

A novel analysis of contamination in Lyman-break galaxy samples at $z \sim 6 - 8$: spatial correlation with intermediate-redshift galaxies at $z \sim 1.3 - 2$

Miftahul Hilmi^{1,2} ,^{1,2}★ Nicha Leethochawalit^{3,1,2} , Michele Trenti^{1,2} and Benjamin Metha^{1,2} 

¹*School of Physics, the University of Melbourne, Parkville, VIC 3010, Australia*

²*ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), Australia*

³*National Astronomical Research Institute of Thailand (NARIT), Mae Rim, Chiang Mai, 50180, Thailand*

Accepted 2024 June 11. Received 2024 June 11; in original form 2023 October 26

ABSTRACT

Potential contamination from low/intermediate-redshift galaxies, such as objects with a prominent Balmer break, affects the photometric selection of high-redshift galaxies through identification of a Lyman break. Traditionally, contamination is estimated from spectroscopic follow-up and/or simulations. Here, we introduce a novel approach to estimating contamination for Lyman-break galaxy (LBG) samples based on measuring spatial correlation with the parent population of lower redshift interlopers. We propose two conceptual approaches applicable to different survey strategies: a single large contiguous field and a survey consisting of multiple independent lines of sight. For a large single field, we compute the cross-correlation function between galaxies at redshift $z \sim 6$ and intermediate-redshift galaxies at $z \sim 1.3$. We apply the method to the CANDELS GOODS-S and XDF surveys and compare the measurement with simulated mock observations, finding that the contamination level in both cases is not measurable and lies below 5.5 per cent (at 90 per cent confidence). For random-pointing multiple field surveys, we measure instead the number count correlation between high-redshift galaxies and interlopers, as a two-point correlation analysis is not generally feasible. We show an application to the LBG samples at redshift $z \sim 8$ and the possible interloper population at $z \sim 2$ in the Brightest of Reionizing Galaxies (BoRG) survey. By comparing the Pearson correlation coefficient with the result from Monte Carlo simulations, we estimate a contamination fraction of 62^{+13}_{-39} per cent, consistent with previous estimates in the literature. These results validate the proposed approach and demonstrate its utility as an independent check of contamination in photometrically selected samples of high-redshift galaxies.

Key words: methods: statistical – surveys – galaxies: high-redshift.

1 INTRODUCTION

Thanks to space-based observatories, thousands of galaxy candidates at redshift $z \gtrsim 6$ have been recently discovered, primarily from several large survey programmes conducted with the *Hubble Space Telescope* (Schmidt et al. 2014; Bouwens et al. 2015; Ishigaki et al. 2015; Morishita et al. 2018; Bowler et al. 2020; Salmon et al. 2020; Roberts-Borsani et al. 2022). The identification of high-redshift galaxy candidates is conducted photometrically through the Lyman-break technique (Steidel et al. 1996), which relies on the identification of a strong spectral break at a wavelength shorter than 1216 Å. This method heavily depends on the colour information of the sources and is therefore subject to contamination from objects with similar photometry, such as cool stars or intermediate-redshift red galaxies. In particular, one of the main sources of contamination for Lyman-break galaxy (LBG) samples are low/intermediate-redshift Balmer break galaxies with a prominent break at 3646 Å rest frame (Atek et al. 2011; van der Wel et al. 2011).

To minimize contamination in photometric catalogues, deep observations at wavelengths shorter than the spectral break are generally required to distinguish between a faint continuum of an interloper and a true non-detection for a high-redshift galaxy. For example Stanway, Bremer & Lehnert (2008) suggest using a set of non-overlapping but adjacent filters to be able to impose a clear colour cut on the selection and thus reduce contamination. Deeper imaging follow-up observations on previously identified candidates also shows that additional photometry blueward of the Lyman break can help discriminating between low- and high-redshift galaxies (e.g. Livermore et al. 2018). Yet, contamination is unavoidable in photometrically selected samples, and thus needs to be understood.

Contamination from intermediate-redshift galaxies can contribute to bias in estimating the high-redshift ultraviolet (UV) luminosity functions (Morishita et al. 2018), in addition to other sources of bias, such as magnification bias (Wyithe et al. 2011; Mason et al. 2015), bias due to the cosmic variance (Trenti & Stiavelli 2008; Moutard et al. 2016; Bowler et al. 2020), and bias due to photometric scatter (Leethochawalit et al. 2022). Previous studies also show that contamination levels becomes higher with increasing redshift. Vulcani et al. (2017) found that the ratio of interlopers to dropouts

* E-mail: mhilmi@student.unimelb.edu.au

grows significantly as a function of redshift. Using a simple model that relies on the dark matter halo mass function, Furlanetto & Mirocha (2023) also found that the expected contamination level increases drastically at $z \gtrsim 10$, requiring stricter selection criteria to identify high-redshift sources robustly.

While spectroscopic observations are the most robust approach to verify high-redshift candidates, they require large investment of telescope time and/or are unfeasible for objects near the detection limit of imaging surveys. Thus, several works have proposed methods to estimate the contamination level. Finkelstein et al. (2015) artificially dim real lower redshift sources to see if this allows them to be selected as high-redshift candidates. This method implicitly assumes that contaminants have similar spectral energy distributions (SEDs) to the known lower redshift sources. They estimate a relatively small contamination fraction of ~ 5 per cent–15 per cent in the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS) Great Observatories Origins Deep Survey (GOODS) fields, which is in agreement with the estimation from the candidates' redshift probability distributions produced by photometric redshift fitting code ($P(z)$ curves). In another work, Rojas-Ruiz et al. (2020) estimate the contamination by downgrading deep images from the *Hubble Frontier Fields* (HFFs) programme (Lotz et al. 2017) to the depth of the shallower *Hubble* images used in their work. By comparing the redshifts determined across the six HFF and the redshifts determined in the counterpart downgraded images, they found one contaminant and concluded that contaminants do not contribute significantly to their sample. Alternatively, Trenti et al. (2011) apply the colour selection used to select their high-redshift sample to a library of SED models of lower redshift galaxies taking into account the depth of the observations for the survey modelled.

This paper proposes a novel conceptual framework to assess contamination, and presents two implementations of the idea. The first approach is based on the spatial correlation between the high-redshift galaxy candidates and known galaxies at the redshift of potential interlopers. It is appropriate for a large contiguous survey. The basic principle is that the angular cross-correlation function of high-redshift and intermediate-redshift galaxies should not indicate any clustering, unless some level of contamination exist. This approach is inspired by Ménard et al. (2013), Schmidt et al. (2014), Rahman, Ménard & Scranton (2016a), and Rahman et al. (2016b), where clustering analyses were proposed to refine photometric redshift estimates. These works typically consider two populations: a reference population with known redshift and angular positions, and the other population with only angular positions known. The redshifts of the second population can be determined when there is a cross-correlation signal with the reference population. The concept of spatial correlation has also been applied in other studies to measure contamination in various samples. Grasshorn Gebhardt et al. (2019) and Farrow et al. (2021) use the cross-correlation function to estimate the contamination fraction of low-redshift [O II] ($z < 0.5$) emitters in the intermediate redshift Ly α emitters sample ($1.9 < z < 3.5$) to estimate the unbiased cosmological parameters. Addison et al. (2019) also suggests the use of cross-correlation function to constrain the contamination fraction in [O III] sources sample due to the misidentification of H α spectral line. Awan & Gawiser (2020) presents a correlation function estimator that can correct for sample contamination, by taking into account the auto and cross-correlation function of the sources and contaminants. In our work, we take this concept to study contamination of LBG samples at high redshift.

The second approach is to quantify the number count correlation between high-redshift galaxies and the possible contaminants at intermediate redshift. It is appropriate for analysing the contami-

nation level of a random-pointing survey with multiple fields. The approach is adapted from the counts-in-cells method proposed by Robertson (2010) to quantify the clustering properties of galaxies for observations that consist of a large number of uncorrelated fields, which has been implemented by Cameron et al. (2019) on BoRG observations. These methods are based on the sources' angular positions and number counts. They therefore minimize the reliance on manipulating/analysing the SEDs of candidates and on simulated high-redshift galaxies, and provide an independent way to cross-check estimates obtained through traditional methods.

This paper is organized as follows. In Section 2, we describe the angular cross-correlation technique to estimate contamination fractions and apply it to CANDELS data. In Section 3, we model the cross-correlation function using mock catalogues generated from IllustrisTNG simulation, to determine what level of contamination this technique is sensitive to. Section 4 discusses the number count analysis based on BoRG samples. We summarize our results and conclusion in Section 5. Throughout this paper, we adopt a cosmological parameter set of $\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$, and $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. All magnitudes are represented in the AB system (Oke & Gunn 1983).

2 CROSS-CORRELATION ANALYSIS

This section explores the spatial correlation between high-redshift galaxies and lower redshift galaxies at the *interloper redshift*, which we define to be the redshift range in which intermediate-redshift galaxies resemble high-redshift galaxies photometrically, and may contaminate the high-redshift galaxies samples. More specifically, this is the redshift where the observed Balmer break of intermediate-redshift galaxies is at the same wavelength as the observed Lyman break of high-redshift galaxies:

$$1216 \text{ \AA} (1 + z_{\text{high}}) = 3646 \text{ \AA} (1 + z_{\text{interloper}}). \quad (1)$$

Here, the Ly α wavelength 1216 $\text{\AA}$ is used instead of the Lyman limit 912 $\text{\AA}$, because for high-redshift galaxies at $z \gtrsim 6$, the continuum between 912 and 1216 $\text{\AA}$ is absorbed by intervening Ly α forest (Madau 1995; Giavalisco 2002).

The method rests on the lack of physical correlation between galaxies at high-redshift and galaxies at interloper redshift since the typical correlation length of dark matter haloes is orders of magnitude smaller than the comoving line-of-sight distance between the two populations. Therefore, the two samples should be uncorrelated unless some galaxies at the lower redshift are misidentified as high-redshift galaxies and contaminate the high- z sample. Based on this, we hypothesize that we should be able to constrain the contamination rate based on the spatial correlation between high-redshift candidates and known galaxies at the interloper redshift. The so-called Schrodinger's galaxy presented in Naidu et al. (2022) is a good illustration of this idea. The SED fitting of the galaxy suggests that the galaxy is at $z \sim 17$ with a small probability to be at $z \sim 5$. However, the galaxy is in the vicinity of three neighbouring galaxies that are at $z \sim 5$. Hence, the authors suggest that the source could also likely be part of the protocluster.

2.1 Data set

To obtain statistically robust spatial correlations, we need large samples of galaxies at both high-redshift and interloper redshift observed in the same survey with a large contiguous area. With this requirement, we use the data set of the GOODS-South and the Hubble eXtreme Deep Field (XDF) fields from the Hubble Legacy Fields Data Release V2.5 (Illingworth et al. 2016; Whitaker

et al. 2019).¹The area of the GOODS-S survey is 64.5 arcmin², while for XDF it is 4.7 arcmin². We decided to use $z \sim 6$ (specifically $z = 5.5 - 6.5$) galaxies as our main high- z sample to ensure the sample is sufficiently large to enable two-point correlation function measurements. The corresponding interloper redshift is $z = 1.2 - 1.5$ with the average redshift equal to $z \sim 1.3$. We use the catalogue from Merlin et al. (2021) to obtain the sample of intermediate-redshift galaxies, and the catalogue from Bouwens et al. (2021) as the high-redshift galaxies sample. Merlin et al. (2021) selected their samples in H_{160} band and used SED fitting to determine the redshifts for objects with no spectroscopic redshifts available. On the other hand, the $z \sim 6$ samples in Bouwens et al. (2021) are detected in $Y_{105} J_{125} J H_{140} H_{160}$ stacked images and are selected based on Lyman-break colour criteria.

Based on the catalogue from Merlin et al. (2021), we select galaxies at redshift $z = 1.2$ to $z = 1.5$, yielding 3379 and 292 galaxies located on GOODS-S and XDF fields, respectively. For high- z galaxies, we select the galaxies at redshift $5.6 < z < 6.5$ from Bouwens et al. (2021). There are 323 and 129 such galaxies on GOODS-S and XDF fields, respectively. Due to the different depths between the edge part and central part of the GOODS-S survey, the completeness of the survey is non-homogeneous and this may introduce systematic errors in our cross-correlation analysis. Therefore, for the GOODS-S area, we restrict the samples to those in the central region with a uniform depth (see Fig. 1).

As the two catalogues are provided by different studies, we investigated if there is any common candidate in the $z \sim 1.3$ and $z \sim 6$ catalogues. We find that there are eight sources in the GOODS-S field and three sources in the XDF field that are reported in both catalogues. We removed those sources from $z \sim 1.3$ catalogue and assigned them only to the high-redshift catalogue, since our study aims to check the quality of the $z \sim 6$ catalogue. Our final samples consist of 1387 $z \sim 1.3$ galaxies and 191 $z \sim 6$ galaxies in the GOODS-S field. The locations of the samples are also shown in Fig. 1. For XDF, our final samples contain 289 and 129 sources in the $z \sim 1.3$ and $z \sim 6$ catalogue, respectively. As an additional note, one of the $z \sim 6$ galaxies in GOODS-S sample is also spectroscopically confirmed to be at $z \sim 1.3$ (Vanzella et al. 2008), but the source is retained in the photometric sample. While it is possible that high-redshift galaxies are misidentified as intermediate-redshift galaxies, the fraction will be very small. The number of galaxies at intermediate-redshift is much higher than those at high-redshift. Therefore, the contamination in intermediate-redshift galaxies sample by high-redshift galaxies is assumed to be zero.

2.2 Analysis

The angular correlation function, $\omega_{\text{cor}}(\theta)$, measures the clustering of galaxies by comparing the observed number of galaxy pairs relative to the expected number of galaxy pairs from a random distribution. The angular correlation function of galaxies at any redshift can generally be described by a power-law function: $\omega_{\text{cor}} = A_{\omega} \theta_i^{-\beta}$ (Lee et al. 2006; Overzier et al. 2006; Barone-Nugent et al. 2014).

To analyse whether there is significant contamination within $z \sim 6$ galaxies sample, we calculate the cross-correlation function between $z \sim 1.3$ and $z \sim 6$ galaxies. Similar to the angular correlation function, the cross-correlation function measures the excess probability of finding a pair of galaxies from two different populations within an angular separation θ . We use the modified Landy–Szalay estimator

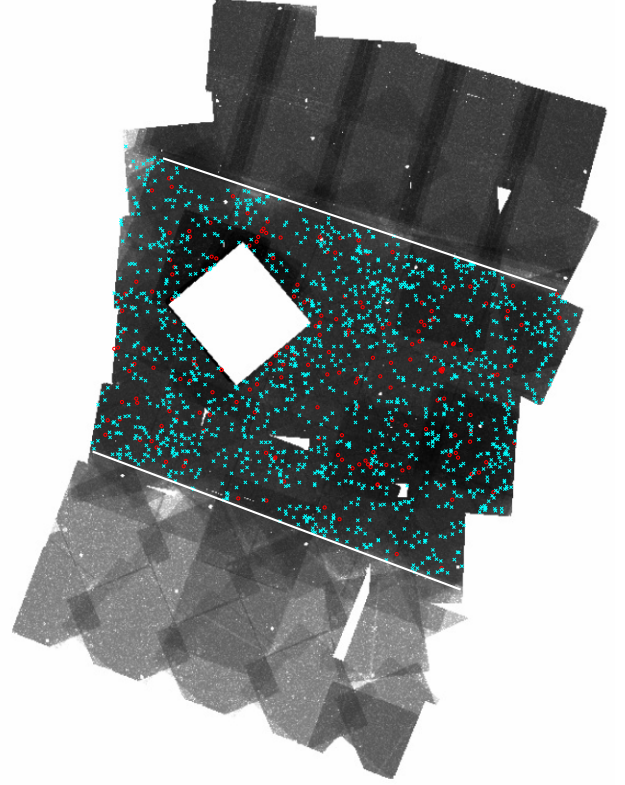


Figure 1. Root mean square (rms) image of GOODS-S field taken in $F125W$ band and shown in logarithmic scale. As we can see from the image, the central region has a different depth compared to the edge regions. Therefore, we only consider galaxies inside the white diagonal lines, with sources marked as cyan (lighter colour) crosses for $z \sim 1.3$ galaxies and red (darker colour) circles for $z \sim 6$ galaxies.

(Landy & Szalay 1993; Blake et al. 2006):

$$\omega_{\text{cross}}(\theta) = \frac{D_1 D_2(\theta) - D_1 R_2(\theta) - D_2 R_1(\theta) + R_1 R_2(\theta)}{R_1 R_2(\theta)}, \quad (2)$$

where $D_1 D_2(\theta)$, $D_1 R_2(\theta)$, $D_2 R_1(\theta)$, and $R_1 R_2(\theta)$ are the number of $z \sim 1.3$ galaxy and $z \sim 6$ galaxy pairs, $z \sim 1.3$ galaxy and random point pairs, $z \sim 6$ galaxy and random point pairs, and random-random point pairs, all measured within an angular separation of $\theta \pm \delta_\theta$, respectively. For our study, the random point catalogue is generated by taking the depths of fields in each filter into account in the same manner as described in detail in Dalmaso, Trenti & Leethochawalit (2024). This process is undertaken to prevent artificial clustering signals induced by non-uniform depth variations. In summary, we randomly inject galaxies with Sérsic light profile in the images of all detection bands. The final random catalogue consists of the injected galaxies that are recovered with the same procedures used for galaxy detection in the GOODS-S (Merlin et al. 2021) and XDF (Bouwens et al. 2021) catalogues. For simplicity, we use the same random catalogue for both galaxy samples ($R_1 = R_2 = R$). We estimate the cross-correlation function in bins of θ , using linear binning with a bin width of $\delta_\theta = 7''.2$. We use Bootstrap re-sampling (Ling, Frenk & Barrow 1986) to estimate errors in the cross-correlation function by re-sampling the data set 10 times. We show the resulting cross-correlation functions with squares and error bars in Fig. 2. Visually, there is no correlation signal in both GOODS-S (left panel) and XDF (right panel) fields.

¹<https://archive.stsci.edu/prepds/hlf/>

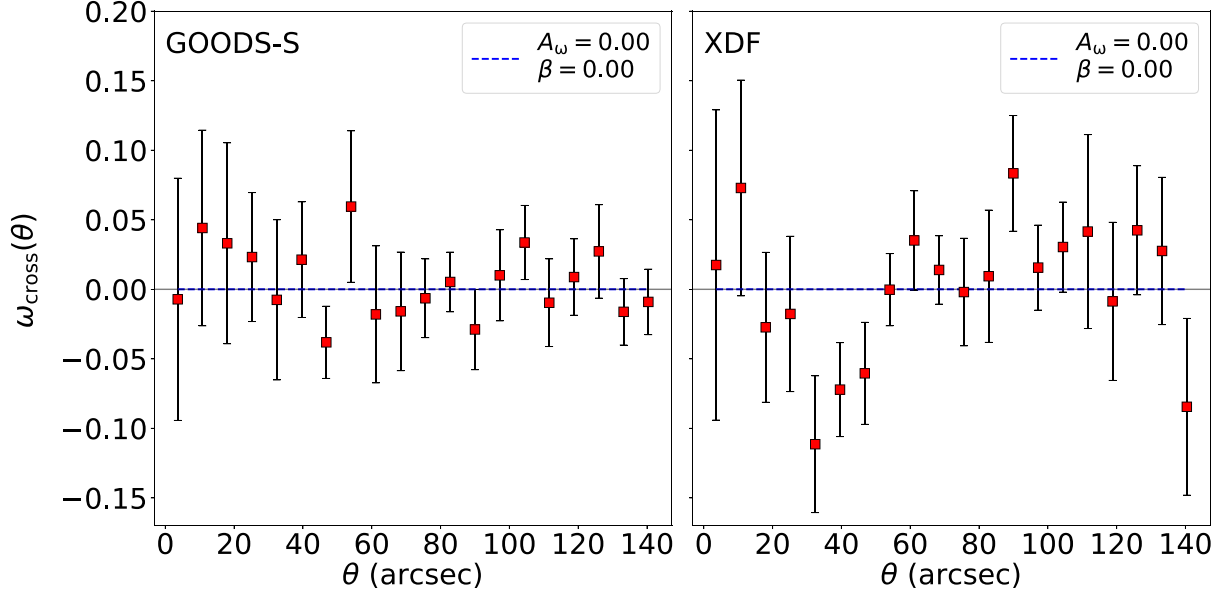


Figure 2. Cross-correlation function of $z \sim 1.3$ and $z \sim 6$ galaxies in the GOODS-S field (left) and XDF (right). Dashed line is the power-law function of $\omega_{\text{cross}}(\theta) = A_{\omega}\theta^{-\beta}$, where A_{ω} and β are the best-fitting parameters constrained by the χ^2 fitting with equation (4).

To investigate the significance of the cross-correlation signal, we conduct a statistical test between the two models of the cross-correlation function. If there is significant contamination, we expect a clustering signal between galaxies at $z \sim 1.3$ and $z \sim 6$. The intrinsic correlation functions of galaxies at both redshifts are power-law functions. Since both populations have similar power-law slopes², we therefore expect the cross-correlation function of these two populations also to take the functional form of a power-law function:

$$\omega_{\text{cross}}(\theta) = A_{\omega}\theta^{-\beta}. \quad (3)$$

However, due to the finite survey area, observed cross-correlation function $\omega_{\text{cross,obs}}(\theta)$ are generally underestimated by a constant factor known as integral constraint (IC):

$$\omega_{\text{cross,obs}}(\theta) = A_{\omega}\theta^{-\beta} - \text{IC}(A_{\omega}, \beta). \quad (4)$$

IC can be calculated by doubly integrating the cross-correlation function $\omega_{\text{cross}}(\theta)$ over the survey area Ω (Roche & Eales 1999):

$$\begin{aligned} \text{IC} &= \frac{1}{\Omega^2} \int_1 \int_2 \omega_{\text{cross}}(\theta) d\Omega_1 d\Omega_2 \\ &= \frac{\sum_i RR(\theta_i) \omega_{\text{cross}}(\theta_i)}{\sum_i RR(\theta_i)} = \frac{\sum_i RR(\theta_i) A_{\omega} \theta_i^{-\beta}}{\sum_i RR(\theta_i)}. \end{aligned} \quad (5)$$

On the other hand, if there is no significant contamination, we expect that the cross-correlation function between galaxies at $z \sim 1.3$ and $z \sim 2$ will follow a random distribution and be given by:

$$\omega_{\text{cross}}(\theta) = \omega_{\text{cross,obs}}(\theta) = 0. \quad (6)$$

²Studies of galaxy correlation function often assume a fixed power-law slope. For example $\beta = 0.8$ is assumed for all galaxies across $z = 0 - 6$ in Arnouts et al. (1999). More recent works (e.g. Barone-Nugent et al. 2014; Dalmasso et al. 2024) use a fixed $\beta = 0.6$ for $z \gtrsim 4$ galaxies. We measured the correlation function for our $z \sim 1.3$ galaxies (see Section 3.3). The measured β values are 0.45 ± 0.19 and 0.81 ± 0.79 for the GOODS-S and the XDF fields, respectively. They are consistent with the power-law slopes assumed for $z = 6$ galaxies in the literature.

To take into account the correlation between measurement in different angular bins, we construct the normalized covariance matrix using the standard estimator:

$$C_{ij} = \frac{1}{N-1} \sum_{l=1}^N [\omega^l(\theta_i) - \bar{\omega}(\theta_i)] [\omega^l(\theta_j) - \bar{\omega}(\theta_j)]. \quad (7)$$

In this equation, the summation is over N independent realizations. However, our Bootstrap samples are not from independent realizations. When the covariance matrix is estimated from the data itself, such as Bootstrap re-sampling, a correction factor of $(N-1)/N$ has to be added, and the covariance matrix becomes:

$$C_{ij} = \frac{N_{\text{boot}} - 1}{N_{\text{boot}}} \sum_{l=1}^{N_{\text{boot}}} [\omega^l(\theta_i) - \bar{\omega}(\theta_i)] [\omega^l(\theta_j) - \bar{\omega}(\theta_j)], \quad (8)$$

where N_{boot} is the total number of Bootstrap samples, $\omega^l(\theta)$ is the measured cross-correlation function from each Bootstrap realization, and $\bar{\omega}(\theta)$ is the mean of cross-correlation function. Due to the relatively small sample size, our resulting covariance matrix is noisy and the inverse of the covariance matrix is ill-conditioned and numerically unstable. Therefore, we apply a ridge regression technique (Hoerl & Kennard 1970; Matthews & Newman 2012) by adding a small value c to the diagonal elements of the covariance matrix to reduce the impact of noise in the off-diagonal elements. We use $c = 0.0001$ as our parameter value (approximately 1 per cent of the median value of the diagonal elements).

Using equations (4) and (5) together, we can estimate the best-fitting parameters A_{ω} and β using the χ^2 minimization method under the conditions that $A_{\omega} \geq 0$ and $\beta \geq 0$:

$$\chi^2 = \sum_{i,j} [\omega(\theta_i) - \omega_{\text{model}}(\theta_i)]^T C_{ij}^{-1} [\omega(\theta_j) - \omega_{\text{model}}(\theta_j)], \quad (9)$$

where $\omega(\theta)$ is the cross-correlation function measured from our data set, $\omega_{\text{model}}(\theta)$ is the cross-correlation function as defined by equation (4), and C_{ij}^{-1} is the inverse of covariance matrix given by the equation (8). We list the best-fitting parameters in the legends of Fig. 2. The best-fitting models are essentially flat straight lines. We do not report

the uncertainties as the system is unbounded i.e. infinite combinations of A_{ω} and β can yield a flat line on the horizontal axis beyond a few arcsec scale.

We find that for both GOODS-S and XDF samples, the best-fitting parameters follow equation (6), suggesting that any contamination that may exist in this sample is too small to be measurable by the cross-correlation analysis. Nevertheless, we know that at least one of 191 sources in the GOODS-S catalogue is spectroscopically confirmed to be a low-redshift source, and therefore the minimum contamination fraction (i.e. the ratio of the number of contaminants to the number objects identified as high-redshift galaxies) in the GOODS-S field is $f_{\text{cont}} \geq 0.5$ per cent. This result suggests to consider what level of contamination would introduce a measurable signal using this technique. We answer this question in the next section, using mock observations from a cosmological simulation.

3 MODELLING THE CROSS-CORRELATION FUNCTION

In this section, we model the cross-correlation function as a function of contamination level. To do so, we perform the same analysis in Section 2 on the mock catalogues generated from *The Next Generation Illustris* simulations (IllustrisTNG; Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018; Springel et al. 2018; Pillepich et al. 2018b) and use a Monte Carlo method to randomly select contaminants from interlopers.

3.1 Illustris mock catalogue

The IllustrisTNG (The Next Generation) simulation suite is a collection of large-volume cosmological magnetohydrodynamical simulations that model galaxy formation, galaxy evolution, and large-scale structure formation within the Λ cold dark matter paradigm. It is the follow-up project of the Illustris simulation series (Genel et al. 2014; Vogelsberger et al. 2014a,b; Nelson et al. 2015; Sijacki et al. 2015). Similarly to Illustris, IllustrisTNG is run using the quasi-Lagrangian moving-mesh code AREPO (Springel 2010), which combines aspects of smooth particle-based hydrodynamical simulations with adaptive mesh-based simulations, in order to avoid numerical issues that each of these other methods possess (Vogelsberger et al. 2013). Subgrid physical prescriptions are used to model a large variety of astrophysical processes that are relevant for galaxy formation and evolution, including stochastic star formation, black hole formation and growth, stellar and active galactic nucleus (AGN) feedback, metal enrichment and cooling, and cosmic magnetic fields. The complete description of the TNG galaxy formation model is presented in the two TNG methods papers (Weinberger et al. 2017; Pillepich et al. 2018a).

In this study, we use the TNG300-1 simulation from the IllustrisTNG simulation suite, with a volume of 302.6 cMpc^3 and a mass resolution of $10^7 M_{\odot}$ per baryonic particle. From this simulation, we generate multiple mock catalogues of high-redshift and intermediate-redshift galaxies. First, we download Snapshots 14 and 43 of this simulation from the TNG public data base³, which correspond to redshift $z = 1.30$ and $z = 5.85$, respectively. Next, we select galaxies within 170 non-overlapping cut-outs from each snapshot to generate position and photometry catalogues. The dimensions of each cut-out box at $z = 1.30$ is $(13.34 \times 8.33 \times 151.90) \text{ cMpc}$, and $(17.44 \times 27.91 \times 165.30) \text{ cMpc}$ at $z = 5.85$. These box sizes were

chosen to be equivalent to a projected 0.12×0.2 degree sky survey in the observer frame, which is approximately the size of the GOODS-S area. In Snapshot 14 at $z = 5.85$, all sub-boxes are oriented such that their long sides align with the Z-axis of the TNG300-1 simulation, with the central X- and Y-positions evenly drawn from a 10×17 grid that avoids the edges of the simulation volume. In Snapshot 43 at $z = 1.30$, we instead orient the cut-out boxes so that the long side aligns with the X-axis of the TNG300-1 simulation, and evenly sample the Y- and Z-positions from a 10×17 grid that avoids both the edges of the simulation, and the area of the simulation volume from which the cut-out boxes at $z = 5.85$ were drawn. This was done to avoid spurious correlations between galaxy clusters observed at $z = 5.85$ and their own progenitors at $z = 1.30$. The generated catalogues contain the position, stellar mass, and photometry of the sources in rest frame U, B, V, K, g, r, i , and z bands.

We convert the i absolute magnitude of galaxies at redshift $z = 1.30$ into apparent magnitude at wavelength $\lambda = 17204 \text{ \AA}$.

$$m = M + 5 \log \left(\frac{D_L}{10 \text{ pc}} \right) - 2.5 \log(1 + z), \quad (10)$$

where M is the absolute magnitude in emitted frame, m is the apparent magnitude in observer frame, D_L is the luminosity distance, and z is the source's redshift. Similarly, we convert the U magnitude of galaxies at $z = 5.85$ into apparent magnitude at $\lambda = 25688 \text{ \AA}$. These wavelengths are the closest wavelengths to the detection bands of the Merlin et al. (2021) and Bouwens et al. (2021) catalogues where the photometry information is available. To simulate with a condition close to the current observation limit, we only select galaxies with an apparent magnitude up to 28.5 for both $z = 1.30$ and $z = 5.85$ galaxies.

To ensure that the mock fields from the simulations match the observation geometrically, we first rotate the mock field to match the position angle of the observation. We then apply the same field of view to ensure that the edges of the mock field have the same shape as those of the observation. Lastly, we apply the segmentation map generated by SExtractor (Bertin & Arnouts 1996) to cut-out the mock galaxies that would have been blocked by foreground galaxies in the real observation. As a sanity check, we measure the average angular correlation function of the simulated galaxies that survive the geometry cut above at both $z = 1.3$ and $z = 5.85$. The average angular correlation functions of all simulated fields are consistent with the observed angular correlation functions of $z \sim 1.3$ and $z \sim 6$ galaxies within 1σ error.

3.2 Monte Carlo simulation

We define sources with apparent magnitude ≥ 24.0 from the mock catalogue of $z = 1.30$ as faint intermediate-redshift galaxies. In our simulation, these sources can contaminate the LBG sample. We introduce a leakage fraction f_{leak} as the probability that a faint $z = 1.30$ galaxy will be misidentified as a $z = 5.85$ galaxy and become a contaminant. We perform a Monte Carlo simulation to study how the cross-correlation function changes as a function of the leakage fraction. We conduct the simulation for $f_{\text{leak}} = 0$ per cent – 10 per cent with a step size of 1 per cent. For $f_{\text{leak}} = 0$ per cent, we calculate the cross-correlation function in the same way as the previous section. We use the bootstrap re-sampling method to derive the mean and the error of the cross-correlation function. For $f_{\text{leak}} > 0$ per cent, we estimate the uncertainty using the Monte Carlo method. For each sub-box, we repeat the process of assigning different $z \sim 1.3$ galaxies as contaminants according to the leakage fraction and remeasuring the cross-correlation function 10 times for each sub-box. The means and

³<https://www.tng-project.org/data/downloads/TNG300-1/>

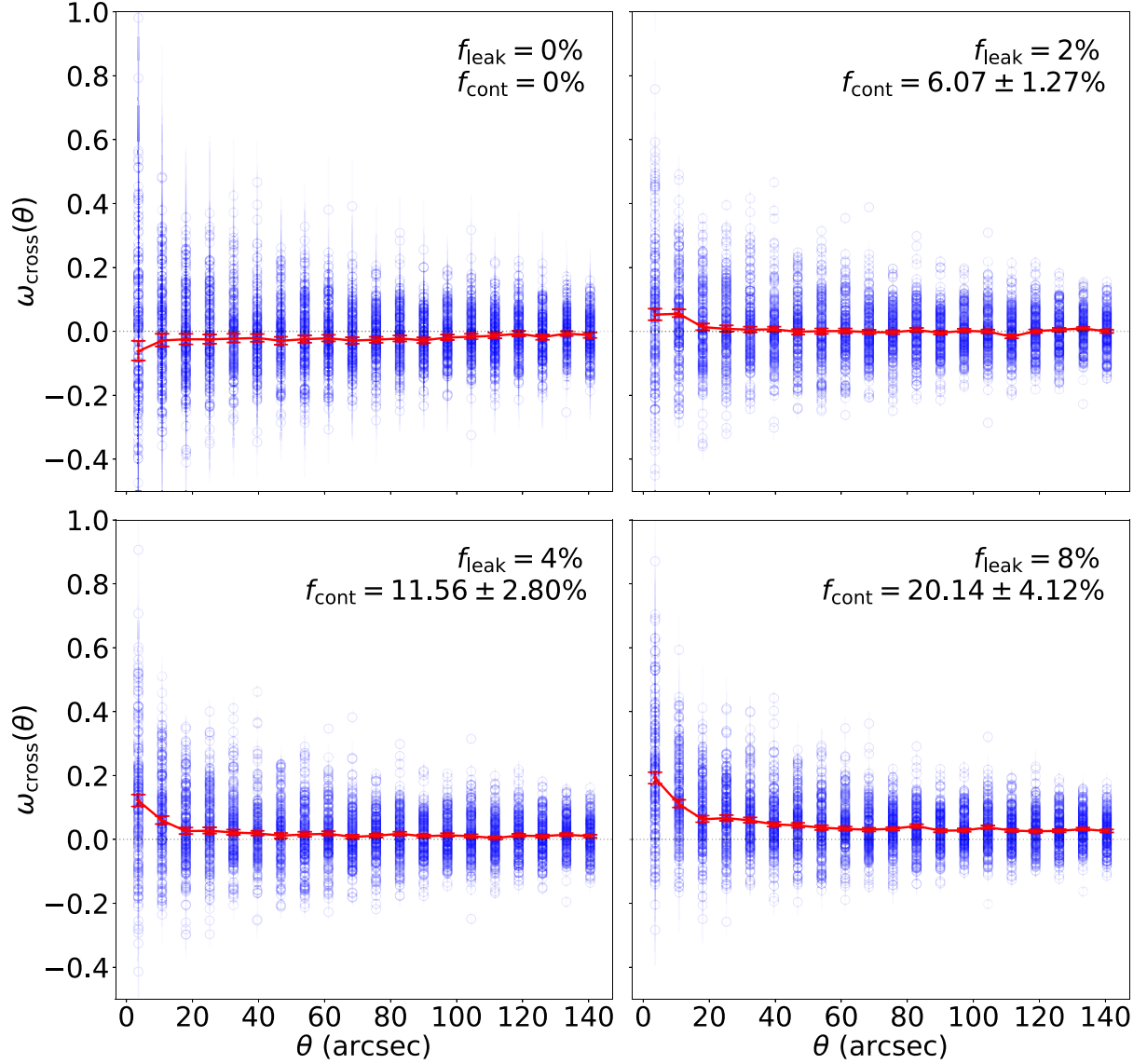


Figure 3. Cross-correlation function of galaxies at redshift $z \sim 1.30$ and $z \sim 5.85$ based on Illustris mock catalogues. The empty blue circles represent each data point generated from the Monte Carlo simulation. The solid red line is the mean and its standard error. Using a Monte Carlo simulation, we adjust the contamination level by increasing the leakage fraction (f_{leak}). As we change f_{leak} from 0 to 8 per cent, the cross-correlation signal increase proportionally.

uncertainties for these individual sub-boxes are shown as empty blue circles in Fig. 3. The weighted mean of all 170 sub-boxes are shown as solid red lines in the same figure. To ease the interpretation, we also convert the leakage fraction into the contamination fraction (f_{cont}) from each simulation. The average value for f_{cont} is displayed for each panel in Fig. 3. Based on Fig. 3, increasing the leakage fraction will increase the clustering signal in the cross-correlation function. Thus, a clustering signal in the cross-correlation function indicates contamination, in agreement with our hypothesis and the results from previous studies (Addison et al. 2019; Grasshorn Gebhardt et al. 2019; Awan & Gawiser 2020; Farrow et al. 2021).

To estimate at what level of contamination we will see a significant cross-correlation signal, we conduct a statistical test for each combination of simulation box. We consider a hypothesis test where the null hypothesis (H_0) is that equations (4) and (6) fit the data equally well – i.e. that there is no cross-correlation between the sample of high-redshift galaxies and the sample at the interloper

redshift. The alternative hypothesis (H_1) is that equation (4) fits the data significantly better than equation (6). Thus, we should use equation (4) where the measured cross-correlation function can be parametrized as a power-law function. To determine which model is preferred (equation 4 or equation 6), we use the Akaike information criterion (AIC; Akaike 1974), taking the number of parameters into account. The model with a smaller AIC value is preferred. For each model, the AIC value is given by:

$$\text{AIC} = 2k - 2 \ln(\mathcal{L}), \quad (11)$$

where k is the number of free parameters, and $\mathcal{L}$ is the likelihood of the model given by:

$$\mathcal{L} = \prod_{i,j} \frac{1}{2\pi^{p/2} |C_{ij}|^{1/2}} \exp \left\{ -\frac{1}{2} [\omega(\theta_i) - \omega_{\text{model}}(\theta_i)]^T C_{ij}^{-1} [\omega(\theta_j) - \omega_{\text{model}}(\theta_j)] \right\}, \quad (12)$$

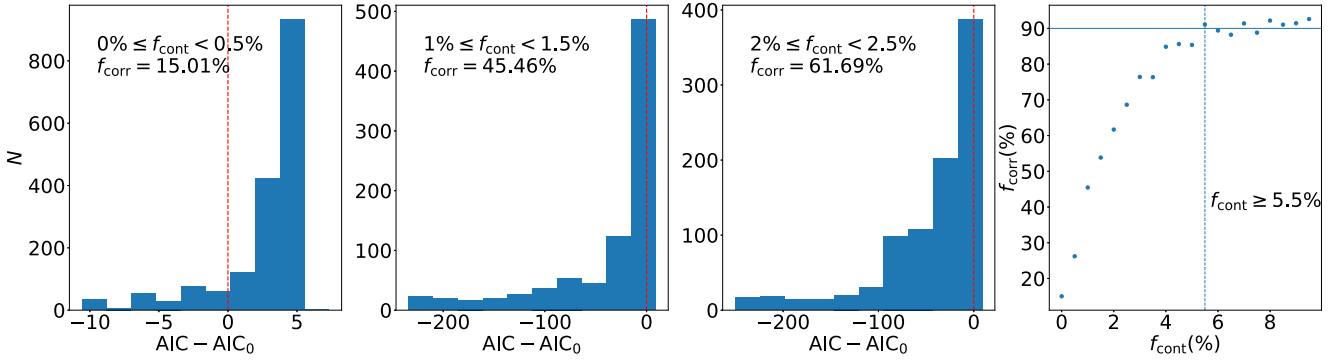


Figure 4. First three left panels: Histograms of $AIC - AIC_0$ for three bins of $f_{\text{cont}} = 0$ per cent – 0.5 per cent, 1 per cent – 1.5 per cent, 2 per cent – 2.5 per cent. Negative value (left side of dashed vertical line) indicating that the simulation shows a cross-correlation signal. Right panel: Fraction of simulation showing a cross-correlation signal (f_{corr}) as a function of contamination fraction. For contamination fractions greater than $f_{\text{cont}} = 5.5$ per cent (dashed vertical line), most simulations show a cross-correlation signal, as indicated by $f_{\text{corr}} \sim 90$ per cent (horizontal line).

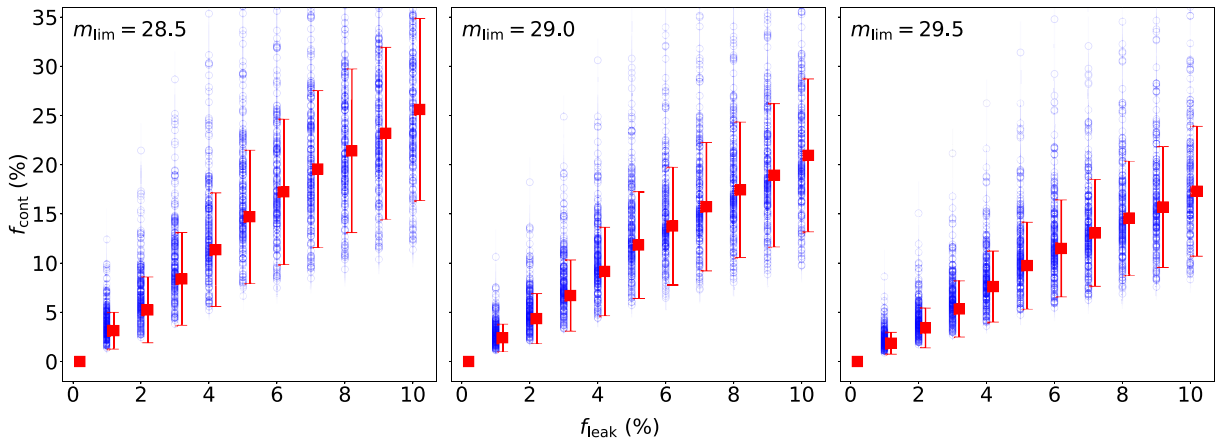


Figure 5. Contamination fraction as a function of leakage fraction for galaxies at redshift $z \sim 1.30$ and $z \sim 5.85$ based on Illustris mock catalogues using three different magnitude cuts. The empty blue circles represent each data point generated from the Monte Carlo simulation. The filled red square is the mean and its standard deviation (slightly shifted to the right for visual clarity). By adjusting the limiting magnitude of our sample in the catalogues, we find that for deeper field, the contamination fraction (f_{cont}) is lower for the same value of leakage fraction (f_{leak}). Therefore, contamination can be minimized with deeper observations.

where p is the number of bins in θ , $|C_{ij}|$ is the determinant of the covariance matrix, $\omega(\theta)$ is the measured cross-correlation function, and $\omega_{\text{model}}(\theta)$ is the modeled cross-correlation function based on equations (4) or (6). Due to the small number of simulated galaxies, the errors in cross-correlation functions are dominated by the Poisson noise. Therefore, we use only the diagonal elements in the covariance matrix, in the same manner as Zheng, Coil & Zehavi (2007) and Harikane et al. (2016). By using the GOODS-S and XDF observational data in Section 2.2, we have tested that the use of the off-diagonal elements in covariance matrix does not change the conclusion of the best-fitting model. Using a fixed IC derived from all of the simulations, we calculate $AIC - AIC_0$ – that is, the difference between the AIC value of model (4) and the AIC value of model (6). If this value is negative, then the simulation produces a significant cross-correlation signal. We calculate how many of the simulations produce significant correlations as a function of contamination fraction. We generate a histogram of $AIC - AIC_0$ for each bin of 0.5 per cent contamination fraction. We show our result in Fig. 4. For $f_{\text{cont}} \geq 5.5$ per cent, ~ 90 per cent of simulations consistently show cross-correlation signal. Therefore, we conclude that the level of contamination in the GOODS-S field is less than 5.5 per cent (at 90 per cent confidence).

To test how the contamination level depends on the depth of the survey, we repeat our Monte Carlo analysis process using two deeper limiting magnitudes of 29.0 and 29.5. We present the results of this experiment in Fig. 5. As we use the fainter magnitude cut, the error bars of the cross-correlation become smaller. The clustering signal for the same value of leakage fraction also becomes smaller, indicating a lower contamination fraction.

We conclude that the contamination level depends on the depth of the survey. This result can be explained as a consequence of the steepening of high-redshift galaxy UV luminosity function towards the faint-end. As the depth of a survey is increased, the number of actual high-redshift galaxies increases more rapidly than the number of intermediate-redshift interlopers.

3.3 Contamination fraction calculation based on previous literature

Awan & Gawiser (2020) introduces a formalism that uses the observed cross-correlation function in the contaminated sample to estimate the true cross-correlation function. Based on their work, the observed cross-correlation function is contributed by four types of pairings. In our context, the four pairings are: (1) between true

$z = 1.3$ galaxies and true $z = 6$ galaxies, (2) true $z = 1.3$ galaxies and observed $z = 6$ galaxies that are actually at $z = 1.3$, (3) true $z = 6$ galaxies and observed $z = 1.3$ galaxies that are actually at $z = 6$, and (3) observed $z = 1.3$ galaxies that are actually at $z = 6$ and observed $z = 6$ galaxies that are actually at $z = 1.3$. The final observed cross-correlation function is:

$$\begin{aligned} \omega_{\text{cross}}^{\text{obs}}(\theta) = & f_{z \sim 1.3}^{\text{true}} f_{z \sim 6}^{\text{true}} \omega_{\text{cross}}^{\text{true}}(\theta) \\ & + f_{z \sim 1.3}^{\text{true}} f_{z \sim 6}^{\text{cont}} \omega_{z \sim 1.3}^{\text{true}}(\theta) \\ & + f_{z \sim 1.3}^{\text{cont}} f_{z \sim 6}^{\text{true}} \omega_{z \sim 6}^{\text{true}}(\theta) \\ & + f_{z \sim 6}^{\text{cont}} f_{z \sim 1.3}^{\text{cont}} \omega_{\text{cross}}^{\text{true}}(\theta), \end{aligned} \quad (13)$$

where $\omega_{\text{cross}}^{\text{obs}}(\theta)$ is the observed cross-correlation function, $\omega_{\text{cross}}^{\text{true}}(\theta)$ is the true cross-correlation function, $\omega_{z \sim 1.3}^{\text{true}}(\theta)$ is the true angular correlation function (ACF) for galaxies at $z \sim 1.3$, $\omega_{z \sim 6}^{\text{true}}(\theta)$ is the true ACF for galaxies at $z \sim 6$, $f_{z \sim 1.3}^{\text{true}}$ is the fraction of galaxies at $z \sim 1.3$ that are not contaminant from $z \sim 6$ galaxies, $f_{z \sim 6}^{\text{true}}$ is the fraction of galaxies at $z \sim 6$ that are not contaminant from $z \sim 1.3$ galaxies, $f_{z \sim 1.3}^{\text{cont}}$ is contamination fraction in $z \sim 1.3$ galaxies sample, and $f_{z \sim 6}^{\text{cont}}$ is the contamination fraction in $z \sim 6$ galaxies sample.

In our case, we assumed that the galaxies sample at lower redshift is not contaminated ($f_{z \sim 1.3}^{\text{cont}} = 0$, $f_{z \sim 1.3}^{\text{true}} = 1$) and the true cross-correlation function should be zero ($\omega_{\text{cross}}^{\text{true}}(\theta) = 0$). Therefore, the first, third, and fourth terms in equation (13) vanish, and the expression can be simplified into:

$$\omega_{\text{cross,obs}}(\theta) = f_{\text{cont}} \omega_{z \sim 1.3}(\theta), \quad (14)$$

where $\omega_{\text{cross,obs}}(\theta)$, f_{cont} , and $\omega_{z \sim 1.3}(\theta)$ are the observed cross-correlation function, the contamination fraction in high-redshift galaxies sample, and the ACF of intermediate-redshift interlopers, respectively.

We estimate the ACF of our $z \sim 1.3$ sample following the power-law form: $\omega_{z \sim 1.3}(\theta) = A_{\omega} \theta^{-\beta}$. The best-fitting parameters are $A_{\omega} = 0.66 \pm 0.18$ and $\beta = 0.45 \pm 0.19$ for the GOODS-S field and $A_{\omega} = 0.48_{-0.48}^{+0.59}$ and $\beta = 0.81 \pm 0.79$ for the XDF field. We conduct the χ^2 minimization method to find the best contamination fraction (f_{cont}) based on the observed cross-correlation function calculated in Section 2.2. Our results for the contamination fraction are $f_{\text{cont,GOODS-S}} = 0.00_{-0.00}^{+2.87}$ per cent for the GOODS-S field and $f_{\text{cont,XDF}} = 0.00_{-0.00}^{+6.76}$ per cent for the XDF field. These results agree with the estimations from our Monte Carlo simulation in Section 3.2, i.e. the contamination level should be less than 5.5 per cent if we do not detect any cross-correlation signal.

4 NUMBER COUNT ANALYSIS

The spatial cross-correlation analysis is most appropriate for surveys with a large contiguous field of view. Still, it may not apply to surveys with several pencil beam observations, such as random-pointing multiple field surveys. In this section, rather than using the spatial cross-correlation analysis presented above, we explore an alternative method to constrain the contamination by using a simple correlation between the number of the targeted population and the number of possible interloper populations. This method is based on the theoretical paper by Robertson (2010) who measures clustering of high-redshift galaxies based on counts-in-cell analysis. The idea is that there should be no correlation in the number counts across the observed fields unless contamination exists. Based on available data in the literature, we test this principle on the BoRG data set, setting $z \sim 8$ galaxies as our high-redshift galaxy sample. The corresponding interloper redshift is $z \sim 2$. We describe the data set in Section 4.1. The correlation analysis is in Section 4.2. We

then discuss the simulation in Section 4.3 and interpret the results in Section 4.4.

4.1 Data set

We use samples of $z \sim 2$ and $z \sim 8$ galaxies from the BoRG survey (Trenti et al. 2011). The BoRG survey is a pure-parallel program on the *Hubble Space Telescope* focused on finding bright galaxy candidates at redshift $z \gtrsim 7$ using the Lyman break technique (Steidel et al. 1996). Specifically, we take $z \sim 8$ galaxies from the catalogues provided by Bradley et al. (2012) and Schmidt et al. (2014) and $z \sim 2$ galaxies from the catalogues Cameron et al. (2019, hereafter C+19).

Both catalogues are based on the first generation of the BoRG data release, which consists of 71 random independent pointings taken with three different near-infrared filters (*WFC3/IR F098M*, *F125W*, and *F160W*) and one optical filter (either *WFC3/UVIS F606W* or *F600LP*). C+19 discards two fields because they affected by star overdensity and significant Galactic dust-reddening. Hence, the number of overlapping search fields between the two catalogues is 69 fields. Due to the nature of the pure-parallel survey, each field has a different exposure time, which leads to 5σ limiting magnitude in *F125W* ranging between 25.6 and 27.5.

The $z \sim 2$ galaxies in C+19 were detected in *H*₁₆₀ band and selected using $Y_{098} - H_{160} > 1.5$ cut. Photometric redshift estimates were obtained with the Bayesian photometric redshift code BPZ (Benítez 2000; Benítez et al. 2004; Coe et al. 2006). The redshift range of C+19 final sample is $1.5 < z < 2.5$. The catalogue consists of 490 galaxies and is expected to be highly complete and not contaminated up to $m_{\text{AB,H}} = 24.5$. On the other hand, the $z \sim 8$ galaxies in Bradley et al. (2012) and Schmidt et al. (2014) were detected in *J*₁₂₅ band and were selected using Lyman break technique that includes objects in the redshift range $7.5 \lesssim z \lesssim 8.5$. The catalogue consists of 42 $z \sim 8$ galaxies with *F125W* magnitude ranging from 25.50 to 27.60.

4.2 Correlation between number counts of galaxies at two redshifts in BoRG data

Due to the different depths for each field in our sample, we may introduce an artificial correlation. Deeper fields may have a higher number count of $z \sim 2$ and $z \sim 8$ galaxies than shallow fields. The median depth among all the fields is 26.75 mag in *F125W* band. We therefore remove all fields with limiting magnitude fainter than 26.75 mag and all galaxies with *F125W* fainter than 26.75 to get a sample of $z \sim 8$ sources that have the same completeness across magnitude bins up to *F125W* = 26.75 mag. The completeness of these $z \sim 8$ galaxies is approximately 60 per cent (Trenti et al. 2012). Our final sample consists of 39 fields with 306 $z \sim 2$ galaxies and 14 $z \sim 8$ galaxies. The number counts between the $z \sim 8$ and $z \sim 2$ galaxies in each field are plotted as filled red circles in Fig. 6.

To quantify the correlation between the number of $z \sim 2$ and $z \sim 8$ galaxies, we measure the Pearson correlation coefficient. We use Fisher's transformation to estimate the confidence interval of Pearson correlation coefficient (a value of 1 indicates a perfect correlation, while a value of 0 indicates no correlation). Our Pearson correlation coefficient is equal to 0.05 ± 0.17 . Although it is positive at face value, it is consistent with zero within 1σ . Regardless, we proceed to measure the corresponding contamination fraction using a Monte Carlo simulation in the following section.

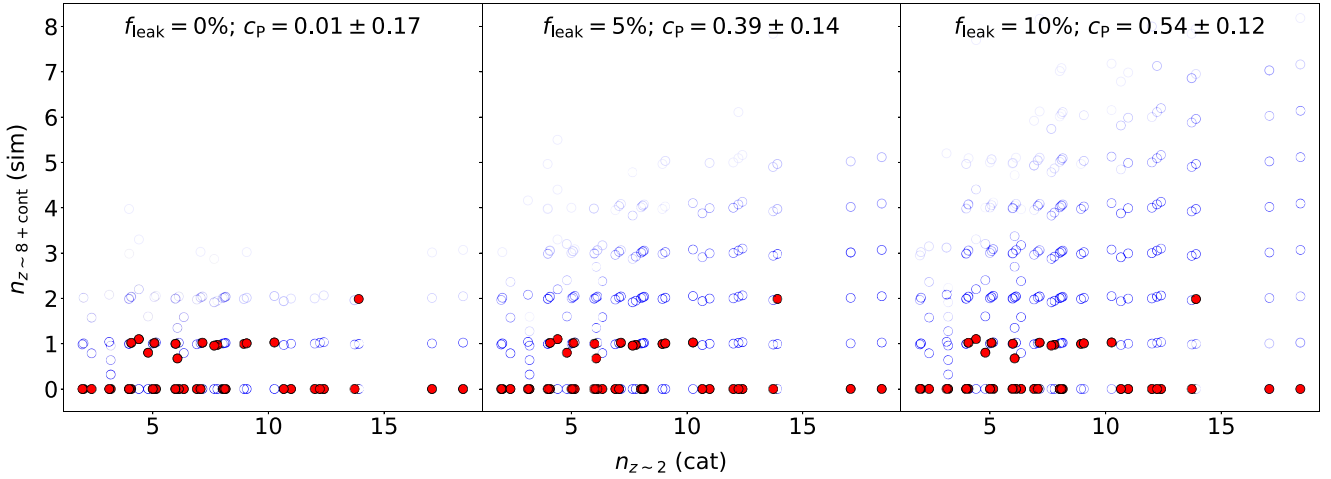


Figure 6. Comparing the number of galaxies observed at intermediate redshift ($z \sim 2$) to the number of galaxies at high redshift ($z \sim 8$) for all 39 fields of the BoRG survey studied in this paper. The filled red circles in every panel show the values taken from the $z \sim 2$ catalogue of C+19 and the $z \sim 8$ catalogues of Bradley et al. (2012) and Schmidt et al. (2014). The empty blue circles are generated from our Monte Carlo simulation, using three different values of the leakage fraction to compute the number count of $z \sim 8$ galaxies plus contaminants. The transparency of empty blue circles represents the frequency of Monte Carlo draws (more transparent means less occurrence). The Pearson correlation coefficient c_p is also shown for each case of leakage fraction. All the presented number counts are rescaled to a median field area of 4.40 arcmin^2 .

4.3 Simulations

To estimate the contamination fraction in the BoRG sample, we perform a Monte Carlo simulation following the methodology outlined in Section 3.2. We first calculate the expected number of faint $z \sim 2$ galaxies and the expected number of $z \sim 8$ galaxies specific to each observed field. Because the sample of $z \sim 2$ galaxies is pure and complete up to $F125W = 24.5$, the interlopers are likely to come from the population with apparent magnitudes fainter than 24.5. We define all $z \sim 2$ galaxies with a magnitude between 24.5 and 26.75 in the $F125W$ band as faint $z \sim 2$ galaxies. These galaxies are not included in the C+19 catalogue.

We estimate the expected number count of faint $z \sim 2$ galaxies in each field of the BoRG survey using the luminosity function of $z \sim 2$ galaxies from Marchesini et al. (2012). Based on this luminosity function, we calculate the ratio of the number of faint galaxies with $24.25 < F125W < 26.75$ to the number of bright galaxies with $F125W \leq 24.5$. We then normalize (multiply) the ratio with the observed number count of bright $z \sim 2$ galaxies from the catalogue to estimate the expected number count of faint $z \sim 2$ galaxies in the field. We also calculate the expected number count of $z \sim 8$ galaxies (i.e. the number count of real $z \sim 8$ galaxies) based on the UV luminosity function with the Schechter parameters from Schmidt et al. (2014) and the assumption that the detection is 60 per cent complete. We present these expected number counts for each field of the BoRG survey that we study in Table 1.

Finally, we perform a Monte Carlo simulation to study how the correlation between the number of $z \sim 2$ and $z \sim 8$ galaxies correlation changes with the leakage fraction. We draw a random number count of faint $z \sim 2$ galaxies and real $z \sim 8$ galaxies following the Poisson distribution $f(k; \lambda) = \lambda^k e^{-\lambda} / k!$, where λ is their expected numbers from the luminosity functions. Then, the number of contaminants is estimated based on the simulated number of faint $z \sim 2$ galaxies and the value of leakage fraction. The number of observed $z \sim 8$ galaxies is the number of true $z \sim 8$ galaxies plus the number of contaminants. As done with the real data, we calculate the Pearson correlation coefficient c_p between the observed number of bright $z \sim 2$ galaxies (which is the same as that of the

real data) and the observed number of $z \sim 8$ galaxies. We conduct the procedure for $f_{\text{leak}} = 0$ per cent – 10 per cent with a step size of 0.2 per cent and repeat the simulation 100 times at each value of the leakage fraction.

4.4 Results and discussion

We present how the Monte Carlo simulation works in Fig. 6. Empty blue circles show the simulated number counts from all 100 simulations for three values of leakage fraction. As the leakage fraction is increased, the simulated number count of $z \sim 8$ galaxies becomes more correlated with the observed number count of $z \sim 2$ galaxies. This is due to more $z \sim 2$ galaxies being misidentified as $z \sim 8$ galaxies. From Fig. 6, we see that the simulation most closely resembles the data (filled red circles) when the leakage fraction is close to zero. The top panel in Fig. 7 shows the Pearson correlation coefficient derived from those simulations. As the leakage fraction increases, the number of contaminants increases. Consequently, the correlation between the number of $z \sim 8$ samples and the number of $z \sim 2$ samples becomes tighter (as indicated by the increasing value of c_p). To facilitate the interpretation, we present contamination fraction (f_{cont} , a percentage of observed $z \sim 8$ galaxies that are low- z interlopers) as a function of f_{leak} in the bottom panel of Fig. 7. The contamination increases rapidly as a function for leakage fraction and plateaus at contamination fraction of ~ 80 per cent. A mere leakage fraction of 1 per cent already corresponds to contamination fraction of 20 – 50 per cent.

To calculate the best-fitting f_{leak} and contamination of the BoRG sample, we compute the weighted average of the leakage fraction from the simulation based on the correlation coefficient of the observations. For each data point in the upper row of Fig. 7 calculated from the simulation, we measure the weight by assuming a Gaussian distribution:

$$g(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right), \quad (15)$$

where x is the c_p value for each data point. In this equation, μ and σ are the Pearson correlation coefficient and its uncertainty

Table 1. The number (n^{cat}) of $z \sim 2$ and $z \sim 8$ galaxies within each of the 39 fields from the BoRG catalogues considered in this analysis. We also tabulate the expected number count of faint $z \sim 2$ galaxies, and the intrinsic number count of $z \sim 8$ galaxies within each field predicted from galaxy luminosity functions (n^{LF}).

Field name	Area	$n^{\text{cat}}_{z \sim 2}$	$n^{\text{LF}}_{\text{faint}, z \sim 2}$	$n^{\text{cat}}_{z \sim 8}$	$n^{\text{LF}}_{\text{int}, z \sim 8}$
0110–0224	13.81	10	20	0	0.83
0228–4102	4.43	4	8	0	0.27
0436–5259	4.33	4	8	0	0.26
0439–5317	4.28	5	10	0	0.26
0440–5244	4.34	5	10	1	0.26
0553–6405	4.00	4	8	1	0.24
0751+2917	4.52	8	16	1	0.27
0846+7654	4.41	11	22	0	0.26
0906+0255	4.39	8	16	0	0.26
0914+2822	4.40	12	24	0	0.26
0952+5304	4.42	4	8	0	0.26
1010+3001	4.54	11	22	0	0.27
1031+5052	5.55	8	16	0	0.33
1033+5051	5.50	6	12	1	0.33
1051+3359	4.26	12	24	0	0.26
1059+0519	4.43	9	18	1	0.27
1103–2330	4.37	9	18	1	0.26
1111+5545	4.31	6	12	0	0.26
1118–1858	4.23	3	6	0	0.25
1119+4026	4.46	7	14	0	0.27
1131+3114	4.41	6	12	1	0.26
1152+5441	4.40	5	10	0	0.26
1209+4543	4.42	6	12	0	0.26
1242+5716	4.29	10	20	1	0.26
1341+4123	4.36	7	14	0	0.26
1358+4326	4.49	14	28	0	0.27
1358+4334	4.32	8	16	0	0.26
1408+5503	4.32	4	8	1	0.26
1416+1638	4.38	17	34	0	0.26
1429–0331	4.35	8	16	0	0.26
1437+5043	6.53	9	18	1	0.39
1459+7146	4.32	12	24	0	0.26
1510+1115	4.43	14	28	2	0.27
1555+1108	4.31	7	14	1	0.26
1632+3733	4.37	2	4	0	0.26
2203+1851	4.60	8	16	1	0.28
2313–2243	5.59	3	6	0	0.33
2345+0054	4.48	2	4	0	0.27
2351–4332	4.30	18	37	0	0.26

estimated from the observation in Section 4.2, i.e. $\mu = 0.05$ and $\sigma = 0.17$. Putting it into equation (15), we calculate the weight of each data point generated from simulation. Then, we measure the weighted mean as our leakage fraction estimation. We calculate $f_{\text{leak}} = 2.90 \pm 2.38$ per cent, corresponds to $f_{\text{cont}} = 62^{+13}_{-39}$ per cent (solid vertical red line in Fig. 7). The uncertainty in our result is high because of the small size of our sample. Our result is consistent with the previous estimate by Bradley et al. 2012 ($f_{\text{cont}} = 42$ per cent) and the follow-up observation of Livermore et al. 2018 ($f_{\text{cont}} \sim 50$ per cent) within the 1σ confidence interval. Bradley et al. (2012) estimate the contamination fraction by degrading a *F606W* image of GOODS-ERS data that is deeper than BoRG data. They then conduct the selection again on the degraded sample and compare it to the original GOODS-ERS catalogue. The contamination fraction can then be calculated by checking which low-redshift galaxy in the original catalogue leaks into the catalogue generated from the degraded images.

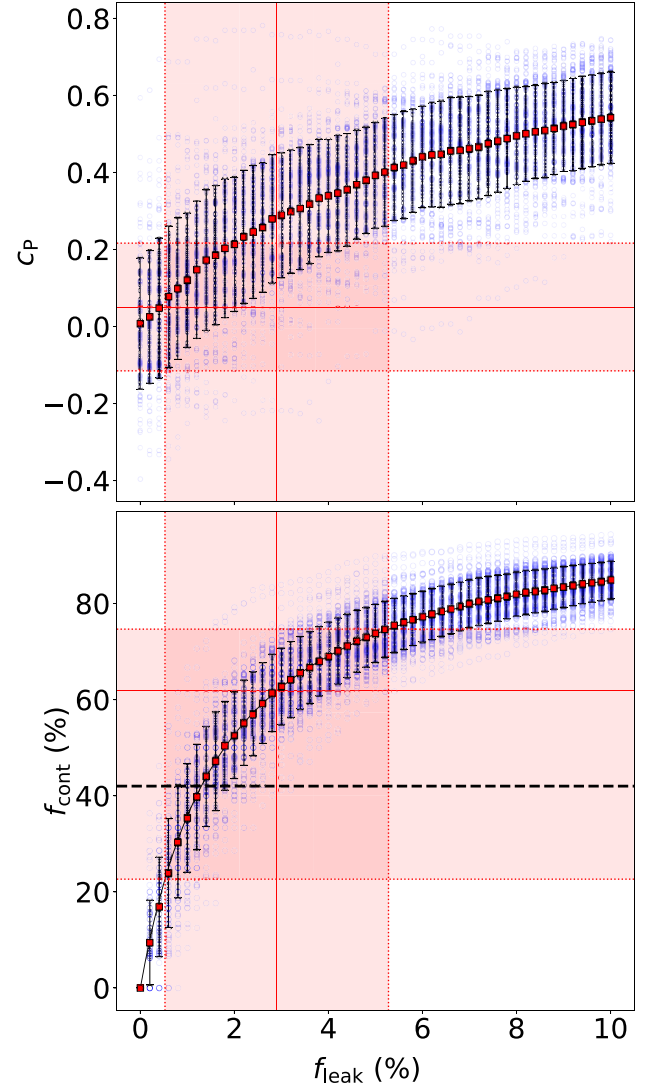


Figure 7. *Top panel:* Pearson correlation coefficient as a function of leakage fraction. The empty blue circles represent each data point generated from the Monte Carlo simulation. The filled red square is the mean and its standard deviation. The solid horizontal red line and its shaded region are Pearson correlation coefficient based on catalogue and its 1σ error, respectively. The solid vertical red line and its shaded region are the estimated leakage fraction and its 1σ error, respectively. *Bottom panel:* Contamination fraction as a function of leakage fraction. The solid blue circles represent each data point generated from the Monte Carlo simulation. The filled red square is the mean and its standard deviation. The solid horizontal red line and its shaded region are our contamination fraction estimates and the 1σ error, respectively. The dashed horizontal black line is the previous estimate from Bradley et al. (2012).

For comparison, we repeat the simulation assuming different limiting magnitudes at which the detection and the redshift determination for the low- z population are complete, specifically at $m_{\text{lim}} = 26.00$ and 28.00 (left and right columns of Fig. 8). Unsurprisingly, our results indicate that deeper fields have less contamination. The faint-end of $z \sim 8$ luminosity function is steeper than $z \sim 2$ luminosity function. Therefore, it is expected that for a deeper observation, we will get more $z \sim 8$ galaxies than $z \sim 2$ galaxies. As the number count of contaminants follows the luminosity function of $z \sim 2$ galaxies, f_{cont} will be lower for the same value of f_{leak} .

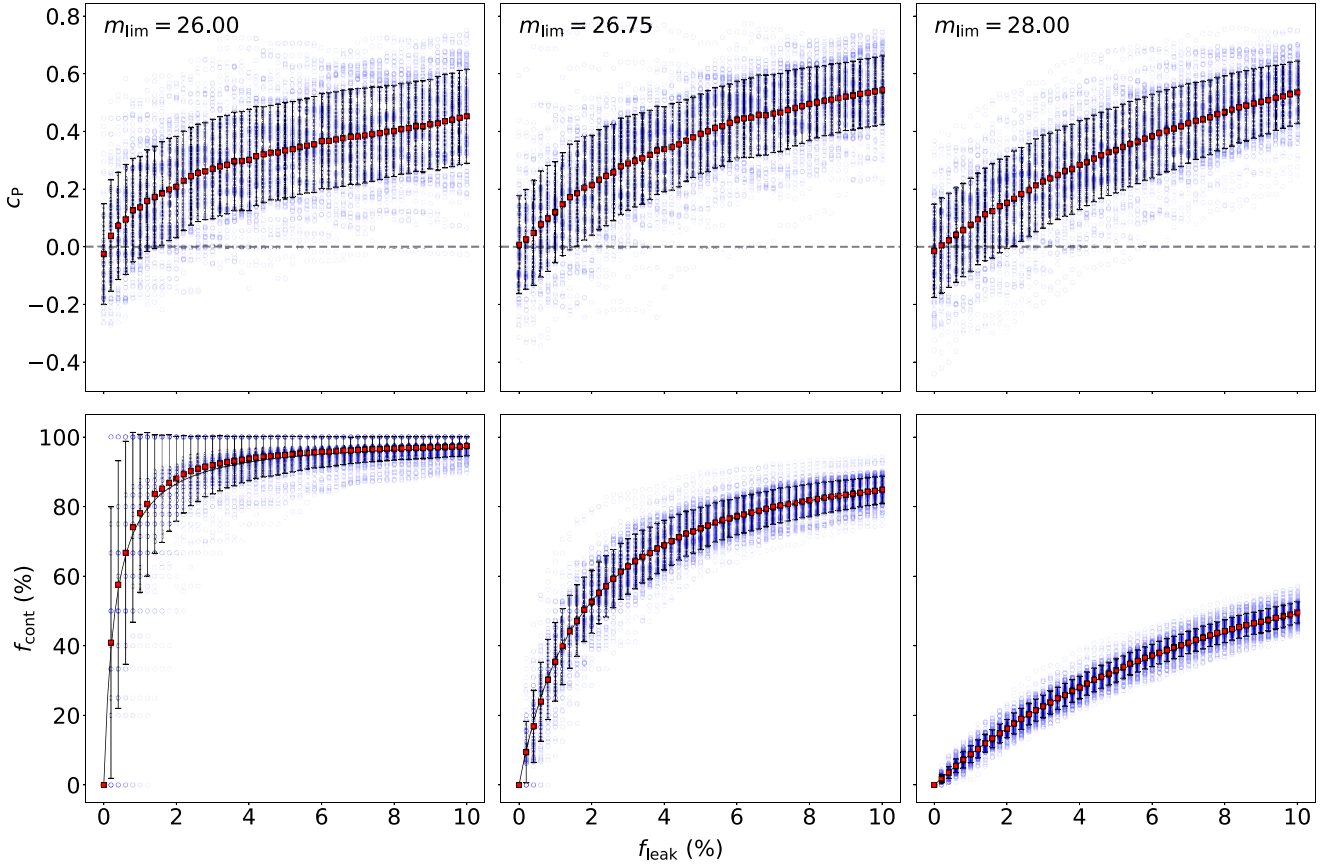


Figure 8. Same as Fig. 7, but applied to two different limiting magnitude cases. *Top panel:* Pearson correlation coefficient c_p as a function of leakage fraction f_{leak} for three different limiting magnitudes of 26.00, 26.75, and 28.00 (left to right, respectively). The empty blue circles represent each data point generated from the Monte Carlo simulation. The filled red square is the mean and its 1σ error. *Bottom panel:* Same as top panel, but the y-axis is the contamination fraction f_{cont} . The solid black curve is the f_{cont} as a function of f_{leak} derived based on the luminosity function.

5 SUMMARY

We presented a novel analysis of contamination in LBG samples at high-redshift by studying the spatial correlation with the intermediate-redshift galaxies. We considered two methods based on the nature of high-redshift surveys: a large-contiguous-field survey and a multiple-field surveys. As a demonstration of the two approaches, we investigated applications to the CANDELS GOODS-S and XDF survey, and to the BoRG random-pointing multiple field survey, respectively. We summarize our results as follows:

- (i) We carried out cross-correlation analysis based on the CANDELS data and performed statistical tests to quantify the contamination level in GOODS-S and XDF fields. Both fields show no significant cross-correlation signal between $z \sim 6$ galaxies and lower redshift galaxies at the redshift of potential contaminants (i.e. $z \sim 1.3$).
- (ii) Using the mock catalogue generated based on IllustrisTNG simulation, we modelled the changes in the cross-correlation function as a function of the contamination fraction. As we increased the contamination, the cross-correlation signal becomes stronger. We estimated that for GOODS-S field, the contamination fraction is below 5.5 per cent at 90 per cent confidence level.
- (iii) Our analysis shows that for a deeper field, the contamination is lower than those with a shallow field. This can be explained based on the luminosity function. The luminosity function for high-

redshift galaxies is steeper towards the faint-end compared to those of intermediate-redshift galaxies. Thus, the number of contaminants increases more slowly than the number of true high-redshift galaxies as the survey depth is increased.

- (iv) We applied a count-in-cell correlation analysis to a survey with a large number of independent lines of sight, using the relatively shallow BoRG data set. We detected evidence of number counts correlation, with a quantitative analysis estimating the contamination fraction for the BoRG $z \sim 8$ sample to be 62^{+13}_{-39} per cent, consistent with the previous calculation by Bradley et al. (2012) within 1σ confidence interval. The large error bar in our estimates is caused by the low average number of counts in each field, which gives rise to large Poisson fluctuations.

Overall, we demonstrated the utility of our novel analysis as an independent check of contamination in LBG samples. We can apply our method to larger data sets expected to become available from upcoming *JWST* observations. For example the number count analysis can be applied in the upcoming PANORAMIC Survey (A Pure Parallel Wide Area Legacy Imaging Survey at 1 – 5 Micron; Williams et al. 2021) and GO 3990: A NIRC2 Pure-Parallel Imaging Survey of Galaxies Across the Universe (Morishita et al. 2023). We expect those surveys will have larger sample sizes than the BoRG survey at $z \sim 8$, enabling higher precision measurements of the contamination fraction from count-in-cell correlation.

ACKNOWLEDGEMENTS

We thank Daniel Joseph Farrow as the referee for useful suggestions and comments that have improved the manuscript. This research was supported by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013. MH acknowledges support from a Melbourne Research Scholarship. BM acknowledges support from an Australian Government Research Training Program (RTP) Scholarship. The authors thank Charlotte A. Mason for useful comments on the manuscript. MH also thanks Yuichi Harikane for advice during the preparation of the manuscript.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

- Addison G. E., Bennett C. L., Jeong D., Komatsu E., Weiland J. L., 2019, *ApJ*, 879, 15
- Akaike H., 1974, *IEEE Trans. Autom. Control*, 19, 716
- Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, *MNRAS*, 310, 540
- Atek H. et al., 2011, *ApJ*, 743, 121
- Awan H., Gawiser E., 2020, *ApJ*, 890, 78
- Barone-Nugent R. L. et al., 2014, *ApJ*, 793, 17
- Benítez N., 2000, *ApJ*, 536, 571
- Benítez N. et al., 2004, *ApJS*, 150, 1
- Bertin E., Arnouts S., 1996, *A&AS*, 117, 393
- Blake C., Pope A., Scott D., Mobasher B., 2006, *MNRAS*, 368, 732
- Bouwens R. J. et al., 2015, *ApJ*, 803, 34
- Bouwens R. J. et al., 2021, *AJ*, 162, 47
- Bowler R. A. A., Jarvis M. J., Dunlop J. S., McLure R. J., McLeod D. J., Adams N. J., Milvang-Jensen B., McCracken H. J., 2020, *MNRAS*, 493, 2059
- Bradley L. D. et al., 2012, *ApJ*, 760, 108
- Cameron A. J., Trenti M., Livermore R. C., van der Velden C., 2019, *MNRAS*, 483, 1922 (C+19)
- Coe D., Benítez N., Sánchez S. F., Jee M., Bouwens R., Ford H., 2006, *AJ*, 132, 926
- Dalmasso N., Trenti M., Leethochawalit N., 2024, *MNRAS*, 528, 898
- Farrow D. J. et al., 2021, *MNRAS*, 507, 3187
- Finkelstein S. L. et al., 2015, *ApJ*, 810, 71
- Furlanetto S. R., Mirocha J., 2023, *MNRAS*, 523, 5274
- Genel S. et al., 2014, *MNRAS*, 445, 175
- Giavalisco M., 2002, *ARA&A*, 40, 579
- Grasshorn Gebhardt H. S. et al., 2019, *ApJ*, 876, 32
- Harikane Y. et al., 2016, *ApJ*, 821, 123
- Hoerl A. E., Kennard R. W., 1970, *Technometrics*, 12, 69
- Illingworth G. et al., 2016, preprint (arXiv:1606.00841)
- Ishigaki M., Kawamata R., Ouchi M., Oguri M., Shimasaku K., Ono Y., 2015, *ApJ*, 799, 12
- Landy S. D., Szalay A. S., 1993, *ApJ*, 412, 64
- Lee K.-S., Giavalisco M., Gnedin O. Y., Somerville R. S., Ferguson H. C., Dickinson M., Ouchi M., 2006, *ApJ*, 642, 63
- Leethochawalit N., Trenti M., Morishita T., Roberts-Borsani G., Treu T., 2022, *MNRAS*, 509, 5836
- Ling E. N., Frenk C. S., Barrow J. D., 1986, *MNRAS*, 223, 21
- Livermore R. C., Trenti M., Bradley L. D., Bernard S. R., Holwerda B. W., Mason C. A., Treu T., 2018, *ApJ*, 861, L17
- Lotz J. M. et al., 2017, *ApJ*, 837, 97
- Madau P., 1995, *ApJ*, 441, 18
- Marchesini D., Stefanon M., Brammer G. B., Whitaker K. E., 2012, *ApJ*, 748, 126
- Marinacci F. et al., 2018, *MNRAS*, 480, 5113
- Mason C. A. et al., 2015, *ApJ*, 805, 79
- Matthews D. J., Newman J. A., 2012, *ApJ*, 745, 180
- Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, preprint (arXiv:1303.4722)
- Merlin E. et al., 2021, *A&A*, 649, A22
- Morishita T. et al., 2018, *ApJ*, 867, 150
- Morishita T. et al., 2023, A NIRCам Pure-Parallel Imaging Survey of Galaxies Across the Universe, JWST Proposal. Cycle 2, ID. #3990
- Moutard T. et al., 2016, *A&A*, 590, A103
- Naidu R. P. et al., 2022, preprint (arXiv:2208.02794)
- Naiman J. P. et al., 2018, *MNRAS*, 477, 1206
- Nelson D. et al., 2015, *Astron. Comput.*, 13, 12
- Nelson D. et al., 2018, *MNRAS*, 475, 624
- Oke J. B., Gunn J. E., 1983, *ApJ*, 266, 713
- Overzier R. A., Bouwens R. J., Illingworth G. D., Franx M., 2006, *ApJ*, 648, L5
- Pillepich A. et al., 2018a, *MNRAS*, 473, 4077
- Pillepich A. et al., 2018b, *MNRAS*, 475, 648
- Rahman M., Ménard B., Scranton R., 2016a, *MNRAS*, 457, 3912
- Rahman M., Mendez A. J., Ménard B., Scranton R., Schmidt S. J., Morrison C. B., Budavári T., 2016b, *MNRAS*, 460, 163
- Roberts-Borsani G., Morishita T., Treu T., Leethochawalit N., Trenti M., 2022, *ApJ*, 927, 236
- Robertson B. E., 2010, *ApJ*, 716, L229
- Roche N., Eales S. A., 1999, *MNRAS*, 307, 703
- Rojas-Ruiz S., Finkelstein S. L., Bagley M. B., Stevans M., Finkelstein K. D., Larson R., Mechtley M., Diekmann J., 2020, *ApJ*, 891, 146
- Salmon B. et al., 2020, *ApJ*, 889, 189
- Schmidt K. B. et al., 2014, *ApJ*, 786, 57
- Sijacki D., Vogelsberger M., Genel S., Springel V., Torrey P., Snyder G. F., Nelson D., Hernquist L., 2015, *MNRAS*, 452, 575
- Springel V., 2010, *MNRAS*, 401, 791
- Springel V. et al., 2018, *MNRAS*, 475, 676
- Stanway E. R., Bremer M. N., Lehnert M. D., 2008, *MNRAS*, 385, 493
- Steidel C. C., Giavalisco M., Pettini M., Dickinson M., Adelberger K. L., 1996, *ApJ*, 462, L17
- Trenti M., Stiavelli M., 2008, *ApJ*, 676, 767
- Trenti M. et al., 2011, *ApJ*, 727, L39
- Trenti M. et al., 2012, *ApJ*, 746, 55
- van der Wel A. et al., 2011, *ApJ*, 742, 111
- Vanzella E. et al., 2008, *A&A*, 478, 83
- Vogelsberger M., Genel S., Sijacki D., Torrey P., Springel V., Hernquist L., 2013, *MNRAS*, 436, 3031
- Vogelsberger M. et al., 2014a, *MNRAS*, 444, 1518
- Vogelsberger M. et al., 2014b, *Nature*, 509, 177
- Vulcani B., Trenti M., Calvi V., Bouwens R., Oesch P., Stiavelli M., Franx M., 2017, *ApJ*, 836, 239
- Weinberger R. et al., 2014, *MNRAS*, 465, 3291
- Whitaker K. E. et al., 2019, *ApJS*, 244, 16
- Williams C. C. et al., 2021, PANORAMIC—A Pure Parallel Wide Area Legacy Imaging Survey at 1–5 Micron, JWST Proposal. Cycle 1, ID. #2514
- Wyithe J. S. B., Yan H., Windhorst R. A., Mao S., 2011, *Nature*, 469, 181
- Zheng Z., Coil A. L., Zehavi I., 2007, *ApJ*, 667, 760

This paper has been typeset from a $\text{\LaTeX}$ file prepared by the author.