PhotoWeb redshift: boosting photometric redshift accuracy with large spectroscopic surveys

M. Shuntov, J. Pasquet, S. Arnouts, O. Ilbert, M. Treyer, E. Bertin, S. de La Torre, Y. Dubois, D. Fouchez, K. Kraljic, et al.

A&A 636, A90 (2020)
https://doi.org/10.1051/0004-6361/201937382

Astronomy & Astrophysics

PhotoWeb redshift: boosting photometric redshift accuracy with large spectroscopic surveys

M. Shuntov[1,2], J. Pasquet[3], S. Arnouts[1], O. Ilbert[1], M. Treyer[1], E. Bertin[2], S. de la Torre[1], Y. Dubois[2], D. Fouchez[3], K. Kraljic[4], C. Laigle[2], C. Pichon[2,5], and D. Vibert[1]

[1] Aix Marseille Université, CNRS, CNES, UMR 7326, Laboratoire d'Astrophysique de Marseille, Marseille, France
[2] Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis bd Arago, 75014 Paris, France
e-mail: shuntov@iap.fr
[3] Aix-Marseille Université, CNRS/IN2P3, Centre de Physique des particules de Marseille, Marseille, France
[4] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[5] Korea Institute for Advanced Study (KIAS), 85 Hoegiro, Dongdaemun-gu, Seoul 02455, Republic of Korea

Received 20 December 2019 / Accepted 6 March 2020

ABSTRACT

Improving distance measurements in large imaging surveys is a major challenge to better reveal the distribution of galaxies on a large scale and to link galaxy properties with their environments. As recently shown, photometric redshifts can be efficiently combined with the cosmic web extracted from overlapping spectroscopic surveys to improve their accuracy. In this paper we apply a similar method using a new generation of photometric redshifts based on a convolution neural network (CNN). The CNN is trained on the SDSS images with the main galaxy sample (SDSS-MGS, $r \leq 17.8$) and the GAMA spectroscopic redshifts up to $r \sim 19.8$. The mapping of the cosmic web is obtained with 680 000 spectroscopic redshifts from the MGS and BOSS surveys. The redshift probability distribution functions (PDF), which are well calibrated (unbiased and narrow, ≤120 Mpc), intercept a few cosmic web structures along the line of sight. Combining these PDFs with the density field distribution provides new photometric redshifts, $z_{\rm web}$, whose accuracy is improved by a factor of two (i.e., $\sigma \sim 0.004(1+z)$) for galaxies with $r \leq 17.8$. For half of them, the distance accuracy is better than 10 cMpc. The narrower the original PDF, the larger the boost in accuracy. No gain is observed for original PDFs wider than 0.03. The final $z_{\rm web}$ PDFs also appear well calibrated. The method performs slightly better for passive galaxies than star-forming ones, and for galaxies in massive groups since these populations better trace the underlying large-scale structure. Reducing the spectroscopic sampling by a factor of 8 still improves the photometric redshift accuracy by 25%. Finally, extending the method to galaxies fainter than the MGS limit still improves the redshift estimates for 70% of the galaxies, with a gain in accuracy of 20% at low z where the resolution of the cosmic web is the highest. As two competing factors contribute to the performance of the method, the photometric redshift accuracy and the resolution of the cosmic web, the benefit of combining cosmological imaging surveys with spectroscopic surveys at higher redshift remains to be evaluated.

Key words. galaxies: distances and redshifts

1. Introduction

Photometric redshifts are a key component for the exploitation of large imaging surveys (see, e.g., Salvato et al. 2019, for a review). They are a cheap alternative to spectroscopic surveys for the measurement of distances of millions of galaxies. They have been widely used to study the evolution of galaxy properties over cosmic time (e.g., Ilbert et al. 2013; Madau & Dickinson 2014; Davidzon et al. 2017) or to link galaxies with their dark matter halos (e.g., Coupon et al. 2015), and are essential to study the nature of dark energy. Weak lensing cosmological probes also need an accurate estimate of the mean redshift of the selected galaxy populations (Knox et al. 2006), while the figure of merit of the baryon acoustic oscillation probe can to some extent be improved by combining dense photometric samples with sparse spectroscopic surveys (Patej & Eisenstein 2018). The derivation of robust redshift probability distribution functions (PDFs) is also necessary to understand the uncertainties attached to any of the above measurements (Mandelbaum et al. 2008).

The highly nonlinear mapping between the photometric space and the redshift space has been performed essentially via two broad techniques. The first, template fitting (e.g., Arnouts et al. 1999; Benítez 2000; Brammer et al. 2008), matches the broad-band photometry of each galaxy to the synthetic magnitudes of a suite of templates across a large redshift interval. This technique does not require a large spectroscopic sample for training, but it is often computationally intensive and involves poorly known parameters, such as dust attenuation, which can lead to degeneracies in color–redshift space. The second group of techniques includes machine learning methods, such as artificial neural networks (Collister & Lahav 2004), k-nearest neighbors (kNN, Csabai et al. 2007), self-organizing maps (SOM, Masters et al. 2015; Davidzon et al. 2019), or random forest techniques (Carliles et al. 2010), which perform better within the limits of the training set (Sánchez et al. 2014), but the lack of spectroscopic coverage in some color–space regions, and at high redshift remain a major issue (Masters et al. 2019). For these reasons, hybrid approaches have emerged to optimize the photometric redshift PDF estimates (e.g., Carrasco Kind & Brunner 2014; Cavuoti et al. 2017; Hemmati et al. 2019).

One limiting factor of these techniques is the information used as input. Magnitudes or colors are affected by choices of aperture size, PSF variations, and overlapping sources (Hildebrandt et al. 2012). In recent years the deep learning techniques, such as

Convolutional Neural Networks (CNN), have bypassed this limitation by dealing directly with multiband galaxy images at the pixel level, without relying on photometric feature extractions (Hoyle 2016; D'Isanto & Polsterer 2018; Pasquet et al. 2019). As shown by Pasquet et al. (2019), who trained a CNN on images from the SDSS spectroscopic sample, this method surpasses current machine learning photometric redshift estimates in the SDSS survey (based on kNN, Beck et al. 2017). CNN photometric redshifts are almost free of bias with respect to disk inclination and galactic reddening E_{B-V}, for example, while color-based photometric redshifts are not. The associated PDFs are also well calibrated and provide a reliable indicator of the redshift uncertainty. However, despite constant improvements in photometric redshift techniques, even the best SED fitting (such as in the COSMOS field imaged in a large number of filters, Laigle et al. 2016) or deep learning methods (Pasquet et al. 2019) hardly reach a redshift uncertainty σ_z below ∼0.01, which corresponds to a distance uncertainty of ∼40 cMpc (at $z \sim 1$).

Redshifts can also be predicted from the spatial distribution of galaxies on a large scale. The spatial cross-correlation between a photometrically selected sample and a reference sample with known spectroscopic redshifts offers an efficient way to infer the redshift distribution $N(z)$ of the photometric sample (known as the clustering redshift technique; e.g., Matthews & Newman 2010; Ménard et al. 2013). This method was extended with a hierarchical Bayesian model to simultaneously constrain $N(z)$ and the redshift of individual galaxies (Leistedt et al. 2016; Sánchez & Bernstein 2019).

With a similar methodology it is possible to improve the individual photo-z estimates by using the known galaxy density field, reconstructed from spectroscopic surveys (Kovač et al. 2010; Aragon-Calvo et al. 2015) or the 3D tomography of the intergalactic medium with neutral hydrogen absorption lines (at high redshift, Schmittfull & White 2016; Lee & White 2016). The large-scale structure formation results from the anisotropic gravitational collapse of the primordial dark matter density fluctuations (Zel'dovich 1970), giving rise to large underdense regions bordered by sheet-like walls, which are framed by filaments connecting density peaks (nodes). These features form the so-called cosmic web (CW; Bond et al. 1996), identified in local galaxy surveys (York et al. 2000; Colless et al. 2003). Galaxies are preferentially found in overdense regions of the underlying density field as a consequence of the biased formation of their dark matter halos (Mo & White 1996). The underdense regions appear almost empty of galaxies (voids account for less than 5% of luminous galaxies; Aragon-Calvo et al. 2015), while they occupy almost 90% of the volume of the universe (Aragón-Calvo et al. 2010; Cautun et al. 2014). The vast majority of galaxies thus lie within the remaining 10% composed of the dense regions distributed in a geometric pattern of walls, filaments, and nodes. The most massive galaxies live preferentially in the nodes (highest density regions), but segregation also occurs in filamentary regions where more massive or passive galaxies are closer to the center of the filaments (Malavasi et al. 2017; Kraljic et al. 2018).

The galaxy density field can therefore provide strong priors on the location of a galaxy and help to narrow its original photometric redshift PDF to a few more probable redshifts corresponding to the spikes of the intercepted density field along the line of sight, as proposed by Kovač et al. (2010). When anchored to the right density peak, the photometric redshift accuracy is increased up to the resolution of the reconstructed density field, usually a few cMpc, i.e., about 10 times better than current individual photometric redshift estimates. This method has been further improved by Aragon-Calvo et al. (2015, named the PhotoWeb redshift method). In addition to the density field, they introduced an extra term to mitigate the influence of the nodes (highest density peaks) in the resulting redshift PDF. For any point along the line of sight, this term scales inversely to the closest distance of any structure, allowing for a better contribution of less dense structures such as filaments and walls. They applied the PhotoWeb redshift to the SDSS sample. They reconstructed the cosmic web with the spectroscopic galaxies up to $z \sim 0.12$ and used the SDSS photometric redshifts of Csabai et al. (2007), based on a k-NN method. Restricting the sample to galaxies with good photometric redshift accuracy ($\Delta z \leq 0.015$), they showed that using the prior knowledge of the cosmic web yields photometric redshifts with Megaparsec accuracy.

In the present paper, we adopt the same strategy as Aragon-Calvo et al. (2015), but we push the analysis further in redshift and magnitude. Our method, the Photo Webredshift technique (hereafter PW-z), relies on the cosmic web extractor DisPerSE (Sousbie 2011), and is applied to the CNN photometric redshifts from Pasquet et al. (2019). We explore the performance of the resulting photometric redshifts $z_{\rm web}$ as a function of CNN PDF width, while making no pre-selection regarding their uncertainties, and we test the reliability of the final PW-z PDF. Furthermore, we analyze the performance of the method with respect to galaxy properties (galaxy types, group memberships) and how the resolution of the CW reconstruction impacts the $z_{\rm web}$ accuracy. Our analysis also extends to galaxies two magnitudes fainter than SDSS using the GAMA survey.

The outline of the paper is as follows. In Sect. 2 we describe the photometric and spectroscopic dataset and show the $z_{\rm CNN}$ measurements. In Sect. 3 we describe the PW-z method. The main results are presented in Sect. 4, followed by the conclusion in Sect. 2. Throughout this paper we adopt a flat cosmology with $\Omega_{\rm m} = 0.307115$ and the Hubble constant $H_0 = 67.77\,{\rm km\,s^{-1}\,Mpc^{-1}}$.

2. Spectroscopic and photometric redshift dataset

2.1. Spectroscopic redshift dataset

To perform this analysis, we use the SDSS and BOSS spectroscopic samples from the data release 12 (DR12, Alam et al. 2015) and the GAMA spectroscopic samples from the data release 3 (Baldry et al. 2018). The characteristics of each sample are as follows:

– We use the main galaxy sample of the SDSS (hereafter MGS) to train and validate our photometric redshift estimates. It is limited to galaxies with dereddened Petrosian magnitudes $r \leq 17.8$. We only use the large contiguous region shown in Fig. 1 (left panel), covering ∼7400 deg^2 and containing ∼480 000 galaxies. The redshift distribution is shown in Fig. 1 (right panel).

– The spectroscopic sample used to reconstruct the CW consists of the MGS sample completed by the luminous red galaxy sample (LRG) and by the BOSS sample for a total of ∼686 000 galaxies up to $z = 0.4$. This redshift limit encompasses all the MGS galaxies. The redshift distribution is shown in Fig. 1 (right panel). In the bottom panel, we show the evolution of the spatial density as a function of redshift, characterized by the mean intergalactic comoving distance[1]. Between $z \sim 0.15$ and $z \sim 0.25$, the

[1] $\langle d_{\rm inter} \rangle = 1/\sqrt[3]{\varphi(z)}$, where $\varphi(z)$ is the selection function taking into account the density variation with the radial distance induced by the flux limits and color selections of the different samples.

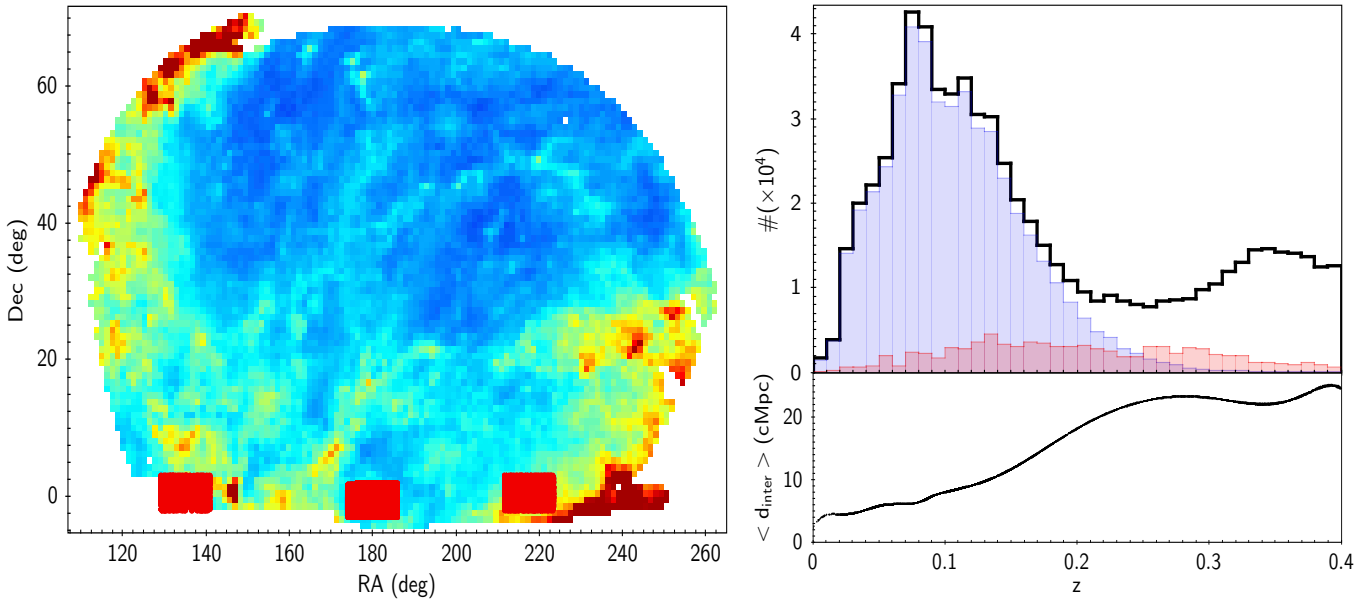

Fig. 1. *Left*: footprints of the SDSS + BOSS sample, color-coded with the galactic reddening excess ($E(B-V)$), and the three equatorial GAMA fields (large red rectangles). *Right top panel*: redshift distributions of the SDSS MGS sample ($r \leq 17.8$, filled light blue histogram), the GAMA sample ($r \leq 19.0-19.8$, filled light red histogram), and the whole spectroscopic sample (MGS + LRGs + BOSS samples; solid black line) used to reconstruct the cosmic web. *Right bottom panel*: mean intergalactic comoving distance of the whole spectroscopic sample (see text).

target density becomes sparse, providing a coarse representation of the CW above $z \sim 0.2$.

– We also derived a second set of photometric redshifts trained with the GAMA spectroscopic survey which is two magnitudes deeper than the MGS sample. It consists of two fields with spectroscopic redshifts down to $r = 19.0$ (namely G09 and G12) and one field down to $r = 19.8$ (G15). These three fields cover 180 deg^2 and overlap the SDSS-BOSS footprint, as shown in Fig. 1 (left panel). As suggested by the GAMA team, we restrict the sample to the most secure redshifts (nQ ≥ 3), namely ~99 500 galaxies. The total redshift distribution is shown in Fig. 1 (right panel). In the deepest field (G15), 4% of the galaxies are located at $z > 0.4$ and will be ignored in the rest of the paper.

2.2. Photometric redshifts with SDSS-MGS

Our first set of photometric redshifts is trained and validated with the SDSS-MGS sample (Pasquet et al. 2019). They are estimated with a convolutional neural network (CNN), which is a special type of multilayered neural network. The input data are 64×64 pixel images centered on the galaxy coordinates in the five bands of the SDSS imaging survey (*ugriz*). The architecture of the CNN is detailed in Pasquet et al. (2019). In brief, it is composed of several convolution and pooling layers followed by fully connected layers. The convolution part of the network is organized in multi-scale blocks called inception modules to treat the signal at different resolution scales (Szegedy et al. 2015). The redshift values are estimated as a classification problem, where each class corresponds to a narrow redshift bin δz (here 180 redshift bins between $0 \leq z \leq 0.4$). The network assigns a probability to each redshift bin, which is used as a probability distribution function (PDF). We define the redshift value as the weighted mean of the PDF ($z_{\rm CNN} = \sum_k z_k {\rm PDF}_k$). The power of this technique relies in the exploitation of all the information available in the images at the pixel level, without any prior feature extraction.

To assess the performance of the method, we adopt the same statistics used by Pasquet et al. (2019):

– the residuals, $\Delta z = (z_{\rm CNN} - z_{\rm spec})/(1 + z_{\rm spec})$;
– the bias, $\langle \Delta z \rangle$, defined as the mean of the residuals;
– the MAD deviation (Median Absolute Deviation)[2], defined as $\sigma_{\rm MAD} = 1.4826 \times {\rm median}(|\Delta z - {\rm median}(\Delta z)|)$.
– the fraction of outliers, η, defined as the fraction of galaxies with $|\Delta z| > 0.05$;

The CNN photometric redshifts are highly accurate at the depth of the SDSS-MGS sample ($r \leq 17.8$), with $\sigma_{\rm MAD}$ lower than 0.01. In Fig. 2 we show the evolution of $\sigma_{\rm MAD}$ as a function of redshift. The behavior of the photometric redshift accuracy is relatively independent of redshift up to $z \sim 0.3$ for the MGS, with less than 2% of the MGS being above this redshift.

Of particular interest is the reliability of the redshift PDF derived by the CNN. To evaluate the predictive power of the PDFs we use the probability integral transform statistic (PIT; Polsterer et al. 2016; Pasquet et al. 2019). For each galaxy the PIT is measured as the cumulative PDF (CDF) up to the spectroscopic redshift, $z_{\rm s}$ $\left({\rm CDF}(z_{\rm s}) = \int_0^{z_{\rm s}} {\rm PDF}(z)\, {\rm d}z\right)$. A flat distribution of the PIT values in a given sample indicates that the PDFs are not biased with respect to the spectroscopic redshifts. They are neither too narrow nor too wide, whereas convex or concave distributions point to under- or overdispersed PDFs, respectively (Polsterer et al. 2016). A negative or positive slope in the PIT distribution indicates a systematic bias (over- or underpredicted redshifts, respectively). We find a nearly flat PIT distribution except at the extreme values that are slightly underpopulated, which suggests that the PDFs are marginally too broad (see Fig. 10 in Pasquet et al. 2019).

2.3. Photometric redshifts with GAMA

We also created a second set of photometric redshifts using GAMA as the training sample, which is two magnitudes deeper than the MGS. The characteristics of the input images remain

[2] Strictly speaking, this is the standard deviation σ estimated from the MAD deviation for normally distributed data: $\sigma \approx 1.4826 \times {\rm MAD}$.

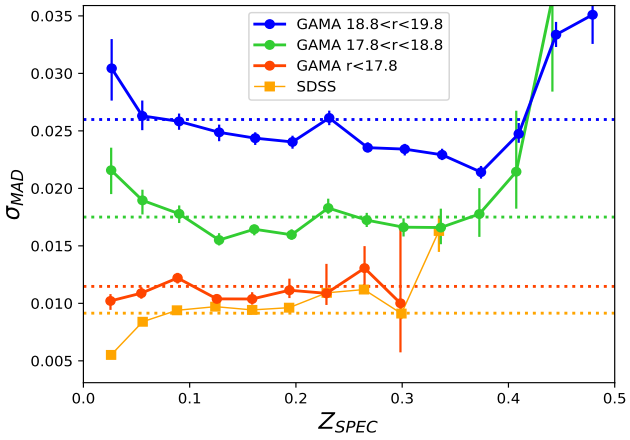

Fig. 2. Accuracy of the photometric redshift point estimates ($\sigma_{\rm MAD}$) for the SDSS-MGS and GAMA surveys as a function of spectroscopic redshift (solid lines) and the complete subsamples (dotted lines).

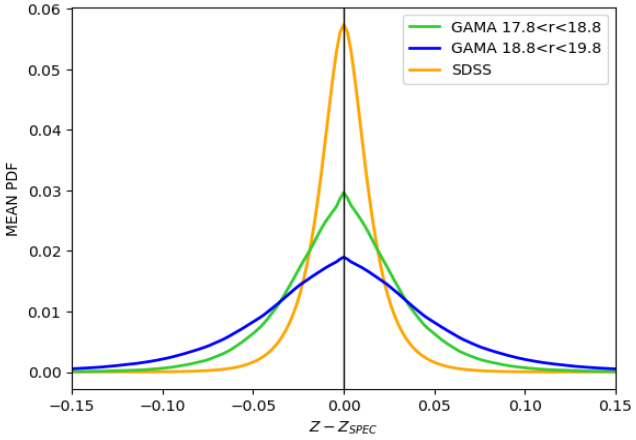

Fig. 3. Mean PDFs, recentered on the individual spectroscopic redshifts, for the SDSS and GAMA surveys. We define three subsamples according to their magnitude range: $r < 17.8$ (orange line), $17.8 < r < 18.8$ (green line), and $18.8 < r < 19.8$ (blue line).

the same as described in Sect. 2.2 (64×64 pixel images from the SDSS imaging survey in five bands *ugriz*).

As the size of this training database is smaller than the MGS sample and extends to higher redshift, we had to adapt the architecture. The new architecture is shallower and alternates five convolution layers and three pooling layers, followed by two fully connected layers. The classifier consists of 300 bins between $0 \leq z \leq 0.6$. The total number of parameters (9 433 196) is reduced compared to the CNN trained on the MGS sample, in order to avoid overfitting. A clear difficulty is the low signal-to-noise ratio of the SDSS images for the GAMA sources fainter than the MGS (up to 2 magnitudes) that degrades the performance. To tackle this problem, we optimized the choice of the activation functions and the pooling operations. We used the hyperbolic tangent as an activation function of the first layer, which saturates the signal at high values, thus narrowing its range in order to facilitate the learning stage. Then we used max pooling instead of average pooling operations in order to give more weight to the flux of the galaxy than to the noise.

Figure 2 shows the CNN redshift precision as a function of spectroscopic redshift for the GAMA training. The lower accuracy obtained for the GAMA sources at bright magnitude ($r < 17.8$) compared to the SDSS-MGS training is due to the smaller size of the training set and the simpler CNN architecture, but it is comparable. At fainter magnitudes, $\sigma_{\rm MAD}$ gradually increases as a result of the decreasing S/N in the five photometric bands, in particular in the u and z bands where the majority of galaxies with $r \geq 19.5$ have a S/N lower than 10. As in Pasquet et al. (2019), we compare our results with the photometric redshifts of Beck et al. (2017) available in SDSS DR12 and estimated with a k-NN method (local linear regression, Csabai et al. 2007). As for the SDSS-MGS, the CNN redshifts performs better, with a MAD deviation $\sigma_{\rm MAD} = 0.017$ and 0.026 in the two faint magnitude bins compared to $\sigma_{\rm MAD} = 0.022$ and 0.034 for the k-NN method.

In Fig. 3, we show the mean PDFs recentered at the spectroscopic values for the same samples. That of the SDSS-MGS sample is significantly narrower than the mean PDF of the GAMA sample, especially at faint magnitude. This broadening of the PDFs with magnitude, in line with the increase in $\sigma_{\rm MAD}$ at lower S/N, reflects the increasing uncertainty on the photometric redshifts in a reliable way since we also find the PIT distribution to be equally flat at all magnitudes.

Finally, we note that the behavior of the photometric redshift accuracy is relatively independent of redshift (Fig. 2) up to $z \sim 0.4$ for GAMA in both magnitude intervals. In the following we restrict our analysis to $z < 0.4$.

In conclusion, the CNN method provides photometric redshifts that are accurate for both the MGS and the GAMA samples and unbiased up to $z \sim 0.3$ and $z \sim 0.4$, respectively. The distance accuracy[3] is 40−112 Mpc at $\langle z \rangle = 0.10$ for $\sigma = 0.009$ (SDSS) and 0.025 (GAMA), respectively, which correspond to the typical size of the largest void. In the following section we investigate whether the combination of the PDFs from the different samples and the knowledge of the cosmic web environment reconstructed with the spectroscopic surveys can further improve the photometric redshift estimates.

3. The PW-*z* method

3.1. Method

As described in Aragon-Calvo et al. (2015), the main idea of the PW-z technique is to exploit the galaxy distribution of a spectroscopic survey in order to improve the photometric redshift of other galaxies that are expected to be embedded in this distribution. The broad PDFs derived from a given photometric redshift technique (here the CNN-based $\mathrm{PDF_{CNN}}(z)$) are combined with the probability distribution function of the density field ($P_{\rm den}(z)$, reconstructed from the spectroscopic survey) along the line of sight (LoS) as follows:

$$\mathrm{PDF_{PW\text{-}}}_z(z) = \mathrm{PDF_{CNN}}(z) \,.\, P_{\rm dens}(z) \,.\, P_{\rm CW}(z). \quad (1)$$

Figure 4 illustrates the method. The original PDF of the galaxy derived from the CNN is shown in the top panel with its mean redshift estimate and uncertainties (68% confidence interval), as well as the spectroscopic redshift. The reconstructed density field, illustrated in the second panel, shows the crossing of several structures along the LoS (alternate low- and high-density regions, spanning a wide dynamical range). To prevent the final

[3] The distance uncertainty can be expressed as $\delta D = [c(1+z)/H(z)].[\delta z/(1+z)]$ (Schmittfull & White 2016), where $H(z) = H_0 \cdot \sqrt{\left(1 + \frac{\Omega_{\rm M}}{\Omega_\Lambda}(1+z)^3\right) / \left(1 + \frac{\Omega_{\rm M}}{\Omega_\Lambda}\right)}$. This leads to $\delta D(\mathrm{Mpc}) = 4500\ \sigma_{\rm MAD}$ at $z \sim 0.1$.

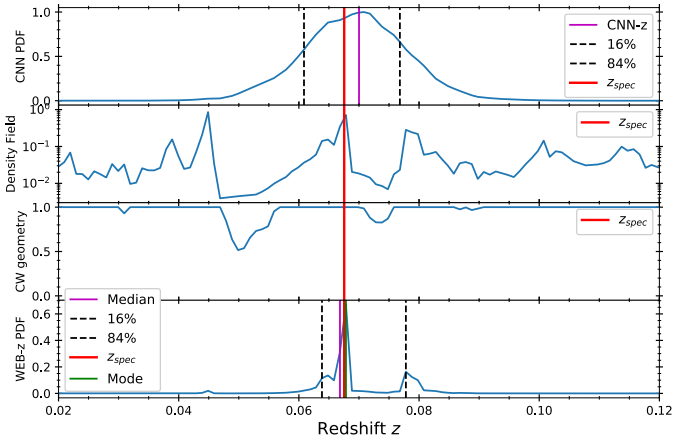

Fig. 4. Illustration of the PW-z technique. *From top to bottom*: (i) initial CNN photometric redshift PDF, (ii) CW density field, (iii) probability taking into account the closest distance of any geometric structure of the CW (Eq. (2)), (iv) final PW-z PDF. The red line indicates the $z_{\rm spec}$, the magenta line the $z_{\rm CNN}$ (*top panel*) and the median of the PW-z PDF (*bottom panel*), the green line indicates the mode of the PW-z PDF, and the dashed lines the 68% confidence interval.

PDF to be anchored on the densest group or cluster regardless of the vicinity of less dense structures (filaments or walls), Aragon-Calvo et al. (2015) introduced an additional term taking into account the geometry of the CW, beyond the density. At each redshift along the LoS, the shortest distance to any of the CW features (walls, filaments, nodes) is estimated and converted into a probability $P_{\rm CW}(z)$[4] as follows:

$$P_{\rm CW}(z) = \begin{cases} 1 & \text{if } d_{\rm ns} \leq 10 \\ \dfrac{(10 - d_{\rm ns})}{20} + 1 & \text{if } 10 < d_{\rm ns} < 30 \\ 0 & \text{if } d_{\rm ns} \geq 30 \end{cases}, \quad (2)$$

where $d_{\rm ns}$ is the 3D Euclidean distance to the nearest CW structure in cMpc. This is illustrated in the third panel. This term alleviates the dominating influence of neighboring nodes on which filaments connect. The density field and CNN PDFs are resampled with $\delta z = 10^{-3}$ in $0 < z < 0.4$ with a linear interpolation, and the $P_{\rm CW}(z)$ is also computed at these points. This results in final $\mathrm{PDF}_{\mathrm{PW}\text{-}z}$ with the same sampling. The resulting PDF ($\mathrm{PDF}_{\mathrm{PW}\text{-}z}(z)$) is shown in the bottom panel with its median, mode, and 68% confidence interval. In that specific case, the original PDF is shrunk around the highest density peak, which happens to correspond to the spectroscopic redshift of this galaxy. Other illustrations of PW-z PDFs are shown in Appendix A.

3.2. Density field and CW reconstruction

To extract the density field from the spectroscopic redshift survey and reconstruct the CW with the complex connectivity of its different components (nodes, filaments, and walls), we use the Discrete Persistent Structure Extractor code (DisPerSE; Sousbie 2011), a geometric 3D robust ridge extractor working directly with the discrete 3D data points. As demonstrated in Sousbie et al. (2011), DisPerSE can identify fairly poorly sampled structures, which will prove critical in what follows.

The underlying density field is computed from the discrete distribution of galaxies using the Delaunay Tessellation Field Estimator (DTFE) technique (Schaap & van de Weygaert 2000). The DTFE is used to generate a simplicial complex, i.e., a geometric complex of cells, faces, edges, and vertices mapping the whole volume. The value of the density field, f, is estimated at each vertex of this complex and scales with the inverse of the volume of each tetrahedron. It naturally maps the anisotropic distribution of galaxies and can be linearly interpolated at any position of the volume, and within holes (unobserved or masked regions) in the spectroscopic survey (see the example in Aragon-Calvo et al. 2015; Malavasi et al. 2017). In Eq. (1) we use the density contrast, defined as $1 + \delta = f/\varphi(z)$, where the local density, f, is normalized by the mean density, $\varphi(z)$, which decreases with radial distance. Along each LoS, $P_{\rm dens}(z)$ is normalized to unity.

To identify the topological structures of the CW (nodes, filaments, and walls), DisPerSE relies on discrete Morse and persistence theories. Morse theory provides a framework in which to extract from f the critical points where the discrete gradient, ∇f, vanishes (e.g., maxima, minima, and saddle points). It then connects critical points via the field lines tangent to ∇f in every point, while relying on a geometrical segmentation of space, known as the discrete Morse complex, within which all the field lines have the same origin and destination. This segmentation defines distinct regions called ascending and descending manifolds. The morphological components of the CW are then identified from these manifolds[5]. The finite sampling of the density field introduces noise to the detection of structural features. DisPerSE makes use of persistent homology to pair the critical points according to the birth and death of the relevant feature. The "persistence" of a feature is assessed by the relative density contrast of the density of the critical pair chosen to pass a certain signal-to-noise threshold. The noise level is defined relative to the variance of persistence values obtained from random sets of points and estimated for each type of critical pair. This thresholding eliminates critical pairs and simplifies the corresponding discrete Morse complex, retaining only its most significant features.

By construction this method is scale invariant and builds a network which adapts naturally to the uneven sampling of observed catalogues. To prevent the spurious detection near the edges of the survey, DisPerSE encloses each field into a larger volume. New particles are added by extrapolating the density field measured at the boundary of the survey (see, e.g., Sousbie 2011; Kraljic et al. 2018, for illustrations). Once the different manifolds are attached to specific topological features, we can estimate the closest structure (node, filament, or wall) from any point along a specific line of sight to derive the associated probability ($P_{\rm CW}(z)$). As in Aragon-Calvo et al. (2015), we find that this additional term is a second-order correction and only affects the results described below at a subpercent level. Finally, we do not correct for the Finger-of-God effect, in contrast to what is done in Aragon-Calvo et al. (2015). The large redshift range considered in this work ($0 < z < 0.4$) introduces a variable sampling of the density field, preventing us from performing an efficient group reconstruction at all z. The "isotropizing" of the groups also introduces an uncertainty in the redshift assignment.

[4] This is our own parameterization since it is not explicitly described in Aragon-Calvo et al. (2015). We choose this function empirically, having in mind the geometry of the reconstructed CW; we tested several versions of this function and found that the exact values do not significantly impact the results.

[5] Ascending 3-manifolds trace the voids, ascending 2-manifolds trace the walls, ascending 1-manifolds trace the filaments, with their end points connected onto the maxima (the peaks of the density field).

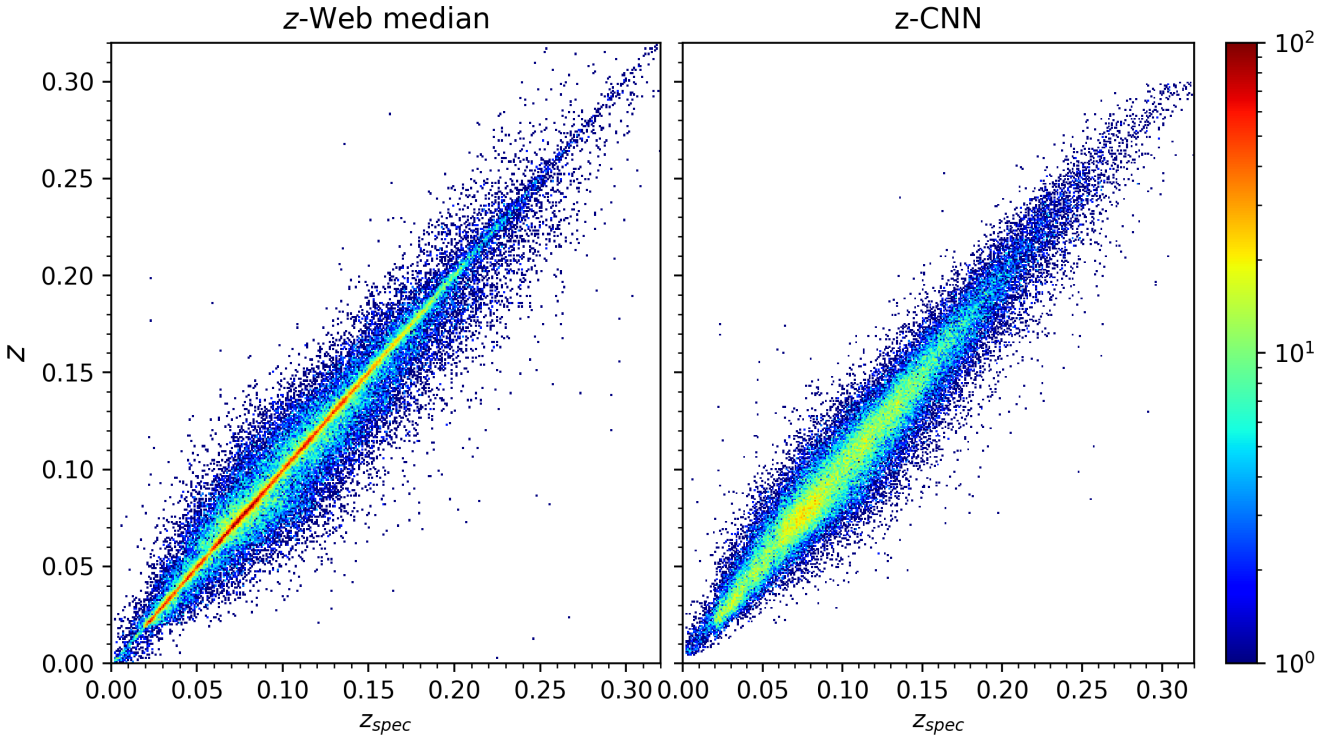

Fig. 5. Comparison between the photometric and spectroscopic redshifts. *Left*: $z_{\rm web}$ defined as the median of the PW-z PDF. *Right*: original $z_{\rm CNN}$.

This may prevent us from getting highly accurate redshift at sub-Mpc scales for some galaxies, but as discussed later it still provides a significant improvement with respect to the original photometric redshifts.

3.3. Adopted strategy

The density field is estimated from the combined SDSS (MGS and LRG) and BOSS samples up to $z \sim 0.4$ and the CW features are reconstructed with a 3σ persistence threshold.

To test the performance of the PW-z method, we select a sample of galaxies that were neither used in the CNN training nor used in the CW reconstruction. In practice, we randomly select ∼17 000 galaxies from each test sample of the cross-validations created by Pasquet et al. (2019), and reconstruct the CW with all the remaining galaxies. In this way the small fraction (∼2%) of galaxies removed has no impact on the CW reconstruction, thus on the results of the PW-z method. We repeat this operation five times for each of the five cross-validations. We find that the results are consistent throughout the five subsamples. In the next sections, the 85 000 test galaxies are used to measure the performances of the PW-z method.

4. Results

The $z_{\rm web}$ redshifts are obtained from the final PW-z PDF derived with Eq. (1). While in Pasquet et al. (2019) we adopted the mean value of the PDF as point estimate, $z_{\rm CNN}$, in the following we consider different definitions for the PW-z redshift, $z_{\rm web}$, with the mode, the mean, and the median of the PW-z PDF, and we explore their relative performance. The mode anchors the $z_{\rm web}$ to the strongest density peak and as such best illustrates the method. The mean and median rely on the full PW-z PDF while still benefiting from the narrowing of the original CNN PDF.

4.1. Global performance of the PW-z method

Figure 5 compares the $z_{\rm web}$ and $z_{\rm CNN}$ redshifts with the spectroscopic redshifts for the full sample. The $z_{\rm web}$ redshifts are significantly improved compared to the $z_{\rm CNN}$, with an increased fraction of sources along the identity line while a modest increase in catastrophic failures is observed. This is quantified in Table 1, and is illustrated in Fig. 6 for the three definitions of $z_{\rm web}$. Figure 6 (top panel) shows the boost of highly accurate $z_{\rm web}$ redshifts from a factor of 2 (for $z_{\rm web}$ mean) to a factor of 6 (for $z_{\rm web}$ mode) when considering the full test sample. The $\sigma_{\rm MAD}$ is reduced by a factor of ∼2.5, 2.0, and 1.4 for $z_{\rm web}$ based on the mode, median, and mean, respectively (Table 1). This translates into a gain in distance uncertainty from ∼40 cMpc (for $z_{\rm CNN}$) down to ∼17 cMpc (for $z_{\rm web}$ mode). About half of the sample (45%) has a redshift accuracy better than 10 cMpc (for $z_{\rm web}$ mode and median), more than twice the fraction for $z_{\rm CNN}$ (20%; see Fig. 6, bottom panel).

As a drawback of the PW-z method, about 25% of $z_{\rm web}$ based on the mode have worse estimates than $z_{\rm CNN}$ (Table 1 and Fig. 6, bottom panel). This happens when the galaxy is associated with the wrong structure of the density field and mainly impacts the $z_{\rm web}$ based on the mode of the PDF where the fraction of less accurate redshifts than $z_{\rm CNN}$ becomes significant. By adopting the $z_{\rm web}$ based on the median instead of the mode, we can mitigate this bias and reduce the number of galaxies with worse redshifts than $z_{\rm CNN}$ to ∼10%, while keeping a high fraction of galaxies with significantly improved redshifts.

The distribution of the $z_{\rm web}$ uncertainties are clearly non-Gaussian, with a high compact core and a lower and broader component, as shown in Fig. 7. Aragon-Calvo et al. (2015) proposed modeling the $z_{\rm web}$ errors with a double-Gaussian function that reflects the small-scale and large-scale errors,

$$f(\Delta z) = C_{\rm S} \exp\left[-\frac{(\Delta z)^2}{2\sigma_{\rm S}^2}\right] + C_{\rm L} \exp\left[-\frac{(\Delta z)^2}{2\sigma_{\rm L}^2}\right], \quad (3)$$

Table 1. Performance of the different $z_{\rm web}$ and $z_{\rm CNN}$ redshift estimates.

Selection	$\sigma_{\rm MAD}$ ($\sigma_{\rm S}$) $\times10^{-3}$	η (%)	$\Delta z_{\rm web}<$ 10 cMpc (%)	$\Delta z_{\rm CNN}$ (%)
$z_{\rm web}$ (mode)				
Full sample	3.8 (0.6)	0.83	48	77
Width < 0.03	2.9 (0.7)	0.13	52	79
Width < 0.02	2.1 (0.6)	0.02	58	83
$z_{\rm web}$ (median)				
Full sample	4.5 (1.2)	0.44	44	88
Width < 0.03	3.5 (1.2)	0.07	49	89
Width < 0.02	2.6 (1.1)	0.02	55	92
$z_{\rm web}$ (mean)				
Full sample	6.6 (2.2)	0.31	31	97
Width < 0.03	5.4 (1.9)	0.04	36	97
Width < 0.02	4.0 (1.5)	0.02	45	98
$z_{\rm CNN}$				
Full sample	9.2 (5.9)	0.28	21	–
Width < 0.03	7.9 (5.3)	0.04	24	–
Width < 0.02	6.3 (4.8)	0.02	29	–

Notes. Performance of the different $z_{\rm web}$ (top three blocks) and $z_{\rm CNN}$ (bottom block) redshift estimates ($\sigma_{\rm MAD}$, η) for the whole sample (85 000 galaxies) and two subsets with CNN PDFs widths ≤0.03 and 0.02 (corresponding to 75% and 40% of the whole sample, respectively). In the second column the $\sigma_{\rm S}$ value from the double-Gaussian modeled residual function is also reported (see text). The last two columns show the fraction of galaxies with residuals $\Delta z \leq 10$ cMpc and with $z_{\rm web}$ accuracy better than the $z_{\rm CNN}$.

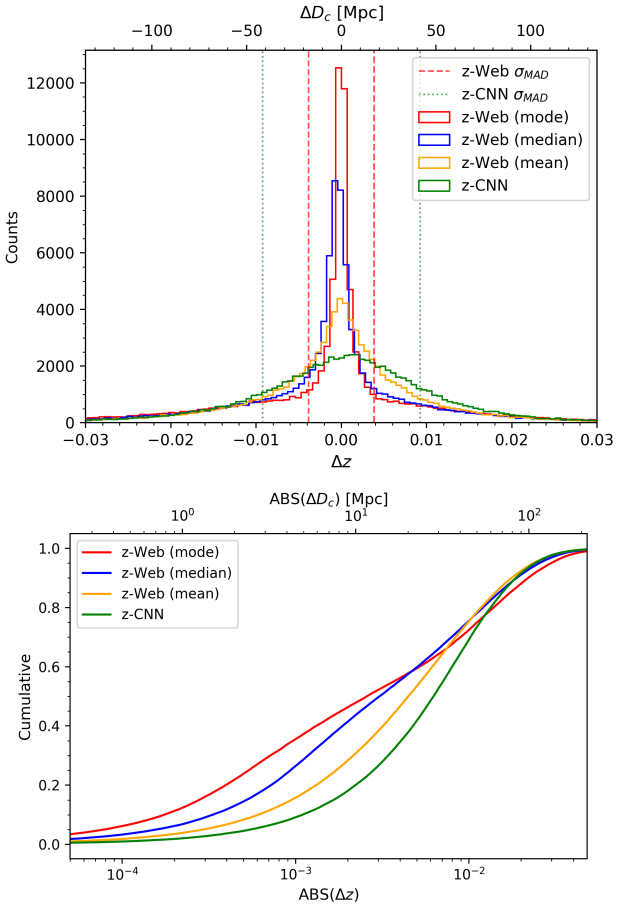

Fig. 6. Differential (*top*) and cumulative (*bottom*) histograms of the residuals for the $z_{\rm web}$ (mode: red, median: blue, mean: orange) and $z_{\rm CNN}$ (green). The dashed and dotted vertical lines indicate the respective $\sigma_{\rm MAD}$. The distance uncertainties in comoving Mpc are shown on the top axis (assuming $\langle z \rangle \sim 0.1$).

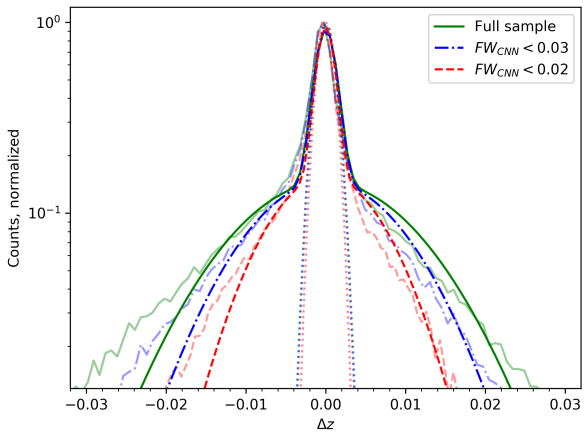

Fig. 7. $z_{\rm web}$ uncertainties (thin lines) modeled with a double-Gaussian fit (thick lines) for the full sample (green) and for galaxies with CNN PDF width <0.03 and 0.02 (blue and red, respectively). The dotted lines represent the $\sigma_{\rm S}$ of the small-scale component fit.

where $C_{\rm S}$, $C_{\rm L}$, $\sigma_{\rm S}$, and $\sigma_{\rm L}$ are the normalization coefficients and standard deviations for the small-scale and large-scale component, respectively. The result of the fit of Eq. (3) to the uncertainties distribution of $z_{\rm web}$ based on the median is presented in Fig. 7 for the full galaxy sample and for galaxies with CNN PDF width <0.03 and 0.02. The fitted values for the full sample are $C_{\rm S} = 1.0$, $\sigma_{\rm S} = 0.00122$, $C_{\rm L} = 0.1$, and $\sigma_{\rm L} = 0.01038$. The small-scale redshift error dispersion $\sigma_{\rm S}$ corresponds to a distance uncertainty of ∼5 Mpc, of the order of the CW reconstruction uncertainties and non-linear processes (e.g., peculiar velocities), while the large-scale error dispersion $\sigma_{\rm L}$ corresponds to ∼46 Mpc, similar to the $z_{\rm CNN}$ uncertainty. Selecting galaxies with smaller CNN PDF widths leads to an improvement of the large-scale $z_{\rm web}$ errors, while the small-scale component is almost unchanged. The values of $\sigma_{\rm S}$ are reported in Table 1.

Finally, the galaxy distribution obtained with the $z_{\rm web}$ (median) and the $z_{\rm CNN}$ redshifts are compared in Fig. 8. The PW-z method performs as expected: the prior information of the spectroscopic CW density field places more galaxies inside the structures, significantly enhancing the CW features, which are barely seen with the CNN redshifts. This clearly illustrates the benefit of adding spatial information from spectroscopic surveys.

4.2. Statistical behavior of the PW-z PDFs and redshift point estimates

The final PDF is significantly modified with respect to the original CNN PDF. We assess the predictive performance of the PW-z PDFs using the PIT test (see Sect. 2.2). The PIT distributions are presented in Fig. 9 in the redshift interval $0 < z < 0.3$, where the majority of our galaxies reside. The PIT distribution is shown for the CNN PDFs (green) and the PW-z PDFs (red). As do the CNN PDFs, the PW-z PDFs exhibit a nearly flat distribution indicating that they are also well-calibrated probability distributions, providing a reliable estimate of the redshift uncertainty. However, they are not exempt from a small bias since a slope is observed, which indicates a slight underestimation of the PW-z redshifts.

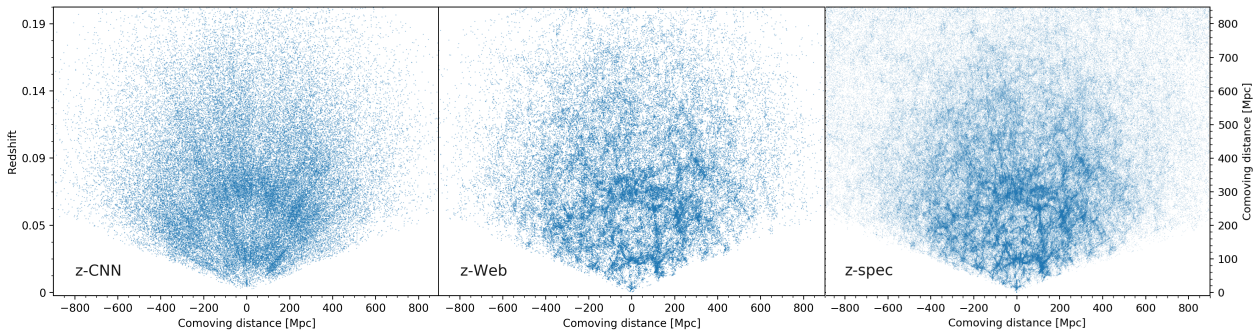

Fig. 8. Galaxy distribution based on $z_{\rm CNN}$ (*left panel*), $z_{\rm web}$ (*median*, *central panel*), and spectroscopic redshifts (*right panel*) with $z \leq 0.2$. The 2D projections include galaxies with $0° < \delta < 45°$ and $109° < \alpha < 264°$.

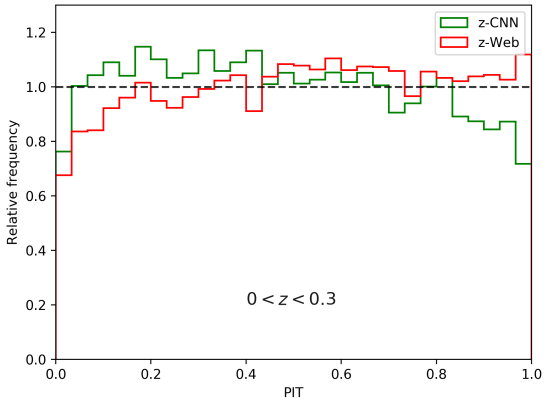

Fig. 9. Probability integral transform distribution of the CNN PDFs (green histogram) and PW-z PDFs (red histogram) in $0 < z < 0.3$.

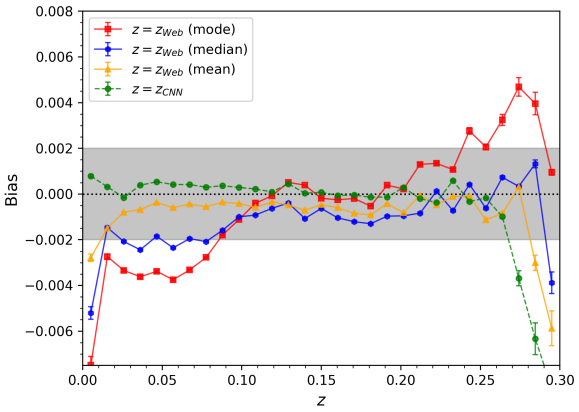

Fig. 10. Mean of the residuals (or bias, $\langle \Delta z \rangle$) as a function of photometric redshift for the $z_{\rm web}$ (mode: red, median: blue, mean: yellow) and $z_{\rm CNN}$ (green). The gray-shaded region ($\langle \Delta z \rangle < 0.002$) is the maximum bias requirement for the Euclid mission in all the photometric redshift bins.

Future cosmological missions request strong constraints on the maximum redshift bias (defined as the mean residual; see Sect. 2.2) in photometric redshift bins used for the tomographic analyses. In particular, the bias requirement for the Euclid mission is $\langle \Delta z \rangle < 0.002$ (Knox et al. 2006). Figure 10 shows the bias, $\langle \Delta z \rangle$, as a function of $z_{\rm web}$ defined as the mode, median, and mean of the PDF and $z_{\rm CNN}$, while the gray-shaded region shows the maximum bias requirement. The mean redshift estimates based on the CNN and PW-z PDFs show a very small bias at all redshifts, fully within the expected cosmological constraint. The $z_{\rm web}$ median redshift shows a small bias still within the constraint, but it appears marginally consistent at low z. However, when using the mode the bias exceeds the tolerance region below $z = 0.10$ and above $z = 0.25$. The anchor onto the main peak of the density field makes this $z_{\rm web}$ estimate less robust for cosmological use.

4.3. Impact of the initial CNN PDF width

The performance of the PW-z method depends on the quality of the initial CNN PDF. If the PDF is narrow enough so it encompasses only a few CW structures, then it increases the probability of finding the structure the galaxy belongs to. In Fig. 11 we show the $z_{\rm web}$ residual (mode) as a function of the CNN PDF width. First, the global trend is that the accuracy of $z_{\rm web}$ improves when the PDF width gets narrower, which is also the case of the $z_{\rm CNN}$ (Table 1). This reflects the reliability and unbiased behavior of the CNN PDF. Then, when the CNN PDF width is narrower than a characteristic scale, or ~80 cMpc, the fraction of greatly improved $z_{\rm web}$ ($\Delta z \leq 0.002$) increases and becomes the

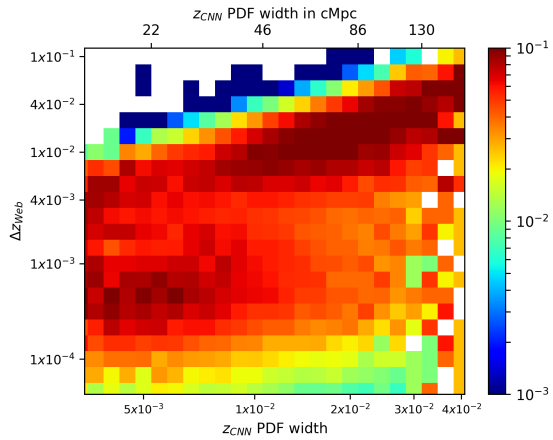

Fig. 11. Two-dimensional distributions of the $z_{\rm web}$ residuals (mode) as a function of CNN PDF width. The histograms are normalized by the area separately for each bin of PDF width. The narrower CNN PDFs properly enclose the true redshift, allowing the boost in redshift accuracy for most sources when combined with the CW information.

majority for galaxies with a width ≤0.01, at which point no further mismatches between structures are possible. The improvement induced by the PW-z method when restricting the sample to CNN PDF widths lower than $\sigma_{\rm CNN} = 0.03$ and 0.02 are reported in Table 1. All the statistical numbers improve, in particular the $\sigma_{\rm MAD}$ decreases by almost a factor of 2 and the fraction of galaxies with residuals lower than 10 cMpc is higher than 55%.

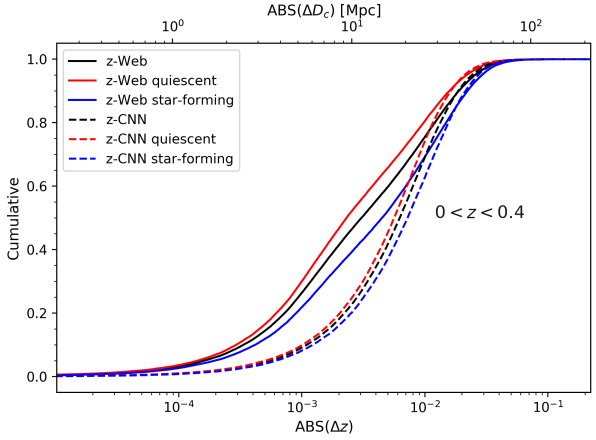

Fig. 12. Cumulative distribution of the residuals for quiescent galaxies (red), star-forming galaxies (blue), and the whole population (black) with $z_{\rm CNN}$ (dashed lines) and $z_{\rm web}$ (median, solid line).

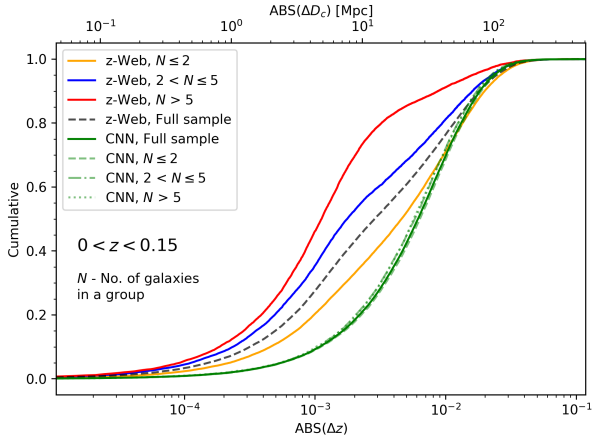

Fig. 13. Cumulative distribution of the residuals $z_{\rm web}$ (colored lines) and $z_{\rm CNN}$ (green lines) for galaxies belonging to different group sizes: one or two members (orange), three to five members (blue), and higher (red).

4.4. PW-z performance with respect to galaxy properties

In this section we examine the performance of the $z_{\rm web}$ for different categories of objects, such as star-forming versus passive galaxies, or as a function of group membership since the prior knowledge from the spectroscopic density field may impact different populations differently.

4.4.1. Galaxy type

We split the active and passive MGS galaxies according to the specific star formation rate values measured by Brinchmann et al. (2004). We consider active galaxies as those with $\log(\mathrm{sSFR}) \geq -11$ and passive otherwise. Figure 12 shows the cumulative distributions of the $z_{\rm web}$ and $z_{\rm CNN}$ residuals for the two subsamples. Passive galaxies show a better redshift accuracy than active ones with the CNN method. This could be due to the brighter magnitude distribution of passive galaxies, but also to the greater diversity of star-forming galaxies (e.g., due to clumpiness, dust lanes, inclination), making the deep learning technique slightly less efficient. After applying the PW-z method, passive galaxies have a greater boost in accuracy than active galaxies. This is a natural consequence of the biased distribution: passive galaxies are preferentially in the high-density regions of the CW compared to star-forming galaxies. This

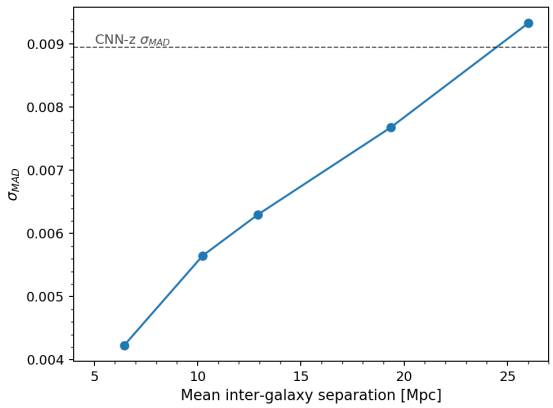

Fig. 14. Mean absolute deviation, $\sigma_{\rm MAD}$, of the $z_{\rm web}$ (based on the median) as a function of the mean intergalactic separation of several sparse spectroscopic samples. The black horizontal line shows the $\sigma_{\rm MAD}$ of the original $z_{\rm CNN}$.

segregation effect was recently quantified with respect to filaments in spectroscopic (Malavasi et al. 2017; Kraljic et al. 2018) and photometric (Laigle et al. 2018) surveys. At least at low redshifts, passive galaxies are statistically closer to filaments than active ones at similar stellar mass. The PW-z method is therefore expected to be more effective for the former population.

4.4.2. Group membership

Almost half of the galaxies in the local universe are part of gravitationally bound systems. These groups are distributed along or at the intersection of the filaments of the CW, and represent the peaks of the galaxy density field. These peaks, however, are slightly diluted along the line of sight due to the peculiar velocities of the galaxies that introduce redshift-space distortions (Fingers of God), which are not corrected for before reconstructing the local density field in the present work. Since most of the groups have a velocity dispersion lower then $\sigma_v = 600\,\mathrm{km\,s^{-1}}$ (Tempel et al. 2014), it will impact the radial distance by less than 10 cMpc, i.e., 4 times lower than the current accuracy achieved by the $z_{\rm CNN}$ redshifts. We match our sample of MGS galaxies with the group catalog of Yang et al. (2007), based on a friends-of-friends algorithm performed up to $z = 0.2$. The MGS galaxy sample is then split according to group size and the redshift residuals for the different subsamples are shown in Fig. 13. As expected, the performance of the PW-z method is highly dependent on the number of group members: the boost in accuracy is most prominent for galaxies belonging to large groups. For the largest group sample ($N > 5$), ~98% have improved $z_{\rm web}$ compared to $z_{\rm CNN}$, and 80% have an error smaller than 10 Mpc. The improvement remains significant for galaxies in groups of intermediate size ($3 < N \leq 5$), with ~90% having better $z_{\rm web}$ than $z_{\rm CNN}$, compared to ~60% for isolated or in paired galaxies. On the contrary, the $z_{\rm CNN}$ residuals are identical for all the subsamples since no prior knowledge about group membership is specified.

4.5. Impact of the spectroscopic CW reconstruction

The spectroscopic sampling of the galaxy density field is the second most important ingredient of the PW-z method after the quality of the original PDF. We evaluate its impact by randomly reducing the number of galaxies in the spectroscopic survey by several factors (25, 12.5, 3.7, 1.6%). We restrict this analysis to

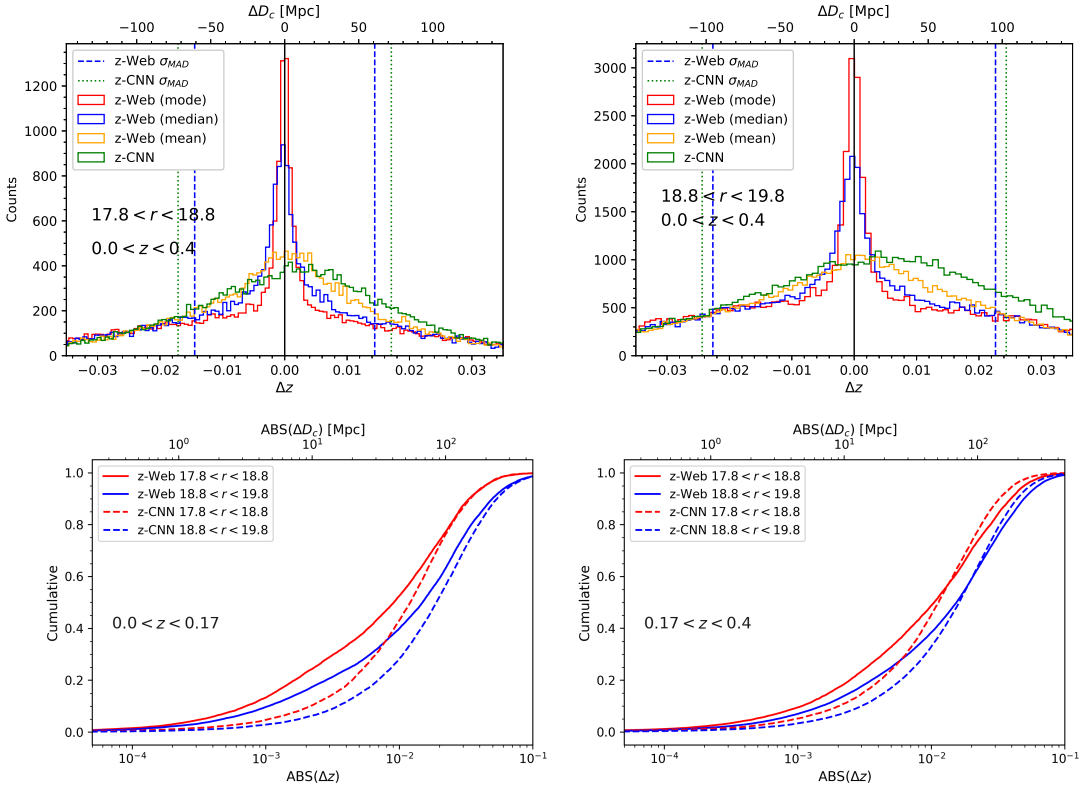

Fig. 15. Redshift residuals of the $z_{\rm CNN}$ and $z_{\rm web}$ for the GAMA survey. *Top*: differential histograms with the full sample for the $z_{\rm CNN}$ (green) and the $z_{\rm web}$ (with the mode, median, and mean of the PDF) at $17.8 \leq r \leq 18.8$ (*left*), and $18.8 \leq r \leq 19.8$ (*right*). *Bottom left*: cumulative histograms for the $z_{\rm CNN}$ (dashed line) and $z_{\rm web}$ (median only, solid line) at low redshift ($z \leq 0.17$) and split into two magnitude bins (red: $17.8 \leq r \leq 18.8$; blue: $18.8 \leq r \leq 19.8$). *Bottom right*: same as left, but at high redshift ($0.17 \leq z \leq 0.4$).

$z \leq 0.15$, where the mean intergalactic separation varies slowly with z-(Fig. 1). The PW-z method is reapplied using the density fields and CW features computed for each of the sparse spectroscopic samples. In Fig. 14, the rms of the median $z_{\rm web}$ residuals, $\sigma_{\rm MAD}$, are shown for the different subsamples, corresponding to different mean intergalactic distances. Decreasing the sampling decreases the performance of the method, but it takes a very sparse sampling to reach the rms value of the original $z_{\rm CNN}$. It can also get worse when the poor reconstruction of the galaxy density field systematically misidentifies the structures that galaxies belong to.

4.6. PW-z performance for the GAMA survey

In this section we push the PW-z method to the fainter galaxy population of the GAMA survey. The CW from the SDSS-BOSS spectroscopic survey is combined with the $z_{\rm CNN}$ photometric redshifts of GAMA. As mentioned in Sect. 2, the low S/N of the SDSS images for GAMA sources ($17.8 \leq r \leq 19.8$) leads to wider $z_{\rm CNN}$ PDFs than for the MGS sample. The impact on the performance are summarized in Fig. 15 and Table 2. For the whole sample, the $z_{\rm web}$ residuals still show a high fraction (~70%) of improved photometric redshifts well centered at $\Delta z = 0$, while the $z_{\rm CNN}$ residual appears slightly biased ($\langle \Delta z \rangle \sim 0.005$; Fig. 15, left panel). In Fig. 15, bottom left and right, we distinguish between low and high redshifts to partly disentangle the impact of the CW reconstruction from the $z_{\rm CNN}$ PDF widths. At low z (bottom left panel), where the CW is better reconstructed, the $z_{\rm web}$ are better than the $z_{\rm CNN}$ in both the bright and faint magnitude bins. It more than doubles the number of galaxies with distance uncertainties better than 10 cMpc (see Table 2). At higher redshift (bottom right panel), about half of the galaxies have improved photometric redshifts in both magnitude bins with a modest gain of highly accurate redshifts ≤10 Mpc). The degradation for the second half of $z_{\rm web}$ with respect to $z_{\rm CNN}$ can be attributed to the sparse CW reconstruction, which introduces associations with the wrong density peaks (as mentioned in Sect. 4.5).

As shown in Table 2, when restricting the sample to the narrowest $z_{\rm CNN}$ PDF widths, the $\sigma_{\rm MAD}$ for $z_{\rm web}$ still improves the accuracy but only for a small fraction of the objects. More practically, we can select a population with a desired redshift accuracy based on their final PW-z PDFs, despite their more complex shapes. In Fig 16, we show the evolution of the $z_{\rm web}$ accuracy as a function of the PW-z PDF width and the relative fraction of galaxies considered. The accuracy deteriorates progressively toward higher PDF widths. With a cut at PDF width ≤0.04 for the low-z sample ($z \leq 0.17$), we can select 70% (50%) of galaxies brighter than 18.8 (19.8), with an accuracy better than $\sigma = 0.007$ (0.008). For the whole GAMA sample, you can select a population with $\sigma = 0.01$ by applying a PDF width cut of 0.038, which will enclose 40% of the population. In conclusion, the method still benefits the photometric redshifts of galaxies two magnitudes fainter than the spectroscopic sample used to reconstruct the density field.

5. Conclusion

In this work, we revisited and extended the study of Aragon-Calvo et al. (2015), who illustrated the benefit of combining photometric redshift PDFs with the knowledge of the CW to boost their accuracy.

Table 2. Performance of PW-z (median) in GAMA survey for $17.8 < r < 18.8$ (top) and $18.8 < r < 19.8$ (bottom), low-redshift samples and different CNN PDF width selections.

GAMA number of galaxies	$\sigma_{\rm MAD}$ ($\sigma_{\rm S}$) $\times 10^{-3}$	η %	$\Delta z_{\rm web}<$ 10 cMpc %	$\Delta z_{\rm CNN}$ %
$17.8 < r < 18.8$				
Full sample 23 987	14.4/17.1 (1.5)	4.0/2.8	27/15	69
$z < 0.17$ 11 457	13.4/16.8 (1.5)	3.2/3.1	25/10	80
Width < 0.03 3513	6.5/9.5 (1.6)	0.5/0.1	35/19	81
Width < 0.02 275	4.7/6.5 (1.1)	0/0	41/26	92
$18.8 < r < 19.8$				
Full sample 65 277	22.6/24.4 (1.9)	9.8/8.2	18/9	72
$z < 0.17$ 16 872	21.1/25.4 (1.6)	9.9/12.0	18/7	97
Width < 0.03 242	7.62/10.7 (1.6)	1.6/1.6	28/18	69

Notes. Values are reported for $z_{\rm web}/z_{\rm CNN}$.

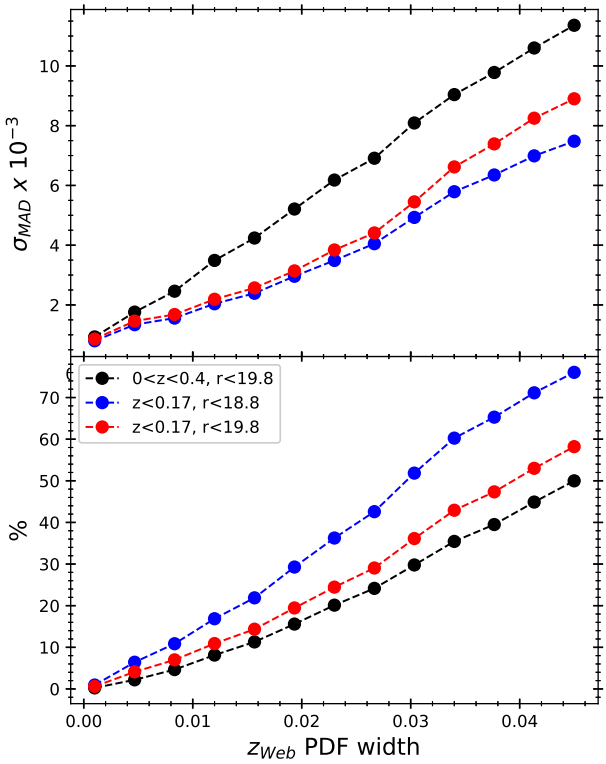

Fig. 16. WEBz performance for GAMA galaxies selected according to their PW-z PDF width. *Top panel*: $\sigma_{\rm MAD}$ for the whole population ($0 < z < 0.4$ and $r < 19.8$; black), the low-z bright ($z < 0.17$ and $r < 18.8$; blue), and the low-z faint ($z < 0.17$ and $r < 19.8$; red) subsamples. *Bottom panel*: cumulative fraction of galaxies for each sample.

Here we make use of the robustness of the cosmic web extractor DisPerSE, and the more accurate and better calibrated photometric redshifts PDFs based on a CNN. The density field and the main components of the cosmic web (nodes, filaments, and walls) are reconstructed with DisPerSE, up to $z \sim 0.4$ using the combined SDSS-MGS and BOSS surveys. The final PDF of each galaxy is obtained by combining their original CNN PDF with the density field and the distance to any CW structures along their line of sight, providing a new estimate of the photometric redshift, $z_{\rm web}$. We first apply this technique to galaxies from the MGS sample ($r \leq 17.8$). Our main conclusions are as follows:

– For the whole population, the method improves the precision of $\sigma_{\rm MAD}$ by a factor of up to 2.5. By using the mode of the final PDF as the new photometric redshift value, the initial distance uncertainty of ~40 cMpc shrinks to ~17 cMpc.
– The $z_{\rm web}$ accuracy is degraded for 10% of the sources that are associated with the wrong structure. This effect can be mitigated by using the mean of the $z_{\rm web}$ PDF rather than the mode, at the price of a lower $z_{\rm web}$ precision.
– The nearly flat PIT distribution shows that the final $z_{\rm web}$ PDFs are also well calibrated and reliable. Although a small bias is observed, it can be kept within the requirements of cosmological missions by choosing the mean or median as redshift point estimates. This allows us to select populations according to their $z_{\rm web}$ uncertainties.
– As expected, the final $z_{\rm web}$ precision depends on the original $z_{\rm CNN}$ PDFs: the narrower the CNN PDF, the lower the number of intercepted structures, the higher the boost in $z_{\rm web}$ accuracy. By selecting a $z_{\rm CNN}$ PDF width narrower than 0.02 (i.e., ~90 cMpc), $\sigma_{\rm MAD}$ is reduced by a factor of ~2 and the fraction of galaxies with residuals lower than 10 cMpc exceeds ~50%.
– The PW-z method performs better for passive galaxies, due to their higher luminosities (S/N) and stronger correlation with the densest regions of the density field compared to star-forming galaxies.
– Using an independent SDSS group catalog, we find that the distance error for 80% of the galaxies in large groups ($N > 5$) is smaller than 10 Mpc.
– Up to a mean intergalactic distance of 20 cMpc, achieved by reducing the sampling for the CW reconstruction, the PW-z method still improves the photometric redshifts.

We then extended the method to galaxies that are two magnitudes fainter than MGS ($r \leq 19.8$) using the GAMA spectroscopic survey. Despite the reduced size of the training sample and the lower S/N of the images, we were able to obtain a CNN photometric redshift precision of $\sigma_{\rm MAD} < 0.026$. We apply the PW-z method in this faint regime while keeping the same CW information as above. We found the following:

– Although the CNN PDFs are significantly wider, 65% of the PW-z redshifts are better than $z_{\rm CNN}$ and twice as many objects (i.e., ~20%) have distance uncertainties lower than 10 cMpc. However, the gain in $\sigma_{\rm MAD}$ is only ~10%. Interestingly, the PW-z method allows us to get rid of a small bias observed for the $z_{\rm CNN}$ in the faintest magnitude bins.
– While the $z_{\rm CNN}$ accuracy depends mainly on the S/N of the images rather than on redshift, a larger fraction of galaxies has improved $z_{\rm web}$ at low redshift ($z \leq 0.17$) than at high redshift ($0.17 < z < 0.4$). This reflects the importance of the resolution of the CW reconstruction.
– The PW-z PDFs can be used to select galaxies with a desired redshift accuracy (e.g., galaxies with PW-z PDF width lower than 0.038 (40%) have an accuracy of $\sigma \sim 0.01$). This will be of interest when the $z_{\rm CNN}$ and $z_{\rm web}$ are extended to $r \leq 19.8$ for the entire SDSS, but with the drawback of a poorly controlled selection function.

Combining the cosmic web with photometric redshift PDFs so as to anchor galaxies to the structures they are most likely to inhabit is a powerful method for improving the original photometric redshifts. The SDSS survey is particularly well suited for such an analysis as it combines highly accurate photometric redshifts ($\sigma \sim 0.01$) and good mapping of the cosmic web (with a resolution better that $\langle d_{\rm inter} \rangle \leq 10$ cMpc). Attaching photometric galaxies to the spectroscopically derived CW improves their photometric redshifts, even for galaxies one or two magnitudes fainter than the spectrocopic limit, as in the case of the GAMA survey. With this technique, constraining the environment of even faint galaxies is now within reach. This will enable extending galactic conformity inside groups (Treyer et al. 2018), for example, or spin alignment (Tempel et al. 2013) studies to the low-mass galaxies.

The applications to future surveys are more tentative as the efficiency of the method depends first on the accuracy of the photometric redshifts and their associated PDF widths, and second on the resolution of the CW based on spectroscopic surveys on the same field. Multiband surveys like PAU (with 30 narrowbands, Eriksen et al. 2019), J-PAS (with 50 narrowbands, Benitez et al. 2014), or low-resolution spectroscopy missions like SPHEREX (Doré et al. 2018) will deliver redshifts with uncertainties below $\sigma_z \leq 0.01$, but cosmological spectroscopic surveys like BOSS (Dawson et al. 2016) and DESI (DESI Collaboration 2016) have, or will have, a moderate resolution (with $\langle d_{\rm inter} \rangle > 15-20$ cMpc), which will hamper the use of the PW-z method.

On the other hand, CW mapping at high redshift with spatial resolution less than $\langle d_{\rm inter} \rangle \sim 10$ cMpc is now within reach with VIPERS ($z \sim 0.8$, Guzzo et al. 2014; Malavasi et al. 2017) or in the near future with PFS (CW traced by the galaxies or by the gas with the tomography technique, Tamura et al. 2018), as well as the spectroscopic survey modes of Euclid-Deep (Laureijs et al. 2011) and WFIRST (Spergel et al. 2015) up to $z \sim 2.5$. However, at such a high redshift the current limitation is the accuracy of the photometric redshifts. Current techniques barely reach $\sigma \sim 0.03$ (Moutard et al. 2016), which again will restrict the use of the PW-z method. Extending the CNN method (Pasquet et al. 2019) at higher redshift should allow us to pass this threshold. Preliminary CNN training on *ugriz* CFHTLS images yields an accuracy below $\sigma \leq 0.02$ at $i_{\rm AB} \leq 22.5$ (Treyer et al., in prep.) and $\sigma \leq 0.015$ at $i_{\rm AB} \leq 23$ on deep HSC images combined with CLAUDS (mimicking the LSST wavelength coverage and depth with *ugrizY* passbands; Sawicki et al. 2019). While very promising at intermediate redshifts ($z \leq 1.5$), it is not yet optimal at higher redshift due to the poor spectroscopic training set and further improvements are needed to fully exploit the combination of the LSST survey with the Euclid and WFIRST missions.

Acknowledgements. We thank the referee for useful comments. This work has been carried out thanks to the support of the CNES, the OCEVU Labex (ANR-11-LABX-0060), the Spin(e) project (ANR-13-BS05-0005, http://cosmicorigin.org), the DEEPDIP project (ANR-19-CE31-0023), the SAGACE project (ANR-14-CE33-0004) and the "Programme National Cosmologie et Galaxies" (PNCG).

References

Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12
Aragón-Calvo, M. A., Platen, E., van de Weygaert, R., & Szalay, A. S. 2010, ApJ, 723, 364
Aragon-Calvo, M. A., van de Weygaert, R., Jones, B. J. T., & Mobasher, B. 2015, MNRAS, 454, 463
Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, MNRAS, 310, 540
Baldry, I. K., Liske, J., Brown, M. J. I., et al. 2018, MNRAS, 474, 3875
Beck, R., Lin, C.-A., Ishida, E. E. O., et al. 2017, MNRAS, 468, 4323
Benítez, N. 2000, ApJ, 536, 571
Benitez, N., Dupke, R., Moles, M., et al. 2014, ArXiv e-prints [arXiv:1403.5237]
Bond, J. R., Kofman, L., & Pogosyan, D. 1996, Nature, 380, 603
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511
Carrasco Kind, M., & Brunner, R. J. 2014, MNRAS, 442, 3380
Cautun, M., van de Weygaert, R., Jones, B. J. T., & Frenk, C. S. 2014, MNRAS, 441, 2923
Cavuoti, S., Tortora, C., Brescia, M., et al. 2017, MNRAS, 466, 2039
Colless, M., Peterson, B. A., Jackson, C., et al. 2003, ArXiv e-prints [arXiv:astro-ph/0306581]
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
Coupon, J., Arnouts, S., van Waerbeke, L., et al. 2015, MNRAS, 449, 1352
Csabai, I., Dobos, L., Trencséni, M., et al. 2007, Astron. Nachr., 328, 852
Davidzon, I., Ilbert, O., Laigle, C., et al. 2017, A&A, 605, A70
Davidzon, I., Laigle, C., Capak, P. L., et al. 2019, MNRAS, 489, 4817
Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44
DESI Collaboration (Aghamousa, A., et al.) 2016, ArXiv e-prints [arXiv:1611.00036]
D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111
Doré, O., Werner, M. W., Ashby, M. L. N., et al. 2018, ArXiv e-prints [arXiv:1805.05489]
Eriksen, M., Alarcon, A., Gaztanaga, E., et al. 2019, MNRAS, 484, 4200
Guzzo, L., Scodeggio, M., Garilli, B., et al. 2014, A&A, 566, A108
Hemmati, S., Capak, P., Pourrahmani, M., et al. 2019, ApJ, 881, L14
Hildebrandt, H., Erben, T., Kuijken, K., et al. 2012, MNRAS, 421, 2355
Hoyle, B. 2016, Astron. Comput., 16, 34
Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55
Knox, L., Song, Y.-S., & Zhan, H. 2006, ApJ, 652, 857
Kovač, K., Lilly, S. J., Cucciati, O., et al. 2010, ApJ, 708, 505
Kraljic, K., Arnouts, S., Pichon, C., et al. 2018, MNRAS, 474, 547
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Laigle, C., Pichon, C., Arnouts, S., et al. 2018, MNRAS, 474, 5437
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
Lee, K.-G., & White, M. 2016, ApJ, 831, 181
Leistedt, B., Mortlock, D. J., & Peiris, H. V. 2016, MNRAS, 460, 4258
Madau, P., & Dickinson, M. 2014, ARA&A, 52, 415
Malavasi, N., Arnouts, S., Vibert, D., et al. 2017, MNRAS, 465, 3817
Mandelbaum, R., Seljak, U., Hirata, C. M., et al. 2008, MNRAS, 386, 781
Masters, D., Capak, P., Stern, D., et al. 2015, ApJ, 813, 53
Masters, D. C., Stern, D. K., Cohen, J. G., et al. 2019, ApJ, 877, 81
Matthews, D. J., & Newman, J. A. 2010, ApJ, 721, 456
Ménard, B., Scranton, R., Schmidt, S., et al. 2013, ArXiv e-prints [arXiv:1303.4722]
Mo, H. J., & White, S. D. M. 1996, MNRAS, 282, 347
Moutard, T., Arnouts, S., Ilbert, O., et al. 2016, A&A, 590, A103
Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26
Patej, A., & Eisenstein, D. J. 2018, MNRAS, 477, 5090
Polsterer, K. L., D'Isanto, A., & Gieseke, F. 2016, ArXiv e-prints [arXiv:1608.08016]
Salvato, M., Ilbert, O., & Hoyle, B. 2019, Nat. Astron., 3, 212
Sánchez, C., & Bernstein, G. M. 2019, MNRAS, 483, 2801
Sánchez, C., Carrasco Kind, M., Lin, H., et al. 2014, MNRAS, 445, 1482
Sawicki, M., Arnouts, S., Huang, J., et al. 2019, MNRAS, 489, 5202
Schaap, W. E., & van de Weygaert, R. 2000, A&A, 363, L29
Schmittfull, M., & White, M. 2016, MNRAS, 463, 332
Sousbie, T. 2011, MNRAS, 414, 350
Sousbie, T., Pichon, C., & Kawahara, H. 2011, MNRAS, 414, 384
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, ArXiv e-prints [arXiv:1503.03757]
Szegedy, C., Wei, L., Yangqing, J., et al. 2015, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1
Tamura, N., Takato, N., Shimono, A., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Proc. SPIE, 10702, 107021C
Tempel, E., Stoica, R. S., & Saar, E. 2013, MNRAS, 428, 1827
Tempel, E., Tamm, A., Gramann, M., et al. 2014, A&A, 566, A1
Treyer, M., Kraljic, K., Arnouts, S., et al. 2018, MNRAS, 477, 2684
Yang, X., Mo, H. J., van den Bosch, F. C., et al. 2007, ApJ, 671, 153
York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579
Zel'dovich, Y. B. 1970, A&A, 5, 84

Appendix A: Additional figure

Additional illustrations of the PW-z method for randomly selected galaxies are presented in Fig. A.1.

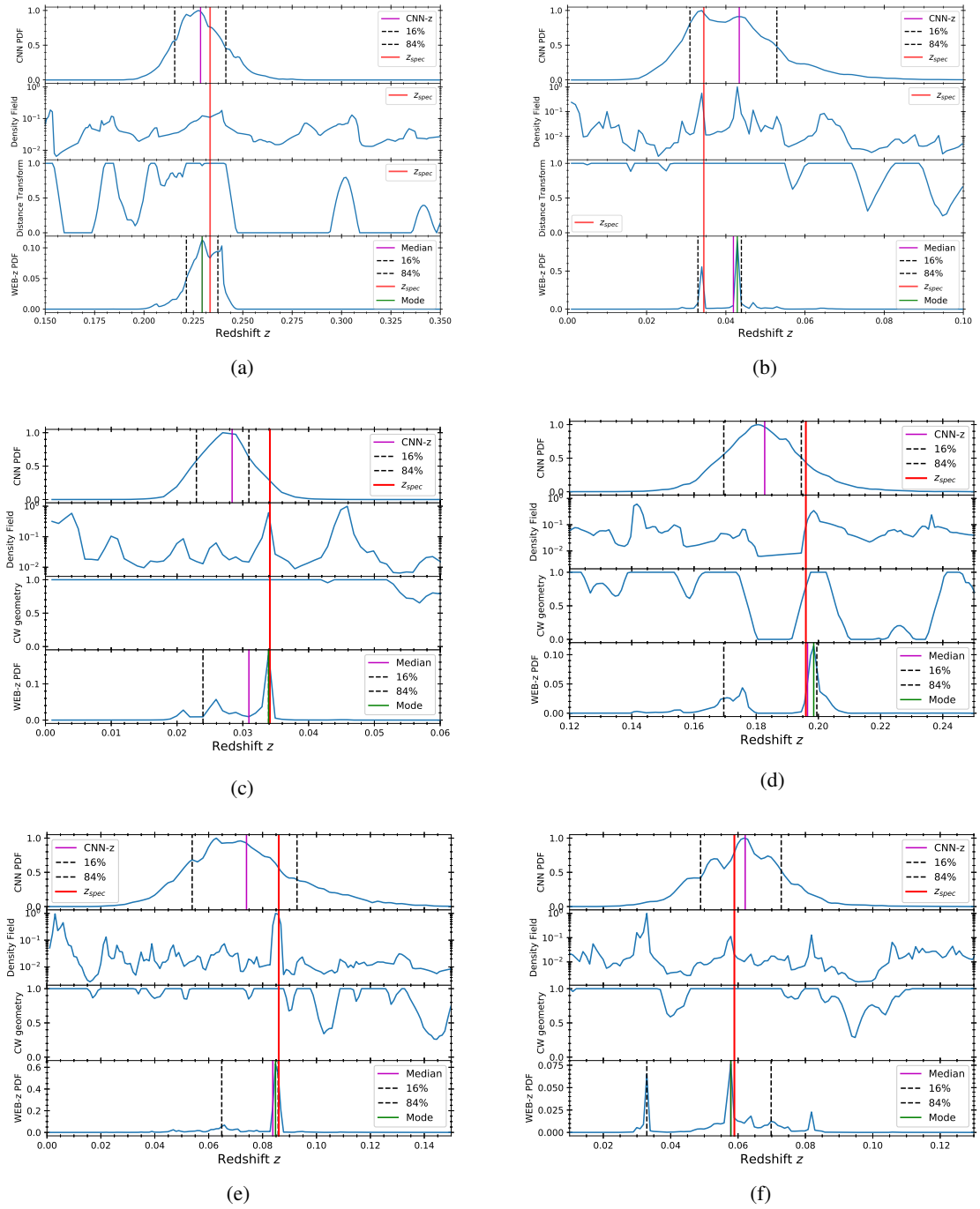

Fig. A.1. Random examples of PDFs obtained with the PhotoWeb method for four sources. The symbols are the same as Fig. 4.