Nancy Grace Roman Space Telescope (Roman)
Technical Report

Title: Roman WFI Co-add Catalog Strategy	Doc #: Roman-STScI-000771, SE-01 Date: 18 April 2025 Rev: -
Authors: Henry C. Ferguson, Eddie Schlafly, Tyler Desjardins, Dario Fadda, Adrea Bellini, Nadia Dencheva, Larry Bradley Phone: 410-338-5098	Release Date: 13 October 2025

Abstract

The Science Operations Center (SOC) for Roman at STScI will be adding together, and re-projecting onto a rectified sky grid, the individual exposures from the Roman Wide Field Instrument (WFI). These co-added images are called "level 3 (L3) data products," and consist of several variants constructed at different times (prompt vs. data-release) and different depth (visits, subsets, and full). The SOC pipeline will then construct catalogs of astronomical objects in these L3 images. The catalogs and associated ancillary data constitute the L4 data products. There are also L4 catalog data products that will be generated from L2 images; these are not discussed in this document.

This technical report summarizes the processing and measurement steps involved in making these catalogs, both for the prompt versions (as the data arrive) and the data-release multi-band versions that will made months later. The purpose is to specify a baseline implementation, identify gaps where research and development work is needed, and identify possible improvements or extensions beyond the baseline implementation. The document refers to sources for algorithm or code for the various steps. Companion documents describe the catalog schema and database architecture.

Operated by the Association of Universities for Research in Astronomy, Inc., for the National Aeronautics and Space Administration under Contract #80GSFC19C0054

Table of Contents

1 Introduction

For *Hubble* and the *James Webb* Space Telescope (JWST), most of the measurements to produce catalogs have been carried out by individual investigation teams. Many users do some custom processing and stacking of the images before making their catalogs. Most observations taken with *Hubble* and *Webb* are dithered – with the telescope pointing moved a small amount between individual exposures. This mitigates the impact of any detector artifacts on the final stacked images that are created by coadding all the exposures – or specific subsets of the images – in a given photometric band. The wide variety of observing strategies for Hubble and Webb have made it impractical to develop a single pipeline that can produce image stacks that would be considered of "science quality" for many of the programs. Among the reasons this is impractical are: (1) the narrow fields of view make image registration challenging and while advances have been made to the process, precisely registering the images before stacking often requires human intervention; (2) the observation structure specified in the proposal planning system does not always map in a straightforward way to a specification of which exposures to combine into which kinds of image stacks; and (3) different observing strategies often require different approaches to outlier rejection and artifact removal. The Hubble Legacy Archive and Hubble Source Catalog now address some of these issues for *Hubble*, but it is still the case that it is often necessary for investigators to return to the individual images and do custom processing.

Roman differs in that it is a survey mission with a wide field of view and a very high data rate. It will be impractical and generally unnecessary for each investigation team to make their own custom image stacks. The wide field of view and stable focal plane means the SOC pipeline can automatically register the individual exposures sufficiently well to meet the astrometric science requirements using *Gaia* reference stars. With a smaller variety of observing modes and observing strategies, it is practical to develop an approach for making stacks of various depth that will probably meet the needs of most science. The approach generally follows standard practices for detection and photometry of stars and galaxies in relatively uncrowded fields. The SOC will of course need to be very responsive to community feedback to ensure that this is the case throughout the mission.

Starting with science-quality stacks, it is now feasible for the SOC pipeline to produce science-quality catalogs. These are intended to have astrometric, photometric, and morphological measurements that are sufficiently accurate and precise for many different investigations. However, they are not specifically aimed at meeting the science requirements of Roman Core

Community Surveys. The Project Infrastructure Teams have been created to achieve that. These teams will be doing custom processing on the Roman data that is focused on their specific science goals, as well as making use of data from other facilities (e.g. Rubin) to augment the Roman data.

The Roman science requirements related to the SOC object catalogs are listed in Appendix 1.

The general community is likely to use a combination of the PIT-provided data products and the SOC-provided data products. The SOC is formally responsible for producing general community-oriented products from the WFI imaging. The PITs are responsible for ensuring that Roman achieves the core science objectives specified in the requirements, but are not responsible for providing general-purpose data products.

The L3 images are relatively small (roughly 5000x5000 pixels) "skycells" that overlap adjacent sky cells by a small amount. These skycells tile larger "projection regions" that each cover about 10 square degrees. All of the skycells within the core portion of a projection region share the same tangent point, making it relatively easy to create large mosaics within a projection region. Figure 1 illustrates relationship between the projection regions and the sky cells. The initial generation of the catalogs is entirely atomic within each skycell: it does not require any of the adjacent skycells or even any metadata about them.

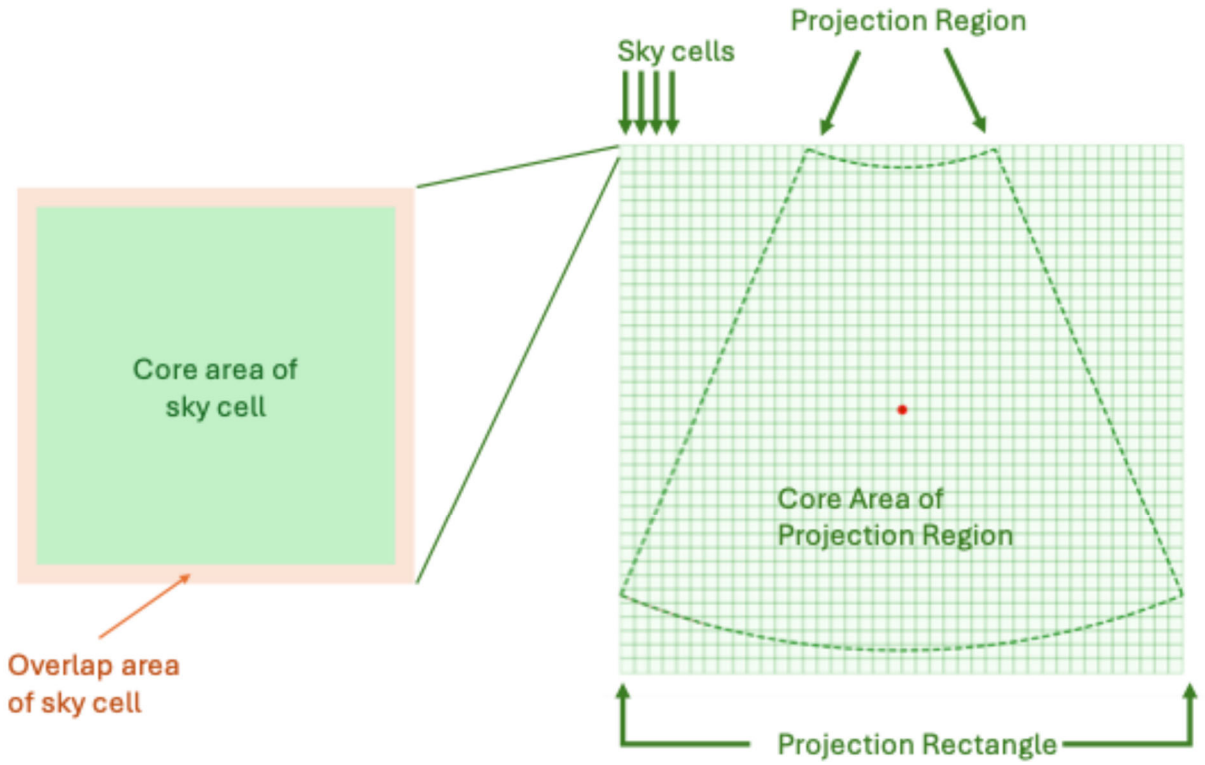

Figure 1. The Roman sky tesselation strategy defines projection regions with boundaries that are lines of constant right ascension and declination, with centers determined from a specific HEALPIX tiling specification. The projection regions are subdivided into a grid of square sky cells are projected onto the

same tangent plane. The sky cells are about 250 arcseconds on a side. The projection regions at the equator are roughly 2 degrees on a side. The L3 images will not be generated when a sky cell is completely outside of its associated projection region (because that same area will be covered by a different tangent-plane projection).

The catalogs for the individual skycells are merged downstream, removing duplications between the overlapping regions. This is discussed further in section 14.

The prompt versions of the L3 images are produced as the data arrive. The current plan is that prompt L3 images will be co-additions of only the exposures within a single visit. This is because it is difficult to anticipate when any other overlapping exposures will arrive. A visit will often have observations at several dither positions, making it valuable to have a co-added image with some additional outlier rejection. The L4 prompt catalogs will be generated for each L3 image as soon as the L3 image is available. These catalogs will be completely independent for each photometric band, with different numbers of detected sources. The prompt versions will not have PSF-matched multi-band photometry, and because they are monochromatic, they will not have photometric redshifts. Furthermore, especially early in the mission, the WFI calibration may change over the course of an observing pass. Different prompt catalogs from the same survey may have different calibrations due to a better understanding of the instrument and/or due to physical changes. Because of the compromises that enable prompt distribution, the L4 prompt catalogs will be archived as individual files but will not be merged and ingested into survey-wide databases. These individual skycell catalogs will saved in Apache parquet format. That same data format is used in other contexts with Roman, so to be specific, we will refer to these henceforth as ***skycell-parquet*** files, which emphasizes they contain data for only a single skycell.

The data-release L3 images will have been reprocessed with consistent calibration across the entire survey dataset. The L4 data-release catalogs will include multi-band PSF-matched photometry based on a single detection image. These catalogs, while originally written into skycell-parquet files, will be de-duplicated across the overlapping skycell and projection region boundaries and merged into survey wide databases available from MAST.

The rest of this document describes the steps in preparing the data for cataloging, doing the source detection, and doing the various measurements. This document focuses primarily on the processing that relies on access to the pixels. There are a variety of quantities that can be derived from these raw measurements (for example converting spatial quantities from pixels to celestial coordinates or converting fluxes from nanoJanskies to magnitudes). These are standard conversions, so are not discussed much here.

The table below lists the processing steps and identifies a baseline approach and possible extensions. This serves as a summary of the more detailed discussions in the sections to follow. From the perspective of software maintenance it may be helpful to have these different steps implemented in separate modules, which can accept inputs either in memory (to save I/O when run as one pipeline) or as files (for running stand alone). Output could be written out as files when run standalone or kept in memory when there is no need to save the intermediate product in the full pipeline. Figure 2 gives a high-level overview of the pipeline data flow.

Step	Baseline	Extensions
Background Subtraction	Iteratively mask sources and spline-fit.	Better treatment of diffraction spikes, ghosts, diffuse wings of galaxies and bright stars and diffuse nebular emission. Use an existing sky map for initial source masking. Scene-dependent decision tree. Machine learning (ML) approaches.
Uncertainty Array Construction	Use the L3 uncertainty array and provide the data necessary to calibrate the resulting bias in detection with artificial source injection.	Create an L3 data product that accurately predicts the variance at the sky level accounting varying sky levels in the individual L2 images, gaps between them in a given skycell, outlier rejection, and large-scale sky non-uniformities.
Object Detection	Use all of the available bands for detection, optimized for a single assumed spectral-energy distribution. Convolve with a single kernel.	More capable decision trees. Give General Astrophysics Survey (GAS) programs the ability to specify the detection band. Multi-kernel detection strategies. Multi-SED detection strategies. ML approaches.
Object de-blending	Photutils watershed algorithm.	Scarlet ML approaches
PSF matching	A single PSF for each band for each survey. Match using photutils routines. PSF matching is only relevant for the data-release catalogs.	Homogenize the PSFs when building the L3 images. Build PSFs when combining the L2 images into the L3 image; multiple PSFs per skycell. PSFs could be extensions in the L3 ASDF file. Use these PSFs for making the kernels. Do the PSF-matched photometry on the L2 images and combine results in catalog space.
Point-source photometry	Photutils PSFPhotometry	Use the L2 images (merging with the forced-photometry time-domain processing).
Extended source photometry	Photutils	Petrosian flux(es). Magnitudes and asinh magnitudes for all apertures.
Non-parametric shapes	Photutils	Petrosian radii M20 parameter Hirata-Seljak-Mandelbaum (HSM) Adaptive Moments ML latent vectors
Parametric shapes	None	Sersic or Spergel parameters Bulge/disk decomposition
Photometric Redshifts	RAIL + LePhare Photometric redshifts are only computed for data-release catalogs.	RAIL + other SED-fitting algorithms RAIL + ML-trained catalog-based algorithms ML image-based algorithms

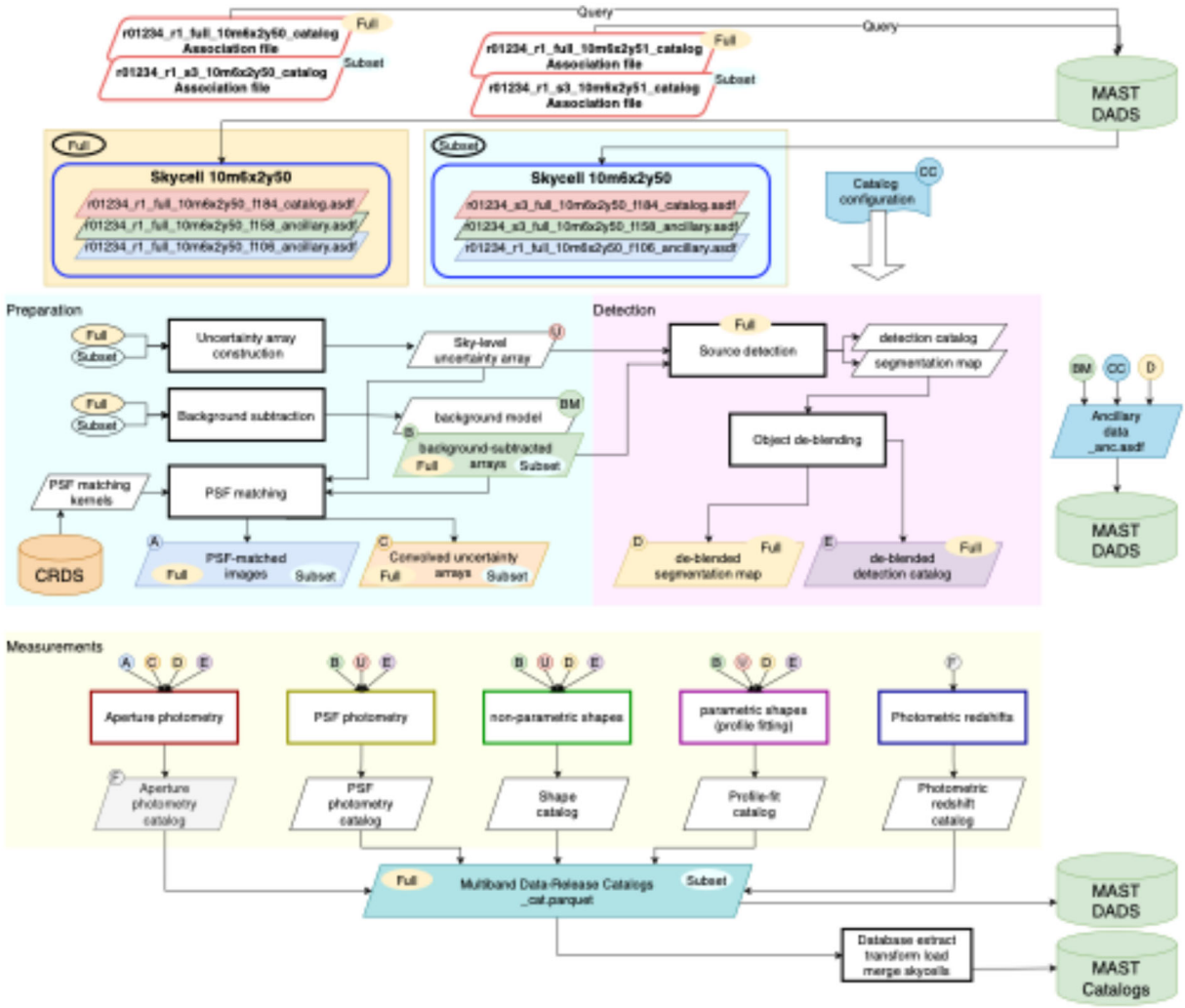

Figure 2: Overview of the catalog processing for the L3 co-added Roman WFI images. The process starts with creating associations that identify the L3 files in the different bands for the same skycell from the same subset of the L3 processing. The images are all prepared by constructing uncertainty arrays and subtracting background. PSFs and PSF kernels relevant to the different bands are retrieved from the Calibration Reference Data System (CRDS) and PSF-matched images made. The source detection and deblending proceeds on the full co-adds. The measurement steps (in the yellow box) use the segments and object centers and Kron parameters from the source detection. The catalogs are merged and ingested into queriable databases in MAST. The ancillary data (segmentation map and background information) are stored in an ASDF-format file for each skycell served from the MAST Data Archiving and Distribution Service (DADS).

1.1 Baseline vs. Extensions Philosophy

This document aims to specify a baseline cataloging strategy that can be developed on a tight schedule, meet the requirements, and deliver catalogs suitable for a wide variety of scientific investigations. Simple extensions to this baseline include adding more measurements. Some of these are measurements that are less frequently used, but might be convenient to have in the catalogs. Others are measurements where the code exists, but is not implemented in the photutils

package (1) (planned for use in generating the baseline catalogs). Still others require research and development and may require calibration before the first data release.

Guiding principles for development of the cataloging software are to (1) to keep the steps modular to facilitate future extensions and (2) to have configurable parameters specified in reference files in the Calibration References Data System (CRDS; as is the case for earlier steps in romancal), so that users can easily tweak even the baseline cataloging strategies without having to modify the source code.

1.2 The Inputs: Description of the Level 3 Images

The Roman WFI Level 3 (L3) images are co-added from the individual exposures (the L2 images). The coaddition process involves the following steps:

- Matching sky backgrounds between the L2 images
- Recording the sky levels in the header
- Rectifying and resampling the images onto the standard L3 skycell projection
- Making a rectified stack of the images to identify outliers
- Drizzling the images, resampling them to the desired final pixel scale, and co-adding them with the appropriate weights
- Creating a context array to record which L2 images contributed to each L3 pixel
- Creating an uncertainty (err) array by combining the L2 uncertainty arrays with appropriate weights. This array is the quadrature sum of the Poisson shot-noise uncertainty (including sky and source), readout noise and flatfield standard deviation.
- Creating a "weight" array. This accounts for different exposure times and sky-background levels when doing the weighted co-addition to make the L3 images. Typically, this would be the inverse variance of the sky background. In practice, for JWST, it is not.

The resulting L3 images have the following arrays: data, uncertainty, context and weight – designated as SCI, ERR, CON and WHT, respectively. The JWST equivalent of the Roman WFI L3 images are the i2d files with more arrays: SCI, ERR, CON, WHT, VAR_POISSON, VAR_RNOISE, and VAR_FLAT, with the last three of these representing the variance due to Poisson counting statistics, readout noise, and statistical (not systematic) uncertainties in the flat field reference file.

1.3 The Outputs

The outputs of the cataloging process are a set of tables and ancillary data files. The tables of "raw" measurements from the pipeline will initially be stored in skycell-parquet files. There will be separate ASDF files for each L3 data product. The skycell-parquet files for one skycell will be separate from those of all the other skycells. Because there are overlaps between skycells, some of the objects will be measured twice. Treatment of these duplications is discussed in section 14. Catalogs for different steps of the pipeline may initially be written into separate intermediate files (depending on how modular the different steps are), but the current concept is that these will be merged into two distinct files as a last step before passing to the archive. The files, and their contents, are:

- [filter]_cat
  - All of the tabular data. For prompt catalogs, these will be separate for the separate filters. For the data-release catalogs, the measurements for the different filters will be together in one file.
- [filter]_anc
  - Ancillary non-tabular data and secondary tables. This could include:
    - An image or parametrized model of the sky background
    - segmentation image accompanying the detection image
    - The detection image itself, or a significance image derived from it
    - Metadata associated with the L2 and L3 files used for this skycell.

The L3 filenames are based on the following template:

r<ppppp>_<L3 product identifier>_<exposure grouping>_<skycell coords>_<optical element>_coadd.asdf

The first field is the program ID. The second is **p** for the prompt products and **r#** for numbered data releases. The third field is the visit number for the visit co-adds, is **full** for the coadds of all available data, and is a string identifying the subset of data for other instances (e.g. **s1** for the first subset, which would have to be documented in the data release). The skycell coords are decimal degrees with a format such as 274p63x31y40 for the projection region centered at roughly 274 degrees RA, +63 degrees Dec, and for the skycell at grid position 31,40 within that projection region.

Example for a full stack skycell L3 file from the first data release of program 12345:

r12345_r1_full_274p63x31y40_f106_coadd.asdf

The file naming convention for L4 files could evolve, but will be based on the L3 product naming to the extent practical. The rootname would follow the L3 convention:

r<ppppp>_<L3 product identifier>_<exposure grouping>_<skycell coords>

For prompt products, this will be followed by an underscore and the filter name (e.g. f106, f129, etc) and type of L4 file (cat or anc). For the data releases, the filters will be combined in one file, so only the file type (cat or anc) is needed.

2 Background Subtraction

The current version of the romancal pipeline performs background matching when combining the individual level 2 (L2) exposures to create a level 3 (L3) co-added, rectified image on a tangent-plane skycell projection. The background levels are recorded in tabular form as metadata in the L3 image header. For images combined from different epochs, the zodiacal background level will be different in the different L2 images. The background matching procedure will take care of this variation to first order, but will not correct for the small gradient in the zodiacal background across each sky cell.

Background subtraction is an important and difficult (often science-dependent) step. If the objects of interest for science are many arcseconds or even arc-minutes in extent then standard background estimation techniques for "blank fields" will adversely affect these objects. On the other hand scattered light patterns (including diffraction spikes) are features with a large extent that ideally should be subtracted before searching for faint stars and galaxies. In crowded stellar fields the wings of the PSF affect the ability to detect and measure nearby stars. If one can assume that all the sources (of interest) are point sources, then an iterative approach to photometry and background estimation can work well. In crowded fields with diffuse emission (e.g. in the disks of nearby galaxies) more complicated iterative approaches are needed. More complex approaches are also needed in clusters of galaxies, where many of the galaxies overlap and there is diffuse light as well.

2.1 Baseline

The baseline approach is similar to the default adopted by SExtractor. The median is computed in 100x100 pixel blocks of the image, after rejecting pixels that are 3-sigma outliers. This low-resolution median map is then median filtered over patches of 3x3 blocks, for an effective smoothing length of 300 pixels. A bi-cubic spline interpolation is used to "zoom" the low-resolution background estimate back to the original pixel scale.

2.2 Extensions

The baseline approach will generally oversubtract the background around bright objects in the image because they influence the median in the patches used to estimate the varying background levels. The first extension would implement a technique to mask out the sources that has been recently developed and tested on a few JWST images.

The improved background subtraction in section 2.2.1 will probably work acceptably for high-latitude fields. It has not yet been tested in crowded fields. Issues to address for further extensions include the following:

- Diffraction spikes and ghosts from bright sources (2.2.2).
- Diffuse wings of galaxies, including intracluster light (2.2.3).
- Overlapping wings of the point-spread function in crowded stellar fields (2.2.4).
- Regions of diffuse nebular emission in Galactic or extragalactic fields (2.2.5).

2.2.1 Identifying Sky Patches

This approach adopts a relatively robust way to identify regions of relatively "clean sky" between the sources in the image, without having to identify the sources individually first on top of the potentially varying background. A strategy that has worked well in a prototype is to measure Moran's I statistic in square patches on the image and designate as "sky" those patches that are below a certain threshold in this statistic. Moran's I is a measure of the spatial autocorrelation in the image. In the absence of correlated sky noise, $I \approx 0$. I will be significantly larger if a star or galaxy is within the patch, or there is a steep gradient across the patch. The definition of I is:

$$I = \frac{N}{W}\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

where $x_{i,j}$ are the intensities in pixels i,j within the patch, w is a set of weights with zeros on the diagonal, which can be tuned to favor certain patterns of correlation, and W is the sum over the weights. The weights can be thought of as a kernel. For testing, we have simply used:

```
kernel = np.array([
[1, 1, 1],
[1, 0, 1],
[1, 1, 1] ])
```

In testing, a patch of ~20x20 pixels with a threshold of I < 0.3 for identifying "sky patches" has worked well for JWST images. One relatively robust way to fix the threshold is to set it so that a fixed fraction of the patches are masked and the rest are used to fit the background. Masking the 60% of the patches with the highest values of the Moran's I statistic works well for the JWST NGDEEP images in all bands and both epochs. Having masked out all portions of the image other than the sky patches, the photutils Background2D procedure can be used to interpolate smoothly between the patches. For testing on JWST NIRCam images, we have created a source mask (True where I > threshold) and used the photutils routine with the default "zoom" spline interpolator and the following parameters:

```
Background2D(
    image,
    box_size      = 20,
    filter_size   = (3,3),
    mask          = source_mask,
    sigma_clip    = SigmaClip(sigma=3.0)
    bkg_estimator = MedianBackground()
  )
```

Figure 3 below shows the masking and background fitting via this procedure for a portion of the NIRCam JWST NGDEEP Deep-field F115W co-added image with particularly bad scattered-light background.

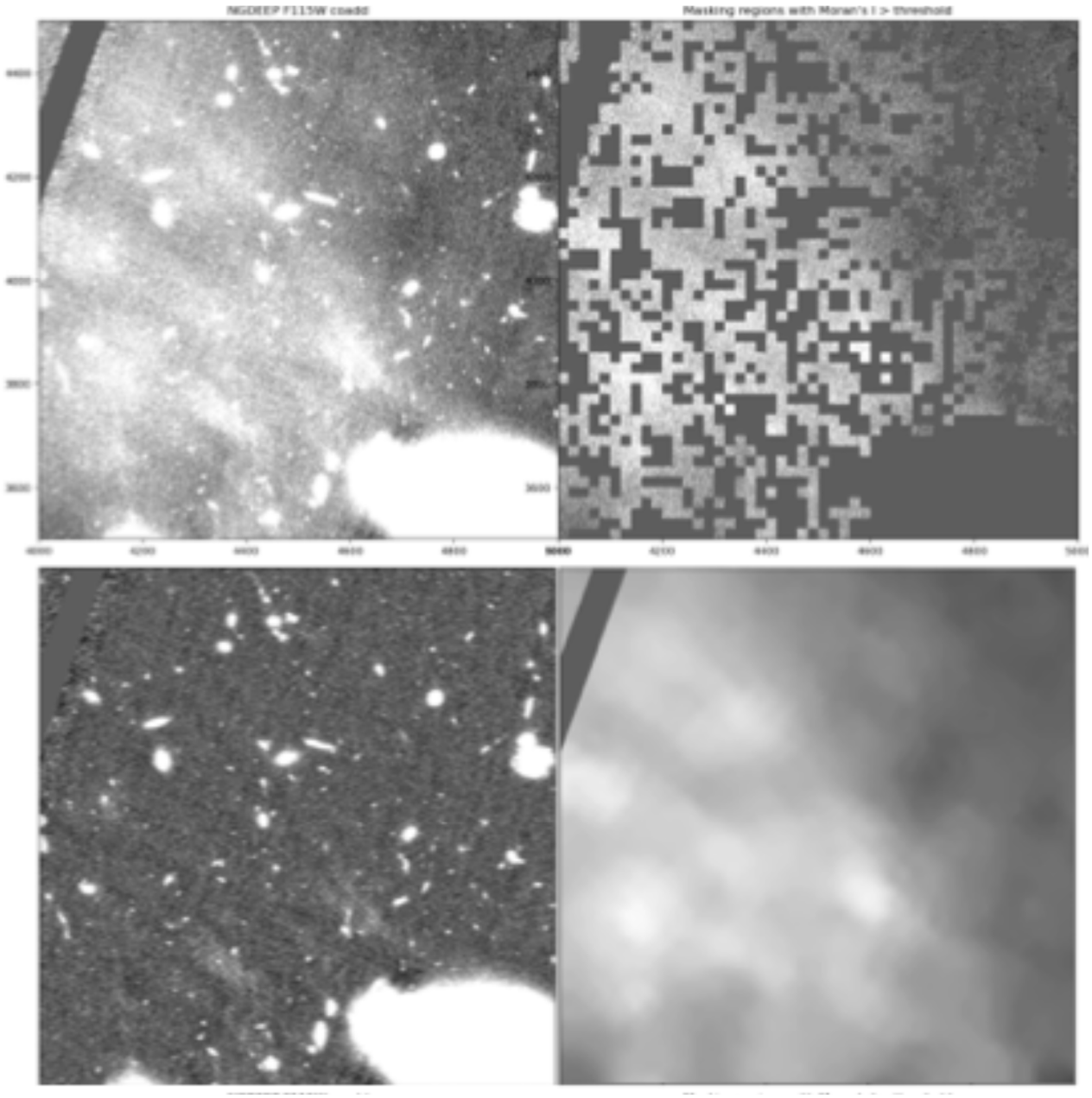

Figure 3: : A portion of the NIRCam JWST NGDEEP Deep-field co-added image with particularly bad scattered-light background. The upper left panel is the original co-added image, even after some scattered light removal was performed on the individual exposures. The upper right panel shows the masking of 20x20-pixel patches that have the values of the Moran's I statistic above a fixed threshold. The lower-right panel shows the fitted background model. Note the lack of enhanced background associated with the galaxy (which was masked as a source), while the scattered light is reasonably well modeled. The lower left panel shows the result of subtracting this sky background – removing most of the scattered light while not over-subtracting the wings of galaxies.

2.2.2 Diffraction spikes and stray light (ghosts, etc.)

The positions and colors of relatively bright stars are already known from all-sky surveys. One of the funded Wide-Field Science (WFS) programs – *ROSALIA: ROman Sky Analyst for Low surface brightness Imaging & Astronomy* – is aimed specifically at modeling the effects of scattered light for Roman. The current version of ROSALIA automatically queries Gaia /

2MASS / WISE to estimate a pseudo-SED for every star and their position relative to the WFI field of view. Future work will include scattered light and diffraction spikes at up to large separations from the stars. Even if this model does not work well enough to fully subtract the diffraction spikes, it will almost certainly be good enough to flag pixels where they might be a source of contamination for the catalogs.

The simplest approach is to build a model of the diffraction spikes and scattered light for every L2 images. Step by step:

1. Locate all the bright sources beforehand from GAIA/2MASS/WISE, etc. catalogs.
2. Using information on the brightness and SED of each star, create an image that is the predicted flux of the diffraction spike in the Level 2 images in each band.
3. Co-add those with the same weights as used when creating the actual L3 images, but with no outlier rejection.
4. If the predictive model for scattered light and diffraction spikes in general does a good job, subtract the model as part of the background removal.
5. Create a flag image that has a series of levels depending on the brightness of the diffraction spike and stray light features in this coadd.
6. When making a catalog, propagate the "worst" flag value through to flag the sources that are potentially affected.

2.2.3 Wings of galaxies and intracluster light

The wings of galaxies and intracluster light are best thought of as a "de-blending" problem rather than a background-subtraction problem. Ideally the light from the wings of galaxies would be assigned back to each galaxy and the intracluster light would be considered as a single object. This will be discussed more in the de-blending section.

However, from the perspective of background subtraction, a fairly simple extension would be to use an existing all-sky catalog (e.g. from 2MASS, *Gaia* and/or WISE) to help in masking out bright, diffuse objects before estimating the Roman sky background.

2.2.4 Overlapping wings of the point-spread function

In crowded stellar fields, an iterative approach is probably desirable, finding, fitting and subtracting progressively fainter sources, while masking regions where the residuals from the PSF-subtraction are liable to bias the background estimate. The "sky patch" approach for identifying relatively clean regions outlined in section 2.2.1 may be useful as part of the procedure.

2.2.5 Regions of diffuse nebular emission

Galactic dust has been mapped to about 15″ resolution (2). With (perhaps substantial) calibration effort it may be possible to use such dust maps to guide a predictive model of the background in the Roman images due to scattering from diffuse Galactic dust. While probably not precise enough for a final background subtraction, this may be helpful for the first step in background subtraction prior to masking sources. Steps in this direction have been made by (3).

For Galactic nebular emission, there is available data for the Galactic plane from from IPHAS and VPHAS+. The data products do not include a map specifically of the diffuse Hα emission, though.

Making catalogs within regions with a mixture of stars, galaxies, nebular emission, and dust obscuration is a more specialized problem. Different scientific goals require different approaches. The baseline approach mentioned above will likely subtract out the relatively smooth nebular emission as background, but leave in small scale structures that might be identified as multiple objects. Fitting those objects with point-spread functions or galaxy profiles will be of little scientific value.

A possible general extension for making better catalogs in such regions would be to develop a decision tree that classifies the scene in each WFI image based on pre-existing knowledge of the sky. The background subtraction and all subsequent steps in generating the catalog could be tailored to the kind of scene. Examples of types of scenes might be "high-latitude field," "Galactic plane," "Crowded stellar field," "Nebula," "Nearby galaxy," etc..

2.2.6 Machine learning

There are a variety of machine learning approaches (often characterized as "de-noising") for training an algorithm to recognize noise in an image and separate it from signal. One would need to be cautious in applying this approach to astronomical scenes from Roman, because the noise properties will vary from image to image (e.g. due to 1/f noise, scattered light, varying background levels, or varying cosmic-ray rates). Nonetheless, deeper images from JWST, along with entirely simulated Roman images, can provide a variety of "truth" images that might enable one to make the algorithm robust. One can in principle take a more nuanced approach than identifying pixels ask "sky" and "non-sky," training the algorithm to return the relative contributions of source and sky to every pixel. This might ultimately be useful for producing mimimally sky-subtracted scenes that leave in a lot of the blended diffuse light from the sources, which can then later be accounted for by advanced de-blending techniques.

3 Uncertainty Array Construction

The algorithm used for image segmentation by SExtractor and Photutils identifies groups of connected pixels that are above a threshold. This threshold is typically a fixed multiple of the root-mean-square (RMS) fluctuations at the sky level. The RMS sky fluctuations will vary across the L3 skycell in the general case because (for example) different L2 images contributed to different parts of the L3 image. The standard procedure is to make an RMS image that predicts the fluctuations in the absence of any sources, and then use this to modulate the detection threshold. (Photutils expects a threshold image, which can be a fixed multiple of the RMS image.)

For *Hubble* high-latitude imaging programs, the standard procedure for making the RMS image has been to calculate the variance at the mean sky level of each exposure from the read noise, an estimate of the mean sky background count rate, and the exposure time. The individual exposures are then stacked into a co-added exposure, weighting by the inverse of this variance.

One of the output arrays is the inverse variance of the resulting stack, accounting for pixels that were rejected as outliers or portions of the image where not all of the individual exposures contributed to the stack.

For JWST, the pipeline produces something similar to this, but not identical. For each input image, it takes the square root of each of the read noise variance array to make an error image. It drizzles the read noise error image onto the output WCS, with drizzle parameters matching those used for the science data, then squares the resampled read noise to make a variance array. For the standard "ivm" weight_type option, it then uses the inverse of these variance array to weight the individual exposures when combining them. The resulting "inverse variance" is in the WHT array of the co-added image (i2d) data product. The readout variance is roughly proportional to the exposure time, so if the sky level is not varying between the individual exposures, this will come close to achieving the relative weights one would get if the sky background were included. However, if the sky background is varying (as it might, for example, if exposures with different zodiacal background levels are combined), then the relative weighting is not quite right. More importantly – from the perspective of creating a threshold image – because it does not include the expected variance of the sky background in the calculation, the JWST WHT array does not directly predict the inverse variance at the sky level.

A common procedure for JWST has been to identify background regions in the images and empirically rescale the weight array so that its inverse predicts the variance at the typical sky level of the images. There are several complications in doing this. (1) This step is ideally done before source detection, but it is difficult to mask the sources before detecting them. (2) Prior to background subtraction, the varying background levels mean that the variance on large scales in the image is much larger than predicted from the variance on small scales. Finally, (3) on small scales of a few pixels, the resampling procedure for creating the L3 images introduces correlations between neighboring pixels which mean that the measured pixel-to-pixel variance is suppressed relative to the true variance.

3.1 Baseline

The current baseline algorithm for creating the output weight array for the L3 co-adds uses the var_rnoise array from the L2 images, as has been done for JWST. This accounts for the change in the effective exposure time in each pixel due to the outlier rejection step in up-the-ramp filtering. Therefore it is getting the desired relative weights of the different input images right, even though it is not actually weighting by the inverse variance at the mean sky level. This offset from the true inverse variance affects the threshold used for considering that a source detection is significant. It does not impact the photometric uncertainties because these are derived from the error array, which includes the Poisson term from both the sky and the sources. This JWST baseline is not ideal because it will result in different effective signal-to-noise (S/N) detection thresholds for different observing strategies.

3.2 Extensions

3.2.1 Using sky patches on the L3 images to correct the weight map to predict the sky variance

The background-subtraction algorithm described in section 2.1 identifies patches of the image that are likely to be free of sources. This suggests a simple and practical way to scale the weight map so that it predicts the inverse of the variance at the sky level – simply measure the variance between these patches, having subtracted off the background to remove the large-scale variations. If the patches are larger than a few pixels on a side, the variance between the block-summed patches (rather than between the pixels within the patches) is a good representation of the sky variance.

The inverse of the weight map, when block-summed to the same gridding and measured in the same patches should be equal to this measured sky variance. Comparison of these two estimates of the sky variance gives us the constant multiplicative correction to the weight map to have it predict the measured sky variance.

3.3 Other Extensions

It is possible to apply use the same strategy for identifying source-free sky patches, outlined in section 2.1, to the individual L2 images prior to combining them to form L3 images (this could either be done as part of L2 processing or an early step in L3 processing). In that case one could simply use the pixel-to-pixel rms within the patches create a weight map by requiring that the measured variance be the sum of the Poisson variance at the median sky level and the variance due to readout noise. As long as the background is not varying enough to affect the sky noise significantly, this gives a way to create an inverse-variance map that accounts for the masking of cosmic rays and bad pixels but does not include the contribution from the sources themselves. Basically a var_poisson array could be created that does not include the contribution from the sources (being either a constant, modulated by the number of readouts used for the ramp fitting in each pixel – which itself is proportional to var_rnoise – or derived from a smooth sky-background fit to the L2 image). This can be added to var_rnoise to predict the variance at the sky level in every pixel. The inverse of that is the weight map to be used for making the L3 co-adds.

Yet another possibility for identifying and masking sources (other than the Moran's I statistic approach outlined in section 2.1) might be to use a pre-existing catalog, e.g. from *Euclid,* to mask all known sources without actually measuring the image.

4 Detection Image Creation

The standard approach to image segmentation is to convolve with a matched filter. The ingredients are

- the image, *I;*
- an image that represents the inverse of the expected variance at the sky level – *i.e.* the weight map described in the previous section $W = 1/\sigma^2_{sky}$; and
- a normalized two-dimensional spatial profile of the source of interest, *P*.

We can then construct a map that represents the significance of detecting a source – with a profile that is the same as the kernel – above the (assumed normally-distributed) sky noise:

$$S(x,y) = \frac{\mathrm{Conv}(I \cdot W, P)}{\sqrt{\mathrm{Conv}(W, P^2)}}$$

where *Conv* represents the convolution operator. This is the algorithm implemented in SExtractor (4).

One way to extend this to use multiple bands is to assume a single spectral-energy distribution for the source of interest, such that the predicted relative flux in each band is f_B. The matched filter for the shape of the source will also in general be wavelength dependent: $P \Rightarrow P_\mathrm{B}$. We can construct a significance image S_B for each band independently using the appropriate normalized kernel P_B for that band. This would typically be the profile of the source of interest, convolved by the band-specific point-spread function.

The combined significance map would weight the significance images for each band according to the expected spectral-energy distribution:

$$S_{\mathrm{combined}}(x,y) = \frac{\sum_\mathrm{B} f_\mathrm{B} \mathrm{Conv}(I_\mathrm{B} W_\mathrm{B}, P_\mathrm{B})}{\sqrt{\sum f_\mathrm{B}^2 \mathrm{Conv}(W_\mathrm{B}, P_\mathrm{B}^2)}}$$

The challenge for making galaxy catalogs is twofold: (1) galaxies have a wide variety of shapes and sizes, and (2) galaxies have a wide variety of colors (spectral-energy distributions; SEDs). The former means that one should in principle use a wide variety of detection kernels. The latter means that one should in principle detect in multiple bands and with a variety of assumed SEDs. In both cases, one could weight "significance" of a detection by a prior to optimize recovery of the sources of interest while suppressing spurious source identification. A variety of approaches are possible. Some are relatively modest extensions to deal with the "multi-scale" problem that galaxies have a wide range of sizes. Others incorporate different strategies for separating objects from background or deblending overlapping objects. Masias et al. 2012 (5) review various options and that review article is cited by many of the strategies developed since then.

The SOC catalogs are to be accompanied by the results of source injection and recovery. Therefore, while this makes it possible to characterize the detection efficiency and measurement biases for complex detection schemes, it can become quite expensive. For galaxies, this characterization is multi-dimensional with the two most important parameters being the total flux and the galaxy size in the detection band. The other dimensions include the ellipticity, shape of the surface-brightness profile, and SED (if the detection uses multiple bands). It becomes rapidly infeasible to characterize completeness and measurement biases across all dimensions. The standard shortcut is adopt distribution functions that are thought to be representative of the true distribution in all dimensions other than flux and size, essentially marginalizing over all of these other parameters.

4.1 Baseline

4.1.1 Source Detection

The SOC pipeline is intended for all WFI imaging data, serving a broad array of science needs. The selection of filters the distribution of observing time between the filters will undoubtedly vary from program to program. An expedient baseline for a general detection strategy is to do the following:

1. Use a single Gaussian detection kernel that is the same in each band. FWHM ~ 0.3" can be optimized based on simulated data.
2. Use all of the bands available for the skycell.
3. Combine the significance maps to optimize for flat spectrum objects (f_ν = constant).

The computation of the weight for (3) involves computing the expected count rate in each band (i.e. takes into account instrument sensitivity). For data releases, the source detection will be done on the full depth co-add (L3 images) of all the available exposures for that data release. The properties of the sources (positions and properties of the apertures) will be used to drive the photometry for the "subset" L3 data products. In other words, there will not be an independent detection for the subsets.

The baseline plan will do source detection individually for each band for the prompt data products.

4.2 Extensions

4.2.1 Detection Kernels

The most expedient baseline is to use a single detection kernel. For fields dominated by stars, the PSF is the optimal kernel. From the JWST CEERS observations, the median FWHM reported by SExtractor for galaxies with magnitudes $26.25 < AB_{F150W} < 26.75$ is 0.3″. This is ~3 times larger than the PSF FWHM in the F146 filter. The galaxies will have different axial ratios and position angles, making it impractical to truly match the kernels. However, detection using a round kernel with a Sersic profile with the same half-light radius can be significantly more efficient at detecting these small galaxies than using a PSF kernel – gaining nearly 1 magnitude in limiting depth for this example. For galaxy detection it is common – even for HST and JWST, where galaxies are reasonably well resolved – to use a Gaussian kernel. So the simplest baseline implementation will be to use a Gaussian kernel. Tests on simulations can be used to decide on the optimal FWHM.

The SOC catalogs will be used to drive spectral extraction, which in turn provides the redshift estimates for the Baryon Acoustic Oscillation measurements. The galaxies with spectra bright enough to provide redshifts are much brighter than the imaging detection limit, even if a PSF kernel were used. Weak-lensing measurements also rely on galaxies well above the detection limit, because reliable shape estimates are needed. Supernova host galaxies will need special treatment in any case, due to the influence of the supernova. Optimization of the detection of the faintest sources is driven more by general astrophysics than by the Roman core cosmology

programs.

4.2.2 Using multiple kernels

A relatively straightforward extension – already possible in the romancal code base – is to create the significance images S_{kernel} separately for several different kernels and then take the maximum of $S_{kernel}(x,y)$ for the detection significance at each pixel. It is important to keep the number of spurious sources low relative to true sources; this may be difficult to achieve when taking the maximum of signficance images. A simpler approach to using multiple kernels might be to detect on the separate convolved images and combine segmentation maps as the next step.

Adding a few circular kernels that represent typical galaxy profiles – convolved with the PSF at each band – would improve the detection efficiency for faint, extended galaxies. These will probably still be relatively small kernels of less than an arcsecond half-light radius. For larger galaxies, the images will often be blended with neighbors, so a better detection approach might be to do an optimized search for extended low-surface brightness objects after subtracting models for all of the stars and compact galaxies. Such an approach is more ambitious than current concepts for the SOC pipeline.

Once the images are segmented it is possible to "grow" the segments to include more of the diffuse light associated with each object using a morphological dilation of the segmentation map. In the simplest baseline we would not do this because the amount of dilation needs to be tuned in some principled way.

Another challenge for detecting low-surface-brightness galaxies is that they are often broken into non-overlapping segments in the detection and de-blending step, even when the detection kernel is a good match to the galaxy profile. A careful census of low-surface brightness objects will require a step to combine nearby segments into a single segment before performing subsequent measurements.

4.2.3 Keeping larger segments rather than maximizing significance for each pixel

Rather than taking the maximum $S_{kernel}(x,y)$, one could carry out the source detection and – more importantly – the deblending step separately for each kernel. This would make it possible to also adjust the minimum number of connected pixels and the de-blending parameters separately for the different detection kernels. The segmentation images can subsequently be merged, keeping the largest segment in the case of overlaps. This might help prevent larger galaxies from being broken into many small pieces. It is essentially the "hot/cold" strategy that has often been used for high-latitude surveys with Hubble and Webb.

4.2.4 Using multiple SEDs

Similarly to extending the detection strategy for multiple kernels, one could consider computing w_B and $S_{combined}$ for different assumed SEDs, and then taking the maximum of $S_{combined}(x,y)$ for the different weights as the significance at each pixel. This could potentially gain a few tenths of a magnitude of survey depth, but would require significantly more effort to optimize and tune to avoid introducing spurious detections.

4.2.5 Customization for the core surveys

Ideally the detection kernel(s), and SED prior(s) will be specified in CRDS reference files. This will facilitate their customization separately for the core surveys and the Galactic Plane survey. It is likely impractical for SOC personnel to customize each GAS program individually.

4.2.6 Customization based on position in the sky

Another relatively straightforward way to customize the kernel selection and SED-prior selection would be based on position in the sky. For example detection strategies optimized for stars could be used based on a stellar-density (and typical SED near the detection limit) prediction from a Milky Way model. However this strategy would lead to varying kernels and SED priors across the Galactic Plane survey, so may introduce more catalog non-uniformity than desired.

4.2.7 Giving GAS teams the ability to customize catalog generation

For the GAS surveys, we could offer the ability to specify the detection band. This would require a new community interface in addition to the Astronomer's Proposal Tool. The disadvantage of allowing GAS users to specify catalog detection bands is that it will be less straightforward to combine catalogs from different surveys. However that may well be the case anyway if the observing strategies are very different or the scene (e.g. the face of a nearby galaxy) is very different from a typical scene. Enabling this requires specifying the configurable cataloging parameters – as far as is practical – in CRDS reference files.

Developing such an interface would have other benefits, as it could also allow customization of the L3 products. A basic interface would be to provide documentation, instructions, examples and help-desk support to allow the teams to construct the requisite association files and pipeline reference files – exactly as the user would do if they were to run these steps in the pipeline themselves. The only difference is that they would deliver those files to the SOC before the data arrive and the SOC would use them to process the GAS data from that program. Because we are planning to provide documentation, instructions, examples and help-desk support to enable users to run the pipeline on their own, the most important missing technical pieces are delivery and validation & testing methods for the files. It is important that the processing actually work when the data arrive. Developing the validation and testing is the challenging part because this has to be done before the data arrive if we are going to be making prompt catalogs. On the other hand, if prompt catalogs could be on a "best effort" basis for the GAS programs, any problems in these files could be worked out before processing for the data releases.

In the L3 case the association files would specify the exposures to be combined to form different subsets for different co-adds. The reference file(s) could specify the drizzle and outlier-rejection parameters for the co-addition.

In the L4 (catalog) case, the association files would specify the L3 images to use for creating the detection band and for doing the multi-band photometry. The reference files would specify parameters for the cataloging processes and the point-spread functions (or parameters that drive the selection of the point-spread functions) to use in matching the spatial resolution of the bands.

The interface to the users could be relatively simple. These are all relatively small ASCII files. Options include: a web form for uploading the files, a drop-box for depositing the files, or an email address for sending the files. The advantage of the first option is that the server behind the website could do some real-time validation before accepting the delivery.

4.2.8 Other Approaches

Prior to developing a strategy using multiple kernels, it will be important to have specific use cases and simulations to guide the development. For both low-surface-brightness galaxies and galaxies that are much more extended than an arcsecond, the detection algorithm is hard to separate from the de-blending algorithm. This is because (a) most galaxies larger than an arcsecond will have some internal structure which may cause them to be catalogued as multiple sources and (b) the mean separation between galaxies brighter than $H_{AB} = 27$ is about 4 arcseconds, so overlaps are common.

Wavelets and curevelets have been used successfully as an approach to multi-scale detection in faint-galaxy images (6) Curvelets are more optimal for creating sparse representations of the data that can be useful for finding elongated or curved structures (7). Beckouche et al. (2013) (8) explore a dictionary-learning approach to filtering images that offers a different approach for generating sparse representations of the data. Such techniques can be particularly valuable in clusters of galaxies, where the multi-scale nature of the detection algorithm can help with the problem that many of the sources overlap. These techniques are in widespread use on galaxy surveys.

Convolutional-neural-network (CNN) based approaches are also attractive (8). These convolve the images with multiple kernels, optimizing the kernels during training. Existing libraries make these relatively easy to develop and train, albeit difficult to optimize in any mathematically formal way. These machine-learning approaches benefit from GPUs and can run very fast once trained. MORPHEUS is existing CNN-based package that does both detection and classification.

4.2.9 Tuning the segmentation map

As mentioned above, the initial segmentation map can be dilated to include more of the diffuse light associated with each object. While the amount of dilation is somewhat science-application dependent, one approach to deciding whether or not to do this dilation (and if so, by how much) would be to study the effect on the Kron aperture sizes as a function of the detection & dilation process. The Kron aperture sizes affect the choice of the Kron aperture for elliptical-aperture photometry. The Kron-aperture fluxes are then typically used to compute “total fluxes” by applying a magnitude- and size-dependent correction to each galaxy based on artificial source-injection simulations.

A tuning procedure for any dilation might be to try to minimize the scatter in this correction based on tests on simulated images. Dilating the segmentation maps might lead to more stable Kron radii and therefore less “noisy” Kron aperture sizes. Too much dilation probably introduces more noise, so there may be an optimum dilation strategy to reduce the scatter in the estimated total magnitudes.

5 Object De-Blending

5.1 Baseline

Photutils has implemented a multi-threshold watershed de-blending algorithm similar to the one used by SExtractor. This does not use any color information and relies on there being a saddle point between the two sources. The two sources are segmented and assigned different labels. There is no ability to assign a pixel proportionally to multiple sources.

These segments are important for later operations. The moments within the segmented regions are used as rough estimates of source sizes (Kron radii), position angles and ellipticities from the analysis of moments of the light intensity distribution.

Because this is already implemented and widely used in the community, this will be the baseline deblending. A typical set of parameters would be npixels=10, nlevels=32, contrast=0.001, but these parameters should be tuned before the first data releases.

5.2 Extensions

5.2.1 Flagging de-blended objects

Internally, photutils currently keeps track of the original segment labels of the objects before deblending. The deblended_labels property returns a list of deblended labels,
the deblended_labels_map property returns a dictionary mapping the deblended labels to the parent labels, and the deblended_labels_inverse_map property returns a dictionary mapping the parent labels to the deblended labels. However, without carrying out all of the photometry and shape measurements on both the parents and the children separately, this information is of somewhat limited use, and is also inconvenient to store in a database.

The current baseline is to simply set a bit in the warning_flags to indicate if the source was deblended from a parent.

5.2.2 Single-Band De-Blending

At the depths of the High-Latitude Wide-Area Survey (HLWAS), the baseline deblending approach is sufficient to serve a wide variety of scientific investigations. However, there are still many galaxies that overlap that might be better de-blended with algorithms other than the standard watershed approach. This is generally true for faint galaxies located near bright ones, and especially true for clusters of galaxies, where disentangling overlapping galaxies becomes essential.

As mentioned earlier, the problems of background estimation, source detection and de-blending are tightly coupled. Working with single-band images, multi-scale object-detection strategies are also applicable to the de-blending problem (e.g. wavelets, curvelets, CNN, etc.). If one considers the de-blending problem to be one of labeling pixels as part of one source or another, the challenge is to come up with a sensible way to apply a set of thresholds to the filtered/de-noised

image. The thresholds can be varied and regions of connected pixels can be labeled.

In the standard approach, each pixel is assigned to only one source (or to the background). Hausen & Robertson (2022) (9) differentiate between a *disjoint deblender*, which assigns all flux in a pixel to a single source exclusively, an *intersecting/discrete deblender*, which assigns the flux to more than one source with uniform weighting across pixels and an *intersecting/continuous* deblender, which assigns the flux to more than one source with variable weighting across pixels. Hausen & Robertson (2022) develop a CNN-based approach that works on a single band, using multi-band segmentation of deep Hubble images via SCARLET (10) for training.

5.2.3 Multiple-Band Deblending

For images taken through multiple filters, it is tempting to use color information to improve the deblending. The scarlet (10) or scarlet-lite packages have been developed for LSST and extended to address the problem of image reconstruction from multiple surveys such as LSST, Roman and Euclid. (11) Scarlet-lite is one of the de-blending approaches adopted for the LSST science pipeline.

Acelin et al. (2020) (12) use a CNN coupled with a Variational Auto Encoder (VAE) to deblend galaxies in multi-band images. They train the method on simulated images built to approximate faint-galaxy shapes by fitting model profiles to galaxies in the *Hubble* COSMOS images. It is feasible to use the same VAE with a different training approach, which might be necessary for the Roman data. The advantage of this approach is that, once trained, the CNN+VAE can be very fast.

The multi-band approach to deblending, while clearly more powerful than the single-band approach, is challenging even if the survey has uniform coverage in all the bands across the full survey area. It is even more challenging if the data are heterogenous, which will be the case for the different GAS programs. There will also be gaps even in the core-community surveys where one filter may end up shallower than the others. Furthermore, it is not straightforward to design training sets and develop objective success criteria for optimizing.

Multi-band approaches to galaxy deblending are an active area of research that may result in modules that could be incorporated into the SOC pipeline in the future.

6 Point-Source Photometry

A standard approach to obtain accurate stellar positions and fluxes is to fit a point spread function to each image. The challenge for the Roman WFI L3 images is that these are are co-added from L2 images that are potentially taken at different times and with the objects placed at different locations in the focal plane. The PSF can thus vary across a single skycell, with discontinuities due to the different spatial coverage of the individual exposures. The subsequent section on “PSF matching” addresses some strategies for dealing with this for extended sources. For point sources, the standard approach is to return to the L2 images and do the PSF fitting in each individual L2 image. The SOC already does PSF fitting to construct catalogs for every L2

image for image registration. These catalogs are completely independent for every L2 image and are thus much shallower than catalogs from the co-added L3 images. The SOC is also making plans for a "Time domain" forced-photometry catalog based on the images from the static data-release catalogs (described in this document) and transients found by the RAPID PIT in their difference-image analysis. Forced photometry starts with object positions in RA & Dec, translates them to positions in the original optically-distorted L2 reference frame, and does the photometry on each individual L2 image with the source position fixed. That catalog is conceived as a time-series catalog rather than one that co-adds all the photometry. The SOC could also make a coadded photometry catalog quite easily, or leave that to the users. In any case, the forced-photometry catalog is not the subject of this document, which is focused on measurements made from the L3 images themselves.

6.1 Baseline

For the baseline catalog based on L3 images, PSF-fit photometry will use a single PSF for each band for the entire area of each survey data release. This is for expedience, with the reasoning that the forced-photometry catalog will be used for precision work.

For the baseline, the source location is allowed to vary independently for each band. There is also no iterative de-blending of point sources from neighboring point sources. (Photutils PSFPhotometry is used for the PSF-fitting photometry).

For the baseline, PSF fits will be done on all sources regardless of star/galaxy classification.

6.2 Extensions

6.2.1 Centering

While the true position of each source should be the same in each band, the varying PSF between the different L3 images in the different bands may mean that the position estimated from the detection image is not the best estimate of the true position. The best estimate is likely to be derived from measuring the L2 images, but one relatively straightforward extension of the baseline plan that would involve only the L3 images would be to simultaneously fit all the bands, requiring that the center be the same in all the bands, with only the fluxes varying.

6.2.2 Addressing PSF variations across the L3 images

Possible ways to address the PSF non-uniformity across and within the L3 images are discussed as extensions in the next section.

7 Point Spread Functions and PSF Matching

In order to measure the spectral-energy distributions of galaxies accurately, it is important to account for the different spatial resolution in the different photometric bands. The standard procedure for doing this is to measure fluxes through the same fixed aperture in each band, but refer the fluxes all to the same reference band by applying empirically-determined corrections. One approach is to convolve all the images to match the image with the broadest PSF. However

this can sacrifice S/N. Another approach is to correct the fluxes in the low-resolution bands to a higher-resolution reference band. Ideally the reference band has high S/N for most galaxies as well as a better PSF. These corrections are based on the assumption that color gradients in the galaxies are unimportant. Consider two bands, one with a narrow PSF, I_1, and one with a broader PSF, I_2. A degraded version of I_1 matched to I_2, I_{1m2} can be created by convolving I_1 with appropriate PSF kernel. A correction for each galaxy can be derived by measuring the flux in I_1, I_{1m2}, and I_2 through a fixed aperture. The correction C is the ratio of the flux measured in I_1 to that in I_{1m2}. The photometry of the actual image I_2 is then multiplied by C to match the reference band.

There are a variety of factors that influence the point-spread function (PSF). At the exposure level (L2 images), these include:

1. The optics, which, even if stable, produce a PSF that varies significantly across the very large WFI focal plane.
2. Temporal variations due to the temperature or evolution (e.g. due to outgassing) of the physical structure of the telescope. If there is regular wavefront sensing and control of Roman, it is likely that the temporal variation is a significantly smaller concern than the spatial variation of the PSF across the focal plane.
3. Variations in the measured PSF due to the sampling by the detector. This means that a star located at the edge of pixel will have a significantly different effective PSF than one located at the edge of a pixel.
4. The WFI bands are sufficiently broad that the PSF varies significantly for stars of different colors.

For point sources, the best approach to dealing with these effects is to do photometry in the L2 images, accounting, where possible, for each of these effects in choosing the appropriate PSF. There is no easy way to mitigate these effects in the L3 images. The co-added images could well come from different parts of the focal plane (which will be the case in adjacent tiles in a mosaic due to the gap-filling dithers). The co-added images may have been taken months or years apart. The PSF can thus vary across a single skycell, with discontinuities due to the different spatial coverage of the individual exposures. The individual dithers sample the sub-pixel phase differently; and generally there are not enough dithers to smooth this out. Finally, the object colors are not known a-priori and it is impractical in any case to choose a different PSF for every object.

For galaxy photometry, the colors are generally much more important than the total flux. These are less sensitive to the PSF spatial variations because the PSF trends across the focal plane are qualitatively similar in the different filters.

7.1 Baseline

The baseline is to adopt a single PSF for each band for the SOC catalogs. These PSFs will be the same across all skycells in a single data release, but could be different for later data releases if calibrations indicate that it should be updated.

The details of how these PSFs are constructed are still to be worked out. The simplest baseline is

to circularize the "average" PSFs in each band and resample them to the resolution needed for the convolution step described in the previous section. This average is both spatial and temporal and over the available sub-pixel phases.

It is important to test this and characterize the photometric precision of the L3-based photometry once we have on-orbit data. This particular calibration should be done with repeat observations taken with sources located at different positions on the detector because source-injection simulations will not be able to reveal the uncertainties due to PSF variations.

7.2 Extensions

7.2.1 PSF-matching when creating the L3 images

Perhaps the simplest approach to mitigating the effect of the PSF variations across the focal plane would be to convolve each L2 image to a common PSF *for that band* as part of making the L3 co-add. This amounts to degrading most of the images slightly to improve the uniformity of PSF-matched multi-band photometry. The variations within a single detector are small enough that a single kernel might suffice, or a coarse grid of kernels. Because this is slightly broadening the PSF, this approach would result in a slight loss of detection sensitivity over the survey area, but in practice this may be negligible because we are convolving with the detection kernel anyway. More noticeable to the user is that the L3 images will look smoothed – in the sense of having more noticeably correlated noise between adjacent pixels – than a typical drizzled image.

If both the baseline set of inhomogenous-PSF L3 images and the homogenized-PSF L3 images are made, we could employ a hybrid strategy where detection is done on the inhomogenous-PSF version to maximize sensitivity, but the photometry is done on the PSF-matched homogenized versions.

7.2.2 Generating L3 PSFs from the L2 images when the L3 images are created

One approach to making a more faithful PSF is to create the L3-image PSFs *when the L3 images are being constructed*. That is the time when the relative weights of the individual L2 images into the final L3 product is known: those relative weights are not stored in the L3 files because the data volume would be large. One could use either a single PSF per L2 image or a coarse grid of PSFs. These would be extracted from the SOC empirical PSF database in a coarse grid across the skycell and then drizzled together with the appropriate weights. The resulting grid of PSF images could be stored as additional arrays in the L3 ASDF files, or as a separate data product. The resulting PSF-grid image could subsequently be used to generate the PSF matching kernels by the cataloging pipeline. Unless this grid is very fine, it would not capture the discontinuous changes in the PSFs across the borders of different L2 files, so this is still a shortcut, but it might be significantly better than adopting a single PSF that is independent of location on the focal plane. Armstrong et al. (13) discuss an approach to mitigating this for weak lensing studies that requires making custom co-adds for small patches, using only images with no gaps or edges within the patch. This approach does not seem practical for Roman.

7.2.3 Forced photometry on the L2 images

Photometry may also be done directly on the L2 images. The PSF is more nearly constant within a single SCA, and it is much easier to construct a spatially-varying PSF for a L2 image because it always represents the same location in the focal plane and a fixed point in time. A possible workflow for this kind of PSF matching would be the following:

- Make the baseline catalogs using the single PSF per band as described above.
- For each source in the catalog, identify the L2 images that contributed to the measurement. This is possible via the L3 metadata, the RA & Dec of the source, and the footprints of the L2 images that were co-added to make the L3 image.
- Make a cutout of the source from each L2 image in all the bands.
- Project the desired apertures onto these images and measure the aperture fluxes in the unconvolved images. Coadding these fluxes gives the "uncorrected" aperture fluxes. (With appropriate weighting and unit conversions.)
- Extract appropriate PSFs for each galaxy at each position in each L2 image from the empirical library or construct them from models.
- Measure PSF-matched fluxes across the bands using the approached described in the introduction to section 7.

7.2.4 Model each galaxy by fitting the L2 images simultaneously

A third approach is to consider that matched-PSF fixed-aperture photometry is generally favored over building a model for each galaxy because it is faster and simpler. It is also often more robust. However, if it becomes almost as complex as building a model for each galaxy, it may be preferable to use the individual L2 images to build a multi-band representation of the two-dimensional image of the galaxy, and use that representation for the photometry. Doing this generally involves some regularization, which is equivalent to building in some assumptions for the behavior of "realistic" galaxy 2D profiles. This is in a sense what IMCOM tries to do in each band separately, or what SCARLET tries to do using the full multi-wavelength information. These packages exist and can be tested to assess whether the approach is feasible and/or preferable to the approaches described above.

A machine-learning approach might be to train a convolutional variational auto-encoder (CVAE) to reproduce images of the "true multi-band object profiles" from a set of L2 images. The training would be based on extensive simulations with realistic object morphologies, e.g., drawn from deep JWST images and from hydrodynamical simulations. The CVAE would then be applied to the Roman data with the latent vectors stored in the catalog or in a separate set of tables. This would not remove the need for more traditional photometry measurements. However, constructing the 2D images from the latent vectors is fast. By design, the images reconstructed from these latent vectors represent the true objects (and uncertainties in their light profiles) better than the L3 images, and by the nature of their construction already correct for the effects of the PSF. This approach is only attractive if sufficient accuracy can be achieved with relatively small latent vectors.

8 Extended-Source Photometry

This section discusses fixed-aperture and scaled-aperture photometry. Fixed apertures have the same size for every source. Scaled apertures are scaled in size and shape based on some prior size and shape measurements of the sources.

8.1 Baseline

The baseline approach for photometry is to have three types of apertures: “aper”, “kron”, and “segment.” The photometry through all of these apertures is measured on the individual L3 images and also on images suitably corrected so that they are PSF matched to a reference band as described in the previous section.

The aperture fluxes are measured in fixed circular apertures centered at the centroid position of each source in the detection band. The Kron (14) apertures are centered at the same locations; they are generally elliptical apertures with position angle determined from the image moments within the segment. The image moments are also used to estimate the semi-major and semi-minor axis lengths: a and b. The Kron aperture has axis lengths that are a factor K_f times a and b. The typical default K_f is 2.5. Below a threshold, a fixed circular aperture is used. The default threshold is when $K_f\,(ab)^{1/2} < T_f$ pixels, with the default $T_f = 1.4$ pixels. In that case the Kron photometry uses a circular aperture of radius T_f pixels. It would be best to have the configuration parameters K_f and T_f exposed in the CRDS config file for making the catalogs. The segment fluxes are measured by summing the fluxes within the segment for each source. These fluxes are good for “isophotal” color estimates for galaxies.

In all cases, the statistical flux uncertainties in the catalog are to be determined using the uncertainty arrays. These are calculated by standard propagation of errors assuming uncorrelated errors. Because the noise is correlated in drizzled images, this is not strictly correct. We do not have the covariance matrix available for each L3 pixel to enable a more rigorous propagation of errors. The source injection and recovery tests described in section 15 will be useful for assessing both the random and the systematic uncertainties, post-facto. The baseline plan for the SOC WFI catalogs does not include using these source-injection tests to modify any of the columns in the catalogs or add additional columns.

In the baseline, neighboring sources that fall within the aperture are masked. There is no attempt to correct for the missing flux. There will be a bit set in warning_flags in the catalog to indicate that this masking was done, but this flag will not tell which of the apertures had this masking applied. While larger apertures are more likely to have such matching applied, a user should be cautious of the small-aperture fluxes anyway for such sources, because there are undoubtedly wings of the neighboring sources that extend beyond the segmentation mask.

8.1.1 Local Background

In the baseline algorithm, a circular annulus is used to provide an estimate the residual local background (after the smoothed background subtraction described in section 2). Sources are masked from this background annulus and a robust estimator is used on the remaining pixels. This background is subtracted in calculating the source flux and also recorded in a separate column in the catalog. The local background is the 3-sigma-clipped median value in the annulus. The background error is the standard error of the median,

$$\sigma\sqrt{\frac{\pi}{2N}}$$

where σ is the standard deviation of the residual flux (i.e. after the global background subtraction) un-rejected pixels, and *N* is the number of un-rejected pixels.

This local background estimate is ***not*** used in estimating any of the fluxes for the any of the columns in the catalog. It is included to allow tests of the global background subtraction, and it could be subtracted post-facto (perhaps only for selected sources based on some selection criteria), should additional adustments of the background prove benefical.

8.2 Extensions

8.2.1 Treatment of neighbors

Neighboring sources may contaminate some of the photometric apertures. There are a variety of ways to deal with this problem.

1. Provide more than a single warning_flag to indicate which apertures included neighbors. Documentation would indicate bits correspond to which apertures.
2. Fill in the masked pixels using the flux in the corresponding pixels reflected across the center of the source. Because most sources are reasonably symmetrical, this will often be better than masking. This is already implemented in photutils as the SourceCatalog aper_mask = correct option (equivalent to SExtractor MASK_TYPE = CORRECT).
3. Use the results of Sersic fitting to subtract the wings of neighboring sources. This makes aperture photometry model dependent, but only in the case of close neighbors.

8.2.2 Petrosian fluxes

One other set of frequently-requested fluxes for galaxies are Petrosian fluxes. The Petrosian ratio, R_P, at a radius r from the center of an object, is defined to be the ratio of the local surface brightness in an annulus at r to the mean surface brightness within r. The Petrosian radius r_P is defined as the radius at which $R_P(r_P)$ equals some specified value $R_{P,lim}$, typically set to 0.2. The annuli for estimating the Petrosian radius can be elliptical (e.g. based on image moments), with axial ratio b/a and position angle held fixed and radius = $(ab)^{1/2}$. In practice, there are a number of complications associated with this definition, because noise, substructure, and the finite size of objects can cause objects to have no Petrosian radius, or more than one. Nevertheless, the flux through an elliptical aperture or set of elliptical apertures scaled to multiples of the Petrosian radius may be a useful addition to the catalog. (15)

9 Magnitudes

9.1 Baseline

The photometry and photometric errors for every aperture can be expressed in magnitudes. The definition is mag = -2.5 $\log_{10}$(flux) + ZP, where the zeropoint, ZP depends on the magnitude system. Popular magnitude systems are AB magnitudes and Vega magnitudes. There needs to be a null value (typically NaN) when the flux is formally negative. Uncertainties that are symmetric in flux are formally asymmetric in log(flux), but the typical procedure is to report a single

magnitude uncertainty computed from the S/N ratio. Rather than store a lot of redundant information in the catalog database, the baseline is the following:

- Report fluxes and uncertainties in nJy for each object in each band.
- Report AB magnitudes and uncertainties only for the Kron aperture when the skycell-parquet files are first generated. AB magnitudes for other apertures will be available from the MAST database versions of the catalogs once all the skycell tables are merged.

9.2 Extensions

The Sloan Digitial Sky Survey (SDSS) also reported asinh magnitudes (16), which behave better than magnitudes at low S/N and hence are helpful for catalog selection based on object colors. This system is useful when the S/N as a function of object flux is relatively uniform across the whole survey. For Roman, one would have to decide whether to leave the "softening parameter" constant for all data products and data releases – which greatly reduces the utility of the asinh magnitudes – or let it vary based on the survey depths, which may be confusing for users.

10 Positions

There are a variety of ways to define or measure positions of astronomical objects:

1. Peak: the pixel containing the highest flux within the segment;
2. PSF: the best-fit coordinates returned from the PSF fit;
3. Centroid: the "center of light" determined from the moments of the pixel fluxes within the segment;
4. Centroid_win: A centroid within a circular Gaussian window instead of the objects isophotal footprint. The Gaussian window is iteratively scaled for each object. The algorithm is implemented in SExtractor and Photutils and described in the documentation;
5. Centroid_adaptive: A different weighting scheme than centroid_win might be better for Roman. This would be a centroid associated with the other adaptive shape measurements described in the next section.

The best estimate of the "position" of each object could differ depending on the profile of the object. PSF positions might be best for stars and centroids might be best for galaxies, where best in this could be assessed as the position that is most consistent across different L3 images of the same patch of sky for that type of object. This "best" position could be listed as RA and Dec in the MAST database version of the catalog to be used for position-based searches.

10.1 Baseline

For the baseline, the strategy is as follows:

Measure positions of type 1-4 in the list above (peak, PSF, centroid and centroid_win). For stars, use an uncertainty-weighted combination of the PSF-fit positions for the different bands. Use the centroid in the detection band as the RA & Dec for galaxies.

Use these positions as the aperture centers for all photometry in all bands. For PSF-fitting

photometry, the positions are included among the free parameters in the fit, after being provided with a first estimate from the centroids.

10.2 Extensions

One straightforward extension would be to include the adaptive-centroid positions (not currently implemented in photutils). The SOC would evaluate which of the galaxy positions give most consistent results across L3 data products and refine the rules for selecting the RA & Dec.

A much more ambitious extension would be to refine the positions based on analysis of the individual L2 images carried out as part of the forced photometry.

11 Non-Parametric Shapes

Non-parametric shapes are those that do not assume a model the source. These are relatively inexpensive to measure and can be measured on the original images and/or the PSF matched images. Non-parametric shapes include the following:

1. Sharpness & roundness1 parameters (defined by DAOphot); (17)
2. Fraction-of-light radii, determined from a fiducial estimate of the total flux.
3. A parameter to indicate if the source is significantly extended relative to a point source.
4. Ellipse parameters determined from image moments within the segment (as implemented in Photutils). These are measured on the unconvolved images in each band.
5. FWHM determined from the image moments (semi-major and semi-minor axis lengths) as the circularized full-width at half maximum of the 2D Gaussian function that has the same second-order central moments as the source (see photutils documentation for the formula). These are measured on the unconvolved images.
6. Gini and M20 by Lotz 2004 (18)
7. Concentration index from circular apertures (the ratio of the fluxes within two apertures).
8. Petrosian radii (15)
9. Concentration asymmetry and smoothness defined by Conselice 2003 (19)
10. Adaptive moments used for weak lensing analysis (20) (21)

11.1 Baseline

A baseline implementation is to measure items 1-5. Items 1 & 2 are measured separately for each band, based on the unconvolved images. Only the half-light radius is provided for the “fraction of light” radii in the baseline. The algorithm for computing item 3 is still under discussion. Currently the field is a Boolean is_extended flag, based on thresholds using other fields in the catalog. Items 4 & 5 are computed from the image moments measured on the detection image. This bare minimum baseline will require users to extract postage stamps for more ambitious measurements, unless we implement the extensions.

11.2 Extensions

The most straightforward extensions are to add more of the parameters mentioned in the list above. Highest priority for implementation is probably item 9, since this will be a useful

complement to the measurements made by the HLIS PIT.

For all of the items, it is useful to have the measurements in all the bands on the original *and* the PSF-matched images. The is_extended parameter should ideally be turned into a floating point probability for rejecting the hypothesis that the object is a single point source, calibrated based on simulations and/or observations of known objects.

12 Parametric Shapes

Parametric models are often used to characterize the shapes of galaxies. These can give more robust estimates of galaxy sizes and shapes than aperture measurements through a limited set of circular apertures or measurements that trace back to the image moments within the segmentation map (the Kron parameters). These can also allow de-blending of overlapping objects via simultaneous model fitting.

The Sérsic profile (22) is most commonly used for faint galaxies. The surface brightness scales with radius as $\log I(r) \sim -b(r/r_e)^{1/n}$. The model is usually fit as an ellipse with $r = (ab)^{1/2}$ with a and b as the semi-major and semi-minor axis lengths. The elliptical model has 6 free parameters: position of the center (2), axis lengths (2), orientation, and n. At low S/N it is common to fix a few of these parameters – for example when fitting in multiple bands, the solution might be restricted to have the same center in all the bands. The Sérsic model is convenient in that it (a) has a scale radius – a generally more robust measure of the physical size of a galaxy than the non-parametric measures – (b) describes the general behavior of the radial surface brightness profiles of exponential disks, bulges, and elliptical galaxies with the variations of the parameter n from 1 to 4.

Tests with simulated data for Hubble and Webb reveal that single Sersic fits are useful for galaxies about 2 magnitudes brighter than the point-source detection limit. (There is also a subtle definitional issue that can be confusing – the total flux to some relatively large truncated radius can still be quite different than implied flux integrated out to infinite radius. Tests on simulated data need to take this into account.)

More complex models could include separate fits for bulge and disk (typically constrained to n=4 for the bulge and n=1 for the disk). Higher order models are used to fit spiral arms and bars (23). Such complex fits are merited for well-resolved galaxies observed at high S/N. However, while scientifically interesting, these galaxies are a small fraction of the total in the catalogs and fitting approaches tend to be more specialized and science-specific than for fitting faint galaxies with simple parametric profiles.

Fitting parametric models can be computationally expensive. The process requires computing the 2D profile and convolving with the PSF for every model to be evaluated. There are various shortcuts that can be adopted. Documentation for Galsim and pysersic offers some discussion. While Sérsic-profile fitting can be done with tools provided by Astropy and photutils, optimizing this for Roman will require some development. Given that – it is worth investigating other options before settling on a baseline implementation.

One other option would be to adopt a different profile: one that is analytic in both real space and Fourier space, thereby offering the possibility to speed up the convolution step or even do the fitting entirely in Fourier space. Spergel (24) proposed the following functional form:

$$I(r) \sim \left(\frac{r}{r_0}\right)^{\nu} K_{\nu}\left(\frac{r}{r_0}\right)$$

where r_0 is the scale radius and K_ν is the modified Bessel function of the second kind. At $\nu = 0.5$, the Spergel profile is equivalent to an exponential profile (Sérsic n=1). At $\nu = -0.6$ (and in the radial range near the half-light radius), the Spergel profile is similar to a de Vaucouleurs profile or n=4 Sérsic profile. According to the Galsim documentation, due to its analytic Fourier transform, it can be much faster to create a 2D rendering of the Spergel profile than the Sérsic profile. The profile accomplishes basically the same goals as the Sérsic profile: to compactly describe galaxy surface brightness profiles. If it can do this much faster for Roman WFI data, it may be a more attractive route than fitting Sérsic profiles.

12.1 Baseline

Given the questions about a performant approach outlined above, the baseline plan is to forego parametric model fitting.

12.2 Extensions

Parametric-model fitting is important enough that there should be an investment in exploring alternatives that can be deployed at scale for Roman. It is probably to best to develop and run this fitting as a stand-alone process initially, incorporating it into the SOC pipeline once fully vetted. Much of the exploratory work can be done by the Roman science community or by the SOC in conjunction with the science community. Items to be investigated include:

- Performance-oriented approaches to fitting parametric profiles. This could well be the most expensive step in cataloging unless high-performance approaches and possibly shortcuts (not just parallelization) are adopted.
- Fitting models other than Sérsic (e.g. Spergel models for performance on faint galaxies, or bulge+disk models or bulge+disk+spiral-feature models for galaxies above some threshold in size and S/N).
- Joint fits using all of the photometric bands. This can offer science advantages over having fits in just the detection band, or having independent fits in all of the bands (especially at low S/N). However, care needs to be taken in designing the algorithm, since the most stable approach will involve some sort of regularization across the bands, which is essentially building in priors to the fitting procedure.
- Joint de-blending. Instead of using single cutouts and masks of surrounding sources, fit groups of neighboring sources simultaneously. The fitting part of this is straightforward, although once again may require some regularization. Scarlet-lite (for example) has much of the infrastructure for doing this simultaneously in all the bands. The LSST pipeline

will be using this for deblending, without adopting parametric models, but the same package can fit parametric models. Research is needed to determine the algorithm and criteria for deciding which sources to fit simultaneously and which to fit individually.

13 Photometric Redshifts

Photometric redshifts (photo-*z*s) are an estimate of an object's redshift (cosmological distance) based only on the measured brightness through each filter. While not as precise as spectroscopic redshifts, they can be obtained for many more galaxies than spectroscopic redshifts. They can also be obtained for fainter galaxies than is practical for spectroscopy. There are a variety of approaches to estimating photometric redshift estimation:

- *Spectral Energy Distribution (SED) template fitting methods*. The redshift is derived by fitting the observed photometry of a galaxy to a set of templates which can be either based on models or based on observed galaxy spectra. These templates are shifted in a fine grid of redshift and the relative fluxes through the different Roman filters computed via synthetic photometry.
- *Empirical Methods*. These are machine learning methods that attempt to map photometric space onto redshift using, in the case of supervised learning, *a-priori* knowledge provided by a subsample of objects for which accurate spectroscopic information is available or, alternatively, proceeding to self-organize the photometric information, identifying regions of the parameter space characterized by similarity factors and then post-facto calibrate the redshift distribution within each region of this map.
- *Clustering Methods*. Because galaxies are intrinsically clustered in three-dimensional space, the cross-correlation of a set of galaxies with unknown redshifts to a set of galaxies with known (spectroscopic) redshifts, can be used to infer the redshift distribution of the former. This can be used as a post-facto calibration step, or as a part of the training for machine learning.

For many scientific applications, including measuring cosmic shear with weak lensing, photo-*z* estimates do not need to be highly accurate on a galaxy-by-galaxy basis but the true redshift distribution of galaxies in bins of photo-*z* must be known to high accuracy. This includes accurately characterizing (at least) the behavior of the point estimates of the redshifts, the dispersion, the skewness of the core distribution, and the outliers arising from catastrophic errors. The narrowness of the core of the redshift probability distribution is also important, although its importance relative to reducing the fraction of outliers varies with the science application. The Roman science requirements document acknowledges that the best photo-*z*s will benefit from adding in complementary ground-based optical data and calibrating using deep spectroscopy. Therefore, we do not expect that photo-*z* estimates derived from the Roman data alone will represent the state-of-the-art. However they may be the only photo-*z* that are available for every object in the catalog, especially at the faint limits of the survey (fainter than likely used for weak-lensing measurements, but still relevant for high-redshift galaxy astrophysics, for example).

There are four science requirements that explicitly mention photometric redshifts – HLIS 2.1.4, 2.1.8, 2.1.9 and 2.1.10 listed in Appendix 1. Only the first of these requirements is specifically driving the SOC pipeline. The other three can be met by some combination of the SOC pipeline

and the Project Infrastructure Teams. Because these teams will likely make use of non-Roman data and survey-level calibrations, fully meeting them is not likely achievable early in the mission anyway. But it is clearly useful to provide a decent photometric redshift estimate early in the mission as part of standard pipeline processing.

The SOC pipeline strategy is informed by this landscape and also by the efforts to develop code and infrastructure for other surveys. Of particular interest is the efforts within the Rubin project and the LSST Dark Energy Science Collaboration (DESC) do develop a photo-z strategy for LSST. In particular, has developed a flexible software library – the Redshift Assessment Infrastructure Layers (RAIL) (25) – which provides a unified interface to multiple photometric codes.

13.1 Baseline

The baseline strategy for the SOC pipeline is to have the photometric redshifts run in a separate module. The input is a de-duplicated catalog of PSF-matched multi-band photometry across the entire area of each survey. Thus, photo-*z* estimation is done only for data-release catalogs and not prompt catalogs, which will not have the PSF-matched multi-band photometry. The SOC pipeline will use RAIL to interface to photometric redshift codes. The current plan for a first implementation is to use LePhare++, a template-fitting photo-z code that is actively maintained and has been integrated into RAIL.

The photo-*z* related fields for the baseline catalog delivered in the skycell-parquet file are subsets of what is provided by the underlying software. Specifically, the baseline includes the "best" point estimate of the photometric redshift (where the meaning of "best" will be documented), a goodness of fit, a code that identifies the best-fit SED, and the 68, 90, and 99% bounds on the photo-*z* estimate.

13.2 Extensions

The table below lists photo-*z* codes with a few comments and a reference. For the SOC pipeline, the simplest to adopt are going to be those that have been incorporated into RAIL, have a robust record of maintenance and/or are written in well-documented Python. As a reference, a table of photo-z and a table of related codes is provided in Appendix 2.

Possible photometric redshift strategies (as opposed to codes) include the following:

- Using complementary data other than from Roman – with Rubin and Subaru being the telescopes most likely to provide deep complementary data at shorter wavelengths. Doing PSF-matched multiwavelength photometry would be a pre-requisite to SED-fitting based techniques. Machine-learning techniques based on the images themselves and an extensive training set might be a shortcut.
- Using ML based photometric codes that use tabular data rather than the images. As opposed to traditional template-fitting approaches, these techniques require an extensive training set that is has nearly identical data properties (bands, S/N, and a distribution of SEDs) as the real data. Once trained, these techniques could end up being faster and more accurate than template fitting.

- Appling deep-learning techniques on the images, rather than using tabulated photometry. The Wide-Field Science program led Brett Andrews is focused specifically on developing such a technique for Roman.

Apart from different codes, there is value in providing a more sophisticated representation of the probability distribution of the redshift estimate. Various approaches to doing this have been used in the past and there is some support for these computations in the RAIL infrastructure.

In addition to the extensions in techniques, there is work beyond the pipeline itself that may need some level of support from the Roman SOC. Specifically:

- Developing (with the Roman science community) a photometric redshift calibration plan.
- Curating, maintaining and archiving the spectroscopic calibration and training sets.

14 De-duplication, Names, Identifiers, and Spatial Indices

14.1 Baseline

14.1.1 De-duplication and flagged_spatial_index baseline

As illustrated in Figure 1, skycells overlap adjacent skycells. There are also skycells along the boundaries of projection regions. However, for both skycells and projection regions, there is a clearly defined core area. A major benefit of doing only one source detection for all of the photometry and all of the different L3 subsets within a data release is that we can remove duplicate measurements of the same source along these boundaries. The algorithm is simple: if the RA and Dec. of the object falls outside of either the core area of the skycell or the core area of the projection region (or both), designate it as an ***overlap object***. Otherwise, consider it a ***primary object***.

The baseline plan is to set a bit in the warning_flags field if the object is an overlap object, and also to have a flagged_spatial_index field in the catalog with the first bit set to 1 for overlap objects and set to 0 for primary objects. The full specification of the flagged-spatial index is shown in the table below. This scheme facilitates spatial matching between SOC catalogs made from different L3 images (e.g. prompt vs. data release, or different data releases). Many objects will end up with the same flagged_spatial_index. Due to uncertainties in coordinates, the same object may end up with a different index in a different catalog. But if the catalogs are sorted by these indices, finding the match should require searching a relatively small number of rows.

Bit(s)	Description
64	Set if this is an overlap object
49-63	Projection Region index (from CRDS)
41-48	skycell y coordinate (index within the grid; currently <60 sky cells across)
33-40	skycell x coordinate
17-32	object y coordinate in units of 0.05 arcseconds/pixel starting from 0,0 pixel within its skycell
1-16	object x coordinate in the same units

14.1.2 Tracking Provenance

The combination of the flagged_spatial_index and the data-release and subset name, as encoded in the first three fields of the skycell-parquet file – e.g. r12345_r1_full – is sufficient e.g. to uniquely identify the data processing run that created a particular row in the MAST catalog tables. The files for the different bands also have this same three-field prefix. For the catalog generation, these associations are explicitly specified in an association file, and that association file is archived.

Tracking the provenance of the individual L3 and L2 files that contributed to a single row of the merged-database possible is thus *mostly* possible if we either record this identifier string on each row, or, more compactly, create a secondary table that lists the input associations and index into that table. The details of this linking via secondary tables is beyond the scope of this document. Nevertheless, the baseline plan is to create this linkage so that a MAST catalog query can return the relevant L3 or L2 files.

The sense in which this only *mostly* records the provenance is that it is impractical to track whether a given L2 file that contributed to a L3 skycell actually contributed to the pixels encompassed by each aperture associated with a given object in the catalog. So a MAST query will return all of the L2 file names and the user will have to use downstream tools to visualize or otherwise determine which L2 files overlap the object of interest.

14.1.3 Other Spatial Indices

Spatial indices allow rapid spatial queries. A common use case is for cone searches to find an object of interest, or set of objects of interest, in the Roman data. Spatial indices can be used to break the database (or file-oriented versions of the catalog, such as in parquet format) into "shards" that can be rapidly searched in parallel. The baseline for MAST is to use Q3C, a quad tree cube sky indexing approach that is very fast in PostgreSQL (26). This open-source algorithm has simpler computations than the Hierarchical Triangular Mesh (HTM) (27) and HEALPIX (28) algorithms. HEALPIX is popular and conceived to be used as a spatial index for the HIPSCat (29) approach to breaking catalogs up on Amazon S3.

The baseline plan is to populate a HEALPIX index in the merged catalog table in MAST, but not when the skycell-parquet files are initially generated.

14.2 Extensions

14.2.1 IAU Names

The International Astronomical Union has a set of specifications for naming objects outside the solar system. The "designation" should consist of the following parts: Acronym ^ Sequence ^ (Specifier), with the specifier being considered obsolete. The acronym is an alphanumeric string that designates the survey or collection, the sequence is a string of characters – normally only numerical -- that uniquely determines the source within the catalog or collection. It is recommended that celestial coordinates be used as the sequence, with a J to indicate Julian equinox 2000.0. (There is no recommendation of a single-letter designation for equinoxes later

than 2000, and there is no indication of the reference system in this single-letter designation.)

IAU names are typically too long to remember, but often too short to fully specify the provenance of the data (e.g. often do not indicate the data release).

The baseline plan is ***not*** to include an IAU name.

14.2.2 Other Object Identifiers

SDSS has an ObjID (used for database indexing) that is a 64-bit-encoded integer of then run, rerun, camera column, field, object label within a given field. When the data is reprocessed (rerun), this number changes. The Hyper Suprime Cam (HSC) Survey has a similar strategy for their object_id field. The Dark Energy Survey data releases have a COADD_OBJECT_ID that serves the same purpose.

Roman's observing plan consists of multiple different surveys (the core surveys as well as the General Astrophysics surveys), each of which will have multiple data releases. Tracking the runs and re-runs via number – at the time of creating the skycell-parquet files – would create a relatively complex interaction between the data archive and the data-processing pipeline. Assigning a sequence number to data-releases after they are delivered to MAST is a simpler approach, but means that type of ObjectID encoding used by these other surveys will not be available in the skycell-parquet files.

15 Artificial Source injection and recovery

An effective way to characterize systematic uncertainties associated with the measurements is to do the same measurements on simulated sources, where the true values are known. There are a lot of strategic choices to be made to make this both scientifically useful and cost effective. We will denote these as "shortcuts," but that is not meant to imply that they are necessarily giving lower-fidelity results than the more costly approaches. Various shortcuts that can be considered include:

1. Injecting objects into the images, rather than making completely simulated images. This is expedient and in many ways better than making complete simulations. However, the noise characteristics of the injected sources will not be identical to the noise characteristics of real sources (although it will be close). The density of inserted sources needs to be small enough that it does not appreciably affect crowding or background estimation.
2. Injecting objects into the L3 images rather than injecting them in at the L2 level and making separate L3 images (or, perhaps more efficiently from the data-storage perspective), separate arrays in the real L3 images. The spatial sampling of the simulated images will be more faithful if they are injected in the L2 images.
3. Truncating the injected galaxy profiles at a practical level, which speeds up the insertion, but does not fully simulate the effects on crowding or background estimation. Care needs to be taken in interpreting "total" fluxes galaxies when this truncation is done.
4. Not re-doing the background estimation on the source-injected images.
5. It is impossible to cover the entirety of the space of "possible" galaxy morphologies. A

standard shortcut is to insert objects with a limited range of parameters.
6. Inject sources only in the detection band or bands used to make the detection image; redo the detection to characterize completeness, but do not redo all the mesurements.

15.1 Baseline

The baseline implementation is the following:

1. Inject sources into every L3 image in all bands.
2. Inject at a density of ~300 simulated sources per skycell so as not to significantly affect crowding or background estimation.
3. Redo background estimation.
4. The injected galaxies are single-Sérsic models and point sources.
5. Draw from a magnitude distribution that runs from 1 magnitude fainter than the nominal point-source detection limit for that survey to 6 magnitudes brighter than the detection limit. (The bright magnitude limit here is notional: the goal is not to invest too much resources characterizing the bright end; at some point the galaxies get large enough that simple Sersic models are hard to justify. Large, bright galaxies will also start to influence background estimation and crowding.)
6. Characterize the distribution of ellipticity, half-light radius and n in the H band in deep JWST images from existing public surveys. A good starting point will be to adopt a linear relation between log(flux) and log(r_e), a flat distribution of ellipticity and a flat distribution of n from 0.8 to 4.5. Broaden these distributions slightly in all dimensions from the empirical ones to better explore the boundaries of parameter space.

In this baseline implementation, the images with injected sources are an intermediate product that is not saved (to reduce storage costs, and because regenerating these images is not costly). The input catalogs are saved. Full output catalogs are generated, but only the objects that match the positions of the objects in the input catalogs to with some (input-size dependent) tolerance are included.

Some testing is needed to refine the details of this implementation.

15.2 Extensions

One extension worth exploring – perhaps before settling on the baseline implementation – could be to use Spergel profiles instead of Sérsic profiles. This an interesting alternative for speed, and should be “plan A” if we decide to fit Spergel profiles instead of Sérsic profiles in the real catalog.

Other than that, the extensions that probably provide the biggest bang for the buck are the following:

1. Inject the sources into the L2 images and re-run the last step of co-addition. Take the difference between this simulated image and the real L3 image, setting pixels to zero below some tolerance, compress (lz4 compression is the fastest option), and save as a (quite small) array together with the real L3 image. This results in a much higher fidelity set of simulated images at relatively modest cost. It is cost-effective to do this while making the L3 images, not later.

2. Train a CVAE on JWST images and simulations to generate more realistic simulated images. This could pay dividends if the same CVAE can be used to characterize the real images. Other ML approaches are possible such as diffusion models, generative adversarial networks or non-variational autoencoders. The CVAE approach is attractive in that by design it has a relatively small latent space, while also perhaps approximating the diversity of image morphology more realistically than parametric models.

16 Flags and Flag maps

Individual Yes/No flags can be efficiently encoded as a bitmask, with the meanings of the bits given in the documentation. As a convenience for users, there should be database views that can select based on the meanings of the bits, rather than forcing the users to figure out which bits are which.

A useful set of flags "external flags", which are propagated through from an input flag image. SExtractor does this propagation, with the output in two catalog entries: IMAFLAGS_ISO and NIMAFLAGS_ISO. A FLAG_TYPE configuration keyword indicates the logical operation to perform on the values of the flags that land within each objects segment: OR, AND, MIN, MAX, or MOST (for the most-frequent non-zero flag). The NIMAFLAGS_ISO catalog parameter contains the numbers of relevant flag map pixels, with a slightly different logic depending on the FLAG_TYPE parameter.

Uses of flag images include:

- Flagging of diffraction spikes, ghosts, or other areas where background subtraction is likely to have left residuals.
- Flagging of gaps in the data, or where the significance image falls below a threshold.
- Saturated pixels

If there are few enough bits of information that need to be encoded, there can be one combined flag image for all the bands and one combined IMAFLAGS_ISO column in the oputput. Otherwise there would be one for each band.

The other set of flags are those associated with the measurements. SExtractor defines the following

Value	Meaning
1	Aperture photometry is likely to be biased by neighboring sources or by more than 10% of bad pixels in any aperture.
2	Object has been deblended
4	At least one object pixel is saturated
8	The isophotal footprint of the detected object is close to an image boundary
16	At least one photometric aperture is incomplete or corrupted (hitting buffer or memory limits).
32	The isophotal footprint is incomplete or corrupted (hitting buffer or memory limits)

64	A memory overflow occurred during deblending
128	A memory overflow occurred during extraction

Photutils does not have the equivalent of external flags, nor does it have SExtractor's list of measurement flags. Photutils PSF fitting routines define some flags as output parameters. There do not appear to be flags in other modules. The sourceCatalog does not have attributes that are flags.

It is likely that flags will be helpful for identifying issues that could affect photometry or shape measurements. This could include:

- Saturated pixels
- Missing bands from the detection image, if a multi-band strategy is used.
- Simultaneous PSF fitting with neighbors
- Simultaneous Sersic (or other profile) fitting with neighbors
- Specific problems (e.g. lack of convergence; bad fits) with PSF or profile fitting. (Because these could use pixels outside the segment, this might not be fully encoded in an equivalent of IMAFLAGS_ISO.)
- Photometric redshift flags (e.g. for missing bands)
- Any errors that were trapped in performing any of the measurements.

16.1 Baseline

The current implementation has warning_flags and psf_flags. The details of the encoding are still under discussion.

17 Neighbors

There are many applications that involve close neighbors. Examples include generating samples of "isolated" sources without close neighbors (often useful as "cleaner" samples of galaxies or for data-quality assessment), finding galaxies that might have been improperly de-blended, finding galaxies with companions, finding gravitational lens candidates, etc.. It could be useful to pre-identify the nearest N neighbors and store their catalog indices separations, and flux ratios (as a separate table or as entries in each table row).

17.1 Baseline

The current implementation identifies the label of the nearest neighbor and the distance to it in the skycell-parquet file. This implementation was inherited from JWST and is used as part of the L2 image-registration step. This means that neighbors in adjacent skycells are not identified in the skycell-parquet files. The skycells have an overlap region of 100 pixels, so nearest neighbors of most objects in the core region of a skycell will be in the overlap region, and hence included in the catalog for that skycell flagged as an overlap object.

17.2 Extensions

The identification of neighbors could be repeated using the merged catalog table once it is in MAST. This need not be limited to the single nearest neighbor and the results do not need to be

stored within the primary table if it is more sensible to store them in a secondary table.

18 Some Perspectives on the Extensions

At a philosophical level, one can consider catalogs as a form of data compression. The L3 images can be thought of as a lossy representation of the information in each of the individual L2 images. The individual columns in the catalog are all lossy representations of information derived from L3 images. Most scientific inference is done on catalogs because it is simpler conceptually and much more practical computationally than dealing with the images. And we measure the quantities from the L3 images instead of from the L2 images because it is simpler conceptually and much more practical computationally.

In the discussion of the extensions, at various points it seems like it might be simpler to achieve the goals by analyzing the L2 images than the L3 images. The SOC plans include doing forced photometry on the L2 images to create time-series catalogs. It may make sense to put some other kinds of measurements into such a "L2 measurement pipeline." A few such areas are highlighted in the discussions of the extensions.

Taking the "data compression" perspective a bit further: assume we have a budget of *N* kilobytes for every row in every L3 catalog. In each row we could put aperture fluxes and errorbars and versions of those in several different magnitude systems. We could include PSF-fit parameters and Sersic fit parameters and various moments of the light distribution to help characterize the image shape and orientation. But what if instead we could put a vector for every image that is *N* kb long from which we could derive all of those quantities – and many others – quickly and to the same accuracy as if we had access to all of the individual L2 images? Would it be a better use of the *N* kb budget to store that vector than a rather imperfect (and somewhat redundant) set of more traditional quantities?

Formulated as a data-compression problem: we would like a compact representation (a "feature vector") for each galaxy that can faithfully reproduce the images in each of the L2 images in every band to within some precision. Traditional strategies for such image compression include Fourier-based approaches as wavelets and shaplets. One machine-learning (ML) approach to this problem is to train a Convolutional Neural-Network Variational Auto Encoder (CVAE) on simulated data and/or real data. From this vector a sample of images that represent the "true image" of the galaxy within the uncertainties can be generated and one can perform all the standard measures on these reconstructed images to get the measurements and uncertainties. One still needs all the individual L2 cutouts as input, but once the latent vector is measured, all the measurements can be done on the reconstructed images. An advantage of this approach is that one need not anticipate all measurements that the users need.

This data-compression approach is well beyond the baseline, but ML approaches to this problem are rapidly moving into the mainstream. The techniques and code for a CVAE exist and there are a variety of other techniques that could be considered. Most of the work would be in developing and refining a network architecture, developing training sets, and validating performance. Research and development for Roman in this area is likely to pay dividends.

19 Acronyms

Acronym	Definition
ASDF	Advanced Scientific Data Format
CRDS	Calibration References Data System
CVAE	Convolutional Variational Auto Encoder algorithm
DMS	Data Management System
G2DP	Grism 2-Dimensional Processing
GAS	General Astrophysics Surveys
L1	Level 1– Reformatted telemetry with metadata
JWST	James Webb Space Telescope
L2	Level 2 – Individual exposures with basic calibration applied
L3	Level 3 – Reprojected onto a tangent plane and possibly co-added
L4	Level 4 – Catalogs and associated ancillary products
L5	Level 5 – Community contributed data products
LePhare	Photometric Analysis for Redshift Estimation software
MAST	Mikulski Archive for Space Telescopes
ML	Machine Learning
NASA	National Aeronautics and Space Administration
PIT	Project Infrastructure Team
PSF	Point-Spread Function
RAIL	Redshift Assessment Infrastructure Layers software
SOC	Science Operations Center
SSC	Science Support Center
STScI	Space Telescope Science Institute
WFI	Wide Field Instrument

20 References

1. **Bradley, Larry and et al.** *astropy/photutils.* s.l. : Zenodo doi 10.5281/zenodo.13989456, 2024.
2. *A Full-sky, High-resolution Atlas of Galactic 12 μm Dust Emission with WISE.* **Meisner, A. M. and Finkbeiner, D. P.** 2014, ApJ, Vol. 781, p. 5.
3. *Fuzzy Galaxies or Cirrus? Decomposition of Galactic Cirrus in Deep Wide- eld Images.* **Liu, Quin et al.** s.l. : ApJ, 2025, Vol. 979, p. 175.
4. *SExtractor: Software for Source Extraction.* **Bertin, E. and Arnouts, S.** s.l. : A&As, 1996, Vol. 117, p. 393.
5. *A review of source detection approaches in astronomical images.* **Masias, M. et al.** 2012, MNRAS, Vol. 422, p. 1674.
6. *A Method for Weak Lensing Observations.* **Kaiser, N., Squires, G. and Broadhurst, T.** s.l. : ApJ, 1995, Vol. 449, p. 460.
7. *Astronomical image representation by the curvelet transform.* **Starck, J. L., Donoho, D. L., & Candes, E. J.** 2003, A&A, Vol. 398, p. 785.
8. *Astronomical image denoising using dictionary learning.* **Beckouche, S., Starck, J. L. & Fadili, J.**

2013, A&A, Vol. 556, p. A132.
9. *Partial-Attribution Instance Segmentation for Astronomical Source Detection and Deblending (https://arxiv.org/pdf/2201.04714).* **Hausen, R. & Robertson, B.** Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).
10. *scarlet: Source separation in multi-band images by Constrained Matrix Factorization.* **Melchior, P., Moolekamp, F., Jerdee, M., et al.** 2018, Astronomy and Computing, Vol. 24, p. 129.
11. *Joint survey processing: combined resampling and convolution for galaxy modelling and deblending.* **Joseph, R., Melchior, P. & Moolekamp, F.** Instrumentation and Methods for Astrophysics, Vol. https://doi.org/10.48550/arXiv.2107.06984 .
12. *Deblending galaxies with Variational Autoencoders: a joint multi-band, multi-instrument approach.* **Arcelin, B., et al.** 2021, MNRAS, Vol. 500, p. 531.
13. **Armstrong, R. et al.** The little coadd that could: Estimating shear from coadded images. *https://arxiv.org/pdf/2407.01771.* [Online] 2024.
14. *Photometry of a complete sample of faint galaxies.* **Kron, R. G.** 1980, ApJS, Vol. 43, p. 305.
15. *Surface Brightness and Evolution of Galaxies.* **Petrosian, V.** 1976, ApJL, Vol. 209, p. 1.
16. *A Modified Magnitude System that Produces Well-Behaved Magnitudes, Colors, and Errors Even for Low Signal-to-Noise Ratio Measurements.* **Lupton, R. H., Gunn, J. E. & Szalay, A. S.** 1999, AJ, Vol. 118, p. 1406.
17. *DAOPHOT: A Computer Program for Crowded-Field Stellar Photometry.* **Stetson, P.** 1987, PASP, Vol. 99, p. 191.
18. *A New Nonparametric Approach to Galaxy Morphological Classification.* **Lotz, J., Primack, J. and Madau, P.** 2004, AJ, Vol. 128, p. 163.
19. *The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories.* **Conselice, C. J.** 2004, ApJS, Vol. 147, p. 1.
20. *Shear calibration biases in weak-lensing surveys.* **Hirata, C. & Seljak, U.** 2003, MNRAS, Vol. 343, p. 459.
21. *Systematic errors in weak lensing: application to SDSS galaxy-galaxy weak lensing.* **Mandelbaum, R. et al.** 2005, MNRAS, Vol. 361, p. 1287.
22. *Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy.* **Sérsic, J. L.** 1963, Boletín de la Asociación Argentina de Astronomía, Vol. 6, p. 41.
23. *A new formula describing the scaffold structure of spiral galaxies.* **Ringermarcher, H. I, Mead, L. R.** 2009, MNRAS, Vol. 397, p. 164.
24. *Analytical Galaxy Profiles for Photometric and Lensing Analysis.* **Spergel, D. N.** 2010, ApJS, Vol. 191, p. 58.
25. **Schmidt, S., et al.** LSSTDESC/RAIL: v0.98.5. [Online] 2023. https://doi.org/10.5281/zenodo.7017551.
26. *Q3C, Quad Tree Cube -- The new Sky-indexing Concept for Huge Astronomical Catalogues and its Realization for Main Astronomical Queries (Cone Search and Xmatch) in Open Source Database PostgreSQL.* **Koposov, S. & Bartunov, O.** San Lorenzo de El Escorial, Spain : s.n., 2006. ADASS.
27. **Szalay, A. S. et al.** Indexing the Sphere with the Hierarchical Triangular Mesh. [Online] 2007. https://ui.adsabs.harvard.edu/link_gateway/2007cs.....1164S/doi:10.48550/arXiv.cs/0701164.
28. ***Gorski, K. M., et al.*** *2005, ApJ, Vol. 622, p. 759.*
29. Hierarchical progressive surveys: Multi-resolution HEALPix data structures for astronomical images, catalogues, and 3-dimensional data cubes. ***Fernique, P. et al.*** *2015, A&A, Vol. 114, p. 1.*
30. Simultaneous Multicolor Detection of Faint Galaxies in the Hubble Deep Field. ***Szalay, A. S., Connolly, A. J. & Szokoky, G. P.*** *1999, AJ, Vol. 117, p. 68.*

21 Appendix 1: Science Requirements Relevant to the SOC Object Catalogs

The following table lists the science requirements from Rev C of the Roman Science Requirements Document.

<table>
<tr><th colspan="2">SRD Requirements driving pipeline extensions</th></tr>
<tr><td>HLIS 2.1.3</td><td>RST shall be capable of producing a catalog containing information for each detected source in the HLIS field, including positions, classifications, photometry (e.g. aperture photometry, model fits, adaptive moment photometry), and limited time domain information for variable sources.</td></tr>
<tr><td>HLIS 2.1.4</td><td>RST shall be capable of producing a catalog containing information for each detected source in the HLIS field, image moments (through at least 2nd order) in each filter at each epoch, and object-appropriate derived data. Examples of object-appropriate derived data include photometric redshifts and morphological parameters for galaxies, parallaxes and proper motions for stars.</td></tr>
<tr><td>HLIS 2.1.5</td><td>The RST shall be capable of including in the HLIS catalog information on the statistical uncertainties for each quantity in the catalog as well as data quality flags where numeric uncertainties are not applicable.</td></tr>
<tr><td>HLIS 2.1.8</td><td>RST shall provide a means for users to determine a redshift probability distribution for an arbitrary sample of objects that reflects a true N(z) with an error on that estimate.</td></tr>
<tr><td>HLIS 2.1.9</td><td>RST shall be capable of providing redshift probability distributions p(z) for galaxies in each tomographic bin of the HLIS lensing sample per the table below on the fraction of probability within |zphot-zspec|/(1+z) of the true redshift. (This way of phrasing the requirement takes an arbitrary p(z) into account and is more closely related to the ultimate N(z) requirement than the typically used σz and outlier fraction measurement. It also reflects the fact that the galaxy population is diverse, and so different populations will have different photo-z properties given the photometry.)
<table>
<tr><th>Fraction of Sample</th><th>68% of p(z) within</th></tr>
<tr><td>~50%</td><td>0.04</td></tr>
<tr><td>~30%</td><td>0.08</td></tr>
<tr><td>~20%</td><td>0.15</td></tr>
</table>
<table>
<tr><th>Fraction of Sample</th><th>90% of p(z) within</th></tr>
<tr><td>~70%</td><td>0.12</td></tr>
<tr><td>~20%</td><td>0.24</td></tr>
<tr><td>~10%</td><td>0.45</td></tr>
</table></td></tr>
<tr><td>HLIS 2.1.10</td><td>RST shall be capable of providing HLIS science data records with the $N(z)$ of each tomographic bin of Δzphot=0.05 such that the systematic uncertainty in the mean redshift of the bin, Δz, is given by $\Delta z/(1+z)<0.002$. (This is the top-level requirement on photo-z calibration. The 0.002 is based on the requirement that the photo-z errors degrade the aggregate precision by a factor of 1.21/2 (i.e. 20% in RSS) for the Reference HLIS survey circa 2021, and assuming that the errors in the photo-z calibration are correlated over a range of Δz =0.2 in redshift.)</td></tr>
</table>

HLIS 2.1.11	RST shall be capable of providing HLIS science data record with S/N ≥ 18 (matched filter detection significance, combining all exposures) per shape/color filter for a galaxy with an exponential disk profile and r_{eff} = 180 mas and mag AB = 24.4/24.3/23.7 (J/H/F184).
HLIS 2.1.14	RST shall provide a simulation package that can inject simulated galaxies (e.g. from x-y-λ data cubes) or stars into the real images and re-run (portions of) the data processing. This is needed to assess completeness/selection effects, the impact of blending on objects of known properties, and the impact of nearby stars on the measurement of galaxy photometric properties. In principle, much of this could be done from observing simulated skies, but these hybrid simulations are useful because they have the correct instrument noise properties and level of crowding by construction.
HLIS2.2.3	RST shall be capable of providing HLIS science data records with the PSF ellipticity, defined by the moment ratios e1=(Ixx-Iyy)/(Ixx+Iyy) and e2=2Ixy/(Ixx+Iyy), determined to an error of ≤ 5.7×10^{-4} RMS per component on angular multipole scales 32 < *l* < 3200.

22 Appendix 2

22.1.1 Photometric Redshift Codes

Code	Language	Repo	Repo type	Latest version	Requires training?	Comments	Reference
HAYATE	?	?	?	2024	Y	Hybrid ML and template fitting	Tanigawa et al. 2024
EAZY	Python	eazy-py	Github	2021	N	Template fitting	Brammer et al. 2008
FAST	IDL	FAST	Github	2019	N	Template fitting	Kreik et al. 2008
Hyperz	C?	Hyperz	webpage	2018	N	Template fitting	Bolzonella et al. 2000
LePHARE	Fortran	LePHARE	webpage	?	N	Template fitting; HLWAS PIT plans to use.	Arnouts et al. 1999
BPZ	Python?	BPZ	webpage	2010 ?	Y?	Bayesian Eigen-template fitting	Benítez 2000
Zebra	C++	Zebra	Bitbucket	2013	N	Template fitting	Feldmann et al. 2006
ANNz	C++	ANNz	webpage		Y	Neural network	Collister & Lahav 2004
ANNz2	Python &	ANNz2	Github	2021	Y	ML (Neural Net, Boosted	Sadeh et al. 2016

	C++					Decision Tree, KNN, ...)	
TPZ	Python & F90	MLZ	Github	2016	Y	Random Forest	Carrasco Kind & Brunner 2013
MLZ	Python & F90	MLZ	Github	2016	Y	SOM; includes TPZ and SparsePz as well.	Carrasco Kind & Brunner 2014
CosmoPhotoz	Python & R	CosmoPhotoz	Github	2012	Y	Generalized Linear Models and PCA based on training set	Elliott et al. 2015
PhotoRApTor	Java	PhotoRApTor	webpage	2015 ?	Y	Neural Network (MLPQNA) ...Also METAPHOR	Cavuoti et al. 2016 (preprint)
SPIDERz	IDL	SPIDERz	sourceforge	2016	Y	Support Vector Machine; uses morphologies as well as photometry; trained on spectroscopic samples	Jones & Singal 2017
CuBANz	C	CuBANz	webpage	2016	Y	Neural Networks with clustering in color space	Saumyadip & Samui Pal, 2017
Photo-z-SQL	C##/ SQLArray?	Photo-z-SQL	Github	2017	N	Template fitting within SQL	Beck et al. 2017
DCMDM	Python (uses Theano)	DCMDM	webpage	2017	Y	Deep Convolutional Density Matrix; computes photo-z PDFs directly from images, not catalogs	D'Isanto & Polsterer 2018
NetZ	?	?	?	?	Y	CNN using galaxy images	Schuldt et al. 2021
iSEDfit	IDL	impro	Github	2013	N	Bayesian template fitting	
DELIGHT	Python	DELIGHT	Github	2022	Y	Builds templates from training set via Gaussian Processes	Leistedt & Hogg 2017
FLEXZBOOST	Python	FLEXCODE	Github	2023	Y	Conditional density estimation (as opposed to regression)	
CMNN	Python	CMNN	Github	2021	Y	Color-matched nearest-neighbor (intended as a survey planning tool rather than best photo-z estimator)	Graham et al. 2020
GPz	Matlab	GPz	Github	2021	Y	Gaussian process + radial basis functions	Almosallam et al. 2016
DEmP	FORTRAN	?	?	2014 ?	Y	Constructs empirical templates based on a training set	Hseih et al. 2014
DNF	Python 2 (2021)	?	?	2015 ?	Y	Nearest Neighbor ML	De Vicente et al. 2015
Phosphoros	C++	Astrorama	Github	2024		Template fitting	Paltani et al. preparation

METAPhoR					Y	Neural network	Cavuoti et al. 2016
SkyNet							Graff et al. 2014
PZFlow	Python/Jax	pzflow	Github	2024	Y	Probabilistic modeling of tabular data with normalizing flows	Crenshaw & Doster 2021

22.1.2 Photometric-Redshift Related Codes

Code	Language	Repo	Repo type	Latest version	Comments
SparsePz	Python	SparsePz	Github	2020	Sparse representation of PDFs
qp	Python	qp	Github	2021	Quantile parametrization for PDFs
RAIL	Python	RAIL	Github	2021	Redshift Assessment Infrastructure Layers
SYNTH-Z	Python	MDN_phoZ	Github	2022	Synthetic templates

23 Appendix 3: Unresolved Issues

The table below lists details still to be firmed up for the baseline implementation.

Topic	Issue
Source injection	Fully specify parameters of the simulated galaxy & star distributions.
Source deblending	Review choice of parameters.
Positions	Finalize which positions to use as the RA & Dec and how to decide stars vs. galaxies.
Positions	Finalize algorithm for computing RA & Dec uncertainties
Photometric Redshifts	Develop procedure & process for tuning
Shapes	Finalize the algorithm for setting the is_extended flag.

24 Appendix 4: Highest Priority Extensions beyond the Baseline

Extension	Comments
Improved background subtraction	The patch-based algorithm described in Section 2.2.1
Diffraction spike treatment	Masking and possibly some mitigation by subtraction as described in section 2.2.2
Improved uncertainty array for detection	Use the patch-based algorithm to scale the weight map as described in section 3.2
Improved user customization	Expose more adjustable parameters to tuning via custom reference files.
Using multiple kernels for source detection	Capability exists in the code. Needs some research to decide on the best kernels.
Galaxy profile fitting	Fit parametric models to galaxy profiles (e.g. Sersic profiles).
Photometric redshifts	Further refinement of probability distribution information.
Is_extended indicator	Turn the Boolean into a calibrated floating point probability