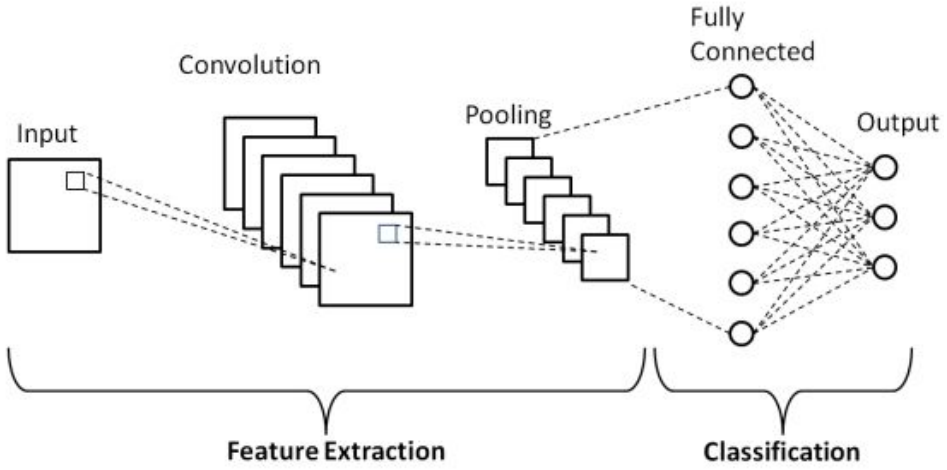
There are three main augmentation applied to the galaxy images.

1. **Random horizontal and vertical flipping of the data.** This helps the model acquire rotational invariance which is important since galaxies are imaged at essentially random orientations relative to our view from Earth.
2. **Gaussian noise applied to the images.** The addition of noise to the model is important since images of different clusters may have varying levels of noise depending on the conditions during the nights the images were taken, allowing for the model to handle different conditions.
3. **Random cropping of the image.** By randomly cropping a portion of the image the center of the image shifted slightly allowing the model to learn some amount of translational invariance.

Overview of the supervised model

The supervised model is a convolutional neural network meaning the majority of layers are convolutions used to reduce the dimensions of the input images. This is done through both convolution operation and subsequent maxpool dividing the length and width of the images by two. ReLU is used as the activation function and batch normalization is included to help stabilize the model.

Once the dimensions reach a small enough size, the images are flattened and passed through linear layers producing the final classification through a sigmoid activation function. Since there are only two classifications, binary cross-entropy is used as the loss function and Adam is used as the optimizer.



DOI:10.3390/app9214500

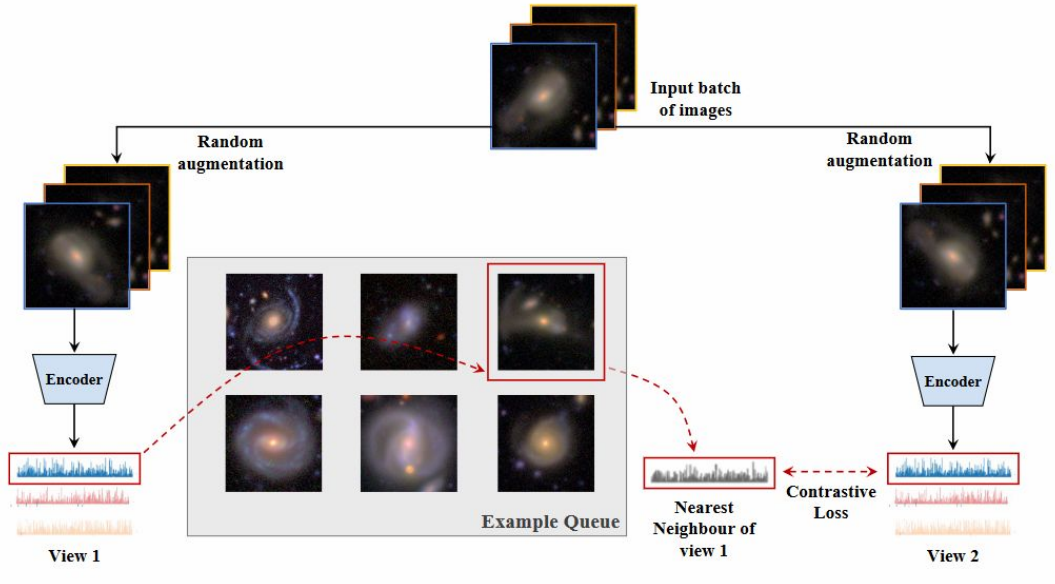
Overview of the self-supervised network

Similar to the supervised model, the self network relies on a convolutional neural network as the encoder which projects the input data into the embedding space. This architecture follows the res-net model, a common framework for convolutional networks. The self-supervised training is done on this model initially allowing it to learn useful embeddings for the galaxy images. After the self-supervised training, a dense network is added to the end of the encoder which takes the embeddings as an input and produces a classification. To train this model supervised fine-tuning is used with labeled training data to train the weights of the linear layers while using the previously trained encoder weights.

Self-supervised training:

In order to train a self-supervised network, we need a way for the model to judge the similarity between two images. To accomplish this, we start by having the model create two augmented versions of an image find the two embedding vectors using the encoder. The embeddings are then assessed using a contrastive loss function which attempts to maximise the projection between the two embeddings.

This trains the model to consider augmentations of the same image as similar. In the NNCLR architecture introduced in Dwebidi et al. 2021, the embeddings are also compared to the most similar embeddings from other images in the dataset which helps the model group images with similar characteristics.



Exploring the feature space

Once the self-supervised is trained, it can be insightful to explore the embedding space it creates. Using UMAP, a tool which projects higher order embedding spaces down into two dimensions, we are able to graph the embeddings created by the encoder. Plotting images near each other in different regions of the projected embedding space shows us insight into the groupings created by the self-supervised model.

What’s next for this project

The notebook has been trained on <5% of the data (some of which still being collected)-- there’s plenty of more to be trained.

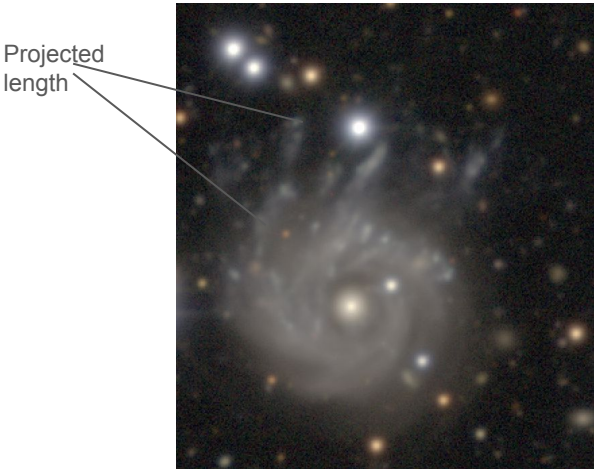
Image scalings can be important–we will be experimenting with normalizations to increase the weight of the lower surface brightness features over (for example) the galaxy nucleus

Extracting the physics–

Once we have the sample of interacting galaxies, measurements of

1) Feature length (gives projected distance)
2) Feature color (gives interaction time)

Can be combined to determine the cluster-centric transverse velocities of the galaxies–this is information that is very hard to get any other ways



(additional spectroscopic measurements give the more precise ages)

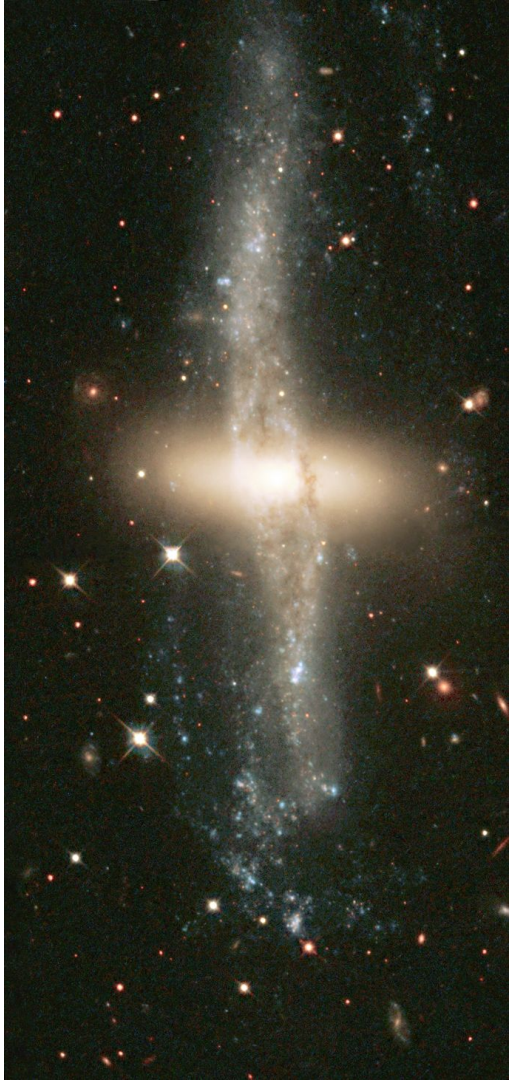
Similarly, measuring where the interacting galaxies are can give information about the conditions (galaxy density, relative velocities, gas density) that drive most interactions.

Unexpected discovery potential

If supervised and unsupervised methods are both effective, why unsupervised?

An unsupervised method finds galaxies that are “different”--you can use the same unsupervised training and then plug in different classifiers as new types of galaxies are discovered and deemed interesting.

(For example, a polar ring galaxy–a long-term remnant of galaxy collision)



Future Survey applications–LSST (starting this year!)

LoVoCCS is big, but there's a much bigger survey just about to begin:

The Vera Rubin Observatory is almost ready to

Begin its 10-year survey of the sky.

The Camera is being

Installed in 1 month, the survey begins at the end of the year.

The telescope is 8.4 meters, the camera has 3.2 Gigapixels, and the

Images are built from a huge string of 30-second "snapshots"

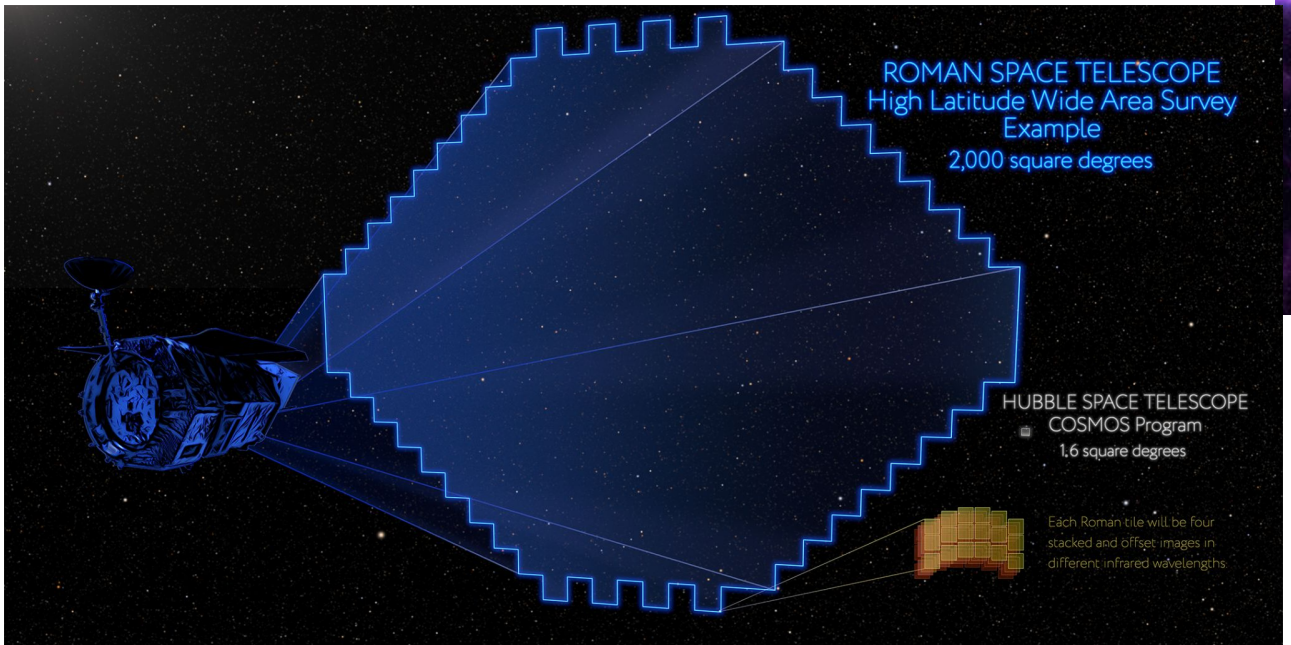
Half the sky imaged every month, 50 billion galaxies* (some too small to classify).

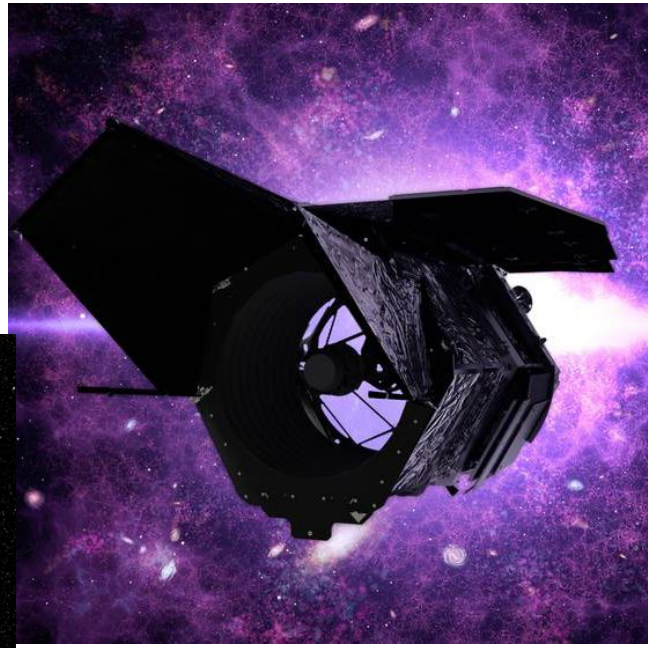
Future Survey Applications–Roman HLWAS

To get around the issue of resolution, we can go to space!

The Roman Space telescope will conduct a survey of 2000

Square degrees (details still being proposed). At 80 galaxies per square arcminute, ~10^9 galaxies to search





In the NIR, can look for tidal rest-frame blue features at z>1, which you can't do from the ground.

Limitations you should know

- The self-supervised methods are prone to finding differences in latent dimensions that represent “nuisance parameters”--augmentation has to be done carefully
- The classifier requires a significant number of classified galaxies–this can be very time consuming to create manually
- Training a large number of images with the self-supervised network can be very slow–make sure to use gpus if you want to train very large datasets.

Conclusions

We hope you've enjoyed this very brief look at one aspect of galaxy classification!

Be sure to submit your classification and training results. See the README file in the shared drive to find the link for this module.

Thank you!