D3.5.1 Prototype SXDS catalogue and first diagnostics performed.

WP3.5: LSST and near-infrared data fusion

Project Acronym	LUSC-B
Project Title	UK Involvement in the Legacy Survey of Space and Time
Document Number	LUSC-B-13

Submission date	31 March 2021
Version	2.0
Status	Draft
Author(s) inc. institutional affiliation	Raphael Shirley (Southampton), Manda Banerji (Southampton)
Reviewer(s)	Nigel Hambly (Edinburgh), David Pinfield (Hertfordshire)

Dissemination level
Public

Version History

Version	Date	Comments, Changes, Status	Authors, Contributors, Reviewers
1.0	01/03/2021	First draft for review	Written by Raphael Shirley with contributions from Manda Banerji
2.0	31/03/2021	Second draft responding to reviewer comments	Written by Raphael Shirley after reviews by Nigel Hambly (Edinburgh and David Pinfield (Hertfordshire)

Table of Contents

List of Figures

List of Tables

1 Executive Summary

Over the past year we have produced alpha versions of the software to jointly process data from the VISTA and Vera C. Rubin telescopes. We are using the existing Hyper Suprime Cam (HSC) Public Data Release 2 (PDR2) as a test data set in place of the upcoming Rubin imaging because it has been processed with the LSST Science Pipelines that will be used for Rubin. Using this first software version we have produced alpha catalogues and images for testing and diagnostics. VISTA aperture fluxes agree with previous catalogues from the same images to ~ 25 millimag depending on the metric used. HSC fluxes agree with the public releases made using the same pipeline but selected from HSC bands to within a millimag indicating that the VISTA offsets may be due to differences in the pipeline itself rather than calibration issues. There remain subtle issues between the different pipelines in both the overall photometric solution and how different flux measurement algorithms are conducted. We believe the pipeline is ready to produce catalogues that are useful for science purposes. We are now working on decisions regarding the final processing runs and data distribution and towards a final test data set over the full overlap of HSC and VISTA.

2 Introduction

For various science cases VISTA near infrared imaging can add value to the upcoming Rubin LSST data sets. For the first years of LSST VISTA will provide the only near infrared $JHKs$ band coverage of the LSST sky. One crucial element of harnessing these two data sets will be forced photometry in both directions. This will lead to three types of object: those detected in LSST but not VISTA, those detected in VISTA but not LSST and those detected in both. All of these will be important so we aim to produce general catalogues which allow investigation of all types. The key aim of the project is to have the software in place to produce these data sets as soon as LSST begins operations. This is currently anticipated to be in October 2023 shortly after the end of the current project.

We determined eight key subtasks for each year. For the first year these were:

- 3.5.1 A Copy HSC and VISTA data (flat files) to IRIS HPC.
- 3.5.1 B Configure VISTA metadata to comply with processing using lsst stack
- 3.5.1 C Install LSST software stack on IRIS HPC and run basic timing tests
- 3.5.1 D Interface with DAC team to ensure HSC and VISTA catalogues and metadata tables are accessible to this workpackage
- 3.5.1 E Develop Beta version of obs_vista package which allows VISTA imaging to be ingested into a generation 2 Butler repository.
- 3.5.1 F Run basic tests on a small 5 patch region of the SXDS field.
- 3.5.1 G Use early tests to estimate times for future pipeline runs.
- 3.5.1 H Run beta software pipeline across SXDS field to create prototype data set.

All of these tasks have been completed. The work presented here is particularly relevant to the following LSST UK Science Requirement Document[1] tasks:

- R5.05: A joint pixel-level analysis pipeline for the combined processing of optical and ground-based near infra-red imaging surveys of comparable seeing together with comprehensive documentation detailing the full pipeline implementation.
- R5.06: Optical+near infra-red (NIR) catalogues produced by joint pixel-level analysis of LSST pre-cursor surveys (DES, HSC) and ground-based near infra-red imaging surveys (UKIDSS-LAS, VHS, VIKING, VIDEO, VEILS). Catalogue delivery will include source-level metadata, detection and measurement image provenance information and workflow provenance information (e.g. configuration files).
- R5.07: Optical+near infra-red (NIR) catalogues produced by joint pixel-level analysis of LSST commissioning and science verification data and ground-based near infra-red imaging surveys (UKIDSS-LAS, VHS, VIKING, VIDEO, VEILS). Catalogue delivery will include source-level metadata, detection and measurement image provenance information and workflow provenance information (e.g. configuration files).
- R5.08: Results of running benchmarking tests on the pipeline in order to scope out future computational requirements.

[1] https://lsst-uk.atlassian.net/wiki/spaces/LUSCSWG/pages/614465537/LSST+UK+Science+Requirements+Document

The key surveys that we are currently processing from the VISTA telescope are the VISTA Hemisphere Survey (VHS) which covers most of the southern sky that will be observed by Rubin, The VISTA Kilo-Degree Infrared Galaxy Survey (VIKING) covering 1200 square degrees of Rubin sky and hundreds of square degrees of the existing HSC data, and finally the VISTA Deep Extragalactic Observations (VIDEO) Survey covering the deep SXDS field and being the principal data set investigated here. These all have coverage of the HSC data sets. We also plan to run on other VISTA data sets that cover Rubin sky when Rubin data becomes available but we have not processed any of these other surveys at this stage.

3 Software development

Software development is chiefly taking place in two GitHub repositories:

- https://github.com/lsst-uk/obs_vista A pure Python module for the LSST Science Pipelines to interact with VISTA data.
- https://github.com/lsst-uk/lsst-ir-fusion The database repository. This is where all the data is stored and pipeline runs are conducted

I will describe the development of each in turn. They were developed concurrently but obs_vista is a prerequisite for the database so I will introduce it first. We have also compiled a minimal data set required to test the software and will make this available in a separate repository shortly.

3.0.1 obs_vista

Any camera team intending to process imaging with the LSST Science Pipelines must develop an observatory or 'Obs' package. The obs_subaru package in particular is highly developed for the public HSC data releases. As far as possible we have attempted to copy the configuration files used there in order to make our processing to some extent standard. We also made use of the obs_necam example package which is a minimal obs package to aid the creation of such packages [6].

3.0.2 The database repository

In conducting a complex data processing task such as this documentation and reproducibility are crucial. For this reason we have adopted an open science framework. Every stage required to conduct the full processing is committed to the repository and described with the intention that a first year graduate student would be able to reproduce the work without expert help. We use the Data Management Unit structure from the GAMA project [3]. The chronologically ordered folders contain each stage of the processing. That is, one can move through the folders running code as described within each in order to conduct a full rerun. This means relative links can be maintained such that the full database can be set up anywhere with all code runnable with minimal set up.

3.1 Documentation

A key aim of the work is that software is well documented in order that new researchers can quickly learn to run and develop the code after the end of the project. We anticipate the

VISTA data to remain useful beyond the first year of LSST operations so rerunning as deeper LSST imaging becomes available will add to the legacy value of the work. Documentaion is an ongoing project with detailed readmes including installation guides in addition to properly formatted and annotated code. Following the release of the full catalogues in the next year we will produce an official beta version of the code with all comments in the code properly formatted and further documentation regarding its running. That said in its current state the documentation is sufficient to install the software and run with the test data set that we will publish shortly.

3.2 Compatibility

The LSST Science Pipelines are developing fast. Our code must be updated in order to remain compatible with the latest version. We aim to update the LSST Science Pipelines at every major release. We will run compatibility tests against the test data set at every major release. At the next major release, 'v22', the second generation Butler is to be deprecated and replaced with the third generation Butler. We will transition to the third generation Butler immediately following the full overlap processing run to be conducted in 2021. In the first instance we will test the capacity of the pipelines to convert the second generation Butler to a third genereation without the need for rerunning. We are then aiming to schedule a full rerun with the third generation Butler around a year later. This will allow us time to test the first catalogues and maintain lists of changes to be implemented for the future run. We anticipate a final run prior to LSST commencement to take place when the SkyMap is finalised and near to the end of the project.

3.3 Astrometric and photometric reference catalogues

A key requirement of the software is reference catalogues. These are used for calibrating images and catalogues. One could use distinct astrometry and photometry references however we follow the HSC example which uses a PanSTAARS/GAIA catalogue. That is the astrometry is from GAIA with PanSTARRS photometry cross matched in for photometry calibration. We use a positional cross match to add VISTA $ZYJHKs$ fluxes leaving all PanSTARRS objects and adding in VISTA where available. For the prototypoe run presented here we used curve of growth corrected aperture fluxes. This is in contrast to the PanSTARRS Point Spread Function, PSF, fluxes and will affect the photometric solution. We originally calibrated against 2MASS using colour terms from [4]. However, we decided to bootstrap to the final VISTA catalogues to increase the number of sources available for calibration and make the photometric solution consistent with the public catalogues. This method is comparable to the Forward Global Model Calibration (FGMC) which is implemented in HSC.

4 Prototype catalogues

We have produced a prototype catalogue on the SXDS field. In order to create this catalogue we needed to create array jobs to submit each stage of the LSST Science Pipelines to the IRIS HPC. One element of this process is checking for failed patches and rerunning them. We have also developed a pipeline for checking for any failed patches, finding the cause of the failure and rerunning with necessary changes. In the first run there were two causes of failures. One was missing HSC data which has now been rectified. The other failures were due to timeouts and have been rectified by splitting the photometry pipeline into stages. The generation 3 Butler also has more sophisticated functionality for managing queuing which checks that input requirements have already been processed at every stage allowing efficient restarting. Figure 1

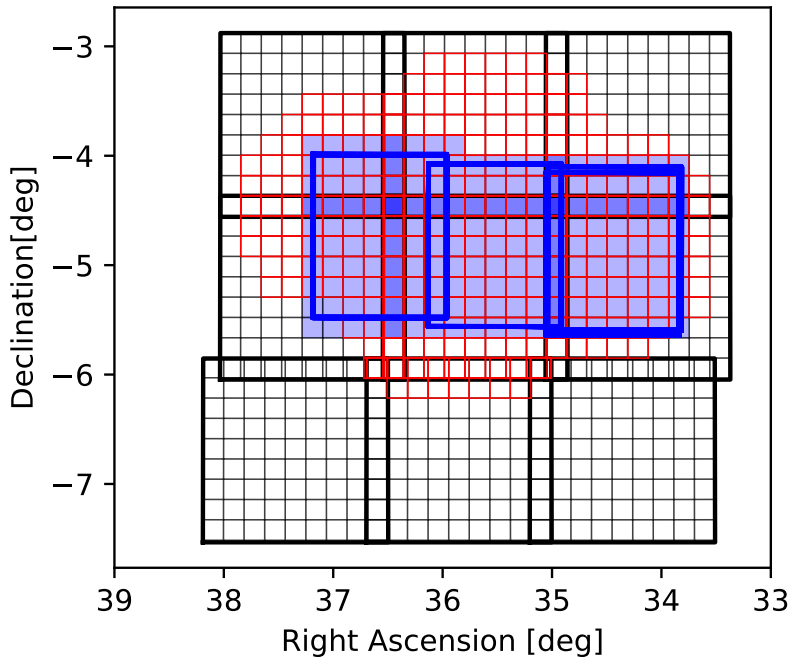

Figure 1: The SXDS fields with HSC deep patches shown in red and VISTA tile pointings shown in blue. The images and patches determine the parellelisation scheme.

shows the overlap of VISTA pointings and HSC patches on the SXDS field. These patches are made into a dictionary such that array jobs can be sent with each patch queued independently. We are currently making such dictionaries and plots for all the overlapping fields in preparation for a full run.

4.1 Slurm array job creation

Slurm is the queing system used by the IRIS high performance computers. We create array jobs which allow us to send thousands of images/patches to a single CPU each in parallel. This means that total wall clock time for runs is often similar to CPU time for a single patch and full runs can be conducted in days. The predominant limit for us is therefore total CPU time used. We us dictionaries with job information that are integer indexed so the Slurm job array, together with the dictionary describes the full job. These files are created by Jupyter Notebooks which are also used for testing the completeness of outputs.

4.2 Timing tests

A key reason for the prototype catalogue was to conduct timing tests in order to plan for required CPU time and understand how many reruns will be possible before commencement of LSST operations. Table 1 shows these times. We also use actual runs to estimate total times for other runs making some assumptions. For the purpose of the estimate we assume that VIKING times will be the same as for VIDEO but the reduced depth will reduce the times. However, moving to an all band detection will increase times due to increasing the numbers of objects.

Table 1: Overview of timing tests on the SXDS and XMM fields. These are used to estimate times for all VHS/HSC and all VIKING/HSC runs. HSC Wide areas are estimates as we are currently waiting for the download to complete. VIKING times are assumed identical to VIDEO which will be an overestimate. Parallelisation is achieved using the Slurm queuing system to send each image and patch job to the queue separately using a job array.

Surveys/fields	Calibration [hr/image]	Coadding [hr/patch]	Photometry [hr/patch]	Stack images	patches	total [hr]
VIDEO/HSC ultradeep SXDS	3	33	12	5 263	220	26×10^3
VHS/HSC wide XMM	2	0.1	2	2 226	4 374	14×10^3
Estimates:						
All VIKING/HSC Wide				20 216	15 557	1×10^6
All VHS/HSC Wide				7 174	13 164	42×10^3
All VHS/LSST				204 996	670 137	2×10^6

These timing tests will therefore be updated following every rerun as a function of decisions made regarding configuration settings and pipeline tasks to be run.

4.3 Diagnostics

At this stage of development diagnostics and quality control are of central importance. We have produced extensive comparison plots using previous catalogues made from the same data sets but with different pipelines. Figure 2 shows astrometry offsets between the protoype catalogue on SXDS and the SExtractor baseline catalogues. These have a median offset of 0.005 arcsec and a standard deviation of 0.1 arcsec. We have been in communication with other UK teams regarding the possibility of issues with the HSC GAIA astrometry catalogues due to Solar reflex and other issues affecting the astrometry solution for the stars used to calibrate each exposure. Nevertheless we necessarily use the HSC astrometric solution in order to maintain consistency with the HSC imaging as will be the case for LSST.

Figure 3 shows the photometric offsets between the Ks band from the prototype catalogue and a catalogue generated using SExtractor. Figure 4 shows the offsets from the same VISTA-Ks selected detection catalogue measurements in the HSC-r band to the public HSC catalogues. These effectively used the same pipeline but with a different driving detection list. This latter figure shows that the forced photometry is working and that these new measurements are in agreement with the public measurements made from the deeper HSC all band detection catalogue. The full diagnostic notebook which includes comparisons between all bands with previous processing runs are available on GitHub[2]. The ∼25 mmag offsets with SExtractor catalogues may be differences between the pipelines and the new pixel scale or due to the photometric solution determined by the reference catalogues. Going forward we may try a different set of calibration measurements. Currently we use curve of growth corrected aperture photometry in 2 arcsec diameter apertures from the public VISTA release. HSC is calibrated against PanSTARRS PSF fluxes. We could also introduce colour terms to correct offsets or make colour terms available to be applied to the final catalogues. Indeed it is a policy of the LSST Science Pipelines to keep images and catalogeus in instrumental flux units and only to apply the photometric calibration at the final stage of catalogue production such that colour terms could be applied without rerunning the pipeline.

[2] https://github.com/lsst-uk/lsst-ir-fusion/blob/master/dmu5/dmu5_SXDS/2_Diagnostics.ipynb

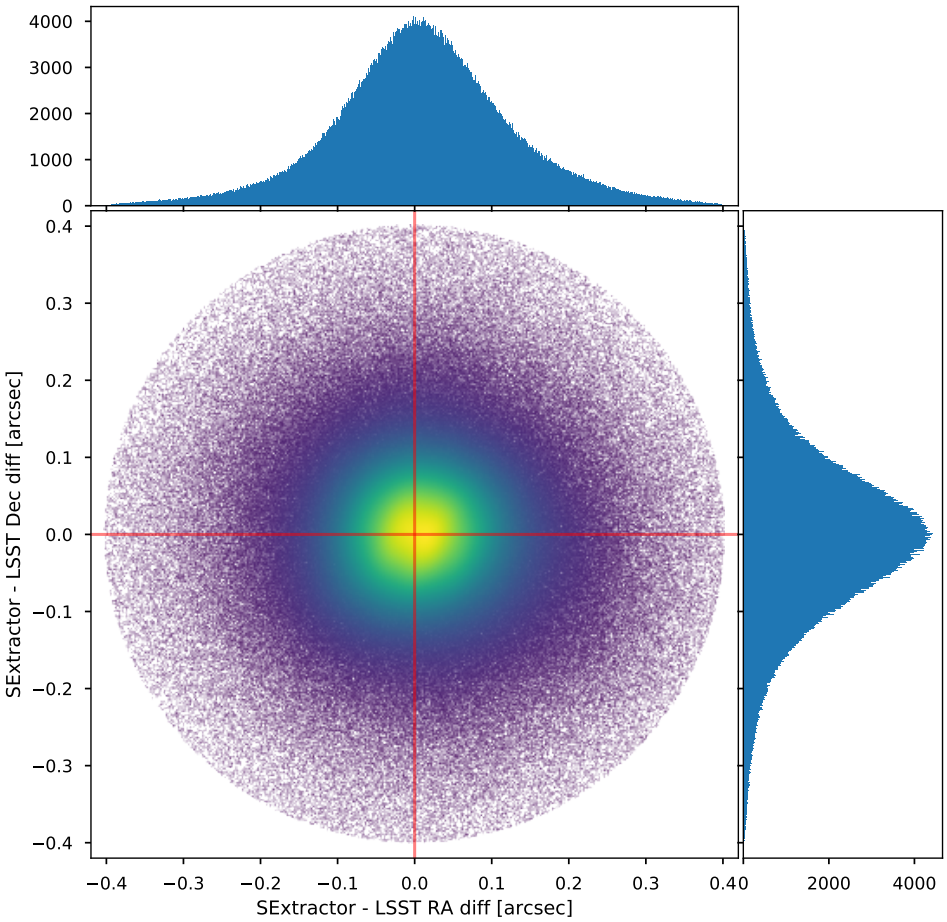

Figure 2: Offsets between the LSST Science Pipelines and the SExtractor catalogue astrometry. A 0.4 arcsec radius was used for the cross match. The median offset is 0.005 arcsec and the standard devation is 0.1 arcsec.

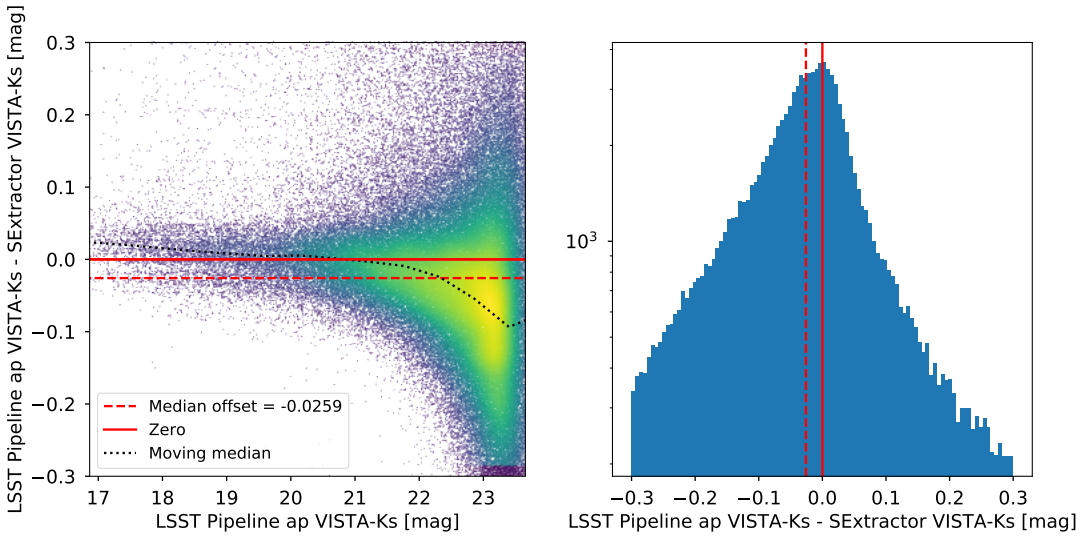

Figure 3: Offsets between the LSST Science Pipelines and the SExtractor catalogues for the Ks band. This is comparing direct aperture magnitudes in a 2 arcsec diameter aperture without 'curve of growth aperture corrections'.

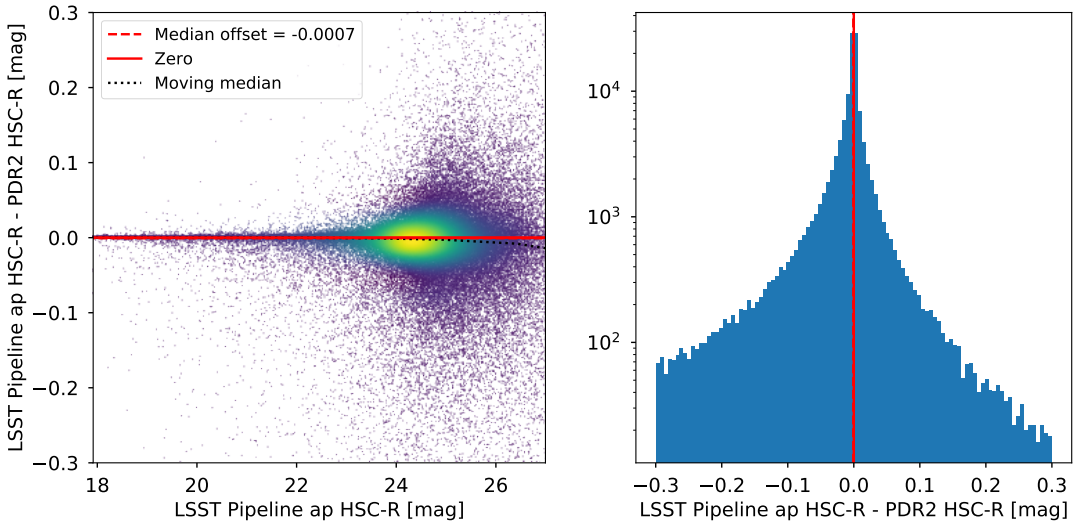

Figure 4: Offsets between the LSST Science Pipelines as run in this work from a VISTA-Ks detection compared to the public HSC all band selected catalogue. This is comparing direct aperture magnitudes in a 2 arcsec diameter aperture without 'curve of growth aperture corrections'. There is a sub millimag offset between the medians but a population of objects that are fainter in the public HSC release. The upcoming all band selected catalogues may impact this population.

4.4 Flux measurements

There are many ways to measure the flux of an object. In the first instance we calculated circular aperture fluxes in 6, 9, 12, and 17 pixel radius circles (1, 1.5, 2, and 2.8 arcseconds), PSF, and Gaussian fluxes. Unlike in the reference SExtractor catalogue the circular aperture fluxes are not 'curve of growth corrected'. Curve of growth correction involves inspecting the flux in a set of reference objects as a function of aperture size. By choosing isolated sources one can observe a levelling off of such a curve in order to estimate the fraction of the PSF in a given aperture and 'correct' aperture measurements such that they reflect the total flux of a point source. The LSST Science Pipelines do not implement this because they believe aperture fluxes to be mainly for profile studies. We therefore copy this practice and leave such a correction to the user so that they can decide such things as on what scale to compute corrections to account for PSF variation across the field of view. Note that we will calculate the 'curve of growth corrections' after processing on each patch independently. However we will supply the values allowing the user to choose whether to apply them. We plan to add in CModel, Kron, and 'convolved aperture fluxes' to future pipeline runs which are useful for many science cases and are included in HSC PDR2. Adding these measurements is already implemented for the next rerun. We will make a decision regarding including them in the full rerun based on timing tests. CModel fluxes in particular might be excluded due to their high cost.

In this first prototype the catalogue was Ks selected. This means we detected pixels and positions in the Ks band and measured those positions in all other bands. In the subsequent catalogues we will use the band merging task to merge detections in all bands in order of signal to noise to create a consistent set of detections which are then also measured in all bands. Conducting this stage of the pipeline is a trivial extension and has already been tested on a small region. This aspect of the pipeline is described in [2].

4.5 Error measurements

The native pixel scale is around twice that of the SkyMap used for the HSC processing which is chosen to be equal to the native HSC pixel scale. This will also be true for LSST. It is necessary to work in a consistent pixel space in order to conduct multiband photometry without solely conducting forced photometry. The main impact of this is that there is correlated noise between adjacent pixels in the noise map. This in turn leads to errors being underestimated. This can be accounted for by multiplying errors by a correction factor. One way to determine this factor is to compare to errors computed in the native scale. Another means is to compare the ratio of root mean square errors in the background pixels in the native vs warped pixel space [1, 5]. This will be applied to the catalogue after processing. Comparing the error measurements for the prototype catalogue to the previous reductions we have show that the factor is around 4 for all VISTA bands. This is in agreement with a naive estimate based on the ratio of pixel scales being around 2. A long term goal, beyond this project, might be to develop the capacity to handle multiple pixel scales within a single photometry pipeline.

4.6 Issues to be resolved

Based on the trial runs we have determined a number of changes to implement for a first full run:

- All photometric reference bands to be from the same survey. We initially calibrated each of VHS, VIKING, and VIDEO against VHS to use a consistent reference however the added depth and extra bands in VIDEO and VIKING should be utilised to reduce and maintain consistency with the previous releases of a given survey.
- Future runs will be based on merging the detections in every band and not just from the Ks band detections as was done in the prototype catalogue. This is already implemented and tested. this will increase the number of objects and run time for photometry measurements.
- Kron, CModel, and convolved aperture measurements to be added to next future trial run. After testing the times for each we will make a decision regarding their inclusion in the final run.
- Measurement of errors needs to be calibrated to account for using the non-native pixel size. This will be done after processing and applied to the final catalogues.
- The pipeline is to be split into smaller sections to avoid timeouts and failed patches. This is already implemented by simply dividing the tasks between multiple jobs.

5 Distribution

At this critical stage of diagnosing the performance of the catalogues we wish to make them widely available in order that other teams can use the catalogues for science and find issues that lie outside our expertise. We have been in contact with the DAC team in Edinburgh and asked them to ingest our prototype catalogues to the VISTA Science Archive. We will make the minor changes requested by them prior to sending the rerun prototype catalogues when complete.

6 Conclusions

We have presented early versions of the software required for joint processing of Vera C. Rubin Observatory and VISTA imaging. Alongside the software we present early data sets made with the software and diagnostic plots which are the subject of ongoing testing. We have identified a number of changes to be made to the code prior to conducting a full run of the overlapping regions of HyperSuprimeCam Public Data Release 2 which should resemble the eventual Rubin data. The key achievements made so far are:

- We have developed a pipeline for conducting detection and forced photometry between HSC and VISTA imaging which will be applicable to upcoming Rubin imaging.
- This was applied to the prototype SXDS field on a Ks band selected sample and is trivially extended to an all-band selection for furture runs.
- Diagnostics show consistency between photometry and astrometry measurements on the same imaging data from preexisting pipelines.

We have also identified the key next steps:

- Use scientific tests to further validate the prototype catalogue. Photometric redshifts and spectral energy distribution fitting are the most important tests for extragalactic astronomy. Making the catalogues available to other groups will help with finding issues in more general science uses.
- Implement changes in response to issues already found and maintain list of changes to be made for a full run. We are currently scheduling future reruns taking in to account limitations on CPU time available and the timeline on which we aim to implement the third generation Butler.
- Maintain software compatibility as the pipeline changes aiming for a final run towards the end of the project
- Determine and apply error offsets due to warping to finer resolution pixel space. Early tests indicate these are around a factor of 4 for all VISTA bands.

References

[1] Banerji et al, Combining Dark Energy Survey science verification data with near-infrared data from the ESO VISTA Hemisphere Survey, MNRAS, 2015.

[2] Bosch et al, The Hyper Suprime-Cam software pipeline, Ast. Soc. Jap., 2018

[3] Driver et al, GAMA: towards a physical understanding of galaxy formation, Ast. Geo., 2009

[4] González-Fernández et al, The VISTA ZYJHKs photometric system: calibration from 2MASS, MNRAS, 2018.

[5] Jarvis et al, The VISTA deep extragalactic observations (VIDEO) survey, MNRAS, 2013.

[6] Mullaney, JR et al 2020, Processing GOTO data with the Rubin Observatory LSST Science Pipelines, Ast. Soc. Aus., 2021.