Extending empirical constraints on the SZ-mass scaling relation to higher redshifts via HST weak lensing measurements of nine clusters from the SPT-SZ survey at z >= 1

Zohren, H.; Schrabback, T.; Bocquet, S.; Sommer, M.; Raihan, F.; Hernández-Martín, B.; ... ; Wright, A.H.

A&A 668, A18 (2022)
https://doi.org/10.1051/0004-6361/202142991

Astronomy & Astrophysics

Extending empirical constraints on the SZ–mass scaling relation to higher redshifts via HST weak lensing measurements of nine clusters from the SPT-SZ survey at $z \gtrsim 1$

Hannah Zohren[1], Tim Schrabback[1], Sebastian Bocquet[2,3], Martin Sommer[1], Fatimah Raihan[1], Beatriz Hernández-Martín[1], Ole Marggraf[1], Behzad Ansarinejad[4], Matthew B. Bayliss[5], Lindsey E. Bleem[6,7], Thomas Erben[1], Henk Hoekstra[8], Benjamin Floyd[9], Michael D. Gladders[7,10], Florian Kleinebreil[1], Michael A. McDonald[11], Mischa Schirmer[12], Diana Scognamiglio[1], Keren Sharon[13], and Angus H. Wright[14]

[1] Argelander Institut für Astronomie, Rheinische Friedrich-Wilhelms-Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
e-mail: hzohren@astro.uni-bonn.de
[2] Faculty of Physics, Ludwig-Maximilians University, Scheinerstr. 1, 81679 München, Germany
[3] Excellence Cluster ORIGINS, Boltzmannstr. 2, 85748 Garching, Germany
[4] School of Physics, University of Melbourne, Parkville, VIC 3010, Australia
[5] Department of Physics, University of Cincinnati, Cincinnati, OH 45221, USA
[6] High Energy Physics Division, Argonne National Laboratory, Argonne, IL 60439, USA
[7] Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA
[8] Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands
[9] Department of Physics and Astronomy, University of Missouri–Kansas City, 5110 Rockhill Road, Kansas City, MO 64110, USA
[10] Department of Astronomy and Astrophysics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA
[11] Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[12] Max-Planck-Institut für Astronomie (MPIA), Königstuhl 17, 69117 Heidelberg, Germany
[13] Department of Astronomy, University of Michigan, 1085 S. University Ave, Ann Arbor, MI 48109, USA
[14] Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing, 44780 Bochum, Germany

Received 23 December 2021 / Accepted 13 August 2022

ABSTRACT

We present a *Hubble* Space Telescope (HST) weak gravitational lensing study of nine distant and massive galaxy clusters with redshifts $1.0 \lesssim z \lesssim 1.7$ ($z_{\rm median} = 1.4$) and Sunyaev Zel'dovich (SZ) detection significance $\xi > 6.0$ from the South Pole Telescope Sunyaev Zel'dovich (SPT-SZ) survey. We measured weak lensing galaxy shapes in HST/ACS $F606W$ and $F814W$ images and used additional observations from HST/WFC3 in $F110W$ and VLT/FORS2 in $U_{\rm HIGH}$ to preferentially select background galaxies at $z \gtrsim 1.8$, achieving a high purity. We combined recent redshift estimates from the CANDELS/3D-HST and HUDF fields to infer an improved estimate of the source redshift distribution. We measured weak lensing masses by fitting the tangential reduced shear profiles with spherical Navarro-Frenk-White (NFW) models. We obtained the largest lensing mass in our sample for the cluster SPT-CL J2040−4451, thereby confirming earlier results that suggest a high lensing mass of this cluster compared to X-ray and SZ mass measurements. Combining our weak lensing mass constraints with results obtained by previous studies for lower redshift clusters, we extended the calibration of the scaling relation between the unbiased SZ detection significance ζ and the cluster mass for the SPT-SZ survey out to higher redshifts. We found that the mass scale inferred from our highest redshift bin ($1.2 < z < 1.7$) is consistent with an extrapolation of constraints derived from lower redshifts, albeit with large statistical uncertainties. Thus, our results show a similar tendency as found in previous studies, where the cluster mass scale derived from the weak lensing data is lower than the mass scale expected in a *Planck* $\nu\Lambda$CDM (i.e. $\nu\Lambda$ cold dark matter) cosmology given the SPT-SZ cluster number counts.

Key words. gravitational lensing: weak – cosmology: observations – galaxies: clusters: general

1. Introduction

Galaxy clusters trace the densest regions of the large-scale structure in the Universe. Studying their number density as a function of mass and redshift, therefore, provides insights into the cosmic expansion and structure formation histories, allowing for constraints of cosmological models (e.g. Haiman et al. 2001; Allen et al. 2011). The expected number of dark matter haloes at a given mass and redshift is predicted by the halo mass function (HMF), which can be obtained from numerical simulations (e.g. Tinker et al. 2008; McClintock et al. 2019; Bocquet et al. 2020). A comparison of these predictions to observations of galaxy clusters as representatives of these haloes and their abundance serves as a probe, which is particularly sensitive to a combination of the cosmological parameters $\Omega_{\rm m}$, the matter energy density of the Universe, and σ_8, the standard deviation of fluctuations in the linear matter density field at scales of $8\,{\rm Mpc}\,h^{-1}$. At the same time, cluster studies can constrain the dark energy equation of state parameter w.

Such studies require samples of galaxy clusters with a well-defined selection function and covering a large redshift range. Common methods for the assembly of such samples include

detection via the overdensity of galaxies in the optical/near-infrared (NIR) regime (e.g. Rykoff et al. 2016), via the X-ray flux (e.g. Piffaretti et al. 2011; Pacaud et al. 2018; Liu et al. 2022), or via the signal from the Sunyaev Zel'dovich (SZ) effect (e.g. Bleem et al. 2015; Planck Collaboration XXIV 2016; Hilton et al. 2021).

The thermal SZ effect (Sunyaev & Zeldovich 1972) describes a distortion of the cosmic microwave background (CMB) blackbody spectrum towards higher energy, caused when CMB photons experience an inverse Compton scattering with the energetic electrons in the intracluster medium. Since the signal is independent of redshift, detecting clusters through the SZ effect enables the assembly of cluster catalogues, which are nearly mass-limited and extend out to very high redshifts. Additionally, the uncertainties in the selection function are relatively low because the SZ-observable provides a mass proxy with a comparably low intrinsic scatter (∼20%, e.g. Angulo et al. 2012). These are promising prerequisites for cosmological studies through the comparison of the observed cluster mass function and the predicted HMF.

However, accurate and precise calibration of the scaling relations between the observable mass proxy and the underlying unobservable halo mass as predicted by the HMF over a wide redshift range is needed to obtain meaningful cosmological constraints. Especially since the remaining uncertainties in the observable-mass scaling relations are the limiting factor hampering the progress to tighter constraints (e.g. Dietrich et al. 2019). It is, therefore, imperative to improve the cluster mass calibration out to the highest redshifts that are now accessible in cluster samples (Bocquet et al. 2019; Schrabback et al. 2018, 2021). Mass measurements from weak gravitational lensing are frequently used as a method to obtain an absolute calibration of the normalisation of these scaling relations (e.g. Okabe et al. 2010; Kettula et al. 2015; Dietrich et al. 2019; Herbonnet et al. 2020; Chiu et al. 2022; Schrabback et al. 2021). Weak gravitational lensing causes a systematic distortion of the shapes of background galaxies when their light travels through the gravitational field of a foreground mass distribution. The weak lensing reduced shear quantifies the tangential distortion with respect to the centre of the mass distribution. The differential projected cluster mass distribution can be inferred from measurements of the reduced shear, without the need for assumptions about the dynamical state of the clusters. This is especially advantageous for high-redshift clusters because these objects are still dynamically young and may not have settled into hydrostatic equilibrium yet. The assumption of hydrostatic equilibrium is an important ingredient for measurements of the cluster mass with X-ray observations.

A complementary method to weak lensing studies with optical/NIR data is CMB cluster lensing, which measures the (stacked) weak lensing signal by galaxy clusters in maps of the temperature and polarisation of the CMB (e.g. Raghunathan et al. 2019; Zubeldia & Challinor 2019; Madhavacheril et al. 2020). Due to the high redshift of the CMB, the mass scale for high-redshift clusters is more easily accessible with this method, and constraints will become increasingly competitive with upcoming instruments such as SPT-3G (Benson et al. 2014) and CMB-S4 (Abazajian et al. 2019).

In the low to intermediate redshift regime, wide-field ground-based surveys such as the Kilo Degree Survey (KiDS, Kuijken et al. 2015), the Dark Energy Survey (DES, The Dark Energy Survey Collaboration 2005) and the Hyper-Suprime-Cam Survey (HSC, Miyazaki et al. 2012) can calibrate cluster masses at the few percent level via weak lensing, but they are not suited to obtain the critically required cluster masses at high redshifts. Their limited depth and ground-based resolution are not sufficient to resolve the shapes of the small and faint background galaxies behind high-redshift clusters.

The aforementioned optical lensing studies have been limited to low to intermediate redshift regimes up to $z \sim 1$. It is important to extend the calibration of scaling relations to higher redshifts because cluster properties (e.g. thermodynamic properties such as density, temperature, pressure, and entropy) evolve over time. Upcoming surveys conducted with *Euclid* (Laureijs et al. 2011), the *Nancy Grace Roman* Space Telescope (formerly known as WFIRST, Spergel et al. 2015), and the *Vera C. Rubin* Observatory (LSST Science Collaboration 2009) will provide improved and critically required constraints on the cluster masses over a wide redshift range, where the exquisite depth of the *Nancy Grace Roman* Space Telescope will be particularly valuable for the very high-redshift regime.

However, until these surveys become available, pointed follow-up studies provide the best option to constrain the cluster mass scale out to high redshifts. With this work, we present the first weak lensing constraints on the mass scale of SZ-selected clusters extending to redshifts above $z \gtrsim 1.2$, using galaxy shape measurements from HST imaging. The median redshift of the sample with nine clusters studied here is $z = 1.4$. This study is an extension of the works by Schrabback et al. (2018, henceforth S18), Dietrich et al. (2019, henceforth D19), Bocquet et al. (2019, henceforth B19), and Schrabback et al. (2021, henceforth S21) to constrain the redshift evolution of the SZ mass scaling relation based on clusters from the 2500 deg^2 South Pole Telescope SZ survey (SPT-SZ survey, Bleem et al. 2015). With our high-redshift sample, we aim to tighten the constraints on the scaling relation parameter $C_{\rm SZ}$, describing its redshift evolution, which in particular helps to break the degeneracy of $C_{\rm SZ}$ with the dark energy equation of state parameter w.

The structure of this paper is as follows: we provide a summary of the studied cluster sample in Sect. 2. We then present the data reduction of our optical observations and describe the photometric calibration steps in Sect. 3. The selection of background galaxies based on four photometric bands and the estimation of the source redshift distribution from photometric redshift catalogues are detailed in Sect. 4. We describe the weak lensing shape measurements in Sect. 5. We present our weak lensing mass constraints including an estimation of the weak lensing mass bias in Sect. 6. We constrain the observable-mass scaling relation incorporating the new lensing results for our high-redshift SPT cluster sample in Sect. 7. Finally, we discuss our results in Sect. 8 and summarise and conclude in Sect. 9.

Unless indicated otherwise, we assume a standard flat Λ cold dark matter (ΛCDM) concordance cosmology throughout this paper with $\Omega_{\rm m} = 0.3$, $\Omega_\Lambda = 0.7$, and $H_0 = 70\,{\rm km\,s^{-1}\,Mpc^{-1}}$, as approximately consistent with CMB constraints (e.g. Planck Collaboration VI 2020). We express masses in terms of $M_{\Delta{\rm c}}$ corresponding to a sphere within which the density is Δ times higher than the critical density at the given redshift.

All reported magnitudes in this work are AB-magnitudes. We correct all magnitude measurements for Galactic extinction with the extinction maps by Schlafly & Finkbeiner (2011).

2. The high-z SPT cluster sample and previous studies

We investigate nine massive and distant galaxy clusters at redshifts $1.0 \lesssim z \lesssim 1.7$ detected by the SPT via their SZ signal.

Table 1. Properties of the galaxy cluster sample.

Cluster name	z_l	ξ	Coordinates centres (deg J2000)				$M_{500c,SZ}$
			SZ α	SZ δ	X-ray α	X-ray δ	$[10^{14}\,M_\odot/h_{70}]$
SPT-CL J0156−5541	1.288 [a]	6.98	29.04490	−55.69801	29.0405	−55.6976	$3.96^{+0.57}_{-0.65}$
SPT-CL J0205−5829	1.322 [b]	10.40	31.44282	−58.48521	31.4459	−58.4849	$5.06^{+0.55}_{-0.68}$
SPT-CL J0313−5334	1.474 [a]	6.09	48.48090	−53.57809	48.4813	−53.5718	$3.31^{+0.55}_{-0.61}$
SPT-CL J0459−4947	1.710 [d]	6.29	74.92693	−49.78724	74.9240	−49.7823	$3.08^{+0.53}_{-0.53}$
SPT-CL J0607−4448	1.401 [a]	6.44	91.89841	−44.80333	91.8940	−44.8050	$3.60^{+0.57}_{-0.63}$
SPT-CL J0640−5113	1.316 [a]	6.86	100.06452	−51.22045	100.0720	−51.2176	$3.89^{+0.58}_{-0.65}$
SPT-CL J0646−6236	0.995 [e]	8.67	101.63906	−62.61360	–	–	$4.97^{+0.64}_{-0.76}$ [f]
SPT-CL J2040−4451	1.478 [c]	6.72	310.24832	−44.86023	310.2417	−44.8620	$3.76^{+0.58}_{-0.63}$
SPT-CL J2341−5724	1.259 [a]	6.87	355.35683	−57.41580	355.3533	−57.4166	$3.58^{+0.51}_{-0.59}$

Notes. We list cluster names, SZ significance ξ, SZ coordinates of the centre and SZ masses as presented in B19. The X-ray coordinates correspond to the centroid positions estimated by McDonald et al. (2017). [a]Spectroscopic redshifts by Khullar et al. (2019). [b]Spectroscopic redshift from Stalder et al. (2013). [c]Spectroscopic redshift from Bayliss et al. (2014). [d]Best redshift constraint currently available (based on a spectral analysis of *XMM-Newton* data, using the 6.7 keV Fe emission line complex, Mantz et al. 2020). [e]Observation design and data reduction followed the same procedures as described in Khullar et al. (2019). More general results will be discussed in a future paper on high-z spectroscopic measurements of SPT clusters. [f]We list the SZ mass recalculated at the updated redshift of the cluster.

They were originally selected to have $z > 1.2$ according to the best redshift estimate available at the time. However, our analysis of more recent spectroscopic observations place the cluster SPT-CL J0646−6236 at lower redshift, $z = 0.995$ (see also note (e) in Table 1). Therefore, only the remaining eight clusters constitute the complete sample of galaxy clusters at high redshifts $z \geq 1.2$ with the strongest detection significance of $\xi \geq 6$ from the 2500 deg^2 SPT-SZ survey (Bleem et al. 2015, see Table 1 for cluster properties). The sample has a median redshift of $z_{\rm med} = 1.4$. Our study represents the first homogeneous weak lensing study of a cluster sample of this size with a clean SZ-based selection function at this high-redshift regime. B19 derive cosmological constraints with galaxy clusters from the 2500 deg^2 SPT-SZ survey and provide updated redshift and SZ mass estimates for the SPT cluster sample, including the clusters studied here (redshift updates for clusters relevant to this work are from Khullar et al. 2019; Mantz et al. 2020). The SZ mass estimates incorporate a weak lensing mass calibration using data from D19 and S18.

The nine clusters in this work are also part of several previous studies. McDonald et al. (2017) examine *Chandra* X-ray data for eight of these clusters and investigate the redshift dependency and compatibility with self-similar evolution of the ICM in a large sample of galaxy clusters. Their study includes an estimation of the positions of the cluster X-ray centres (see also Table 1) and the X-ray-based masses (derived from the $M_{\rm gas}$−M relation from Vikhlinin et al. 2009), as well as density profiles and morphologies of the clusters. Ghirardini et al. (2021) investigate thermodynamic properties, for example, density, temperature, pressure, and entropy with combined *Chandra* and *XMM-Newton* X-ray observations of seven clusters in our sample and compare them with the corresponding properties of low-redshift clusters. Additionally, Bulbul et al. (2019) include two of the clusters in their analysis of X-ray properties of SPT-selected galaxy clusters observed with *XMM-Newton*. They constrain the scaling relations between the X-ray observables of the ICM (luminosity $L_{\rm X}$, ICM mass $M_{\rm ICM}$, emission-weighted mean temperature $T_{\rm X}$, and integrated pressure $Y_{\rm X}$), redshift, and halo mass. Further X-ray studies investigating astrophysical properties featuring one or more clusters from our sample include McDonald et al. (2013), Sanders et al. (2018), and Mantz et al. (2020). There have also been efforts to obtain precise spectroscopic redshifts for the majority of clusters in our sample (Stalder et al. 2013; Bayliss et al. 2014; Khullar et al. 2019; Mantz et al. 2020), where some studies specifically investigate the galaxy kinematics and velocity distributions (Ruel et al. 2014; Capasso et al. 2019). Several multi-wavelength studies of cluster samples (including one or more clusters from our sample) with varying size investigate different cluster components such as the baryon content (Chiu et al. 2016, 2018), the properties, growth and star formation in brightest cluster galaxies (BCGs, McDonald et al. 2016; DeMaio et al. 2020; Chu et al. 2021), the mass-richness relation (Rettura et al. 2018), environmental quenching of the galaxy populations in clusters (Strazzullo et al. 2019), and AGN-feedback (Hlavacek-Larrondo et al. 2015; Bîrzan et al. 2017). The cluster SPT-CL J2040−4451 was already studied in a weak lensing analysis by Jee et al. (2017), using infrared imaging from the Wide Field Camera 3 (WFC3) on the HST for shape measurements. We compare their analysis strategy and ours in detail in Sect. 8.

3. Data and data reduction

3.1. HST *ACS and WFC3 data*

We used high-resolution imaging from the HST to measure galaxy shapes for the weak lensing analysis as detailed in Sect. 5. The observational data analysed in our study were obtained during Cycles 19, 21, 23, and 24 as part of the SPT follow-up GO programmes 12477 (PI: F. High), 13412 (PI: T. Schrabback), 14252 (PI: V. Strazzullo), and 14677 (PI: T. Schrabback) in the filters $F606W$ and $F814W$ with the ACS/WFC instrument and $F110W$ with the WFC3/IR instrument. We measured the shapes of galaxies for our weak lensing analysis in the $F606W$ and $F814W$ imaging, which have a field of view of $202'' \times 202''$ at a pixel scale of $0\farcs05$/pixel. The ACS (Advanced Camera for Surveys) observations were obtained in a single pointing except for SPT-CL J0205−5829 for which an additional larger 2×2 mosaic was obtained in $F606W$ as part of GO programme 12477. The field of view of WFC3 is $136'' \times 123''$ with a pixel scale of roughly $0\farcs128$/pixel (the pixels are not exactly square

Table 2. Summary of the integration times, image quality, and depth from our observations with HST/ACS, HST/WFC3, and VLT/FORS2.

Cluster name	$F606W$			$F814W$			$F110W$			$U_{\rm HIGH}$		
	$t_{\rm exp}$ [ks]	IQ [″]	Depth [mag]	$t_{\rm exp}$ [ks]	IQ [″]	Depth [mag]	$t_{\rm exp}$ [ks] [(b)]	IQ [″]	Depth [mag] [(b)]	$t_{\rm exp}$ [ks]	IQ [″]	Depth [mag]
SPT-CL J0156−5541	5.5	0.10	27.0	4.9	0.10	26.6	0.6	0.29	26.3	4.8	0.73	26.9
SPT-CL J0205−5829	3.7 [(a)]	0.10	27.1 [(a)]	3.7	0.08	26.5	0.6	0.29	26.3	4.8	0.85	26.8
SPT-CL J0313−5334	3.7	0.10	26.9	3.7	0.09	26.1	0.6	0.29	26.3	4.8	0.80	27.1
SPT-CL J0459−4947	2.3	0.11	26.7	4.8	0.10	26.5	0.6	0.28	26.3	6.0	0.81	26.9
SPT-CL J0607−4448	2.3	0.10	26.7	4.8	0.10	26.3	0.6	0.28	26.4	4.8	0.97	26.4
SPT-CL J0640−5113	5.6	0.10	26.7	3.3	0.10	26.2	0.6	0.26	26.1	2.4	0.97	26.3
SPT-CL J0646−6236	4.0	0.10	26.8	4.0	0.10	26.1	0.6	0.27	26.1	4.8	1.07	26.3
SPT-CL J2040−4451	2.1	0.10	26.6	4.8	0.10	26.1	0.6	0.28	26.1	4.8	0.88	26.5
SPT-CL J2341−5724	5.3	0.10	26.5	4.8	0.10	26.2	0.6	0.29	26.1	4.8	0.92	26.9
HUDF	–	–	–	–	–	–	–	–	–	6.6	1.03	26.6

Notes. For the image quality (IQ), we report the full width at half maximum of the PSF, based on measurements with `Source Extractor`. The depth corresponds to 5σ limiting magnitudes, computed from the standard deviation of 1000 non-overlapping apertures without flux from detected sources. We used apertures with diameters of 0.″7 for HST bands and 1.″2 for $U_{\rm HIGH}$. [(a)]For the cluster SPT-CL J0205−5829 a 2×2 ACS mosaic from GO programme 12477 and one single ACS pointing from GO programme 14677 are available in the $F606W$ band. We have eight overlapping exposures in the region with the biggest overlap with our observations in the other bands. We report the integration time and depth based on this region. [(b)]The $F110W$ stacks are mosaics of eight exposures. The highest/intermediate/lowest depth is achieved, where eight/four/two exposures overlap, respectively. Since regions with only two overlapping exposures make up the most area in the stacks, we report integration times and depths equivalent to two exposures.

shaped). We observed 2×2 mosaics in the $F110W$ filter (with partial overlap between pointings), which roughly match the field of view of the ACS observations. We used the observations in the $F110W$ filter exclusively for the photometric selection of the background galaxies carrying the weak lensing signal. The integration times range between 2.3 and 5.5 ks ($F606W$), 3.3 and 4.9 ks ($F814W$), and 2.4 ks ($F110W$, spread out over a 2×2 mosaic to reach a minimum depth of 0.6 ks over the full ACS footprint; see Table 2).

We performed the basic image reduction steps for the HST/ACS imaging data with the ACS calibration pipeline `CALACS`[1]. However, we deviated from the standard processing steps regarding the correction for charge transfer inefficiency (CTI). As in previous studies by S18 and S21, we performed the CTI correction with the algorithm by Massey et al. (2014) and applied it to both the HST/ACS imaging data and the respective dark frames. Furthermore, we performed a quadrant-based sky background subtraction, improved the bad pixel masks by manually flagging satellite trails and cosmic ray clusters, and computed accurately normalised rms maps following the prescription by Schrabback et al. (2010).

The HST/WFC3 imaging data reduction was performed similarly to the HST/ACS imaging data reduction. We downloaded the pre-reduced `flt` frames, which had already undergone basic image processing via the WFC3 calibration pipeline `calwf3`[2], but we did not perform a quadrant-based sky background subtraction because it is not suitable for the parallel read-out mechanism of WFC3. Instead, we used `Source Extractor` (version 2.23.1, Bertin & Arnouts 1996) to obtain a background model, which we subtracted. This allowed us to account properly for gradients in the background level. These occasionally occur in particular due to a variable airglow line of He I in the lower exosphere at 10 830 Å, which mostly affects the bands $F105W$ and $F110W$ (see Chapter 7.9.5 of the WFC3 instrument handbook[3] and WFC3 ISR 2014-03). We did not perform a CTI correction for the WFC3 data, as they are not affected by this.

Subsequently to the initial data reduction, we employed the software `DrizzlePac`[4] for aligning and combining HST images in particular using the tasks `TweakReg` and `AstroDrizzle`. This typically involved combining 4 to 10 exposures for HST/ACS imaging or 8 exposures in a 2×2 mosaic for WFC3 imaging. For the stacking with `AstroDrizzle`, we used the `lanczos3` kernel at the native pixel scale of 0.″05 (0.″128) of the ACS (WFC3) images to distribute the flux onto the output image. Additionally, we employed the rms image as the weighting image. We produced the stack for the imaging in the $F606W$ band first and subsequently used this stack as the astrometric reference image for the stacks in the $F814W$ and $F110W$ bands to ensure optimal astrometric alignment between the stacks.

3.2. Very Large Telescope (VLT) FORS2 data

We used additional observations from VLT/FORS2 in the $U_{\rm HIGH}$ passband obtained as part of the ESO programme 0100.A-0204(A) (PI: Schrabback) between November 18 and November 20, 2017. Together with the HST imaging, these observations facilitated a robust photometric selection of background galaxies. The images were taken with the two blue-sensitive $2k\times4k$ E2V CCDs in standard resolution with 2×2 binning, providing observations over a field of view of 6.′8 × 6.′8 at a pixel scale of 0.″25/pixel. We observed the nine galaxy clusters in our sample and additionally one pointing centred on the *Hubble* Ultra Deep Field (HUDF, Beckwith et al. 2006), which we used to assess the photometric calibration of the $U_{\rm HIGH}$ band. The integration times per cluster ranged between 2.4 ks and 6.6 ks (see Table 2).

We reduced the data with the software `THELI`[5] (Erben et al. 2005; Schirmer 2013). We performed a bias subtraction,

[1] https://hst-docs.stsci.edu/acsdhb, Chapter 3.
[2] https://hst-docs.stsci.edu/wfc3dhb, Chapter 3.
[3] https://hst-docs.stsci.edu/wfc3ihb
[4] https://www.stsci.edu/scientific-community/software/drizzlepac.html
[5] https://www.astro.uni-bonn.de/theli/

flat-field correction, and a subtraction of a background model. The latter was obtained by taking advantage of the dither pattern applied between exposures. The images were median combined, resulting in the background model. This enabled us to distinguish features at a fixed detector position from sky-related signals. The background model was rescaled to the illumination level of the individual exposures and then subtracted from them. We applied a sky background subtraction using `Source Extractor` (Bertin & Arnouts 1996). We obtained the astrometric calibration based on the *Gaia* DR1 catalogue (Gaia Collaboration 2016a,b) as reference. Finally, the images were co-added. We did not match the astrometry of the VLT observations to the one of the HST data. Checking for offsets between HST and VLT astrometrty with `Source Extractor` detected sources, we found small offsets of the order of 0.1 arcsec, which we corrected for in the photometric analysis.

3.3. Photometry

3.3.1. Photometric measurements with LAMBDAR

We performed photometric measurements on our fully reduced images with the Lambda Adaptive Multi-Band Deblending Algorithm in R (`LAMBDAR`[6], Wright et al. 2016). This algorithm can perform consistent and matched aperture photometry across images with varying pixel scales and resolutions. Therefore, it is ideally suited for our analysis, which requires accurate and precise colour measurements between the HST and VLT imaging with very different resolutions. In the following, we give a brief summary of the `LAMBDAR` algorithm. We refer the reader to Wright et al. (2016) for a more in-depth description.

`LAMBDAR` requires at least two inputs: a FITS image and a catalogue of object locations and aperture parameters. Additionally, we provide a point-spread function (PSF) model for the FITS image. These files are read in as the first step, then the aperture priors from the catalogue are transferred onto the same pixel grid as the input FITS image, using the image's astrometric solution (stored in the FITS header). Subsequently, the aperture priors are convolved with the input PSF, and object deblending is executed based on the convolved aperture priors. Images are deblended via multiplication with a deblending function. For this, it is assumed that the total flux in a pixel equals the sum of the fluxes from sources with aperture models overlapping that pixel. The flux per source is distinguished with the help of the deblending function. This function is calculated using the second assumption that the PSF convolved aperture model is a tracer of the emission profile of each source. Taking into account the estimation of the local sky-backgrounds and noise correlation using random/blank apertures, `LAMBDAR` calculates the object fluxes with the help of the deblended convolved aperture priors. Here, the code accounts for aperture weighting and/or missed flux through an appropriate normalisation of fluxes. Finally, flux uncertainties in relation to all of the above steps are determined.

For our purposes, using `LAMBDAR` has two main advantages. Firstly, we can comfortably perform matched aperture photometry across our images with varying PSF sizes between 0.″08 and 1.″07. Secondly, the prior aperture definitions derived from high-resolution optical imaging allow for deblending of sources leading to more accurate flux measurements, in particular in comparison to conventional fixed aperture photometry.

For each cluster, we ran `Source Extractor` on the $F606W$ image to obtain the input catalogue with object locations and aperture parameters. We set the detection and analysis thresholds to 1.4σ. We required a minimum of 8 pixels for a source detection and set the deblending threshold to 32 with a minimum contrast parameter of 0.005. Before the detection, the images were smoothed with a Gaussian filter of 5 pixels with a full width at half maximum (FWHM) of 2.5 pixels. We checked for residual shifts in the astrometry of our images with respect to the $F606W$ detection image and corrected for them with a linear shift if necessary to avoid biases in the flux measurements with `LAMBDAR`. For the HST images, we used `TinyTim` (Krist et al. 2011) to obtain a PSF model for the photometric analysis. For the ACS images (i.e. in $F606W$ and $F814W$), we looked up the average focus from the duration of the observation at the HST Focus Model tool[7]. Since this tool does not offer an estimate for WFC3/IR (i.e. for the images in $F110W$), we assumed a focus offset of 0.0 microns as default[8]. We used the central chip position as the position of reference for the estimation of the PSF model. In the case of the ACS instrument with two chips, we took the central pixel of chip 1 as a reference. For our VLT/FORS2 images, we obtained a PSF model with the help of the software `PSFEx` (Bertin 2011).

Some of our fully reduced images exhibited slight residual gradients in the background level. Therefore, we performed an initial run with `Source Extractor` to obtain a background-subtracted image. We used these as the FITS input images to be analysed with `LAMBDAR`.

3.3.2. Photometric zeropoints

The photometric calibration for the HST bands is straightforward. We obtained a photometric zeropoint (ZP) for each coadd from the header keywords `PHOTFLAM` and `PHOTPLAM`:

$$\begin{aligned} \mathrm{ZP_{AB}} = &-2.5\log_{10}(\mathrm{PHOTFLAM}) \\ &-5.0\log_{10}(\mathrm{PHOTPLAM}) - 2.408. \end{aligned} \quad (1)$$

`PHOTFLAM` is the inverse sensitivity, which facilitates the transformation from an instrumental flux in units of electrons per second to a physical flux density and `PHOTPLAM` denotes the pivot wavelength in units of Å[9]. Afterwards, we accounted for Galactic extinction with the extinction maps by Schlafly & Finkbeiner (2011)[10].

The challenge in the photometric calibration of the U_{HIGH} band is the lack of an adequate reference catalogue with well-calibrated U band magnitudes for our cluster fields. In such a case, a common method for calibration is to use a stellar locus (High et al. 2009). It is based on the assumption that stars occupy a well-defined region, the stellar locus, in colour-colour space, independent of the line of sight. Then, the photometry can be calibrated by matching the photometry of stars in an observation to the universal stellar locus. However, we found that direct use of a stellar locus does not work well for our analysis due to the limited number of stars in the small fields of view. Additional large scatter resulted in substantial uncertainties of the stellar locus approach. We, therefore, developed a calibration strategy based on a galaxy locus, where we

[6] https://github.com/AngusWright/LAMBDAR

[7] http://focustool.stsci.edu/cgi-bin/control.py

[8] To cross-check this assumption, we measured the photometry with an alternative PSF model with a very different focus offset of 4.0 microns. We found that both measurements differ by 0.001 mag (median), which is negligible for our purposes.

[9] https://www.stsci.edu/hst/instrumentation/acs/data-analysis/zeropoints

[10] Obtained from the website https://irsa.ipac.caltech.edu/applications/DUST/

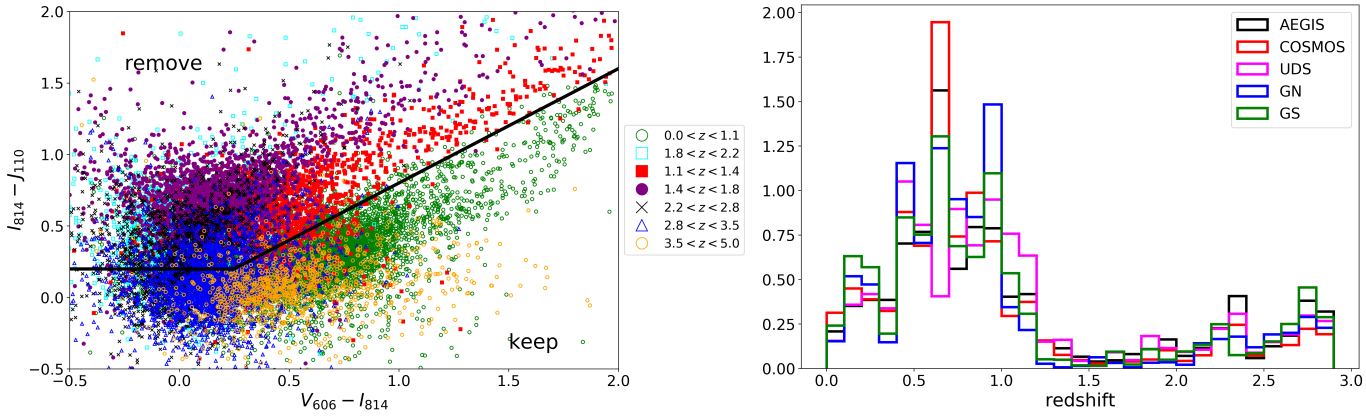


Fig. 1. Removal of galaxies in the cluster redshift regime from the galaxy locus at magnitudes of $24.2 < V_{606} < 27.0$. *Left*: cut in the VIJ plane to remove galaxies with photometric redshifts $1.2 \lesssim z \lesssim 1.7$ according to the catalogues by Skelton et al. (2014) (that is galaxies in the regime of cluster redshifts of the sample studied here), illustrated for galaxies from the GOODS-South field with photometry from Skelton et al. (2014). Red and purple symbols roughly correspond to galaxies in the cluster redshift regime. *Right*: redshift distribution of galaxies in our chosen galaxy locus from the five CANDELS/3D-HST fields.

made use of the fact that galaxies have a characteristic distribution in colour-colour space, similar to stars occupying the stellar locus. We identified a reference galaxy locus from the 3D-HST photometric catalogues as presented in Skelton et al. (2014). They summarised photometric measurements in the five CANDELS/3D-HST fields (AEGIS, COSMOS, GOODS-North [abbreviated GN], GOODS-South [abbreviated GS], and UDS) over a total area of ~900 arcmin^2. Among others, this includes the following bands relevant for our reference galaxy locus: the HST bands $F606W$ and $F814W$ and U bands from various instruments such as CFHT/MegaCam (AEGIS, COSMOS, and UDS), KPNO 4 m/Mosaic (GOODS-North), and VLT/VIMOS (GOODS-South). We describe in Sect. 3.3.3 how we accounted for the differences in these effective band-passes. Compared to the CANDELS/3D-HST fields our cluster fields are overdense at the cluster redshift, changing the local galaxy distribution in colour-colour space. To account for this, we applied a preselection, which used the well-calibrated HST-only colours to remove galaxies at the cluster redshift (see Fig. 1). In addition, the galaxy distribution varies locally due to line of sight variations. We reduced the impact of these by limiting the calibration with the galaxy locus to relatively faint galaxies in the regime $24.2 < V_{606} < 27.0$. We note that we optimised the calibration to be most accurate in this magnitude regime since it is the regime we used for the selection of background source galaxies (see Sect. 4). Together, this allowed us to calibrate $U - V_{606}$ colour estimates in the cluster fields by matching the galaxy distribution of the VIJ-selected galaxies in the $V_{606} - I_{814}$ versus $U - V_{606}$ colour-colour space.

For the calibration, we first accounted for Galactic extinction with the extinction maps by Schlafly & Finkbeiner (2011). We then smoothed the distribution of the galaxies in the UVI colour-colour space with a Gaussian kernel[11] both for the galaxies of the reference galaxy locus and the galaxies in our observation. We identified the peak position of the highest density and applied a shift to the $U_{\rm HIGH}$ magnitudes according to the difference in $U - V_{606}$ of the peak positions. We quantified and propagated the statistical uncertainty of 0.08 mag of our colour calibration scheme (see Appendix B for a robustness test of the $U_{\rm HIGH}$ band zeropoint calibration with the help of the reference galaxy locus; see Table 3 for the effect of this statistical uncertainty on the average geometric lensing efficiency).

3.3.3. Defining a common photometric system

When we investigate colour cuts for a suitable selection of background galaxies, we need to make sure to work in a consistent photometric framework. Regarding the U bands, we have measurements from four different instruments at hand: $U_{\rm HIGH}$ from VLT/FORS2 (our observations), $U_{\rm MEGACAM}$ from CFHT/MegaCam, $U_{\rm KPNO}$ from KPNO 4 m/Mosaic, and $U_{\rm VIMOS}$ from VLT/VIMOS (the latter three filters are employed in different CANDELS/3D-HST fields in Skelton et al. 2014). All of these have different effective filter curves. We, therefore, had to make sure that we employed these different bands to select consistent source populations, in particular regarding the $U - V_{606}$ colour. Comparing the $U - V_{606}$ colour of these populations, we found that there are small offsets among the CANDELS/3D-HST fields. We quantified these by identifying the peak position of the galaxy loci after smoothing the distribution with a Gaussian kernel (galaxies with $24.2 < V_{606} < 27.0$, where galaxies in the cluster redshift regime $1.2 \lesssim z \lesssim 1.7$ are excluded according to a cut in the VIJ colour plane; see Sect. 3.3.2). We applied a shift to the U bands to make the peak positions coincide with the peak position of the galaxy locus in GOODS-South as an anchor. We used this field as an anchor because we have observations of our own in $U_{\rm HIGH}$ in the HUDF situated within GOODS-South. We list the applied shifts in Table B.1. As a cross-check, we compared the peak positions in the $U - V_{606}$ colour distribution for differently selected galaxy subsamples in Fig. 2. Here, we generally found good agreement. For example, for the full population of galaxies with $24.2 < V_{606} < 27.0$, we measured a standard deviation of the density peak positions between the five CANDELS/3D-HST fields of 0.045 mag. We conclude that the photometry is sufficiently comparable as a basis for the selection of background galaxies (we summarise systematic and statistical uncertainties connected to the photometry at the end of Sect. 4.2). In addition to these considerations for the U bands, we used HST bands for which we have available observations for our cluster fields, that is, $F606W$, $F814W$, and $F110W$. Since not all reference catalogues have magnitude information on the galaxies in all of these bands, we needed to apply a few interpolations to estimate the fluxes and

[11] Using scipy.stats.gaussian_kde in python.

magnitudes of galaxies in our photometric system of filters. In this case, we performed an interpolation based on the closest available filters in effective wavelength, where one filter is redder (R) and one is bluer (B) than the missing filter (X):

$$F_X = s(\lambda_{\mathrm{eff,X}} - \lambda_{\mathrm{eff,B}}) + F_B,$$
$$m_X = -2.5\log_{10}(F_X) + \mathrm{ZP},$$
$$\text{with} \quad s = \frac{(F_R - F_B)}{(\lambda_{\mathrm{eff,R}} - \lambda_{\mathrm{eff,B}})}, \tag{2}$$

where F denotes the flux, m denotes the magnitude, ZP is the zeropoint (it is fixed to ZP = 25.0 for all bands in the Skelton et al. (2014) CANDELS/3D-HST photometric catalogues), and λ_{eff} is the effective wavelength of the respective filter. In a catalogue that covers the sources in all bands, we can gauge how well the interpolation typically represents the measured magnitude. Overall, there is a good match between the interpolated and the measured magnitudes. We do, however, see that the interpolation becomes increasingly noisy and asymmetric for fainter magnitudes. This is likely related to the (potentially different) depths of the available bands.

None of the available reference catalogues provides measurements in the band $F110W$. Options for interpolation are to use a combination of either $F105W$ and $F125W$, or $F850LP$ and $F125W$, or $F814W$ and $F125W$. Depending on the method used, we found that a small median offset of the order of 0.04 mag with a standard deviation of 0.07 mag can be introduced. We did not attempt to correct for such differences but we investigated the impact of systematic photometric offsets on the estimate of the average lensing efficiency in Appendix C, finding that the impact of such a systematic offset can well be neglected given our current statistical uncertainties. We also checked how well our photometry compares to measurements from Skelton et al. (2014) in Appendix A. From this, we concluded that slight offsets in photometry can occur, and we included the expected uncertainties in the overall error budget of our analysis (summarised at the end of Sect. 4.2).

4. Photometric selection of source galaxies and estimation of the source redshift distribution

For a robust weak lensing analysis, it is important to preferentially select the galaxies at redshifts higher than the cluster redshifts. Only these galaxies carry the weak lensing signal that we are interested in. We need to estimate the expected source redshift distribution of the selected galaxies to quantify the average geometric lensing efficiency $\langle\beta\rangle$ defined as

$$\langle\beta\rangle = \frac{\sum\beta(z_i)w_i}{\sum w_i}, \tag{3}$$

with the shape weights w_i (see Schrabback et al. 2018, and Sect. 5) and

$$\beta = \frac{D_{\mathrm{ls}}}{D_{\mathrm{s}}}H(z_{\mathrm{s}} - z_{\mathrm{l}}), \tag{4}$$

where D_{l}, D_{s}, and D_{ls} denote the angular diameter distances to the lens at redshift z_{l}, the source at redshift z_{s}, and between lens and source, respectively. The Heavyside step function is defined as $H(x) = 1$ if $x > 0$ and $H(x) = 0$ if $x \leq 0$. It is sufficient to estimate the averages $\langle\beta\rangle$ and $\langle\beta^2\rangle$ to tie the measured weak lensing shear signal to the cluster mass (e.g. Bartelmann & Schneider 2001).

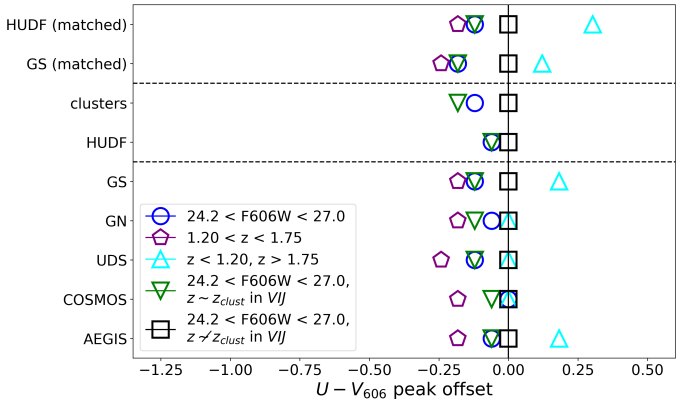


Fig. 2. Offsets between different populations of galaxies and the reference galaxy locus (galaxies with $24.2 < V_{606} < 27.0$, where galaxies at the cluster redshifts $1.2 \lesssim z \lesssim 1.7$ are excluded according to a cut in the VIJ colour plane; represented by black squares). Overall the populations exhibit quite similar offsets in $U - V_{606}$ colour despite relying on different U bands. *Top section:* comparison of directly matched galaxies in the HUDF region based on our measurements and the catalogue in GOODS-South by Skelton et al. (2014). *Mid section*: comparison of cluster fields (measurements from all nine cluster fields combined) and our measurements in the HUDF area, where we have U_{HIGH} imaging. Since we do not have photometric redshifts available, the populations relying on these are missing (purple pentagons and cyan triangles). *Bottom section*: comparisons for five CANDELS/3D-HST fields.

A straightforward but prohibitively observationally expensive way to identify the background galaxies is based on their spectroscopic redshifts. High-quality photometric redshifts can also be helpful if examined carefully for systematic outliers. Such redshift information is, however, not available for the galaxies in our observed cluster fields. Instead, we aim to use only the photometry from our observations to identify background galaxies. For this, we need reference catalogues of galaxies providing redshift and magnitude information in different bands. This allows us to understand how to distinguish background galaxies from contaminating foreground and cluster galaxies solely based on their colours. This is a commonly used strategy for weak lensing studies covering various redshift regimes (e.g. Klein et al. 2019; Schrabback et al. 2018, 2021). In the following section, we first describe the reference catalogues used in this work. After that, we present suitable cuts in colour space to preferentially select background galaxies for the weak lensing analyses. These cuts identified in photometric redshift reference catalogues can be safely applied to the cluster fields, because gravitational lensing is a colour-indifferent effect.

4.1. Redshift catalogues

4.1.1. UVUDF

The HUDF is a region of the sky that has been studied extensively both spectroscopically and in various photometric filters by the HST. Rafelski et al. (2015, henceforth R15) conducted a joint analysis of imaging ranging from near-ultraviolet (NUV) bands $F225W$, $F275W$, and $F336W$ (UVUDF, Teplitz et al. 2013), over optical bands $F435W$, $F606W$, $F775W$, and $F850LP$ (Beckwith et al. 2006), to near-infrared (NIR) bands $F105W$, $F125W$, $F140W$, and $F160W$ (UDF09 and UDF12, Oesch et al. 2010a,b; Bouwens et al. 2011; Koekemoer et al. 2013; Ellis et al. 2013). These data sets cover an area of

12.8 arcmin^2, but only 4.6 arcmin^2 have full NIR coverage. R15 provide photometric redshifts obtained with the code BPZ (Benítez 2000), which are highly robust due to the exquisite depth and high wavelength coverage of the data sets (e.g. demonstrated in Brinchmann et al. 2017, who found a median of $|(z_{\rm MUSE} - z_{\rm p})/(1 + z_{\rm MUSE})| < 0.05$ from a comparison of photometric redshifts to high quality redshifts from the MUSE integral field spectrograph). Given their accuracy, the R15 photo-zs provide an important benchmark for our computation of the average lensing efficiency. However, the small area covered in the sky leads to a substantial impact of sampling variance. Consequently, we also need to incorporate other data sets, which are shallower but cover a larger footprint in the sky (see Sect. 4.1.2).

4.1.2. 3D-HST

Skelton et al. (2014, henceforth S14) present catalogues with photometric measurements in filters covering a wide wavelength range and photometric redshifts for galaxies from the CANDELS/3D-HST fields over a total area of ∼900 arcmin^2. Their aim is to homogeneously combine various data sets available for these fields. Firstly, this includes the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS, Grogin et al. 2011; Koekemoer et al. 2011). It is an imaging survey conducted with HST/WFC3 and HST/ACS in five fields of the sky, namely AEGIS, COSMOS, GOODS-North, GOODS-South, and UDS. Secondly, the 3D-HST program (Brammer et al. 2012) provides slitless spectroscopy obtained with the WFC3 G141 grism for galaxies across nearly 75% of the CANDELS area and thus includes redshifts and spatially resolved spectral lines. Additionally, the WFC3 G141 grism spectroscopy data products are presented in Momcheva et al. (2016), who also developed software to optimally extract spectra for the objects from the S14 photometric catalogues. S14 combined the photometric data sets from the CANDELS and 3D-HST programmes with available ancillary data sets in the five CANDELS/3D-HST fields by using a common WFC3 detection image, conducting consistent PSF-homogenised aperture photometry, and estimating photometric redshifts and redshift probability distributions with the code EAZY (Brammer et al. 2008). The S14 photometric redshift catalogues form an excellent basis to estimate the redshift distribution for our weak lensing study. They cover a large area on the sky distributed over five independent lines-of-sight. This helps to combat sampling variance when estimating the average lensing efficiency. Additionally, the wide wavelength coverage, especially including deep NIR observations, is particularly valuable for robust redshift measurements out to high redshifts, as required for this study.

However, S18 and Raihan et al. (2020, henceforth R20) show that the photometric redshifts by S14 suffer from catastrophic outliers, which can significantly bias weak lensing mass measurements. Through the comparison of photometric redshift measurements from S14 and R15, R20 found that these outliers led to a systematic underestimation of the mean geometric lensing efficiency by −13.2% (for clusters at a redshift of 0.9) with a catastrophic outlier fraction of 5%. R20 were able to mitigate this by recomputing the photometric redshifts using the code BPZ instead of EAZY. In particular, the interpolation of the implemented spectral energy distribution (SED) template set helped reduce the bias[12]. For our weak lensing study, we used the updated R20 photometric redshift catalogues in the five CANDELS/3D-HST fields to estimate the average redshift distribution and lensing efficiency of our samples of selected background galaxies.

Additionally, S18 found some systematic deviations between the R15 photometric redshifts and the grism redshifts (Brammer et al. 2012; Momcheva et al. 2016). Upon revisiting this comparison, now including MUSE spectroscopic redshifts (Inami et al. 2017, see Sect. 4.1.4 below for details), R20 identified the affected redshift regimes and corrected the respective bias by subtracting the median offset. This bias amounts to 0.081 (0.162) for the photo-z regime $1.0 < z < 1.7$ ($2.6 < z < 3.2$). The resulting 'fixed' redshift catalogues do not suffer from issues with catastropic redshift outliers and are denoted as R15_fix catalogues.

4.1.3. HDUV

The *Hubble* Deep UV Legacy Survey (HDUV, Oesch et al. 2018, henceforth Oe18) is an imaging programme, which expands on the S14 catalogues with deeper UV observations in the WFC3/UVIS bands $F275W$ and $F336W$. It targets ∼100 arcmin^2 within the GOODS-North and GOODS-South fields. Oe18 conducted photometry consistent with S14 regarding the detection image and flux measurements and recalculated photometric redshifts with the EAZY code including their deeper UV images.

4.1.4. MUSE

The MUSE *Hubble* Ultra Deep Field Survey (Bacon et al. 2015; Inami et al. 2017; Brinchmann et al. 2017) comprises spectroscopic redshift measurements of almost 1400 sources in the HUDF region. This increases the number of available spectroscopic redshifts in this region by a factor of eight. It was conducted with the Multi Unit Spectroscopic Explorer (MUSE) at the Very Large Telescope. Inami et al. (2017) provide spectroscopic redshifts for sources with a completeness of 50% at 26.5 mag in $F775W$. The redshift distribution includes sources beyond $z > 3$ and up to a $F775W$ magnitude of ∼30 mag. This spectroscopic redshift catalogue is an excellent reference to judge the reliability of the photometric redshift catalogues used for the colour selection of background galaxies.

4.2. Selection of background galaxies through colour cuts

4.2.1. Defining the colour and magnitude cuts

We aim to find criteria based on colours and magnitudes that help us distinguish the background galaxies of interest from the contaminating foreground and cluster galaxies. To this end, we made use of the S14/R20 catalogues providing photometry and photometric redshifts for the largest number of galaxies. First, we decided to focus on the magnitude regime $24.2 < V_{606} < 27.0$ for the selection strategy. Inspecting the redshift distributions of galaxies in the CANDELS/3D-HST fields, we found that there is no significant amount of background galaxies at redshifts $z \gtrsim 1.8$ present at magnitudes brighter than $V_{606} < 24.2$. By focusing on galaxies fainter than this limit, we could exclude

[12] When recomputing the photo-zs, R20 employed an approximately homogeneous subset of broad-band filters (between U and H band), which are available for all five CANDELS fields. Since they dropped additional bands, this may increase the scatter in some of the photo-z estimates compared to the S14 catalogue. However, for our analysis it is more important to have accurate estimates of the overall redshift distribution of colour-selected high-z lensing source galaxies, as provided by the R20 catalogues.

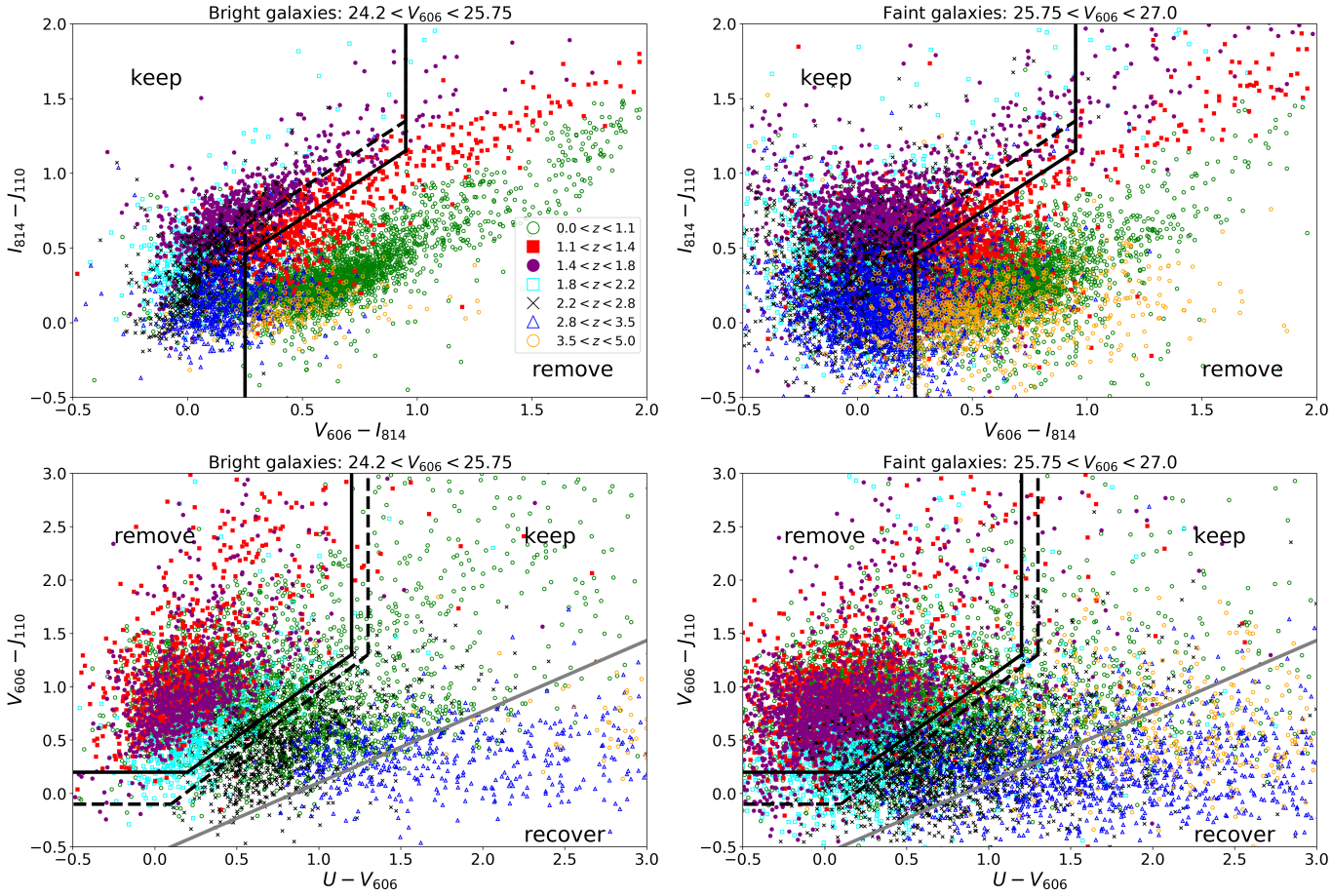


Fig. 3. Colour selection for galaxy clusters at redshift $1.2 \lesssim z \lesssim 1.7$. The selected source galaxies are at redshift $z \gtrsim 1.7$. We display galaxies based on their photometry from S14 in the GOODS-South field. *Top*: first selection step in the VIJ plane for bright galaxies on the left and faint galaxies on the right. *Bottom*: second selection step in the UVJ plane for bright galaxies on the left and faint galaxies on the right. The solid black lines indicate cuts applied for bright galaxies, the dashed black lines show cuts for faint galaxies. Galaxies below the diagonal grey line are recovered in both the bright and the faint regime.

many bright foreground galaxies. Additionally, our cluster fields roughly reach limiting magnitudes of 27 mag in the $F606W$ band.

Second, we inspected the colour-colour plots of different combinations of colours to identify a suitable strategy. We found that a combination of the colour plane including V_{606}, I_{814}, and J_{110} and the colour plane including U, V_{606}, and J_{110} provided a useful basis for a selection of background galaxies, that is galaxies at redshifts higher than the cluster redshifts of $1.2 \lesssim z \lesssim 1.7$. We developed a selection consisting of two steps. For the first step, a strategic cut in the colour plane $V_{606} - I_{814}$ and $I_{814} - J_{110}$ (short VIJ plane) allowed us to remove a significant fraction of foreground galaxies at $0.0 < z < 1.1$. We discarded all galaxies to the right of this cut (redder in $V_{606} - I_{814}$, see the black line in upper panels of Fig. 3). With this cut, we did, however, still keep a lot of galaxies at the cluster redshift while discarding a substantial fraction of background galaxies at $z > 2.2$. The second step using the colour plane $U - V_{606}$ and $V_{606} - J_{110}$ (short UVJ plane) helped us to refine the selection. Here, we could remove almost all galaxies at the cluster redshift (galaxies that are blue in $U - V_{606}$ and red in $V_{606} - J_{110}$, occupying the upper left corner of the UVJ plane in Fig. 3), and at the same time recover high-redshift sources we had discarded in the first selection step (galaxies that are red in $U - V_{606}$, occupying the lower right corner of the UVJ plane in Fig. 3). Additionally, we slightly varied these cuts depending on whether the galaxies were bright ($24.2 < V_{606} < 25.75$) or faint ($25.75 < V_{606} < 27.0$). Fainter galaxies typically exhibit a larger photometric scatter than brighter galaxies. We could, therefore, apply slightly tighter cuts for brighter galaxies without a high risk of contamination by cluster galaxies due to scatter. Figure 3 illustrates our cuts in the two colour planes and for the bright and faint magnitude regimes for clusters at redshift $1.2 \lesssim z \lesssim 1.7$. For clarity, we summarise the selection strategy as follows: we selected all galaxies below the grey line in the UVJ plane and all galaxies that are both to the left of the black line in the VIJ plane and to the right of the black line in the UVJ plane.

We also investigated if it is possible to optimise the selection depending on the cluster redshift. For instance galaxies at redshift $1.3 < z < 1.7$ could be used for a cluster at redshift $z = 1.2$, but have to be removed for a cluster at redshift $z = 1.7$. Unfortunately, such an optimisation was not possible with the available filters because all the galaxies in the redshift regime $1.2 \lesssim z \lesssim 1.7$ occupy a similar location in the UVJ plane (see red and purple symbols in Fig. 3). We investigated two alternative selection strategies in Appendix D, which did not improve the signal-to-noise ratio (S/N) of the lensing analysis. We, therefore, decided to use common selection criteria for background galaxies, independent of the cluster redshift for the majority of our cluster sample. The only exception is the cluster

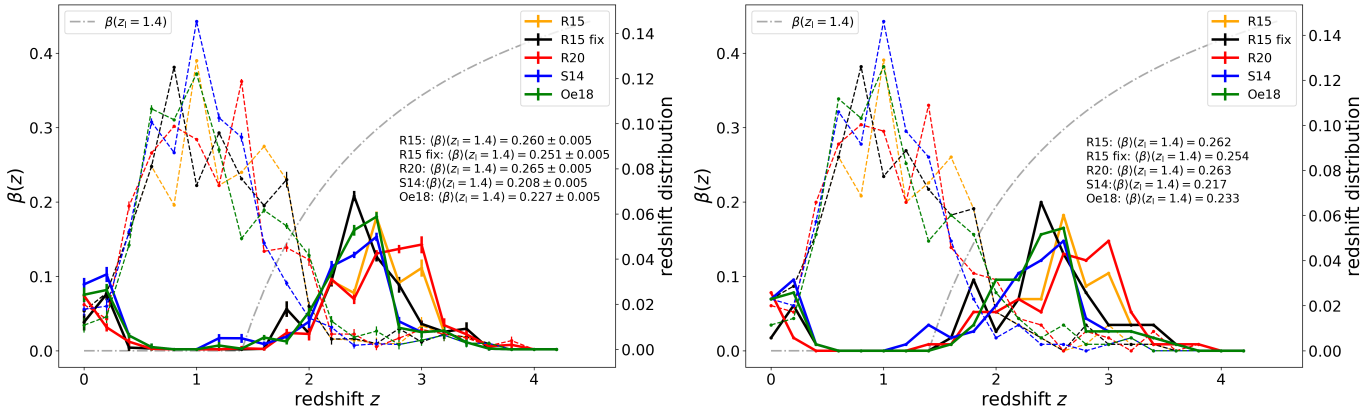


Fig. 4. Redshift distributions resulting from the colour selection for galaxy clusters at redshift $1.2 \lesssim z \lesssim 1.7$ in the HUDF region. The selected source galaxies (solid lines) are mostly at redshift $z \gtrsim 1.7$. Removed galaxies (dashed lines) are mostly at redshifts $z \lesssim 1.7$. The distributions only show galaxies matched between the five reference redshift catalogues (R15, R15_fix, R20, S14, and Oe18) and the photometric catalogue from this work. We additionally display the average lensing efficiency curve as a function of redshift (grey dash-dotted line) at the median lens redshift of the clusters at $z_l = 1.4$. *Left*: redshift distributions for the five redshift catalogues and employing a colour selection based on the S14 photometry. The uncertainties represent the standard deviations from 50 noise realisations of the U band in the S14 photometry. *Right*: redshift distributions for the five redshift catalogues and employing a colour selection based on the LAMBDAR photometry measured from our observations of HUDF in $U_{\rm HIGH}$ and from the S14 stacks in different HST-bands (see text).

SPT-CL J0646−6236 at the lowest redshift of $z = 0.995$. We used an optimised selection strategy for this particular cluster, which we describe in Appendix E.

Additionally, we investigated how beneficial the use of the U band is for an efficient source selection since it is the band introducing the largest uncertainties. We found that it is possible to select sources with a similar average geometric lensing efficiency only based on the bands $F606W$, $F814W$, and $F110W$. However, the resulting source density of such a selection is significantly lower. In conclusion, the signal-to-noise ratio of the lensing measurement (proportional to the product of the average lensing efficiency and the square root of the source density) is about 1.4 times higher when the U band is included for the source selection.

4.2.2. Comparison of selections based on the S14 and the LAMBDAR photometry

We calculated the average lensing efficiency $\langle\beta\rangle$ for the selection based on the S14 photometry and for five catalogues with photometric redshift information, namely the original S14 redshifts, the updated R20 redshifts by R20, the redshifts given in R15, a modified version of the R15 redshifts from R20 called R15_fix, and the redshifts from Oe18. Throughout this section, we used the median lens redshift of our cluster sample of $z_l = 1.4$ for the calculation of $\langle\beta\rangle$. In addition to the selection as described in Sect. 4.2.1, we employed a signal-to-noise (S/N) threshold of $S/N_{\rm flux,606} > 10$ as applied for the shape measurements of galaxies (the signal-to-noise ratio is defined via the ratio of FLUX_AUTO and FLUXERR_AUTO from Source Extractor; see also Sect. 5). We note that R20 optimised the redshifts for a source selection targeting background galaxies behind clusters of $0.6 \lesssim z \lesssim 1.1$ (the cluster sample from S18). They apply a cut based on $V - I$ colour at $V - I < 0.3$ and a magnitude cut of $V_{606} < 26.5$. Even though these settings differ from ours, we found that the R20 catalogues are still applicable for our analysis because on average 84% of galaxies in our selection in the cluster fields also fulfil the condition $V - I < 0.3$. Additionally, we found that the average lensing efficiency calculated based on R20 photo-zs for our colour-selected galaxies in the HUDF was not significantly affected by a change of the magnitude limit from $V_{606} < 27.0$ to $V_{606} < 26.5$.

The five redshift catalogues (denoted R15, R15_fix, R20, S14, and Oe18) overlap in the HUDF region. We matched the sources from our five reference catalogues based on their coordinates through the function associate from the LDAC tools[13]. For a match, we required a distance smaller than 0.″3. In Fig. 4, we show the redshift distribution of the galaxies that we selected with our strategy. We note that the S14 U band (5σ depth = 27.9) is considerably deeper than our observations in the $U_{\rm HIGH}$ band in the HUDF (5σ depth = 26.6). To account for this difference, we added Gaussian noise to the S14 U band photometry and show the average redshift distribution derived from 50 noise realisations of galaxies in the HUDF for a $U_{\rm HIGH}$ band depth of 26.6 mag in Fig. 4. We note that, when we estimated the average lensing efficiency for the cluster fields, we added Gaussian noise to both the U band and HST photometry from the S14 catalogues to account for the difference between the depths in the respective cluster fields and in the CANDELS/3D-HST fields. When we calculated the average lensing efficiency, we employed the shape weights from S18 that depend on the signal-to-noise ratio (FLUX_AUTO/FLUXERR_AUTO) in V_{606}. Since the S14 catalogues do not provide measurements of FLUX_AUTO, we used the listed total fluxes and respective errors instead[14]. The redshift distributions show that S14 and Oe18 have an excess of galaxies at the cluster redshifts and in the foreground at $z < 0.4$ compared to the other catalogues. This is connected to the reported contamination by catastrophic redshift outliers (see Sect. 4.1.2). We can see this effect as well in Fig. 4 where the S14 and Oe18 redshift catalogues yield lower values of the average lensing efficiency than the other redshift catalogues. In contrast to that, the average lensing efficiency results from the R20 redshift catalogues are in

[13] https://marvinweb.astro.uni-bonn.de/data_products/THELIWWW/LDAC/

[14] As a cross-check, we calculated the average lensing efficiency with the shape weights based on the total fluxes in the S14 catalogues and the AUTO fluxes in catalogues by S18. They have analysed shallower stacks in the CANDELS/3D-HST fields, including measurements of FLUX_AUTO, which allowed us to draw a direct comparison. We found that the difference between both options is less than 1%.

good agreement with the robust photometric redshift catalogues R15 and R15_fix. According to these catalogues, we expect nearly no contamination by cluster galaxies for our selection strategy (only ∼1% of selected galaxies are within the cluster redshift range). Figure 4 displays a small residual contribution of foreground galaxies in our source selection. This is, however, not a concern as long as the redshift distribution is modelled accurately. From a comparison of the average lensing efficiency based on R20 and R15_fix we infer a systematic uncertainty of $\Delta(\langle\beta\rangle)/\langle\beta\rangle_{\mathrm{R15\_fix}} = 5.6\%$.

Since we measured fluxes in our observations with LAMBDAR, we additionally inspected the redshift distributions that we obtained when we used the LAMBDAR photometry measured from our observations of the HUDF in U_{HIGH} and from the S14 stacks in the HST filters $F606W$, $F814W$, $F850LP$ and $F125W$[15] (we interpolated between the latter two filters to estimate the magnitude in the filter $F110W$). The resulting distribution is shown on the right-hand side of Fig. 4. This corresponds to a systematic uncertainty of $\Delta(\langle\beta\rangle)/\langle\beta\rangle_{\mathrm{R15\_fix}} = 3.5\%$. Overall, the average lensing efficiency results based on S14 and LAMBDAR photometry agree within the uncertainties (see Fig. 4).

4.2.3. Comparison of selections based on photo-zs and spec-zs

As a cross-check for the photometric redshift catalogues, we retrieved spectroscopic/grism redshifts from the MUSE and 3D-HST catalogues, respectively, for all galaxies matched by their coordinates in the HUDF field. As a reference, we then calculated the average lensing efficiency of the colour-selected sources based on the spectroscopic/grism redshifts. Here, we only used the MUSE spec-zs with the highest quality flags 3 (secure redshift, determined by multiple features) and 2 (secure redshift, determined by a single feature, see Inami et al. 2017). In the case of galaxies with both spectroscopic redshifts from MUSE and grism redshifts from 3D-HST, we used the former for the calculation of $\langle\beta_{\mathrm{spec}}\rangle$. To estimate the uncertainty, we bootstrapped the colour-selected galaxies and recalculated the average lensing efficiency 1000 times. Figure 5 shows how the average lensing efficiency calculated from the five photometric redshift catalogues compares to the one calculated based on spectroscopic/grism redshifts. We did not find a bias within the uncertainties, but we notice that the average lensing efficiency based on R15_fix, R20 and Oe18 matches closest to the result based on the spectroscopic/grism redshifts. It also has to be noted that the spectroscopic/grism redshifts are only complete in comparison to the full sample of matched galaxies in the HUDF region up to a magnitude of $V_{606} \lesssim 25.0$ mag (see Fig. 6). We still decided to correct our measurements of the average lensing efficiency by the roughly three percent offset between the R20 redshift-based and the spectroscopic redshift-based lensing efficiency for all clusters except SPT-CL J0646−6236. For the specific source selection used for this cluster, such an offset did not occur.

4.2.4. Differences between the five CANDELS/3D-HST fields

Since we estimate the average lensing efficiency from all CANDELS fields, we want to evaluate the expected systematic uncertainties arising from differences in the depths, available

[15] https://archive.stsci.edu/prepds/3d-hst/;
($F606W + F850LP$: GO programme 9425 with PI M. Giavalisco, $F814W$: GO programme 12062 with PI S. Faber, $F125W$: GO programme 13872 with PI G. Illingworth).

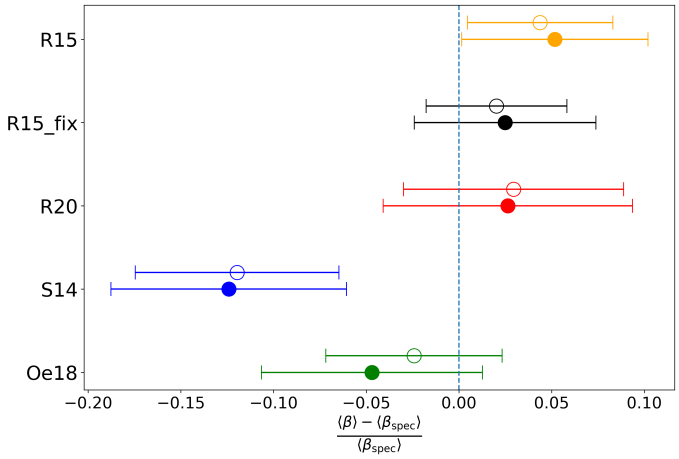


Fig. 5. Relative bias in the average lensing efficiency normalised to the result based on spectroscopic/grism redshifts for the five reference redshift catalogues (R15, R15_fix, R20, S14, and Oe18). We performed the colour selection for all coordinate-matched galaxies in the HUDF with available spectroscopic/grism redshifts. The uncertainties represent the scatter from 1000 bootstrap resamples. Filled symbols represent source selections based on the S14 photometry, open symbols represent source selections based on the LAMBDAR photometry.

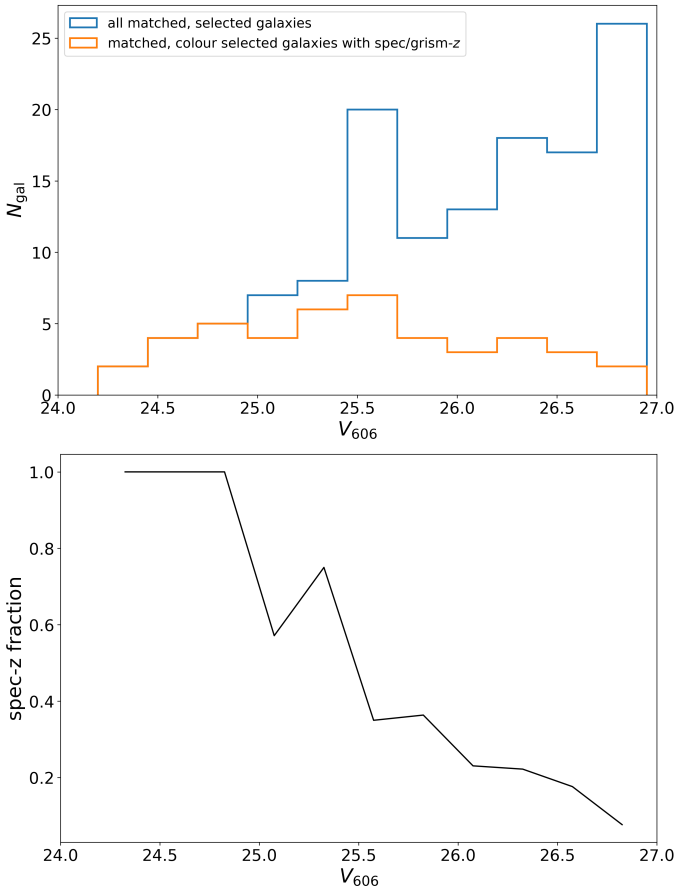


Fig. 6. Overview about available spectroscopic redshift information as a function of magnitude. *Top*: histogram of all matched and colour-selected galaxies within the HUDF region (blue). The orange histogram shows how many of these have a robust spec-z from MUSE or grism-z. *Bottom*: fraction of matched and colour-selected galaxies within the HUDF region with a robust spec-z from MUSE or grism-z, corresponding to the ratio of the orange and blue histogram in the *top panel*.

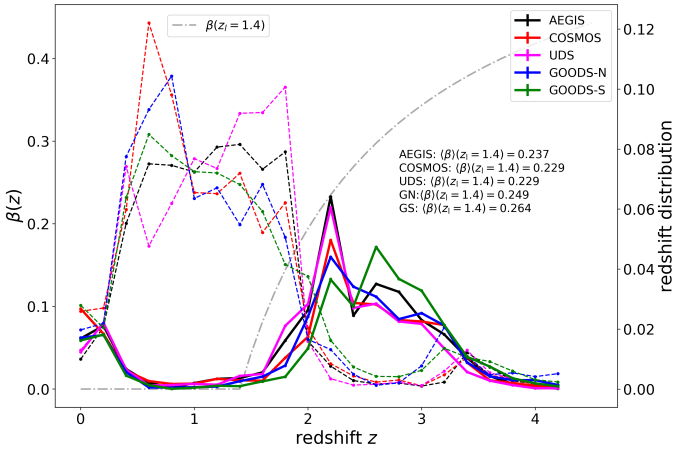


Fig. 7. Redshift distribution of the galaxies in the CANDELS/3D-HST fields for the colour selection for clusters at $1.2 \lesssim z \lesssim 1.7$, employing the R20 photometric redshift catalogues. The selected source galaxies (solid lines) are mostly at redshift $z \gtrsim 1.7$. Removed galaxies (dashed lines) are mostly at redshifts $z \lesssim 1.7$. We additionally display the average lensing efficiency curve as a function of redshift (grey dash-dotted line) at the median lens redshift of the clusters at $z_l = 1.4$.

filters, and calibrations in the five CANDELS/3D-HST fields. Additionally, we expect statistical sampling variance due to line of sight variations.

We quantified the systematic uncertainties by measuring the average lensing efficiency for colour-selected galaxies independently in the five CANDELS/3D-HST fields (see Fig. 7). We obtained a mean of the average lensing efficiencies of $\langle\beta\rangle_{\rm mean} = 0.242$ with a standard deviation of $\sigma(\langle\beta\rangle) = 0.014$ between the $N = 5$ fields (using the photometric redshifts from R20). This translates into a systematic uncertainty of $\sigma(\langle\beta\rangle)/\langle\beta\rangle_{\rm mean} = 5.7\%$. We calculated this more conservative systematic uncertainty without dividing by $\sqrt{N-1}$ because we noticed that the value of the GOODS-South field is notably higher, and thus, one field might not automatically be a good representation of the average of all. We added this uncertainty in quadrature to our systematic error budget (see Table 3). We note that this uncertainty also contains a statistical contribution as each CANDELS/3D-HST field represents a different line of sight. However, since the fields are each much larger than the small subpatches studied in the paragraph below, we conservatively assume that the variations between the CANDELS/3D-HST fields are dominated by systematic uncertainties.

We gauged the expected statistical uncertainty from line of sight variations in the average lensing efficiency by placing non-overlapping apertures with the same area as the field of view of our observations (about 11 arcmin^2) in the CANDELS/3D-HST fields. We can fit exactly eight apertures in each of the fields. We calculated the average lensing efficiency independently for all of the apertures, where we obtained the mean $\langle\beta\rangle_{\rm mean} = 0.243$ with a scatter of $\sigma(\langle\beta\rangle) = 0.017$. Hence, we added a statistical uncertainty of $\sigma(\langle\beta\rangle)/\langle\beta\rangle_{\rm mean} = 6.9\%$ to our statistical error budget (see Table 3).

Regarding uncertainties of the source redshift distribution, we estimated a total statistical uncertainty of 8.0% on the average lensing efficiency corresponding to 12.1% on the mass scale. This includes uncertainties in the $U_{\rm HIGH}$ band calibration (see Appendix B) and line of sight variations (this section), which we added in quadrature. Furthermore, we estimated a total systematic uncertainty of 8.6% on the average geometric lensing efficiency. Here, we took into account systematics for the $F110W$ band (interpolation versus direct measurement, aperture photometry versus LAMBDAR photometry, see Appendices A and C), uncertainties in the measurement of $V - I$ colours (see Appendices A and C), uncertainties of the R20 redshift catalogues (see Sect. 4.2.2), and variations between the CANDELS/3D-HST fields (differences of the filters, depths, availability of U bands, and usage of different bands to interpolate the J_{110} magnitudes, see this section). Again, we added these contributions in quadrature. All of these uncertainties are summarised in Table 3.

Table 3. Summary of our systematic and statistical error budget.

Source of systematic uncertainties	Rel. error signal	Rel. error M_{500c}	Sect./ App.
Redshift distribution:			
– R20 vs. R15_fix comp.	5.6%	8.4%	4.2.2
– Variations between CANDELS/3D-HST fields	5.7%	8.6%	4.2.4
– $F110W$ band (LAMBDAR/S14, interp.)	2.2%	3.3%	A/C
– $V - I$ colour (LAMBDAR/S14)	2.2%	3.3%	A/C
Shape measurements:			
– Shear calibration	2.3%	3.4%	5
Mass model:			
– $c(M)$ relation		4.0%	
– Miscentring for			
X-ray centres		3.8%/	6.3
SZ centres		9.2%	6.3
Total (added in quadrature)		14.4%/ 16.7%	

Source of statistical uncertainties	Rel. error signal	Rel. error M_{500c}	Sect./ App.
Redshift distribution:			
– Line of sight variations	6.9%	10.4%	4.2.4
– $U_{\rm HIGH}$ band calibration	4.1%	6.2%	B/C
Total (added in quadrature)		12.1%	

Notes. In the upper part of the table, we list all systematic uncertainties, which ultimately translate into an uncertainty in the weak lensing mass measurement, where we added the individual contributions in quadrature to obtain an estimate for the total uncertainty. We report the relative uncertainties in percent in the second column, the resulting relative uncertainty on the mass in the third column, and refer the reader to the respective sections or appendices listed in the last column for more detailed information about the contributions to the error budget. In the lower table, we list statistical uncertainties in the redshift distribution, which affect the calculation of the average geometric lensing efficiency $\langle\beta\rangle$ for individual cluster fields. We note that the final statistical uncertainties reported in Tables 6 and 7 do include additional contributions from shape noise and uncorrelated large-scale structure projections.

4.3. Check for cluster member contamination

We aim to preferentially select background galaxies with our magnitude and colour cuts both in the cluster fields and the CANDELS/3D-HST fields. Investigating the source density of the selected galaxies and their radial dependence allows us to test if we have a substantial amount of contamination by cluster galaxies and if our method provides a consistent selection in the cluster fields and the CANDELS/3D-HST fields in the presence of noise (S18).

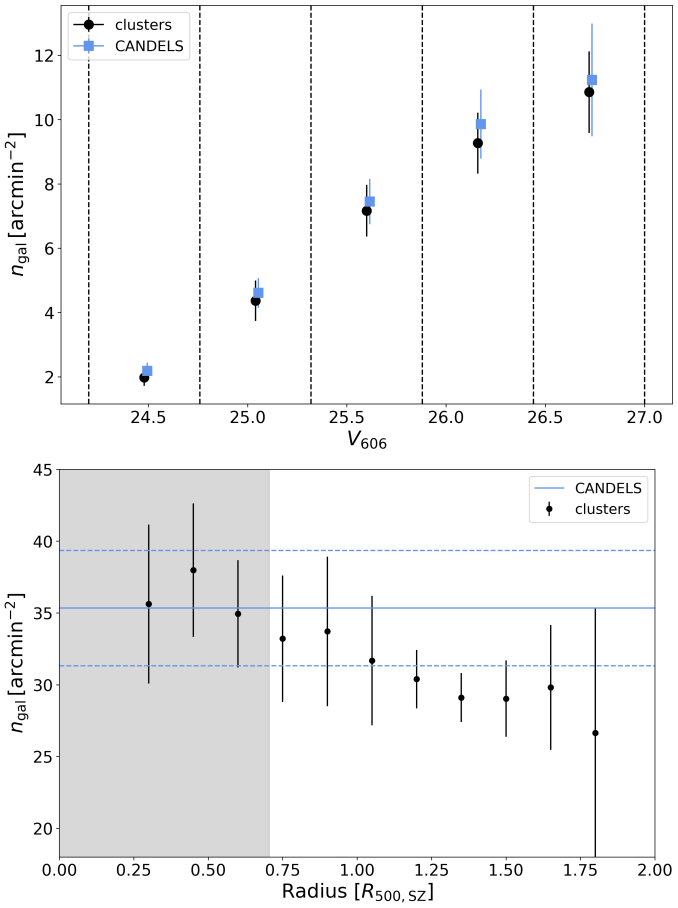


Fig. 8. Number density profiles of selected source galaxies. *Top*: number density of selected galaxies $n_{\rm gal}$ averaged over the nine cluster fields (black symbols) and averaged over the five 3D-HST/CANDELS fields (blue symbols) as a function of magnitude. We took into account the masks, for example, from bright stars in the images, and we only considered photometrically selected galaxies, that is, no flags from shape measurements or signal-to-noise ratio cuts were considered here. The error bars correspond to the uncertainty of the mean from the variation between the contributing cluster fields or 3D-HST/CANDELS fields, respectively. *Bottom*: average density of selected sources as a function of the distance to the X-ray cluster centre (except for the cluster SPT-CL J0646−6236, where we used the SZ centre). These distances are given in units of the radius $R_{\rm 500c,SZ}$ based on the SZ mass $M_{\rm 500c,SZ}$. Blue lines indicate the average density and 1σ uncertainties from the five 3D-HST/CANDELS fields. The error bars correspond to the uncertainty of the mean from the variation between the contributing cluster fields or 3D-HST/CANDELS fields, respectively. We excluded the grey-shaded region when we measured weak lensing masses. It corresponds to 500 kpc or about 0.71 R_{500} for a cluster with R_{500} = 700 kpc.

To this end, we added Gaussian noise to the S14 photometric catalogues according to the difference between the depth of the cluster observations and the depth of the CANDELS/3D-HST fields. This may vary depending on the field and filter. We only added Gaussian noise if the CANDELS/3D-HST observation in a filter were deeper than the corresponding observation in the cluster field. Occasionally, the cluster observations were slightly deeper than some of the CANDELS/3D-HST observations, but only by ∼0.2 mag. We considered this negligible for the validity of this test.

We measured the source density of selected sources accounting for masks, for example, due to bright stars for the cluster fields and CANDELS/3D-HST fields. We only considered photometrically selected galaxies and did not consider potential flags from the shape measurement pipeline. We also did not apply the signal-to-noise ratio cut $S/N_{\rm flux,606} > 10$ as mentioned in Sects. 4.2 and 5 for this test, since the quantities `FLUX_AUTO` and `FLUXERR_AUTO` required to calculate the signal-to-noise ratio are not available in the CANDELS/3D-HST catalogues. In Fig. 8 (top panel), we show the average source density of selected galaxies as a function of the V_{606} band magnitude. We found a good agreement over the full magnitude range of interest in this study.

Additionally, we investigated the radial dependence of the source density of selected galaxies. In principle, an increase of the number density towards the cluster centre can indicate cluster member contamination. However, the profile can also be affected by blending and/or masking of background galaxies by cluster member galaxies, magnification, or selection effects. We accounted neither for blending and/or masking by cluster galaxies nor magnification in our analysis. The blending/masking by cluster galaxies should be less important than for clusters at lower redshifts since the cluster galaxies are more cosmologically dimmed. Additionally, we conservatively excluded the core region $r < 500$ kpc, when we measured the weak lensing masses so that this effect should not play a significant role. Regarding magnification, for S21 the application of a magnification correction had only a minor impact on the source density profile. Given the higher redshifts of our clusters, the lensing strength and, therefore, the expected impact of magnification is even lower, which is why we ignore it here.

Figure 8 (bottom panel) displays the radial distance from the X-ray centre (except for the cluster SPT-CL J0646−6236, where we used the SZ centre) in units of the radius $R_{\rm 500c,SZ}$, which we derived from the SZ mass $M_{\rm 500c,SZ}$. We found a very slight trend of a higher source density towards the centres of the clusters. However, the profile is consistent with flat within the uncertainties. Together, both measurements provide an important confirmation for the success of the photometric background selection and cluster member removal.

5. Shape measurements

The shape of a galaxy can be quantified by its ellipticity, as a complex number $\epsilon = \epsilon_1 + {\rm i}\epsilon_2$. The observed ellipticity $\epsilon_{\rm obs}$ of a background galaxy can be related to the intrinsic ellipticity $\epsilon_{\rm orig}$ and reduced shear g via (Bartelmann & Schneider 2001)

$$\epsilon_{\rm obs} = \frac{\epsilon_{\rm orig} + g}{1 + g^* \epsilon_{\rm orig}}. \tag{5}$$

According to the cosmological principle, the intrinsic orientation of galaxies should have no preferred direction[16]. Therefore, the expectation value for an average over many galaxies $\langle \epsilon_{\rm orig} \rangle = 0$ vanishes. In conclusion, we can estimate the reduced shear, that is, the main observable for weak lensing studies, from the ensemble-averaged PSF-corrected ellipticities of the background galaxies via

$$\langle \epsilon_{\rm obs} \rangle = g. \tag{6}$$

We measured galaxy shapes in the ACS $F606W$ (V) and $F814W$ (I) images using the KSB+ formalism (Kaiser et al. 1995; Luppino & Kaiser 1997; Hoekstra et al. 1998) as

[16] Despite this principle, intrinsic alignments of galaxies due to various physical effects can pose a challenge for weak lensing analyses, especially for cosmic shear studies. See for example Troxel & Ishak (2015) for a review. These intrinsic alignments are, however, not a concern for this work.

implemented by Erben et al. (2001) and Schrabback et al. (2007). We modelled the spatially and temporally varying ACS point-spread function using an interpolation based on principal component analysis, as calibrated on dense stellar fields (Schrabback et al. 2010, 2018). We corrected for shape measurement and selection biases as a function of the KSB+ galaxy signal-to-noise ratio from Erben et al. (2001). This correction was derived by Hernández-Martín et al. (2020), who analysed custom `Galsim` (Rowe et al. 2015) image simulations with ACS-like image characteristics. Importantly, Hernández-Martín et al. (2020) tuned their simulated source samples such that the measured distributions in galaxy size, magnitude, signal-to-noise ratio, and ellipticity dispersion closely matched the corresponding measured distributions of the magnitude and colour-selected source samples from S18, while also incorporating realistic levels of blending. Varying the properties of the simulations, Hernández-Martín et al. (2020) estimated a (post-correction) multiplicative shear calibration uncertainty of the employed KSB+ pipeline of ∼1.5%. Our data are very similar to those analysed by S18. Therefore, we expect the Hernández-Martín et al. (2020) shear calibration to be directly applicable for our analysis. However, our colour selection selects galaxies at slightly higher redshifts on average compared to the $V-I$ selection from S18. Some of our image stacks are also slightly deeper. We, therefore, conservatively increased the shear calibration uncertainty in our systematic error budget by a factor ×1.5 (see Table 3).

Given their greater average depth (see Table 2), we based our shear catalogue primarily on the $F606W$ stacks. Here, we included galaxies with a measured flux signal-to-noise ratio $S/N_{\rm flux,606} > 10$[17] (defined as the ratio of the `FLUX_AUTO` and `FLUXERR_AUTO` parameters from `Source Extractor`). This single-band selection matches the one employed in Sect. 4.2 in the computation of the average geometric lensing efficiency. For galaxies that additionally have $S/N_{\rm flux,814} > 10$, we combined the shape measurements from both filters to reduce the impact of measurement noise.

In order to compute shape weights and filter-combined estimates of the reduced shear, we made use of the $\log_{10} S/N_{\rm flux}$-dependent fits computed by S18, see their appendix A for the total ellipticity dispersion $\sigma_{\epsilon,V/I}$, the intrinsic ellipticity dispersion $\sigma_{{\rm int},V/I}$, and the ellipticity measurement noise $\sigma_{{\rm m},V/I}$ of $V-I$ colour selected galaxies in custom CANDELS (Grogin et al. 2011) V ($F606W$) and I ($F814W$) band stacks of approximately single-orbit depth[18]. With the complex reduced shear estimates $\epsilon_{V/I}$ obtained in the V band and the I band, respectively, and the shape weights

$$w_{V/I} = \left(\sigma_{\epsilon,V/I}\right)^{-2}, \tag{7}$$

we computed the filter-combined reduced shear estimate as

$$\epsilon_{\rm comb} = \frac{w_V \epsilon_V + w_I \epsilon_I}{w_V + w_I}. \tag{8}$$

The measurement noise is independent between the stacks in the different filters, which is why the combined ellipticity measurement variance reads

$$\sigma^2_{\rm m,comb} = \frac{(w_V \sigma_{{\rm m},V})^2 + (w_I \sigma_{{\rm m},I})^2}{(w_V + w_I)^2}. \tag{9}$$

In the relevant S/N or magnitude regime, differences are small between $\sigma_{{\rm int},V}$ and $\sigma_{{\rm int},I}$ for the colour-selected source samples from S18. In addition, Jarvis & Jain (2008) found that intrinsic shapes are highly correlated between HST images of galaxies in different optical filters. Therefore, as an approximation, we interpolated the intrinsic ellipticity dispersion between the filters

$$\sigma_{\rm int,comb} = \frac{w_V \sigma_{{\rm int},V} + w_I \sigma_{{\rm int},I}}{w_V + w_I}, \tag{10}$$

allowing us to compute shape weights for the combined shear estimate as

$$w_{\rm comb} = \left(\sigma^2_{\rm int,comb} + \sigma^2_{\rm m,comb}\right)^{-1}. \tag{11}$$

We reached an average final source density after all photometry and shape cuts of 13.1 arcmin^{-2} (see Table 4) for the clusters with $1.2 \lesssim z \lesssim 1.7$. We note that this is substantially lower than the values shown in Fig. 8 because we now included the signal-to-noise ratio and lensing cuts[19].

Table 4. Number densities of selected source galaxies measured in the cluster fields.

Cluster name	$n_{\rm gal}$ [arcmin^{-2}]
SPT-CL J0156−5541	14.3
SPT-CL J0205−5829	12.7
SPT-CL J0313−5334	20.1
SPT-CL J0459−4947	10.7
SPT-CL J0607−4448	13.3
SPT-CL J0640−5113	10.2
SPT-CL J2040−4451	11.2
SPT-CL J2341−5724	12.6
Average	13.1
SPT-CL J0646−6236	26.9

Notes. We apply the source selection as described in Sect. 4.2.1 including only sources that pass the lensing selections and have a signal-to-noise ratio $S/N_{\rm flux,606} > 10$, leading to lower numbers compared to Fig. 8. The cluster SPT-CL J0646−6236 is listed separately because we applied a different selection strategy for this cluster (see Appendix E).

6. Weak lensing results

Our pipeline used to obtain weak lensing constraints largely follows S18 and S21 to which we refer the reader for more detailed descriptions.

[17] With the aim to potentially reduce statistical uncertainties in our analysis, we also computed results using an alternative signal-to-noise ratio cut of $S/N_{\rm flux} > 7$. While this did increase the source number density, we found that it only marginally changes the constraints of our SZ–mass scaling relation analysis, likely due to the low shape weights and the increased photometric scatter of the additional faint galaxies. In an interest to keep our study consistent with previous studies, for example, by S21, we chose to use the cut of $S/N_{\rm flux} > 10$.

[18] We employ the $\log_{10} S/N_{\rm flux}$-dependent fits instead of the magnitude-dependent fits provided by S18 in order to account for the slightly higher depth of some of our stacks and the significant dependence of the measurement noise on $\log_{10} S/N_{\rm flux}$. For comparison, the dependence of $\sigma_{{\rm int},V/I}$ on $\log_{10} S/N_{\rm flux}$ is weak in the regime covered by most of our sources.

[19] While the number density is affected by a change of the signal-to-noise ratio cut, we found that the average geometric lensing efficiency is not sensitive to it. The change is smaller than ∼1% comparing the results with or without the cut at $S/N_{\rm flux,606} > 10$.

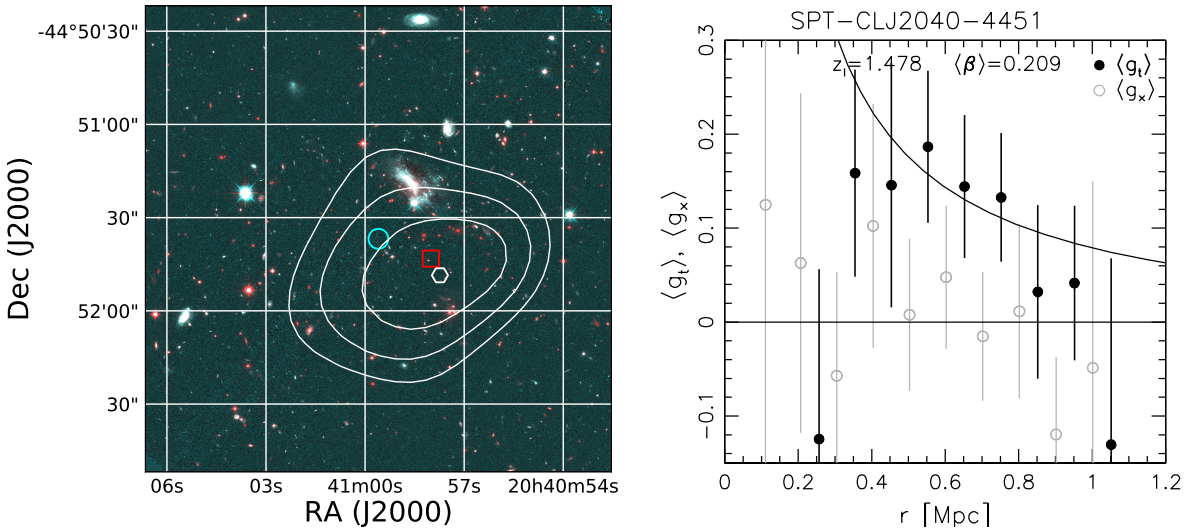


Fig. 9. Weak lensing results for SPT-CL J2040−4451. *Left*: signal-to-noise ratio contours of the mass reconstruction, starting at 2σ in steps of 0.5σ, overlaid on a $F606W/F814W/F110W$ colour image ($2\farcm5 \times 2\farcm5$ cutout). The peak in the S/N map is indicated by the hexagon (excluding potential spurious secondary peaks near the edge of the field of view). The cyan circle and the red square show the locations of the SZ peak and the X-ray centroid, respectively. *Right*: reduced shear profile around the X-ray centre, including the tangential component (solid black circles including the best-fitting NFW model) and the cross component (open grey circles), which has been shifted along the x-axis for clarity. The results for the other clusters are shown in Appendix G.

Table 5. Constraints on the peaks in the mass reconstruction signal-to-noise ratio maps including their locations (α, δ), positional uncertainties $(\Delta\alpha, \Delta\delta)$ as estimated by bootstrapping the galaxy catalogue (we note that this underestimates the true uncertainty as found by Sommer et al. 2022), and their peak signal-to-noise ratios $(S/N)_{\rm peak}$.

Cluster	α [deg J2000]	δ [deg J2000]	$\Delta\alpha$ [arcsec]	$\Delta\delta$ [arcsec]	$\Delta\alpha$ [kpc]	$\Delta\delta$ [kpc]	$(S/N)_{\rm peak}$
SPT-CL J0156−5541	29.04676	−55.69426	9.1	4.8	76	41	2.0
SPT-CL J0459−4947	74.92771	−49.77739	8.1	9.6	69	81	2.2
SPT-CL J0640−5113	100.08319	−51.21488	6.4	5.5	53	46	2.6
SPT-CL J0646-6236	101.63130	−62.62127	1.5	2.2	12	18	5.5
SPT-CL J2040−4451	310.24056	−44.86349	4.6	3.7	39	31	3.4
SPT-CL J2341−5724	355.34768	−57.41418	7.7	8.1	64	68	2.2

Notes. We excluded unreliable peaks close to the edge of the field of view (compare Figs. 9 and G.2–G.4).

6.1. Mass reconstructions

The weak lensing convergence κ and shear γ are both second-order derivatives of the lensing potential (e.g. Bartelmann & Schneider 2001). As a result, it is possible to reconstruct the convergence distribution from the shear field up to a constant, which is also known as the mass-sheet degeneracy (Kaiser & Squires 1993; Schneider & Seitz 1995). Here, we employed the Wiener-filtered reconstruction algorithm from McInnes et al. (2009) and Simon et al. (2009), where we fixed the mass-sheet degeneracy by setting the average convergence inside the observed fields to zero. We computed S/N maps of the reconstruction, where the noise map is computed as the root mean square (rms) image of the κ field reconstructions of 500 noise shear fields, which were created by randomising the ellipticity phases in the real source catalogue. Given the limited field of view and our choice to set the average convergence to zero, we expect to slightly underestimate the true S/N levels (S21).

The obtained S/N reconstructions are shown as contours in the left panels of Figs. 9 and G.2–G.4. SPT-CL J0646−6236 and SPT-CL J2040−4451 show clear peaks in the mass reconstruction signal-to-noise ratio maps with $S/N_{\rm peak} > 3$ (see Table 5 for details). We find tentative counterparts to the clusters with $2 < S/N_{\rm peak} < 3$ for SPT-CL J0156−5541, SPT-CL J0459−4947, SPT-CL J0640−5113, and SPT-CL J2341−5724. The other clusters either show no significant peak in their corresponding mass reconstruction S/N maps, or only a peak close to the edge of the field of view, which is less reliable and likely spurious. While some of the clusters remained undetected in the reconstructed mass maps, we note that these maps are only for illustration purposes. We still took the tangential reduced shear profiles of all clusters in our sample into account for the likelihood analysis (see Sect. 7).

6.2. Fits to the tangential reduced shear profiles

When measuring the reduced shear signal with respect to the centre of a mass concentration such as a cluster, it is helpful to distinguish the tangential component $g_{\rm t}$ and the cross component $g_\times$:

$$g_{\rm t} = -g_1 \cos 2\phi - g_2 \sin 2\phi,$$
$$g_\times = +g_1 \sin 2\phi - g_2 \cos 2\phi, \quad (12)$$

where ϕ indicates the azimuthal angle with respect to the centre. We computed the tangential component ('t') and the cross

Table 6. Weak lensing mass constraints derived from the fit of the tangential reduced shear profiles around the X-ray centres using spherical NFW models assuming the $c(M)$ relation from Diemer & Kravtsov (2015) with updated parameters from Diemer & Joyce (2019) for two different over-densities $\Delta \in \{200c, 500c\}$.

Cluster	$M_{200c}^{\mathrm{biased,ML}}$ $[10^{14}\,M_\odot]$	$\hat{b}_{200c,\mathrm{WL}}$	$\sigma(\ln b_{200c,\mathrm{WL}})$	$M_{500c}^{\mathrm{biased,ML}}$ $[10^{14}\,M_\odot]$	$\hat{b}_{500c,\mathrm{WL}}$	$\sigma(\ln b_{500c,\mathrm{WL}})$
SPT-CL J0156−5541	$4.5^{+3.5}_{-2.9} \pm 1.0 \pm 0.5$	0.88 ± 0.02	0.35 ± 0.03	$3.1^{+2.5}_{-2.1} \pm 0.7 \pm 0.3$	0.92 ± 0.03	0.28 ± 0.05
SPT-CL J0205−5829	$0.1^{+2.8}_{-2.4} \pm 0.5 \pm 0.0$	0.76 ± 0.03	0.41 ± 0.05	$0.1^{+1.9}_{-1.6} \pm 0.3 \pm 0.0$	0.79 ± 0.03	0.41 ± 0.04
SPT-CL J0313−5334	$2.8^{+3.3}_{-2.4} \pm 1.1 \pm 0.3$	0.86 ± 0.03	0.44 ± 0.04	$1.9^{+2.4}_{-1.7} \pm 0.8 \pm 0.2$	0.83 ± 0.03	0.37 ± 0.05
SPT-CL J0459−4947	$4.4^{+6.8}_{-4.4} \pm 1.5 \pm 0.5$	0.85 ± 0.05	0.51 ± 0.08	$3.0^{+5.0}_{-3.0} \pm 1.1 \pm 0.4$	0.79 ± 0.05	0.43 ± 0.10
SPT-CL J0607−4448	$0.6^{+3.4}_{-2.2} \pm 0.7 \pm 0.1$	0.86 ± 0.03	0.46 ± 0.04	$0.4^{+2.4}_{-1.5} \pm 0.4 \pm 0.0$	0.82 ± 0.04	0.45 ± 0.06
SPT-CL J0640−5113	$6.6^{+5.1}_{-4.5} \pm 1.1 \pm 0.7$	0.93 ± 0.03	0.27 ± 0.08	$4.6^{+3.8}_{-3.2} \pm 0.8 \pm 0.5$	0.85 ± 0.04	0.37 ± 0.05
SPT-CL J2040−4451	$16.4^{+5.8}_{-5.7} \pm 1.6 \pm 1.9$	0.89 ± 0.04	0.44 ± 0.06	$12.0^{+4.5}_{-4.4} \pm 1.3 \pm 1.4$	0.74 ± 0.04	0.48 ± 0.06
SPT-CL J2341−5724	$5.7^{+3.9}_{-3.5} \pm 1.1 \pm 0.6$	0.88 ± 0.03	0.35 ± 0.04	$4.0^{+2.9}_{-2.5} \pm 0.8 \pm 0.4$	0.87 ± 0.03	0.25 ± 0.05

Notes. The maximum likelihood mass estimates $M_\Delta^{\mathrm{biased,ML}}$ are given in $10^{14}\,M_\odot$, where errors correspond to statistical 68% uncertainties from shape noise (asymmetric errors), followed by uncorrelated large-scale structure projections, the calibration of the U_{HIGH} band, and variations in the redshift distribution between different lines of sight (for systematic uncertainties see Table 3). Statistical corrections for mass modelling biases have not yet been applied for $M_\Delta^{\mathrm{biased,ML}}$. They are characterised by $\hat{b}_{\Delta,\mathrm{WL}} = \exp\left[\langle \ln b_{\Delta,\mathrm{WL}} \rangle\right]$ and $\sigma(\ln b_{\Delta,\mathrm{WL}})$, which relate to the mean and the width of the estimated mass bias distribution (see Sect. 6.3).

Table 7. As Table 6, but for the analysis centring the shear profiles around the SZ centres.

Cluster	$M_{200c}^{\mathrm{biased,ML}}$ $[10^{14}\,M_\odot]$	$\hat{b}_{200c,\mathrm{WL}}$	$\sigma(\ln b_{200c,\mathrm{WL}})$	$M_{500c}^{\mathrm{biased,ML}}$ $[10^{14}\,M_\odot]$	$\hat{b}_{500c,\mathrm{WL}}$	$\sigma(\ln b_{500c,\mathrm{WL}})$
SPT-CL J0156−5541	$3.9^{+3.4}_{-2.8} \pm 1.1 \pm 0.4$	0.74 ± 0.02	0.41 ± 0.04	$2.7^{+2.5}_{-1.9} \pm 0.8 \pm 0.3$	0.73 ± 0.02	0.36 ± 0.04
SPT-CL J0205−5829	$0.3^{+3.1}_{-2.3} \pm 0.5 \pm 0.0$	0.76 ± 0.03	0.38 ± 0.05	$0.2^{+2.2}_{-1.6} \pm 0.4 \pm 0.0$	0.72 ± 0.03	0.40 ± 0.05
SPT-CL J0313−5334	$4.3^{+3.8}_{-3.1} \pm 1.2 \pm 0.4$	0.80 ± 0.03	0.33 ± 0.06	$3.0^{+2.8}_{-2.2} \pm 0.8 \pm 0.3$	0.76 ± 0.03	0.34 ± 0.05
SPT-CL J0459−4947	$6.9^{+7.0}_{-5.7} \pm 1.7 \pm 0.8$	0.83 ± 0.07	0.49 ± 0.12	$4.9^{+5.3}_{-4.1} \pm 1.2 \pm 0.6$	0.67 ± 0.06	0.65 ± 0.09
SPT-CL J0607−4448	$2.4^{+4.0}_{-2.5} \pm 1.0 \pm 0.3$	0.76 ± 0.04	0.23 ± 0.11	$1.7^{+2.9}_{-1.7} \pm 0.7 \pm 0.2$	0.72 ± 0.03	0.34 ± 0.07
SPT-CL J0640−5113	$3.4^{+5.1}_{-3.4} \pm 1.0 \pm 0.4$	0.66 ± 0.03	0.56 ± 0.05	$2.3^{+3.7}_{-2.3} \pm 0.7 \pm 0.3$	0.70 ± 0.03	0.36 ± 0.07
SPT-CL J0646−6236	$12.1^{+3.3}_{-3.3} \pm 1.3 \pm 1.1$	0.78 ± 0.02	0.41 ± 0.03	$8.6^{+2.4}_{-2.5} \pm 0.9 \pm 0.8$	0.78 ± 0.02	0.39 ± 0.03
SPT-CL J2040−4451	$15.7^{+5.8}_{-5.8} \pm 1.5 \pm 1.8$	0.77 ± 0.04	0.40 ± 0.07	$11.5^{+4.5}_{-4.4} \pm 1.2 \pm 1.3$	0.71 ± 0.04	0.47 ± 0.07
SPT-CL J2341−5724	$3.8^{+3.8}_{-3.0} \pm 1.0 \pm 0.4$	0.71 ± 0.03	0.46 ± 0.04	$2.6^{+2.7}_{-2.1} \pm 0.7 \pm 0.3$	0.70 ± 0.03	0.41 ± 0.05

component ('×') of the reduced shear in linear bins of width 100 kpc (see the right panels of Figs. 9 and G.2–G.4) around both the X-ray centroids (when available) and the SZ centres of the targeted clusters. We fitted the tangential reduced shear profiles using spherical Navarro-Frenk-White (NFW, Navarro et al. 1997) models following Wright & Brainerd (2000), employing the concentration–mass relation from Diemer & Kravtsov (2015) with updated parameters from Diemer & Joyce (2019). When deriving mass constraints, we excluded the cluster cores ($r < 500$ kpc), since the inclusion of smaller scales would both increase the intrinsic scatter and systematic uncertainties related to the mass modelling (see e.g. Sommer et al. 2022; Grandis et al. 2021). We note that weak lensing mass constraints can also be derived this way for clusters, which were undetected in the reconstructed mass maps (see Sect. 6.1). We summarise the resulting fit constraints in Tables 6 and 7. For clusters with both X-ray and SZ centres, we regarded the X-ray-centred analysis as our primary result given the smaller expected mass modelling biases (see Sect. 6.3).

6.3. Estimation of the weak lensing mass modelling bias

Weak lensing mass estimates can suffer from systematic biases caused by deviations of the cluster from an NFW profile, triaxial or complex mass distributions (e.g. due to mergers), both correlated and uncorrelated large-scale structure, and miscentring of the fitted shear profile. The measured weak lensing mass $M_{\Delta,\mathrm{WL}}$ at an overdensity Δ is typically smaller than the true mass of the halo $M_{\Delta,\mathrm{halo}}$ by a factor

$$b_{\Delta,\mathrm{WL}} = \frac{M_{\Delta,\mathrm{WL}}}{M_{\Delta,\mathrm{halo}}}. \tag{13}$$

This bias also depends on the specific properties of the sample such as mass and redshift and the measurement setup regarding the employed concentration–mass relation and radial fitting range.

In this study, we obtained an estimate for the weak lensing mass bias distribution following the method described by Sommer et al. (2022). They showed that the traditional, simplifying assumption of a log-normal bias distribution according to

$$\ln\left(\frac{M_{\Delta,\mathrm{WL}}}{M_{\Delta,\mathrm{halo}}}\right) \sim \mathcal{N}(\mu, \sigma^2) \tag{14}$$

is a suitable choice in the absence of miscentring. Here, $\mathcal{N}(\mu, \sigma^2)$ is the log-normal distribution with expectation value $\mu = \langle \ln b_{\Delta,\mathrm{WL}} \rangle$ and variance σ^2. The expectation value μ in log-space translates to a measure of the bias in linear space via the estimator

$$\hat{b}_{\Delta,\mathrm{WL}} = \exp[\langle \ln b_{\Delta,\mathrm{WL}} \rangle]. \tag{15}$$

Following Sommer et al. (2022), we used snapshots of the Millennium XXL simulations (MXXL, Angulo et al. 2012) at redshift $z = 1$ to estimate the weak lensing mass bias distribution. We obtained an estimate for each cluster individually by incorporating the given SZ mass and uncertainties of the measured radial tangential shear profile as input information. First, we used all haloes in the MXXL simulations with a halo mass within 2σ of the SZ mass of the respective cluster (see Table 1). Their mass distributions were projected along three mutually orthogonal axes increasing the effective sample size. We note that we did include a line of sight integration length of $200\,h^{-1}$ Mpc and not the full line of sight. Consequently, this method takes into account only correlated but not uncorrelated large-scale structure. However, integration along a line of sight twice as long changes the mean results only marginally (Becker & Kravtsov 2011). The projected mass distributions of the massive haloes served to calculate the shear and convergence fields on a grid with four arcsecond resolution. We converted the shear to the reduced shear using the same average lensing efficiency as in the respective cluster observations. This reduced shear field was azimuthally averaged in the same range and bins as in the cluster analysis to obtain a reduced shear profile. As the centre, we used either the 3D halo centre (most bound particle) or an offset centre drawn from an empirical miscentring distribution. We added noise to the reduced shear profile in each radial bin matching the corresponding uncertainties of the actual cluster tangential reduced shear estimates. We then obtained a weak lensing mass estimate by fitting the tangential reduced shear profile with an NFW profile, analogous to the analysis in our actual cluster observations. Subsequently, the comparison of the obtained weak lensing mass with the true halo mass provided the estimate for the weak lensing mass bias distribution for our specific setup. The full probability distribution $P(M_{\Delta,\mathrm{WL}}|M_{\Delta,\mathrm{halo}})$ was modelled with the help of Bayesian statistics as described in Sommer et al. (2022), where the SZ-derived mass estimates ($M_{200\mathrm{c,SZ}}$ and $M_{500\mathrm{c,SZ}}$) from B19 served as a prior for the mass estimation. Thus, we did not take into account any mass dependence of the bias other than using the SPT-SZ masses as a prior.

We incorporated miscentring into the estimation of the weak lensing mass bias distribution by applying an offset in a random direction before obtaining the reduced shear profile and subsequently fitting the masses. The offset was drawn from a miscentring distribution derived from the Magneticum Pathfinder Simulation (Dolag et al. 2016) measuring the offset between X-ray (or SZ) peaks from the simulation as a proxy for the centre and the position of the most bound particle (see S21, for a detailed description). We note that the log-normal assumption does not hold anymore for the weak lensing mass bias distribution in case of miscentring. However, the deviation is at the 3−5% level regarding the true mass. Therefore, we could still obtain meaningful estimates of the mean bias and scatter from a log-normal fit.

We found that the weak lensing mass bias distribution is nearly independent of mass within the 2σ bounds of the given SZ-derived mass of the respective clusters. Thus, we averaged the bias and scatter over this mass range and report the results in Tables 6 and 7. We found that the clusters exhibit a weak lensing mass bias $\hat{b}_{\Delta,\mathrm{WL}}$ between 0.74 and 0.92 in the presence of miscentring (using X-ray centres) with a scatter σ between 0.25 and 0.48 regarding the weak lensing masses $M_{500\mathrm{c}}$. On average, the masses computed with the X-ray centres are slightly less biased with a slightly smaller scatter when compared to the masses computed with the SZ centre (see Tables 6 and 7). This is a result of the on average smaller offsets of the X-ray miscentring distribution compared to the offsets of the SZ miscentring distribution (S21).

We note that we have derived these estimates from the MXXL snapshot at $z = 1$. S21 report weak lensing mass bias estimates, which are interpolated between results at $z = 0.25$ and $z = 1$ according to the given cluster redshift. We found that the results using the $z = 0.25$ snapshot are very similar to those at $z = 1$. This suggests that there is no strong redshift evolution, and we decide to report the results from the $z = 1$ snapshot, closest to the redshift range of our sample.

7. Constraints on the SPT observable-mass scaling relation

In this section, we present how we combined the weak lensing mass measurements of our nine high-redshift SPT clusters with results for clusters at lower redshifts, namely weak lensing mass measurements of 19 SPT clusters with redshifts $0.29 \leq z \leq 0.61$ based on Magellan/Megacam observations (D19, sample Megacam-19) and of 30 SPT clusters with redshifts $0.58 \leq z \leq 1.13$ based on HST observations (S21, sample HST-30). We used this sample of in total 58 SPT clusters (we refer to it as HST-39 + Megacam-19) with weak lensing mass measurements to constrain the SPT observable-mass scaling relation. Thereby, we extended the previous studies (S18; D19; B19; S21) out to redshifts of up to $z = 1.7$.

7.1. Likelihood formalism for the observable-mass scaling relation

In this section, we briefly summarise our likelihood formalism. It follows the definitions in D19, B19, and S21, which we refer the reader to for further details.

The SPT observable-mass scaling relation is based on the measured detection significance ξ as a mass proxy. Its relation to the unbiased detection significance ζ can be quantified from simulations (Vanderlinde et al. 2010) or analytically (Zubeldia et al. 2021) and exhibits a scatter given by a Gaussian of unit width

$$P(\xi|\zeta) = \mathcal{N}\left(\sqrt{\zeta^2 + 3}, 1\right). \tag{16}$$

Further following B19 and S21, we define the scaling relation between the unbiased detection significance ζ and the mass $M_{500\mathrm{c}}$ as a power-law in mass and the dimensionless Hubble parameter $E(z) \equiv H(z)/H_0$:

$$\langle \ln \zeta \rangle = \ln\left[\gamma_{\mathrm{field}} A_{\mathrm{SZ}} \left(\frac{M_{500\mathrm{c}}}{3 \times 10^{14}\, M_\odot/h}\right)^{B_{\mathrm{SZ}}} \left(\frac{E(z)}{E(0.6)}\right)^{C_{\mathrm{SZ}}}\right], \tag{17}$$

where A_{SZ}, B_{SZ}, and C_{SZ} parametrise the normalisation, mass slope, and redshift evolution, respectively, and γ_{field} characterises the effective depth of the individual SPT fields. Since we want to constrain this relation with the help of weak lensing mass measurements, we additionally need to consider the relation between lensing mass and true mass (see Eq. (13)). We set $\Delta = 500c$ and omit this notation in this section for readability, so that the relation reads

$$\ln\langle M_{\mathrm{WL}} \rangle = \ln b_{\mathrm{WL}} + \ln M. \tag{18}$$

Combining both relations, we therefore obtain the joint relation

$$P\left(\begin{bmatrix} \ln \zeta \\ \ln M_{\mathrm{WL}} \end{bmatrix} | M, z\right) = \mathcal{N}\left(\begin{bmatrix} \langle \ln \zeta \rangle(M,z) \\ \langle \ln M_{\mathrm{WL}} \rangle(M,z) \end{bmatrix}, \Sigma_{\zeta - M_{\mathrm{WL}}}\right), \tag{19}$$

where the covariance matrix $\Sigma_{\zeta-M_{\rm WL}}$ summarises how the logarithms of the observables ζ and $M_{\rm WL}$ scatter. It is given by

$$\Sigma_{\zeta-M_{\rm WL}} = \begin{pmatrix} \sigma^2_{\ln\zeta} & \rho_{\rm SZ-WL}\sigma_{\ln\zeta}\sigma_{\ln M_{\rm WL}} \\ \rho_{\rm SZ-WL}\sigma_{\ln\zeta}\sigma_{\ln M_{\rm WL}} & \sigma^2_{\ln M_{\rm WL}} \end{pmatrix}. \quad (20)$$

The quantities $\sigma_{\ln\zeta}$ and $\sigma_{\ln M_{\rm WL}}$ denote the widths of the normal distributions, which characterise the intrinsic scatter in $\ln\zeta$ and $\ln M_{\rm WL}$, respectively. They are assumed to be independent of redshift and mass. Correlated scatter between the SZ and the weak lensing observable is described by the correlation coefficient $\rho_{\rm SZ-WL}$.

We note that the weak lensing observable is not the mass $M_{\rm WL}$, but rather the tangential reduced shear $g_{\rm t}$. Therefore, the likelihood for each cluster reads

$$P(g_{\rm t}|\xi, z, \boldsymbol{p}) = \iiint {\rm d}M\,{\rm d}\zeta\,{\rm d}M_{\rm WL} \times [P(\xi|\zeta)P(g_{\rm t}|M_{\rm WL}, N_{\rm source}(z), \boldsymbol{p}) \times P(\zeta, M_{\rm WL}|M, z, \boldsymbol{p})P(M|z, \boldsymbol{p})]. \quad (21)$$

Here, $P(\zeta, M_{\rm WL}|M, z, \boldsymbol{p})$ is the joint scaling relation introduced in Eq. (19) and $P(M|z, \boldsymbol{p})$ denotes the halo mass function by Tinker et al. (2008). It represents a weighting required to account for Eddington bias. The vector $\boldsymbol{p}$ summarises the astrophysical and cosmological modelling parameters. Furthermore, the source redshift distribution is given by $N_{\rm source}(z)$ and the terms $P(\xi|\zeta)$ and $P(g_{\rm t}|M_{\rm WL}, N_{\rm source}(z), \boldsymbol{p})$ contain information about the intrinsic scatter and observational uncertainties in the observables[20]. Finally, the total log-likelihood corresponds to the sum of logarithms of the individual cluster likelihoods

$$\ln\mathcal{L} = \sum_{i=1}^{N_{\rm cl}}\sum_{j=1}^{N_{\rm bin}} \ln P(g_{{\rm t},ij}|\xi_{ij}, z_{ij}, \boldsymbol{p}), \quad (22)$$

where $N_{\rm cl} = 58$ is the total number of clusters considered to obtain constraints on the SPT observable-mass scaling relation and $N_{\rm bin}$ is the number of radial bins for the reduced shear profiles. We note that we naturally accounted for the selection function of the sample because we applied the established likelihood formalism only to the clusters from the SPT-SZ survey. Furthermore, the subsamples of clusters with weak lensing measurements were assembled randomly, independent of their lensing signal, so that the likelihood function is complete and does not suffer from biases due to weak lensing selections (D19; B19). In particular, this means that we also included the clusters that were not detected with a peak in the mass maps (see Sect. 6.1), because we would otherwise have introduced unwanted selection effects.

We cannot constrain all parameters in this relation equally well with the current weak lensing mass measurements. In particular, our data set does not allow for meaningful constraints for $B_{\rm SZ}$ and $\sigma_{\ln\zeta}$ (S21). Thus, we introduced the following priors. Regarding the slope parameter, we used a Gaussian prior $B_{\rm SZ} \sim \mathcal{N}(1.53, 0.1^2)$, which is motivated by the cosmological study in B19. We assumed $\sigma_{\ln\zeta} \sim \mathcal{N}(0.13, 0.13^2)$ as used by de Haan et al. (2016) and derived based on mock observations of hydrodynamic simulations from Le Brun et al. (2014). Additionally, we implemented the weak lensing mass modelling bias and corresponding scatter obtained in Sect. 6.3 and adopted a flat prior for the correlation coefficient, that is $\rho_{\rm SZ-WL} \in [-1, 1]$.

We conducted the likelihood analysis with an updated version of the pipeline used in B19 and S21, which is embedded in the `COSMOSIS` framework (Zuntz et al. 2015) and where the likelihood is explored with the MULTINEST sampler (Feroz et al. 2009). The full, updated pipeline will be made available along with a future publication by Bocquet et al. (in prep.).

We tested the likelihood machinery with mock cluster data. We simulated an SPT cluster catalogue with SZ detection significances and redshifts. We chose a number density and shape noise resembling the optical observations and implement an average source redshift distribution to simulate weak lensing cluster observations. These served as a basis to generate mock shear profiles, which we used as input for the likelihood analysis. Running the analysis on these mock data, we found that the resulting constraints on the scaling relation meet the expectation, thereby providing a valuable consistency check of our pipeline.

7.2. Redshift evolution of the ζ–mass relation

We applied the likelihood setup to our full cluster sample of 58 clusters with weak lensing mass measurements to constrain the ζ–mass relation. We present our results in Table 8. With our analysis, we constrained the scaling relation parameters $A_{\rm SZ} = 1.71 \pm 0.19$ and $C_{\rm SZ} = 1.34 \pm 1.00$, while the parameter $B_{\rm SZ}$ is dominated by the prior. Figure 10 displays the redshift evolution of the scaling relation, now for the first time extending out to redshifts up to $z \sim 1.7$ (red band, result of the fiducial analysis). For comparison, we show the constraints from S21 based on the HST-30 + Megacam-19 samples in blue, demonstrating that our findings in this study are fully consistent with these previous results. This was expected because we added only nine clusters to the previously used sample. In addition, our clusters are at the high-redshift end and therefore the statistical uncertainties are larger compared to clusters at lower and intermediate redshifts. Furthermore, the diagonally hatched region represents the scaling relation constraints from B19, who analysed weak lensing measurements from the Megacam-19 sample and 13 clusters from S18 in combination with X-ray measurements and cluster abundance information. They marginalised over cosmological parameters for a flat $\nu\Lambda$CDM cosmology. For comparison, we also show results computed for a joint analysis of *Planck* primary CMB anisotropies (TT,TE,EE+low-E, Planck Collaboration V 2020) and the SPT cluster abundance as the vertically hatched region. Again, this includes a marginalisation over cosmological parameters assuming a flat $\nu\Lambda$CDM cosmology. This analysis does not incorporate any weak lensing mass measurements.

As also found in S21, we observe an offset between the red and vertically hatched regions implying that the mass scale preferred from our analysis with the weak lensing data sets is lower than the mass scale that would be consistent with the *Planck* $\nu\Lambda$CDM cosmology by a factor of $0.72^{+0.09}_{-0.14}$ (at our pivot redshift of $z = 0.6$).

Analogous to S21, we wanted to check if the simple scaling relation model is applicable over the full, wide redshift range investigated here by performing a binned analysis, where the amplitude $A_{\rm SZ}$ is allowed to vary individually for each bin. Therefore, we added a bin of $1.2 < z < 1.7$ to the bins that were already used before in S21 (namely $0.25 < z < 0.5$, $0.5 < z < 0.88$, and $0.88 < z < 1.2$). We kept the redshift evolution parameter fixed to the value from the fiducial analysis at $C_{\rm SZ} = 1.34$. From Fig. 10, we can see that the results

[20] We note that we already included the shape noise of the tangential reduced shear profiles when we quantified the mass modelling bias in Sect. 6.3. However, the scatter $\sigma(\ln b_{\rm 500c,WL})$ of the weak lensing mass modelling bias changes only marginally for a noiseless estimation of the bias, so that our scaling relation results are not affected.

Table 8. Fit results for the parameters of the ζ–mass relation, analogously to table 12 in S21, now including the weak lensing measurements for the nine high-z SPT clusters from this work.

Parameter	Prior	HST-39 + Megacam-19		SPTcl ($\nu\Lambda$CDM)	*Planck* + SPTcl ($\nu\Lambda$CDM)
		fiducial	binned	(B19)	(no WL mass calibration)
$\ln A_{\rm SZ}$	Flat	1.71 ± 0.19	–	1.67 ± 0.16	$1.27^{+0.08}_{-0.15}$
$\ln A_{\rm SZ}(0.25 < z < 0.5)$	Flat	–	1.74 ± 0.23	–	–
$\ln A_{\rm SZ}(0.5 < z < 0.88)$	Flat	–	1.58 ± 0.31	–	–
$\ln A_{\rm SZ}(0.88 < z < 1.2)$	Flat	–	1.85 ± 0.43	–	–
$\ln A_{\rm SZ}(1.2 < z < 1.7)$	Flat	–	1.89 ± 0.81	–	–
$C_{\rm SZ}$	Flat/fixed	1.34 ± 1.00	1.34	$0.63^{+0.48}_{-0.30}$	$0.73^{+0.17}_{-0.19}$
Prior-dominated parameters in our analysis:					
$B_{\rm SZ}$	$\mathcal{N}(1.53, 0.1^2)$	1.56 ± 0.09	1.57 ± 0.10	1.53 ± 0.09	1.68 ± 0.08
$\sigma_{\ln\zeta}$	$\mathcal{N}(0.13, 0.13^2)$	$0.16^{+0.06}_{-0.13}$	$0.15^{+0.04}_{-0.13}$	0.17 ± 0.08	$0.16^{+0.07}_{-0.12}$

Notes. SPTcl ($\nu\Lambda$CDM) denotes the results from the B19 study, which combined SPT cluster counts with weak lensing and X-ray mass measurements. The results from the analysis denoted as *Planck* + SPTcl ($\nu\Lambda$CDM) are based on a combination of measurements from the *Planck* CMB anisotropies (TT,TE,EE+low-E, Planck Collaboration V 2020) and SPT cluster counts.

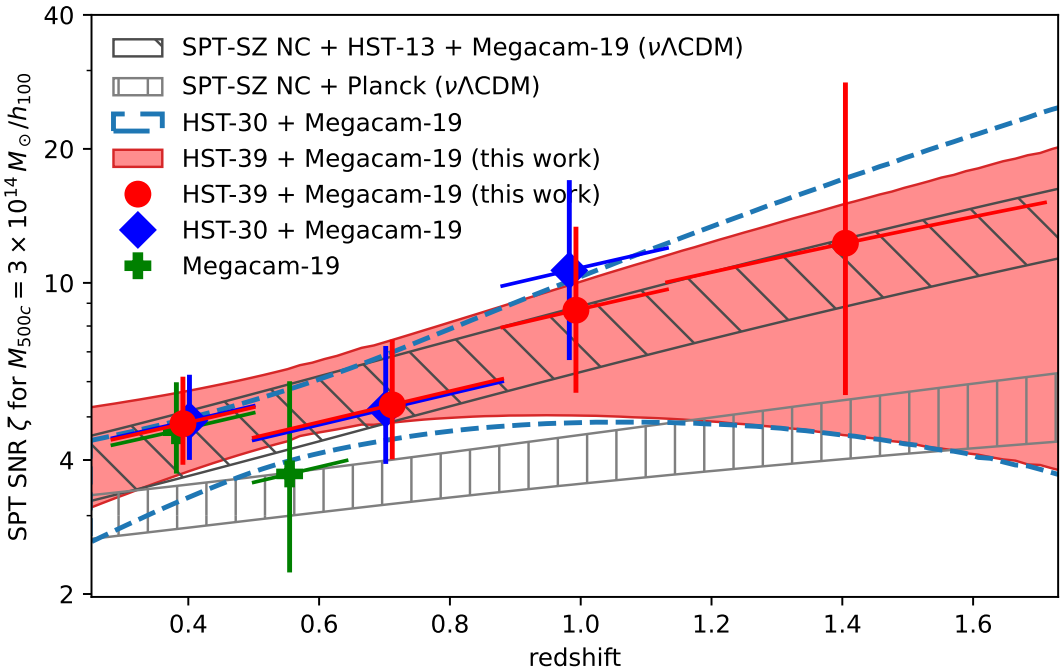


Fig. 10. Evolution of the unbiased SPT detection significance ζ at the pivot mass $3 \times 10^{14}\, M_\odot/h_{100}$ as a function of redshift. The red band indicates the main result of this work. The blue dashed curves indicate the corresponding 1σ band from the S21 analysis for comparison. The red and blue data points represent the corresponding binned analyses. They are placed in the centre of the bins. Horizontal error bars represent the bin widths. The redshift evolution parameter is fixed to $C_{\rm SZ} = 1.34$ for our binned analysis. The diagonally hatched and vertically hatched bands correspond to the relations from the B19 study and the SPT cluster counts in combination with a flat *Planck* $\nu\Lambda$CDM cosmology, respectively. The displayed uncertainties correspond to the 68% credible interval (bands for the full relation and error bars for the binned analysis).

in our new high-redshift bin are consistent with the scaling relation results from the full unbinned analysis. Additionally, we found that our results in the lower redshift bins are very similar to the results from the binned analysis in S21. This is also expected because the bins contain the same clusters except for SPT-CL J0646−6236, which was added to the third redshift bin and causes a small shift towards a higher cluster mass scale due to its large cluster mass.

8. Discussion

Weak lensing studies of galaxy clusters with ever higher redshifts face the increasingly difficult challenge to identify background galaxies carrying the lensing signal (e.g. Mo et al. 2016; Jee et al. 2017; Finner et al. 2020). In a simplified consideration, the signal-to-noise ratio of a lensing measurement scales with the product of the average geometric lensing efficient $\langle\beta\rangle$ and the square root of the source number density $\sqrt{n}$. For comparison purposes, we define the weak lensing sensitivity factor $\tau_{\rm WL}$ as the product of these two quantities: $\tau_{\rm WL} = \langle\beta\rangle\sqrt{n}$[21]. The average

[21] In principle, the signal-to-noise ratio of a lensing measurement also depends on other parameters such as cluster mass and fit range. However, the signal-to-noise ratio still scales with the weak lensing sensitivity factor $\tau_{\rm WL}$. We use it to represent how the source selection affects the lensing signal-to-noise ratio and compare this quantity for different studies.

geometric lensing efficiency is tied to the purity of the source sample, that is, the fraction of true background source galaxies. A higher purity is desirable as it also increases the average geometric lensing efficiency. At the same time, cuts to identify true background source galaxies should not be too rigorous as this might reduce the overall source density potentially at the cost of also excluding true background galaxies. Additionally, a lower source density is more subject to shot noise, consequently reducing the lensing signal-to-noise ratio.

Some previous weak lensing studies were conducted with HST/WFC3 in infrared bands to measure masses of clusters at redshifts $z \gtrsim 1.5$. They introduced varying techniques to select source galaxies for the lensing measurements. For their weak lensing analysis of cluster SpARCS1049+56 at redshift $z = 1.71$, Finner et al. (2020) selected sources via a magnitude cut of $H_{F160W} > 25.0$ mag and specific shape cuts aiming to remove galaxies with high uncertainty in the ellipticity measurement and objects that are too small or too elongated to be galaxies. Applying this method to their observations, they achieved a source density of 105 arcmin^{-2} and estimated an average geometric lensing efficiency of $\langle\beta\rangle = 0.107$. This translates into a signal-to-noise ratio of $\tau_{\rm WL} \sim 1.10$. Alternatively, Jee et al. (2017) performed a weak lensing study of clusters SPT-CL J2040−4451 and IDCS J1426+3508 at redshifts $z = 1.48$ and $z = 1.75$, respectively. They selected source galaxies requiring that they are bluer than the cluster red-sequence combined with a bright magnitude and shape measurement uncertainty cut. They obtained a source density of ∼240 arcmin^{-2} with an average lensing efficiency of $\langle\beta\rangle = 0.086$ and $\langle\beta\rangle = 0.120$ for IDCS J1426+3508 and SPT-CL J2040−4451, respectively. This corresponds to $\tau_{\rm WL} \sim 1.33$ and $\tau_{\rm WL} \sim 1.86$, respectively.

Mo et al. (2016) conducted a weak lensing study of IDCS J1426+3508 prior to Jee et al. (2017) using HST/ACS and HST/WFC3 data from the bands $F606W$, $F814W$, and $F160W$. They measured galaxy shapes with the $F606W$ imaging selecting source galaxies with $24.0 < V_{F606W} < 28.0$ (the latter is roughly the 10σ depth limit of their observations), $0\farcs27 < FWHM$[22] $< 0\farcs9$ (to exclude too large/small galaxies either because they are likely foreground galaxies or to avoid PSF problems, respectively), and $I_{F814W} - H_{F160W} < 3.0$ (to exclude cluster red-sequence galaxies). They achieved an average lensing efficiency of $\langle\beta\rangle = 0.086$ at a source density of 89 arcmin^{-2}, resulting in $\tau_{\rm WL} \sim 0.81$.

In conclusion, both NIR studies (Jee et al. 2017; Finner et al. 2020) achieved higher source densities, but lower average geometric lensing efficiencies than our study, which has an average source density of 13.1 arcmin^{-2} and an average geometric lensing efficiency of $\langle\beta\rangle = 0.244$, and thus $\tau_{\rm WL} \sim 0.88$. The studies by Jee et al. (2017) and Finner et al. (2020) owe the high signal-to-noise ratios mainly to very deep observations enabling high source densities. In contrast, our study focuses on a high purity as visible in Figs. 4 and 7, which display that we selected almost only high-z sources at $z \gtrsim 2$ with high lensing efficiency, while keeping the contamination of foreground, cluster, and near background galaxies low. This strategy resulted in an average lensing efficiency more than twice as high, and it helps to keep systematic uncertainties low for several reasons. First, excluding galaxies at the cluster redshift minimises uncertainties related to the correction for cluster member contamination. Second, galaxies in the near background are located in a regime where $\beta(z)$ is a steep function of z. Thus, systematic redshift uncertainties lead to larger systematic uncertainties in $\langle\beta\rangle$ than for the distant background galaxies selected in our approach. Finally, the efficient removal of foreground galaxies minimises the impact that catastrophic redshift outliers scattering between low and high redshifts have on the computation of $\langle\beta\rangle$ (see S18; R20). While we found that the uncertainties in the redshift distribution (R20 versus R15_fix comparison and variations between CANDELS/3D-HST fields) dominate the systematic error budget (see Table 3), our comparatively low number density introduced high statistical uncertainties, which (together with other statistical uncertainties) outweigh the systematic ones in our current analysis. However, we stress that our approach, which aims to limit systematic uncertainties by using data of moderate depth and applying a stringent background selection, could directly be applied to similar data sets obtained for larger cluster samples.

In combination with the considerable measurement uncertainties and the substantial expected intrinsic scatter (see Sect. 6.3), the best-fitting cluster mass estimates in our study are, therefore, expected to scatter significantly. This likely explains the relatively low mass estimate of SPT-CL J0205−5829, which remained undetected in the weak lensing data despite its high SZ-inferred mass, and the comparably high best-fitting mass estimate for SPT-CL J2040−4451. Still, we emphasise that our study aims to provide mass constraints that are accurate on average for our sample of nine galaxy clusters. Indeed, the median ratio of lensing mass to SZ mass from SPT is close to unity. We found a median ratio of bias corrected weak lensing mass to SZ mass $M_{\rm 500c,WL,corr}/M_{\rm 500c,SZ}$ of 1.048 ± 0.372 or 1.064 ± 0.462 using the weak lensing masses with X-ray centres (8 clusters) or SZ centres (9 clusters), respectively. We estimated the uncertainties via bootstrapping of the cluster sample.

Deviations between the X-ray or SZ mass and the lensing mass for individual clusters can, for instance, be caused by their different sensitivities to large-scale structure projections, triaxiality, and variations in density profiles. For example, we measured the highest weak lensing mass for the cluster SPT-CL J2040−4451, which is notably higher than the expectation from the SZ or X-ray mass estimates. However, taking the statistical uncertainties of the weak lensing, SZ and X-ray mass estimates into account, as well as the mass modelling bias and scatter, we found that the bias-corrected weak lensing mass agrees with its SZ (X-ray) mass estimate at the 1.2σ (1.2σ) level. We used the SZ mass listed in Table 1 and the X-ray mass $M_{\rm 500c,X\text{-}ray} = 3.10^{+0.79}_{-0.47} \times 10^{14}\,M_\odot$ from McDonald et al. (2017) as reference. We quantified the expected discrepancy between the SZ or X-ray mass and the weak lensing mass further in Appendix F. For this particular cluster, Jee et al. (2017) found a weak lensing mass of $M_{\rm 200c} = 8.6^{+1.7}_{-1.4} \times 10^{14}\,M_\odot$ (not corrected for mass modelling bias), which is also higher than the X-ray and SZ mass estimates of the cluster. Our weak lensing mass constraint of $M^{\rm biased,ML}_{\rm 200c} = 16.4^{+5.8}_{-5.7} \pm 1.6 \pm 1.9 \times 10^{14}\,M_\odot$ (for comparability with Jee et al. 2017 not corrected for mass modelling bias) deviates only by 1.2σ from the result by Jee et al. (2017), so that our results confirm the generally higher lensing mass for SPT-CL J2040−4451 (albeit with larger statistical uncertainties), suggesting potential line of sight effects. This conclusion is additionally supported by a high dynamical mass measurement (albeit with large uncertainties) by Bayliss et al. (2014).

Several differences in the analyses especially regarding the source selection strategies and fit ranges may explain the difference between the lensing masses from Jee et al. (2017) and our study. Jee et al. (2017) obtained their weak lensing mass constraint from HST/WFC3 imaging in $F105W$, $F140W$, and $F160W$. They fitted a spherical NFW profile assuming the concentration–mass relation of Dutton & Macciò (2014) and

[22] Measured with `Source Extractor`.

centred at their measured X-ray peak position (from *Chandra* data), including weak lensing sources outside of a minimum radius $r_{\rm min} = 26$ arcsec, corresponding to 218 kpc at the cluster redshift. The WFC3/IR observations by Jee et al. (2017) provide a full azimuthal coverage out to $r \lesssim 60$ arcsec, while we have $r \lesssim 90$ arcsec ($r \lesssim 72$ arcsec) around the SZ (X-ray) centre in our observations. We note that our inner fit limit ($r_{\rm min} = 500$ kpc) corresponds to an angular radius of 59 arcsec. Accordingly, our analysis primarily employs reduced shear measurements at larger scales compared to the analysis of Jee et al. (2017).

Additionally, we measured the weak lensing mass assuming the concentration–mass relation by Diemer & Kravtsov (2015) with updated parameters from Diemer & Joyce (2019), we centred the fit around the X-ray centroid from McDonald et al. (2017), which has a distance of 8.1 arcsec to the X-ray peak employed by Jee et al. (2017), and we used galaxies outside a minimum radius of $r_{\rm min} = 500$ kpc. We excluded any scales smaller than this to minimise systematic mass modelling uncertainties and the impact of a potential residual cluster member contamination (below the detection limit). Since the X-ray peak and centroid positions are relatively close to each other, it is reasonable to compare the weak lensing mass results without applying the statistical mass modelling correction.

The largest difference between the Jee et al. (2017) study and ours is the source selection strategy. Jee et al. (2017) based their work on imaging that is significantly deeper (with a limiting magnitude of $F140W \sim 28$ mag) than ours but limited to a smaller field of view. Their selection of background galaxies focussed on the exclusion of red-sequence galaxies (galaxies at $F105W-F140W < 0.5$ are selected) and resulted in a source number density of $\sim$240 arcmin^{-2} with a fraction of non-background sources (with $z \leq z_{\rm cluster}$) of approximately 45%. Additionally, the inclusion of scales at 218 kpc $< r <$ 500 kpc likely shrinks statistical uncertainties since the lensing signal is high in the inner regions of the cluster. This allowed them, in turn, to achieve small statistical uncertainties of their weak lensing mass constraints. However, the inclusion of such core regions usually increases the intrinsic scatter and mass modelling uncertainties (Sommer et al. 2022, see also Sect. 6.3). Our more strict selection strategy for the background galaxies based on magnitudes/colours from four bands is contaminated by 17 to 20% of non-background galaxies. The shallower data finally resulted in a source number density of 11.2 arcmin^{-2} for SPT-CL J2040−4451 so that our analysis exhibits substantially larger statistical uncertainties in the weak lensing mass constraints.

Jee et al. (2017) reported the detection of the cluster in their weak lensing mass map at the location $\alpha = 20^{\rm h}40^{\rm m}57\overset{\rm s}{.}85$ and $\delta = -44°51'42\overset{''}{.}4$ with 6σ significance. In our mass map, we detected a peak at 3.4σ, with a separation of 6.6 arcsec from the location in Jee et al. (2017). While this offset is slightly larger than our estimate of the positional uncertainty derived using bootstrapping (see Table 5), we note that Sommer et al. (2022) found that bootstrapping substantially underestimates the true uncertainty. The peaks from both studies are close to the X-ray centroid position from McDonald et al. (2017) so that they are overall in agreement. We also note that the peak in our weak lensing mass reconstruction for SPT-CL J2040−4451 closely coincides with the X-ray centroid. Accordingly, the shear profile is approximately centred on the position that maximises the lensing signal. This likely scatters the mass result high, especially if the statistical correction for mass modelling bias is applied.

While several studies undoubtedly confirmed SPT-CL J2040−4451 as one of the most massive high-redshift clusters known, our study shows that based on our weak lensing measurements, the SPT cluster population is less massive than what one would expect in a *Planck* ΛCDM cosmology, also at very high redshifts (see Sect. 7).

With our cluster sample and analysis, we enabled constraints on the SZ–mass scaling relation and its redshift evolution for the first time out to the redshift regime of $z > 1.2$. While lensing studies at lower redshifts can be calibrated more precisely and systematics are generally smaller, high-redshift clusters are particularly sensitive to probe, for example, models with massive neutrinos (Ichiki & Takada 2012), or deviations from standard ΛCDM expectations, such as early dark energy (Klypin et al. 2021). Therefore, exploring the high-redshift regime is worthwhile to understand the cosmological ΛCDM model and its possible extensions. Our study provides a first step towards constraints from clusters at redshifts $z > 1.2$.

9. Summary and conclusions

In this work, we studied the gravitational lensing signal of a sample of nine clusters with high redshifts $z \gtrsim 1.0$ in the SPT-SZ survey. They all exhibit a strong SZ signal with a high SZ detection significance $\xi > 6.0$. We obtained weak lensing mass constraints from shape measurements of galaxies with high-resolution HST/ACS imaging in the $F606W$ and $F814W$ bands. With the help of additional HST imaging using WFC3/IR in $F110W$ and VLT/FORS2 imaging in $U_{\rm HIGH}$, we applied a strategy to photometrically select background galaxies, even for clusters at such challenging high redshifts.

Using updated photometric redshift catalogues computed by R20 for the CANDELS/3D-HST fields as a reference, we estimated the source redshift distribution and calculated the average geometric lensing efficiency, applying the same selection criteria in the reference photometric redshift catalogues as in the cluster observations. We also added Gaussian noise to the reference catalogues if they were deeper than our cluster observations. We carefully investigated sources of systematic and statistical uncertainties for estimates of the average geometric lensing efficiency. We found consistent results in the HUDF field comparing our photometric measurements employing the algorithm LAMBDAR for adaptive aperture photometry and the S14 photometric measurements based on fixed aperture photometry. A comparison based on photometric and spectroscopic redshifts revealed a ∼3% difference in calculating the average geometric lensing efficiency, which we accounted for in the weak lensing analysis.

We reconstructed the projected cluster mass distributions based on the shear measurements of the selected galaxies. In the resulting mass maps, we detected two of the clusters with a peak at $S/N > 3$, four clusters with $S/N > 2$, and three clusters were not detected. We obtained weak lensing mass constraints by fitting the tangential reduced shear profiles with spherical NFW models, employing a fixed concentration–mass relation by Diemer & Kravtsov (2015) with updated parameters from Diemer & Joyce (2019). We reported statistical uncertainties from shape noise, uncorrelated large-scale structure projections, line of sight variations in the source redshift distribution, and uncertainties in the calibration of the $U_{\rm HIGH}$ band. We also estimated mass modelling biases using simulated clusters from the Millennium XXL simulations accounting for miscentring. Masses based on the X-ray centre were less biased ($\hat{b}_{\Delta c,{\rm WL}}$) and exhibited a slightly smaller scatter of the mass bias ($\sigma(\ln b_{\Delta c,{\rm WL}})$) than masses obtained using SZ centres. This is

consistent with findings in previous studies (e.g. Sommer et al. 2022; S21).

We carefully investigated the sources of systematic uncertainties in our study. The total systematic uncertainty of our weak lensing mass estimates amounts to 14.4% (16.7%) for the analyses centring the reduced shear profiles around the X-ray (SZ) centres. Here, the largest contribution (12.9%) comes from uncertainties related to the source selection and calibration of the source redshift distribution (see Table 3).

Our weak lensing mass constraints for SPT-CL J2040−4451 are higher, but still consistent with the earlier results obtained by Jee et al. (2017). Given the limited depth of our data and the high redshifts of the targeted clusters, our weak lensing mass estimates are relatively noisy. However, on average they are consistent with the SZ-inferred mass estimates from B19, which employ a weak lensing mass calibration based on data from Dietrich et al. (2019) and Schrabback et al. (2018). We found a median ratio of 1.048 ± 0.372 or 1.064 ± 0.462 using the weak lensing masses with X-ray centres (8 clusters) or SZ centres (9 clusters), respectively.

Finally, we used the obtained weak lensing mass measurements in a joint analysis with measurements for clusters at lower (D19) and intermediate (S21) redshifts to constrain the scaling relation between the debiased SPT cluster detection significance ζ and cluster mass, thereby expanding the previous studies by B19 and S21 to higher redshifts $z > 1.2$. Our binned analysis of the redshift evolution of the ζ–mass scaling relation revealed that the new highest redshift bin at $1.2 < z < 1.7$ is consistent with the scaling relation behaviour predicted from lower redshifts, albeit with large statistical uncertainties. Even with these large uncertainties at the high redshift end, our results for the full, unbinned analysis support previous findings where the mass scale preferred in an analysis including the weak lensing measurements is lower than the mass scale required for consistency with the *Planck* $\nu\Lambda$CDM cosmology presented in Planck Collaboration V (2020).

In our pilot study, we developed an approach for weak lensing mass measurements of high-z clusters with well-controlled systematics, thereby obtaining such measurements for a first significant sample of SZ-selected clusters at $z \gtrsim 1.2$. However, the small sample size and limited depth of the data imply large statistical uncertainties, which can be addressed by applying the approach to new weak lensing data of additional high-redshift clusters. While statistical uncertainties dominate in our study, there also remain notable systematic uncertainties, which need to be reduced in the future. Our study shows that the largest systematic uncertainty for lensing studies of high-redshift galaxy clusters arises from the calibration of the source redshift distribution. Here, surveys such as the planned *James Webb* Space Telescope Advanced Deep Extragalactic Survey[23] (JADES) will help to calibrate the redshift distributions especially for high-redshift clusters, which are observed with deep imaging data. This survey will provide imaging and spectroscopy to unprecedented depth and infer photometric and spectroscopic redshifts over an area of 236 arcmin^2 in the GOODS-South and GOODS-North fields. Additionally, direct calibration methods and those utilizing the stacked redshift probability distribution functions of galaxies already show promising results and need to be further explored to help reduce systematic uncertainties in the redshift calibration (e.g. Euclid Collaboration 2021). Furthermore, in-depth analyses of hydrodynamical simulations will help to better understand and reduce systematics due to the concentration–mass relation, the weak lensing mass modelling, and miscentring distribution uncertainties.

[23] https://pweb.cfa.harvard.edu/research/james-webb-space-telescope-advanced-deep-extragalactic-survey-jades

Acknowledgements. This research is based on observations made with the NASA/ESA *Hubble* Space Telescope obtained from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-26555. These observations are associated with GO programmes 12477, 13412, 14252, and 14677 (observations of the nine clusters targeted for this study), as well as 14043, 9425, 12062, and 13872 (archival data in the GOODS-South region). This work is based on observations taken by the 3D-HST Treasury Program (HST-GO-12177 and HST-GO-12328) with the NASA/ESA *Hubble* Space Telescope. This work made use of HDUV Data Release 1.0 data products (Oesch et al. 2018). The team members involved in the HDUV survey are: P. Oesch, M. Montes, N. Reddy, R. J. Bouwens, G. D. Illingworth, D. Magee, H. Atek, C. M. Carollo, A. Cibinel, M. Franx, B. Holden, I. Labbe, E. J. Nelson, C. C. Steidel, P. G. van Dokkum, L. Morselli, R. P. Naidu, S. Wilkins. This work is based on observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere under ESO programme 0100.A-0204(A). This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular, the institutions participating in the *Gaia* Multilateral Agreement. The Bonn group acknowledges support from the German Federal Ministry for Economic Affairs and Climate Action (BMWK) provided through DLR under projects 50OR1803, 50OR2002, 50OR2106, and 50QE2002, as well as support provided by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 415537506. H.Z., F.R., and D.S. are members of and received financial support from the International Max Planck Research School (IMPRS) for Astronomy and Astrophysics at the Universities of Bonn and Cologne. D.S. acknowledges support from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 776247. H.Ho. acknowledges support from Vici grant 639.043.512 financed by the Netherlands Organization for Scientic Research. AHW is supported by an European Research Council Consolidator Grant (No. 770935). Argonne National Laboratory's work was supported by the US Department of Energy, Office of High Energy Physics, under contract DE-AC02-06CH11357. This work was performed in the context of the South Pole Telescope scientific programme. S.P.T. is supported by the National Science Foundation through grants OPP-1852617. Partial support is also provided by the Kavli Institute of Cosmological Physics at the University of Chicago. The authors would like to thank Peter Schneider and the anonymous referee for useful comments, which helped to improve this manuscript. Data availability: The full, updated pipeline, which we used for the likelihood analysis in this work, will be made available along together with an upcoming publication by Bocquet et al. (in prep.). The data underlying this article will be shared upon reasonable request to the corresponding author.

References

Abazajian, K., Addison, G., Adshead, P., et al. 2019, ArXiv e-print [arXiv:1907.04473]
Allen, S. W., Evrard, A. E., & Mantz, A. B. 2011, ARA&A, 49, 409
Angulo, R. E., Springel, V., White, S. D. M., et al. 2012, MNRAS, 426, 2046
Bacon, R., Brinchmann, J., Richard, J., et al. 2015, A&A, 575, A75
Bartelmann, M., & Schneider, P. 2001, Phys. Rep., 340, 291
Bayliss, M. B., Ashby, M. L. N., Ruel, J., et al. 2014, ApJ, 794, 12
Becker, M. R., & Kravtsov, A. V. 2011, ApJ, 740, 25
Beckwith, S. V. W., Stiavelli, M., Koekemoer, A. M., et al. 2006, AJ, 132, 1729
Benítez, N. 2000, ApJ, 536, 571
Benson, B. A., Ade, P. A. R., Ahmed, Z., et al. 2014, SPIE Conf. Ser., 9153, 91531P
Bertin, E. 2011, ASP Conf. Ser., 442, 435
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bîrzan, L., Rafferty, D. A., Brüggen, M., & Intema, H. T. 2017, MNRAS, 471, 1766
Bleem, L. E., Stalder, B., de Haan, T., et al. 2015, ApJS, 216, 27
Bocquet, S., Dietrich, J. P., Schrabback, T., et al. 2019, ApJ, 878, 55
Bocquet, S., Heitmann, K., Habib, S., et al. 2020, ApJ, 901, 5
Bouwens, R. J., Illingworth, G. D., Oesch, P. A., et al. 2011, ApJ, 737, 90
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Brammer, G. B., van Dokkum, P. G., Franx, M., et al. 2012, ApJS, 200, 13
Brinchmann, J., Inami, H., Bacon, R., et al. 2017, A&A, 608, A3
Bulbul, E., Chiu, I. N., Mohr, J. J., et al. 2019, ApJ, 871, 50

Capasso, R., Saro, A., Mohr, J. J., et al. 2019, MNRAS, 482, 1043
Chiu, I., Mohr, J., McDonald, M., et al. 2016, MNRAS, 455, 258
Chiu, I., Mohr, J. J., McDonald, M., et al. 2018, MNRAS, 478, 3072
Chiu, I. N., Ghirardini, V., Liu, A., et al. 2022, A&A, 661, A11
Chu, A., Durret, F., & Márquez, I. 2021, A&A, 649, A42
de Haan, T., Benson, B. A., Bleem, L. E., et al. 2016, ApJ, 832, 95
DeMaio, T., Gonzalez, A. H., Zabludoff, A., et al. 2020, MNRAS, 491, 3751
Diemer, B., & Joyce, M. 2019, ApJ, 871, 168
Diemer, B., & Kravtsov, A. V. 2015, ApJ, 799, 108
Dietrich, J. P., Bocquet, S., Schrabback, T., et al. 2019, MNRAS, 483, 2871
Dolag, K., Komatsu, E., & Sunyaev, R. 2016, MNRAS, 463, 1797
Dutton, A. A., & Macciò, A. V. 2014, MNRAS, 441, 3359
Ellis, R. S., McLure, R. J., Dunlop, J. S., et al. 2013, ApJ, 763, L7
Erben, T., Van Waerbeke, L., Bertin, E., Mellier, Y., & Schneider, P. 2001, A&A, 366, 717
Erben, T., Schirmer, M., Dietrich, J. P., et al. 2005, Astron. Nachr., 326, 432
Euclid Collaboration (Ilbert, O., et al.) 2021, A&A, 647, A117
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Finner, K., James Jee, M., Webb, T., et al. 2020, ApJ, 893, 10
Gaia Collaboration (Brown, A. G. A., et al.) 2016a, A&A, 595, A2
Gaia Collaboration (Prusti, T., et al.) 2016b, A&A, 595, A1
Ghirardini, V., Bulbul, E., Kraft, R., et al. 2021, ApJ, 910, 14
Grandis, S., Bocquet, S., Mohr, J. J., Klein, M., & Dolag, K. 2021, MNRAS, 507, 5671
Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35
Haiman, Z., Mohr, J. J., & Holder, G. P. 2001, ApJ, 553, 545
Herbonnet, R., Sifón, C., Hoekstra, H., et al. 2020, MNRAS, 497, 4684
Hernández-Martín, B., Schrabback, T., Hoekstra, H., et al. 2020, A&A, 640, A117
High, F. W., Stubbs, C. W., Rest, A., Stalder, B., & Challis, P. 2009, AJ, 138, 110
Hilton, M., Sifón, C., Naess, S., et al. 2021, ApJS, 253, 3
Hlavacek-Larrondo, J., McDonald, M., Benson, B. A., et al. 2015, ApJ, 805, 35
Hoekstra, H., Franx, M., Kuijken, K., & Squires, G. 1998, ApJ, 504, 636
Ichiki, K., & Takada, M. 2012, Phys. Rev. D, 85, 063521
Inami, H., Bacon, R., Brinchmann, J., et al. 2017, A&A, 608, A2
Jarvis, M., & Jain, B. 2008, JCAP, 2008, 003
Jee, M. J., Ko, J., Perlmutter, S., et al. 2017, ApJ, 847, 117
John, S. 1982, Commun. Stat. – Theor. Meth., 11, 879
Kaiser, N., & Squires, G. 1993, ApJ, 404, 441
Kaiser, N., Squires, G., & Broadhurst, T. 1995, ApJ, 449, 460
Kettula, K., Giodini, S., van Uitert, E., et al. 2015, MNRAS, 451, 1460
Khullar, G., Bleem, L. E., Bayliss, M. B., et al. 2019, ApJ, 870, 7
Klein, M., Israel, H., Nagarajan, A., et al. 2019, MNRAS, 488, 1704
Klypin, A., Poulin, V., Prada, F., et al. 2021, MNRAS, 504, 769
Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, 197, 36
Koekemoer, A. M., Ellis, R. S., McLure, R. J., et al. 2013, ApJS, 209, 3
Krist, J. E., Hook, R. N., & Stoehr, F. 2011, SPIE Conf. Ser., 8127, 81270J
Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015, MNRAS, 454, 3500
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
Le Brun, A. M. C., McCarthy, I. G., Schaye, J., & Ponman, T. J. 2014, MNRAS, 441, 1270
Liu, A., Bulbul, E., Ghirardini, V., et al. 2022, A&A, 661, A2
LSST Science Collaboration (Abell, P. A., et al.) 2009, ArXiv e-prints [arXiv:0912.0201]
Luppino, G. A., & Kaiser, N. 1997, ApJ, 475, 20
Madhavacheril, M. S., Sifón, C., Battaglia, N., et al. 2020, ApJ, 903, L13
Mantz, A. B., Allen, S. W., Morris, R. G., et al. 2020, MNRAS, 496, 1554
Massey, R., Schrabback, T., Cordes, O., et al. 2014, MNRAS, 439, 887
McClintock, T., Rozo, E., Becker, M. R., et al. 2019, ApJ, 872, 53
McDonald, M., Benson, B. A., Vikhlinin, A., et al. 2013, ApJ, 774, 23
McDonald, M., Stalder, B., Bayliss, M., et al. 2016, ApJ, 817, 86
McDonald, M., Allen, S. W., Bayliss, M., et al. 2017, ApJ, 843, 28
McInnes, R. N., Menanteau, F., Heavens, A. F., et al. 2009, MNRAS, 399, L84
Miyazaki, S., Komiyama, Y., Nakaya, H., et al. 2012, SPIE Conf. Ser., 8446, 84460Z
Mo, W., Gonzalez, A., Jee, M. J., et al. 2016, ApJ, 818, L25
Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, ApJS, 225, 27
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493
Oesch, P. A., Bouwens, R. J., Carollo, C. M., et al. 2010a, ApJ, 725, L150
Oesch, P. A., Bouwens, R. J., Carollo, C. M., et al. 2010b, ApJ, 709, L21
Oesch, P. A., Montes, M., Reddy, N., et al. 2018, ApJS, 237, 12
Okabe, N., Zhang, Y. Y., Finoguenov, A., et al. 2010, ApJ, 721, 875
Pacaud, F., Pierre, M., Melin, J. B., et al. 2018, A&A, 620, A10
Piffaretti, R., Arnaud, M., Pratt, G. W., Pointecouteau, E., & Melin, J. B. 2011, A&A, 534, A109
Planck Collaboration XXIV. 2016, A&A, 594, A24
Planck Collaboration V. 2020, A&A, 641, A5
Planck Collaboration VI. 2020, A&A, 641, A6
Rafelski, M., Teplitz, H. I., Gardner, J. P., et al. 2015, AJ, 150, 31
Raghunathan, S., Patil, S., Baxter, E., et al. 2019, ApJ, 872, 170
Raihan, S. F., Schrabback, T., Hildebrandt, H., Applegate, D., & Mahler, G. 2020, MNRAS, 497, 1404
Rettura, A., Chary, R., Krick, J., & Ettori, S. 2018, ApJ, 867, 12
Rowe, B. T. P., Jarvis, M., Mandelbaum, R., et al. 2015, Astron. Comput., 10, 121
Ruel, J., Bazin, G., Bayliss, M., et al. 2014, ApJ, 792, 45
Rykoff, E. S., Rozo, E., Hollowood, D., et al. 2016, ApJS, 224, 1
Sanders, J. S., Fabian, A. C., Russell, H. R., & Walker, S. A. 2018, MNRAS, 474, 1065
Schirmer, M. 2013, ApJS, 209, 21
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Schneider, P., & Seitz, C. 1995, A&A, 294, 411
Schrabback, T., Erben, T., Simon, P., et al. 2007, A&A, 468, 823
Schrabback, T., Hartlap, J., Joachimi, B., et al. 2010, A&A, 516, A63
Schrabback, T., Applegate, D., Dietrich, J. P., et al. 2018, MNRAS, 474, 2635
Schrabback, T., Bocquet, S., Sommer, M., et al. 2021, MNRAS, 505, 3923
Simon, P., Taylor, A. N., & Hartlap, J. 2009, MNRAS, 399, 48
Skelton, R. E., Whitaker, K. E., Momcheva, I. G., et al. 2014, ApJS, 214, 24
Sommer, M. W., Schrabback, T., Applegate, D. E., et al. 2022, MNRAS, 509, 1127
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, ArXiv e-prints [arXiv:1503.03757]
Stalder, B., Ruel, J., Šuhada, R., et al. 2013, ApJ, 763, 93
Strazzullo, V., Pannella, M., Mohr, J. J., et al. 2019, A&A, 622, A117
Sunyaev, R. A., & Zeldovich, Y. B. 1972, Comments Astrophys. Space Phys., 4, 173
Teplitz, H. I., Rafelski, M., Kurczynski, P., et al. 2013, AJ, 146, 159
The Dark Energy Survey Collaboration 2005, ArXiv e-prints [arXiv:astro-ph/0510346]
Tinker, J., Kravtsov, A. V., Klypin, A., et al. 2008, ApJ, 688, 709
Troxel, M. A., & Ishak, M. 2015, Phys. Rep., 558, 1
Vanderlinde, K., Crawford, T. M., de Haan, T., et al. 2010, ApJ, 722, 1180
Vikhlinin, A., Burenin, R. A., Ebeling, H., et al. 2009, ApJ, 692, 1033
Wright, A. H., Robotham, A. S. G., Bourne, N., et al. 2016, MNRAS, 460, 765
Wright, C. O., & Brainerd, T. G. 2000, ApJ, 534, 34
Zubeldia, Ì., & Challinor, A. 2019, MNRAS, 489, 401
Zubeldia, Ì., Rotti, A., Chluba, J., & Battye, R. 2021, MNRAS, 507, 4852
Zuntz, J., Paterno, M., Jennings, E., et al. 2015, Astron. Comput., 12, 45

Appendix A: Comparison of S14 and LAMBDAR photometry

While we measured fluxes in our observations with the LAMBDAR software, we only had the S14 photometry available when we estimated the redshift distribution from the CANDELS/3D-HST fields. Therefore, we checked how consistent we expect our measurements to be with the S14 photometry. We can perform this check in the central region of the GOODS-South field covering the HUDF, which we observed in the VLT FORS2 $U_{\rm HIGH}$ band. In addition to our stack in the $U_{\rm HIGH}$ band, we downloaded the stacks[24] the S14 team used in the bands F606W, F814W, F850LP, and F125W (F606W + F850LP: GO programme 9425 with PI M. Giavalisco, F814W: GO programme 12062 with PI S. Faber, F125W: GO programme 13872 with PI G. Illingworth) and measured the photometry on these stacks with LAMBDAR. We used the PSF models provided on the 3D-HST website. We then matched the galaxies in our catalogue with the galaxies in the S14 photometric catalogue with the `associate` function from the `LDAC` tools, requiring a distance of not more than 0.″3 for a match. We interpolated the magnitude J_{110} from our measurements in the filters F850LP and F125W.

In this appendix, we define all offsets of the magnitudes or colours in terms of S14 photometry minus LAMBDAR photometry. In Fig. A.1, we show how our magnitude measurements with LAMBDAR compare to the S14 photometry. We found a negative shift with a median offset of up to ~ -0.1 mag between S14 and LAMBDAR in all of the *HST* bands with a scatter of $\sim$0.3 mag. In part, this negative shift is caused by sources with a `Source Extractor` detection flag of FLAG > 0 (based on our detection in the F606W band). For these sources, `Source Extractor` recognises, for instance, contamination by nearby sources or blending. We found that the magnitude differences of these sources are predominantly negative in the direct comparison of S14 and LAMBDAR, meaning that S14 measurements are systematically brighter than LAMBDAR measurements. This is consistent with the expectation given the measurement techniques. S14 utilise aperture photometry, where fluxes are measured within apertures of fixed size with a diameter of 0.″7 for *HST* images. In contrast to that, LAMBDAR actively deblends photometry and thus measures fainter magnitudes for blended sources. But also for sources with FLAG = 0, we found a slight asymmetry skewed towards more negative magnitude differences between the S14 and LAMBDAR photometry.

For the $U_{\rm HIGH}$ band, we found a median offset of -0.062 mag with a scatter of 0.703 mag, which is a considerably larger scatter than for the *HST* bands. This is likely connected to the difference in depth between the $U_{\rm VIMOS}$ stack from S14 (5σ depth 27.4 mag) and our $U_{\rm HIGH}$ stack (5σ depth 26.6 mag) and the difference of the seeing (0.″8 for $U_{\rm VIMOS}$ versus 1.″0 for $U_{\rm HIGH}$). We found that including a conversion from the $U_{\rm VIMOS}$ band to the $U_{\rm HIGH}$ band based on the respective filter curves does not reduce this scatter. However, Fig. A.1 reveals that the scatter is a strong function of magnitude, suggesting that it is indeed related to the shallower depth of the $U_{\rm HIGH}$ data. When limited to bright $V_{606} < 25$ galaxies, it reduces to 0.426 mag.

Regarding the comparisons of colour measurements (see Fig. A.2), we found slightly positive shifts for all colours based on *HST* bands. In particular, these colours typically exhibited small shifts of up to $\sim$0.04 mag with a scatter of up to $\sim$0.11 mag. The shift for $U_{\rm HIGH} - V_{606}$ is -0.005 mag with a scatter of 0.712 mag. Systematic shifts of this order will only mildly impact the estimates of the average lensing efficiency $\langle\beta\rangle$, as we show in Appendix C. We additionally reduced a data set in the filter F110W (GO programme 14043, PI: F. Bauer) located within the GOODS-South field and compared our F110W photometry with the results from the S14 photometric catalogues. We found only mild offsets of -0.010 mag and -0.022 mag between the S14 and our photometry for the colours $V_{606} - J_{110}$ and $I_{814} - J_{110}$, respectively.

When we calculated the average lensing efficiency for the cluster fields, we could, in principle, apply the scatter that we measured when comparing the S14 and LAMBDAR photometry to all CANDELS/3D-HST catalogues to account for the different measurement techniques. However, we have to keep in mind that the comparison, which we presented here, is limited in some respects: the U bands we compared here have different depths so that we cannot clearly distinguish between effects due to depth and due to the different filter curves of $U_{\rm HIGH}$ and $U_{\rm VIMOS}$. Additionally, the CANDELS/3D-HST fields employed different U bands, and also each field has different depths in different filters. Therefore, we decided to account for differences in depth in a consistent way for all five CANDELS/3D-HST fields by adding Gaussian noise based on the difference to the depths in our cluster fields (see Table 1). However, we did investigate how shifts in the photometry as presented in this section can affect the average lensing efficiency and added the related uncertainties to our error budget (see Table 3 and Appendix C).

Appendix B: Robustness of the photometric zeropoint estimation via the galaxy locus method

For our U band calibration purposes, we defined the galaxy locus to comprise all galaxies in the magnitude range $24.2 < V_{606} < 27.0$, but excluding galaxies approximately at the cluster redshift ($1.2 \lesssim z \lesssim 1.7$) through a cut in the VIJ colour plane (see Fig. 1). As described in Sect. 3.3.3, we corrected for small shifts in the U band photometry among the five CANDELS/3D-HST fields based on the peak position of highest density in the UVI colour plane. These shifts are listed in Table B.1.

In order to estimate how well the zeropoint calibration of the $U_{\rm HIGH}$ band works for the observations of our cluster fields, we tested the zeropoint estimation in the CANDELS/3D-HST fields using only subsets of galaxies that approximately match the number of galaxies available in the cluster fields. Our cluster field observations roughly cover a field of view of 11 arcmin^2. We, therefore, only used galaxies from a region of this size from a random position in the respective CANDELS/3D-HST fields. A number of around 400 to 600 galaxies per subsample belongs to our galaxy locus (as defined by the magnitude and colour cuts in Sect. 3.3.2), which approximately equals the expected number of locus galaxies in our cluster fields. Since we had already applied a shift to the U bands in the CANDELS/3D-HST fields as explained above, this means that we measured the residual zeropoint offset for 100 different (possibly overlapping) subsamples and report the average residual zeropoint offset and scatter in Table B.1. Overall, we found that the offsets did not exceed a value of ~ -0.04 mag with a scatter of 0.08 mag. The impact of such offsets is studied in Appendix C.

[24] https://archive.stsci.edu/prepds/3d-hst/

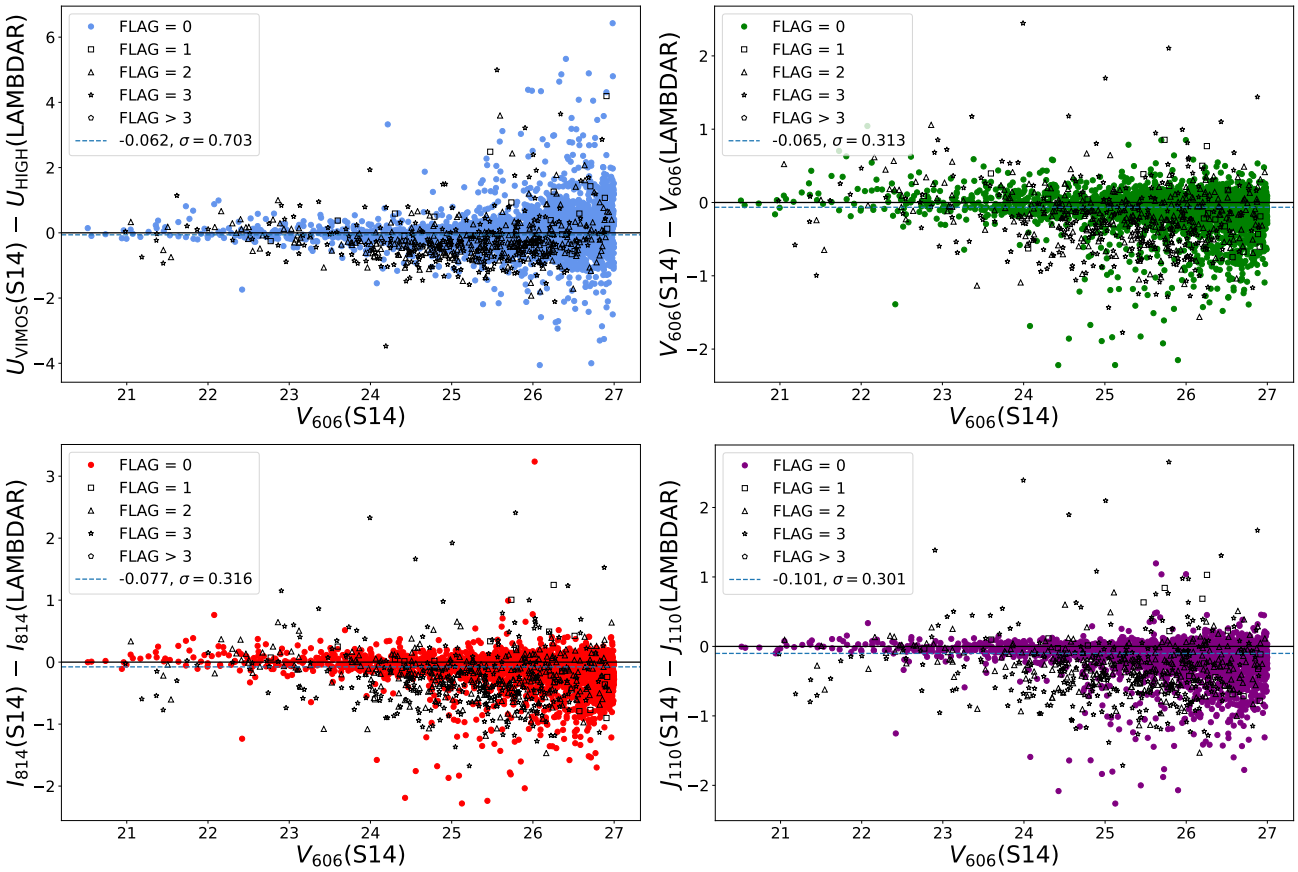


Fig. A.1. Magnitude differences between S14 and LAMBDAR photometry for the $U_{\rm HIGH}$, V_{606}, I_{814}, and J_{110} magnitudes. The blue dashed lines represent the median, and we indicate the scatter of the respective bands in the legend label. We show all matched galaxies down to $V_{606} < 27.0$ mag. We note the different scales on the y-axis for the U magnitudes and the *HST*-based magnitudes.

Table B.1. Overview about absolute and residual zeropoint offsets between CANDELS/3D-HST fields.

Field	Zeropoint offsets	
	full [mag]	100 samples [mag]
AEGIS	0.121	$-0.013\,, \sigma = 0.053$
COSMOS	0.121	$-0.021\,, \sigma = 0.062$
UDS	0.121	$-0.037\,, \sigma = 0.076$
GOODS-North	−0.040	$-0.020\,, \sigma = 0.080$
GOODS-South	0.0	$-0.027\,, \sigma = 0.055$

Notes. *First column*: Names of the CANDELS/3D-HST fields. *Second column*: Overview about the measured zeropoint offsets in the U band between the galaxy loci from the five CANDELS/3D-HST catalogues from S14 with respect to the locus in the GOODS-South field, which serves as an anchor. *Third column*: Average residual offset computed from 100 subsamples in the CANDELS/3D-HST fields (drawn from areas with a similar field of view as *HST*/ACS) after applying the 'full' correction (second column). The values correspond to the average and scatter.

Appendix C: Effect of systematic offsets in the photometry on $\langle\beta\rangle$

In order to estimate how systematic shifts in the photometry affect the average lensing efficiency, we applied different systematic shifts to the colours $U - V_{606}$, $V_{606} - I_{814}$, $V_{606} - J_{110}$, and $I_{814} - J_{110}$ from the S14 photometry. We then calculated $\langle\beta\rangle$ based on the photometric redshifts for the colour-selected galaxies. Since we applied a Gaussian noise to the U band from the GOODS-South field, we evaluated five noise realisations. A summary of the uncertainty level of the photometric shifts (based on our results presented in Appendix A and B) and the consequential uncertainties of the average lensing efficiency are presented in Table C.1.

Table C.1. Impact of expected photometric uncertainties of relevant colours on the average lensing efficiency.

Colour	Expected uncert.	$\left(\frac{\Delta\langle\beta\rangle}{\langle\beta\rangle}\right)_{\rm HUDF,R20}$	$\left(\frac{\Delta\langle\beta\rangle}{\langle\beta\rangle}\right)_{\rm CAND}$
$U - V_{606}$	±0.08 mag	2.7 %	4.1 %
$V_{606} - I_{814}$	±0.02 mag	2.9 %	2.2 %
$V_{606} - J_{110}$	±0.05 mag	2.7 %	2.2 %
$I_{814} - J_{110}$	±0.05 mag	0.3 %	0.1 %

Notes. We quantified this by calculating the difference $\Delta\langle\beta\rangle$ between the results for $\langle\beta\rangle$ (at reference redshift $z_{\rm l} = 1.4$) based on the S14 photometry shifted by the expected uncertainty in a positive and negative direction. We divide this by the average lensing efficiency $\langle\beta\rangle$ without shift of the photometry. *First column:* Colour. *Second column:* Expected uncertainty of the colour. *Third column:* Impact on the average lensing efficiency for matched galaxies in the HUDF region. We report the value based on the R20 photometric redshifts. *Fourth column:* Average impact on the average lensing efficiency for galaxies in the five CANDELS/3D-HST fields using the R20 photometric redshifts.

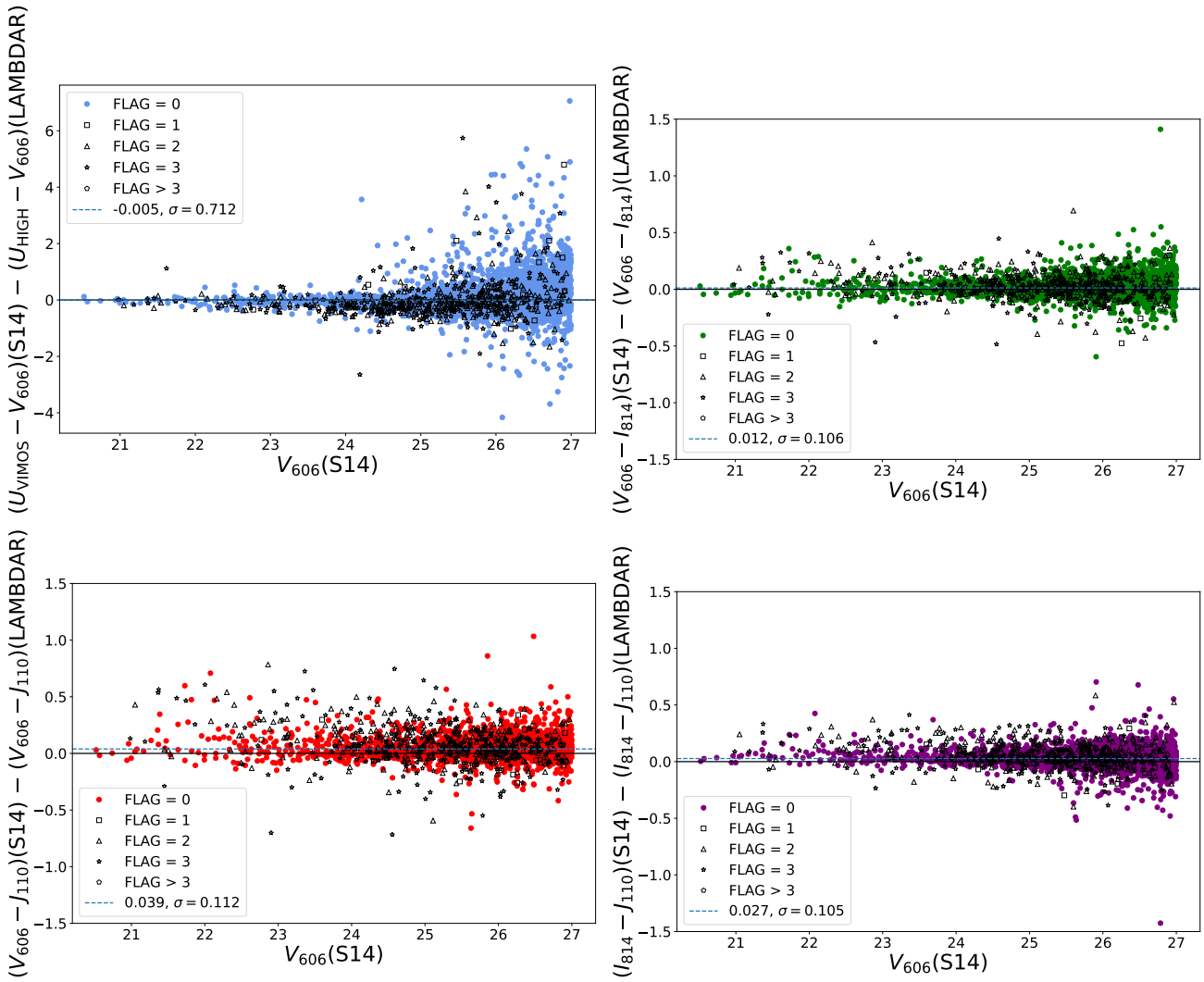


Fig. A.2. Colour differences between S14 and LAMBDAR photometry for the colours $U_{\rm HIGH} - V_{606}$, $V_{606} - I_{814}$, $V_{606} - J_{110}$, and $I_{814} - J_{110}$. The blue dashed lines represent the median and we indicate the scatter of the respective colours in the legend label. We show all matched galaxies down to $V_{606} < 27.0$ mag. We note the different scales on the y-axis for the $U - V_{606}$ colour and the *HST*-based colours.

Appendix D: Alternative colour selection strategies for clusters at $z \sim 1.2$

As mentioned before, galaxies at redshift $1.3 < z < 1.7$ could, in principle, be used for a lensing analysis for a cluster at redshift $z \sim 1.2$, but have to be removed for a cluster at redshift $z \sim 1.7$. We explored two alternative colour selection strategies of background galaxies for a cluster at redshift $z \sim 1.2$ aiming to add the galaxies at $1.3 < z < 1.7$ into the selection, which would increase the signal-to-noise ratio of the lensing measurement. In our first alternative, we left the first step of the selection in the *VIJ* colour plane unchanged, because it serves the removal of (the same) foreground galaxies as in the default selection strategy. However, we noticed that the galaxies at the cluster redshift in the *UVJ* colour plane occupy a smaller space in the upper left corner. Therefore, we modified the cuts slightly so that fewer background galaxies are cut from this corner (see Fig. D.1). At a lens redshift of $z = 1.2$ and using the matched sources from the HUDF region as in Sect. 4.2.1, the default selection strategy achieved an average lensing efficiency of $\langle\beta\rangle = 0.324$ with a number density of $n = 13.4\,\mathrm{arcmin}^{-2}$, resulting in a weak lensing sensitivity factor of $\tau_{\rm WL} = 1.19$. In comparison to that the alternative strategy achieved $\langle\beta\rangle = 0.317$ with a number density of $n = 15.3\,\mathrm{arcmin}^{-2}$, resulting in a weak lensing sensitivity factor of $\tau_{\rm WL} = 1.23$. In conclusion, this alternative provides only a negligible improvement of the weak lensing sensitivity factor, which would be even less for clusters at higher redshifts $1.2 < z \lesssim 1.6$.

As a second alternative selection strategy, we made use of the fact that the galaxies at the cluster redshift for a cluster at $z = 1.2$ are concentrated more towards the lower right of the *VIJ* colour plane than for a cluster, for instance, at $z = 1.7$. In this strategy, we used the *VIJ* plane to cut not only the foreground but also the galaxies at the cluster redshift (see Fig. D.2). To cut all galaxies at the cluster redshift this way, the cuts need to be extended further towards bluer $V - I$ colour (to the left in the *VIJ* plane). Consequently, cutting the galaxies at the cluster redshift in the upper left corner of the *UVJ* colour plane is not necessary anymore, which allows us to keep more background galaxies (mainly the close background galaxies indicated by cyan symbols in Fig. D.2). With this strategy, we found an average lensing efficiency of $\langle\beta\rangle = 0.276$ with a number density of $n = 16.9\,\mathrm{arcmin}^{-2}$, resulting in a weak lensing sensitivity factor of $\tau_{\rm WL} = 1.13$. Thus, we found we cannot increase

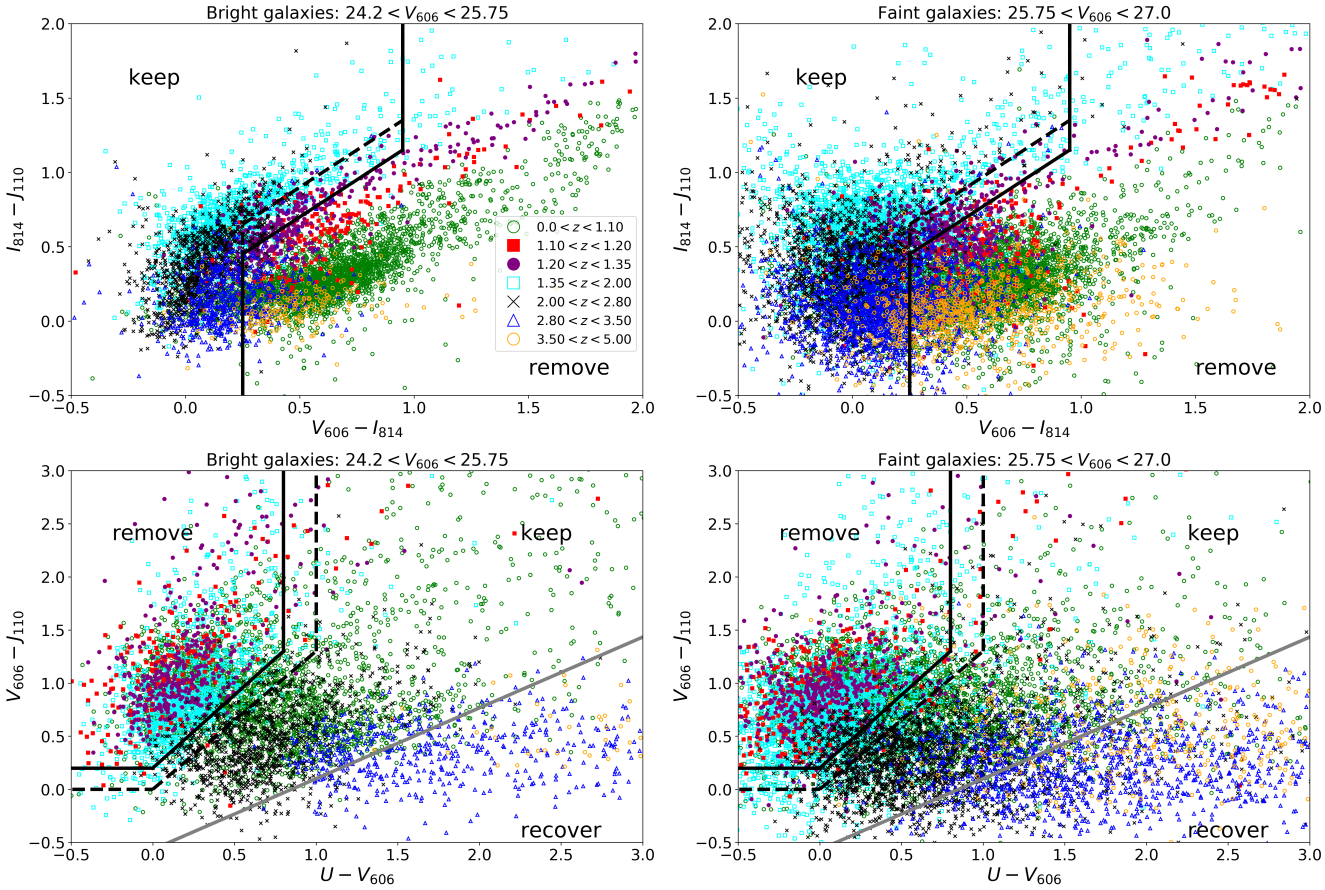


Fig. D.1. First alternative colour selection for galaxy clusters at redshift $z \sim 1.2$. The selected source galaxies are at redshift $z \gtrsim 1.3$. We display galaxies based on their photometry from S14 in the GOODS-South field. *Top:* First selection step in the VIJ plane for bright galaxies on the left and faint galaxies on the right. *Bottom:* Second selection step in the UVJ plane for bright galaxies on the left and faint galaxies on the right. The solid black lines indicate cuts applied for bright galaxies, the dashed black lines show cuts for faint galaxies. Galaxies below the diagonal grey line are recovered in both the bright and the faint regime.

the weak lensing sensitivity factor with this strategy. While the number density did increase mainly in the regime of near background galaxies, we also lost a notable fraction of the far background galaxies at high redshift due, to the more extended cut in the VIJ plane. As a result, the average geometric lensing efficiency decreased strongly and this could not be compensated by the higher source number density.

From exploring these two alternative background source selection strategies, we concluded that it is not beneficial to introduce a selection strategy that is optimised based on the cluster redshift for clusters with redshifts between $1.2 \lesssim z \lesssim 1.7$. We, therefore, applied the selection strategy presented in Sect. 4.2.1 for all clusters in our sample with $1.2 \lesssim z \lesssim 1.7$. However, for the cluster SPT-CL $J0646{-}6236$, which is located at a lower redshift of $z = 0.995$, an alternative selection strategy did increase the weak lensing sensitivity factor noticeably as presented in Appendix E.

Appendix E: Colour selection strategy for the cluster SPT-CL $J0646{-}6236$ at $z = 0.995$

The cluster SPT-CL $J0646{-}6236$ has the lowest redshift in our sample with $z = 0.995$. With the default background source selection strategy presented in Sect. 4.2.1, we do miss the galaxies in the redshift regime $1.1 \lesssim z \lesssim 1.7$, which we could incorporate for the lensing analysis of this cluster. In contrast to the alternative background source selection strategies presented in Appendix D, we found that it is possible to achieve a significantly higher weak lensing sensitivity factor with a modification of the default selection strategy for this cluster. The original cut in the VIJ plane already removed the majority of the galaxies at the cluster redshift $z \sim 1$, so that we could omit the cut of sources in the upper left corner of the UVJ plane (see Fig. E.1). As a result, we achieved a number density of selected background source galaxies, which was two times higher ($27.4\,\mathrm{arcmin}^{-2}$) than for the default selection while the average geometric lensing efficiency only mildly decreased. At a lens redshift of $z = 0.995$, we found $\langle\beta\rangle = 0.392$ for the default selection and $\langle\beta\rangle = 0.336$ for the optimised selection. As a consequence, the weak lensing sensitivity factor increased by about 23 per cent from $\tau_{\mathrm{WL}} = 1.43$ for the default selection strategy to $\tau_{\mathrm{WL}} = 1.76$ for the optimised strategy. Therefore, we used this optimised strategy in the lensing analysis of the cluster SPT-CL $J0646{-}6236$ at $z = 0.995$.

Appendix F: Consistency of weak lensing mass results with SZ or X-ray masses

Some of the clusters in our sample have a lensing mass that has scattered high or low with respect to the reference mass measured from SZ or X-ray data (see Table 1 for SZ masses, McDonald et al. 2017 for X-ray masses). This concerns

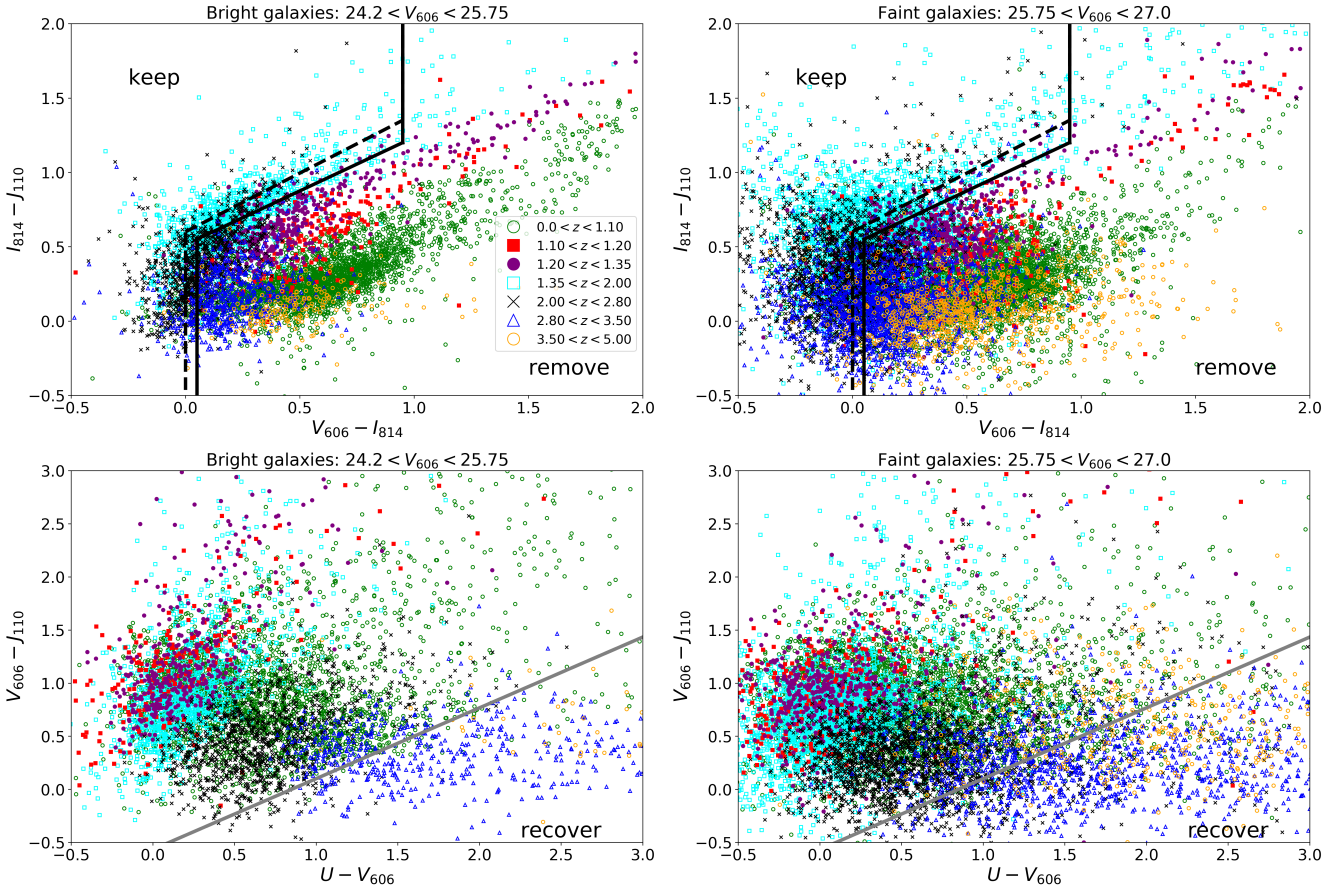


Fig. D.2. Second alternative colour selection for galaxy clusters at redshift $z \sim 1.2$. The selected source galaxies are at redshift $z \gtrsim 1.3$. We display galaxies based on their photometry from S14 in the GOODS-South field. *Top:* First selection step in the VIJ plane for bright galaxies on the left and faint galaxies on the right. The solid black lines indicate cuts applied for bright galaxies, the dashed black lines show cuts for faint galaxies. *Bottom:* Second selection step in the UVJ plane for bright galaxies on the left and faint galaxies on the right. Galaxies below the diagonal grey line are recovered in both the bright and the faint regime.

in particular the clusters SPT-CL $J2040{-}4451$ and SPT-CL $J0205{-}5829$. To quantify the tension between weak lensing mass and SZ or X-ray mass for individual targets, we employed a simple model to test for the level at which the mass ratios are consistent with unity. To this end, we randomly drew 10,000 weak lensing masses $M_{\mathrm{WL,rand},i}$ from a Normal distribution $\mathcal{N}(M_{500c}^{\mathrm{biased,ML}}, \sigma_{\mathrm{stat}}(M_{500c}^{\mathrm{biased,ML}}))$ given the best-fit weak lensing mass estimates and statistical uncertainties (see Sect. 6.2 and Table 6). We divided these by correction factors randomly drawn from the corresponding log-normal mass bias distributions (described in Sect. 6.3). Similarly, we drew 10,000 SZ (or X-ray) masses $M_{\mathrm{SZ,rand},i}$ (or $M_{\mathrm{X,rand},i}$) from the best-fit values in conjunction with their uncertainties (Table 1 for SZ masses, McDonald et al. 2017 for X-ray masses), using a Normal distribution. In case of asymmetric uncertainties, a two-piece Normal distribution (e.g. John 1982)s was employed. We proceeded to take ratios of the weak lensing and SZ (or X-ray) mass distributions $M_{\mathrm{WL,rand},i}/M_{\mathrm{SZ,rand},i}$ (or $M_{\mathrm{WL,rand},i}/M_{\mathrm{X,rand},i}$). For a given target, the resulting ratio distribution was analysed for its consistency with unity. In particular, we constructed confidence intervals based on the shortest possible interval containing a given fraction (the confidence level) of the distribution. In this way, we found the lowest level of confidence making the mass ratio consistent with one. For SPT-CL $J2040{-}4451$, which has a best-fit weak lensing mass noticeably higher than the SZ mass (X-ray mass), we found this confidence level to be 70 per cent (75 per cent), corresponding to a probability of 0.3 (0.25) of seeing an outlier with this degree or more of discrepancy (for an individual cluster). Similarly, for SPT-CL $J0205{-}5829$, the probability of an outlier with or exceeding the observed degree of discrepancy is 0.09 for the SZ mass (0.21 for the X-ray mass). We conclude that the observed scatter between lensing masses and SZ or X-ray masses is well within the expectation given the large statistical uncertainties of our study, and given that these two clusters are the most extreme outliers within our sample of nine clusters.

Appendix G: Weak lensing results: mass maps and tangential reduced shear profiles

We show the weak lensing results, including the mass maps and tangential reduced shear profiles for the studied cluster sample in Figs. G.2 to G.4. In addition, we display the stacked profile of the cluster sample in Fig. G.1. Following S18 (their Sect. 7.3), we stacked the lensing signal of the clusters in terms of the differential surface mass density $\Delta\Sigma(r)$, where we computed Σ_{crit} based on the average lensing efficiency $\langle\beta\rangle$ from the individual clusters, respectively. Since the clusters vary in mass, we rescaled them to an approximately similar signal amplitude with the help of the SZ masses listed in Table 1. Based on this mass and assuming the concentration–mass relation by Diemer & Kravtsov (2015) with updated parameters from Diemer & Joyce (2019), we

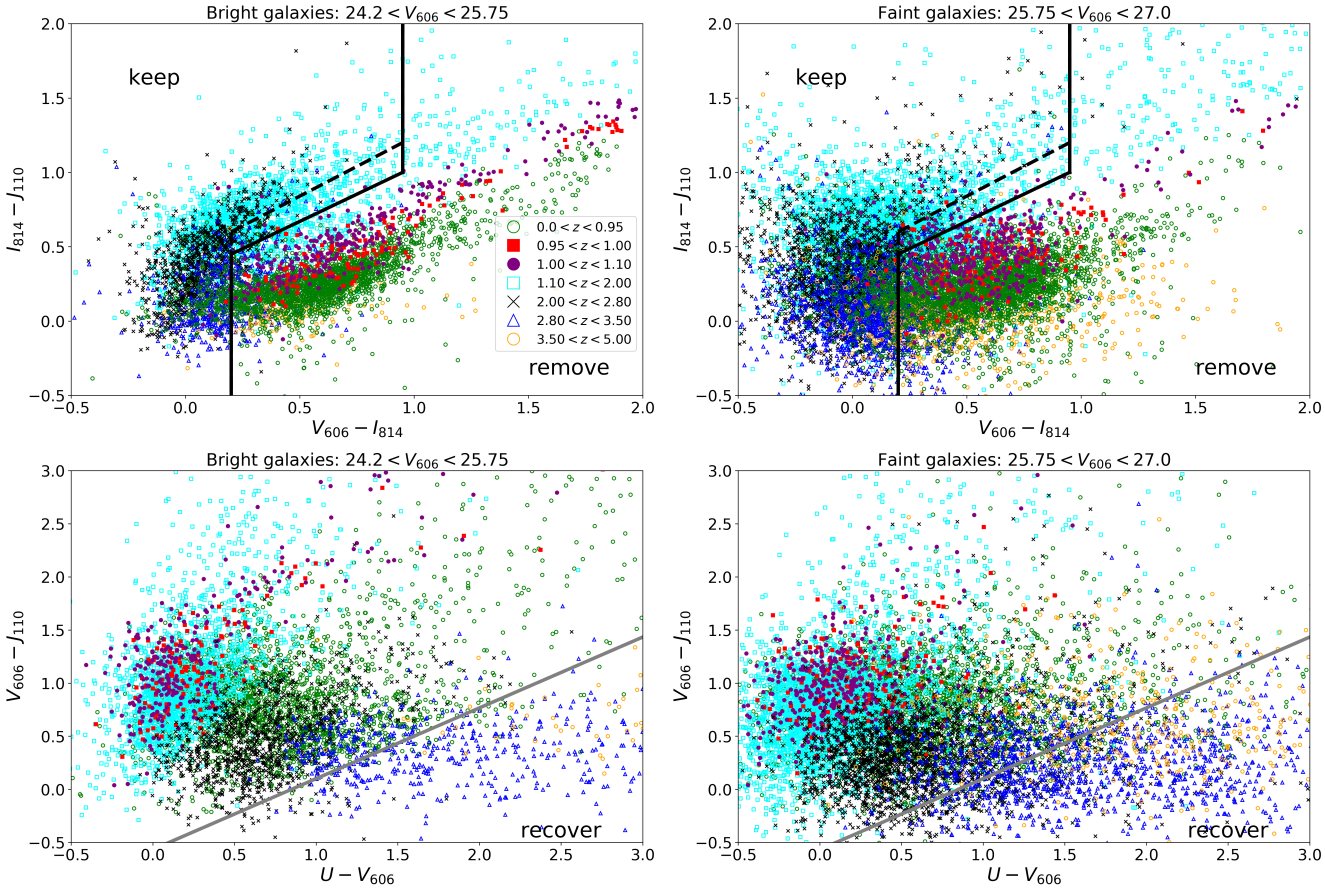


Fig. E.1. Colour selection for galaxy clusters at redshift $z \sim 1.0$. The selected source galaxies are at redshift $z \gtrsim 1.1$. We display galaxies based on their photometry from S14 in the GOODS-South field. *Top:* First selection step in the VIJ plane for bright galaxies on the left and faint galaxies on the right. The solid black lines indicate cuts applied for bright galaxies, the dashed black lines show cuts for faint galaxies. *Bottom:* Second selection step in the UVJ plane for bright galaxies on the left and faint galaxies on the right. Galaxies below the diagonal grey line are recovered in both the bright and the faint regime.

computed a theoretical NFW model for the differential surface mass density $\Delta\Sigma_{\rm model}$. We then rescaled the cluster lensing signal by a factor s according to

$$\Delta\Sigma^*(r) = s\Delta\Sigma(r) \equiv \frac{\langle\Delta\Sigma_{\rm model}(800\,{\rm kpc})\rangle}{\Delta\Sigma_{\rm model}(800\,{\rm kpc})}\Delta\Sigma(r), \tag{G.1}$$

where we used $r = 800\,$kpc as the reference scale to evaluate the theoretical model. The weighted average then reads

$$\langle\Delta\Sigma^*\rangle(r_j) = \sum_{i\in{\rm clusters}} \Delta\Sigma_i^*(r_j)\hat{W}_{ij} \Big/ \sum_{i\in{\rm clusters}} \hat{W}_{ij}, \tag{G.2}$$

with $\hat{W}_{ij} = \left[s\sigma(\Delta\Sigma(r_j))\right]^{-2}$ and $\sigma(\Delta\Sigma(r_j))$ as the 1σ uncertainty of $\Delta\Sigma(r_j)$.

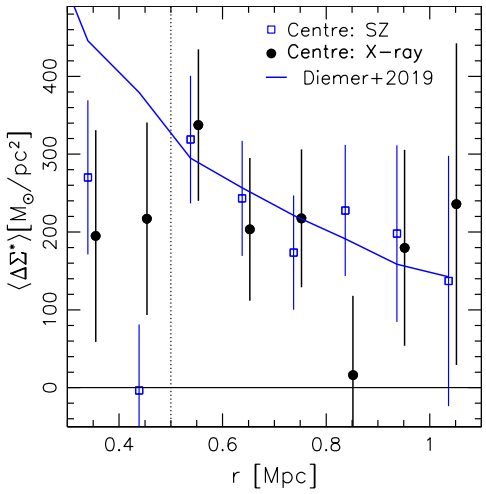


Fig. G.1. Weighted average of the rescaled differential surface mass density profiles for the clusters in our sample. The black points and blue squares refer to measurements using the X-ray (all clusters except SPT-CL $J0646{-}6236$ for which we do not have X-ray measurements) and SZ centres, respectively. The blue line shows the average weighted model NFW function assuming a fixed concentration–mass relation following Diemer & Kravtsov (2015) with updated parameters from Diemer & Joyce (2019) for measurements from the SZ centres. The vertical dotted line indicates the lower limit of our fit range.

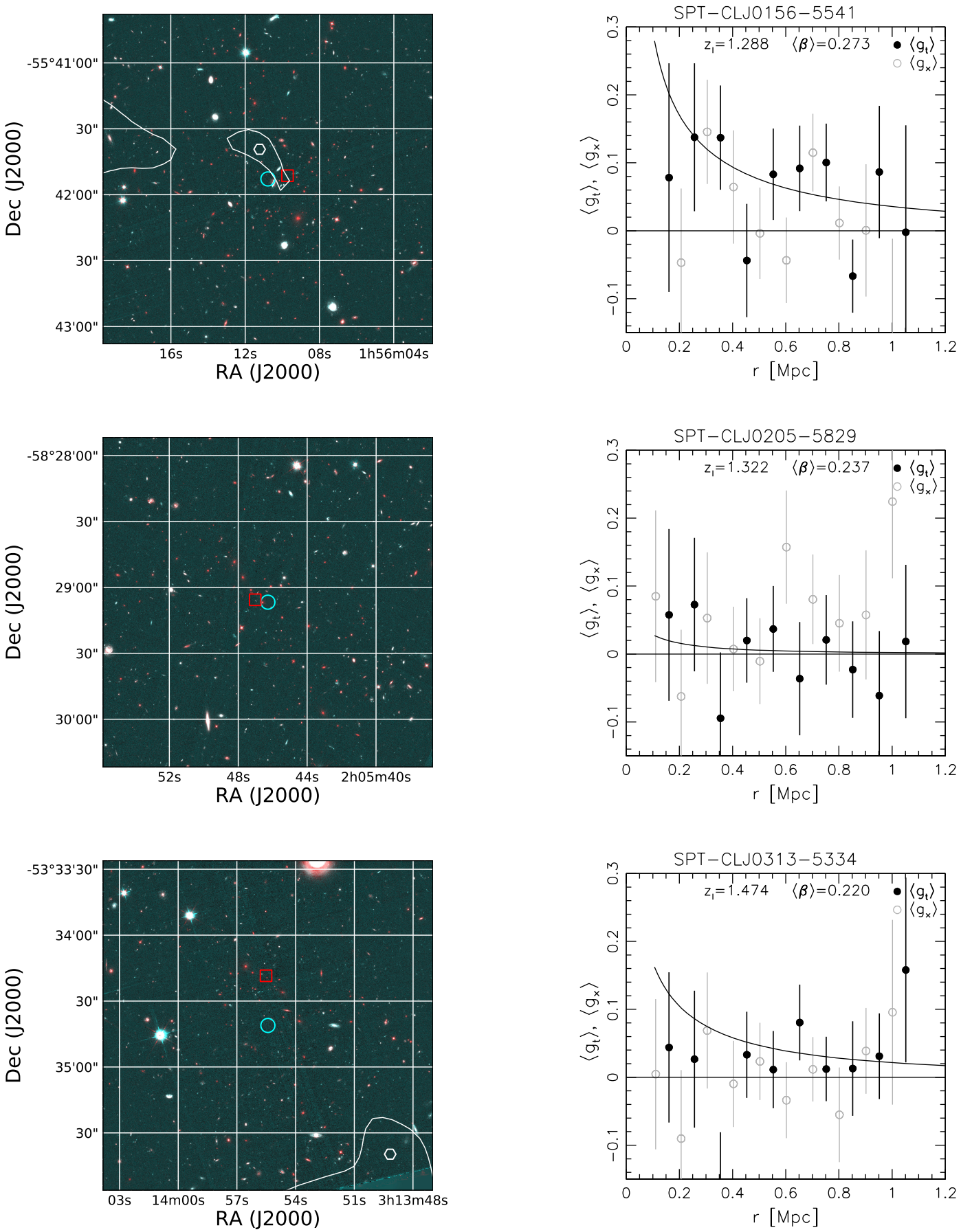
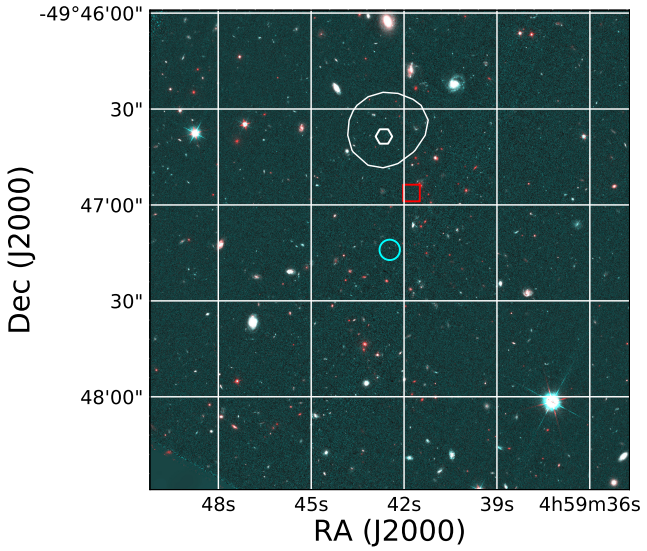
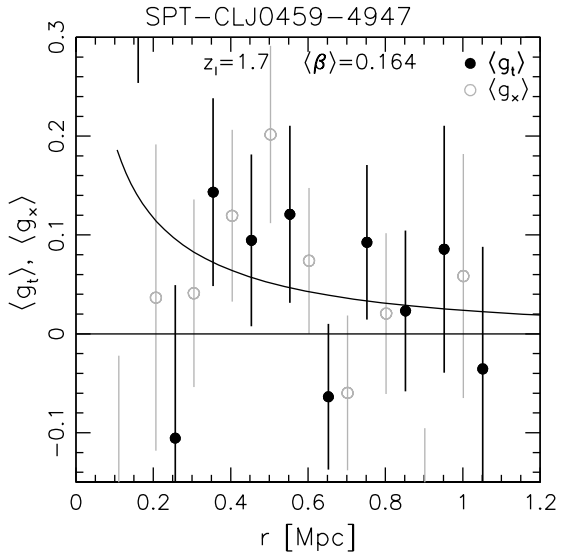
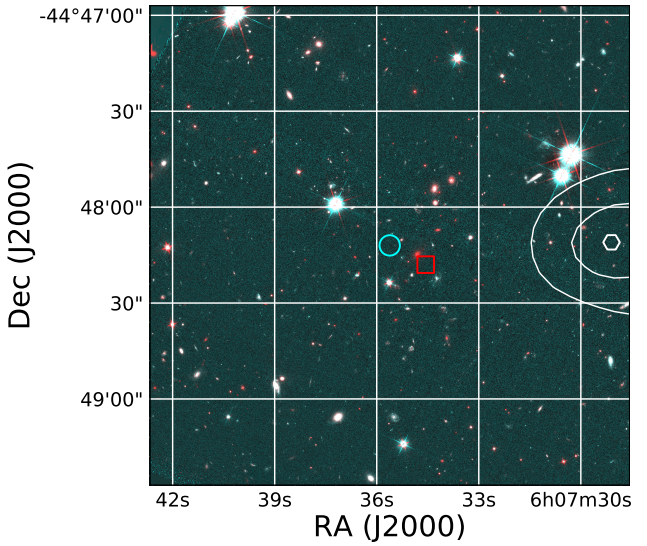
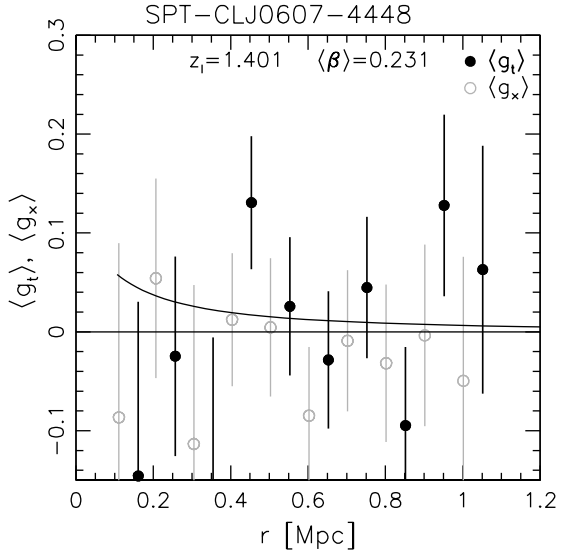


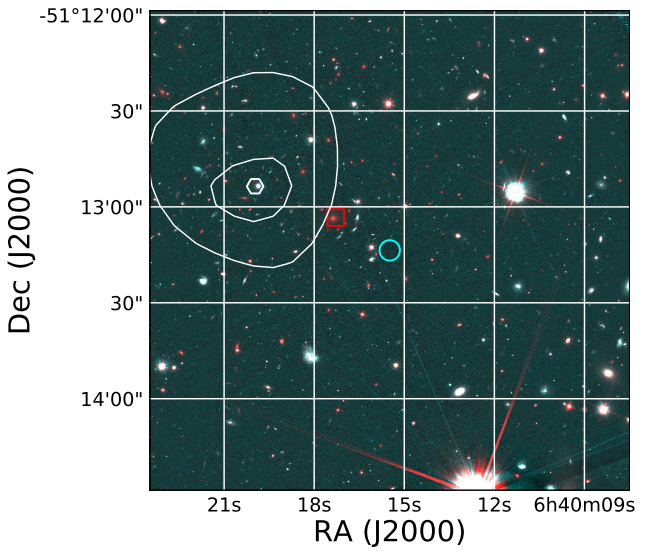
Fig. G.2. Weak lensing results for the clusters in our sample (see the caption of Fig. 9 for details).











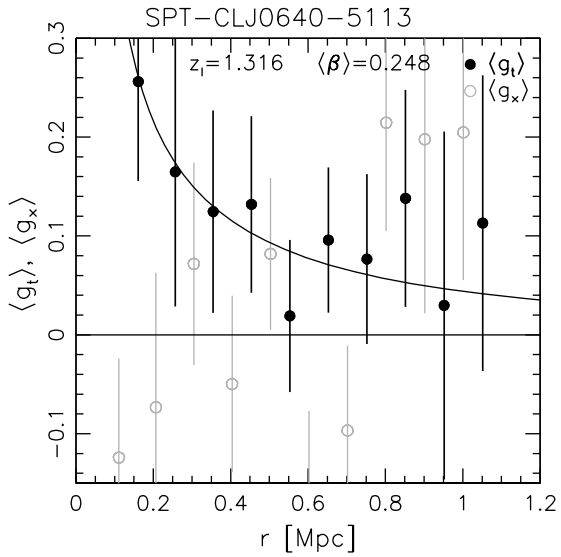


Fig. G.3. Weak lensing results for the clusters in our sample (continued, see the caption of Fig. 9 for details).

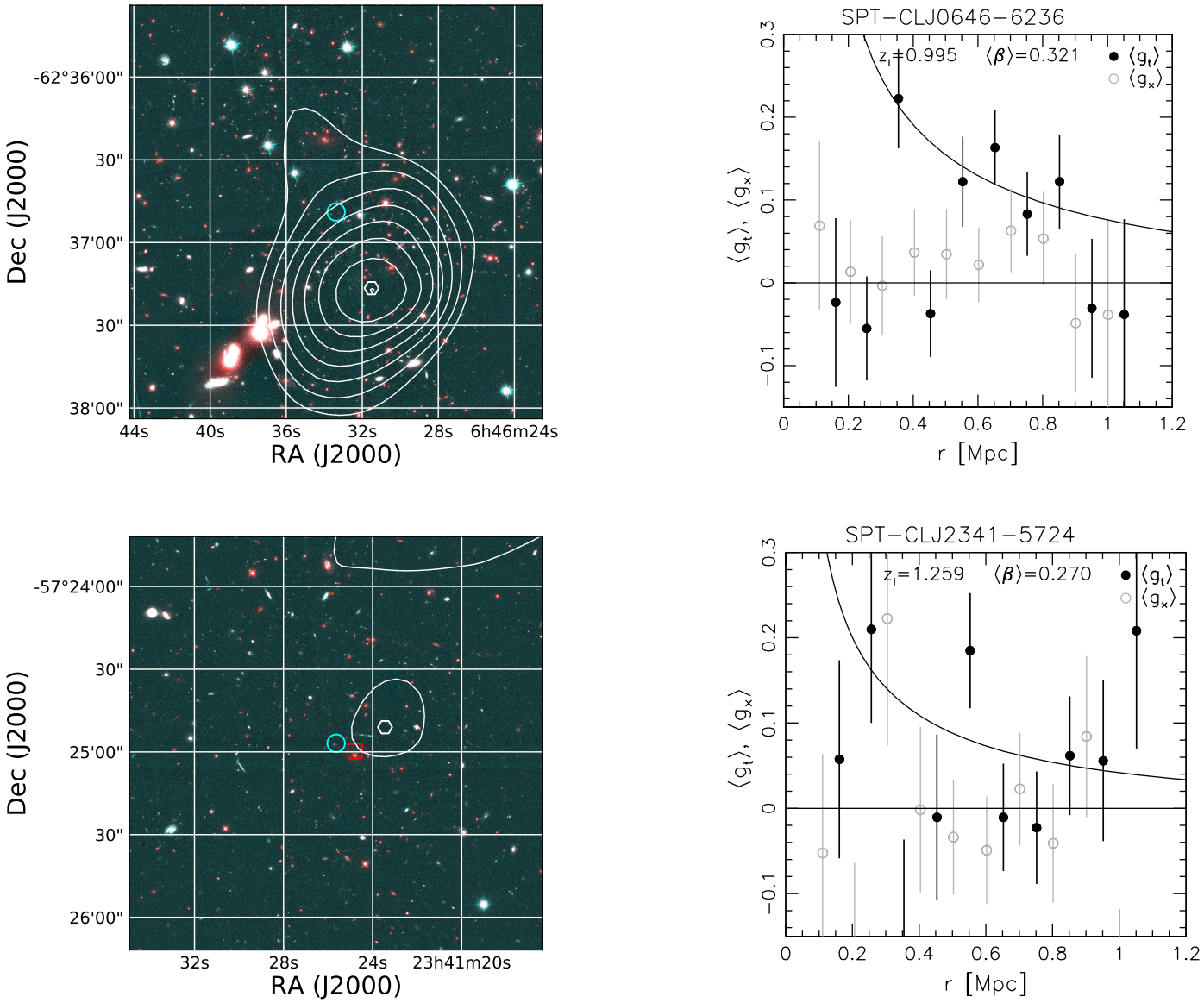


Fig. G.4. Weak lensing results for the clusters in our sample (continued, see the caption of Fig. 9 for details). For SPT-CL *J*0646−6236 the reduced shear profile was computed with respect to the SZ centre.