D3.5.2 First wide area catalogues from joint HSC+VISTA processing

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-27

Submission date	27 June 2022
Version	2.0
Status	Final
Author(s) inc. institutional affiliation	Raphael Shirley (Southampton) and Manda Banerji (Southampton)
Reviewer(s)	David Pinfield (Hertfordshire) and Nigel Hambly (Edinburgh)

Dissemination level
Public

Version History

Version	Date	Comments, Changes, Status	Authors, Contributors, Reviewers
1.0	23/03/2022	First draft for review	Written by Raphael Shirley and Manda Banerji
2.0	05/05/2022	Second draft responding to reviewer comments	Updated by Raphael Shirley following reviews by David Pinfield (Hertfordshire) and Nigel Hambly (Edinburgh)
3.0	27/06/2022	Final version for publishing	Written by Raphael Shirley and Manda Banerji

Table of Contents

List of Figures

List of Tables

1 Executive Summary

This work package was begun in July 2020. In the first year we built a prototype version of the software to conduct joint photometry with Visible and Infrared Survey Telescope for Astronomy (VISTA) Visible and InfraRed CAMera (VIRCAM) and Vera C. Rubin Observatory Legacy Survey of Space and Time Camera (LSSTCam) imaging. This involved early prototype runs over the SXDS deep field using VISTA Deep Extragralactic Observations (VIDEO) survey and Hyper Suprime-Cam (HSC) Public Data Release 2 (PDR2) Deep surveys totalling a few square degrees. Over the last year we have continued development that was undertaken in the first year of the project. Here we present the first full runs over the overlapping regions of the VISTA Hemispere Survey (VHS), VISTA KIlo degree Infrared Galaxy survey (VIKING), and VIDEO surveys using the alpha version of the obs_vista code. This first wide area prototype run was conducted using the second generation 'Butler' middleware. This full wide run revealed issues with the pipeline that we discuss here including how they are currently being mitigated. In this document we present the results of this large area run. We will present issues discovered and motivate the work to be completed in the final year of the project. Recent development has concentrated on implementing the major refactoring of the LSST Science Pipelines that has taken place. The prototype run here, with all the issues that are discussed is publicly available in the form of catalogues. This was done partly to test the publishing technologies under development and partly to encourage collaborators to help with early testing prior to the final data set being published. Images and all the intermediate files have been deleted in preparation for upcoming reruns with the generation 3 Butler middleware. The central aim remains to have all software required ready to conduct full overlap runs when early LSST data is available which is currently anticipated in early 2024.

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 the LSST survey 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. 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. Given the additional depth of LSST compared to the VHS survey for most of the sky LSST detected objects will dominate the catalogues. All of these will be important scientifically. While we hope that the catalogues will be of general use they are targeted toward extragalactic science. This is due to the particular issues of high extinction and high source density in galactic fields which we do not specifically address. 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 April 2024. At the close of this Phase B project in March 2023 we will present the first release of the software developed during the preceding three years.

We determined eight key subtasks for each year. For this, the second year, these were:

- WP 3.5.2 A Begin developing generation 3 Butler capability.
- WP 3.5.2 B Produce larger area XMM test catalogues using generation 2 pipelines to compare depths between surveys.
- WP 3.5.2 C Compare scientific results obtained with previous HSC-VISTA forced photometry and the current generation 2 outputs.
- WP 3.5.2 D Prepare SLURM queuing machinery for all sky pipeline runs.
- WP 3.5.2 E Conduct preliminary generation 3 pipeline tests on SXDS field.
- WP 3.5.2 F Download HSC PDR3 and check image compatibility and/or any software updates required.
- WP 3.5.2 G Compute and validate photometric redshifts and spectral energy distribution models for preexisting generation 2 prototype catalogues with collaborators.
- WP 3.5.2 H Implement new variance plane from confidence maps and exposure numbers.

This deliverable is concentrated on the large area prototype run related to aim WP 3.5.2 B conducted using the generation 2 middleware. In order to retain standalone readability and coherence some material in this document is reproduced from the previous deliverable D 3.5.1.

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-

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

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.

The key large area surveys that we are currently processing from the VISTA telescope are VHS [9]; which covers most of the southern sky that will be observed by Rubin and hundreds of square degrees of HSC, VIKING [6]; covering 1200 square degrees of Rubin sky and hundreds of square degrees of the existing HSC data, and finally VIDEO [8] covering three deep fields which overlap with LSST and one covered by HSC. 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 at this stage.

3 Software development

As detailed in deliverable 3.5.1 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

They have both been extended since the previous deliverable and are developed concurrently. obs_vista is the main product of all the work and is required for processing VISTA data with the LSST Science Pipelines. The database repository contains documentation, notebooks for summarising input and prototype datasets, and scripts required for submitting large jobs to the High Performance Computing (HPC) available to us. We have also compiled a minimal data set required to test and develop the software on a laptop which is briefly described in section 3.0.3.

3.0.1 obs_vista

Any camera team intending to process imaging with the LSST Science Pipelines must develop an observatory or 'obs' package. It contains all the camera specific code required to process raw imaging. The obs_subaru package in particular is highly developed for the public HSC data releases. As far as possible we have attempted to be consistent with the configuration files used there in order to make our processing to some extent standard and comparable with HSC. We also made use of the obs_necam example package which is a minimal obs package to aid the creation of such packages [10]. Furthermore obs_lsst, obs_decam, obs_sdss, obs_megacam, and others have been used as reference points. The git hash for the obs_vista version for the prototype run presented here, which is referred to as run p2021.1, is:

- 3e1f31778536083b34a32d4b9a19c3f6a5d8e148

One crucial difference between the obs_vista package presented here and others is that it is designed to work with data from additional obs packages. Going forward it may be possible to instead develop multi instrument pipelines allowing any obs package to be used alongside any other. Currently this is not possible so obs_vista is specifically tailored to utilising additional data sets.

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 [5]. 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. This set up assumes a traditional locally accessible disk space with a hierarchical folder structure. Going forward to large runs it may be necessary to use an object store such that the Butler database will be referenced from all database structure but not located as a directory within it.

In order to permit a large run there is significant preprocessing required. We must catalogue all the input images, check their coverage and compare measurements in addition to construct calibration catalogues from previous processed photometric surveys. We have attempted to explain much of this preprocessing and development in notebooks with added text and equations to explain the reasons behind any decisions made.

3.0.3 Test dataset

In order to run the software any outside party will need to have access to the raw datasets. These are all available publicly but for ease I have collected a minimal test set to enable people to run the software on a small test region. This is is the same structure as the database repository but with the required 'data' folders which are excluded from the main GitHub repository. This test data set is available on Google Drive[2]. It can then be copied over the database repository from the same directory level.

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. Documentation is an ongoing project with detailed readmes including installation guides in addition to properly formatted and annotated code in the obs_vista package itself. Leading up to the end of the project we will ensure python style compliance with the LSST guidelines and remove comments that were solely for development purposes while testing configuration options. In its current state the documentation should be sufficient to install the software and run with the test data set provided.

[2] https://drive.google.com/file/d/1fLFjtHErYV3CXE5avS6w5vAR2Z52TEiN/view?usp=sharing

3.2 Compatibility

The LSST Science Pipelines are developing fast. Our code must be updated in order to remain compatible with the latest version. The first version of the software was made with the stable release version v21 of the science pipelines. The wide area runs produced here were made with v22. We are now running using the weekly versions incremented on v23. We do this so that any breaking changes are caught immediately. The crucial issue going forward is deprecating the second generation code presented here which will become defunct in the next major release. This generation refers to the 'Butler' middelware version which manages all task inputs and outputs. This has required a significant refactoring of obs_vista over the last year.

3.3 The Instrument Signature Removal Task

One key aspect of the CASU 'stack' images is that they are produced alongside a 'confidence' image. This confidence image allows for the production of a variance plane which accounts for bad pixels in addition to the effectively lower exposure times in the edge pixels due to dithering. We are therefore developing a VistaIsrTask which inherits from lsst.ip.isrTask and overwrites the variance methods to incorporate these confidence images which must be independently ingested into the Butler. This has not been applied in these runs but is now finished and will be applied in all future runs.

4 Prototype catalogues

We previously presented a prototype catalogue on the deep VIDEO SXDS field. One element of this process is checking for failed patches and rerunning them. We 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 due to download errors which required downloading missing files and rerunning. The other failures were due to timeouts or memory limits and have been rectified by splitting the photometry pipeline into stages and/or resubmitting jobs with additional resources. The generation 3 Butler has more sophisticated functionality for managing queuing which checks that input requirements have already been processed at every stage allowing efficient restarting. It also permits allowing resubmits with additional memory or time up to a given limit.

4.1 Slurm array job creation

In order to create this and further catalogues we created 'Slurm' array jobs to submit each stage of the LSST Science Pipelines to the IRIS HPC. Slurm is the queuing system available on the HPC we have access to; the Cambridge Service for Data-Driven Discovery (CSD3). Figures 1–6 show the overlap of VISTA pointings and HSC wide survey tracts on all the HSC PDR2 fields. For the second generation Butler run presented here we created array jobs which allow us to send thousands of images/patches to a single CPU each in parallel using job dictionaries created in the database repository. The wide area processing is a so-called 'embarrassingly parallel' problem due to the fact that each patch on the sky is not dependant on adjacent patches and so can be easily parellelised with entirely independent processes. For this reason total wall clock time for runs is often similar to CPU time for a single patch and full runs can be conducted in days, albeit with the use of millions of total CPU hours. The predominant limit for us is therefore total CPU time used. We used dictionaries with job information that are integer indexed so

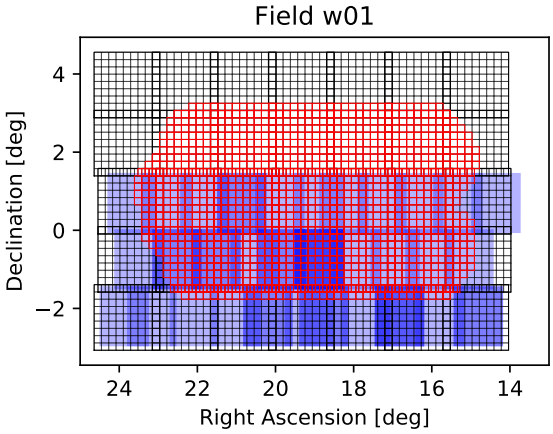

Figure 1: The w01 or XMM field with HSC wide patches shown in red and VISTA VHS tile pointings shown in blue. The images and patches determine the parellelisation scheme. This is the main test field since it is small enough to run in a week of wall clock time and is covered by the three main VISTA wide surveys.

the Slurm job array, together with the dictionary these describe the full job details. These files are created by Jupyter Notebooks which are also used for testing the completeness of outputs. Throughout the entire project notebooks are only ever used for producing the job setup data and investigating reduced outputs. At no point are they used at scale for large processing runs.

With the third generation Butler this is no longer possible and parellelisation must be conducted using the Batch Processing System (BPS) developed specifically for the LSST Science Pipelines. We will not describe that future development here. It will also be possible to produce a single very large job for the entire processing run which can then be restarted repeatedly after any time out until it is completed.

The generation 2 processing required the construction of two types of job dictionary for each survey; an input image dictionary and a coadded image patch dictionary. These two stages of parellelisation had to be separated because each image covered multiple patches and each patch is covered by multiple images. In figures 1–6 we show the tracts and VISTA tile pointings that contain both VISTA and HSC imaging that go into the Slurm array jobs. In total there were 30 thousand images files each with 16 detector images and 29 thousand patches processed. Each patch is around 700 arcseconds squared and are arranged in to tracts each containing 9 by 9 patches with overlap. A full description of the HSC sky map used is available online[3].

4.2 CPU time and disk space requirements

A key reason for the prototype catalogues was to conduct timing tests and calculate disk space requirements. Table 1 shows the times based on the latest run. These times increased by around 50% from the early estimates in the previous deliverable due to the addition of Kron, CModel, and convolved aperture measurements. The convolved aperture fluxes are measured on convolved images with a consistent PSF with a full width half maximum 3.5, 5.0, and 6.5 pixels separately. The disk space requirements shown in Table 2 are based on an early test generation

[3] https://hsc-release.mtk.nao.ac.jp/doc/index.php/available-data__pdr3/

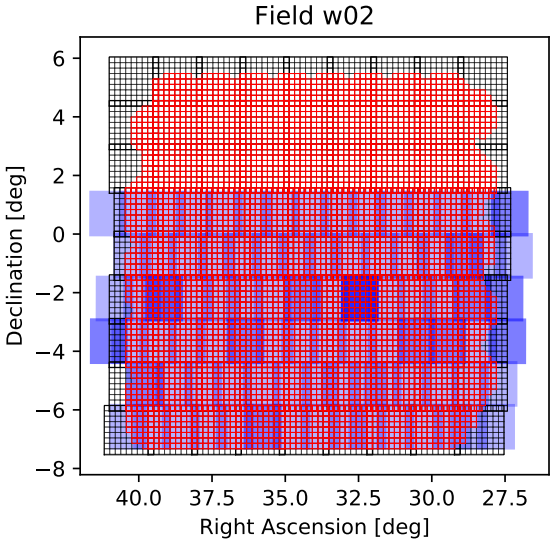

Figure 2: The w02 or XMM field with HSC wide patches shown in red and VISTA VHS tile pointings shown in blue. The images and patches determine the parellelisation scheme. This is the main test field since it is small enough to run in a week of wall clock time and is covered by the three main VISTA wide surveys.

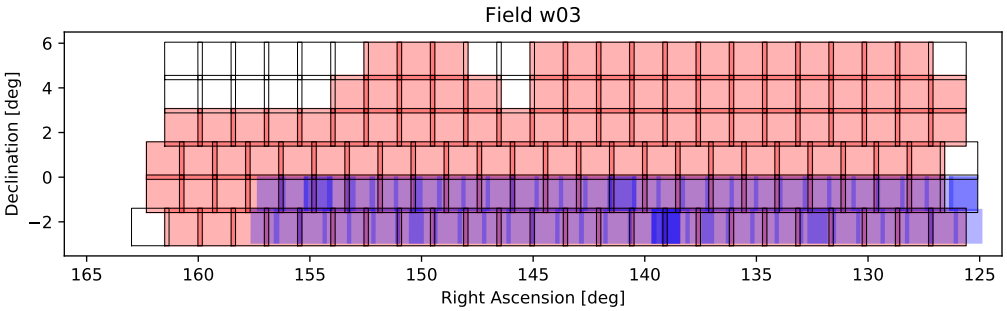

Figure 3: The w03 or XMM field with HSC wide patches shown in red and VISTA VHS tile pointings shown in blue. The images and patches determine the parellelisation scheme. This is the main test field since it is small enough to run in a week of wall clock time and is covered by the three main VISTA wide surveys.

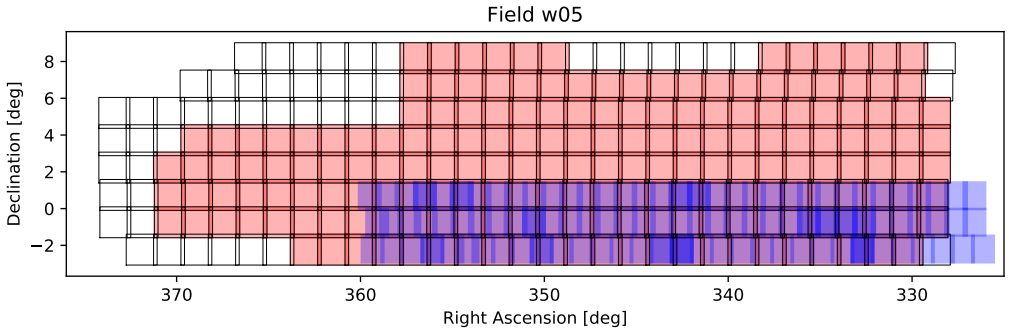

Figure 4: The w05 or XMM field with HSC wide patches shown in red and VISTA VHS tile pointings shown in blue. The images and patches determine the parellelisation scheme. This is the main test field since it is small enough to run in a week of wall clock time and is covered by the three main VISTA wide surveys.

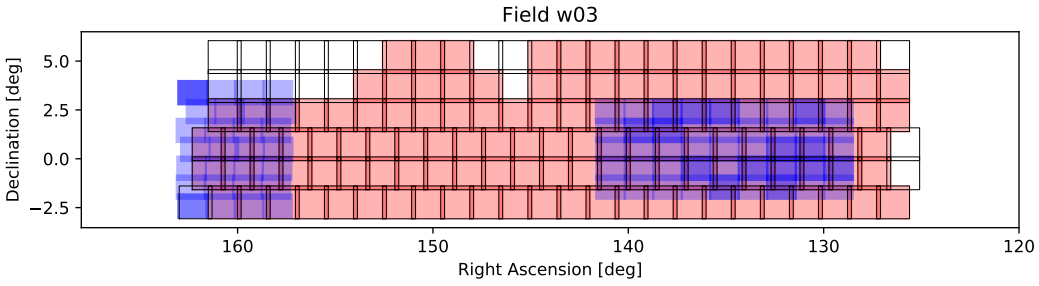

Figure 5: The w03 or XMM field with HSC wide patches shown in red and VISTA VIKING tile pointings shown in blue. The images and patches determine the parellelisation scheme. This is the main test field since it is small enough to run in a week of wall clock time and is covered by the three main VISTA wide surveys.

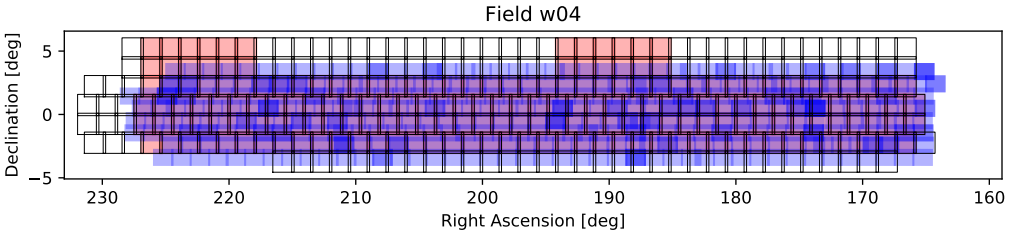

Figure 6: The w04 or XMM field with HSC wide patches shown in red and VISTA VIKING tile pointings shown in blue. The images and patches determine the parellelisation scheme. This is the main test field since it is small enough to run in a week of wall clock time and is covered by the three main VISTA wide surveys.

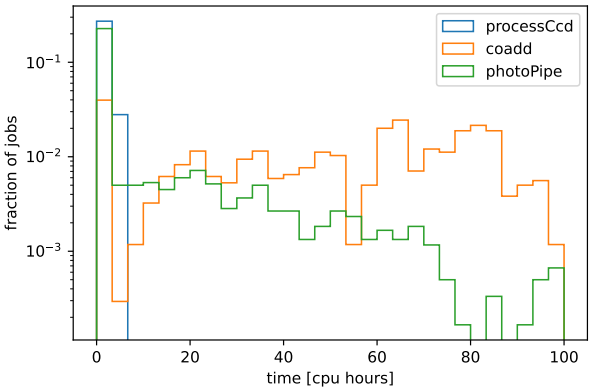

Figure 7: The distribution of total CPU time in hours for each stage of the pipeline for the VIDEO deep processing for a single image or patch. For the VHS and VIKING runs the coadd times do have have such a spread but both share the long tail of the photometry pipeline which leads to significant numbers timing out and requiring rerunning with the generation 2 pipeline.

3 run because the new middleware includes additional intermediate data sets which increase disk space requirements. For long term storage intermediate data sets could be deleted if required.

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. We have also compared measurements of the same objects from different surveys in order to understand the effect of the survey depth, measurement algorithms and different detection bands. Figure 8 shows astrometry offsets between the prototype 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. The astrometry is currently calibrated against GAIA DR1. We may expect minor improvements when we move to the GAIA DR2 astrometric calibration catalogues which are currently under development for the main LSST processing run. We should

Table 1: Overview of cpu time requirements on all the major runs to be conducted. These times are based on the wide area generation 2 runs presented here. These have increased by around 50% from the early tests presented in the first deliverable to the addition of model and other photometry measurements. These provide lower limits on actually used resources since testing and rerunning will be likely required as the science pipelines evolve. The times should not change significantly between the middleware generations because the pipe tasks have not changed.

Surveys/fields	Calibration [hr/image]	Coadding [hr/patch]	Photometry [hr/patch]	Stack images	patches	total [hr]
VIDEO/HSC PDR2 udeep	2	52	44	5 263	220	32×10^3
VIKING/HSC PDR2 wide	2	2	3	17 800	15 557	100×10^3
VHS/HSC PDR2 wide	2	1	3	7 174	13 164	60×10^3
Estimates:						
All VIDEO/LSST				13 476	1 458	176×10^3
All VIKING/LSST				41 615	58 320	350×10^3
All VHS/LSST				204 996	670 137	2.7×10^6
Total						3.2×10^6

Table 2: Overview of disk space requirements based on latest generation 3 middleware early tests including all intermediate data products. We present the values for the latest generation 3 test which includes additional intermediate data sets. The VIDEO area consists of 11 bands in total, VIKING of 9, and VHS of 9.

Surveys/fields	Raw [MiB/image]	Processed [MiB/image]	Coadd [MiB/patch]	Measurements [MiB/patch]	Images	patches	total [TiB]
VIDEO/HSC test	87	1400	221	175			
Estimates:							
All VIDEO area					7 174	13 164	23
All VIKING area					41 615	58 320	183
All VHS area					204 996	670 137	1802
Total							2008

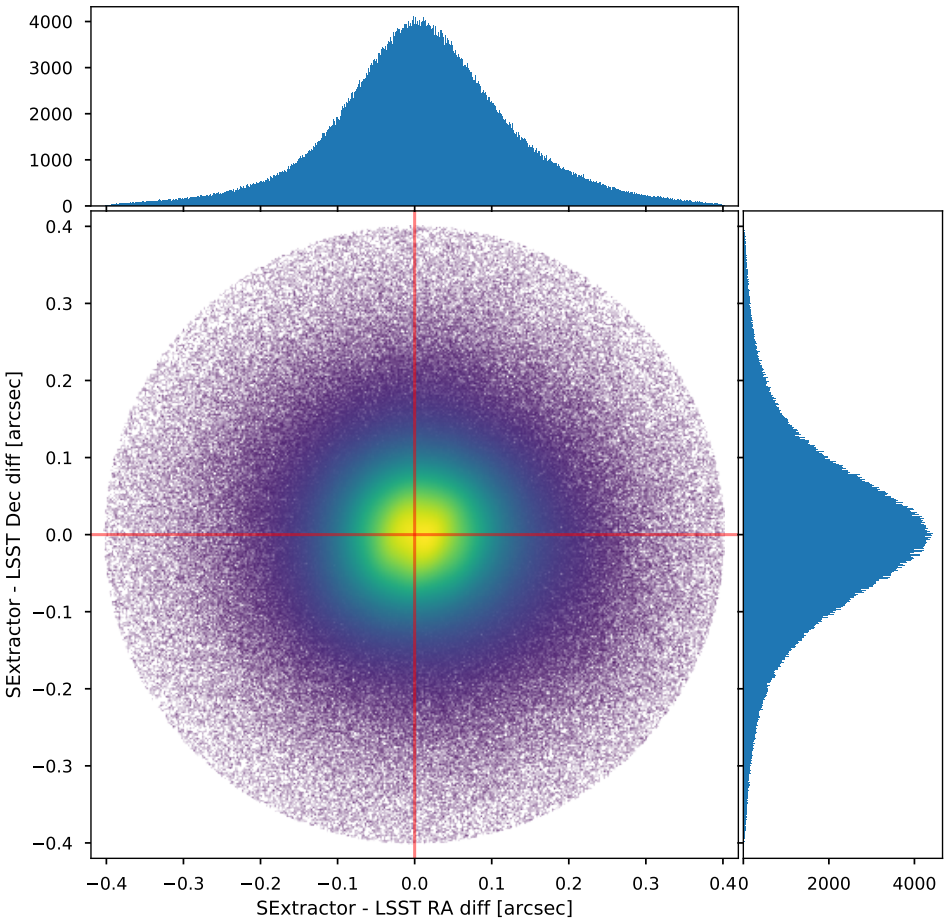

Figure 8: Offsets between the LSST Science Pipelines and the SExtractor catalogue astrometry. A 0.4 arcsec radius was used for the cross match. The median offset is 0.005 arcsec and the standard devation is 0.1 arcsec.

therefore treat 5 marcsec precision as a limit in all future checks and diagnostics.

Figure 9 shows the photometric offsets between the Ks band from the VIDEO prototype catalogue p2021.1 and the public VIDEO DR5 catalogue generated using SExtractor. There are still features of the VIDEO comparison that are not understood. These may be limitations of the calibrators used or real differences between the measurement algorithms. We are currently updating the calibration catalogue and working on understanding the causes of discrepancies. Determining the final photometric calibration catalogue is therefore the most critical issue to be resolved before conducting the ultimate runs and presenting the software and prototype data at the end of the work package.

Figure 10 shows the offsets from the same all band 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. This is further evidence that offsets we see in VISTA bands are due to the catlogue of photometric calibrators used since in the case of the HSC bands the calibrators are identical in each of the catalogues compared.

Figure 11 shows the comparison between the wide area VHS run presented here and the public VHS DR6 reduction which was used to calibrate the images and catalogues. These provide a qualitative test of photometric accuracy in comparison to a different reduction of the same data.

In this final year of the project we will provide the ultimate run with the various updates that have been provided. The current VHS photometric performance is within 5 mmag of the previous reductions in terms of the median photometry offsets measured in a 2 arcsec diameter aperture. Furthermore, the LSST Science Pipelines are developing an extensive set of performance metrics which used the relation between the measurements and the calibrator objects to provide per patch performance metrics[4]. These will also allow comparison of performance as a function of location on the sky and the ability to flag problematic regions. A focus of work for the final year of the work package will be improving the diagnostics provided and providing per patch statistics in order to track performance on a smaller spatial scale than across an entire field.

The full diagnostic notebooks for each survey which include comparisons between all bands with previous processing runs are available on GitHub[5]. All median offsets between aperture measurements compared to the VHS DR6 catalogues are less than 5 mmag. There are different calibration options available and here all images are calibrated against the public VHS magnitudes; i.e. against the previous reduction of the same underlying data set. We are comparing aperture magnitudes in a 2 arcsec diameter with no curve of growth correction.

All calibration is done using the LSST PSF magnitudes. PSF magnitudes are not computed for previous VISTA reductions. We use curve of growth corrected aperture magnitudes in a 2 arcsecond diameter circle as a proxy for PSF magnitude for the calibration using point sources. HSC is calibrated against PanSTARRS PSF magnitudes with colour terms applied. The original VISTA processing is calibrated against 2MASS with colour terms applied. The LSST Science Pipelines keep images and catalogues in instrumental flux units and apply the photometric calibration at the final stage of catalogue production such that colour terms and calibration can be applied without rerunning the full pipeline. The run presented here has the VISTA measurements calibrated against the public VSA available catalogue corrected aperture measurements. We have decided that in the subsequent and ultimate run we will calibrate against the original 2MASS calibrators to avoid any dependence on the public VISTA catalogues. We will also test the Forward Global Photometric Calibration (FGPC [4]) method to utilise the additional depth available in the final coadds to calibrate the individual input images.

4.4 Flux measurements

Following the early version of the obs_vista package we added three new measurement algorithms: Kron, CModel and convolved aperture. We also moved from a Ks selected catalogue to an all band merged catalogue. This aspect of the pipeline is described in [2]. This can be summarised with the slogan 'detected in any band measured in every band'. That is, each of all the HSC and VISTA bands have detection performed. These detections are merged and those merged objects are measured in every band.

4.5 Error measurements

One of the key issues we identified in the first year was the effect of correlated noise on error measurements in the native vs warped pixel space. Previously this has been dealt with by measuring noise in randomly placed locations [1, 8]. The pipeline we implement performs all measurements on 100 sky positions per patch which can be used to measure background noise in an aperture. These can be used for aperture measurements to measure a constant error for low

[4]See the Faro package https://lsst.ncsa.illinois.edu/~czw/pipelines_lsst_io/_build/html/modules/lsst.faro/index.html

[5]e.g. https://github.com/lsst-uk/lsst-ir-fusion/blob/master/dmu5/dmu5_VHS/1.1_Diagnostics_P2021.1.ipynb

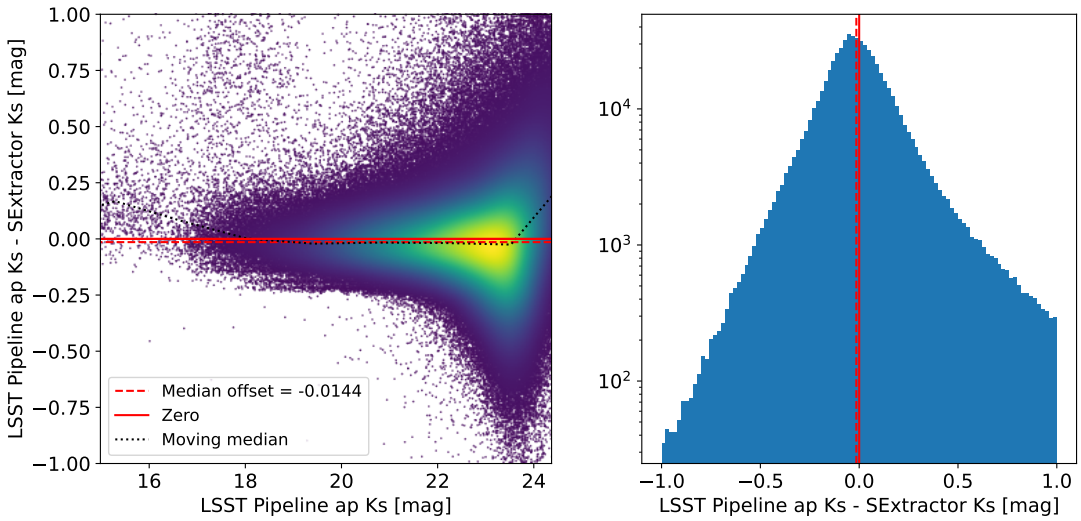

Figure 9: Offsets between the LSST Science Pipelines VIDEO run from p2021.1 and the public DR5 VIDEO SExtractor catalogues for the Ks band. This is comparing direct aperture magnitudes in a 2 arcsec diameter aperture without 'curve of growth aperture corrections'.

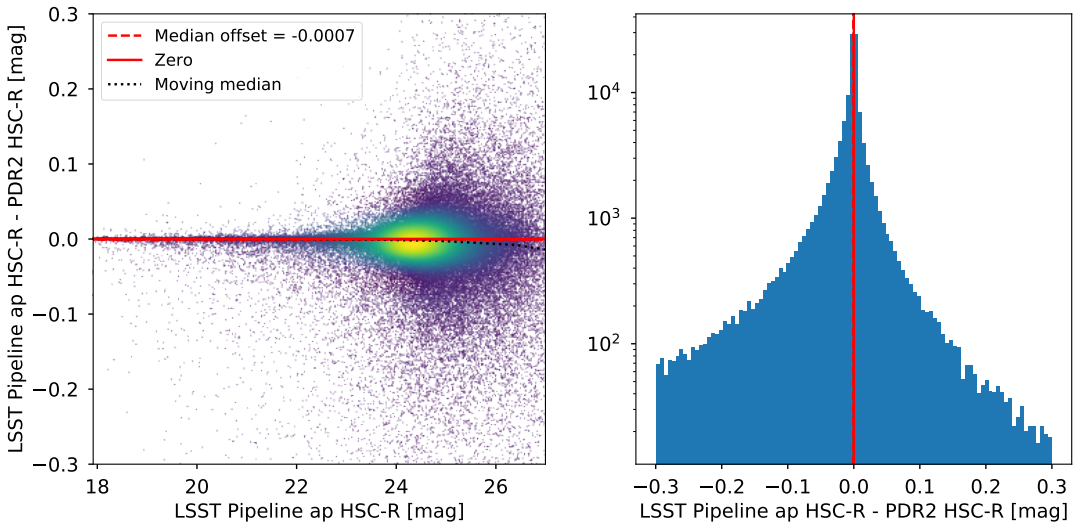

Figure 10: Offsets between the LSST Science Pipelines as run in this work from a VISTA-Ks detection compared to the public HSC all band selected catalogue. This is comparing direct aperture magnitudes in a 2 arcsec diameter aperture without 'curve of growth aperture corrections'. There is a sub millimag 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. We see very similar low offsets between all depths of public HSC catalogues.

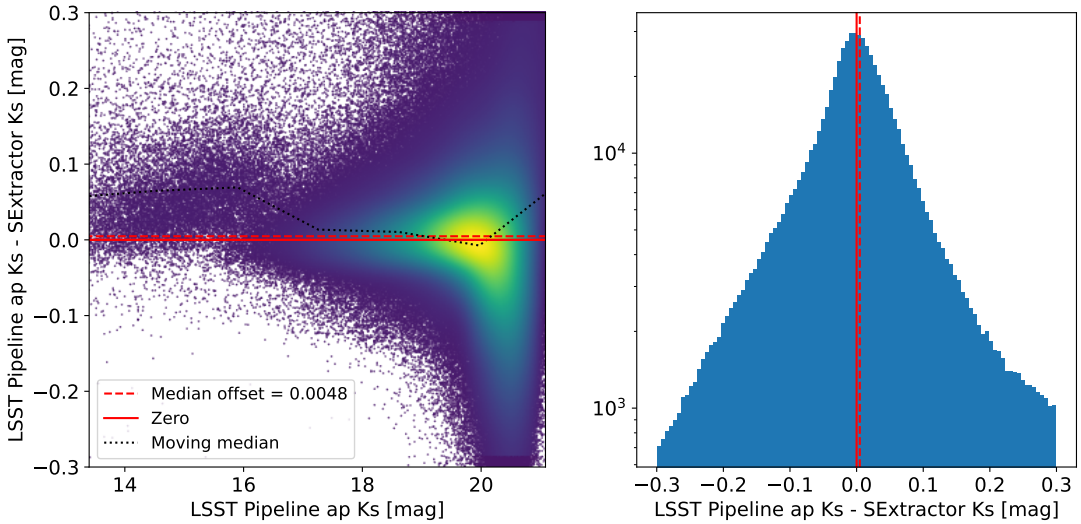

Figure 11: Offsets between the LSST Science Pipelines VHS run p2021.1 presented here and the public VHS SExtractor catalogues for the Ks band. This is comparing direct aperture magnitudes in a 2 arcsec diameter aperture without 'curve of growth aperture corrections'.

signal to noise objects. Other error measurements can be scaled using the value of the aperture errors as measured compared to the sky positions. This is an inherent feature of warping to a non native pixel scale and we leave the user to decide on which error measurement to use depending on their particular science case.

4.6 Future work

Following the early trial run in the first year we made a number of changes to the pipeline. Following the wide area runs conducted in the second year we have already conducted a number of updates. These changes already implemented but applied in the run presented here, in addition to the final stages of development that remain are:

4.6.1 Deprecating generation 2 middleware code

Deprecating the second generation Butler is currently underway. We don't want the code base to contain redundant code. We will still have the previous versions in Git should we wish to access the old Butlers so all generation 2 code should be retired. This task is already underway.

4.6.2 Applying the VISTA confidence maps

We made the decision to use the VISTA CASU pipeline stack images which have already undergone some processing. These are provided alongside 'confidence maps' which provide information about bad pixel regions and number of exposures going into each pixel. These have been applied in the generation 3 code which has been developed over the last months and will be used in all future processing runs.

4.6.3 Job completeness checks and rerunning failed jobs

One of the key difficulties encountered in conducting the wide area runs was in checking for job completeness including identifying necessary reruns. The entire machinery for doing this has been drastically updated with the generation 3 middleware. In the wide area runs presented here we checked for the presence of output files and where absent we reran those patches with more memory or time if required. This was done manually and therefore could lead to wasting resources while testing memory and cpu requirements. This is now implemented in the standard batch processing system available in generation 3 which builds a 'quantum graph' of all inputs and outputs and creates a job which can be restarted until the graph is complete including rerunning jobs which run out of memory with more memory available. Current work is focused on testing this new system which we are currently running.

4.6.4 Extend performance metrics and diagnostic comparisons

There are ongoing discrepancies between the various comparison data sets produced using previous pipelines. Building up to the presentation of the final prototype HSC-VISTA data set which will be published at the end of the work package in March 2023 we will extend the performance metrics and diagnostic plots making use of the LSST Science Pipelines Faro package designed to test performance of the LSST data. These will be produced for each patch in addition to across an entire survey in order to understand how performance varies with spatial location. These will be used to make a final decision regarding which measurements to use for photometric calibration.

4.6.5 Account for stellar fields

All of the prototype runs have been in extragalactic fields at high galactic latitudes. There is a concern that in stellar fields the timing estimates will be significantly low. Additionally, since the objects will be dominated by stars it will be inefficient to run the entire set of measurements including model measurements on all objects. We will therefore develop a production run strategy which excludes crowded fields from the main VHS Wide Fast Deep overlap production run. We will therefore aim to produce a secondary run on the crowded fields with fewer measurements conducted in order to optimise CPU usage. We will also perform a test run using VHS data only on a crowded field in order to provide an estimate of run times and to test the outputs in that environment.

4.6.6 Test 2MASS photometric calibration

In the data presented here all photometry is calibrated against the public VISTA reductions of the same underlying data set. We have now produced a 2MASS calibration catalogue and will calibrate photometry to this in future runs prior to making a recommendation for the final HSC-VISTA prototypes and subsequent LSST-VISTA runs. There remain issues with the photometric calibration and determining the ultimate photometric calibration catalogue is of critical importance for the final pipeline to be delivered in March 2023.

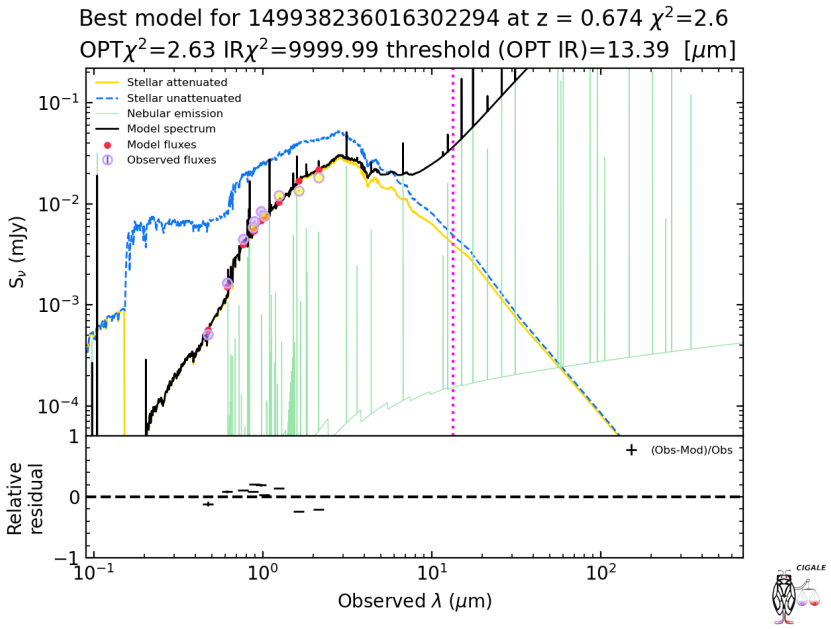

Figure 12: Example CIGALE fit of a VIDEO galaxy with a spectroscopic redshift. These early tests of spectroscopic samples were to provide further sanity checks of the photometric solution.

4.7 Photometric redshifts and Spectral Energy Distribution modelling

Alongside collaborators we have started to test the wide catalogues presented here for photometric redshift performance and SED modelling. This work will be of crucial importance for the exploitation of the data sets and will be the main focus following the completion of the upcoming third generation runs. Figure 12 shows an example fit with the CIGALE code [3] for an object with a spectroscopic redshift. These early tests show photometric consistency but further metrics must be determined to check for the differences with previous catalogues and to provide metrics for comparing redshift performance with and without the VISTA fluxes.

5 Distribution

The first trial runs have been ingested into the VSA as catalogues allowing them to be queried publicly. One feature of the LSST Science Pipelines is the very large number of columns in the final catalogues as can be seen in the public HSC releases. We have provided example queries to get typical reduced column sets. Furthermore, we make the column descriptions available to allow users to select columns of interest. A link to the VSA alongside an example query is available on the GitHub pages [6].

Going forward we will move to the LSST QServ table serving package. We will ingest the current generation 2 catalogue to test this before the upcoming generation 3 runs. These tables will be in the same format regardless of the pipeline generation. We will provide reduced column sets for the full sky final catalogues where disk space will be an issue for distribution. Full column data will still be available through the Butler.

[6] https://github.com/lsst-uk/lsst-ir-fusion/tree/master/dmu5

6 Conclusions

In the first deliverable D 3.5.1 we presented the first early version of the obs_vista software alongside test runs performed on the SXDS field. Here we extended the Slurm machinery to conduct wide area runs across the VISTA wide surveys VHS and VIKING where they overlap with HSC PDR2. The key results of this are.

- A complete run over the wide XMM field was conducted. We identified changes that should be made to various configuration settings.
- We then processed the entirety of the HSC PDR2 wide area that overlaps with VHS or VIKING. This revealed issues with ensuring completeness in the generation 2 pipelines which are being corrected in the third generation pipeline.
- The full output measurement and forced measurement catalogues are made available on the VISTA Science Archive.

Based on issues found with these catalogues, we have made changes as part of the major refactoring to update the pipelines to be compatible with the generation 3 Butler middleware. The major next steps required before the end of the Phase B work package are:

- Complete transition to the third generation compatible obs_vista package and rerun of the latest version of the software across the entire overlap of HSC and VISTA using the latest HSC PDR3 release.
- Publish the resulting catalogues via the upcoming LSST table publishing software QServ in addition to on the VISTA Science Archive as any issues with the early LSST table access software are resolved.
- Make the full generation 3 Butler for the upcoming run available to the UK Rubin Science Platform. This may require using the S3 Simple Storage Solution data store that will eventually be used for the large LSST runs.
- Determine a production strategy which excludes dense stellar fields for the main run in order to avoid wasting computing resources to calculate expensive model magnitudes on these fields. Determine an additional reduced run on these fields in order to ensure imaging and base measurements are still available there.
- Finalise the photometric calibration catalogue and ensure photometric precision as measured by median offset is below 5 mmag on all bands in the 2 arcsec aperture measurements compared to the public VISTA data releases.

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] Boquien et al, CIGALE: a python Code Investigating GALaxy Emission, AAP, 2019.

[4] D. L. Burke et al, Forward Global Photometric Calibration of the Dark Energy Survey, AJ, 2018.

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

[6] Edge et al, The VISTA Kilo-degree Infrared Galaxy (VIKING) Survey: Bridging the Gap between Low and High Redshift, The Messenger, 2013.

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

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

[9] McMahon et al, First Scientific Results from the VISTA Hemisphere Survey (VHS), The Messenger, 2013.

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