Forside

Eksamensinformationer
NFYB05035E - Bachelorprojekt i de fysiske fag, Niels Bohr Institutet - ID:twn176 (Kimi Cardoso Kreilgaard)

Besvarelsen afleveres af
Kimi Cardoso Kreilgaard
twn176@alumni.ku.dk

Administration
Eksamensteam, tel 35 33 64 57
eksamen@science.ku.dk

Bedømmere
Johan Peter Uldall Fynbo
Eksaminator
jfynbo@nbi.ku.dk
+4535325983

Francesco Maria Valentino
Eksaminator
Francesco.Valentino@nbi.ku.dk
+4535333806

Hans Kjeldsen
Censor
hans@phys.au.dk

Besvarelsesinformationer

Tro og love-erklæring: Ja

UNIVERSITY OF COPENHAGEN
NIELS BOHR INSTITUTE

A Shot in the Dark: Finding Heavily Obscured Galaxies in Multi-Wavelength Surveys

Bachelor Thesis

Supervised by Iary Davidzon,
Francesco Maria Valentino
and Johan Peter Uldall Fynbo

Kimi Cardoso Kreilgaard

Submission Date
16.06.2021

Acknowledgements

First and foremost, I wish to express my sincere gratitude to my main supervisors Iary Davidzon and Francesco Valentino for their engagement and exceptional guidance during this project. Finally, I would also like to thank Linea Hedemark for her great spirits and support during the COVID-19 lockdown where our thesis workdays proved invaluable for my productivity.

Abstract

Recent work revealed a significant population of "hidden galaxies" in the early universe (redshift $z > 3$), which systematically escaped optical and near IR telescope imaging due to high dust content. These peculiar "NIR-dropout galaxies" can now be identified in state-of-the-art infrared surveys, where the impact of dust is less strong. This project proposes a semi-automated method for the identication of the most reliable candidates of "NIR-dropout galaxies". The technique takes advantage of unsupervised and semi-supervised machine learning to classify galaxies. In particular, we use t-distributed Stochastic Neighbour Embedding (t-SNE) for dimensionality reduction on unlabeled data, consisting of cutouts of telescope images in various bands. With information from the COSMOS2020 catalogue we construct a combined model of the telescope image. Subtracting the model from the the scientific image we produce a residual map that is used for the detection of the objects constituting our sample. Through visual inspection of the residual image, a sample of respectively promising and unfit candidates is labelled to allow for semi-supervised classification of galaxies according to the vote of k neighbours. With this method we are able to classify 128 robust candidates from the entire sample of 2024 unlabelled galaxies, thus reducing the sample to be visually inspected by a factor of ~ 15. This method will allow identification and classification of distant, optically faint galaxies among the millions of object that will be detected in next-generation surveys such as *Euclid*. Upon confirmation of the candidates, this might lead to a challenge of our understanding of galaxy evolution and/or provide insight into the physical features of intrinsically faint, distant galaxies.

Link to the repository containing the code used in this project:
https://github.com/KimiKreil/Bachelor_Thesis_Code

Contents

1 Introduction

Reaching a complete understanding of the formation and evolution of galaxies is one of the central challenges of modern astronomy. To be able to study the early stages of galaxy evolution, we look for progenitors of present-day galaxies. In an astrophysical context, studying the early universe implies looking for distant galaxies with a high redshift, i.e. galaxies which emitted the light we observe at an earlier point in cosmic time. For example, a galaxy at redshift $z = 5$ (corresponding to a luminosity distance of 3 Gpc) emitted light ~ 1 billion years after the Big Bang [1].

One method of identifying a distant galaxy is observing the spectrum and directly measure the redshift, using the following cosmological redshift formula:

$$\lambda_{\text{obs}} = (1 + z)\lambda_{\text{emit}} \tag{1}$$

where λ_{obs} is the observed wavelength, z is the redshift and λ_{emit} is the rest-frame wavelength. For a large sample of galaxies this is, however, quite expensive and thus only photometric data are available. For that reason, we will adopt another approach utilising broad band photometric data in looking for so-called "dropouts", a technique first introduced in [2]. The intrinsic luminosity of dropout galaxies falls off sharply below a specific wavelength and the galaxy will thus be visible in bands with wavelengths longer than the cutoff value, while they will drop out at shorter wavelengths. The cutoff wavelength can either be a Lyman break (912 AA) or a Balmer break (3646 AA). As an example, using eq. 1, we would observe the Balmer limit for a galaxy at $z = 5$ at $\lambda_{obs} = 21876\,\text{AA} \approx 2.19\,\mu$m. Dropouts, with a large break in their continuum flux, have two possible explanations for their red colours: (i) either dust extinction causes strong reddening of star-forming galaxies, since shorter (bluer) wavelengths are attenuated the most, or (ii) the dropout galaxy is actually a quiescent galaxy that is intrinsically faint below the given wavelength, due to ceased star formation [3]. The observed cut off wavelength for a dropout galaxy in the early universe is longer than that of a low redshift one. A near infrared (NIR)-dropout is therefore not visible in optical or near infrared wavelengths, but can be detected in surveys at wavelengths above $2\,\mu$m. In this thesis, we therefore look for NIR-dropouts, which in principle will lead to objects with higher redshifts than if we looked for, e.g. optical dropouts.
Identifying a dropout galaxy by itself does not guarantee that we have found a galaxy at a high redshift, since it could be either an intrinsically faint galaxy or a galaxy faint due to dust. Nonetheless both findings would be interesting for future research. Finding intrinsically faint galaxies could provide new information about early death (quenching) of massive galaxies, while a dusty galaxy can provide insight into our knowledge of the physical features of galaxies with unprecedented dust contents. Both scenarios will enlighten us on aspects of galaxy evolution that have not been investigated thoroughly so far.

The photometric data used for this thesis are from the COSMOS2020 catalogues (Classic and Farmer), which are still under development by a team of astronomers led by NBI researchers (Weaver et. al. [4]). The catalogue is an upgraded version of COSMOS2015 [5], which incorporates new imaging and spectroscopic data in the COSMOS field. Due to the increase in exposure time, the new catalogue is almost one magnitude deeper in the Visible Infrared Survey Telescope for Astronomy (VISTA) bands, and includes all *Spitzer*/IRAC data ever taken in COSMOS. It contains source detection and multi-wavelength photometry for 1.7 million sources across the $2\,\mathrm{deg}^2$ field. Due to the rareness of the kind of galaxies we are looking for, a large survey is preferable, since we will be able to identify a bigger sample of reliable candidates. On the other hand, we also look for a deep survey due to the faintness of the targets. COSMOS2020, although not a catalogue based on the deepest surveys available, has the desired balance between the two requirements, thus providing us with a formidable tool to detect a larger sample of fainter sources. Since we are interested in NIR-dropouts, we will primarily be working with the UltraVISTA bands (H at $1.65\,\mu$m and Ks at $2.16\,\mu$m), along with the IRAC bands: $ch1$ ($\lambda = 3.57\,\mu$m) and $ch2$ ($\lambda = 4.51\,\mu$m).

This thesis has two separate aims. One is exploring whether a new, better residual map can be produced from the Farmer catalogue, surpassing classical aperture photometry. The other main purpose is to devise a new technique for a faster selection of NIR-dropouts.
In the article by Weaver et. al. [4], a new tool for profile-fitting photometric extraction, *The Farmer*, is introduced, which provides additional parameters that we can use for fitting the light profiles of sources, when creating a residual map. The process of developing a residual map of the IRAC $ch1$ image from the 2020 catalogue is described in detail in section 2.1 and the results are reported in section 3.1.
Recent work ([6] and [7]) has started a systematic investigation of a few tens of candidate NIR-dropouts. The candidates are found to be mostly massive, dusty galaxies from the early universe, which challenges our understanding of massive galaxy formation. This thesis aims to find such NIR-dropouts among the million of galaxies detected in the COSMOS field. We therefore propose a semi-automated method for the identification of the most reliable candidates of NIR-dropouts. The technique combines unsupervised and semi-supervised machine learning to classify galaxies, utilising in particular t-distibuted Stochastic Neighbour Embedding (t-SNE) for dimensionality reduction and k Nearest Neighbour (kNN) voting for classification. The method for classification of the objects is described in section 2.2.

2 Method

In the two sections below the methods adopted for respectively creating a residual map and for classification of NIR-dropouts are described.

2.1 Producing a Residual Map with Profile-Fitting

Since we are interested in objects that are visible in the IRAC bands and not detected in the NIR bands, we aim to produce a residual map of the IRAC $ch1$ image. However, due to limited computing power we refrain from attempting to create a residual map of the entire IRAC image, the full extent of which is 5.2 arcmin^2. Instead we will work with tile 7-7, which is a smaller region in the image, spanning ~ 8 arcmin^2.

In general a residual map is produced by creating a model of the image and subtracting it from the scientific image. To produce a model of the entire image one first needs to perform source detection, so it is known which objects that should be subtracted. Furthermore, one needs to apply either aperture photometry or model-based photometry, which is in general terms a way of estimating the flux from a source, so we can subtract appropriate values from the scientific image. The details of the steps described here, are elaborated in the subsections below.

2.1.1 Source Detection and Aperture Photometry

In the COSMOS2020 Classic catalogue, the object photometry is carried out using SEP [8], a Python wrapper for SExtractor [9]. To understand how the software works, we performed the source detection on the telescope images from three bands: two optical bands (g and i) from the Hypersurprime Camera (HSC) mounted on *Subaru* and the NIR band Ks from the VISTA telescope. The code is available in this notebook (link). First, we need to produce background estimation, since each pixel in a telescope image is the sum of background noise and flux from the object that we are interested in. The background estimation is a way of mapping the background flux level in different areas of the image. Subtracting this background map, we can thus produce a "clean" image, where the background is ~ 0 so the pixel flux values within a source on average only includes the flux we are interested in. There will still be statistical fluctuations present.

There are two main parameters that control the detection of sources: the deblending threshold and the flux threshold. The deblending threshold controls when to split objects that are very close in the image and the flux threshold determines the signal to noise ratio (S/N) required for a detection. For example, a very low flux threshold will lead to the detection of spurious objects while a high threshold will overlook fainter objects. These steps provide the coordinates of the detected objects along with a peak flux corresponding to the value of the pixel with the highest flux. To find the integrated flux and the magnitude of the objects, aperture photometry is used. The COSMOS2020 Classic catalogue uses fixed aperture diameters of respectively $2''$ and $3''$ and computes the aperture flux as the sum of all pixels within the aperture in the cleaned image (where the background map is subtracted).

Camera	**Zero Point [mag]**
HSC	31.4
VISTA	30
IRAC	21.58

Table 1. Zero point magnitudes to convert flux densities (in the image's native units) into magnitudes for the listed bands.

Another option is to adapt the apertures to the size of the sources. Both methods are attempted in the linked notebook. Here we computed the aperture diameter using the limits xmax, xmin, ymax and ymin from the SEP output to find the mean extent of the source along the two axes. A visual representation of the detected sources marked with their adaptive aperture can be found in fig. 10 and 11 in the appendix. The AB magnitude is found with the formula [3]:

$$m_X = -2.5 \log_{10}(F_X) + m_{0,X} \tag{2}$$

where m_X is the computed magnitude, F_X is the measured flux density in the band in native units and $m_{0,X}$ is the zero-point in band X, which converts the native flux density units into physical units. In this project, we have used the zero-point magnitudes listed in table 1. Computing the magnitudes (obtained from fixed $2''$ aperture) with this method, we are able to compare our results with the $2''$ aperture magnitudes reported in the catalogue. The distribution of the COSMOS2020 Classic magnitudes and ours are both illustrated in fig. 1, along with the difference in magnitude between corresponding objects plotted in a residual histogram. The residuals are centred at 0 and have a small dispersion, indicating that our measurements are consistent with the ones in the catalogue.

The centroids of the sources found by SEP in the IRAC *ch*1 band can then passed to the software IRACCLEAN [10] to model the IRAC *ch*1 emission. The centroids from the SEP algorithm and the total flux found from fixed aperture photometry is used to create point source models of the detected objects. The point source model is an approximation, since the light profiles of the galaxies will differ from a perfect point source. However, due to the low resolution of the IRAC images and their large point spread functions (PSFs), much larger than the expected sizes of our candidates and of the most distant objects, it is reasonable to use the point-like approximation. The COSMOS team used IRACCLEAN software and a similar approach as described above, but performed on both *ch*1 and *ch*2 in IRAC, to produce what we will refer to in this project as the classic residual map. There are, however, some limitations to aperture photometry and the residual map it produces, which is why we will also explore another approach of creating a residual map.

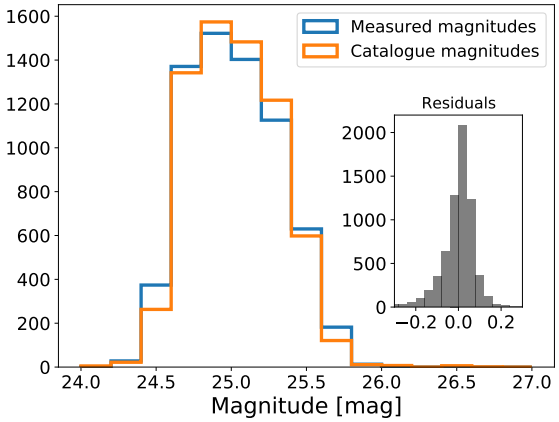

Figure 1. The distribution of magnitudes pertaining to objects in the COSMOS field. In blue, the magnitude measurements found by running SEP is plotted. In orange the 2" magnitudes reported in the COSMOS2020 Classic catalogue are displayed. In grey a residual plot is shown, mapping the difference between the measured and reported magnitude for each object.

2.1.2 Model-Based Photometry

An alternative to aperture photometry is model-based photometry, a method from which the Farmer tool is developed. The Farmer tool [4], like the previous method, also starts out with a source detection algorithm (SEP) reporting the centroids of each source detected. Next it identifies particularly crowded regions where there are multiple objects nearby each other that could present overlapping flux. Contrary to the previous methods, these sources are modelled simultaneously. Furthermore this method does not require enforcing a fixed aperture size. Instead the total flux is determined by fitting an intrinsic light profile to the source, and integrating the area covered by the model below the surface. The Farmer tool fits one of 5 available models (after PSF convolution) to an observed source. Common for all the models is that they are parametrised using the centroid position of the source as a fixed parameter and the total flux as a free parameter. The models are:

1. **Point Source** models are taken directly from the point spread function (PSF) used. It is similar to the method applied when using aperture photometry.

2. **Simple Galaxy** models are circular symmetric, exponential light profile with a fixed 0.45" effective radius.

3. **Exponential Galaxy** models are exponential light profiles. In addition to the common parameters these are parametrised also by effective radius, axis ratio and position angle.

4. **DeVaucouleurs Galaxy** models are de Vaucouleurs light profiles. These models require similar parameters as the exponential galaxy model.

5. **Composite Galaxy** models are combines a Exponential Galaxy profile with a DeVaucouleur Galaxy profile. The profiles are concentric, and the model, like the others, is therefore only parametrised by one pair of centroids coordinates. Each component is parametrised by the free parameters described in the two previous models. In addition there is a "fraction of total flux" parameter that distributes the flux between the two components.

Except for the Point Source model, they can all be described analytically by changing the parameter n in a 2d Sérsic profile. $n = 1$ corresponds to a Exponential Galaxy and $n = 4$ creates a De Vaucouleurs Galaxy. The general profile is given by [11] in eq. 3,

$$I(R) = I_e \cdot \exp\left(-b_n \cdot \left(\left(\frac{R}{R_e}\right)^{\frac{1}{n}} - 1\right)\right) \tag{3}$$

where I_e is the intensity at the effective radius R_e that encloses half of the total light from the model. The constant b_n is defined from n, to make sure the total luminosity is obtained when integrating the profile. Since four models are all analytical, they can easily be integrated to obtain the total flux. The real challenge lies in choosing the right model and obtaining a converging fit. The Farmer tool includes a decision tree for selecting the most appropriate model type to fit a given

source with. The models are tested in the order they are described above (1 to 5), and the quality of their residuals are ranked, so the best model is chosen. When the most appropriate model type is selected for all sources, optimisation of the free parameters is performed as the final step, so that the parameters chosen for one object will not directly affect the parameters of nearby objects. While this method does improve on some of the drawbacks from aperture photometry, it is almost impossible to make all model fittings converge. There are a few possible reasons for the failure to converge. Sometimes a bright source is simply not well described by a smooth profile. It could also be due to a very blended pair that was not successfully separated in the SEP detection, and thus not described correctly by a single light profile.
With this process, the Farmer tool can provide us with the parameters needed for creating intrinsic light profiles of the sources detected in the image. The models describe how we expect the light to look intrinsically, but when the light is observed through a telescope it is warped due to the way the imaging system responds to a point source. A point spread function (PSF) describes the response of an imaging system to a point source, mapping how the light will appear in the image. Therefore, to create a residual image we need to model the objects as they would appear in the image. We obtain the final models from a convolution of the intrinsic light profiles with the PSF of the image. The residual map is then produced by subtracting the convolved models from the telescope image, carefully placing the models right according to the detected centroid coordinates.

For the next part of the project we need a catalogue containing the coordinates and total flux of objects detected in the residual image. For this purpose we use SEP source detection and aperture photometry to estimate the total flux. This catalogue will likely contain different objects of various level of interest: spurious objects, left-over flux from subtracted sources due to imperfections in the residual image, and dropout galaxies.

2.2 Classification with Semi-supervised Machine Learning

In this thesis, we develop a semi-automated method of selecting NIR-dropout galaxies based on t-SNE [12] and kNN voting. This method uses as input: 4 telescope image cutouts of each galaxy in the sample, from respectively the H, Ks, $ch1$ and $ch2$ bands. The method is semi-supervised, and hence will need some parameter tuning along the way and some labels (to be defined in section 2.2.2). With this, the technique will predict whether each galaxy in the sample is a NIR dropout or not.

2.2.1 Dimensionality Reduction with t-SNE

As mentioned, the input that the method is given for each object in the sample is four small telescope images centred at the object in question. We use cutouts of the size 21×21 pixels, meaning that we in total will have $21 \times 21 \times 4 = 1764$ features for each object in the sample. Inherently, this is a very high-dimensional classification problem, and it is particularly difficult to produce

meaningful predictions from, when the learning is unsupervised. A way to simplify the problem is embedding the high-dimensional data in a 2-d representation. There are many approaches to obtaining such an embedding, but for our purposes we will adopt the method of t-SNE, which has gained a lot of popularity since its release in 2008. A recent NBI paper [13] proposed this novel technique for classification of galaxies from photometric data, which has inspired the approach applied here. In particular they used t-SNE to classify quiescent galaxies at $z = 1$ from their spectral energy distributions (SEDs).

t-SNE takes a set of high dimensional vectors, in this case the pixels originating from four different images, and computes the euclidean distance between each vector. From this it calculates a probability that two vectors should be considered neighbours. This probability considers the user-specified hyper parameter *perplexity*, which determines the sizes of the neighbourhoods based on the density of the data in the respective regions. The hyper parameter can effectively be understood as a estimate for the number of neighbours that should be considered similar. Values between 1 and 50 are suggested by the author to be the most appropriate [12]. A low perplexity emphasizes local structure in the embedding, while a high perplexity will preserve more of the global structure of the samples. A universally good perplexity value cannot be determined, but should be selected manually for each data set. This is because the density of the objects varies depending on the sample size, and thus the number of optimal neighbours will vary. The perplexity is chosen, in this thesis, by performing the t-SNE embedding for various perplexities and determining which structure is best suited for the purpose of classifying dropouts.
t-SNE is stochastic in its nature, and will thus compute different embeddings for the same data set unless the same random seed is specified every time. To be able to reproduce our results, this was therefore implemented. Its stochastic nature lies in the fact that it will try out different configurations for the embedding and evaluate the probabilities, which it will try to improve in the next iteration. Due to the huge dimension of the data, this makes the process much faster, but also means that the embedding is not reversible, and new objects cannot be added individually to the embedding space. All data thus has to be given to the algortihm at once. We can specify the maximum number of iterations for the procedure, and the algorithm will output the number of iterations it actually used to produce the embedding. This means we can also use the number of iterations as a tool for choosing the right perplexity. If the number of iterations is smaller than the maximum number of iterations, this means the embedding has converged in the sense that it could not find a better configuration for the embedding during the next 300 iterations (since the default value for the parameter *n_iter_without_progress* is 300).
Since the t-SNE algortithm is a form of unsupervised machine learning, the embedding map will be produced without any direct knowledge of what a dropout galaxy looks like. It will however place objects it considers similar close to each other, even though the algorithm is unaware that the features are pixels from a series of telescope images. This is also one of the reasons t-SNE is so popular: the flexibility in its applications makes it a very powerful tool.

2.2.2 Visual Classification of Tracers

Thus far the machine learning has been completely unsupervised, but now that a 2-d representation of the sample is obtained, we need to understand in which region dropout galaxies are located with another method, since t-SNE, by itself, is not enough to produce predictions. As mentioned in section 2.2.1, t-SNE has no direct knowledge of the features it is mapping, and while it is able to place similar objects close to one another, it is not able to define which categories different regions belong to. Even if the the method performs extremely well for our purpose, in the sense that it would output 3+ tight clusters that are well separated from each other, it would still not be enough, since it would not know which cluster contains what. We say 3+ clusters here, since we know there will be at least three kinds of objects: spurious sources, left-over sources and dropout galaxies.

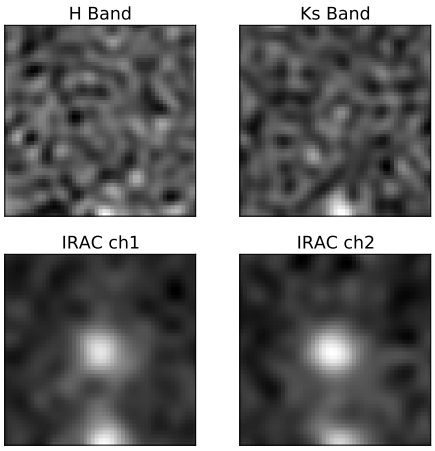

Figure 2. An example of a tracer for NIR-dropouts. The object is clearly visible in the IRAC bands, while it drops out in the NIR bands.

This is therefore where the semi-supervised learning has to enter our method. We select "tracers" in our sample, which are what is determined to be a reliable NIR-dropouts based on a visual inspection of the objects in the four bands. We look for objects that are visible in IRAC $ch1$ and $ch2$, and which do not appear in the NIR bands H and Ks. Since the objects we look for are similar to those found in [6] and [7], we can also use these as a reference. An example of which objects we would classify as a tracer is displayed in fig. 2. We can then map where the tracers are placed in the embedding by the t-SNE algorithm. If the tracers cluster well, this is a good indication of a promising region for finding NIR-dropouts.
For the method adopted in this thesis, we also need "anti-tracers", which are indicators of non NIR-dropouts. These are also selected from a visual inspection of the four bands, this time, however, we look for objects that look nothing alike the tracer displayed in fig. 2.
Since we devise the semi-supervised method for selecting NIR-dropouts in this thesis, we will also need to validate it. This means we need a sample of tracers and anti-tracers for producing predictions, and additionally a similar sample for validating. In case of future use of the method this is, however, not necessary, thus reducing the number of objects that has to be visually classified.

2.2.3 Classification with kNN voting

Now that we have an embedding that is more easily visualised and interpreted, and we have tracers and anti-tracers indicating the interesting regions and the non-interesting regions, we can construct the semi-supervised algorithm used to produce predictions on the sample. An object can either be classified as a NIR-dropout or a non NIR-dropout galaxy. The method takes basis in the principle

of k nearest neighbours and voting. The code performing the classification is divided into the following modules:

1. First, the distances between an unknown object and all tracers are computed, from which the k nearest tracers are selected (a range of values $1 < k < 50$ will be tested).

2. The votes of the k neighbours are then collected. A neighbour that has been previously identified as a tracer will vote for a NIR dropout prediction, while an anti-tracer will vote against.

3. It is checked if the fraction of positive votes is above the threshold f_{min}, which defines the minimum fraction required for predicting a dropout. In the implementation we will experiment different values of f_{min}, ranging from 0 to 1.

4. On behalf of the above, the predictions are computed.

An alternate strategy would be collecting votes from all the neighbours within a fixed radius in the 2-d space. The choice of using nearest neighbours instead of this other option is based on two considerations. First of all the two dimensions of the t-SNE embedding are arbitrary, thus defining a physically meaningful radius is challenging. Furthermore, the embedding represents an approximation of the higher dimensional distribution of objects, and therefore we cannot be certain that a euclidean distance in one region of the embedding is consistent with the distance measured in another region. Assuming a uniform density of objects in the embedding, using the k nearest neighbours, would in turn provide a more consistent voting. By introducing the threshold f_{min}, we are able to modify our selection of NIR dropouts based on the desired balance of quality and quantity. One can then select a parameter of f_{min} that identifies fewer, more robust dropout candidates or be less conservative and obtain predictions of more NIR-dropouts that in turn may be less reliable.

We will evaluate the performance of the model based on the purity of the sample, defined as $TP/(TP + FP)$ where TP is the number of true positives and FP is the number of false positives. We will also produce a receiver operator characteristic (ROC) curve, useful for visualising the performance for different values of f_{min}, and typically an adopted score measure in machine learning. The ROC curve is defined as the true positive rate (TPR) as a function of the false positive rate (FPR) [14].

3 Results

3.1 Residual Map of an IRAC Image

We produced a residual map for a ~ 8 arcmin^2 region in the IRAC $ch1$ telescope image, by adopting the model-based photometry method described in section 2.1.2 for creating intrinsic light profiles of

Model type	Total number of sources	Number of removed sources due to unphysical parameters	Number of sources modelled	Number of flagged objects (of those that are modelled)
Point Source	6844	1635	5209	69
Simple	188	57	131	6
Exponential	1782	432	1350	17
De Vaucouleur	517	198	319	17
Composite	169	5	164	0
Sum	9500	2327	7173	109

Table 2. Overview of the number of sources modelled in tile 7-7 of the IRAC ch1 telescope image with parameters extracted from Farmer. Columns from left to right: (1) The model type used, (2) Total number of detected sources, (3) Number of objects that had to be removed when performing a sanity check, (4) Number of sources left, that were actually modelled, (5) The number of objects that were modelled even though the fitting process in Farmer failed to converge.

sources. The parameters used for the Sérsic profiles were extracted by Farmer from the optical-NIR images of COSMOS2020 and then included in the catalogue and the PSF for the IRAC camera was used for convolution. In total we detected 9500 sources in the image and modelled 7173 of them, the vast majority of them being point source galaxies. The number of sources modelled with each profile type is reported in table 2. Here we also report the number of objects we had to remove from the modelling process due to unphysical parameters. Examples of these parameters include negative fluxes, effective radii of zero extent (in the models using this parameter) or axis ratios above 10, which produced unusually elliptical galaxies appearing as narrow elongated lines. These are sources where either the modelling did not converge or the model converged to a wrong value. The number of flagged objects that were modelled is also reported, these are the objects that were already known to be pathological, for example big saturated stars. In the table we see that the vast majority of the sources are modelled as point sources. Furthermore, it is clear that a non-negligible fraction of sources are not modelled. Nonetheless, we produced a combined model image with the valid models, only including the sources (not the background), and placing the centre of each source model at the centroid coordinates reported in the COSMOS2020 Classic catalogue. A 1×1 arcmin^2 section of the resulting residual image is displayed alongside the same region in the IRAC *ch*1 telescope image in fig. 3. From this section of the image it is not immediately clear that we have overlooked some sources, since this would have resulted in bright spots in the residual image. Comparing the two images in the figure, we can see that some models performed well and left an almost empty residual, while others sources leave patterns in the residual map due to imperfect modelling.

An example where the modelling performed well and produced a good residual, for each model type, is displayed in fig. 5. Similarly, two examples, where the profile-fitting did not perform as well, are displayed in fig. 4.

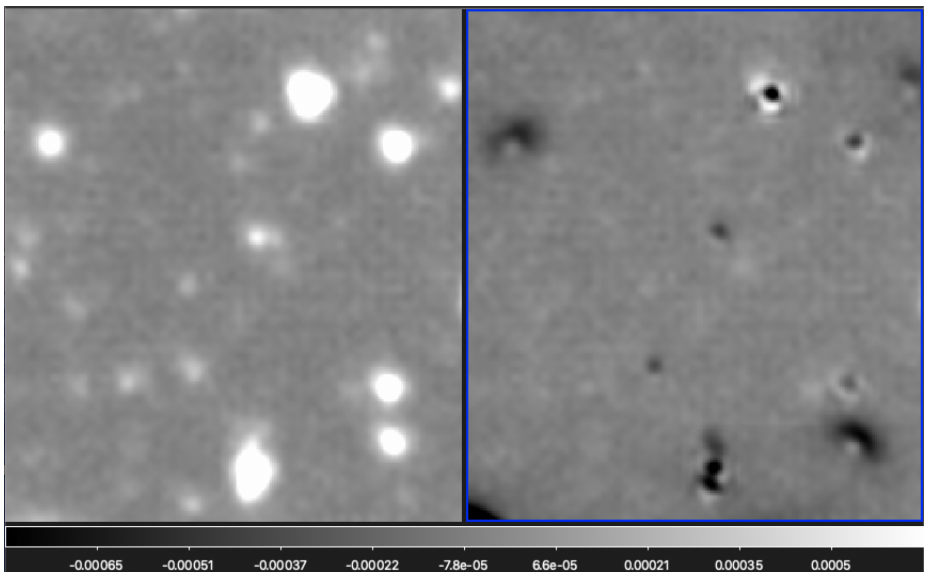

Figure 3. The left plot is a zoomed $\sim$ arcmin^2 region in the IRAC *ch*1 image. To the right, the same region is displayed in the residual tab created utilising the Farmer tool. Colour, limits and scales are all matched between the two.

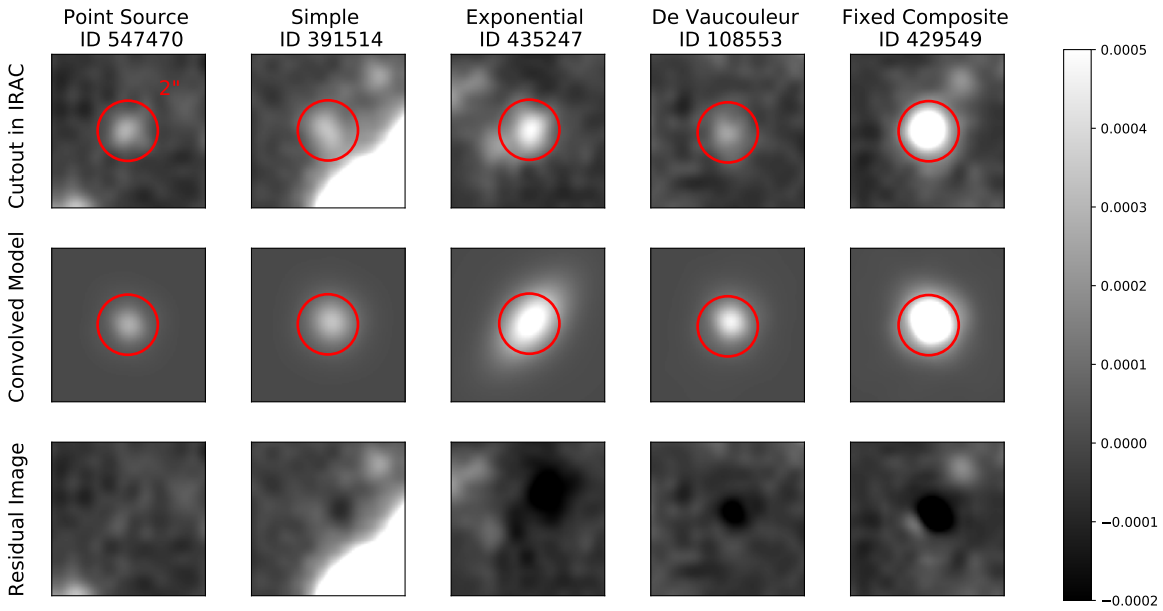

Figure 5. A good example of source modelling, for each of the five Farmer model types. The first row displays a cutout from the telescope image. The intrinsic profiles, are convolved with the IRAC PSF in the second row. The bottom row displays the residual computed by subtracting the second row from the first. All apertures marked are 2". And the cutouts are 21×21 pixels corresponding to a sidelength of 3.15".

3.2 Dimensionality Reduction

Performing source detection with SEP on an IRAC residual map, we identify 3139 objects. We used the classic residual map (described in section 2.1.1), since we did not have time, nor computational power, to produce an IRAC-size robust residual map. Furthermore a few problems arose in the residual map that was created following the description in section 2.1.2 (this is discussed in section 4).

A residual catalogue is created, containing information about the detected sources from telescope images (cutouts of 21×21 pixels corresponding to 3.15"×3.15") in four bands: H, Ks, $ch1$ and $ch2$. These cutouts are used as the input for the t-SNE method, described in section 2.2.1 to create a 2d embedding of the higher dimensional space (21×21×4) in which each object is represented. To select an appropriate perplexity value, the embedding is computed for multiple values of the hyper parameter. The embedding results, from all perplexities explored, are displayed in fig. 12, 13, and 14 in the appendix, where the colours between each consecutive band are mapped. In fig. 6 a few select embeddings are plotted. Here the colour $Ks - ch1$ is used to colour map the sample, since this is where we search for dropouts. The colour is defined as the difference in magnitude between two bands [1]. Due to negative flux values, the magnitudes are sometimes undefined. These objects are assigned another colour in the plot, depending on which band the problem occurred in, so as to be able to distinguish them from the non problematic cases. We chose to continue with a perplexity of 40, since it is in the suggested range between 1 and 50 and it produces a useful clustering with a well separated cluster to the right. Furthermore the embedding was obtained with 6349 iterations which is lower than the maximum number of iterations set to 1000, hence the method has converged.

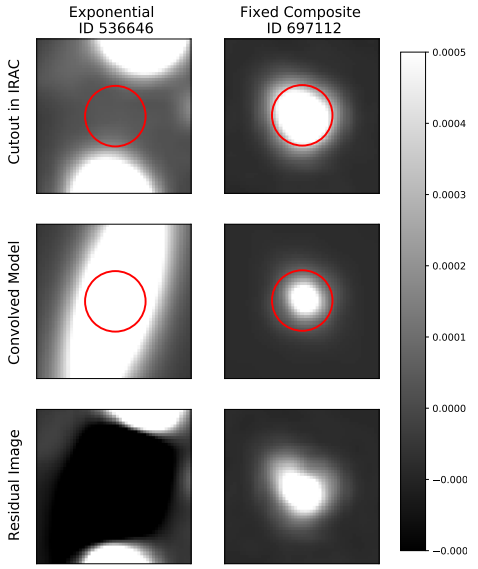

Figure 4. Two examples where the modelling of the sources was not as successful. All apertures marked are 2". And the cutouts are 21×21 pixels corresponding to a sidelength of 3.15".

From manual visual inspection of the four bands, we labelled the best tracers and anti-tracers of NIR dropouts. Furthermore we matched the coordinates of the detected objects in the residual map with the coordinates from the sources in the COSMOS 2020 Classic catalogue, to identify detection due to leftover flux in the residual map. The dense cluster located to the right in the middle plot in fig. 6, is also explored, and mapping the fluxes we find that these are particularly bright sources. In fig. 7 these regions are explored by plotting a cutouts from objects in each region. This figure also shows the cutouts for an object that was mapped far away from the others, it is however deemed

[1] reference to MBW book

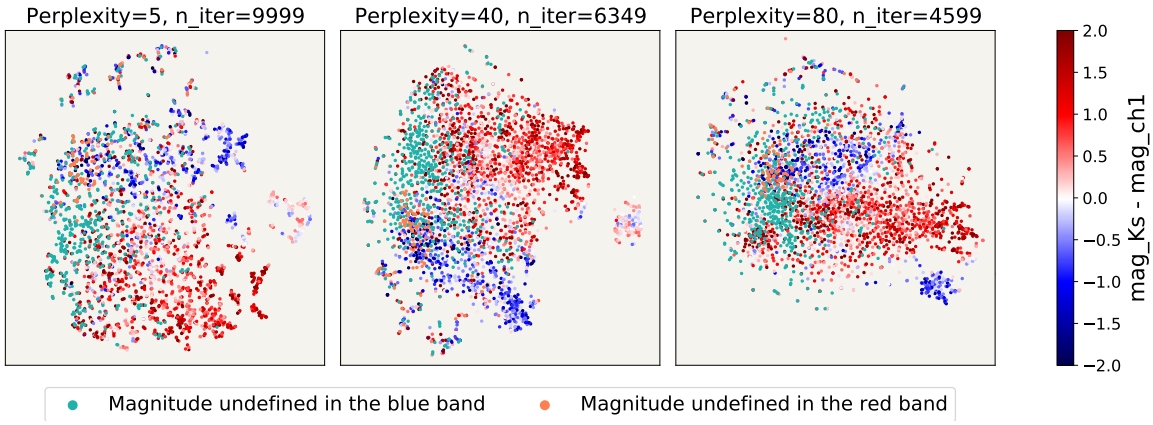

Figure 6. The t-SNE embedding results for three different perplexities. All embedding has a maximum number of iterations set to 1000. n_iter reports the number of iterations actually used.

to be of no interest for our purpose. Looking at the cutout of the tracer, in green, one can see that it is not visible in the NIR bands but appears in the IRAC bands, as is expected for a tracer of a NIR-dropout. The example object from the bright sources region, appears much brighter than the other objects, which is also an indication that this region is in fact bright objects. These cutouts suggests that t-SNE is actually performing a meaningful embedding, where different regions have different features.

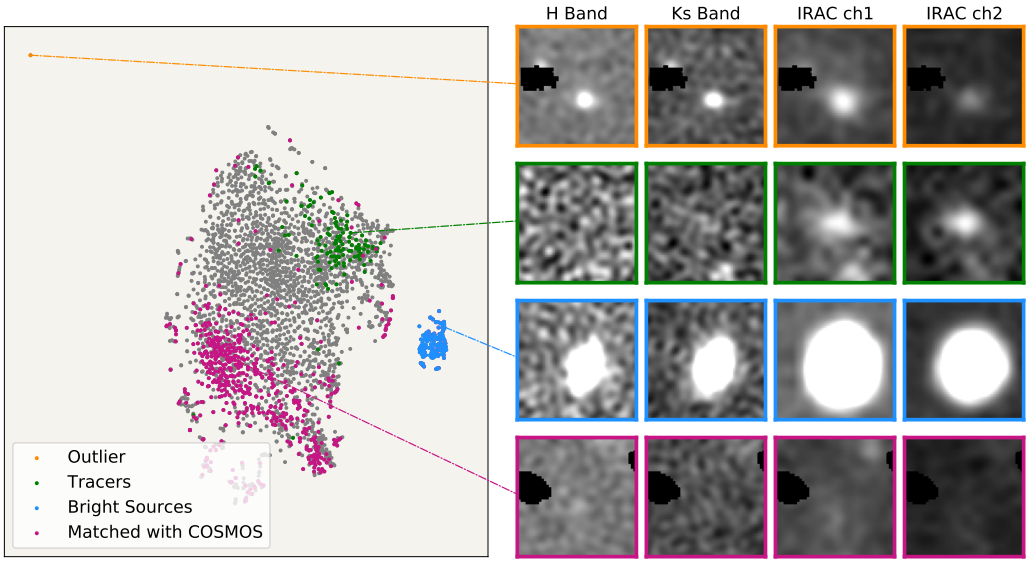

Figure 7. The main plot shows the t-SNE embedding for perplexity 40, where objects in different regions are marked. The telescope images of an object in each region is also plotted, where the colourbar is restrained to the minimum and maximum flux detected for the object shown from the tracer region.

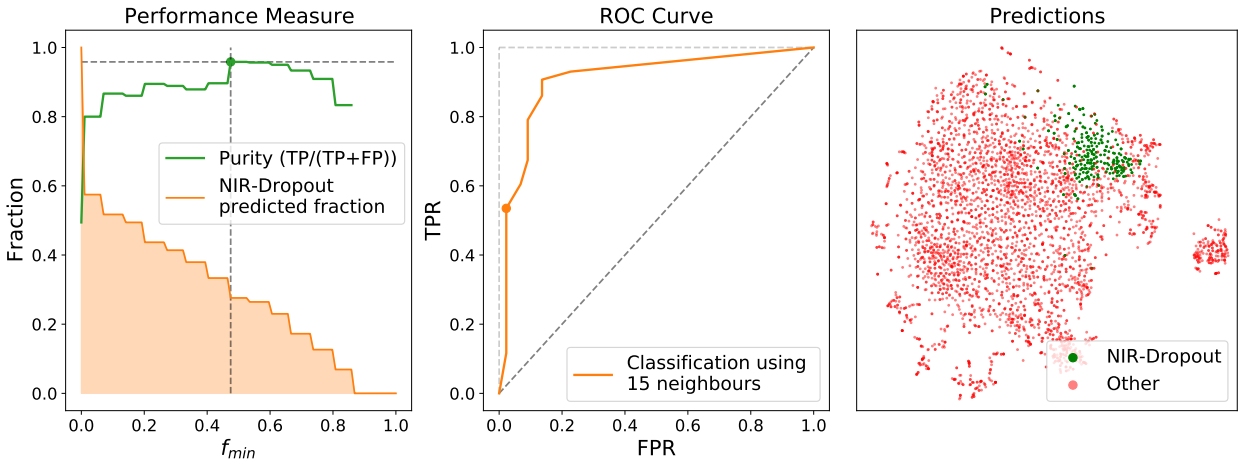

Figure 8. From left to right: (1) Performance measure as a function of the threshold f_{min}. In green the purity is plotted defined as the number of true positive (TP) divided by the number of total number of predicted positives (TP+FP) where FP is the number of false positives. The maximum purity achieved is marked with a green dot. (2) The receiver operating characteristic (ROC) curve. The value of f_{min} corresponding to the maximum purity is marked with an orange dot. (3) The predictions produced with $f_{min} = 0.47$ and $k = 15$ which lead to the highest purity.

3.3 Predictions

Before we produce predictions, we need to select a sample of both the tracers and the anti-tracers for the validation. In total we use 1069 objects anti-tracers and 133 tracers. To produce predictions, respectively 1025 and 90 non dropouts and tracers are used, while respectively 44 and 43 are used for validation of the method. The validation sample is selected so it balances the two possible outputs. We ran the classification code on the training sample, which in addition to the labelled objects also includes unknown objects.

Testing out various k number of neighbours between 1 and 30, we chose $k = 15$, since it produces the best results. With $k = 15$ neighbours and 100 values of f_{min} between 0 and 1, we are able produce the results displayed in the first two left plots in fig. 8. The purity of the NIR dropout predictions is computed for each run of the classification. The parameters $f_{min} = 0.47$ and $k = 15$ produces the highest purity, found to be 96%, and are therefore selected to produce the predictions displayed in the far right plot in fig. 8. Out of 2024 objects, 128 are predicted to be NIR-dropouts.

To inspect the results we plot the SED of the predicted NIR-dropouts, shown in fig. 9. For comparison of different regions in the embedding, the SED for respectively the matched and bright objects shown in fig. 7 are also shown. Due to the definition of a NIR dropout galaxy we expect the predicted dropouts to decrease in magnitude (increase in brightness) between bands Ks and $ch1$, which is also what the figure displays.

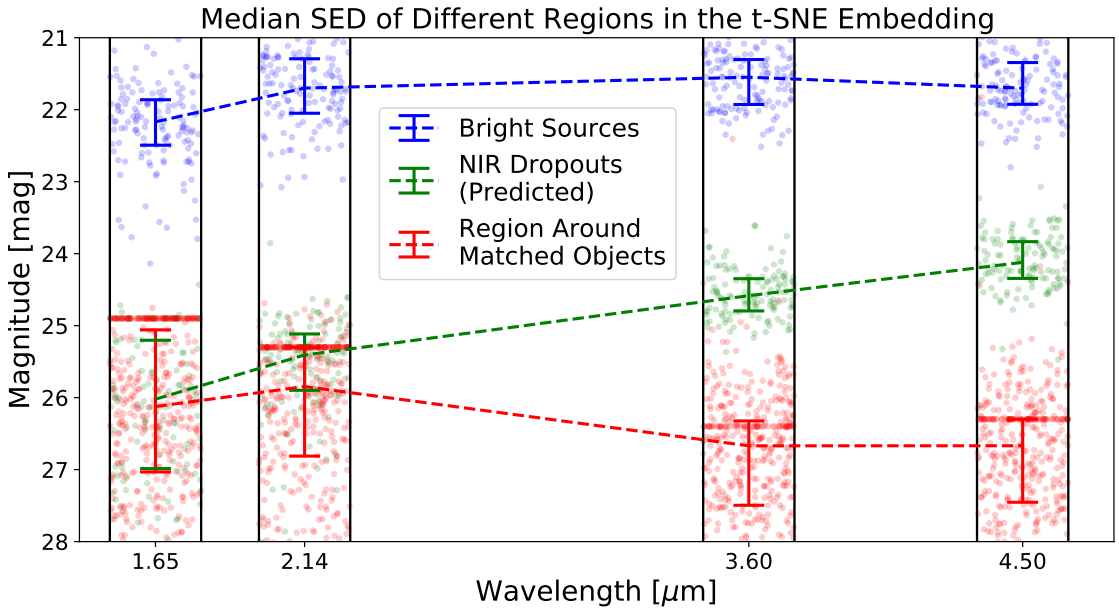

Figure 9. Spectral energy distribution for objects that are found in the bright source region (blue), objects in the densest regions of sources matched with the COSMOS2020 Classic catalogue (red) and objects that are predicted to be NIR-dropouts (green). Undefined magnitudes are replaces with the upper limit of the band reported in [4]. The errorbars on the plot are made using the 25th and 75th percentile of the distributions of magnitudes.

4 Discussion

Throughout this thesis we have mentioned two residual maps: the classic residual map created with IRACCLEAN as described in section 2.1.1, and the Farmer residual map created for a smaller section of the full IRAC image, using the Farmer tool as described in section 2.1.2. Both are mentioned throughout this work because they each have their advantages and drawbacks.
The classic residual map is an older map that has been well tested and uses an established software as its base, and is therefore a very robust residual image. The drawback of this residual map arises from the limitations related to aperture photometry, which is the basis for this map. All sources in the IRAC telescope image are modelled individually with a point source model, where the flux is a free parameter that rescales the profile to fit the source. The flux used is the aperture flux, hence the aperture size is an assumption that is enforced on all sources, even though an adaptive aperture may have been more appropriate. This is not always a problem as long as the aperture size is relatively large, including all flux intrinsic to the source, and the surrounding is background which values are centred around zero. However, since we are working with very faint sources in this project, even a little contamination in the measure of flux might have a significant effect. The problem especially arises if there is a blended pair of galaxies, and the intrinsic flux cannot be properly estimated or divided between the sources. There are thus multiple aspects, we would like to improve in the classic residual map, which was the motivation behind attempting to create a new residual image using the recently developed Farmer tool.
In section 2.1.2 we described how the Farmer overcomes some of these drawbacks, for example by adopting simultaneous modelling of sources in crowded regions. On the other hand, we encountered entirely new problems with this method. For a significant fraction of the models unphysical parameters were listed, and we had to remove them from the modelling process all together. These unphysical parameters are either a result of a failure to converge or the result of converging to a wrong value. An explanation might be that the Farmer during the fitting process fell into a wrong local minimum, so an idea for a solution could be narrowing the range of allowed values to those that are physically meaningful. This problem, of not being able to model all sources, is one of the main reason we had to use the classic residual map for the prediction of NIR dropouts. The Farmer is however still not at its final version, and this is likely something that will be improved in version 2. The possibility that the Farmer tool, in the future, can be used to create a new and better residual image than what can be done with aperture photometry is still there.

Another main result we showed was our semi-supervised classification of NIR-dropouts. While we could have hoped that t-SNE would have produced an embedding with more distinct clusters, the results seem to be robust. We see this especially when considering fig. 7, where t-SNE is able to perform anomaly detection, placing an outlier far away from the main cluster, during the unsupervised part of our technique. This plot in general tells us that the resulting embedding is meaningful with different regions displaying different features.

The technique adopted for producing predictions of dropouts still requires human intervention, as we need to provide the algorithm with tracers and anti-tracers from a visual classification of the cutouts. While it is not ideal, that one needs to manually inspect images to be able to produce an output, the method is after all still much faster than relying only on visual classification, which is an important result when trying to make classifications from larger surveys. An example of such a survey is the *Euclid* which is a wide and deep survey that will perform imaging of very large patches of the sky in optical and NIR down to $\sim 26 - 27$ mag. In such a situation, we will be observing over a 10 times larger area, and classical visual inspection is next to impossible. In such a scenario this machine learning method is definitely the preferred technique. The predictions produced with out method are also quite good, reaching a purity of up to 96%. One can however still tune the parameter f_{min} depending on what the predictions are to be used for. Say one wants to perform spectroscopic follow up on the candidates, a smaller sample that is more reliable is preferable.

5 Conclusion

A novel technique has been devised for the semi-automated classification of NIR-dropouts based on photometric data sampled from four different bands in the COSMOS field. In particular the UltraVISTA bands (H at $1.65\,\mu$m and Ks at $2.16\,\mu$m), along with the IRAC bands: $ch1$ ($\lambda = 3.57\,\mu$m) and $ch2$ ($\lambda = 4.51\,\mu$m) from the COSMOS2020 catalogue. The method takes advantage of semi-supervised machine learning, in particular it uses t-SNE and kNN for producing predictions. It does, however, need to be provided with a sample of labelled objects that are found from visual inspection of the cutout images in each band. With the number of neighbours set to $k = 15$ and the minimum positive voting fraction selected as $f_{min} = 0.47$, the technique predicts 128 NIR-dropouts from the entire unknown sample of 2024 candidates, with a purity of 96%. This machine learning technique for classifying dropouts is strongly preferred for large surveys, where it produces predictions much faster than other techniques, such as manual classification. Suggested future work for the devised technique is to make sure that all steps are easily adaptable to also produce predictions for candidates from other fields than the COSMOS. An example of a large and deep, next generation survey perfect for detecting NIR-dropouts is the *Euclid*.

The sample of candidates the technique produces predictions on, is found from source detection in a residual image. It was explored whether it is possible to produce a new, better residual map of the IRAC $ch1$ image, utilising model-based photometry, in particular the Farmer tool [4], rather than a classical approach based on aperture photometry. It is found, that with the current version of the Farmer tool, this is not within reach. It is however plausible that a future version improving the way that the model optimisation is performed, can produce a better, more robust residual map.

References

[1] E. L. Wright. "A Cosmology Calculator for the World Wide Web". In 118.850 (Dec. 2006), pp. 1711–1715. DOI: 10.1086/510102. arXiv: astro-ph/0609593 [astro-ph].

[2] Charles C. Steidel et al. "Spectroscopic Confirmation of a Population of Normal Star-forming Galaxies at Redshifts Z ¿ 3". In 462 (May 1996), p. L17. DOI: 10.1086/310029. arXiv: astro-ph/9602024 [astro-ph].

[3] Houjun Mo, Frank van den Bosch, and Simon White. *Galaxy Formation and Evolution*. Cambridge University Press, 2010. DOI: 10.1017/CBO9780511807244.

[4] J. R. Weaver et al. "COSMOS2020: A next-generation catalog to explore the 1 ¡ z ¡ 8 universe". In *American Astronomical Society Meeting Abstracts*. Vol. 53. American Astronomical Society Meeting Abstracts. Jan. 2021, p. 215.06.

[5] C. Laigle et al. "THE COSMOS2015 CATALOG: EXPLORING THE 1 ¡ z ¡ 6 UNIVERSE WITH HALF A MILLION GALAXIES". In *The Astrophysical Journal Supplement Series* 224.2 (June 2016), p. 24. ISSN: 1538-4365. DOI: 10.3847/0067-0049/224/2/24. URL: http://dx.doi.org/10.3847/0067-0049/224/2/24.

[6] Belén Alcalde Pampliega et al. "Optically Faint Massive Balmer Break Galaxies at z ¿ 3 in the CANDELS/GOODS Fields". In *The Astrophysical Journal* 876.2 (May 2019), p. 135. ISSN: 1538-4357. DOI: 10.3847/1538-4357/ab14f2. URL: http://dx.doi.org/10.3847/1538-4357/ab14f2.

[7] T. Wang et al. "A dominant population of optically invisible massive galaxies in the early Universe". In *Nature* 572.7768 (Aug. 2019), pp. 211–214. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1452-4. URL: http://dx.doi.org/10.1038/s41586-019-1452-4.

[8] Kyle Barbary. *SEP: Source Extraction and Photometry*. Nov. 2018. ascl: 1811.004.

[9] E. Bertin and S. Arnouts. "SExtractor: Software for source extraction." In 117 (June 1996), pp. 393–404. DOI: 10.1051/aas:1996164.

[10] Bau-Ching Hsieh et al. "THE TAIWAN ECDFS NEAR-INFRARED SURVEY: ULTRA-DEEP J AND K S IMAGING IN THE EXTENDED CHANDRA DEEP FIELD-SOUTH". In *The Astrophysical Journal Supplement Series* 203.2 (Nov. 2012), p. 23. DOI: 10.1088/0067-0049/203/2/23. URL: https://doi.org/10.1088/0067-0049/203/2/23.

[11] Jose Luis Sersic. *Atlas de Galaxias Australes*. 1968.

[12] Laurens van der Maaten and Geoffrey Hinton. "Viualizing data using t-SNE". In *Journal of Machine Learning Research* 9 (Nov. 2008), pp. 2579–2605.

[13] Charles L. Steinhardt et al. "A Method to Distinguish Quiescent and Dusty Star-forming Galaxies with Machine Learning". In *The Astrophysical Journal* 891.2 (Mar. 2020), p. 136. ISSN: 1538-4357. DOI: 10.3847/1538-4357/ab76be. URL: http://dx.doi.org/10.3847/1538-4357/ab76be.

[14] Tom Fawcett. "An introduction to ROC analysis". In *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. URL: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

Appendix A Source Detection in Optical and NIR bands

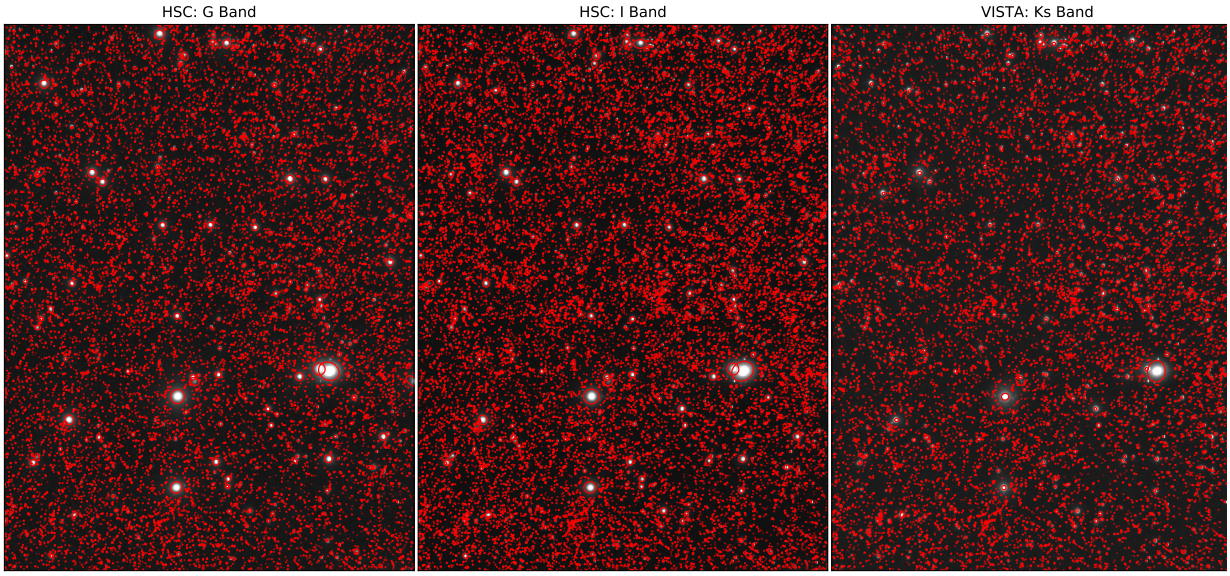

Figure 10. Results from source detection with SEP. An adaptive aperture is plottet at the detected centroid for each source in the entire image from respectively the g, i (optical) or Ks (NIR) band.

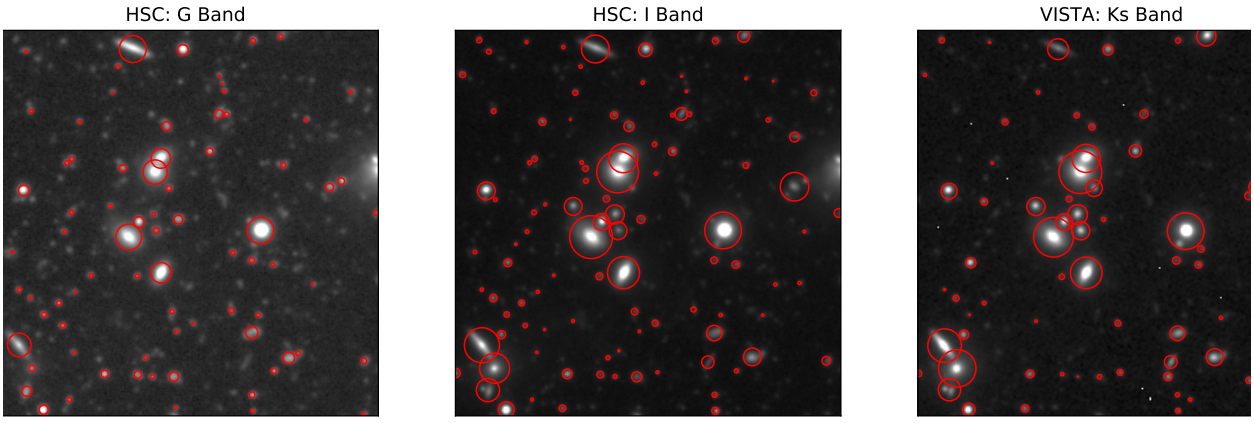

Figure 11. Zoomed version of fig. 10. The images are 500×500 pixels.

Appendix B Color Mapped t-SNE Embeddings

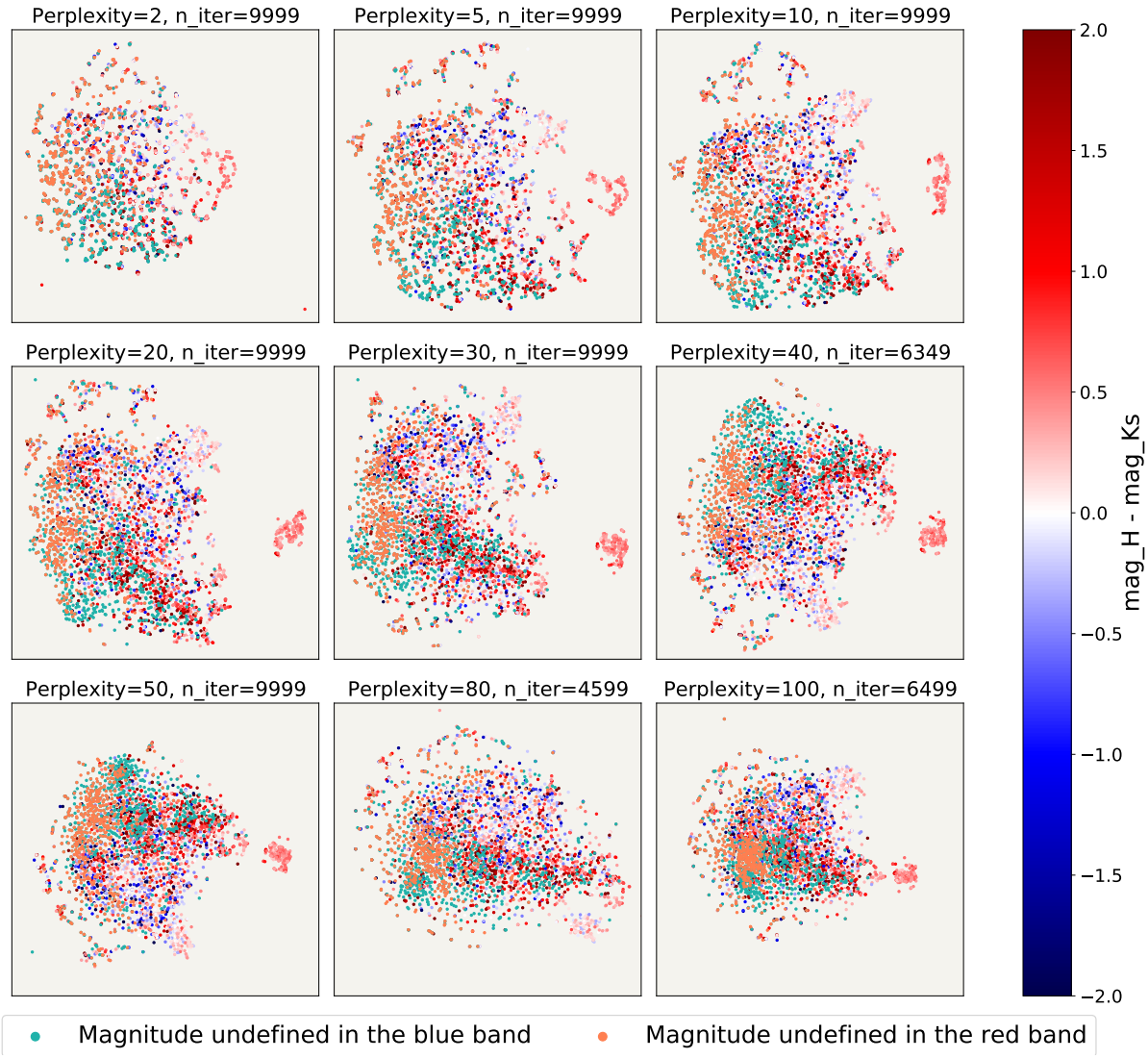

Figure 12. The t-SNE embedding results for nine different perplexities. All embedding has a maximum number of iterations set to 1000. n_iter reports the number of iterations actually used. The samples are colour mapped with the colour defined as $m_H - m_{Ks}$

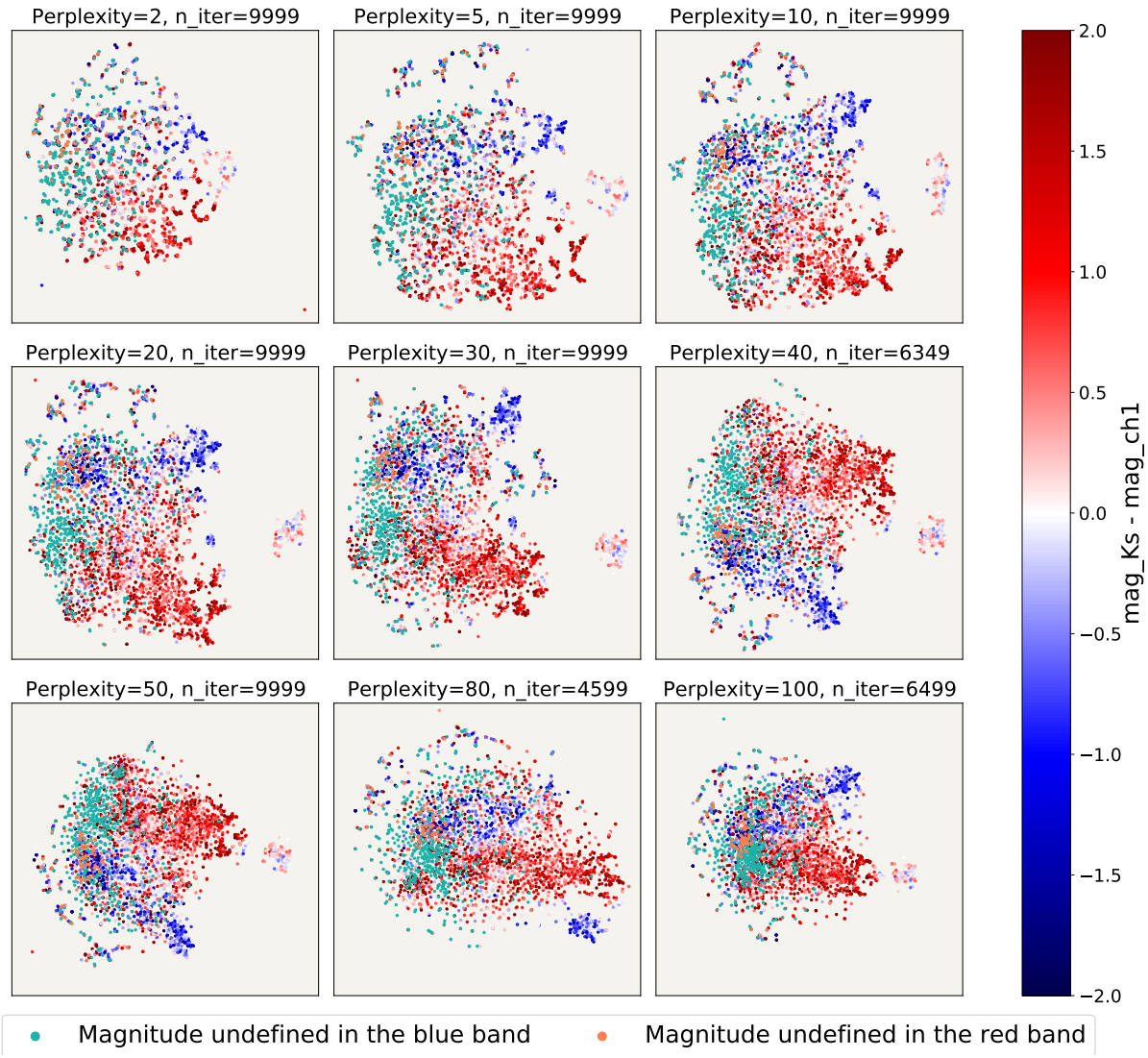

Figure 13. The t-SNE embedding results for nine different perplexities. All embedding has a maximum number of iterations set to 1000. n_iter reports the number of iterations actually used. The samples are colour mapped with the colour defined as $m_{Ks} - m_{ch1}$

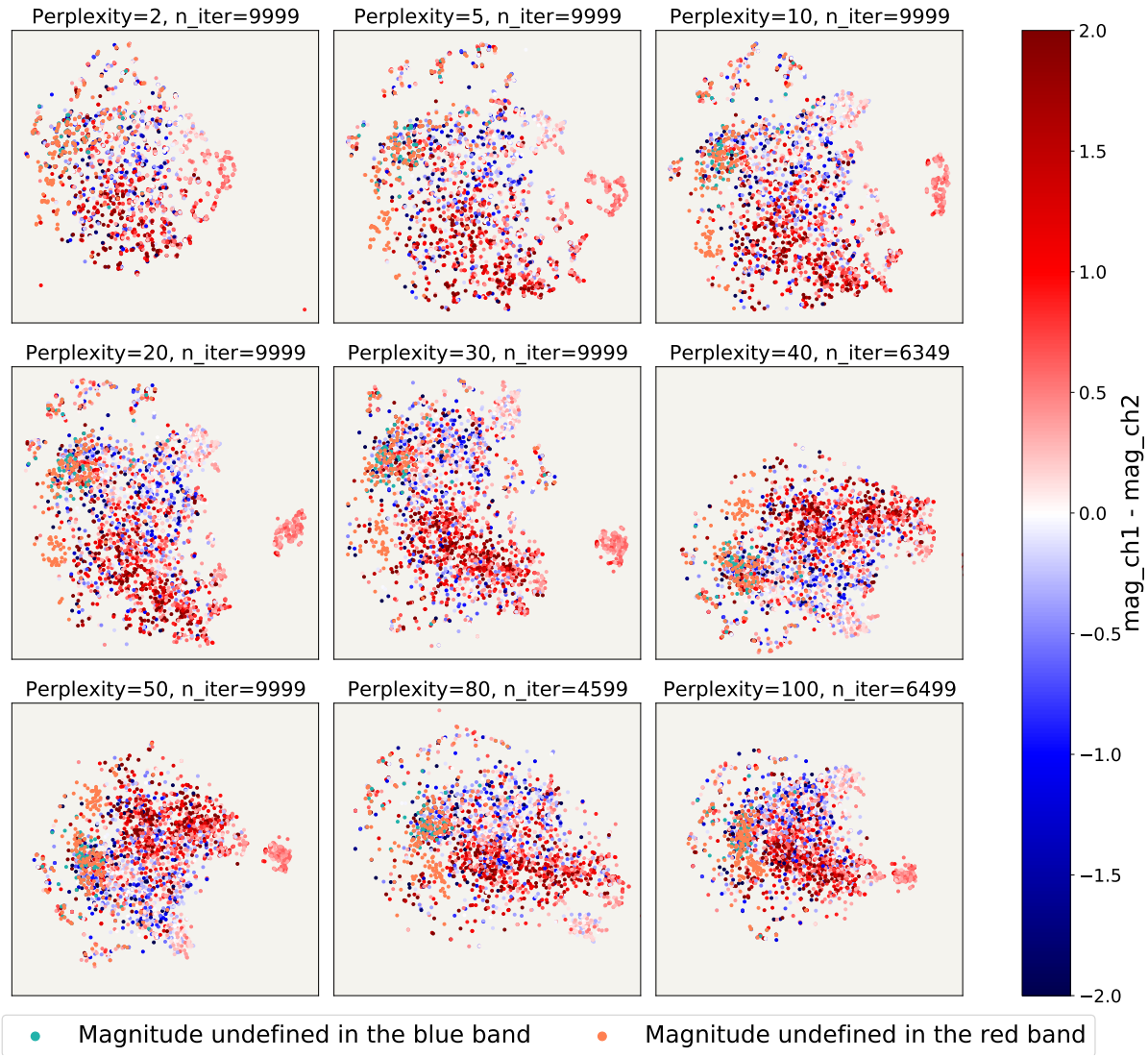

Figure 14. The t-SNE embedding results for nine different perplexities. All embedding has a maximum number of iterations set to 1000. n_iter reports the number of iterations actually used. The samples are colour mapped with the colour defined as $m_{ch1} - m_{ch2}$