**Preprocessing**

**Downstream**

* [**Catalina Real-Time Transient Survey**](http://crts.caltech.edu/)
  + Metadata
    - Fits images
    - 1 square degree up to 19 square degree
    - **Grayscale**
    - **5280 x 5280 pixels** (according to xml file)
  + [Deep-Learnt Classification of Light Curves](https://arxiv.org/pdf/1709.06257.pdf)
    - Preprocessing
      * Difference in magnitude and difference time
      * Binned for range of dm and dt values
      * Image intensity is normalized
    - Downstream
      * Convert time series to two-dimensional light curve
* [**DESI Sky Survey**](https://data.desi.lbl.gov/doc/releases/)
  + Metadata
    - **41M** images
    - **152 x 152**
    - **197,979 pairs**
    - grayscale/rgb
  + [A brief review of contrastive learning applied to astrophysics](https://academic.oup.com/rasti/article/2/1/441/7238495)
    - Similarity Search based on He et al (2019) MoCo algorithm
  + [Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation](https://arxiv.org/pdf/1107.3160.pdf)
    - Use weak gated experts to derive photometric redshifts
      * Redshifts for:
        + Optical galaxies with spectroscopic redshifts
        + Optical quasars with spectroscopic confirmation and redshift
        + Optical + ultraviolet quasars spectroscopically confirmed
      * Weak gated experts method
        + Preprocessing

Used fuzzy k-means

* + [Astroclip: Connecting Diverse Observational Modalities in Astrophysics](https://arxiv.org/pdf/2310.03024.pdf)
    - Preprocessing
      * Center crop images to 96 x 96
      * Cross-match galaxy spectra from the DESI Early Data Release
    - Downstream
      * Maximize mutual information about object constructing embeddings of multiple modalities
      * Used to align representations from different modalities around shared semantics
  + [Self-supervised similarity search for large scientific datasets](https://arxiv.org/pdf/2110.13151.pdf)
    - Preprocessing
      * Chose 42,272,646 galaxies with a z-band magnitude < 20
      * Centered on each galaxy, extracted 152 x 152 pixel cutout in the three optical bands (g, r, z)
    - Downstream
      * Create interactive semantic similarity search tool
      * Discover rare objects given only a single example
  + [Transfer learning for galaxy morphology from one survey to another](https://watermark.silverchair.com/sty3497.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAA38wggN7BgkqhkiG9w0BBwagggNsMIIDaAIBADCCA2EGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMykRbv6AiHHOzc56LAgEQgIIDMvKE-TSKjHyRz9JJU5SeV0hnHKqBOy_XGTiPiWse8FHNI-3bUWc98j03ebFoXQqbgh3Dfj9a8jDb5xH3_ehARbde76lyfP5pj3cKzjyejt8VFFAXnqrm8HwJ8sZi3ngdOkOIgvLkAT5GuEy5ga4kwuMs9UkUqXjQCpivOtvp5tya-xK2_zSBmjw5zTKydhzSuyWqhYt7vfncyrjlVlX8sDPO3k-mD0hfmsj8GUKEPBUzDC_FGZcQq1uYdot3JbtVe8_VW1S7nwfKNtQcBT78zI060QMpHo24OY0vaeB-SZ8TN1TLk-idTJ39Uy65hp8VzjIMk-MdFWe4nGmThrcHwPdbz-h5SC4uV8ndYDHbUlVtNduPGdNHiD8DBiWnMqUu0RBakXNG7bemxu2MvGW3BPwBTxSj9YNmaS1lcQBxUFGWIRhORvTe2HNnANLR663rCUp7fFao7bvpKPb5IaWr8JR8QJ3L2HRterDQrgWKkIeK2pHg9ko6ZwqsQT_0cOlbZyGJsdbvcxk89TxDLiwh2yQf-K80rbXHmFQcq1JwBmVtNzzVQPffbxq7-6aOM5dFsz1Hh-_T_xVkAUgsyn-XMBztvQsiMcHBOCf45KiGS9MJ85n2TgfXi3x4LKZQI_lTb1OZBVUROZ6hnNWuh6kUGOzxv_wNNtwF1G7IMTC2L5_GUsmeUbQfYaRUaXDuDuY4EzStH8keQZKIJJ5rjBmF6H_0VBPfCOLX4VAiFejULYnQKyptaZ1E0osSFIQyO99BMax4bFMrkWAUuEGbEDDXPa_7mVGBo1IgTtxiqAbXPUCQzpmZfyZA4vJ6iDQQcwOxBvoGWzLLexRFI4-MKfUpPkYvrivanLYndUsIx2TogzBhm4wTodmb1L40b_PV8zRzv5AmicqkNfK7ptsEpI7QMBKHZXQTKyrVcd11lDlzP4k1WDygExpkeyQiUWwNN-PWSJWmOrwswxrOxgb_kxFZkRU_vmcLz3iviFYzZBZ2c6MHP2tv4w_gOq-yfBjPq6DsJu3s3-Ah-k8HODMOGm9aJ8H7mK5gj43GvS37CwmVKFpKyxSZwzJef-pEOqQMGgsgc9r-)
    - Preprocessing
      * (also uses SDSS)
      * Not discussed in paper
    - Downstream
      * Morphological classification of galaxies
* [**HTRS-Survey**](https://sites.google.com/site/htrupublications/htru-discoveries)
  + Has primarily medlat and hilate
  + Metadata
    - **>100** pulsar discoveries
    - Tabular data
  + [Pulsar Candidate Identification Using Semi-Supervised Generative Adversarial Networks](https://arxiv.org/pdf/2010.07457.pdf)
    - Preprocessing
      * Plot various two-dimensional and one-dimensional projections
      * Used pre-processing code made available by [Zhu et al](https://iopscience.iop.org/article/10.1088/0004-637X/781/2/117)
    - Downstream
      * Pulsar candidate identification
  + [Separation of pulsar signals from noise using supervised machine learning algorithms](https://arxiv.org/pdf/1704.04659.pdf)
    - Preprocessing
      * Feature selection
        + Feature selection

Mutual information between a feature and its corresponding class label

p = probability density function

x, y = continuous random variables

Rank features

* + - Downstream
      * Separation of pulsars from radio frequency interference
      * Binary classification
    - Metrics
      * Recall
      * Precision
      * Accuracy
      * F1 score
      * Log loss
      * G-Mean
      * Area under Receiver Operating Characteristics
        + Best value is 1 and worst is .5
        + TP vs FP rate
      * False Positive Rate
* [**LOFAR**](https://lofar-surveys.org/releases.html)
  + Metadata
    - **4,396,228** radio sources
  + [Deep learning based detection of cosmological diffuse radio sources](https://academic.oup.com/mnras/article-pdf/480/3/3749/25519514/sty2102.pdf)
    - Preprocessing
      * Created suite of large cosmological simulations of extragalactic magnetic fields
        + ASKAP as basis
      * Applied random rotations
      * Tiling based approach that divides image into small square tiles
      * Labeled tiles
    - Downstream
      * Detect extended extragalactic radio sources
  + [Convolutional Deep Denoising Autoencoders for Radio Astronomical Images](https://arxiv.org/pdf/2110.08618.pdf)
    - Preprocessing
      * Images are synthetically generated
      * Random rotations to each of the different redshift slices
      * Created noise image image, clean images, and dirty images
      * Divided into square tiles
    - Downstream
      * Detect faint, diffused radio sources predicted to characterize the radio cosmic web
* [**Palomar-Quest Digital Synoptic Sky Survey**](https://arxiv.org/pdf/0801.3005.pdf)
  + Metadata
    - Area coverage
    - per clear night
    - **2048 x 2048 pixels** (ChatGPT)
    - Single channel (ChatGPT)
  + [New Approaches to Object Classification in Synoptic Sky Surveys](https://arxiv.org/pdf/0810.4945.pdf)
    - Preprocessing
      * Multilayer perceptron with a softmax activation function and cross-entropy error
    - Downstream
      * Morphological object classification
      * Removing instrument-related artifacts
* [**Pan-STARRS1**](https://outerspace.stsci.edu/display/PANSTARRS/)
  + MetaData
    - **2885** light curves with host-galaxy redshifts; **557** have spectroscopically-confirmed types
    - **60** Orthogonal Transfer Arrays devices (OTA); each device has 8 x 8 array of “cells”; single OTA format is **4846 x 4868** pixel array; each device has **64** cells where each cell is **590 x 598** pixels
    - Seems grayscale (1 channel) when processed through a bandpass; full display uses 3 channels
  + [ParSNIP: Generative Models of Transient Light Curves with Physics-Enabled Deep Learning](https://arxiv.org/pdf/2109.13999.pdf)
    - Preprocessing
      * Make rough estimate of the time of maximum light of the light curve by taking the median time of the five highest signal-to-noise observations in the light curve
      * Biweight estimator to estimate background level for each bandpass
      * Subtract estimate of background level from each light curve
      * Normalization of brightness
      * Add error floor of .01
    - Downstream
      * Model unknown intrinsic diversity of different transients
      * Predicts time-varying spectra of transients
* [**PLAsTiCC**](https://www.kaggle.com/code/michaelapers/the-plasticc-astronomy-starter-kit)
  + Metadata
    - Tabular data or time series
    - First table: **12** features; Second Table: 6 features
    - Test set: **~3.5M** objects; Training Data: **8000** sources, maybe **~3.5M** objects
* [**Sloan Digital Sky Survey -- DR7**](https://classic.sdss.org/dr7/)
  + Metadata
    - Approximately **357 million** unique celestial objects (ChatGPT)
      * Each image is **2048 x 1489** (from DR9)
      * Without bands it seems to follow the RGB channel standard
      * Bands: **u**, **g**, **r**, **i**, **z**
        + **1** channel
    - Spectrum: infrared (DR18)
  + [A brief review of contrastive learning applied to astrophysics](https://academic.oup.com/rasti/article/2/1/441/7238495)
    - Morphology Classification - Wei et al. (2022)
    - Predict spectra from images (combined with multimodal conditional diffusion models) - Chen et al (2020a)
    - Domain Adaptation and Clustering - Chen et al (2020a)
  + [Classification of Astronomical Bodies by Efficient Layer Fine-Tuning of Deep Neural Networks](https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9672430&casa_token=1ajTnZylzWcAAAAA:r2B7Crg6drIdJ-lTmZK0DWnMWWVOf2z509EfUNVJ_nWgpPHJmlIYD5ovuW8MF_RZz6w4Rxxm)
    - Preprocessing
      * Resized from 2048 x 2048 x 3 to 512 x 512 x 3 to reduce number of training parameters
      * Feature Selection
        + Uses images
    - Downstream
      * classification
  + [Deep Learning Approach to Photometric Redshift Estimation](https://arxiv.org/pdf/2310.16304.pdf)
    - Preprocessing
      * Sigma-clipping on the redshift values (3 standard deviation)
      * Feature selection
        + 5 wavelengths of light: *u, g*, *r*, *i*, and *z*
        + Removed redshift values less than 0
    - Downstream
      * Redshift predictions
  + [Fanaroff-Riley classification of radio galaxies using group-equivariant convolutional neural networks](https://arxiv.org/pdf/2102.08252.pdf)
    - Preprocessing
      * Some labeled objects were excluded
        + 40 objected denoted as 3 - unclassifiable
        + 28 objects which had an angular extent greater than selected image size
        + 4 objects with structure that was found to overlap the edge of the sky area
        + Single object in 3-digit category 103
      * image pixels set to zero if value is below a threshold of three times the local rms noise
      * Clipped to about 150 x 150 pixels
      * All pixels outside a square central region with extent equal to the largest angular size of the radio galaxy are set to zero
      * Normalized
      * Feature Selection
        + Images
    - Downstream
      * Classification of astronomical objects
  + [Galaxy Morphological Classification with Efficient Vision Transformer](https://arxiv.org/pdf/2110.01024.pdf)
    - Preprocessing
      * Downloaded from Kaggle
        + Size is **424 x 424 x 3** with **g**, **r**, **i** filters from SDSS
      * Apply thresholds on series of voting questions answered by participants in GZ2
      * 64% training, 16% validation, and 20% testing
      * Crop images into 224 x 224 x 3
      * Flipping and rotating data augmentation techniques
      * Normalize by mean ([.094, .0815, .063]) and standard deviation ([.1303, .11, .0913])
      * Feature Selection
        + Images with *g*, *r*, and *I* filters
    - Downstream
      * Galaxy morphological classification
  + [Photometric redshifts for Quasars in multi band surveys](https://arxiv.org/pdf/1305.5641.pdf)
    - Preprocessing
      * Used four datasets SDSS, GALEX, UKIDSS, and WISE
      * Selected quasars which had reliable measure of the spectroscopic redshifts
      * Used maximum radius r = 1.5” to associate the optical quasars to counterparts in each of the three catalogs
      * Feature selection
        + 15 bands from various datasets
        + 43 different features
        + Best combination of features

Five SDSS psfMag

Two isophotal magnitudes of GALEX

Four HallMag for UKIDSS

Four magnitudes for WISE

* + - Downstream
      * Evaluating photometric redshifts for quasars
  + [Self-Supervised Representation learning for Astronomical Images](https://arxiv.org/pdf/2012.13083.pdf)
    - Preprocessing
      * Galactic extinction
        + Artificial sampling by sampling a reddening value from and applying corresponding per-channel extinction according to photometric calibration from Schlafly, D. & Finkbeiner
      * Point spread function (PSF)
        + PSF augmentation using Aaussian smoothing
      * Rotation
        + Random rotation from distribution
      * Random jitter & crop
        + Jitter center of image along respective axis
      * Gaussian noise
        + Add Gaussian noise for each color channel
      * Feature Selection
        + Used images for morphology classification
        + for redshift estimation
    - Downstream
      * Similarity search
      * Galaxy morphologies
      * Redshift predictions
      * Anomaly detection
      * Strong lens finding
      * Low brightness gals
  + [Towards Galaxy Foundation Models with Hybrid Contrastive Learning](https://arxiv.org/pdf/2206.11927.pdf)
    - Preprocessing (seems to be all they did)
      * Images grouped into surveys according to telescope used and operating conditions
      * Volunteers labeled images
    - Downstream
      * Purely-contrastive approaches in general (paper is not super clear)
  + [Photometric Redshifts from SDSS Images with an Interpretable Deep Capsule Network](https://arxiv.org/pdf/2112.03939.pdf)
    - *Note: uses a subsection of the SDSS that Amirezza uses*
    - Preprocessing
      * Pre-processed images and spectroscopic redshifts used by Pasquet et al
        + Stacked and re-sampled images to common 64 x 64 x 5 pixel grid
        + Background subtracted and photometrically calibrated with same zero point
      * Feature Selection
    - Downstream
      * Photometric redshift estimation
    - Metrics
      * Hinge loss
        + Tj represents class labels
        + M+ = .9
        + M- = .1
        + Lambda = .5
      * Output of decoder network for sum of squared errors
      * Squared error for redshift regression network
      * For margin loss and total squared reconstruction (weighted)
      * Same as above except with photo-z loss
      * Photo-z evaluation metrics

* [**VLA FIRST**](https://science.nrao.edu/vlass)
  + Metadata
    - Tile Size: 512 x 512
    - HiPS image
  + [FIRST](https://sundog.stsci.edu/)
    - Metadata
      * 1150 x 1550 pixels
      * 34.5 arcmin x 46.5 arcmin - .45 square degree area of the sky
    - [Classifying Radio Galaxies with Convolutional Neural Network](https://arxiv.org/pdf/1705.03413.pdf)
      * Downstream
        + Morphological classification of categories of radio sources
      * Preprocessing
        + Sigma-clipped statistics of each image are estimated in order to calculate the background noise and flux level

Pixels above certain sigma level from the median are discarded or nulled

Below 3 std. Were cut-off

* + - * + Augmentations

Images were 300 x 300; rotated in small angles in steps of either 1, 2, or 3 degrees

150 x 150 patch centered on source was cut out from main image

* + - [Weight Pruning and Uncertainty in Radio Galaxy Classification](https://arxiv.org/pdf/2111.11654.pdf)
      * Preprocessing
        + Describe in Anna M. M. Scaife and Fiona Porter. Fanaroff-Riley classification of radio galaxies using group-equivariant convolutional neural networks
      * Downstream
        + Could not find; maybe uncertainty quantification
      * Metrics
        + Uncertainty calibration error
  + [FR-Deep](https://zenodo.org/records/4715983)
    - Metadata
      * 1360 images
      * 150 x 150 pixels
      * 1 channel
    - [Attention-gating for improved radio galaxy classification](https://arxiv.org/pdf/2012.01248.pdf)
      * Preprocessing
        + Clip image

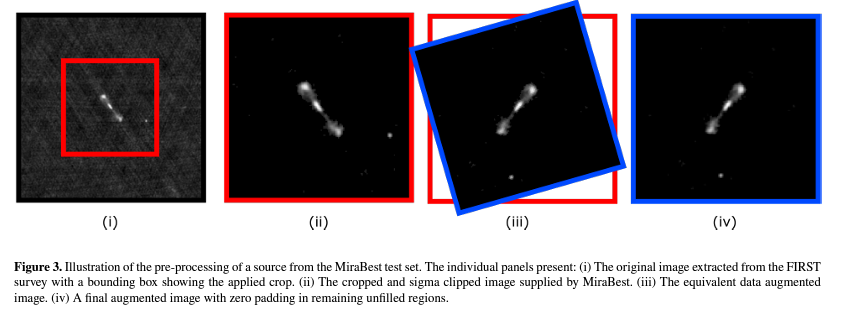
Pixel value set to 0 if value is below threshold of 3x RMS of the local noise

Removes most artifacts and leaves behind cleander images with clear sources

* + - * + Clip size to 150 x 150 pixel

Each pixel corresponds to 1.8’’ x 1.8’’

* + - * + Normalize image



* + - * + Augmentation

Such that is processed input image; ;

* + - * Downstream Task
        + Classify radio galaxies
    - [Radio Galaxy Zoo: Leveraging latent space representations from variational autoencoder](https://arxiv.org/pdf/2311.08331.pdf)
      * Preprocessing
        + Selected resolution of 64 x 64
      * Downstream
        + Classifying galaxies in labeled dataset
        + Similarity search
  + [MiraBest](https://arxiv.org/abs/2305.11108)
    - Metadata (assume same as FR-Deep since both came from VLA FIRST)
      * 1360 images
      * 150 x 150 pixels
      * 1 channel
    - [Attention-gating for improved radio galaxy classification](https://arxiv.org/pdf/2012.01248.pdf)
      * Preprocessing (same as FR-Deep)
      * Downstream Task (same as FR-Deep)
    - [Quantifying Uncertainty in Deep Learning Approaches to Radio Galaxy Classification](https://arxiv.org/pdf/2201.01203.pdf)
      * Preprocessing
        + Pixels below 3 level of the background noise were set to 0
        + Clipped to 150 x 150 pixels
        + Normalization
        + For integrity of dataset, removed objects

40 unclassifiable objects

28 objects with extent greater than chosen image size of 150 x 150

4 objects with overlapping regions of the FIRST survey

1 object in category 103

* + - * Downstream Task
        + Radio galaxy classification
    - [Radio Galaxy Zoo: towards building the first multi-purpose foundation model for radio astronomy with self-supervised learning](https://arxiv.org/pdf/2305.16127.pdf)
      * Preprocessing
        + Each sample has *confident* or *uncertain* tag
        + 104 *confident* samples for test set
        + 25% of remaining data for validation
        + 75% for training
        + Used RGZ DR1 dataset for unlabelled dataset

Removed any overlap between MiraBest and RGZ DR1

* + - * + Augmentation

Rotate input to random orientation

Center crop 128 x 128 to 70 x 70 image

Reduce range of random cropping to 80-100% of the image

Randomly flip horizontal and vertically

Color jitter

blur

* + - * Downstream
        + Image analysis of resolved extragalactic continuum images
        + Classification in a label scare regime with data from MIGHTEE survey
* [**The Zwicky Transient Facility**](https://www.ztf.caltech.edu/ztf-public-releases.html)
  + Metadata
    - **494 x 495 pixels**
    - **Grayscale**
    - **16 CCDS**
    - **4 readout** channels of **3k x 3k pixels**
    - **~700** observations
  + [Machine learning for the Zwicky Transient Facility](https://arxiv.org/ftp/arxiv/papers/1902/1902.01936.pdf)
    - Preprocessing
      * Image difference via ZOGY algorithm
      * Made initial cuts
        + Detection signal to noise (S/N): S/N>5; this S/Nis from the ZOGY point source match-filtered image;
        + Photometric S/N>5; based on an 8-pixel diameter circular aperture
        + Detection is >10 pixels from an image edge
        + Source elongation (A/B from fitted elliptical profile) 2
        + Ratio of fluxes, R, satisfying: 0<R 1.5 where R=flux in 8-pixel diameter aperture/flux in 18-pixel diameter aperture
        + Number of negative pixels in a 5×5 pixel area 13
        + Number of bad pixels in a 5×5 pixel area 7
        + Absolute difference between PSF and aperture photometry 1mag.
      * Real-bogus classifier that scores individual sources
    - Downstream
      * Classification of objects into various classes
* **Other**
  + [**Deep Probabilistic Imaging: Uncertainty Quantification and Multi-modal Solution Characterization for Computational Imaging**](https://arxiv.org/pdf/2010.14462.pdf)
    - Downstream
      * Estimate posterior distribution of an unobserved image
  + [Gravity Spy: Integrating Advanced LIGO Detector Characterization, Machine Learning, and Citizen Science](https://arxiv.org/pdf/1611.04596.pdf)
    - Preprocessing
      * Chose glitches that satisfied a certain criteria
      * Represented glitches with Omega Scans
        + Uses a combination of sine-Gaussians
        + Scans all Q templates (time-frequency tilings) and identifies the template that gives loudest SNR value
        + Normalize color scale
      * Consulted experts
      * Classified glitches based on morphology of the glitch in its Omega Scan
    - Downstream
      * Categorizing glitches
  + [A transformer-based embedding for the representation of light curves](https://arxiv.org/pdf/2205.01677.pdf)
    - Preprocessing
      * Unlabelled Dataset
        + Discard some light curves that show white noise behavior
      * Labeled dataset
        + Updated labels
      * Set a maximum number of observations for all sequences, padding with zero values if necessary
      * 200 as maximum light curve length
        + Longer than 200 => sample temporal windows starting from a random position
        + Shorter than 200 => used padding
        + Independently subtract mean from each sample
    - Downstream
      * Foundation model for classification and regression
      * Modeling of masked light curves
  + [Transfer learning for radio galaxy classification](https://arxiv.org/pdf/1903.11921.pdf)
    - Preprocessing
      * Used NVSS and FIRST radio surveys
      * Pixel-value re-scaling
        + Zeroed pixel values below 3 \* rms
        + Rescaled
      * Image rotation
        + Augment dataset by rotating image by 1 deg, 2 deg, 3 deg, etc
      * Image clipping
        + Clipped to 150 x 150
    - Downstream
      * Radio galaxy classification
    - Metrics
      * Confusion matrix
      * F1 score
      * AUC